Gaussian Processes for Survival Analysis

Tamara Fernández Aguilar

Department of Statistics

University of Oxford

A thesis presented for the degree of

Doctor of Philosophy

Merton College July, 2017

For Victor and Momo.

Acknowledgements

First and foremost, I would like to thank my family for their encouragement and love throughout all these years. I especially thank my mother Janet, my sisters Liz, Nico and Tania, and my brother Felipe. Also, I would like to acknowledge the unconditional support of my grandparents, Rene and Inelia. I am deeply grateful to my father Victor Fernández as from him I learned the importance of study. I hope my work honours his memory.

I would like to express my sincere gratitude to my supervisor Yee-Whye Teh for his continuous support and motivation. Besides my supervisor, I would like to thank my whole research group. Thanks for all the words of support and shared moments. Finally, I would like to thank all the people that, at any point, contributed with their valuable comments and suggestions about my work.

I am completely in debt with my amazing friends and superb proof readers Jessie Wu, Jovana Mitrovic, Luke Kelly, Francisca Santana and Frauke Harms. Special thanks to my Chilean family in Oxford, Nicolas Raab, Francisca Santana and Andrea Tartakowsky, for our wonderful Sunday lunches and their friendship.

Finally, I would like to acknowledge my momo, Nico, for being my family in Oxford, and hopefully in life. Thanks for being an endlessly source of inspiration, companionship and love, but above all, thanks for making me believe in hard work and to teach me to not be afraid of problems. Definitely this thesis would not have been possible without you.

This thesis was supported by Becas Chile.

Abstract

Survival analysis is an old area of statistics dedicated to the study of time-to-event random variables. Typically, survival data have three important characteristics. First, the response is a waiting time until the occurrence of a predetermined event. Second, the response can be *censored*, meaning that we do not observe its actual value but a bound for it. Last, the presence of covariates. While there exists some feasible parametric methods for modelling this type of data, they usually impose very strong assumptions on the real complexity of the response and how it interacts with the covariates. While these assumptions allow us to have tractable inference schemes, we lose inference power and overlook important relationships in the data. Due to the inherent limitations of parametric models, it is natural to consider nonparametric approaches.

In this thesis, we introduce a novel Bayesian nonparametric model for survival data. The model is based on using a positive map of a Gaussian process with stationary covariance function as prior over the so-called hazard function. This model is thoughtfully studied in terms of prior behaviour and posterior consistency. Alternatives to incorporate covariates are discussed as well as an exact and tractable inference scheme.

Contents

List of Figures

Chapter 1

Introduction

"La inspiración existe, pero tiene que encontrarte trabajando".

Pablo Picasso

1.1 Survival analysis

Survival analysis is an area of statistics concerned with the study of *time-to-event* data. Typically, survival data has three characteristics. First, the response is a waiting time until the occurrence of a predetermined event of interest. Second, the response can be *censored*, meaning that we do not observe its actual value but a bound for it. Last, the presence of covariates.

Data of this kind can be encountered in a broad range of areas. Some examples include, but are not limited to: failure times in mechanical systems, time to death of patients in a clinical trial or duration of unemployment in a population. Hence, the applicability and the study of new survival methods remains of keen interest for the statistical and machine learning communities. This thesis aims to contribute to the literature of survival analysis by introducing and studying a novel nonparametric Bayesian model for survival data.

1.1.1 Elements of survival analysis

Let T be a positive continuous random variable denoting the time until some event of interest occurs. Additionally, denote by $f(t)$ and $F(t)$ the associated probability density function (p.d.f.) and cumulative distribution function (c.d.f.), respectively. The primary component in survival analysis is the *survival function* defined as

$$S(t) \quad = \quad 1 - F(t). \tag{1.1.1}$$

This function denotes the probability that the event of interest will not happen before a specified time t. Typically, this is the object we want to estimate.

Another key element of interest in survival analysis is the so-called *hazard function* $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$, which is defined as

$$\lambda(t) \quad = \quad \lim_{dt \to 0} \frac{\mathbf{P}(t \leq T < t + dt | T \geq t)}{dt}.$$

Here, the numerator stands for the conditional probability that the event happens in the time interval $[t, t + dt)$ and the denominator corresponds to the width of such an interval. The division of these terms gives us the notion of a rate of occurrence per unit of time. Furthermore, taking the limit when the width of the interval goes to zero, results in an instantaneous rate of occurrence. In our particular case of continuous random variables this limit can be expressed as the division between the probability density function and survival function, i.e.,

$$\lambda(t) = \frac{f(t)}{S(t)}. \tag{1.1.2}$$

Based on this, we define the *cumulative hazard function* (or cumulative risk)

$\Lambda : \mathbb{R}_+ \to \mathbb{R}_+$ as the integral up to time t of the hazard function, i.e.,

$$\Lambda(t) = \int_0^t \lambda(s)ds. \tag{1.1.3}$$

Observe from equation (1.1.1) that the hazard function $\lambda(t)$ corresponds to the derivative of $-\log S(t)$, thus an alternative way of writing the survival function in terms of the hazard function is given by the following equation

$$S(t) = \exp\{-\Lambda(t)\}. \tag{1.1.4}$$

This representation and the fact that $F(t)$ denotes a cumulative distribution function allows us to deduce $F(t) \overset{t\uparrow\infty}{\to} 1$ if and only if $\Lambda(t) \overset{t\uparrow\infty}{\to} \infty$. Finally, from equations (1.1.2) and (1.1.4), it holds

$$f(t) \quad = \quad \lambda(t)\exp\{-\Lambda(t)\}. \tag{1.1.5}$$

1.1.2 Censoring

At a superficial level, survival analysis might look as a standard problem with data in a particular range and it is natural to think that the problem can be handled directly by well-understood statistical methods for analysing continuous (or discrete) data. The reason why standard methods do not work is due to a fundamental problem that generally occurs when dealing with survival data, namely, incomplete information. For example, suppose a company wants to estimate the time until some product fails. For that, suppose that they allow a two-year period to follow-up some costumers. During this follow-up period a product can either fail or not, but the fact that we do not observe the product failing does not ensure us that it may not fail right after we stop observing it. Additionally, it may also happen that some costumers become unavailable, making impossible to follow the state of those products. Finally, it is quite likely that costumers do not remember

the exact day of the failure and only have an estimate of it. Hence, the data we collect is a mixture of complete and incomplete information and this is the major difference compared to most other statistical data problems.

All these types of incomplete observations are termed censored survival times. Throughout this thesis, we consider four types of censoring: uncensored data, right censoring, left censoring and interval censoring. In particular, these types of censoring involve the notion of a censor time that bounds the survival time, as shown in Figure 1.1. Additionally, we assume non-informative censoring, i.e., we assume that for each observation the censor time is independent of the event time.

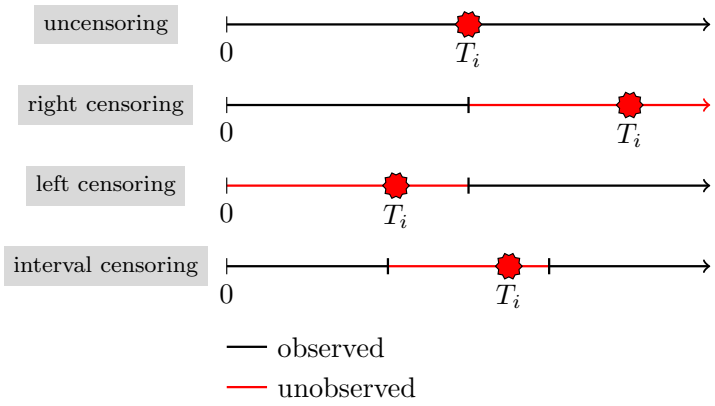

Figure 1.1: Different types of censoring for a random survival time T_i.

Formally, let $T_1, \ldots, T_n$ be a sequence of independent and identically distributed (i.i.d.) random survival times, then each T_i can behave as follows: the simplest scenario is the uncensored case in which we observe the actual value of the random variable T_i. Right censored stands for when we do not observe the actual value of the random variable T_i but a lower bound t for it, i.e., $T_i > t$. Observe that by the non-informative censoring assumption the lower bound t can be either, the realization of a random variable G_i which is independent of T_i, or a constant value. Respectively, left censoring and interval censoring arise when we observe either an upper bound t of T_i, or an interval (t_l, t_r) in which T_i falls.

1.1.3 Related methods

There exists a vast literature of statistical models arising in survival analysis. These modelling approaches will be different depending on whether we do or do not have covariates. This is due to the fact that the presence of covariates typically changes the target of the inference mechanism. While for a setting without covariates the main objective is to give answer to distributional questions about the survival data, in the setting with covariates we usually want to assess the effect of a particular covariate on the distribution of the survival times, but we are not necessarily interested on the distribution itself. In this brief review, we give an insight of the most widely-known modelling approaches, distinguishing between frequentist and Bayesian methods.

A first approach to survival analysis comes from the counting processes theory. Here, the two principal representatives are the Kaplan-Meier [34] and the Nelson-Aalen [1] estimators, which are, arguably, the most popular estimators in survival analysis. The counting processes approach ([27],[3]) to survival analysis is an extension of the theory of empirical processes to censored data. Assuming standard conditions on the data, by a simple application of this theory, we can deduce the Nelson-Aalen estimator to estimate the cumulative hazard function, and the Kaplan-Meier estimator to estimate the survival function. Furthermore, this theory allows us to study several properties of these models such as bias, large sample properties, consistency, etc.. Among these properties, it is worth to mention the impressive result that both, the Kaplan-Meier and the Nelson-Aalen are the nonparametric maximum likelihood estimators for their corresponding quantities ([32],[3, section IV.1.5]). Even though the counting process approach is quite technical, the Kaplan-Meier and Nelson-Aalen estimators are quite simple to implement, making them an excellent starting point for practitioners. Beyond the two classical estimators, the counting process theory allows us to derive and study several nonparametric tests straightforwardly, examples include the log-rank test

[45], the Gehan-Breslow test [23], [8], and the Harrington-Flemming test [28]. In particular, we remark that the log-rank test is the most popular test among practitioners. A drawback of the counting process approach is that it is not clear how to treat covariates.

Another approach to model survival data is to assume a parametric family for the probability distribution function P associated with the survival time T, that is, to consider $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^d$ denotes the parameter space. Compared with the previous approaches, this one has the immediate drawback that we constrain our solution to the family given by $\mathcal{P}$. Nevertheless, it gives us a way of introducing covariates by simply indexing the distributions on them, i.e., by considering $P_{\theta,\boldsymbol{x}}$. Observe that specific choices of certain parametric families can add further constraints to the model. This is not necessarily a critical shortcoming if the assumptions made are feasible for the data, nevertheless, extra care must be taken. Examples of this approach include any type of regression model for survival data, e.g. [44], and the so-called *accelerated failure rate models* [60]. Accelerated failure rate models consider a parametric form to model survival times but set a rescaling factor for the time, depending on the covariates, which accelerates or decelerates the hazard rate. Another example is the *threshold regression model* [40]. This model has a nice physical interpretation as the survival times are represented by hitting times in which some latent process, associated with the event of interest, reaches some boundary. For instance, the latent process can be the overall development of a fetus and the boundary is a minimum level of hormones and physical development that triggers labour. Both frequentist and Bayesian paradigms can be adopted for making inference in these modelling approaches.

Another well-known method that successfully incorporates covariates is the Cox proportional hazards model [13]. Arguably, this model together with the Kaplan-Meier estimator and the Log-rank test constitute the three pillars of the practice of survival analysis. This model differentiates from the approaches above

as it usually focuses on assessing the effect of a covariate on the distribution over the survival times while ignoring the distribution itself. It achieves this by imposing a proportional relation between the hazard functions of different levels of a covariate. Indeed, in this setting, the hazard function takes the form $\lambda_{\boldsymbol{x}}(t) = \lambda_0(t)e^{\beta^\intercal \boldsymbol{x}}$ for a time t and a set of covariates $\boldsymbol{x}$. The estimation of the parameter β can be performed without any consideration of the hazard function by using a partial likelihood instead of a full likelihood. See [12] for more details.

We proceed to review some well-known Bayesian nonparametric methods in survival analysis. Since the introduction of the Dirichlet process as prior over probability measures, multiple nonparametric discrete priors have been introduced and applied to survival models. Some examples are the Dirichlet processes [54], neutral-to-the-right (NTR) processes [16], the extended Gamma process [19], the beta process [29], and the beta-Stacy process [59]. A comprehensive account of these priors can be found in [30].

In general, these priors have appealing properties given right censored data. Indeed, NRT processes [20], the extended Gamma process [19], the beta process [29] and the beta-Stacy process [59] are all known to be conjugate under right censoring. For the Dirichlet process, we do not have conjugacy as the posterior given right censored data corresponds to a mixture of Dirichlet processes [6]. Nevertheless, the Dirichlet process is conjugate under right censoring when seen as a member of the class of NTR priors. Furthermore, NTR priors are still conjugate given right censored data when considering a proportional hazard term for covariates. This can be understood as a Bayesian nonparametric version of the Cox model ([33], [29],[39]). An even stronger version of the previous result, including independent left censoring, is proved in [36].

A more recent work which gives a concrete and fast inference scheme is the ANOVA-DDP proposed in [14]. Here, the authors propose a prior over distribution functions by mixing ideas of ANOVA and Dirichlet processes. This results in a flexible model that can incorporate covariates in an ANOVA-type fashion and still

have a good performance estimating the distribution of the survival times.

Within the context of Gaussian processes a few models have been considered, e.g. [46] and [31]. While these models propose an explicit framework to introduce covariates, they are constrained to the proportional hazards assumption. Another more flexible model that was recently proposed is [4]. In this work, the authors propose an accelerated failure model in which the dependence of the failure times on covariates is modelled by rescaling the time by a function of the covariates. The nonparametric part comes to play when considering a Gaussian process prior over said function.

1.2 Nonparametric Bayesian inference

Let $T_1, \ldots, T_n \in \mathbb{R}_+$ be a set of realisations of some collection of random variables associated with an experiment of interest. A first step towards building a statistical model is to consider a joint probability distribution $P^{(n)}$ for the data and assume that this probability distribution is contained within a known class of probability distributions $\mathcal{P}$.

Typically, we characterise such a class of distributions by indexing it by a parameter θ belonging to some measurable parameter space Θ, that is, we consider $\mathcal{P} = \{P_\theta^{(n)} : \theta \in \Theta\}$. In particular, the nature of the parameter θ allows us to make a distinction between two types of models. We say a model is parametric when the parameter θ denotes a finite-dimensional object, e.g. a real number or a finite-dimensional vector. On the other hand, we say that a model is nonparametric when θ represents an infinite-dimensional object, e.g. a probability density or a hazard function.

The next step is to perform inference about the unknown parameters. This can be done by following either a frequentist or a Bayesian approach. In the frequentist setting, we assume the existence of an unknown true parameter value $\theta_0 \in \Theta$ which we need to learn from the data. For the Bayesian counterpart, there

is no notion of a true parameter, hence the uncertainty about the distribution of the data is quantified by defining a prior probability distribution over the space Θ. The choice of the prior probability is a crucial step in Bayesian modelling as it represents the initial belief of the Bayesian statistician before observing the data.

Let Π denote a prior probability distribution on Θ. A Bayesian statistician updates her belief by considering the conditional distribution of θ given the observations $\boldsymbol{T^{(n)}} = (T_1, \ldots, T_n)$. This updated probability distribution over θ can be obtained by using Bayes' formula

$$\Pi(\theta \in B | \boldsymbol{T^{(n)}} \in A) \;=\; \frac{\int_B P_\theta^{(n)}(A)\Pi(d\theta)}{\int_\Theta P_\theta^{(n)}(A)\Pi(d\theta)}, \tag{1.2.1}$$

and it is commonly referred to as the posterior distribution of θ given the event $\boldsymbol{T^{(n)}} \in A$ for any measurable set A. By assuming $P_\theta^{(n)}$ is dominated by the Lebesque measure on $\mathbb{R}^n$ for θ in a set of positive prior probability, we write this posterior in terms of the density $p_\theta^{(n)}$

$$\Pi(\theta \in B | \boldsymbol{T^{(n)}}) \;=\; \frac{\int_B p_\theta^{(n)}(\boldsymbol{T^{(n)}})\Pi(d\theta)}{\int_\Theta p_\theta^{(n)}(\boldsymbol{T^{(n)}})\Pi(d\theta)}. \tag{1.2.2}$$

As Bayesian statisticians, we expect that the posterior distribution is well-behaved as the data grows. A known frequentist method to assesses the behaviour of a posterior distribution is posterior consistency. In particular, this methodology assumes the existence of a true parameter θ_0 that generates the data and then assesses whether the posterior concentrates around this true parameter asymptotically. While this validation method will not make sense to a fully Bayesian statistician, as for her the notion of a true parameter that generates the data is incorrect, it gives us a frequentist validation of our methods. Below, we consider the case of i.i.d. random variables unless we explicitly state otherwise.

1.2.1 Posterior consistency

We say a posterior distribution is consistent around a true parameter $\theta_0 \in \Theta$ with respect to a metric d on the parameter space Θ, if for all $\epsilon > 0$, it holds that

$$\Pi(d(\theta, \theta_0) > \epsilon | \boldsymbol{T^{(n)}}) \to 0 \quad \text{as} \quad n \to \infty, \tag{1.2.3}$$

either in $P_{\theta_0}^{(n)}$-probability or $P_{\theta_0}^{(n)}$-almost surely.

Doob's theorem [17] gives the first well-known result regarding posterior consistency for nonparametric models. It states that the posterior is consistent at every possible true parameter $\theta_0 \in \Theta$, except for a set of Π-measure zero. Despite this is good result, consistency can fail when the parameter belongs to this set of null probability and this theorem gives no guidance about the shape of this null set.

A solution to the problem above is given by Schwartz's theorem [53]. Arguably, this theorem is the most important and most used result for proving posterior consistency for nonparametric Bayesian models. Schwartz's theorem gives sufficient conditions on the prior Π on Θ and the true parameter $\theta_0 \in \Theta$ for proving posterior consistency. This theorem is stated below.

Theorem 1.2.1. *[37] Suppose that p_{θ_0} belongs to the Kullback-Leibler (KL) support of the prior Π, and for every neighbourhood $\mathcal{U}$ (with respect to d) of p_{θ_0} there exists a sequence of test functions ϕ_n such that $\mathbf{E}_{\theta_0}\phi_n \leq e^{-Cn}$ and $\sup_{p_\theta \in \mathcal{U}^c} \mathbf{E}_\theta(1 - \phi_n) \leq e^{-Cn}$ for some positive constant C. Then, the posterior distribution of p, that is, $\Pi(\cdot|X_1, \ldots, X_n)$ for the model $X_1, \ldots, X_n | p \overset{i.i.d}{\sim} p$ and $p \sim \Pi$, is $P_{\theta_0}^{(n)}$-almost surely consistent at p_{θ_0}.*

The first condition asserts that the density under the true parameter θ_0 must belong to the KL-support of the prior, that is, for every $\epsilon > 0$,

$$\Pi\left(\theta : \int_0^\infty \log \frac{p_{\theta_0}}{p_\theta}(t) p_{\theta_0}(t) dt < \epsilon\right) > 0. \tag{1.2.4}$$

The second condition requires the existence of statistical tests ϕ_n with exponentially small type I and II error. A variation of Schwartz's theorem providing an alternative KL condition is given in [5]. Extensions to the case of independent but not identically distributed observations can be found in [11] and in [10], for the case of consistency in probability and almost surely consistency, respectively.

Results in posterior consistency

Here, we review some well-known results regarding posterior consistency for Bayesian nonparametric models. For the case of discrete process priors, a few but powerful results are known. In particular, [26] shows posterior consistency given right censored data under a Dirichlet process prior. A more general result was obtained in [35]. In this work, the authors give sufficient conditions for which we achieve posterior consistency for NTR priors given right censored data. From the latter result, it is deduced that most of the popular priors such as the Dirichlet process, the beta process and the Gamma process have consistent posteriors in the survival analysis setting.

Up to the best of our knowledge, within the setting of continuous priors in survival analysis, we are not aware of results stating posterior consistency. Nevertheless, in the setting of density estimation, which can be seen as a simplification of the problem of survival analysis given uncensored data, there are a few known results.

In [43], a prior over probability measures was constructed by convolving a continuous kernel with the Dirichlet process. Posterior consistency for this prior was proved in [24] in the framework of density estimation. Another example, is given in [56]. In this work, the authors prove posterior consistency for a density estimation model with a Gaussian process prior introduced in [41]. There are some alternatives approaches for modelling conditional distributions, i.e., data with covariates. In [49], the authors provide sufficient conditions to prove posterior consistency for

a broad family of priors, which are formulated as predictor-dependent mixture of Gaussian kernels.

General results for Gaussian processes priors are accounted in [58] and [57]. In particular, in [57], the authors study an abstract model and prove posterior consistency for it. Such result can be applied to density estimation, regression and classification problems among others. Other relevant works proving posterior consistency for models under Gaussian process priors, and which are not related with the problem of density estimation or survival analysis, are Choi and Schervish [10] in the setting of regression and Ghosal and Roy [25] in the setting of logistic regression.

1.3 Thesis contributions and organization

This thesis proposes a novel Bayesian nonparametric model for survival data. This is done by constructing a prior over hazard functions in terms of a stochastic process which, in turn, is defined as a positive map of a stationary Gaussian process.

Each chapter of this thesis contributes to the study of the proposed model by covering a particular subject related to the definition, implementation or properties of this model. Since this thesis covers a wide range of different topics in statistics, we make each chapter as self-contained as possible, that is, any result or theorem used in this thesis will be reviewed and cited in the respective chapter it is used. This thesis is mainly focused on methodological aspects of the proposed model. Nevertheless, an experimental section with empirical results is included for completeness. We expect, throughout the chapters of this thesis, to give the reader a better understanding of the benefits and limitations of the proposed model.

1.3.1 Chapter 2

In this chapter, we introduce a nonparametric Bayesian model for survival data. This model is based on defining a prior over the so-called hazard function. This prior is constructed by multiplying a parametric baseline hazard function λ_0 by a positive map of a stationary Gaussian process, i.e., $\lambda(t) = \lambda_0(t)\sigma(l(t))$, where $(l(t))_{t\geq 0}$ is a stationary Gaussian process and σ denotes the sigmoid function. Given a sequence $T_1, \ldots, T_n$ of i.i.d. random survival times, this can be expressed in the following hierarchical model

$$\begin{aligned} T_i|\lambda &\overset{i.i.d.}{\sim} f_\lambda, \quad i = 1, \ldots, n, \\ f_\lambda(t) &= \lambda(t)e^{-\int_0^t \lambda(s)ds}, \\ \lambda(t) &= 2\lambda_0(t)\sigma(l(t)), \\ (l(t))_{t\geq 0} &\sim \mathcal{GP}(0, K), \end{aligned}$$

where K denotes a stationary covariance function.

We propose two ways of including covariates. The first is motivated by a proportional hazards model, where covariates are included by multiplying a solely time dependent hazard function by a positive map of a linear predictor. The second approach considers defining the hazard function as a positive map of a linear predictor with time varying coefficients, i.e., for a covariate $\boldsymbol{x} \in \mathbb{R}^d$, we define $\lambda_{\boldsymbol{x}}(t) = \sigma(\beta_0(t) + \beta_1(t)x_1 + \ldots + \beta_d(t)x_d)$, and consider independent Gaussian process priors for each time dependent coefficient. The latter model can be interpreted as a linear regression with time-varying coefficients.

1.3.2 Chapter 3

In this chapter, we study properties of the prior. In particular, we find sufficient conditions over the kernel of a stationary Gaussian process and a baseline hazard function $\lambda_0(t)$ for which the prior defines proper hazard functions with

probability one. For our setting, a hazard function $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ is said to be proper if λ is continuous and $\Lambda(t) \overset{t\uparrow\infty}{\to} \infty$ almost surely.

This result is of crucial importance as it ensures the model is well-defined. Indeed, $\Pi(\Lambda(t) \overset{t\uparrow\infty}{\to} \infty) < 1$ implies that there exists a set of positive prior probability for which $\lim_{t\to\infty} F(t) < \infty$, i.e., a random survival time T can be such that $T = \infty$ with positive probability. While this property is not totally unreasonable, for instance, some people never marry and thus this event never occurs, we are interested on building a prior over events that are bound to happen. Therefore, this property must be satisfied with probability one.

The proof technique proposed for proving this result uses the continuity of the Gaussian process, which needs to be checked, together with concentration inequalities for the cumulative hazard function. We conclude the chapter by studying some properties of the random moments of the random survival times T which can be derived by using the same proof techniques.

1.3.3 Chapter 4

In this chapter, we address all the computational issues related with sampling from the posterior distribution of the proposed model. In particular, the inference scheme is based on a sampler for modulated renewal processes proposed by [51]. In turn, the latter result corresponds to a refined version of the algorithm presented in [2] for sampling from Gaussian Cox processes. Based on these results, we develop a tractable exact inference scheme for sampling from the posterior distribution of the hazard functions.

Additionally, we adapt our version of the algorithm to perform inference and to deal with right censoring, left censoring and interval censoring. Left and interval censoring require the augmentation of the actual survival times. We propose two algorithms to do so, one based on a rejection sampler and another based on a Gibbs sampler.

Later, to make the algorithm scalable, we introduce random Fourier features to approximate the Gaussian process and we supply the corresponding inference algorithm. Posterior distributions are also derived in this chapter.

This chapter is based on our paper [21].

1.3.4 Chapter 5

In chapter 5, we aim to give an empirical evaluation of the model based on simple yet informative experiments. Our first experiment considers a density estimation problem with synthetic data. For our second evaluation, we replicate the experiment proposed in [14], in which the authors propose a setting of crossing hazards. Additionally, we consider the same experiment with right censored data. Finally, we evaluate our method in a real dataset, and compare with competing approaches.

This chapter is also based on our paper [21].

1.3.5 Chapter 6

In Chapter 6, we identify sufficient conditions to establish posterior consistency on our model. We prove that under certain assumptions there exists a true distribution from which the survival times originated, then with enough data we can recover such distribution. This is a basic statistical property that any well-founded model should have as it guarantees that we obtain reliable estimates as the data grows.

From a theoretical point of view, the main difficulty and novelty of this work is the consideration of survival lifetimes on $\mathbb{R}_+$, removing the requirement for practitioners to know and specify a compact lifetime domain in advance. To the best of our knowledge, most of the previous works proving posterior consistency for models under a Gaussian process priors consider data on a compact set, and most of the techniques used can not be extended to our framework. Therefore,

our work not only proves posterior consistency of the model but also gives a new proof technique for data in $\mathbb{R}_+$.

This chapter is based on our yet unpublished manuscript [22].

1.4 Notation

The chapters of this thesis are mostly self-contained and they introduce their own notation, nevertheless some notation is shared among all the chapters.

We usually use the letters s and t to denote times in $\mathbb{R}_+ = [0, \infty)$. Bold letters $\mathbf{x} = (x_1, \ldots, x_d)$ and $\mathbf{y} = (y_1, \ldots, y_d)$ refer to d-dimensional covariates associated with the data. We usually denote the probability density function, cumulative density function and survival function of a random variable by f, F and S, respectively. Furthermore, we write the hazard function and the cumulative hazard function with the letters λ and Λ, respectively.

In this thesis, we use the following asymptotic notation. Given two functions $g, h : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} = \mathbb{R}_+$ or $\mathbb{N}$, we say that $g = \mathcal{O}(h)$ if and only if it exists a constant $C > 0$ and $T \in \mathcal{X}$ such that for all $t \geq T$,

$$\left| \frac{g(t)}{h(t)} \right| \leq C,$$

and, we say that $g = o(h)$, if and only if

$$\lim_{t \to \infty} \left| \frac{g}{h} \right| = 0.$$

We use the notation $g \ll h$ to denote $g = o(h)$. Lastly, we say that $g = o(h)$ when $t \to 0$ to indicate that $g(t)/h(t)$ tends to 0 as t approaches 0. We denote by Θ the parameter space and by Π the prior distribution on Θ. We use $\mathbf{P}$, $\mathbf{E}$, $\mathbf{Var}$ and $\mathbf{Cov}$ to denote probability, expected value, variance and covariance, respectively. We denote by $\mathcal{GP}(0, K)$ a Gaussian process with zero mean and kernel K.

Chapter 2

A model for survival data

2.1 Model

In this chapter, we introduce a novel Bayesian nonparametric model for survival data. Our approach is motivated by the aim of giving practitioners flexible, yet interpretable statistical models that can be related to ubiquitous distributions in survival analysis.

A common way of building a statistical model given a sequence of i.i.d. random variables is to assume some parametric family for their common density function. In survival analysis, this is equivalent to assume a parametric family for the hazard function. While parametric approaches usually impose strong assumptions about the true nature of the data, they are a very useful and appealing resource in the case when there is enough information about the true nature of the data, so the imposed constraints can be analysed and then verified. Indeed, the main reason to prefer parametric models over other methods is their ease of implementation and ease of understanding, making them popular among practitioners.

On the other hand, when the data is too complex or not completely understood, parametric approaches can be unsuitable for modelling. In this setting, nonparametric approaches become an appealing alternative as they usually im-

pose very few constraints, and then, not much knowledge about the data is required. While nonparametric methods propose flexible and interesting models, they typically lack of interpretation and are harder to understand and to implement.

This chapter proposes a statistical model for survival data lying in a middle ground between the parametric and the nonparametric approach. In particular, we work under the Bayesian paradigm in which prior knowledge about the distribution of the data, or other constraints, can be incorporated by specifying an appropriate prior. The flexibility of the model can be assured by specifying a prior with large support.

Our proposed model considers a parametric family for the hazard function multiplied by a nonparametric *correction* term. In a nutshell, we consider hazard functions of the form $\lambda(t) = \lambda_0(t)v(t)$, where λ_0 is a parametric hazard function chosen by the user, and v is a random positive function with mean 1. This choice of v ensures that on average $\lambda(t)$ is centred at $\lambda_0(t)$. Here, v represents the nonparametric part of the model.

While there are many ways to choose the random function v, it is hard to select one that makes sense in a practical way, and such that it is also computationally tractable. In an attempt to stay close to the classical and well-behaved processes, we choose $v(t) = 2\sigma(l(t))$, where $(l(t))_{t\geq 0}$ is a continuous Gaussian process with mean 0, and $\sigma(x) = (1+e^{-x})^{-1}$ is the sigmoid function. By the symmetry of σ, it is not hard to see that $\mathbf{E}(2\sigma(l(t))) = 1$ for all $t \in \mathbb{R}_+$. The idea above is expressed in the following hierarchical model

$$\begin{aligned} T_i|\lambda &\overset{ind.}{\sim} f_\lambda, \quad i = 1, \ldots, n, \\ f_\lambda(t) &= \lambda(t)e^{-\int_0^t \lambda(s)ds}, \\ \lambda(t) &= 2\lambda_0(t)\sigma(l(t)), \\ (l(t))_{t\geq 0} &\sim \mathcal{GP}(0, K), \end{aligned} \tag{2.1.1}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$, denotes the sigmoid function. The second line in the above model is given by the relationship between the density and hazard function given in equation (1.1.5), that is, $f(t) = \lambda(t)e^{-\int_0^t \lambda(s)ds}$.

The main advantage of this model is that it can be easily explained and interpreted even by practitioners that are not familiar with Bayesian nonparametric methods. Indeed, the model we propose is just a noisy version of their preferred parametric model denoted by $\lambda_0(t)$. In other words, the proposed model resembles their preferred parametric model for survival data but it also accounts for more flexibility by considering the noise term.

Remark 2.1.1 (Well-definiteness of the prior)**.** The above model relies on defining a prior probability distribution on the space of hazard functions. Since a hazard function λ is a well-known mathematical object, we need to ensure that the prior we are considering takes into account the nature of λ.

For our setting we consider two key properties to be satisfied by random hazard functions. In particular, a random hazard function λ must be a continuous and positive function, and it must satisfy $\Lambda(t) = \int_0^t \lambda(s)ds \overset{t\uparrow\infty}{\to} \infty$, in order to ensure that $S(t) \overset{t\uparrow\infty}{\to} 0$. For now, we assume that the previously defined prior effectively generates proper hazard functions and we focus on the modelling aspects. Sufficient conditions to ensure this result will be given later in Chapter 3.

2.2 Likelihood

We give an expression for the likelihood based on a sequence $T_1, \ldots, T_n$ of i.i.d. random survival times. Observe that since censoring is a property of the data, the likelihood must account for this. For the uncensored case, the likelihood function

is

$$L(\lambda|\{T_i\}_{i=1}^n) = \prod_{i=1}^n f_\lambda(T_i) = \prod_{i=1}^n \lambda(T_i) e^{-\int_0^{T_i} \lambda(s)ds}. \tag{2.2.1}$$

In the case of right censored data, the probability of observing a time event beyond T_i is best represented by the survival function evaluated at this time, that is, the contribution to the likelihood is given by $S(T_i)$. Following this line of thought, the contribution of a left censored observation is $F(T_i)$. Lastly, for an interval censored observation, the contribution of the likelihood is given by $S(T_i^l) - S(T_i^u)$, where T_i^l and T_i^u denote a lower and upper bound for the actual time of the event.

In order to ease the notation, we introduce an indicator variable δ_i that denotes the different types of censoring. In particular, we define $\delta_i = 1$ for an uncensored survival time T_i, $\delta_i = 0$ for right censoring, $\delta_i = 2$ for left censoring and lastly $\delta_i = 3$ for interval censoring. Hence, the general form of the likelihood function under censoring is given by

$$\begin{aligned} L(\lambda|\{(T_i,\delta_i)\}_{i=1}^n) &= \prod_{i:\delta_i=1} \lambda(T_i) e^{-\int_0^{T_i} \lambda(s)ds} \prod_{i:\delta_i=0} e^{-\int_0^{T_i} \lambda(s)ds} \prod_{i:\delta_i=2} 1 - e^{-\int_0^{T_i} \lambda(s)ds} \\ &\times \prod_{i:\delta_i=3} e^{-\int_0^{T_i^l} \lambda(s)ds} - e^{-\int_0^{T_i^u} \lambda(s)ds}. \end{aligned} \tag{2.2.2}$$

2.3 Adding covariates

So far we have assumed that all the survival times are identically distributed. This assumption is unrealistic in many applications. For instance, the survival times for patients with lung cancer might be different across different ages and depending on their smoking history. We propose two alternatives for modelling the relationship between time and covariates.

2.3.1 Proportional hazards approach

Recalling the proportional hazards model of Cox [13], a first approach to introduce dependence on a covariate $\boldsymbol{x} \in \mathbb{R}^d$ is to assume a hazard function of the following form

$$\lambda_{\boldsymbol{x}}(t) = 2\lambda_0(t)\sigma(l(t))e^{\beta^\intercal \boldsymbol{x}}. \tag{2.3.1}$$

Here, $\beta = (\beta_1, \ldots, \beta_d)^\intercal$ is the vector of regression coefficients, and $2\lambda_0(t)\sigma(l(t))$ accounts for the hazard function. The main difference with the standard implementation of the Cox proportional hazards model is that in this set-up we jointly estimate the hazard function and the coefficients β.

2.3.2 A Gaussian process on time and covariates

An alternative way of introducing dependence on covariates is by considering a Gaussian process whose kernel is indexed in both time and covariates. In general, a nice and simple way to generate kernels in this way, is to construct kernels for each covariate and time separately and then perform basic operation on them, i.e., addition and multiplication. We refer to the work of [18] for details of the construction and interpretation of operations between kernels.

In particular, we use the following construction that considers operations between a general stationary kernel for the time and a linear kernel for the covariates. Indeed, let $\boldsymbol{x}$ and $\boldsymbol{y}$ be covariates taking values on $\mathbb{R}^d$, and let t and s denote times taking values on $\mathbb{R}_+$. Then, we define a kernel K, in time and covariates, as

$$K((t, \boldsymbol{x}), (s, \boldsymbol{y})) \;=\; K_0(t, s) + \sum_{j=1}^{d} x_j y_j K_j(t, s), \tag{2.3.2}$$

where $K_j : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ is a stationary kernel function for $j = 1, \ldots, d$. Observe that the above expression effectively defines a kernel as it corresponds to the kernel

of a well-defined Gaussian process. Indeed, let $(l_j(t))_{t\geq 0}$ with $j = 0, \ldots, d$ denote a collection of independent Gaussian processes with stationary kernels K_j, then

$$l(t, \boldsymbol{x}) \quad = \quad l_0(t) + \sum_{j=1}^{d} x_j l_j(t), \tag{2.3.3}$$

is a Gaussian process with kernel function defined by equation (2.3.2).

We decided to use this particular choice of kernel as it is simple, yet rich enough to capture interesting relationships of the data. Indeed, as this model defines a hazard function that depends on the covariates through a linear regression equation with time-varying coefficients, it is possible to model interactions of time and covariates, e.g. crossing hazards. More complex dependences can be incorporated to the model through the linear regressor, e.g. interactions between two or more covariates.

Chapter 3

Properties of the prior

3.1 Hazard function

The model defined in equation (2.1.1) of Chapter 2 builds on defining a prior on the space of hazard functions. This prior is constructed by considering a fixed baseline hazard function λ_0 and a multiplicative nonparametric noise given by a centred and stationary Gaussian process $(l(t))_{t\geq 0}$, with covariance function denoted by K, i.e.,

$$\lambda(t) = 2\lambda_0(t)\sigma(l(t)), \quad t \geq 0, \tag{3.1.1}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$.

So far, we have assumed that this prior is well-defined, nevertheless, no formal result has been given about this statement. This chapter aims to provide a proper study of the prior and, through this, to identify conditions over the Gaussian process $(l(t))_{t\geq 0}$ and the baseline hazard function λ_0 for which the prior generates proper survival functions.

We say a hazard function $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ is proper if λ is continuous and the cumulative hazard function satisfies $\Lambda(t) \overset{t\uparrow\infty}{\to} \infty$. This result is very important as it ensures the model is well-defined. Indeed, $\Pi(\Lambda(t) \overset{t\uparrow\infty}{\to} \infty) < 1$ implies that there

exists a set of positive prior probability for which $\lim_{t\to\infty} F(t) < 1$, i.e., a random survival time T can be such that $T = \infty$ with positive probability. While this property is not totally unreasonable, for instance some people never marry and thus this event never occurs, we are interested in building a prior over events that are bound to happen. Therefore, this property must be satisfied with probability one.

The first property we verify is the continuity of the random hazard functions sampled from the prior. In particular, by equation (3.1.1), sufficient conditions for generating continuous sample hazard functions are that the baseline hazard function λ_0 and the Gaussian process $(l(t))_{t\geq 0}$ are continuous. Results proving continuity of stationary Gaussian processes can be found in [42, Chapter 2]. In particular, example 2.2 [42, Chapter 2] gives the following useful proposition, which covers most of the interesting and common cases.

Proposition 3.1.1. *Let $(l(t))_{t\geq 0}$ be a Gaussian process with stationary kernel $K(t) = \mathbf{Cov}(l(t+s), l(s))$. If K satisfies*

$$K(t) = K(0) - C|t|^{\alpha} + o(|t|^{\alpha}), \tag{3.1.2}$$

for some $C > 0$, as $t \to 0$, and $\alpha \in (0, 2]$, then the Gaussian process $(l(t))_{t\geq 0}$ has continuous sample paths [1].

Example 3.1.2. Consider a stationary Gaussian process $(l(t))_{t\geq 0}$ with zero mean and stationary kernel function given by $K(t) = 1/(1+t)$. Then, by Taylor, the kernel K admits the following representation

$$K(t) \quad = \quad \sum_{n=0}^{\infty} (-t)^n = 1 - t + o(t), \tag{3.1.3}$$

as t approaches zero. Therefore, by proposition 3.1.1, we conclude that $(l(t))_{t\geq 0}$ has continuous sample paths.

[1] It means that there exits a continuous version of the Gaussian process $(l(t))_{t\geq 0}$.

Proposition 3.1.1 gives us a very intuitive result. The notion of continuity of a stationary Gaussian process is related to the behaviour of the kernel $K(t)$ when $t \to 0$. Indeed, for a fixed $s > 0$, $\epsilon > 0$ and $t \to 0$

$$\mathbf{P}(|l(s+t) - l(s)| > \epsilon) \leq \frac{\mathbf{E}(|l(s+t) - l(s)|^2)}{\epsilon^2} = \frac{2(K(0) - K(t))}{\epsilon^2}. \tag{3.1.4}$$

Then, if $K(t)$ approaches $K(0)$ fast enough as $t \to 0$, $l(s+t)$ and $l(s)$ will be close enough with high probability. On the other hand, by rearranging the terms of equation (3.1.2) and taking $\alpha = 1$, we have

$$\frac{K(t) - K(0)}{|t|} = -C + o(1) \tag{3.1.5}$$

as $t \to 0$ for some $C > 0$. Thus, a sufficient condition to satisfy this result is that K is differentiable. Important examples of stationary kernels satisfying this condition and thus generating continuous Gaussian processes are the squared exponential $K_b(t) = e^{-|t|^2/b^2}$ and the Ornstein-Uhlenbeck $K_b(t) = e^{-|t|/b^2}$ among others.

Next, we prove the second property, that is, $\Lambda(t) \overset{t\uparrow\infty}{\to} \infty$. In order to do this, we make two essential assumptions. The first assumption is that λ_0 must be a proper hazard function, i.e., continuous and such that $\Lambda_0(t) \overset{t\uparrow\infty}{\to} \infty$, which implies that the survival function satisfies $S_0(t) \overset{t\uparrow\infty}{\to} 0$. The second assumption is about the kernel function associated with the Gaussian process. In particular, we assume that the stationary kernel function K is strictly decreasing. Observe that for the kernel defined in example 3.1.2 this assumption is trivially satisfied, as well as for the squared exponential and the Ornstein-Uhlenbeck kernels.

Based on the conditions stated above, we prove that the random cumulative hazard function $\Lambda(t)$ is concentrated around the baseline cumulative hazard function $\Lambda_0(t)$ and thus it exhibits a similar behaviour. The proofs are based on the

first and second moments method. Hence, we must study the mean and variance of the functions $\lambda(t)$ and $\Lambda(t)$ for all $t \in \mathbb{R}_+$. Additionally to the well-definiteness of the random survival function $S(t)$, we prove that under some reasonable assumptions over the baseline hazard function λ_0, if the moments associated with the baseline hazard function are finite, then the random moments associated with the random hazard functions are finite as well.

As the results we intend to prove depend on the structure of the baseline hazard function λ_0, we consider a particular yet general family of functions. In particular, we consider the family given by

$$\lambda_0(t) \quad = \quad \Omega(t+s)^{\beta-1}, \tag{3.1.6}$$

for $\Omega > 0$, $s \geq 0$ and $\beta > 0$, or $\Omega > 0$ and $s > 0$ in the case of $\beta = 0$. This general class includes some popular choices in survival analysis. In the simplest scenario, we can take a constant baseline hazard function $\lambda_0(t) = \Omega$ which is simply the hazard function of a Exponential random variable with mean $1/\Omega$. Another choice might be $\lambda_0(t) = \Omega t^{\beta-1}$ with $\Omega > 1$ and $\beta > 1$, which corresponds to the hazard function of a Weibull distribution, another popular default distribution in survival analysis.

We proceed to study the expectation and variance of the random hazard function λ. By the symmetry of the sigmoid function, it is not difficult to see that for any fixed $t \in \mathbb{R}_+$

$$\mathbf{E}(\lambda(t)) \quad = \quad 2\lambda_0(t)\mathbf{E}(\sigma(l(t))) = \lambda_0(t),$$

meaning that the random hazard function λ is centred on the baseline hazard function λ_0. Moreover, since the sigmoid function σ is upper bounded by 1, we have that $\lambda(t) \leq 2\lambda_0(t)$ for all $t \geq 0$ almost surely. From the latter result we deduce that $\Lambda(t) \leq 2\Lambda_0(t)$, and thus $S_0(t)^2 = e^{-2\Lambda_0(t)} \leq e^{-\Lambda(t)} = S(t)$ for all

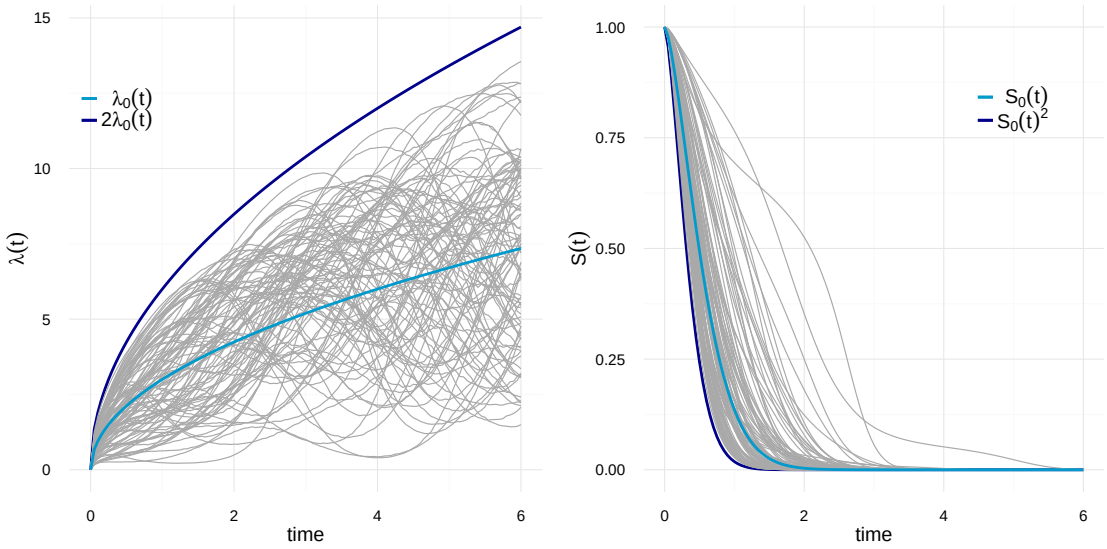

Figure 3.1: Samples from the prior for the hazard function and survival function, respectively. In the first plot the random hazard functions are upper bounded by $2\lambda_0(t)$, while in the second plot the random survival functions are lower bounded by $S_0(t)^2$.

$t \geq 0$. This suggests that the faster decay we can achieve for a random survival function is lower bounded by $S_0(t)^2$, as shown in Figure 3.1.

The covariance function for λ is given by

$$\mathbf{Cov}(\lambda(t), \lambda(s)) = 4\lambda_0(t)\lambda_0(s)\mathbf{Cov}(\sigma(l(t)), \sigma(l(s))) \leq 3\lambda_0(t)\lambda_0(s)$$

as $\mathbf{Cov}(\sigma(l(t)), \sigma(l(s))) \leq \frac{3}{4}$ for all $t, s \in \mathbb{R}_+$, $t \geq s$. Observe that the last result is trivially obtained by upper bounding the sigmoid function by 1. Indeed, even if $(l(t))_{t\geq 0}$ would be any stochastic process, the bound would remain the same. Therefore, we can work towards a tighter bound by considering the properties of the Gaussian process when upper bounding the covariance. This leads us to the following lemma.

Lemma 3.1.3. *For any $t, s \in \mathbb{R}_+$, such that $t - s > 1$, we have*

$$\mathbf{Cov}(\lambda(t), \lambda(s)) \quad \leq \quad 2\lambda_0(t)\lambda_0(s)\frac{K(t-s)}{K(0) - K(1)}. \tag{3.1.7}$$

Remark 3.1.4. The denominator is well-defined as K is strictly decreasing.

We defer the proof of lemma 3.1.3 to the appendix section of this chapter.

3.2 Cumulative hazard function

Recall the definition of the random cumulative hazard function

$$\Lambda(t) \quad = \quad \int_0^t 2\lambda_0(s)\sigma(l(s))ds.$$

Using the results obtained for the mean and covariance of the hazard function λ and the characterization of the baseline hazard function λ_0 given in equation (3.1.6), we give an estimate for the mean and variance of Λ. This result is established in the following lemma.

Lemma 3.2.1. *Let $t \in \mathbb{R}_+$, then*

$$\mathbf{E}(\Lambda(t)) \quad = \quad \Lambda_0(t) = \begin{cases} \Omega\beta^{-1}[(t+s)^\beta - s^\beta] & \beta > 0 \\ \Omega[\log(t+s) - \log(s)] & \beta = 0 \end{cases}.$$

Furthermore,

$$\mathbf{Var}(\Lambda(t)) \quad = \quad \begin{cases} \mathcal{O}\left(\frac{\Lambda_0(t)^2}{t}\int_0^t K(z)dz\right) & \beta \geq 1 \\ \mathcal{O}\left(\frac{\Lambda_0(t)^2}{t} + \Lambda_0(t)\int_0^t K(z)dz\right) & \beta \in (1/2, 1) \\ \mathcal{O}\left(\Lambda_0(t)\int_0^t K(z)dz\right) & \beta \in [0, 1/2] \end{cases}.$$

We defer the proof of lemma 3.2.1 to the appendix section of this chapter. Observe that the mean of Λ is simply given by the baseline cumulative hazard function Λ_0. On the other hand, the estimation of the variance is slightly more complicated as it depends on the integral of the kernel, that is, on $\int_0^t K(z)dz$.

Since K is strictly decreasing, it holds that $\int_0^t K(z)dz \leq K(0)t$. Additionally, if $K(z) \overset{z\uparrow\infty}{\to} 0$, then $\int_0^t K(z)dz = o(t)$. From these results, we can deduce that the

variance of the cumulative random hazard function Λ is relatively small compared with its mean Λ_0, suggesting that Λ is concentrated around Λ_0. We proceed to derive concentration inequalities for the cumulative random hazard function $\Lambda(t)$ around its mean $\Lambda_0(t)$ for any fixed $t \geq 0$. Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a positive function and define the event

$$E_t = \{|\Lambda(t) - \Lambda_0(t)| > \Lambda_0(t) f(t)\}.$$

Then, by Chebyshev's inequality and lemma 3.2.1,

$$\mathbf{P}(E_t) \leq \frac{\mathbf{Var}(\Lambda(t))}{\Lambda_0(t)^2 f(t)^2} = \begin{cases} O\left(\frac{\int_0^t K(z)dz}{tf(t)^2}\right) & \beta \geq 1 \\ O\left(\frac{1}{tf(t)^2} + \frac{\int_0^t K(z)dz}{\Lambda_0(t)f(t)^2}\right) & \beta \in (1/2, 1) \\ O\left(\frac{\int_0^t K(z)dz}{\Lambda_0(t)f(t)^2}\right) & \beta \in [0, 1/2] \end{cases} . \quad (3.2.1)$$

For $\beta \geq 1$, if we select $f(t) \gg \sqrt{\frac{\int_0^t k(z)dz}{t}}$, then $\mathbf{P}(E_t) \overset{t\uparrow\infty}{\to} 0$. The same result holds for $\beta \in (1/2, 1)$ if we choose $f(t) \gg \sqrt{\frac{1}{t} + \frac{\int_0^t k(z)dz}{\Lambda_0(t)}}$ and for $\beta \in [0, 1/2]$ by choosing $f(t) \gg \sqrt{\frac{\int_0^t k(z)dz}{\Lambda_0(t)}}$. Even though this is an asymptotic result, it allows us to gain intuition about the region in which the prior over the cumulative hazard function puts most of it mass (for fixed t). It turns out that in the long run this region seems to be very concentrated around the mean Λ_0.

Example 3.2.2. Let $(l(t))_{t\geq 0}$ be a centred Gaussian process with squared exponential kernel given by $K(t) = \exp\left\{-\frac{|t|^2}{2}\right\}$ for $t \in \mathbb{R}_+$. Additionally, choose $\lambda_0(t) = 3t^{0.5}$. This choice falls under the case $\beta \geq 1$. By equation (3.2.1), if we select the squared exponential kernel, we have that $\mathbf{P}(E_t) = O\left(\frac{1}{tf(t)^2}\right)$. Furthermore, by choosing $f(t) \gg \sqrt{\frac{1}{t}}$, we get $\mathbf{P}(E_t) \overset{t\uparrow\infty}{\to} 0$. Thus, in the limit when $t \to \infty$, $\Lambda(t)$ falls within a distance $f(t)\Lambda_0(t)$ from its mean $\Lambda_0(t)$.

Example 3.2.3. Let $(l(t))_{t\geq 0}$ and $K(t)$ be the same as in Example 3.2.2. Alternatively, take $\lambda_0(t) = 3(t+1)^{-1}$. This choice falls under the case of $\beta = 0$ and thus

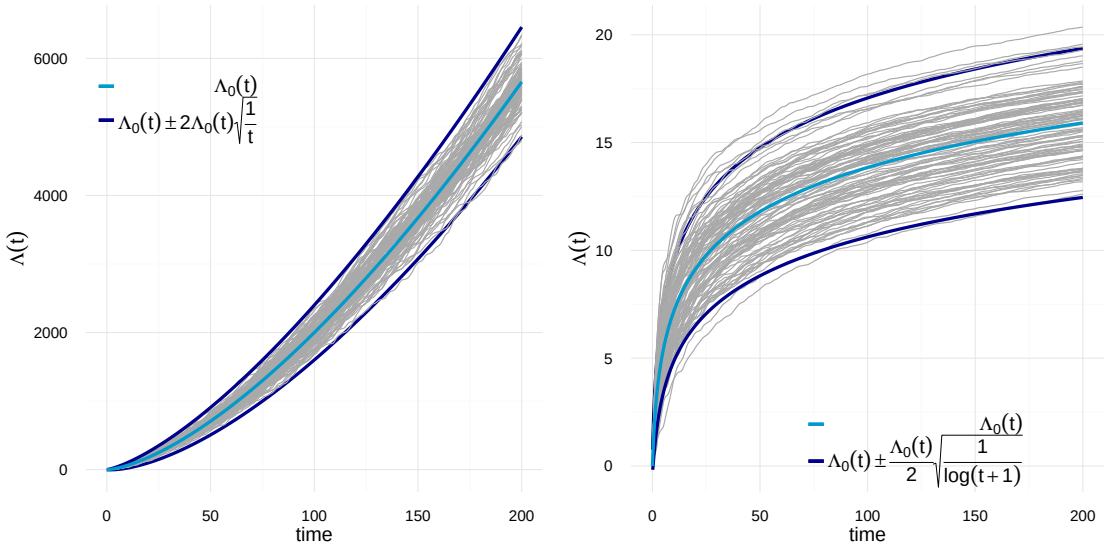

Figure 3.2: **Left panel**: Random cumulative hazard functions for example 3.2.2.; **Right panel**: Random cumulative hazard functions for example 3.2.3.

$\mathbf{P}(E_t) = O\left(\frac{1}{\log(t+1)f(t)^2}\right)$. We conclude that we need to have $f(t) \gg \sqrt{\frac{1}{\log(t+1)}}$ in order to guarantee that $\mathbf{P}(E_t) \overset{t\uparrow\infty}{\to} 0$.

3.3 Survival function

Suppose we want to compute the probability that for a given time t, the cumulative hazard function falls outside the interval $[\frac{1}{2}\Lambda_0(t), \frac{3}{2}\Lambda_0(t)]$. By equation (3.2.1), we have that for all fixed $t \in \mathbb{R}_+$

$$\mathbf{P}\left(|\Lambda(t) - \Lambda_0(t)| > \frac{1}{2}\Lambda_0(t)\right) = \begin{cases} O\left(\frac{\int_0^t K(z)dz}{t}\right) & \beta \geq 1 \\ O\left(\frac{1}{t} + \frac{\int_0^t K(z)dz}{\Lambda_0(t)}\right) & \beta \in (1/2, 1) \\ O\left(\frac{\int_0^t K(z)dz}{\Lambda_0(t)}\right) & \beta \in [0, 1/2] \end{cases} \quad (3.3.1)$$

For now, assume that the probability of equation 3.3.1 reaches 0 as time approaches to infinity for any possible value of β. Then, we can always find a sequence of times $\{t_n\}_{n\geq 1}$ such that $\mathbf{P}(B_{t_n}) = \mathbf{P}(|\Lambda(t_n) - \Lambda_0(t_n)| > \frac{1}{2}\Lambda_0(t_n)) \leq n^{-2}$ and thus $\sum_{n\geq 1}\mathbf{P}(B_{t_n}) < \infty$. Therefore, by the Borel-Cantelli lemma, it holds

that the probability that infinitely many events B_{t_n} occur is 0, that is, there exits $N \geq 1$ such that for all $n \geq N$,

$$|\Lambda(t_n) - \Lambda_0(t_n)| \leq \frac{1}{2}\Lambda_0(t_n), \tag{3.3.2}$$

implying that $\Lambda(t_n) \geq \frac{1}{2}\Lambda_0(t)$ for all $n \geq N$. Using this and the fact that the exponential function is increasing, we get the desired result. Indeed,

$$\lim_{t\to\infty} S(t) = \lim_{n\to\infty} S(t_n) \leq \lim_{n\to\infty} e^{-\frac{1}{2}\Lambda_0(t_n)} = 0. \tag{3.3.3}$$

We formalise this result by finding sufficient conditions over the integral of the kernel so that the probability defined in equation (3.3.1) reaches 0 as time goes to infinity. This is accomplished in the following proposition.

Proposition 3.3.1. *Let $(l(t))_{t\geq 0}$ be a centred Gaussian process with stationary kernel K and suppose that K satysfies*

$$\int_0^t K(z)dz \;=\; \begin{cases} o(t) & \beta \geq 1 \\ o(\Lambda_0(t)) & \beta \in [0,1) \end{cases}, \tag{3.3.4}$$

then $S(t) \overset{t\uparrow\infty}{\to} 0$ with probability one.

Remark 3.3.2. Observe that since we chose the covariance function K as a strictly decreasing function, the condition $\int_0^t K(z)dz = o(t)$ is satisfied trivially.

3.4 Moments

An alternative interpretation of the prior we have defined is in terms of the survival times. Indeed, since the random hazard function λ is trivially upper bounded by twice the baseline hazard function λ_0, we can think about the prior as a probability distribution over hazard functions that generates delayed random

survival times T. In particular, they are delayed with respect the model with baseline hazard function $2\lambda_0$. Hence, it is clear that the moments of T will always be lower bounded by the moments of a random variable with hazard function $2\lambda_0$. For $\beta > 0$, with β the parameter of the baseline hazard function defined in equation (3.1.6), we prove the following result.

Lemma 3.4.1. *Let λ_0 be defined as in equation (3.1.6) and let $\int_0^t K(z)dz = o(\log(t))$. For $\alpha > 0$ and $\beta > 0$, $\mathbf{E}(T^\alpha) < \infty$ almost surely.*

Proof. Let $\alpha > 0$, then

$$\begin{aligned}\mathbf{E}(T^\alpha|l) &= \int_0^\infty \mathbf{P}(T^\alpha > t|l)dt \\ &= \int_0^\infty S_l(t^{1/\alpha})dt.\end{aligned}$$

Hence, it suffices to show that there exists $T^\star > 0$ such that for all $t > T^\star$, $S_l(t^{1/\alpha}) < C/t^{1+\gamma}$ for some $\gamma > 0$ and some constant $C > 0$.

Consider the following sequence of times $\{t_n = e^{\alpha n^2}\}_{n\geq 1}$ and define the event $E^\alpha_{t_n} = \left\{|\Lambda(t_n^{1/\alpha}) - \Lambda_0(t_n^{1/\alpha})| > \frac{1}{2}\Lambda_0(t_n^{1/\alpha})\right\}$. By equation (3.2.1) and the assumption $\int_0^t K(z)dz = o(\log(t))$, it follows that $\mathbf{P}(E^\alpha_{t_n}) = \frac{o(n^2)}{e^{n^2}}$ for $\beta \geq 1$ and $\mathbf{P}(E^\alpha_{t_n}) = \frac{o(n^2)}{e^{\beta n^2}}$ for $\beta \in (0,1)$. Hence, there exits $N^\star > 0$ such that

$$\sum_{n=1}^{\infty} \mathbf{P}(E^\alpha_{t_n}) \leq N^\star + \sum_{n=N^\star}^{\infty} \frac{Cn^3}{e^{\beta n^2}} < \infty$$

for some positive constant C. By the Borel-Cantelli lemma, we deduce that there exits $N > 0$ such that for all $n \geq N$, $\Lambda(e^{n^2}) \geq \frac{1}{2}\Lambda_0(e^{n^2})$ and thus $S(t_n^{1/\alpha}) = e^{-\Lambda(e^{n^2})} \leq e^{-\frac{1}{2}\Lambda_0(e^{n^2})} = e^{-\frac{\Omega}{2\beta}[(e^{n^2}+s)^\beta - s^\beta]}$.

Observe that the survival function $S(t)$ is a non-increasing function. Hence,

by upper bounding the integral, we conclude

$$\begin{aligned}\int_0^\infty S(t^{1/\alpha})dt &= C' + \int_{t_N}^\infty S(t^{1/\alpha})dt \\ &\leq C' + \sum_{n=N}^{\infty}(t_{n+1}-t_n)S(t_n^{1/\alpha}) \\ &\leq C' + \sum_{n=N}^{\infty} e^{(n+1)^2} e^{-\frac{\Omega}{2\beta}[(e^{n^2}+s)^\beta - s^\beta]} < \infty.\end{aligned}$$

□

Remark 3.4.2 (Covariates.)**.** In chapter 2, we introduced two ways of modelling the dependencies on covariates. The first is a proportional hazards approach, where the random hazard function $\lambda_{\boldsymbol{x}}$ is written as the product of a random hazard function λ which only depends on time and the function $e^{\beta^\intercal \boldsymbol{x}}$. Therefore, it is sufficient to have λ being a well-defined object as $e^{\beta^\intercal \boldsymbol{x}}$ does not depend on time. The second method of introducing dependence is by the kernel given in equation (2.3.2). There for each fixed covariate $\boldsymbol{x}$, we have an stationary kernel. Thus, we can use the same analysis to check whether $S_{\boldsymbol{x}}(t) = \mathbf{P}(T > t|\boldsymbol{x}) \overset{z\uparrow\infty}{\to} 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$. Therefore, both of these cases are covered by the analysis of this chapter.

Appendix

3.A Proofs

3.A.1 Proof of Lemma (3.1.3)

Proof. Since $(l(t))_{t\geq 0}$ is a centred stationary Gaussian process and σ is the sigmoid function, $\mathbf{E}(\sigma(l(t))) = 1/2$ for all $t \geq 0$. We proceed to estimate the covariance. Let $t, s \in \mathbb{R}_+$ with $t - s > 1$, then

$$\begin{aligned}
\mathbf{E}(\sigma(l(t))\sigma(l(s))) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sigma(x)\sigma(y)\frac{e^{-\frac{K(0)(x^2+y^2)-2K(t-s)xy}{2(K(0)^2-K(t-s)^2)}}}{2\pi(K(0)^2-K(t-s)^2)^{1/2}}dxdy \\
(\text{since } x^2+y^2 \geq 2xy) \quad &\leq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sigma(x)\sigma(y)\frac{e^{-\frac{(K(0)-K(t-s))(x^2+y^2)}{2(K(0)^2-K(t-s)^2)}}}{2\pi(K(0)^2-K(t-s)^2)^{1/2}}dxdy, \\
&\leq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sigma(x)\sigma(y)\frac{e^{-\frac{(x^2+y^2)}{2(K(0)+K(t-s))}}}{2\pi(K(0)^2-K(t-s)^2)^{1/2}}dxdy, \\
&= \left(\frac{K(0)+K(t-s)}{K(0)-K(t-s)}\right)^{1/2}\mathbf{E}(\sigma(l(0)))^2 \\
&\leq \frac{K(0)+K(t-s)}{4[K(0)-K(t-s)]}.
\end{aligned}$$

Since $\mathbf{E}(\sigma(l(t))) = 1/2$, using the previous result, we obtain

$$\begin{aligned}
\mathbf{Cov}(\sigma(l(t)),\sigma(l(s))) &\leq \frac{1}{4}\frac{K(0)+K(t-s)}{K(0)-K(t-s)} - \mathbf{E}(\sigma(l(t)))\mathbf{E}(\sigma(l(s))) \\
&\leq \frac{1}{2}\frac{K(t-s)}{K(0)-K(t-s)}.
\end{aligned}$$

Since K is strictly decreasing and $t - s > 1$, we have

$$\mathbf{Cov}(\lambda(t), \lambda(s)) = 4\lambda_0(t)\lambda_0(s)\mathbf{Cov}(\sigma(l(t)), \sigma(l(s))) \leq 2\lambda_0(t)\lambda_0(s)\frac{K(t-s)}{K(0)-K(1)}.$$

□

3.A.2 Proof of Lemma (3.2.1)

Proof. We start by computing the mean value of $\Lambda(t)$ for $t \in \mathbb{R}_+$. By Tonelli's theorem, we have that

$$\mathbf{E}(\Lambda(t)) = \int_0^t 2\lambda_0(z)\mathbf{E}(\sigma(l(z)))dz = \Lambda_0(t).$$

Hence, the random cumulative hazard function Λ is centred around the baseline cumulative hazard function Λ_0. Moreover, by the definition of λ_0, given in equation (3.1.6), we have

$$\Lambda_0(t) = \begin{cases} \frac{\Omega}{\beta}[(t+s)^\beta - s^\beta] & \beta > 0 \\ \Omega[\log(t+s) - \log(s)] & \beta = 0 \end{cases}.$$

This proves the first part of the lemma. The rest of the proof focuses on finding good estimates for the variance. Using Tonelli's theorem, we have

$$\begin{aligned} \mathbf{Var}(\Lambda(t)) &= \int_0^t \int_0^t 4\lambda_0(z)\lambda_0(z')\mathbf{Cov}(\sigma(l(z)), \sigma(l(z')))dz'dz \\ &= I_1 + I_2, \end{aligned}$$

where

$$I_1 = \int_{|z-z'|\leq 1} 4\lambda_0(z)\lambda_0(z')\mathbf{Cov}(\sigma(l(z)), \sigma(l(z')))dz'dz$$

and

$$I_2 \;=\; \int_{1<|z-z'|\leq t} 4\lambda_0(z)\lambda_0(z')\mathbf{Cov}(\sigma(l(z)),\sigma(l(z')))dz'dz.$$

For the first integral we upper bound the covariance by

$$\mathbf{Cov}(\sigma(l(z)),\sigma(l(z'))) \;\leq\; \sqrt{\mathbf{Var}(\sigma(l(z)))\mathbf{Var}(\sigma(l(z'))}.$$

Since $(l(t))_{t\geq 0}$ is a stationary Gaussian process, the variance does not depend on t and thus $\mathbf{Cov}(\sigma(l(z)),\sigma(l(z'))) \leq \mathbf{Var}(\sigma(l(z)))$. In particular, for all $z \in \mathbb{R}_+$

$$\mathbf{Var}(\sigma(l(z))) \;=\; \mathbf{E}(\sigma(l(z))^2) - \mathbf{E}(\sigma(l(z)))^2 \leq 3/4$$

since the sigmoid function is upper bounded by 1 and the expected value of $\sigma(l(z))$ is $1/2$. Hence,

$$I_1 \;\leq\; 6\int_0^t \int_z^{z+1} \Omega^2 (z'+s)^{\beta-1}(z+s)^{\beta-1}dz'dz. \tag{3.A.1}$$

We distinguish between three cases for values of β.

1. For $\boldsymbol{\beta \geq 1}$,

$$I_1 \;\leq\; 6\Omega^2 \int_0^t (z+s+1)^{2\beta-2}dz \leq 6\Omega^2 \frac{(t+s+1)^{2\beta-1}}{2\beta-1}. \tag{3.A.2}$$

2. For $\boldsymbol{\beta \in [0, 1/2)}$,

$$I_1 \;\leq\; 6\Omega^2 \int_0^t (z+s)^{2\beta-2}dz \leq 6\Omega^2 \frac{s^{2\beta-1}}{1-2\beta}. \tag{3.A.3}$$

3. For $\boldsymbol{\beta \in (1/2, 1)}$,

$$I_1 \;\leq\; 6\Omega^2 \int_0^t (z+s)^{2\beta-2}dz \leq 6\Omega^2 \frac{(t+s)^{2\beta-1}}{2\beta-1}. \tag{3.A.4}$$

4. For $\boldsymbol{\beta = 1/2}$,

$$I_1 \leq 6\Omega^2 \int_0^t (z+s)^{-1} dz \leq 6\Omega^2 \log(t+s). \tag{3.A.5}$$

We proceed by finding an upper bound for the second integral I_2. Notice that

$$\begin{aligned} I_2 &= \int_{1<|z-z'|\leq t} 4\lambda_0(z)\lambda_0(z')\mathbf{Cov}(\sigma(l(z)),\sigma(l(z')))dz'dz \\ &= 8\int_1^t \int_0^{z'-1} \lambda_0(z)\lambda_0(z')\mathbf{Cov}(\sigma(l(z)),\sigma(l(z')))dzdz'. \end{aligned}$$

By Lemma (3.1.3), we have that for $z' - z > 1$, $\mathbf{Cov}(\sigma(l(z')),\sigma(l(z))) \leq \frac{K(z'-z)}{2(K(0)-K(1))}$. Hence,

$$\begin{aligned} I_2 &\leq 4\int_1^t \int_0^{z'-1} \lambda_0(z)\lambda_0(z')\frac{K(z'-z)}{K(0)-K(1)}dzdz' \\ &\leq \frac{4\Omega^2}{K(0)-K(1)}\int_1^t \int_0^{z'-1} K(z'-z)(z+s)^{\beta-1}(z'+s)^{\beta-1}dzdz'. \end{aligned} \tag{3.A.6}$$

We again distinguish between three cases for values of β. Consider the following change of variable $w' = z'$ and $w = z' - z$.

1. For $\boldsymbol{\beta \geq 1}$,

$$\begin{aligned} I_2 &\leq \frac{4\Omega^2}{K(0)-K(1)}\int_1^t \int_w^t K(w)(w'+s)^{\beta-1}(w'-w+s)^{\beta-1}dw'dw \\ &\leq \frac{4\Omega^2}{K(0)-K(1)}\int_1^t \int_w^t K(w)(w'+s)^{2\beta-2}dw'dw \\ &\leq \frac{4\Omega^2}{K(0)-K(1)}\int_1^t K(w)dw \int_0^t (w'+s)^{2\beta-2}dw' \\ &\leq \frac{4\Omega^2}{K(0)-K(1)}\frac{(t+s)^{2\beta-1}}{2\beta-1}\int_1^t K(w)dw. \end{aligned} \tag{3.A.7}$$

2. For $\boldsymbol{\beta \in (0,1)}$,

$$\begin{aligned} I_2 &\leq \frac{4\Omega^2}{K(0)-K(1)} \int_1^t \int_w^t K(w)(w'+s)^{\beta-1}(w'-w+s)^{\beta-1} dw' dw \\ &\leq \frac{4\Omega^2}{K(0)-K(1)} \int_1^t \int_w^t K(w)(w'+s)^{\beta-1} s^{\beta-1} dw' dw \\ &\leq \frac{4\Omega^2}{K(0)-K(1)} \int_1^t K(w) dw \int_0^t (w'+s)^{\beta-1} s^{\beta-1} dw' \\ &\leq \frac{4\Omega^2}{K(0)-K(1)} \frac{s^{\beta-1}(t+s)^{\beta}}{\beta} \int_1^t K(w) dw. \end{aligned} \tag{3.A.8}$$

3. For $\boldsymbol{\beta = 0}$,

$$\begin{aligned} I_2 &\leq \frac{4\Omega^2}{K(0)-K(1)} \int_1^t \int_w^t K(w)(w'+s)^{-1} s^{-1} dw' dw \\ &\leq \frac{4\Omega^2}{K(0)-K(1)} \int_1^t K(w) dw \int_0^t (w'+s)^{-1} s^{-1} dw' \\ &\leq \frac{4\Omega^2}{K(0)-K(1)} s^{-1} \log(t+s) \int_1^t K(w) dw. \end{aligned} \tag{3.A.9}$$

Finally, putting everything together we deduce the following upper bounds for the variance $\mathbf{Var}(\Lambda(t))$,

1. For $\boldsymbol{\beta \geq 1}$. From equations (3.A.2) and (3.A.7), it holds

$$\begin{aligned} \mathbf{Var}(\Lambda(t)) &\leq 2\Omega^2 \frac{(t+s)^{2\beta-1}}{2\beta-1} \left(3 + \frac{2}{K(0)-K(1)} \int_1^t K(z) dz \right) \\ &= \mathcal{O}\left(t^{2\beta-1} \int_0^t K(z) dz \right). \end{aligned} \tag{3.A.10}$$

2. For $\boldsymbol{\beta \in (1/2, 1)}$. From equations (3.A.4) and (3.A.8), it holds

$$\begin{aligned} \mathbf{Var}(\Lambda(t)) &\leq 2\Omega^2 \left(\frac{3(t+s)^{2\beta-1}}{2\beta-1} + \frac{2}{K(0)-K(1)} \frac{s^{\beta-1}(t+s)^{\beta}}{\beta} \int_1^t K(z) dz \right) \\ &= \mathcal{O}\left(t^{2\beta-1} + t^{\beta} \int_0^t K(z) dz \right). \end{aligned} \tag{3.A.11}$$

3. For $\boldsymbol{\beta = 1/2}$. From equations (3.A.5) and (3.A.8), we obtain

$$\begin{aligned}\mathbf{Var}(\Lambda(t)) &\leq 2\Omega^2\left(3\log(t+s) + \frac{4}{K(0)-K(1)}s^{-1/2}(t+s)^{1/2}\int_1^t K(z)dz\right)\\ &= \mathcal{O}\left(t^\beta \int_0^t K(z)dz\right). \qquad (3.A.12)\end{aligned}$$

4. For $\boldsymbol{\beta \in (0, 1/2)}$. From equations (3.A.3) and (3.A.8), it holds

$$\begin{aligned}\mathbf{Var}(\Lambda(t)) &\leq 2\Omega^2\left(\frac{3s^{2\beta-1}}{1-2\beta} + \frac{2}{K(0)-K(1)}\frac{s^{\beta-1}(t+s)^\beta}{\beta}\int_1^t K(z)dz\right)\\ &= \mathcal{O}\left(t^\beta \int_0^t K(z)dz\right). \qquad (3.A.13)\end{aligned}$$

5. For $\boldsymbol{\beta = 0}$. From equations (3.A.3) and (3.A.9), we get

$$\begin{aligned}\mathbf{Var}(\Lambda(t)) &\leq 2\Omega^2\left(\frac{3s^{2\beta-1}}{1-2\beta} + \frac{4}{K(0)-K(1)}s^{-1}\log(t+s)\int_1^t K(z)dz\right)\\ &= \mathcal{O}\left(\log(t)\int_0^t K(z)dz\right). \qquad (3.A.14)\end{aligned}$$

□

Chapter 4

Posterior computation

In this chapter we propose an inference scheme to sample from the posterior distribution of the model defined in Chapter 2.

4.1 Data augmentation scheme

Recall the simplest version of the model proposed in equation (2.1.1). Observe that the probability density function defined in this hierarchical model depends on the integral of a Gaussian process. Indeed, the density function of our model is given by $f_\lambda(t) = \lambda(t)e^{-\int_0^t \lambda(s)ds}$, where $\lambda(t) = 2\lambda_0(t)\sigma(l(t))$, $(l(t))_{t\geq 0}$ is a Gaussian process, and $\lambda_0(t)$ is a fixed baseline hazard function chosen by the user. The integral $\int_0^\infty 2\lambda_0(t)\sigma(l(t))dt$ is not analytically tractable and thus an exact evaluation of the likelihood cannot be performed.

In order to avoid the use of a sampling mechanism based on an approximation of the likelihood, we resort to a data augmentation scheme based on the relationship between a survival time and the first jump of a Poisson process. This relation is as follows. Let T be a random survival time with associated hazard function λ, and let J be the first jump of a Poisson process with intensity function λ, then the random variables T and J are equal in distribution[1].

[1] This follows from a simple computation.

In our case, the intensity of the Poisson process is a function of a Gaussian process, obtaining what is called a Gaussian Cox process. The main difficulty of working with Gaussian Cox processes remains in the fact that the likelihood for the first jump J still deals with the integral of the random infinite-dimensional object λ. This does not come as a surprise since $T \stackrel{d}{=} J$. Nevertheless, interpreting our model in terms of a Gaussian Cox process allows us to use the result proposed in [2]. In [2], the authors developed an algorithm for exact tractable inference of Gaussian Cox processes by exploiting a nice trick which allows them to make inference without sampling the whole Gaussian process but only a finite number of points.

The notion behind the algorithm proposed in [2], and its subsequent refined version, given in [51], can be explained by the simple idea of Poisson thinning. The idea of thinning is explained in the following theorem.

Theorem 4.1.1 (Poisson thinning)**.** *Let $(N_t)_{t\geq 0}$ be a Poisson process with rate function $\lambda(t)$. Suppose that each point landing at t is kept independently with probability $q(t)$. Then the resulting set of points is an inhomogeneous Poisson process with rate $\lambda(t)q(t)$.*

Suppose we want to sample the first jump of a Poisson process with intensity given by $\lambda(t) = 2\lambda_0(t)\sigma(l(t))$. By using Poisson thinning, since λ is upper bounded by $2\lambda_0$ (recall that σ denotes the sigmoid function), we can sample points from the Poisson process with intensity $2\lambda_0$ and then thin (reject) them with probability $p(t) = 1-\lambda(t)/2\lambda_0(t)$. The first of the resulting kept (accepted) points corresponds to the first jump of the Poisson process of intensity λ. An illustration of this procedure is shown in Figure 4.1. Moreover, the joint distribution of the set of rejected jump points G (which may possible be an empty set) and the first accepted jump point T is given in the following proposition.

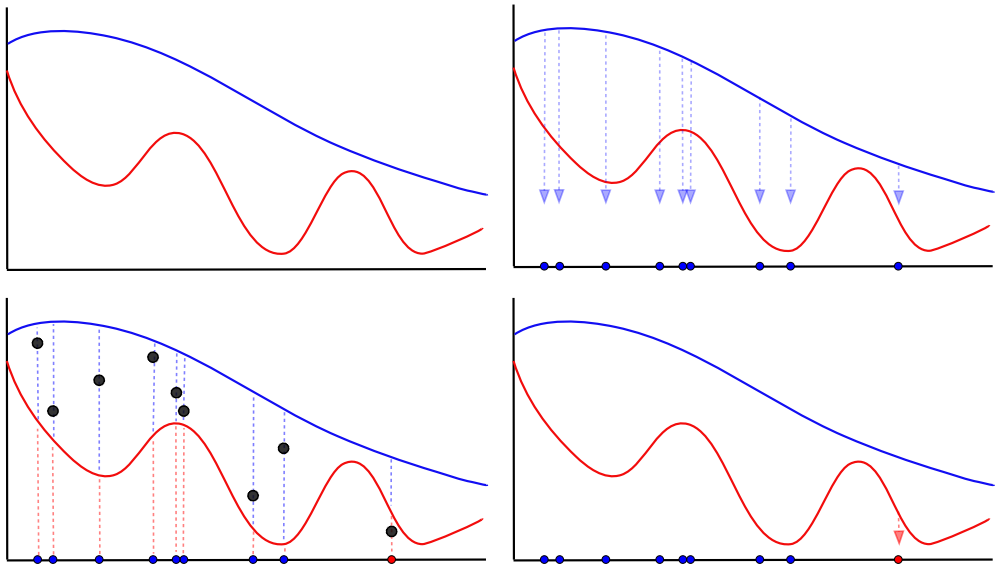

Figure 4.1: Poisson thinning: The blue curve represents $2\lambda_0$ while the red curve denotes the random hazard function λ. We start by sampling points from a Poisson process with intensity $2\lambda_0$, then each point is thinned (painted blue) with probability $1-p(t) = 1-\lambda(t)/\lambda_0(t)$. We repeat this process until we accept (keep) the first point and then we paint it red.

Proposition 4.1.2. *Let* $\Lambda_0(t) = \int_0^T \lambda_0(t)dt$, *then*

$$p(G, T|\lambda_0, l) \quad = \quad \left(2\lambda_0(T)\prod_{g\in G} 2\lambda_0(g)\right) e^{-2\Lambda_0(T)} \left(\sigma(l(T))\prod_{g\in G}(1-\sigma(l(g)))\right).$$

Notice that in order to evaluate the above joint probability distribution we just need to know the value of the Gaussian process in the points $G\cup\{T\}$ instead of the whole interval $[0, T]$. Therefore, as long as the user is using a well-known parametric form for λ_0, so Λ_0 is analytically tractable (in the Weibull case this can be easily done), then we have a tractable complete likelihood.

Let $\{T_i\}_{i=1}^n$ be a sequence of survival times and let $\{G_i\}_{i=1}^n$ be the corresponding sequence of sets of rejected jump times of the thinned Poisson process, generated independently for each T_i. Then by using the proposition above, the

model of equation (2.1.1) can be reformulated as

$$\begin{aligned}(G_i, T_i)|\lambda_0, l &\overset{ind.}{\sim} 2\lambda_0(T_i)\left(\prod_{g\in G_i} 2\lambda_0(g)\right) e^{-2\Lambda_0(T_i)}\sigma(l(T))\left(\prod_{g\in G_i} 1-\sigma(l(g))\right)\\ (l(t))_{t\geq 0} &\sim \mathcal{GP}(0, K).\end{aligned} \tag{4.1.1}$$

To perform inference we need the data $\{G_i \cup \{T_i\}\}_{i=1}^n$, whereas we only receive the survival times $\{T_i\}_{i=1}^n$. Thus, we need to sample the missing data $\{G_i\}_{i=1}^n$ given $\{T_i\}_{i=1}^n$. The next proposition gives us a way to do this.

Proposition 4.1.3. *[51] Let T be a survival time and let G be its set of rejected points. Then set of points G given (T, λ_0, l) is distributed as a non-homogeneous Poisson process with intensity $2\lambda_0(t)(1-\sigma(l(t)))$ on the interval $[0, T]$.*

Therefore, as long as sampling from a Poisson process with intensity $2\lambda_0(t)$ is tractable, we can use the same trick of thinning to sample from the distribution of rejected points G. In particular, the Mapping theorem for Poisson processes gives us a way of sampling from a Poisson process with intensity $2\lambda_0(t)$ by using an homogeneous Poisson process with intensity 1. This is done by sampling from the Poisson process with intensity 1 on the time $t^\star = 2\Lambda_0(t)$, then we rescale the points applying the inverse of $2\Lambda_0$, the resulting points are distributed according the Poisson process with intensity $2\lambda_0(t)$. The claim above is formally stated in the following theorem.

Theorem 4.1.4 (Mapping theorem)**.** *Let $(N(t))_{t\geq 0}$ be a Poisson process with intensity λ and define $\Lambda(t) = \int_0^t \lambda(s)ds$. Additionally, let $(N^\star(t))_{t\geq 0}$ be a Poisson process with intensity 1. Then for any $t \geq 0$ it holds that $N^\star(\Lambda(t))$ has the same distribution as $N(t)$.*

4.1.1 Covariates

In Chapter 2, we proposed two different alternatives to include covariates. The first alternative comes from defining a proportional hazards approach, in which the hazard function is defined as $\lambda_{\boldsymbol{x}}(t) = \lambda(t)e^{\beta^\intercal \boldsymbol{x}}$, where $\lambda(t) = 2\lambda_0(t)\sigma(l(t))$ is the stochastic process previously defined. Observe that conditioned in the parameter β, $\lambda_{\boldsymbol{x}}(t)$ is upper bounded by $2\lambda_0(t)e^{\beta^\intercal \boldsymbol{x}}$. Then we adapt our inference scheme by augmenting the model as previously described, but considering a different bound for each T_i. For the second alternative, as the dependence on covariates is introduced in the Gaussian process by defining a kernel in both time a covariates, the proposed approach works without any major modification. Indeed, $2\lambda_0(t)$ remains as an upper bound for $\lambda_{\boldsymbol{x}}(t)$ regardless the value of $\boldsymbol{x}$. The only objects that change are the rejection probabilities (they depend on both, the survival time T_i and associated covariates $\boldsymbol{x}_i$).

4.2 Adding censoring

As mentioned in Chapter 1, we can encounter three types of censoring: right, left and interval censoring. We assume that each data point T_i is associated with an (observable) indicator δ_i which tells us the type of censoring. We describe how the data augmentation scheme described before can easily handle any type of censorship.

Right censorship: In presence of right censoring, the likelihood for a survival time T_i is $S(T_i)$. The related event in terms of the rejected points correspond to not accepting any location in $[0, T_i)$. Hence, we can treat right censorship in the same way as the uncensored case, by just sampling from the distribution of the rejected jump times prior T_i. In this case, T_i is not an accepted location.

Left censorship: In this set-up, the likelihood for a survival time T_i corresponds to $F(T_i)$. Treating this type of censorship is slightly more difficult than

the previous case because the event is more complex. We ask for at least accepting one jump time prior T_i, which may lead us to have a larger set of latent variables G_i. In order to avoid this, we decided to proceed by imputing the actual survival time $T_i^\star$ by sampling from its truncated distribution on $[0, T_i]$ and proceed like in the uncensored case.

Interval censorship: If we know that the actual survival time T_i is in the interval $[s, t]$ we can deal with interval censoring in the same way as left censoring but imputing T_i in the interval $[s, t]$.

4.2.1 Imputing censored random variables

In this subsection we describe two sampling algorithms to impute random survival times in the case of having left or interval censoring.

Rejection sampling

This sampling algorithm is based on the fact that the hazard function of a survival time T does not change if we condition on the event $\{T > t_l\}$. Indeed, the conditional probability density function of T given the event $\{T > t_l\}$ is denoted by $f_{t_l}(t) = f(t)/S(t_l)$ for $t > t_l$, where f and S are the probability density function and survival function of T. From this, we compute the survival function as $S_{t_l}(t) = S(t)/S(t_l)$ and finalise by concluding that the hazard function remains the same for $t > t_l$, that is,

$$\lambda_{t_l}(t) = \begin{cases} 0 & t \leq t_l \\ \lambda(t) & t > t_l \end{cases} . \tag{4.2.1}$$

We can use the idea of Poisson thinning, previously explained in this chapter, to sample from the density f_{t_l} given the hazard function λ and the time t_l. As for an upper bound t_u, we can just stop the procedure when the proposed jump is greater than this value.

The probability of not accepting any jump, and thus not generating a sample from f_{t_l}, corresponds to

$$\mathbf{P}(\text{Reject all}) = \mathbf{P}(N_\lambda(t_u) - N_\lambda(t_u) = 0) = e^{-\int_{t_l}^{t_u} \lambda(s)ds}. \quad (4.2.2)$$

It is not hard to see that in the case that $t_u = \infty$, then $\mathbf{P}(\text{Reject all}) = 0$ provided that $\Lambda(\infty) = \infty$. If the latter assumption is not satisfied our sampling algorithm could potentially take an infinite amount of time to end.

Gibbs sampling

Another option in order to avoid a rejection sampling algorithm is to consider a Gibbs sampler. We illustrate the algorithm for the case of interval censorship as left censorship follows naturally as a particular case of it. We start considering an initial accepted time T within the interval of interest, then as a first step in our algorithm, see an illustration in Figure 4.2, we resample the jumps times before T from a non-homogeneous Poisson process with rate $2\lambda_0 - \lambda$, and after T from an inhomogeneous Poisson process with rate $2\lambda_0$. Here λ and λ_0 are defined as in the model of equation 2.1.1. We call the set of ordered resulting jump times $\{z_j\}_{j=1}^n$. Notice this is a non-empty set as it at least contains T.

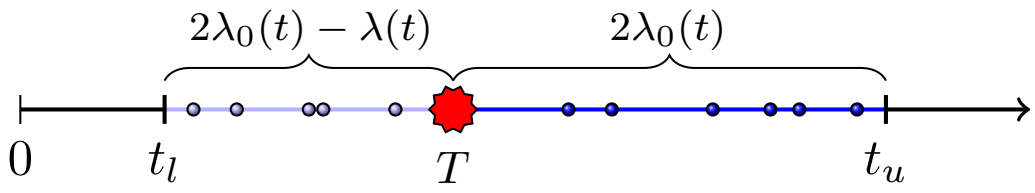

Figure 4.2: Illustration of Gibbs sampler augmentation scheme.

Algorithm 1: Pseudo-code

Input: λ, λ_0 t_l, t_u and a initial value of T

Sample $\boldsymbol{z}|\boldsymbol{T}$

1 Sample a set of points A on (t_l, T) from a Poisson process with rate $2\lambda_0 - \lambda$

2 Sample a set of points B on (T, t_u) from a Poisson process with rate $2\lambda_0$

3 Define $\boldsymbol{z} = A \cup T \cup B$ ordered in ascending order.

Sample $\boldsymbol{T}|\boldsymbol{z}$

4 Choose $T = z_j$ with probability given by 4.2.3

Figure 4.3: Gibbs sampler for imputing censored random variables.

In a second step, given the set of points $\{z_j\}_{j=1}^n$, we resample T with probability given by

$$P(T = z_j | \{z_j\}_{j=1}^n) = \frac{\frac{\lambda(z_j)}{2\lambda_0(z_j)} \prod_{k=1}^{j-1} 1 - \frac{\lambda(z_k)}{2\lambda_0(z_k)}}{\sum_{l=1}^n \frac{\lambda(z_l)}{2\lambda_0(z_l)} \prod_{k=1}^{l-1} 1 - \frac{\lambda(z_k)}{2\lambda_0(z_k)}}. \tag{4.2.3}$$

4.3 Posterior distributions

We derive the posterior distributions for the parameters of the model under the two covariate approaches proposed in Chapter 2. For the purpose of these computations, we only consider right censoring as for left and interval censoring we use the data augmentation schemes proposed in the previous section.

Let T_i be a random survival time with associated covariate $\boldsymbol{x}_i$. Additionally, consider an indicator variable δ_i taking the value $\delta_i = 1$ when T_i is uncensored or $\delta_i = 0$ when T_i is right censored. Following the proposed data augmentation scheme (4.1) to perform exact inference, we augment the data by sampling G_i, the set of rejected jump times associated with T_i. Then, the data is denoted by $\mathcal{D} = \{(T_i, \delta_i, \boldsymbol{x}_i, G_i)\}_{i=1}^n$.

4.3.1 Likelihood

Proportional hazards model

Recall the definition of the hazard function for the proportional hazards model defined in equation (2.3.1), i.e., $\lambda_{\boldsymbol{x}} = 2\lambda_0(t)\sigma(l(t))e^{\beta^\intercal \boldsymbol{x}}$. Then the likelihood for the observed data $\mathcal{D} = \{(T_i, \delta_i, \boldsymbol{x}_i, G_i)\}_{i=1}^n$ under this model corresponds to

$$\begin{aligned} L(\lambda_0, \beta, l|\mathcal{D}) &= \prod_{i=1}^{n} [2\lambda_0(T_i)e^{\beta^\intercal \boldsymbol{x}_i}]^{\delta_i} \left(\prod_{g \in G_i} 2\lambda_0(g)e^{\beta^\intercal \boldsymbol{x}_i} \right) e^{-2\Lambda_0(T_i)e^{\beta^\intercal \boldsymbol{x}_i}} \sigma(l(T_i)) \\ &\times \left(\prod_{g \in G_i} 1 - \sigma(l(g)) \right). \end{aligned}$$

Kernel in time and covariates

For the second approach, we consider a Gaussian process defined in both time and covariates as in equation (2.3.3). In this case, the likelihood for the observed data $\mathcal{D} = \{(T_i, \delta_i, \boldsymbol{x}_i, G_i)\}_{i=1}^n$ corresponds to

$$L(\lambda_0, l|\mathcal{D}) = \prod_{i=1}^{n} [2\lambda_0(T_i)]^{\delta_i} \left(\prod_{g \in G_i} 2\lambda_0(g) \right) e^{-2\Lambda_0(T_i)} \sigma(l(T_i, \boldsymbol{x}_i)) \prod_{g \in G_i} (1 - \sigma(l(g, \boldsymbol{x}_i))).$$

For the common parameters, i.e., λ_0 and the Gaussian process l we derive a general form for the posterior. This is achieved by taking into account the similarities of the structure of the likelihood in both models.

4.3.2 Posterior distribution for β

This parameter is exclusive of the proportional hazards approach. Assume a prior probability ν over β, then the posterior distribution is proportional to

$$p(\beta|\lambda_0, l) \propto \prod_{i=1}^{n} e^{(\delta_i + |G_i|)\beta^\intercal \boldsymbol{x}_i} e^{-2\Lambda_0(T_i)e^{\beta^\intercal \boldsymbol{x}_i}} \nu(\beta), \tag{4.3.1}$$

here $|G_i|$ denotes the cardinality of the set G_i. Notice there is not a clear choice of the prior distribution ν such that the posterior distribution has a tractable closed form.

4.3.3 Posterior distribution for λ_0

Assuming a prior probability ν over λ_0, or over some of its parameters, the posterior distribution is proportional to

$$p(\lambda_0|l) \propto \prod_{i=1}^{n} \left(\lambda_0(T_i)^{\delta_i} \prod_{g \in G_i} \lambda_0(g) e^{-C_i \Lambda_0(T_i)} \right) \nu(\lambda_0), \tag{4.3.2}$$

where C_i is a constant taking the value $C_i = 2e^{\beta^\intercal \boldsymbol{x_i}}$ for the proportional hazards model and $C_i = 2$ for the second approach. Notice that in the second approach C_i does not depend on i. Moreover, for both approaches, the posterior distribution of λ_0 does not depend on the Gaussian process l.

We provide two examples of the computation of this posterior for two well-known hazard functions in survival analysis. In order to ease the notation we do not consider covariates.

Example 4.3.1 (The Exponential case)**.** A particular case corresponds to when the baseline hazard function is chosen to be constant, that is, $\lambda_0(t) = \Omega$ for all $t \geq 0$. This choice gives rise to the Exponential model, as the chosen hazard function corresponds to the hazard function of an Exponential random variable with mean Ω^{-1}. Assuming a Gamma prior over the baseline hazard, $\Omega \sim Gamma(\alpha, \beta)$, the posterior has a closed form. Indeed,

$$\begin{aligned} p(\Omega|l) &\propto \left(\prod_{i=1}^{n} \Omega^{|G_i|+\delta_i} e^{-2\Omega T_i} \right) \Omega^{\alpha-1} e^{-\beta\Omega} \\ &\propto \Omega^{(\sum_{i=1}^{n} |G_i|+\delta_i)+\alpha-1} e^{-\Omega(2\sum_{i=1}^{n} T_i+\beta)}, \end{aligned} \tag{4.3.3}$$

where $|G_i|$ denotes the cardinality of the set G_i. Thus the posterior distribution

for Ω is a Gamma distribution with shape parameter $(\sum_{i=1}^{n} |G_i| + \delta_i) + \alpha$ and rate parameter $2\sum_{i=1}^{n} T_i + \beta$.

Example 4.3.2 (The Weibull case)**.** An alternative approach is to choose λ_0 as the hazard function of a Weibull random variable with shape parameter $k > 0$ and scale parameter $b > 0$, i.e., $\lambda_0(t) = bkt^{k-1}$. If we put a Gamma prior over the scale parameter, $b \sim Gamma(\alpha, \beta)$, the posterior has a closed form

$$\begin{aligned} p(b|k,l) &\propto \left(\prod_{i=1}^{n} b^{|G_i|+\delta_i} e^{-2bT_i^k} \right) b^{\alpha-1} e^{-\beta b} \\ &\propto b^{(\sum_{i=1}^{n} |G_i|+\delta_i)+\alpha-1} e^{-b(2\sum_{i=1}^{n} T_i^k+\beta)}. \end{aligned} \tag{4.3.4}$$

Thus the posterior distribution for the scale parameter b is a Gamma distribution with shape parameter $(\sum_{i=1}^{n} |G_i| + \delta_i) + \alpha$ and rate parameter $2\sum_{i=1}^{n} T_i^k + \beta$. In the case of the shape parameter, we do not have a tractable closed form for the posterior distribution

$$p(k|b,l) \propto \prod_{i=1}^{n} \left(kT_i^{k-1} \prod_{g \in G_i} kg^{k-1} e^{-2bT_i^k} \right) \nu(k), \tag{4.3.5}$$

where ν denotes a probability density function on $\mathbb{R}_+$.

4.3.4 Posterior distribution for l

In this case, we derive a general form for the posterior distribution of the Gaussian process under the two proposed covariate approaches. This can be done as the two posteriors have the same structure. The only difference is the index space in which the Gaussian process is evaluated. Indeed, for the first approach the index space of the Gaussian process is just the time, i.e. $(l(t))_{t\geq 0}$, while for the second approach is both time and covariates, i.e. $(l(t, \boldsymbol{x}))_{t\geq 0, \boldsymbol{x}\in\mathbb{R}^d}$. For the second approach, in order to ease the notation, we write $l(t)$ instead of $l(t, \boldsymbol{x})$, but we understand from the context that the Gaussian process is indexed both time

and covariates.

Consider the sets of points $\boldsymbol{T} = \bigcup_{i=1}^{n} T_i$, representing a collection of independent survival times for n observations, and $\boldsymbol{G} = \bigcup_{i=1}^{n} G_i$, the associated rejected jump times for each observation T_i. Define the set of points $\chi = \boldsymbol{G} \cup \bar{\boldsymbol{T}}$, where $\bar{\boldsymbol{T}} = \{T_i \in \boldsymbol{T} : \delta_i = 1\}$, then we compute the posterior distribution for the finite-dimensional object denoted by $l(\chi)$ as

$$p(l(\chi)|\lambda_0, \chi) \propto \left[\prod_{i=1}^{n} \sigma(l(T_i))^{\delta_i} \left(\prod_{g \in G_i} 1 - \sigma(l(g))\right)\right] N_{|\chi|}(l(\chi); 0, K(\chi, \chi)), \tag{4.3.6}$$

where $N_d(\boldsymbol{z}, 0, \Sigma)$ denotes the density function of a d-variate Normal distribution with zero mean and covariance function Σ, evaluated at $\boldsymbol{z} \in \mathbb{R}^d$. Notice this posterior is independent from λ_0.

Conditional distribution on a set of new locations

The posterior distribution of $(l(t))_{t \geq 0}$ is not a Gaussian process, but a different stochastic process on time. Nevertheless, it inherits the good properties of a Gaussian process. Let A denote a set of new locations such that $A \cap \chi = \emptyset$, then

$$\begin{aligned} p(l(A)|l(\chi), \chi) &= \frac{\prod_{i=1}^{n} \sigma(l(T_i))^{\delta_i} \left(\prod_{g \in G_i} 1 - \sigma(l(g))\right)}{\prod_{i=1}^{n} \sigma(l(T_i))^{\delta_i} \left(\prod_{g \in G_i} 1 - \sigma(l(g))\right)} \\ &\quad \times \frac{N_{|\chi \cup A|}(l(\chi); 0, K(\chi \cup A, \chi \cup A))}{N_{|\chi|}(l(\chi); 0, K(\chi, \chi))} \\ &= \frac{N_{|\chi \cup A|}(l(\chi); 0, K(\chi \cup A, \chi \cup A))}{N_{|\chi|}(l(\chi); 0, K(\chi, \chi))}. \end{aligned} \tag{4.3.7}$$

From this, we conclude that the posterior conditional distribution of l instantiated in a set of new locations A given $l(\chi)$ is a $|A|$-variate Gaussian distribution, with mean $\mu' = K(A, \chi)K(\chi, \chi)^{-1}l(\chi)$ and covariance matrix given by $\Sigma' = K(A, A) - K(A, \chi)K(\chi, \chi)^{-1}K(\chi, A)$. Notice this property allows us to

easily impute the sets of rejected jump times $\{G_i\}_{i=1}^n$.

4.4 Inference algorithm

In this section we describe the inference algorithm to sample from the posterior distributions. So far, we have described two ways of incorporating covariates. Nevertheless, from this moment onwards, we focus on the second approach, i.e., the model that incorporates covariates by indexing the Gaussian process in the index space $\mathbb{R}_+ \times \mathbb{R}^d$. We do this as this approach interests us more because it proposes a more flexible model. Indeed, notice that this model does not impose a proportional hazards constrain. A version of the algorithm for the proportional hazards model is not difficult to obtain from this analysis.

The data augmentation scheme proposed in section 4.1 suggests the following inference algorithm. For each data point $\{(T_i, \delta_i, \boldsymbol{x}_i)\}$ sample $G_i|(T_i, \delta_i, \boldsymbol{x}_i, \lambda_0, l)$, then $l|(\{(T_i, \delta_i, \boldsymbol{x}_i, G_i)\}_{i=1}^n, \lambda_0)$ and finalise sampling from $\lambda_0|(\{T_i, \delta_i, \boldsymbol{x}_i, G_i\}_{i=1}^n, l)$, where n is the number of data points. The sampling of l can be seen as a Gaussian process binary classification, where the points G_i and T_i are the set of rejected and accepted points, respectively. A variety of MCMC techniques can be used to sample l, see [48] for details.

For our algorithm we use the following notation. The datasets are denoted as $\{(T_i, \delta_i, \boldsymbol{x}_i)\}_{i=1}^n$. G_i will refer to the set of rejected points of T_i, and $\boldsymbol{G} = \bigcup_{i=1}^n G_i$ and $\boldsymbol{T} = \{T_1, \ldots, T_n\}$. For a point $t \in G_i \cup \{T_i\}$ we denote $l(t)$ instead of $l(t, \boldsymbol{x}_i)$, but remember that each point has an associated covariate. For a set of points A we denote $l(A) = \{l(a) : a \in A\}$. Also $2\Lambda_0(t)$ refers to $2\int_0^t \lambda_0(s)ds$ and $(2\Lambda_0(t))^{-1}$ denotes its inverse function. Finally, N denotes the number of iterations we are going to run our algorithm. The pseudo code of our algorithm is given in Figure 4.4.

Lines 2 to 11 sample the set of rejected points G_i for each survival time T_i. Particularly lines 3 to 5 used the Mapping theorem, which tells us how to map

Algorithm 2: Inference Algorithm.

Input: Number of data points n, set of (uncesored) times $\boldsymbol{T}$ and the Gaussian proces l instantiated in the data points.

1 **for** *q=1:N* **do**
  Update $\boldsymbol{G}$:
2 **for** *i=1:n* **do**
3 $n_i \sim \text{Poisson}(1; 2\Lambda_0(T_i))$;
4 $\tilde{C}_i \sim U(n_i; 0, 2\Lambda_0(T_i))$;
5 Set $A_i = (2\Lambda_0)^{-1}(\tilde{C}_i)$;
6 Set $\boldsymbol{A} = \cup_{i=1}^{n} A_i$
7 Sample $l(\boldsymbol{A})|l(\boldsymbol{G} \cup \boldsymbol{T}), \lambda_0$
8 **for** *i=1:n* **do**
9 $U_i \sim U(n_i; 0, 1)$
10 set $G_{(i)} = \{a \in A_i \text{ such that } U_i < 1 - \sigma(l(a))\}$
11 Set $\boldsymbol{G} = \cup_{i=1}^{n} G_i$
12 **Update parameters of** $\lambda_0(t)$
13 **Update** $l(\boldsymbol{G} \cup \boldsymbol{T})$ and hyperparameter of the kernel.

Figure 4.4: Pseudo-code for inference algorithm

a homogeneous Poisson process into a non-homogeneous with the appropriate intensity. Observe it makes use of Λ_0 and its inverse function, which shall be provided or be easily computable. The other lines evaluate the Gaussian process in the new locations, and decide which locations to keep in G_i (as in proposition 4.1.3). Line 7 is used to sample the Gaussian process in the set of points $\boldsymbol{A}$ given the values in the current set $\boldsymbol{G} \cup \boldsymbol{T}$. Observe initially $\boldsymbol{G} = \emptyset$. Right censoring is easy to handle. Indeed, if T_i is right censored we just do not include it in the set $\boldsymbol{T}$.

4.5 Approximation scheme

As it is shown is algorithm 2, in line 7 we need to sample the Gaussian process l in the set of points A and also update the values of l in the set $\boldsymbol{G} \cup \boldsymbol{T}$, see line 13. Both lines require matrix inversion and thus it scales poorly for massive datasets

that generate a large set $\boldsymbol{G}$. In order to help inference we use a random feature approximation of the Kernel.

4.5.1 Random Fourier feature approximation

In order to speed-up the inference we use the random feature approximation of the kernel proposed in [50]. A very simple explanation of this approach comes from the following ideas. On one hand, from the Euler's formula, we have that for every real number x, $e^{-ix} = \cos(x) + i\sin(x)$. Then, for a random variable X, the characteristic function can be written as $\psi_X(t) = \mathbf{E}(e^{itX}) = \mathbf{E}(\cos(tX)) + i\mathbf{E}(\sin(tX))$ for $t \in \mathbb{R}$. On the other hand, we have the relationship between a kernel K, and a characteristic function given by Bochner's theorem.

Theorem 4.5.1 (Bochner [52])**.** *A continuous kernel $k(x, y) = k(x - y)$ on $\mathbb{R}^d$ is positive definite if and only if $k(\delta)$ is the Fourier transform of a positive measure μ.*

Hence, Bochner's theorem guarantees that if we have a continuous and stationary kernel K, then, under the appropriate re-scaling, it can be written as the Fourier transform of a proper probability measure P. Mixing both ideas, we can obtain an approximation of the kernel K, which is associated with the probability measure P, by using a Monte Carlo approximation based on a random sample from P. This idea is illustrated in the following examples.

Example 4.5.2. Consider the squared exponential kernel denoted by $K(t) = e^{-\frac{1}{2l^2}|t|^2}$. Let $X \sim N(0, \sigma^2)$. It is well known that the real part of the characteristic function of X corresponds to $\text{Re}(\psi_X(t)) = e^{-\frac{1}{2}\sigma^2|t|^2}$. Then, we make a Monte Carlo approximation of the kernel by taking a random sample $\{s_k\}_{k=1}^m \sim N(0, l^{-2})$, indeed,

$$K(t) = e^{-\frac{1}{2l^2}t^2} = \int \cos(tx)\frac{l}{\sqrt{2\pi}}e^{-\frac{l^2}{2}x^2}dx \approx \sum_{k=1}^{m}\cos(s_k t)$$

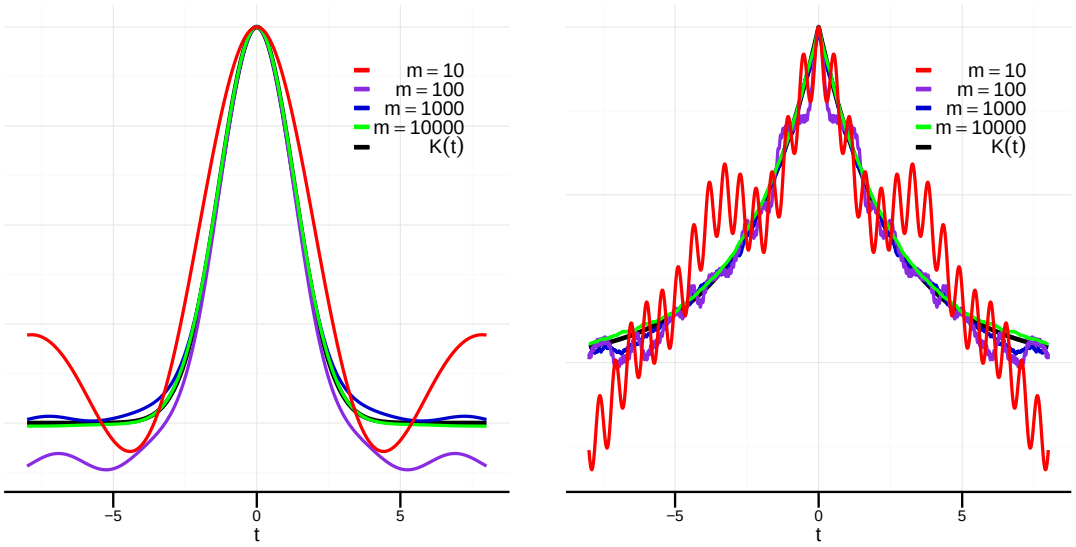

Figure 4.5: Random Fourier feature approximations of the squared exponential and Ornstein-Uhlenbeck kernel respectively, where m in the legend denotes the number of random features.

Example 4.5.3. Another example is the Ornstein-Uhlenbeck kernel, $K(t) = e^{-\frac{1}{2l}|t|}$. For this setting, we take a collection of i.i.d. Cauchy random variables $\{s_k\}_{k=1}^m$, with $s_k \sim \text{Cauchy}(0, \frac{1}{2l})$ and then

$$K(t) = e^{-\frac{1}{2l}|t|} = \int \cos(tx) \frac{2l}{\pi} \left[\frac{1}{\frac{x^2}{4l^2} + 1} \right] dx \approx \sum_{k=1}^{m} \cos(s_k t).$$

A Gaussian process with an approximated kernel

We use the previous approximation scheme to define an alternative Gaussian process which approximates the covariance function of the original Gaussian process. We give a construction for the squared exponential kernel but recall that, from the above result, this can be done for any continuous and stationary kernel (as long as we can identify the probability measure P).

Let $\{s_k\}_{k=1}^m$ be a random sample from $N(0, l^{-2})$ and let each $\{a_k\}_{k=1}^m$ and

$\{b_k\}_{k=1}^m$ be independent random samples from $N(0, \sigma^2/m)$. Then, conditioned on the values of s_k, we define the Gaussian process $(g(t))_{t\geq 0}$ as

$$g(t) = \sum_{k=1}^{m} a_k \cos(s_k t) + b_k \sin(s_k t). \tag{4.5.1}$$

Notice that g is a Gaussian process as, conditioned on the values s_k, it is written as a sum of independent and normally distributed random variables. Moreover, it approximates the covariance of the squared exponential kernel as

$$\begin{aligned}
\mathbf{Cov}(g(t), g(s)|\{s_k\}_{k=1}^m) &= \sum_{k=1}^{m}\sum_{k'=1}^{m'} \mathbf{Cov}(a_k \cos(s_k t) + b_k \sin(s_k t), a_{k'} \cos(s_{k'} s) \\
&+ b_{k'} \sin(s_{k'} s)|\{s_k\}_{k=1}^m) \\
&= \sum_{k=1}^{m} \cos(s_k t)\cos(s_k s)\mathbf{Var}(a_k) + \sin(s_k t)\sin(s_k s)\mathbf{Var}(b_k) \\
&= \sigma^2 \sum_{k=1}^{m} \cos(s_k(t-s)) \approx \sigma^2 e^{-\frac{1}{2l^2}|t-s|^2}.
\end{aligned} \tag{4.5.2}$$

for any $t, s \in \mathbb{R}_+$.

4.5.2 Approximation scheme

Consider the times $t, s \in \mathbb{R}_+$, and covariates $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. Then, the kernel used in our experiments corresponds to

$$K((t, \boldsymbol{x}), (s, \boldsymbol{y})) = K_0(t, s) + \sum_{j=1}^{d} x_j y_j K_j(t, s), \tag{4.5.3}$$

where each K_j is a square exponential kernel with overall variance σ_j^2 and length scale parameter ϕ_j. Based on the above approximation, for a fixed $m \in \mathbb{N}$, we

propose the following Gaussian process which approximates the kernel of interest

$$g(t, \boldsymbol{x}) = g_0(t) + \sum_{j=1}^{d} x_j g_j(t). \tag{4.5.4}$$

Here, each $g_j(t) = \sum_{k=1}^{m} a_k^j \cos(s_k^j t) + b_k^j \sin(s_k^j t)$, and each a_k^j and b_k^j are independent samples of $\mathcal{N}(0, \sigma_j^2)$ where σ_j^2 is the overall variance of the kernel K_j. Moreover, s_k^j are independent samples of $\mathcal{N}(0, \phi_j^{-2})$ where ϕ_j is the length scale parameter of the kernel K_j. These features remains fixed as g is a Gaussian process conditioned on these values.

The inference algorithm for this scheme is practically the same, except for two small changes. The values $l(A)$ in line 7 are easier to evaluate because we only need to know the values of the a_k^j and b_k^j, and no matrix inversion is needed. In line 13 we just need to update all values a_j^k and b_j^k. Since they are independent variables there is no need for matrix inversion.

Chapter 5

Experiments

In this chapter, we give an empirical evaluation of the proposed methods. This evaluation is performed on Synthetic data and real data.

5.1 Synthetic data

5.1.1 Half Normal

The Half-Normal distribution is a probability measure on $\mathbb{R}_+$ obtained by considering the absolute value of a Normal random variable with zero mean and variance σ^2. Let X be a Normal random variable with zero mean and variance σ^2, then $Y = |X|$ follows a half Normal distribution with probability density function given by

$$f(y) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} e^{-\frac{y^2}{2\sigma^2}}, \quad y \geq 0. \tag{5.1.1}$$

Consider a random sample $T_1, \ldots, T_n$ from the distribution above with $n = 10, 25, 50, 100$ and 200. For our first experiment, we estimate the survival function, the hazard function and the cumulative hazard function for each random sample. Then, we assess how these estimations vary as the sample size n increases. Observe

this setting falls under the uncensored case.

In order to give the required estimates, we adopt the model proposed in equation (2.1.1). In particular, we consider a constant baseline hazard function $\lambda_0(t) = \Omega$ and a Gamma prior distribution over Ω. For the specification of the Gaussian process, we choose a squared exponential kernel $K_b(t) = e^{-|t|^2/(2b)^2}$ with fixed length-scale parameter $b = 1$. Inference in this model is performed by using the exact MCMC inference scheme described in algorithm (4.4).

For the results shown below, we run $N = 30500$ iterations of the MCMC sampler, from which we burn the initial 5000, and then save one sample every 50 iterations, obtaining 500 samples in total. The parameter Ω is initialized on the maximum likelihood under the assumption of an Exponential distribution for the data. This is related to the choice of the baseline hazard function as the hazard function of an Exponential random variable. The choice of the hyper-parameters associated with the Gamma prior over Ω is such that the mean is also centred on the maximum likelihood. The estimators for the survival function, the hazard function and cumulative hazard function are shown in Figures 5.1, 5.2 and 5.3, respectively.

Results

We start by setting the range of the x-axis in each plot of Figures 5.1, 5.2 and 5.3, between zero and twice the maximum value of the random sample $\max\{T_1, \ldots, T_n\}$. The purpose is to show the behaviour of our estimates of the survival function, the hazard function and the cumulative hazard function, both in the range in which observe data and where we do not observe observations any more.

In the case of the survival function, we observe that as the sample size increases, the quality of the estimation improves and that the posterior seems to be shrinking towards the true survival function, both in the range with and with-

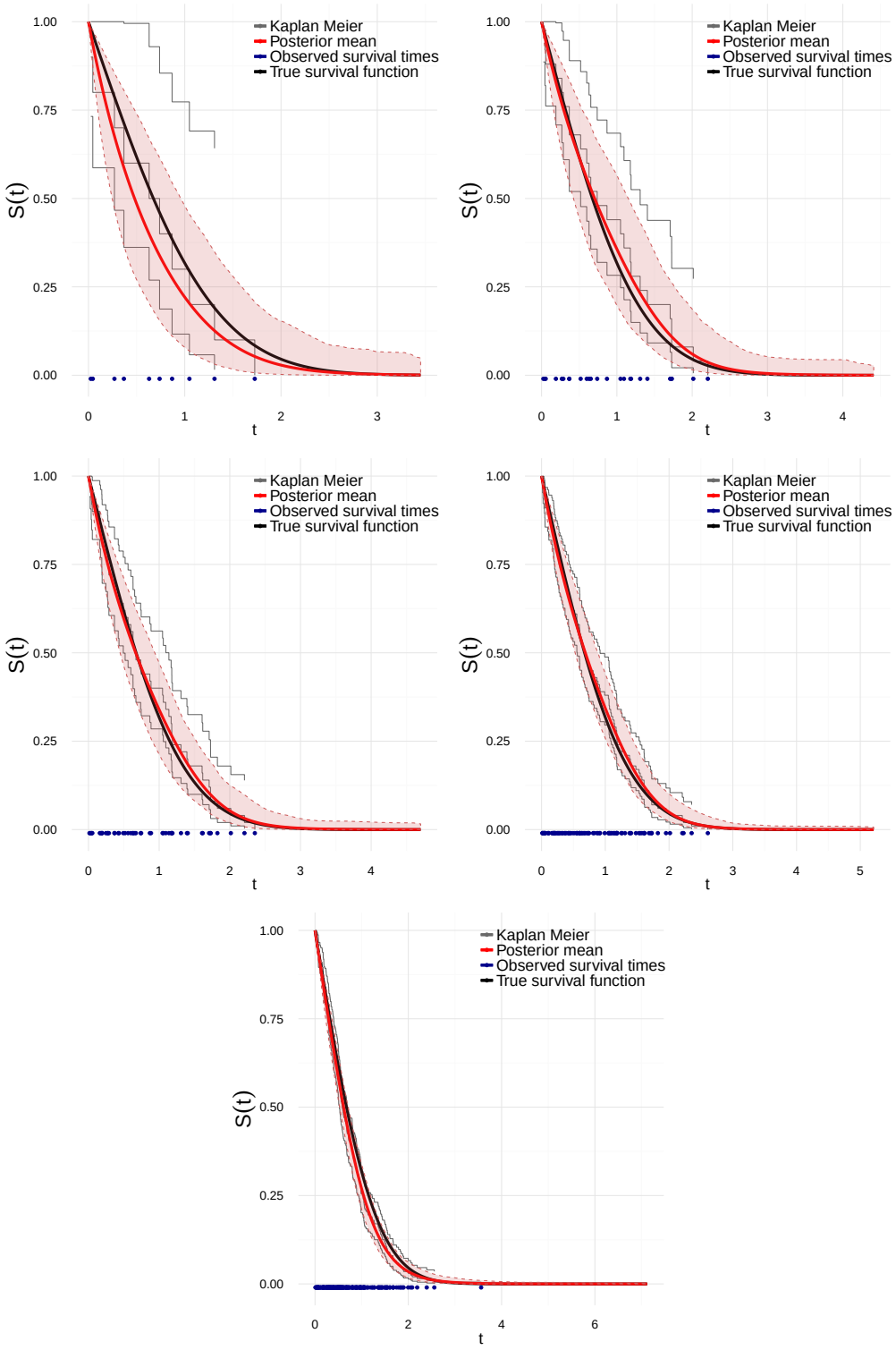

Figure 5.1: Estimated survival functions. From left-to-right and top-to-bottom, sample size $n = 10, 25, 50, 100$ and 200.

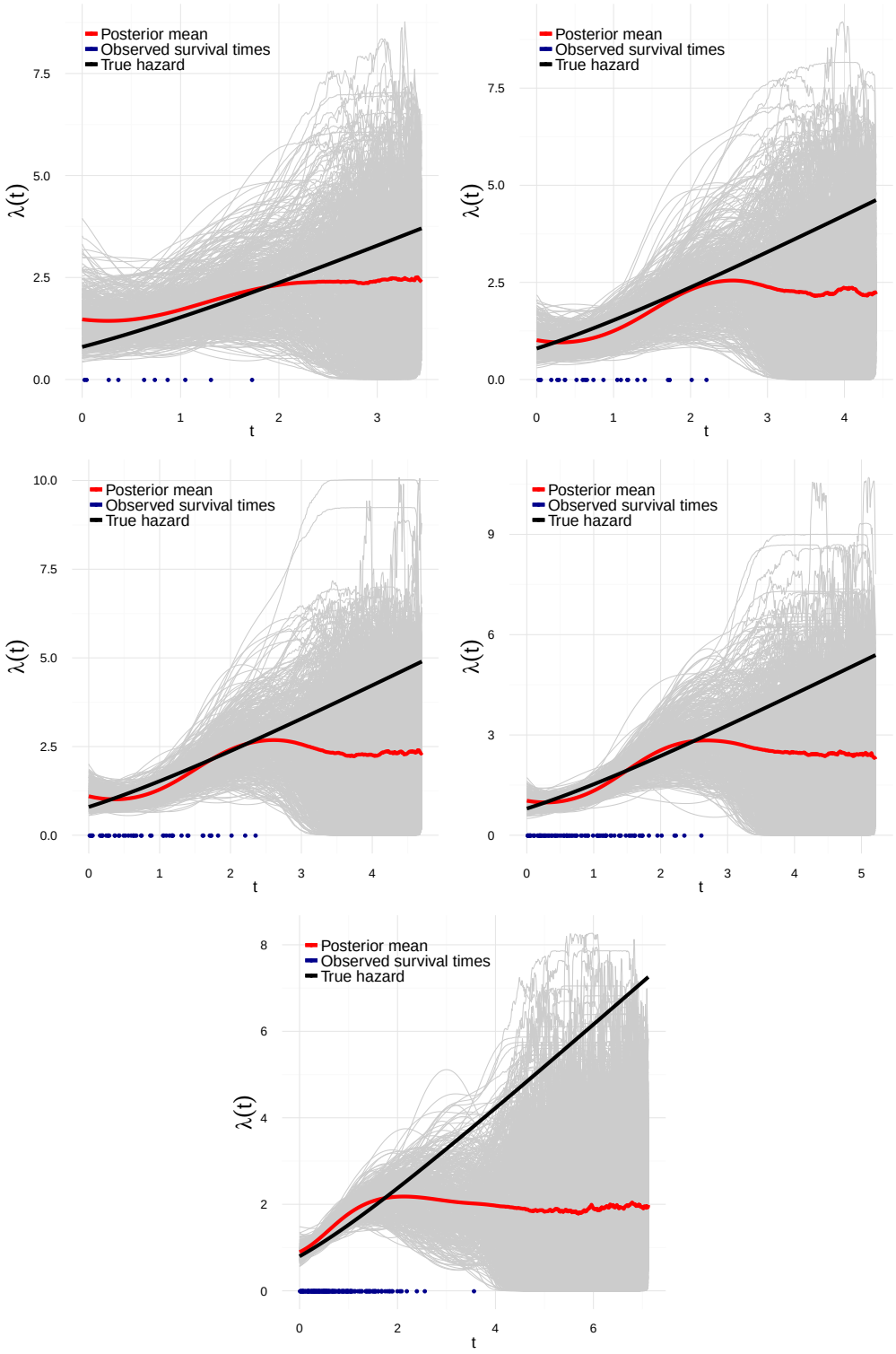

Figure 5.2: Estimated hazard functions. From left-to-right and top-to-bottom, sample size $n = 10, 25, 50, 100$ and 200.

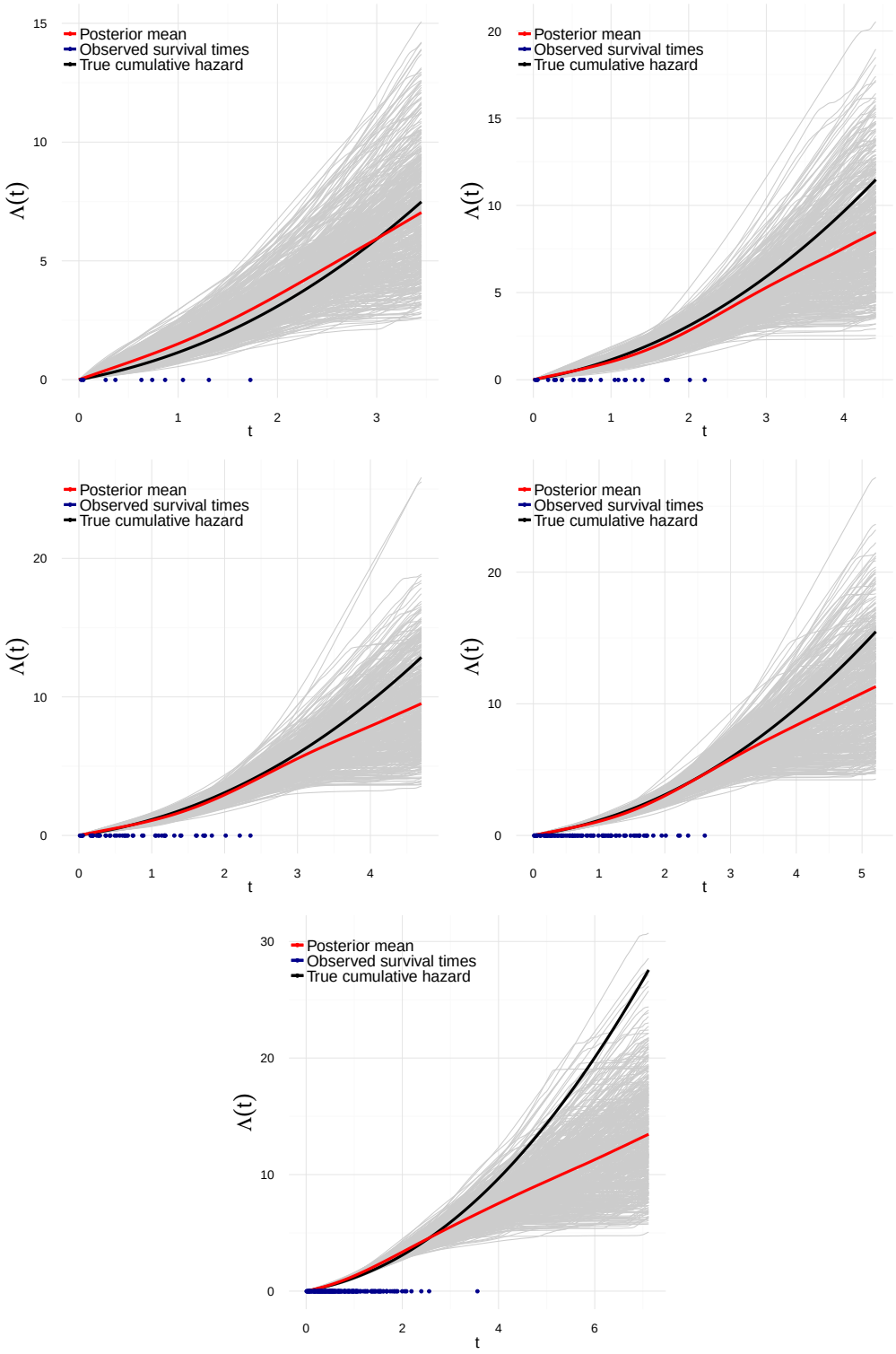

Figure 5.3: Estimated cumulative hazard functions. From left-to-right and top-to-bottom, sample size $n = 10, 25, 50, 100$ and 200.

out observations. This does not seems to be true for the case of the hazard and cumulative hazard functions. In particular, for these functions the estimations seems to be relatively accurate in the range in which we observe data but quickly deteriorate towards the range without observations. This behaviour is even more accentuated for the hazard function. The reason for this, is that in the region without observations the posterior approaches to the prior, which for this example, is underestimating the value of the true hazard.

The effect described above can be explicitly seen by observing the trace plot of the parameter Ω for $n = 200$ (Figure 5.4) and in the plot showing the estimate of the hazard function for $n = 200$ (Figure 5.2). In particular, we observe that after the last observation occurs, the hazard function quickly centres at the value 2. This coincides with the mean value taken by the parameter Ω in the trace plot. This makes sense as the prior for the hazard function is defined as $\lambda(t) = 2\Omega\sigma(l(t))$, where $\sigma(l(t))$ is a stochastic process centred on $1/2$. This experiment suggests us a notion of convergence of survival functions but not for hazard or cumulative hazard functions.

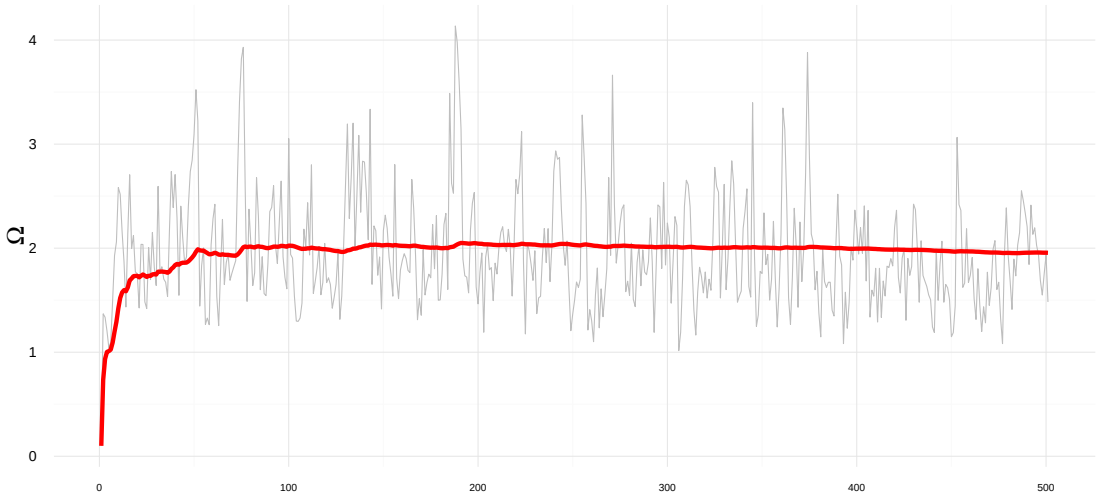

Figure 5.4: Trace plot associated with Ω for $n = 200$.

5.1.2 Crossing hazards

In this section, we reproduce the experiment proposed in [14]. In this work, the authors generate data from two different groups following each of the following densities

$$f_0(x) = \frac{1}{\sqrt{2\pi}0.8} e^{-\frac{1}{2\times 0.8^2}(x-3)^2}, \tag{5.1.2}$$

and

$$f_1(x) = 0.4\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-4)^2} + 0.6\frac{1}{\sqrt{2\pi}0.8} e^{-\frac{1}{2\times 0.8^2}(x-2)^2}. \tag{5.1.3}$$

We simulate $n = 25, 50, 100$ and 150 points from each density, restricted to $\mathbb{R}_+$. The data contains the sample points and a covariate indicating if such points were sampled from the p.d.f. f_0 or f_1. Additionally, to each data point, we add 3 noisy covariates taking random values in the interval $[0, 1]$. We report the results for the model proposed in equation (2.1.1) for a Weibull and a Exponential baseline hazard function. All the experiments are performed using our approximation scheme of equation (4.5.4) with a value of $m = 50$. The implementation is exactly the same as in Algorithm 2.

For the Weibull baseline hazard function $\lambda_0(t) = 2\beta t^{\alpha-1}$, we consider a Gamma prior over β and a uniform prior $U(0, 2.3)$ over α. The posterior distribution for β is conjugate, thus we can easily sample from it. For the prior distribution of α, we constrain the support to $(0, 2.3)$ as the expected number of augmented data points greatly increases with this parameter, making inference slower. For the Exponential baseline hazard function $\lambda_0(t) = 2\Omega$, we consider a Gamma prior over the parameter Ω. We refer to both models as the Weibull model (W-SGP) and the Exponential model (E-SGP). As the tuning of initial parameters can be hard, we propose to use the maximum likelihood estimator as initial parameters, and choose hyper-parameters such that the priors are centred on these values.

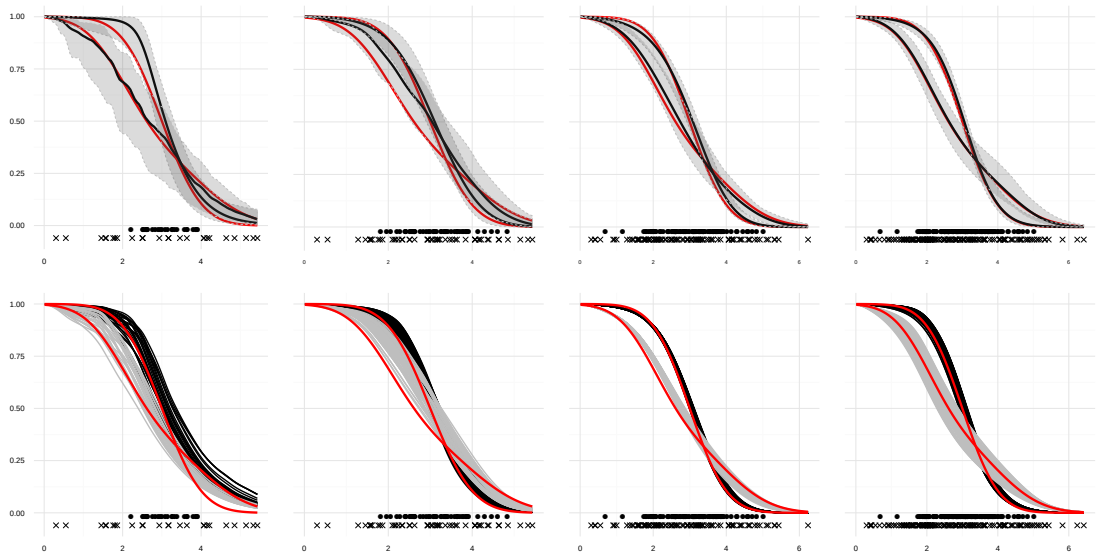

Figure 5.5: Weibull Model. First row: clean data, Second row: data with noise covariates. Per columns we have 25, 50, 100 and 150 data points per each group (shown in X-axis) and data is increasing from left to right. Dots indicate data is generated from p_0, crosses, from p_1. In the first row a credibility interval is shown. In the second row each curve for each combination of noisy covariate is given.

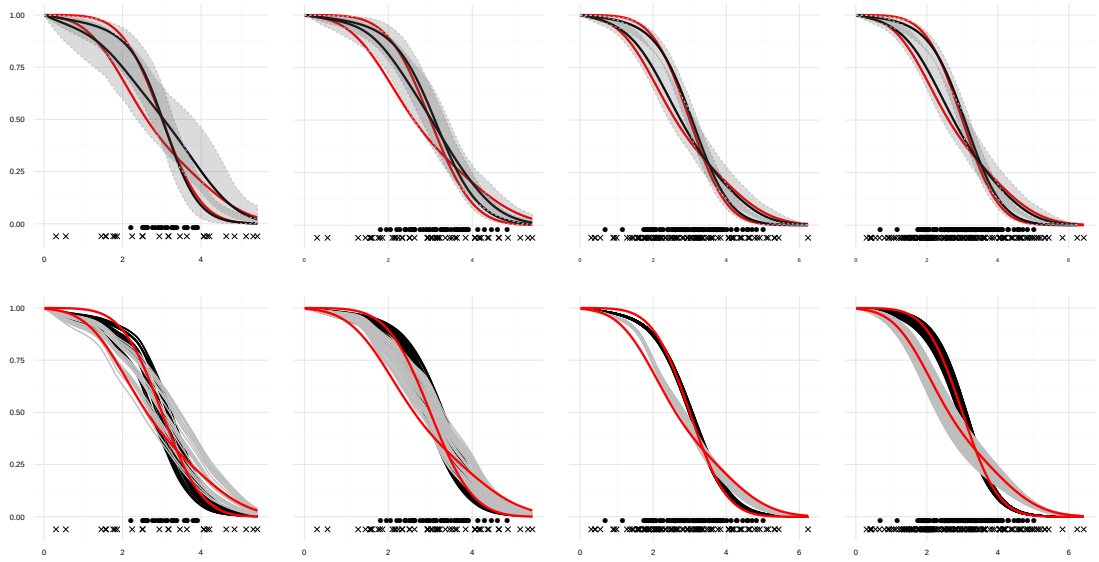

Figure 5.6: Exponential Model. First row: clean data, Second row: data with noisy covariates. Per columns we have 25,50,100 and 150 data points per each group (shown in X-axis) and data is increasing from left to right. Dots indicate data is generated from density p_0, crosses, from p_1. In the first row a confidence interval for each curve is given. In the second row each curve for each combination of noisy covariate is shown.

For the experiments, we set all the kernels K_j to be squared exponential kernels with length-scale parameter ϕ_j and overall variance σ_j^2, i.e. $K_j(t) = \sigma_j^2 e^{-t^2/\phi_j^2}$. For the length-scale parameters ϕ_j we use a log-Normal prior, and for the variance σ_j^2 a Gamma prior (an Inverse-Gamma is also useful since it is conjugate). The resulting survival functions are shown in Figure 5.5, for the Weibull model and Figure 5.6, for the Exponential model. These results were obtained by running $N = 8050$ iterations, from which the initial 3000 were burned, and 1 observation was saved every 100 samples, obtaining 50 samples from the posterior.

In the clean case, the more data the better the approximation. In particular, we can detect almost perfectly the cross in the survival function. In the noisy dataset, it appears that with few data points, the noisy covariates seem to have an effect in the precision of our estimation, nevertheless the more points the more precise. For this set of experiments, we do not report a significant difference between both models.

Increasing right censoring

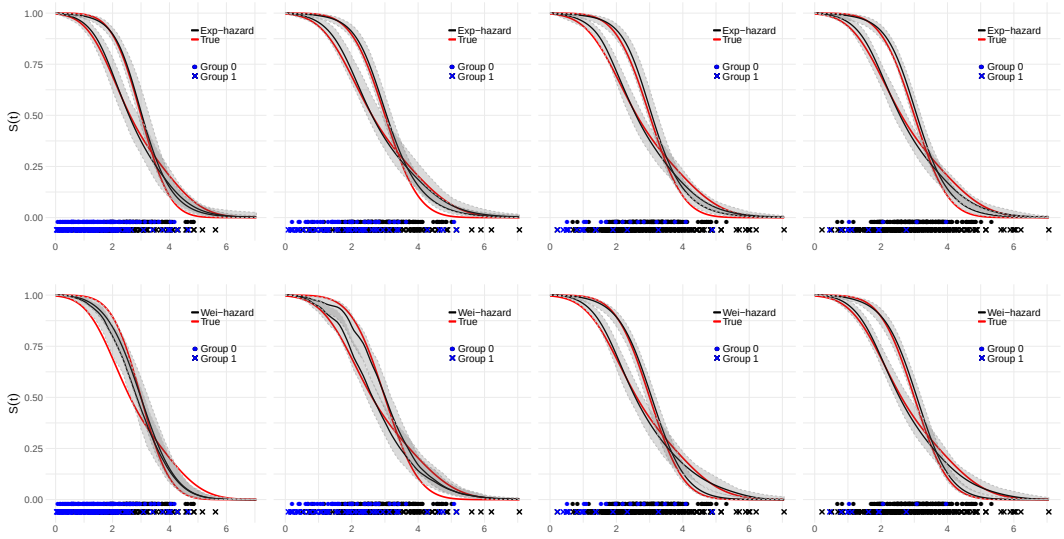

Figure 5.7: From left to right: estimates for the survival curves for approximately 50, 25, 10 and 5 percent of censored observations (censored observations appears in blue). First row: E-SGP model. Second row: W-SGP model.

We repeat the experiment for the clean data with $n = 150$ but considering independent right censoring. In particular, we sample the censored random variables from an Exponential distribution in such a way the resulting data contains approximately $50\%, 25\%, 10\%$ and 5% censored observations in average. As the percentage of censored observations decreases, our estimates improve. Nevertheless, there is a clear negative effect of censoring on the estimates.

5.2 Real data experiments

C-Index

To compare our models we use the so-called concordance index (C-Index). The concordance index is a standard measure in survival analysis which estimates how good the model is at ranking survival times. We consider a set of survival times with their respective censoring indices and set of covariates $(T_1, \delta_1, \boldsymbol{x}_1), \ldots, (T_n, \delta_n, \boldsymbol{x}_n)$. On this particular context, we only consider right censoring.

To compute the C-index, consider all possible pairs $(T_i, \delta_i, \boldsymbol{x}_i; T_j, \delta_j, \boldsymbol{x}_j)$ for $i \neq j$. We call a pair admissible if it can be ordered. If both survival times are right-censored i.e. $\delta_i = \delta_j = 0$ it is impossible to order them, we have the same problem if the smallest of the survival times in a pair is censored, i.e. $T_i < T_j$ and $\delta_i = 0$. All the other cases under this context will be called admissible.

Given just the covariates $\boldsymbol{x}_i, \boldsymbol{x}_j$ and the status δ_i, δ_j, the model has to predict if $T_i < T_j$ or the other way around. We compute the C-index by considering the number of pairs which were correctly sorted by the model, given the covariates, over the number of admissible pairs. A larger C-index indicates the model is better at predicting which patient dies first by observing the covariates. If the C-index is close to 0.5, it means the prediction made by the model is close to random.

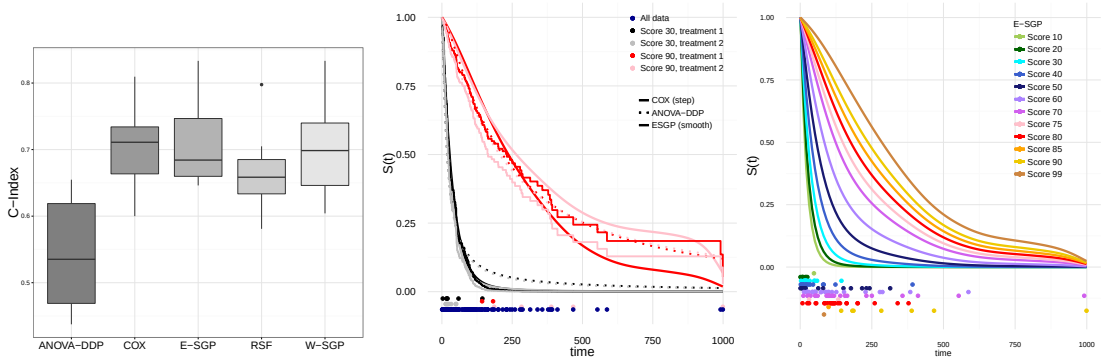

Figure 5.8: **Left panel**: C-Index for ANOVA-DDP, COX, E-SGP, RSF, W-SGP; **Middle panel**: Survival curves obtained for the combination of score: 30, 90 and treatments: 1 (standard) and 2 (test); **Right panel**: Survival curves, using W-SGP, across all scores for fixed treatment 1, diagnosis time 5 moths, age 38 and no prior therapy. (Best viewed in colour)

Veteran data

We run experiments on the Veteran data, available in the R-package survival [55]. Veteran consists of a randomized trial of two treatment regimes for lung cancer. It has 137 samples and 5 covariates: **treatment** indicating the type of treatment of the patients, their **age**, the **Karnofsky performance score**, and indicator for **prior treatment** and **months from diagnosis**. It contains 9 censored times, corresponding to right censoring.

In the experiments, we run our model with a Weibull (W-SGP) and Exponential (E-SGP) baseline hazard function, respectively. In particular, the experiments are performed using our approximation scheme of equation (4.5.4) with a value of $m = 50$. The implementation is exactly the same as in Algorithm 2. We also use ANOVA DDP, Cox proportional hazards and random survival forest.

For the Weibull baseline hazard function $\lambda_0(t) = 2\beta t^{\alpha-1}$, we consider a Gamma prior over β and a Uniform prior $U(0, 2.3)$ over α. For the Exponential baseline hazard function $\lambda_0(t) = 2\Omega$, we consider a Gamma prior over the parameter Ω. In particular, we use the maximum likelihood estimates related to each baseline hazard function as initial parameters for each model. The hyper-parameters of

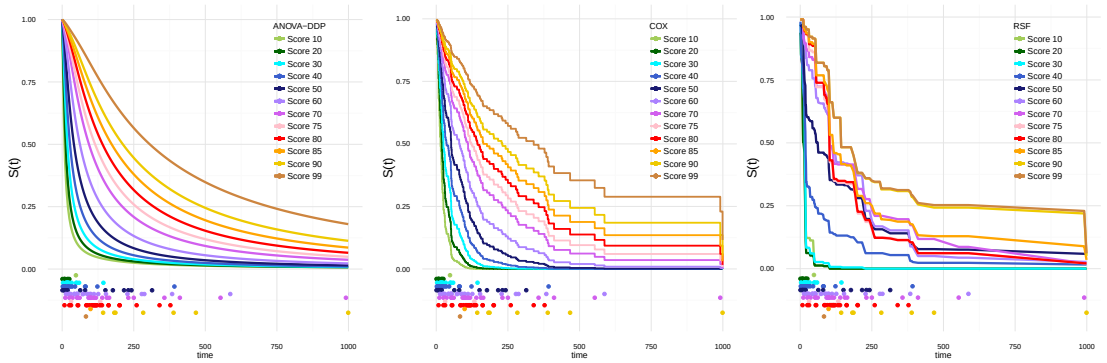

Figure 5.9: Survival curves across all scores for fixed treatment 1, diagnosis time 5 months, age 38 and no prior therapy. **Left:** ANOVA-DDP; **Middle:** Cox proportional; **Right:** Random survival forests.

the prior are chosen in such a way that the mean of the prior is centred on these maximum likelihood estimates. For the length-scale parameters ϕ_j of the squares exponential kernels K_j we use a Log-Normal prior, and for the variance σ_j^2 a Gamma prior .

We perform 10-fold cross validation and compute the C-index for each fold. Figure 5.8 reports the results. For this dataset the only significant variable corresponds to the **Karnofsky performance score**. In particular as the values of this covariate increases, we expect an improved survival time. All the studied models achieve such behaviour and suggest a proportionality relation between the hazards. This is observable in the C-Index boxplot as we can observe good results for the proportional hazards model. Nevertheless, our method detect some differences between the treatments when the **Karnofsky performance score** is 90, as it can be seen in figure 5.8.

For the other competing models we observe an overall good result. In the case of ANOVA-DDP we observe the lowest C-INDEX. In figure 5.9 we see that ANOVA-DDP seems to be overestimating the Survival function for lower scores. Arguably, our survival curves are more visually pleasant than Cox proportional hazards and Random Survival Trees.

Chapter 6

Posterior consistency

6.1 Introduction

In this chapter we verify posterior consistency for a simplified version of the model proposed in Chapter 2 which assumes survival times on $\mathbb{R}_+$ and covariates on $[0,1]^d$. The relationship between time and covariates is modelled by using the kernel defined in equation (2.3.2). The posterior consistency result we prove asserts that

$$\lim_{n\to\infty} \Pi(d(\theta,\theta_0) > \epsilon | \mathcal{D}_n) \quad = \quad 0, \quad \mathbf{P}^n_{\theta_0} \quad a.s., \tag{6.1.1}$$

that is, we have an in-probability convergence of θ to θ_0 (with respect the metric d) $\mathbf{P}^n_{\theta_0}$ almost surely on the data $\mathcal{D}_n$, i.e., assuming the data $\mathcal{D}_n$ was jointly generated under $\mathbf{P}^n_{\theta_0}$. In particular, the metric d studied corresponds to the largest deviation of probability distributions functions, indexed in θ and θ_0 respectively, over the class of all possible rectangles on $\mathbb{R}_+ \times [0,1]^d$. Specifically, the considered metric corresponds to

$$d(\theta,\theta_0) = \sup_{A\in\mathcal{A}} |\mathbf{P}_{\theta_0}((T,\mathbf{X}) \in A) - \mathbf{P}_{\theta}((T,\mathbf{X}) \in A)|, \tag{6.1.2}$$

where $\mathcal{A}$ denotes the class of rectangles. We base our proof on the extension of Schwartz's theorem for non-i.i.d. observations given in [10]. Hence, our result is based on the existence of tests with exponentially small type I and type II error, and the positivity of Kullback-Leibler neighbourhoods of the true parameter. In particular, we use results from the Vapnik-Chervonenkis (VC) theory for constructing the required tests, and properties of the reproducing kernel Hilbert space attached to a Gaussian process for verifying the result regarding the Kullback-Leibler neighbourhoods. Complete reviews for both theories can be found on [15] and [58] respectively.

6.2 Simplified model

Consider $T_1, \dots, T_n$ a sequence of random survival times taking values on $\mathbb{R}_+$, and let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ be the collection of associated covariates, each one taking values on $[0,1]^d$. Then, we introduce a simplified version of the model introduced in (2.1.1). In particular, we define

$$\begin{aligned} T_i|\lambda, \boldsymbol{x}_i &\overset{ind}{\sim} f_{\boldsymbol{x}_i}(t), \\ f_{\boldsymbol{x}_i}(t) &= \lambda_{\boldsymbol{x}_i}(t) e^{-\int_0^t \lambda_{\boldsymbol{x}_i}(s)ds}, \\ \lambda_{\boldsymbol{x}_i}(t) &= \Omega\sigma\left(\eta_0(t) + \sum_{j=1}^{d} x_{i,j}\eta_j(t)\right), \\ \Omega &\sim \nu, \\ (\eta_j(t))_{t\geq 0} &\overset{ind.}{\sim} \mathcal{GP}(0, K_j), \quad j = 1, \dots, d, \end{aligned} \tag{6.2.1}$$

where $\sigma(x) = (1+e^{-x})^{-1}$ is the sigmoid function, ν denotes a probability measure on $\mathbb{R}_+$, and K_j is a stationary covariance function for all $j = 0, \dots, d$. We assume the kernels are such that they generate a continuous Gaussian process. Observe that the boundedness of the link function σ is a necessary condition for applying the inference scheme proposed in Chapter 4, but not necessarily for giving a correct

definition of a prior over hazard functions. For simplicity, we adopt this particular definition of σ, but at the end of this chapter, we discuss about its relevance for our results. We also adopt the constraints for the kernel given in Chapter 3 to have a well-defined model for survival data.

6.3 Posterior consistency

Consider the parameter space $\Theta = \mathbb{R}_+ \times \mathcal{F}^{d+1}$, where $\mathcal{F}$ denotes a space of functions. We assume that there exists a true parameter $\theta_0 = (\Omega_0, \eta_{0,0}, \ldots, \eta_{d,0})$ on Θ, which determines the true survival function $S_x^0(t)$. The aim of this chapter is to find conditions over the prior Π on Θ and true parameter $\theta_0 \in \Theta$, to ensure this model achieves posterior consistency of the survival functions defined on $\mathbb{R}_+$.

Schwartz [53] gave conditions for proving posterior consistency at θ_0 under a loss function $d(\theta, \theta_0)$ in the context of independent and identically distributed random variables. These conditions include the existence of a sequence of tests with exponentially small type I and type II error, and prior positivity of Kullback-Leibler neighbourhoods of the true parameter θ_0. A slightly more involved scenario is when the observations are not identically distributed. In this setting, Choudhuri et al. [11] extended Schwartz's theorem in the case of convergence in-probability. The almost surely convergence result was given by Choi et al. in [10]. We proceed to enunciate such result.

Theorem 6.3.1 (Choi and Schervish)**.** *Let $\{Z_i\}_{i=1}^{\infty}$ be independently random variables with densities $(f_i(\cdot, \theta))_{i=1}^{\infty}$, with respect to a common σ-finite measure, where the parameter θ belongs to an abstract measurable space Θ. The densities $f_i(\cdot; \theta)$ are assumed to be jointly measurable. Let $\theta_0 \in \Theta$ and let P_{θ_0} stand for the joint distribution of $\{Z_i\}_{i=1}^{\infty}$ when θ_0 is the true value of θ. Define the loss function $d(\theta, \theta_0)$ and let $U_\epsilon = \{\theta : d(\theta, \theta_0) \leq \epsilon\}$ be a set of Θ for all $\epsilon > 0$. Let θ have a*

prior Π *on* Θ. *Define*

$$\begin{aligned} \Upsilon(\theta_0, \theta) &= \log \frac{f_i(Z_i; \theta_0)}{f_i(Z_i; \theta)}, \\ KL_i(\theta_0, \theta) &= \mathbf{E}_{\theta_0}(\Upsilon(\theta_0, \theta)), \\ V_i(\theta_0, \theta) &= \mathbf{Var}_{\theta_0}(\Upsilon(\theta_0, \theta)). \end{aligned}$$

(T1) Prior positivity of neighbourhoods.

Suppose that for all $\varepsilon > 0$ *there exits a set* B_ε *with* $\Pi(B_\varepsilon) > 0$ *such that*

(i) $\sum_{i=1}^{\infty} \frac{V_i(\theta_0,\theta)}{i^2} < \infty, \forall \theta \in B_\varepsilon$,

(ii) $\Pi(B_\varepsilon \cap \{\theta : KL_i(\theta_0, \theta) < \varepsilon \text{ for all } i\}) > 0$

(T2) Existence of tests.

Suppose that there exist test functions $(\Phi_n)_{n=1}^{\infty}$ *and constants* $C_1, c_1 > 0$ *such that*

(i) $\sum_{n=1}^{\infty} \mathbf{E}_{\theta_0} \Phi_n < \infty$,

(ii) $\sup_{\theta \in U_\varepsilon^c} \mathbf{E}_\theta (1 - \Phi_n) \leq C_1 e^{-c_1 n}$,

Then for all $\varepsilon > 0$

$$\Pi(\theta \in U_\varepsilon^c | Z_1, \ldots, Z_n) \to 0 \qquad a.s.[\mathbf{P}_{\theta_0}].$$

Remark 6.3.2. It is worth noticing that for i.i.d. samples, we recover Schwartz's theorem by replacing condition (T1) with the following condition.

(G1) The parameter $\theta_0 \in \Theta$ is such that for every $\varepsilon > 0$,

$$\Pi\left(\theta : \mathbf{E}_{\theta_0}\left(\log \frac{f(Z, \theta_0)}{f(Z, \theta)} < \varepsilon\right)\right) > 0.$$

in which $f(Z, \theta)$ does not depend on i, as we have i.i.d samples. Moreover, we do not need any condition on the variance.

6.4 Framework

In this section we introduce all the necessary elements before introducing the main result.

6.4.1 Notation

In order to introduce the notation, we first specify some general assumptions on how the covariates arise in the model. In particular, we adopt the approach given in [25], and consider both random and fixed design for the covariates.

- **[RD] Random design:** We assume the covariates X are distributed according a probability distribution Q on $[0,1]^d$.
- **[NRD] Non random design:** In this case, we assume the covariates are fixed. In particular, given the sequence of covariates $(\boldsymbol{x}_i)_{i\geq 1}$, we state our results as function of their empirical distribution $Q_n = n^{-1}\sum_{i=1}^n \delta_{\boldsymbol{x}_i}$.

For the random design, notice that conditioned on $X = \boldsymbol{x}$, the probability density function for the survival lifetimes is fully specified by the parameter $\theta = (\Omega, \eta_0, \eta_1, \ldots, \eta_d) \in \Theta$, then we use $f_{\boldsymbol{x}}(t,\theta)$ when referring to it. We use the same notation for cumulative density function $F_{\boldsymbol{x}}(t,\theta)$, the hazard function $\lambda_{\boldsymbol{x}}(t,\theta)$ and the cumulative hazard function $\Lambda_{\boldsymbol{x}}(t,\theta)$ given the covariate $X = \boldsymbol{x}$. We adopt the same notation for the fixed design case, but understand from the context that $f_{\boldsymbol{x}}(t,\theta)$ makes reference to the distribution of times under the fixed covariate $\boldsymbol{x}$, i.e $\boldsymbol{x}$ is not random, nor does $f_{\boldsymbol{x}}(t,\theta)$ denote a conditional density function.

Furthermore, for the random design case, we denote the join probability measure on pairs $\{(T_i, X_i)\}_{i\geq 1}$ under the parameter θ, by $\mathbf{P}_\theta$. We also denote by $\mu_\theta(A) = \mathbf{P}_\theta((T_i, X_i) \in A)$, for any i (remember the pairs (T_i, X_i) are i.i.d), and $A \subseteq \mathbb{R}_+ \times [0,1]^d$. For the fixed design case, our probability measure depends on θ as well as on the covariates $\boldsymbol{x} = (\boldsymbol{x}_i)_{i\geq 1}$. We denote by $\mathbf{P}_\theta(\cdot|\boldsymbol{x})$ the joint distribution of $(T_i)_{i\geq 1}$ associated with the fixed sequence of covariates

$(\boldsymbol{x}_i)_{i\geq 1}$. Sometimes, to shorten the notation we write $\tilde{\mathbf{P}}_\theta(\cdot) = \mathbf{P}_\theta(\cdot|\boldsymbol{x})$. We define $\tilde{\mu}_\theta^{\boldsymbol{x}_i}([a,b]) = \mathbf{P}_\theta(T_i \in [a,b]|\boldsymbol{x}_i)$. Note that all measures $\tilde{\mu}_\theta^{\boldsymbol{x}_i}$ are potentially different since the distribution of T_i depends on its covariate $\boldsymbol{x}_i$.

6.4.2 Metric for survival analysis

We provide a different metric for each design of covariates, **[RD]** and **[NRD]**. As we already stated, our object of interest are the survival functions $S_{\boldsymbol{x}}(t)$ for $\boldsymbol{x} \in [0,1]^d$. Under the further assumption that the covariates X follow a distribution Q on $[0,1]^d$ (**[RD]**), a sensible way of assessing the quality of our estimators is by considering

$$d_B(\theta,\theta_0) \quad = \quad \sup_{t\in\mathbb{R}_+} \left| \int_B (S_{\boldsymbol{x}}(t,\theta) - S_{\boldsymbol{x}}(t,\theta_0)) Q(d\boldsymbol{x}) \right|, \tag{6.4.1}$$

with $B \subseteq [0,1]^d$. This choice is related with the idea of stratifying the space of covariates and measuring how good is the estimator in the different strata under the distribution Q.

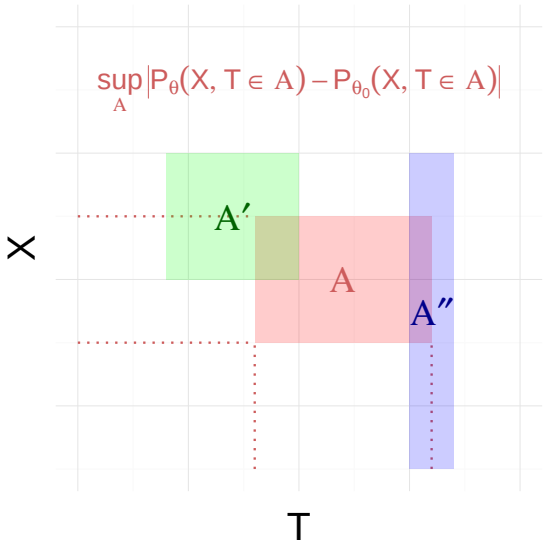

Figure 6.1: The proposed metric compares survival functions over the class $\mathcal{A}$ of rectangles in $\mathbb{R}_+ \times [0,1]^d$.

In order to obtain a proper metric, it is important to check the value of $d_B(\theta, \theta_0)$ over a large class of subsets B on $[0,1]^d$. Let $\mathcal{A}$ be the class of rectangles in $\mathbb{R}_+ \times [0,1]^d$, then we define the metric to be

$$d(\theta, \theta_0) = \sup_{A \in \mathcal{A}} |\mu_\theta(A) - \mu_{\theta_0}(A)|. \tag{6.4.2}$$

On the other hand, under the assumption of fixed covariates ([**NRD**]), we replace the distribution Q by the empirical distribution $Q_n = \sum_{i=1}^n \delta_{\boldsymbol{x}_i}$. In this way, for a set $A = A_1 \times A_2 \subseteq \mathbb{R}_+ \times [0,1]^d$, the loss function becomes

$$\begin{aligned} d_n(\theta, \theta_0) &= \sup_{A \in \mathcal{A}} \left| \int_{A_2} (\tilde{\mu}_\theta^{\boldsymbol{x}}(A_1) - \tilde{\mu}_{\theta_0}^{\boldsymbol{x}}(A_1)) Q_n(d\boldsymbol{x}) \right| \\ &= \sup_{A \in \mathcal{A}} \left| \sum_{i=1}^n \frac{\delta_{\boldsymbol{x}_i}(A_2)}{n} (\tilde{\mu}_\theta^{\boldsymbol{x}_i}(A_1) - \tilde{\mu}_{\theta_0}^{\boldsymbol{x}_i}(A_1)) \right|. \end{aligned} \tag{6.4.3}$$

6.4.3 Non-stationary Gaussian processes

The main difficulty of this work is that the index space $\mathcal{T}$ of the Gaussian processes $(\eta_j(t))_{t \in \mathcal{T}}$ defined in equation (6.2.1) is $\mathbb{R}_+$. Up to our knowledge, most of the results proving posterior consistency under a Gaussian process prior use the fact that these processes are defined in a L^∞ space of functions, condition which is not satisfied in our case. A typical scenario is to put a Gaussian prior on the space of continuous functions on a closed interval, as seen [10], [25] and [56]. The reason for this, is that the results need to compute small ball probabilities. These have been vastly studied in the setting of Gaussian measures on L^∞, providing good tools for finding appropriate bounds for these probabilities.

To bypass the problem that our processes are not in $L^\infty(\mathbb{R}_+)$ (it is easy to prove that a stationary Gaussian process in $\mathbb{R}_+$ is unbounded), we consider a simple modification of the original processes, this modification belongs to the space of bounded functions allowing us to use machinery for bounded Gaussian processes, in particular the theory of RKHS to deal with small ball probabilities

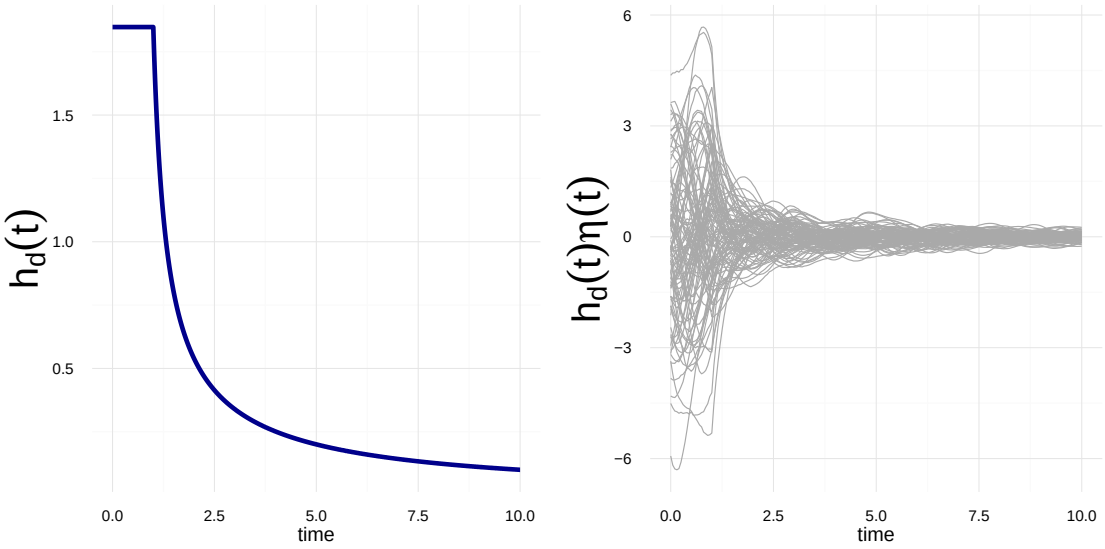

Figure 6.2: Plot of $h_d(t)$ with $d = 0$.

which are fundamental in our analysis. After that, we translate the results of the modified processes to the original ones.

Define the positive and continuous piecewise function

$$h_d(t) = \begin{cases} \frac{d+1}{1+\log(1-e^{-1})} & t \leq 1 \\ \frac{d+1}{t+\log(1-e^{-t})} & t > 1 \end{cases}, \tag{6.4.4}$$

where d represents the dimension of the covariate $\boldsymbol{x}$. Then, we define the non-stationary Gaussian processes $(\hat{\eta}_j(t))_{t\geq 0}$, as function of the original Gaussian processes $(\eta_j(t))_{t\geq 0}$ defined in equation (6.2.1) as

$$\hat{\eta}_j(t) = h_d(t)\eta_j(t), \tag{6.4.5}$$

for all $t \geq 0$ and all $j \in \{0, \ldots, d\}$. We prove that the new Gaussian processes $(\hat{\eta}_j(t))_{t\geq 0}$ lie in the separable Banach space of continuous functions tending to zero, endowed with the uniform norm.

Lemma 6.4.1. *Let $\mathcal{C}_0$ denote the space of all the continuous functions from $\mathbb{R}_+$ to $\mathbb{R}$ tending 0 as t approaches infinity. Then the Banach space $\mathbb{B} = (\mathcal{C}_0, \|\|_\infty)$ of*

functions in $\mathcal{C}_0$ endowed with the uniform norm is separable.

Lemma 6.4.2. *The set of sample paths of $\hat{\eta}_j$ is subset of the separable Banach space $\mathbb{B} = (\mathcal{C}_0, \|\cdot\|_\infty)$ for every $j \in \{0, \dots, d\}$ with probability 1.*

We defer the proofs of Lemma 6.4.1 and Lemma 6.4.2 for the appendix 6.A of this chapter.

Lemma 6.4.2 is equivalent to say that the sample paths of the original process $(\eta_j(t))_{\mathbb{R}_+}$ belongs to the space $(\mathcal{C}, \|\cdot\|_{h_d,\infty})$, where $\mathcal{C}$ is the set of continuous functions in $\mathbb{R}_+$ such that for $f \in \mathcal{C}$, $fh_d(t) \overset{t\to\infty}{\to} 0$, and the norm is given by

$$\|f\|_{h_d,\infty} = \sup_{t\in\mathbb{R}_+} |h_d(t)f(t)|.$$

From Lemma 6.4.1 we obtain the space $(\mathcal{C}, \|\cdot\|_{h_d,\infty})$ is a separable Banach space. Since the sample paths of $(\eta_j(t))_{t\in\mathbb{R}_+}$ belong to this space with probability 1, it induces a Gaussian measure on this space, i.e., a measure over sample paths such that marginally they are normally distributed.

RKHS and support

Here we analyse the support of the Gaussian processes $(\hat{\eta}_j(t))_{t\in\mathbb{R}_+}$ in the space $(\mathcal{C}_0, \|\cdot\|_\infty)$ or equivalently, the support of original Gaussian processes $(\eta_j(t))_{t\in\mathbb{R}_+}$ in $(\mathcal{C}, \|\cdot\|_{h_d,\infty})$. We recall that the support of a probability measure μ on a metric space (X, ρ) is given by the set of points $x \in X$ such that for all $\varepsilon > 0$ it holds

$$\mu(B_{x,\varepsilon}) > 0,$$

where $y \in B_{x,\varepsilon}$ if and only if $\rho(x, y) < \varepsilon$. This support is usually easy to compute for real-valued random variables (the measure associated with the random variable), but it is not clear how to obtain the support for stochastic processes. Nevertheless, for the case of Gaussian process the theory of reproducing kernel Hilbert spaces helps us to describe such support.

The reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ attached to a Gaussian process $(\eta(t))_{t\in R_+}$ with mean zero is the completion of the linear space of functions

$$f(t) = \mathbf{E}(\eta(t)H),$$

where $H = \sum_{i=1}^{n} a_i\eta(s_i)$, with $a_i \in \mathbb{R}$ and $s_i \in \mathbb{R}_+$, with respect the inner product

$$\langle \mathbf{E}(\eta(\cdot)H_1), \mathbf{E}(\eta(\cdot)H_2)\rangle_{\mathbb{H}} = \mathbf{E}(H_1H_2).$$

Several properties of Gaussian proesses can be obtained from the RKHS $\mathbb{H}$ and we refer [58] for a detailed account of the topic. The following result relates the RKHS $\mathbb{H}$ with the support of the Gaussian process as measure on a separable Banach space.

Lemma 6.4.3. *([58, Lemma 5.1]) The support of a mean-zero Gaussian process in a separable Banach space $\mathbb{B}$ is the closure of its RKHS in $\mathbb{B}$.*

Example Using the theorem above we compute the support of the Gaussian process Gaussian processes $(\hat{\eta}_j(t))_{t\in\mathbb{R}_+}$ in the space $(\mathcal{C}_0, \|\cdot\|_\infty)$, and as corollary, the support of original Gaussian processes $(\eta_j(t))_{t\in\mathbb{R}_+}$ in $(\mathcal{C}, \|\cdot\|_{h_d,\infty})$ when the original kernel is the square exponential kernel $K(s,t) = \exp(-(s-t)^2)$. For that we need to find the closure of the reproducing kernel Hilbert space in the Banach space. Observe that the kernel of $(\hat{\eta}_j(t))_{t\in\mathbb{R}_+}$ is given by

$$\hat{K}(s,t) = h(s)h(t)K(s,t),$$

and thus the family of functions of the form

$$\begin{aligned} f(t) &= \mathbf{E}\left(h(t)\eta(t)\sum_{i=1}^{n} a_i h(s_i)\eta(s_i)\right) \\ &= h(t)\sum_{i=1}^{n} a_i h(s_i)K(t,s_i) = h(t)\sum_{i=1}^{n} b_i K(t,s_i) \end{aligned} \tag{6.4.6}$$

with $a_i, b_i \in \mathbb{R}$ and $s_i \in \mathbb{R}_+$. To find the support, instead of finding the RKHS and then its closure in the Banach space, we closure the family of functions given in equation (6.4.6) directly in the Banach space. We will see that the closure gives the full space $\mathcal{C}_0$.

Let $g \in (\mathcal{C}_0, \|\cdot\|_\infty)$, we will approximate g by functions of the form of 6.4.6. By definition, for every $\varepsilon \geq 0$, it exists $T \geq 0$ such that $|g(t)| \leq \varepsilon/4$ for all $t \geq T$. It is well-known that any continuous function f in a compact space $[a, b]$ can be arbitrarily approximated with respect to the supremum norm by functions of the form

$$\sum_{i=1}^{n} a_i K(t, s_i) = \sum_{i=1}^{n} a_i e^{-(t-s_i)^2},$$

with $a_i \in \mathbb{R}$, $s_i \in \mathbb{R}_+$ and $n \in \mathbb{N}$ (see [9]). Let q be a function of the form above such that $\sup_{t\in[0,T]} |g(t)/h(t) - q(t)| \leq \varepsilon/(4h_d(0))$. Note that the natural extension of q to $\mathbb{R}_+$ is such that it is decreasing from T onwards and qh belongs to the family of equation 6.4.6. Then

$$\begin{aligned}
\sup_{t\in\mathbb{R}_+} |g(t) - h_d(t)q(t)| &\leq \varepsilon/4 + \sup_{t\geq T} |g(t) - h(t)q(t)| \\
&\leq \varepsilon/4 + \sup_{t\geq T} |g(t)| + \sup_{t\geq T} |h_d(t)q(t)| \\
&\leq \varepsilon/4 + \varepsilon/4 + h(T)q(T) \\
&\leq \varepsilon/2 + g(T) + \varepsilon h_d(T)/(4h(0)) \leq \varepsilon.
\end{aligned}$$

We conclude that a subset of the RKHS is enough to approximate all the functions of $(\mathcal{C}_0, \|\cdot\|_\infty)$ and by Lemma 6.4.3 the support of the Gaussian processes $(\hat{\eta}_j(t))_{t\in\mathbb{R}_+}$ is $(\mathcal{C}_0, \|\cdot\|_\infty)$. As corollary, we obtain that the support of $(\eta_j(t))_{t\in\mathbb{R}_+}$ in $(\mathcal{C}_0, \|\cdot\|_\infty)$ are all the continuous functions f such that $fh_d \in \mathcal{C}_0$.

Remark 6.4.4. The exercise above can be repeated with almost all standard kernels as they approximate continuous function in compact sets.

6.4.4 Assumptions

In this section, we make assumptions over the prior Π and the parameter space Θ, which allows us to prove posterior consistency. Assumptions (A1), (A2) and (A4) will be common to both covariates designs, while the third assumption will be design-specific. Indeed, (A3) will be used for the random design case ([**RD**]), and (A3') will be used for the fixed design case ([**NRD**]).

(A1) The stationary covariance function $K_j(t)$ is such that $(K_j(0)-K_j(2^{-n}))^{-1} \geq n^6$ for all $j \in \{0,\ldots,d\}$.

(A2) In addition, ν assigns positive probability to every neighbourhood of the true parameter Ω_0, i.e., for every $\delta > 0$, $\nu\left(\left|\frac{\Omega}{\Omega_0} - 1\right| < \delta\right) > 0$.

(A3) Finite first moment under the distribution of the true parameter, i.e.,

$$\mathbf{E}_{\theta_0}(T) = \mathbf{E}(\mathbf{E}_{\theta_0}(T)|X = \boldsymbol{x})) = \int_{[0,1]^d}\int_0^\infty t f_{\boldsymbol{x}}(t,\theta_0)dtQ(d\boldsymbol{x}) < M < \infty.$$

(A3') We assume that given $\delta > 0$, there exits $M \in [0,\infty)$ such that

$$\mathbf{E}_{\theta_0}(T^2\mathbf{1}\{T > M\}|\boldsymbol{x}) \;=\; \int_M^\infty t^2 f_{\boldsymbol{x}}(t,\theta_0)dt < \delta$$

for all $\boldsymbol{x} \in [0,1]^d$, i.e., the class of random variables T^2 indexed in the covariate $\boldsymbol{x} \in [0,1]^d$ is uniformly integrable under the true parameter θ_0.

(A4) The true parameters $\eta_{j,0}$ belong to the support of $(\eta_j(t))_{t\geq 0}$ in the Banach space $(\mathcal{C}, \|\cdot\|_{h_d,\infty})$, for all $j \in \{1,\ldots,d\}$.

As an example of (A4), recall that for the square exponential function this support is the set of continuous functions f such that $fh_d \in \mathcal{C}_0$.

6.5 Main result

6.5.1 Random design case

We first state the posterior consistency results under the loss function given by equation (6.4.2), for the case where the covariates are drawn from the probability distribution Q. Observe that there is not an unique interpretation for this result under this assumption. For instance, a first approach is to consider the covariates X as data generated from our model under the distribution Q. In this way, the first posterior consistency result we can think of, is in terms of the joint probability distribution $\mathbf{P}_{\theta_0}$ of the pairs $((T_i, X_i))_{i\geq 1}$. This is stated in the following theorem.

Theorem 6.5.1. *Let $((T_i, X_i))_{i=1}^{\infty}$ be a sequence of independent random variables, where $T_i \in \mathbb{R}_+$ denotes a random survival time, and $X_i \in [0,1]^d$ its associated covariate. Suppose that the covariates arise according a random design with probability distribution Q on $[0,1]^d$, and that conditioned on X_i, the random failure times T_i are generated according the model described in equation (6.2.1). Let $\mathbf{P}_{\theta_0}$ denote the joint probability measure of $((T_i, X_i))_{i=1}^{\infty}$ under the true parameter θ_0. Assume that assumptions (A1), (A2), (A3) and (A4) are satisfied, then for every $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pi\left(d(\theta,\theta_0) > \epsilon | (T_1, X_1), \ldots, (T_n, X_n)\right) = 0, \quad \mathbf{P}_{\theta_0}\ a.s,$$

with $d(\theta,\theta_0)$ defined in equation (6.4.2).

Another approach is by taking covariates that effectively follow a distribution Q but are not considered to be generated by the model. To make an explicit distinction between this case and the former, we denote the posterior distribution by a different letter Π_2. Notice that the likelihood under this approach is different than the one in the first case, as we consider the conditional distribution of the times T given the covariates X, but under the reasonable assumption that

the distribution Q is unrelated to the parameter θ, we recover exactly the same posterior distribution. A direct consequence of this, is that Theorem 6.5.1 can be used to deduce the following corollary.

Corollary 6.5.2. *By the definition of conditional expectation, the latter result can be interpreted as that for every* $\epsilon > 0$,

$$\mathbf{E}\left(\mathbf{P}_{\theta_0}\left(\lim_{n\to\infty}\Pi_2\left(d(\theta,\theta_0) > \epsilon|(T_1,X_1),\ldots,(T_n,X_n)\right) = 0\middle|(X_n)_{n\geq 1}\right)\right) = 1,$$

which implies

$$Q\left(\mathbf{P}_{\theta_0}\left(\lim_{n\to\infty}\Pi_2\left(d(\theta,\theta_0) > \epsilon|(T_1,X_1),\ldots,(T_n,X_n)\right) = 0\middle|(X_n)_{n\geq 1}\right) = 1\right) = 1,$$

i.e., Q almost surely convergence of the joint conditional distribution of the times $(T_i)_{i\geq 1}$ given the covariates $(X_i)_{i\geq 1}$ under the assumption of the true parameter θ_0.

The last approach is to consider a random design in which the distribution Q is an unknown parameter. Following the Bayesian approach, we specify a prior over Q and assume that the distribution Q is unrelated to θ. It follows that the likelihood for θ can be separated out from that of Q. Then, if θ and Q are independent *a priori*, so they are *a posteriori*. Assuming a true parameter Q_0, the independence of θ and Q allows us to exactly replicate the proof of Theorem 6.5.1 for this case, when assumption (A3) holds for Q_0. Notice that in this case we assume the covariates are generated by Q_0.

6.5.2 Non-random design case

We consider the version of the theorem for the non-random design approach.

Theorem 6.5.3. *Let $(T_i)_{i\geq 1}$ be a sequence of independent random variables on $\mathbb{R}_+$ denoting survival lifetimes. Additionally consider the sequence of associated*

fixed covariates $(\boldsymbol{x}_i)_{i\geq 1}$. *Suppose that the distribution of random survival lifetimes* T *for a fixed value of a covariate* $\boldsymbol{x}$ *is generated according the model described in equation* (6.2.1). *Let* $\tilde{\mathbf{P}}_{\theta_0}$ *define the joint probability measure of* $(T_i)_{i\geq 1}$ *under the true parameter* θ_0. *Assume that assumptions (A1), (A2), (A3') and (A4) are satisfied, then for every* $\epsilon > 0$,

$$\lim_{n\to\infty} \Pi\left(d_n(\theta,\theta_0) > \epsilon | (T_1,\boldsymbol{x}_1),\ldots,(T_n,\boldsymbol{x}_n)\right) = 0, \quad \tilde{\mathbf{P}}_{\theta_0}\ a.s,$$

with $d_n(\theta,\theta_0)$ *defined in equation* (6.4.3).

6.6 Overview of proofs

We give an overview of the proofs and lemmas used to prove Theorem 6.5.1 and Theorem 6.5.3. As many of these results will be used by both proofs, we classify them into two general subsections related to conditions (T1) and (T2) of Theorem 6.3.1: Existence of tests and Prior Positivity of neighbourhoods.

It is worth noticing that for the random design approach ([**RD**]), the pairs $((T_i,X_i))_{i\geq 1}$ are i.i.d. samples from the joint distribution of time and covariates and hence Schwartz's Theorem is enough for proving Theorem 6.5.1. In particular, for this design, we exchange condition (T1) of Theorem 6.3.1 by the alternative condition (G1). For the non-random design ([**NRD**]), we need the extended version given in Theorem 6.3.1.

6.6.1 Existence of tests

Given the metrics in equations (6.4.2) and (6.4.3), a natural choice of a test is to consider the largest deviation of the joint empirical distribution μ_n with respect the true joint distribution $\mathbf{P}_{\theta_0}$ over the class of rectangles $\mathcal{A}$ on $\mathbb{R}^+\times[0,1]^d$.

Formally, for $A = A_1 \times A_2 \in \mathcal{A}$, we define the empirical distribution as

$$\mu_n(A) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \delta_A(T_i, X_i). \tag{6.6.1}$$

Then we define the sequence of tests

$$\phi_n((T_1, X_1), \ldots, (T_n, X_n)) \coloneqq \mathbf{1}\left\{\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu_{\theta_0}(A)| > \frac{\epsilon}{4}\right\}, \tag{6.6.2}$$

for the random design case associated with the loss function given in equation (6.4.2) and

$$\tilde{\phi}_n((T_1, \boldsymbol{x}_1), \ldots, (T_n, \boldsymbol{x}_n)) \coloneqq \mathbf{1}\left\{\sup_{A\in\mathcal{A}} \left|\mu_n(A) - \int_{A_2} \tilde{\mu}^{\boldsymbol{x}}_{\theta_0}(A_1) Q_n(d\boldsymbol{x})\right| > \frac{\epsilon}{4}\right\}, \tag{6.6.3}$$

for the fixed design case associated with the loss function given in equation (6.4.3).

We use the bounded difference inequality and results from the Vapnik Chervonenkis (VC) theory for providing the exponentially small bounds for the sequence of probabilities $\mathbf{E}_{\theta_0}(\phi_n)$ and $\tilde{\mathbf{E}}_{\theta_0}(\tilde{\phi}_n)$, with $n \in \mathbb{N}$. The next theorem states the bounded difference inequality as given in [15, Theorem 2.2]. This inequality is a common tool to prove concentration of a random variable around its mean.

Theorem 6.6.1 (The bounded difference inequality)**.** *Consider a set A, and let $g : A^n \to \mathbb{R}$ be a function satisfying the bounded difference assumption, i.e.,*

$$\sup_{\substack{x_1,\ldots,x_n \\ x_i'\in A}} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i, \tag{6.6.4}$$

for $1 \leq i \leq n$. Then, for all $t > 0$,

$$\mathbf{P}(g(X_1, \ldots, X_n) - \mathbf{E}g(X_1, \ldots, X_n) \geq t) \leq e^{-2t^2/\sum_{i=1}^n c_i^2}, \tag{6.6.5}$$

and

$$\mathbf{P}(g(X_1,\ldots,X_n) - \mathbf{E}g(X_1,\ldots,X_n) \leq -t) \leq e^{-2t^2/\sum_{i=1}^n c_i^2}. \tag{6.6.6}$$

For our application we choose

$$g((T_1,X_1),\ldots,(T_n,X_n)) \;=\; \sup_{A\in\mathcal{A}}|\mu_n(A) - \mu_{\theta_0}(A)|,$$

where the dependence on the data $(T_1,X_1),\ldots,(T_n,X_n)$ is only through the empirical function μ_n. Then, the above quantity will change by at most $c_i = 1/n$ when changing the value of T_i. This satisfies the bounded difference assumption and hence we can use the Theorem 6.6.1 to derive the following concentration result

$$\mathbf{P}_{\theta_0}\left(\left|\sup_{A\in\mathcal{A}}|\mu_n(A) - \mu_{\theta_0}(A)| - \mathbf{E}_{\theta_0}\sup_{A\in\mathcal{A}}|\mu_n(A) - \mu_{\theta_0}(A)|\right| \geq \frac{\epsilon}{8}\right) \leq 2e^{-\frac{\epsilon^2}{32}n}. \tag{6.6.7}$$

This quantity is very similar to what we expect to prove. Indeed, it only differs by the nuisance term $\mathbf{E}_{\theta_0}\sup_{A\in\mathcal{A}}|\mu_n(A) - \mu_{\theta_0}(A)|$. It is here where we apply the VC theory to obtain an upper bound for this quantity. In particular, we need to find an upper bound that tends to zero, as n increases, in order to have an useful result. We do this by using the following result which appears in [15, Theorem 3.1].

Theorem 6.6.2.

$$\mathbf{E}\left\{\sup_{A\in\mathcal{A}}|\mu_n(A) - \mu(A)|\right\} \;\leq\; 2\sqrt{\frac{\log(2S_{\mathcal{A}}(n))}{n}}, \tag{6.6.8}$$

where $S_{\mathcal{A}}$ is called the shatter coefficient, and it is defined as

$$S_{\mathcal{A}}(n) \;=\; \max_{x_1,\ldots,x_n\in\mathbb{R}^d}|\{\{x_1,\ldots,x_n\}\cap A; A\in\mathcal{A}\}|, \tag{6.6.9}$$

i.e., $S_{\mathcal{A}}$ denotes the maximal number of different subsets of n points which can be separated by elements of $\mathcal{A}$.

These utility of the proposed bound depends on obtaining an appropriate bound for the shatter coefficient $S_{\mathcal{A}}(n)$ related to the class of sets $\mathcal{A}$. In turn, this result is related to the finiteness of the VC dimension V of the class of sets $\mathcal{A}$. From [15, Chapter 4], we introduce the definition of the VC dimension V as the largest integer n such that

$$S_{\mathcal{A}}(n) \quad = \quad 2^n, \tag{6.6.10}$$

then, for a finite V we can obtain an upper bound for the shattering coefficient $S_{\mathcal{A}(n)}$. We enunciate this result in corollary 6.6.3 as it appears in [15, Corollary 4.1].

Corollary 6.6.3. *Let $\mathcal{A}$ be a class of sets with Vapnik-Chervonenkis dimension $V < \infty$. Then for all n:*

$$S_{\mathcal{A}}(n) \quad \leq \quad (n+1)^V, \tag{6.6.11}$$

and for all $n \geq V$:

$$S_{\mathcal{A}}(n) \quad \leq \quad \left(\frac{ne}{V}\right)^V. \tag{6.6.12}$$

For our particular choice of $\mathcal{A}$, the class of rectangles in $\mathbb{R}_+ \times [0,1]^d$, V is bounded by $2(d+1)$, [15, Lemma 4.1]. Hence, we conclude

$$\mathbf{E}_{\theta_0} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu_{\theta_0}(A)| \quad \leq \quad 2\sqrt{\frac{\log(2(n+1)^{2(d+1)})}{n}}, \tag{6.6.13}$$

which tends to zero as n approaches infinity. This result allows us to get rid of this quantity and to prove the exponentially small bounds for the type I and II

error.

A direct consequence of our proof technique, is that we can modify our results and include a larger class of sets $\mathcal{A}$, as long as this new class has a finite VC dimension. Nevertheless, some families of sets have infinite VC dimension, for example the total variation metric is the particular instance when the family $\mathcal{A}$ of sets are all the measurable sets of $\mathbb{R}_+ \times [0,1]^d$ and this family has infinity VC dimension.

Lemma 6.6.4. *Let $\epsilon > 0$ and $\mathbf{P}_{\theta_0}$ be the true joint probability measure defined by the true parameter θ_0, then*

$$\mathbf{E}_{\theta_0}(\phi_n) \leq C_1 e^{-C_2 \epsilon^2 n} \quad , \quad \sup_{\theta \in \Theta, d(\theta_0, \theta) > \epsilon} \mathbf{E}_{\theta}(1 - \phi_n) \leq C_1 e^{-C_2 \epsilon^2 n},$$

for some positive constants C_1 and C_2.

Lemma 6.6.5. *Let $\epsilon > 0$ and $\tilde{\mathbf{P}}_{\theta_0}$ be the true joint probability measure defined by the true parameter θ_0, then*

$$\tilde{\mathbf{E}}_{\theta_0}(\tilde{\phi}_n) \leq \tilde{C}_1 e^{-\tilde{C}_2 \epsilon^2 n} \quad , \quad \sup_{\theta \in \Theta, d_n(\theta_0, \theta) > \epsilon} \tilde{\mathbf{E}}_{\theta}(1 - \tilde{\phi}_n) \leq \tilde{C}_1 e^{-\tilde{C}_2 \epsilon^2 n},$$

for some positive constants $\tilde{C}_1$ and $\tilde{C}_2$.

We defer the proofs of Lemma 6.6.4 and Lemma 6.6.5 for the appendix 6.A of this chapter.

6.6.2 Prior positivity of neighbourhoods

We proceed to prove the conditions related to Kullback-Leibler neighbourhoods of the true parameter θ_0. To this end, we define a set $B_{\delta,\tau}$, in which conditions (G1) and (T1) will be satisfied for both settings. We finish our result by proving the positivity of such set.

Definition 6.6.6. Let $\delta > 0$ and $\tau > 1$. Define

$$B_{\delta,\tau} := \left\{ \theta : \left| \frac{\Omega}{\Omega_0} - 1 \right| < \delta; \sup_{t\in[0,\tau]} |\eta_j(t) - \eta_{j0}(t)| \leq \frac{\delta}{1+\tau}, \right.$$
$$\left. \inf_{t>\tau} \{\eta_j(t) h_d(t)\} > -1, \forall j \right\}. \tag{6.6.14}$$

Observe this set is a function of δ and τ. We prove that for a given $\epsilon > 0$, we can always find some $\delta(\epsilon) > 0$ and $\tau(\epsilon) > 1$ such that $KL(\theta_0, \theta) < \epsilon$ for all $\theta \in B_{\delta(\epsilon),\tau}$, for all $\tau \geq \tau(\epsilon)$, in the case of condition (G1). The intuition behind this proof is that since the KL is finite (and defined as an integral), from some point onwards, the tail of the Gaussian process becomes irrelevant. We quantify the notion of *irrelevance* by multiplying the Gaussian process by the decreasing function h_d from some time $\tau(\epsilon)$. The same applies for condition (T1).This result is formalised in the following lemmas.

Lemma 6.6.7. *Assume the random design **[RD]** for the covariates and assumption (A3). Let $0 < \epsilon < 1$, then there exits $0 < \delta(\epsilon) < \frac{1}{2}$ sufficiently small and $\tau(\epsilon) > 1$ sufficiently large such that for all $\tau \geq \tau(\epsilon)$ and for all $\theta \in B_{\delta(\epsilon),\tau}$,*

$$\mathbf{E}_{\theta_0} \left(\log \frac{f_X(T, \theta_0)}{f_X(T, \theta)} \right) < \epsilon.$$

We defer the proof of lemma 6.6.7 for the appendix 6.A of this chapter.

Lemma 6.6.8. *Assume the fixed design **[NRD]** for the covariates and assumption (A3'). Let $0 < \epsilon < 1$, then there exits $0 < \delta(\epsilon) < \frac{1}{2}$ sufficiently small and $\tau(\epsilon) > 1$ sufficiently large such that for all $\tau \geq \tau(\epsilon)$ and for all $\theta \in B_{\delta(\epsilon),\tau}$,*

(i) $KL_i(\theta_0, \theta) < \epsilon$ *for all* i

(ii) $\sum_{i=1}^{\infty} \frac{V_i(\theta_0,\theta)}{i^2} < \infty$.

We defer the proof of lemma 6.6.8 for the appendix 6.A of this chapter.

Positivity of $B_{\delta,\tau}$

We proceed to verify the prior positivity of the set $B_{\delta,\tau}$. Since this set is common to both, Lemma 6.6.7 and Lemma 6.6.8, we just give a general proof and use it in the two settings.

Observe that the set $B_{\delta,\tau}$ given in equation (6.6.14) depends in two parameters: a constant $\delta > 0$ and a time $\tau > 1$. Then Lemma 6.6.7 and Lemma 6.6.8 state that given $\epsilon > 0$ (related to a ϵ-KL neighbourhood of θ_0), we can always find $\delta(\epsilon)$ and $\tau(\epsilon) > 1$ such that there exist a collection of sets $B_{\delta(\epsilon),\tau}$, with $\tau > \tau(\epsilon)$ satisfying conditions (G1) or (T1) (depending in which case we are) of Theorem 6.3.1. The next natural step is to prove whether these sets have positive prior measure.

In this section we prove that given $\delta(\epsilon) > 0$, we can find $\tau^\star > 1$ such that $\Pi(B_{\delta(\epsilon),\tau^\star}) > 0$ for all $\tau > \tau^\star$. In this way, by taking $\tau \geq \max\{\tau(\epsilon), \tau^\star\}$, we are able to pick a set $B_{\delta(\epsilon),\tau}$ (since we have a collection of them), with positive prior probability and satisfying lemmas 6.6.7 or 6.6.8 (depending in which setting we are)

By independence of the Gaussian processes and the parameter Ω,

$$\begin{aligned}\Pi(B_{\delta,\tau}) &= \prod_{j=0}^{d} \Pi\left(\sup_{t\in[0,\tau]} |\eta_j(t) - \eta_{j,0}(t)| \leq \frac{\delta}{1+\tau}, \inf_{t>\tau}\{\eta_j(t)h_d(t)\} > -1\right) \\ &\times \nu\left(\left|\frac{\Omega}{\Omega_0} - 1\right| < \delta\right) \end{aligned} \tag{6.6.15}$$

with $\nu\left(\left|\frac{\Omega}{\Omega_0} - 1\right| < \delta\right) > 0$ by assumption (A3). Hence, we just need to check the positivity of the sets

$$\mathcal{G}^j_{\delta,\tau} = \left\{\sup_{t\in[0,\tau]} |\eta_j(t) - \eta_{j,0}(t)| \leq \frac{\delta}{1+\tau}, \inf_{t>\tau}\{\eta_j(t)h_d(t)\} > -1\right\}, \tag{6.6.16}$$

for every $j \in \{0, \ldots, d\}$.

The proposed proof uses the modified Gaussian processes proposed in equation (6.4.5) to lower bound the probabilities of these sets. Indeed, we consider the sets

$$\hat{\mathcal{G}}^j_{\delta,\tau} = \left\{ \sup_{t\in[0,\tau]} |\hat{\eta}_j(t) - \hat{\eta}_{j,0}(t)| \leq \frac{\delta h_d(\tau)}{1+\tau}, \sup_{t>\tau} |\hat{\eta}_j(t) - \hat{\eta}_{j,0}(t)| \leq \frac{1}{3} \right\} \quad (6.6.17)$$

and prove that there exits some $\tau_{S_j} > 1$ such that for all $\tau > \tau_{S_j}$, it holds

$$\hat{\mathcal{G}}^j_{\delta,\tau} \subseteq \mathcal{G}^j_{\delta,\tau}. \quad (6.6.18)$$

Therefore, we transform the problem above to compute probabilities for the non-stationary Gaussian processes $(\hat{\eta}_j(t))_{t\geq 0}$. This is easier as these stochastic processes belong to a $L^\infty(\mathbb{R}_+)$ space of functions.

By using the structure of the sets $\mathcal{G}^j_{\delta,\tau}$ and $\hat{\mathcal{G}}^j_{\delta,\tau}$, we break the proof in two parts: for $t \in [0,\tau]$ and $t > \tau$. For the first part, by considering assumption (A4), i.e $\eta_{j,0} = \frac{\hat{\eta}_{j,0}}{h_d}$, it holds

$$\begin{aligned} \left\{ \sup_{t\in[0,\tau]} |\hat{\eta}_j(t) - \hat{\eta}_{j,0}(t)| \leq \frac{\delta h_d(\tau)}{1+\tau} \right\} &= \left\{ \sup_{t\in[0,\tau]} |\eta_j(t) - \eta_{j,0}(t)| \frac{h_d(t)}{h_d(\tau)} \leq \frac{\delta}{1+\tau} \right\} \\ &\subseteq \left\{ \sup_{t\in[0,\tau]} |\eta_j(t) - \eta_{j,0}(t)| \leq \frac{\delta}{1+\tau} \right\} \quad (6.6.19) \end{aligned}$$

since $\frac{h_d(t)}{h_d(\tau)} \geq 1$ for all $t \in [0,\tau]$ (h_d is a positive and decreasing function).

We proceed to prove the second part, i.e., when $t > \tau$. By Lemma 6.4.2, the processes $(\hat{\eta}_j(t))_{t\geq 0}$ can be seen as probability measures on the separable Banach space $\mathbb{B} = (\mathcal{C}_0, \|\|_\infty)$. Then, for a function $\hat{\eta}_{j,0}$ in the support of $\hat{\eta}_j$, there exits $\tau_{S_j} > 1$ such that

$$\hat{\eta}_{j,0}(t) \geq -\frac{1}{2}, \qquad \forall t \geq \tau_{S_j}.$$

Then, for $\tau > \tau_{S_j}$

$$\begin{aligned}\left\{\sup_{t>\tau}|\hat{\eta}_j(t) - \hat{\eta}_{j,0}(t)| \leq \frac{1}{3}\right\} &\subseteq \left\{\inf_{t>\tau}\{\hat{\eta}_j(t) - \hat{\eta}_{j,0}\} > -\frac{1}{3}\right\} \\ &\subseteq \left\{\inf_{t>\tau}\{\hat{\eta}_j(t)\} > -\frac{1}{3} - \frac{1}{2}\right\} \\ &\subseteq \left\{\inf_{t>\tau}\{\eta_j(t)h_d(t)\} > -1\right\}. \qquad (6.6.20)\end{aligned}$$

By putting together equation (6.6.19) and equation (6.6.20) we conclude the result given in equation (6.6.18). Observe that if we choose $\tau > \tau_S$, with $\tau_S = \max\{\tau_{S_0}, \ldots, \tau_{S_d}\}$, the previous result of equation (6.6.18) happens for all j at the same time.

We continue by computing the probabilities of the sets

$$\hat{\mathcal{G}}^j_{\delta,\tau} = \left\{\sup_{t\in[0,\tau]}|\hat{\eta}_j(t) - \hat{\eta}_{j,0}(t)| \leq \frac{\delta h_d(\tau)}{1+\tau}, \sup_{t>\tau}|\hat{\eta}_j(t) - \hat{\eta}_{j,0}(t)| \leq \frac{1}{3}\right\}$$

In order to do this, we define a centred version of these events and prove that given $\delta > 0$, there exits a $\tau_{C_j}(\delta) > 1$ such that the centred version has positive prior probability.

Lemma 6.6.9. *Define the centred event*

$$C^j_{\delta,\tau} = \left\{\sup_{t\in[0,\tau]}|\hat{\eta}_j(t)| \leq \frac{\delta h_d(\tau)}{2(1+\tau)}, \sup_{t>\tau}|\hat{\eta}_j(t)| \leq \frac{1}{6}\right\},$$

then there exits $\tau_C > 1$ such that

$$\Pi(C^j_{\delta,\tau}) > 0.$$

for all $\tau \geq \tau_C$ and all $j \in \{0, \ldots, d\}$.

We defer the proof of lemma 6.6.9 for the appendix 6.A of this chapter.

Theorem 6.6.10 (Positivity of the set B)**.** *Consider $\hat{\mathcal{G}}^j_{\delta,\tau}$ as defined in equation (6.6.17), then there exists $\tau_P > \tau_C$ such that*

$$\Pi\left(\hat{\mathcal{G}}^j_{\delta,\tau}\right) \;>\; 0.$$

for all $\tau > \tau_P$ for all j.

Proof of Theorem 6.6.10. By Lemma 6.4.2, the processes $(\hat{\eta}_j(t))_{t\geq 0}$ can be seen as probability measures on the separable Banach space $\mathbb{B} = (\mathcal{C}_0, \|\|_\infty)$. Hence, by Lemma 5.1 of Van der Vaart and van Zanten (2008) [58], the support of $(\hat{\eta}_j)_{t\geq 0}$ is equal to the closure of the reproducing kernel Hilbert space $\mathbb{H}_j \subset \mathbb{B}$ in the Banach space $\mathbb{B}$.

Let $\delta_B = \frac{\delta h_d(\tau)}{2(1+\tau)}$ and take $\tau_B > 0$ such that $0 < \delta_B < \frac{1}{6}$ for all $\tau > \tau_B$. Notice this is possible as $h_d(t)$ is a strictly decreasing for $t > 1$. Consider $g_j \in \mathbb{H}_j$ with $\|g_j - \hat{\eta}_{0,j}\|_\infty \leq \delta_B$, by triangular inequality,

$$\|\hat{\eta}_j - \hat{\eta}_{j,0}\|_\infty \;\leq\; \|\hat{\eta}_j - g_j\|_\infty + \delta_B, \tag{6.6.21}$$

hence

$$\begin{aligned}
\Pi\left(\hat{\mathcal{G}}^j_{\delta,\tau}\right) &= \Pi\left(\sup_{t\in[0,\tau]} |\hat{\eta}_j - \hat{\eta}_{j,0}|(t) \leq \frac{\delta h_d(\tau)}{1+\tau}, \sup_{t>\tau} |\hat{\eta}_j - \hat{\eta}_{j,0}|(t) \leq \frac{1}{3}\right) \\
&\geq \Pi\left(\sup_{t\in[0,\tau]} |\hat{\eta}_j - g_j|(t) \leq \frac{\delta h_d(\tau)}{2(1+\tau)}, \sup_{t>\tau} |\hat{\eta}_j - g_j|(t) \leq \frac{1}{3} - \delta_B\right) \\
&\geq e^{-\frac{1}{2}\|g_j\|^2_{\mathbb{H}_j}} \Pi\left(\sup_{t\in[0,\tau]} |\hat{\eta}_j|(t) \leq \frac{\delta h_d(\tau)}{2(1+\tau)}, \sup_{t>\tau} |\hat{\eta}_j|(t) \leq \frac{1}{6}\right),
\end{aligned}$$

where last inequality comes from the result of Kuelbs et al. [38],[58, Lemma 5.2], and from using the fact that $\delta_B \leq \frac{1}{6}$. On the other hand, by Lemma 6.6.9, there

exits $\tau_C > 1$ such that for all $\tau \geq \tau_C$,

$$\Pi\left(\sup_{t\in[0,\tau]} |\hat{\eta}_j|(t) \leq \frac{\delta h_d(\tau)}{2(1+\tau)}, \sup_{t>\tau} |\hat{\eta}_j|(t) \leq \frac{1}{6}\right) > 0, \tag{6.6.22}$$

for all $j \in \{0, \ldots, d\}$. Finally, we conclude

$$\Pi(\mathcal{G}^j_{\delta,\tau}) > 0.$$

for all $\tau > \tau_P = \max\{\tau_B, \tau_C\}$ and for all $j \in \{0, \ldots, d\}$. □

Given $\epsilon > 0$, by Lemma 6.6.7 (or alternatively Lemma 6.6.8 if we are in the fixed covariates design), there exits $\delta(\epsilon) > 0$ and $\tau(\epsilon) > 1$ such that for all $\tau > \tau(\epsilon)$, the set $B_{\delta(\epsilon),\tau}$ is contained in a Kullback-Leibler neighbourhood of θ_0 of radius ϵ. On the other hand, by equation (6.6.18) and Lemma 6.6.10, for the same $\delta(\epsilon) > 0$, there exists τ_S and τ_P such that for all $\tau > \max\{\tau_S, \tau_P\}$, $\Pi(B_{\delta(\epsilon),\tau}) > 0$. Therefore, we conclude the proof of condition (G1) (or (T1) respectively) by taking any set $B_{\delta(\epsilon),\tau}$ with $\tau > \{\tau_S, \tau_P, \tau(\epsilon)\} > 1$.

Appendix

6.A Main results

6.A.1 Proof of Lemma 6.4.1

Proof. Define the set

$$D_0^n = \left\{ g : \mathbb{R}^+ \to \mathbb{R} : g(t) = \left\{ \begin{array}{cc} g(t) & t \in [0,n] \\ \frac{g(n)}{t+1-n} & t \geq n \end{array} \right\} \right.,$$

where $g(x) \in \mathcal{C}[0,n]$, the set of continuous functions on the compact $[0,n]$. Moreover, define D_0 as the countable union of these sets,

$$D_0 \quad = \quad \bigcup_{n \geq 1} D_0^n.$$

Take $f \in \mathcal{C}_0$. Then by definition, given $\delta > 0$ there exits $T > 0$ such that for all $t \geq T$,

$$\sup_{t \geq T} |f(t)| \leq \delta.$$

Let $\epsilon > 0$, and define,

$$g_N(t) = \left\{ \begin{array}{cc} f(t) & t \in [0,N] \\ \frac{f(N)}{t+1-N} & x \geq N \end{array} \right. ,$$

where N is any natural number such that for $N > T^\star$ and $T^\star > 0$ we have that $\sup_{t \geq T^\star} |f(t)| \leq \frac{\epsilon}{3}$. Notice that by construction g_N belongs to D_0, and

$$\begin{aligned} \sup_{t \in \mathbb{R}^+} |f(t) - g_N(t)| &= \sup_{t \geq N} |f(t) - g_N(t)| \\ &\leq \sup_{t \geq N} |f(t)| + |g_N(t)| \leq \frac{2}{3}\epsilon. \end{aligned}$$

In order to complete the argument, we need to prove that D_0^n is separable. This fact along with the fact that a countable union of countable sets is countable completes the proof.

It is well known that the space $\mathcal{C}[0, n]$ endowed with the uniform norm is separable. Then we have a countable collections of functions $\{h_m^n\}_{m=1}^\infty$ that approximate any element on $\mathcal{C}[0, n]$. Consider the sequence $\{h_{0,m}^n\}_{m=1}^\infty$ defined as,

$$h_{0,m}^n = \begin{cases} h_m^n(t) & t \in [0, n] \\ \frac{h_m^n(n)}{t+1-n} & t \geq n \end{cases}.$$

By construction, for any $g_n \in D_0^n$ and for all $\epsilon > 0$ there exists $h_{0,m}^n$ such that,

$$\sup_{t \in [0,n]} |g_n - h_{0,m}^n| = \sup_{t \in [0,n]} |g_n - h_m^n| < \epsilon.$$

On the other hand, the latter implies $|g_n(n) - h_{0,m}(n)| < \epsilon$. Hence

$$\sup_{t \geq n} \left| g_n(t) - h_{0,m}^n(t) \right| = \sup_{t \geq n} \left| \frac{g_n(n)}{t+1-n} - \frac{h_m^n(n)}{t+1-n} \right| \leq |g_n(n) - h_{0,m}(n)| < \epsilon.$$

Therefore

$$\sup_{t \in \mathbb{R}^+} |g_n - h_{0,m}^n| < \epsilon,$$

concluding that D_0^n is separable. □

6.A.2 Proof of Lemma 6.4.2

Proof. Consider the collection of Gaussian processes $(\eta_j(t))_{t\geq 0}$ for $j \in \{0, \ldots, d\}$, with zero mean and covariance function denoted by K_j. In particular K_j were chosen such that each of them generates Gaussian processes with continuous sample paths. Hence, since h_d is continuous, each non-stationary Gaussian processes $(\hat{\eta}_j(t))_{t\geq 0}$, defined in equation (6.4.5), has continuous sample paths.

We proceed to prove

$$\Pi\left(\lim_{\tau\to\infty}\sup_{t\geq\tau}|\eta_j(t)h_d(t)| = 0\right) \;=\; 1.$$

for every $j \in \{1, \ldots, d\}$. Since each K_j satisfies assumption (A1) we prove this result for an arbitrary j. By Lemma 6.B.5 and assumption (A1), there exists $\tau^\star > 1$ such that for all $\tau \geq \tau^\star$,

$$\Pi\left(\sup_{t\geq\tau}|\eta_j(t)h_d(t)| \geq 1\right) \;\leq\; p(\tau),$$

where $p(\tau) = p(\tau, d, K_j(0))$ is the function defined in defined in equation (6.B.6) of lemma 6.B.5. Moreover, by Lemma 6.B.5 and fixed $d \in \mathbb{N}$,

$$\lim_{\tau\to\infty} p(\tau) \;=\; 0. \tag{6.A.1}$$

On the other hand, consider an increasing sequence of times $\{\tau_i\}_{i=1}^{\infty}$ in the following way. Take $\tau_1 = \tau > \tau^\star$ and choose τ_i such that, for the event

$$E_i \;=\; \left\{\sup_{t\geq\tau_i}|\eta_j(t)h_d(t)| \geq \frac{1}{i}\right\},$$

it holds

$$\begin{aligned}\Pi(E_i) &\leq p(\tau_i) \\ &\leq \frac{p(\tau)}{i^2}.\end{aligned}$$

Observe this can be done since, for fixed d, the function $p(\tau)$ decreases to zero in τ. Hence

$$\begin{aligned}\sum_{i=1}^{\infty}\Pi(E_i) &\leq \sum_{i=1}^{\infty}\frac{p(\tau)}{i^2} \\ &= \frac{p(\tau)\pi^2}{6} < \infty.\end{aligned}$$

Thus, by the Borel-Cantelli lemma

$$\limsup_{i\to\infty}\sup_{t\geq\tau_i}|\eta_j(t)h_d(t)| \leq \limsup_{i\to\infty}\frac{1}{i} = 0,$$

P-a.s.. Since this proof holds for an arbitrary j, we conclude the result. □

6.A.3 Proof of Lemma 6.6.4 and 6.6.5

Proof. We proceed to prove Lemma 6.6.4 and Lemma 6.6.5. Let $\epsilon > 0$, by a direct application of the bounded difference inequality [15, Theorem 2.2], we obtain

$$\begin{aligned}&\mathbf{P}_{\theta_0}\left(\left|\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu_{\theta_0}(A)| - \mathbf{E}_{\theta_0}\left(\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu_{\theta_0}(A)|\right)\right| > \tfrac{\epsilon}{8}\right) \\ &\leq \\ &2e^{-n\epsilon^2/32}. \qquad (6.A.2)\end{aligned}$$

For the case of Lemma 6.6.5, the same result can be applied, since the function

$$\begin{aligned} g(T_1,\ldots,T_n) &= \sup_{A=A_1\times A_2\in\mathcal{A}} \left|\mu_n(A) - \int_{A_2} \mu^x_{\theta_0}(A_1)Q_n(dx)\right| \\ &= \sup_{A=A_1\times A_2\in\mathcal{A}} \left|\sum_{i=1}^n \frac{\delta_{x_i}(A_2)}{n}\left(\delta_{T_i}(A_1) - \mu^{x_i}_{\theta_0}(A_1)\right)\right|, \end{aligned}$$

changes by at most $1/n$ when changing T_i, for a fixed sequence $(x_i)_{i\geq 1}$ and regardless of what $\mathcal{A}$ is. Therefore, by the bounded difference inequality, and for sets $A = A_1 \times A_2 \in \mathcal{A}$ and

$$\bar{\mu}^{\theta_0}_n(A) = \mu_n(A) - \int_{A_2} \mu^x_\theta(A_1)Q_n(dx), \tag{6.A.3}$$

then

$$\begin{aligned} &\tilde{\mathbf{P}}_{\theta_0}\left(\left|\sup_{A\in\mathcal{A}}\left|\bar{\mu}^{\theta_0}_n(A)\right| - \tilde{\mathbf{E}}_{\theta_0}\left(\sup_{A\in\mathcal{A}}\left|\bar{\mu}^{\theta_0}_n(A)\right|\right)\right| > \tfrac{\epsilon}{8}\right) \\ &\leq \\ &2e^{-n\epsilon^2/32}. \end{aligned} \tag{6.A.4}$$

Going back to Lemma 6.6.4, the result in [15, Theorem 3.1] gives a bound for the expectation in equation (6.A.2), in terms of the VC shatter coefficient $S_{\mathcal{A}}(n)$ defined in equation (6.6.10). Indeed,

$$\mathbf{E}_{\theta_0}\left(\sup_{A\in\mathcal{A}}|\mu_n(A) - \mu_{\theta_0}(A)|\right) \leq 2\sqrt{\frac{\log 2S_{\mathcal{A}}(n)}{n}}.$$

For the case of Lemma 6.6.5, the proof of [15, Theorem 3.1] can be replicated for obtaining the same bound for the expectation in equation (6.A.4) for a fixed sequence of covariates $(x_i)_{i\geq 1}$, then

$$\tilde{\mathbf{E}}_{\theta_0}\left(\sup_{A\in\mathcal{A}}\left|\mu_n(A) - \int_{A_2}\mu^{\boldsymbol{x}}_{\theta_0}(A_1)Q_n(d\boldsymbol{x})\right|\right) \leq 2\sqrt{\frac{\log 2S_{\mathcal{A}}(n)}{n}}.$$

As in both cases we have exactly the same bounds in terms of the shatter coefficient of $S_{\mathcal{A}}(n)$, we just prove Lemma 6.6.4 and argue the proof for Lemma 6.6.5 is exactly the same.

By [15, Lemma 4.1], the VC dimension of the class of subsets $\mathcal{A}$ (rectangles in $\mathbb{R}^+ \times [0,1]^d$) is upper bounded by $2(d+1)$. Furthermore, [15, Corollary 4.1] provides a bound for the VC shatter coefficient of the class $\mathcal{A}$ given by

$$S_{\mathcal{A}}(n) \quad \leq \quad (n+1)^{2(d+1)}.$$

Combining the last results, we get

$$\mathbf{E}_{\theta_0}\left(\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu_{\theta_0}(A)|\right) \quad \leq \quad 2\sqrt{\frac{\log(2(n+1)^{2(d+1)})}{n}},$$

which decreases to 0 as n goes to infinity. From this result and the bounded difference inequality, Theorem 6.6.1, we conclude that there exists $N > 0$ large enough such that for all $n \geq N$ it holds

$$\mathbf{E}_{\theta_0}(\phi_n) = \mathbf{P}_{\theta_0}\left(\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu_{\theta_0}(A)| > \frac{\epsilon}{4}\right) \leq 2e^{-n\frac{\epsilon^2}{32}}.$$

We proceed to prove the exponentially small bound for the type II error. By definition of the supremum there exists a rectangle $A^\star \in \mathcal{A}$ such that

$$|\mu_{\theta_0}(A^\star) - \mu_{\theta_1}(A^\star)| > \frac{2\epsilon}{3},$$

with $\theta_1 \in \{\theta : d(\theta,\theta_0) > \epsilon\}$. On the other hand,

$$\begin{aligned}\mathbf{E}_{\theta_1}(1-\phi_n) &= \mathbf{P}_{\theta_1}\left(\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu_{\theta_0}(A)| \leq \frac{\epsilon}{4}\right)\\ &\leq \mathbf{P}_{\theta_1}\left(|\mu_n(A^\star)-\mu_{\theta_0}(A^\star)| \leq \frac{\epsilon}{4}\right). \qquad (6.A.5)\end{aligned}$$

Define the event,

$$E = \left\{ \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu_{\theta_1}(A)| \leq \frac{\epsilon}{4} \right\},$$

then by the first part of this proof, it holds $\mathbf{P}_{\theta_1}(E^c) \leq 2e^{-n\epsilon^2/32}$. We proceed to compute the probability of equation (6.A.5) using the law of total probability

$$\begin{aligned}
(6.A.5) &= \mathbf{P}_{\theta_1}\left(|\mu_n(A^\star) - \mu_{\theta_0}(A^\star)| \leq \frac{\epsilon}{4} \cap E \right) \\
&\quad + \mathbf{P}_{\theta_1}\left(|\mu_n(A^\star) - \mu_{\theta_0}(A^\star)| \leq \frac{\epsilon}{4} \cap E^c \right) \\
&\leq \mathbf{P}_{\theta_1}\left(|\mu_n(A^\star) - \mu_{\theta_0}(A^\star)| \leq \frac{\epsilon}{4} \cap \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu_{\theta_1}(A)| \leq \frac{\epsilon}{4} \right) \\
&+ \mathbf{P}_{\theta_1}\left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu_{\theta_1}(A)| > \frac{\epsilon}{4} \right) \\
&\leq \mathbf{P}_{\theta_1}\left(|\mu_n(A^\star) - \mu_{\theta_0}(A^\star)| \leq \frac{\epsilon}{4} \cap |\mu_n(A^\star) - \mu_{\theta_1}(A^\star)| \leq \frac{\epsilon}{4} \right) + 2e^{-n\frac{\epsilon^2}{32}}.
\end{aligned} \tag{6.A.6}$$

By the triangle inequality, we have

$$|\mu_{\theta_0}(A^\star) - \mu_{\theta_1}(A^\star)| - |\mu_n(A^\star) - \mu_{\theta_1}(A^\star)| \leq |\mu_n(A^\star) - \mu_{\theta_0}(A^\star)|.$$

Therefore,

$$\begin{gathered}
\left\{ |\mu_n(A^\star) - \mu_{\theta_0}(A^\star)| \leq \tfrac{\epsilon}{4} \cap |\mu_n(A^\star) - \mu_{\theta_1}(A^\star)| \leq \tfrac{\epsilon}{4} \right\} \\
\subseteq \\
\left\{ |\mu_{\theta_0}(A^\star) - \mu_{\theta_1}(A^\star)| - |\mu_n(A^\star) - \mu_{\theta_1}(A^\star)| \leq \tfrac{\epsilon}{4} \right\} \cap \left\{ |\mu_n(A^\star) - \mu_{\theta_1}(A^\star)| \leq \tfrac{\epsilon}{4} \right\} = \emptyset.
\end{gathered}$$

since $|\mu_{\theta_0}(A^\star) - \mu_{\theta_1}(A^\star)| > \frac{2\epsilon}{3}$ and $|\mu_n(A^\star) - \mu_{\theta_1}(A^\star)| \leq \frac{\epsilon}{4}$. By plugging-in this result in equation (6.A.6), it holds

$$(6.A.6) \leq \mathbf{P}_{\theta_1}(\emptyset) + 2e^{-n\frac{\epsilon^2}{32}},$$

Finally,

$$\mathbf{E}_{\theta_1}(1-\phi_n) \;\leq\; 2e^{-n\frac{\epsilon^2}{32}}.$$

By taking the supremum over the $\{\theta_1 \in \Theta : d(\theta_0,\theta_1) > \epsilon\}$ we conclude the result. □

6.A.4 Proof of Lemma 6.6.7 and 6.6.8

Proof. Define

$$Y_{\boldsymbol{x}}(t) \;=\; \eta_0(t) + \sum_{j=1}^{d} x_j \eta_j(t)$$

and $Y_{\boldsymbol{x}0}(t)$, when replacing θ by θ_0. The function

$$f_{\boldsymbol{x}}(t;\theta) \;=\; \Omega\sigma(Y_{\boldsymbol{x}}(t))e^{-\int_0^t \Omega\sigma(Y_{\boldsymbol{x}}(s))ds}$$

can be interpreted as the conditional distribution of a survival time T given the covariate $X = \boldsymbol{x}$ in the random design. But it also can be interpreted as the distribution of the time T for a fixed valued of the covariate $\boldsymbol{x}$ in the fixed design. We split the proof in two parts, one for the random design and another part considering a fixed design.

♣ **[Random design, condition (G1)]**

Using the towering property of conditional expectation, we have

$$\mathbf{E}_{\theta_0}\left(\log\frac{f_X(T,\theta_0)}{f_X(T,\theta)}\right) \;=\; \int_{[0,1]^d} \mathbf{E}_{\theta_0}\left(\log\frac{f_{\boldsymbol{x}}(T,\theta_0)}{f_{\boldsymbol{x}}(T,\theta)}\middle| X=\boldsymbol{x}\right) Q(d\boldsymbol{x}).$$

Moreover, by definition

$$\begin{aligned}\mathbf{E}_{\theta_0}\left(\log\frac{f_{\boldsymbol{x}}(T,\theta_0)}{f_{\boldsymbol{x}}(T,\theta)}\middle|X=\boldsymbol{x}\right) &= \int_0^\tau \log\frac{f_{\boldsymbol{x}}(t,\theta_0)}{f_{\boldsymbol{x}}(t,\theta)}f_{\boldsymbol{x}}(t,\theta_0)dt \\ &+ \int_\tau^\infty \log\frac{f_{\boldsymbol{x}}(t,\theta_0)}{f_{\boldsymbol{x}}(t,\theta)}f_{\boldsymbol{x}}(t,\theta_0)dt. \qquad (6.A.7)\end{aligned}$$

The first integral breaks into three parts,

$$\begin{aligned}\int_0^\tau \log\frac{f_{\boldsymbol{x}}(t,\theta_0)}{f_{\boldsymbol{x}}(t,\theta)}f_{\boldsymbol{x}}(t,\theta_0)dt &\leq \int_0^\tau \log\frac{\Omega_0}{\Omega}f_{\boldsymbol{x}}(t,\theta_0)dt \\ &+ \int_0^\tau \left|\log\frac{\sigma(Y_{\boldsymbol{x}0}(t))}{\sigma(Y_{\boldsymbol{x}}(t))}\right| f_{\boldsymbol{x}}(t,\theta_0)dt \\ &+ \int_0^\tau \left|\int_0^t \Omega\sigma(Y_{\boldsymbol{x}}(s)) - \Omega_0\sigma(Y_{\boldsymbol{x}0}(s))ds\right| f_{\boldsymbol{x}}(t,\theta_0)dt \\ &= I_1 + I_2 + I_3. \qquad (6.A.8)\end{aligned}$$

We proceed to bound each of the integrals in equation (6.A.8). For $\theta \in B_{\delta,\tau}$,

$$I_1 = \int_0^\tau \log\frac{\Omega_0}{\Omega}f_{\boldsymbol{x}}(t,\theta_0)dt \leq \int_0^\tau \left(\frac{\Omega_0}{\Omega}-1\right)f_{\boldsymbol{x}}(t,\theta_0)dt \leq \frac{\delta}{1-\delta}.$$

For the second integral, by equation (6.B.2),

$$I_2 = \int_0^\tau \left|\log\frac{\sigma(Y_{\boldsymbol{x}0}(t))}{\sigma(Y_{\boldsymbol{x}}(t))}\right| f_{\boldsymbol{x}}(t,\theta_0)dt \leq \int_0^\tau \frac{d+1}{\tau+1}\delta f_{\boldsymbol{x}}(t,\theta_0)dt < \frac{d+1}{\tau+1}\delta,$$

since $f_{\boldsymbol{x}}(\cdot,\theta_0)$ a p.d.f..

Finally, we break the last integral of equation (6.A.8) into two terms,

$$\begin{aligned} I_3 &= \int_0^\tau \left|\int_0^t \left(\Omega\sigma(Y_{\boldsymbol{x}}(s)) - \Omega_0\sigma(Y_{\boldsymbol{x}}(s)) + \Omega_0\sigma(Y_{\boldsymbol{x}}(s)) - \Omega_0\sigma(Y_{\boldsymbol{x}0}(s))\right)ds\right| f_{\boldsymbol{x}}(t,\theta_0)dt \\ &\leq \int_0^\tau \left|\int_0^t (\Omega\sigma(Y_{\boldsymbol{x}}(s)) - \Omega_0\sigma(Y_{\boldsymbol{x}}(s)))ds\right| f_{\boldsymbol{x}}(t,\theta_0)dt \\ &+ \int_0^\tau \left|\int_0^t (\Omega_0\sigma(Y_{\boldsymbol{x}}(s)) - \Omega_0\sigma(Y_{\boldsymbol{x}0}(s)))ds\right| f_{\boldsymbol{x}}(t,\theta_0)dt \\ &= I_{3_1} + I_{3_2}. \end{aligned}$$

For the first part we use the fact that $|\Omega - \Omega_0| \leq \delta\Omega_0$ and that $\int_0^t \sigma(Y_{\boldsymbol{x}}(t)) \leq t$ (since σ is the sigmoid function and thus upper bounded by 1), hence

$$I_{3,1} \leq \int_0^\tau |\Omega - \Omega_0| t f_{\boldsymbol{x}}(t, \theta_0) dt \leq \delta\Omega_0 \mathbf{E}_{\theta_0}(T|X = \boldsymbol{x}).$$

For the second term, by equation (6.B.1)

$$\begin{aligned} I_{3,2} &\leq \int_0^\tau \Omega_0 \int_0^t |\sigma(Y_{\boldsymbol{x}}(s)) - \sigma(Y_{\boldsymbol{x}0}(s))| \, ds f_{\boldsymbol{x}}(t, \theta_0) dt \\ &\leq \int_0^\tau \Omega_0 \int_0^t \frac{d+1}{\tau+1} \delta ds f_{\boldsymbol{x}}(t, \theta_0) dt \\ &\leq \Omega_0 \frac{\tau(d+1)}{\tau+1} \delta \leq \Omega_0 (d+1)\delta. \end{aligned}$$

Putting everything together,

$$\begin{gathered} \int_0^\tau \log \tfrac{f_{\boldsymbol{x}}(t,\theta_0)}{f_{\boldsymbol{x}}(t,\theta)} f_{\boldsymbol{x}}(t, \theta_0) dt \\ \leq \\ I_1 + I_2 + I_3 \\ \leq \\ \delta \left(\tfrac{1}{1-\delta} + \tfrac{d+1}{\tau+1} + \Omega_0(d+1) + \Omega_0 \mathbf{E}_{\theta_0}(T|X = \boldsymbol{x}) \right). \end{gathered} \tag{6.A.9}$$

For $\delta \in (0, 1/2)$ and $\tau \geq 1$, we have

$$\begin{aligned} \mathbf{E}\left(\int_0^\tau \log \frac{f_{\boldsymbol{x}}(t, \theta_0)}{f_{\boldsymbol{x}}(t, \theta)} f_{\boldsymbol{x}}(t, \theta_0) dt \right) \leq \delta \Bigg(& 2 + \frac{d+1}{2} + \Omega_0(d+1) \\ & + \Omega_0 \int_{\boldsymbol{x} \in [0,1]^d} \mathbf{E}_{\theta_0}(T|X = \boldsymbol{x}) Q(d\boldsymbol{x}) \Bigg). \end{aligned} \tag{6.A.10}$$

Using assumption A3 we have that the latter integral is finite, hence we can choose δ small enough such that the whole term is less than $\epsilon/2$.

Now take the second integral of equation (6.A.7) and break it into two integrals,

$$\begin{aligned}\int_\tau^\infty \log \frac{f_{\boldsymbol{x}}(t,\theta_0)}{f_{\boldsymbol{x}}(t,\theta)} f_{\boldsymbol{x}}(t,\theta_0)dt &\leq \int_\tau^\infty \log f_{\boldsymbol{x}}(t,\theta_0) f_{\boldsymbol{x}}(t,\theta_0)dt \\ &+ \int_\tau^\infty |\log f_{\boldsymbol{x}}(t,\theta)| f_{\boldsymbol{x}}(t,\theta_0)dt \\ &= I_1 + I_2. \end{aligned} \tag{6.A.11}$$

For the first integral, since $\log(x) \leq x - 1$,

$$\begin{aligned} I_1 &\leq \int_\tau^\infty (f_{\boldsymbol{x}}(t,\theta_0) - 1) f_{\boldsymbol{x}}(t,\theta_0)dt \\ &\leq \int_\tau^\infty f_{\boldsymbol{x}}(t,\theta_0)^2 dt. \end{aligned}$$

For the second integral, we have that

$$\begin{aligned} I_2 &= \int_\tau^\infty \left|\log\left(\Omega\sigma(Y_{\boldsymbol{x}}(t)) e^{-\Omega\int_0^t \sigma(Y_{\boldsymbol{x}}(s))ds}\right)\right| f_{\boldsymbol{x}}(t,\theta_0)dt \\ &\leq \int_\tau^\infty |\log\Omega| f_{\boldsymbol{x}}(t,\theta_0)dt + \int_\tau^\infty |\log\sigma(Y_{\boldsymbol{x}}(t))| f_{\boldsymbol{x}}(t,\theta_0)dt \\ &+ \int_\tau^\infty \left|\Omega\int_0^t \sigma(Y_{\boldsymbol{x}}(s))ds\right| f_{\boldsymbol{x}}(t,\theta_0)dt \\ &= I_{2,1} + I_{2,2} + I_{2,3}. \end{aligned}$$

Recall that for $\theta \in B_{\delta,\tau}$ we have that $\Omega_0(1-\delta) \leq \Omega \leq \Omega_0(1+\delta)$ and $\delta \in (0, 1/2)$, hence the first integral is finite and bounded by,

$$I_{2,1} \leq K_0 \mathbf{P}_{\theta_0}(T > \tau | X = \boldsymbol{x}),$$

where K_0 is a constant depending on Ω_0 and δ.

For $\theta \in B_{\delta,\tau}$ we have that $\eta_j(t)h_d(t) > -1$ for all $t \geq \tau$ and all $j \in \{0, \ldots, d\}$.

Then, by the definition of $h_d(t)$, for all $t \geq \tau$

$$Y_{\boldsymbol{x}}(t) \geq \sum_{j=0}^{d} \eta_j(t) \geq -t - \log(1 - e^{-t}). \tag{6.A.12}$$

Furthermore, notice that the inverse $\sigma^{-1}(e^{-t}) = -t - \log(1 - e^{-t})$, which finally implies $0 \geq \log \sigma(Y_{\boldsymbol{x}}(t)) \geq -t$ since the function σ is upper bounded by one and covariates $\boldsymbol{x} \in [0,1]^d$. Using this last argument, we get a bound for

$$I_{2,2} \leq \int_{\tau}^{\infty} t f_{\boldsymbol{x}}(t, \theta_0) dt = \mathbf{E}_{\theta_0}(T 1_{\{T>\tau\}} | X = \boldsymbol{x}).$$

Finally, for the last integral we have

$$I_{2,3} \leq \int_{\tau}^{\infty} \Omega_0(1+\delta) t f_{\boldsymbol{x}}(t, \theta_0) dt = \Omega_0(1+\delta) \mathbf{E}_{\theta_0}(T 1_{\{T>\tau\}} | X = \boldsymbol{x}).$$

By putting everything together, we have

$$\begin{aligned} \int_{\tau}^{\infty} \log \frac{f_{\boldsymbol{x}}(t, \theta_0)}{f_{\boldsymbol{x}}(t, \theta)} f_{\boldsymbol{x}}(t, \theta_0) dt &\leq \int_{\tau}^{\infty} f_{\boldsymbol{x}}(t, \theta_0)^2 dt + K_0 \mathbf{P}_{\theta_0}(T > \tau | X = \boldsymbol{x}) \\ &+ (1 + \Omega_0(1+\delta)) \mathbf{E}_{\theta_0}(T 1_{\{T>\tau\}} | X = \boldsymbol{x}). \end{aligned} \tag{6.A.13}$$

By definition, $f_{\boldsymbol{x}}(t, \theta_0)$ is bounded by Ω_0 for all $t \geq 0$ and $\boldsymbol{x} \in [0,1]^d$. Using this fact and by taking expectation, we get

$$\begin{aligned} \mathbf{E}\left(\int_{\tau}^{\infty} \log \frac{f_{\boldsymbol{x}}(t, \theta_0)}{f_{\boldsymbol{x}}(t, \theta)} f_{\boldsymbol{x}}(t, \theta_0) dt \right) &\leq (\Omega_0 + K_0) \mathbf{E}(\mathbb{P}_{\theta_0}(T > \tau | X = \boldsymbol{x})) \\ &+ (1 + \Omega_0(1+\delta)) \\ &\times \mathbf{E}(\mathbf{E}_{\theta_0}(T 1_{\{T>\tau\}} | X = \boldsymbol{x})). \end{aligned} \tag{6.A.14}$$

We conclude by using that

$$\lim_{n \to \infty} \mathbf{E}(\mathbf{E}_{\theta_0}(T 1_{\{T>n\}} | X = \boldsymbol{x})) = 0 \tag{6.A.15}$$

which is true for every random variable T with finite expectation, i.e., assumption (A2). Also, by Markov inequality notice that

$$\mathbf{E}(\mathbf{P}_{\theta_0}(T > \tau | X = \boldsymbol{x})) \quad \leq \quad \mathbf{E}(\mathbf{E}_{\theta_0}(T 1_{\{T>\tau\}} | X = \boldsymbol{x})).$$

The above allows us to deduce that the integral of the tail, i.e., the integral given in equation (6.A.11), is finite and hence we can choose $\tau \geq 1$ large enough, such that equation (6.A.14) sum less than $\epsilon/2$. The above together with equation (6.A.10) allows us to conclude that

$$\mathbf{E}_{\theta_0}\left(\log \frac{f_X(T,\theta_0)}{f_X(T,\theta)}\right) \quad \leq \quad \epsilon. \tag{6.A.16}$$

♣ **[Fixed design, condition (T1) (i)]**

Recall that in this setting, the covariates are fixed for each data point T_i. Therefore, for each i we have a different covariates $\boldsymbol{x_i}$, but for ease of notation, we just write $\boldsymbol{x}$. Take $f_{\boldsymbol{x}}(t,\theta_0)$ as the distribution of times under the true parameter θ_0 and covariate $\boldsymbol{x}$, then

$$\begin{aligned} KL_i(\theta_0,\theta) &= \mathbf{E}\left(\Upsilon(\theta_0;\theta)\right) \\ &= \int_0^\infty \log \frac{f_{\boldsymbol{x}}(t,\theta_0)}{f_{\boldsymbol{x}}(t,\theta)} f_{\boldsymbol{x}}(t,\theta_0) dt. \end{aligned}$$

which has the same form as equation (6.A.7), hence we replicate exactly the same proof of the random design case, but we justify with assumption (A3') instead of assumption (A3). Indeed, notice that the main assumption we are using to conclude the result in the random design case is that $\mathbf{E}_Q(\mathbf{E}_{\theta_0}(T|X)) < M < \infty$. Since in the fixed design case we are working with expressions of the form $\mathbf{E}_{\theta_0}(T|\boldsymbol{x})$ (without integrating over the covariates), the uniformly integrability required in assumption (A3') is a sufficient condition to replace the latter assumption.

♣ **[Fixed design, condition (T1) (ii)]**

Using the definition of variance,

$$V_i(\theta_0,\theta) = \mathbf{Var}_{\theta_0}(\Upsilon(\theta_0,\theta)) \leq \mathbf{E}_{\theta_0}(\Upsilon^2(\theta_0,\theta))$$

By definition

$$\begin{aligned}\mathbf{E}_{\theta_0}(\Upsilon^2(\theta_0,\theta)) &= \int_0^\tau \log^2 \frac{f_{\boldsymbol{x}}(t,\theta_0)}{f_{\boldsymbol{x}}(t,\theta)} f_{\boldsymbol{x}}(t,\theta_0)dt + \int_\tau^\infty \log^2 \frac{f_{\boldsymbol{x}}(t,\theta_0)}{f_{\boldsymbol{x}}(t,\theta)} f_{\boldsymbol{x}}(t,\theta_0)dt \\ &= I_1 + I_2. \end{aligned} \tag{6.A.17}$$

Consider the first integral of equation (6.A.17). By replacing the p.d.f. we have

$$I_1 = \int_0^\tau \left(\log\frac{\Omega_0}{\Omega} + \left|\log\frac{\sigma(Y_{\boldsymbol{x}0}(t))}{\sigma(Y_{\boldsymbol{x}}(t))}\right| + \left|\int_0^t \Omega\sigma(Y_{\boldsymbol{x}}(s)) - \Omega_0\sigma(Y_{\boldsymbol{x}0}(s))ds\right|\right)^2 f_{\boldsymbol{x}}(t,\theta_0)dt.$$

Notice that for all $a,b,c \in \mathbb{R}$ hold that $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, then

$$\begin{aligned} I_1 &\leq \int_0^\tau 3\left(\log^2\frac{\Omega_0}{\Omega} + \left|\log\frac{\sigma(Y_{\boldsymbol{x}0}(t))}{\sigma(Y_{\boldsymbol{x}}(t))}\right|^2 + \left|\int_0^t \Omega\sigma(Y_{\boldsymbol{x}}(s)) - \Omega_0\sigma(Y_{\boldsymbol{x}0}(s))ds\right|^2\right) f_{\boldsymbol{x}}(t,\theta_0)dt \\ &= \int_0^\tau 3\log^2\frac{\Omega_0}{\Omega} f_{\boldsymbol{x}}(t,\theta_0)dt + \int_0^\tau 3\left|\log\frac{\sigma(Y_{\boldsymbol{x}0}(t))}{\sigma(Y_{\boldsymbol{x}}(t))}\right|^2 f_{\boldsymbol{x}}(t,\theta_0)dt \\ &+ \int_0^\tau 3\left|\int_0^t \Omega\sigma(Y_{\boldsymbol{x}}(s)) - \Omega_0\sigma(Y_{\boldsymbol{x}0}(s))ds\right|^2 f_{\boldsymbol{x}}(t,\theta_0)dt \\ &= I_{1,1} + I_{1,2} + I_{1,3}. \end{aligned} \tag{6.A.18}$$

Take the first integral of equation (6.A.18). Since $\theta \in B_{\delta,\tau}$ and $\delta \in (0,1/2)$ we have

$$I_{1,1} \leq \int_0^\tau 3\left(\frac{\Omega_0}{\Omega} - 1\right)^2 f_{\boldsymbol{x}}(t,\theta_0)dt \leq 12\delta,$$

since for $\Omega \in B_{\delta,\tau}$, $\frac{\Omega_0}{\Omega} - 1 \leq \frac{\delta}{1-\delta} \leq 2\delta$. Next, consider the second integral of

equation (6.A.18). By equation (6.B.2) and since $\tau \geq 1$

$$I_{1,2} \leq \int_0^\tau 3\left(\frac{d+1}{1+\tau}\delta\right)^2 f_{\boldsymbol{x}}(t,\theta_0)dt \leq \frac{3(d+1)^2}{2}\delta.$$

Finally, we take the last integral of equation (6.A.18), using $(a+b)^2 \leq 2(a^2+b^2)$,

$$\begin{aligned} I_{1,3} &= \int_0^\tau 3\left|\int_0^t (\Omega-\Omega_0)\sigma(Y_{\boldsymbol{x}}(s)) + \Omega_0(\sigma(Y_{\boldsymbol{x}}(s)) - \sigma(Y_{\boldsymbol{x}0}(s)))ds\right|^2 f_{\boldsymbol{x}}(t,\theta_0)dt \\ &\leq \int_0^\tau 6\left|\int_0^t \Omega\sigma(Y_{\boldsymbol{x}}(s)) - \Omega_0\sigma(Y_{\boldsymbol{x}}(s))ds\right|^2 f_{\boldsymbol{x}}(t,\theta_0)dt \\ &+ \int_0^\tau 6\left|\int_0^t \Omega_0\sigma(Y_{\boldsymbol{x}}(s)) - \Omega_0\sigma(Y_{\boldsymbol{x}0}(s))ds\right|^2 f_{\boldsymbol{x}}(t,\theta_0)dt \\ &= I_{1,3,1} + I_{1,3,2}. \end{aligned}$$

Given that $\theta \in B_{\delta,\tau}$, we have that $|\Omega - \Omega_0| \leq \Omega_0\delta$. Moreover, recall that $\int_0^t \sigma(Y_{\boldsymbol{x}}(s))ds \leq t$ since $\sigma(s) \leq 1$ for all $s \in \mathbb{R}$, hence

$$I_{1,3,1} \leq K_0\delta\int_0^\tau t^2 f_{\boldsymbol{x}}(t,\theta_0)dt \leq K_0\delta\mathbf{E}_{\theta_0}(T^2|\boldsymbol{x}) \tag{6.A.19}$$

for some constant K_0.

For the second term we have, by equation (6.B.1),

$$\begin{aligned} I_{1,3,2} &\leq \int_0^\tau 6\Omega_0^2\left(\int_0^t |\sigma(Y_{\boldsymbol{x}}(s)) - \sigma(Y_{\boldsymbol{x}0}(s))|\, ds\right)^2 f_{\boldsymbol{x}}(t,\theta_0)dt \\ &\leq \int_0^\tau 6\Omega_0^2\left(\int_0^t \frac{d+1}{1+\tau}\delta ds\right)^2 f_{\boldsymbol{x}}(t,\theta_0)dt \\ &\leq K_0'\delta \end{aligned}$$

for some constant K_0'.

Putting everything together, and using assumption (A3') we have

$$\begin{aligned} I_1 &= \int_0^{\tau} \log^2 \frac{f_{\boldsymbol{x}}(t,\theta_0)}{f_{\boldsymbol{x}}(t,\theta)} f_{\boldsymbol{x}}(t,\theta_0)dt \\ &= \delta\left(12 + \frac{3(d+1)^2}{2} + K_0 \mathbf{E}_{\theta_0}(T^2|\boldsymbol{x}) + K_0'\right) < \infty. \end{aligned}$$

Now, we take the second integral of equation (6.A.17)

$$\begin{aligned} I_2 &\leq \int_{\tau}^{\infty} \left(|\log f_{\boldsymbol{x}}(t,\theta_0)| + |\log f_{\boldsymbol{x}}(t,\theta)|\right)^2 f_{\boldsymbol{x}}(t,\theta_0)dt \\ &\leq \int_{\tau}^{\infty} 2\left|\log f_{\boldsymbol{x}}(t,\theta_0)\right|^2 f_{\boldsymbol{x}}(t,\theta_0)dt + \int_{\tau}^{\infty} 2\left|\log f_{\boldsymbol{x}}(t,\theta)\right|^2 f_{\boldsymbol{x}}(t,\theta_0)dt \\ &= I_{2,1} + I_{2,2}. \end{aligned} \tag{6.A.20}$$

By inspecting the first integral of equation (6.A.20) we have,

$$\begin{aligned} I_{2,1} &\leq \int_{\tau}^{\infty} 2\left|f_{\boldsymbol{x}}(t,\theta_0) - 1\right|^2 f_{\boldsymbol{x}}(t,\theta_0)dt \\ &\leq \int_{\tau}^{\infty} 2 f_{\boldsymbol{x}}(t,\theta_0)^3 dt - \int_{\tau}^{\infty} 4 f_{\boldsymbol{x}}(t,\theta_0)^2 dt + 2 \\ &\leq \int_{\tau}^{\infty} 2 f_{\boldsymbol{x}}(t,\theta_0)^3 dt + 2. \end{aligned}$$

Recall that $f_{\boldsymbol{x}}(t,\theta_0)$ is bounded by Ω_0 for all $\boldsymbol{x} \in [0,1]^d$, hence

$$I_{2,1} \leq \int_{\tau}^{\infty} 2\Omega_0^2 f_{\boldsymbol{x}}(t,\theta_0)dt + 2 \leq 2\Omega_0^2 + 2 < \infty.$$

For the second integral, we have

$$\begin{aligned} I_{2,2} &\leq \int_{\tau}^{\infty} 2|\log \Omega\sigma(Y_x(t))e^{-\int_0^t \Omega\sigma(Y_x(s))ds}|^2 f_x(t,\theta_0)dt \\ &\leq \int_{\tau}^{\infty} 6|\log\Omega|^2 f_x(t,\theta_0)dt + \int_{\tau}^{\infty} 6|\log\sigma(Y_x(t))|^2 f_x(t,\theta_0)dt \\ &+ \int_{\tau}^{\infty} 6\left(\int_0^t \Omega\sigma(Y_x(s))ds\right)^2 f_x(t,\theta_0)dt. \end{aligned} \tag{6.A.21}$$

Recall that for $\theta \in B_{\delta,\tau}$ we have $\Omega_0(1-\delta) \leq \Omega \leq \Omega_0(1+\delta)$, hence the first integral of equation (6.A.21) is finite. Using the same argument as before, equation (6.A.12), for $\theta \in B_{\delta,\tau}$ we have $-t < \log \sigma(Y_{\boldsymbol{x}}(t)) \leq 0$ for all $\boldsymbol{x} \in [0,1]^d$. Hence, the second integral of equation (6.A.21)

$$\int_\tau^\infty 6 \log^2 \sigma(Y_{\boldsymbol{x}}(t)) f_{\boldsymbol{x}}(t,\theta_0) dt \quad \leq \quad 6\mathbb{E}_{\theta_0}(T^2|\boldsymbol{x}). \tag{6.A.22}$$

Finally, we bound the last integral of equation (6.A.21) by,

$$\int_\tau^\infty 6 \left(\int_0^t \Omega \sigma(Y_{\boldsymbol{x}}(s)) ds \right)^2 f_{\boldsymbol{x}}(t,\theta_0) dt \quad \leq \quad 6K_2 \mathbf{E}_{\theta_0}(T^2|\boldsymbol{x}) \tag{6.A.23}$$

where K_2 is some constant that depends on Ω_0. Finally, by the uniformly integrability of T^2 over $\boldsymbol{x} \in [0,1]^d$, assumption $(A3')$

$$V_i(\theta_0, \theta) \quad < \quad \infty \tag{6.A.24}$$

uniformly in i. □

6.A.5 Proof of Lemma 6.6.9

Proof. Consider

$$\Pi(C^j_{\delta,\tau}) \quad = \quad \Pi\left(\sup_{t\in[0,\tau]} |\hat{\eta}_j(t)| \leq \frac{\delta h_d(\tau)}{2(1+\tau)}, \sup_{t>\tau} |\hat{\eta}_j(t)| \leq \frac{1}{6} \right),$$

since both events are symmetric and convex, by the Gaussian correlation inequality, Lemma 6.B.4 enunciated in Appendix B, it holds

$$\Pi(C^j_{\delta,\tau}) \quad \geq \quad \Pi\left(\sup_{t\in[0,\tau]} |\hat{\eta}_j(t)| \leq \frac{\delta h_d(\tau)}{2(1+\tau)} \right) \Pi\left(\sup_{t\geq\tau} |\hat{\eta}_j(t)| \leq \frac{1}{6} \right).$$

We continue by lower bounding each term separately. By Lemma 6.B.6, which

we prove in appendix B, we have that for any $\tau > 1$ and $\delta > 0$, it holds

$$\Pi\left(\sup_{t\in[0,\tau]}|\hat{\eta}_j(t)| \leq \frac{\delta h_d(\tau)}{1+\tau}\right) \;>\; 0,$$

for all $j \in \{0,\ldots,d\}$.

On the other hand

$$\Pi\left(\sup_{t\geq\tau}|\hat{\eta}_j(t)| \leq \frac{1}{6}\right) \;=\; 1 - \Pi\left(\sup_{t\geq\tau}|\hat{\eta}_j(t)| \geq \frac{1}{6}\right).$$

Then, by lemma 6.B.5, there exists $\tau_j^\star > 1$ such that for $\tau > \tau_j^\star$,

$$\Pi\left(\sup_{t\geq\tau}|\hat{\eta}_j(t)| \leq \frac{1}{6}\right) \;\geq\; 1 - p_j(\tau), \tag{6.A.25}$$

where $p_j(\tau) = p_j(\tau, d)$ decreases to zero for fixed d, for each $j \in \{0,\ldots,d\}$. From the latter result, we conclude there exists $\tau_j > \tau_j^\star$ large enough such that for all $\tau \geq \tau_j$

$$\Pi\left(\sup_{t\geq\tau}|\hat{\eta}_j(t)| \leq \frac{1}{6}\right) > 0,$$

for each $j \in \{0,\ldots,d\}$. By considering $\tau_C = \max\{\tau_0,\ldots,\tau_d\}$ we conclude

$$\Pi(C^j_{\delta,\tau}) > 0,$$

for all $\tau \geq \tau_C$ and for all j at the same time. □

6.B Other results on Gaussian processes

6.B.1 Lemma 6.B.1

Lemma 6.B.1. *Let $\delta > 0$, $\tau > 1$ and $\boldsymbol{x} \in [0,1]^d$. Define*

$$B_{\delta,\tau} = \left\{ (\eta_0, \ldots, \eta_d) : \sup_{t\in[0,\tau]} |\eta_j(t) - \eta_{j,0}(t)| \leq \frac{\delta}{1+\tau}, \forall j \in \{0,\ldots,d\} \right\}$$

and

$$Y_{\boldsymbol{x}}(t) = \eta_0(t) + \sum_{j=1}^{d} \boldsymbol{x}_j \eta_j(t)$$

and $Y_{\boldsymbol{x}0}(t)$ as above but replacing θ by θ_0. Let $\sigma(x)$ be the sigmoid function, then for all $\delta > 0$ and $(\eta_0, \ldots, \eta_d) \in B_{\delta,\tau}$,

$$\max_{\boldsymbol{x}\in[0,1]^d} \sup_{t\in[0,\tau]} |\sigma(Y_{\boldsymbol{x}}(t)) - \sigma(Y_{\boldsymbol{x}0}(t))| \leq \frac{d+1}{1+\tau}\delta, \tag{6.B.1}$$

and

$$\max_{\boldsymbol{x}\in[0,1]^d} \sup_{t\in[0,\tau]} |\log\sigma(Y_{\boldsymbol{x}}(t)) - \log\sigma(Y_{\boldsymbol{x}0}(t))| \leq \frac{d+1}{1+\tau}\delta. \tag{6.B.2}$$

Proof of Lemma 6.B.1

Proof. Let $a = Y_{\boldsymbol{x}}(t)$ and $b = Y_{\boldsymbol{x}0}(t)$, by the mean value theorem, there exits c between a and b such that

$$\begin{aligned} |\sigma(a) - \sigma(b)| &= \left| \frac{e^c}{(e^c+1)}(a-b) \right| \\ &\leq |a-b| \\ &= \left| (\eta_0 - \eta_{0,0})(t) + \sum_{j=1}^{d} \boldsymbol{x}_j (\eta_j - \eta_{j,0})(t) \right|. \end{aligned}$$

By taking supremum on both sides of the above equation, we get equation (6.B.1), indeed

$$\begin{aligned}\sup_{t\in[0,\tau]}|\sigma(Y_{\boldsymbol{x}}(t))-\sigma(Y_{\boldsymbol{x}0}(t))| &\leq \sup_{t\in[0,\tau]}\left|(\eta_0-\eta_{0,0})(t)+\sum_{j=1}^{d}\boldsymbol{x}_j(\eta_j-\eta_{j,0})(t)\right| \\ &\leq \sup_{t\in[0,\tau]}|(\eta_0-\eta_{0,0})(t)| \\ &+ \sum_{j=1}^{d}|\boldsymbol{x}_j|\sup_{t\in[0,\tau]}|\eta_j-\eta_{j0}(t)| \\ &\leq \frac{\delta}{1+\tau}+\sum_{j=1}^{d}|\boldsymbol{x}_j|\frac{\delta}{1+\tau}.\end{aligned}$$

The we conclude

$$\max_{\boldsymbol{x}\in[0,1]^d}\sup_{t\in[0,\tau]}|\sigma(Y_{\boldsymbol{x}}(t))-\sigma(Y_{\boldsymbol{x}0}(t))| \leq \frac{d+1}{1+\tau}\delta$$

The proof for equation (6.B.2) follows by the same argument. □

6.B.2 Lemma 6.B.2

Lemma 6.B.2. *Let $\tau>0$ and $\eta(t)$ be a continuous function on the interval $[0,\tau]$, then for $t_k^n=\frac{k\tau}{2^n}$, with $k\in\{0,\dots,2^n\}$ and $n\geq 1$*

$$\sup_{t\in[0,\tau]}|\eta_t|\leq|\eta(0)|+|\eta(\tau)|+\sum_{n=1}^{\infty}\max_{0\leq k\leq 2^n-1}|\eta(t_{k+1}^n)-\eta(t_k^n)|.$$

Proof of Lemma 6.B.2

Proof. Define a partition of the interval $[0,\tau]$ as $A_n=\{t_k^n=\frac{k\tau}{2^n},k=0,\dots,2^n\}$. Let $A_0=\{0,\tau\}$, then

$$\sup_{t\in A_0}|\eta(t)| \leq |\eta(0)|+|\eta(\tau)|.$$

For $A_1 = \{0, \tau/2, \tau\}$, we want to show that

$$\sup_{t\in A_1}|\eta(t)| \leq |\eta(0)| + |\eta(\tau)| + \max\{|\eta(\tau) - \eta(\tau/2)|, |\eta(\tau/2) - \eta(0)|\}.$$

If the supremum is at $\{0, \tau\}$ we are done. Assume the supremum is at $\tau/2$, and let $x \in \{0, \tau\}$

$$\begin{aligned} |\eta(\tau/2)| &= |\eta(\tau/2) - \eta(x) + \eta(x)| \\ &\leq |\eta(\tau/2) - \eta(x)| + |\eta(x)| \\ &\leq \max\{|\eta(\tau) - \eta(\tau/2)|, |\eta(\tau/2) - \eta(0)|\} + |\eta(x)|, \end{aligned}$$

thus the result follows. By induction, assume that

$$\sup_{t\in A_n}|\eta(t)| \leq |\eta(0)| + |\eta(\tau)| + \sum_{i=1}^{n} \max_{0\leq k\leq 2^i-1} |\eta(t_{k+1}^i) - \eta(t_k^i)|,$$

then we need to prove this result for A_{n+1}. Explicitly, we want to show that

$$\sup_{t\in A_{n+1}}|\eta(t)| \leq |\eta(0)| + |\eta(\tau)| + \sum_{i=1}^{n+1} \max_{0\leq k\leq 2^i-1} |\eta(t_{k+1}^i) - \eta(t_k^i)|,$$

i.e.,

$$\begin{aligned} \max\left\{\sup_{t\in A_n}|\eta(t)|, \sup_{t\in A_{n+1}/A_n}|\eta(t)|\right\} &\leq |\eta(0)| + |\eta(\tau)| + \sum_{i=1}^{n} \max_{0\leq k\leq 2^i-1} |\eta(t_{k+1}^i) - \eta(t_k^i)| \\ &+ \max_{0\leq k\leq 2^{n+1}-1} |\eta(t_{k+1}^{n+1}) - \eta(t_k^{n+1})|. \end{aligned}$$

If the supremum is in A_n we are done. Assume the supremum is in $A_{n+1}/A_n = \left\{\frac{(2k+1)\tau}{2^{n+1}}, k = 0, \ldots, 2^n - 1\right\}$. Let the supremum be at $\frac{(2m+1)\tau}{2^{n+1}}$ and let x be such that

$$x \in \left\{\frac{2m\tau}{2^{n+1}}, \frac{(2m+2)\tau}{2^{n+1}}\right\} \subset A_n,$$

with $m \in \{0, \dots, 2^n - 1\}$ then,

$$\begin{aligned}\left|\eta\left(\frac{(2m+1)\tau}{2^{n+1}}\right)\right| &\leq \left|\eta\left(\frac{(2m+1)\tau}{2^{n+1}}\right) - \eta(x)\right| + |\eta(x)| \\ &\leq \max_{0\leq k\leq 2^{n+1}-1} |\eta(t_{k+1}^{n+1}) - \eta(t_k^{n+1})| + |\eta(0)| + |\eta(\tau)| \\ &+ \sum_{i=1}^{n} \max_{0\leq k\leq 2^i-1} |\eta(t_{k+1}^i) - \eta(t_k^i)|.\end{aligned}$$

Finally, since η is continuous and $[0, \tau]$ is compact, there exits a $x \in [0, \tau]$ such that $\sup_{t\in[0,\tau]} |\eta(t)| = \eta(x)$. Then, for all $\epsilon > 0$ it exits n and $k \in \{0, \dots, 2^n\}$ such that

$$\eta(x) \leq \eta\left(\frac{k\tau}{2^n}\right) + \epsilon,$$

since

$$\eta\left(\frac{k\tau}{2^n}\right) \leq |\eta(0)| + |\eta(\tau)| + \sum_{i=1}^{\infty} \max_{0\leq k\leq 2^i-1} |\eta(t_{k+1}^i) - \eta(t_k^i)|,$$

we obtained the desired conclusion. □

6.B.3 Gaussian correlation inequality

Theorem 6.B.3 (Gaussian Correlation Inequality [47])**.** *For any $n \geq 1$, if μ is a mean zero Gaussian measure on $\mathbb{R}^n$, then for C_1, C_2 convex closed subsets of $\mathbb{R}^n$ symmetric around the origin, we have,*

$$\mu(C_1 \cap C_2) \geq \mu(C_1)\mu(C_2)$$

6.B.4 Lemma 6.B.4

Lemma 6.B.4. *Let $(l(t))_{t\geq 0}$ denote a Gaussian Process with zero mean, stationary covariance function K and continuous sample paths, then*

$$\mathbf{P}\left(\sup_{t\in[0,\tau]}|l(t)|\leq M_1,\ \sup_{t\in(\tau,\tau_2]}|l(t)|\leq M_2\right) \geq \mathbf{P}\left(\sup_{t\in[0,\tau]}|l(t)|\leq M_1\right) \times\mathbf{P}\left(\sup_{t\in(\tau,\tau_2]}|l(t)|\leq M_2\right) \quad (6.B.3)$$

for $\tau\leq\tau_2$ and M_1, M_2 constants. The same result holds for τ_2 equal to infinity.

Proof of Lemma 6.B.4

Proof. Consider the finite collection of indices $A_n = \left\{t_k = \frac{k\tau_2}{2^n}, k=0,\ldots,2^n\right\}$, then by definition of a Gaussian process, $l(A_n)$ is a zero mean (since $(l(t))_{t\geq 0}$ has zero mean) multivariate Gaussian random variable in $\mathbb{R}^{|A_n|}$. Moreover, let $t_{k^\star}$ the largest element in A_n such that $t_{k^\star}\leq\tau$, then the sets

$$\begin{aligned} C_1 &= \{|l(t_0)|\leq M_1,\ldots,|l(t_{k^\star})|\leq M_1, l(t_{k^\star+1})\in\mathbb{R},\ldots,l(\tau_2)\in\mathbb{R}\} \\ C_2 &= \{l(t_0)\in\mathbb{R},\ldots,l(t_{k^\star})\in\mathbb{R}, |l(t_{k^\star+1})|\leq M_2,\ldots,|l(\tau_2)|\leq M_2\} \end{aligned} \quad (6.B.4)$$

are convex closed subsets of $\mathbb{R}^{|A_n|}$. Therefore, we can apply the Gaussian correlation inequality, Theorem 6.B.3, and have that

$$\mathbf{P}\left(C_1\cap C_2\right) \geq \mathbf{P}\left(C_1\right)\mathbf{P}\left(C_2\right),$$

which can be rewritten as

$$\begin{aligned} &\mathbf{P}\left(\sup_{t\in[0,\tau]\cap A_n}|l(t)|\leq M_1,\ \sup_{t\in(\tau,\tau_2]\cap A_n}|l(t)|\leq M_2\right) \quad (6.B.5) \\ \geq\ &\mathbf{P}\left(\sup_{t\in[0,\tau]\cap A_n}|l(t)|\leq M_1\right)\mathbf{P}\left(\sup_{t\in(\tau,\tau_2]\cap A_n}|l(t)|\leq M_2\right). \end{aligned}$$

Moreover, notice that by construction $A_n \subset A_{n+1}$ and then

$$\begin{aligned} & \left\{ \sup_{t\in[0,\tau]\cap A_{n+1}} |l(t)| \leq M_1, \sup_{t\in(\tau,\tau_2]\cap A_{n+1}} |l(t)| \leq M_2 \right\} \\ \subseteq & \left\{ \sup_{t\in[0,\tau]\cap A_n} |l(t)| \leq M_1, \sup_{t\in(\tau,\tau_2]\cap A_n} |l(t)| \leq M_2 \right\}, \end{aligned}$$

and then, by a standard properties of measures,

$$\begin{aligned} & \lim_{n\to\infty} \mathbf{P}\left(\sup_{t\in[0,\tau]\cap A_n} |l(t)| \leq M_1, \sup_{t\in(\tau,\tau_2]\cap A_n} |l(t)| \leq M_2 \right) \\ = & \mathbf{P}\left(\bigcap_{n\in\mathbb{N}} \left\{ \sup_{t\in[0,\tau]\cap A_n} |l(t)| \leq M_1, \sup_{t\in(\tau,\tau_2]\cap A_n} |l(t)| \leq M_2 \right\} \right) \\ = & \mathbf{P}\left(\sup_{t\in[0,\tau]} |l(t)| \leq M_1, \sup_{t\in(\tau,\tau_2]} |l(t)| \leq M_2 \right), \end{aligned}$$

where the last step is because the continuity of the Gaussian process $(l(t))_{t\geq 0}$. Then, taking limits in both sides of equation (6.B.5), we get the desired result. By the same arguments, we have the result when τ_2 is infinity. □

6.B.5 Lemma 6.B.5

Lemma 6.B.5. *Let $(\eta(t))_{t\geq 0}$ be a Gaussian process with continuous paths, zero mean and stationary covariance function K, satisfying $(K(0) - K(2^{-n}))^{-1} \geq n^6$, i.e., Assumption (A1). Then there exits some $\tau^\star > 1$, such that for all $\tau > \tau^\star$, $M > 0$ and $d \in \mathbb{N}$*

$$\Pi\left(\sup_{t\geq\tau} |\eta(t) h_d(t)| \geq M \right) \leq p(\tau),$$

where $p(\tau) = p(\tau, d, M, K(0))$ *is a positive function that depends on the dimension* d*, constant* M *and the variance* $K(0)$*, and takes the form of*

$$p(\tau) \quad = \quad \sum_{j\geq 0} q_1 \exp\{-q_2(\tau + j)^2\}, \tag{6.B.6}$$

where $q_1 = q_1(d, M, K(0))$ *and* $q_2(d, M, K(0))$ *are positive constants that depend on the dimension* d*, constant* M *and variance* $K(0)$*. Furthermore, for fixed* d *and* M*, it holds*

$$\lim_{\tau\to\infty} p(\tau) \quad = \quad 0.$$

Proof of Lemma 6.B.5

Proof. Define the sets

$$S \quad = \quad \left\{ \sup_{t\geq\tau} |\eta(t)h_d(t)| \geq M \right\},$$

and

$$S_j \quad = \quad \left\{ \sup_{t\in[\tau+j,\tau+j+1]} |\eta(t)| \geq h_d(\tau + j)^{-1} M \right\}.$$

Then since $h_d(t)$ is strictly decreasing for $t \geq \tau > 1$, we have

$$\Pi(S) \leq \Pi\left(\bigcup_{j\geq 0} S_j\right) \leq \sum_{j\geq 0} \Pi(S_j).$$

Consider the partition $A_n^j = \{t_{j,n,k} = \tau + j + \frac{k}{2^n}, k = 0, \ldots, 2^n\}$. Since the paths of the Gaussian process $(\eta(t))_{t\geq 0}$ are continuous, by Lemma 6.B.2 we have

$$\sup_{t\in[\tau+j,\tau+j+1]} |\eta(t)| \leq |\eta(\tau + j)| + \sum_{n=1}^{\infty} \max_{0\leq k\leq 2^n-1} |\eta(t_{j,n,k+1}) - \eta(t_{j,n,k})| + |\eta(\tau + j + 1)|.$$

On the other hand, define the set

$$\begin{aligned} Z_j \;\coloneqq\; & \left\{ |\eta(\tau+j)| \geq \frac{h_d(\tau+j)^{-1}M}{4} \right\} \\ \cup\; & \left\{ \bigcup_{n\geq 1} \left\{ \max_{0\leq k\leq 2^n-1} |\eta(t_{j,n,k+1}) - \eta(t_{j,n,k})| \geq \frac{3h_d(\tau+j)^{-1}M}{\pi^2 n^2} \right\} \right\} \\ \cup\; & \left\{ |\eta(\tau+j+1)| \geq \frac{h_d(\tau+j)^{-1}M}{4} \right\}, \end{aligned}$$

then we prove

$$S_j \subseteq Z_j,$$

for all $j \geq 0$. We proceed by contradiction. Take a function η in S_j and suppose it is not in Z_j, then by the definition of Z_j

$$|\eta(\tau+j)| + \sum_{n=1}^{\infty} \max_{0\leq k\leq 2^n-1} |\eta(t_{j,n,k+1}) - \eta(t_{j,n,k})| + |\eta(\tau+j+1)| < h_d(\tau+j)^{-1}M,$$

and hence, by Lemma 6.B.2,

$$\sup_{t\in[\tau+j,\tau+j+1]} |\eta(t)| < h_d(\tau+j)^{-1}M.$$

Nevertheless, for η in S_j it holds

$$\sup_{t\in[\tau+j,\tau+j+1]} |\eta(t)| \geq h_d(\tau+j)^{-1}M.$$

which is a contradiction. By the union bound

$$\begin{aligned} \Pi(S_j) \;\leq\; & \Pi(Z_j) \\ \leq\; & \Pi\left(|\eta(\tau+j)| \geq \frac{h_d(\tau+j)^{-1}M}{4} \right) + \Pi\left(|\eta(\tau+j+1)| \geq \frac{h_d(\tau+j)^{-1}M}{4} \right) \\ +\; & \sum_{n\geq 1} \sum_{0\leq k\leq 2^n-1} \Pi\left(|\eta(t_{j,n,k+1}) - \eta(t_{j,n,k})| \geq \frac{3h_d(\tau+j)^{-1}M}{\pi^2 n^2} \right). \end{aligned} \tag{6.B.7}$$

We continue by using a well-known concentration inequality for Gaussian random variables, that is, *"the Gaussian concentration inequality"*. For a Normal random variable X with mean μ and variance σ^2, the Gaussian concentration inequality [7] states

$$\mathbb{P}(|X-\mu| \geq t) \leq 2\exp\left\{-\frac{t^2}{2\sigma^2}\right\}. \tag{6.B.8}$$

Then, since η denotes a centred Gaussian process, we use this inequality to bound the following probabilities

$$\Pi\left(|\eta(\tau+j)| \geq \frac{h_d(\tau+j)^{-1}M}{4}\right) \leq 2\exp\left\{\frac{h_d(\tau+j)^{-2}M^2}{32K(0)}\right\} \tag{6.B.9}$$

and

$$\Pi\left(|\eta(\tau+j+1)| \geq \frac{h_d(\tau+j)^{-1}M}{4}\right) \leq 2\exp\left\{\frac{h_d(\tau+j)^{-2}M^2}{32K(0)}\right\}. \tag{6.B.10}$$

Additionally, define the events $I_{j,n,k} = \left\{|\eta(t_{j,n,k+1}) - \eta(t_{j,n,k})| \geq \frac{3h_d(\tau+j)^{-1}M}{\pi^2 n^2}\right\}$ and notice $\eta(t_{j,n,k+1}) - \eta(t_{j,n,k+1})$ is Gaussian random variable with zero mean and variance given by $2(K(0) - K(2^{-n}))$. Then, by using the Gaussian concentration inequality, equation (6.B.8), it holds

$$\begin{aligned}\sum_{n\geq 1}\sum_{0\leq k\leq 2^n-1}\Pi\left(I_{j,n,k}\right) &\leq \sum_{n\geq 1} 2^{n+1}\exp\left\{-\frac{9h_d(\tau+j)^{-2}M^2}{4\pi^2 n^4(K(0)-K(2^{-n}))}\right\} \\ &= 2\sum_{n\geq 1}\exp\left\{-\frac{9h_d(\tau+j)^{-2}M^2(K(0)-K(2^{-n}))^{-1}}{4\pi^4 n^4} + n\log 2\right\}.\end{aligned}$$

By assumption (A1) it holds $(K(0) - K(2^{-n}))^{-1} \geq n^6$, then

$$\begin{aligned}\sum_{n\geq 1}\sum_{0\leq k\leq 2^n-1}\Pi\left(I_{j,n,k}\right) &\leq 2\sum_{n\geq 1}\exp\left\{-\frac{9h_d(\tau+j)^{-2}M^2 n}{4\pi^4} + n\log 2\right\} \\ &= 2\sum_{n\geq 1}\exp\left\{-C_j\right\}^n, \end{aligned} \tag{6.B.11}$$

where $C_j = \frac{9h_d(\tau+j)^{-2}M^2}{4\pi^4} - \log 2$. Since $h_d(\tau)$ is strictly decreasing for $\tau \geq 1$, it holds that $\exp\{-C_j\}$ is strictly decreasing for $j \geq 0$ and fixed τ. Moreover, by the same argument, we can always pick a $\tau^\star > 1$ in such way that for all $\tau \geq \tau^\star$,

$$\exp\{-C_0\} = \exp\left\{-\frac{9h_d(\tau)^{-2}M^2}{4\pi^4} + \log 2\right\} < 1. \tag{6.B.12}$$

Using the latter result, the geometric series in equation (6.B.11) converges for all $j \geq 0$, as $\exp\{-C_j\} \leq \exp\{-C_0\} < 1$. Indeed

$$(6.B.11) \quad = \quad 2\frac{\exp\{-C_j\}}{1-\exp\{-C_j\}} \leq 2\frac{\exp\{-C_j\}}{1-\exp\{-C_0\}} \leq q_0 \exp\{-C_j\}, \tag{6.B.13}$$

for all $j \geq 0$ and some constant q_0 (which depends on d and M). Plugging-in the result of equations (6.B.9), (6.B.10) and (6.B.13) in equation (6.B.7), we have

$$\Pi(S_j) \quad \leq \quad 4\exp\left\{-\frac{h_d(\tau+j)^{-2}M^2}{32K(0)}\right\} + q_0\exp\{-C_j\}.$$

Finally,

$$\begin{aligned}\Pi(S) &\leq \sum_{j\geq 0}\Pi(S_j) \\ &\leq \sum_{j\geq 0}\left(4\exp\left\{-\frac{h_d(\tau+j)^{-2}M^2}{32K(0)}\right\} + q_0\exp\left\{-\frac{9h_d(\tau+j)^{-2}M^2}{4\pi^4} - \log 2\right\}\right),\end{aligned}$$

with

$$h_d(t)^{-2} \quad = \quad \frac{(t+\log(1-e^{-t}))^2}{(d+1)^2}.$$

Then there exists some positive constants $q_1 = q_1(d, M, K(0))$, where $K(0)$ denoted the kernel evaluated at time 0, and $q_2 = q_2(d, M, K(0))$, such that

$$\Pi(S) \quad \leq \quad \sum_{j\geq 0} q_1 \exp\{-q_2(\tau+j)^2\}. \tag{6.B.14}$$

From an standard application of dominated convergence, we deduce that the limit goes to zero as τ goes to infinity for fixed M and d. □

6.B.6 Lemma 6.B.6

Lemma 6.B.6. *Let $(\eta(t))_{t\geq 0}$ be a Gaussian process with continuous sample paths, zero mean and stationary covariance function K, satisfying $(K(0)-K(2^{-n}))^{-1} \geq n^6$, i.e., Assumption (A1). Let $\tau \geq 1$ and $\delta > 0$, then*

$$\Pi\left(\sup_{t\in[0,\tau]} |\eta(t)h_d(t)| \leq \frac{\delta h_d(\tau)}{1+\tau}\right) \quad > \quad 0.$$

Proof of Lemma 6.B.6

Proof. Recall the definition of function h_d given in equation (6.4.4)

$$h_d(t) \quad = \quad \begin{cases} \frac{d+1}{1+\log(1-e^{-1})} & t \leq 1 \\ \frac{d+1}{t+\log(1-e^{-t})} & t > 1 \end{cases},$$

hence $h_d(t)$ is strictly decreasing and positive for $t \geq 1$. Consider $\tau \geq 1$, then

$$\left\{\sup_{t\in[0,\tau]} |\eta(t)h_d(t)| \leq \frac{\delta h_d(\tau)}{1+\tau}\right\} \quad \supseteq \quad \left\{\sup_{t\in[0,\tau]} |\eta(t)| \leq \frac{\delta h_d(\tau)}{h_d(1)(1+\tau)}\right\}.$$

Consider the partition $A_n = \left\{t_k^n = \frac{k\tau}{2^n}, k = 0, \ldots, 2^n\right\}$, by Lemma 6.B.2

$$\sup_{t\in[0,\tau]} |\eta(t)| \quad \leq \quad |\eta(0)| + |\eta(\tau)| + \sum_{n=1}^{\infty} \max_{0\leq k\leq 2^n-1} \left|\eta(t_{k+1}^n) - \eta(t_k^n)\right|.$$

For simplicity, let's call $\psi_\delta(\tau) = \delta\frac{h_d(\tau)}{h_d(1)(1+\tau)}$ and define the event

$$S \quad = \quad \left\{\sup_{t\in[0,\tau]} |\eta(t)| \leq \psi_\delta(\tau)\right\}.$$

On the other hand, define the event

$$\begin{aligned} Z &:= \left\{|\eta(0)| \leq \frac{\psi_\delta(\tau)}{4}\right\} \cap \left\{ |\eta(\tau)| \leq \frac{\psi_\delta(\tau)}{4}\right\} \\ &\cap \left\{\bigcap_{n=1}^{\infty}\left\{\max_{0\leq k\leq 2^n-1}\left|\eta(t_{k+1}^n)-\eta(t_k^n)\right| \leq \frac{3\psi_\delta(\tau)}{n^2\pi^2}\right\}\right\}. \end{aligned}$$

Then, for any function η in Z, it holds

$$\sup_{t\in[0,\tau]}|\eta(t)| \leq |\eta(0)|+|\eta(\tau)|+\sum_{n=1}^{\infty}\max_{0\leq k\leq 2^n-1}\left|\eta(t_{k+1}^n)-\eta(t_k^n)\right| \leq \psi_\delta(\tau),$$

which means it also belongs to S and then

$$\Pi(S) \geq \Pi(Z).$$

Additionally, define finite version of the event Z, that is,

$$\begin{aligned} Z_N &:= \left\{|\eta(0)| \leq \frac{\psi_\delta(\tau)}{4}\right\} \cap \left\{ |\eta(\tau)| \leq \frac{\psi_\delta(\tau)}{4}\right\} \\ &\cap \left\{\bigcap_{n=1}^{N}\left\{\max_{0\leq k\leq 2^n-1}\left|\eta(t_{k+1}^n)-\eta(t_k^n)\right| \leq \frac{3\psi_\delta(\tau)}{n^2\pi^2}\right\}\right\}. \end{aligned}$$

Consider the collection of random variables $\{\alpha_k^n = \eta(t_{k+1}^n)-\eta(t_k^n), k \in \{0,\dots,2^n-1\}, n \in \{1,\dots,N\}\}$, $\eta(0)$ and $\eta(\tau)$. Observe there are in total $2+\sum_{n=1}^{N} 2^n = 2^{N+1}$ variables, then the joint distribution of them is a zero mean, 2^{N+1}-variate Gaussian distribution, i.e., they lie on $\mathbb{R}^{2^{N+1}}$.

Therefore, by replicating the arguments given in the proof of Lemma 6.B.4 (since we have convex, closed and symmetric around the origin events on $\mathbb{R}^{2^{N+1}}$), we use the Gaussian correlation inequality, to prove

$$\Pi(Z_N) \geq \Pi\left(|\eta_0| \leq \frac{\psi_\delta(\tau)}{4}\right)\Pi\left(|\eta_\tau| \leq \frac{\psi_\delta(\tau)}{4}\right)\prod_{n=1}^{N}\prod_{k=0}^{2^n-1}\Pi\left(\left|\eta_{t_{k+1}}-\eta_{t_k}\right| \leq \frac{3\psi_\delta(\tau)}{n^2\pi^2}\right).$$

Additionally, since the events $Z_N \supseteq Z_{N+1}$ for all $N \geq 1$,

$$\lim_{N\to\infty} \Pi\left(Z_N\right) = \lim_{N\to\infty} \Pi\left(\bigcap_{n=1}^{N} Z_n\right) = \Pi\left(\lim_{N\to\infty} \bigcap_{n=1}^{N} Z_n\right) = \Pi(Z),$$

and hence

$$\begin{aligned}\Pi(S) &\geq \Pi(Z) \\ &\geq \Pi\left(|\eta_0| \leq \frac{\psi_\delta(\tau)}{4}\right) \Pi\left(|\eta_\tau| \leq \frac{\psi_\delta(\tau)}{4}\right) \prod_{n=1}^{\infty} \prod_{k=0}^{2^n-1} \Pi\left(\left|\eta_{t_{k+1}} - \eta_{t_k}\right| \leq \frac{3\psi_\delta(\tau)}{n^2\pi^2}\right).\end{aligned}$$

Recall that for a Normal random variable X with mean μ and variance σ^2, the Gaussian concentration inequality [7] states

$$\mathbf{P}(|X-\mu| > t) \leq 2\exp\left\{-\frac{t^2}{2\sigma^2}\right\}. \tag{6.B.15}$$

Using this result and the assumption (A1) $(K(0) - K(2^n))^{-1} \geq n^6$, we have

$$\begin{aligned}\Pi(S) &\geq \Pi\left(|\eta_0| \leq \frac{\psi_\delta(\tau)}{4}\right)^2 \prod_{n=1}^{\infty} \left(1 - \exp\left\{-\frac{9\psi_\delta(\tau)^2 (K(0) - K(2^{-n}))^{-1}}{4n^4\pi^2}\right\}\right)^{2^n} \\ &\geq \Pi\left(|\eta_0| \leq \frac{\psi_\delta(\tau)}{4}\right)^2 \prod_{n=1}^{\infty} \left(1 - \exp\left\{-\frac{9\psi_\delta(\tau)^2 n^2}{4\pi^2}\right\}\right)^{2^n}.\end{aligned}$$

Now use $\exp\left\{-\frac{x}{1-x}\right\} \leq 1 - x$

$$\begin{aligned}\mathbf{P}(S) &\geq \mathbf{P}\left((|\eta_0| \leq \frac{\psi_\delta(\tau)}{4}\right)^2 \prod_{n=1}^{\infty} \exp\left\{-\frac{2^n \exp\left\{-\frac{9\psi_\delta(\tau)^2 n^2}{4\pi^2}\right\}}{1 - \exp\left\{-\frac{9\psi_\delta(\tau)^2 n^2}{4\pi^2}\right\}}\right\} \\ &\geq \mathbf{P}\left((|\eta_0| \leq \frac{\psi_\delta(\tau)}{4}\right)^2 \exp\left\{-\sum_{n=1}^{\infty} \frac{\exp\left\{-\frac{9\psi_\delta(\tau)^2 n^2}{4\pi^2} + n\log 2\right\}}{1 - \exp\left\{-\frac{9\psi_\delta(\tau)^2 n^2}{4\pi^2}\right\}}\right\} > 0,\end{aligned}$$

since the latter series converges. □

Chapter 7

Conclusions and future work

The main aim of this thesis has been to contribute to the existing literature of survival analysis by proposing and studying a novel Bayesian nonparametric model for survival data. To this end, our first major contribution was the proposal and study of a prior distribution over hazard functions, based on a Gaussian process. In this regard, we studied the well-definiteness of the hazards sampled from the prior and the moments of the random variables associated with it. The second major contribution of this thesis was the proposal of an exact and tractable inference scheme to sample from the posterior distribution. Our scheme is able to deal with different types of censoring as well as covariates. Lastly, our final major contribution was to provide sufficient conditions to ensure posterior consistency of our model. In addition to this, we believe our proof technique is quite valuable since, up to our knowledge, most of the previous works proving posterior consistency for models under a Gaussian process prior consider data on a compact set and most of the techniques used cannot be extended to our framework. Therefore, our work not only proves the result but also gives a new technique for data in $[0, \infty+)$ under Gaussian process priors.

In Chapter 2, we introduced our model and proposed two covariate schemes. In one hand, we believe that an appealing property of our model is that it can

be interpreted in terms of a parametric baseline hazard. Indeed, as the prior is constructed in such way we can centre it in a parametric family, we can explain this model as a noisy version of such hazard. On the other hand, a drawback of our model is that when we include an incorrect prior knowledge, by specifying a not suitable baseline hazard function λ_0, the inference scheme may need a lot of time and data to perform well. This is because the prior seems to be too concentrated around its mean, i.e., λ_0. An option to avoid this problem is to consider a broad parametric family for λ_0 and to learn the appropriate parameters. Another option is to modify our model such a way it is less concentrated around λ_0.

In Chapter 3, we analysed frequentist properties of our prior and their impact in the model. In particular, we proved that the prior defines proper hazard functions with probability one. Additionally, we studied the (random) moments of the random variables associated with those hazards. All the results in this chapter rely on properties of the link function, in our case, sigmoid function $\sigma(x) = (1+e^{-x})^{-1}$. We believe that our analysis can be extended to other bounded, non-negative functions without major modifications. The result for unbounded functions seems to be more difficult to analyse and thus constitutes an interesting research direction. Part of our analysis included to prove that, for large t, the random cumulative hazard function $\Lambda(t)$ is concentrated around $\Lambda_0(t)$. We believe this holds for all large t at the same time, but we were not able to provide a proof.

In Chapter 4, we introduced an exact inference scheme to perform inference in the model. While our approach proposed an effective way of sampling from the posterior distributions, it was computationally expensive. In order to give a solution to this problematic, we proposed an approximation scheme based on random Fourier features. Future research directions are to find less expensive, but still exact, inference algorithms. Also, we believe it would be interesting to explore the use of alternative inference schemes based on approximations of the likelihood of the original (not augmented) model.

In Chapter 5, we assessed our model empirically obtaining overall good results

in the synthetic dataset as well as in the real dataset. Future work consist on assessing our model in a larger real dataset.

In Chapter 6, we proved posterior consistency for a simplified version of the model proposed in Chapter 2, and under a metric that considers the maximum difference between two survival functions over a large class of stratifications of covariates $\boldsymbol{x} \in [0,1]^d$ and time. In general, we believe the assumptions listed in section 6.4.4 are quite reasonable and natural for the type of data we were dealing with. We proceed to review the more important of them.

A1) This assumption ask for stationary kernels K such that $(K(0)-K(2^{-n}))^{-1} \geq n^6$. This is not a restrictive assumption since most of the standard kernels satisfy such property.

A4) Here we assume that the true parameters $\eta_{j,0}$ belongs to the support of $(\eta_j(t))_{t\geq 0}$ in the Banach space $(\mathcal{C}, \|\cdot\|_{h_d,\infty})$. This assumption can be very complicated to understand, but in practice, as we saw in Remark 6.4.4 and in the example above it, for the commonly used stationary kernels, this assumption is reduced to check that the true parameter is a continuous function such that the quotient between the true parameter and $1+t$ tends to 0 as $t \to \infty$ (This follows from the definition of the function h_d).

Additionally, we think our proof can be extended to other models that do not consider the sigmoid link function, but other non-negative bounded and Lipchitz functions. The reason for this is because our proof just uses these particular properties of the sigmoid function. Finally, another avenue for future work is to extend this result for the censored case with covariates, as we think the case of right censoring without covariates should follow straightforwardly from our analysis as there is a broad literature of existing results that can provide useful tests for survival data. It is also future work to consider other metrics to prove this result. A final open problem is to obtain convergence rates.

Bibliography

[1] Odd Aalen. Nonparametric estimation of partial transition probabilities in multiple decrement models. *The Annals of Statistics*, pages 534–545, 1978.

[2] Ryan Prescott Adams, Iain Murray, and David JC MacKay. Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 9–16. ACM, 2009.

[3] Per Kragh Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

[4] James E Barrett and Anthony CC Coolen. Gaussian process regression for survival data with competing risks. *arXiv preprint arXiv:1312.1591*, 2013.

[5] Andrew R Barron. Discussion: On the consistency of bayes estimates. *The Annals of statistics*, 14(1):26–30, 1986.

[6] J Blum and V Susarla. On the posterior distribution of a dirichlet process given randomly right censored observations. *Stochastic Processes and their Applications*, 5(3):207–211, 1977.

[7] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[8] Norman Breslow. A generalized kruskal-wallis test for comparing k samples subject to unequal patterns of censorship. *Biometrika*, pages 579–594, 1970.

[9] Craig Calcaterra. Linear combinations of gaussians with a single variance are dense in l2. In *Proceedings of the World Congress on Engineering*, volume 2, 2008.

[10] Taeryon Choi and Mark J Schervish. On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98(10):1969–1987, 2007.

[11] Nidhan Choudhuri, Subhashis Ghosal, and Anindya Roy. Bayesian estimation of the spectral density of a time series. *Journal of the American Statistical Association*, 99(468):1050–1059, 2004.

[12] David R Cox. Partial likelihood. *Biometrika*, pages 269–276, 1975.

[13] DR Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[14] Maria De Iorio, Wesley O Johnson, Peter Müller, and Gary L Rosner. Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, 65(3):762–771, 2009.

[15] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

[16] Kjell Doksum. Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, pages 183–201, 1974.

[17] J.L. Doob. Application of the theory of martingales. *In Le Calcul des Probabilités et ses Applications,Colloques Internationaux du Centre National de la Rechercher Scientifique*, no. 13:23–27, 1949.

[18] David K Duvenaud, Hannes Nickisch, and Carl E Rasmussen. Additive gaussian processes. In *Advances in neural information processing systems*, pages 226–234, 2011.

[19] RL Dykstra and Purushottam Laud. A bayesian nonparametric approach to reliability. *The Annals of Statistics*, pages 356–367, 1981.

[20] Thomas S Ferguson and Eswar G Phadia. Bayesian nonparametric estimation based on censored data. *The Annals of Statistics*, pages 163–186, 1979.

[21] Tamara Fernández, Nicolás Rivera, and Yee Whye Teh. Gaussian processes for survival analysis. In *Advances in Neural Information Processing Systems*, pages 5021–5029, 2016.

[22] Tamara Fernández and Yee Whye Teh. Posterior consistency for a non-parametric survival model under a gaussian process prior. *arXiv preprint arXiv:1611.02335*, 2016.

[23] Edmund A Gehan. A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52(1-2):203–223, 1965.

[24] Subhashis Ghosal, Jayanta K Ghosh, RV Ramamoorthi, et al. Posterior consistency of dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158, 1999.

[25] Subhashis Ghosal and Anindya Roy. Posterior consistency of gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, pages 2413–2429, 2006.

[26] Jayanta K Ghosh and RV Ramamoorthi. Consistency of bayesian inference for survival analysis with or without censoring. *Lecture Notes-Monograph Series*, pages 95–103, 1995.

[27] Richard D Gill. Censoring and stochastic integrals. *Statistica Neerlandica*, 34(2):124–124, 1980.

[28] David P Harrington and Thomas R Fleming. A class of rank test procedures for censored survival data. *Biometrika*, pages 553–566, 1982.

[29] Nils Lid Hjort. Nonparametric bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, pages 1259–1294, 1990.

[30] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G Walker. *Bayesian nonparametrics*, volume 28. Cambridge University Press, 2010.

[31] Heikki Joensuu, Aki Vehtari, Jaakko Riihimäki, Toshirou Nishida, Sonja E Steigen, Peter Brabec, Lukas Plank, Bengt Nilsson, Claudia Cirilli, Chiara Braconi, et al. Risk of recurrence of gastrointestinal stromal tumour after surgery: an analysis of pooled population-based cohorts. *The lancet oncology*, 13(3):265–274, 2012.

[32] Søren Johansen. The product limit estimator as maximum likelihood estimator. *Scandinavian Journal of Statistics*, pages 195–199, 1978.

[33] John D Kalbfleisch. Non-parametric bayesian analysis of survival time data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 214–221, 1978.

[34] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[35] Yongdai Kim and Jaeyong Lee. On posterior consistency of survival models. *Annals of Statistics*, pages 666–686, 2001.

[36] Yongdai Kim, Jaeyong Lee, et al. Bayesian analysis of proportional hazard models. *The Annals of Statistics*, 31(2):493–511, 2003.

[37] Bas Kleijn, Aad van der Vaart, and Harry van Zanten. *Lectures on Nonparametric Bayesian Statistics*. Unpublished, 2012.

[38] James Kuelbs, Wenbo V Li, and Werner Linde. The gaussian measure of shifted balls. *Probability Theory and Related Fields*, 98(2):143–162, 1994.

[39] Purushottam W. Laud, Paul Damien, and Adrian F. M. Smith. *Bayesian Nonparametric and Covariate Analysis of Failure Time Data*, pages 213–225. Springer New York, New York, NY, 1998.

[40] Mei-Ling Ting Lee and George A Whitmore. Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statistical Science*, pages 501–513, 2006.

[41] Peter J Lenk. The logistic normal distribution for bayesian, nonparametric, predictive densities. *Journal of the American Statistical Association*, 83(402):509–516, 1988.

[42] Georg Lindgren. Lectures on stationary stochastic processes.

[43] Albert Y Lo et al. On a class of bayesian nonparametric estimates: I. density estimates. *The annals of statistics*, 12(1):351–357, 1984.

[44] Gilbert Mackenzie. Regression models for survival data: the generalized time-dependent logistic family. *The Statistician*, pages 21–34, 1996.

[45] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports. Part 1*, 50(3):163–170, 1966.

[46] Sara Martino, Rupali Akerkar, and Håvard Rue. Approximate bayesian inference for survival models. *Scandinavian Journal of Statistics*, 38(3):514–528, 2011.

[47] Yashar Memarian. The gaussian correlation conjecture proof. *arXiv preprint arXiv:1310.8099*, 2013.

[48] Iain Murray and Ryan P Adams. Slice sampling covariance hyperparameters of latent gaussian models. In *Advances in Neural Information Processing Systems*, pages 1732–1740, 2010.

[49] Debdeep Pati, David B Dunson, and Surya T Tokdar. Posterior consistency in conditional distribution estimation. *Journal of multivariate analysis*, 116:456–472, 2013.

[50] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2007.

[51] Vinayak Rao and Yee W. Teh. Gaussian process modulated renewal processes. In *Advances in Neural Information Processing Systems*, pages 2474–2482, 2011.

[52] Walter Rudin. *Fourier analysis on groups*. Courier Dover Publications, 2017.

[53] Lorraine Schwartz. On bayes procedures. *Probability Theory and Related Fields*, 4(1):10–26, 1965.

[54] V Susarla and John Van Ryzin. Nonparametric bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, 71(356):897–902, 1976.

[55] Terry M Therneau and Thomas Lumley. Package survival, 2015.

[56] Surya T Tokdar and Jayanta K Ghosh. Posterior consistency of logistic gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, 137(1):34–42, 2007.

[57] Aad W van der Vaart and J Harry van Zanten. Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, pages 1435–1463, 2008.

[58] Aad W van der Vaart, J Harry van Zanten, et al. Reproducing kernel hilbert spaces of gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pages 200–222. Institute of Mathematical Statistics, 2008.

[59] Stephen Walker and Pietro Muliere. Beta-stacy processes and a generalization of the pólya-urn scheme. *The Annals of Statistics*, pages 1762–1780, 1997.

[60] Lee-Jen Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879, 1992.