# Statistical Models for Survival Data

## *Theory and Applications*

Alessandro Morosini

*Supervised by*

Prof. Antonio Lijoi

Bocconi University

July 2024

# Contents

# Chapter 1

# Introduction

## 1.1 Overview of Survival Analysis

Survival analysis is a branch of statistics dedicated to studying the time until an event of interest occurs. The time from the start of the observation period to the occurrence of the event serves as the outcome variable, which we analyze, and it can be represented in years, months, days, age, or some general time measure. The event of interest can vary widely, ranging from the death of a patient to the bankruptcy of a firm or the malfunctioning of a device.

Because of its adaptability, survival analysis has been used in several different fields, earning it many titles: in economics, it is known as duration analysis; in the social sciences, it is referred to as event history analysis; and in engineering, it is termed reliability theory. The name survival analysis comes from its predominant application in healthcare and medicine, where it is used to model the survival times of patients.

## 1.2 Survival Data

### 1.2.1 Censoring and Notation

Censoring is a key problem in survival analysis, and it occurs when the exact survival time of individuals is not known. Most survival analyses must consider this issue to accurately model

the data and avoid biased conclusions.

The most commonly encountered type of censoring is right censoring, which occurs when the event of interest has not happened by the end of the study period. The reasons for right censoring are many, but it mainly happens when participants are still alive by the end of the study or when they drop out of the study before the event actually occurs.

Other, less common, types of censoring are left and interval censoring. Left censoring happens when the event has already occurred before the subject is observed in the study, while interval censoring occurs when the event is known to have occurred within a specific time interval, but the precise time is not known. Throughout the course of this work, we will assume that censoring is non-informative, which means that censoring is independent of the event that would have been observed otherwise.

Another common issue related to survival data is truncation, which refers to the fact that we only observe individuals whose events happen within a specific time frame and we do not have information about events outside of this window.

We now introduce some notation that will be used throughout this work to represent survival data. We let $y$ be the time to the event of interest starting from the beginning of the study, and $c$ be the point at which the observation is censored if the event has not yet occurred; the observed time is then defined as $t = \min(y, c)$, i.e. it is the minimum of the failure time and the censoring time. We also define the censoring indicator $\delta = \mathbb{1}_{\{y \leq c\}}$, which is 1 if the event is observed and 0 if it is censored. The observed full data $D$ is then represented as $D = (n, \mathbf{t}, X, \delta)$, where $n$ is the number of subjects, $\mathbf{t}$ is the vector of observed times $(t_1, \ldots, t_n)$, $X$ is the matrix of covariates with each row $\mathbf{x_i}$ representing the covariates for the $i$-th subject, and $\delta$ is the vector of censoring indicators $(\delta_1, \ldots, \delta_n)$.

### 1.2.2 Core Survival Functions

Having presented the fundamental concept of censoring and some common notation, we now present the core functions used to describe and analyze survival data.

The time to the event of interest is modeled as the realization of a continuous non-negative

random variable $T$, and we let $f(t)$ be its probability density function (pdf). Then, cumulative distribution function (cdf) of $T$ is

$$F(t) = P(T \leq t) = \int_0^t f(u)du, \tag{1.1}$$

which is also known as the failure function in the context of survival analysis.

The failure function is strictly related to the survival function $S(t)$, which represents the probability for an individual to survive up to a certain time $t$. The survival function is

$$S(t) = Pr(T > t) = 1 - F(t). \tag{1.2}$$

By differentiation, one also easily obtains

$$f(t) = -\frac{dS(t)}{dt}. \tag{1.3}$$

Both the failure function $F(t)$ and the survival function $S(t)$ are probabilities, hence they are included between zero and one. Moreover, we mention three important and intuitive properties of the survival function. First, it is monotonically decreasing in $t$, meaning that the probability of survival decreases as time goes on. Second, $S(0) = 1$, representing the fact that at the beginning of the study no event has occurred. Lastly, $\lim_{t\to\infty} S(t) = 0$, meaning that in an infinitely long term no individual will survive. Figure 1.1 shows the relationship among the three functions.

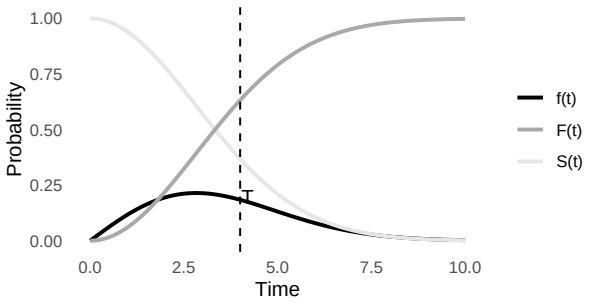


Figure 1.1: The figure shows the probability density function $f(t)$, the cumulative distribution function $F(t)$, and the survival function $S(t)$, illustrating their relationship.

Another fundamental function is the hazard rate $h(t)$, also known as instantaneous death rate or conditional failure rate. It is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T \leq t + \Delta t \mid T > t)}{\Delta t}. \tag{1.4}$$

For an infinitesimally small interval of time, $h(t)\Delta t$ represents the approximate probability of the event occurring at time $t$, conditional on survival up to time $t$.

Combining Equation 1.4 and the rules of conditional probability, the following fundamental relationship is obtained:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t \mid T > t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{P(T > t)\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{S(t)\Delta t} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{1.5}$$

In words, the hazard rate $h(t)$ is defined as the ratio of the density of events at $t$ and the probability of surviving to that time without experiencing the event. We stress that the hazard rate is not a probability, hence it has no upper constraint of one and can have a variety of shapes. However, it follows by the properties of $f(t)$ and $S(t)$ that $h(t) \geq 0$.

An important property worth mentioning is that there exists a one-to-one relationship between the definition of survival function and hazard rate, meaning that whatever $h(t)$ is chosen, it is possible to uniquely obtain $S(t)$, hence $F(t)$ and $f(t)$. This property is obtained combining Equations 1.1, 1.2 and 1.5:

$$\begin{aligned} h(t) &= \frac{f(t)}{1 - F(t)} \\ &= \frac{-\frac{d}{dt}(1 - F(t))}{1 - F(t)} \\ &= -\frac{d}{dt} \log(1 - F(t)), \end{aligned} \tag{1.6}$$

and integrating both sides we have

$$\begin{aligned}\int_0^t h(u)du &= \int_0^t -\frac{d}{dt}\log(1-F(u))du \\ &= -\log(1-F(u))\Big|_{u=0}^{u=t} \\ &= -\log(1-F(t)) \\ &= -\log(S(t)),\end{aligned} \tag{1.7}$$

where we used the fact that $S(0) = 1$ and $F(0) = 0$. Exponentiating both sides, one finally obtains the following relation:

$$S(t) = \exp\Big(-\int_0^t h(u)du\Big) = \exp(-H(t)), \tag{1.8}$$

where $H(t) = \int_0^t h(u)du$ is known as the cumulative hazard. Moreover, by Equation 1.5 one also has

$$f(t) = h(t)\exp(-H(t)). \tag{1.9}$$

Up to this point we have assumed implicitly that the event of interest can happen at any point in time, implying that the outcome variable is continuous. However, this is not always true. In some scenarios, survival times have been grouped and should therefore be modeled as discrete variables, or it might also be that the underlying process itself is discrete by nature. In these cases, a slight modification of the framework that was presented is necessary.

## 1.3 Thesis Outline

Traditional survival analysis models are widely used to estimate survival functions and assess the impact of covariates on survival times. Some of the most commonly used models include the non-parametric Kaplan-Meier estimator, the semi-parametric Cox proportional hazards model, as well as various parametric models such as the Exponential and the Weibull. These techniques will be discussed in Chapter 2.

While traditional survival analysis models are powerful and widely used, they have several limitations. These motivate the investigation of alternative approaches. In this work, we aim to explore the use of Bayesian approaches in survival analysis. The basics of Bayesian statistics will be presented in Chapter 3, which will also highlight the main differences with the frequentist approach and describe some of its most relevant advantages.

A very powerful, yet not widely used in survival analysis, non-parametric Bayesian model is then introduced in Chapter 4: the Dirichlet process mixture model. The reader will be presented with an introduction to the Dirichlet distribution and Dirichlet process, and the discussion will then be extended to Dirichlet process mixture models and their application in survival analysis.

Chapter 5 will provide a real world application of the discussed techniques. Both the frequentist and Bayesian models will be applied to estimate the effectiveness of chemotherapy drugs on the treatment of lung cancers. A comparative performance of the models will also be provided.

Finally, Chapter 6 summarizes the strengths and weaknesses of each model discussed, presenting the key findings of the analysis.

# Chapter 2

# Popular Frequentist Models in Survival Analysis

Some of the most widely used frequentist models in survival analysis are presented in this chapter. First, the Kaplan-Meier estimator and a test to compare survival curves are discussed. Then, semi-parametric models are introduced, with a focus on Cox regression and the assumption of proportional hazards. Finally, parametric models are presented, addressing both proportional hazards and accelerated failure time specifications. An illustrative example of the widely used Weibull model is also provided.

## 2.1 Non-Parametric Modeling

Non-parametric models play a significant role in survival analysis because they provide an estimation of the survival function without making any specific assumption about the distribution of the outcome variable. This class of models is especially useful when little information about the data is available, and the most common assumptions cannot be made.

Non-parametric methods excel in their flexibility and require minimal assumptions. Nonetheless, they have some drawbacks. First, the effects of the covariates on the outcome variable cannot be properly quantified. Additionally, due to their lack of assumptions, non-parametric models do not have a way to incorporate information. Hence, when such information is avail-

able, other models that can exploit it might be preferred.

### 2.1.1 Kaplan-Meier Estimator

The most popular non-parametric approach is related to the Kaplan-Meier (KM) estimator [11], which provides an empirical estimate for the survival function.

To construct the estimator, we order the set of $n$ observed lifetimes from smallest to largest, obtaining $t_1 < t_2 < \ldots < t_n$. We let $d_i$ and $n_i$, with $d_i \leq n_i$, denote respectively the number of events and the number of individuals at risk at time $t_i$. The conditional probability of survival after a certain time $t_i$ is then estimated by

$$p(t_i) = \frac{n_i - d_i}{n_i}. \tag{2.1}$$

The Kaplan-Meier estimator is the product over all times $t_i$ less than $t$, that is

$$\hat{S}(t) = \prod_{t_i<t} p(t_i) = \prod_{t_i<t} \left( \frac{n_i - d_i}{n_i} \right), \tag{2.2}$$

This estimator comes in handy when comparing the survival times of two or more groups, as it can be fitted on the different groups separately to then compare the estimated survival curves. However, a statistical test is necessary to ascertain whether the group differences are statistically significant. Hence, we need a procedure to test the null hypothesis $H_0$ that all groups have equal survival against the alternative hypothesis $H_1$ that they have different survivals. For this task, a non-parametric hypothesis test called the log-rank test is usually used.

### 2.1.2 Log-Rank Test

The construction of the log-rank test [15] involves computing the observed and expected number of events within a specific group at each event time, and then adding these values for all event times to obtain an overall summary of the data.

We let $K$ be the total number of groups, $n_{ki}$ be the number of individuals at risk in group $k$

at time $t_i$, and $d_{ki}$ be the number of events in group $k$ at time $t_i$. Then, the expected number of events at time $t_i$ for group $k$, under the null hypothesis that the hazard rates are the same in all groups, is calculated as

$$E_{ki} = \frac{d_i}{n_i} n_{ki}, \tag{2.3}$$

Here, $n_i = \sum_{k=1}^{K} n_{ki}$ is the total number of individuals at risk at time $t_i$ across all groups, and $d_i = \sum_{k=1}^{K} d_{ki}$ is the total number of events at time $t_i$ across all groups.

The test statistic for the log-rank test is then calculated as

$$s = \sum_{k=1}^{K} \frac{\left(\sum_i (d_{ki} - E_{ki})\right)^2}{\sum_i E_{ki}}. \tag{2.4}$$

We note that $d_{ki}$ can be zero, which occurs when there are no events in group $k$ at time $t_i$. This situation is handled naturally by the test, since the deviations from expectation are squared and summed across all event times.

Under the null hypothesis, the test statistic follows a chi-square distribution with degrees of freedom equal to the number of groups $K$ minus one. We write:

$$s \sim \chi^2_{(K-1)}. \tag{2.5}$$

A significant result, namely a small p-value, suggests that the null hypothesis should be rejected, indicating a difference in survival times between groups.

### 2.1.3 Nelson-Aalen Estimator

The Nelson-Aalen (NA) estimator [1] is another widely used non-parametric method. It directly estimates the cumulative hazard function, from which the survival function can then be recovered. Similar to the Kaplan-Meier estimator, the Nelson-Aalen estimator does not assume any specific underlying distribution for the time until the event occurs.

The estimate of the cumulative hazard is computed by accumulating the hazard contributions

for each event time:

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}. \tag{2.6}$$

The corresponding estimate of the survival function $\hat{S}(t)$ can be obtained by combining this result with Equation 1.8:

$$\hat{S}(t) = \exp\left(-\sum_{t_i \leq t} \frac{d_i}{n_i}\right). \tag{2.7}$$

## 2.2 Semi-Parametric Modeling

A semi-parametric model is one that has both parametric and non-parametric components. The most common semi-parametric model for survival data is Cox regression [3], which is used to explore the relationship between survival times and one or more covariates.

### 2.2.1 Cox Regression

Cox regression is based on the proportional hazards (PH) assumption, which states that the effect of the covariates on the hazard is multiplicative and constant over time. Mathematically, the hazard function for individual $i$ with covariates $\mathbf{x_i} = (x_{i1}, \ldots, x_{ip})$ is defined as

$$h(t \mid \mathbf{x_i}) = h_0(t) \exp(\beta_1 x_{i1} + \ldots + \beta_p x_{ip}) = h_0(t) \exp(\beta' \mathbf{x_i}). \tag{2.8}$$

Here, $h_0(t)$ is the baseline hazard function, which represents the hazard when all covariates are zero, and is common to all individuals as it does not depend on $\mathbf{x_i}$. Instead, the term $\exp(\beta' \mathbf{x_i})$ is a function of the covariates which is individual-specific and does not depend on time. This formulation also implies that the regressors have a linear effect on the log-hazard ratio:

$$\log(HR_0) = \log\left(\frac{h(t \mid \mathbf{x_i})}{h_0(t)}\right) = \log\left(\exp(\beta' \mathbf{x_i})\right) = \beta' \mathbf{x_i}. \tag{2.9}$$

A consequence of the PH assumption is that the hazard ratio between any two individuals $i$

and $j$ is constant over time, since it only depends on the covariates:

$$HR = \frac{h(t \mid \mathbf{x_i})}{h(t \mid \mathbf{x_j})} = \frac{h_0(t)\exp(\beta' \mathbf{x_i})}{h_0(t)\exp(\beta' \mathbf{x_j})} = \frac{\exp(\beta' \mathbf{x_i})}{\exp(\beta' \mathbf{x_j})}. \tag{2.10}$$

The violation of this assumption can lead to biased estimates, so it is important to always test the validity of the model using both graphical and statistical tests.

Coefficients $(\beta_1, \ldots, \beta_p)$ have a straightforward interpretation: a one-unit increase in the covariate $x_k$ is associated with a $\exp(\beta_k)$ increase in the hazard rate. An alternative interpretation is that a one-unit increase in $x_k$ is associated with a $\beta_k$ increase in the log-hazard ratio.

A peculiarity of the Cox model is that it does not require the baseline hazard function $h_0(t)$ to be specified. This flexibility is powerful because it eliminates the need to assume a form for the hazard function. However, the lack of a specified density makes it impossible to estimate the parameters via maximum likelihood, so we need to resort to an alternative procedure.

### 2.2.2 Partial Likelihood

In Cox regression, the coefficients are estimated via partial likelihood. The individual contributions to the likelihood are the probabilities that the individual experiencing the event at a specific time is indeed the one who had the event among all those at risk in that moment. This implicitly assumes that there is only one event occurring at each event time, and focuses on the order of events rather than their exact timing.

Consider $n$ individuals, where each individual $i$ has an observed failure time $t_i$, a censoring indicator $\delta_i$ and a covariate vector $\mathbf{x}_i$. Moreover, let $\mathscr{R}_i$ be the risk set at time $t_i$. Then, the partial likelihood is constructed as

$$\begin{aligned}
\mathscr{L}(\beta \mid D) &= \prod_{i=1}^{n} \left( \frac{h(t_i \mid \mathbf{x_i})}{\sum_{j \in \mathscr{R}_i} h(t_i \mid \mathbf{x_j})} \right)^{\delta_i} \\
&= \prod_{i=1}^{n} \left( \frac{h_0(t_i)\exp(\beta' \mathbf{x_i})}{\sum_{j \in \mathscr{R}_i} h_0(t_i)\exp(\beta' \mathbf{x_j})} \right)^{\delta_i} \\
&= \prod_{i=1}^{n} \left( \frac{\exp(\beta' \mathbf{x_i})}{\sum_{j \in \mathscr{R}_i} \exp(\beta' \mathbf{x_j})} \right)^{\delta_i},
\end{aligned} \tag{2.11}$$

where $D = (n, \mathbf{t}, X, \delta)$ represents the full data. This leads to the partial log-likelihood:

$$\log \mathscr{L}(\beta \mid D) = \sum_{i=1}^{n} \delta_i \left( \beta \mathbf{x_i} - \log \sum_{j \in \mathscr{R}_i} e^{\beta \mathbf{x_j}} \right). \tag{2.12}$$

The estimators are obtained by maximization of the partial log-likelihood with respect to $\beta$. Mathematically, this amounts to solving for the parameters that make the derivative of the partial log-likelihood equal to zero, while ensuring the second-order conditions for a maximum are met. However, as this process usually has no closed-form solutions, computational iterative algorithms such as Newton-Raphson are typically employed.

Once the regression coefficients are estimated, we can then proceed to estimate the cumulative baseline hazard function $\hat{H}_0(t)$ using a form of the Nelson-Aalen estimator that includes regressors:

$$\hat{H}_0(t) = \sum_{i:t_i \leq t} \frac{d_i}{\sum_{l \in \mathscr{R}_i} e^{\hat{\beta}' \mathbf{x_l}}}. \tag{2.13}$$

Moreover, by Equation 1.8, the following estimator for the baseline survivor function is obtained:

$$\hat{S}_0(t) = \exp\left( \sum_{i:t_i \leq t} \frac{d_i}{\sum_{l \in \mathscr{R}_i} e^{\hat{\beta} \mathbf{x_l}}} \right) \tag{2.14}$$

We then obtain the estimates for the survival function of an individual with covariates $\mathbf{x_i}$ as

$$\hat{S}_i(t) = \hat{S}_0(t)^{\exp(\hat{\beta} \mathbf{x_i})}. \tag{2.15}$$

The previously described Cox model is not fully parametric but rather semi-parametric. This is because, even after estimating the parameters, we do not have a form for the baseline hazard $h_0(t)$, meaning we still do not fully know the underlying distribution. While the strength of the Cox model lies in its flexibility, as it allows us to avoid making assumptions about the distribution, this also has its limitations. Specifically, it does not allow us to incorporate information about the distribution into the model and does not directly provide forms for the density, survival, and hazard functions, requiring us to use additional estimators. Moreover, there exist several applications in which the proportional hazards assumption does not hold true.

## 2.3 Parametric Modeling

Parametric models are those in which the outcome variable, in survival analysis time, is assumed to follow a specific family of distributions with unknown parameters. Once these parameters are estimated, the exact distribution of the outcome variable is determined, enabling precise inference thanks to the one-to-one relationship between the survival functions.

The parametric specification also allows us to fit the model via maximum likelihood estimation, as opposed to Cox regression in which we resort to partial likelihood estimation.

### 2.3.1 Proportional Hazards and Accelerated Failure Time

Parametric models can be classified into two broad categories: proportional hazards (PH) and accelerated failure time (AFT) models. The first class relies on the assumption introduced in Equations 2.8 and 2.9. The second class assumes that the effect of covariates is multiplicative with respect to survival time, implying that the ratio of survival times is constant.

A parametric proportional hazards model is similar to a Cox proportional hazards model, with the only difference that we now specify a parametric form for the baseline hazard function. Therefore, for individual $i$ with covariates $\mathbf{x_i}$, the hazard function at time $t$ is defined as in Equation 2.8, where $h_0(t)$ now has a parametric form. Nonetheless, the interpretation of the coefficients remains the same as in the Cox model, due to the fact that the hazard ratio preserves its interpretation too. Indeed, one can notice that Equation 2.10 is true no matter the form of $h_0(t)$, as long as the PH assumption is met.

The accelerated failure time model is another way of analyzing survival data with a focus on the survival time itself rather than the hazard rate. In AFT models, the survival time of subject $i$ with covariates $\mathbf{x_i}$ is assumed to be $t_i = t_0 \exp(\beta' \mathbf{x_i})$, where $t_0$ is the baseline survival time.

In terms of survival function, the model is expressed as

$$S(t) = S_0\Big(t \exp(\beta' \mathbf{x})\Big), \tag{2.16}$$

where $S_0$ is the baseline survival and $\exp(\beta'\mathbf{x})$ is referred to as the acceleration factor.

Another common way to formulate the AFT model is through the log-transformed survival time, in which case we have

$$\log(t) = \beta_0 + \beta'\mathbf{x} + \sigma\varepsilon, \tag{2.17}$$

where $\beta_0$ is the intercept, $\varepsilon$ is the error term and $\sigma$ is an unknown scale parameter.

In AFT models, a one-unit increase in $x_k$ is associated with the survival time being scaled by $\exp(\beta_k)$. Thus, AFT models provide a useful framework for understanding the direct impact of covariates on survival time, differing from the PH model which focuses on hazard rates.

### 2.3.2 Full Likelihood

The availability of the main survival functions allows us to estimate the model via maximum likelihood. For the data points in the set $D_{\text{observed}}$, the contribution is simply the density $f(t)$; for the censored data points in the set $D_{\text{censored}}$, the contribution is the probability of survival up to the censoring time, namely $S(t)$. Hence, the full likelihood is defined as

$$\mathscr{L}(\theta \mid D) = \prod_{j \in D_{\text{observed}}} f(t_j) \prod_{k \in D_{\text{censored}}} S(t_k), \tag{2.18}$$

which can be reformulated more compactly as through the use of censoring indicators $\delta_i$ as

$$\mathscr{L}(\theta \mid D) = \prod_{i=1}^{n} \Big[ S(t_i)h(t_i) \Big]^{\delta_i} S(t_i)^{1-\delta_i}. \tag{2.19}$$

Taking the logarithm of the likelihood we get the following log-likelihood:

$$\log \mathscr{L}(\theta \mid D) = \sum_{i=1}^{n} \Big[ \delta_i \log(S(t_i)h(t_i)) + (1-\delta_i) \log(S(t_i)) \Big]. \tag{2.20}$$

The maximum likelihood estimator is then obtained by maximizing the log-likelihood. As in the case for the partial log likelihood, this is achieved by setting the first derivative to zero and solving for the parameters, while checking the second order conditions. However, in most cases, numerical algorithms are employed.

### 2.3.3 The Weibull Model

We proceed to illustrate an example of the general parametric model fitting procedure by presenting the Weibull model, which is the most commonly used in parametric survival analysis.

The Weibull density function, parameterized by a shape parameter $\alpha$ and a scale parameter $\lambda$, is defined as

$$f(t \mid \alpha, \lambda) = \frac{\alpha}{\lambda} t^{\alpha-1} e^{-\frac{t^\alpha}{\lambda}} \mathbb{1}_{[0,\infty)}(t), \tag{2.21}$$

where $\alpha, \lambda > 0$. The Weibull is denoted by $\text{Weib}(\alpha, \lambda)$ and we write $T \sim \text{Weib}(\alpha, \lambda)$.

The cumulative distribution function can be obtained by integration of the density:

$$\begin{aligned} F(t \mid \alpha, \lambda) &= \int_0^t f(u \mid \alpha, \lambda)\, du \\ &= \int_0^t \frac{\alpha}{\lambda} u^{\alpha-1} e^{-\frac{u^\alpha}{\lambda}}\, du \\ &= -e^{-\frac{u^\alpha}{\lambda}} \Big|_{u=0}^{u=t} \\ &= 1 - e^{-\frac{t^\alpha}{\lambda}}. \end{aligned} \tag{2.22}$$

Once the CDF is specified, by Equation 1.2 one easily obtains the survival function as

$$S(t) = 1 - \left(1 - e^{-\frac{t^\alpha}{\lambda}}\right) = e^{-\frac{t^\alpha}{\lambda}}, \tag{2.23}$$

and by Equation 1.5 we also obtain a parametric form for the hazard:

$$h(t) = \frac{\frac{\alpha}{\lambda} t^{\alpha-1} e^{-\frac{t^\alpha}{\lambda}}}{e^{-\frac{t^\alpha}{\lambda}}} = \frac{\alpha}{\lambda} t^{\alpha-1}. \tag{2.24}$$

From Equation 2.20 the log-likelihood is then

$$\begin{aligned} \log \mathscr{L}(\alpha, \lambda | \mathscr{D}) &= \sum_{i=1}^n \left[ \delta_i \log\left( e^{-\frac{t_i^\alpha}{\lambda}} \frac{\alpha}{\lambda} t_i^{\alpha-1} \right) + (1 - \delta_i) \log\left( e^{-\frac{t_i^\alpha}{\lambda}} \right) \right] \\ &= \sum_{i=1}^n \left[ -\frac{t_i^\alpha}{\lambda} + \delta_i \left( \log(\alpha) - \log(\lambda) + (\alpha - 1) \log(t_i) \right) \right], \end{aligned} \tag{2.25}$$

which is maximized as discussed earlier. The regressors and the respective coefficients are usually introduced into the model through a reparametrization of scale parameter $\lambda$. Based on the form of the reparametrization, the Weibull model can either assume the shape of a proportional hazards or an accelerated failure time model.

The Weibull is the most common model in survival analysis, as it offers flexibility, interpretability and a relatively straightforward mathematical formulation. However, there are other several alternatives, each suitable for different assumptions about the hazard. In simple cases in which the hazard is deemed to be constant, the Exponential model is used. In more complicated settings, more complex distributions such as the Log-Normal and the Log-Logistic are used. Table 2.1 reports the survival and the hazard functions for these models. The densities can then be easily recovered through Equation 1.5.

| **Distribution** | $\mathbf{S(t)}$ | $\mathbf{h(t)}$ | **Assumption** |
|---|---|---|---|
| Exponential | $\exp(-\frac{t}{\lambda})$ | $\frac{1}{\lambda}$ | PH, AFT |
| Weibull | $\exp(-\frac{t^\alpha}{\lambda})$ | $\frac{\alpha}{\lambda}x^{\alpha-1}$ | PH, AFT |
| Log-Logistic | $\frac{1}{1+\lambda t^p}$ | $\frac{\lambda p t^{p-1}}{1+\lambda t^p}$ | AFT |
| Log-Normal | $1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)$ | $\frac{\phi\left(\frac{\log t-\mu}{\sigma}\right)}{\sigma t\left(1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)\right)}$ | AFT |

Table 2.1: Survival and hazard functions for Exponential, Weibull, Log-Logistic and Log-Normal models.

# Chapter 3

# Basic Elements of Bayesian Statistics

In this chapter, we outline the basic assumptions of the frequentist approach and compare them with the Bayesian ones. The Bayesian framework is then introduced, showing the concepts of prior and posterior through the Bayes' theorem. The chapter ends with a discussion about the importance of the choice of priors and introduces a common Monte Carlo Markov Chain method for sampling from the posterior.

## 3.1 Frequentist vs Bayesian Thinking

Frequentist and Bayesian statistics are two fundamental paradigms in statistical inference, each offering distinct approaches to interpreting data and probabilities.

Frequentist statistics interprets probability as the long-run frequency of events in repeated sampling. Model parameters are treated as fixed and unknown quantities. The core assumption is that the observed data is a realization of random variables that are independent and identically distributed (i.i.d.) from a distribution $f_\theta$, $\theta$ being the vector of parameters of interest. We write:

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_\theta, \qquad \theta \in \Theta$$

where $\Theta$ denotes the parameter space. The drawback of the assumptions in this approach is that the outcome of the first $n-1$ observations is not considered for the prediction of the $n^{\text{th}}$ draw.

On the other side, Bayesian statistics removes the strict i.i.d. assumption and replaces it with the notion of conditional independence and identity in distribution:

$$X_1, \ldots, X_n \mid \theta \overset{\text{i.i.d.}}{\sim} f(\cdot \mid \theta)$$
$$\theta \sim \pi,$$

where $\pi$ denotes a density function for the parameter. In this case, parameters are viewed as random variables rather than as fixed quantities.

In this framework, a prior distribution for $\theta$ is specified, summarizing the prior beliefs about the parameter. Then, the observed data is used to update the prior, producing a posterior distribution. This makes the method flexible and effective, particularly in those situations where updating and quantifying uncertainty is critical and prior knowledge is available.

## 3.2 The Bayesian Framework

### 3.2.1 Bayes' Theorem

At the core of Bayesian statistics is the assumption that the vector of unknown parameters $\theta$ is random, hence it is defined by a density function. We let $\pi(\theta)$ be the prior distribution, which reflects our knowledge about the parameters before the data is observed. Moreover, we let $\mathscr{L}(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$ be the likelihood, which specifies the probability of observing the data given the parameters. The posterior distribution of $\theta$, which reflects an updated belief about the parameters based on the observed data, is then obtained through Bayes' theorem:

$$\pi(\theta \mid x_1, \ldots, x_n) = \frac{\mathscr{L}(x_1, \ldots, x_n \mid \theta)\pi(\theta)}{\int_\Theta \mathscr{L}(x_1, \ldots, x_n \mid \theta)\pi(\theta)\, d\theta}. \tag{3.1}$$

One can note that the denominator $m(x_1, \ldots, x_n) = \int_\Theta \mathscr{L}(x_1, \ldots, x_n \mid \theta)\pi(\theta)\, d\theta$, which is the marginal distribution of the data, does not depend on the parameter vector $\theta$. This implies that

the posterior distribution is proportional to the product of the likelihood and the prior. We write:

$$\pi(\theta \mid x_1, \ldots, x_n) \propto \mathscr{L}(x_1, \ldots, x_n \mid \theta)\pi(\theta), \tag{3.2}$$

where the proportionality constant is the reciprocal of the marginal distribution of the data, also known as the normalizing constant. The normalizing constant ensures that the posterior distribution integrates to one, making it a valid probability distribution.

In most instances, evaluating $m(x_1, \ldots, x_n)$ can be challenging, which means that $\pi(\theta \mid x_1, \ldots, x_n)$ may not have an explicit form. In such cases, we generally resort to computational methods such as Monte Carlo Markov Chains (MCMC) to sample from the posterior, as it will be briefly discussed in Section 3.3.

Once the posterior is available, we can then obtain the posterior predictive distribution for a future observation $x_{n+1}$ as

$$\pi(x_{n+1} \mid x_1 \ldots, x_n) = \int_\Theta f(x_{n+1} \mid \theta)\pi(\theta \mid x_1, \ldots, x_n)\, d\theta \tag{3.3}$$

We can generate predictions about future observations thanks to the posterior predictive distribution. This is important not only for predictive purposes, but also for model validation. In fact, we can evaluate the model's fit and appropriateness by comparing the data produced by the posterior predictive distribution with the actual observed data. Any differences may point to the need for model improvement.

### 3.2.2 The Importance of Priors

When prior knowledge is limited, non-informative priors are typically employed. These aim to have minimal influence on the posterior distribution, allowing the data to drive the inference process and the belief update. An example of a non-informative prior is a uniform distribution, implying that all values for the parameter are equally likely a priori.

On the other hand, if substantial prior knowledge is available, informative priors are used. An example of an informative prior is a low-variance normal distribution centered around the

value which we believe the parameter to be close to.

A common way to incorporate prior knowledge into the model is through the power prior, which directly exploits historical data $D_0$. The power prior is defined as

$$\pi(\theta \mid D_0, a_0) \propto \mathscr{L}(D_0 \mid \theta)^{a_0} \pi_0(\theta) \tag{3.4}$$

where $\mathscr{L}(D_0 \mid \theta)$ is the likelihood based on $D_0$ and $\pi_0(\theta)$ is the initial prior. The parameter $a_0$ ranges from 0 to 1 and controls the influence of historical data on the current study. The influence of past data on the prior is null when $a_0 = 0$ and increases as $a_0$ approaches one. Eventually, when $a_0 = 1$, the prior of the current study is actually the posterior of a past study.

A convenient characteristic of certain priors is that they lead to a posterior that belongs to the same family of distributions. This property simplifies the computation of the posterior, making these priors particularly useful when dealing with computationally intensive calculations. These priors are referred to as conjugate priors.

In summary, the choice of prior is critical as it affects the posterior distribution, influences the inference drawn from the model, and impacts computational efficiency. Carefully selecting an appropriate prior ensures robust and reliable Bayesian analysis.

## 3.3 Sampling from the Posterior

As previously mentioned, in most cases the posterior distribution does not have a closed form due to the fact that the normalizing constant is hard to evaluate. This represents an important problem, as it makes direct sampling from the posterior distribution infeasible. Numerical methods such as Markov Chain Monte Carlo (MCMC) are generally employed to overcome this issue, as they are able to generate samples that approximate the posterior distribution.

One of the most widely used MCMC methods is Gibbs sampling [8], which is particularly useful when direct sampling from the joint distribution is difficult, but sampling from the conditional distributions is more practical. Gibbs sampling iteratively samples each parameter from its conditional distribution, given the most recent values of the other parameters:

- **Step 0**: Choose an arbitrary starting value $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)})$, and set $i = 0$.
- **Step 1**: At iteration $i+1$, generate $\theta^{(i+1)} = (\theta_1^{(i+1)}, \theta_2^{(i+1)}, \ldots, \theta_p^{(i+1)})$ as follows:
  - Draw $\theta_1^{(i+1)} \sim \pi(\theta_1 \mid \theta_2^{(i)}, \theta_3^{(i)}, \ldots, \theta_p^{(i)}, x_1, x_2, \ldots, x_n)$
  - Draw $\theta_2^{(i+1)} \sim \pi(\theta_2 \mid \theta_1^{(i+1)}, \theta_3^{(i)}, \ldots, \theta_p^{(i)}, x_1, x_2, \ldots, x_n)$
  - $\vdots$
  - Draw $\theta_k^{(i+1)} \sim \pi(\theta_k \mid \theta_1^{(i+1)}, \theta_2^{(i+1)}, \ldots, \theta_{k-1}^{(i+1)}, \theta_{k+1}^{(i)}, \ldots, \theta_p^{(i)}, x_1, x_2, \ldots, x_n)$
  - $\vdots$
  - Draw $\theta_p^{(i+1)} \sim \pi(\theta_p \mid \theta_1^{(i+1)}, \theta_2^{(i+1)}, \ldots, \theta_{p-1}^{(i+1)}, x_1, x_2, \ldots, x_n)$
- **Step 2**: Set $i = i+1$, and go to Step 1.

This process is repeated for a finite number of iterations $S$. It can be proven that, under certain conditions discussed in [8], the estimated vector sequence $\{\theta^{(i)},\ i = 1, \ldots, S\}$ has a stationary distribution $\pi(\theta \mid x_1, \ldots, x_n)$.

In practice, it is common to discard a certain number of initial samples generated by the Gibbs sampler. This is done because these samples may not be representative of the target distribution, as the chain might not have converged to the stationary distribution yet. This initial set of samples is known as the burn-in. The remaining samples after the burn-in are then used as an approximation of the actual posterior realizations.

# Chapter 4

# Dirichlet Process Mixture Models

This chapter offers a presentation of the Dirichlet distribution and of Dirichlet processes, together with their properties. The knowledge of these is necessary to introduce the concept of Dirichlet process mixture models, a class of non-parametric models widely used in Bayesian statistics. A formulation of Dirichlet process mixture models for survival analysis is then discussed.

## 4.1 Dirichlet Distribution

### 4.1.1 Definition

The Dirichlet distribution is a fundamental concept in probability and statistics, particularly useful for modeling probabilities of outcomes in multinomial experiments. It is defined on the $(k-1)$-dimensional probability simplex $\Delta_k$, which is the set of $k$-dimensional vectors where each component is non-negative and the components sum to one. Mathematically:

$$\Delta_k = \left\{ \mathbf{q} \in \mathbb{R}^k \mid q_i \geq 0 \ \forall i, \quad \sum_{i=1}^{k} q_i = 1 \right\}. \tag{4.1}$$

Let $\mathbf{q} = (q_1, \ldots, q_k)$ be a random vector on this simplex, hence $q_i \geq 0 \ \ \forall i$ , $\sum_{i=1}^{k-1} q_i \leq 1$ and $q_k = 1 - (\sum_{i=1}^{k-1} q_i)$. Moreover, let $\nu = (\nu_1, \ldots, \nu_k)$ be a vector of positive real numbers. The

Dirichlet probability density function, parameterized by $\nu$ and defined over $\mathbf{q}$, is

$$f(\mathbf{q};\nu) = \frac{\Gamma(\sum_{i=1}^{k}\nu_i)}{\prod_{i=1}^{k}\Gamma(\nu_i)}\prod_{i=1}^{k}q_i^{\nu_i-1}. \tag{4.2}$$

We denote this distribution by $\mathrm{Dir}(\nu)$ and write $\mathbf{q} \sim \mathrm{Dir}(\nu)$.

The Dirichlet can be seen as a distribution over the set of $k$-dimensional discrete distributions, with the parameter vector $\nu$ shaping its behavior. Each component $\nu_i$ determines the mean of the corresponding component $q_i$, with $\mathbb{E}[q_i] = \frac{\nu_i}{\sum_{j=1}^{k}\nu_j}$. If all $\nu_i > 1$, the distribution is concentrated around the mean, reflecting greater certainty. If all $\nu_i < 1$, the distribution spreads out, showing higher uncertainty with more weight on extreme values. When $\nu_i = 1$ for all $i$, the distribution is uniform, implying no preference for any outcome.

The Dirichlet is a generalization of the Beta distribution to multiple dimensions. Indeed, letting $k = 2$ and $q_2 = 1 - q_1$, one obtains the density of the latter:

$$f(\mathbf{q};\nu) = \frac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)}q_1^{\nu_1-1}(1-q_1)^{\nu_2-1}\mathbb{1}_{(0,1)}(q_1), \tag{4.3}$$

The Dirichlet distribution is particularly useful in Bayesian analysis of categorical data, as it is the conjugate prior for the multinomial distribution. This means that if the prior distribution is a Dirichlet, the posterior distribution remains a Dirichlet, though with different parameters. To understand this property, we first introduce the multinomial probability mass function, parameterized by an integer $n$ and a probability mass function $\mathbf{q} = (q_1,\ldots,q_k)$, defined as

$$f(x_1,x_2,\ldots,x_k \mid n,\mathbf{q}) = \frac{n!}{x_1!x_2!\cdots x_k!}\prod_{i=1}^{k}q_i^{x_i}. \tag{4.4}$$

Now, consider a Bayesian model where the probabilities $\mathbf{q}$ are assigned a Dirichlet prior:

$$\begin{aligned}\mathbf{q} = (q_1,\ldots,q_k) &\sim \mathrm{Dir}(\nu),\\ (X_1,\ldots,X_n) \mid \mathbf{q} &\sim \mathrm{Mult}_k(n,\mathbf{q}).\end{aligned} \tag{4.5}$$

Applying Bayes' theorem, we obtain the posterior distribution of $\mathbf{q}$ as

$$\begin{aligned}\pi(\mathbf{q} \mid x_1, x_2, \ldots, x_k) &\propto \left( \frac{n!}{x_1! x_2! \cdots x_k!} \prod_{i=1}^{k} q_i^{x_i} \right) \left( \frac{\Gamma\left(\sum_{i=1}^{k} \nu_i\right)}{\prod_{i=1}^{k} \Gamma(\nu_i)} \prod_{i=1}^{k} q_i^{\nu_i - 1} \right) \\ &\propto \prod_{i=1}^{k} q_i^{\nu_i + x_i - 1}. \end{aligned} \tag{4.6}$$

The right-hand side above represents the kernel of a Dirichlet distribution, thus we have

$$\pi(\mathbf{q} \mid x_1, \ldots, x_k) = \frac{\Gamma\left(\sum_{i=1}^{k} (\nu_i + x_i)\right)}{\prod_{i=1}^{k} \Gamma(\nu_i + x_i)} \prod_{i=1}^{k} q_i^{\nu_i + x_i - 1}, \tag{4.7}$$

indicating that

$$\mathbf{q} \mid x_1, \ldots, x_k \sim \mathrm{Dir}(\nu + \mathbf{x}). \tag{4.8}$$

### 4.1.2 Sampling

The stick-breaking process provides a convenient method for sampling from the Dirichlet distribution, leveraging the ability to simulate samples from a Beta distribution. This process iteratively divides a unit-length stick into $k$ segments, where the length of each segment follows the Dirichlet distribution. The procedure is outlined as follows:

1. **Step 0**: The first segment $q_1$ is generated as $q_1 = b_1$ with $b_1 \sim \mathrm{Beta}(\nu_1, \sum_{i=2}^{k} \nu_i)$.

2. **Step 1**: For $j = 2, \ldots, k-1$, given that $j-1$ segments of the stick have already been created, the remaining length is $l_j = \prod_{i=1}^{j-1}(1-u_i)$. Simulate $b_j \sim \mathrm{Beta}(\nu_j, \sum_{i=j+1}^{k} \nu_i)$ and let $q_j = b_j l_j$.

3. **Step 2**: After $k-1$ iterations, the remaining length of the stick is $l_k = \prod_{i=1}^{k-1}(1-u_i)$, and we set $q_k = l_k$.

This process results in $k$ segments of the original unit-length stick, with lengths $(q_1, \ldots, q_k)$. It can be demonstrated that these lengths follow the Dirichlet distribution $\mathrm{Dir}(\nu_1, \ldots, \nu_k)$.

## 4.2 Dirichlet Process

### 4.2.1 Definition

Having established a solid understanding of the Dirichlet distribution and its properties, we now extend our exploration to the Dirichlet process (DP). The Dirichlet process generalizes the concept of the Dirichlet distribution to an infinite-dimensional setting, allowing us to model distributions over potentially infinite categories.

To proceed with the definition of the Dirichlet process, we introduce distribution $G_0$ defined over the measurable space $\Omega$, and a positive real number $\kappa$. We say that the random distribution $G$ is a Dirichlet process [6] with base distribution $G_0$ and concentration parameter $\kappa$ if

$$(G(B_1),\ldots,G(B_r)) \sim \mathrm{Dir}(\kappa G_0(B_1),\ldots,\kappa G_0(B_r)), \tag{4.9}$$

for any finite measurable partition $\{B_i\}_{i=1}^r$ of $\Omega$, and we write $G \sim DP(\kappa, G_0)$. The base distribution $G_0$ acts as the mean of the Dirichlet process and represents any prior knowledge about the distribution $G$, as $E[G \mid \kappa, G_0] = G_0$. The concentration parameter $\kappa$ controls the variance of $G$ around the base distribution $G_0$: the larger the $\kappa$, the smaller the variance, and as $\kappa \to \infty$, $G$ converges to $G_0$.

An important property of the Dirichlet process is its conjugacy. Let the observed data be the realization of a random distribution, on which we put a Dirichlet process prior. Mathematically:

$$\begin{aligned} X_1,\ldots,X_n \mid G &\overset{\text{i.i.d.}}{\sim} G \\ G &\sim DP(\kappa, G_0). \end{aligned} \tag{4.10}$$

Then, the posterior distribution given observed data remains a Dirichlet process. Specifically, given a measurable partition $\{B_i\}_{i=1}^r$ of $\Omega$ and letting $n_k$ be the number of observations in $B_k$, we have

$$(G(B_1),\ldots,G(B_r)) \sim \mathrm{Dir}(\kappa G_0(B_1)+n_1,\ldots,\kappa G_0(B_r)+n_r). \tag{4.11}$$

It follows that

$$G \mid x_1, \ldots, x_n \sim DP\left(\kappa + n, \frac{\kappa G_0 + \sum_{i=1}^{n} \delta_{x_i}}{\kappa + n}\right), \tag{4.12}$$

where $\delta_{x_i}$ is the Dirac measure at $x_i$, defined as $\delta_{x_i}(B) = 1$ if $x_i \in B$, and 0 otherwise. The updated base measure of the posterior distribution can also be reformulated as

$$E[G \mid x_1, \ldots, x_n] = \frac{\kappa}{\kappa + n} G_0 + \frac{n}{\kappa + n} \frac{\sum_{i=1}^{n} \delta_{x_i}}{n}, \tag{4.13}$$

which provides an intuitive explanation of the role of the concentration parameter, related to the fact that the updated base measure is a weighted average of the prior base measure and the empirical distribution of the data. Specifically, a larger $\kappa$ places more weight on $G_0$, leading to a posterior distribution closer to $G_0$, while a smaller $\kappa$ allows the data to have more influence on the posterior distribution, resulting in greater deviation from $G_0$ to better fit the observed data.

The posterior base distribution is also the predictive distribution for $x_{n+1}$. Indeed:

$$f(x_{n+1} \mid x_1, \ldots, x_n) = \int G(x_{n+1}) \pi(G \mid x_1, \ldots, x_n) dG = E[G \mid x_1, \ldots, x_n], \tag{4.14}$$

hence, by the properties of the Dirichlet process:

$$f(x_{n+1} \mid x_1, \ldots, x_n) = \frac{\kappa}{\kappa + n} G_0 + \frac{n}{\kappa + n} \frac{\sum_{i=1}^{n} \delta_{x_i}}{n}. \tag{4.15}$$

### 4.2.2 Sampling

The formulation of the posterior distribution presented in Equation 4.15 underlies a method for sampling from a Dirichlet process known as the Pólya urn, or Chinese restaurant, process [2].

Let $\Omega$ be a continuum of colors, and let $X \sim G$ be balls whose color corresponds to the drawn value. We start with an empty urn and pick the first ball, whose color is drawn from the base distribution $G_0$, i.e., $X_1 \sim G_0$. For each subsequent ball, the process is as follows: with probability $\frac{\kappa}{\kappa+n}$, the color of the new ball is a new color drawn from $G_0$; with probability $\frac{n}{\kappa+n}$, the color of the new ball is the same as the color of a randomly selected ball from the urn,

representing the empirical distribution of colors that are already present in the urn. If we repeat this process, we obtain the random sequence $(X_1, X_2, \ldots)$ with distribution:

$$\begin{aligned} X_1 &\sim G_0 \\ X_{n+1} \mid X_1, \ldots, X_n &\sim \frac{\kappa}{\kappa + n} G_0 + \frac{n}{\kappa + n} \frac{\sum_{i=1}^{n} \delta_{x_i}}{n}. \end{aligned} \tag{4.16}$$

An important property of the Dirichlet process, which is evident from the Pólya urn process, is that draws from $G$, with $G \sim \mathrm{DP}(\kappa, G_0)$, will be point masses regardless of the smoothness of $G_0$. This implies that realizations from a Dirichlet process are discrete distributions. This issue can be mitigated by introducing a continuous kernel, as we shall see in Section 4.3.

Understanding the Dirichlet process as a distribution whose realizations are discrete random distributions paves the way for another widely used sampling method, which exploits the stick-breaking process. Mathematically, we say that draws from a Dirichlet process can be represented as a weighted sum of point masses with probability one [18] and we write:

$$G = \sum_{k=1}^{\infty} w_k \delta_{\phi_k}, \tag{4.17}$$

where $w_k$ such that $\sum_{k=1}^{\infty} w_k = 1$ are the weights, $\delta_{\phi_k}$ is the Dirac measure at $\phi_k$, and $\phi_k$ are drawn from the base measure $G_0$. The weights are constructed using by stick-breaking, leading to the model:

$$\begin{aligned} \beta_k &\overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(1, \kappa) \\ \phi_k &\overset{\text{i.i.d.}}{\sim} G_0 \\ G &= \sum_{k=1}^{\infty} \left[ \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \right] \delta_{\phi_k}. \end{aligned} \tag{4.18}$$

For this construction to represent a Dirichlet process, an infinite number of atoms and weights must be drawn. However, because it is not feasible to simulate an infinite series directly, truncation is commonly introduced as a practical device. By truncating the summation at a parameter

$K$, it is possible to obtain a finite formulation for $G$ which is computationally manageable and still approximates the Dirichlet process closely.

## 4.3 Dirichlet Process Mixture Model

### 4.3.1 Definition

Probability distributions sampled from a Dirichlet process are discrete. While this property facilitates convenient sampling, it implies that the Dirichlet process is not suitable as a prior for modeling continuous distributions. This issue is addressed by the Dirichlet process mixture model (DPMM).

In the DPMM, the generative process is described as follows: data points $x_i$ are generated from a density function $k(x_i \mid \theta_i)$, hence $\int k(x \mid \theta)dx = 1$, where $k(\cdot \mid \cdot)$ represents a parametric kernel, usually continuous. Each data point $x_i$ is associated with some latent parameter vector $\theta_i$, which is independently and identically distributed according to $G$, conditionally on $G$. The distribution $G$ is assigned a Dirichlet process prior with concentration parameter $\kappa$ and base distribution $G_0$. Formally, this hierarchical model is expressed as

$$\begin{aligned} x_i \mid \theta_i &\overset{\text{ind.}}{\sim} k(x_i \mid \theta_i), \quad i = 1, \ldots, n \\ \theta_i \mid G &\overset{\text{i.i.d.}}{\sim} G, \quad i = 1, \ldots, n \\ G &\sim \mathrm{DP}(\kappa, G_0). \end{aligned} \tag{4.19}$$

To better understand the overall structure of the mixture model, we can reformulate it in terms of a mixing distribution. Instead of focusing on the individual parameters $\theta_i$ for each data point, we consider the DP mixture distribution $F$, which describes the distribution over the data points directly:

$$F = \int_{\Theta} K(\cdot \mid \theta) G(d\theta). \tag{4.20}$$

Here, $G \sim \mathrm{DP}(\kappa, G_0)$ and $K(\cdot \mid \theta)$ is the cumulative distribution corresponding to the kernel, i.e.

$K(x \mid \theta) = \int_{-\infty}^{x} k(s \mid \theta) ds$. This formulation captures the concept that $F$ is a mixture distribution created by integrating the kernel distribution $K(\cdot \mid \theta)$ over all possible values of $\theta$ according to the distribution $G$. This model can then be reformulated in terms of a density function as

$$f(x) = \int_{\Theta} k(x \mid \theta) G(d\theta). \tag{4.21}$$

The observed data is then described as a realization of the density of the mixing distribution, that is

$$X_1, \dots X_n \overset{\text{i.i.d.}}{\sim} f. \tag{4.22}$$

In practical terms, this implies that the data points $x_i$ are generated from a mixture model where the mixing distribution $G$ is derived from a Dirichlet process. This perspective shows that the Dirichlet process mixture models are a generalization of finite mixture models, as they allow for an infinite number of potential mixture components. Such flexibility makes Dirichlet process mixture models particularly useful in non-parametric statistics, where the number of mixture components is not fixed but instead inferred from the data.

### 4.3.2 DPMM in Survival Analysis

The definition of a Dirichlet process mixture model begins with the choice of the kernel. In survival analysis, the observed data belongs to $\mathbb{R}^+$. This consideration, combined with the fact that we want to build a model able to approximate both monotone and non-monotone hazards, suggests that good choices for the kernel are the Gamma and the Weibull distributions. Following the suggestions in [13], we use a Weibull kernel, whose density is defined in Equation 2.21, as mixing both on the scale parameter $\alpha$ and the shape parameter $\lambda$ allows to practically approximate any function on $\mathbb{R}^+$.

Following [5], we place a Gamma prior on $\kappa$ to facilitate sampling. Mathematically, we write $\kappa \sim \text{Ga}(a_\kappa, b_\kappa)$, where $a_\kappa, b_\kappa > 0$ and the density is defined as

$$f(\kappa \mid a_\kappa, b_\kappa) = \frac{b_\kappa^{a_\kappa}}{\Gamma(a_\kappa)} \kappa^{a_\kappa - 1} e^{-b_\kappa \kappa} \mathbb{1}_{(0,\infty)}(\kappa). \tag{4.23}$$

As for the base measure, we follow the form proposed in [19]:

$$G_0(\alpha,\lambda) \propto \text{Ga}(\alpha \mid a_\alpha, b_\alpha)\mathbb{1}_{(f(\lambda),\infty)}(\alpha)\text{Ga}(\lambda \mid a_\lambda, b_\lambda) \tag{4.24}$$

which means that we place a bivariate prior on $(\alpha,\lambda)$ made by the product of two Gamma distributions, the first being parametrized by $a_\alpha$ and $b_\alpha$ while the second by $a_\lambda$ and $b_\lambda$. The role of $f(\lambda)$ in the base distribution is to keep both $\alpha$ and $\lambda$ away from values close to zero to ensure well behavior for the mixture model.

Higher model flexibility is obtained assuming that $b_\lambda$ is a random variable itself, on which we put a prior. Specifically, we let $b_\lambda \sim \text{Ga}(b_\lambda \mid a_{b_\lambda}, b_{b_\lambda})$.

Assuming that the rest of the parameters for the priors are constant, we obtain the following full Bayesian hierarchical model:

$$\begin{aligned}
t_i \mid \alpha_i,\lambda_i &\overset{\text{ind.}}{\sim} \text{Weib}(t_i \mid \alpha_i,\lambda_i), \quad i = 1,\dots,n \\
(\alpha_i,\lambda_i) \mid G &\overset{\text{i.i.d.}}{\sim} G, \quad i = 1,\dots,n \\
G \mid \nu, b_\lambda &\sim DP(\kappa, G_0) \\
\kappa &\sim \text{Ga}(\kappa \mid a_\kappa, b_\kappa) \\
b_\lambda &\sim \text{Ga}(b_\lambda \mid a_{b_\lambda}, b_{b_\lambda}).
\end{aligned} \tag{4.25}$$

The choice of prior hyperparameters is usually a data-dependent process, which is especially delicate when limited knowledge is available. Although it is true that informative values might enhance inference, it is also true that they might introduce wrong information and bias the model estimates. On the other side, using completely uninformative priors is also not optimal, as the possibility to introduce information into the model is not exploited. In our implementation we follow the Low Information Omnibus (LIO) prior discussed in [19], which aims to find a balance between the two extremes.

In the LIO model specification, a transformation is applied to the observed data. Taking $c$ be the 95-th percentile of $t_1,\dots t_n$, we let $z_i = \frac{10y_i}{c}$ and we use the mixture of Weibulls to model

$z_1, \ldots, z_n$. Inference in the initial scale can then be recovered as $t_i \mid \alpha_i, \lambda_i \sim \text{Weib}(\alpha_i, \lambda_i(\frac{10}{c})^{\alpha_i})$.

The purpose of the transformation is to scale the data so that we are able to specify the hyperparameters as well as the form for $f(\lambda)$ in a way that is not dependent on the data. Specifically, following the assumption that after the transformation most of the observations will be smaller than 25, we have

$$f(\lambda) = \max\left(0, \frac{\log(\log(20)/\lambda)}{\log(25)}\right). \tag{4.26}$$

As for the specific hyperparameters, these are chosen to provide a wide variety of components in the DPMM, so that the model generates flexible distributions for the transformed data.

# Chapter 5

# Illustration with Real Data

In this section, we apply the previously discussed model in a real world scenario. We will work on a dataset taken from [21], which contains data about lung cancer treatments. We will first focus on frequentist approaches, fitting the Kaplan-Meier estimator and then the Cox PH Regression and AFT Weibull models. We will then switch to a Bayesian approach, applying a DPM of Weibulls.

The software used in this chapter is available in the form of R code in this GitHub repository.

## 5.1 Lung Cancer Data Description

The available dataset contains information about a lung cancer study. Patients are divided into two groups, arm A and arm B, each receiving different treatments. Time to death, which is measured in days, is recorded. Information about the age of the patient is also available.

At the time in which the data was collected, it was common practice to treat lung cancers by providing patients with etoposide, followed by cisplatin (EP), which are two common chemotherapy drugs. Patients who receive this treatment are part of group B, which is the control group. On the other side, patients from Group A received the inverted combination of drugs, namely cisplatin followed by etoposide (CE); this group is considered to be the treatment. Group A contains 62 observations, of which 15 censored, while group B contains 59 individuals, of which 9 censored.

## 5.2 Models

### 5.2.1 KM Estimator

We start the analysis by fitting a Kaplan-Meier estimator to the data. This is done both on the entire dataset with no distinction among groups, and on the two groups separately to observe potential differences among the survival curves. Figure 5.1 shows the results, including confidence intervals for the estimators.

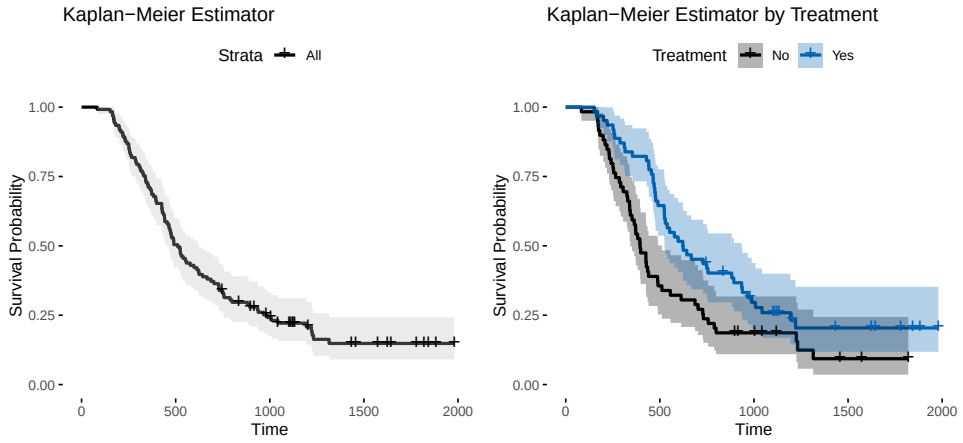


Figure 5.1: On the left, Kaplan-Meier estimator fitted on the whole dataset. On the right, Kaplan-Meier estimator fitted on control and treatment group separately.

It is clear that the estimated survival function for the treatment group is higher than the one for the control group. However, we have an overlap in the confidence intervals for the two different survival curve estimates. This is because the confidence intervals are significantly wide, especially for higher times where fewer individuals are present in the study.

In order to validate the efficacy of the treatment, we apply a log-rank test. The results prove a statistically significant difference in survival time between the two groups, as the p-value equals 0.008. This confirms that the treatment has a positive effect on survival, supporting the evidence provided by the plots in Figure 5.1.

Moreover, we also note the interesting fact that the survival curves do not approach zero as time increases. Although this might seem in contrast with the theoretical definition of survival function, it is actually common in real world scenarios and it is due to structure of the estimator.

In fact, failure times will not be observed for all individuals, as some of them will eventually recover and drop the study. These censored observations result in the estimated survival function not converging to 0 as $t \to \infty$.

### 5.2.2 Cox PH Regression

The Kaplan-Meier estimator combined with the log-rank test is a powerful tool that allowed us to conclude that the treatment has a positive and statistically significant impact on survival. However, the estimator does not provide the possibility to quantify such impact. For this purpose, we proceed by fitting a Cox proportional hazard model.

Although our variable of interest is treatment, variable age is also included in the model to control for potential confounding effects. Table 5.1 shows the regression results.

| **covariate** | **coef** | **exp(coef)** | **SE** | **z** | **p** | **95% CI for exp(coef)** |
|---|---|---|---|---|---|---|
| Treatment | -0.5140 | 0.5981 | 0.2041 | -2.519 | 0.0118* | (0.4009, 0.8922) |
| Age | 0.0280 | 1.0284 | 0.0129 | 2.181 | 0.0292* | (1.0028, 1.0547) |

Table 5.1: Cox proportional hazards model results, including treatment and age as covariates. Significance codes: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Again, the results of the Cox proportional hazards model show that the treatment has a statistically significant positive effect. Being the treatment a categorical (binary) variable, the coefficient has a very intuitive explanation: let $x$ be the regressor for an individual in the control group and let $x^*$ be the regressor for an individual in the treatment group; then $x = 1$ and $x^* = 0$. Assuming that the two individuals have the same age, from Equation 2.10 we have

$$HR = \frac{h(t|\mathbf{x})}{h(t|\mathbf{x}^*)} = \exp(\beta'\mathbf{x} - \beta'\mathbf{x}^*) = \exp(\beta'\mathbf{x}). \tag{5.1}$$

Therefore, we can conclude that the hazard ratio is equal to $\exp(-0.5140) = 0.5981$. This suggests that patients in the treatment group have roughly a 40.19% lower hazard compared to the patients in the control group. This result is statistically significant, as evidenced by the 0.0118 p-value for the coefficient. Moreover, the 95% confidence interval for the hazard ratio

does not include 1, further confirming the significance of the treatment effect.

Testing the proportional hazards assumption is crucial for validating the Cox model's results. Indeed, violation of this assumption can lead to biased estimates and wrong conclusions. We proceed to validating the assumption by employing the Schoenfeld residuals test [17], whose results are reported in Table 5.2.

| **covariate** | **chi-square** | **degrees of freedom** | **p** |
|---|---|---|---|
| Treatment | 4.144 | 1 | 0.042* |
| Age | 0.441 | 1 | 0.507 |
| Global | 4.495 | 2 | 0.106 |

Table 5.2: Results of the Schoenfeld residuals test for the proportional hazards assumption. Significance codes: $^{***}\, p < 0.001$; $^{**}\, p < 0.01$; $^{*}\, p < 0.05$

The global test has a p-value of 0.106, which is not statistically significant at the confidence level of 5%. This suggests that the proportional hazards assumption holds for the overall model. However, the PH assumption does not appear to hold for the covariate treatment, as indicated by the p-value of 0.042. This implies the impact of the treatment on the risk of death due to lung cancer is not constant over time. Following a similar reasoning, we instead conclude that the assumption holds for the covariate age. To further investigate these findings, we analyze the Schoenfeld residuals reported in Figure 5.2.

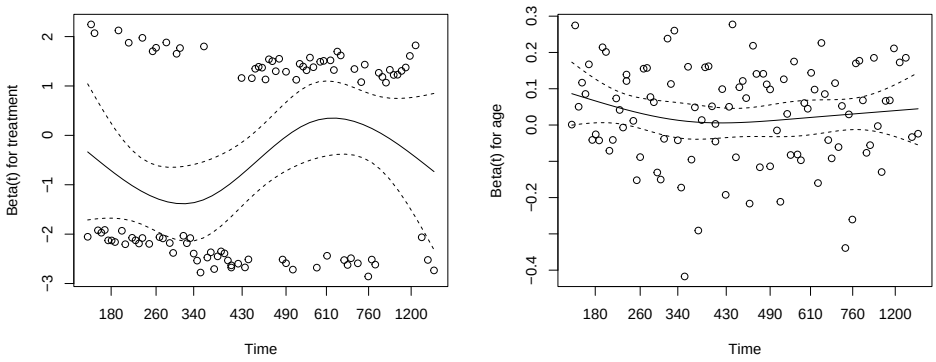


Figure 5.2: Schoenfeld residuals for treatment (left) and age (right) covariates. Points are residuals at different times, solid lines are smoothed estimates, and dashed lines are confidence intervals.

The plots reveal that the treatment line is not horizontal, which indicates that treatment effects are varying with time, i.e. there is a violation of the PH assumption. On the other side, the plot for age displays a horizontal smoothed line, meaning that the assumption is satisfied. These results are in line with the test's outcomes.

In summary, although the overall model satisfies the PH assumption, the treatment covariate does not. This suggests that the impact of treatment is not constant over time and requires to implement some adjustments to the model.

### 5.2.3 AFT Weibull

When the proportional hazards assumption does not hold, a commonly used solution is to adjust for time-dependent covariate. However, we proceed in an alternative way, fitting an accelerated failure time model under a Weibull distribution. Indeed, we believe it's plausible that the treatment effect acts on the survival rather than on the hazard. The results from a Weibull model with the AFT specification are reported in Table 5.3.

| covariate | coef | exp(coef) | SE | z | p | 95% CI for exp(coef) |
|---|---|---|---|---|---|---|
| Treatment | 0.3993 | 1.4908 | 0.14308 | 2.79 | 0.0053** | (1.1225, 1.9812) |
| Age | -0.0203 | 0.9798 | 0.00906 | -2.25 | 0.0244* | (0.9613, 0.9987) |
| Log(scale) | -0.3493 | - | 0.07904 | -4.42 | 9.9e-06*** | - |

Table 5.3: Weibull AFT regression results, including treatment and age as covariates. Significance codes: $^{***}\, p < 0.001$; $^{**}\, p < 0.01$; $^{*}\, p < 0.05$

The results show that treatment has a statistically significant effect on survival at a 5% significance level, as the p-value is 0.0053. The coefficient is 0.3993, which implies that the acceleration factor is $\exp(0.3993) = 1.4908$. This suggests that patients in the treatment group have a survival time approximately 49.08% longer compared to those in the control group. The 95% confidence interval for the acceleration factor $(1.1225, 1.9812)$ which does not include 1, further indicates the positive effect of treatment.

Fitting a parametric model has the advantage that, once the parameters are estimated, forms for the other descriptive survival functions readily become available. Figure 5.3 reports a

smoothed visualization of the estimated survival, cumulative distribution, hazard and density.

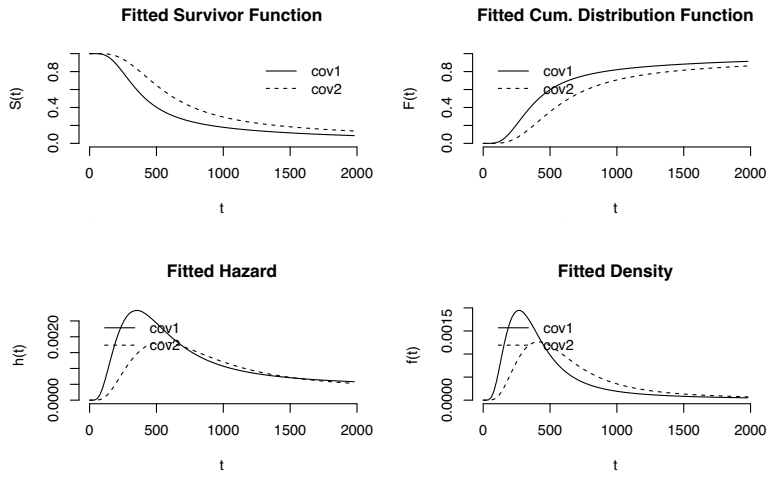


Figure 5.3: Smoothed survival, cumulative density, hazard and density function for the fitted Weibull model. Control is represented by cov1 and treatment by cov2.

### 5.2.4 DPM of Weibulls

Fitting a parametric model such as an AFT Weibull allows for powerful inference as all descriptive survival functions become available once the hazard is specified and the parameters estimated. However, as previously discussed, this method relies on the strong assumption that data can be described by a specific class of distributions, which is not always the case. This assumption can be removed by fitting a semi-parametric model such as Cox regression, but then we must resort to estimators for the hazard and survival functions, introducing another level of complexity and uncertainty into the model. Therefore, there is usually a tradeoff between the depth of assumptions and the flexibility of the model.

Fitting a Dirichlet process mixture model with Weibull kernels on survival data allows us to strike a balance between these two forces. Indeed, as it is a non-parametric model, no assumption is made on the distribution on the data, which makes is extremely flexible. We recall, in fact, that a DPMM can be thought of as a generalization of mixture models that allows for infinite components. At the same time, however, there exists methods such as the ones discussed in [13] to obtain draws from the posterior of the most important functionals.

We proceed to fit a DPMM on the lung cancer data, following the specification presented in Section 4.3.2. The model is fitted on group A and group B separately to make inference about the effectiveness of the treatment. The estimated survival and hazard functions are reported in Figure 5.4.

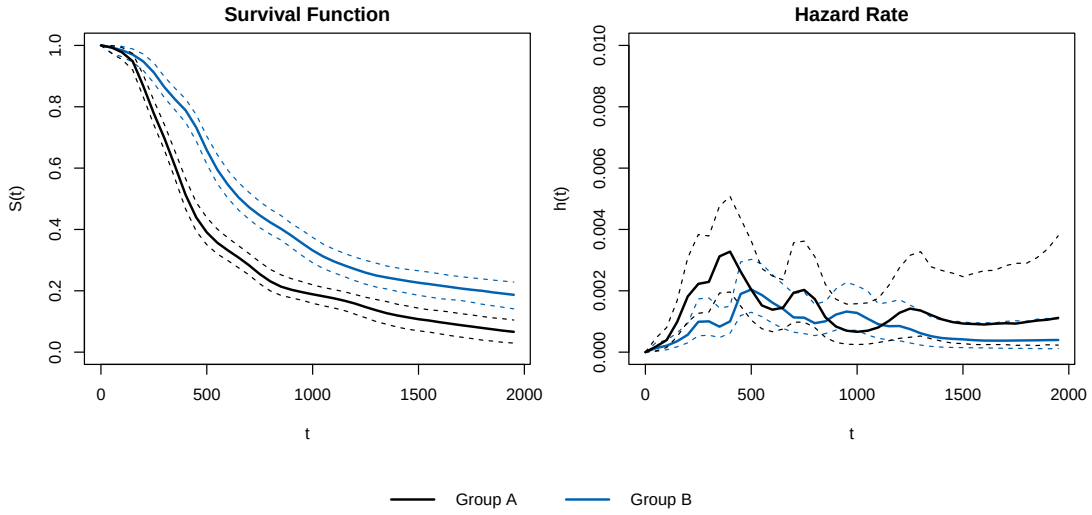


Figure 5.4: Posterior point estimates (solid line) and 95% confidence interval estimates (dashed lines) of survival and hazard rate, respectively on left and right.

The posterior estimates of the DPMM confirm previous findings while also revealing some new and interesting details.

The survival function estimates demonstrate higher survival for the treatment group, in line with the results from previous models. We note that the 95% confidence intervals for the survival of the two groups do not overlap, as opposed to the case of the KM estimator, indicating higher confidence in the significance of the treatment.

From the hazard plot, it is clear that the lines are not parallel, indicating that the proportional hazards assumption does not hold. At the same time, it is also evident that the parametric nature of the previously fitted Weibull model is not suitable for this context. Indeed, the DPMM reveals a multimodal hazard function, showcasing its power to capture the complexity of the underlying process. We also note that the wider confidence intervals at the extremes reflect increased uncertainty due to fewer observations, a common issue in survival analysis.

An interesting observation is that both hazards initially increase for about a year, possibly reflecting that participants enter the study in an advanced phase of the tumor, hence they have a relatively high probability of death. Following this period, the hazards start decreasing, likely showing that the treatments are effective, with the treatment group performing better. As the hazard keeps decreasing, some fluctuations are observed. These might be due to responses to the treatment, such as side effects or health complications. By the end of the study the hazard is almost zero, indicating a very low chance of death, especially for the treatment group.

Overall, the DPMM offers insightful information that other models were unable to provide, thanks to its non-parametric nature that manages to capture complex relationships in the data.

# Chapter 6

# Conclusion

In conclusion, this work discussed frequentist and Bayesian methods for analyzing survival data. We demonstrated their application in a real-world scenario modeling deaths due to lung cancer to understand the effectiveness of a change in chemotherapy medicine combinations.

By employing the KM estimator, we showed that the estimated survival curve for the treatment group was significantly higher than the one for control group. This observation was confirmed by the low p-value of the log-rank test. To quantify the effect of the treatment, we first employed a Cox PH model and then an AFT Weibull model. Both reported coherent results, the first showing how treatment reduced the hazard rate, the second how it increased survival time.

While frequentist survival analysis models are powerful and widely used, they have several limitations. The Kaplan-Meier estimator does not easily incorporate covariates, hence it has a limited ability to account for the effects of explanatory variables. Cox regression, although flexible, relies on the proportional hazards assumption, which may not always hold in practice, as in the case of the lung cancer data. On the other side, parametric models require strong assumptions about the distribution of survival times, which can be restrictive and may not accurately reflect the true underlying distribution.

These limitations motivate the investigation of alternative approaches. There exist several alternative frequentist models that could have been explored, such as an extension of the Cox model with time-dependent covariates, flexible parametric models using splines, different para-

metric AFT specifications with different distributions, and many others. Additionally, machine learning techniques, such as survival trees, survival forests and survival neural networks, could have been applied to capture complex relationships in the data.

However, we decided to focus on a Bayesian approach. Bayesian statistics offers several advantages, from the incorporation of external knowledge, to the possibility of continuously monitoring evidence and the ability to quantify uncertainty about the data-generating process. Moreover, recent advantages in Monte Carlo Markov Chain sampling methods allow us to make inference even when sample sizes are limited, without needing to do asymptotic approximations.

Therefore, we introduced a fully non-parametric model widely used in Bayesian statistics, the Dirichlet process mixture model. Although its use in survival analysis is not common, its application on the lung cancer dataset was successful. The results provided alternative views on the interpretation of the lung cancer data and confirmed the efficacy of the proposed chemotherapy treatment.

# Bibliography


[1] Odd Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6, 2007.

[2] David Blackwell and James B. MacQueen. Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1, 2007.

[3] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 34, 1972.

[4] Frank Emmert-Streib and Matthias Dehmer. Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, 1, 2019.

[5] Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90, 1995.

[6] Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1, 2007.

[7] Bela A. Frigyik, Amol Kapila, and Maya R. Gupta. Introduction to the dirichlet distribution and related processes. *University of Washington*, 2010.

[8] Alan E. Gelfand and Adrian F.M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 1990.

[9] Maya R. Gupta. A measure theory tutorial. *University of Washington*, 2006.

[10] Joseph G. Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. Bayesian survival analysis. *Springer Series in Statistics*, 2001.

[11] E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 1958.

[12] David G. Kleinbaum and Mitchel Klein. Survival analysis: A self-learning text. *Statistics for Biology and Health*, 1996.

[13] Athanasios Kottas. Nonparametric bayesian survival analysis using mixtures of weibull distributions. *Journal of Statistical Planning and Inference*, 136, 2006.

[14] Antonio Lijoi. An introduction to bayesian statistics. *Bocconi University*, 2023.

[15] N. Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports. Part 1*, 50, 1966.

[16] Valerie Poynor and Athanasios Kottas. Nonparametric bayesian inference for mean residual life functions in survival analysis. *Biostatistics*, 20, 2019.

[17] David Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69, 1982.

[18] J. Sethuraman. A constructive definition of dirichlet priors, 1994.

[19] Yushu Shi, Michael Martens, Anjishnu Banerjee, and Purushottam Laud. Low information omnibus (lio) priors for dirichlet process mixture models. *Bayesian Analysis*, 14, 2019.

[20] Yee Whye Teh. Dirichlet process. *University College London*, 2023.

[21] Z. Ying, S. H. Jung, and L. J. Wei. Survival analysis with median regression models. *Journal of the American Statistical Association*, 90, 1995.