Reliable survival analysis based on the Dirichlet Process

Francesca Mangili[1], Alessio Benavoli, Cassio P. de Campos, Marco Zaffalon

IPG-IDSIA, Galleria 2, 6928 Manno-Lugano, Switzerland

Istituto Dalle Molle di studi sull'Intelligenza Artificiale (IDSIA), Scuola universitaria professionale della Svizzera italiana (SUPSI), Universit della Svizzera italiana (USI), Switzerland

Abstract

We present a robust Dirichlet process for estimating survival functions from samples with right-censored data. It adopts a prior near-ignorance approach to avoid almost any assumption about the distribution of the population lifetimes, as well as the need of eliciting an infinite dimensional parameter (in case of lack of prior information), as it happens with the usual Dirichlet process prior. We show how such model can be used to derive robust inferences from right-censored lifetime data. Robustness is due to the identification of the decisions that are prior-dependent, and can be interpreted as an analysis of sensitivity with respect to the hypothetical inclusion of fictitious new samples in the data. In particular, we derive a nonparametric estimator of the survival probability and a hypothesis test about the probability that the lifetime of an individual from one population is shorter than the lifetime of an individual from another. We evaluate these ideas on simulated data and on the Australian AIDS survival dataset. The methods are publicly available through an easy-to-use `R package`.

Keywords: Prior near-ignorance; Dirichlet Process; Survival analysis; Censored data; Bayesian nonparametrics; Log-rank test.

1. Introduction

Studies are conducted daily to compare the survival time of individuals of different medical, demographic, environmental, and behavioral characteristics. For instance, consider the Australian AIDS dataset [1, 2], where analyses suggested that a difference in survival time existed when discriminating individuals with

[1]Corresponding author: e-mail: francesca@idsia.ch

AIDS by the use (or not) of drugs (for whom a different survival was arguably expected), but also suggested that individuals with AIDS from the Queensland region in Australia have significantly worse survival time than those from the New South Wales region. Even if the latter conclusion has been questioned in those studies because survival difference associated with Australian regions was at first not expected, no formal analysis was used to assess the reliability of the result. This is a common situation in practice, partially related to the lack of methods with such purpose. Reliability of conclusions often come at a later stage by comparing the obtained results with those of other studies. The overall goal of this work is to provide interpretable and easy-to-use reliable methods for survival estimation and for testing differences in survival time. We address this problem from a Bayesian nonparametric perspective. In particular, we consider a Bayesian nonparametric approach based on a robust Dirichlet prior process.

Bayesian nonparametric procedures have appeared after [3], who introduced the Dirichlet process (DP). DP priors are behind the most popular Bayesian nonparametric models, and a number of authors have proposed nonparametric approaches based on them to estimate survival functions with censored data [4, 5, 6]. In particular, [4] developed an estimator for survival functions that converges to the Kaplan-Meier estimator as the prior strength of the DP goes to zero. An extension to the so-called "neutral to the right" priors was considered by [7]. Other authors focus on the hazard function; for instance, [8] uses Beta processes to develop a Bayesian estimator for the cumulative hazard.

A DP is completely characterized by its prior strength (or precision), which is a positive scalar number, and a prior base probability measure, which is an infinite-dimensional "parameter". Without detailed prior information, the subjective choice of the infinite-dimensional prior base measure may be difficult and may considerably affect inferences, leading to non-robust decisions. Generalizing earlier ideas developed in Bayesian parametric robustness [9, 10] and near-ignorance models [9, 11, 12], our first contribution is to adopt a viewpoint that models incomplete prior information by considering, instead of a single prior probability measure, the family of all probability measures that are compatible with the available information (13; 14). In this view, one models the lack of prior information by considering the set of all DP priors with fixed prior strength, and with normalized base measure free to vary within the set of all probability measures. The only required parameter is the prior strength (a positive scalar), and thus we refer to this prior as near-ignorance DP (or IDP) (13; 14). On the one hand, this viewpoint simplifies prior elicitation; on the other hand, it provides an extra level of robustness, because for all inferences it is able to estimate bounds for the inferences that

are independent of the prior base probability measure.

Based on the IDP, we propose a simple and robust estimator for the survival probability under a scenario with right-censored data. This can be seen as a generalization to continuous spaces of the work of [15], who takes a different approach for survival analysis and considers a finite partition of the continuous timeline (which depends on the observed events) in order to apply the (discrete and parametric) Imprecise Dirichlet Model [16]. Our IDP approach naturally reduces to that model when only a finite partition of the space is considered, yet it allows us to model survival data in continuous-time domains in a natural way. The probabilistic interpretation of posterior distributions provided by the IDP approach gives a straightforward way to quantify the uncertainty of the predictions. When it comes to estimating the prediction performance, we show on artificial data that despite the actual curves differ little from those estimated with Kaplan-Meier, the credible intervals of IDP are more reliable than Kaplan-Meier's linear (and log-transformed) confidence intervals obtained from Greenwood's formula.

Our second contribution is to use the IDP survival function estimator to develop a hypothesis test for the probability $P(X < X')$ that an individual from a population survives for a time X shorter than the survival time X' of an individual from another population. We call this test *generalized IDP rank-sum test* (or *IDP test* for short), since it extends the procedure proposed in [13] to right-censored data. Many authors have considered the problem of comparing two survival curves and proposed several tests. [17] provide a review of the different tests and compare them on a number of artificial datasets. The most commonly used tests are the log-rank test [18] and some generalizations of the Mann-Witney-Wilcoxon rank-sum test [19, 20], which are more powerful in case of early differences in the hazard function, since they weight more the first part of the survival curve. However, there are no clear recommendations as to which test is the best with respect to differences in the hazard function and in the setting of samples (with respect to number of observations and censoring distributions). Moreover, those tests are designed to detect generic differences between the two curves, whereas often we are interested in verifying whether the probability of survival is significantly larger in one population when compared to another. Those tests might be less reliable for such purpose, which is for instance prominent in case of crossing hazards. IDP allows deriving the posterior probability distribution of $P(X < X')$ and thus it is well suited to test directional hypothesis such as $P(X < X') > 1/2$, or equivalently $P(X < X') > P(X \geq X')$ (meaning that it is more probable that an individual of the first population has shorter life than an individual of the second population rather than longer or equal). When it comes to decision making, we can interpret the IDP

model in two ways. A first way, more strictly related to prior near-ignorance models, verifies whether decisions are prior-dependent or not given a predefined value of the prior strength. Decisions are taken only when they do not depend on the choice of the prior. A second, perhaps more intuitive, interpretation considers the upper and lower bounds given by the IDP model as obtained by adding in the most favorable/unfavorable positions a number of *fictitious* data equal to the strength of the DP prior. This is possible because we demonstrate that upper and lower posterior bounds are obtained by considering discrete priors made of a finite number of Dirac delta functions. This interpretation of the IDP shares some similarity with the method of influence curves used in robust analysis (21, 22) to study how an estimator varies when a new observation is added. In light of this interpretation, we consider as an indicator of robustness of the IDP test decision the minimum number of individuals that, if added to the available samples, could make the test contradict its decision (that is, the minimum prior strength that would make the decision prior-dependent).

The IDP test is widely applicable in clinical trials, and can be of interest also in reliability analysis of industrial components or social sciences, where X and X' represent the time to an event of interest. We will show through numerical simulations that the test is more consistent than traditional ones, especially in the case of non-proportional hazards, and that the IDP test can identify when decisions taken by the log-rank and Wilcoxon-type tests are not robust. We will also show that the results of the IDP test cannot be easily reproduced by the simple decrease of the significance level of the traditional tests (typically used to increase robustness); indeed, the IDP test provides a more sensible way to achieve reliable results.

Besides experiments using simulated data, we analyze the Australian AIDS dataset [1, 2]. Previous analyses suggested that a difference in survival time of AIDS patients existed when considering use (or not) of drugs, disease transmission, as well as Australian regions [1] (even though it is acknowledged that this last finding required further investigation), but not when considering gender. Our IDP test confirms those findings and outputs how reliable they are. In particular, the IDP test confirms that the survival time difference identified by traditional tests between individuals of distinct Australian regions is not reliable and that further information is needed, thus providing a mathematical model and reasoning that clarifies previous results.

2. Robust estimation of survival probabilities from right censored data

Let $X_1,\ldots,X_n$ be independent random variables describing the lifetimes of n individuals which are censored on the right by n independent follow up times $Y_1,\ldots,Y_n$ which are also independent of $X_1,\ldots,X_n$. The X_i are identically distributed (as X) with cumulative distribution function (CDF) $F(t)$. It will be slightly more convenient in what follows to deal with survival functions $S(t) = 1-F(t)$. In practice, we can only observe the event (either death or censoring) that occurs first, and thus each observation of the sample $\mathbf{Z}_n = \{(Z_1,d_1),\ldots,(Z_n,d_n)\}$ includes the time of the observation, $Z_i = \min(X_i,Y_i)$, and the indicator d_i which is 1 if $X_i < Y_i$ and 0 otherwise. In this work we only consider non-informative censoring, which occurs if individuals leave the trial independently of their actual state. Our main goal is to estimate the survival function $S(t)$. Let us define $\tilde{X}_1,\ldots,\tilde{X}_{n^d}$ as the subset of n^d uncensored observations (for which $d_i = 1$) and $\tilde{Y}_1,\ldots,\tilde{Y}_{n^c}$ the subset of $n^c = n - n^d$ censored observations (for which $d_i = 0$). Without loss of generality, we assume that Z_i, $\tilde{X}_i$ and $\tilde{Y}_i$ are ordered. For ease of presentation, it is assumed that X and Y are continuous and there are no ties; however, if ties are present, one can still use the results of this paper by introducing between tied observations a fictitious distance going down to zero [23].

A traditional nonparametric estimator for the survival function from lifetime data is the Kaplan-Meier (KM) estimator [24]:

$$\hat{S}(t) = \prod_{\tilde{X}_i \le t} \frac{n_i^+}{n_i}, \tag{1}$$

where $n_i = n - i + 1$ and $n_i^+ = n - i$ are the numbers of individuals at risk at time $\tilde{X}_i$ and just after $\tilde{X}_i$, respectively (that is, the individuals that have not yet been removed from the sample by censoring or death). The variance of this estimator is (usually) approximated by Greenwood's formula:

$$E[(\hat{S}(t) - E[\hat{S}(t)])^2] = \sum_{\tilde{X}_i \le t} \frac{\hat{S}^2(t)}{n_i n_i^+}, \tag{2}$$

and pointwise confidence intervals are based on the assumption of normality for the distribution of $\hat{S}(t)$. However, this is a rather rough assumption if n is small and $\hat{S}(t)$ is close to 1 or 0. A better approximation assumes normality for the logarithmic transformation of the hazard function [25].

For the purpose of Bayesian nonparametric estimation, the Dirichlet process [3] can be used. Let $\mathbb{X}$ be a standard Borel space with Borel σ-field $\mathscr{B}_{\mathbb{X}}$ and $\mathbb{P}$ be

the space of probability measures on $(\mathbb{X}, \mathscr{B}_{\mathbb{X}})$ equipped with the weak topology and the corresponding Borel σ-field $\mathscr{B}_{\mathbb{P}}$. Let $\mathbb{M}$ be the class of all probability measures on $(\mathbb{P}, \mathscr{B}_{\mathbb{P}})$. We call the elements $\mu \in \mathbb{M}$ nonparametric priors. An element of $\mathbb{M}$ is called a DP distribution $\mathrm{Dp}(\alpha)$ with base measure α if for every finite measurable partition $B_1, \ldots, B_m$ of $\mathbb{X}$, the vector $(P(B_1), \ldots, P(B_m))$ has a Dirichlet distribution with parameters $(\alpha(B_1), \ldots, \alpha(B_m))$, where $\alpha(\cdot)$ is a finite positive Borel measure on $\mathbb{X}$. Said $s = \alpha(\mathbb{X})$ the prior strength of the DP and $\alpha^* = \alpha/s$ the normalized base measure, we will use $\mathrm{Dp}(s, \alpha^*)$ as an alternative notation equivalent to $\mathrm{Dp}(\alpha)$; moreover, if $\mathbb{X} = \mathbb{R}$, we shall also describe $P \sim \mathrm{Dp}(s, \alpha^*)$ by saying $P \sim \mathrm{Dp}(s, G)$, where G stands for the cumulative distribution function (CDF) of α^*. As $F(t) = P(0,t]$ and $S(t) = 1 - P(0,t]$, hereafter we will refer indifferently to $\mathrm{Dp}(s, \alpha^*)$ as the prior of P, $F(t)$ and $S(t)$.

Consider the partition B_1 and $B_1^c = \mathbb{X} \backslash B_1$; then, if $P \sim \mathrm{Dp}(s, \alpha^*)$, from the definition of DP we have that $(P(B_1), P(B_1^c)) \sim Dir(s\alpha^*(B_1), s(1 - \alpha^*(B_1^c)))$, which is a Beta distribution. From the moments of the Beta distribution, we can thus derive that:

$$\mathscr{E}[P(B_1)] = \alpha^*(B_1), \qquad \mathscr{E}[(P(B_1) - \mathscr{E}[P(B_1)])^2] = \frac{\alpha^*(B_1)(1 - \alpha^*(B_1))}{(s+1)}, \quad (3)$$

where we have used the calligraphic letter $\mathscr{E}$ to denote expectation with respect to the Dirichlet process. This shows that the normalized measure α^* of DP reflects the prior expectation of P, while the scaling parameter s controls how much P is allowed to deviate from its mean.

Let f be a real-valued bounded function on $\mathbb{X}$. We call $E[f] = \int f dP$ a predictive inference about X, where P is a probability measure on $(\mathbb{X}, \mathscr{B}_{\mathbb{X}})$. If $P \sim \mathrm{Dp}(s, \alpha^*)$, then the expectation with respect to the Dirichlet process of $E[f]$ is

$$\mathscr{E}\big[E(f)\big] = \mathscr{E}\left[\int f dP\right] = \int f d\mathscr{E}[P] = \int f d\alpha^*. \quad (4)$$

One of the most remarkable properties of DP priors is that the posterior distribution of P is again a DP. Let $X_1, \ldots, X_n$ be independent and identically distributed samples from P and $P \sim \mathrm{Dp}(s, \alpha^*)$. Then the posterior distribution of P given the observations is

$$P|X_1, \ldots, X_n \sim \mathrm{Dp}\left(s+n, \frac{s}{s+n}\alpha^* + \frac{1}{s+n}\sum_{i=1}^{n} \delta_{X_i}\right), \quad (5)$$

where δ_{X_i} is an atomic probability measure centered at X_i. This means that the Dirichlet process satisfies a property of conjugacy, in the sense that the posterior

for P is again a Dirichlet process with updated unnormalized base measure $\alpha + \sum_{i=1}^{n} \delta_{X_i}$. From (3) and (5) we can easily derive the posterior mean and variance of P.

If we use $\mathrm{Dp}(s, \alpha^*)$ as prior for the distribution of $S(t)$, then the posterior distribution given the sample $\mathbf{Z}_n$ of randomly censored observations is a mixture of Dirichlet processes [5] and thus the conjugacy property is not satisfied anymore when data are censored. The p-th order moment of the posterior distribution of $S(t)$ is [4]

$$\mathscr{E}[S(t)^p|\mathbf{Z}_n] = \prod_{j=0}^{p-1} \frac{\alpha(t,\infty) + n_t^+ + j}{s+n+j} \prod_{\tilde{Y}_i \le t} \frac{\alpha[\tilde{Y}_i,\infty) + n_i + j}{\alpha[\tilde{Y}_i,\infty) + n_i^+ + j}, \tag{6}$$

where n_t and n_t^+ are the numbers of individuals at risk at time t and just after t. Here, as in the rest of the paper, the product is taken to be one if the number of factors is zero (that is, if $t < \tilde{Y}_1$). All prior inferences about $S(t)$ are fully determined by the choice of α^*, which reflects our prior guess about the distribution of X, and s, which reflects the strength of our belief in such guess. In the absence of prior information about the distribution of X, choosing the infinite dimensional parameter α^* may be hard and arbitrary, and may affect the reliability of posterior inferences. The only solution to this problem that has been proposed so far is the limiting DP obtained when the prior strength goes to zero. This model has been subject to some controversy, since it is not actually non-informative and assigns zero posterior probability to any set that does not include the observations (see [26] for a detailed discussion). This situation is aggravated in survival analysis with censored data. In this case, the posterior survival distribution may assign positive probability to survival times later than that of the latest observed lifetime; since there are no observed deaths there, the posterior distribution will depend on the prior base measure even if the prior strength goes to zero. As a consequence, this choice of prior does not always remove the dependence of the posterior inferences from the prior base measure. Therefore, we propose a different choice, which calls back to the ideas of sets of prior probabilities and near-ignorance models [9, 11, 12].

Definition 1. *Let $\mu \in \mathbb{M}$ be a nonparametric prior on P and $\mathscr{E}_\mu[E(f)]$ be the expectation of $E[f]$ with respect to μ. A class of nonparametric priors $\mathscr{T} \subset \mathbb{M}$ is called a prior ignorance model for predictive inferences about X if, for any*

real-valued bounded function f on $\mathbb{X}$, it satisfies:

$$\underline{\mathscr{E}}[E(f)] = \inf_{\mu\in\mathscr{T}} \mathscr{E}_\mu[E(f)] = \inf_{x\in\mathbb{X}} f(x),\ \ \overline{\mathscr{E}}[E(f)] = \sup_{\mu\in\mathscr{T}} \mathscr{E}_\mu[E(f)] = \sup_{x\in\mathbb{X}} f(x)\,, \quad (7)$$

where $\underline{\mathscr{E}}[E(f)]$ and $\overline{\mathscr{E}}[E(f)]$ are the lower and upper bounds of $\mathscr{E}_\mu[E(f)]$, respectively. ■

From (7) it can be observed that the range of $\mathscr{E}_\mu[E(f)]$ under the class $\mathscr{T}$ is the same as the original range of f. In other words, by specifying the class $\mathscr{T}$, we are not giving any information on the value of the expectation of f. Here, we have focused on the expectation of f as most of the nonparametric predictive inferences about a variable X can be expressed in terms of expectations of f, i.e., $\int f(X)dP_X = E[f]$. For instance, in this paper we are interested in $S(t) = \int_t^\infty dF(x) = E[f_t(x)]$ with $f_t(x) = I_{(t,\infty)}(x)$. Therefore, being prior ignorant on this kind of inferences is important in nonparametric statistics. Note that, if $\mathscr{T}$ is a prior ignorance model, then all sets $\mathscr{T}^*$ including $\mathscr{T}$ as a subset are prior ignorance models. However, a set of priors should not be too large otherwise the posterior inferences would become too little informative to be of any practical use. For instance, if we take $\mathscr{T} = \mathbb{M}$, that is, the largest possible set of priors, we have a prior ignorance model whose posterior inferences remain vacuous, or, in other words, a model that cannot learn from data [12, Sec. 7.3.7]. Such model would be useless in practice.

We are now ready to define the IDP.

Definition 2. *We call prior near-ignorance DP (or IDP) the class $\mathscr{T}$ of DPs such that*

$$\mathscr{T} = \{\mathrm{Dp}(s,\alpha^*):\ \forall\alpha^* \in \mathbb{P}\}\,.$$

■

Thus, the IDP set of priors is obtained by fixing the prior strength s and letting the normalized base measure α^* vary in the set of all probability distributions $\mathbb{P}$. We say that the IDP is a model of *near*-ignorance because to define it the modeler has to choose s. By considering the DP priors with base measures concentrated on the values of X that give the inf and sup of f it is easy to see from (4) that $\mathscr{T}$ verifies (7).

Given independent and identically distributed $X_1,\ldots,X_n$ from $P \sim \mathrm{Dp}(s,\alpha^*)$,

the posterior lower and upper bounds of $\mathscr{E}_\mu[E(f)|X_1,\ldots,X_n]$ are [13]:

$$\begin{aligned}\underline{\mathscr{E}}\big[E(f)|X_1,\ldots,X_n\big] &= \frac{s}{s+n}\inf_{x\in\mathbb{X}} f(x) + \frac{\sum_{i=1}^n f(X_i)}{s+n},\\ \overline{\mathscr{E}}\big[E(f)|X_1,\ldots,X_n\big] &= \frac{s}{s+n}\sup_{x\in\mathbb{X}} f(x) + \frac{\sum_{i=1}^n f(X_i)}{s+n},\end{aligned} \tag{8}$$

and the effect of prior near-ignorance vanishes asymptotically, that is, it holds (for any finite s)

$$\overline{\mathscr{E}}\big[E(f)|X_1,\ldots,X_n\big] - \underline{\mathscr{E}}\big[E(f)|X_1,\ldots,X_n\big] = \frac{s}{s+n}(\sup_{x\in\mathbb{X}} f(x) - \inf_{x\in\mathbb{X}} f(x)) \to 0, \quad \text{for } n\to\infty.$$

Moreover, from (8), it also follows that if we collect n i.i.d. random variable $X_1,\ldots,X_n$ the IDP estimates converge to Σ as $n\to\infty$, where $\Sigma = \lim_{n\to\infty}\frac{\sum_{i=1}^n f(X_i)}{s+n}$. Therefore, IDP provides an asymptotically consistent estimator of Σ.

2.1. Survival curve estimator

We apply the IDP to obtain a robust estimator of the survival function $S(t)$ and provide credible intervals for its value. As the IDP is a prior near-ignorance model, we have that prior inferences for the survival function $S(t) = E[I_{(t,\infty)}(x)]$ at a given time t are vacuous. Theorem 1, whose proof is given in the supplementary material, gives the posterior moments of the survival function $S(t)$ when using the IDP approach.

Theorem 1. *Given the IDP prior, the posterior lower and upper p-th order moments of $S(t)$ are*

$$\underline{\mathscr{E}}[S(t)^p|\mathbf{Z}_n] = \prod_{j=0}^{p-1}\frac{n_t^+ + j}{n_t^+ + s + j}\prod_{\tilde{X}_i\le t}\frac{n_i^+ + s + j}{n_i + s + j}, \quad \overline{\mathscr{E}}[S(t)^p|\mathbf{Z}_n] = \prod_{j=0}^{p-1}\prod_{\tilde{X}_i\le t}\frac{n_i^+ + s + j}{n_i + s + j}, \tag{9}$$

and are obtained for the prior DP with base probability measure $\alpha^* = \delta_{X_0}$, where $X_0 = t$ for the lower expectation, and $X_0 > t$ for the upper expectation. ■

Hereafter, we will denote by $\underline{\mu}_t$ and $\overline{\mu}_t$ the prior measures that give, respectively, the lower and the upper moments of $S(t)$. In the interpretation of the prior δ_{X_0} as fictitious data (deaths in this case), we can see that the upper bound for $S(t)$ is found when all fictitious deaths happen after the time of interest, i.e. $X_0 > t$, whereas the lower is obtained when they happen at $X_0 = t$. From (9) we can see

that the imprecision $\overline{\mathscr{E}}[S(t)^p|\mathbf{Z}_n] - \underline{\mathscr{E}}[S(t)^p|\mathbf{Z}_n]$ increases at the increase of s and that if $s \to \infty$ then $\underline{\mathscr{E}}[S(t)^p|\mathbf{Z}_n] \to 0$ and $\overline{\mathscr{E}}[S(t)^p|\mathbf{Z}_n] \to 1$. When $s \to \infty$ we are in the situation discussed in Section 2 where the set of priors $\mathscr{T} = \mathbb{M}$ is too large and thus the model cannot learn from data.

From (9) we can also derive upper and lower bounds for the expectation of $S(t)$,

$$\underline{\mathscr{E}}[S(t)|\mathbf{Z}_n] = \frac{n_t^+}{s+n_t^+} \prod_{\tilde{X}_i \le t} \frac{s+n_i^+}{s+n_i}, \quad \overline{\mathscr{E}}[S(x)|\mathbf{Z}_n] = \prod_{\tilde{X}_i \le t} \frac{s+n_i^+}{s+n_i}, \tag{10}$$

and state the following result:

Corollary 1. *The upper and lower bounds in (10) for the expectation of $S(t)$ always encompass the KM estimate $\hat{S}(t)$.* ■

We can see that the difference between the bounds of $\mathscr{E}_\mu[S(t)|\mathbf{Z}_n]$, which is

$$\overline{\mathscr{E}}[S(t)|\mathbf{Z}_n] - \underline{\mathscr{E}}[S(t)|\mathbf{Z}_n] = \frac{s}{s+n_t^+} \prod_{\tilde{X}_i \le t} \frac{s+n_i^+}{s+n_i},$$

goes to zero for $n_t^+ \to \infty$. From this and from the result in Corollary 1, it follows that the IDP estimator converges to the Kaplan-Meier one, which, for $t \le \max_i\{Z_i\}$, is a uniformly consistent estimator of $S(t)$ [27].

The posterior distribution of $S(t)$ derived from the IDP is used to build credible intervals for the survival function without resorting to the Gaussian approximation. We define the pointwise two-sided $100(1-\gamma)\%$ credible interval $[S_*(t), S^*(t)]$ for $S(t)$, as the symmetric interval having lower probability $1-\gamma$ of including the true value; that is, we want to ensure probability at least $1-\gamma$ that $S_*(t) \le S(t) \le S^*(t)$. Such interval is obtained by solving the equations:

$$\underline{\mathscr{E}}\left[I_{S(t)\le S^*(t)}\middle|\mathbf{Z}_n, \mathbf{Z'}_{n'}\right] = 1-\gamma/2, \text{ and } \underline{\mathscr{E}}\left[I_{S(t)\ge S_*(t)}\middle|\mathbf{Z}_n, \mathbf{Z'}_{n'}\right] = 1-\gamma/2, \tag{11}$$

with I the indicator function. This implies that

$$\begin{aligned}
\mathscr{E}_\mu\left[I_{S_*(t)\le S(t)\le S^*(t)}\middle|\mathbf{Z}_n, \mathbf{Z'}_{n'}\right] &= 1 - \left(1 - \mathscr{E}_\mu\left[I_{S(t)\ge S_*(t)}\middle|\mathbf{Z}_n, \mathbf{Z'}_{n'}\right]\right) \\
&\quad - \left(1 - \mathscr{E}_\mu\left[I_{S(t)\le S^*(t)}\middle|\mathbf{Z}_n, \mathbf{Z'}_{n'}\right]\right) \\
&= \mathscr{E}_\mu\left[I_{S(t)\ge S_*(t)}\middle|\mathbf{Z}_n, \mathbf{Z'}_{n'}\right] + \mathscr{E}_\mu\left[I_{S(t)\le S^*(t)}\middle|\mathbf{Z}_n, \mathbf{Z'}_{n'}\right] - 1 \\
&\ge 1-\gamma.
\end{aligned}$$

Theorem 2. *Given the IDP prior, the posterior lower bounds of* $\underline{\mathscr{E}}_{\mu}\left[I_{S(t)\leq a}\middle|\mathbf{Z}_n,\mathbf{Z}'_{n'}\right]$ *and* $\underline{\mathscr{E}}_{\mu}\left[I_{S(t)\geq a}\middle|\mathbf{Z}_n,\mathbf{Z}'_{n'}\right]$ *are found for the same priors that give, respectively, the upper and the lower moments of* $S(t)$*. That is,*

$$\begin{aligned}\underline{\mathscr{E}}\left[I_{S(t)\leq a}\middle|\mathbf{Z}_n,\mathbf{Z}'_{n'}\right] &= \underline{\mathscr{E}}_{\overline{\mu}_t}\left[I_{S(t)\leq a}\middle|\mathbf{Z}_n,\mathbf{Z}'_{n'}\right], \text{ and}\\ \underline{\mathscr{E}}\left[I_{S(t)\geq a}\middle|\mathbf{Z}_n,\mathbf{Z}'_{n'}\right] &= \underline{\mathscr{E}}_{\underline{\mu}_t}\left[I_{S(t)\geq a}\middle|\mathbf{Z}_n,\mathbf{Z}'_{n'}\right].\end{aligned}$$

■

Deriving analytical expressions for these expectations is usually a difficult task, especially as the number of observations increases, since conjugacy does not hold in case of censored data. Nevertheless, we have developed an efficient method for the numerical computation of posterior upper and lower distributions of $S(t)$ by Monte Carlo sampling, based on the fact that they correspond to the atomic priors in Theorem 1 (theoretical derivations and additional technical details are given in the supplementary material).

2.2. Comparison of survival curves

Let X and X' be lifetimes of individuals from two populations, F_1 and F_2 their distribution functions and S_1 and S_2 their survival functions; we want to test the hypothesis $P(X<X')\leq 1/2$ against $P(X<X')>1/2$, which resembles a standard Wilcoxon sum-rank test. For this, we need to estimate

$$P(X<X') = \int F_1(t)dF_2(t) = \int \left[1-S_1(t)\right]d[-S_2(t)]$$

based on samples $\mathbf{Z}_n$ and $\mathbf{Z}'_{n'}$ of n and n' observations from S_1 and S_2, respectively, which are assumed to be independent. We take as prior distributions for S_1 and S_2 the Dirichlet Processes $\mu_1=\mathrm{Dp}(s,G_1)$ and $\mu_2=\mathrm{Dp}(s,G_2)$, respectively (we use the same s for ease of presentation, but that is not a limitation). Then, by the properties of the IDP, it follows that a-priori: $\mathscr{E}_{\mu_1,\mu_2}[P(X<X')]=\int G_1(t)dG_2(t)$. The IDP satisfies the condition of prior near-ignorance, hence a-priori we are fully ignorant about the hypothesis $P(X<X')\leq 1/2$. Using IDP, the hypothesis test can be performed in two steps. First, we define a loss function [28, Ch 4.4]

$$L(P,a) = \begin{cases} k_0\cdot I_{\{P(X<X')>\frac{1}{2}\}}, & \text{if } a=0,\\ k_1\cdot I_{\{P(X<X')\leq\frac{1}{2}\}}, & \text{if } a=1,\end{cases} \tag{12}$$

where a is our action and k_0 is the loss we incur by taking the action $a = 0$ (that is, declaring that $P(X < X') \leq \frac{1}{2}$) when actually $P(X < X') > \frac{1}{2}$, while k_1 gives the loss we incur by taking the action $a = 1$ (that is, declaring that $P(X < X') > \frac{1}{2}$) when actually $P(X < X') \leq \frac{1}{2}$. Second, we compute the expected value of this loss and we choose $a = 1$ if $\mathscr{E}_{\mu_1,\mu_2}[L(P,0)|\mathbf{Z}_n, \mathbf{Z'}_{n'}] \geq \mathscr{E}_{\mu_1,\mu_2}[L(P,1)|\mathbf{Z}_n, \mathbf{Z'}_{n'}]$, i.e.,

$$\begin{aligned} & k_0 \mathscr{E}_{\mu_1,\mu_2}\left[I_{\{P(X<X')>\frac{1}{2}\}}|\mathbf{Z}_n, \mathbf{Z'}_{n'}\right] \geq k_1 \mathscr{E}_{\mu_1,\mu_2}\left[I_{\{P(X<X')\leq\frac{1}{2}\}}|\mathbf{Z}_n, \mathbf{Z'}_{n'}\right] \\ & \Rightarrow \mathscr{E}_{\mu_1,\mu_2}\left[I_{\{P(X<X')>\frac{1}{2}\}}|\mathbf{Z}_n, \mathbf{Z'}_{n'}\right] \geq \frac{k_1}{k_0+k_1}, \end{aligned} \tag{13}$$

and $a = 0$ otherwise. When the above inequality is satisfied, we can declare that $P(X < X') > \frac{1}{2}$ with probability at least $\frac{k_1}{k_0+k_1}$. For comparison with the traditional test we will take $\frac{k_1}{k_0+k_1} = 1-\gamma$, where γ is the significance level of the test (usually the notation α is used, but α has another meaning here); notice however that, while a principled way of choosing γ is lacking in the traditional tests, we can set it in a more informed way based on the losses k_0 and k_1 expected in case of error. Finally, according to the decision rule in (13), we verify whether $\underline{\mathscr{E}}\left[I_{\{P(X<X')>\frac{1}{2}\}}\middle|\mathbf{Z}_n, \mathbf{Z}'_{n'}\right] > 1-\gamma$, and whether $\overline{\mathscr{E}}\left[I_{\{P(X<X')>\frac{1}{2}\}}\middle|\mathbf{Z}_n, \mathbf{Z}'_{n'}\right] > 1-\gamma$, and then:

1. If both inequalities are true, we declare that the probability of longer survival is higher for the first group than for the second group with probability greater than $1-\gamma$.
2. If both inequalities are false, we declare that the probability of longer survival is higher for the first group than for the second group with probability lower than the desired probability of $1-\gamma$.
3. If the left-hand inequality is false but the right-hand inequality is true, we are in an *indeterminate* situation.

Thus, to perform the test we need to compute lower and upper bounds for the probability that $P(X < X') > \frac{1}{2}$.

Lower bounds for the expectation of $P(X < X')$ and the probability of $P(X < X') > \frac{1}{2}$ in the case of censored data are given in the two following theorems. Similar results can be found for the upper bounds.

Theorem 3. *The lower bound of the expectation of $P(X < X')$ is found using the DP priors $\overline{\mu}_1 = \mathrm{Dp}(s, \alpha^*_{10})$ and $\underline{\mu}_2 = \mathrm{Dp}(s, \alpha^*_{20})$ with base probability measure*

$$d\alpha_{10} = s\delta_{Z_0}, \quad d\alpha_{20} = s\textstyle\sum_{i=0}^{n'} \pi_i \delta_{Z'^+_i}, \tag{14}$$

where $Z_0 > \max\{Z_1,\dots,Z_n,Z'_1,\dots,Z'_{n'}\}$, $Z'_i < Z'^{+}_i < Z_{i+1}$, *with* $i=0,\dots,n'$, $Z'_0=0$ *and* $Z'_{n'+1}=\infty$, *and where* $\pi=(\pi_0,\pi_1,\dots,\pi_{n'})$ *is a vector of weights verifying* $\pi_i=0$ *if* $d'_i=1$ *and* $\sum_{i=0}^{n'}\pi_i=1$. *The vector* π *is obtained by minimizing the expectation*

$$\mathscr{E}_{\overline{\mu}_1,\underline{\mu}_2}[P(X<X')|\mathbf{Z}_n,\mathbf{Z'}_{n'}] = \sum_{i=1}^{n^d}\left(\overline{\mathscr{E}}[S_1(\tilde{X}_{i-1})|\mathbf{Z}_n]-\overline{\mathscr{E}}[S_1(\tilde{X}_i)|\mathbf{Z}_n]\right)\mathscr{E}[S_2(\tilde{X}_i)|\mathbf{Z'}_{n'}], \tag{15}$$

where $\tilde{X}_0=0$, $\overline{\mathscr{E}}[S_1(t)|\mathbf{Z}_n]$ *is given by (10), and the posterior expectation of* $S_2(t)$ *is derived from (6) and is equal to:*

$$\mathscr{E}_{\underline{\mu}_2}[S_2(t)|\mathbf{Z'}_{n'}] = \prod_{Z'_i\le t}\frac{\alpha_{20}(Z'_i,+\infty)+n'_i-d'_i}{\alpha_{20}[Z'_i,+\infty)+n'_i}.$$

■

Theorem 4. *The infimum of the probability that* $P(X<X')>\frac{1}{2}$ *is found for the DP priors* $\overline{\mu}_1=\mathrm{Dp}(s,\alpha^*_{10})$ *and* $\underline{\mu}^*_2=\mathrm{Dp}(s,\alpha^*_{20})$ *with base probability measure*

$$d\alpha_{10}=s\delta_{Z_0},\quad d\alpha_{20}=s\textstyle\sum_{i=0}^{n'}\pi'_i\delta_{Z'^{+}_i}, \tag{16}$$

where $Z_0 > \max\{Z_1,\dots,Z_n,Z'_1,\dots,Z'_{n'}\}$, $Z'_i < Z'^{+}_i < Z_{i+1}$, *with* $i=0,\dots,n'$, $Z'_0=0$ *and* $Z'_{n'+1}=\infty$, *and where* $\pi'=(\pi'_0,\pi'_1,\dots,\pi'_{n'})$ *is a vector of weights that verifies* $\pi'_i=0$ *if* $d'_i=1$ *and* $\sum_{i=0}^{n'}\pi'_i=1$. *The vector* π' *is obtained by minimizing* $\mathscr{E}_{\pi'}[I_{P(X<X')>\frac{1}{2}}|\mathbf{Z}_n,\mathbf{Z'}_{n'}]$. ■

From these theorems we see that the extreme values of the expectation of $P(X<X')$ and of the probability of $P(X<X')>\frac{1}{2}$ are found for discrete priors. Hence, the priors can still be easily interpreted as additional deaths and the posterior distribution of $P(X<X')$ at these priors efficiently computed by Monte Carlo sampling (see proofs and results in the supplementary material). However, while minimizing (15) can be reduced to a convex optimization problem, so that the prior $\underline{\mu}_2$ can be easily found, the minimization problem in Theorem 4 is much more computationally expensive since for each vector π' the objective function $\mathscr{E}_{\pi'}[I_{P(X<X')>\frac{1}{2}}|\mathbf{Z}_n,\mathbf{Z'}_{n'}]$ has to be computed by Monte Carlo sampling. Thus, finding the exact extrema in this case is not tractable. Conservative bounds for

the value of $\mathscr{E}_{\mu_1,\mu_2}[I_{P(X<X')>\frac{1}{2}}|\mathbf{Z}_n,\mathbf{Z'}_{n'}]$ can be efficiently computed based on the following result.

Theorem 5. *Let $S_{\underline{\mu}_t,i}$, $i=1,2$, be a process such that at each time t the distribution of $S_{\underline{\mu}_t,i}(t)$ is the posterior distribution of $S_i(t)$ when the prior base measure is δ_t (that is, the one that gives the lower expectation of $S_i(t)$) and $S_{\overline{\mu},i}$ the posterior distribution of $S_i(t)$ when the prior base measure is δ_{X_0}, $X_0 > t$. Then,*

$$\begin{aligned}
\mathscr{E}_{\mu_1,\mu_2}\left[I_{\{P(X<X')>\frac{1}{2}\}}\middle|\mathbf{Z}_n,\mathbf{Z}'_{n'}\right] &\geq \mathscr{E}\left[I_{\int S_{\underline{\mu}_t,2}(t)d[-S_{\overline{\mu},1}(t)]>\frac{1}{2}\}}\middle|\mathbf{Z}_n,\mathbf{Z}'_{n'}\right] \\
\mathscr{E}_{\mu_1,\mu_2}\left[I_{\{P(X<X')>\frac{1}{2}\}}\middle|\mathbf{Z}_n,\mathbf{Z}'_{n'}\right] &\leq 1-\mathscr{E}\left[I_{\int S_{\underline{\mu}_t,1}(t)d[-S_{\overline{\mu},2}(t)]>\frac{1}{2}\}}\middle|\mathbf{Z}_n,\mathbf{Z}'_{n'}\right].
\end{aligned} \tag{17}$$

■

We have verified in several case studies that, when s is small, the bounds in (17) are close to the exact ones (see the supplementary material). Thus, for small s, they do not increase too much the fraction of indeterminate instances. Instead, for large s these bounds are too rough and in most cases do not lead to determinate decisions. For this reason, we propose to use a less conservative approach that approximates the probability obtained using $\underline{\mu}_2^*$ with that obtained for $\underline{\mu}_2$. In the absence of censoring this result is exact as we have the equality between these two probabilities.

Theorem 6. *In the absence of censoring, the upper and lower bounds of $\mathscr{E}_{\mu_1,\mu_2}[I_{\{P(X<X')>\frac{1}{2}\}}|\mathbf{Z}_n,\mathbf{Z}'_{n'}]$ are found for the priors that give, respectively, the upper and lower expectation of $P(X<X')$.* ■

Unfortunately, this result does not hold in case of censored data (see the supplementary material for a counterexample). Then, the lower and upper probabilities obtained using $\underline{\mu}_2$ are inner approximations of the correct ones, but we have verified empirically that the difference between them is very small (see the supplementary material for more details). Moreover, results of the simulations in Section 3 show that a test performed using this approximation is calibrated and more reliable than the traditional tests. Therefore, the use of the priors $\overline{\mu}_1, \underline{\mu}_2$ allows an efficient computation of the lower and upper probabilities and, at the same time, provides a good approximation of the true lower and upper probabilities. For this reason we suggest to use this approximation.

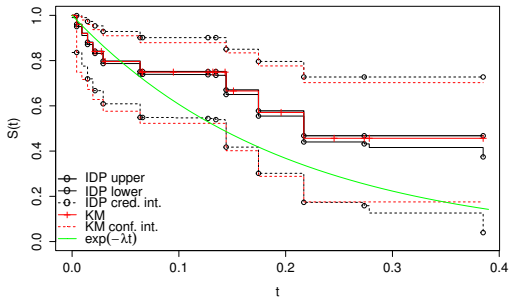
Although the IDP test shares several similarities with a standard Bayesian approach, it embodies a change of paradigm when it comes to taking decisions. The IDP test has the advantage of producing an *indeterminate* outcome when the decision about the hypothesis being tested is *prior-dependent*, that is, the IDP test is able to tell whether the decision which minimizes the expected loss would change depending on the DP base measure one chooses. Therefore, the IDP test is robust in providing a determinate decision only when all DP priors in the class represented by the IDP agree. Indeterminate outcomes indicate that collecting additional data could provide valuable further evidence to the decision maker.

To perform a sensitivity analysis of the test decision with respect to the possible values of the prior strength s, we propose to evaluate the maximum value (s_{max}) for which the decision remains determinate. Then, we are guaranteed that by repeating the test with s_{max} more observations in each group we will never contradict the decision taken with the available data. In this view, s_{max} can be seen as a measure of robustness of the test decision. Notice that values of s_{max} greater than 1 indicate already a quite robust decision: for example, if $s_{max} = 2$ we are implicitly considering the very unlikely event that all the 4 fictitious observations (2 for each group) fall in the most unfavorable position for the hypothesis under test.

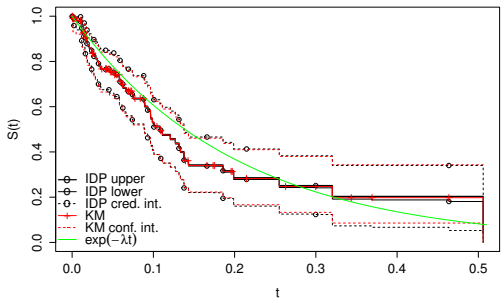
3. Simulations

Using simulated data, we compare the IDP estimator for survival functions with the traditional KM estimator, where we employ Greenwood's formula plus the Gaussian approximation for the estimation of confidence intervals regarding the survival probability. For obtaining the KM estimator, we use the `R package survival` [29] and we consider both the linear *plain* confidence interval (which is routinely constructed by most statistical packages), and the interval built using a *log-log* transformation of the cumulative hazard rate [25]. We consider simulated cases where lifetime data and follow up times are generated from an exponential distribution, such that $X_1, \ldots, X_n, Y_1, \ldots, Y_n \sim Exp(\lambda = 5)$. As an illustration, Figure 1 shows examples of survival curves estimated for $n = 25$ and $n = 100$ with 95% confidence intervals for the log-log KM and the 95% credible interval defined in (11) for the IDP with $s = 0.25$. While a (somewhat small) difference between the two estimators can be noticed for $n = 25$, the same is unperceivable with $n = 100$. Notice however that the lower bound for the IDP credible interval accounts better for the observed censored data. The lower bound of the IDP decreases with censoring, thus reflecting the fact that, after censoring, we have less

individuals under observation and thus the estimate we can (robustly) provide has wider uncertainty and, thus, wider credible intervals.



(a) $n = 25$



(b) $n = 100$

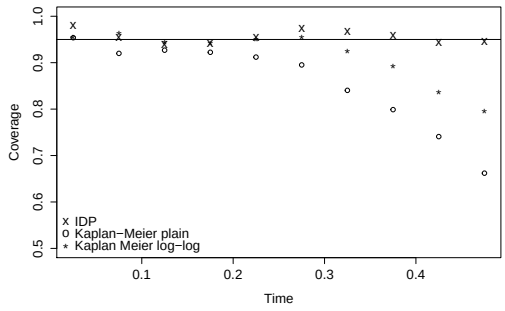
Figure 1: Survival curve estimates from the IDP ($s = 0.25$) and the log-log KM approaches after n observations. The continuous lines represent the KM estimator (red line) and the lower (thin black line) and upper (thick black line) IDP estimates; the dotted lines represent the confidence/credible intervals estimated by the KM (red) and the IDP (black).

Figure 2 shows the coverage of the 95% intervals obtained by the IDP and the KM estimators (both plain and log-log) over 1000 repetitions (and varying time instants). The plain KM confidence intervals provide a coverage lower than the target value of 0.95, as it is known that Greenwood's estimator tends to underestimate the true variance of KM for small to moderate samples [25]. The log-log KM estimator provides a better coverage than plain KM, but still below the target. Note that coverage decreases at later time instants, which is exactly when fewer data are available for the estimation (fewer individuals at risk). On the other hand, IDP is more robust and, in general, achieves a coverage slightly larger than 0.95.

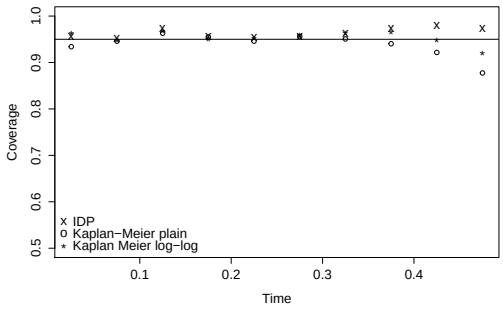
Table 1: Coverage of pointwise KM's confidence intervals and IDP's credible intervals.

	KM (plain)	KM (log-log)	IDP(s=0.01)	IDP(s=0.25)	IDP(s=0.5)
n=25	0.877	0.932	0.928	0.960	0.975
n=100	0.931	0.948	0.942	0.956	0.966

If one chooses a very low value of s, for instance $s = 0.01$, then IDP cannot reach the target coverage either. Table 1 presents coverages (averaged over all time instants) for different values of s. The greater the choice of s the more conservative is IDP and the width of its credible intervals increases. In these results, as it can be seen in Table 1, values for s above 0.5 are already too conservative. We take $s = 0.25$ for most experiments from now on, because in general it already achieves the desired coverage.



(a) $n = 25$



(b) $n = 100$

Figure 2: Coverage of the Kaplan-Meier and IDP credible intervals with $s = 0.25$.

Table 2: Fraction of H1 decisions for the different tests in the cases where the IDP is determinate or indeterminate (round brackets). Into square brackets in the IDP column, we report the fraction of indeterminate outcomes.

	Log-rank	Peto-Peto	IDP($s = 0.25$)
SH	0.0270 (0.288)	0.0291 (0.329)	0.0259 [0.073]
PH	0.7038 (0.623)	0.6842 (0.437)	0.6879 [0.183]
EHD	0.7714 (0.231)	0.8617 (0.878)	0.8347 [0.147]
LHD	0.4315 (0.957)	0.2061 (0.137)	0.2299 [0.117]

For the purpose of comparing the survival of individuals of two independent populations, we have applied the IDP test to a number of artificial datasets simulating the different scenarios proposed by [17], that is, same hazard (SH), proportional hazards (PH), early hazard difference (EHD) and late hazard difference (LHD), and we have compared it against the log-rank test and the (Peto-Peto modification of the) Gehan-Wilcoxon test, both implemented in the `R package survival`. Tests are one-sided, always analyzing whether $P(X < X') > \frac{1}{2}$ (which is considered as the alternative hypothesis), with X and X' denoting the corresponding variables of interest in two independent populations. Table 2 gives the fraction of rejection of the null hypothesis over the cases where IDP yields a determinate decision and over the cases where it cannot take a decision (into round brackets); the fraction of indeterminate cases for the IDP test is also shown (square brackets). The SH line regards same hazards, so the null hypothesis is true and the fraction of rejections represents the type-I error (which should be lower than $\gamma = 0.05$). In the other lines, the fraction of rejections measures the power of the test. Further details about the simulation setting are presented in the supplementary material.

Let us first consider the determinate cases. We see that all tests perform similar for SH. As expected, the Peto-Peto test outperforms the log-rank test only in the EHD scenario. Overall, IDP presents similar power to Peto-Peto, and is outperformed by the log-rank for LHD. Even though the IDP test is never the most powerful, it outperforms the log-rank in the EHD scenario and the Peto-Peto test in the LHD scenario (it ties with Peto-Peto for PH), thus it can be seen as a compromise between the other two tests when the scenario under study is unknown (which is often the case). The IDP test also gives us an extra information: it suggests that part of the cases (indicated between square brackets) are *indeterminate* and could be better decided if more data were available (with few more data, some of those decisions could easily become a rejection of the null hypothesis). While

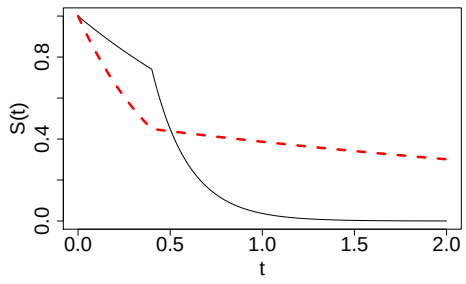
traditional tests always issue a determinate decision, the IDP test acknowledges that a reliable decision cannot be made without collecting more data. In the cases where the IDP test is indeterminate we can see from Table 2 that the type-I error in the SH scenario is much larger than γ both for the log-rank and Peto-Peto tests. In the PH scenario the power of the traditional tests reduces with respect to that obtained when the IDP test is determinate. In the EHD scenario the power of the Peto-Peto test is much larger than that of the Log-rank test since the former is biased toward H1 in such scenario, whereas the log-rank test is biased toward H0. The opposite situation happens for the LHD scenario. Indeed those two tests carry some assumptions that are usually hard to verify a priori. The IDP test is capable of detecting the situations where those assumptions mostly affect the test decisions; the fact that the two tests usually disagree in those cases shows that the decision is difficult and may not be reliable. Moreover, we will see in Example 2 that the assumptions underlying the log-rank and Peto-Peto tests may affect the reliability of their decisions under the null hypothesis. Therefore, for the robustness of the decision it is important to isolate those instances where the decision strongly relies on those uncertain assumptions.

We believe that robustness is an important feature of a test, as it may contribute to avoid misleading decisions and false discoveries which affect the replicability of results. In this context, we show an example where log-rank and Peto-Peto tests are not reliable (a reliable test should not reject the null hypothesis with probability higher than $\gamma = 0.05$ when it is true).

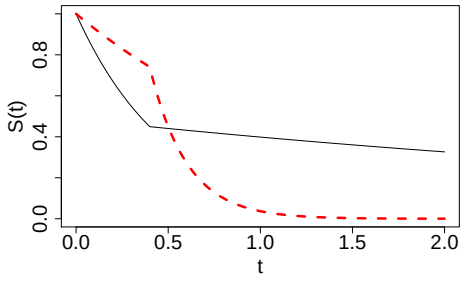
Example 1. *Consider two groups with different sizes ($n = 20$ and $n' = 100$) from the same population. Lifetime data are taken from the same exponential distribution $Exp(\lambda = 1)$; censoring times are sampled from the uniform with support $[0,2]$. Due to the very different sample size, the type-I error, evaluated over 1000 repetitions, is greater than the desired value $\gamma = 0.05$ both for log-rank and Peto-Peto tests (0.065 and 0.061, respectively). If we limit to consider only the samples where the IDP test is determinate (89% of the cases), the Type-I error is 0.032 both for the log-rank and the Peto-Peto tests and 0.030 for the IDP test. Considering instead only the samples where the IDP is indeterminate, the error become 0.300 for the log-rank and 0.336 for the Peto-Peto test.*

Besides the robustness that is provided by the IDP test in general, Example 2 shows that the IDP test is particularly useful in situations where hazards cross (as well as in situations where one is unsure about such crossing, if he wants to take reliable decisions), as most classical tests become unreliable.

Example 2. *Consider two scenarios A and B, each of which with two groups of* $n = 50$ *and* $n' = 25$ *samples. Lifetime data are sampled from two populations with survival functions given in Figure 3.*



(a) Scenario A



(b) Scenario B

Figure 3: Survival functions for control (continuous thin line) and treated (dashed thick line) populations.

The survival curves in the two scenarios are very similar, but curves of control and treatment have their shape interchanged. The hazard function is designed so that in both scenarios $P(X < X')$ *is slightly smaller than* $\frac{1}{2}$ *(with* X *representing the control group); hence the null hypothesis is true in both scenarios. The type-I error of the three tests, evaluated on 1000 different repetitions, is shown in Table 3. The log-rank has a large Type-I error in scenario A, while Peto-Peto has very large Type-I error in scenario B, and both are way greater than the target of* 0.05*. Indeed, these tests are designed to respectively weight more late and early differences in the curves and, thus, cannot account correctly for the overall behavior*

Table 3: Type-I error of different tests for significance level $\gamma = 0.05$ and $\gamma = 0.01$ over the determinate cases. Into round brackets we report the Type-I error for Log-rank and Peto-Peto in the IDP indeterminate instances, and into square brackets the fraction of indeterminate outcomes for IDP.

	Log-rank		Peto-Peto		IDP ($s = 0.25$)
	$\gamma = 0.05$	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.01$	$\gamma = 0.05$
Scenario A	0.327 (1.000)	0.133 (1.000)	0.027 (0.227)	0.007 (0.000)	0.031 [0.022]
Scenario B	0.002 (0.000)	0.001 (0.000)	0.076 (0.959)	0.019 (0.000)	0.042 [0.049]

of the functions. The IDP test does not assume any characteristic for the survival function; its Type-I error has been always smaller than 0.05*. When the IDP is indeterminate the error is 1 or almost 1 for the cases where the tests are biased toward the alternative hypothesis (that is, scenario A for log-rank and scenario B for Peto-Peto test). In the opposite scenario, tests are biased toward favoring the null hypothesis and thus the error is much smaller; this is particularly evident for the log-rank, which is more strongly biased and thus never rejects the null hypothesis apart for only 2 extreme cases where also the IDP and the Peto-Peto do reject it. The log-rank does not achieve error below* 0.05 *in Scenario A even if we set the significance level of the test at* 0.01*. Peto-Peto test, instead, obtains an error below* 0.05 *if we set* $\gamma = 0.01$ *(but not below the target* γ *itself), but such choice is completely arbitrary, as one cannot know (or rely on) which would be the* γ *to achieve type-I error of* 5% *(besides the sharp drop in power that the test would face). Into round brackets we report the error over the indeterminate cases identified by an IDP test with* $\gamma = 0.05$*. We expect that, in those cases which are critical for a decision criteria with* $\gamma = 0.05$*, a more conservative test using* $\gamma = 0.01$ *would not reject the null hypothesis thus achieving a very small Type-I error error. This is always true except for the log-rank test in scenario A (with an error of 1) thus confirming once more the strong bias of this test in case of crossing hazards.*

4. Australian AIDS survival data

In this section, we consider the Australian AIDS survival dataset [1, 2] of cases reported to the Australian National Centre in HIV Epidemiology and Clinical Research. The dataset contains records of 2843 individuals diagnosed from 1982 to 1991, of which 1787 died prior to the end of the study.

We consider the five pairs of groups listed in Table 4. The first considers gender, the next two consider populations from distinct regions of Australia (*NSW*,

Table 4: p-values (Log-rank and Peto-Peto tests) and posterior probabilities (IDP test) for the AIDS dataset. (H0)/(H1) indicates the decision taken at the level $\gamma = 0.05$, while (I) indicates an indeterminate outcome.

	Group 1	Group 2	N. Data	Log-rank	Peto-Peto	IDP ($s = 0.25$)	s_{max}
1	Male	Female	85/1002	0.182 (H0)	0.525 (H0)	0.579/0.732 (H0)	5.63
2	NSW	VIC	835/343	0.228 (H0)	0.024 (H1)	0.911/0.920 (H0)	1.25
3	QLD	NSW	835/193	0.046 (H1)	0.027 (H1)	0.927/0.967 (I)	-
4	No Drug	Drug	975/112	0.011 (H1)	0.019 (H1)	0.986/0.990 (H1)	3.125
5	Blood	Haemoph.	46/78	0.046 (H1)	0.007 (H1)	0.991/0.996 (H1)	2.08

VIC, and *QLD* mean, respectively, New South Wales including the Australian Capital Territory, Victoria, and Queensland, the three most populated regions in the country), the fourth considers users of drugs (or not), and the fifth considers the reported transmission between *Blood* (that is, receipt of blood, blood components or tissue) and *Haemophilia* (coagulation disorder). These tests have been chosen to confirm (or to find evidence against) the results reported long ago by the responsible for the dataset himself [1]. For each pair, we test the hypothesis that the survival time is shorter for the first group of individuals than for the second group. More precisely, let X be the survival time of individuals from Group 1 (for example, *Male* in the first line of Table 4) and X' that of individuals from Group 2. We test $P(X < X') \leq \frac{1}{2}$ (null hypothesis, or H0) against $P(X < X') > \frac{1}{2}$ (alternative hypothesis, or H1) using the log-rank and the Peto-Peto modification of the Gehan-Wilcoxon test, as well as the IDP test. In Table 4, we present p-values for the log-rank and Peto-Peto tests and posterior probabilities for the IDP test. We use the significance level $\gamma = 0.05$; then the null hypothesis is rejected (denoted H1) if the p-value is smaller than γ (log-rank and Peto-Peto tests) or if the lower probability of the alternative hypothesis is greater than $1 - \gamma$ (IDP test); if, instead, the upper probability is greater than $1 - \gamma$ but the lower probability is not, then we declare it *indeterminate*. Figure 4 shows the survival curves related to the tests under consideration.

According to the previous study [1], no difference in survival has been verified between *Male* and *Female*, and between *NSW* and *VIC*, while a difference in survival time was identified for *QLD* versus *NSW*, *No-drug* versus *drug* usage, and *Blood* versus *Haemophilia* (the last comparison was not directly performed, but can be inferred from their conclusions). Table 4 shows that our results are consistent with those findings. The IDP test confirms that gender is not significant to discriminate survival time even if the strength of its prior were as large as 5.63, that is, even if more than 5 fictitious additional data samples were placed in

each group in the most adversarial positions in time, the conclusion would remain unchanged. Log-rank and Peto-Peto do not reject H0 either.

For *NSW* against *VIC*, log-rank and Peto-Peto tests provide very different p-values. The Peto-Peto test weights more the first part of the survival curves, where the curve of *NSW* lies below that of *VIC*. In the second part, the curves appear interchanged, but this is somehow neglected by the test. This is a clear case where the proportional hazards assumption fails, and traditional tests are less reliable. IDP indicates that up to $s = 1.25$ we can reliably *not* reject H0. These results are supported by data and discussions presented in Australian technical reports [30, 31].

Interestingly, both log-rank and Peto-Peto reject H0 for the comparison *QLD* versus *NSW*, in accordance with previous results [1]. However, [1] clearly acknowledge that the poor survival of *QLD* with respect to *NSW* is *"obscure and require further investigation"*. Indeed recent Queensland reports [32, 33] (as well as the previously mentioned Australian reports) show no reason to believe that survival of AIDS patients in that region is shorter than in New South Wales. The IDP test identifies such doubtful situation and responds an *indeterminate* outcome (with an indeterminacy that exists even if $s \to 0$), suggesting that further data should be collected for reaching a better decision.

Finally, both *No-drug* versus *Drug* and *Blood* versus *Haemophilia* tests have H0 rejected with all the three tests, even though p-values of log-rank and Peto-Peto are quite different in the latter. This might be explained by the possible late cross in the curves, as shown in Figure 4. In both comparisons, IDP provides a robust decision of rejecting H0 together with a measure of its robustness, given by the maximum strength s such that the decision of rejecting H0 would yet take place. For *Blood* versus *Haemophilia*, even if $s_{max} = 2.08$ additional fictitious data samples were included in each of the groups ($s_{max} = 3.125$ fictitious data in the *No-drug* versus *Drug*), we would still reject H0, so the results are indeed quite reliable.

From these results we can see that the IDP test does not simply issue an indeterminate outcome when the p-value of the traditional tests is close to the significance level γ: for instance, the p-value of the log-rank test is very close to the confidence level γ both in comparison *QLD* versus *NSW* and *Blood* versus *Haemophilia* (the p-value is 0.046 in both cases), but only in the first comparison the IDP test is indeterminate. The same situation can be seen for the Peto-Peto test with comparisons *NSW* versus *VIC* (p-value$= 0.024$) and *QLD* versus *NSW* (p-value$= 0.027$). Neither can the IDP test be mimicked by issuing an indeterminate outcome when the decision of the other tests differ: they both rejected H0 in

the comparison *QLD* versus *NSW*, while IDP test is indeterminate and they take different decisions in the comparison *NSW* versus *VIC* while IDP issues H0.

5. Conclusions

In this paper we have employed a model of prior near-ignorance for nonparametric inferences based on the Dirichlet process to develop robust methods for survival analysis with right-censored data. The approach has the benefits of a Bayesian inference while avoiding completely the need of specifying the infinite-dimensional parameter of the Dirichlet Process. The only free parameter is the strength of the prior, which has a clear interpretation as the number of additional fictitious data (placed in the most adversarial way) that we impose to the model while checking whether the decision would remain unchanged. This makes the elicitation of the prior very easy, and allows us to compute posterior inferences for which no closed form expression exists by a simple Monte Carlo sampling from the Dirichlet distribution, thus avoiding more demanding sampling approaches typically used for the Dirichlet process (for example, stick breaking). Based on this prior near-ignorance model, we have developed an estimator for the survival curves which can provide reliable credible intervals for the probability of survival. We have also proposed a general, simple and conservative approach for testing the difference in survival of individuals from two independent populations called IDP test, which is a robust alternative to log-rank and other weighted Wilcoxon rank-sum tests. The IDP test is able to identify whether the decision is prior-dependent, and gives the possibility of evaluating the size of the difference between survival times through the posterior distribution of $P(X < X')$. Moreover, the IDP test allows us to perform an analysis of robustness with respect to the only parameter that has to be elicited in the IDP test, the prior strength, by computing its maximum value for which the IDP test remains determinate. This can be interpreted as a measure of robustness of the decision. Results have shown that this test has similar power than that of classical tests and yet is more reliable. The study of the Australian AIDS dataset has demonstrated that the IDP test was able to identify the low reliability of a previous conclusion that the survival of individuals with AIDS from the Queensland region in Australia is significantly worst than for individuals from New South Wales, a result that had been questioned by the curator of the dataset himself.

6. Supplementary Material and Software

Supplementary material is available online at the journal website. It contains details about the numerical approximation of the posterior distribution of $S(t)$ and $P(X < X')$ and about the simulations and proofs of all results presented in this paper.

The `R package IDPsurvival` available from the CRAN Repository provides the software in the form of R code, together with a sample input data set and complete documentation.

Acknowledgments

The authors would like to thank the anonymous Reviewer for its comments and suggestions which have been very helpful to improve the quality of the paper.

This work has been partially supported by the Swiss NSF grant n. 200020-137680/1.

References

[1] B. Ripley, P. Solomon, A note on Australian AIDS survival, Tech. rep., University of Adelaide, department of Statistics Research Report 94/3 (1994).

[2] W. Venables, B. Ripley, Modern applied statistics with S, Springer, 2002.

[3] T. Ferguson, A Bayesian analysis of some nonparametric problems, The Annals of Statistics 1 (2) (1973) 209–230.

[4] V. Susarla, J. Van Ryzin, Nonparametric Bayesian estimation of survival curves from incomplete observations, J. of the American Statistical Association 71 (356) (1976) 897–902.

[5] J. Blum, V. Susarla, On the posterior distribution of a Dirichlet process given randomly right censored observations, Stochastic Processes and their Applications 5 (3) (1977) 207–211.

[6] M. Zhou, Nonparametric Bayes estimator of survival functions for doubly/interval censored data, Statistica Sinica 14 (2) (2004) 533–546.

[7] T. Ferguson, E. Phadia, Bayesian nonparametric estimation based on censored data, The Annals of Statistics 7 (1) (1979) 163–186.

[8] N. Hjort, Nonparametric Bayes estimators based on Beta processes in models for life history data, Annals of statistics 18 (3) (1990) 985–1500.

[9] J. Berger, An overview of robust Bayesian analysis with discussion, Test 3 (1) (1994) 5–124.

[10] J. Berger, D. Rios Insua, F. Ruggeri, Bayesian robustness, in: D. Rios Insua, F. Ruggeri (Eds.), Robust Bayesian Analysis, Vol. 152 of Lecture Notes in Statistics, Springer New York, 2000, pp. 1–32.

[11] L. Pericchi, P. Walley, Robust Bayesian credible intervals and prior ignorance, Int. Statistical Review 59 (1991) 1–23.

[12] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, 1991.

[13] A. Benavoli, F. Mangili, F. Ruggeri, M. Zaffalon, Imprecise Dirichlet process with application to the hypothesis test on the probability that $X < Y$, ArXiv:1402.2755arXiv:1402.2755.

[14] A. Benavoli, G. Corani, F. Mangili, M. Zaffalon, F. Ruggeri, A Bayesian Wilcoxon signed-rank test based on the Dirichlet process, in: Proceedings of the 31st International Conference on Machine Learning, Beijing, 2014.

[15] F. Coolen, An imprecise Dirichlet model for Bayesian analysis of failure data including right-censored observations, Reliability Engineering & System Safety 56 (1) (1997) 61–68.

[16] P. Walley, Inferences from multinomial data: learning about a bag of marbles, J. of the Royal Statistical Society. Series B (Methodological) 58 (1) (1996) 3–57.

[17] E. Letòn, P. Zuluaga, Relationships among tests for censored data, Biometrical J. 47 (3) (2005) 377–387.

[18] N. Mantel, Evaluation of survival data and two new rank order statistics arising in its consideration., Cancer chemotherapy reports. Part 1 50 (3) (1966) 163–170.

[19] H. Mann, D. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Annals of Mathematics and Statistics 18(1) (1947) 50–60.

[20] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin 1(6) (1945) 80–83.

[21] F. Hampel, The influence curve and its role in robust estimation, J. of the American Statistical Association 69 (346) (1974) 383–393.

[22] C. Huber, Robust versus nonparametric approaches and survival data analysis, in: Advances in Degradation Modeling, Springer, 2010, pp. 323–337.

[23] F. Coolen, K. Yan, Nonparametric predictive comparison of two groups of lifetime data, in: International Symposium on Imprecise Probabilities and Their Applications. (ISIPTA), Vol. 3, 2003, pp. 148–161.

[24] E. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, J. of the American Statistical Association 53 (282) (1958) 457–481.

[25] M. Moeschberger, J. Klein, Survival analysis: Techniques for censored and truncated data: Statistics for Biology and Health, Springer, 2003.

[26] D. Rubin, The Bayesian bootstrap, The Annals of Statistics 9 (1) (1981) 130–134.

[27] J.-G. Wang, A note on the uniform consistency of the kaplan-meier estimator, The Annals of Statistics 15 (3) (1987) 1313–1316. doi:10.1214/aos/1176350507.
URL http://dx.doi.org/10.1214/aos/1176350507

[28] J. O. Berger, Statistical decision theory and Bayesian analysis, Springer-Verlag, 1993.

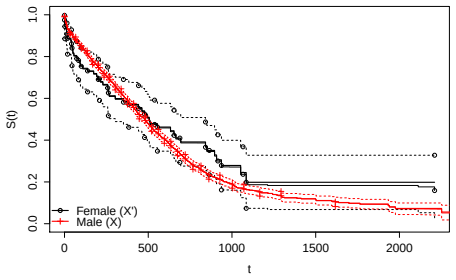
[29] T. Therneau, A Package for Survival Analysis in S, r package version 2.37-7 (2014).
URL http://CRAN.R-project.org/package=survival

[30] National Centre in HIV Epidemiology and Clinical Research, Mapping HIV outcomes: geographical and clinical forecasts of numbers of people living with HIV in Australia, Tech. rep., University of New South Wales (2013).

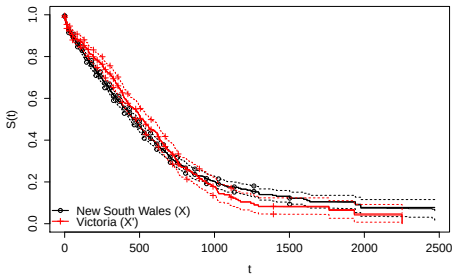
[31] Kirby Institute, HIV, viral hepatitis and sexually transmissible infections in Australia Annual Surveillance Report 2013, Tech. rep., University of New South Wales (2013).

[32] Queensland Health, 2011 HIV/AIDS Report: Epidemiology and Surveillance, Tech. rep., Queensland Government (2011).

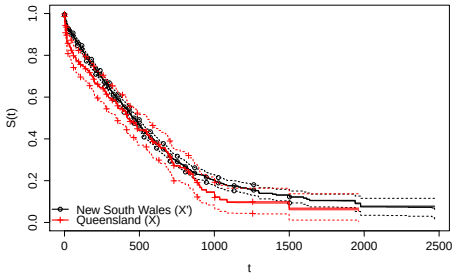
[33] Queensland Health, HIV in Queensland 2012, Tech. rep., Queensland Government (2012).



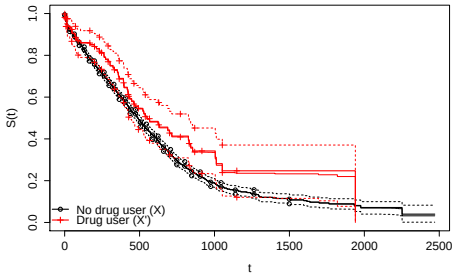
(a) *Male* and *Female*.



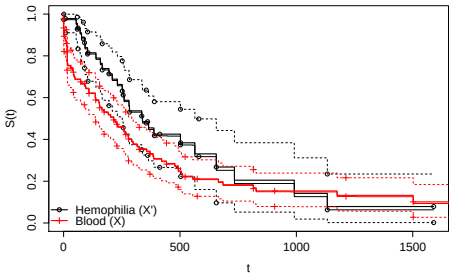
(b) *NSW* and *VIC*.



(c) *NSW* and *QLD*.



(d) *No-drug* and *Drug*.



(e) *Blood* and *Haemophilia*.

Figure 4: Survival curves estimated with IDP ($s = 0.25$) for the groups in Table 4. Continuous lines represent the IDP lower (thin line) and upper (thick line) expectations of $S(t)$; the dotted lines represent the upper and lower bounds of the credible intervals.