
A Predictive Approach to Bayesian Nonparametric Survival Analysis

Edwin Fong
University of Oxford

Brieuc Lehmann
University College London

Abstract

Bayesian nonparametric methods are a popular choice for analysing survival data due to their ability to flexibly model the distribution of survival times. These methods typically employ a nonparametric prior on the survival function that is conjugate with respect to right-censored data. Eliciting these priors, particularly in the presence of covariates, can be challenging and inference typically relies on computationally intensive Markov chain Monte Carlo schemes. In this paper, we build on recent work that recasts Bayesian inference as assigning a predictive distribution on the unseen values of a population conditional on the observed samples, thus avoiding the need to specify a complex prior. We describe a copula-based predictive update which admits a scalable sequential importance sampling algorithm to perform inference that properly accounts for right-censoring. We provide theoretical justification through an extension of Doob’s consistency theorem and illustrate the method on a number of simulated and real data sets, including an example with covariates. Our approach enables analysts to perform Bayesian nonparametric inference through only the specification of a predictive distribution.

1 INTRODUCTION

Survival data, also known as time-to-event data, is ubiquitous in a number of domains including economics, engineering, biology, and medicine. Common examples include the time to failure of a mechanical component, or the time to death of an individual following treatment. The overarching aim of survival

analysis is to study the distribution of these survival times. In survival regression, the aim is to assess the effect of covariates on survival time.

A characteristic feature of survival data is that it is often censored - that is, we may not know the survival time exactly. In the case of *right-censoring*, we only observe the information $Y > c$, where Y is the time-to-event of interest and c is the observed censoring time. Right-censoring can occur, for example, if a subject leaves a study before the event of interest has occurred. The partial nature of the information associated with the observed data poses some challenges to statistical inference.

A primary goal in survival analysis is to *predict* the survival time for a new individual, perhaps taking into account known covariates (e.g. age) for said individual. In other words, the aim is to learn a predictive distribution $p(y_{n+1}|x_{n+1}, \{x_i, y_i\}_{i=1:n})$, where $\{x_i, y_i\}_{i=1:n}$ is an observed training set. To reduce notational burden, we henceforth omit reference to covariates x . The standard Bayesian approach to this problem is to first specify a data-generating distribution $f_\theta(y)$, depending on a (potentially infinite-dimensional) parameter θ , and prior $\pi(\theta)$. The predictive distribution is then taken to be the posterior predictive distribution. In the uncensored case, this is

$$p(y_{n+1}|y_{1:n}) = \int f_\theta(y_{n+1}) \pi(\theta|y_{1:n}) d\theta, \quad (1)$$

where $\pi(\theta|y_{1:n}) \propto \pi(\theta) \prod_{i=1}^n f_\theta(y_i)$ is the posterior distribution, which is often also of interest.

Here, we take a more direct approach to prediction and posterior inference by explicitly specifying a predictive distribution instead of the usual likelihood and prior. In particular, we extend the notion of *martingale posterior distributions* (Fong et al., 2021) to right-censored data, appropriately accounting for the partially observed nature of the censored values. In doing so, we leverage one of the key advantages of the martingale posterior framework in replacing the standard Markov chain Monte Carlo (MCMC) approach to posterior computation with a GPU-friendly and parallelisable optimisation-based algorithm. Our main contri-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

butions are as follows: a) we describe a class of copula-based predictive updates that are suitable under right-censoring; b) we extend Doob’s consistency theorem to the setting with right-censored observations, confirming the conceptual equivalence of standard Bayesian inference and the martingale posterior in this setting; c) to perform inference, we develop a sequential importance sampling (IS) procedure, avoiding the need for more computationally intensive MCMC algorithms.

2 RELATED WORK

There is a rich history of Bayesian nonparametric methods for the analysis of survival data. These typically employ a neutral-to-the-right (NTR) process (Doksum, 1974) prior on the survival function, chosen for its conjugacy property with respect to censored data (Ferguson & Phadia, 1979). Some examples of such priors include the extended gamma process (Kalbfleisch, 1978), the beta process (Hjort et al., 1990), and the beta-Stacy process (Muliere & Walker, 1997b). Muliere and Walker (1997a) offered a generalisation of the beta process based on a Pólya tree prior. Yet another alternative approach was taken by Kottas (2006), who modelled the distribution of survival times using a Dirichlet process mixture model (DPMM) with a Weibull kernel. Our copula-based predictive update is intimately linked to the DPMM (see Section 3.2).

Building on these foundations, extensions to survival regression have been developed based on proportional hazard models, for example by Hjort et al. (1990), Kalbfleisch (1978), and Kim and Lee (2003). Riva-Palacio et al. (2021) relax the restriction of proportionality through the use of a vector of completely random measures. De Iorio et al. (2009) developed a dependent Dirichlet process (DDP) mixture model for survival regression that also permits survival curves to cross in the context of a treatment effect analysis. Further examples can be found in Ghosal and van der Vaart (2017, Chapter 13).

The idea of focusing directly on the specification of a predictive distribution goes back to at least Hill (1968), who posited a uniform distribution on the intervals between the order statistics of the observations. Extensions of Hill’s predictive distribution to censored data have been proposed by Berliner and Hill (1988) and Coolen and Yan (2004). We build on recent work that proposes to relax the assumption of exchangeability in favour of conditionally identically distributed (Berti et al., 2004) sequences, thus allowing for more flexible specifications of the predictive distribution (Berti et al., 2021). In particular, we focus on one-step-ahead predictive updates based on bivariate copulas, initially proposed in Hahn et al. (2018) for the uncensored

case. As noted in Fong et al. (2021), there are also connections between this predictive approach and the Bayesian bootstrap (Rubin, 1981) and its extensions to censored data (Arfê & Muliere, 2020; Lo et al., 1993).

3 BACKGROUND

In this section, we describe the martingale posterior distribution framework in the *uncensored* independently and identically distributed (i.i.d.) data setting, as introduced by Fong et al. (2021). In this work, Bayesian inference is reframed as an imputation problem, where the task is to elicit the joint predictive density on the missing information, which is the remainder of the population $y_{n+1:\infty}$ given an observed sample $y_{1:n}$ in the i.i.d. case. The joint density of interest can be written as a product of 1-step-ahead predictive densities,

$$p(y_{n+1:N} | y_{1:n}) = \prod_{i=n+1}^N p_{i-1}(y_i), \quad (2)$$

where we write $p_i(y) := p(y | y_{1:i})$ with corresponding cumulative distribution function (CDF) $P_i(y)$. Intuitively, a general statistic of interest can then be recovered from the population $y_{1:N}$, which is written as $\theta(y_{1:N})$. The predictive uncertainty in $y_{n+1:N}$ then induces a distribution on $\theta(y_{1:N})$. We will formalize these notions later on.

For the parametric Bayesian with sampling density $f_\theta(y)$ and prior $\pi(\theta)$, the posterior predictive density $p_i(y)$ is defined as in (1). The statistic is then an estimate of θ indexing the sampling density, e.g. the posterior mean, $\bar{\theta}_N = E[\Theta | y_{1:N}]$, where Θ is the Bayesian random parameter that is marginally distributed according to the prior π . With this choice, it can be shown through Doob’s consistency theorem (Doob, 1949) that the above scheme is equivalent to posterior sampling in the limit of $N \rightarrow \infty$, that is $\bar{\theta}_\infty \sim \pi(\theta | y_{1:n})$, where $\bar{\theta}_\infty := \lim_{N \rightarrow \infty} \bar{\theta}_N$. Through this result, parameters are viewed as functions of the population of observables, and Bayesian uncertainty can intuitively be seen to arise from subjective uncertainty on the missing remainder of the population.

3.1 Martingale Posterior Distributions

The martingale posterior distribution considers more general sequences of predictive distributions than that induced by the likelihood and prior, and is hence a generalisation of standard Bayesian inference. Given a sequence of predictive CDFs $P_n(y), P_{n+1}(y), \dots$, one can impute the remainder of the infinite population through the scheme

$$Y_{n+1} \sim P_n(y), \quad Y_{n+2} \sim P_{n+1}(y), \quad \dots$$

This sequential imputation scheme is termed *predictive resampling*, as it is inspired by the Pólya urn scheme (Blackwell & MacQueen, 1973) for the Bayesian bootstrap. In practice, it is infeasible to work with infinite populations, so we terminate predictive resampling at Y_N for some large $N > n$. However, there are still constraints on this sequence P_i needed to ensure a notion of predictive coherence, and in particular for the random limiting empirical distribution to exist so that we can compute a functional of interest. The limiting empirical distribution is given by

$$F_\infty(y) = \lim_{N \rightarrow \infty} \frac{1}{N} \left\{ \sum_{i=1}^n \mathbb{1}(y_i \leq y) + \sum_{i=n+1}^N \mathbb{1}(Y_i \leq y) \right\}.$$

A sufficient condition for the existence of F_∞ is that the sequence P_i implies a *conditionally identically distributed* (c.i.d.) sequence of random variables (r.v.s), as investigated in Berti et al. (2004). The sequence Y_{n+1}, Y_{n+2}, \dots is c.i.d. if

$$P(Y_{i+k} \leq y \mid y_{1:i}) = P_i(y), \quad \forall k > 0.$$

This ensures that the sequence of predictive distributions P_i is a martingale, and that predictive resampling returns a well-defined empirical distribution.

Moreover, the parameter of interest no longer needs to index a sampling density. Instead, it can be defined as

$$\theta_0 := \theta(F_0) = \arg \min_{\theta} \int \ell(\theta, y) dF_0(y)$$

where F_0 is the true sampling density and the loss function $\ell(\theta, y)$ is elicited by the analyst. After predictive resampling, a sample from the martingale posterior can be recovered by computing $\theta_\infty = \theta(F_\infty)$.

3.2 Bivariate Copula Updates

We now discuss a concrete example of a c.i.d. sequence of predictive densities that is both computationally feasible and no longer relies on the likelihood and prior. A useful class of predictive densities depends on the *bivariate copula density*, and was introduced in Hahn et al. (2018). Briefly, the bivariate copula density is a bivariate probability density function with uniform marginals, that is $d : [0, 1]^2 \rightarrow \mathbb{R}$ where $\int d(u, v) du = \int d(u, v) dv = 1$. See Nelsen (2007) for more details. For univariate data, a sequence of predictive densities can be defined recursively through

$$p_i(y) = d_i \{P_{i-1}(y), P_{i-1}(y_i)\} p_{i-1}(y). \quad (3)$$

Here, d_i is a sequence of bivariate copula densities which models the dependence between Y_i and Y_{i+1} . Hahn et al. (2018) showed that all Bayesian predictives have an update of this form, although it is usually intractable for non-conjugate models. A tractable

sequence of copula densities is then introduced, inspired by the DPMM, which does not correspond to a Bayesian model. Exploiting the c.i.d. property of this update, Fong et al. (2021) explore their use in the context of martingale posteriors and provide further extensions to multivariate data and regression.

3.2.1 Bivariate Copula Updates on \mathbb{R}^+

The updates introduced in Hahn et al. (2018) are applicable to data with support on the entire real line \mathbb{R} , and are motivated by the DPMM with the normal kernel. Survival times, however, are typically strictly positive, so we will now introduce a copula update for data supported on the positive reals, \mathbb{R}^+ .

We begin by introducing said copula update in the absence of censoring, which is motivated by the first posterior predictive update of the DPMM with an exponential kernel. The DPMM can be written as

$$f_G(y) = \int \text{Exp}(y \mid \theta) dG(\theta)$$

$$G \sim \text{DP}(c, G_0), \quad G_0 = \Gamma(\theta \mid a, 1).$$

where $\text{Exp}(y \mid \theta)$ is the exponential density with rate θ , and $\Gamma(\cdot \mid a, 1)$ is the gamma density with shape a and rate 1. In Appendix B, we show that this inspires the update

$$p_i(y) = [1 - \alpha_i + \alpha_i d_a \{P_{i-1}(y), P_{i-1}(y_i)\}] p_{i-1}(y)$$

$$d_a(u, v) = \frac{a + 1}{a} \frac{(1 - u)^{-\frac{a+1}{a}} (1 - v)^{-\frac{a+1}{a}}}{\left\{ (1 - u)^{-\frac{1}{a}} + (1 - v)^{-\frac{1}{a}} - 1 \right\}^{a+2}}. \quad (4)$$

In fact, the above update corresponds exactly to the DPMM update from $p_0 \rightarrow p_1$, but is different for p_i with $i > 1$. The sequence α_i should in general be $\mathcal{O}(i^{-1})$ to approach the independent copula for consistent estimation, and the specific suggestion of $\alpha_i = (2 - 1/i)/(i + 1)$ is motivated in Fong et al. (2021). Note that the above is a mixture of the independence copula and the Clayton copula (Clayton, 1978), as was also pointed out in Hahn et al. (2018). See Balakrishnan and Lai (2009, Chapter 2.9) for more details. The update for the CDF $P_i(y)$ is similarly tractable and is derived in Appendix B.

Here, $a > 0$ acts as a bandwidth term, where smaller values indicates a stronger peak; the update in (4) is analogous to a kernel density estimate but on \mathbb{R}^+ . This is illustrated in Figure 1a in which we plot the copula kernel $d_a(u_{i-1}, v_{i-1}) p_{i-1}(y)$ for decreasing values of a , where $u_{i-1} = P_{i-1}(y), v_{i-1} = P_{i-1}(y_i)$. The updated density is a weighted mixture of p_i (dashed) and the copula kernel (solid), which is shown in Figure 1b.

In the case of regression with covariates $\mathbf{x} \in \mathbb{R}^d$, a similar argument based on the DDP mixture model

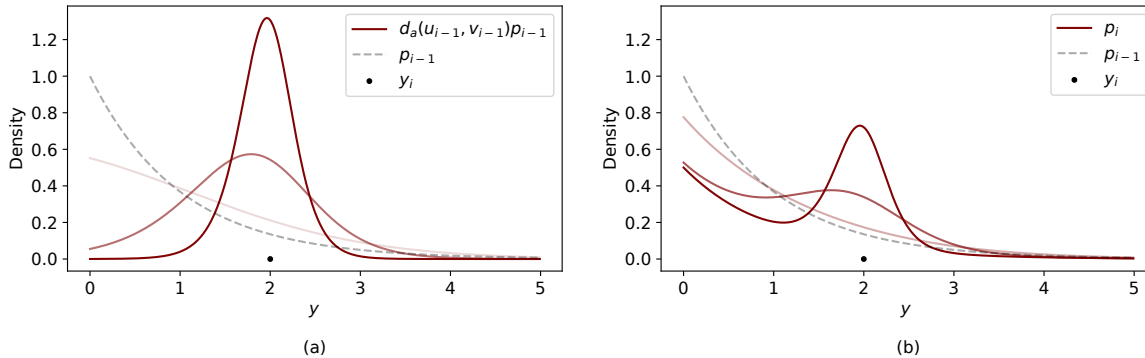


Figure 1: Plot of the (a) copula kernel $d_a(u_{i-1}, v_{i-1}) p_{i-1}(y)$ and (b) updated predictive $p_i(y)$, for $a = 2, 0.5, 0.2$ (—, —, —).

can be used to derive an update for the conditional density $p_i(y | \mathbf{x})$. This takes on the form

$$p_i(y | \mathbf{x}) = \{1 - \alpha_i(\mathbf{x}, \mathbf{x}_i) + \alpha_i(\mathbf{x}, \mathbf{x}_i) d_a(q_{i-1}, r_{i-1})\} p_{i-1}(y | \mathbf{x}) \quad (5)$$

where $q_{i-1} = P_{i-1}(y | \mathbf{x})$, $r_{i-1} = P_{i-1}(y_i | \mathbf{x}_i)$. The exact form of the function $\alpha_{i+1}(\mathbf{x}, \mathbf{x}_{i+1})$, provided in Appendix B, can be derived from the multivariate copula update (Fong et al., 2021). Intuitively, $\alpha_i(\mathbf{x}, \mathbf{x}_i)$ weights the copula kernel based on the distance between the covariate of interest \mathbf{x} and the updating datum \mathbf{x}_i .

3.2.2 Practical Details

We now review the practical details discussed in Fong et al. (2021). To estimate $p_n(y)$ before predictive resampling, we need to begin with $p_0(y)$, which acts as our prior guess of the true density. A choice that works well in practice is to set $p_0(y) = \text{Lomax}(a, 1)$, which matches the DPMM. Here a is the bandwidth parameter, which we can set by maximizing the prequential log-likelihood (Dawid, 1984), $\sum_{i=1}^n \log p_{i-1}(y_i)$. Fitting the copula method has a computational complexity of $\mathcal{O}(n^2)$, as we must first compute the overhead terms $P_{i-1}(y_i)$ for $i = 1, \dots, n$. Given these terms, computing $p_n(y)$ at any value of interest is then $\mathcal{O}(n)$.

A key property of the copula methods is the convenience of predictive resampling. The copula update for p_{n+1} only depends on Y_{n+1} through $P_n(Y_{n+1})$, and predictive resampling involves drawing $Y_{n+1} \sim P_n$. As a result, we only need to draw $P_i(Y_{i+1}) \stackrel{\text{iid}}{\sim} \mathcal{U}[0, 1]$ for $i = n, \dots, N - 1$, and compute the copula update (4) appropriately. Drawing a sample of $p_N(y)$ at some test point is thus $\mathcal{O}(N - n)$. In the regression case, we can draw $\mathbf{X}_{n+1:N}$ from the Bayesian bootstrap, and $P_i(Y_{i+1} | \mathbf{X}_{i+1})$ can be similarly drawn from the uniform distribution as in the no-covariate case.

4 PREDICTIVE RESAMPLING UNDER RIGHT-CENSORING

The above description assumed that we observed each of the survival times exactly. We are now ready to extend the predictive resampling framework to right-censored data. Suppose we have observed the dataset $\mathcal{D}_n := \{y_{1:k}, Y_{k+1:n} \geq c_{k+1:n}\}$, where for convenience we have ordered the data such that the first k are observed and the remaining are right-censored. Throughout the remainder of this work, we assume that the censoring mechanism is *non-informative* - that is, we treat $c_{k+1:n}$ as constants in conditional probability statements. See Berliner and Hill (1988) and Appendix A for more details on the relevant assumptions in the Bayesian and predictive cases. Once again, the Bayesian requires $y_{1:N}$ (for $N \rightarrow \infty$) to compute any statistic of interest, and so it is natural that the Bayesian elicits the predictive density

$$p(y_{k+1:N} | \mathcal{D}_n) \quad (6)$$

on $Y_{k+1:N}$ which is missing. In contrast to the uncensored case however, $Y_{k+1:n}$ is partially observed. The key is to factorize (6) into

$$p(y_{n+1:N} | y_{1:n}) p(y_{k+1:n} | \mathcal{D}_n), \quad (7)$$

and so predictive resampling consists of the following:

1. Impute $Y_{k+1:n} \sim p(y_{k+1:n} | \mathcal{D}_n)$.
2. Predictive resample $Y_{n+1:N} \sim p(y_{n+1:N} | y_{1:n})$ as before.
3. Compute a statistic of interest $\theta(Y_{1:N})$.

The distribution of $\theta(Y_{1:N})$ is then approximately our martingale posterior distribution $\pi_\infty(\theta | \mathcal{D}_n)$, where the subscript is used to distinguish from the regular Bayesian posterior. We note that the exact martingale posterior distribution would involve computing

the functional of the limiting empirical distribution; see Appendix A for more details. We also highlight the connection to the multiple imputation framework of Rubin (1996), where the full posterior $\pi(\theta \mid y_{1:n})$ is replaced by $p(y_{n+1:N} \mid y_{1:n})$ and the imputing predictive is given by $p(y_{k+1:n} \mid y_{1:k}, Y_{k+1:n} \geq c_{k+1:n})$.

4.1 Doob's Consistency Theorem for Right-censored Observations

As discussed above for the i.i.d. setting with fully observed data, it follows from Doob's consistency theorem (Doob, 1949) that predictive resampling with the parametric posterior predictive distribution is equivalent to posterior sampling in the limit of $N \rightarrow \infty$. We now extend this result to the case where some of the observations are right-censored, as is typical for survival data.

Assume that for all N , the r.v.s $[\Theta, Y_1, \dots, Y_N]$ have joint density

$$p(\theta, y_{1:N}) = \pi(\theta) \prod_{i=1}^N f_{\theta}(y_i).$$

Denote the posterior mean as $\bar{\theta}_N = E[\Theta \mid Y_{1:N}]$ (for Θ in a linear space), and $f_{\theta}^c(y) = \mathbb{1}\{y \geq c\} f_{\theta}(y) / \bar{F}_{\theta}(c)$ to be the density of a data point right-censored at c , where \bar{F}_{θ} is the survival function of f_{θ} .

We draw $Y_{k+1:n} \sim p(y_{k+1:n} \mid \mathcal{D}_n)$ where

$$p(y_{k+1:n} \mid \mathcal{D}_n) = \int \prod_{i=k+1}^n f_{\theta}^{c_i}(y_i) \pi(\theta \mid \mathcal{D}_n) d\theta,$$

and $\pi(\theta \mid \mathcal{D}_n) \propto \pi(\theta) \prod_{i=1}^k f_{\theta}(y_i) \prod_{i=k+1}^n \bar{F}_{\theta}(c_i)$, which follows from the non-informative censoring.

We then draw $Y_{n+1:N} \sim p(Y_{n+1:N} \mid y_{1:n})$ where

$$p(y_{n+1:N} \mid y_{1:n}) = \int \prod_{i=n+1}^N f_{\theta}(y_i) \pi(\theta \mid y_{1:n}) d\theta,$$

and compute $\bar{\theta}_N$ from $Y_{1:N}$. The following result establishes the equivalence of predictive resampling and standard Bayesian inference as $N \rightarrow \infty$.

Theorem 1. *Assume $E[|\Theta| \mid \mathcal{D}_n] < \infty$. Under regularity conditions on f_{θ} , we have that*

$$\lim_{N \rightarrow \infty} \bar{\theta}_N = \Theta \quad \text{a.s. } P^{\infty}(\cdot \mid \mathcal{D}_n). \quad (8)$$

where P^{∞} is over Θ and $Y_{k+1:\infty}$.

Proof. See Appendix A. \square

Similarly to Fong et al. (2021), the above theorem directly links Bayesian uncertainty in the parameter,

represented by $\Theta \sim \pi(\theta \mid \mathcal{D}_n)$, to the statistical uncertainty in the *partially* observed $Y_{k+1:n}$ and unobserved $Y_{n+1:\infty}$. This can be seen by considering the following two distinct methods of sampling Θ from the posterior. The first is the standard Bayesian approach to draw $\Theta \sim \pi(\theta \mid \mathcal{D}_n)$ directly. The second, predictive resampling, begins by first imputing the partially observed data points $Y_{k+1:n}$ from the joint density $p(y_{k+1:n} \mid \mathcal{D}_n)$ followed by the completely unseen observables $Y_{n+1:\infty}$ from the sequence of predictive densities

$$Y_{n+1} \sim p(\cdot \mid y_{1:n}), Y_{n+2} \sim p(\cdot \mid y_{1:n+1}), \dots,$$

until we have the complete information $Y_{1:\infty}$. Given $Y_{1:\infty}$, we can then compute the limiting estimate $\bar{\theta}_{\infty} = \lim_{N \rightarrow \infty} \bar{\theta}_N$, which is the posterior mean, on the entire dataset. By the above theorem, this returns $\bar{\theta}_{\infty} \sim \pi(\theta \mid \mathcal{D}_n)$.

We emphasize that the purpose of Theorem 1 as outlined above is to provide a conceptual illustration that, in the Bayesian *parametric* case, the uncertainty in a point estimator $\bar{\theta}_N$ computed from imputed observations is equivalent to uncertainty in the Bayesian random parameter Θ . The choice of the posterior mean $\bar{\theta}_N$ as the estimator is one of mathematical convenience, allowing us to directly leverage the result of Doob (1949); it may not be of practical use when the posterior mean is not analytically available. In the more general martingale posterior case, the c.i.d. property guarantees the existence of the limiting empirical distribution, F_{∞} , under our imputation and predictive resampling scheme, again relying on martingales in an analogous way to Doob's theorem. We can then compute the functional of interest, $\theta(F_{\infty})$, to obtain a posterior sample. To show this in the c.i.d. case, we condition on $Y_{k+1:n}$ and utilize the properties of the original c.i.d. sequence in a similar way to Theorem 1. Further details can be found in Appendix A.

5 COPULA UPDATES UNDER RIGHT-CENSORING

The copula updates introduced in Section 3.2 assumed that observations were fully known. We now extend these methods to the right-censored case. For the purposes of exposition, we will continue to treat the first k data points $y_{1:k}$ as uncensored, with the remaining $y_{k+1:n}$ as right-censored at $c_{k+1:n}$. In practice however, a random ordering is usually preferred, and we highlight that the copula methods are not exchangeable. See Appendix B for further discussion on ordering.

If the aim is to predict survival outcomes for a new individual given right-censored observations, the quan-

tity of interest is the predictive density

$$p(y_{n+1} \mid \mathcal{D}_n). \quad (9)$$

This can be written as

$$\int p(y_{n+1} \mid y_{1:n}) p(y_{k+1:n} \mid \mathcal{D}_n) dy_{k+1:n}, \quad (10)$$

which can be computed via Monte Carlo, where $p(y_{n+1} \mid y_{1:n})$ is available through (4). To obtain both the martingale posterior and predictive density, we will now develop a method to simulate from

$$p(y_{k+1:n} \mid \mathcal{D}_n). \quad (11)$$

5.1 Importance Sampling

To simulate from (11) given a prescribed sequence $\{p_{i-1}(y_i)\}_{i=1:n}$, we draw inspiration from Kong et al. (1994), which considered this problem for fully missing data by sequential imputation followed by importance reweighting. We now introduce the methodology in the particular case of right-censored data.

For the first k data points, the update (4) can be used recursively to obtain $p_k(y_{k+1})$. When we reach the first censored datum $Y_{k+1} \geq c_{k+1}$, we cannot directly update the predictive density as it requires the value of Y_{k+1} . An intuitive, but incorrect, solution is as follows: impute $Y_{k+1} \sim p(y_{k+1} \mid Y_{k+1} \geq c_{k+1}, y_{1:k})$, then treat the sampled Y_{k+1} as an observed value to update to p_{k+1} via (4). Then, draw Y_{k+2} from $p_{k+1}(y_{k+2}) = p(y_{k+2} \mid y_{1:k}, Y_{k+1}, Y_{k+2} \geq c_{k+2})$ and continue on in a sequential manner until we have $Y_{k+1:n}$.

However, this sample is not drawn from (11). In short, this is because we have not used the future censored information $\{Y_j \geq c_j\}_{j>i}$ when imputing Y_i , for $i = k+1, \dots, n$. To correct for this, we can use IS, treating $Y_{k+1:n}$ as a proposal sample. Assuming non-informative censoring, the importance weights can be derived through the factorization of (11) into

$$\begin{aligned} & \frac{\prod_{i=k+1}^n p(y_i, Y_i \geq c_i \mid y_{1:i-1})}{p(Y_{k+1:n} \geq c_{k+1:n} \mid y_{1:k})} \\ & \propto \underbrace{\prod_{i=k+1}^n p(y_i \mid Y_i \geq c_i, y_{1:i-1})}_{\text{Proposal}} \underbrace{\prod_{i=k+1}^n P(Y_i \geq c_i \mid y_{1:i-1})}_{\text{Unnormalized IS weights}}. \end{aligned} \quad (12)$$

In the above, we have used the notation

$$\begin{aligned} p(x, Y \geq c) & := p(x \mid Y \geq c) P(Y \geq c) \\ & = P(Y \geq c \mid x) p(x), \end{aligned}$$

to represent the mixed joint density of the observed values and censored events, where x is a continuous

r.v. and $\mathbb{1}(Y \geq c)$ can be considered as a discrete r.v.. The proposal is the joint density of $Y_{k+1:n}$ drawn from the scheme above, and the IS weight can be computed as we sequentially impute, since it depends only on p_i .

For the copula method specifically, the proposal is efficient to simulate from sequentially in a rejection-free manner. Writing $p_{i-1}^{c_i}(y) = p(y \mid Y_i \geq c_i, y_{1:i-1})$, we note that $p_{i-1}^{c_i}(y) \propto \mathbb{1}(y \geq c_i) p_{i-1}(y)$ from non-informative censoring. Working in the space of CDFs, simulating $Y_i \sim p_{i-1}^{c_i}$ is equivalent to drawing

$$U_i \sim \mathcal{U}[P_{i-1}(c_i), 1], \quad Y_i = P_{i-1}^{-1}(U_i).$$

However, we note that the update (4) depends on Y_i only through $U_i = P_{i-1}(Y_i)$, and so we can utilize U_i directly without computing P_{i-1}^{-1} . The term $P_{i-1}(c_i)$ is used both in the proposal and the IS weight, and can be computed exactly as a tractable update also exists for the CDF sequence.

Given B samples from the proposal $\{Y_{k+1:n}^{(j)}\}_{j=1:B}$ and self-normalized IS weights given by

$$w^{(j)} = \prod_{i=k+1}^n \left[1 - P_{i-1}^{(j)}(c_i) \right], \quad \tilde{w}^{(j)} = w^{(j)} / \sum_{j=1}^B w^{(j)},$$

we can then approximate (9) with

$$\hat{p}(y_{n+1} \mid \mathcal{D}_n) = \sum_{j=1}^B \tilde{w}^{(j)} p_n^{(j)}(y_{n+1}),$$

where $p_n^{(j)}$ is the random predictive density computed from $\{y_{1:k}, Y_{k+1:n}^{(j)}\}$ through (4). Similarly, we can approximate the martingale posterior through

$$\hat{\pi}_\infty(\theta \mid \mathcal{D}_n) = \sum_{j=1}^B \tilde{w}^{(j)} \delta_{\theta_N^{(j)}}$$

where $\theta_N^{(j)} = \theta(y_{1:k}, Y_{k+1:n}^{(j)})$ and the unobserved $Y_{n+1:N}^{(j)} \sim p(y_{n+1:N} \mid y_{1:k}, Y_{k+1:n}^{(j)})$ are simulated through regular predictive resampling after imputing $Y_{k+1:n}^{(j)}$.

5.2 Sequential Monte Carlo

If the number of missing data points $n - k$ is large, IS may fail due to the dimensionality of the proposal. To mitigate the exponential variance increase of vanilla IS, we can use sequential Monte Carlo (SMC), noting that the importance weights have a straightforward online update. This induces additional resampling steps of $\{w^{(1:B)}, Y_{k+1:i}^{(1:B)}\}$ at time points i when the effective sample size (ESS) is too low, e.g. less than 50% of the original number of particles. In practice, we find

SMC to drastically improve the performance of our method for larger values of $n - k$ for a minor increase in computation. See Doucet, Johansen, et al. (2009) for more details.

5.3 Algorithm and Practical Details

In practice, we find that randomizing the order of data greatly increases the ESS in comparison to ordering the uncensored data before the censored data. However, the IS weights in this case have a slightly different form to take into account observed data points between censored data points. This is shown in Algorithm 1, using the notation $\delta_i = 1$ to indicate that y_i is observed and $\delta_i = 0$ to indicate that Y_i is right-censored at c_i . See Appendix B for the derivation and more details on the impact of ordering on the ESS.

To select the bandwidth a , we can maximize the joint likelihood of the observations, $p(\mathcal{D}_n)$, which can be computed with SMC (Appendix B). As we are still required to compute $\{P_{i-1}(\delta_i y_i + (1 - \delta_i)c_i)\}_{i=1:n}$ for each particle, the total complexity of Algorithm 1 is $\mathcal{O}(Bn^2)$, followed by $\mathcal{O}(Bn)$ for each prediction. Details on the selection of the number of forward samples N to sufficiently approximate the infinite population are given in Appendix B.

Algorithm 1: Survival Copula Imputation

```

Initialize  $p_0$  and  $w^{(1:B)} = 1$ 
for  $i \leftarrow 1$  to  $n$  do
    for  $j \leftarrow 1$  to  $B$  do
        if  $\delta_i = 1$  then
            Update  $w^{(j)} = w^{(j)} \times p_{i-1}^{(j)}(y_i)$ 
            Update  $p_i^{(j)} \leftarrow \{p_{i-1}^{(j)}, y_i\}$  from (4)
        end
        if  $\delta_i = 0$  then
            Sample  $Y_i^{(j)} \sim p_{i-1}^{c_i}$ 
            Update  $w^{(j)} = w^{(j)} \times [1 - P_{i-1}^{(j)}(c_i)]$ 
            Update  $p_i^{(j)} \leftarrow \{p_{i-1}^{(j)}, Y_i^{(j)}\}$  from (4)
        end
    end
    if  $ESS(w^{(1:B)}) < 0.5B$  then
        Resample  $\{1/B, \bar{p}_n^{(1:B)}\} \leftarrow \{w^{(1:B)}, p_n^{(1:B)}\}$ 
    end
end
Return  $\{w^{(1:B)}, p_n^{(1:B)}\}$ 
    
```

5.4 Covariates

In the context of survival regression, we are interested in the effect of observed covariates $\mathbf{x}_{1:n}$ on survival outcomes. The density of interest is

now $p(y_{k+1:n} | \mathcal{D}_n, \mathbf{x}_{1:n})$. Given a tractable sequence of conditional densities, $p_i(y | \mathbf{x})$, the importance reweighting method above generalizes easily (Appendix B). With (ignorable) missing covariates, our reweighting scheme can be combined with that of Kong et al. (1994). Note that the terms $\alpha_i(\mathbf{x}, \mathbf{x}_i)$ also depend on a hyperparameter ρ_x . We can fit $\{a, \rho_x\}$ jointly by maximizing the conditional prequential log-likelihood.

6 EXAMPLES

We illustrate our approach on a simulated data example and three real data examples, including two with covariates, comparing our approach to common Bayesian nonparametric survival analysis methods.

All copula examples are implemented in JAX (Bradbury et al., 2018) and run on an Azure NC6 Virtual Machine with a one-half Tesla K80 GPU card (with compilation times < 5 s). The DPMM examples are run on a 2.4 GHz 8-Core Intel Core i9-9980HK using the R packages `dirichletprocess` (Ross & Markwick, 2018) and `ddpanova` (De Iorio et al., 2004). For all examples, we have $B = 2000$ IS/MCMC samples and set $N = 2000 + n$ for the number of future samples which is sufficiently large for convergence (Appendix C). For the copula methods, we use a single random permutation of the data for each run and fit the bandwidth automatically by maximizing the prequential log-likelihood. The code and data used is available online¹. Further details, such as evaluation of the ESS and standardization, are provided in Appendix C.

6.1 Simulated Data

We begin by providing an empirical illustration of Theorem 1, that is Doob's consistency theorem for right-censored observations, through a simulated data example with a Bayesian *parametric* predictive. We generated data

$$Y_i \sim \text{Exp}(1), \quad C_i \sim \text{Exp}(2), \quad i = 1, \dots, n,$$

and right-censor if $y_i \geq c_i$, with $n = 50$. Around 76% of data points were right-censored. We consider fitting this data with an $\text{Exp}(1/\theta)$ sampling density under a conjugate inverse-gamma prior $\text{IG}(\theta | a_0, b_0)$. The posterior is then $\pi(\theta | \mathcal{D}_n) = \text{IG}(a_n, b_n)$, where $a_n = a_0 + k$, $b_n = b_0 + \sum_{i=1}^k y_i + \sum_{i=k+1}^n c_i$, and the posterior predictive is also analytically tractable as the $\text{Lomax}(a_n, b_n)$ distribution (Appendix C).

For the inverse-gamma prior, we set $b_0 = 1$ and select $a_0 = 1.2$ by maximizing the marginal likelihood. We perform the imputation of the censored data points as

¹https://github.com/edfong/survival_mp

in Algorithm 1, noting that the importance weights are available via the Lomax CDF. This is then followed by regular predictive resampling (see Appendix B). Figure 2 illustrates the close agreement between the standard Bayesian posterior and the martingale posterior induced by the Lomax posterior predictive distribution, as expected from Theorem 1.

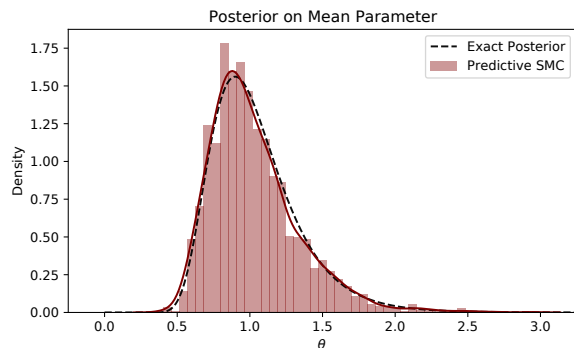


Figure 2: Exact Bayesian posterior vs. martingale posterior with parametric Lomax predictive generated via Algorithm 1.

6.2 Primary Biliary Cirrhosis

We now shift attention to survival data from a randomized clinical trial on $n = 312$ patients with primary biliary cirrhosis (PBC) (Dickson et al., 1989), available in R through the `survival::pbc` dataset. A total of 158 patients received D-penicillamine, while the remaining 154 patients received a placebo. We compared our predictive resampling approach using the nonparametric exponential copula update with a DPMM using an exponential kernel. In particular, we focus on the predictive accuracy of these methods, evaluating performance on each of the trial arms separately. We first applied 10 random 50-50 train-test splits to the data, fit each model to the training set, then computed the mean log-likelihood of the resulting fit on the test set, which contains both censored and uncensored data points. The predictive accuracy of the two methods is almost identical (Table 1), and full details can be found in the Appendix C.

In Figure 3a, we also plot the posterior mean and 95% credible intervals of the survival function for the placebo arm, fitted to all 154 data points. We see that the results are similar, but the copula method has wider credible intervals. See Appendix C for posterior plots of the nonparametric density. In Figure 3b, we plot posterior samples of the median, which is again similar. The copula method required 3.6s to optimize for hyperparameters and fit the data, and a further 0.9s for predictive resampling on a y -grid of size 149. In contrast, the DPMM took around 2 minutes.

Table 1: Average Test Log-likelihood with Standard Errors (in Brackets) on the Two Arms of the PBC Dataset.

Dataset	Copula (exp)	DPMM (exp)
PBC (treatment)	-0.44 (0.02)	-0.43 (0.02)
PBC (placebo)	-0.39 (0.02)	-0.39 (0.03)

6.3 Survival Regression

Next, we illustrate our method for survival data in the presence of covariates. For this purpose, we analysed two datasets. First, we analysed data on $n = 205$ patients in Denmark with malignant melanoma, using tumour thickness as a covariate. Second, we analysed survival data of $n = 863$ kidney transplant patients at The Ohio State University Transplant Center from 1982 to 1992, using patient age as a covariate. These datasets are available in R from the MASS (Venables & Ripley, 2002) and KMsurv (Klein et al., 2012) packages respectively.

For the baselines, we fit a DPMM with a log-normal kernel and an accelerated failure time (AFT) model with log-normal noise. For a fair comparison, we utilize a variant of the copula update, substituting the Clayton copula in (4) with the Gaussian copula, and setting $p_0(y) = \text{Log-normal}(y | 0, 1/(1 - \rho))$. This corresponds to a copula update based on the log-normal DPMM; more details can be found in Appendix B.

Once again, we carried out 10 random 50-50 train-test splits and evaluated the predictive log-likelihood on the test set (Table 2). We see that the copula method performs the best for the melanoma dataset, but slightly worse than the other methods for the kidney dataset. Optimization, fitting and prediction for the copula method required around 3s and 14s for the melanoma and kidney dataset respectively, compared to 10s and 76s respectively for the DDP mixture, for each train-test split. It is also possible to predictively resample in the regression context (Appendix B).

Table 2: Average Test Log-likelihood with Standard Errors (in Brackets) on the Melanoma and Kidney Datasets.

Dataset	Copula	DPMM	AFT
Melanoma	-0.22 (0.03)	-0.25 (0.02)	-0.23 (0.02)
Kidney	-0.11 (0.004)	-0.10 (0.004)	-0.10 (0.003)

For the melanoma dataset, we also visually evaluate the fit of the copula method on all 205 data points (Figure 4). We follow the setup of Riva-Palacio et al. (2021) and plot the predictive survival function for various tumour thicknesses x , comparing to the Kaplan-Meier (KM) estimator fit on windows centered around each x value. The copula method matches reasonably

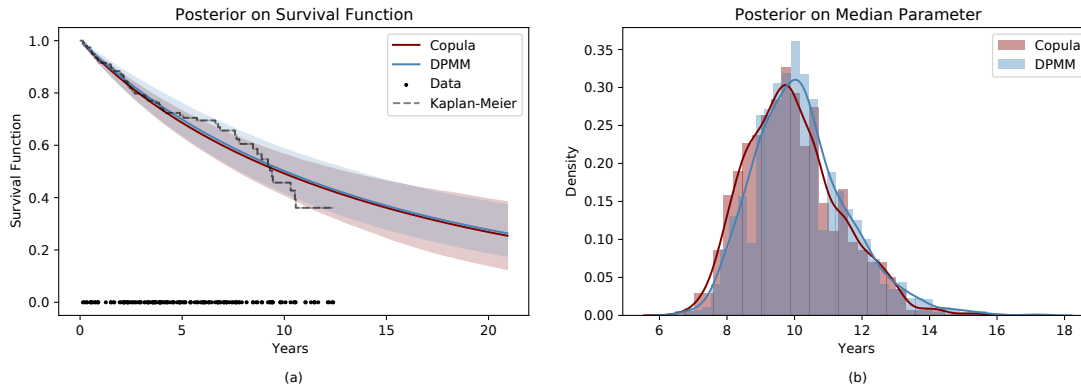


Figure 3: (a) Posterior mean and 95% credible interval of the survival function and (b) samples of median for the PBC placebo arm.

closely with the stratified KM estimator. Plots of the nonparametric median function can be found in Appendix C.

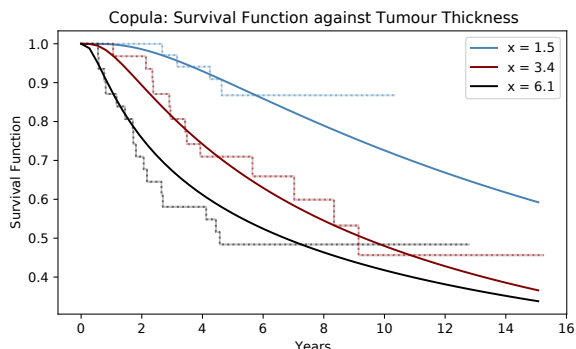


Figure 4: Survival function for copula method (—) for $x = \{1.5, 3.4, 6.1\}$ with KM (.....) fit to windows $\{(1.255, 1.75), (2.7, 4.1), (4.1, 8.1)\}$.

7 DISCUSSION

In this work, we have built on the martingale posterior distribution framework to the case in survival analysis where some of the observations are right-censored. We make use of one-step-ahead bivariate copula updates to perform inference, which admit a straightforward sequential importance sampling algorithm, thus avoiding the need for the likelihood/prior construction or MCMC. Our method is competitive with other Bayesian nonparametric survival models, both in terms of predictive accuracy and computation time. We note that a similar approach could be applied to other types of data with partially observed information. In future work, we hope to generalize by replacing imputation of the right-censored data with the imputation of $p(y_{\text{mis}} | y_{\text{obs}})$, where $y_{1:n} = \{y_{\text{obs}}, y_{\text{mis}}\}$.

There are a number of practical details regarding the implementation of our predictive resampling scheme that merit further investigation. Firstly, the compu-

tational complexity of the copula updates is $\mathcal{O}(n^2)$, which may be overly onerous for large datasets. Approximate methods such as subsampling may be one path towards reducing this computational cost. Secondly, we have found that predictive performance is, perhaps unsurprisingly, sensitive to the specification on the initial predictive density p_0 . In our current scheme, the bandwidth parameter a also controls the tails of p_0 , which we set to be adaptive as a default value is difficult under right-censoring. Finally, we use grid-based optimisation to select the copula update hyperparameters. Although this approach was suitable for the examples studied here, it would not scale well to settings with a larger number of hyperparameters. Potential alternatives include stochastic gradient descent methods or a theoretically-justified plug-in selection procedure.

We conclude by relating the martingale posterior framework to the *operational subjective* approach to statistical inference (Lad, 1996). The operational subjectivist specifies uncertain knowledge about quantities of interest through personal probability assertions, termed “previsions”, relying on the fundamental theorem of prevision (de Finetti, 1937; Lad et al., 1990; Lad et al., 1992) to ensure the coherence of a set of assertions. By assuming the exchangeability of observations, De Finetti’s representation theorem provides a concise characterisation of the predictive via a common sampling density (de Finetti, 1937). In contrast, Fong et al. (2021) proposes to elicit the predictive directly via one-step-ahead copula updates. Under right-censoring, computational difficulty arises from the partial nature of the information associated with censored observations, which we resolve by using a Monte Carlo approximation that first imputes the missing information. While not pursued here, we note that it may be possible to bypass the Monte Carlo approximation by specifying a predictive that directly accounts for censored observations.

Acknowledgements

We thank the area chairs and the reviewers for their constructive feedback comments that have greatly improved this manuscript. We gratefully acknowledge invaluable guidance and support, especially in the initial stages of this work, from Chris Holmes. We also thank Stephen Walker for helpful discussions on predictive resampling for censored data, George Deligianidis for assistance with the extension of Doob's consistency theorem, and Andrew Yiu for useful comments on the manuscript. E.F. was funded by The Alan Turing Institute Doctoral Studentship, under the EPSRC grant EP/N510129/1, and is currently employed at Novo Nordisk. B.L. was supported by the UK Engineering and Physical Sciences Research Council through the Bayes4Health programme (grant number EP/R018561/1) and gratefully acknowledges funding from Jesus College, Oxford.

References

- Arfè, A., & Muliere, P. (2020). A general Bayesian bootstrap for censored data based on the beta-stacy process. *arXiv preprint arXiv:2002.04081*.
- Balakrishnan, N., & Lai, C. D. (2009). *Continuous bivariate distributions*. Springer Science & Business Media.
- Berliner, L. M., & Hill, B. M. (1988). Bayesian Nonparametric Survival Analysis. *Journal of the American Statistical Association*, *83*(403), 772–779. <https://doi.org/10.1080/01621459.1988.10478660>
- Berti, P., Dreassi, E., Pratelli, L., & Rigo, P. (2021). A class of models for Bayesian predictive inference. *Bernoulli*, *27*(1), 702–726. <https://doi.org/10.3150/20-BEJ1255>
- Berti, P., Pratelli, L., & Rigo, P. (2004). Limit theorems for a class of identically distributed random variables. *The Annals of Probability*, *32*(3), 2029–2052.
- Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, *1*(2), 353–355.
- Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., & Wanderman-Milne, S. (2018). *JAX: Composable transformations of Python+NumPy programs* (Version 0.2.5). <http://github.com/google/jax>
- Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, *65*(1), 141–151.
- Coolen, F., & Yan, K. (2004). Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference*, *126*(1), 25–54.
- Dawid, A. P. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, *147*(2), 278–290.
- De Iorio, M., Johnson, W. O., Müller, P., & Rosner, G. L. (2009). Bayesian Nonparametric Non-proportional Hazards Survival Modeling. *Biometrics*, *65*(3), 762–771. <https://doi.org/10.1111/j.1541-0420.2008.01166.x>
- De Iorio, M., Müller, P., Rosner, G. L., & MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, *99*(465), 205–215.
- de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives'. Paris: Institut Henri Poincaré. English translation by Kyburg, HE (1980). Prevision: Its logical laws, its subjective sources. *Studies in subjective probability*, *2nd edn*. New York: Wiley.
- Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., & Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology*, *10*(1), 1–7.
- Doksum, K. (1974). Tailfree and Neutral Random Probabilities and Their Posterior Distributions. *The Annals of Probability*, *2*(2), 183–201. <https://doi.org/10.1214/aop/1176996703>
- Doob, J. L. (1949). Application of the theory of martingales. *Actes du Colloque International Le Calcul des Probabilités et ses applications (Lyon, 28 Juin–3 Juillet 1948)*, Paris CNRS, 23–27.
- Doucet, A., Johansen, A. M. et al. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, *12*(656–704), 3.
- Ferguson, T. S., & Phadia, E. G. (1979). Bayesian Nonparametric Estimation Based on Censored Data. *The Annals of Statistics*, *7*(1), 163–186. <https://doi.org/10.1214/aos/1176344562>
- Fong, E., Holmes, C., & Walker, S. G. (2021). Martingale posterior distributions. *arXiv preprint arXiv:2103.15671*.
- Ghosal, S., & van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press. <https://doi.org/10.1017/9781139029834>
- Hahn, P. R., Martin, R., & Walker, S. G. (2018). On recursive Bayesian predictive distributions.

- Journal of the American Statistical Association*, 113(523), 1085–1093.
- Hill, B. M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63(322), 677–691. <http://www.jstor.org/stable/2284038>
- Hjort, N. L. et al. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, 18(3), 1259–1294.
- Kalbfleisch, J. D. (1978). Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(2), 214–221. <http://www.jstor.org/stable/2984758>
- Kim, Y., & Lee, J. (2003). Bayesian analysis of proportional hazard models. *The Annals of Statistics*, 31(2), 493–511. <https://doi.org/10.1214/aos/1051027878>
- Klein, Moeschberger, & modifications by Jun Yan. (2012). *KMsurv: Data sets from Klein and Moeschberger (1997), survival analysis* [R package version 0.1-5]. <https://CRAN.R-project.org/package=KMsurv>
- Kong, A., Liu, J. S., & Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American statistical association*, 89(425), 278–288.
- Kottas, A. (2006). Nonparametric Bayesian survival analysis using mixtures of Weibull distributions. *Journal of Statistical Planning and Inference*, 136(3), 578–596. <https://doi.org/10.1016/j.jspi.2004.08.009>
- Lad, F. (1996). *Operational subjective statistical methods: A mathematical, philosophical, and historical introduction*. Wiley.
- Lad, F., Dickey, J., & Rahman, M. (1990). The fundamental theorem of prevision. *Statistica*, 50, 19.
- Lad, F., Dickey, J. M., & Rahman, M. A. (1992). Numerical application of the fundamental theorem of prevision. *Journal of Statistical Computation and Simulation*, 40(3-4), 135–151. <https://doi.org/10.1080/00949659208811372>
- Lo, A. Y. et al. (1993). A Bayesian bootstrap for censored data. *The Annals of Statistics*, 21(1), 100–123.
- Muliere, P., & Walker, S. (1997a). A Bayesian non-parametric approach to survival analysis using Polya trees. *Scandinavian Journal of Statistics*, 24(3), 331–340. <http://www.jstor.org/stable/4616459>
- Muliere, P., & Walker, S. (1997b). Neutral to the right processes from a predictive perspective: A review and new developments.
- Nelsen, R. (2007). *An Introduction to Copulas*. Springer New York.
- Riva-Palacio, A., Leisen, F., & Griffin, J. (2021). Survival Regression Models With Dependent Bayesian Nonparametric Priors. *Journal of the American Statistical Association*, 1–10. <https://doi.org/10.1080/01621459.2020.1864381>
- Ross, G. J., & Markwick, D. (2018). Dirichletprocess: An R package for fitting complex Bayesian nonparametric models.
- Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9(1), 130–134. <https://doi.org/10.1214/aos/1176345338>
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434), 473–489.
- Therneau, T. M. (2021). *A package for survival analysis in R* [R package version 3.2-13]. <https://CRAN.R-project.org/package=survival>
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth) [ISBN 0-387-95457-0]. Springer. <http://www.stats.ox.ac.uk/pub/MASS4>

Supplementary Material: A Predictive Approach to Bayesian Nonparametric Survival Analysis

A THEORY

A.1 Non-informative Censoring

To illustrate the idea of non-informative censoring, consider the example of a single censored datum, $Y_1 \geq C_1$ where we observe $C_1 = c_1$. The usual random censoring assumption is

$$Y_1 \sim F_\theta, \quad C_1 \sim G_\lambda.$$

with Y_1, C_1 independent. Under this assumption, the censoring mechanism is already non-informative for the maximum likelihood estimate (MLE) of θ , as the estimate of θ does not depend on G_λ . In the Bayesian case the additional assumption of prior independence, $\pi(\theta, \lambda) = \pi(\theta)\pi(\lambda)$, is sufficient for the censoring mechanism to be non-informative. It is straightforward to show that the posterior in this case does not rely on G_λ , as it takes the form

$$\begin{aligned} \pi(\theta \mid Y_1 \geq C_1, c_1) &\propto p(\theta, Y_1 \geq C_1, c_1) \\ &= \int p(\theta, \lambda, Y_1 \geq C_1, c_1) d\lambda \\ &= \pi(\theta) \bar{F}_\theta(c_1) \int \pi(\lambda) g_\lambda(c_1) d\lambda \\ &\propto \pi(\theta) \bar{F}_\theta(c_1). \end{aligned}$$

In the above, we have used the notation $p(x, Y \geq c) = p(x \mid Y \geq c)P(Y \geq c) = P(Y \geq c \mid x)p(x)$ to represent the mixed joint probability density function of the observed values and censored events, where x is a continuous r.v. and $\mathbf{1}(Y \geq c)$ can be considered as a discrete r.v.. We will continue to do so for the remainder of the Appendix in other contexts.

In predictive resampling, $p(y_{2:\infty} \mid y_1)$ by definition does not depend on the censoring. The censoring mechanism can thus only affect the imputing density $p(y_1 \mid Y_1 \geq C_1, c_1)$. Defining $p(y_1 \mid Y_1 \geq c_1) \propto \mathbf{1}(y_1 \geq c_1)p(y_1)$, the Bayesian assumptions for non-informative censoring implies

$$p(y_1 \mid Y_1 \geq C_1, c_1) = p(y_1 \mid Y_1 \geq c_1).$$

We can see this through the following:

$$\begin{aligned} p(y_1 \mid Y_1 \geq C_1, c_1) &\propto p(y_1, Y_1 \geq C_1, c_1) \\ &= P(Y_1 \geq C_1 \mid y_1, c_1) p(y_1, c_1) \\ &= \mathbf{1}(y_1 \geq c_1) p(y_1, c_1) \\ &= \mathbf{1}(y_1 \geq c_1) p(y_1) p(c_1) \\ &\propto \mathbf{1}(y_1 \geq c_1) p(y_1). \end{aligned}$$

In the above, $p(y_1, c_1) = \int f_\theta(y_1) g_\lambda(c_1) d\pi(\theta, \lambda)$. The key is that $p(y_1, c_1)$ factorizes into $p(y_1)p(c_1)$ under the assumptions of random censoring and independence in the prior.

In the absence of the likelihood and prior, a sufficient condition for $p(y_1 \mid Y_1 \geq C_1, c_1) = p(y_1 \mid Y_1 \geq c_1)$ is the factorization $p(y_1, c_1) = p(y_1)p(c_1)$, that is Y_1 and C_1 are a priori independent. More generally for the martingale posterior, a sufficient assumption for non-informative censoring is that under our predictive distribution, the vector $Y_{1:N}$ is independent of $C_{1:N}$ for all N . For the remainder of the Appendix, we continue to assume this and will treat the censoring times c_i as constants.

A.2 Doob's Consistency Theorem for Right-censored Observations

In this section, we prove Doob's consistency theorem for the setting where some of the observations may be censored. As a reminder, we have $\mathcal{D}_n := \{y_{1:k}, Y_{k+1:n} \geq c_{k+1:n}\}$. As $y_{1:k}$ are fully observed, we consider their values as fixed constants. For completeness, we include a repeat of the setup here.

Assume that for all N , the r.v.s $[\Theta, Y_1, \dots, Y_N]$ have joint density

$$p(\theta, y_{1:N}) = \pi(\theta) \prod_{i=1}^N f_\theta(y_i).$$

We will make use of the standard Doob's consistency theorem (for uncensored observations) and so require the usual identifiability and measurability assumptions on the parametric sampling density f_θ , which can be found in Doob (1949) or Ghosal and van der Vaart (2017, Theorem 6.9, Proposition 6.10). Specifically, the identifiability condition is such that $F_\theta \neq F_{\theta'}$ whenever $\theta \neq \theta'$, where F_θ is the cumulative distribution function of f_θ . This ensures that the parameter Θ can be recovered from the infinite sample.

We assume that Θ lies in a linear space so the expectation is well-defined, and all regular conditional probability measures exist. We then write the posterior mean as $\bar{\theta}_N = E[\Theta | Y_{1:N}]$, and denote $f_\theta^c(y) = \mathbb{1}\{y \geq c\} f_\theta(y) / \bar{F}_\theta(c)$ to be the density of a data point right-censored at c , where \bar{F}_θ is the survival function of f_θ .

We draw $Y_{k+1:n} \sim p(y_{k+1:n} | \mathcal{D}_n)$ where

$$p(y_{k+1:n} | \mathcal{D}_n) = \int \prod_{i=k+1}^n f_\theta^{c_i}(y_i) \pi(\theta | \mathcal{D}_n) d\theta,$$

and $\pi(\theta | \mathcal{D}_n) \propto \pi(\theta) \prod_{i=1}^k f_\theta(y_i) \prod_{i=k+1}^n \bar{F}_\theta(c_i)$, which follows from the non-informative censoring assumption.

We then draw $Y_{n+1:N} \sim p(Y_{n+1:N} | y_{1:n})$ where

$$p(y_{n+1:N} | y_{1:n}) = \int \prod_{i=n+1}^N f_\theta(y_i) \pi(\theta | y_{1:n}) d\theta,$$

and compute $\bar{\theta}_N$ from $\{y_{1:k}, Y_{k+1:N}\}$. The following result establishes the equivalence of predictive resampling and standard Bayesian inference as $N \rightarrow \infty$.

Theorem 1. *Assume $E[|\Theta| | \mathcal{D}_n] < \infty$. Under regularity conditions on f_θ , we have that*

$$\lim_{N \rightarrow \infty} \bar{\theta}_N = \Theta \quad \text{a.s. } P^\infty(\cdot | \mathcal{D}_n), \quad (13)$$

where P^∞ is over Θ and $Y_{k+1:\infty}$.

Proof. For each $y_{k+1:n} \in \mathbb{R}^{n-k}$ such that $E[|\Theta| | y_{1:n}] < \infty$, Doob's consistency theorem gives us

$$\lim_{N \rightarrow \infty} \bar{\theta}_N = \Theta \quad \text{a.s. } P^\infty(\cdot | y_{1:n}). \quad (14)$$

Note that the tower rule gives us

$$E[E[|\Theta| | y_{1:n}] | \mathcal{D}_n] = E[|\Theta| | \mathcal{D}_n] < \infty,$$

so $E[|\Theta| | y_{1:n}] < \infty$ for $P(\cdot | \mathcal{D}_n)$ -almost all $y_{k+1:n}$. This implies that (14) holds for $P(\cdot | \mathcal{D}_n)$ -almost all $y_{k+1:n}$. Finally, we have the following:

$$\begin{aligned} P^\infty \left(\lim_N \bar{\theta}_N = \Theta | \mathcal{D}_n \right) &= \int \mathbb{1} \left\{ \lim_N \bar{\theta}_N = \Theta \right\} dP^\infty(\Theta, y_{k+1:\infty} | \mathcal{D}_n) \\ &= \int \underbrace{\int \mathbb{1} \left\{ \lim_N \bar{\theta}_N = \Theta \right\} dP^\infty(\Theta, y_{n+1:\infty} | y_{1:n})}_{=1 \text{ a.s. } P(y_{k+1:n} | \mathcal{D}_n)} dP(y_{k+1:n} | \mathcal{D}_n) \\ &= 1, \end{aligned}$$

which is exactly statement (13). □

A.3 Conditionally Identically Distributed Sequences for Right-censored Observations

In this section, we first review the c.i.d. properties in the fully observed case, before discussing the implications of the c.i.d. property when there is right-censoring in the observations. In the fully observed case, assume that the sequence of r.v.s $[Y_{n+1}, \dots, Y_N]$ is c.i.d., where $P_i(y)$ is the usual predictive cumulative distribution function of Y_{i+1} conditional on $Y_{1:i}$ for $i \geq n$. Following Berti et al. (2004) and Fong et al. (2021), the sequence of predictive cumulative distribution functions is a martingale as it satisfies

$$E[P_i(y) \mid y_{1:i-1}] = P_{i-1}(y)$$

almost surely for each $y \in \mathbb{R}$, for $i \geq n$. We highlight that P_i is the predictive distribution conditional on $y_{1:i}$, so the expectation is over Y_i . From the properties of the c.i.d. sequence (Berti et al., 2004, Lemma 2.1, 2.4), we have that the predictive distribution converges weakly to a random probability distribution, P_∞ , almost surely, that is

$$P_N(y) \rightarrow P_\infty(y) \quad \text{a.s. } P^\infty(\cdot \mid y_{1:n})$$

for each $y \in \mathbb{R}$, where $P^\infty(\cdot \mid y_{1:n})$ is over $Y_{n+1:\infty}$. Furthermore, we have that $E[P_\infty(y) \mid y_{1:n}] = P_n(y)$ almost surely for each $y \in \mathbb{R}$, which is the unbiasedness coherence condition from Fong et al. (2021). The empirical distribution,

$$F_N(y) = \frac{1}{N} \left\{ \sum_{i=1}^n \mathbb{1}(y_i \leq y) + \sum_{i=n+1}^N \mathbb{1}(Y_i \leq y) \right\},$$

also satisfies the same property (Berti et al., 2004, Theorem 2.2), that is

$$F_N(y) \rightarrow F_\infty(y) \quad \text{a.s. } P^\infty(\cdot \mid y_{1:n})$$

for each $y \in \mathbb{R}$, and in fact $F_\infty = P_\infty$ almost surely.

Returning to the right-censored case where $Y_{k+1:n}$ is in fact random and drawn from $p(y_{k+1:n} \mid \mathcal{D}_n)$, we note that the above convergence holds for each $y_{k+1:n}$. We can therefore write

$$\begin{aligned} P^\infty \left(\lim_N F_N(y) = F_\infty(y) \mid \mathcal{D}_n \right) &= \int \mathbb{1} \left\{ \lim_N F_N(y) = F_\infty(y) \right\} dP^\infty(y_{k+1:\infty} \mid \mathcal{D}_n) \\ &= \int \underbrace{\int \mathbb{1} \left\{ \lim_N F_N(y) = F_\infty(y) \right\} dP^\infty(y_{n+1:\infty} \mid y_{1:n})}_{=1} dP(y_{k+1:n} \mid \mathcal{D}_n) \\ &= 1. \end{aligned}$$

A similar result can be shown for the limiting predictive distribution P_∞ . From the above, we see that the limiting empirical distribution exists under the imputation and predictive resampling scheme due to the martingale property of the c.i.d. sequence. We can thus compute $\theta(F_\infty)$ to obtain a posterior sample from the martingale posterior, $\pi_\infty(\theta \mid \mathcal{D}_n)$, where we use the subscript in π_∞ to distinguish from the regular Bayesian posterior. Note here that F_∞ is unknown but we can obtain samples through predictive resampling, in contrast to the parametric case of Doob's theorem where $\theta_\infty = \Theta$ is known.

Interestingly, the unbiasedness coherence condition of Fong et al. (2021) is also satisfied in the right-censored case. We can compute the posterior mean of P_∞ as

$$\begin{aligned} E[P_\infty(y) \mid \mathcal{D}_n] &= E[E[P_\infty(y) \mid y_{1:n}] \mid \mathcal{D}_n] \\ &= E[P_n(y) \mid \mathcal{D}_n] \\ &= P(Y_{n+1} \leq y \mid \mathcal{D}_n), \end{aligned}$$

which is the cumulative distribution function of (9). Note that the outer expectation in the first line is over $Y_{k+1:n} \sim p(y_{k+1:n} \mid \mathcal{D}_n)$, whereas the inner expectation is over $Y_{n+1:\infty} \sim p(y_{n+1:\infty} \mid y_{1:n})$. Once again, the posterior mean of P_∞ is our best estimate of the distribution of Y_{n+1} given \mathcal{D}_n , and our imputation and predictive resampling scheme has incurred no bias. To summarize, we inherit the nice coherency properties of Fong et al. (2021) as we have the c.i.d. sequence conditional on the imputed $Y_{k+1:n}$.

A final point is that even with $Y_{k+1:n}$ marginalized out, the sequence $Y_{n+1:\infty}$ remains c.i.d. as the marginalized predictive distribution is just a mixture of the fully observed P_i . However, the predictive distribution when $Y_{k+1:n}$ is marginalized is not tractable, which is why we introduce the sequential Monte Carlo scheme in our paper. Eliciting the marginalized predictive directly would be an interesting avenue of future work.

B METHODOLOGY

B.1 Predictive Resampling in the Uncensored Case

Predictive resampling for the uncensored case, as described in Fong et al. (2021), is given by Algorithm 2. A slight intricacy is that the limiting empirical F_∞ may be continuous but F_N is always discrete. We opt instead to use the final random predictive P_N as an estimate of the limiting empirical F_∞ , as it can be continuous and in fact converges to F_∞ in the limit of $N \rightarrow \infty$ (Berti et al., 2004) as discussed in Appendix A.3. A martingale posterior sample can then be computed as $\theta_N = \theta(P_N)$.

Algorithm 2: Predictive Resampling (Fong et al., 2021)

```

Compute  $p_n$  from  $y_{1:n}$ 
for  $j \leftarrow 1$  to  $B$  do
    for  $i \leftarrow n + 1$  to  $N$  do
        Sample  $Y_i \sim P_{i-1}$ 
        Update  $P_i \leftarrow \{P_{i-1}, Y_i\}$ 
    end
    Evaluate  $\theta_N^{(j)} = \theta(P_N)$ 
end
Return  $\{\theta_N^{(1)}, \dots, \theta_N^{(B)}\}$ 
    
```

B.2 Copula Updates

The copula update for the first step of the DPMM has been derived previously in Hahn et al. (2018) and Fong et al. (2021). We will restate the key details that are not included in the main paper here.

From the main paper, the copula update for the densities is

$$p_{i+1}(y) = [1 - \alpha_{i+1} + \alpha_{i+1} d_a \{P_i(y), P_i(y_{i+1})\}] p_i(y). \quad (15)$$

If the kernel of the DPMM has density $f_\theta(y)$, and the base measure of the DP has centering measure $\pi(\theta)$, then the bivariate copula density is

$$d_a(u, v) = \frac{\int f_\theta \{P_0^{-1}(u)\} f_\theta \{P_0^{-1}(v)\} \pi(\theta) d\theta}{p_0 \{P_0^{-1}(u)\} p_0 \{P_0^{-1}(v)\}}, \quad (16)$$

where $p_0(y) = \int f_\theta(y) \pi(\theta) d\theta$ and a is a hyperparameter that depends on the specification of the likelihood and prior. We then have $P_0(y) = \int^y p(y') dy'$ and P_0^{-1} is the inverse CDF.

Note that the update (15) requires the CDF $P_i(y)$. Fortunately this update is typically tractable, and involves integrating (15):

$$\begin{aligned}
 P_{i+1}(y) &= (1 - \alpha_{i+1}) P_i(y) + \alpha_{i+1} \int_0^y d_a \{P_i(y'), P_i(y_{i+1})\} p_i(y') dy' \\
 &= (1 - \alpha_{i+1}) P_i(y) + \alpha_{i+1} \int_0^{P_i(y)} d_a \{u', P_i(y_{i+1})\} du' \\
 &= (1 - \alpha_{i+1}) P_i(y) + \alpha_{i+1} I_a \{P_i(y), P_i(y_{i+1})\}.
 \end{aligned}$$

The second line follows from the change of variables $u' = P_i(y')$, and we have that

$$I_a(u, v) = \int_0^u d_a(u', v) du'. \quad (17)$$

If π is conjugate to f_θ , then the forms of I_a and d_a are typically tractable.

B.3 Exponential Copula Update

In this section, we derive the copula density corresponding to the DPMM with the exponential kernel and gamma centering measure, that is

$$f_\theta(y) = \theta \exp(-\theta y), \quad \pi(\theta) = \text{Gamma}(\theta \mid a, b) \quad (18)$$

for $y \geq 0$. We can derive the copula by considering

$$\begin{aligned} \int f_\theta(y) f_\theta(y_1) d\pi(\theta) &= \frac{b^a}{\Gamma(a)} \int_0^\infty \theta^{a+1} \exp\{-(b+y+y_1)\theta\} d\theta \\ &= \frac{b^a}{\Gamma(a) (b+y+y_1)^{a+2}} \int_0^\infty x^{a+1} \exp(-x) dx \\ &= \frac{a(a+1)}{b^2} \left(1 + \frac{y+y_1}{b}\right)^{-(a+2)} \end{aligned} \quad (19)$$

where in the second line we have used the substitution $x = \theta(b+y+y_1)$, and the third line uses $\Gamma(a+2) = a(a+1)\Gamma(a)$. We also have

$$p_0(y) = \int f_\theta(y) d\pi(\theta) = \frac{a}{b} \left(1 + \frac{y}{b}\right)^{-(a+1)} \quad (20)$$

which is the Lomax(a, b) density. The copula density then takes the form

$$d_{a,b}(y, y_1) = \frac{a+1}{a} \frac{\left(1 + \frac{y}{b}\right)^{a+1} \left(1 + \frac{y_1}{b}\right)^{a+1}}{\left(1 + \frac{y+y_1}{b}\right)^{a+2}}.$$

We would like the density as a function of (u, v) . To this end, note that the marginal CDF and inverse CDF are

$$P_0(y) = 1 - \left(1 + \frac{y}{b}\right)^{-a}, \quad P_0^{-1}(u) = b \left\{ (1-u)^{-\frac{1}{a}} - 1 \right\}. \quad (21)$$

Finally, this gives us the copula density

$$d_a(u, v) = \frac{a+1}{a} \frac{(1-u)^{-\frac{a+1}{a}} (1-v)^{-\frac{a+1}{a}}}{\left\{ (1-u)^{-\frac{1}{a}} + (1-v)^{-\frac{1}{a}} - 1 \right\}^{a+2}} \quad (22)$$

where $u = P(y)$ and $v = P(y_1)$.

To derive $I_a(u, v)$, we compute the integral in (17). Substituting $x = (1-u')^{-\frac{1}{a}} + (1-v)^{-\frac{1}{a}} - 1$ gives us

$$dx = du' \times \frac{1}{a} (1-u')^{-\frac{a+1}{a}}, \quad c_1 = (1-v)^{-\frac{1}{a}}, \quad c_2 = (1-u)^{-\frac{1}{a}} + (1-v)^{-\frac{1}{a}} - 1.$$

Plugging this into (17) gives us

$$\begin{aligned} I_a(u, v) &= (a+1)(1-v)^{-\frac{a+1}{a}} \int_{c_1}^{c_2} x^{-(a+2)} dx \\ &= 1 - \frac{(1-v)^{-\frac{a+1}{a}}}{\left\{ (1-u)^{-\frac{1}{a}} + (1-v)^{-\frac{1}{a}} - 1 \right\}^{a+1}}. \end{aligned}$$

In practice, we set $b = 1$ and so $p_0(y) = \text{Lomax}(a, 1)$.

B.4 Log-normal Copula Update

We derive the copula update for the DPMM with a log-normal kernel with a normal base measure, that is

$$f_\theta(y) = \frac{\mathcal{N}(\log(y) \mid \theta, 1)}{y}, \quad \pi(\theta) = \mathcal{N}(\theta \mid 0, \tau^{-1})$$

for $y \geq 0$. Working with $z = \exp(y)$, we get

$$\int f_\theta(y) f_\theta(y_1) d\pi(\theta) = \frac{\int \mathcal{N}(z | \theta, 1) \mathcal{N}(z_1 | \theta, 1) d\mathcal{N}(\theta | 0, 1)}{\log(z) \log(z_1)}$$

Similarly, we have

$$p_0(y) = \frac{\int \mathcal{N}(z | \theta, 1) d\mathcal{N}(\theta | 0, 1)}{\log(z)}.$$

Plugging the above into (16) gives us the bivariate Gaussian copula density $c_\rho(u, v)$, as the $\log(z)$ terms cancel out. We write c_ρ instead of d_a to remain consistent with Hahn et al. (2018) and Fong et al. (2021).

The Gaussian copula density is

$$c_\rho(u, v) = \frac{\mathcal{N}_2\{\Phi^{-1}(u), \Phi^{-1}(v) | 0, 1, \rho\}}{\mathcal{N}\{\Phi^{-1}(u) | 0, 1\} \mathcal{N}\{\Phi^{-1}(v) | 0, 1\}} \quad (23)$$

where $\rho \in (0, 1)$, Φ is the standard normal CDF, and $\mathcal{N}_2(0, 1, \rho)$ is the bivariate normal density with mean 0, variance 1 and correlation ρ . We can similarly compute $H_\rho(u, v) = \int_0^u c_\rho(u', v) du'$, which is

$$H_\rho(u, v) = \Phi \left\{ \frac{\Phi^{-1}(u) - \rho \Phi^{-1}(v)}{\sqrt{1 - \rho^2}} \right\}. \quad (24)$$

Although this implies the same copula update as the one on \mathbb{R} as introduced in Hahn et al. (2018), the key difference is $p_0(y)$, which can be shown to be

$$p_0(y) = \text{Log-normal} \left(y | 0, \frac{1}{1 - \rho} \right).$$

B.5 Ordering for Copula Method

Kong et al. (1994) suggests ordering the data such that the observed data comes before the missing data, which is to ensure the proposal is close to the target for importance weight stability. In the right-censoring case however, we have found that this intuition does not extend. In practice, randomizing the order of data greatly increases the ESS in comparison to ordering the uncensored data before the censored data. Although one can average the results over different permutations, we find that a single permutation works well in practice. In the random order case, the IS weights have a slightly different form to take into account the observed data points between censored data points - this is provided in Algorithm 1 and derived in Section B.6.

We postulate that ordering the data is undesirable due to the nature of right-censoring: as the uncensored $y_{1:k}$ will tend to take on smaller values, a density estimate constructed from $y_{1:k}$ will not be sufficiently right-skewed compared to the target distribution, which has support on the larger values $y_{k+1:n}$ that have been right-censored. This results in the proposal being too light-tailed with respect to the target distribution, leading to IS weights with high variance. We recommend randomizing the order as it results in a heavier-tailed proposal, and this works much better in practice.

We demonstrate this in the parametric example of Section 6.1, where the joint density on $Y_{1:N}$ is exchangeable, so the ordering only affects importance weight stability. We compare the ESS of the IS weights for random data ordering versus the ordering $\{y_{1:k}, Y_{k+1:n} \geq c_{k+1:n}\}$, which is computed as

$$\text{ESS}(w^{(1:B)}) = 1 / \sum_{j=1}^B \{w^{(j)}\}^2. \quad (25)$$

We carry out Algorithm 1 followed by predictive resampling *without* the SMC resampling steps for the two orderings. As we see in Figures 5a and 6a, the random ordering case is quite close to the truth even without SMC resampling, but the uncensored/censored ordering case is a poor approximation. As expected, the ESS for the random and uncensored/censored cases are 967 and 8 respectively for $B = 2000$. To visualize the cause, we see in Figure 6b that the proposal has poor support over the true posterior of θ as it is peaked and not sufficiently right-skewed. On the other hand, the random ordering case proposal in Figure 5b has a heavy right tail - we do not plot the true posterior here as it is significantly more peaked than the proposal.

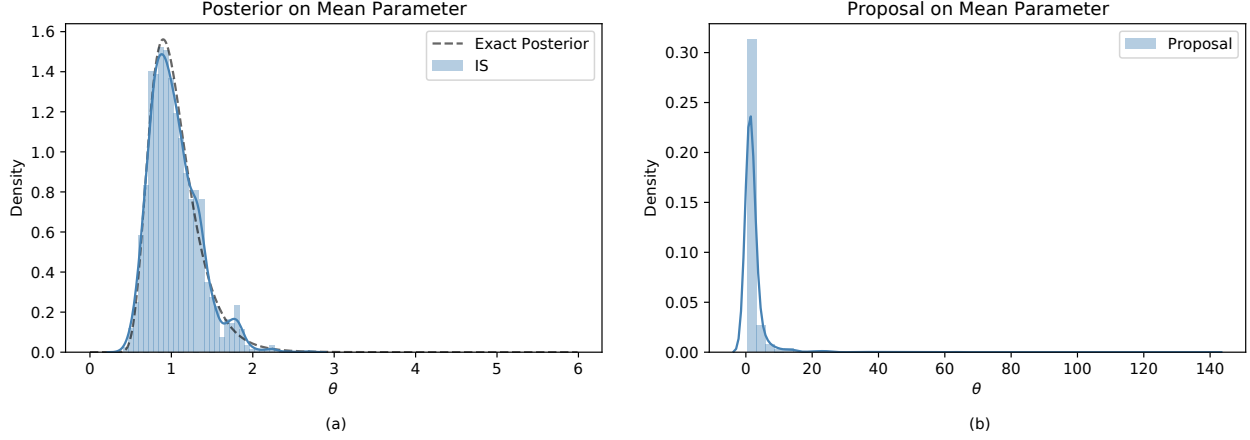


Figure 5: Random ordering: (a) Martingale posterior generated via Algorithm 1 (without SMC resampling) and predictive resampling for θ ; (b) Proposal distribution before IS reweighting

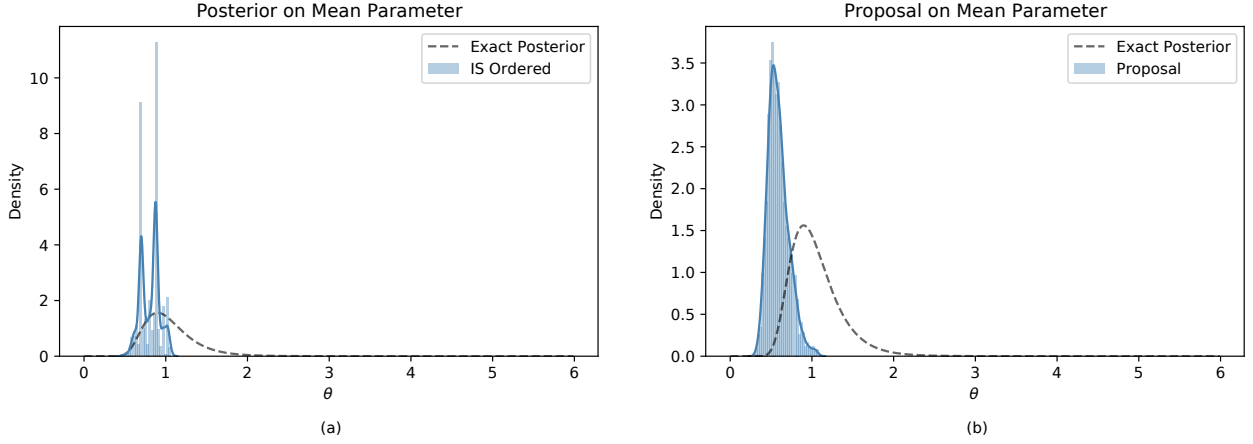


Figure 6: Uncensored/censored ordering: (a) Martingale posterior generated via Algorithm 1 (without SMC resampling) and predictive resampling for θ ; (b) Proposal distribution before IS reweighting

B.6 Derivation of Algorithm 1

As discussed in the main paper and above, the IS weights take on a slightly different form under random ordering, which we now derive. For a dataset \mathcal{D}_n , denote the indices of observed data points as \mathcal{I}_o and censored data points as \mathcal{I}_c , so that $\mathcal{I}_o \cup \mathcal{I}_c = \{1, \dots, n\}$ and $\mathcal{I}_o \cap \mathcal{I}_c = \emptyset$, and $\mathcal{D}_n = \{y_{\mathcal{I}_o}, Y_{\mathcal{I}_c} \geq c_{\mathcal{I}_c}\}$. Our sequential imputation scheme in Algorithm 1 gives the proposal density

$$q(y_{\mathcal{I}_c}) = \prod_{i \in \mathcal{I}_c} p(y_i | Y_i \geq c_i, y_{1:i-1}).$$

Our target however is the conditional density

$$p(y_{\mathcal{I}_c} | Y_{\mathcal{I}_c} \geq c_{\mathcal{I}_c}, y_{\mathcal{I}_o}) \propto p(y_{\mathcal{I}_c}, Y_{\mathcal{I}_c} \geq c_{\mathcal{I}_c}, y_{\mathcal{I}_o}).$$

We can factorize the mixed joint density into

$$\begin{aligned} p(y_{\mathcal{I}_c}, Y_{\mathcal{I}_c} \geq c_{\mathcal{I}_c}, y_{\mathcal{I}_o}) &= \prod_{i \in \mathcal{I}_c} p(y_i, Y_i \geq c_i | y_{1:i-1}) \prod_{j \in \mathcal{I}_o} p(y_j | y_{1:j-1}) \\ &= \prod_{i \in \mathcal{I}_c} p(y_i | Y_i \geq c_i, y_{1:i-1}) P(Y_i \geq c_i | y_{1:i-1}) \prod_{j \in \mathcal{I}_o} p(y_j | y_{1:j-1}) \\ &= q(y_{\mathcal{I}_c}) \prod_{i \in \mathcal{I}_c} P(Y_i \geq c_i | y_{1:i-1}) \prod_{j \in \mathcal{I}_o} p(y_j | y_{1:j-1}). \end{aligned}$$

Dividing the above by $q(y_{\mathcal{I}_c})$ gives us the unnormalized importance weights

$$\begin{aligned} w &= \prod_{i \in \mathcal{I}_c} P(Y_i \geq c_i | y_{1:i-1}) \prod_{j \in \mathcal{I}_o} p(y_j | y_{1:j-1}) \\ &= \prod_{i=1}^n [\delta_i p(y_i | y_{1:i-1}) + (1 - \delta_i) P(Y_i \geq c_i | y_{1:i-1})] \end{aligned} \quad (26)$$

where $\delta_i = 1$ if $i \in \mathcal{I}_o$ and $\delta_i = 0$ if $i \in \mathcal{I}_c$. The above is exactly the importance weight in Algorithm 1.

B.7 Selecting Hyperparameters

In the copula densities above, we set the bandwidth parameters a or ρ by maximizing the (mixed) joint ‘marginal likelihood’ $p(\mathcal{D}_n)$. Assuming no SMC resampling steps for now, we can estimate $p(\mathcal{D}_n)$ through IS:

$$\hat{p}(\mathcal{D}_n) = \sum_{j=1}^B w^{(j)}$$

where $w^{(j)}$ are the unnormalized IS weights from (26). This can be shown to be a valid estimate, as we are approximating the expectation

$$\begin{aligned} \int \prod_{i \in \mathcal{I}_c} P(Y_i \geq c_i | y_{1:i-1}) \prod_{j \in \mathcal{I}_o} p(y_j | y_{1:j-1}) q(y_{\mathcal{I}_c}) dy_{\mathcal{I}_c} &= \int p(y_{\mathcal{I}_c}, Y_{\mathcal{I}_c} \geq c_{\mathcal{I}_c}, y_{\mathcal{I}_o}) dy_{\mathcal{I}_c} \\ &= p(Y_{\mathcal{I}_c} \geq c_{\mathcal{I}_c}, y_{\mathcal{I}_o}), \end{aligned}$$

which is exactly $p(\mathcal{D}_n)$. When SMC resampling steps are present, we can then approximate the ratio Z_i/Z_{i-1} at each time step and compute the product to get Z_n (Doucet, Johansen, et al., 2009, Section 3.5). Here $Z_i = p(\mathcal{D}_i)$, where \mathcal{D}_i are the data points up to and including datum i .

To maximize $\hat{p}(\mathcal{D}_n)$, we optimize across a pre-specified grid of hyperparameter values. We note that, although a gradient-based approach may be possible, it is likely to be slow due to the large number of particles B and potentially unstable due to the need to differentiate through an IS estimate.

B.8 Initializaton and Standardization

As motivated by the copula derivations above, we initialize the exponential copula update with $p_0(y) = \text{Lomax}(y | a, 1)$, where a is the same hyperparameter as the bandwidth. We set a in an adaptive way, as a default value is difficult to set - we cannot gauge the tail behaviour from the observed sample in the presence of right-censoring. An equivalent argument applies for ρ in the log-normal copula update, which controls the variance of $p_0 = \text{Log-normal}(y | 0, 1/(1 - \rho))$.

However, to prevent a wildly inappropriate p_0 , we opt to normalize the observed times in a heuristic manner to ensure the times are of the right order of magnitude. To illustrate this, we briefly use the alternative notation of observed times and censoring indicators $\{t_i, \delta_i\}_{i=1:n}$ for convenience, which corresponds one-to-one to the notation $\mathcal{D}_n = \{y_{\mathcal{I}_o}, Y_{\mathcal{I}_c} \geq c_{\mathcal{I}_c}\}$. From the Bayesian model in (18), we see that the prior expected value of θ is a/b . The hyperparameter a is a prior pseudo-count, so we aim for a default target value of $a \approx 1$. As we set $b = 1$, this suggests that we are aiming for a target $E[\theta] \approx 1$ under the exponential model. Finally, we highlight that the MLE of the rate θ for the exponential sampling density takes the form

$$\hat{\theta} = \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n t_i}.$$

We thus opt to multiply the times $t_{1:n}$ by $\hat{\theta}$ above to ensure an MLE of θ of 1. In the log-normal case, the MLE is not tractable unlike in the exponential case. As a result, we prefer to also multiply by $\hat{\theta}$ above in this case which works well in practice.

B.9 Diagnostics

We now briefly discuss the assessment of the computational accuracy of our method, as we rely on Monte Carlo and truncation approximations. We provide these diagnostics for our experiments in Section C. For the simulation of $y_{\mathcal{I}_c}$, we report the usual diagnostics for SMC - that is, we plot the ESS as computed in (25) against time in order to observe the number of resample steps. We also track the number of unique particles of $y_{\min(\mathcal{I}_c)}$ as a measure of particle degeneracy. See Doucet, Johansen, et al. (2009) for more details.

To diagnose the convergence of predictive resampling in the nonparametric case, we track the L_1 distance between the starting CDF P_n and forward simulated CDF P_N , i.e. we compute $\int |P_N(y) - P_n(y)| dy$ which we can approximate numerically. Note that this is the 1-Wasserstein metric. We use this as the survival function is of primary interest; we expect the Wasserstein-1 distance to converge to a constant as $N \rightarrow \infty$. In the parametric case, we simply observe the value of $\bar{\theta}_N$, which will also converge to a constant from Doob's result.

B.10 Copula Regression

For conditional density estimation, the copula update takes on the form

$$p_{i+1}(y | \mathbf{x}) = \{1 - \alpha_{i+1}(\mathbf{x}, \mathbf{x}_{i+1}) + \alpha_{i+1}(\mathbf{x}, \mathbf{x}_{i+1}) c_\rho(q_i, r_i)\} p_i(y | \mathbf{x}), \quad (27)$$

where $q_i = P_i(y | \mathbf{x})$, $r_i = P_i(y_i | \mathbf{x}_i)$. The update above corresponds to the conditional density update in the multivariate copula update (Fong et al., 2021). A simplification is also suggested in Fong et al. (2021) for the form of $\alpha_{i+1}(\mathbf{x}, \mathbf{x}_{i+1})$, which is

$$\alpha_i(\mathbf{x}, \mathbf{x}') = \frac{\alpha_i \prod_{j=1}^d c_{\rho_x} \{ \Phi(x^j), \Phi(x'^j) \}}{1 - \alpha_i + \alpha_i \prod_{j=1}^d c_{\rho_x} \{ \Phi(x^j), \Phi(x'^j) \}}. \quad (28)$$

Here, c_ρ is the Gaussian copula density as in (23). We initialize $p_0(y | \mathbf{x}) = p_0(y)$ which may be the Lomax or log-normal density as described above, independent of \mathbf{x} . We also standardize the data in the same way as in Section B.8, ignoring the covariates. For predictive resampling, as mentioned in the main paper, we draw $X_{n+1:N}$ through the Bayesian bootstrap.

C EXPERIMENTS

In this section, we provide further details on each of the individual experiments.

C.1 Simulated Data

For the simulated data example, the aim was to show the equivalence between predictive resampling and posterior sampling in the *parametric* case. Our (well-specified) model is

$$f_\theta(y) = \frac{\exp(-y/\theta)}{\theta}, \quad \pi(\theta) = \text{Inverse-gamma}(\theta \mid a_0, b_0)$$

where we have reparametrized so that θ is the mean of the population. Once again, it is convenient to use the $\{t_i, \delta_i\}_{i=1:n}$ notation for the observed data. Under non-informative censoring, the posterior is simply $\pi(\theta \mid \mathcal{D}_n) = \text{IG}(a_n, b_n)$, where $a_n = a_0 + \sum_{i=1}^n \delta_i$, $b_n = b_0 + \sum_{i=1}^n t_i$. The posterior predictive is also analytically tractable as the Lomax(a_n, b_n) distribution, with density and CDF given in (20) and (21). It is also helpful to derive the marginal likelihood, which takes on the form

$$\begin{aligned} p(\mathcal{D}_n) &= \int \prod_{i=1}^n [\delta_i f_\theta(y_i) + (1 - \delta_i) \bar{F}_\theta(c_i)] \pi(\theta) d\theta \\ &= \frac{b_0^{-k} \Gamma(k + a_0)}{\Gamma(a_0)} \left(1 + \frac{\sum_{i=1}^n t_i}{b_0}\right)^{-(k+a_0)} \\ &= \frac{\Gamma(k + a_0)}{\Gamma(a_0)} \frac{b_0^{a_0}}{(b_0 + \sum_{i=1}^n t_i)^{k+a_0}} \end{aligned}$$

where $k = \sum_{i=1}^n \delta_i$. Setting $b_0 = 1$, we maximize the above using gradient descent to elicit a_0 , yielding $a_0 = 1.46$ for our particular example.

In Algorithm 1, for a censored datum $Y_i \geq c_i$, we wish to simulate $Y_i \sim p_{i-1}^{c_i}$, where $p_{i-1}^{c_i}(y) = p(y \mid Y_i \geq c_i, y_{1:i-1})$. Once again, we work in the space of CDFs and draw

$$U_i \sim \mathcal{U}[P_{i-1}(c_i), 1], \quad Y_i = P_{i-1}^{-1}(U_i).$$

In this case, we require P_{i-1}^{-1} , which is tractable and easy to compute, as given in (21). Updating the predictive then involves computing $a_i = a_{i-1} + 1$ and $b_i = b_{i-1} + Y_i$, and the IS weight update involves the Lomax(a_{i-1}, b_{i-1}) CDF at c_i .

To predictively resample $Y_{n+1:N}$, we draw

$$U_i \sim \mathcal{U}[0, 1], \quad Y_i = P_{i-1}^{-1}(U_i),$$

followed by the same updates for a_i and b_i . For the limiting parameter estimate, we utilize the posterior mean, which takes the form

$$\bar{\theta}_N = \frac{b_N}{a_N - 1}$$

for the Inverse-gamma(a_N, b_N) posterior. As an aside, note that $\bar{\theta}_N = \sum_{i=1}^N Y_i / N$ would also work as it is a strongly consistent estimator.

In Figure 7a, we plot the ESS as Algorithm 1 progresses. Although there are two resampling steps, resulting in a decrease in the number of unique particles each time, we still have approximately 600 particles at the end, which is sufficient for estimating $p(y_{n+1} \mid \mathcal{D}_n)$ accurately. In Figure 7b, we plot the paths of $\bar{\theta}_i$ for a few predictive resampling chains, where we see that $N = 2000 + n$ is sufficient for convergence.

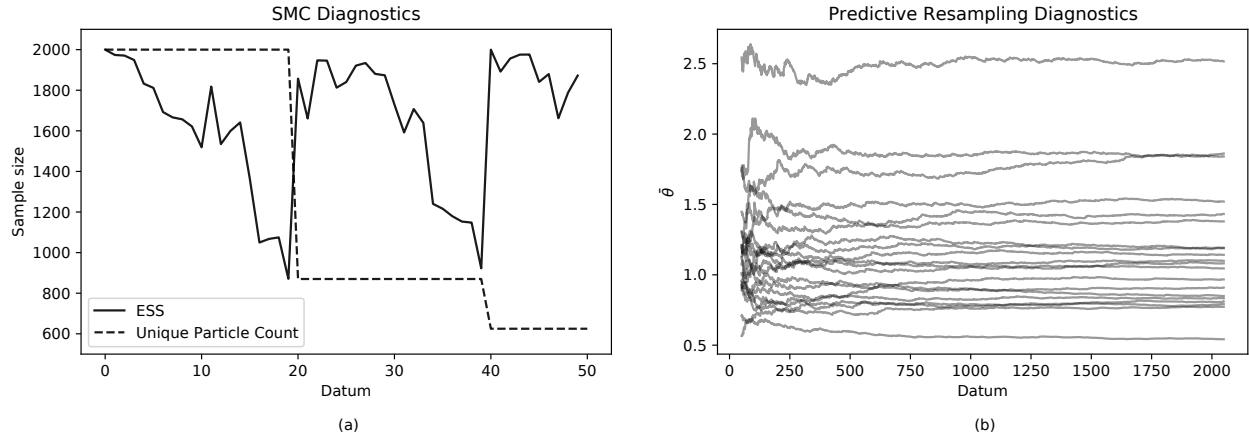


Figure 7: (a) ESS and unique particles; (b) Trajectories of $\bar{\theta}$ from predictive resampling

C.2 Primary Biliary Cirrhosis

For the copula method fit to the full dataset, we set the bandwidth a by maximizing $\hat{p}(\mathcal{D}_n)$ on the grid of $[0.5, 0.6, 0.7, 0.8, 0.9]$ and $[1.1, 1.2, 1.3, 1.4, 1.5]$ for the treatment and placebo datasets respectively. This gives $a = 0.8$ and $a = 1.2$ for the treatment and placebo respectively. For the cross-validation runs with the 50-50 train-test split, we however use the grid $[1.1, 1.2, 1.3, 1.4, 1.5]$ for both treatment and placebo; the difference in grid for placebo is due to the 50-50 train-split preferring a different value of a . For the baseline, we elicit a DPMM with the exponential kernel and Gamma prior, that is the DPMM with kernel and centering measure given in (18) with $b = 1$. We fit the DPMM using the R package `dirichletprocess` (Ross & Markwick, 2018). Like the copula method, we set $a = 0.8$ and $a = 1.2$ in the Gamma centering measure of the DPMM for the treatment and placebo datasets respectively for both the full and the cross-validation fits.

For computing Figure 3 in the main paper, predictive resampling was carried out on a grid of size 149 (not 100 as incorrectly stated in the main paper) between 0 and 21 years. In Figure 8a, the ESS/particle count plots show that no resampling steps were required. In Figure 8b, the 1-Wasserstein distance between P_n and P_N has roughly converged at $N = 2000 + n$ forward steps as implemented in the paper.

For reference, we also plot the posterior mean and 95% credible intervals of the random limiting density p_N for the copula method and the DPMM in Figure 9. We see that while the posterior means are similar, the uncertainty bands are noticeably different.

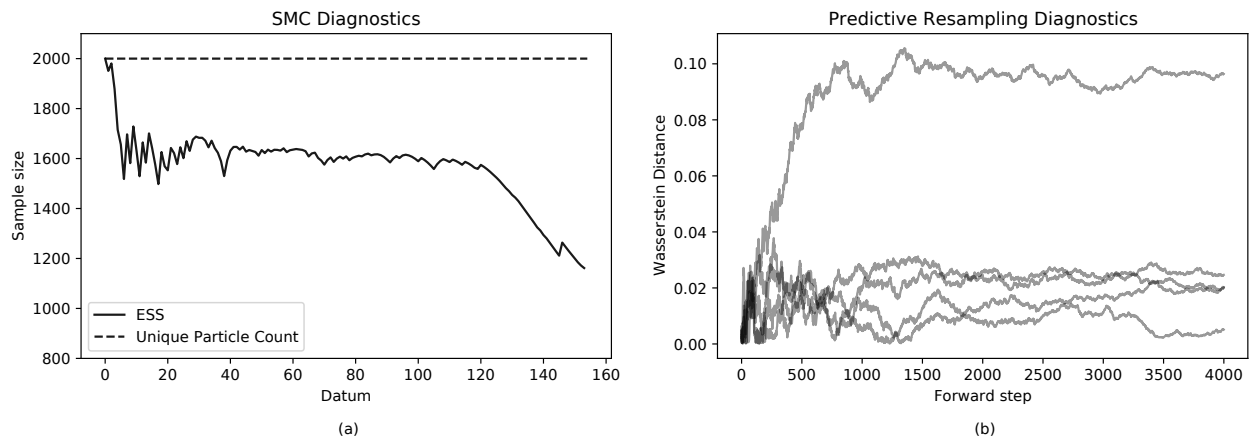


Figure 8: (a) ESS and unique particles; (b) Trajectories of Wasserstein-1 Distance from predictive resampling

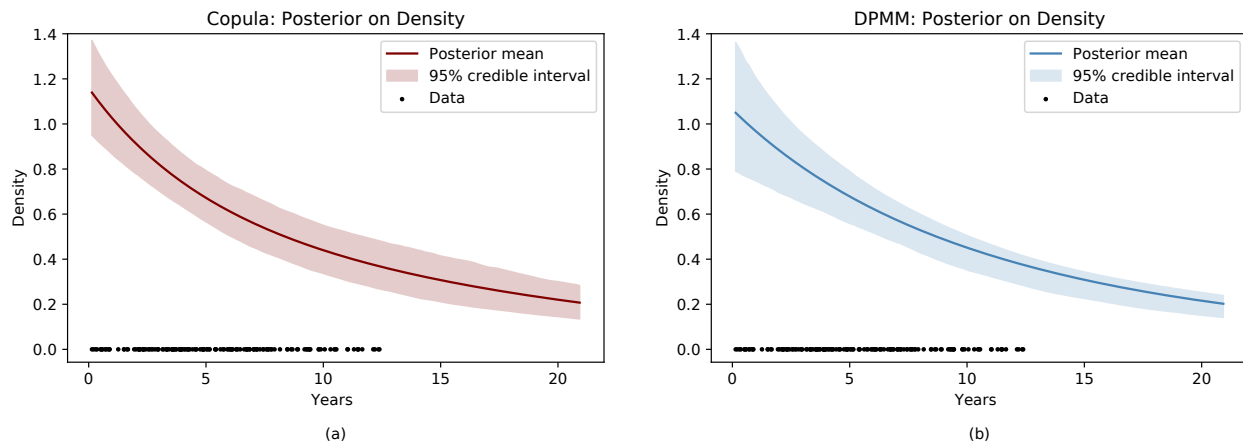


Figure 9: Posterior mean and 95% credible interval of density for (a) copula method and (b) DPMM.

C.3 Survival Regression

For both the melanoma and kidney examples, we set the bandwidth a on the grid $[0.5, 0.6, 0.7, 0.8, 0.9]$ for the copula method. For the nonparametric baseline, we use the `ddpsurvival` function with default settings from the `ddpanova` package (De Iorio et al., 2004). The package uses the methodology introduced in De Iorio et al. (2009) which extends the ANOVA dependent Dirichlet process (DDP) of De Iorio et al. (2004). This method however employs a Gaussian kernel DDP, so we log transform the survival times before fitting, which is equivalent to running a DDP with a log-normal kernel. We also compare to the linear AFT with the log-normal distribution, using the `survreg` function in the `survival` package (Therneau, 2021) in R.

For the melanoma dataset, we also provide additional results for fitting to the full dataset. For the copula method, optimizing the hyperparameters and fitting the model gives us $\rho = 0.6, \rho_x = 0.8$. For Figure 4 in the main paper, we compute the predictive density on a grid of size 56 between 0 and 5565 days (largest datum). Similarly, Figure 10a shows the equivalent plot for the DDP - the fit with the KM plots for the DDP is not as close as the copula method. We also compute the median survival time as a function of x on a x -grid of size 40, which is shown in Figure 10b. The median function of the DDP is smoother than that of the copula method, where the latter is controlled by the value of ρ_x .

To demonstrate predictive resampling, we consider the conditional density/survival function at $x = 3.4$. For the copula method, we carry out $N = 10000 + n$ forward samples with $B = 2000$, which takes 5.6s. However, we point out that this needs to be run for each \mathbf{x} of interest, which may be costly. We plot the posterior mean and 95% credible intervals for the copula estimate of the conditional density and survival function in Figure 11, with the DDP posterior mean functions overlaid. Finally, in Figure 12a, we see that 3 resampling steps reduce the unique particle count to ≈ 400 , and Figure 12b demonstrates that $N = 10000 + n$ is sufficient for convergence.

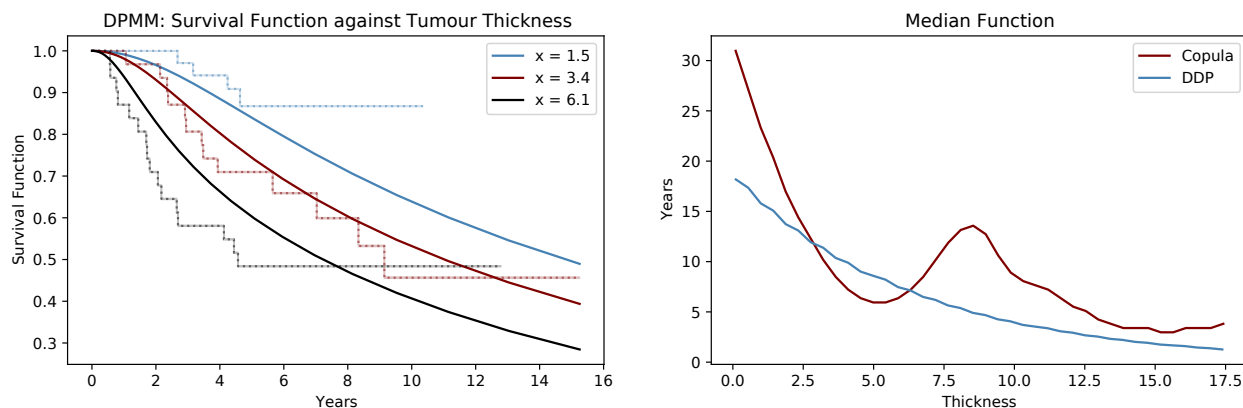


Figure 10: (a) Survival function for DDP method (—) for $x = \{1.5, 3.4, 6.1\}$ with KM (····) fit to windows $\{(1.255, 1.75), (2.7, 4.1), (4.1, 8.1)\}$; (b) Median survival time against tumour thickness x

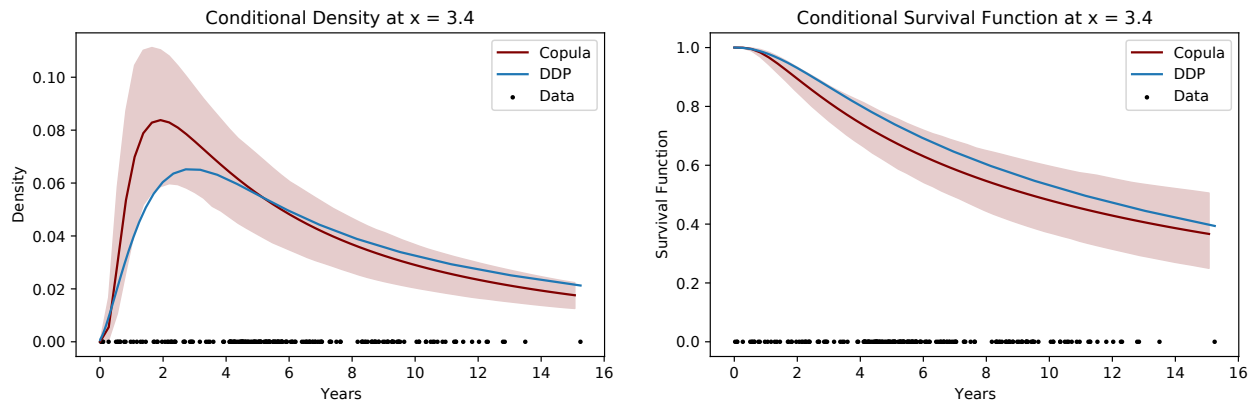


Figure 11: Posterior mean and 95% credible interval (for copula method only) of density for (a) density and (b) survival function.

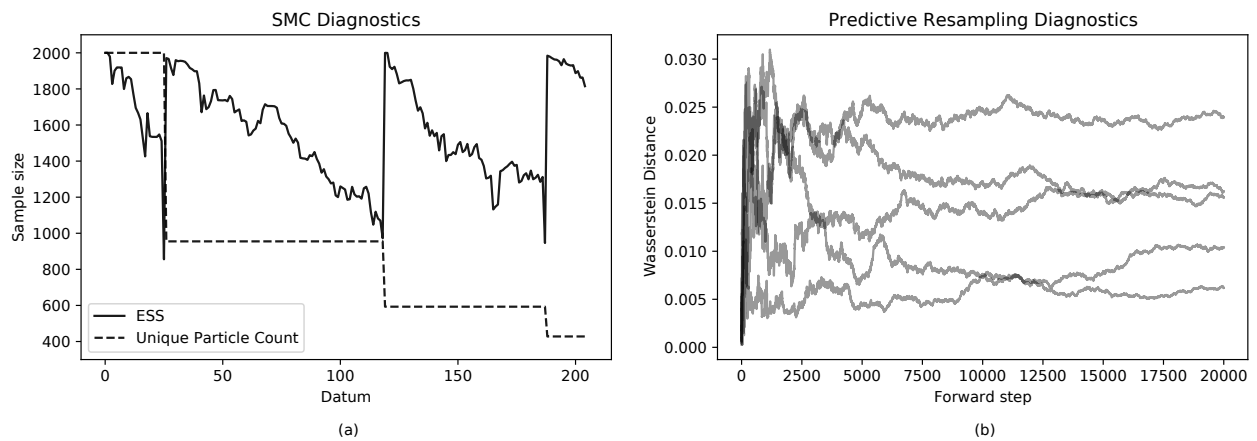


Figure 12: (a) ESS and unique particles; (b) Trajectories of Wasserstein-1 Distance from predictive resampling.