BAYESIAN VARIABLE SELECTION FOR SURVIVAL ANALYSIS MODELS

By

Leriou Ilias

Supervisor: Ioannis Ntzoufras

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfillment of the requirements for
the degree of PhD in Statistics

Athens, Greece
September 2025

ΜΠΕΫΖΙΑΝΗ ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΓΙΑ ΜΟΝΤΕΛΑ ΕΠΙΒΙΩΣΗΣ

Λερίου Ηλίας

ΔΙΑΤΡΙΒΗ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Διδακτορικού Διπλώματος στη Στατιστική

Αθήνα
Σεπτέμβριος 2025

DEDICATION

Dedicated to my parents and my dearly missed grandmother.

Αφιερωμένο στους γονείς μου και στην κεκοιμημένη γιαγιά μου Όλγα, που μου λείπει πολύ.

ACKNOWLEDGEMENTS

I would like to start by thanking Jesus Christ for things that I cannot put into words. Secondly, I would like to thank my parents for their unconditional love and support through difficult times. Thirdly, I thank my supervisor Professor Ioannis Ntzoufras for his patience, guidance and support. Finally, I would like to thank the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT) for supporting this work.

The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under the HFRI PhD Fellowship grant (GA. no. 186758/$I2$/01.11.2017).

ABSTRACT

Ilias Leriou

BAYESIAN VARIABLE SELECTION FOR SURVIVAL ANALYSIS MODELS

September 2025

Selecting among different models has been gaining significant scientific concern over the last decades. This stems from the fact that due to the increased complexity of natural phenomena, the ability to understand and make sense of them is crucially important since this implies understanding the underlying mechanism that produces that phenomena and hence implies understanding and capturing the uncertainty behind them.

Models that attempt to capture the uncertainty of natural phenomena can be drawn from a wide pool of distributions that can be both simple and complex depending on the problem under study. For a wide variety of problems, the Normal distribution has been used a default option due to its lack of complexity, the thorough understanding of its behavior, and the fact that it can be approximated under the Central Limit Theorem (CLT) provided a large enough sample. Unfortunately, such distributions, can be a useful modeling tool only for a small class of phenomena leading to the need of more complex models to explain real world issues or concerns.

Among the real world issues is the modeling of time-to-event data. These recorded times can be of general interest, from modeling sport events (goal arrival times) to modeling medically related times (time until an active treatment is effective, or time until the death of a patient). These problems, by nature, entail complexity especially due to the wide variety of factors that can affect the outcome variable, the fact that the response (time-to-event) is strictly non negative and due to the existence of censoring.

In this thesis, we focus on the class of Accelerated Failure Time (AFT) survival models as the means of the likelihood of these data, while providing default priors for objective Bayesian variable selection problems that to complete the Bayesian setup of modeling such data. More specifically, we focus on modeling data using the Weibull distribution while also starting the conversation regarding how this methodology can be extended by using more general distributions.

The proposed prior has been applied in a wide variety of simulated data to fully understand its general and limiting behavior. Finally, the model selection consistency property has been demonstrated through simulated data.

ΠΕΡΙΛΗΨΗ

Ηλίας Λερίου

ΜΠΕΫΖΙΑΝΗ ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΓΙΑ ΜΟΝΤΕΛΑ ΕΠΙΒΙΩΣΗΣ

Σεπτέμβριος 2025

Η επιλογή μεταξύ διαφορετικών μοντέλων έχει αποκτήσει τις τελευταιες δεκαετίες σημαντικό επιστημονικό ενδιαφέρον. Αυτό πηγάζει από την ύπαρξη συνεχούς αυξανόμενης πολυπλοκότητας των φυσικών φαινομένων, και η ικανότητα μας να βγάλουμε νόημα από αυτά είναι εξαιρετικά σημαντική εφόσον αυτό προϋποθέτει την κατανόηση του μηχανισμού που παράγει αυτά τα φαινόμενα και επομένως την κατανόηση της αβεβαιότητας πίσω από αυτά.

Μοντέλα που προσπαθούν να συλλάβουν την αβεβαιότητα των φυσικών φαινομένων μπορούν να παρθούν από μία μεγάλη γκάμα κατανομών που μπορούν να είναι απλές και σύνθετες ανάλογα με το πρόβλημα που μελετούμε. Για τα περισσότερα προβλήματα, η κανονική κατανομή αποτελεί την προεπιλεγμένη επιλογή λόγω της μαθηματικής ευκολίας και της ευρείας κατανόησης της συμπεριφοράς της. Επιπλέον, μπορεί να προσεγγιστεί μέσω του Κεντρικού Οριακού Θεωρήματος (ΚΟΘ) δεδομένου μεγάλου δείγματος.

Ανάμεσα στα πραγματικά προβλήματα είναι η μοντελοποίηση του χρόνου μέχρι την εμφάνιση ενός γεγονότος. Αυτοί οι χρόνοι μπορεί να είναι γενικού ενδιαφέροντος, από την μοντελοποίηση γεγονότων που αφορούν τον αθλητισμό (χρόνοι μέχρι την πραγματοποίηση ενός γκολ) μέχρι την μοντελοποίηση χρόνων που σχετίζονται με την ιατρική (χρόνος μέχρι ο ενεργή θεραπεία να είναι αποτελεσματική, ή χρόνος μέχρι τον θάνατο ενός ασθενούς). Αυτά τα προβλήματα, από την φύση τους, μπορούν να είναι πολύπλοκα ειδικά λόγω του μεγάλου αριθμού των παραγόντων που μπορούν να επηρεάσουν την μεταβλητή απόκρισης και λόγω του γεγονότος ότι η μεταβλητή απόκρισης (χρόνος - μέχρι - το γεγονός) είναι αυστηρά μη αρνητικός, και από την ύπαρξη λογοκριμένων παρατηρήσεων.

Σε αυτή τη δουλειά, επικεντρωνόμαστε στην κλάση των μοντέλων επιβίωσης που λέγονται μοντέλα Επιταχυνόμενου Χρόνου Αποτυχίας (ΕΧΑ) ως προς την πιθανοφάνεια των δεδομένων, ενώ ταυτόχρονα παραθέτουμε εκ των προτέρων κατανομές χαμηλής πληροφορίας για την επιλογή μεταβλητών ολοκληρώνοντας έτσι την Μπεϋζιανή μοντελοποίηση. Πιο συγκεκριμένα, επικεντρωνόμαστε σε δεδομένα Weibull κατανομής ανοίγοντας όμως και την συζήτηση για επεκτάσεις εφαρμογής της προτεινόμενης προσέγγισης σε πιο γενικές κατανομές.

Η εκ των προτέρων κατανομές που προτείνουμε δοκιμάστηκαν σε προσομοιωμένα δεδομένα που προέρχονται απο μεγάλο αριθμό διαφορετικών σεναρίων για την πλήρη κατανόηση της γενικής και οριακής της συμπεριφοράς. Τέλος, μέσω προσομοιώσεων, η εκ των προτέρων κατανομή που προτείνουμε οδηγεί σε μια Μπεϋζιανή διαδικασία επιλογής μεταβλητών που ικανοποιεί την ιδιότητα της συνέπειας όσον αφορά την επιλογή του τελικού μοντέλου.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction to Survival Analysis

Survival analysis is the statistical sector that is mainly concerned on modeling time-to-event data T_i i.e. the time until the event i of interest (e.g. death) for individual i. The range of sciences where survival analysis is applicable includes epidemiology, engineering, sports and statistical process control and hence this allows for inference that can relate to disease progression, product failure, or even the time until a goal is scored in soccer match.

A differentiating factor that makes survival analysis unique among other statistical methodologies is the presence of *censoring*. Censoring is the phenomenon according to which the event time of one or multiple subjects in the data is not fully observed because it occurs outside our observation interval. The most common types of censoring include:

- *Right censoring*: The event occurs after the observations period. Therefore we simply know that $T_i > c_i$; where T is the time of the event and c_i is the censoring time or the end of the observation for i individual's observation time.
- *Left censoring*: The event occurs before the beginning of the observation. Hence $T_i < c_i^0$ where c_i^0 is the beginning of observation of the i individual.
- *Interval censoring*: The event occurs in between a specific range of observation times i.e. $c_i^0 < T_i < c_i$, $\forall i$.

The main endpoint of survival analysis tools is to create a framework under which the information that multiple observations can be subject to censoring, is properly utilized in the analysis. Part of those tools are specific functions that summarize critical aspects of these data. These functions include the survivor function $S(t)$, the hazard function $h(t)$ and the probability density function $f(t)$. The *survivor function* $S(t) = P(T > t)$, gives the probability that the event times occurs after a specific time. The *hazard function*

$h(t) = \frac{P(t \leq T < t+\Delta t | T \geq t)}{\Delta t}$ represents the instantaneous risk of a subject experiencing an event in a particular time t and *probability density function* $f(t)$, that represents the probability density at a specific time.

$$f(t) = h(t)exp\left(-\int_0^t h(s)ds\right)$$

which also relates to the survivor function as:

$$f(t) = h(t)S(t)$$

The Cox Proportional Hazards (PH) model is defined by the hazard function for an individual i with covariate vector $\boldsymbol{x}_i$ as:

$$h(t|x_i) = h_0(t)\exp(x_i^T\beta), \tag{1.1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients. The semi-parametric nature of this model, where the baseline hazard $h_0(t)$ is unspecified, complicates Bayesian analysis.

The Accelerated Failure Time (AFT) model offers a parametric alternative to the semi-parametric Cox model, as it proposes a log-linear relationship between the failure time T and the covariates. Unlike the Proportional Hazards model, which assumes covariates multiply the hazard function, the AFT model assumes covariates accelerate or decelerate the event time. Mathematically, the model is typically formulated on the log-time scale:

$$Y_i = \log(T_i) = \alpha + x_i^T\beta + \sigma\epsilon_i, \quad i = 1, \ldots, n, \tag{1.2}$$

where α is the intercept, σ is a scale parameter, and ϵ_i are independent and identically distributed error terms with a specific density $f_\epsilon(\cdot)$ (e.g., Normal, Extreme Value, or Logistic).

A coefficient β_j implies that a unit increase in the covariate x_j scales the expected survival time by a factor of $\exp(\beta_j)$ and is generally called the acceleration factor. Commonly used distributions for T include the Log-normal (where $\epsilon \sim N(0,1)$), Weibull (where ϵ follows an Extreme Value distribution), and Log-logistic (where ϵ follows a Logistic distribution).

1.2 Bayesian approach and priors

Bayesian inference is a statistical paradigm that treats parameters as random variables and uses Bayes' theorem to combine prior beliefs with the observed data, producing an updated posterior distribution that formally quantifies the revised beliefs (Bayes, 1763; Bernardo, 1979). The prior distribution is crucial in this process because it provides a formal mathematical mechanism to make use of various sources of information like the scientific knowledge, historical data, and expert opinions into the current analysis (Spiegelhalter et al., 2004). This ability to make use of past information using methods like power priors is a powerful tool especially when new data is limited or can be generated only once. However, the careful specification of the prior is critical since a well-calibrated prior enhances analytical efficiency, an inappropriate prior can introduce subjective bias. Depending on the inferential goal, general choices of priors include informative

priors that are designed to synthesize past evidence and objective priors constructed to introduce minimal influence so that the data dominates.

Since Bayesian analysis requires the specification of prior distributions as it is its foundational element that determines how how prior beliefs are combined with observed data to conduct posterior inference. The gradual development of variable selection methods for survival analysis models, requires an initial broader understanding of the hierarchy of prior distributions for estimation, starting from those that are based on historical data, to those that rely less on subjective prior beliefs.

In Bayesian survival analysis, the choice between the Cox PH and AFT models dictates how covariate are interpreted and how prior distributions are specified. The Cox PH model is a semi-parametric approach which assumes that covariates act multiplicatively on the underlying hazard function (Cox, 1972), requires specifying priors for the regression coefficients as well as a non-parametric prior for the unspecified baseline hazard (Kalbfleisch, 1978; Ibrahim et al., 2001). In contrast, the AFT model as a fully parametric approach assumes that covariates directly accelerate or decelerate the time to event by a constant factor and because AFT models rely on parametric distributions, they utilize the full likelihood and require proper or improper priors for the distributional scale parameters alongside the chosen priors for the regression coefficients.

1.2.1 Non-informative priors

The search for prior distributions that exert minimal influence on posterior inference often referred to as non-informative or objective priors has been a central pursuit in Bayesian statistics for centuries (Berger, 2006; Bernardo, 1979). The motivation behind objective Bayesian analysis was and is to make inference depending almost exclusively on the underlying statistical model and the available data, rather than introducing subjectivity (Berger et al., 2009). Even though a prior cannot be completely non-informative, the goal is to define priors that are dominated by the likelihood function.

Jeffreys (Jeffreys, 1946) introduced a class of priors based on the Fisher Information matrix, $I(\theta)$. The Fisher Information quantifies the expected amount of information that the data y provides about the parameter θ:

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(y|\theta)\right]. \tag{1.3}$$

The Jeffreys prior is defined as proportional to the square root of the determinant of the Fisher Information matrix Jeffreys (1946); Kass and Wasserman (1996):

$$\pi_J(\theta) \propto \sqrt{\det(I(\theta))}. \tag{1.4}$$

The definition (1.4) ensures the property of invariance because the square root of the determinant transforms exactly as required to cancel the Jacobian term in the change-of-variables formula. Common examples of Jeffreys priors include the prior for the location parameter where the location family $f(y|\theta) = f(y-\theta)$, the Fisher information is constant, yielding the flat prior $\pi_J(\theta) \propto 1$ and the prior on the scale parameter, in which, since for the scale family $f(y|\sigma) = \frac{1}{\sigma}f(\frac{y}{\sigma})$ the Fisher information is $I(\sigma) \propto \frac{1}{\sigma^2}$, yielding the prior

$\pi_J(\sigma) \propto \sqrt{\frac{1}{\sigma^2}} = \frac{1}{\sigma}$. This is equivalent to a uniform prior on $\log(\sigma)$. This prior will be used later in this thesis.

The Reference prior, which was first introduced by Bernardo (Bernardo, 1979) and further developed by his joint work with Berger (Berger and Bernardo, 1992) formalized the definition of an objective prior as one that maximizes the missing information about the parameter of interest, maximizing the divergence between the prior and the posterior (Berger et al., 2009). More specifically, it is described as maximizing the expected Kullback-Leibler (KL) divergence between the posterior $\pi(\theta|y)$ and the prior $\pi(\theta)$ as the amount of data $k \to \infty$.

The main limitation in the context of model selection is that objective priors for estimation can be improper (i.e., $\int \pi(\theta)d\theta = \infty$) (Berger and Pericchi, 2004; Taboga, 2021). While improper priors yield proper posteriors in many estimation problems, they cannot be used directly for calculating Bayes Factors between models of differing dimensions due to arbitrary constants that appear in the Bayes Factor and do not cancel out, rendering the model comparison undefined (Berger and Pericchi, 1996). This yields to the need for the development of specialized priors for selection, such as intrinsic priors or g-priors, which are discussed in later sections.

Consonni et al. (Consonni et al., 2018) provide an analytical review of prior distributions used under an objective Bayesian framework. More specifically, in their work they focus on: priors for estimation or prediction, priors for model selection, and priors for high-dimensional models with the goal of reviewing methods on objective Bayesian model comparison between normal linear models. Finally, they refer to a wide variety of model selection topics, including hierarchical models, non-parametric models and objective priors for high-dimensional spaces.

1.2.2 Informative Priors

The use of informative priors explicitly acknowledges that a scientifically analysis rarely depends on no prior information on the phenomena under testing. Informative priors represent formal mathematical incorporation of existing knowledge derived from expert opinion/knowledge, theoretical constraints, or historical data into a probability distribution $\pi(\theta)$ before the current data y are observed.

For computational tractability and mathematical convenience, informative priors are often chosen from the family of *conjugate priors*. A prior $\pi(\theta)$ is conjugate to a likelihood function $L(\theta|y)$ if the resulting posterior $\pi(\theta|y)$ belongs to the same distributional family as the prior. Among the advantages of using conjugate priors is is their interpretation. For example, consider a Gaussian model where data $y_1, \ldots, y_n \sim N(\theta, \sigma^2)$ with known variance σ^2. If we specify a Normal prior $\theta \sim N(\mu_0, \tau^2)$, the posterior distribution is:

$$\theta|y \sim N\left(\frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\tau^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right). \tag{1.5}$$

from the above formula 1.5 it is obvious that the posterior is of the same distributional family as the prior used and hence the normal prior on θ satisfies the conditions to make it a conjugate prior.

In high-dimensional estimation settings, informative shrinkage priors are often used shrinking small coefficients as zero. The Spike-and-Slab prior (George and McCulloch, 1993) models each coefficient β_j as a mixture of a point mass at zero (the spike) and a diffuse distribution (the slab). Using a latent binary indicator γ_j, the prior is:

$$\beta_j|\gamma_j \sim (1-\gamma_j)\delta_0(\beta_j) + \gamma_j N(0, c^2), \quad \gamma_j \sim Bernoulli(w), \tag{1.6}$$

where δ_0 is a Dirac mass at zero and c^2 is large.

1.2.3 Informative priors via Power priors

The power prior, introduced by Ibrahim and Chen (Ibrahim and Chen, 2000), provides a formalized mathematical framework for constructing informative priors using historical data. This approach has become a standard method for leveraging historical in a variety of statistical settings where D_0 is considered to be the historical data that is available to inform the analysis of the current data D (Ibrahim and Chen, 2000).

The basic formulation of the power prior for a parameter vector θ is defined as:

$$\pi(\theta|D_0, a_0) \propto L(\theta|D_0)^{a_0}\pi_0(\theta) \tag{1.7}$$

where $L(\theta|D_0)$ is the likelihood function of the historical data, $\pi_0(\theta)$ is an initial prior, and $0 \leq \alpha_0 \leq 1$ is a scalar (power) parameter (Ibrahim and Chen, 2000). The parameter α_0 controls the influence of the historical data. More specifically, when $\alpha_0 = 0$ the historical data are ignored, while when $\alpha_0 = 1$ they are given a full weight. Among the desirable theoretical properties that the power prior poccesses, is the fact that it minimizes a convex sum of Kullback-Leibler divergences between the posterior with no borrowing and the posterior with full borrowing (Ibrahim et al., 2003).

While α_0 can be fixed based on expert judgment or empirical measures of similarity, it can also be treated as a random variable with its own prior distribution (the normalized power prior) (Duan et al., 2006). This allows the data to determine the appropriate degree of borrowing, although it introduces computational complexity because the normalizing constant of the historical likelihood raised to the power α_0 must be computed (Ibrahim et al., 2015). A recent work (Chen et al., 2025) from include the "borrowing-by-parts" power prior, which allows different power parameters for different subsets of parameters. In the context of variable selection, the power prior is particularly attractive because it automatically specifies consistent priors for model-specific parameters across different models.

1.3 Priors for Survival Models

The presence of censoring in survival analysis, presents challenges for prior specification especially combined with the complex nature of many semi-parametric models. Unlike standard regression where each observation typically contributes a complete data point, survival data consist of time-to-event pairs (t_i, δ_i), where δ_i indicates whether the event was observed ($\delta_i = 1$) or censored ($\delta_i = 0$). That means that the information

content is heterogeneous across observations as an uncensored unit provides exact information about the failure time, whereas a censored unit only indicates that the failure time exceeds t_i. Standard priors borrowed directly from the linear regression literature often do not account for this data asymmetry, by assuming a uniform information contribution from all sample units (Castellanos et al., 2021).

1.3.1 Non-informative Priors for Cox PH models

In the frequentist setting, inference is typically based on the partial likelihood, which eliminates $h_0(t)$ (Cox, 1972). The partial likelihood is given by:

$$L_p(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(x_i^T \beta)}{\sum_{j \in R(t_i)} \exp(x_j^T \beta)} \right]^{\delta_i}, \tag{1.8}$$

where $R(t_i)$ is the risk set at time t_i (Nikooienejad et al., 2020). From a Bayesian perspective, obtaining the partial likelihood as a marginal posterior requires integrating out the baseline hazard using a non-parametric prior, such as a Gamma process or Dirichlet process, often with limiting non-informative hyperparameters (Kalbfleisch, 1978; Sinha and Dey, 1997). For instance, modeling the cumulative baseline hazard $H_0(t)$ with a Gamma process $\mathcal{GP}(c_0 H^*(t), c_0)$ having shape function $c_0 H^*(t)$ and scale parameter c_0, (Kalbfleisch, 1978) showed that the marginal posterior for β approaches the partial likelihood as the prior weight $c_0 \to 0$. However based on (Ibrahim et al., 2001) improper priors on β in the Cox model, such as $\pi(\beta) \propto 1$, can lead to improper posteriors if not carefully constructed.

Recent objective approaches include the use of Test-Based Bayes Factors (TBF) based on the partial likelihood deviance, essentially employing a generalized g-prior framework on the regression coefficients without explicitly modeling the baseline hazard (Held et al., 2016). If LR_γ denotes the partial likelihood ratio statistic for model M_γ, the Bayes factor against the null model can be approximated as:

$$BF_{\gamma 0} \approx \int_0^\infty (1+g)^{-p_\gamma/2} \exp\left(\frac{LR_\gamma}{2} \frac{g}{1+g} \right) \pi(g) dg, \tag{1.9}$$

implicitly placing a g-prior $\beta_\gamma \sim N(0, g\mathcal{I}^{-1}(\hat{\beta}_\gamma))$ scaled by the observed Fisher information $\mathcal{I}$ (Held et al., 2016). Other approaches involve Jeffreys priors adapted for survival data. In De Santis et al. (2001) it was demonstrated that standard Jeffreys priors derived for uncensored data are suboptimal when censoring is present. In their work, they derived specific Jeffreys priors for survival models that explicitly depend on the expected number of uncensored observations, $E[n_u]$, where n_u is the number of uncensored observations, rather than the total sample size n, arguing that any prior ignoring the censoring mechanism introduces bias.

1.3.2 Informative Priors for Cox PH models

In the context of the Cox Proportional Hazards model, informative priors are essential when the goal is to incorporate historical data. Unlike the objective approach, which relies on the asymptotic behavior of the partial likelihood, informative priors are specified directly on the regression coefficients β, often treating the baseline hazard as a nuisance parameter or modeling it independently.

A common approach is to elicit priors based on expected hazard ratios (HR). Since the regression coefficients in the Cox model represent log-hazard ratios, $\beta_j = \log(HR_j)$, it is natural to specify a multivariate Normal prior on β:

$$\beta \sim N_p(\mu_0, \Sigma_0). \tag{1.10}$$

Here, the hyperparameters μ_0 and Σ_0 can be defined based on how informative or not the prior is intended to be. The posterior for β, conditional on the partial likelihood $L_p(\beta)$, is then proportional to:

$$\pi(\beta|y) \propto L_p(\beta) \times \exp\left(-\frac{1}{2}(\beta - \mu_0)^T \Sigma_0^{-1}(\beta - \mu_0)\right). \tag{1.11}$$

Power Priors for Historical Data When specific historical datasets D_0 are available, the power prior, in the semi-parametric Cox framework, the power prior for β is constructed in the usual way but by raising the partial likelihood instead of the regular likelihood of the historical data, $L_p(\beta|D_0)$, to a power $a_0 \in$:

$$\pi(\beta|D_0, a_0) \propto [L_p(\beta|D_0)]^{a_0} \pi_0(\beta), \tag{1.12}$$

where $\pi_0(\beta)$ is an initial prior (often vague) and a_0 is the power parameter (**?**). If the baseline hazard is also of interest, the power prior can be defined on the full likelihood $L(\beta, h_0|D_0)$, typically requiring a Gamma process prior for the cumulative baseline hazard $H_0(t)$ (Ibrahim et al., 2001).

1.3.3 Non-informative Priors for AFT models

In the context of Bayesian estimation for AFT models, the standard objective prior is often the reference prior for location-scale families. For the regression coefficients β, the intercept α, and the scale σ, the widely accepted non-informative prior is the independent Jeffrey's prior:

$$\pi^N(\alpha, \beta, \sigma) \propto \frac{1}{\sigma}. \tag{1.13}$$

This prior is invariant to location and scale transformations and is commonly used as a baseline for Bayesian inference in log-linear survival models. The marginal likelihood $m(y)$ becomes defined only up to an arbitrary constant, necessitating objective model selection strategies such as Intrinsic Bayes Factors (IBF) or the use of specific proper objective priors like g-priors.

1.3.4 Informative Priors for AFT models

When historical data or expert knowledge is available, informative priors can be employed to improve estimation precision and handle identifiability issues in complex AFT models. The power prior framework extends naturally to AFT models due to their log-linear structure. Let $D_0 = (y_0, X_0, \delta_0)$ denote the historical data, where $y_0 = \log(t_0)$. The power prior for the parameters $\theta = (\alpha, \beta, \sigma)$ is defined as:

$$\pi(\alpha, \beta, \sigma|D_0, a_0) \propto [L(\alpha, \beta, \sigma|D_0)]^{a_0} \pi_0(\alpha, \beta, \sigma), \tag{1.14}$$

where $L(\cdot|D_0)$ is the likelihood of the historical AFT model and a_0 is the power parameter. For a Log-normal AFT model, if π_0 is the standard non-informative prior $1/\sigma$, the conditional power prior for β given

σ becomes a multivariate normal centered at the historical estimates, with precision scaled by a_0. Specifically, if we focus on the regression coefficients, the power prior creates the following regularization:

$$\beta|\sigma, D_0, a_0 \sim N\left(\hat{\beta}_0, \frac{\sigma^2}{a_0}(X_0^T X_0)^{-1}\right), \tag{1.15}$$

essentially borrowing the correlation structure and effect sizes from the historical data while inflating the variance by $1/a_0$ to account for heterogeneity (Chen et al., 2025).

1.4 The Basics of Model and Variable Selection

In Bayesian variable selection, the uncertainty about which variables belong in the model is treated explicitly by assigning prior probabilities to the candidate models. The Bayes Factor is the primary tool for comparing two models, M_i and M_j. It is defined as the ratio of their marginal likelihoods:

$$B_{ij} = \frac{m_i(y)}{m_j(y)} = \frac{\int f(y|\theta_i, M_i)\pi(\theta_i|M_i)d\theta_i}{\int f(y|\theta_j, M_j)\pi(\theta_j|M_j)d\theta_j} \tag{1.16}$$

where $m_i(y)$ is the marginal likelihood of the data under model M_i. The Bayes Factor quantifies the evidence in the data favoring one model over another, irrespective of the prior model probabilities (Kass and Raftery, 1995). In high-dimensional settings where the model space is vast, selecting a single best model may be unstable. Instead, researchers often focus on the marginal importance of each variable. The Posterior Inclusion Probability for a variable x_k is defined as the sum of the posterior probabilities of all models that include x_k:

$$PIP(x_k) = \sum_{M_\gamma:\gamma_k=1} P(M_\gamma|y) \tag{1.17}$$

Variables with PIPs exceeding 0.5 are often comprised in the Median Probability Model.

The Lindley's Paradox is described as a situation where frequentist and Bayesian methods yield contradictory conclusions when testing a point null hypothesis ($H_0 : \theta = 0$) against an alternative ($H_1 : \theta \neq 0$) with a very large sample size (Lindley, 1957). This occurs because the likelihood under the alternative model becomes spread out over the prior space as n increases, penalizing the alternative model for its complexity/-vagueness compared to the sharp null. This highlights the sensitivity of Bayes Factors to the prior variance of the parameters under the alternative hypothesis, necessitating careful calibration of objective priors in variable selection.

1.5 Objective Approaches for Variable Selection

In the Gaussian linear model framework, the response vector Y of length n is modeled as $Y = X_\gamma\beta_\gamma + \epsilon$, where $\epsilon \sim N_n(0, \sigma^2, I_n)$, X_γ is the $n \times p_\gamma$ design matrix for model M_γ and β_γ is the vector of the coefficients. The challenge lies in defining $\pi(\beta_\gamma, \sigma^2|M_\gamma)$.

1.5.1 Zellner's g-Prior and Mixtures

The most prevalent objective approach is Zellner's g-prior (Zellner, 1986), which utilizes the correlation structure of the design matrix to define the prior covariance. The prior is specified as:

$$\beta_\gamma|\sigma^2, g, M_\gamma \sim N_{p_\gamma}(0, g\sigma^2(X'_\gamma X_\gamma)^{-1}), \quad \pi(\sigma^2) \propto 1/\sigma^2. \tag{1.18}$$

Here, the hyperparameter g acts as a penalty that controls the shrinkage of the coefficients. This specification leads to a closed-form marginal likelihood for model M_γ:

$$m(Y|M_\gamma, g) \propto (1+g)^{-p_\gamma/2}\left[1 - \frac{g}{1+g}R^2_\gamma\right]^{-(n-1)/2}, \tag{1.19}$$

where R^2_γ is the coefficient of determination. While simple, fixing g can lead to the information paradox (where support for the true model does not go to infinity as $R^2 \to \infty$) and Lindley's paradox (Liang et al., 2008). To resolve these issues, mixtures of g-priors have been proposed, where a prior $\pi(g)$ is placed on g. A prominent example is the Zellner-Siow prior (Zellner and Siow, 1980), which implies a Cauchy prior on β_γ by placing an inverse-gamma prior on g. More recently, Liang et al. (Liang et al., 2008) introduced the hyper-g prior family:

$$\pi(g) = \frac{a-2}{2}(1+g)^{-a/2}, \quad g > 0, \tag{1.20}$$

which retains computational tractability via Gaussian hypergeometric functions while avoiding the information paradox.

1.5.2 Generalizing g-Priors to GLMs

Extending objective variable selection to Generalized Linear Models (GLMs) presents challenges bacause unlike the normal case, the marginal likelihood is generally not available in closed form. In a GLM, the density of Y_i belongs to the exponential family:

$$f(y_i|\theta_i, \phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}, \tag{1.21}$$

where the canonical parameter θ_i is linked to the linear predictor $\eta_i = x_i^T\beta_\gamma$ via a link function $g(\mu_i) = \eta_i$.

A direct extension of the g-prior to GLMs is difficult because the information matrix depends on the parameters. Several approaches have been proposed. Sabanés Bové and Held (Sabanés Bové and Held, 2011) utilize an asymptotic approximation of the Fisher information matrix to define a g-prior for GLMs:

$$\beta_\gamma|g, \phi, M_\gamma \sim N_{p_\gamma}(0, g\phi(X_\gamma^T W X_\gamma)^{-1}), \tag{1.22}$$

where W is the weight matrix evaluated at the maximum likelihood estimate (MLE) or the prior mode. To handle the unknown g, they employ an integrated Laplace approximation. This involves approximating the conditional marginal likelihood $m(Y|M_\gamma, g)$ via Laplace expansion and then numerically integrating over the prior $\pi(g)$. This method, termed the "Compound Hypergeometric Information Criterion" (CHIC) when specific priors are used as presented in Li and Clyde (2018), provides a unified framework consistent with linear model results.

1.5.3 Test-Based Bayes Factors (TBF)

Another objective strategy for GLMs eliminates explicit prior specification by using test statistics. Johnson (Johnson, 2005, 2008) proposed Test-Based Bayes Factors (TBF), which replace the data with a summary statistic, such as the deviance difference or the Wald statistic. If D_γ is the deviance of model M_γ and D_0 is the deviance of the null model, the Bayes Factor can be approximated as a function of the likelihood ratio statistic $R_\gamma = D_0 - D_\gamma$. Using the geometry of the likelihood, standard TBFs effectively assume a prior on the non-centrality parameter of the test statistic distribution under the alternative hypothesis. Held et al. **?** showed that TBFs in GLMs correspond implicitly to using a g-prior on the regression coefficients. For a model with p_γ covariates, the TBF against the null model takes the form:

$$BF_{\gamma 0} \approx \int_0^\infty (1+g)^{-p_\gamma/2} \exp\left(\frac{R_\gamma}{2}\frac{g}{1+g}\right)\pi(g)dg. \tag{1.23}$$

This approach allows for computationally efficient model comparison in logistic and Poisson regression without complex MCMC schemes, relying instead on standard GLM output.

1.5.4 Power-Expected-Posterior (PEP) Priors

Fouskakis et al. (Fouskakis et al., 2018) developed the Power-Expected-Posterior (PEP) prior to generalize the Intrinsic Bayes Factor methodology to GLMs while avoiding any instability issues of the minimal training samples. The PEP prior is constructed using an imaginary training sample y^* as:

$$\pi^{PEP}(\beta_\gamma|M_\gamma) = \int \pi^N(\beta_\gamma|y^*,\delta)m(y^*)dy \tag{1.24}$$

Here, the imaginary data y^* is integrated out with respect to the prior predictive distribution of a reference model (usually the null model). By setting $\delta = n$, the prior information content is equivalent to a single unit of information, leading to an objective prior that is consistent and parsimonious. In Fouskakis et al. (2018) they implement this method for logistic and Poisson regression.

1.6 Bayesian variable selection methods for survival models

1.6.1 Censoring-Adjusted g-Priors

Applying standard objective priors to survival data is challenging due to the heterogeneity of information content because of censoring. The standard g-priors, defined as $\beta|\sigma,g \sim N(0, g\sigma^2(X^TX)^{-1})$, are a sub-optimal option for modeling the coefficients due to implicitly assuming that all n observations contribute equally to the precision of the estimates. Castellanos et al. (Castellanos et al., 2021) demonstrated that this assumption leads to bias in survival settings, as censored observations carry significantly less information than uncensored ones.

To address this, Castellanos et al. Castellanos et al. (2021) proposed a generalization of the g-prior for log-Normal AFT models that explicitly incorporates censoring information into the prior covariance

structure. The proposed prior takes the form:

$$\pi(\beta_\gamma|\alpha,\sigma,g) = N_{p_\gamma}\left(0, g\Sigma_M\right), \tag{1.25}$$

where the covariance matrix Σ_M is constructed to reflect the effective information content of the sample. Unlike the standard $(X^TX)^{-1}$ which treats all rows of the design matrix symmetrically, Σ_M weights the contributions of censored individuals based on the censoring times c. Specifically, for the Log-normal AFT model, the matrix Σ_M is derived using predictive matching arguments.

1.6.2 Regularization

In variable selection contexts, particularly with high-dimensional genomic data ($p \gg n$), informative priors are used to enforce sparsity (i.e., the belief that most β_j are zero). This corresponds to penalized partial likelihood estimation. For example, the Bayesian Lasso implies independent Laplace priors on the coefficients:

$$\pi(\beta|\lambda) = \prod_{j=1}^{p} \frac{\lambda}{2} \exp(-\lambda|\beta_j|), \tag{1.26}$$

which puts high density at zero but has heavier tails than the Normal distribution (Antoniadis et al., 2010). However, standard sparsity priors like the Laplace or Cauchy peak at zero, which can lead to slow contraction rates for true non-zero coefficients. To address this, Johnson and Rossell (Johnson and Rossell, 2010) designed non-local priors that assign exactly zero probability density at the null value, making this functional form keep the posterior away from zero for true effects, without excessively penalizing large coefficients. Finally, Nikooienejad et al. (Nikooienejad et al., 2020) proposed using Non-Local Priors (NLP) for the Cox model, such as the product inverse-moment (piMOM) prior which is mathematically formulated as:

$$\pi(\beta_j|\tau,r) \propto (\beta_j^2)^{-r} \exp\left(-\frac{\tau}{\beta_j^2}\right). \tag{1.27}$$

Unlike local priors, the piMOM density is exactly zero at the null value ($\beta_j = 0$). This creates a penalty that keeps the posterior away from zero for small coefficients, effectively separating noise from signal and providing strong model selection consistency even in the presence of censoring.

1.6.3 Intrinsic Priors

Alternative objective approaches include the use of Intrinsic Priors. For the Exponential AFT model, the intrinsic prior $\pi^I(\theta)$ can be derived analytically using minimal training samples (MTS). However, defining an MTS in survival analysis is non-trivial since a set of observations is minimal only if it contains at least one uncensored event to ensure posterior propriety (Berger and Pericchi, 2004). Based on (De Santis et al., 2001), the resulting intrinsic priors often take the form of specific Inverse Gamma or Beta-Prime distributions depending on the parameterization, providing a proper objective baseline for computing Bayes Factors without subjective input.

1.7 Discussion

In this chapter we started by highlighting the distinction between priors suitable for parameter estimation and those required for model selection. For estimation, the literature provides robust frameworks, ranging from reference priors (Bernardo, 1979; Berger et al., 2009) to informative power priors that uses historical data (Ibrahim and Chen, 2000; Chen et al., 2025). The power prior, in particular, offers a theoretically sound alternative that minimizing Kullback-Leibler divergence to incorporate external sources of information, such as experts' opinion (Ibrahim et al., 2015). However, as highlighted in the discussion of Lindley's Paradox (Lindley, 1957; Sprenger, 2013), priors that are inappropriate for estimation (like the vague priors) can be suboptimal for model selection. As the sample size $n \to \infty$, fixed vague priors force the Bayes Factor to favor the null hypothesis excessively (Bartlett, 1957). This leads to the need for carefully using specific priors, such as Zellner's g-prior (Zellner, 1986) and its mixtures (Liang et al., 2008), which adapt the prior scale to the information content of the design matrix.

The central methodological problem that is addressed in this work is the adaptation of these objective selection priors to survival data. While methods like the Intrinsic Bayes Factors (IBF) Berger and Pericchi (1996) provide a sound mechanism for converting improper priors into proper ones using training samples or likelihood fractions, they do not account for the information loss due to censoring. The work of Castellanos et al. Castellanos et al. (2021) aimed to provide an answer to this problem by demonstrating that simply borrowing g-prior structures from linear models is insufficient because it implicitly assumes that all observations contribute uniformly to the likelihood. In survival analysis, censored observations contribute significantly less information than uncensored ones Castellanos et al. (2021). This insight aligns with the broader literature on effective sample size (ESS), where the weight of a censored sample is a fraction of a fully observed one (Berger et al., 2014). The construction of the prior covariance matrix Σ_M by Castellanos et al., which weights censored observations, effectively generalizes the logic of Fouskakis and Ntzoufras's Power-Expected-Posterior (PEP) priors (Fouskakis et al., 2018) from the GLM context of survival models.

The Test-Based Bayes Factor (TBF) approach (Johnson, 2005; Held et al., 2016) offers an alternative for Cox Proportional Hazards models by utilizing the asymptotic distribution of the deviance statistic. This allows for objective BVS without explicitly modeling the baseline hazard. On the other hand, Non-Local Priors (NLPs), such as the product inverse-moment prior, address variable selection by separating between the null and alternative parameter spaces (Nikooienejad et al., 2020).

While semi-parametric Cox models are widely used, they rely on the proportional hazards assumption. Parametric Accelerated Failure Time (AFT) models, such as the Weibull and Log-normal, offer direct interpretation of time-to-event ratios and are often superior for prediction. However, the choice between these parametric families introduces model uncertainty and therefore the primary aim of our work is the extension of the objective BVS framework to the Generalized F (GF) distribution (Cox, 2008). The GF nests the Weibull, Log-normal, and Generalized Gamma distributions, providing a unified framework. Extending the censoring-adjusted g-prior Σ_M to the GF model would allow for simultaneous selection of covariates and the structural form of the baseline hazard. Furthermore, validating this extension requires moving beyond

Predictive Matching to testing for Model Selection Consistency, ensuring that the posterior probability of the true model converges to 1 asymptotically even under the complex, non-Gaussian likelihoods like the Generalized F distribution.

1.7.1 Aim and scope of this Thesis

The work of Castellanos et al. (2021) provided a simple method to apply Bayesian Variable Selection under censoring by showing that borrowing priors from the 'uncensored' literature may lead to unsatisfactory results because they inherently fail to account for the heterogeneous information content introduced by censoring Castellanos et al. (2021); Berger and Pericchi (2004). Their initial attempt to solve the problem led to the development a generalized g-prior structure (Σ_M) specifically for the Accelerated Failure Time (AFT) model with lognormal errors **?**. As an initial step in understanding how AFT models of interest behave under a practical survival framework, our first goal is to conduct a case study analysis using a variate of AFT models for general Bayesian survival inference in Chapter 2. Moving on, our goal for the primary development in this work (presented in Chapter 3) is to confirm the robustness of this objective BVS approach by applying it to the Weibull distribution Collett (2015); Klein and Moeschberger (2003); Lawless (2003). The Weibull model is analytically essential as it is the only parametric model that is both an AFT model and a Proportional Hazards model (Cox et al., 2007), hence being a reference model in parametric survival studies.

To provide robust theoretical validation for the methodology when applied to the Weibull AFT model, we showed through simulations that the Model Selection Consistency property Casella et al. (2009), (Held et al., 2016; Berger and Pericchi, 2001), (Bayarri et al., 2012) is valid under our proposed methodology. Model Selection Consistency is a paramount theoretical requirement for objective Bayesian procedures, guaranteeing that the posterior probability concentrates of the true model tends to 1 as the sample size increases (Casella et al., 2009; Held et al., 2016; Johnson and Rossell, 2012). Proving consistency for the Weibull likelihood, represents a difficult analytical problem compared to the lognormal base case. This core theoretical work on rigorously applying the Σ_M prior construction to the Weibull model and demonstrating its asymptotic consistency constitutes the main content of Chapter 3 (Castellanos et al., 2021).

Proceeding to Chapter 4, we extend the fundamental derivation to the Generalized Gamma Distribution (GGD) (Stacy, 1962), (Lawless, 2003), a flexible three-parameter model that mathematically includes the Weibull and lognormal distributions as special cases (Lawless, 2003). Within this chapter, the major theoretical focus shifts to providing the necessary derivations for prior scaling by executing all relevant calculations for the effective sample size. Finally, in Chapter 4 we also introduce the ultimate goal of extending the methodology to the broadest possible class of AFT models, the Generalized F (GF) Distribution (Prentice, 1975; Cox et al., 2007). The GF distribution encompasses the GGD, but its four-parameter structure and non-Gaussian nature significantly complicate computation (Prentice, 1975; Johnson et al., 1983). Therefore, the treatment of the GF model in Chapter 4 is framed as a conversation starter, outlining the analytical challenges and necessary adaptations for future research in applying the robust objective BVS methodology to such complex models. This thesis ends with Chapter 5 with a final discussion on results and possible work

for the future.

Chapter 2

Application to Football

2.1 Introduction

Association Football (Soccer) has gathered increasing interest from researchers due to the great uncertainty involved in the final outcome of each game. Most of the football analytics articles focus of on the modeling/prediction of either the final outcome in the form of win/draw/loss using multinomial logistic regression models (Hvattum, 2017) and other competing methods for the final score of the game using Poisson-based models and their extensions; see for example in Maher (1982), Dixon and Coles (1997), Karlis and Ntzoufras (2003, 2009). Here we are taking a different direction. We focus on modeling the goal arrival times within a football game.

Modeling the number of goals in a soccer game has been thoroughly studied over the last years. Maher (1982) was among the first to present a well-fitted and appealing Poisson regression model for such outcomes. Extensions of the basic double Poisson model have been proposed by Dixon and Coles (1997) and Karlis and Ntzoufras (2003) in order to account for the excess of zeros and for possible correlations between the goals scored by the two opponents. Additional models have been proposed to account for other characteristics of the model such as over-dispersion, a phenomenon which was first brought in the foreground in soccer by Karlis and Ntzoufras (2000) and then addressed by Karlis and Ntzoufras (2003). Another important issue in the use of dynamic time-dependent parameters used to capture the performance of the teams (Rue and Salvesen, 2000; Owen, 2011; Koopman and Lit, 2015; Egidi and Gabry, 2018). Additional research includes the article of Karlis and Ntzoufras (2009) in which a Bayesian version of the Skellam distribution is proposed as a model for the goal differences between the home and away team while Štrumbelj and Šikonja (2010) introduces the notion of using the odds derived from the bookmakers as a forecasting tool by analyzing their effectiveness in 10699 soccer matches. Additionally, the implementation of information drawn from historical data and bookmakers' odds has also been thoroughly researched by Egidi et al. (2018). Finally, for a comprehensive and detailed review of statistical and machine learning methods, see in Tsokos et al. (2019) and references therein. With regards to the modeling of goal arrival times that we deal with in this article, a very limited amount of research work has appeared in the relevant bibliography. One of the first

related research papers is the one by Dixon and Robinson (1998) who have taken the goal arrival times to be coming from two Poisson birth processes when analyzing 4000 games from English competitions. They have concluded that the rate of the processes increases during the game and that it is influenced by the current score. Analysis of inter-arrival scoring times in Ice Hockey was conducted by Thomas (2007) using Weibull and Plateau-Hazard distributions. Finally, a direct approach to survival modeling in football was presented in Nevo and Ritov (2013) where the effect of the first and second goal was assessed using the Cox proportional hazards model in 760 Premier League games (2 seasons 2008-2010). This approach, although promising, is limited to the modeling of the first two goals independently of the scoring team, and by further considering the assumption of the proportionality of hazards.

Survival analysis models and the commonly applied Poisson models for predicting the final score capture different aspects of the data. Poisson models (and their extensions) are designed to model count data, meaning that in football, they answer the question: "*How many goals will a team score?*". On the other hand, survival analysis models focus on modeling the time until an event (goal) occurs, rather than predicting the final outcome (Kleinbaum and Klein, 2012). The final score is estimated as a byproduct of these models. Thus, survival models address the question: *"When will a goal be scored?*". This can be particularly informative for in-play soccer (live) betting or for understanding the dynamics of scoring frequency across opposing teams in a league. This discussion highlights that survival models require and model "richer" data as response inputs and, consequently, they answer a wider range of football-related questions compared to score-based models. Another argument for using survival models instead of Poisson-based models is that survival analysis allows for the estimation of hazard functions, which are crucial for understanding the actual risk of each event occurring (Cox, 1972). Finally, most survival models (with the well-known exception of the exponential model) assume non-constant hazards, which implies that the rate of scoring varies over time. This is not the case for Poisson models, which are based on the less realistic assumption of constant hazards, as they imply the exponential distribution between goal arrival times (Andersen and Gill, 1982).

In this work, we propose a general approach for modeling all goal inter-arrival times within a football match. We assess several assumptions such as the shape of the assumed distribution and potential dependence structures before we decide on the final proposed model. For this model, we assess the fit and the predictive performance of the selected model and it is compared with standard models such as the standard vanilla double Poisson model with respect to the prediction of the final score (although the primary aim of such model is to focus on the prediction of goal arrival times). Our survival analysis model models the inter-arrival times between subsequent goals. It is based on the following important assumptions: (a) when a team scores, then the scoring times of both teams restart to zero, (b) an event (goal) for a team defines the scoring time for the scoring team and a censored time of the opponent, (c) when no goal occurs (in the occasion of $0-0$ or at the end of the game) then both times are considered as censored, and (d) all inter-arrival goal times are considered independent given the parameter values.

The implementation of our approach requires goal arrival times data that are not easily available via usual football data repositories such as `https://www.football.co.uk/`. For this reason, we were forced to explore

other sources to extract the appropriate information about each of the matches considered. One of the richest electronic resources offering relevant information is the understat website (`https://understat.com/`), where the interested researcher can find an extensive amount of features of each game regarding major leagues such as the English Premier League, the Spanish La Liga and the German Bundesliga.

In what follows, this chapter is organized into seven distinct sections. In Section 2.2, we formally explain the data structure and formulation. Section 2.3 provides modeling details and compares the fit for football data among the most popular parametric event-time models. Section 2.3.2 discusses various model extensions incorporating additional assumptions. It also compares these models using the Deviance Information Criterion (DIC) to assess their suitability for the given data. Section 2.4 focuses on the results providing parameter estimates, their practical interpretation, and the assessment of the goodness of fit of our model via simulations. Section 2.5 refers to additional modeling issues including the red card effect, model comparisons and censoring at half-time. We conclude this chapter with a short discussion in Section 2.6 about the limitations of our approach and the possible future extensions.

2.2 Goal Arrival Times Data in Football

2.2.1 Data Formulation

Modeling multiple event times in sports is not a trivial task since theoretically, an event can occur in any interval (no matter how small it is) with a positive probability. Hence, the actual time of the event is rarely captured precisely, and it is rounded up to the closest integer unit of time (that is, minute in football). Therefore, strictly speaking, the time t recorded for any goal scored is bi-variately censored, since it actually means that the goal occurred within the time interval $[t, t+1)$. In our case, the times obtained from the box scores are indeed recorded with a precision of one minute.

In order to implement our approach we need to record the i goal arrival time by $\mathcal{T}_i$ which represents the periods (in minutes) between two subsequent events (goals). Moreover, the time between the last goal scored in a game and the end of the game is also recorded as a censored goal arrival time.

Therefore, the recorded intervals of each game with a score of $GO_1 - GO_2$ will be equal to the number of goals scored, plus one for the final goalless interval, that is, $GO_1 - GO_2 + 1$. The index $i \in \{1, \ldots, n\}$ denotes the observation number, which corresponds to a recorded interval while n denotes the total number of recorded interval times under consideration and is set equal to the total number of goals scored plus the number of games considered in the dataset.

Additionally, to $\mathcal{T}_i$, we also use two binary indicators in order to specify the type of goal arrival interval: (a) the home team indicator $\mathcal{H}_i$ taking the value of one if $\mathcal{T}_i$ refers to the time of a goal scored by the home team or zero otherwise, and (b) a censoring indicator $\mathcal{C}_i$ which takes the value of one when the corresponding time refers to a censored time due to the end of the game and zero otherwise. The indicator $\mathcal{C}_i$ is used to identify (for $\mathcal{C}_i = 1$) a time interval with no goal, which is censored for both opponent teams. This interval is observed once in each game and corresponds to the last recorded interval of each game. Moreover, when

the game ends with no score (i.e., a 0-0 draw), it will be the only recorded interval.

Our approach is based on the idea that when a goal is scored, then the recorded time resets to zero, that is we model the gap arrival times. Hence, in our modeling approach, we assume two random variables T_{i1} and T_{i2} which reflect to the goal arrival times for the first and the second team in a game m. Then, we denote with t_{i1} and t_{i2} the corresponding i observed pair of times which are defined as

$$t_{i1} = \begin{cases} \mathcal{H}_i\mathcal{T}_i & \text{if } \mathcal{C}_i = 0 \\ \mathsf{NA} & \text{if } \mathcal{C}_i = 1 \end{cases} \tag{2.1}$$

$$t_{i2} = \begin{cases} (1-\mathcal{H}_i)\mathcal{T}_i & \text{if } \mathcal{C}_i = 0 \\ \mathsf{NA} & \text{if } \mathcal{C}_i = 1 \end{cases} . \tag{2.2}$$

The use of NA in the above data specification is due to the way we model censored observations in Bayesian analysis and the relevant MCMC software (i.e. in OpenBUGS or JAGS). Specifically, censored observations are considered as missing data where the censoring time point is known. Then, these (partially) "missing" data are generated from the corresponding (truncated) predictive distribution; see in Ntzoufras (Bayesian Modeling using WinBUGS page 297 - 298) for a detailed description.

To implement our approach in a Bayesian software such as OpenBUGS or JAGS, we needed to separately define the corresponding censoring times which were specified as

$$c_{i1} = (1-\mathcal{H}_i)(1-\mathcal{C}_i)\mathcal{T}_i + \mathcal{C}_i\mathcal{T}_i \tag{2.3}$$

$$c_{i2} = \mathcal{H}_i(1-\mathcal{C}_i)\mathcal{T}_i + \mathcal{C}_i\mathcal{T}_i \ . \tag{2.4}$$

In the above formulation, if a team scores, we set its censoring indicator equal to zero (indicating no censoring) and the censoring indicator for the opposing team equal to the observed interval time (i.e., the time elapsed since the previous goal). However, if an end-of-game interval with no goals is observed, both censoring indicators will be set to the corresponding interval time.

Let us explain this by considering, as an example, the data layout for the first match in our dataset, between Manchester UND and Leicester, which ended with a score of 2-1. For this game, the recorded times are $\mathcal{T} = (2, 80, 9, 4)$, the home indicator values are given by $\mathcal{H} = (1, 1, 0, 0)$, and end-of-game censoring indicator takes values $\mathcal{C} = (0, 0, 0, 1)$. These data are re-arranged as shown in Table 2.1.

From Table 2.1, it is clear that the survival times are presented in an unusual manner. As we have already mentioned, an important assumption underlying our approach and therefore also the above structure the data is that the recorded event time resets after every goal. This data handling is equivalent to modeling the goal arrival times between each event. To be more precise, in the game of Manchester UND versus Leicester presented in Table 2.1 the the goals were scored in the 2nd, 82nd, 91st minute of the game with 5 minutes of added time due to delays (i.e. the final duration of the game was 95 minutes). If we denote by $\mathcal{T}^* = (0, 2, 82, 91, 95)$ the original scoring times with by additionally appending the value of zero at the beginning of the vector and the total duration of the game, then the gap goal arrival times are simply calculated as $\mathcal{T}_i = \mathcal{T}^*_{i+1} - \mathcal{T}^*_i$ for $i = 1, 2, 3, 4$. Hence, the value of $t_{21} = 80$ means that 80 minutes have passed since the previous goal was scored. The missing times (t_{41} and t_{42}) in the last line of the game (as it

	Scoring	Time	Home	End-of-game		Goal Arrival Times		Censoring Times	
	Time	Interval	Team	Censoring		Home	Away	Home	Away
Game	$\mathcal{T}_i^*$	$\mathcal{T}_i$	$\mathcal{H}_i$	$\mathcal{C}_i$		t_{i1}	t_{i2}	c_{i1}	c_{i2}
1	2	2	1	0		2	NA	0	2
1	82	80	1	0	$\Rightarrow$	80	NA	0	80
1	91	9	0	0		NA	9	9	0
1	95	4	0	1		NA	NA	4	4

Table 2.1: Example of data layout for survival modeling: Data refers to the first match of English Premier League 2018-2019 (Manchester United – Leicester).

appears in Table 2.1) represents the inability to observe at what time would a team have scored from last scored goal due to the fact that the game was ended.

Finally, only one line of data will be recorded in the case of a 0-0 draw. The corresponding goal arrival times will be unobserved and thus will be set to NA, i.e., $(t_{i1}, t_{i2}) = (\mathsf{NA}, \mathsf{NA})$. Additionally, both censoring times will be set equal to the total duration of the game (90 minutes plus any extra injury or stoppage time added due to in-game delays). For instance, $(c_{i1}, c_{i2}) = (90, 90)$ for a game with no extra injury or stoppage time.

2.3 Bayesian Weibull Model Formulation

2.3.1 A Weibull Vanilla Model

Let us denote by T_{ij} the random variable that represents the goal arrival time concerning the i-th row and the j-th column (before observing it); where $i = 1, 2, \dots n$ and $j = 1, 2$

$$\log(T_{ij}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \sigma \epsilon_{ij} \tag{2.5}$$

where $\boldsymbol{\beta}$ is the vector of coefficients associated with the covariates $\boldsymbol{x}_{ij}^T$, σ is a scale parameter, and $\epsilon_{1j}, \dots, \epsilon_{nj}$ are independent and identically distributed according to the standard Gumbel distribution.

By setting $\gamma = 1/\sigma$ and making the relevant change of variables, it follows that

$$T_{ij} \sim \mathcal{W}eibull\left(\gamma, \boldsymbol{x}_{ij}^T \boldsymbol{\beta}\right) \tag{2.6}$$

Hence,

$$T_{ij} \sim \mathcal{W}eibull(\gamma, \lambda_{ij}), \quad \text{for } i = 1, 2, \dots, n, \text{ and } j = 1, 2,$$

with density $f(t) = \gamma \lambda t^{\gamma - 1} e^{-\lambda t^\gamma}$, expected value $E(T) = \lambda^{-\frac{1}{\gamma}} \Gamma\big(1 + 1/\gamma\big)$, median $Md(T) = \lambda^{-\frac{1}{\gamma}} (\log 2)^{1/\gamma}$, and hazard function $h(t|\lambda, \gamma) = \lambda \gamma t^{\gamma - 1}$ with $\gamma, \lambda > 0$.

The Weibull distribution is a more flexible counterpart to the Exponential distribution mainly due to the

non-constant hazards. More specifically, the hazard is monotonically increasing (decreasing) for $\gamma > (<)1$ while for $\gamma = 1$ the Weibull distribution becomes identical to the Exponential distribution leading to constant hazard.

Moreover, we use the standard vanilla formulation adopted from the basic football models of Karlis and Ntzoufras (2003) given by

$$\begin{array}{rcl} \log\left\{E(T_{i1})\right\} & = & \mu + home + att_{HT_i} + def_{AT_i} \\ \log\left\{E(T_{i2})\right\} & = & \mu + att_{AT_i} + def_{HT_i} \end{array},$$

where HT_i and AT_i denote the home and away team for i observed event time.

This model can be also written in a log-linear form in terms of the λ_{ij} parameters given by

$$\begin{array}{rcl} \log \lambda_{i1} & = & \mu^* + home^* + att^*_{HT_i} + def^*_{AT_i} \\ \log \lambda_{i2} & = & \mu^* + att^*_{AT_i} + def^*_{HT_i} \end{array}.$$

In this model formulation, $\mu^* = -\gamma\{\mu - \Gamma(1+1/\gamma)\}$, $home^* = -\gamma\, home$, $att^*_k = -\gamma\, att_k$ and $def^*_k = -\gamma\, def_k$ for $k = 1, \ldots, K$; where K is the number of teams in our dataset. Note, that the second formulation, which simply implements a vanilla structure on the λ parameters of the Weibull distribution rather on the expectations is just a re-parametrization of the first formulation.

Concerning the model parameters, μ is the intercept appearing in both Weibull distributions, the parameter home depicts the home effect, att_k and def_k are the parameters representing the attacking and defensive abilities of team k with $k = 1, 2, ...K$. For the attacking the the defensive parameters we need to consider a constraint in order to make the model identifiable. We have selected to use the sum-to-zero constraints given by

$$\sum_{k=1}^{K} att_k = \sum_{k=1}^{K} def_k = 0$$

to measure the differences in the log-expected times of scoring the next goal in comparison to a team of average attacking or defensive ability.

Parameter γ is essential for the behaviour of the goal-arrival times. When $\gamma < 1$ then the goal arrival rate decreases with time. Hence as the time passes the less likely it becomes to score. Accordingly, $\gamma > 1$ means that the goal arrival rate increases with time while for the value of one the distribution decays to the exponential distribution where the goal arrival rate is constant across time.

To complete the Bayesian formulation we need to specify our prior distributions for the model parameters. Here, we use low-information prior by considering normal priors with large prior variances for the parameters involved in the linear predictor, that is $att_k, def_k, \mu, home \sim Normal(0, 10^{-3})$ while for the positive parameter γ we use $\gamma \sim Gamma(10^{-3}, 10^{-3})$ distribution with mean equal to one and variable equal to 1000. In this work, we focus on using low-information prior distributions to allow the data to speak for themselves. The analysis could be further enhanced by incorporating expert knowledge or historical data into the prior, which is often available in football. However, we did not explore this aspect further, as the primary aim of this chapter is to present and validate a model for goal arrival times by solely using our dataset. For a more detailed discussion on the use of prior information, see Section 2.6.

For the model introduced in Section 2.3.1, the likelihood can be written as

$$f(\boldsymbol{y}|\boldsymbol{\vartheta}) = \prod_{i}^{n} f(\boldsymbol{y}_i|\boldsymbol{\vartheta}) = \prod_{i}^{n} f(t_{i1}, c_{i1}, t_{i2}, c_{i2}|\boldsymbol{\vartheta})$$

where $\boldsymbol{y}_i = (t_{i1}, c_{i1}, t_{i2}, c_{i2})$ is the vector of time data related with the i time interval under consideration, $\boldsymbol{y}$ is a $n \times 4$ matrix with all data times and $\boldsymbol{\vartheta} = (\mu, home, \boldsymbol{att}, \boldsymbol{def}, \gamma)$ is the vector parameters for the Weibull goal arriving time model defined in Section 2.3.1.

The probability density $f(\boldsymbol{y}_i|\boldsymbol{\vartheta})$ is the contribution of each observed time interval i to the likelihood. Under the data and model formulation described in Sections 2.2 and 2.3.1, this density can be defined by three different cases. The first case occurs when the home team scores ($\mathcal{H}_i = 1$ and $\mathcal{C}_i = 0$). Consequently, the goal-scoring time for the home team is observed to be t_{i1}, while the goal-scoring time for the away team is unobserved, but we know that it is greater than $c_{i2} = t_{i1}$. Therefore, the likelihood component is given by

$$f(\boldsymbol{y}_i|\boldsymbol{\vartheta}) = f(t_{i1}|\gamma, \lambda_{i1})S(c_{i2}|\gamma, \lambda_{i1}).$$

Similarly, in the second case, the away team scores ($\mathcal{H}_i = 0$ and $\mathcal{C}_i = 0$). The scoring time for the away team is now observed to be equal to t_{i2}, while the scoring time for the home team is censored at $c_{i1} = t_{i2}$. This results in

$$f(\boldsymbol{y}_i|\boldsymbol{\vartheta}) = S(c_{i1}|\gamma, \lambda_{i1})f(t_{i2}|\gamma, \lambda_{i1}).$$

Finally, the third case is when the game finishes with no goals by either of the competing teams ($\mathcal{H}_i = 0$ and $\mathcal{C}_i = 1$). In this case, both scoring times are censored, and the likelihood component is given by

$$f(\boldsymbol{y}_i|\boldsymbol{\vartheta}) = S(c_{i1}|\gamma, \lambda_{i1})S(c_{i2}|\gamma, \lambda_{i1}).$$

Hence, for any set of observed values $\boldsymbol{y}_i = (t_{i1}, c_{i1}, t_{i2}, c_{i2})$ for i time interval under consideration, the likelihood contribution of time time interval i is given by

$$f(\boldsymbol{y}_i|\boldsymbol{\vartheta}) = \begin{cases} f(t_{i1}|\gamma, \lambda_{i1})S(c_{i2}|\gamma, \lambda_{i1}) & \text{if} \quad \mathcal{H}_i = 1 \ \& \ \mathcal{C}_i = 0 \\ S(c_{i1}|\gamma, \lambda_{i1})f(t_{i2}|\gamma, \lambda_{i1}) & \text{if} \quad \mathcal{H}_i = 0 \ \& \ \mathcal{C}_i = 0 \\ S(c_{i1}|\gamma, \lambda_{i1})S(c_{i1}|\gamma, \lambda_{i1}) & \text{if} \quad \mathcal{H}_i = 0 \ \& \ \mathcal{C}_i = 1 \end{cases}.$$

The full likelihood can be now written as

$$f(\boldsymbol{y}|\boldsymbol{\vartheta}) = \prod_{i=1}^{n} \left[\left\{ f(t_{i1}|\gamma, \lambda_{i1})S(c_{i2}|\gamma, \lambda_{i1}) \right\}^{(1-\mathcal{C}_i)\mathcal{H}_i} \left\{ f(t_{i2}|\gamma, \lambda_{i1})S(c_{i1}|\gamma, \lambda_{i1}) \right\}^{(1-\mathcal{C}_i)(1-\mathcal{H}_i)} \times \left\{ S(c_{i1}|\gamma, \lambda_{i1})S(c_{i2}|\gamma, \lambda_{i1}) \right\}^{\mathcal{C}_i} \right].$$

The posterior distribution can be obtained by combining the likelihood described above with the prior distributions, as discussed in Section 2.3.1. A simple MCMC algorithm or a Gibbs sampler can then be implemented to obtain posterior summaries of the parameters of interest. In this analysis, we used OpenBUGS (v3.2.3 rev1012) to implement the MCMC algorithm.

In OpenBUGS, a data augmentation approach—commonly employed in survival models with censored observations—is used to simplify the simulation of the posterior distribution. This approach generates all

censored observations from their corresponding posterior predictive distributions, thereby creating a complete dataset of event times at each iteration of the MCMC algorithm. As a result, the likelihood simplifies to its usual form, given by

$$f(\boldsymbol{t}|\boldsymbol{\vartheta}) = \prod_{i=1}^{n} f(t_{i1}|\gamma, \lambda_{i1}) f(t_{i2}|\gamma, \lambda_{i1}) \ ,$$

where $\boldsymbol{t}$ is a $n \times 2$ matrix containing the (augmented) goal arrival times of i time interval in each row. The missing/censored goal-arrival times are generated from the $\mathcal{W}eibull(\boldsymbol{\vartheta}_{ij})$ distribution with the constraint that $T_{ij} > c_{ij}$. Hence, in the MCMC algorithm the following step is added

$$\begin{aligned} \text{If } \mathcal{C}_i = 1 \text{ or } \mathcal{H}_i = 0 \text{ then Sample } T_{i,1} &\sim \mathcal{W}eibull(\gamma, \lambda_{i1}) \mathcal{I}(T_{i1} > c_{i1}) \\ \text{If } \mathcal{C}_i = 1 \text{ or } \mathcal{H}_i = 1 \text{ then Sample } T_{i2} &\sim \mathcal{W}eibull(\gamma, \lambda_{i2}) \mathcal{I}(T_{i2} > c_{i2}) \ , \end{aligned}$$

where $\mathcal{I}(A)$ is an indicator function taking the value of one if A is true and zero otherwise.

Note that, this is exactly the way we define our Weibull model in OpenBUGS. Hence, we specify the likelihood by

$$T_{i,1} \sim \mathcal{W}eibull(\gamma, \lambda_{i1}) \mathcal{I}(T_{i1} > c_{i1}) \text{ and } T_{i2} \sim \mathcal{W}eibull(\gamma, \lambda_{i2}) \mathcal{I}(T_{i2} > c_{i2})$$

for all $i = 1, \ldots, n$, with c_{i1} and c_{i2} taking the value of zero when t_{i1} or t_{i2} is observed, respectively. Predicting future observations using MCMC is a key feature of Bayesian statistics. Consider a future game between the home team h and the away team a. We aim to predict a future goal arrival time, denoted as $\boldsymbol{t}^{\text{pred}}_{(h,a)}$. This prediction can be obtained directly from the posterior predictive distribution which is given by

$$f\big(\boldsymbol{t}^{pred}_{(h,a)} \,\big|\, \boldsymbol{y}\big) = \int f\big(t^{pred}_{(h,a)} \,\big|\, \boldsymbol{\vartheta}\big) f\left(\boldsymbol{\vartheta}|\boldsymbol{y}\right) d\boldsymbol{\vartheta}. \tag{2.7}$$

The probability density $f\big(\boldsymbol{t}^{pred}_{(h,a)}|\boldsymbol{\vartheta}\big)$ is given as the product of the probability densities of the Weibull distribution for the home and away goal arrival times. This depends only on parameters $(\gamma, \mu, home, att_h, att_a, def_h, def_a)$ which are related to the teams h and a competing each other in the game we wish to predict.

When using an MCMC algorithm, it is straightforward to generate a single set of future/predicted goal arrival times $\boldsymbol{t}^{pred}_{(h,a)}$ from the corresponding predictive distribution by adding the following simple steps in the MCMC sampler.

- Calculate $\lambda_1^{pred} = \mu + home + att_h + def_a + H$ and $\lambda_2^{pred} = \mu + att_a + def_h$
- Generate $t_1^{pred} \sim \mathcal{W}eibull(\gamma, \lambda_1^{pred})$ and $t_2^{pred} \sim \mathcal{W}eibull(\gamma, \lambda_2^{pred})$, respectively.
- If $t_1 < t_2 \leq 90 + t_{inj}$, the consider that home team scores. Otherwise, if $t_2 < t_1 \leq 90 + t_{inj}$ consider that the away team scores

In the above procedure, t_{inj} is the extra injury time played in each game. This can be generated using a simple distribution - for example a uniform in $(3, 7)$ interval or use a more a distribution estimated by past data. Moreover, model parameters $\boldsymbol{\vartheta}$ will be equal to their corresponding values generated in each iteration of the MCMC algorithm.

The above procedure simply generates a single set of goal arrival times. To generate a set of goal arrival times representing a full game, then we need to repeat the above procedure until the $T_K = \sum_{k=1}^{K} \min\{t_{k1}, t_{k2}\} > 90 + t_{inj}$ with $T_{K-1} \leq 90 + t_{inj}$; where t_{k1} and t_{k2} are the k-th generated goal survival time for the game with teams h vs. a. This procedure will be used to reconstruct the full league or predict future games under the Weibull model for goal arrival times; see Section 2.4.3 for details.

2.3.2 Model Extensions and Further Assumptions

In this section, we indirectly check for the validity of two main assumptions of the Weibull model presented in Section 2.3.1: (a) the validity of the Weibull distributional assumption and (b) the independence between goal arrival times. For the first, in Section 2.3.2.1, we consider three alternative distributions which are popular in the survival analysis context. We fit them using the Bayesian approach and we compare them using the deviance information criterion (DIC) of Spiegelhalter et al. (2002). For the independence assumption, we use a bivariate distribution as well as random effects models and compare them with the independent Weibull vanilla model (Section 2.3.2.2).

2.3.2.1 Evaluation of Different Distributional Assumptions for Goal Arrival Times

In order to check for the appropriateness of the Weibull assumption we compare this model with other similar models using distributions which are also standard in survival analysis. Hence we compare the Weibull formulation introduced in Section 2.3.1 with three additional distributions/models: (a) the exponential distribution, (b) the log-Normal distribution and (c) the log-logistic distribution.

a) The Exponential distribution with density $f(t|\lambda) = \lambda e^{-\lambda t}$ with mean $1/\lambda$ and hazard function $h(t|\lambda) = \lambda$ with $\lambda > 0$ and $t \geq 0$. Even though this distribution has its use in the survival analysis literature, its constant hazard constrains survival modeling to non-realistic scenarios. Note that the exponential is a special case of the weibull distribution for $\gamma = 1$.

b) The log-Normal distribution with density

$$f(t|\mu, \sigma) = \frac{\phi\left(\frac{\log(t)-\mu}{\sigma}\right)}{\sigma t},$$

with mean $e^{\mu+\sigma^2/2}$, median e^{μ} and hazard function given by

$$h(t|\mu, \sigma) = \frac{\frac{\phi\left(\frac{\log(t)-\mu}{\sigma}\right)}{\sigma t}}{1 - \Phi\left(\frac{\log(t)-\mu}{\sigma}\right)},$$

with $\mu \in \mathcal{R}, \sigma > 0, t > 0$ where $\phi(\cdot)$ is the probability density function and $\Phi()$ is the cumulative distribution function of a standard normal variant. The hazard in this case makes an initial increase from zero to a maximum and then as $t \to \infty$ tends approximately to zero, creating upside-down bathtub-shaped hazards. This hazard behaviour makes the log-Normal model among the most popular AFT models in survival analysis.

c) The log-logistic distribution with density

$$f(t|a,b) = \frac{b}{a}\frac{(t/a)^{b-1}}{\left(1+(t/a)^b\right)^2}$$

with mean $\frac{a\pi/b}{\sin(\pi/b)}$, median the scale parameter a and hazard function

$$h(t|a,b) = \frac{\left(\frac{b}{a}\right)\left(\frac{t}{a}\right)^{b-1}}{1+\left(\frac{t}{a}\right)^b},$$

with $t \geq 0$ and $a, b > 0$. This model allows for a compromise between the Weibull and the log-Normal model due to the flexibility of its hazards. In particular, the hazard decreases monotonically as t increases when $b < 1$ while accounting for uni-modal hazards if $b > 1$ similar to the log-Normal case.

We compare and evaluate the predictive ability of the fitted models using DIC. Results are presented in Table 2.2 for the English Premier League 2018-19 data we consider in this article. The DIC values indicate that the Weibull model clearly outperforms its competing models in terms of the predictive performance.

	Distribution			
	Exponential	log-Normal	log-logistic	Weibull
DIC	11040	11040	11180	**11010**

Table 2.2: Diagnostic criteria to assess each models' predictive performance and goodness of fit. MCMC chain run for 10000 iterations with 1000 iterations taken as burnin.

2.3.2.2 Assessing Dependence Between Goal Arrival Times

In this section, we introduce and assess the dependence between the goal arrival times of the two opposing teams by considering (a) the Marshall-Olkin Bivariate Weibull Model and (b) random effects models.

Among the popular choices for modeling bivariate survival times, is the Marshall Olkin Bivariate Weibull distribution. It is based on the Weibull distribution and can be assumed as a natural extension of it in order to account for dependence. The distribution can be derived by assuming independent Weibull random variables U_0, U_1 and U_2 assuming that they have the same shape parameter γ and scale parameters λ_0, λ_1 and λ_2, respectively. Then for $T_1 = U_0 \wedge U_1, T_2 = U_0 \wedge U_2$, the assumed joint distribution of (T_1, T_2) is the Marshall Olkin Bivariate Weibull distribution. To mathematically specify of the Marshall Olkin Bivariate Weibull model, let

$$(T_1, T_2) \sim MOBW(\gamma, \lambda_0, \lambda_1, \lambda_2)$$

with joint probability density function given by

$$f_{T_1,T_2}(t_1,t_2) = \begin{cases} f_W(t_1;\gamma,\lambda_1) f_W(t_2;\gamma,\lambda_0+\lambda_2) & \text{if } 0 < t_1 < t_2 \\ f_W(t_1,\gamma,\lambda_0+\lambda_1) f_W(t_2,\gamma,\lambda_2) & \text{if } 0 < t_2 < t_1 \\ \frac{\lambda_0}{\lambda_0+\lambda_1+\lambda_2} f_W(t;\gamma,\lambda_0+\lambda_1+\lambda_2) & \text{if } 0 < t_1 = t_2 = t \end{cases}$$

where $f_W(x;\gamma,\lambda) = \gamma\lambda x^{\gamma-1}e^{-\lambda x^\gamma}$. The *joint survivor function* has a similar closed form formulation and is given by: $S_{T_1,T_2}(t_1,t_2) = S_W(t_1;\gamma,\lambda_1)S_W(t_2;\gamma,\lambda_2)S_W(t_1 \wedge t_2\ \gamma,\lambda_0)$; where $S_W(t;\gamma,\lambda) = e^{-\lambda t^\gamma}$, $\lambda_0,\lambda_1,\lambda_2,\gamma,t_1,t_2 > 0$. The parameter λ_0 is connected with the correlation between T_1 and T_2 although there is no closed form expression between the two quantities (Lai et al., 2017). Within the Bayesian paradigm, using MCMC, we can easily estimate the correlation induced by such a model using the posterior predictive distribution.

The Weibull model for the goal arrival times can be considered a special MOBW model case since it can be derived for $\lambda_0 = 0$. Hence, we may assess whether the additional parameter λ_0, which introduces correlation between the goal arrival times of the two competing teams, is necessary for improving the model's fit and predictive performance by comparing the DIC measure of these two models.

The model was fitted using R2OpenBUGS (Sturtz et al., 2020) using the zeros trick since this distribution was not explicitly available on this platform. Under this approach, the log-likelihood is calculated as

$$\begin{aligned}\mathcal{L} = \sum_{i=1}^{n}\mathcal{L}_i \text{ with } \mathcal{L}_i \quad = \quad & (1-\mathcal{C}_i)\left\{\mathcal{H}_i \log f_W(\mathcal{T}_i|\lambda_1,\gamma) + \mathcal{H}_i \log S_W(\mathcal{T}_i|\lambda_0+\lambda_2,\gamma)\right. \\ & +(1-\mathcal{H}_i)\log f_W(\mathcal{T}_i|\lambda_0+\lambda_1,\gamma) + (1-\mathcal{H}_i)\log S_W(\mathcal{T}_i|\lambda_2,\gamma)\} \\ & +2\mathcal{C}_i \log S_W(\mathcal{T}_i|\lambda_0+\lambda_1+\lambda_2,\gamma).\end{aligned}$$

Moreover, we have introduced dependence using random effects. We have considered two random effects models: The first one accounts for effects concerning the game itself while the second was fitted by additionally considering team random effects. As a natural follow-up of investigating and accounting for possible correlations that are arising within each game event-by-event but also as the League progresses, we induced random effects as additive terms in the linear predictor in Section 2.3.1.
To be more specific, the log mean goal arrival time between goals in Section 2.3.1 is linked to both fixed and random effects as:

$$\begin{aligned}\log\left\{E(T_{i1})\right\} &= \mu + home + att_{HT_i} + def_{AT_i} + \epsilon_{G_i} \\ \log\left\{E\big(T_{i2}\big)\right\} &= \mu + att_{AT_i} + def_{HT_i} + \epsilon_{G_i}\end{aligned}$$

where G_i is a variable which indicates the game where the i event took place, ϵ_g are the within-game random effects for $g = 1,2,\ldots,n_G$ where n_G is the number of games (or equivalently ϵ_{G_i} for $i = 1,\ldots,n$). These random effects are assumed to be exchangeable, and hence:

$$\epsilon_g \sim N\left(0,\sigma_G^2\right) \quad \text{with} \quad \sigma_G^{-2} \sim Gamma\left(10^{-1},10^{-1}\right);\ \text{for } g = 1,\ldots,n_G$$

To investigate for potential negative correlations between goal arrival times, a reversed signs game effect model was employed. The rationale behind incorporating this model is to capture cases where within each game, the leading team forces goal arrival times to reduce while the trailing team has the opposite statistical behaviour.

Finally, a team-specific - random effects model was fitted to capture the potential heterogeneity between teams and its impact on the overall timing of goals.

In this case, the log mean of goal arrival times between goals has a slightly different formulation presented as:

$$\log\{E(T_{i1})\} = \mu + home + att_{HT_i} + def_{AT_i} + \epsilon_{G_i} + \tilde{\epsilon}_{HT_i} + \tilde{\tilde{\epsilon}}_{AT_i}$$
$$\log\{E(T_{i2})\} = \mu + att_{AT_i} + def_{HT_i} + \epsilon_{G_i} + \tilde{\epsilon}_{AT_i} + \tilde{\tilde{\epsilon}}_{HT_i}$$

where HT_i, G_i are defined as above, but now the random effect concerns each of the teams rather than solely the game where the i-th event took place. With $\tilde{\epsilon}_{i,j}$ we are referring the centered random effect for $i = 1, 2, ..., n$ and $j = 1, 2, ..., n_G$.

According to the DIC values in Table 2.3, the Weibull fixed-effects model outperforms the game-only random-effects models. To explore potential negative correlations between goal arrival times, a reversed-signs random-effects model (labelled as "Different signs" model in column 5 of Table 2.3) was examined. For the random-effects model incorporating team effects, the DIC is identical to that of the fixed-effects model, indicating no improvement in the predictive performance.

	Model				
	Independent	*Bivariate*	*Random Effects*		
	Weibull	MOBW	Game Effects	Different Signs	Team Effects
DIC	**11010**	13910	11150	11150	11010

Table 2.3: Diagnostic criteria to assess each models predictive performance and goodness of fit. MCMC chain run for 11000 iterations with 1000 iterations taken as burnin

2.4 English Premier League 2018-19 Data

The English Premier League 2018-2019 data has received a reasonable amount of attention as it is considered a major annual soccer event. This data consisted of 20 teams and $n_G = 380$ games in total while the total number of observations following the data structure presented in Table 2.1, was $n = 1452$ (source: English Premier League understat webpage: `https://understat.com/league/EPL`). Among the 380 games, 48% (34%) were home (away) wins while 19% were draws. Event time-focused descriptives include the median scoring time of the home (away) team which was measured to be 50 (51) with a standard deviation of 27 (26) implying a rough empirical estimate of the median arrival time driven score for this data to be $1 - 1$.

2.4.1 Model Based Inference

The model was fit to the data using Markov Chain Monte Carlo (MCMC) via R2OpenBUGS software and it ran until convergence was assured (11000 iterations with 1000 iterations as burn-in).

Starting from estimates regarding the attacking and defensive abilities of each team, the posterior means accompanied with forests plots of the posterior team abilities are presented in Figure 2.1 while the estimates regarding the remaining parameters of the model are given in Table 2.4.

				Posterior Quantiles	
	Mean	Median	SD	2.5%	97.5%
Home Advantage (*home*)	-0.193	-0.199	0.053	-0.286	-0.086
Intercept (μ)	4.025	4.028	0.035	3.952	4.090
Weibull Shape (γ)	1.131	1.130	0.028	1.078	1.186

Table 2.4: Posterior summaries for the 2018-19 English Premier League under the Weibull fixed effects model - the log(Mean) case.

2.4.2 Interpretation of the model

This section focuses on the analysis and the interpretation of the estimated model parameters for fixed effects Weibull model fitted on the data from the matches of the English Premier League, season 2018/19; see Section 2.3 for the detailed description of the proposed full Bayesian formulation.

Table 2.4, presents the basic model parameters of the fitted model. From this table, we observe that both the home effect *home* and the constant parameter μ appear to be away from zero indicating that should be included in the model formulation. Specifically, there is a clear home advantage (posterior mean of -0.19) suggesting that the home team experiences a decrease in mean or median goal arrival time of -0.19 (on a logarithmic scale) and a 17.3% reduction in median goal arrival time (on the original scale) compared to the corresponding goal arrival times when playing away against the same team (posterior credible interval: $8.25\% - 25\%$). Concerning the constant parameter μ, the log mean goal arrival times in away games is expected to be 4.025 which corresponds to around 56 minutes when two teams of average performance competing against each other.

Regarding the shape parameter γ, the data suggest a posterior mean of 1.13 and a 95% posterior interval ranging from 1.078 to 1.186. The posterior distribution of γ is clearly greater than one, implying (a) evidence against the exponential model (where $\gamma = 1$) and in favour of the Weibull model, and (b) that the goal-scoring rate increases as the match progresses. This result aligns with the findings reported by Silva and Swartz (2016), which indicate that goal scoring is not constant during a match. Moreover, the Weibull model was also found better than the exponential one in terms of DIC (11010 vs. 11040, respectively); see also Section 2.3.2.1 and Table 2.2 for details.

Before proceeding with the analysis of the team ability parameters, we should note that the attacking abilities of teams in goal arrival time models are a decreasing function of their actual attacking strength. Better teams tend to score more frequently, meaning they have shorter time intervals between their goals. Consequently, the Weibull ability parameters will be lower for stronger teams. Equivalently, the defensive team parameters will be an increasing function of each team's defensive quality, indicating that teams with better defensive skills will delay the scoring time of their opponents.

Figure 2.1 presents the 95% posterior credible intervals of the team attacking and defensive ability parameters (att_k and def_k for $k = 1, \ldots, 20$) of the Weibull model. The model appears to capture in a reasonable way the attacking and defensive team abilities following their final league rankings. From this

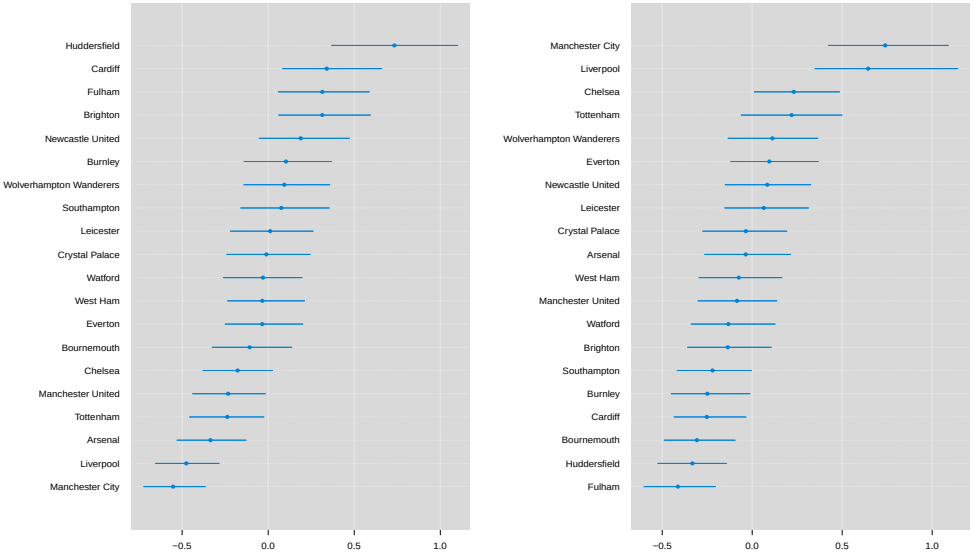

Figure 2.1: 95% Posterior intervals for the attacking (left) and defensive (right) abilities of each of the teams in the dataset along with posterior estimates of each of the teams' final points using both the posterior mean and the posterior median (25000 iterations, 1000 burn-in).

figure, we observe that Manchester City, Liverpool, Arsenal, Tottenham, and Manchester United not only have the lowest (i.e., the best) attacking abilities (sorted from smallest to largest), but their 95% posterior intervals do not include the value of zero, indicating their superiority over a team with "average" attacking ability. This pattern does not exactly match the final league table rankings, as these teams finished 1st, 2nd, 5th, 4th, and 6th. However, it aligns perfectly with their ranking by the number of goals scored. For this reason, although Chelsea finished 3rd in the league, it has the 6th-best attacking ability, as this was Chelsea's ranking in terms of goals scored. In terms of defense, Manchester City and Liverpool are once again dominant, followed by Chelsea, with the highest defensive ability parameters. For these three teams, the posterior distributions of the defensive parameters are well place above zero, indicating a clearly greater defensive performance over a team with "average" defensive ability. Regarding the below 'average' worst-performing teams, Huddersfield, Fulham, Cardiff, and Brighton had the poorest attacking abilities, while Fulham, Huddersfield, Bournemouth, Cardiff, Burnley, and Southampton had the weakest defensive abilities (from worst to best).

2.4.3 League Reconstruction

Assessing the in-sample predictive ability of our model was what naturally followed after the previously mentioned clear indication of our model's decency. The predictive distribution was used to conduct such model assessment through MCMC sampling. In particular, we used the predictive distribution to regenerate the League's rank under our model, and make comparisons to the actual rank. This was a two-step procedure since we first had to simulate goal arrival times from our model's predictive distribution given by (2.7). The

pseudocode used for this purpose is presented in Algorithm 1.

Algorithm 1 Algorirthm for League Regeneration using Weibull Survival Model.

- **Input**: $\left(\gamma^{(t)},\ \eta_{1,G}^{(t)},\ \eta_{2,G}^{(t)}\right)$: MCMC values of the parameters of the model for each MCMC iteration t and games G.
- **Output**: $L^{(t)} = Res[,]$

  for t = 1 **to** T_{MCMC} **do**
  - $L^{(t)} = 0$
  - **for** $G = 1$ to N_G **do**
    - $t_{inj} \sim \mathcal{U}[3, 7]$
    - $gt = g_1 = g_2 = 0$;
    - **while**(gt $< 90 + t_{inj}$) **do**
      - **for** $k = 1$ to 2 **do**
        $$st_k \sim Weibull\Big(e^{-\gamma^{(t)}\left\{\eta_{k,G}^{(t)} - \log\Gamma\left(1+1/\gamma^{(t)}\right)\right\}},\ \gamma^{(t)}\Big)$$
      - $gt = gt + min(st_1, st_2)$
      - **if** $(gt > 90 + t_{inj})$ **stop**
      - **if** $(st_1 < st_2)$ $g_1 = g_1 + 1$; else $g_2 = g_2 + 1$
      - $Res[G,] = (g_1, g_2)$
  - $L^{(t)} = Res[,]$

  Return League results $L^{(t)}$ for $t = 1, 2, ..., T_{MCMC}$

  Indexes: $G = 1, 2, ..., N_G$*;* N_G*: number of games* $t = 1, 2, ..., T_{MCMC}$*;* T_{MCMC}*: number of MCMC iterations*

Table 2.5 contains posterior, median, standard deviations and the 95% predictive credible intervals for the points of each team under the fixed effects Weibull model. In terms of the finally predicted rank, it is evident that the model exactly captures 10 out of 20 while all other actual points are within the 95% posterior credible interval of the predicted points. This result points to an outstanding agreement between our model and the actual rank since the structural assumptions of our model are minimal and the main modeling target is the scoring goal arrival times rather than the actual goals. More specifically, the model captures the top two highest-rated teams and the four lowest-rated teams accurately. All other deviations between our model and reality are minor since the maximum misplacement of a team in the final rank is observed is two ranks and the difference between points can be translated as misspredicting a maximum of two matches. Overall, our model appears to be predicting the exact position of the top and worst teams in the league while making slight misspredictions in the medium-ability teams where the posterior ranking probabilities are too close to making clear ranking discrimination and thus adds to the plausibility of this parsimonious fixed effects model for in-sample predictions especially when the endpoint is predicting the

	Ranking		Obs.	Posterior Predicted Points				
	Obs.	Pred.	Points	Median	Mean	SD	$P_{2.5\%}$	$P_{97.5\%}$
Man. City	1	1	98	90	90.12	7.05	76	103
Liverpool	2	2	97	89	88.89	7.22	74	102
Chelsea	3	4	72	67	66.63	8.67	49	83
Tottenham	4	3	71	69	69.33	8.57	52	86
Arsenal	5	5	70	65	65.16	8.68	48	82
Man. Und	6	6	66	59	58.78	8.79	42	76
Wolves	7	8	57	51	51.35	8.66	35	69
Everton	8	10	54	55	55.09	8.70	38	72
West Ham	9	7	52	48	48.15	8.91	31	66
Leiceister	10	12	52	55	54.52	8.65	38	72
Watford	11	9	50	48	48.00	8.53	32	65
Crystal Palace	12	11	49	51	50.77	8.76	34	68
Newcastle	13	13	45	46	46.10	8.25	30	63
Bournemouth	14	14	45	44	44.32	8.65	28	62
Burnley	15	16	40	40	39.87	8.21	25	57
Southampton	16	15	39	42	41.82	8.28	26	59
Brighton	17	17	36	36	36.15	8.05	21	53
Cardiff	18	18	34	32	31.91	7.56	18	47
Fulham	19	19	26	28	28.18	7.40	15	44
Huddersfield	20	20	16	21	21.68	6.37	10	35

Notes: *Obs.: Observed; Pred.:Predicted; SD: Standard deviation;* $P_{a\%}$*: a% percentile; Predicted ranking is based on posterior mean predicted points*

Table 2.5: Reconstructed league for the EPL 2018-2019 data under the Weibull model.

top/worst teams' rankings.

Table 2.6 presents the Root Mean Square Error (RMSE) of the posterior estimates for the mean and median points from the reproduced final league table under the fixed-effects Weibull model. The results further validate the model's goodness-of-fit, as it reproduces the final league standings with an average error of approximately four points and one ranking position. This indicates that, on average, the model maintains accurate predictions with an error, on average, of one win and one draw for each team.

Another useful insight drawn from Table 2.6 is that the RMSE for the posterior mean estimate of ranks is calculated to be 1.4 positions in the final league table. This implies that, on average, our model predicts final league positions with a discrepancy of just one or two positions. This result further supports the model's strong fit to the data, particularly given the simplicity of the proposed fixed-effects model and its indirect approach to modeling the goals scored by each team.

	Mean Points	Median Points	Rank
RMSE	4.168	4.153	1.4

Table 2.6: Root Mean Square Error (RMSE) of the posterior mean and median points along with the posterior mean of ranks after 12000 replications of the League under the proposed model.

2.4.4 Out-of-sample prediction

The out-of-sample predictive ability of the model is also of major scientific concern, especially due to its intriguing nature and its' applications to betting. In this case, the predictive ability of the model is expected to lessen compared to the in-sample measures presented in the previous subsections, however, the loss of out-of-sample predictive ability of a model with that impressive in-sample predictive performance is expected to be minor.

Under the Bayesian framework, the only adjustment required to assess the out-of-sample ability of our model is to assume that a portion of the data is known, and the rest of the data is simulated using the predictive distribution. This is to say that in the in-sample model assessment conducted previously, we assumed that the entirety of the data (League) is known and each new League was simulated from the posterior distribution under the assumption that the observed teams' abilities hold true for the teams' considered in the simulation of each new league.

2.4.4.1 Second half of the season prediction

The choice of having as training dataset the first half of the season can be reasonably justified in both purely statistical terms, and by the nature of the sport under the League considered. From a purely statistical point of view, having half the data as a training dataset can be considered to be sufficient in terms of predicting the other half of the data since we are using half of the entirety of the information contained in the original dataset for learning purposes. The additional argument from the nature of the sport and the League is that during the first half of the Season, each team is playing against each other team. That means that all teams have competed with each other at least once and hence if the endpoint was to maximize the amount of information contained in the training dataset while minimizing the length of the data used for its construction, that would be during the half of the English Premier League season.

Among the measures that assess the out-of-sample predictive ability of our model is the percentage of agreement in predicting the games of the second half of the season. Since Bayesian theory allows us to make inference on the entirety of its distribution of that percentage, and not consider it as a point estimate, the posterior distribution presented on the left panel of Figure 2.2, contains all the information we need to get an initial idea of how are model performs. More specifically, the posterior median (mean) is 46.84% (46.77%) with a 95% credible interval of $40.53\% - 53.16\%$. Those posterior measures are indications that our model performs better than expected especially considering the fact that if we were to choose the outcome of matches at random, we would expect to make accurate predictions in 33% of the time. Since that percentage is not included in the 95% credible interval, our parsimonious fixed effects model with minimum assumptions,

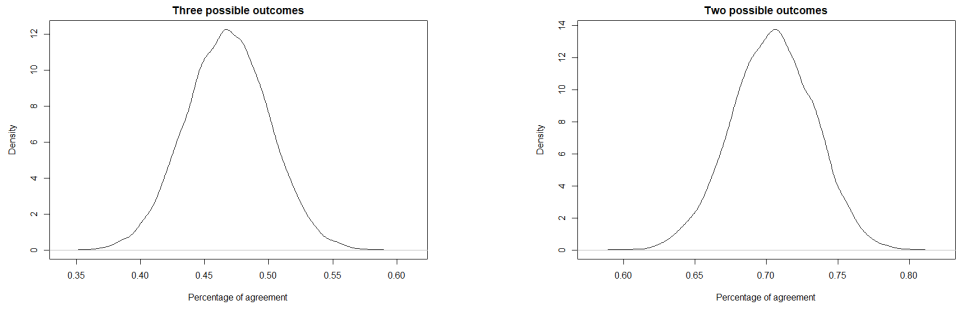

Figure 2.2: Posterior distributions of the percentage of the agreement in predicted games for the second half of the season for English Premier League 2018/19 under the proposed fixed effects Weibull model. The posterior distribution of the percentage of agreement on the left applies when the events of interest are three (aka Win/Draw/Loss) while on the right, the posterior distribution applies when the event of interest is considered to be two (Win & Loss / Draw).

provides significantly better outcome prediction even in the out-of-sample case.

This results can be made more comprehensible if we consider two events instead of three, thus making comparisons with a percentage of 50% which is the expected times of correct predictions if we were to choose between two outcomes completely at random. Therefore, if we assume that Win or Loss of a match is one event and draw is the other event, then the posterior distribution presented on the right panel of Figure 2.2 reveals very promising results with regards to the predictive ability of our model making it more easily understandable since the posterior median (mean) are now 70.53% (70.45%) with a 95% credible interval of $64.74\% - 75.79\%$.

2.4.4.2 Brier Score

Another useful metric for measuring the predictive ability of our model by understanding match-specific surprising outcomes is the Brier penalty (BP). This metric has been used by Spiegelhalter and Ng (2009) as a modification to the original Brier Score introduced by Brier (1950), and can be expressed as the squared distances between a vector of probabilities and the actual outcome.

The Brier Score for categorical data is given by

$$BS = \frac{1}{N}\sum_{k=1}^{N}\left[\sum_{j=1}^{J}\left(p_k(j) - a_{kj}\right)^2\right],$$

where N is the number of predicted outcomes, J is the number of categories, $p_k(j)$ is the predicted probability for k outcome and j category, and a_{kj} is the binary variable taking the value of one if category j observed and zero otherwise.

In our case, we consider the categorical variable Y_i denoting the final outcome of game i (for $i = 1, \ldots, N$) in the form of (home) win ($Y_k = 1$), draw ($Y_k = 2$) and loss ($Y_k = 3$); where N denotes the number of games under consideration. Hence, $\mathbf{a}_k = (a_{k1}, a_{k2}, a_{k3})$ will be equal to $(1, 0, 0)^T$ for a home win, $(0, 1, 0)^T$ for a

draw, and $(0,0,1)^T$ for the home loss. We can rewrite the Brier score as

$$BS = \frac{1}{N}\sum_{k=1}^{N} BS_k \quad \text{and} \quad BS_k = \sum_{j=1}^{3} \left(p_k(j) - a_{kj}\right)^2 .$$

The Brier Score for k game decomposes to

$$\begin{aligned} BS_k &= \sum_{j=1}^{3} \left(p_k(j)^2 + a_{kj}^2 - 2a_{kj}p_k(j)\right) \\ &= \sum_{j=1}^{3} p_k(j)^2 + 1 - 2p_k(Y_k) \end{aligned}$$

This expression is exactly the modified Brier Score named as Brier Penalty (BP) in Spiegelhalter and Ng (2009) which is association football (soccer) can be written in the following simplified form:

$$BP_k = 1 + P(\text{Home Win})^2 + P(\text{Draw})^2 + P(\text{Away Win})^2 - 2P(\text{Actual Outcome}).$$

All probability outcomes in the above formula refer to the k-th game. The support of this metric is $[0,2]$ with BP $\to 0$ indicating perfect predictions while BP $\to 2$ indicating the opposite.

We applied this metric to all 190 matches simulated from the predictive distribution obtaining a Brier Penalty for each of them. Even though the bibliography regarding this metric is unclear with regards to the threshold above which the model's predictive ability can be problematic, we concentrated our attention on matches for which the Brier Penalty was above 1 to account for the symmetry of this metric interpretational property.

One direct positive insight after computing this metric was that only 31 out of a total of 190 matches have gotten a Brier Penalty of above 1. This meant that our model provided decent out-of-sample predictions for around 83.7% of cases hence providing more evidence that adds we introduced an adequate predictive model in both the in-sample and the out-of-sample case.

Since the Brier Penalties can be calculated for each match, it can be insightful in investigating surprising results in specific matches of the League. In our case, these matches are presented in Table 2.7 along with the final rank difference of the teams participating. The ranking differences presented in Table 2.7 act as a tool for making sense of the matches that our model did not capture. In particular, Table 2.7 clearly indicates that those miss-predictions can be justified by the fact that in most cases, the final outcome is the opposite of what would have logically been predicted if we were to use the final ranks of teams for prediction of outcomes. The most obvious example of the existence of this phenomenon is the match between Newcastle United and Manchester City. The winner of this League lost to a team with a rank distance of 12 positions in the final rank with a score of 2-1. This surprising aspect can be observed for the majority of matches in Table 2.7 with 20/31 matches having an absolute ranking difference of above or equal to 5 ranks (with the magnitude of that metric being above 10 for 11 matches).

Moving beyond the specifics of the miss-predictions and their justifications, we compared the Brier penalties concerning the out-of-sample predictions under our model with the Brier penalties if we were to assign equal probabilities to each of the outcomes. That would mean that instead of assigning the

Win/Draw/Loss probabilities according to our model, we would assign a probability of 0.33 to each, making it equivalent to comparing our model to random chance mimicking what the analysis we conducted in the in-sample case. According to Spiegelhalter and Ng (2009) this comparison can be made by summing over all Brier penalties calculated under our model and then comparing it to the summation of all Brier penalties calculated under random chance. The Brier penalty concerning the 190 matches of the second half of the Season under our model is calculated to be ≈ 105.35 whereas the Brier penalty under random chance is calculated to be ≈ 126.67 making which can be translated into a 16.82% difference in favour of our model.

Taking into consideration the information presented in this subsection, we can draw the safe conclusion that the miss-predictions of our proposed model, are justified since any discrepancies between what the model predicted in terms of probability that translate to a Brier penalty, and reality, are what it was a-prior expected provided information concerning the final ranking of the teams. Therefore, our model appears to be performing exceptionally well even in out-of-sample scenarios.

2.5 Further modeling issues

2.5.1 Comparison to the Double Poisson model

To further affirm both the validity of our methods, especially those concerning the transcription of the goal arrival times to actual goal events, and its is-sample predictive ability, our results concerning the regeneration of the EPL league were compared to a well-established and widely used model for modeling goals events as integers, the double Poisson model.

Comparing Table 2.8 to Table 2.5, it is evident that both models seem to capture the final ranking of EPL in a compatible manner and very precise manner. To be more specific, both models seem to affirm that the median points of each team are systematically lower than the actual points while all actual points are captured accurately by each 95% credible interval. To provide further evidence of the surprising comparability of the two models, the percentage of ranks that are predicted exactly is 50%. This is a surprising finding given the fact that our model makes limited assumptions and models the actual events (goals) indirectly. A final surprising remark concerning the proposed model is that the posterior standard deviation of each team's points is systematically lower than that in the double Poisson case, providing evidence that our model's predictions are slightly more accurate with regard to the prediction of the rank of each team.

Match Opponents		Final	Probability			Rank
			Home		Away	
Home	Away	Score	Win	Draw	Win	Difference
Tottenham	Wolves	1–3	0.69	0.19	0.11	-3
Leicester	Cardiff	0–1	0.65	0.19	0.15	-9
Burnley	West Ham	2–0	0.20	0.21	0.59	5
Chelsea	Southampton	0–0	0.79	0.13	0.08	-13
Tottenham	Man. Und	0–1	0.68	0.16	0.16	-2
Leicester	Southampton	1–2	0.60	0.22	0.18	-7
Watford	Bournemouth	0–0	0.67	0.18	0.15	-4
Bournmouth	Chelsea	4–0	0.17	0.18	0.65	11
Man. Und	Burnley	2–2	0.80	0.11	0.09	-9
Newcastle	Man City	2–1	0.05	0.13	0.82	12
Liverpool	Leicester	1–1	0.77	0.17	0.06	-7
West Ham	Liverpool	1–1	0.06	0.15	0.79	8
Brighton	Burnley	1–3	0.61	0.21	0.18	2
Man. Und	Liverpool	0–0	0.08	0.14	0.78	4
Burnley	Tottenham	2–1	0.05	0.10	0.85	11
Tottenham	Arsenal	1–1	0.58	0.19	0.23	-1
Everton	Liverpool	0–0	0.07	0.15	0.78	6
Southampton	Tottenham	2–1	0.10	0.13	0.77	12
Chelsea	Wolves	1–1	0.65	0.22	0.13	-4
Bournemouth	Burnley	1–3	0.66	0.17	0.17	-1
Leicester	Newcastle	0–1	0.56	0.28	0.17	-4
Fulham	Everton	2–0	0.16	0.17	0.67	11
Chelsea	Burnley	2–2	0.87	0.09	0.04	-12
Bournemouth	Fulham	0–1	0.65	0.17	0.17	-5
Arsenal	Crystal Palace	2–3	0.68	0.18	0.14	-7
Tottenham	West Ham	0–1	0.72	0.16	0.12	-6
Huddersfield	Man. Und	1–1	0.14	0.16	0.70	14
Bournemouth	Tottenham	1–0	0.14	0.16	0.70	10
Arsenal	Brigton	1–1	0.67	0.17	0.15	-12
Tottenham	Everton	2–2	0.69	0.16	0.15	-4
Man. Und	Cardiff	0–2	0.75	0.14	0.11	-12

Table 2.7: Assessed probabilities for 31 score-specific matches of the second half of the season with Brier penalties above one, using the Weibull model.

2.5.2 The red card effect

The impact of red cards on factors influencing goal-scoring rates has attracted scientific attention (Červenỳ et al., 2018). Specifically, it is expected that an increase in the number of red cards received by a team

	Ranking		Obs.	Posterior Predicted Points				
	Obs.	Pred.	Points	Median	Mean	SD	$P_{2.5\%}$	$P_{97.5\%}$
Man City	1	1	98	92	91.74	7.33	76	105
Liverpool	2	2	97	91	90.49	7.53	75	104
Tottenham	4	3	71	70	69.70	9.35	51	87
Chelsea	3	4	72	68	67.55	9.49	49	86
Arsenal	5	5	70	66	65.40	9.46	47	84
Man United	6	6	66	60	59.55	9.52	41	78
Everton	9	7	54	58	57.89	9.62	39	76
Leicester	7	8	52	55	54.70	9.66	36	74
Wolves	8	9	57	53	53.16	9.47	34	72
Crystal Palace	11	10	49	52	51.75	9.68	33	71
West Ham	12	11	52	51	51.25	9.61	33	70
Watford	10	12	50	49	49.22	9.60	31	68
Newcastle	13	13	45	48	48.52	9.55	30	68
Bournemouth	14	14	45	46	46.11	9.50	28	65
Southampton	16	15	39	41	41.51	9.21	24	60
Burnley	15	16	40	40	40.18	9.12	23	58
Brighton	17	17	36	37	37.21	8.96	21	55
Cardiff	18	18	34	32	32.32	8.50	17	50
Fulham	19	19	26	27	27.74	8.01	14	45
Huddersfield	20	20	16	21	21.41	7.13	9	37

Notes: *Obs.: Observed; Pred.:Predicted; SD: Standard deviation;* $P_{a\%}$*: a% percentile; Predicted ranking is based on posterior mean predicted points*

Table 2.8: Reconstructed league for the EPL 2018-2019 data under the Double Poisson model.

reduces its probability of scoring. Consequently, the scoring time for that team should increase compared to the opposing team's scoring time, assuming the latter does not also receive a red card during the same period. During the EPL 2018/19 season, a total of 47 red cards were issued across 380 matches. Subsequently, the red card covariate (x_{ij}) for period i and team j is incorporated into the model formulation in the following way:

$$x_{ij} = \begin{cases} 0, & \text{no red card during time period } i \\ -1, & \text{If the home team } (j=1) \text{ plays with fewer players for period } i \\ 1, & \text{If the away team } (j=2) \text{ plays with fewer players for period } i \end{cases},$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2$. The values of x_{ij} reflect that when a team receives a red card, it is expected to negatively impact this team and positively benefit the opposing team. However, if no red card is issued

during the match, it should have no impact on either team. If a red card is issued during period i, the covariate x_{ij} will be set equal to the proportion of time that the team received the red card plays with fewer players.

Here we consider two different models regarding the home effect. The first model assumes a common effect for both home and away teams while the the second assumes different home and away effects. Hence the red card indicator is inserted into the linear predictor of Model 1 as follows:

$$\begin{aligned} \log\{E(T_{i1})\} &= \mu + home + att_{HT_i} + def_{AT_i} + \beta x_{ij} \\ \log\{E(T_{i2})\} &= \mu + att_{AT_i} + def_{HT_i} - \beta x_{ij} \end{aligned},$$

while in the second one β is substituted by β_1 and β_2 and hence the linear predictor is given by

$$\begin{aligned} \log\{E(T_{i1})\} &= \mu + home + att_{HT_i} + def_{AT_i} + \beta_1 x_{ij} \\ \log\{E(T_{i2})\} &= \mu + att_{AT_i} + def_{HT_i} - \beta_2 x_{ij} \end{aligned}.$$

Note the covariate (x_{ij}) is assigned a plus sign or a minus sign depending on whether the team that received the red card is the home team or the away team in order to capture the red card effect for both teams. For all red card coefficients, we have used a normal low information prior with a variance equal to 0.001. Posterior summaries for all red card effects are presented in Table 2.9 and Figure 2.3.

Models	Red Card Effect	Mean	Median	Sd	2.5%	97.5%
1	Common (β)	-0.259	-0.26	0.157	-0.571	0.039
2	Home (β_1)	-0.108	-0.117	0.204	-0.490	0.316
	Away (β_2)	-0.423	-0.421	0.209	-0.827	-0.021

Model 1: Common red card effect; Model 2: Separate red card effect for home and away teams

Table 2.9: Posterior summaries for the red card effects for the 2018-19 English Premier League under the Weibull fixed effects model.

For Model 1 (common red card effect), the negative posterior mean of this effect suggests that its presence decreases the goal arrival time for the team with the extra player. This is consistent with the intuition that the red card will benefit the team by playing with an extra player. However, this effect does not seem to be of importance since the 95% posterior interval includes the value of zero which corresponds to a no-card effect. Regarding Model 2, which assumes different red card effects for the home and away teams, the picture is slightly different since, although the home red card effect still seems to be of no importance, the posterior distribution for the away team suggests that the red card effect is away from zero having a major impact in their scoring times. Finally, the DIC for both models (common β and separate β's) are the same and equal to the corresponding value of the Weibull fixed effects model.

2.5.3 Half-time censoring

A meaningful alternative to the data setup presented in Table 2.1 can be constructed assuming that the goal arrival times are not only bi-variately censored at the end of each match but also at half-time.

In practical terms, the alternative version of Table 2.1 can be constructed as:

	Scoring	Time	Home	End-of-half		Goal Arrival Times		Censoring Times	
	Time	Interval	Team	Censoring		Home	Away	Home	Away
Game	$\mathcal{T}_i^*$	$\mathcal{T}_i$	$\mathcal{H}_i$	$\mathcal{C}_i$		t_{i1}	t_{i2}	c_{i1}	c_{i2}
1	2	2	1	0		2	NA	0	2
1	NA	43	NA	1	$\Rightarrow$	NA	NA	43	43
1	82	37	1	0		37	NA	0	37
1	91	9	0	0		NA	9	9	0
1	NA	4	NA	0		NA	NA	4	4

Table 2.10: Example of data layout assuming that both goal arrival times are censored at half (regular) time (45 minutes) for survival modeling: Data refers to the first match of English Premier League 2018-2019 (Manchester United – Leicester).

Table 2.10 differentiates from Table 2.1 only in the sense that there is now an extra line in the data because of which time resets one additional time at the $45th$ minute of the regular time.

The validity of Table 2.1 and Table 2.10 as data modeling structures, was tested by fitting the same fixed effects model for which the results were presented in Table 2.4 and comparing the DIC values, and it was found that the data fixed effects Weibull model under Table 2.10 received a DIC value of 11070 which is higher than the DIC value for the same model but under the data structure of Table 2.1. This implies that

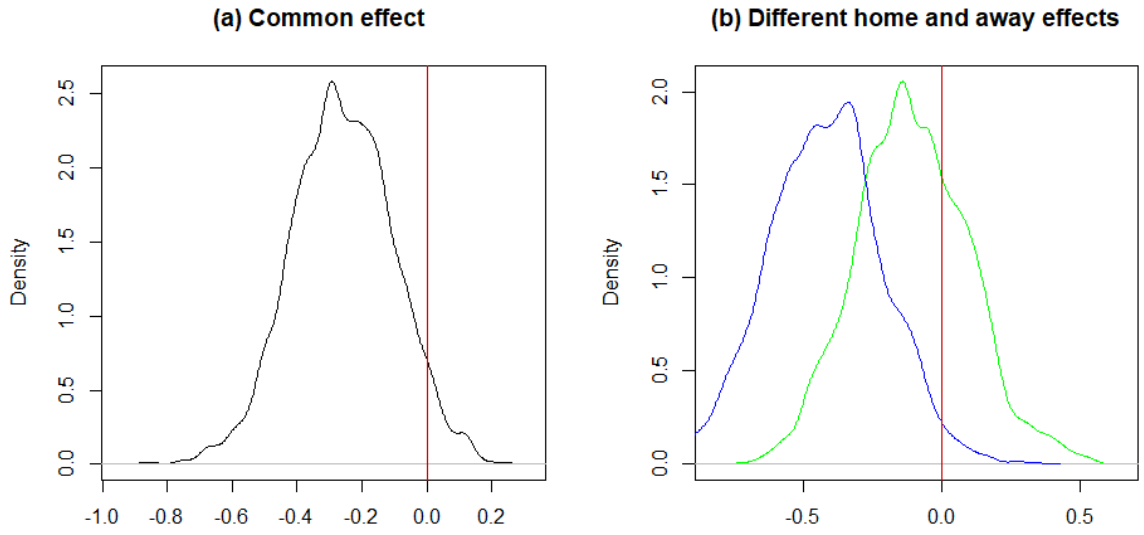

Figure 2.3: Posterior densities for the red card coefficients for the models (a) with common effect (b) different home and away effects; the red vertical line at zero refers to no red card effect; MCMC details: 10000 iterations, 1000 burn-in.

the model fitted using Table 2.1 is preferable to the model using the data of Table 2.10.

2.5.4 Assessing the goal scoring rate

Although the main focus of this chapter is on the expected goal arrival times, in this section, following a comment by a referee, we study and compare two additional models with an alternative structure on the scoring rate parameter γ. The first model considers different rate parameters γ for the home and away teams, while the second model assumes different shape parameters γ (or goal arrival rates) based on which team is ahead in the score. In this latter model, we use three parameters for γ: one for the leading team, one for when the game is tied, and one for the trailing team. Both of these models are compared with the standard model presented earlier in this chapter, which assumes the same γ for both opponents.

Table 2.11 presents the estimated shape parameters γ which captures the goal scoring rate for the three different models under comparison. From this table, it is clear that the model assuming different goal-scoring rates for home and away teams is not notably different than the standard model with a common γ (DIC equal to 11010 for both models). Additionally, the posterior distributions for the home and away γ are quite similar and they fall within a similar range as the common γ parameter of the standard model. The 95% posterior interval for the difference $\gamma_1 - \gamma_2$ ranges from -0.071 to 0.132, with the value of zero located quite close to the centre of the posterior distribution. This result also indicates that this difference between home and away γ parameters is negligible and can be safely assumed to be equal to zero.

On the other hand, Model 2 shows an improved DIC value compared to the standard model with a common γ (10870 vs. 11010), indicating that it better captures the overall rate of goal arrival times. Examining the posterior distributions (see Table 2.11), we observe that the posterior distributions of γ for the leading team, or when both teams are tied, are centered around one, suggesting a constant rate of goal arrival. However, for the trailing team, the posterior distribution of γ is clearly higher than one, confirming the findings of Silva and Swartz (2016), which reported that the trailing team is more likely to score the next goal.

2.6 Discussion

With the present chapter, we are introducing the Weibull fixed effects model to the literature for modeling the gap scoring times of soccer data under a Bayesian survival analysis framework. The existence of the censored observations in the data is handled by the proposed structure of the data in order to be able to model them using the predictive distribution. The predictive distribution was also used for converting the goal arrival goal arrival times into actual goals under a first proposed simple algorithm based on discriminating between goal arrival times. The model proposed appears to provide a very satisfying fit to the data since it appears to be able to replicate the final ranking of the teams in a precise manner. Additionally, the model appears to be performing exceptionally well in both in-sample and out-of-sample predictions especially considering its' simplicity and the fact that it requires minimal assumptions with regards to the scoring goal arrival times

Model	Shape parameters	Posterior Mean	Posterior Median	Posterior Sd	Credible Intervals 2.5%	97.5%	DIC
0	Common γ	1.131	1.130	0.028	1.078	1.186	11010
1	Home γ_1	1.145	1.145	0.036	1.074	1.216	11010
	Away γ_2	1.115	1.115	0.041	1.035	1.198	
2	Leading team (γ_1)	0.943	0.943	0.029	0.888	1.003	10870
	No team ahead (γ_2)	1.044	1.042	0.050	0.951	1.145	
	Trailing team (γ_3)	1.731	1.727	0.076	1.588	1.889	

Model 0: Standard model with common γ; Model 1: Model with different γ for home and away teams; Model 2: Model with different γ depending on which team is leading and which team is behind in score

Table 2.11: Trailing team inducing variant scoring rate case.

that it models.

Even though the finally proposed model appears to have an overwhelming fit and predictive ability, further extensions like simultaneous modeling of the goal arrival goal arrival times with the use of a bivariate distribution that captures both positive and negative correlations between the scoring times may lead to further predictive ability increase. However, assessing both the bivariate Marshall Olkin model and the random effects models, did not yield to a satisfactory model fit.

In our proposed modeling approach, the total number of goals (and intervals/events) n is unknown prior to observing the game. Consequently, our analysis is conditional on the observed number of goals. This approach is used for practical reasons in order to apply a simplified modeling approach. The implications of this assumption should be further investigated in future work by considering more sophisticated modeling approaches. One potential approach could involve extending the model by introducing an additional hierarchical layer; however, this would necessitate a considerably more sophisticated method for modeling the survival times. Given the complexity of this task, we have not pursued this issue further in this work.

Although in this chapter we have worked with the use of low-information prior distributions, resulting in an analysis of an "objective" fashion where Bayesian inference relies solely on data, in football, sports fans and experts often possess good knowledge of the game. This expertise can be incorporated into informative prior distributions, potentially improving the model's precision. Nevertheless, eliciting such prior beliefs poses a challenge, as it requires transforming common knowledge into meaningful mathematical prior distributions. For instance, while experts might easily rank the relative strength of teams, it would be considerably more difficult for them to express this knowledge in terms of scoring times in football matches. Alternatively, informative prior distributions might be constructed through the use of historical data and possibly the formulation of power priors (e.g. Ibrahim and Chen, 2000). For example, Egidi et al. (2018) explored incorporating historical data into modeling by representing the scoring rates of each team as convex combinations of parameters estimated from betting odds and historical data.

Finally, further work is required for the incorporation of in-play covariates that add to the predictive power of any chosen model which can be achieved by applying Bayesian variable selection techniques (e.g. Dellaportas et al., 2002).

Chapter 3

Bayesian Variable Selection using g-prior for Weibull Models with censoring

3.1 Introduction

The idea of the g-prior as a distribution to characterize the regression coefficients, can be extended to GLM models. Since the g-prior can be derived based on the Hessian matrix, the form of the likelihood plays a crucial role in the final form of the prior. Extending the previously mentioned application and theory on how to derive it for a well established survival model (the log-Normal distribution) (Castellanos et al., 2021) we proceed to derive all the relative formulas for the Weibull accelerated failure time model.

The main difference between the two approaches is the use of different Hessian matrix that can include non normal quantities that lead to issues with divergence or overflow/underflow since crucial integrals do not have a closed forms leading to the need for approximation methods. In addition to that, essential quantities for Bayesian variable selection (BVS) such as the marginal likelihood and hence the Bayes factors (BF) do not have a closed form leading to computational issues which can be handled by approximations.

All the above differences that stem from the non-normal likelihood require deep understanding of the behavior of the prior in both extreme and regular situations in terms of how heavy the censoring is and how it affects the posterior inclusion probabilities. Castellanos et al. (2021) open the discussion on how to handle the problem of variable selection under a log-normal AFT model. In what follows we contribute to this discussion and hence to the relevant literature by exploring how a prior can be constructed when the log-response is not normal by handling the entailed difficulties.

3.2 Likelihood under censoring

Assume that individual i has observed survival time y_i that is generated from a distribution with probability density function $f_Y(y)$. The vector observed $\boldsymbol{y} = (y_1, ..., y_n)$ is observed under the condition all $y_i \leq c_i$ where c_i is the i-th censoring time. Hence the response vector is observed based on the rule that the observation y_i is less than the value c_i. Otherwise instead of the value y_i, we observe c_i for $i = 1, 2, \ldots n$.

In survival analysis, the likelihood function is defined by considering both the probability density function of the model and its survivor function. This contribution of each of those two quantities depend whether each subjects' response is non-censored or censored. To be more specific, assuming that all model parameters are expressed by the $\boldsymbol{\theta}$ parameter, and the responses are expresses by $\boldsymbol{y}$, the general form of the likelihood is explicitly expressed by:

$$L(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{\delta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})^{\delta_i} S(c_i|\boldsymbol{\theta})^{1-\delta_i}$$

where $S()$ is the survivor function of the model considered, c_i are the log-censored times and $\boldsymbol{\delta}$ is the vector containing indicators defined as:

$$\delta_i = \begin{cases} 1, & \text{if} \quad y_i \leq c_i \\ 0, & \text{if} \quad y_i > c_i \end{cases}$$

3.2.1 Incorporation of covariates

Let $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ be the number of potential covariates that when used in their entirety will consist the full (saturated) model. Suppose that $\mathcal{M}_{\text{FULL}}$ is the full model while $\mathcal{M}_{NULL}$ is the null model containing only the intercept.

These model structures can be expressed as:

$$\mathcal{M}_{FULL} : Y_i = \log(T_i) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i + \sigma\epsilon_i, \quad \epsilon_i \sim Gumbel(0, 1)$$

and

$$\mathcal{M}_{NULL} : Y_i = \log(T_i) = \beta_0 + \sigma\epsilon_i, \quad \epsilon_i \sim Gumbel(0, 1)$$

Note that the intercept β_0 is common for all models and since it is generally assumed that the covariates are centered, the interpretation of this common parameter is consistent throughout.

3.3 The Weibull AFT model

Let T follow a Weibull distribution with shape parameter $k > 0$ and scale parameter $\lambda > 0$. The probability density function (PDF) of T is defined as follows:

$$f_T(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} e^{-\left(\frac{t}{\lambda}\right)^k}$$

In general, the Weibull AFT model is defined as follows:

$$Y_i = \log(T_i) = \mu_i + \sigma\epsilon, \quad \epsilon \sim Gumbel(0,1)$$

where $\mu_i = \log(\lambda_i)$ and $\sigma = \frac{1}{k}$.

We will be working with Y_i instead of directly modeling T_i allowing the support of the response to be real valued and for computational reasons. To be more specific, the pdf of Y is derived to be

$$f_Y(y) = f_T(e^y)\left|\frac{dt}{dy}\right| = \frac{k}{\lambda}\left(\frac{e^y}{\lambda}\right)^{k-1} e^{-\left(\frac{e^y}{\lambda}\right)^k} e^y = \frac{k}{\lambda^k} e^{yk} e^{-\left(\frac{e^y}{\lambda}\right)^k} \quad -\infty < y < \infty, \quad k > 0, \quad \lambda > 0$$

Now incorporating $i = 1, 2, ..., n$ observations, the likelihood under this transformed model assuming no censoring is written as:

$$L(k, \lambda_1, \ldots, \lambda_n|\boldsymbol{y}) = \prod_{i=1}^{n} \frac{k}{\lambda_i^k} e^{y_i k} e^{-\left(\frac{e^{y_i}}{\lambda_i}\right)^k}$$

Hence the log-likelihood takes the following form:

$$\log L(k, \lambda_1, \ldots, \lambda_n|\boldsymbol{y}) = \sum_{i=1}^{n} \log\left(\frac{k}{\lambda_i^k} e^{y_i k} e^{-\left(\frac{e^{y_i}}{\lambda_i}\right)^k}\right) = \sum_{i=1}^{n}\left\{\log(k) - k\log(\lambda_i) + y_i k - \left(\frac{e^{y_i}}{\lambda_i}\right)^k\right\}$$

Now substituting $k = \frac{1}{\sigma}$ and $\lambda_i = e^{\mu_i}$ the log-likelihood takes the form:

$$\log L(k, \lambda_1, \ldots, \lambda_n|\boldsymbol{y}) = \log L(\sigma, \mu_1, \ldots, \mu_n|\boldsymbol{y}) = \sum_{i=1}^{n} l(\sigma, \mu_i|\boldsymbol{y})$$

with

$$l(\sigma, \mu_i|\boldsymbol{y}) = \log L(\sigma, \mu_i|\boldsymbol{y}) = \log\left(\frac{1}{\sigma}\right) - \frac{1}{\sigma}\mu_i + y_i\frac{1}{\sigma} - \left(\frac{e^{y_i}}{e^{\mu_i}}\right)^{\frac{1}{\sigma}} = -\log(\sigma) + \frac{y_i - \mu_i}{\sigma} - e^{\frac{y_i - \mu_i}{\sigma}}$$

Now letting $z_i = \frac{y_i - \mu_i}{\sigma}$ log-likelihood becomes:

$$l(\sigma, \mu_i|\boldsymbol{z}) = -\log(\sigma) + z_i - e^{z_i}$$

In order to incorporate censoring into the likelihood, the survivor function of Y has to be taken into account and hence it is calculated through the CDF as follows:

$$S_Y(y) = P(Y > y) = 1 - P(Y \leq y) = 1 - P(\log(T) \leq y) = 1 - P(T \leq e^y) = 1 - F_T(e^y)$$

Now since the cdf is given by

$$F_T(t) = 1 - e^{-\left(\frac{t}{\lambda_i}\right)^k}$$

The survivor function becomes:

$$S_Y(y_i) = 1 - F_T(e^{y_i}) = 1 - 1 + e^{-\left(\frac{e^{y_i}}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}} = e^{-e^{\frac{y_i - \mu_i}{\sigma}}} = e^{-e^{z_i}}$$

Now, using all the above derived quantities, and since the survival likelihood requires both $L(k, \lambda_i|\boldsymbol{y})$ and the survivor function $S(y)$, it takes the following form:

$$L_S(k, \lambda_i) = \prod_{i=1}^{n} L(k, \lambda_i|\boldsymbol{y})^{\delta_i} S(c_i)^{1-\delta_i} \quad \forall\, i = 1, 2, ..., n$$

Now since the total number of observations is n contains both the censored and the uncensored observations, let n_c and n_u be the number of censored and uncensored observations respectively. It has been established from the previous arguments that if an observation is not censored ($\delta_i = 1$), then its contribution comes from the probability density function $f(y|k, \lambda)$ while if an observation is censored ($\delta_i = 0$), the the contribution comes from the survivor function.

Hence the model likelihood can be written in terms of the number of censored and uncensored observations assuming that the first n_u observations are uncensored and then rest $n - n_u$ are censored as follows:

$$\mathcal{L} = \mathcal{L}_u \times \mathcal{L}_c \text{ with } \mathcal{L}_u = \prod_{i \in \mathcal{U}} f(y_u|k, \lambda_i) \text{ and } \mathcal{L}_c = \prod_{i \in \overline{\mathcal{U}}} S(c_i)$$

where $\mathcal{U} = \{i : \delta_i = 1\}$.
Hence the total likelihood $\mathcal{L}$ can be splitted into the "uncensored" likelihood $\mathcal{L}_u$ and the "censored" likelihood $\mathcal{L}_c$.

Using these results, the total log-likelihood can be written as:

$$l = \log \mathcal{L} = l_u + l_c,$$

where $l_u = \log \mathcal{L}_u$ and $l_c = \log \mathcal{L}_c$
Therefore, the total log-likelihood takes the following form:

$$\log(\mathcal{L}) = \sum_{i=1}^{n_u} l(\sigma, \mu_i|\boldsymbol{y}) + \sum_{i=n_u+1}^{n} \log S(c_i)$$

Now let $l_i^u = l(\sigma, \mu_i|\boldsymbol{y})$ be the i observational component of the "uncensored" part of the likelihood and let $l_i^c = \log S(c_i)$ be the coresponding component of the "censored" part of the likelihood.

Hence, using the quantities that we derived from above,

$$l_i^u = l(\sigma, \mu_i|\boldsymbol{y}) = -\log(\sigma) + z_i - e^{z_i} \text{ and } l_i^c = -e^{z_i}$$

These will be the quantities that we will be working with to derive the Hessian matrix that will be our guide to derive the weights necessary for constructing prior on $\boldsymbol{\beta}$.

3.4 Derivation of weights

The key idea for deriving the weights is that the total likelihood can be splitted into two parts diversifying this way the contribution of censored and uncensored observations. Hence, we will calculate the required

quantities (elements) that define the W matrix presented in Section 3.2 in order to construct a covariance matrix that will take into account the data contributions coming from uncensored and censored observations.

3.4.1 Partial derivatives (Uncensored Case)

The partial derivative with respect to β_j is:

$$\begin{aligned}\frac{\partial l_i^u}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j}\left[-\log(\sigma) + z_i - e^{z_i}\right] \\ &= \frac{\partial}{\partial \beta_j}\left[\frac{y_i - \beta_0 - \boldsymbol{\beta^T} x_i}{\sigma} - e^{\frac{y_i-\beta_0-\boldsymbol{\beta^T} x_i}{\sigma}}\right] \\ &= -\frac{1}{\sigma}x_{i,j} + e^{\frac{y_i-\beta_0-\boldsymbol{\beta}^T x_i}{\sigma}}\frac{1}{\sigma}x_{i,j}\end{aligned}$$

Now the partial derivative with respect to β_j and β_k is:

$$\begin{aligned}\frac{\partial^2 l_i^u}{\partial \beta_k \partial \beta_j} &= \frac{\partial}{\partial \beta_k}\left[\frac{\partial l_i^u}{\partial \beta_j}\right] \\ &= \frac{\partial}{\partial \beta_k}\left[-\frac{1}{\sigma}x_{i,j} + e^{\frac{y_i-\beta_0-\boldsymbol{\beta}^T x_i}{\sigma}}\frac{1}{\sigma}x_{i,j}\right] \\ &= -\frac{1}{\sigma^2}x_{i,j}x_{i,k}e^{\frac{y_i-\beta_0-\beta^T x_i}{\sigma}}\end{aligned}$$

Hence, we obtain

$$\boxed{\frac{\partial^2 l_i^u}{\partial \beta_k \partial \beta_j} = -\frac{1}{\sigma^2}x_{i,j}x_{i,k}e^{z_{iu}}} \tag{3.1}$$

3.4.2 Partial derivatives (Censored Case)

With regards to the derivative of the likelihood contribution of a censored observation, we have the following:

$$\frac{\partial l_i^c}{\partial \beta_j} = \frac{\partial}{\partial \beta_j}\left[-e^{z_i}\right] = \frac{1}{\sigma}e^{\frac{c_i-\beta_0-\beta^T \mathbf{x}_i}{\sigma}}$$

Hence, the second derivative is given by:

$$\frac{\partial^2 l_i^c}{\partial \beta_k \partial \beta_j} = \frac{\partial}{\partial \beta_k}\left[\frac{1}{\sigma}e^{\frac{c_i-\beta_0-\beta^T \mathbf{x}_i}{\sigma}}\right] = -\frac{1}{\sigma^2}x_{i,j}x_{i,k}e^{\frac{c_i-\beta_0-\beta^T \mathbf{x}_i}{\sigma}}$$

Therefore, we obtain

$$\boxed{\frac{\partial^2 l_i^c}{\partial \beta_k \partial \beta_j} = -\frac{1}{\sigma^2}x_{i,j}x_{i,k}e^{z_{ic}}} \tag{3.2}$$

3.4.3 Fisher Information Matrix (Uncensored Case)

The Fisher information matrix requires us to calculate the expectation of minus the double partial derivative that we have just derived while also calculating the probability of being uncensored.

To be more specific, the quantity of interest that need to be calculated is:

$$\mathcal{I}_{j,k} = E\left(-\frac{\partial^2 l_i^u}{\partial\beta_k\partial\beta_j}\right)P(\delta_i=1) + E\left(-\frac{\partial^2 l_i^c}{\partial\beta_k\partial\beta_j}\right)P(\delta_i=0)$$

In particular, the quantity that has to be calculated is the following:

$$E\left(-\frac{\partial^2 l_i^u}{\partial\beta_k\partial\beta_j}\right)P(\delta_i=1)$$

Here, $P(\delta_i=1) = P(Y_i \le c_i) = F(c_i) = 1 - S(c_i) = 1 - e^{-e^{z_{ic}}}$.

As for the expectation term, we have the following:

$$E\left(-\frac{\partial^2 l_i^u}{\partial\beta_k\partial\beta_j}\middle| Y_i \le c_i\right) = E\left(-\left[-\frac{1}{\sigma^2}x_{i,j}x_{i,k}e^{\frac{Y_i-\beta_0-\beta^T x_i}{\sigma}}\right]\middle| Y_i \le c_i\right) = \frac{1}{\sigma^2}x_{i,j}x_{i,k}E\left[e^{\frac{Y_i-\beta_0-\beta^T x_i}{\sigma}}\middle| Y_i \le c_i\right]$$

Now in order to calculate the expectation required we follow the procedure below:

$$E\left[e^{\frac{Y_i-\beta_0-\beta^T x_i}{\sigma}}\middle| Y_i \le c_i\right] = E\left[e^{\frac{Y_i-\mu_i}{\sigma}}\middle| Y_i \le c_i\right] = E\left[T_i^{\frac{1}{\sigma}}e^{-\frac{\mu_i}{\sigma}}\middle| T_i \le c_i^*\right]$$

where $c_i^* = e^{c_i}$.

Now, in order to compute this conditional expectation, we first have to calculate the conditional pdf, namely

$$f_{T_i|T_i\le c_i^*}(t) = \frac{f_{T_i}(t)}{P(T_i \le c_i^*)} = \frac{f_{T_i(t)}}{F_{T_i}(c_i^*)} = \frac{\frac{1}{\sigma e^{\mu_i}}\left(\frac{t}{e^{\mu_i}}\right)^{\frac{1}{\sigma}-1}e^{-\left(\frac{t}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}}{1-e^{-\left(\frac{c_i^*}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}}$$

Hence, the conditional expectation is given by

$$E\left[T_i^{\frac{1}{\sigma}}e^{-\frac{\mu_i}{\sigma}}\middle| T_i \le c_i^*\right] = E\left[T^{\frac{1}{\sigma}}e^{-\frac{\mu_i}{\sigma}}\middle| T \le c_i^*\right] = \frac{1}{1-e^{-\left(\frac{c_i^*}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}}\int_0^{c_i^*}\left(\frac{t}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}\frac{\frac{1}{\sigma}}{e^{\mu_i}}\left(\frac{t}{e^{\mu_i}}\right)^{\frac{1}{\sigma}-1}e^{-\left(\frac{t}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}dt$$

By setting $u = \left(\frac{t}{e^{\mu_i}}\right)^{\frac{1}{\sigma}} \Rightarrow du = \left(\frac{1}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}\frac{1}{\sigma}t^{\frac{1}{\sigma}-1}dt$, we obtain

$$E\left[T^{\frac{1}{\sigma}}e^{-\frac{\mu_i}{\sigma}}\middle| T \le c_i^*\right] = \frac{1}{1-e^{-\left(\frac{c_i^*}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}}\int_0^{\left(\frac{c_i^*}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}} ue^{-u}du$$

$$= \frac{1}{1-e^{-\left(\frac{c_i^*}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}}\left(-\left(\frac{c_i^*}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}e^{-\left(\frac{c_i^*}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}} - e^{-\left(\frac{c_i^*}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}} + 1\right)$$

$$= 1 - \frac{\left(\frac{c_i^*}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}e^{-\left(\frac{c_i^*}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}}{1-e^{-\left(\frac{c_i^*}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}}$$

Thus, the conditional expectation of the second derivative of the log-likelihood given that it is uncensored is given by

$$E\left[-\frac{\partial l_i}{\partial\beta_k\partial\beta_j}\middle|\delta_i=1\right]=E\left[-\frac{\partial^2 l_i}{\partial\beta_k\partial\beta_j}\right]=\frac{1}{\sigma^2}x_{i,j}x_{i,k}\left(1-\frac{\left(\frac{c_i^*}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}e^{-\left(\frac{c_i^*}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}}{1-e^{-\left(\frac{c_i^*}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}}\right)=\frac{1}{\sigma^2}x_{i,j}x_{i,k}\left(1-\frac{\left(\frac{e^{c_i}}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}e^{-\left(\frac{e^{c_i}}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}}{1-e^{-\left(\frac{e^{c_i}}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}}\right)$$

While when multiplied by, $P(\delta_i=1)=1-e^{-\left(\frac{e^{c_i}}{e^{\mu}}\right)^{\frac{1}{\sigma}}}$

$$\begin{aligned}E\left[-\frac{\partial^2 l_{ui}}{\partial\beta_k\partial\beta_j}\right]P(\delta_i=1)&=\frac{1}{\sigma^2}x_{i,j}x_{i,k}\left(1-\frac{\left(\frac{e^{c_i}}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}e^{-\left(\frac{e^{c_i}}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}}{1-e^{-\left(\frac{e^{c_i}}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}}\right)\left(1-e^{-\left(\frac{e^{c_i}}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}\right)\\&=\frac{1}{\sigma^2}x_{i,j}x_{i,k}\left(1-e^{-\left(\frac{e^{c_i}}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}-\left(\frac{e^{c_i}}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}e^{-\left(\frac{e^{c_i}}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}\right)\end{aligned}$$

Now setting $z_i^c=\frac{c_i-\mu_i}{\sigma}$ we get the following:

$$E\left[-\frac{\partial^2 l_i^u}{\partial\beta_k\partial\beta_j}\right]P(\delta_i=1)=\frac{1}{\sigma^2}x_{i,j}x_{i,k}\left(1-e^{-e^{z_i^c}}-e^{z_i^c-e^{z_i^c}}\right)$$

Similarly, for the censored case and assuming that, c_i are fixed and known a priori for all $i=1,2,\dots n,$,

$$E\left[-\frac{\partial^2 l_i^c}{\partial\beta_k\partial\beta_j}\right]P(\delta_i=0)=\frac{1}{\sigma^2}x_{i,j}x_{i,k}\left(\frac{e^{c_i}}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}e^{-\left(\frac{e^{c_i}}{e^{\mu_i}}\right)^{\frac{1}{\sigma}}}=\frac{1}{\sigma^2}x_{i,j}x_{i,k}\left(e^{z_i^c-e^{z_i^c}}\right)$$

Therefore, the j,k-th component of the Fisher information matrix that concerns $\boldsymbol{\beta}$ can be written as:

$$\mathcal{I}_{j,k}=\frac{1}{\sigma^2}x_{i,j}x_{i,k}\left[1-e^{-e^{z_i^c}}-e^{z_i^c-e^{z_i^c}}+e^{z_i^c-e^{z_i^c}}\right]=\frac{1}{\sigma^2}x_{i,j}x_{i,k}\left[1-e^{-e^{z_i^c}}\right]$$

Finally,

$$\boxed{\mathcal{I}_{j,k}=\frac{1}{\sigma^2}x_{i,j}x_{i,k}\left(1-e^{-e^{z_i^c}}\right)}$$

Now, using the components of the expected Fisher information matrix, we move on to deriving the elements of the diagonal of the weight matrix W. Note that the expected Fisher information matrix concerning $(\beta_0,\boldsymbol{\beta})$ is the following:

$$\mathcal{I}=\frac{1}{\sigma^2}\begin{pmatrix}\mathbf{1}^{\boldsymbol{T}}W(\beta_0,\boldsymbol{\beta},\sigma)\mathbf{1} & \mathbf{1}^{\boldsymbol{T}}W(\beta_0,\boldsymbol{\beta},\sigma)\boldsymbol{X}\\ \boldsymbol{X}^{\boldsymbol{T}}W(\beta_0,\boldsymbol{\beta},\sigma)\mathbf{1} & \boldsymbol{X}^{\boldsymbol{T}}W(\beta_0,\boldsymbol{\beta},\sigma)\boldsymbol{X}\end{pmatrix}$$

Therefore, after all the relative derivations, the weight matrix that we are proposing, is the explicitly defined as:

$$\boxed{W(\beta_0,\boldsymbol{\beta},\sigma)=Diag\left(1-e^{-e^{z_i}}\right),\quad\forall i=1,2,...,n}$$

where Diag(ξ_i) is a diagonal matrix with ξ_i is the being the its diagonal elements.

3.5 The proposed prior for $\boldsymbol{\beta}$

To complete the definition of the proposed Normal prior, we have to formally define the covariance matrix of the prior. To derive that covariance matrix in our case, we first present the block of the inverse of Fisher information matrix that corresponds to $\boldsymbol{\beta}$ and that is:

$$\mathcal{I}_{\boldsymbol{\beta}}^{-1} = \sigma^2 \left(\boldsymbol{X}^T \left\{ W(\beta_0, \boldsymbol{\beta}, \sigma) - W(\beta_0, \boldsymbol{\beta}, \sigma) \frac{11^T}{tr(W(\beta_0, \boldsymbol{\beta}, \sigma))} W(\beta_0, \boldsymbol{\beta}, \sigma) \right\} \boldsymbol{X} \right)^{-1}$$

Since this quantity depends on $\boldsymbol{\beta}$ we cannot properly use this matrix as a prior variance covariance matrix.

There are two possible directions:

- Use the MLEs in the $\mathcal{I}_\beta$ making this prior empirical.
- Exploit the information of the two models under comparison.

Specifically, we follow the local prior approach where our prior is centered around the parameter values under the null model/hypothesis (that is $\boldsymbol{\beta} = 0$). Hence we may use the same approach also for the specification of the prior variance-covariance matrix and use the approximate variance under the parameter values of the null model. Hence, we plug in the values of $\boldsymbol{\beta} = 0$.

3.5.1 Effective Sample Size

Here, we follow the approach of using a unit-type information prior. Since the censored data are incomplete observations we need to use the "effective" sample size to be used as the value of g. Hence, the effective sample size borrowing information under the null that we propose is:

$$\boxed{n_e = \sum_{i=1}^{n} \left(1 - e^{-e^{z_{i0}}}\right) = n_u + \sum_{i \in \overline{\mathcal{U}}} (1 - e^{z_{i0}})}$$

We define the effective sample size by correctly accounting for the contribution of the censored observations. This effective sample size will allow us to adjust for the censored observations, it will be intuitively less than or equal the total number of observations n depending on how heavy the censoring is.

where $z_{i0} = \frac{c_i - \beta_0}{\sigma}$.
Note that the quantity inside the summation is exactly the CDF of the Y evaluated at c_i assuming that $\boldsymbol{\beta} = \mathbf{0}$ meaning that the effective sample size can be written as a function of the CDF of Y that is:

$$n_e = \sum_{i=1}^{n} F_Y(c_i) = n_u + \sum_{i \in \overline{\mathcal{U}}} F_Y(c_i)$$

which is the sum of the probability of each individual i being uncensored. This finding will be useful both in terms of interpretation and in terms of desired properties that the effective sample size should have for variant censoring intensity.

3.5.2 Properties of the effective sample size

Focusing on what is included in the summation of the effective sample size, namely the quantity

$$w_i = 1 - e^{-e^{\frac{c_i - \beta_0}{\sigma}}} \tag{3.3}$$

it would be useful to see how this quantity behaves under extreme censoring and when censoring is absent.

Weights under no censoring

We express no censoring as $c_i \to \infty$ meaning that the censoring time tends to infinity and hence we expect approximately all observations to be uncensored since the time of censoring tends to infinity. Under this scenario, we have

$$\lim_{c_i \to \infty} w_i = \lim_{c_i \to \infty} \left(1 - e^{-e^{\frac{c_i - \beta_0}{\sigma}}}\right) = \left(1 - e^{-e^{\infty}}\right) = \left(1 - \frac{1}{e^{e^{\infty}}}\right) = (1 - 0) = 1$$

Weights under extreme censoring

On a similar note, to express extreme amount of censoring, we let c_i tend to $-\infty$ essentially leaving no time for any observation to be less than that value and hence leaving no room of uncensored observations. In this case we have:

$$\lim_{c_i \to -\infty} w_i = \lim_{c_i \to -\infty} \left(1 - e^{-e^{\frac{c_i - \beta_0}{\sigma}}}\right) = \left(1 - e^{-e^{-\infty}}\right) = \left(1 - \frac{1}{e^{e^{-\infty}}}\right) = (1 - 1) = 0$$

The behavior of w_i between extremes

3.5.3 w_i as a weight function

The property that w_i are weights can be proved by showing that:

- $0 \leq w_i \leq 1$ for all c_i.
- w_i is a non-decreasing function of c_i.

Both of these points are automatically proven due to the fact that w_i is the CDF of Y evaluated at c_i and hence it is both non-decreasing and has its range is $[0, 1]$ for all c_i.

Total effective sample size behavior under no censoring

Under no censoring ($c_i \to \infty$), the total effective sample size tends to become equal to the total number of observations (censored and uncensored) n, since:

$$\lim_{c_i \to \infty} n_{eff_{LW}} = \lim_{c_i \to \infty} \left(\sum_{i=1}^{n} \left\{ 1 - e^{-e^{\frac{c_i - \beta_0}{\sigma}}} \right\} \right) = \sum_{i=1}^{n} \left(\lim_{c_i \to \infty} \left\{ 1 - e^{-e^{\frac{c_i - \beta_0}{\sigma}}} \right\} \right) = \sum_{i=1}^{n} 1 = n$$

Total effective sample size behavior under extreme censoring

Similarly, under extreme censoring, $c_i \to -\infty$, the total effective sample size is negligible, since:

$$\lim_{c_i \to -\infty} n_{eff_{LW}} = \lim_{c_i \to -\infty} \left(\sum_{i=1}^{n} \left\{ 1 - e^{-e^{\frac{c_i - \beta_0}{\sigma}}} \right\} \right) = \sum_{i=1}^{n} \left(\lim_{c_i \to -\infty} \left\{ 1 - e^{-e^{\frac{c_i - \beta_0}{\sigma}}} \right\} \right) = \sum_{i=1}^{n} 0 = 0$$

Total effective sample size general behavior

Under all the above subsection relating to the effective sample size that we have derived, it is evident that

$$0 \leq n_{eff_{LW}} \leq n$$

and hence the effective sample size that we are proposing is a value that is less than or equal to the total number of observations n depending on the amount of censoring in the data.

3.5.4 The final form of the proposed covariance matrix Σ_{LW}

Under all the above derivations, our proposal for the covariance matrix of the prior of $\boldsymbol{\beta}$ has the following form:

$$\boxed{\Sigma_{LW} = \sigma^2 n_e \left[X^T \left\{ W(\beta_0, \boldsymbol{\beta} = 0, \sigma) - W(\beta_0, \boldsymbol{\beta} = 0, \sigma) \frac{1 1^T}{n_e} W(\beta_0, \boldsymbol{\beta} = 0, \sigma) \right\} X \right]^{-1}}$$

where

$$W(\beta_0, \boldsymbol{\beta} = 0, \sigma) = Diag(\boldsymbol{w})$$

with

$$\boldsymbol{w} = (w_1, w_2, \ldots, w_n)^T, \quad w_i = 1 - e^{-e^{z_{i0}}}, \quad z_{i0} = \frac{c_i - \beta_0}{\sigma}$$

and

$$n_e = \sum_{i=1}^{n} w_i = n_u + \sum_{i \in \overline{\mathcal{U}}} w_i$$

3.6 The final form of the proposed prior on $\boldsymbol{\beta}$

The final form of the prior that we are proposing on $\boldsymbol{\beta}$ of a model $\mathcal{M}_j$ is given below:

$$\boxed{\boldsymbol{\beta_j} | \mathcal{M}_j, \beta_0, \sigma \sim N_{p_j}(\boldsymbol{\beta_j}; 0, \Sigma_{LW_{\mathcal{M}_j}})}$$

where $\boldsymbol{\beta}_j$ is a sub-vector of $\boldsymbol{\beta}$ that concerns only the active covariates in model $\mathcal{M}_j$.
The covariance matrix $\Sigma_{\mathcal{M}_j}$ is defined as:

$$\Sigma_{LW_{\mathcal{M}_j}} = \sigma^2 n_e \left[X_j^T \left\{ W(\beta_0, \boldsymbol{\beta}=0, \sigma) - W(\beta_0, \boldsymbol{\beta}=0, \sigma) \frac{11^T}{n_e} W(\beta_0, \boldsymbol{\beta}=0, \sigma) \right\} X_j \right]^{-1}$$

with X_j being the sub-matrix of X including only the active columns of the X depending on model $\mathcal{M}_j$ and p_j is the number of covariates in model $\mathcal{M}_j$. Finally, $n_{eff_{LW}}$ will remain unaffected since it only depends on β_0, c_i and σ which are common among all models. Note that in our proposed prior g has been replaced with n_e. In order to make our prior to be a unit information prior.

3.7 The marginal likelihood $m(\boldsymbol{y})$

The marginal likelihood is a crucial quantity that is generally used to derive the Bayes Factors and hence derive the posterior model probabilities and hence the posterior inclusion probabilities that will be the tool that we will use for model selection. Unfortunately, the marginal likelihood is not always available in closed form and hence numerical methods have to be applied to simulate from it. In our case, the marginal likelihood under a model that is not the null model has the following form:

$$m(\boldsymbol{y}) = \int_{\mathbb{R}^{|}} \int_{-\infty}^{\infty} \int_0^{\infty} \left\{ \prod_{i=1}^{n} f(y_i|k, \beta_0, \boldsymbol{\beta})^{\delta_i} S(c_i)^{1-\delta_i} \right\} f_{N_p}\left(\boldsymbol{\beta}; 0, \sigma^2 n_e \Sigma_{LW}\right) \frac{1}{\sigma} d\sigma d\beta_0 d\boldsymbol{\beta}$$

while the marginal under the null model is given by:

$$m_0(\boldsymbol{y}) = \int_0^{\infty} \int_{-\infty}^{\infty} \left\{ \prod_{i=1}^{n} f(y_i|k, \beta_0, \boldsymbol{\beta})^{\delta_i} S(c_i)^{1-\delta_i} \right\} \frac{1}{\sigma} d\beta_0 d\sigma$$

We move to **proving that if not all y_i are equal and $n_u > 1$** then $0 < m_0(\boldsymbol{y}) < \infty$.
First notice that

$$m_0(\boldsymbol{y}) = \int_0^{\infty} \int_{-\infty}^{\infty} \left\{ \prod_{i=1}^{n} f(y_i|\sigma, \beta_0, \boldsymbol{\beta})^{\delta_i} S(c_i)^{1-\delta_i} \right\} \frac{1}{\sigma} d\beta_0 d\sigma \leq \int_0^{\infty} \int_{-\infty}^{\infty} \prod_{i \in \mathcal{U}} f(y_i|\sigma, \beta_0) \frac{1}{\sigma} d\beta_0 d\sigma$$

since $S(c_i)$ is a probability and hence it has a range of $[0, 1]$.
Now, if we prove that

$$\int_0^{\infty} \int_{-\infty}^{\infty} \prod_{i \in \mathcal{U}} f(y_i|\sigma, \beta_0) \frac{1}{\sigma} d\beta_0 d\sigma < \infty$$

then the marginal under the null is finite.
We have the following:

$$f(\boldsymbol{y_u}|\sigma, \beta_0) = \frac{1}{\sigma^{n_u}} exp\left\{ \sum_{i \in \mathcal{U}} \frac{y_i}{\sigma} \right\} exp\left\{ -\frac{n_u \beta_0}{\sigma} \right\} exp\left\{ -\exp\left\{ -\frac{\beta_0}{\sigma} \right\} \sum_{i \in \mathcal{U}} exp\left\{ \frac{y_i}{\sigma} \right\} \right\}$$

Hence,

$$\int_{-\infty}^{\infty} f(\boldsymbol{y_u}|\sigma, \beta_0) d\beta_0 = \int_{-\infty}^{\infty} \frac{1}{\sigma^{n_u}} exp\left\{ \sum_{i \in \mathcal{U}} \frac{y_i}{\sigma} \right\} exp\left\{ -\frac{n_u \beta_0}{\sigma} \right\} exp\left\{ -exp\left\{ -\frac{\beta_0}{\sigma} \right\} \sum_{i \in \mathcal{U}} exp\left\{ \frac{y_i}{\sigma} \right\} \right\} d\beta_0 \quad (3.4)$$

Now let $v = exp\left\{-\frac{\beta_0}{\sigma}\right\} \Rightarrow dv = -\frac{1}{\sigma}exp\left\{-\frac{\beta_0}{\sigma}\right\}d\beta_0 \Rightarrow d\beta_0 = -\sigma exp\left\{\frac{\beta_0}{\sigma}\right\}dv = -\frac{\sigma}{\nu}$

Then, for $\beta_0 \to -\infty \Rightarrow v \to \infty$ and $\beta_0 \to \infty \Rightarrow v \to 0$.

Therefore, the integral (3.4) becomes equal to

$$\begin{aligned}
\int_{-\infty}^{\infty} f(\boldsymbol{y_u}|\sigma,\beta_0)d\beta_0 &= \int_0^{\infty} \frac{1}{\sigma^{n_u}} exp\left\{\sum_{i\in\mathcal{U}}\frac{y_i}{\sigma}\right\} v^{n_{unc}} exp\left\{-v\left(\sum_{i\in\mathcal{U}} exp\left(\frac{y_i}{\sigma}\right)\right)\right\}\sigma\frac{1}{v}dv \\
&= \int_0^{\infty} \frac{\sigma}{\sigma^{n_u}} exp\left\{\sum_{i\in\mathcal{U}}\frac{y_i}{\sigma}\right\} v^{n_u-1} exp\left\{-v\left(\sum_{i\in\mathcal{U}} exp\left(\frac{y_i}{\sigma}\right)\right)\right\}dv \\
&= \frac{1}{\sigma^{n_u-1}} exp\left\{\sum_{i\in\mathcal{U}}\frac{y_i}{\sigma}\right\} \int_0^{\infty} v^{n_u-1} exp\left\{-v\left(\sum_{i\in\mathcal{U}} exp\left(\frac{y_i}{\sigma}\right)\right)\right\}dv \\
&= \frac{1}{\sigma^{n_u-1}} exp\left\{\sum_{i\in\mathcal{U}}\frac{y_i}{\sigma}\right\} \frac{\Gamma(n_u)}{\left[\sum_{i\in\mathcal{U}} exp\left(\frac{y_i}{\sigma}\right)\right]^{n_u}}
\end{aligned}$$

Since,

$$\sum_{i\in\mathcal{U}} \exp\left\{\frac{y_i}{\sigma}\right\} < n_u \exp\frac{y_{max}}{\sigma}$$

where n_{max} is the number of responses (y_i) that are equal to the maximum value of the response vector $\boldsymbol{y}_u$.

Therefore,

$$n_u e^{\frac{y_{min}}{\sigma}} \leq \sum_{i\in\mathcal{U}} e^{\frac{y_i}{\sigma}} \leq n_u e^{\frac{y_{max}}{\sigma}}$$

Hence,

$$\frac{e^{\sum_{i\in\mathcal{U}}\frac{y_i}{\sigma}}}{\left(e^{\frac{y_i}{\sigma}}\right)^{n_u}} \leq \frac{e^{\sum_{i\in\mathcal{U}}\frac{y_i}{\sigma}}}{\left(n_{min}e^{\frac{y_{min}}{\sigma}}\right)^{n_u}} = n_{min}^{-n_u} e^{-\frac{(n_u y_{min} - \sum_{i\in\mathcal{U}} y_i)}{\sigma}} = n_{min}^{-n_u} e^{\frac{\sum_{i\in\mathcal{U}} y_i - n_u y_{min}}{\sigma}}$$

Note that since not all y_i are equal by definition $n_u y_{min} - \sum_{i\in\mathcal{U}} y_i > 0$

Hence, the entire integral is bounded by:

$$\frac{\Gamma(n_u)}{\sigma^{n_u}} exp\{-\frac{n_u y_{min} - \sum_{i\in\mathcal{U}} y_i}{\sigma}\} \tag{3.5}$$

Hence, the integral converges since the above quantity is integrable with respect to σ.

Now, as $\sigma \to \infty$ we have the following:

$$exp\left\{\frac{\sum_{i\in\mathcal{U}} y_i}{\sigma}\right\} \approx 1 + \sum_{i\in\mathcal{U}}\frac{y_i}{\sigma} \Rightarrow \sum_{i\in\mathcal{U}} exp\left\{\frac{y_i}{\sigma}\right\} \approx \sum_{i\in\mathcal{U}}\left(1+\frac{y_i}{\sigma}\right)$$

Therefore, as $\sigma \to \infty$,

$$\frac{exp\left(\sum_{i\in\mathcal{U}}\frac{y_i}{\sigma}\right)}{\left[\sum_{i\in\mathcal{U}} exp\left(\frac{y_i}{\sigma}\right)\right]^{n_u}} \approx \frac{1+\frac{\sum_{i\in\mathcal{U}} y_i}{\sigma}}{\left(n_u + \frac{\sum_{i\in\mathcal{U}} y_i}{\sigma}\right)^{n_u}} \approx \frac{1}{n_u^{n_u}}$$

Therefore, for $\sigma \to \infty$ the integral behaves like

$$\sigma^{-n_u}\frac{1}{n_u^{n_u}}\Gamma(n_u)$$

which is integrable if $n_u > 1$.

Therefore, if $n_{unc} > 1$ and not all y_i are equal (not all log-survival times are tied), then the marginal under the null converges since

$$\int_0^\infty \int_{-\infty}^\infty \prod_{i \in \mathcal{U}} f(y_i|\sigma, \beta_0) \frac{1}{\sigma} d\beta_0 d\sigma$$

is convergent and

$$m_0(\boldsymbol{y}) \leq \int_0^\infty \int_{-\infty}^\infty \prod_{i \in \mathcal{U}} f(y_i|\sigma, \beta_0) \frac{1}{\sigma} d\beta_0 d\sigma$$

3.8 Laplace Approximation vs Bridge Sampling

The Laplace approximation, is an approximation of integrals and here the marginal distribution, that uses a Normal distribution as an approximation of the posterior. To be more specific, it expands the product of the likelihood and the prior (which is the integrand) around the posterior mode.

Bridge sampling, on the other hand, handles the approximation in different manner, since it samples first from the posterior distribution, and then from an "bridge" distribution, an example of which is the normal approximation of the posterior.

While both methods are generally widely used, there are differences in terms of accuracy. In particular, the Laplace approximation, even though its approximation is improving as the sample size increase, it has an error of approximation of order $\mathcal{O}(n^{-1})$. On the other hand, the accuracy of the bridge sampling is substantially improved due to the fact that it samples from two separate sources and hence it leads to better accuracy especially compared to methods that rely solely on an approximation based on Taylor expansions as the sample size increases.

The Laplace approximation method is generally less effective when the posterior cannot be approximated by a normal density. Another source of limitation of this method includes the case of multi-modality of the posterior or the mode of a parameter to be in the limits of the parameter space. Bridge sampling can also be less accurate when the posterior is multi-modal while also having the limitation of being significantly more time consuming.

We have used both methods for approximating the marginal under each possible model and compared them in terms of their actual estimate of the log-marginal. The relevant results can be found in Table 3.1. Both methods we applied in an example were the true parameters were $\boldsymbol{\beta} = (0.2, 0.8, 0)$, $\beta_0 = 7$ $n = 100$, $\sigma = 0.5$, and the censoring percentage was set to 40%.

From Table 3.1 it is evident that the Bridge Sampling method estimation of the log-marginal is significantly close to the Laplace approximation of the same quantity. This finding can be used as an argument for the using the Laplace approximation for computational speed reasons since it takes significantly less time to be applied while also being robust to the posterior. Since, and according to the findings of Table 3.1 we expect the same inference in terms of Bayesian Variable Selection, we moved on suggesting (and using) the Laplace approximation for all the relevant to Bayesian Variable Selection quantities especially due to computational efficiency. In particular the computational times for each of those methods are presented in

Table 3.1: Marginal likelihood estimation comparison between Laplace Approximation and Bridge Sampling. The true values of the parameters were: $\sigma = 0.4$, $\beta_0 = 7$, $n = 100$, $\boldsymbol{\beta} = (0.2, 0.8, 0)$ and the percentage of censoring observations was set to a typical 40%.

	Laplace	Bridge
Model 1	-172.828	-173.099
Model 2	-160.170	-160.332
Model 3	-177.018	-177.368
Model 4	-156.082	-156.545
Model 5	-174.439	-175.060
Model 6	-162.131	-162.652
Model 7	-158.136	-158.962
Null Model	-175.460	-175.478

the Table below.

Table 3.2: Computational time (in seconds) for three runs of each method.

Method	Min	Q1	Mean	Median	Q3	Max
Bridge	58.32	58.74	58.98	59.16	59.31	59.46
Laplace	7.78	7.79	7.94	8	8.02	8.25

The time difference is evident with the Laplace approximation being approximately 7 times faster.

3.9 Complete controlled example

In this section, we implement a simulation study by considering a varying number of parameters. The aim is to study the behaviour of the BVS based on our proposed prior for $\boldsymbol{\beta}$ and illustrate its ability to identify the correct true model via the posterior inclusion probabilities.
The true model (in the uncensored case) is mathematically presented in simple terms below:

$$y_i = \log(t_i) = \beta_0 + \boldsymbol{\beta}^T x_i + \sigma\epsilon_i, \quad \epsilon_i \sim Gumbel(0, 1)$$

for $i = 1, 2, ..., 70$ ($n = 70$) with $\boldsymbol{\beta} = (0.2, 0.25, 0.9, -0.15, 0, 0, 0, 0)$, $p = 8$, $\beta_0 = 7$, $\sigma = 0.5$ and the amount of censoring in the data, was defined to be 50% that represents medium size censoring. The size of vector $\boldsymbol{\beta}$ in this case implies a total of $2^8 = 256$ models under consideration for which the marginals were approximated

using the Laplace approximation, as well as the induced Bayes Factors and inclusion probabilities. The inclusion probabilities are presented in Table 3.3 which illustrates that our proposed prior captures the

Table 3.3: Posterior inclusion probabilities.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1.000	1.000	1.000	1.000	0.068	0.049	0.056	0.049

correct model considering that half of the observations are censored. While X_5 and X_7 are only marginally included in the model, the correct model is identified in a probability threshold of 7% meaning that under this controlled experiment and with a relatively large amount of models to be considered while still having a considerable amount of censoring the correct model is identified. This is a promising finding supporting our choice of the prior and our method in general while still requiring further simulations to fully justify it. This work has been done in the following sections.

3.10 The Censoring Effect

Investigating whether and how the amount of censoring affects the posterior inclusions probabilities under our proposed prior is important for a correctly specified variable selection method. Extreme cases of censoring (70% and upwards) are rare in practice and as expected can cause problems since the amount of missing information will dominate the inference and fully specified uncensored observations will be sparse.

However, to investigate how censoring affects our inclusion probabilities we ran a simulation for varying percentage of censoring.

Figure 3.1 shows that when the censoring percentage as the censoring percentage increases from 5% to 60% the posterior inclusions probabilities identify the correct model. The opposite is true when censoring starting from 60% up to 80% exists in the data. In that case the true model is not clearly identified. This can possibly be due to the relatively small total sample size. We will discuss and address this issue in a later section on Model Selection Consistency.

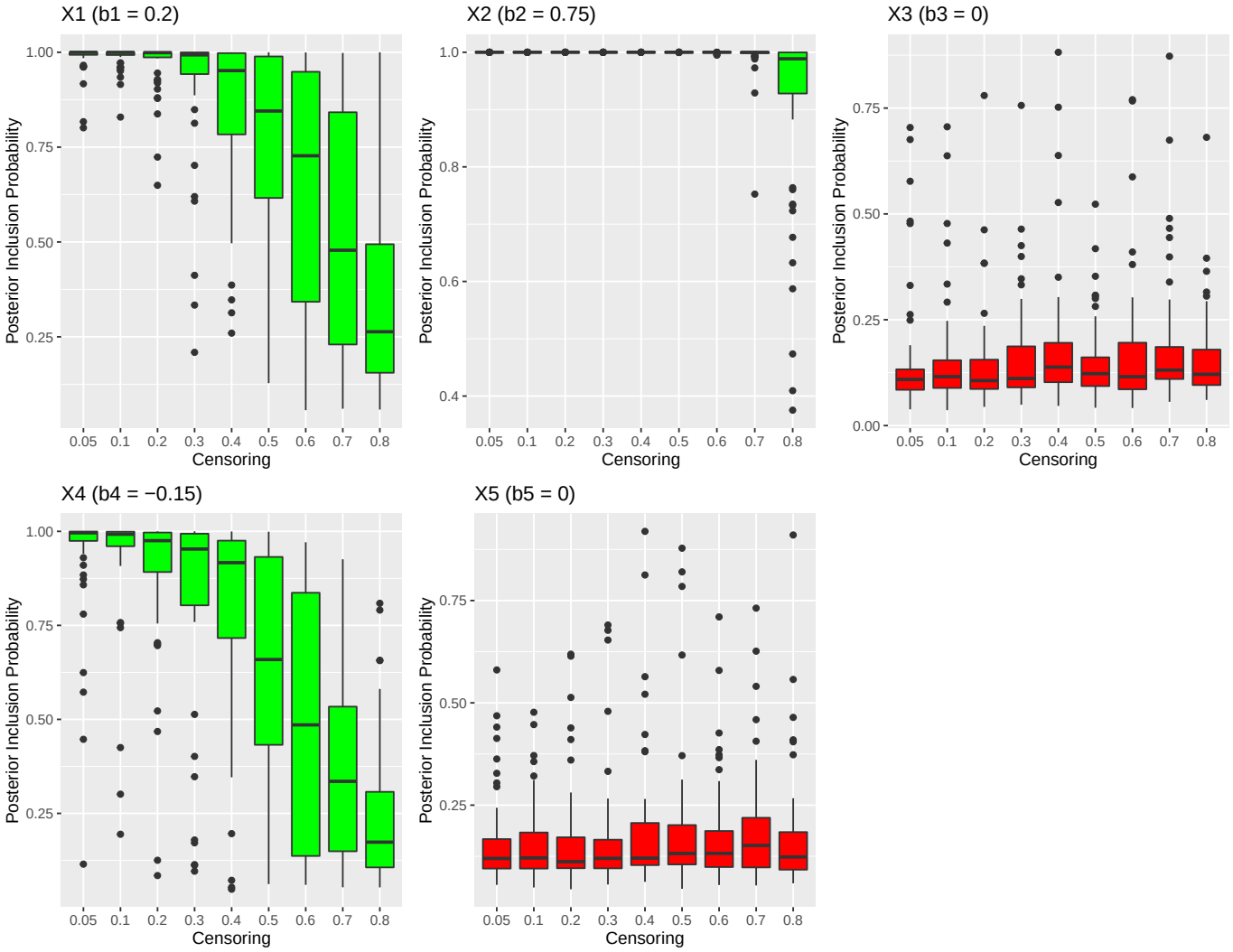

Figure 3.1: Posterior Inclusion Probabilities plotted against a varying censoring intensity for each of the covariates considered for 50 simulated datasets. True coefficient values $\boldsymbol{\beta} = (0.2, 0.75, 0, -0.15, 0)$, and true parameter values $\sigma = 0.8$, $\beta_0 = 7$, $n = 70$

3.11 The Sample Size Effect

In this section we assess the effect of the sample size on posterior inclusion probabilities is assessed under difference censoring schemes. We consider sample sizes $n \in \{50, 100, 200, 500\}$ under three censoring schemes:

- low censoring(5%)
- medium censoring (40%) and
- high censoring(70%)

In all comparison we consider 50 simulated datasets.

3.11.1 Censoring Cases for varying sample size

Figures 3.2, 3.3, 3.4, 3.5 3.6 suggest that overall, when the censoring is small, the model identifies the true model as n increases in terms since the inclusion probabilities tend to one for the non-zero effects and to zero for the zero effects. This effective is done for most datasets for $n = 50$ and for all datasets for $n = 100$.

We observe that the variability of the inclusion probabilities across samples is considerably higher for $n = 50$ but it considerably decreases for $n = 100$ where in very few datasets (2) two variables with zero effects are missclassified as important. In the case of high censoring, the effect is severe. For 50 observations only one variable (out of five) is identified as important, while for $n = 100$ there is a considerable number of datasets where the importance of variables is missclassified. It takes at least 200 observations for 5 covariates to reach posterior inclusion probabilities that agree with the true model.

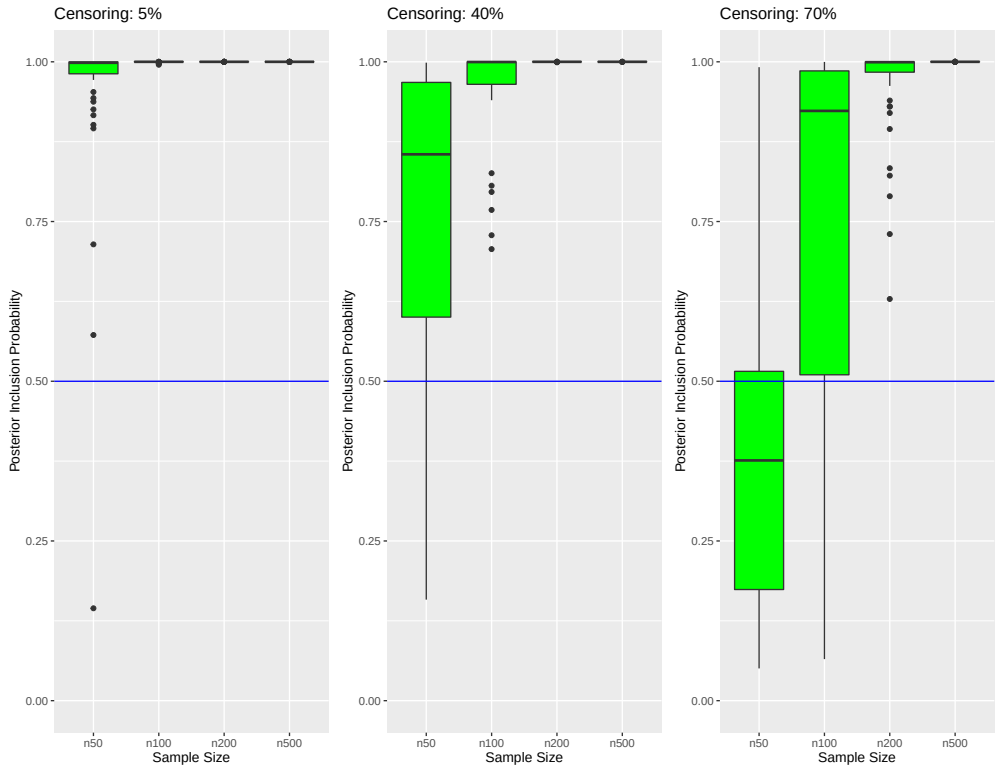

Figure 3.2: Posterior Inclusion Probabilities for X1 assuming for a censoring percentage of 5%, 40% and 70% for $n = 50, 100, 200, 500$ and the true parameters being $\boldsymbol{\beta} = (0.20, 0.75, 0, -0.15, 0)$, $\sigma = 0.8$, $\beta_0 = 7$, datasets $= 50$

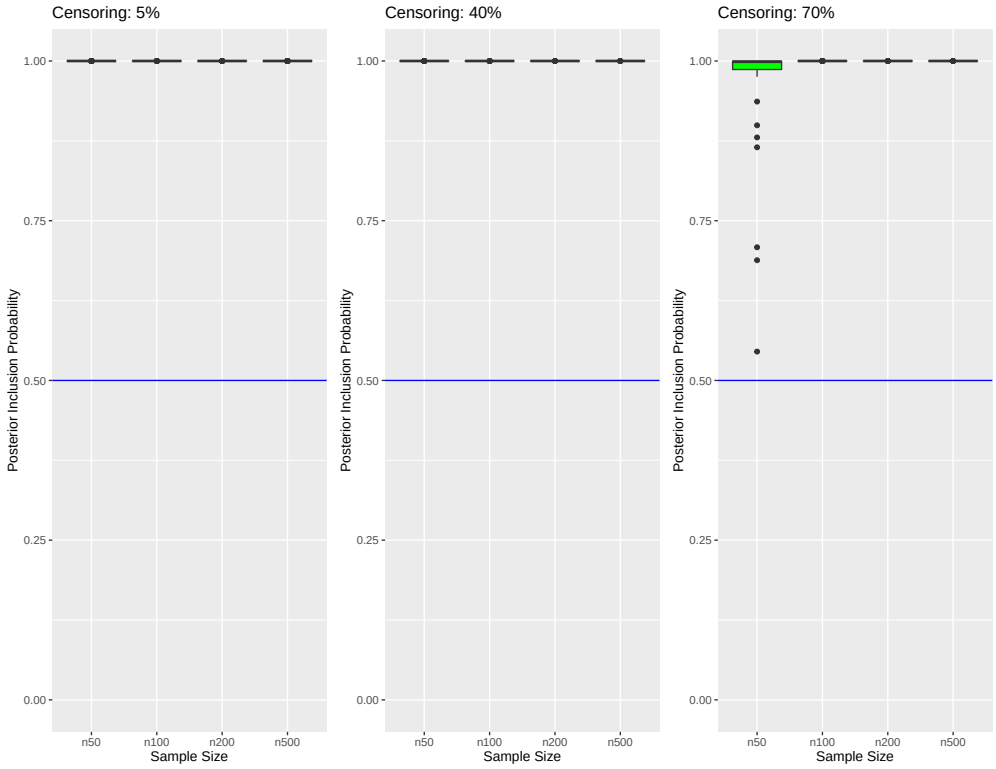

Figure 3.3: Posterior Inclusion Probabilities for X2 assuming for a censoring percentage of 5%, 40% and 70% for $n = 50, 100, 200, 500$ and the true parameters being $\boldsymbol{\beta} = (0.20, 0.75, 0, -0.15, 0)$, $\sigma = 0.8$, $\beta_0 = 7$, datasets $= 50$

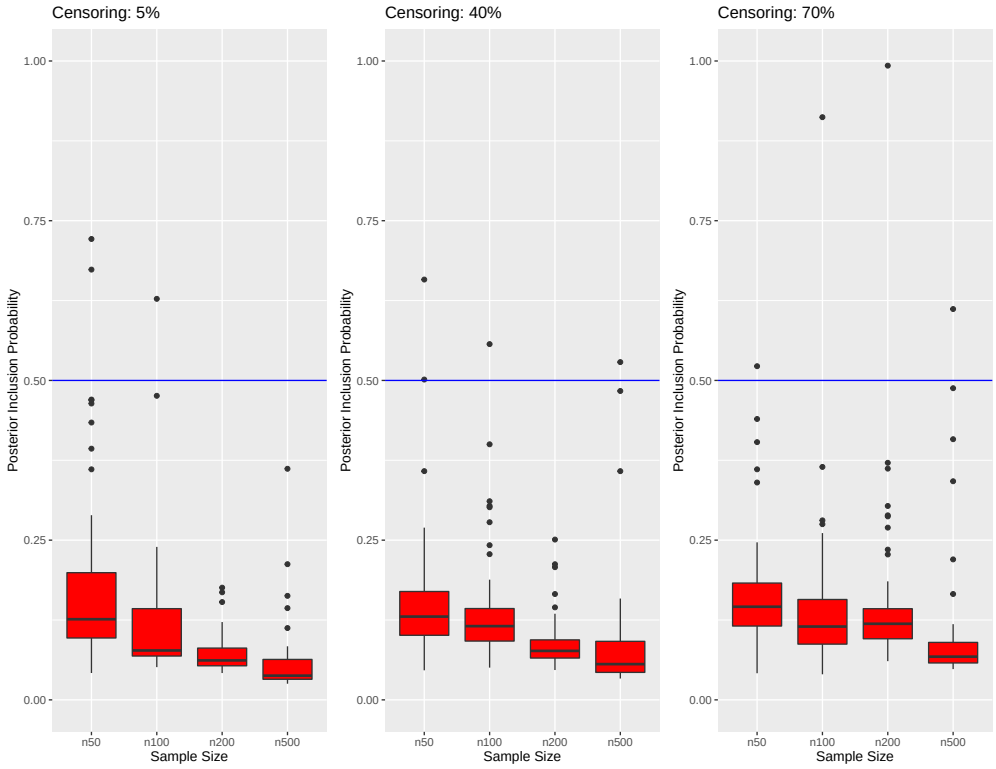

Figure 3.4: Posterior Inclusion Probabilities for X3 assuming for a censoring percentage of 5%, 40% and 70% for $n = 50, 100, 200, 500$ and the true parameters being $\boldsymbol{\beta} = (0.20, 0.75, 0, -0.15, 0)$, $\sigma = 0.8$, $\beta_0 = 7$, datasets $= 50$

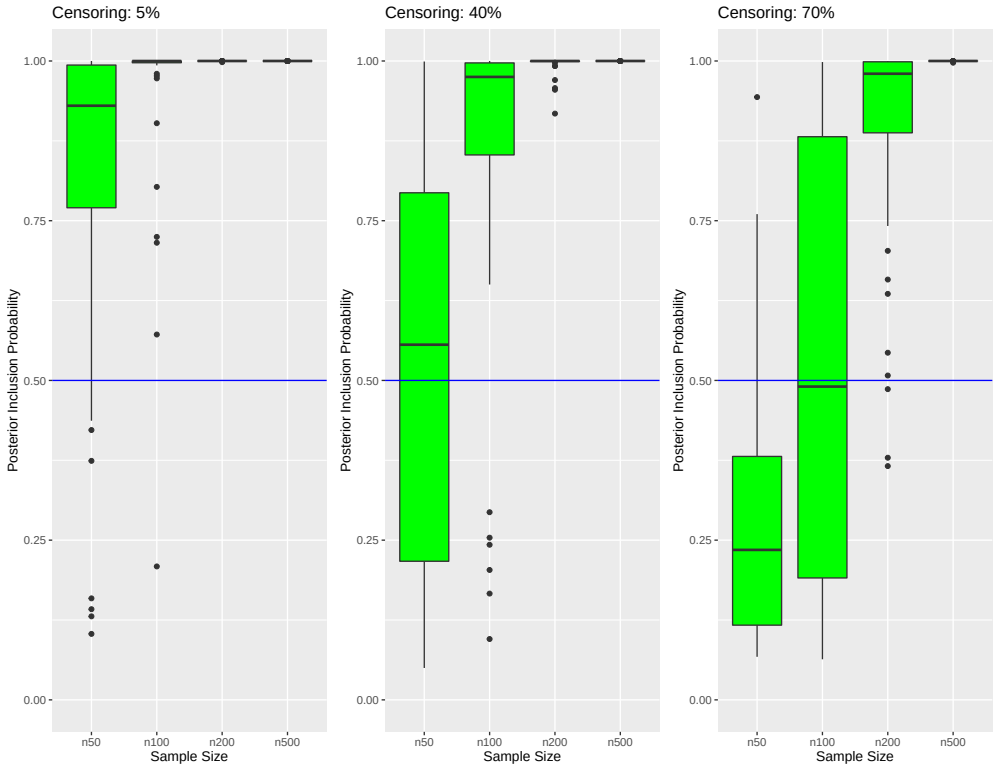

Figure 3.5: Posterior Inclusion Probabilities for X4 assuming for a censoring percentage of 5%, 40% and 70% for $n = 50, 100, 200, 500$ and the true parameters being $\boldsymbol{\beta} = (0.20, 0.75, 0, -0.15, 0)$, $\sigma = 0.8$, $\beta_0 = 7$, datasets $= 50$

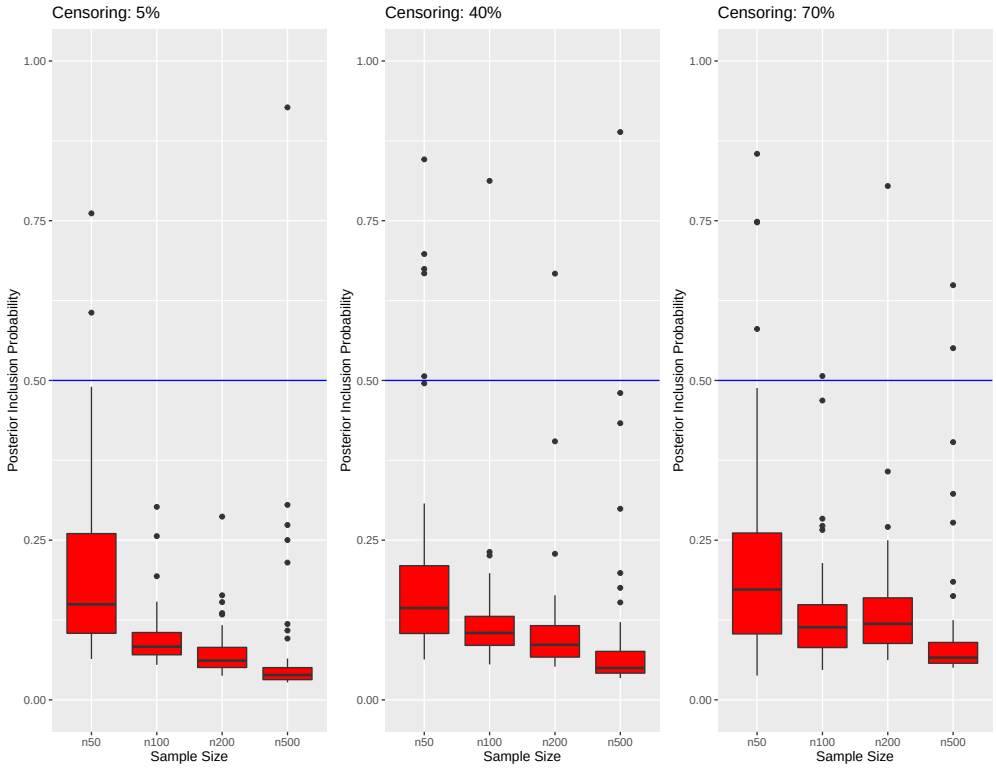

Figure 3.6: Posterior Inclusion Probabilities for X5 assuming for a censoring percentage of 5%, 40% and 70% for $n = 50, 100, 200, 500$ and the true parameters being $\boldsymbol{\beta} = (0.20, 0.75, 0, -0.15, 0)$, $\sigma = 0.8$, $\beta_0 = 7$, datasets $= 50$

The overall conclusion is that the BVS under our suggested prior is relatively robust for small and medium amount of censoring.

3.12 Model selection consistency

Model selection consistency is a crucial property should be desired when constructing a prior for Bayesian Variable Selection. In particular, it refers to the ability of a Bayesian variable selection procedure to identify the true model as the sample size goes to infinity. This crucial property can be fairly simply understood by explaining that if the data is generated from a specific true model, then as the amount of data increases, the procedure of variable selection will point to the true model. This is extremely important since if there is no guarantee that, if the true model is known, the variable selection mechanism will choose that model, the gather more data will not lead to spotting the true model.

A formal mathematical definition of this property was presented by Bayarri et al. (2012), as:
If $\mathcal{M}_T$ is the true model then

$$P\left(\mathcal{M}_T|y\right) \xrightarrow{p} 1$$

as $n \to \infty$.

This essentially means that the posterior true model probability tends to 1 as n becomes extremely large.

To show that the prior that we are proposing satisfies the model selection consistency, we first check how the posterior inclusions probabilities behave for significant and insignificant covariates to check if, as the sample size increase the posterior inclusion probabilities point to what we would expect if the procedure had found the true model. Additionally, after strictly following how this property has been formulated by Bayarri et al. (2012), we check whether the posterior model probability of the true model actually converges to 1 (or a neighborhood of 1) as the sample size increases. The results are presented in a series of boxplots below after simulations have ran for 20 datasets.

3.13 Model selection consistency based on Posterior inclusion probabilities

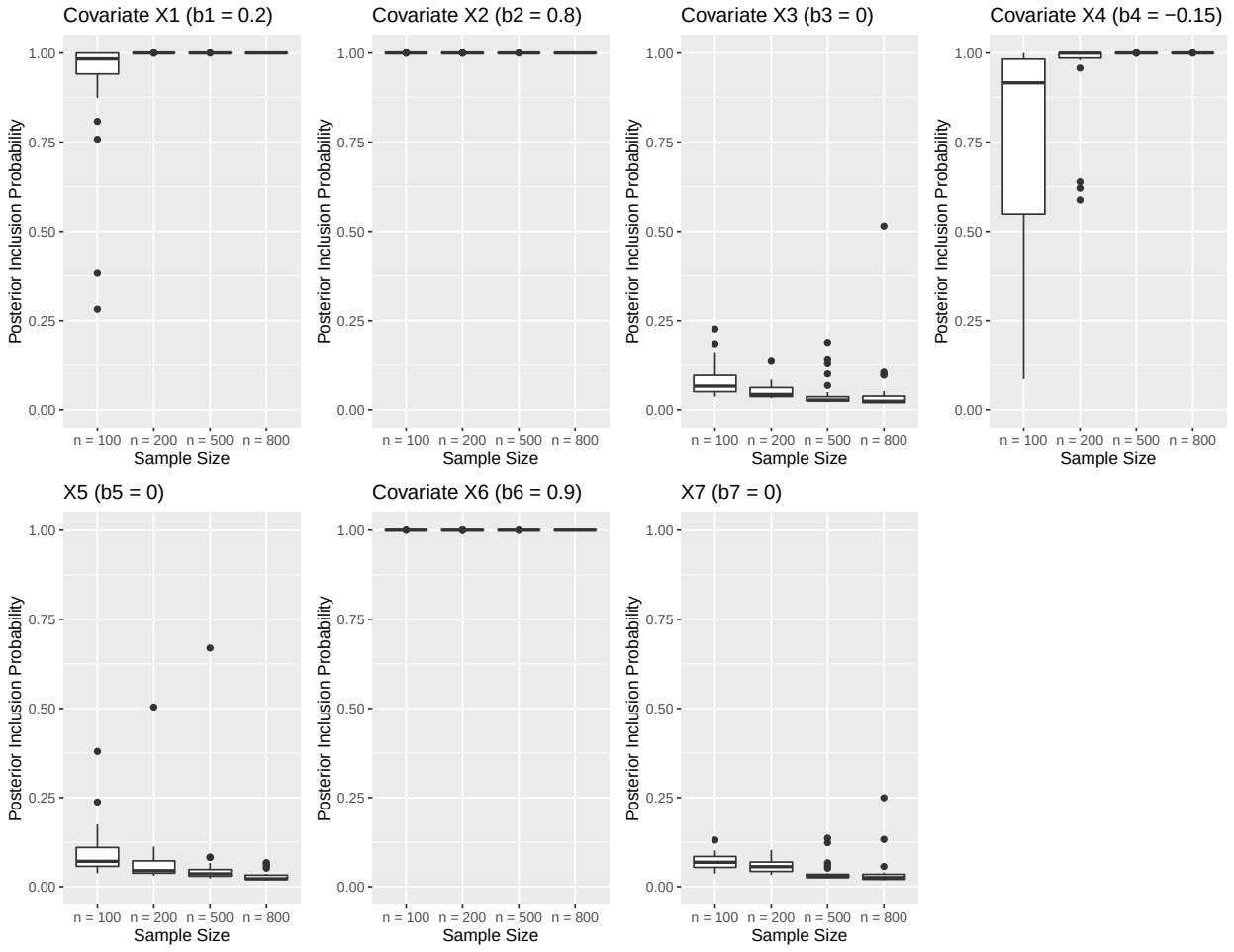

Figure 3.7: Posterior Inclusion Probabilities for all covariates under consideration, as n increases from 100 to 800. The true values of the parameters were: $\boldsymbol{\beta} = (0.2, 0.8, 0, -0.15, 0, 0.9, 0)$, $\sigma = 0.9$, $\beta_0 = 7$. The number of datasets that this simulation ran on was 20.

The pattern that follows Figures 3.7, is the posterior inclusion probabilities of both significant and insignificant covariates converge to what we expect if our suggested prior was following the notion of model selection consistency in the sense that the covariates' posterior inclusions probabilities converge to values close to 1 for significant covariates and to 0 for insignificant covariates as the sample size increase.

3.14 Model selection consistency based on Posterior Model probabilities

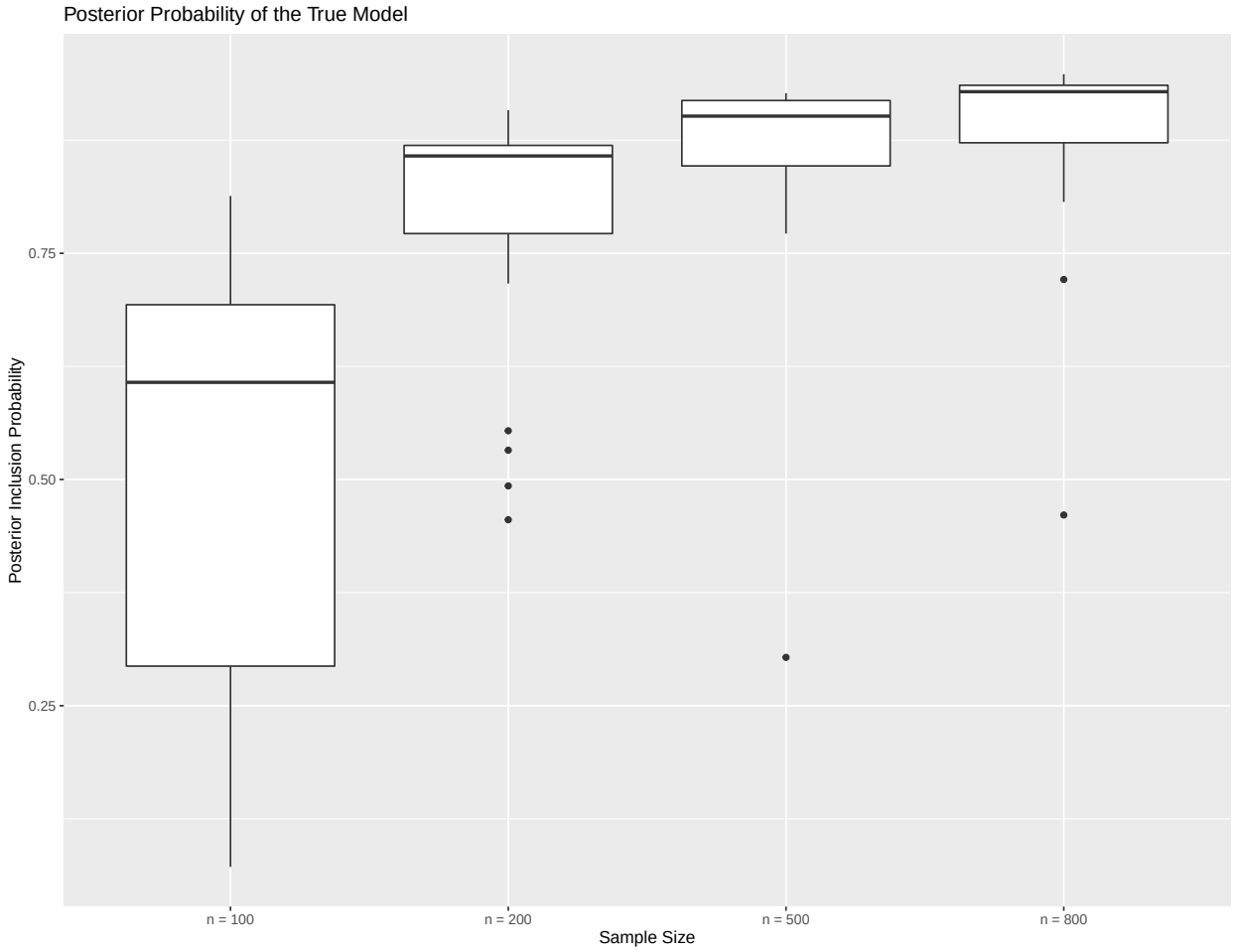

Figure 3.8: Posterior True Model Probability, as n increases from 100 to 800. The true values of the parameters were $\boldsymbol{\beta} = (0.2, 0.8, 0, -0.15, 0, 0.9, 0)$, $\sigma = 0.9$, $\beta_0 = 7$. The number of datasets that this simulation ran on was 20.

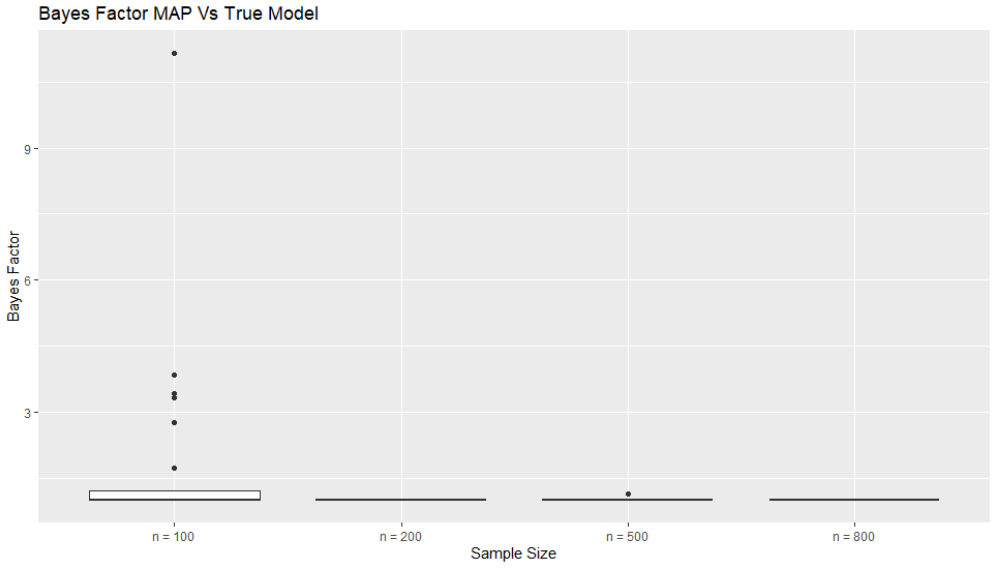

Figure 3.9: Bayes Factor comparison between the MAP and the True model as n increases from 100 to 800. The true values of the parameters were $\boldsymbol{\beta} = (0.2, 0.8, 0, -0.15, 0, 0.9, 0)$, $\beta_0 = 7$. The number of datasets that this simulation ran on was 20.

From Figure 3.8 we observe that the posterior probability of the true model icreases with n starting from low values for $n = 100$ and rising to 0.92 (median) for $n = 800$. This is a clear indication that model is true as defined by Bayarri et al. (2012). This is because the True model has a posterior inclusion probability that is increasing consistently and for $n = 800$ is close to a neighborhood of one and more specifically a median of 0.9286 which we expect to reach one as the sample size increases further.

Focusing on the Bayes factors presented in Figure 3.9 is is evident that the median Bayes Factor when comparing the MAP model to the True model for $n = 100$ is 1.84 meaning that even if the true model is not selected, its Bayesian support is close to the MAP.

3.15 The Primary Biliary Cirrhosis (PBC) dataset

In this section we analyse the real world dataset called Primary Biliary Cirrhosis dataset containing survival data from the Mayo Clinic trial that was conducted to study the liver disease called primary biliary cirrhosis and was conducted between 1974 and 1984. The dataset includes 418 patients, of whom 312 participated in a randomized trial comparing an active treatment to placebo.

In our analysis we focused solely on the continuous numerical variables of the dataset. The predictors that we included in the saturated (full) model were: age (age), serum bilirubin (bili), serum cholesterol (chol), albumin (albumin), copper (copper), alkaline phosphatase (alk.phos), aspartate aminotransferase (ast), triglycerides (trig), platelet count (platelet), prothrombin time (protime).

The original censoring indicator in the PBC dataset, takes three values: 0 if the observation censored, one (1) if the subject has received a transplant and two (2) if the subject has died. For simplicity and for

consistency throughout this thesis, we recorded this variable as binary with 0 representing the value of the censoring indicator when the respective subject is censored and 1 if the subject experienced an event (either transplant or death).

The dataset contains missing values across several covariates. After excluding observations with missing values, the final sample size was $n = 270$. We proceed under the assumption that the missing data are Missing Completely at Random, ensuring that the remaining subset maintains a representative distribution for demonstration of our methodology.

3.15.1 Results

Since we are working with real data rather than simulations, there is no true model for which variables should be included in the model. Instead, we assess the performance of our BVS method through comparison with established frequentist variable selection approaches. Our validation strategy employs two complementary perspectives:

We first fit a standard AFT Weibull model and identify variables that are statistically significant based on p-values ($\alpha = 0.05$). While p-values measure evidence against the null hypothesis rather than probability of inclusion, comparing our posterior inclusion probabilities (PIPs) against this classical criterion provides insight into the agreement between Bayesian and frequentist paradigms.

To obtain a frequentist quantity more directly comparable to PIPs, we employ a bootstrap resampling approach. Over 1000 bootstrap samples, we fit the AFT Weibull model and record which covariates are statistically significant in each replicate. The resulting bootstrap inclusion frequency the proportion of replicates in which each variable is selected serves as a frequentist analog to our Bayesian PIPs. This metric captures both the strength of association and the stability of selection across different data subsamples.

Table 3.15.1 presents the results of this comparison across all ten candidate variables in the PBC dataset.

Variable	**PIP**	**Bootstrap Freq**	p**-value**
Age	0.824	0.785	0.002
Bilirubin	1.000	0.992	0.000
Cholesterol	0.092	0.102	0.814
Albumin	0.999	0.988	0.000
Copper	0.991	0.860	0.000
Alk. Phosphatase	0.082	0.134	0.745
AST	0.310	0.426	0.286
Triglycerides	0.082	0.117	0.722
Platelets	0.145	0.174	0.414
Prothrombin	0.939	0.943	0.000

Table 3.4: Variable selection comparison for PBC dataset. PIP = Posterior inclusion probability; Bootstrap Freq = Proportion of 1,000 bootstrap samples with $p < 0.05$.

The results in Table 3.15.1 reveal strong agreement between the Bayesian and bootstrap frequencies of selected variables approaches. The close agreement between PIPs and bootstrap frequencies is immediately apparent: variables with high PIPs consistently show high bootstrap frequencies, and variables with low PIPs consistently show low bootstrap frequencies. This agreement validates that our Bayesian variable selection method produces results consistent with frequentist stability analysis.

3.16 Discussion

This chapter introduced a unit information prior that accounts for censoring for Bayesian variable selection Weibull survival analysis models. This approach is inspired by the idea of how a prior on the coefficients is constructed in a GLM framework and is based on the work of Castellanos et al. (2021). We started by explaining the general concept of g priors according to the well established theory on prior suggestions, presented by Liang et al. (2008) concerning the Normal likelihood models, and then move on to how these are extended so far to GLMs and log-Normal survival models with censoring (Castellanos et al., 2021).
We then moved to describe how the those well established methods are not exactly applicable on survival analysis framework due to the presence of censored observations and derived the relevant quantities that are needed for constructing a prior that accounts for censoring by introducing a quantity that weights the information between the censored and the uncensored observations mimicking in this way the behavior of the well established Zellner's g-prior.

We tested our prior in a variety of censoring situations and have established that out prior is robust as long as the percentage of censored observations is not unrealistically large. This finding makes a good argument for using our prior under a censored framework using the Weibull AFT model in real world scenarios.

Indeed our Bayesian variable selection method performed excellently on the PBC dataset, correctly identifying the five clinically established predictors (bilirubin, albumin, copper, prothrombin time, and age) with high PIPs while excluding weak predictors. The near-perfect correlation of 0.992 between PIPs and bootstrap frequencies validates the method's stability and alignment with frequentist approaches. Importantly, the method appropriately captured uncertainty for borderline variables rather than forcing binary decisions, demonstrating both methodological rigor and practical effectiveness.

The proposed methodology can be applied to distributions such as the Generalized Gamma. The prior that will be derived under this general distribution is expected to have as a special case the prior that we derived for the Weibull AFT case in the chapter. In this way, the method proposed here can serve as a conversation starter for future work.

Expanding on future work, a further extension of our work would require assigning a distribution on the c_i instead of assuming they are fixed allowing for further hierarchical Bayesian modeling and hence better understanding of the censoring mechanism. Additionally, by including other types of censoring, we may add an interesting complexity on how to derive priors for BVS even in the cases where censoring is informative.

Constructing a prior that is robust in terms of larger number of covariates than those we considered in

this chapter would certainly find big data applications or even clinical trials where multiple measures are taken from a single subject.

Finally, constructing a prior under censoring but assuming a semi parametric model is a challenging future task but we are positive that this methodology will prove to be a useful starting point.

Chapter 4

The Generalized Gamma and Generalized F models

4.1 Introduction

The focus of Chapter 3 was to provide a sensible default prior for BVS. This prior was found to be robust to realistic levels of censoring by weighting the information between the non censored and censored observations. In this Chapter the aim is to provide a similar prior for models that are model general and consider it as a special case.

To be more specific, we will introduce the Generalized Gamma and the Generalized F distributions which include many of the well known AFT models as special cases. We will follow the same steps to deriving the prior weights and the prior under this general model. By doing this, we allow the reader to actually use the generalized form of the prior depending on the AFT model under consideration. Hence, the proper general prior will adopt the desired sub-prior depending on specific parameter values or setups.

We start by providing a formal definition of the models and explain how they can be written as AFT models. After that, we will focus on its properties. Additionally, we derive the desired prior weights and check their properties and how simpler AFT models can be derived as special cases. Finally we construct the proposed prior and present our final prior proposal to $\boldsymbol{\beta}$ under this model. We will close this chapter by presenting a simple example of how well our proposed BVS for these distributions in terms of whether the posterior inclusion probabilities are able to point to the true effects. We conclude this chapter with a brief discussion of our findings.

4.2 The Generalized Gamma Case

Let T be a non negative continuous random variable, and λ, k, s_1 be positive parameters, then the T follows a Generalized Gamma distribution if its pdf can be written as:

$$f\left(t|\lambda,k,s_1\right)=\frac{kt^{ks_1-1}}{\lambda^{ks_1}\Gamma(s_1)}exp\left\{-\left(\frac{t}{\lambda}\right)^k\right\} \tag{4.1}$$

This distribution is connected to other commonly used AFT models as shown in the flow chart of Figure 4.1

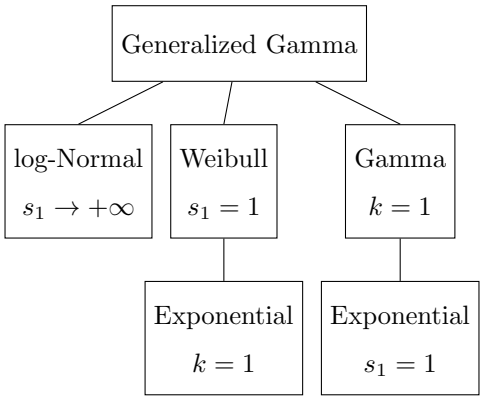

Figure 4.1: Generalized Gamma Flow Chart

The Generalized Gamma distribution can be written as an AFT model in the following way:

$$Y_i=\log(T_i)=\log(\lambda_i)+\frac{1}{k}\epsilon,\quad \epsilon\sim\log Gamma(s_1,1)$$

we prove the above result in the following subsection.

4.2.1 Derivation of the distribution of the error term ϵ

Let

$$Y=\log(T)=\mu+\sigma\epsilon,\quad \mu\in\mathbb{R},\quad \sigma>0,\quad \epsilon\sim\mathbb{D}()$$

where $\mu=\log(\lambda)$, $\sigma=\frac{1}{k}$

Now,

$$f_Y(y)=f_T(e^y)\left|\frac{dt}{dy}\right|=\frac{k(e^y)^{ks_1-1}}{\lambda^{ks_1}\Gamma(s_1)}exp\left\{-\left(\frac{e^y}{\lambda}\right)^k\right\}e^y=\frac{k(e^y)^{ks_1}}{\lambda^{ks_1}\Gamma(s_1)}exp\left\{-\left(\frac{e^y}{\lambda}\right)^k\right\}$$

Generally, if $S\sim Gamma(p,q)$

$$f_S(s)=\frac{1}{\Gamma(p)q^p}s^{p-1}e^{-\frac{s}{q}}$$

and hence for $Gamma(p,1)$ the Gamma pdf takes the following form:

$$f_S(s)=\frac{1}{\Gamma(p)}s^{p-1}e^{-s}$$

Now consider $V = \log(S)$ Then

$$f_V(v) = f_S(e^v)\left|\frac{ds}{dv}\right| = \frac{1}{\Gamma(p)}(e^v)^{p-1}e^{-e^v}s = \frac{1}{\Gamma(p)}(e^v)^{p-1}e^{-e^v}e^v = \frac{1}{\Gamma(p)}e^{pv}e^{-e^v} \tag{4.2}$$

Now for the error term,

$$\begin{aligned} f_\epsilon(\epsilon) &= f_Y\left(\log(\lambda) + \frac{1}{k}\epsilon\right)\left|\frac{dy}{d\epsilon}\right| = \frac{k(e^{\log(\lambda)+\frac{1}{k}\epsilon})^{ks_1}}{\lambda^{ks_1}\Gamma(k)}exp\left\{-\left(\frac{e^{\log(\lambda)+\frac{1}{k}\epsilon}}{\lambda}\right)^k\right\}\frac{1}{k} \\ &= \frac{k\lambda^{ks_1}e^{s_1\epsilon}}{\lambda^{ks_1}\Gamma(s_1)}exp\left\{-\left(\frac{\lambda^k e^\epsilon}{\lambda^k}\right)\right\}\frac{1}{k} = \frac{ke^{s_1\epsilon}}{\Gamma(s_1)}exp\left\{-e^\epsilon\right\}\frac{1}{k} = \frac{e^{s_1\epsilon}}{\Gamma(s_1)}exp\left\{-e^\epsilon\right\} \end{aligned}$$

Therefore,

$$\boxed{f_\epsilon(\epsilon) = \frac{e^{s_1\epsilon}}{\Gamma(s_1)}exp\left\{-e^\epsilon\right\}}$$

which has the exact same form like the form as Equation 4.2.

Therefore, the AFT formulation of the Generalized Gamma distribution is:

$$Y = \log(T) = \log(\lambda) + \frac{1}{k}\epsilon, \quad \epsilon \sim \log Gamma(s_1, 1).$$

log-likelihood component of the uncensored case

In this subsection we proceed with the derivation of all the relevant quantities required to define the inverse of the Fisher Information matrix for the $\boldsymbol{\beta}$ components.

We have

$$f_Y(y) = \frac{k}{\lambda^{ks_1}\Gamma(s_1)}\exp\left(ks_1y - e^{b(y-\log(\lambda))}\right)$$

this leads to the following

$$\begin{aligned} l_u = \log\left[f_Y(y)\right] &= \log\left[\frac{k}{\lambda^{ks_1}\Gamma(s_1)}\exp\left(ks_1y - e^{k(y-\log(\lambda))}\right)\right] \\ &= \log(k) - ks_1\log(\lambda) - \log\left[\Gamma(s_1)\right] + ks_1y - e^{k(y-\log(\lambda))} \\ &= \log\left(\frac{1}{\sigma}\right) - \frac{1}{\sigma}s_1\mu - \log\left[\Gamma(s_1)\right] + \frac{1}{\sigma}s_1y - e^{\frac{1}{\sigma}(y-\mu)} \\ &= -\log(\sigma) + \frac{1}{\sigma}(s_1y - s_1\mu) - \log\left[\Gamma(s_1)\right] - e^{\frac{1}{\sigma}(y-\mu)} \\ &= -\log(\sigma) + \frac{s_1}{\sigma}(y-\mu) - \log\left[\Gamma(s_1)\right] - e^{\frac{1}{\sigma}(y-\mu)} \end{aligned}$$

The finally derived form is given by

$$\boxed{l_{ui} = -\log(\sigma) + \frac{s_1}{\sigma}(y_i-\mu) - \log\left[\Gamma(s_1)\right] - e^{\frac{1}{\sigma}(y_i-\mu)}} \tag{4.3}$$

this is because $k = \frac{1}{\sigma}$ and $\log(\lambda) = \mu$ where $\mu = \beta_0 + \beta^T x$.

Censored case

We know that

$$l_c = \log\left(S_Y(c)\right) \tag{4.4}$$

The survivor function

The survivor function is given by

$$S_Y(y) = 1 - F_Y(y)$$

Thus, we have that

$$F_Y(y) = P(Y \leq y) = P\left(\mu + \sigma\epsilon \leq y\right) = P\left(\epsilon \leq \frac{y-\mu}{\sigma}\right),$$

where

$$\epsilon \sim \log Gamma(s_1, 1)$$

Note that $\epsilon \sim \log Gamma(s_1, 1) \Rightarrow \epsilon = \log(W)$ where $W \sim Gamma(s_1, 1)$.

Hence, the cumulative probability function is given by

$$F_Y(y) = P\left(\epsilon \leq \frac{y-\mu}{\sigma}\right) = P\left(\log(W) \leq \frac{y-\mu}{\sigma}\right) = P\left(W \leq e^{\frac{y-\mu}{\sigma}}\right)$$

where $\gamma(a, b)$ is the lower incomplete Gamma function. In general, if $X \sim Gamma(a, b)$ then the cumulative probability function is given by $F_X(x) = \frac{1}{\Gamma(a)}\gamma\left(a, \frac{x}{b}\right)$

Therefore we have that,

$$F_Y(y) = P\left(W \leq e^{\frac{y-\mu}{\sigma}}\right) = \frac{1}{\Gamma(k)}\gamma\left(k, e^{\frac{y-\mu}{\sigma}}\right)$$

where $\gamma(a, b)$ is the lower incomplete gamma function.

Finally, for the survivor function of the Generalized Gamma distribution we obtain:

$$\begin{aligned} S_Y(y) = 1 - F_Y(y) &= 1 - \frac{1}{\Gamma(s_1)}\gamma\left(s_1, e^{\frac{y-\mu}{\sigma}}\right) \\ &= \frac{\Gamma(s_1) - \gamma(s_1, e^{\frac{y-\mu}{\sigma}})}{\Gamma(s_1)} \\ &= \frac{\int_0^{\infty} t^{s_1-1}e^{-t}dt - \int_0^{e^{\frac{y-\mu}{\sigma}}} t^{s_1-1}e^{-t}dt}{\Gamma(s_1)} \\ &= \frac{\int_{e^{\frac{y-\mu}{\sigma}}}^{\infty} t^{s_1-1}e^{-t}dt}{\Gamma(s_1)} \\ &= \frac{\Gamma\left(s_1, e^{\frac{y-\mu}{\sigma}}\right)}{\Gamma(s_1)} \end{aligned}$$

The log-likelihood component of the censored cases

Following Equation 4.4 the log-likelihood component for the censored observation i is derived from,

$$l_c = \log\left(S_Y(c)\right) = \log\left[\frac{\Gamma\left(s_1, e^{\frac{c-\mu}{\sigma}}\right)}{\Gamma(s_1)}\right] = \log\left[\Gamma\left(s_1, e^{\frac{c-\mu}{\sigma}}\right)\right] - \log\left[\Gamma(s_1)\right]$$

And hence we have,

$$\boxed{l_{ci} = \log\left[\Gamma\left(k, e^{\frac{c_i-\mu}{\sigma}}\right)\right] - \log\left[\Gamma(s_1)\right]} \tag{4.5}$$

4.2.2 Derivative functions of the log-likelihood components

The Uncensored Case

For the uncensored case, we have that the log-likelihood component is given by 4.3.

Hence, the first derivative is given by

$$\begin{aligned}
\frac{\partial l_{ui}}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j}\left[-\log(\sigma) + \frac{s_1}{\sigma}(y_i - \mu) - \log\left[\Gamma(s_1)\right] - e^{\frac{1}{\sigma}(y_i-\mu)}\right] \\
&= \frac{\partial}{\partial \beta_j}\left[-\log(\sigma) + \frac{k}{\sigma}\left(y_i - (\beta_0 + \beta^T x)\right) - \log\left[\Gamma(s_1)\right] - e^{\frac{1}{\sigma}\left(y_i-(\beta_0+\beta^T x)\right)}\right] \\
&= \frac{\partial}{\partial \beta_j}\left[\frac{s_1}{\sigma}\left(y_i - (\beta_0 + \beta^T x)\right) - e^{\frac{1}{\sigma}\left(y_i-(\beta_0+\beta^T x)\right)}\right] \\
&= -\frac{s_1}{\sigma}x_{ij} + \frac{1}{\sigma}x_{ij}e^{\frac{1}{\sigma}\left(y_i-(\beta_0+\beta^T x)\right)} \\
&= \frac{1}{\sigma}x_{ij}\left(e^{\frac{1}{\sigma}\left(y_i-(\beta_0+\beta^T x)\right)} - s_1\right)
\end{aligned}$$

Therefore, the second derivative is given by we obtain

$$\boxed{\frac{\partial l_{ui}}{\partial \beta_j} = \frac{1}{\sigma}x_{ij}\left(e^{\frac{1}{\sigma}\left(y_i-(\beta_0+\beta^T x)\right)} - s_1\right)}$$

Now, for the second derivative we have

$$\begin{aligned}
\frac{\partial^2 l_{ui}}{\partial \beta_j \partial \beta_k} &= \frac{\partial}{\partial \beta_k}\left[\frac{1}{\sigma}x_{ij}\left(e^{\frac{1}{\sigma}\left(y_i-(\beta_0+\beta^T x)\right)} - k\right)\right] \\
&= \frac{1}{\sigma}x_{ij}\frac{\partial}{\partial \beta_k}\left[e^{\frac{1}{\sigma}\left(y_i-(\beta_0+\beta^T x)\right)}\right] \\
&= \frac{1}{\sigma}x_{ij}\left[-\frac{1}{\sigma}x_{ik}e^{\frac{1}{\sigma}\left(y_i-(\beta_0+\beta^T x)\right)}\right] \\
&= -\frac{1}{\sigma^2}x_{ij}x_{ik}e^{\frac{1}{\sigma}\left(y_i-(\beta_0+\beta^T x)\right)}
\end{aligned}$$

Therefore, the second derivative is given by

$$\boxed{\frac{\partial^2 l_{ui}}{\partial \beta_j \partial \beta_k} = -\frac{1}{\sigma^2}x_{ij}x_{ik}e^{\frac{1}{\sigma}\left(y_i-(\beta_0+\beta^T x)\right)}}$$

The Censored Case

With regards to the censored case, since the log-likelihood component is given by Equation 4.5, the first derivative is given by

$$\begin{aligned}
\frac{\partial l_{ci}}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j}\left\{\log\left[\Gamma\left(s_1, e^{\frac{c_i-\mu}{\sigma}}\right)\right] - \log\left[\Gamma(s_1)\right]\right\} \\
&= \frac{\partial}{\partial \beta_j}\log\left[\Gamma\left(s_1, e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right)\right] \\
&= \frac{1}{\Gamma\left(s_1, e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right)}\frac{\partial}{\partial \beta_j}\Gamma\left(s_1, e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right)
\end{aligned}$$

From the properties of the Gamma function, we have that:

$$\frac{\partial \Gamma(s,x)}{\partial x} = -x^{s-1}e^{-x}$$

Thus the first derivative of the log-likelihood component of a censored observation is given by,

$$\begin{aligned}
\frac{\partial l_{ci}}{\partial \beta_j} &= \frac{1}{\Gamma\left(s_1, e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right)}\left[-\left(e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right)^{s_1-1} e^{-\left(e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right)}\right]\frac{\partial}{\partial \beta_j}\left(e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right) \\
&= \frac{1}{\Gamma\left(s_1, e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right)}\left[-\left(e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right)^{s_1-1} e^{-\left(e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right)}\right]\left(-\frac{x_{ij}}{\sigma}\right)\left(e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right) \\
&= \frac{x_{ij}}{\sigma}\frac{1}{\Gamma\left(s_1, e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right)}\left[\left(e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right)^{s_1} e^{-\left(e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right)}\right]
\end{aligned}$$

Taking into consideration that $z_i^c = \frac{c_i-(\beta_0+\beta^T x)}{\sigma}$, we end up

We have,

$$\boxed{\frac{\partial l_{ci}}{\partial \beta_j} = \frac{x_{ij}}{\sigma}\frac{e^{z_i^c s_1}e^{-e^z}}{\Gamma(s_1, e^z)}}$$

$$\frac{\partial^2 l_i^c}{\partial \beta_j \partial \beta_k} = -\frac{1}{\sigma^2}x_{ij}x_{ik}$$

$$\frac{e^{\left(\frac{c_i-(\beta_0+\beta^T x)}{\sigma}\right)s_1}e^{-e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}}\left[\left(s_1 - e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right)\Gamma\left(s_1, e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right) + e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}s_1}e^{-e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}}\right]}{\Gamma\left(s_1, e^{\frac{c_i-(\beta_0+\beta^T x)}{\sigma}}\right)^2}$$

Additionally if we consider the,

$$z_i^c = \frac{c_i-(\beta_0+\beta^T x)}{\sigma}$$

then

$$\boxed{\frac{\partial^2 l_c}{\partial \beta_j \partial \beta_k} = -\frac{1}{\sigma^2}x_{ij}x_{ik}\frac{e^{z_i^c s_1}e^{-e^{z_i^c}}\left[(s_1 - e^{z_i^c})\Gamma(s_1, e^{z_i^c}) + e^{z_i^c s_1}e^{-e^{z_i^c}}\right]}{\Gamma(s_1, e^{z_i^c})^2}}$$

4.3 The Fisher Information Matrix

As already discussed in Chapter 3, the Fisher information matrix comprises of the following parts

- The uncensored part
- The censored part

The Uncensored component

The uncensored component of the Fisher information matrix is

$$E\left[\frac{\partial^2 l_{ui}}{\partial \beta_j \partial \beta_k}\right]P(\delta_i = 1),$$

where

$$\frac{\partial^2 l_{ui}}{\partial\beta_j\partial\beta_k} = -\frac{1}{\sigma^2}x_{ij}x_{ik}e^{\frac{1}{\sigma}\left(y_i-(\beta_0+\beta^T x)\right)},$$

and

$$P(\delta_i = 1) = F_Y(y) = \frac{1}{\Gamma(s_1)}\gamma\left(s_1, e^{\frac{y-(\beta_0-\beta^T x)}{\sigma}}\right).$$

Now, The expectation of the second mathematically presented as

$$E\left[\frac{\partial^2 l_{ui}}{\partial\beta_j\partial\beta_k}\right] = E\left[-\frac{1}{\sigma^2}x_{ij}x_{ik}e^{\frac{1}{\sigma}\left(y_i-(\beta_0+\beta^T x)\right)}\right] = -\frac{1}{\sigma^2}x_{ij}x_{ik}E\left[e^{\frac{1}{\sigma}\left(y_i-(\beta_0+\beta^T x)\right)}\right]$$

For the expectation part, we have that

$$\begin{aligned}
E\left[e^{\frac{1}{\sigma}\left(y_i-(\beta_0+\beta^T x)\right)}\right] &= E\left[e^{\frac{1}{\sigma}\log(t_i)-\frac{1}{\sigma}(\beta_0+\beta^T x)}\right] \\
&= E\left[e^{\log(t_i)\frac{1}{\sigma}}e^{-\frac{1}{\sigma}(\beta_0+\beta^T x)}\right] \\
&= E\left[\left(\frac{t_i}{e^{\beta_0+\beta^T x}}\right)^{\frac{1}{\sigma}}\right]
\end{aligned}$$

Since we are considering the uncensored part, this expectation is conditional.

Hence, we obtain

$$E\left[\left(\frac{t_i}{e^{\beta_0+\beta^T x}}\right)^{\frac{1}{\sigma}}\right] = E\left[\left(\frac{t_i}{e^{\beta_0+\beta^T x}}\right)^{\frac{1}{\sigma}}\middle| t_i \le c_i^*\right] = E\left[\left(\frac{t_i}{e^{\mu}}\right)^{\frac{1}{\sigma}}\middle| t_i \le c_i^*\right]$$

where $c_i^* = e^{c_i}$ with c_i being the log-censoring times.

The conditional pdf of T is given by:

$$f_{T|T\le c_i^*}(t) = \frac{f_T(t)}{P(T \le c_i^*)}$$

where

$$P(T \le c_i^*) = P\left(e^{\mu+\sigma\epsilon} \le c_i^*\right) = P(\mu+\sigma\epsilon \le \log(c_i^*)) = P\left(\epsilon \le \frac{\log(c_i^*)-\mu}{\sigma}\right) = F_\epsilon\left(\frac{c_i-\mu}{\sigma}\right)$$

Since the error term follows a log-Gamma distribution, then

$$\epsilon = \log(J), \quad J \sim Gamma(k,1)$$

Therefore the probability function of ϵ is given by,

$$F_\epsilon\left(\frac{c_i-\mu}{\sigma}\right) = P\left(\epsilon \le \frac{c_i-\mu}{\sigma}\right) = P\left(\log(J) \le \frac{c_i-\mu}{\sigma}\right) = P\left(J \le e^{\frac{c_i-\mu}{\sigma}}\right) = F_J\left(e^{\frac{c_i-\mu}{\sigma}}\right)$$

where $J \sim Gamma(k,1)$ and

$$F_J(j) = \frac{1}{\Gamma(k)}\gamma(k,j)$$

Therefore, we get that

$$\boxed{F_J\left(e^{\frac{c_i-\mu}{\sigma}}\right) = \frac{\gamma\left(k, e^{\frac{c_i-\mu}{\sigma}}\right)}{\Gamma(k)}}$$

Following this result, the conditional expectation of interest is given by,

$$E\left[\left(\frac{t_i}{e^{\mu}}\right)^{\frac{1}{\sigma}} \middle| t_i \leq c_i^*\right] = \frac{1}{P\left(T \leq c_i^*\right)} \int_0^{c_i^*} \left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}} f_T(t)dt$$

For the above integral we get the following result:

$$\begin{aligned}
\int_0^{c_i^*} \left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}} f_T(t)dt &= \int_0^{c_i^*} \left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}} \frac{\frac{1}{\sigma} t^{\frac{1}{\sigma}s_1 - 1}}{(e^{\mu})^{\frac{1}{\sigma}s_1}\Gamma(s_1)} \exp\left(-\left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}}\right) dt \\
&= \int_0^{c_i^*} \frac{1}{t\sigma\Gamma(s_1)} \left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}} \frac{t^{\frac{1}{\sigma}s_1}}{(e^{\mu})^{\frac{1}{\sigma}s_1}} \exp\left(-\left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}}\right) dt \\
&= \int_0^{c_i^*} \frac{1}{t\sigma\Gamma(s_1)} \left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}} \left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}s_1} \exp\left(-\left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}}\right) dt \\
&= \int_0^{c_i^*} \frac{1}{t\sigma\Gamma(s_1)} \left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}(s_1+1)} \exp\left(-\left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}}\right) dt
\end{aligned}$$

Let us consider

$$u = \left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}} \Rightarrow du = \frac{1}{\sigma}\left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}-1} dt = \frac{1}{\sigma}\left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}} \frac{1}{t} dt$$

Additionally, for $t \to 0 \Rightarrow u \to 0$ and for $t \to c_i \Rightarrow u \to \left(\frac{c_i^*}{e^{\mu}}\right)^{\frac{1}{\sigma}}$

Hence, the integral of interest takes the following form,

$$\begin{aligned}
\int_0^{c_i^*} \left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}} f_T(t)dt &= \int_0^{c_i^*} \frac{1}{t\sigma\Gamma(s_1)} \left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}(s_1+1)} \exp\left(-\left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}}\right) dt \\
&= \int_0^{\left(\frac{c_i^*}{e^{\mu}}\right)^{\frac{1}{\sigma}}} \frac{1}{\Gamma(s_1)} u^{s_1} exp(-u)du = \frac{1}{\Gamma(s_1)} \int_0^{\left(\frac{c_i^*}{e^{\mu}}\right)^{\frac{1}{\sigma}}} u^{(s_1+1)-1} exp(-u)du
\end{aligned}$$

The above integral is computed to be equal to the following quantity,

$$\int_0^{c_i^*} \left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}} f_T(t)dt = \frac{1}{\Gamma(s_1)} \int_0^{\left(\frac{c_i^*}{e^{\mu}}\right)^{\frac{1}{\sigma}}} u^{(s_1+1)-1} exp(-u)du = \frac{\gamma\left(s_1+1, \left(\frac{c_i^*}{e^{\mu}}\right)^{\frac{1}{\sigma}}\right)}{\Gamma(s_1)}$$

The conditional expectation of interest, is given by,

$$\begin{aligned}
E\left[\left(\frac{t_i}{e^{\mu}}\right)^{\frac{1}{\sigma}}\middle| t_i \leq c_i^*\right] &= \frac{1}{P\left(T \leq c_i^*\right)}\int_0^{c_i^*}\left(\frac{t}{e^{\mu}}\right)^{\frac{1}{\sigma}} f_T(t)dt \\
&= \frac{1}{P\left(T \leq c_i^*\right)}\frac{\gamma\left(s_1+1,\left(\frac{c_i^*}{e^{\mu}}\right)^{\frac{1}{\sigma}}\right)}{\Gamma(s_1)} \\
&= \frac{\Gamma(s_1)}{\gamma\left(s_1, e^{\frac{\log(c_i^*)-\mu}{\sigma}}\right)}\frac{\gamma\left(s_1+1,\left(\frac{c_i^*}{e^{\mu}}\right)^{\frac{1}{\sigma}}\right)}{\Gamma(s_1)} \\
&= \frac{\gamma\left(s_1+1,\left(\frac{c_i^*}{e^{\mu}}\right)^{\frac{1}{\sigma}}\right)}{\gamma\left(s_1, e^{\frac{\log(c_i^*)-\mu}{\sigma}}\right)} \\
&= \frac{\gamma\left(s_1+1, e^{\frac{\log(c_i^*)-\mu}{\sigma}}\right)}{\gamma\left(s_1, e^{\frac{\log(c_i^*)-\mu}{\sigma}}\right)}
\end{aligned}$$

Finally, the expectation of the uncensored component is given by

$$\begin{aligned}
E\left[\frac{\partial^2 l_{ui}}{\partial\beta_j\partial\beta_k}\right]P(\delta_i=1) &= -\frac{1}{\sigma^2}x_{ij}x_{ik}\frac{\gamma\left(s_1+1, e^{\frac{\log(c_i^*)-\mu}{\sigma}}\right)}{\gamma\left(s_1, e^{\frac{\log(c_i^*)-\mu}{\sigma}}\right)}P(\delta_i=1) \\
&= -\frac{1}{\sigma^2}x_{ij}x_{ik}\frac{\gamma\left(s_1+1, e^{\frac{\log(c_i^*)-\mu}{\sigma}}\right)}{\gamma\left(s_1, e^{\frac{\log(c_i^*)-\mu}{\sigma}}\right)}P(T \leq c_i^*) \\
&= -\frac{1}{\sigma^2}x_{ij}x_{ik}\frac{\gamma\left(s_1+1, e^{\frac{\log(c_i^*)-\mu}{\sigma}}\right)}{\gamma\left(s_1, e^{\frac{\log(c_i^*)-\mu}{\sigma}}\right)}\frac{\gamma\left(s_1, e^{\frac{c_i-\mu}{\sigma}}\right)}{\Gamma(s_1)} \\
&= -\frac{1}{\sigma^2}x_{ij}x_{ik}\frac{\gamma\left(s_1+1, e^{\frac{\log(c_i^*)-\mu}{\sigma}}\right)}{\Gamma(s_1)} \iff
\end{aligned}$$

$$\boxed{E\left[\frac{\partial^2 l_{ui}}{\partial\beta_j\partial\beta_k}\right]P(\delta_i=1) = -\frac{1}{\sigma^2}x_{ij}x_{ik}\frac{\gamma\left(s_1+1, e^{\frac{\log(c_i^*)-\mu}{\sigma}}\right)}{\Gamma(s_1)}} \tag{4.6}$$

The Censored part

For the censored component, we have to calculate

$$E\left[\frac{\partial^2 l_c}{\partial\beta_j\partial\beta_k}\right]P(\delta_i=0), \tag{4.7}$$

where

$$\frac{\partial^2 l_c}{\partial\beta_j\partial\beta_k} = -\frac{1}{\sigma^2}x_{ij}x_{ik}\frac{e^{zs_1}e^{-e^z}\left[(s_1-e^z)\Gamma(s_1,e^z)+e^{zs_1}e^{-e^z}\right]}{\Gamma(s_1,e^z)^2}$$

and

$$P(\delta_i = 0) = P(T > c_i^*) = 1 - P(T \leq c_i^*) = 1 - F_T(t) = 1 - \frac{\gamma\left(s_1, e^{\frac{\log(c_i^*)-\mu}{\sigma}}\right)}{\Gamma(s_1)} = \frac{\Gamma\left(s_1, e^{\frac{\log(c_i^*)-\mu}{\sigma}}\right)}{\Gamma(s_1)} = \frac{\Gamma\left(s_1, e^z\right)}{\Gamma(s_1)}.$$

Since the log-censoring times c_i are assumed to be fixed, the expression (4.7) becomes

$$E\left[\frac{\partial^2 l_c}{\partial\beta_j\partial\beta_k}\right]P(\delta_i = 0) = \left(\frac{\partial^2 l_c}{\partial\beta_j\partial\beta_k}\right)P(\delta_i = 0) = -\frac{1}{\sigma^2}x_{ij}x_{ik}\frac{e^{zs_1}e^{-e^z}\left[(s_1 - e^z)\Gamma(s_1, e^z) + e^{zk}e^{-e^z}\right]}{\Gamma(s_1, e^z)^2}\frac{\Gamma\left(s_1, e^z\right)}{\Gamma(s_1)}$$

Thus, (4.7) is finally given by

$$\boxed{E\left[\frac{\partial^2 l_c}{\partial\beta_j\partial\beta_k}\right]P(\delta_i = 0) = -\frac{1}{\sigma^2}x_{ij}x_{ik}\frac{e^{zs_1}e^{-e^z}\left[(s_1 - e^z)\Gamma(s_1, e^z) + e^{zs_1}e^{-e^z}\right]}{\Gamma(s_1, e^z)\Gamma(s_1)}} \tag{4.8}$$

Hence,the elements of the expected Fisher information matrix for the β block is given by adding the minus of the expectation for the censored component and the uncensored component given by (4.8) and (4.6) respectively.

$$\boxed{\mathcal{I}(\beta)_{jk} = \frac{1}{\sigma^2}x_{ij}x_{ik}\left[\frac{\gamma\left(s_1 + 1, e^z\right)}{\Gamma(k)} + \frac{e^{zs_1}e^{-e^z}\left[(s_1 - e^z)\Gamma(s_1, e^z) + e^{zs_1}e^{-e^z}\right]}{\Gamma(s_1, e^z)\Gamma(s_1)}\right]}$$

4.3.1 The proposed weight function

Following the same arguments as in the Weibull case in Chapter 3 ,the weight function is given by:

$$w(z) = \frac{\gamma\left(s_1 + 1, e^z\right)}{\Gamma(s_1)} + \frac{e^{zs_1}e^{-e^z}\left[(s_1 - e^z)\Gamma(s_1, e^z) + e^{zs_1}e^{-e^z}\right]}{\Gamma(s_1, e^z)\Gamma(s_1)}$$

where $z = \frac{c-\beta_0-\beta^T x}{\sigma}$

Now using the prior information under the null model as in the Weibull case in Chapter 3 in Section (3.5), the weights are simplified to::

$$w_{i_{GG}} = \frac{\gamma\left(s_1 + 1, e^{z_{i0}}\right)}{\Gamma(s_1)} + \frac{e^{z_{i0}s_1}e^{-e^{z_{i0}}}\left[(s_1 - e^{z_{i0}})\Gamma(s_1, e^{z_{i0}}) + e^{z_{i0}s_1}e^{-e^{z_{i0}}}\right]}{\Gamma(s_1, e^{z_{i0}})\Gamma(s_1)}.$$

4.3.2 Properties of the weight function

The above weight seems to be bounded by zero and s_1:

$$0 \leq w_{i_{GG}} \leq s_1$$

which is confirmed by simulations. Therefore, to constrain it to $[0,1]$ we may simply divide them by $s_1 > 0$. Hence, our finally proposed weight function under the Generalized Gamma AFT model is given by:

$$\boxed{w^*_{i_{GG}} = \frac{w_{i_{GG}}}{s_1} = \frac{\gamma\left(s_1+1, e^{z_{i0}}\right)}{s_1\Gamma(s_1)} + \frac{e^{z_{i0}s_1}e^{-e^{z_{i0}}}\left[(s_1 - e^{z_{i0}})\Gamma(s_1, e^{z_{i0}}) + e^{z_{i0}s_1}e^{-e^{z_{i0}}}\right]}{s_1\Gamma(s_1, e^{z_{i0}})\Gamma(s_1)}}$$

where as the reader may recall,

$$z_{i0} = \frac{c_i - \beta_0}{\sigma} \tag{4.9}$$

and c_i are the log censoring times.

The rescaled weights $w^*_{i_{GG}}$ in the implemented simulations behave like a non decreasing function. Hence both the desired properties of $0 \leq w^*_{i_{GG}} \leq 1$ and the fact that $w^*_{i_{GG}}$ is a non decreasing function seem to be satisfied.

4.3.3 The effective sample size

The effective sample size in this case can be formulated following the same arguments as the in the Weibull case as:

$$\boxed{n_{e_{GG}} = \sum_{i=1}^{n} w^*_{i_{GG}}}.$$

Similarly to the Weibull case, when all the observations are censored then this effective sample size is zero while when there are no censored observations then it tends to become n. In mathematical terms, this can be written as

$$\lim_{c_i \to \infty} n_{e_{GG}} = \sum_{i=1}^{n} \lim_{c_i \to \infty} w^*_{i_{GG}} = n$$

and

$$\lim_{c_i \to -\infty} n_{e_{GG}} = \sum_{i=1}^{n} \lim_{c_i \to -\infty} w^*_{i_{GG}} = 0$$

4.3.4 The final form of the proposed covariance matrix Σ_{GG}

The prior's covariance matrix that we are proposing under the Generalized Gamma AFT model, has the following form:

$$\boxed{\Sigma_{GG} = \sigma^2 n_{e_{GG}}\left[X^T\left\{W(\beta_0, \boldsymbol{\beta}=0, \sigma, s_1) - W(\beta_0, \boldsymbol{\beta}=0, \sigma, s_1)\frac{11^T}{n_{e_{GG}}}W(\beta_0, \boldsymbol{\beta}=0, \sigma, s_1)\right\}X\right]^{-1}}$$

where

$$W(\beta_0, \boldsymbol{\beta}=0, \sigma, s_1) = \frac{\gamma\left(s_1+1, e^{z_{i0}}\right)}{s_1\Gamma(s_1)} + \frac{e^{z_{i0}s_1}e^{-e^{z_{i0}}}\left[(s_1 - e^{z_{i0}})\Gamma(s_1, e^{z_{i0}}) + e^{z_{i0}s_1}e^{-e^{z_{i0}}}\right]}{s_1\Gamma(s_1, e^{z_{i0}})\Gamma(s_1)}, \quad i = 1, 2, ..., n$$

with z_{i0} given by (4.9) and

$$n_e = \sum_{i=1}^{n} w^*_{i_{GG}}$$

4.4 The final form of the proposed prior on $\boldsymbol{\beta}$

The final form of the prior that we are proposing on $\boldsymbol{\beta}$ of a model $\mathcal{M}_j$ is given below:

$$\boxed{\boldsymbol{\beta_j}|\mathcal{M}_j, \beta_0, \sigma, s_1 \sim N_{p_j}(\boldsymbol{\beta_j}; 0, \Sigma_{GG})}$$

4.4.1 Priors on extra parameters

Since in the Generalized Gamma AFT model we are introduced to the extra parameter $s_1 > 0$ we assign this parameter with a positive low information Gamma prior namely:

$$s_1 \sim Gamma(0.01, 0.01)$$

which is centered around one but has large variance. To the other parameters we follow the exact same prior scheme as the Weibull case and assign them with $\pi(\beta_0, \sigma) = \frac{1}{\sigma}$.

4.5 Weibull prior weights as a special case of the Generalized Gamma weights

Since the Generalized Gamma distribution is a generalization of the Weibull case, we use the generalized weights to derive the Weibull weights shown in Chapter 3, by simply setting $s_1 = 1$.

In particular, starting with the weights derived from the Generalized Gamma distribution:

$$w^*_{i_{GG}} = \frac{\gamma\left(s_1 + 1, e^{z_{i0}}\right)}{s_1\Gamma(s_1)} + \frac{e^{z_{i0}s_1}e^{-e^{z_{i0}}}\left[(s_1 - e^{z_{i0}})\Gamma(s_1, e^{z_{i0}}) + e^{z_{i0}s_1}e^{-e^{z_{i0}}}\right]}{s_1\Gamma(s_1, e^{z_{i0}})\Gamma(s_1)}$$

Now setting $s_1 = 1$,

$$\begin{aligned} w^*_{i_{GG}}\Big|_{s_1=1} &= \frac{\gamma\left(2, e^{z_{i0}}\right)}{1\Gamma(1)} + \frac{e^{z_{i0}1}e^{-e^{z_{i0}}}\left[(1 - e^{z_{i0}})\Gamma(1, e^{z_{i0}}) + e^{z_{i0}1}e^{-e^{z_{i0}}}\right]}{1\Gamma(s_1, e^{z_{i0}})\Gamma(1)} \\ &= 1 - (e^{z_{i0}} + 1)e^{-e^{z_{i0}}} + e^{z_{i0}}e^{-e^{z_{i0}}}(1 - e^{z_{i0}}) + e^{2z_{i0}}e^{-e^{z_{i0}}} \\ &= 1 - e^{z_{i0}}e^{-e^{z_{i0}}} - e^{-e^{z_{i0}}} + e^{z_{i0}}e^{-e^{z_{i0}}} - e^{2z_{i0}}e^{-e^{z_{i0}}} + e^{2z_{i0}}e^{-e^{z_{i0}}} \\ &= 1 - e^{-e^{z_{i0}}} \end{aligned}$$

which is exactly the weight derived in Chapter 3 in equation (3.3) concerning the Weibull AFT model.

4.6 Toy Example

In this section we will demonstrate the behavior of BVS under the proposed unit information prior in terms of its ability to spot the covariates that we expect to be included in the model based on the posterior inclusion

probabilities. The true model again and similarly to the Weibull AFT case, (in the uncensored case) is mathematically presented in simple terms below:

$$y_i = \log(t_i) = \beta_0 + \boldsymbol{\beta}^T x_i + \sigma\epsilon_i, \quad \epsilon_i \sim \log Gamma(s_1, 1)$$

for $i = 1, 2, ..., 80$ with $\boldsymbol{\beta} = (0.3, 0, 0.2, 0)$, $\beta_0 = 5$, $\sigma = 0.5$, $s_1 = 4$ and the amount of censoring in the data, was defined to be 40%. The vector of $\boldsymbol{\beta}$ in this case defined $2^4 = 16$ models in total for which the marginals had to be approximated using the Laplace approximation.

Table 4.1 illustrates that out proposed prior captures in an impressive manner the correct model considering

Table 4.1: Posterior inclusion probabilities under the Generalized Gamma proposed prior.

X_1	X_2	X_3	X_4
1.000	0.051	0.982	0.052

that half of the observations are censored. While X_2 and X_4 are only marginally included in the model, the correct model is identified in a probability threshold of 5.5% meaning that under this controlled experiment and with a relatively large amount of models to be considered while still having a considerable amount of censoring the correct model is identified. This is a promising finding supporting our choice of the prior and our method in general while still requiring further simulations to fully justify it.

4.7 The Generalized F case

Definition

Let us now consider survival times

$$T \sim GeneralizedF(\lambda, k, s_1, s_2)$$

with pdf defined as follows:

$$f_T(t) = \frac{k}{t}\frac{1}{B(s_1, s_2)}\frac{\left(\frac{1}{s_2}\right)^{s_1}\left(\frac{t}{\lambda}\right)^{ks_1}}{\left[1 + \frac{1}{s_2}\left(\frac{t}{\lambda}\right)^k\right]^{s_1+s_2}}, \quad t, \lambda, k, s_1, s_2 > 0.$$

Now the general form of this distribution allows us to have the Generalized Gamma distribution as its special case as shown in Figure (4.2):

What is evident from the tree diagram of Figure 4.2 is that the Generalized F distribution gives us potential access to many more distributions of simpler form. Particularly it gives us access to distributions like the log-logistic, or the BurrIII and BurrXII which are relatively common in financial modeling.

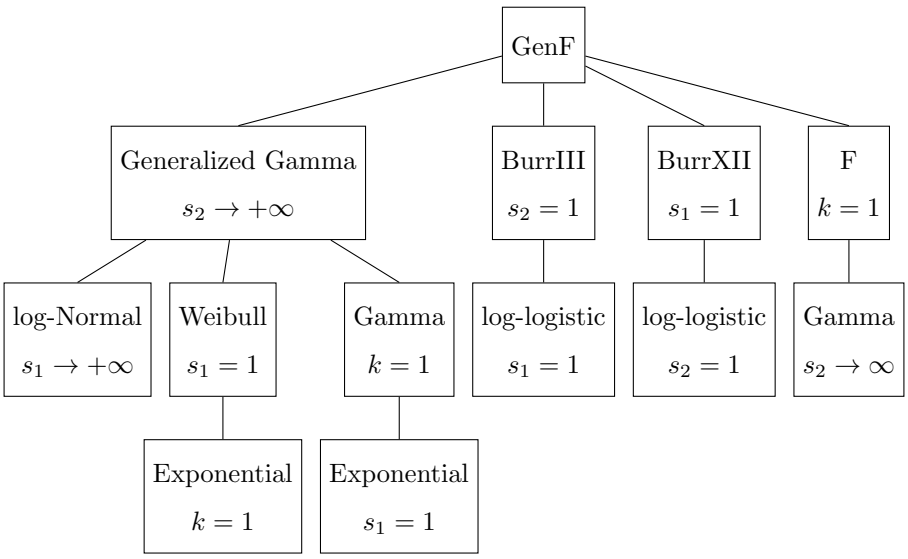

Figure 4.2: Generalized F distribution tree diagram

Assuming that $T \sim GeneralizedF(\lambda, k, s_1, s_2)$ as defined above, the AFT formulation of the model is derived to be the following:

$$\log(T) = \log(\lambda) + \frac{1}{k}\epsilon, \quad \epsilon \sim \mathbb{D}()$$

Derivation of the distribution of the error term ϵ

Let us consider

$$Y = \log(T) = \mu + \sigma\epsilon, \quad \mu \in \mathbb{R}, \quad \sigma > 0, \quad \epsilon \sim \mathbb{D}()$$

which is the standard formulation of the AFT models. Now,

$$f_Y(y) = f_T(e^y)\left|\frac{dt}{dy}\right| = \frac{k}{e^y}\frac{1}{B\left(s_1, s_2\right)}\frac{\left(\frac{1}{s_2}\right)^k\left(\frac{e^y}{\lambda}\right)^{ks_1}}{\left[1+\frac{1}{s_2}\left(\frac{e^y}{\lambda}\right)^k\right]^{s_1+s_2}}e^y = \frac{k}{B\left(s_1, s_2\right)}\frac{\left(\frac{1}{s_2}\right)^{s_1}\left(\frac{e^y}{\lambda}\right)^{ks_1}}{\left[1+\frac{1}{s_2}\left(\frac{e^y}{\lambda}\right)^k\right]^{s_1+s_2}}$$

Immediately it follows that,

$$f_\epsilon(\epsilon) = f_Y\left(\log(\lambda)+\frac{1}{k}\epsilon\right)\left|\frac{dy}{d\epsilon}\right| = \frac{k}{B\left(s_1,s_2\right)}\frac{\left(\frac{1}{s_2}\right)^{s_1}\left(\frac{e^{\left(\log(\lambda)+\frac{1}{k}\epsilon\right)}}{\lambda}\right)^{ks_1}}{\left[1+\frac{1}{s_2}\left(\frac{e^{\left(\log(\lambda)+\frac{1}{k}\epsilon\right)}}{\lambda}\right)^{k}\right]^{s_1+s_2}}\frac{1}{k}$$

$$= \frac{1}{B\left(s_1,s_2\right)}\frac{\left(\frac{1}{s_2}\right)^{s_1}\left(\frac{e^{\left(\log(\lambda)+\frac{1}{k}\epsilon\right)}}{\lambda}\right)^{ks_1}}{\left[1+\frac{1}{s_2}\left(\frac{e^{\left(\log(\lambda)+\frac{1}{k}\epsilon\right)}}{\lambda}\right)^{k}\right]^{s_1+s_2}}$$

$$= \frac{s_2^{-s_1}}{B\left(s_1,s_2\right)}\frac{\left(\frac{\lambda^{ks_1}e^{s_1\epsilon}}{\lambda^{ks_1}}\right)}{\left[1+\frac{1}{s_2}\left(\frac{\lambda^k e^\epsilon}{\lambda^k}\right)\right]^{s_1+s_2}}$$

$$= \frac{s_2^{-s_1}}{B\left(s_1,s_2\right)}\frac{e^{s_1\epsilon}}{\left[1+\frac{1}{s_2}e^\epsilon\right]^{s_1+s_2}}$$

Therefore, the probability density function of the error term is given by

$$\boxed{f_\epsilon(\epsilon) = \frac{s_2^{-s_1}}{B\left(s_1,s_2\right)}\frac{e^{s_1\epsilon}}{\left[1+\frac{e^\epsilon}{s_2}\right]^{s_1+s_2}}}$$

We will refer to it as the log-Generalized F distribution. As for the limiting case,

$$\lim_{s_2\to\infty} f_\epsilon(\epsilon) = e^{s_1\epsilon}\lim_{s_2\to\infty}\left[\frac{s_2^{-s_1}}{B\left(s_1,s_2\right)}\frac{1}{\left[1+\frac{e^\epsilon}{s_2}\right]^{s_1+s_2}}\right] = \frac{e^{s_1\epsilon}}{\Gamma(s_1)}exp(-e^\epsilon) \tag{4.10}$$

since

$$\lim_{s_2\to\infty}\frac{s_2^{-s_1}}{B\left(s_1,s_2\right)} = \frac{1}{\Gamma(s_1)}, \quad \text{and} \quad \lim_{s_2\to\infty}\frac{1}{\left[1+\frac{e^\epsilon}{s_2}\right]^{s_1+s_2}} = \frac{1}{exp\left(e^\epsilon\right)} = exp\left(-e^\epsilon\right).$$

The expression (4.10) the probability density function of the log-gamma distribution which corresponds to the distribution of the error term in the Generalized Gamma case for survival times.

Derivations of the log-likelihood components

The Uncensored case

Since

$$f_Y(y) = \frac{k}{B\left(s_1,s_2\right)}\frac{\left(\frac{1}{s_2}\right)^{s_1}\left(\frac{e^y}{\lambda}\right)^{ks_1}}{\left[1+\frac{1}{s_2}\left(\frac{e^y}{\lambda}\right)^{s_1}\right]^{s_1+s_2}} = \frac{\frac{1}{\sigma}}{B(s_1,s_2)}s_2^{-s_1}\frac{\left(\frac{e^y}{e^\mu}\right)^{\frac{s_1}{\sigma}}}{\left(1+\frac{1}{s_2}\left(\frac{e^y}{e^\mu}\right)^{\frac{1}{\sigma}}\right)^{s_1+s_2}} \tag{4.11}$$

$$= \frac{\frac{1}{\sigma}}{B(s_1,s_2)}s_2^{-k}\frac{\left(e^{\frac{y-\mu}{\sigma}}\right)^{s_1}}{\left(1+\frac{1}{s_2}e^{\frac{y-\mu}{\sigma}}\right)^{s_1+s_2}} \tag{4.12}$$

The log-probability function for an uncensored observation is given by

$$\begin{aligned} l_u = \log\left[f_Y(y)\right] &= \log\left[\frac{\frac{1}{\sigma}}{B(s_1,s_2)} s_2^{-s_1} \frac{\left(e^{\frac{y-\mu}{\sigma}}\right)^{s_1}}{\left(1+\frac{1}{s_2}e^{\frac{y-\mu}{\sigma}}\right)^{s_1+s_2}}\right] \\ &= \log\left(\frac{1}{\sigma}\right) - \log\left[B\left(s_1,s_2\right)\right] + \log\left(s_2^{-s_1}\right) + s_1\left(\frac{y-\mu}{\sigma}\right) - \left(s_1+s_2\right)\log\left(1+\frac{1}{s_2}e^{\frac{y-\mu}{\sigma}}\right) \\ &= -\log\left(\sigma\right) - \log\left[B\left(s_1,s_2\right)\right] - s_1\log\left(s_2\right) + s_1\left(\frac{y-\mu}{\sigma}\right) - \left(s_1+s_2\right)\log\left(1+\frac{1}{s_2}e^{\frac{y-\mu}{\sigma}}\right) \end{aligned}$$

Therefore,

$$\boxed{l_{ui} = -\log\left(\sigma\right) - \log\left[B\left(s_1,s_2\right)\right] - s_1\log\left(s_2\right) + s_1\left(\frac{y_i-\mu}{\sigma}\right) - \left(s_1+s_2\right)\log\left(1+\frac{1}{s_2}e^{\frac{y_i-\mu}{\sigma}}\right)}$$

this is because $k = \frac{1}{\sigma}$ and $\log(\lambda) = \mu$ where $\mu = \beta_0 + \beta^T x$.

The Censored case

For the log-pdf of a log-censored time c, we have that

$$l_c = \log\left(S_Y(c)\right) \quad \text{with} \quad S_Y(y) = 1 - F_Y(y)$$

For the cumulative distribution function we have that

$$F_Y(y) = P(Y \leq y) = P(\log\left(T\right) \leq y) = P(T \leq e^y) = F_T\left(e^y\right),$$

while for the cdf of the Generalized F we have the following: Hence,

$$\begin{aligned} F_T(t) = \int_0^t f_X(x)dx &= \int_0^t \frac{k}{x}\frac{1}{B\left(s_1,s_2\right)}\frac{\left(\frac{1}{s_2}\right)^k\left(\frac{x}{\lambda}\right)^{ks_1}}{\left[1+\frac{1}{s_2}\left(\frac{x}{\lambda}\right)^k\right]^{s_1+s_2}}dx \\ &= \int_0^t \frac{b}{x}\frac{1}{B\left(s_1,s_2\right)}\frac{\left[\left(\frac{1}{s_2}\right)\left(\frac{x}{\lambda}\right)^k\right]^{s_1}}{\left[1+\frac{1}{s_2}\left(\frac{x}{\lambda}\right)^k\right]^{s_1+s_2}}dx \end{aligned}$$

Now by considering the transformation

$$\begin{aligned} u = \frac{1}{s_2}\left(\frac{x}{\lambda}\right)^k &\Rightarrow (s_2u)^{\frac{1}{k}} = \frac{x}{\lambda} \\ &\Rightarrow \lambda(s_2u)^{\frac{1}{k}} = x \\ &\Rightarrow \lambda s_2^{\frac{1}{k}}u^{\frac{1}{k}} = x \\ &\Rightarrow \lambda s_2^{\frac{1}{k}}\frac{1}{k}u^{\frac{1}{k}-1}du = dx \end{aligned}$$

For the limits of the integral we have that for $x = 0$, then $u = 0$ and for $x = t$ then $u = \frac{1}{s_2}\left(\frac{t}{\lambda}\right)^k$. Thus the integral of the pdf becomes

$$\begin{aligned}
F_T(t) &= \int_0^t \frac{k}{x}\frac{1}{B\left(s_1,s_2\right)}\frac{\left[\left(\frac{1}{s_2}\right)\left(\frac{x}{\lambda}\right)^k\right]^{s_1}}{\left[1+\frac{1}{s_2}\left(\frac{x}{\lambda}\right)^k\right]^{s_1+s_2}}dx \\
&= \int_0^{\frac{1}{s_2}\left(\frac{t}{\lambda}\right)^k} \frac{k}{\lambda(s_2u)^{\frac{1}{k}}}\frac{1}{B\left(s_1,s_2\right)}\frac{\left[\left(\frac{1}{s_2}\right)\left(\frac{\lambda(s_2u)^{\frac{1}{k}}}{\lambda}\right)^k\right]^{s_1}}{\left[1+\frac{1}{s_2}\left(\frac{\lambda(s_2u)^{\frac{1}{k}}}{\lambda}\right)^k\right]^{s_1+s_2}}\lambda s_2^{\frac{1}{k}}\frac{1}{k}u^{\frac{1}{k}-1}du \\
&= \int_0^{\frac{1}{s_2}\left(\frac{t}{\lambda}\right)^k} \frac{1}{B\left(s_1,s_2\right)}\frac{1}{u}\frac{u^{s_1}}{(1+u)^{s_1+s_2}}du \\
&= \mathbb{I}\left(\frac{\frac{1}{s_2}\left(\frac{t}{\lambda}\right)^k}{1+\frac{1}{s_2}\left(\frac{t}{\lambda}\right)^k};s_1,s_2\right)
\end{aligned}$$

where $\mathbb{I}\left(\frac{x}{1+x};s_1,s_2\right)$ is the regularized incomplete beta function.

Therefore,

$$\boxed{F_T(t) = \mathbb{I}\left(\frac{\frac{1}{s_2}\left(\frac{t}{\lambda}\right)^k}{1+\frac{1}{s_2}\left(\frac{t}{\lambda}\right)^k};s_1,s_2\right)}$$

Therefore,

$$\boxed{F_Y(y) = F_T(e^y) = \mathbb{I}\left(\frac{\frac{1}{s_2}\left(\frac{e^y}{\lambda}\right)^k}{1+\frac{1}{s_2}\left(\frac{e^y}{\lambda}\right)^k};s_1,s_2\right)}$$

and hence

$$S_Y(y) = 1 - F_Y(y) = 1 - \mathbb{I}\left(\frac{\frac{1}{s_2}\left(\frac{e^y}{\lambda}\right)^k}{1+\frac{1}{s_2}\left(\frac{e^y}{\lambda}\right)^k};s_1,s_2\right)$$

where $b = \frac{1}{\sigma}$, $\lambda = e^{\mu}$, $\mu = \beta_0 + \beta^T x$. Therefore,

$$\boxed{S_Y(y) = 1 - F_Y(y) = 1 - \mathbb{I}\left(\frac{\frac{1}{s_2}\left(e^{\frac{y-\beta_0-\beta^T x}{\sigma}}\right)}{1+\frac{1}{s_2}\left(e^{\frac{y-\beta_0-\beta^T x}{\sigma}}\right)};s_1,s_2\right)}$$

Derivation of l_c

As mentioned above,

$$\boxed{l_c = \log\left(S_Y(c)\right) = \log\left[1 - \mathbb{I}\left(\frac{\frac{1}{s_2}\left(e^{\frac{c-\beta_0-\beta^T x}{\sigma}}\right)}{1+\frac{1}{s_2}\left(e^{\frac{c-\beta_0-\beta^T x}{\sigma}}\right)};s_1,s_2\right)\right] = \log\left[1 - \mathbb{I}\left(\frac{\frac{1}{s_2}e^z}{1+\frac{1}{s_2}e^z};s_1,s_2\right)\right],}$$

4.7.1 Derivative functions of the log-likelihood components

The derivation of the the relevant log-likelihood components require the calculation of the second derivatives of the uncensored and censored log-likelihoods respectively.

Uncensored Case

For the uncensored part and without providing all the mathematical details, the first derivative is given by

$$\boxed{\frac{\partial l_{ui}}{\partial \beta_j} = \frac{1}{\sigma} x_{ij} \left(\frac{s_2(e^z - s_1)}{e^z + s_2} \right)},$$

while the second derivative is given by

$$\boxed{\frac{\partial^2 l_{ui}}{\partial \beta_j \partial \beta_k} = -\frac{1}{\sigma^2} x_{ij} x_{ik} \frac{s_2(s_2 + s_1)e^z}{(e^z + s_2)^2}} \tag{4.13}$$

Censored Case

For the censored part, the first derivative takes the form:

$$\boxed{\frac{\partial l_{ci}}{\partial \beta_j} = -\frac{x_{ij}}{\sigma} s_1 s_2^{s_2} \frac{\left(\frac{1}{s_2 + e^z}\right)^{s_2}}{{}_2F_1\left(s_1, 1 - s_2, s_1 + 1, \frac{1}{1 + s_2 e^z}\right)}}$$

while the second

$$\boxed{\frac{\partial^2 l_{ci}}{\partial \beta_j \partial \beta_k} = -\frac{1}{\sigma^2} x_{ij} x_{ik} \frac{\left(\frac{s_2}{e^z + s_2}\right)^{1+s_2} \left(\frac{e^z}{e^z + s_2}\right)^{s_1} \left(s_2 \left(\frac{s_2}{e^z + s_2}\right)^{s_2 - 1} \left(\frac{e^z}{e^z + s_2}\right)^{s_1} + (e^z - s_1)\, s_2 B(s_1, s_2) \left(\mathbb{I}\left(\frac{e^z}{e^z + s_2}; s_1, s_2\right) - 1 \right) \right)}{s_2 B(s_1, s_2)^2 \left(\mathbb{I}\left(\frac{e^z}{e^z + s_2}; s_1, s_2\right) - 1 \right)^2}} \tag{4.14}$$

Fisher Information Matrix

Similarly to Chapter 3, the Fisher information matrix comprises of the following parts:

- The uncensored part
- The censored part

The uncensored part

The uncensored component of the Fisher information matrix is

$$E\left[\frac{\partial^2 l_{ui}}{\partial \beta_j \partial \beta_k}\right] P(\delta_i = 1) \quad \text{where} \quad \frac{\partial^2 l_{ui}}{\partial \beta_j \partial \beta_k} \quad \text{is given in (4.13)}$$

and

$$P(\delta_i = 1) = F_Y(y) = P\left(Y \leq y\right) = P\left(\log(T) \leq y\right) = P\left(T \leq e^y\right) = F_T(e^y) = \mathbb{I}_{\left(\frac{\frac{1}{s_2}\left(\frac{e^y}{\lambda}\right)^k}{1 + \frac{1}{s_2}\left(\frac{e^y}{\lambda}\right)^k}\right)}(s_1, s_2)$$

Now the expectation of the second derivative is given by,

$$E\left[\frac{s_2(s_2+s_1)e^z}{(e^z+s_2)^2}\right] = s_2(s_2+s_1)E\left[\frac{\left(\frac{t_i}{\lambda}\right)^k}{\left(\left(\frac{t_i}{\lambda}\right)^k+s_2\right)^2}\middle| t_i \le c_i^*\right]$$

where $c_i^* = e^{c_i}$.

The conditional pdf of T given that T represents the uncensored times, is derived as follows:

$$f_{T|T\le c_i^*}(t) = \frac{f_T(t)}{P(T \le c_i^*)}$$

where

$$P\left(T \le t\right) = F_T(t) = \mathbb{I}\left(\frac{\frac{1}{s_2}\left(\frac{t}{\lambda}\right)^{ks_1}}{1+\frac{1}{s_2}\left(\frac{t}{\lambda}\right)^k}; s_1, s_2\right)$$

Therefore, the conditional expectation becomes

$$s_2(s_2+s_1)E\left[\frac{\left(\frac{t_i}{\lambda}\right)^k}{\left(\left(\frac{t_i}{\lambda}\right)^k+s_2\right)^2}\middle| t_i \le c_i^*\right] = \frac{s_2(s_2+s_1)}{P\left(T \le c_i^*\right)}\int_0^{c_i^*}\frac{\left(\frac{t}{\lambda}\right)^k}{\left(\left(\frac{t}{\lambda}\right)^k+s_2\right)^2}f_T(t)dt$$

Therefore, the initially required quantity is given by

$$\boxed{E\left[\frac{\partial^2 l_{ui}}{\partial\beta_j\partial\beta_k}\right]P(\delta_i=1) = -\frac{1}{\sigma^2}x_{ij}x_{ik}\frac{s_1s_2}{s_1+s_2+1}\mathbb{I}\left(\frac{\frac{1}{s_2}e^z}{1+\frac{1}{s_2}e^z}; s_1+1, s_2+1\right)}$$

Censored part

Now, for the censored part, we have to calculate

$$E\left[\frac{\partial^2 l_c}{\partial\beta_j\partial\beta_k}\right]P(\delta_i=0)$$

where $E\left[\frac{\partial^2 l_c}{\partial\beta_j\partial\beta_k}\right]$ is given by (4.14) and

$$P(\delta_i=0) = P(T > c_i^*) = 1-\mathbb{I}\left(\frac{\frac{1}{s_2}e^z}{1+\frac{1}{s_2}e^z}; s_1, s_2\right) = 1-\mathbb{I}\left(\frac{e^z}{s_2+e^z}; s_1, s_2\right)$$

Since the c_i are assumed to be fixed, the expectation is equal to the second derivative itself.

Therefore, we have

$$E\left[\frac{\partial^2 l_c}{\partial\beta_j\partial\beta_k}\right]P(\delta_i=0) = \left[\frac{\partial^2 l_c}{\partial\beta_j\partial\beta_k}\right]P(\delta_i=0)$$

Therefore,

$$\left[\frac{\partial^2 l_c}{\partial\beta_j\partial\beta_k}\right]P(\delta_i=0) = -\frac{1}{\sigma^2}x_{ij}x_{ik}\frac{\left(\frac{s_2}{e^z+s_2}\right)^{1+s_2}\left(\frac{e^z}{e^z+s_2}\right)^{s_1}\left(s_2\left(\frac{s_2}{e^z+s_2}\right)^{s_2-1}\left(\frac{e^z}{e^z+s_2}\right)^{s_1}+\left(e^z-s_1\right)s_2B(s_1,s_2)\left(\mathbb{I}\left(\frac{e^z}{e^z+s_2}; s_1, s_2\right)-1\right)\right)}{s_2B(s_1,s_2)^2\left(1-\mathbb{I}\left(\frac{e^z}{e^z+s_2}; s_1, s_2\right)\right)}$$

Thus

$$\left[\frac{\partial^2 l_c}{\partial\beta_j\partial\beta_k}\right]P(\delta_i=0)=-\frac{1}{\sigma^2}x_{ij}x_{ik}\frac{\left(\frac{s_2}{e^z+s_2}\right)^{1+s_2}\left(\frac{e^z}{e^z+s_2}\right)^{s_1}\left(s_2\left(\frac{s_2}{e^z+s_2}\right)^{s_2-1}\left(\frac{e^z}{e^z+s_2}\right)^{s_1}+(e^z-s_1)\,s_2B(s_1,s_2)\left(\mathbb{I}\left(\frac{e^z}{e^z+s_2};s_1,s_2\right)-1\right)\right)}{s_2B(s_1,s_2)^2\left(1-\mathbb{I}\left(\frac{e^z}{e^z+s_2};s_1,s_2\right)\right)}$$

Hence, the the elements of the Fisher information matrix for the β block are given by:

$$I(\beta)_{jk}=\frac{1}{\sigma^2}x_{ij}x_{ik}\left[\frac{s_1s_2}{s_1+s_2+1}\mathbb{I}\left(\frac{\frac{1}{s_2}e^z}{1+\frac{1}{s_2}e^z};s_1+1,s_2+1\right)+\frac{\left(\frac{s_2}{e^z+s_2}\right)^{1+s_2}\left(\frac{e^z}{e^z+s_2}\right)^{s_1}\left(s_2\left(\frac{s_2}{e^z+s_2}\right)^{s_2-1}\left(\frac{e^z}{e^z+s_2}\right)^{s_1}+(e^z-s_1)\,s_2B(s_1,s_2)\left(\mathbb{I}\left(\frac{e^z}{e^z+s_2};s_1,s_2\right)-1\right)\right)}{s_2B(s_1,s_2)^2\left(1-\mathbb{I}\left(\frac{e^z}{e^z+s_2};s_1,s_2\right)\right)}\right]$$

4.7.2 The proposed weight function

The weight function according to this formulation of the Fisher information matrix is:

$$w_{GENF_i}=\left[\frac{s_1s_2}{s_1+s_2+1}\mathbb{I}\left(\frac{\frac{1}{s_2}e^z}{1+\frac{1}{s_2}e^z};s_1+1,s_2+1\right)+\frac{\left(\frac{s_2}{e^z+s_2}\right)^{1+s_2}\left(\frac{e^z}{e^z+s_2}\right)^{s_1}\left(s_2\left(\frac{s_2}{e^z+s_2}\right)^{s_2-1}\left(\frac{e^z}{e^z+s_2}\right)^{s_1}+(e^z-s_1)\,s_2B(s_1,s_2)\left(\mathbb{I}\left(\frac{e^z}{e^z+s_2};s_1,s_2\right)-1\right)\right)}{s_2B(s_1,s_2)^2\left(1-\mathbb{I}\left(\frac{e^z}{e^z+s_2};s_1,s_2\right)\right)}\right]$$

Now using information under the null model, in a similar manner as in the Weibull case in Chapter 3, the weights take the following form:

$$w_{GENF_i}=\left[\frac{s_1s_2}{s_1+s_2+1}\mathbb{I}\left(\frac{\frac{1}{s_2}e^{z_{i0}}}{1+\frac{1}{s_2}e^{z_{i0}}};s_1+1,s_2+1\right)+\frac{\left(\frac{s_2}{e^{z_{i0}}+s_2}\right)^{1+s_2}\left(\frac{e^{z_{i0}}}{e^{z_{i0}}+s_2}\right)^{s_1}\left(s_2\left(\frac{s_2}{e^{z_{i0}}+s_2}\right)^{s_2-1}\left(\frac{e^{z_{i0}}}{e^{z_{i0}}+s_2}\right)^{s_1}+(e^{z_{i0}}-s_1)\,s_2B(s_1,s_2)\left(\mathbb{I}\left(\frac{e^{z_{i0}}}{e^{z_{i0}}+s_2},s_1,s_2\right)-1\right)\right)}{s_2B(s_1,s_2)^2\left(1-\mathbb{I}\left(\frac{e^{z_{i0}}}{e^{z_{i0}}+s_2},s_1,s_2\right)\right)}\right]$$

where $z_{i0}=\frac{c_i-\beta_0}{\sigma}$.

4.7.3 Properties of the weight function

The above weight is bounded as:

$$0\leq w_{i_{GENF}}\leq\frac{s_1s_2}{s_1+s_2+1}$$

This is confirmed by our simulation runs.

Therefore, to rescale the weights to $[0,1]$ we divide it by its upper bound $\frac{s_1s_2}{s_1+s_2+1}>0$ and hence the weight function under the Generalized F AFT model is:

$$w^*_{i_{GENF}}=\frac{w_{i_{GENF}}}{\frac{s_1s_2}{s_1+s_2+1}}=\left[\mathbb{I}\left(\frac{\frac{1}{s_2}e^{z_{i0}}}{1+\frac{1}{s_2}e^{z_{i0}}};s_1+1,s_2+1\right)+\frac{(s_1+s_2+1)\left(\frac{s_2}{e^{z_{i0}}+s_2}\right)^{1+s_2}\left(\frac{e^{z_{i0}}}{e^{z_{i0}}+s_2}\right)^{s_1}\left(s_2\left(\frac{s_2}{e^{z_{i0}}+s_2}\right)^{s_2-1}\left(\frac{e^{z_{i0}}}{e^{z_{i0}}+s_2}\right)^{s_1}+(e^{z_{i0}}-s_1)\,s_2B(s_1,s_2)\left(\mathbb{I}\left(\frac{e^{z_{i0}}}{e^{z_{i0}}+s_2},s_1,s_2\right)-1\right)\right)}{s_1s_2^2B(s_1,s_2)^2\left(1-\mathbb{I}\left(\frac{e^{z_{i0}}}{e^{z_{i0}}+s_2},s_1,s_2\right)\right)}\right]$$

where $z_{i0}=\frac{c_i-\beta_0}{\sigma}$

It can also be shown through simulations, that $w^*_{i_{GENF}}$ is a non decreasing function.Hence both desired properties are satisfied. Finally, the same argument regarding the range of $w^*_{i_{GENF}}$ approximating zero and one depending on whether all observations are censored or all observations are uncensored can be mathematically proven.

For illustration, we provide Figure 4.3 of $w^*_{i_{GENF}}$ as a function of z_{i0} after fixing $s_1=1$ and $s_2\to\infty$ in our case we took $s_2=20000$.

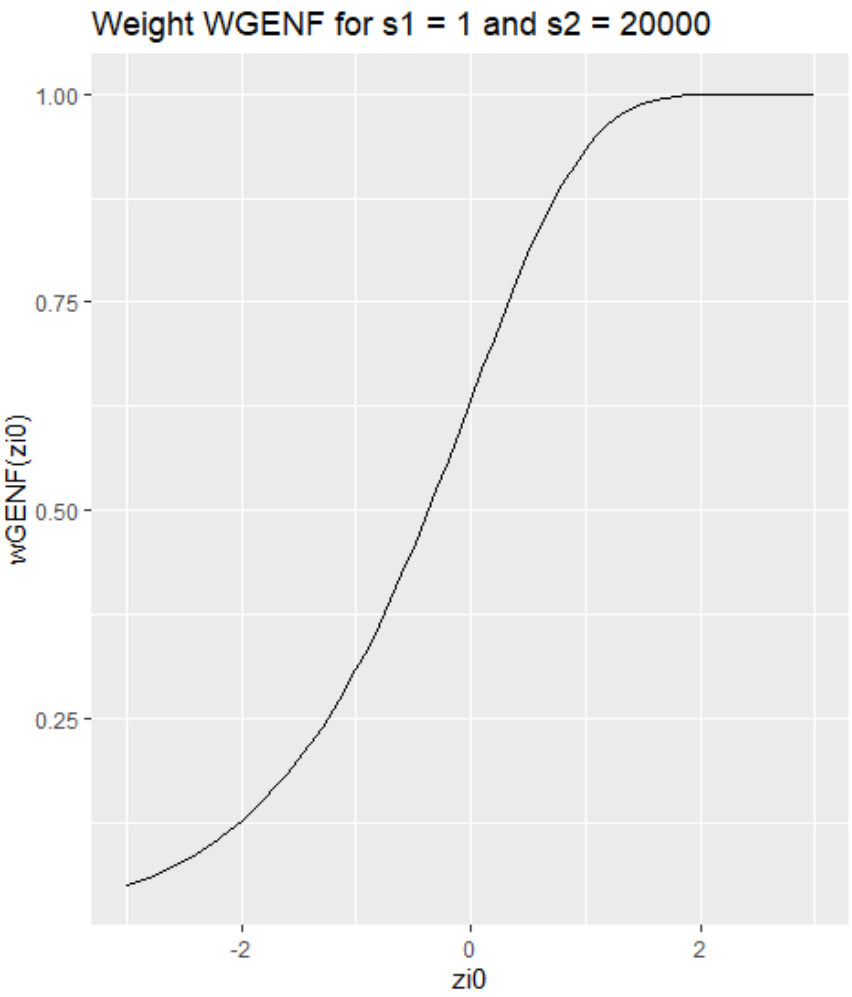

Figure 4.3: The behavior of $w^*_{i_{GENF}}$ over different values of z_{i0} after fixing $s_1 = 1$ and $s_2 = 20000$

What is evident from Figure 4.3 is the fact that as $s_2 \to \infty$ and $s_1 = 1$ confirms what is expected to be true based on the tree plot presented in Subsection 4.6. In particular, when $s_2 \to \infty$ and $s_1 = 1$ the weight function is indeed non decreasing and behaves exactly as the Weibull weight that we derived earlier in terms of its range being $[0, 1]$.

4.7.4 The effective sample size

The effective sample size in this case can be formulated using the same arguments as in the Weibull case.

Hence it is given by

$$\boxed{n_{e_{GF}} = \sum_{i=1}^{n} w^*_{i_{GENF}}}$$

Similarly to the Weibull case, since the $0 \leq w^*_{i_{GENF}} \leq 1$ then again, it holds true that all the observations are censored then this effective sample size is zero while when there are no censored observations then it tends to becomes equal to one. In mathematical terms, we have that

$$\lim_{c_i \to \infty} n_{e_{GF}} = \sum_{i=1}^{n} \lim_{c_i \to \infty} w^*_{i_{GF}} = n$$

and

$$\lim_{c_i \to -\infty} n_{e_{GF}} = \sum_{i=1}^{n} \lim_{c_i \to -\infty} w^*_{i_{GF}} = 0$$

4.7.5 The final form of the proposed covariance matrix Σ_{GF}

The prior's covariance matrix that we are proposing under the Generalized Gamma AFT model, has the following form:

$$\boxed{\Sigma_{GF} = \sigma^2 n_{e_{GF}} \left[X^T \left\{ W(\beta_0, \boldsymbol{\beta} = 0, \sigma, s_1, s_2) - W(\beta_0, \boldsymbol{\beta} = 0, \sigma, s_1, s_2) \frac{11^T}{n_{e_{GF}}} W(\beta_0, \boldsymbol{\beta} = 0, \sigma, s_1, s_2) \right\} X \right]^{-1}}$$

where

$$W_i(\beta_0, \boldsymbol{\beta} = 0, \sigma, s_1, s_2) = \mathbb{I}\left(\frac{\frac{1}{s_2} e^{z_{i0}}}{1 + \frac{1}{s_2} e^{z_{i0}}}; s_1 + 1, s_2 + 1 \right) + \frac{(s_1+s_2+1)\left(\frac{s_2}{e^{z_{i0}}+s_2}\right)^{1+s_2} \left(\frac{e^{z_{i0}}}{e^{z_{i0}}+s_2}\right)^{s_1} \left(s_2 \left(\frac{s_2}{e^{z_{i0}}+s_2}\right)^{s_2-1} \left(\frac{e^{z_{i0}}}{e^{z_{i0}}+s_2}\right)^{s_1} + (e^{z_{i0}} - s_1) s_2 B(s_1, s_2) \left(\mathbb{I}\left(\frac{e^{z_{i0}}}{e^{z_{i0}}+s_2}; s_1, s_2\right) - 1 \right) \right)}{s_1 s_2^2 B(s_1, s_2)^2 \left(1 - \mathbb{I}\left(\frac{e^{z_{i0}}}{e^{z_{i0}}+s_2}, s_1, s_2\right)\right)}$$

for $i = 1, 2, ..., n$, where z_{i0} is given in (4.9) and

$$n_{e_{GF}} = \sum_{i=1}^{n} W_i(\beta_0, \boldsymbol{\beta} = 0, \sigma, s_1, s_2)$$

4.8 The final form of the proposed prior on $\boldsymbol{\beta}$

The final form of the prior that we are proposing on $\boldsymbol{\beta}$ of a model $\mathcal{M}_j$ is given below:

$$\boxed{\boldsymbol{\beta_j} | \mathcal{M}_j, \beta_0, \sigma, s_1, s_2 \sim N_{p_j}(\boldsymbol{\beta_j}; 0, \Sigma_{GF)}}$$

4.8.1 Priors on extra parameters

Since in the Generalized F AFT model we are introduced to the extra parameter $s_1 > 0$ and the parameter $s_2 > 0$ one can assign the following objective priors to them:

$$s_1 \sim Gamma(0.01, 0.01), \quad \text{and} \quad s_2 \sim Gamma(0.01, 0.01)$$

which is centered around 1 but has large variance. To the other parameters we follow the exact same prior scheme as the Weibull and the Generalized Gamma case one can assign them with $\pi(\beta_0, \sigma) = \frac{1}{\sigma}$.

4.8.2 Potential issues on the marginals

Due to the complexity of the prior under the Generalized F model, computational issues can arise both due to the complexity of model structure hence the regular techniques of approximating the marginals may not be efficient or precise at all. Because of that, it would be safer to approximate the marginals under the Bayesian framework that we have established by utilizing more sophisticated techniques that do not depend on the posterior being unimodal or that it can be well approximated by the normal distribution.

4.8.3 Overflow and Underflow problems

Similarly to the Generalized Gamma case, the covariance matrix and of the proposed prior under this model require the calculation of quantities that may lead to overflow or underflow issues. More specifically, the incomplete beta function that is found in the covariance matrix of the prior that we suggest has to be carefully handled especially if the software that is used to apply our methodology is the R programming language. However, there are packages in R that are able to push through the limits of precision of R but they have to be very carefully handled for the method to be applicable.

4.9 Discussion

In this chapter, we extended the application of the method that we applied on the Weibull model to two more general distributions namely the Generalized Gamma distribution and the Generalized F distributions. Both of these distributions include the Weibull distribution as its special case.

With regards to the Generalized Gamma distribution, we presented a way to write this complex model as an AFT model allowing for a linear relationship between the log-Survival times and the covariates. The addition of one extra parameter coming from the generalized nature of the distribution considered allowed us to assigned it an "objective" prior hence adding the necessary Bayesian structure to our modeling procedure.

Following the same procedure as in Chapter 3, we computed the relevant weights that are used to drive the computation of the effective sample size and then proposed our final version of the prior. We assessed its ability to spot the correct covariates that are to be included in the model based on a simple toy example for which the true model was known and the results looked promising.

We also went even further to discuss how our methodology can be applied on an even bigger model that has the Generalized Gamma as its special case. It appeares that to that model also leading to weight quantities that are more complex in terms of mathematical expression but relatively simple to understand.

However the application of our methods to the Generalized F distribution was only for illustration purposes and for proposing a generalized version of the already generalized g-prior that we proposed under the Generalized Gamma model. Further source of understanding of its this distribution could include comparing the distribution of the error term with a known distribution making simulations easier and more robust.

Implementing BVS for the Generalized F version under the proposed prior, can be seen as a challenge for future work since it contains intractable quantities that require approximations and further approximations.

The same additional work that can be done in the Generalized Gamma case, can be done in the Generalized F case in terms of simulations. Particularly it would be interesting to study how the posterior inclusion probabilities are influenced by the sample size and the different levels of censoring. It would also be insightful to check whether model selection consistency is valid for a prior constructed under this model.

Overall the applicability of our methodology to more complex models is a promising contribution and opens the door for further work on aspect that go beyond just handling the censored observations. Finding the limits of the applicability of our method is a challenge for the future especially if the assumptions that we made get to be challenged.

This chapter can serve as a basis for future work. Extending the simulations like those conducted in the Weibull case, while also exploring different objective priors on the extra parameters. Additionally, exploring faster but accurate methods for approximating the marginal distributions required for the Bayes Factors can make the implementation of BVS more effective.

Finally, the applicability of the method we provided under for Cox PH models can be a direct consequence if the Generalized F distribution can be written in a Cox PH form. That would mean that the methodology we provided for constructing a prior under censoring provides a partial (due to the existence of the baseline hazard) answer to the problem for semi parametric models.

Chapter 5

General Findings - Limitations - Future Work

5.1 General Findings

The goal of this thesis was to investigate how the Bayesian Variable Selection techniques can be implemented to models beyond the well established GLMs by viewing the problem from a survival analysis perspective. The presence of censored observations added a layer of complexity in the sense of understanding how the information from the observations should be weighted in a coherent way.

We started by providing a general introduction to survival analysis models. Then moved on to an application of survival models in a non-standard field using soccer data. In this application we established a way of formulating the soccer data so that the goal arrival times are handled under a Bayesian survival analysis framework. Our finally proposed model had a good fit and did well with regards to its prediction abilities especially considering its simplicity.

We then moved to constructing a default unit information prior for Bayesian variable selection under a survival analysis framework under the model that we used in the soccer application in Chapter 2. In particular, we used the accelerated failure time formulation of the Weibull model as our basis for deriving unit information priors under censoring. In particular, we followed the well established theory on how to derive the g-prior covariance matrix for GLMs by redefining the likelihood to account for censored observations. This was achieved by breaking down the likelihood contribution of the censored and uncensored observations. Our suggested prior seems to offer a coherent way to adjust for censoring leading to reasonable performance in terms of identifying true covariates. We tested our suggested prior by conducting simulations and assess its ability to identify true non zero effects under censoring. Our results indicate that the BVS under our proposed prior is robust under realistic levels of censoring. Therefore, the suggested prior seems to handle appropriately censoring. Secondly we have the behavior of BVS under different levels of sample size. The aim of this exercise was to check if our prior leads to a BVS which is model selection consistent. Simulations

indicate a behavior which is compliant with this important property since as $n \to \infty$ the posterior probability of the true model increases towards one.

As a final step, we have applied the same approach to more general distributions that has it as a special case. In particular, we have considered the Generalized Gamma and the Generalized F distributions as the basis for building a default unit information prior for BVS. We have specifically shown how the quantities should be derived in this case in order to correctly account for the likelihood contribution of censored and non censored observations. Our final proposed covariance matrix contains complex quantities like the incomplete gamma function which is something that is to be expected as a natural follow up of increasing the complexity of the model considered. To complete the argument of possibly using this methodology to distributions like these, we conducted a toy example to illustrate the behaviour of the implied BVS. Results seem promising since our prior was able to correctly identify the correct model structure even for a relatively small sample and under medium level of censoring.

5.2 Limitations

Even though our methodology on constructing default priors appears to be delivering on its purpose, limitations such as the assumption that the censoring times are known a priori shadow its applicability on problems with random censoring. However clinical trials with a priori known follow up period are an example among many, where our prior is still applicable.

Another source of limitation for our methodology is the large p problems for which we are not certain on whether our proposed prior would be able to identify the correct model. Increasing the number of covariates would not allow for full enumeration of the models considered. Therefore, clever methods should be applied.

Additionally, in our work we have not studied the effect of collinearity in BVS under our proposed prior. The existence of highly collinear covariates adds an additional layer of difficulty because it causes instability to the estimated covariate effects. An additional issue that requires careful handling due to the presence of collinear covariates is that, using MCMC for variable selection would require an increased number of iterations for convergence. Furthermore, more clever or adaptive methods might be needed to avoid potential failure due to poor sampling.

A further extension of this limitation could include the existence of interactions between the different covariates making the modeling process increase its complexity.

An additional limitation regarding the nature of the covariates used in this thesis is the fact that our approach focused on purely continuous covariates. Including categorical covariates would include the existence of extra dummy variables should be treated accordingly. Also the use of different parametrization may have an affect.

Finally, our method is still limited to accelerate failure time models. Therefore, a reasonable future step would have been to blued a unit information prior for semi-parametric Cox PH models.

5.3 Future Work

The uniqueness of our approach to weighting information in AFT models that do not resemble the normal model, opens the door for future work and applications.

Extensive simulations regarding the prior derived from the Generalized F distribution, and checking how robust it is to censoring appears to be the initial step to justifying its propriety. As a natural follow up to that, model selection consistency has to be properly checked through simulations.

Additionally, we could add an extra hierarchical layer in modeling by assigning a distribution of the censoring times. This will add more structure to how this problem is handled and will give us access to the entire distribution of the censoring times allowing for further modeling.

Going beyond the univariate case like we did in this thesis could be the next step in research on the matter. Extending the method even by one dimension would require careful bivariate handling of survival times, careful handling of the correlation between them and special handling of the double censoring scheme.

Incorporating methods that account for other types of censoring (left censoring or interval censoring) would extend the applicability and use of our proposed methodology to further survival analysis scenarios. Additionally, there is room for future work regarding the nature of the censoring itself. More precisely, since in this thesis we considered only non informative censoring, a possible extension could be consider the case of informative censoring (where the censoring event provides information with regards to future survival) for completion purposes although the case of informative censoring is generally considered uncommon. An example of informative censoring under a medical framework would be that a patient is no longer followed (and hence is censored) because he has fully recovered. This will require generalizing the survival analysis methods themselves for handling such extreme cases.

Writing the Generalized F AFT model as a Cox PH model could provide insights on how our generalized prior for beta's affect or refer to the generalized semi parametric cases also although that would require careful handling in prior allocation regarding the baseline hazard.

Moving further than just model selection consistency would add more evidence that our prior is practically more theoretically sound in terms of useful properties. In particular, mathematically proving the entirety of the desiderata would add to the already stable ground of our proposed prior.

Bibliography

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, **10**(4):1100–1120.

Antoniadis, A., Fryzlewicz, P., and Letué, F. (2010). The dantzig selector in cox's proportional hazards model. *Scandinavian Journal of Statistics*, 37(4):531–552.

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

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550–1577.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418.

Berger, J. and Pericchi, L. (2004). Training samples in objective model selection. *Annals of Statistics*, 32:841–869.

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

Berger, J. O., Bayarri, M. J., and Pericchi, L. R. (2014). The effective sample size. *Econometric Reviews*, 33(1-4):197–217.

Berger, J. O. and Bernardo, J. M. (1992). On the development of the reference prior method. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 4*, pages 35–60. Oxford University Press.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, 37(2):905–938.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122.

Berger, J. O. and Pericchi, L. R. (2001). Objective bayesian methods for model selection: Introduction and comparison. *Lecture Notes-Monograph Series*, 38:135–207.

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

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, **78**(1):1–3.

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

Castellanos, M. E., Garcia-Donato, G., and Cabras, S. (2021). A model selection approach for variable selection with censored data. *Bayesian Analysis*, 16(1):271–300.

Červenỳ, J., van Ours, J. C., and van Tuijl, M. A. (2018). Effects of a red card on goal-scoring in world cup football matches. *Empirical Economics*, **55**:883–903.

Chen, M.-H., Guan, Z., Lin, M., and Sun, M. (2025). Power priors for leveraging historical data: Looking back and looking forward. *Journal of Data Science*, 23(1):1–30.

Collett, D. (2015). *Modelling Survival Data in Medical Research*. CRC Press, Boca Raton, FL, 3rd edition.

Consonni, G., Fouskakis, D., Liseo, B., and Ntzoufras, I. (2018). Prior distributions for objective bayesian analysis. *Bayesian Analysis*, 13(2):627–679.

Cox, C. (2008). The generalized f distribution: An umbrella for parametric survival analysis. *Statistics in Medicine*, 27(21):4301–4312.

Cox, C., Chu, H., Schneider, M. F., and Muñoz, A. (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine*, 26(23):4352–4374.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

De Santis, F., Mortera, J., and Nardi, A. (2001). Jeffreys priors for survival models with censored data. *Journal of Statistical Planning and Inference*, 99:193–209.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On Bayesian model and variable selection using mcmc. *Statistics and Computing*, **12**(1):27–36.

Dixon, M. and Robinson, M. (1998). A birth process model for association football matches. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **47**(3):523–538.

Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **46**(2):265–280.

Duan, Y., Ye, K., and Smith, E. P. (2006). Evaluating water quality using power priors to incorporate historical information. *Environmetrics*, 17(1):95–106.

Egidi, L. and Gabry, J. (2018). Bayesian hierarchical models for predicting individual performance in soccer. *Journal of Quantitative Analysis in Sports*, **14**(3):143–157.

Egidi, L., Pauli, F., and Torelli, N. (2018). Combining historical data and bookmakers' odds in modelling football scores. *Statistical Modelling*, **18**(5-6):436–459.

Fouskakis, D., Ntzoufras, I., and Perrakis, K. (2018). Power-expected-posterior priors for generalized linear models. *Bayesian Analysis*, 13(3):721–748.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Held, L., Gravestock, I., and Sabanés Bové, D. (2016). Objective bayesian model selection for cox regression. *Statistics in Medicine*, 35(29):5376–5390.

Hvattum, L. M. (2017). Ordinal versus nominal regression models and the problem of correctly predicting draws in soccer. *Journal homepage: http://iacss. org/index. php? id*, 16(1).

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, pages 46–60.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). The power prior: theory and applications. *Statistics in Medicine*, 34(28):3724–3749.

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

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2003). On optimality properties of the power prior. *Journal of the American Statistical Association*, 98(461):204–213.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A*, 186:453–461.

Johnson, R. A., Evans, J. W., and Green, D. W. (1983). Some models and procedures for censored survival data. *Biometrics*, 39(3):575–584.

Johnson, V. E. (2005). Bayes factors based on test statistics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):689–701.

Johnson, V. E. (2008). Properties of bayes factors based on test statistics. *Scandinavian Journal of Statistics*, 35(2):354–368.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170.

Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660.

Kalbfleisch, J. D. (1978). Non-parametric bayesian analysis of survival time data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(2):214–221.

Karlis, D. and Ntzoufras, I. (2000). On modelling soccer data. *Student*, **3**(4):229–244.

Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(3):381–393.

Karlis, D. and Ntzoufras, I. (2009). Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, **20**(2):133–145.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 2nd edition.

Kleinbaum, D. G. and Klein, M. (2012). *Survival analysis: A self-learning text*. Springer, 3rd edition. Introduction to Survival Analysis.

Koopman, S. J. and Lit, R. (2015). A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 167–186.

Lai, C.-D., Dong Lin, G., Govindaraju, K., and Pirikahu, S. (2017). A simulation study on the correlation structure of Marshall Olkin bivariate weibull distribution. *Journal of Statistical Computation and Simulation*, **87**(1):156–170.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, Hoboken, NJ, 2nd edition.

Li, Y. and Clyde, M. A. (2018). Mixtures of g-priors in generalized linear models. *Journal of the American Statistical Association*, 113(524):1828–1845.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1-2):187–192.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, **36**(3):109–118.

Nevo, D. and Ritov, Y. (2013). Around the goal: examining the effect of the first goal on the second goal in soccer using survival analysis methods. *Journal of Quantitative Analysis in Sports*, **9**(2):165–177.

Nikooienejad, A., Wang, W., and Johnson, V. E. (2020). Bayesian variable selection for survival data using inverse moment priors. *The Annals of Applied Statistics*, 14(2):809–834.

Owen, A. (2011). Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics*, **22**(2):99–113.

Prentice, R. L. (1975). Discrimination among some parametric models. *Biometrika*, 62(3):607–614.

Rue, H. and Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **49**(3):399–418.

Sabanés Bové, D. and Held, L. (2011). Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6(3):387–410.

Silva, R. M. and Swartz, T. B. (2016). Analysis of substitution times in soccer. *Journal of Quantitative Analysis in Sports*, **12**(3):113–122.

Sinha, D. and Dey, D. K. (1997). Semiparametric bayesian analysis of survival data. *Journal of the American Statistical Association*, 92(439):1195–1212.

Spiegelhalter, D. and Ng, Y.-L. (2009). One match to go! *Significance*, **6**(4):151–153.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, **64**(4):583–639.

Sprenger, J. (2013). Testing a precise null hypothesis: The case of lindley's paradox. *Philosophy of Science*, 80(5):733–744.

Stacy, E. W. (1962). A generalization of the gamma distribution. *The Annals of Mathematical Statistics*, 33(3):1187–1192.

Štrumbelj, E. and Šikonja, M. R. (2010). Online bookmakers' odds as forecasts: The case of european soccer leagues. *International Journal of Forecasting*, **26**(3):482–488.

Sturtz, S., Ligges, U., and Gelman, A. (2020). R2OpenBUGS: A Package for Running OpenBUGS from R. *R2OpenBUGS vignette*.

Taboga, M. (2021). Uninformative prior. Lectures on probability theory and mathematical statistics. Kindle Direct Publishing.

Thomas, A. C. (2007). Inter-arrival times of goals in ice hockey. *Journal of Quantitative Analysis in Sports*, **3**(3).

Tsokos, A., Narayanan, S., Kosmidis, I., Baio, G., Cucuringu, M., Whitaker, G., and Király, F. (2019). Modeling outcomes of soccer matches. *Machine Learning*, **108**(1):77–95.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. In Goel, P. K. and Zellner, A., editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. North-Holland/Elsevier.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In Bernardo, J. M. and et al., editors, *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, pages 585–603. University of Valencia.