Bayesian Analysis (2022) **TBA**, Number TBA, pp. 1–27

Bayesian Nonparametric Bivariate Survival Regression for Current Status Data*

Giorgio Paulon[†,¶], Peter Müller[‡], and Victor G. Sal y Rosas[§]

Abstract. We consider Bayesian nonparametric inference for event time distributions based on current status data. We show that under dependent censoring conventional mixture priors, including the popular Dirichlet process mixture prior, lead to biologically uninterpretable results as they unnaturally skew the probability mass for the event times toward the extremes of the observed data. Simple assumptions on dependent censoring can fix the problem. We then extend the discussion to bivariate current status data with partial ordering of the two outcomes. In addition to dependent censoring, we also exploit some minimal known structure relating the two event times. We design a Markov chain Monte Carlo algorithm for posterior simulation. Applied to a recurrent infection study, the method provides novel insights into how symptoms-related hospital visits are affected by covariates.

Keywords: survival regression, current status data, Bayesian nonparametrics, joint modeling, race model, recurrent infections.

1 Introduction

We develop Bayesian nonparametric survival regression for bivariate event times that are subject to a single censoring time. In particular, we consider bivariate current status data (Groeneboom and Wellner, 1992), referring to situations where the only available information on each event time is whether or not it exceeds a monitoring time that is common to the two outcomes. Data of this type are often collected in studies on the prevalence of recurrent infectious diseases such as partner studies of HIV infections (Jewell and Shiboski, 1990), or in carcinogenicity testing when a tumor under investigation is occult (Dunson and Dinse, 2002). Wang and Ding (2000) show that the distribution for bivariate current status data is not identifiable using nonparametric maximum likelihood estimation. The goal of this article is twofold: first (Section 3), we propose a dependent censoring scheme that is useful for modeling univariate event time data; second (Sec-

*Dr. Müller acknowledges partial support from grant NSF/DMS 1952679 from the National Science Foundation, and under R01 CA132897 from the U.S. National Cancer Institute. Dr. Sal y Rosas Celi was supported by Dirección de Gestión de la Investigación at the PUCP through grant DGI-2017-496.

[†]Department of Statistics and Data Sciences, University of Texas at Austin, 2317 Speedway D9800, Austin, TX 78712-1823, USA, giorgio.paulon@utexas.edu

[‡]Department of Mathematics, University of Texas at Austin, 2515 Speedway C1200, Austin, TX 78712-1202, USA, pmueller@math.utexas.edu

[§]Sección Matemáticas, Departamento de Ciencias, Pontificia Universidad Católica del Perú, Av. Universitaria 1801, San Miguel 15088, Peru, vsalyrosas@pucp.edu.pe

[¶]Corresponding Author

 https://doi.org/10.1214/22-BA1346

tion 4), we embed such dependent censoring within a flexible model that can estimate the joint distribution of bivariate outcomes with the aid of weak structural assumptions.

Our goal is to develop a flexible model whose components have a biologically meaningful interpretation. Bayesian models are especially useful in such scenarios because of their ability to accommodate prior information. Nonparametric priors are often used to flexibly model a baseline survival function, usually completed with a parametric component that relates survival to a number of predictors. For example, Bayesian extensions of the proportional hazards (PH) model (Cox, 1972) have been proposed in Kalbfleisch (1978) and in Hjort (1990). Generalizations of the accelerated failure times (AFT) model (Buckley and James, 1979) based on a Dirichlet process prior appear in Christensen and Johnson (1988), Kuo and Mallick (1997), Kottas and Gelfand (2001), Hanson and Johnson (2004), or alternatively using Polya trees, for example in Hanson and Johnson (2002). In other cases the main inference target is the hazard function. Sparapani et al. (2016), for instance, construct nonparametric survival regression using a Bayesian additive regression tree (BART) model (Chipman et al., 2010) by adding time as an ordinal predictor to a BART-probit model for the hazard function.

In general, censored observations contribute limited information, via the distribution function or survival function as the corresponding factors of the joint likelihood. This becomes problematic in the case of current status data, as we shall demonstrate. Some proposals have been put forward to tackle these issues. In the case of univariate survival regression, generalizations of the PH model for current status data have been introduced in Cai et al. (2011), Wang et al. (2015) and in Huang (1996). Xue et al. (2004) propose a partly linear AFT model for univariate current status data. More similar to our approach, Wang and Ding (2000) model dependence between bivariate event times via a copula function. Dunson and Dinse (2002) use a Bayesian probit model with normal frailties to induce dependence among multivariate current status data. Nevertheless, there remains a gap in the literature concerning flexible nonparametric regression models for bivariate current status data under dependent censoring.

The motivating case study is inference for the Partner Notification Study (Golden et al., 2005). The goal of the study is to understand the times of development of infection and symptoms for recurrent episodes of gonorrhea and/or chlamydial infections. The study design includes a single follow-up visit for each individual. During this visit the presence of symptoms and infection was recorded, leading to all censored data with shared censoring times for the two outcomes.

Let S denote the time of the onset of symptoms, I the time of infection, and C the time of the hospital visit. Thus, four responses are possible: presence of both disease and symptoms ($I < C, S < C$), absence of both ($I > C, S > C$), absence of symptoms and presence of disease ($I < C, S > C$), and symptoms without disease ($I > C, S < C$). The latter can be explained by the fact that the surveyed symptoms are very generic and might also arise due to other underlying causes. This setup yields data that are bivariate in nature as two outcomes are registered. However, the censoring times, i.e. the hospital visit times, are restricted to a lower dimensional subspace, with a single follow-up visit to assess the presence of both symptoms and disease. Additional complexity arises from the partial ordering of the two outcomes: the infection time is a priori unlikely to follow the symptoms time. This can only occur when the symptoms arise due to other causes.

Our model introduces features to reflect this consideration. We use a mixture model with one submodel being subject to an order constraint, representing symptoms due to the infection of interest, and another submodel without such constraint, allowing for symptoms due to other causes. While our discussion is motivated by a specific application, we note that similar data formats arise frequently in any study that involves data collection during follow-up visits. For example, doctors might record tumor recurrence using a CT scan and symptoms as reported by patients.

In the first part of this article, we demonstrate with simple examples the problems arising from the use of standard techniques with current status data. We then introduce structural assumptions that allow us to correctly estimate a meaningful distribution of the latent bivariate outcomes. We propose a Bayesian nonparametric (BNP) approach for modeling the joint distribution under these assumptions. An important feature of BNP models is their large support, allowing us to approximate essentially arbitrary distributions (Ishwaran and James, 2001). To handle covariates, our approach is based on the dependent Dirichlet process (DDP) prior introduced by MacEachern (1999). See also the discussion in De Iorio et al. (2004) for the special case of categorical covariates.

In summary, the main features of our approach here are (i) using BNP priors together with informative monitoring times to address the challenges that arise with parametric inference for univariate current status data; (ii) exploiting known structural dependence of the two event times to allow for some borrowing of strength; and (iii) exploiting heterogeneity if known. While our motivating application gives rise to particular assumptions for (ii) and (iii), the same framework remains valid in more generality. Current status data on any two event times that are likely to be ordered give rise to the same setup.

The rest of this article is organized as follows. Section 2 describes the clinical study that motivates this article. Section 3 develops the proposed inference approach starting from a simple univariate case. Section 4 uses the univariate model as a building block for bivariate outcomes and outlines a Markov chain Monte Carlo (MCMC) strategy for estimation. Section 5 presents the results of the proposed method applied to the Partner Notification Study. Section 6 finishes with concluding remarks. Additional details, including proofs, the MCMC scheme, convergence diagnostics and simulation studies are deferred to the supplementary materials (Paulon et al., 2022).

2 The Partner Notification Study

The Partner Notification Study (Golden et al., 2005) enrolled men and women who received a diagnosis of gonorrhea or genital chlamydia at most 14 days prior to enrollment. It was conducted in King County Seattle (Washington state, U.S.A.) from September 1998 to March 2003. Researchers contacted clinicians who diagnosed and treated the infections to seek permission to contact their patients. To minimize the likelihood of reinfection before randomization, patients who could not be contacted within 14 days after treatment were not eligible for the study, yielding a total of $n = 1864$ participants. The study was designed to gather current status data of recurrent gonorrhea or chlamydial infection in patients 3 to 19 weeks after randomization to standard (control group,

933 individuals) or expedited partner therapy (intervention group, 931 individuals). The primary outcome was persistent or recurrent gonorrhea and/or chlamydial infection in the original participants within 90 days after enrollment, although actual follow-up times varied considerably (19 to 161 days) due to both difficulty scheduling follow-up visits and anticipated hospitalizations due to symptoms. A scheme illustrating the trial follow-up timing is shown in Figure 1. The issue of patient noncompliance is handled by our model via a dependent censoring mechanism. Sal y Rosas and Hughes (2011) previously analyzed data on infection times from the same study, explicitly allowing for outcome misclassification.

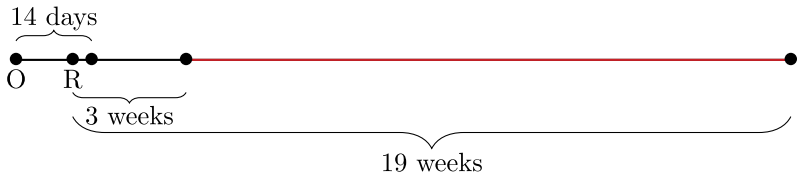


Figure 1: "O" marks the date of the original diagnosis for the recurrent infection studied in the trial, "R" is the randomization date. The randomization must occur in the 14 days following the original infection. The red segment represents the time window for the follow-up visit.

When visiting the hospital, two outcomes were recorded for each patient: presence of reinfection (I_i) and of symptoms (S_i). Thus, two latent event times (I_i, S_i) correspond to a common censoring time C_i, i.e. the time of the hospital visit. The data record for each patient C_i, and whether the patient has already experienced the infection $\Delta_{I_i} = \mathbb{1}(I_i < C_i)$ and some symptoms $\Delta_{S_i} = \mathbb{1}(S_i < C_i)$. While in general symptoms should follow the onset of infection, the definition of symptoms in this study is very generic and they might also be due to other causes. In the case $I_i < S_i$ it is impossible to tell whether symptoms are due to the disease of interest or any other cause, while when $I_i > S_i$ the symptoms are known to be due some other cause. Importantly, the protocol, and thus the data did not include recording of actual event times for symptoms, even in the case of $S_i < C_i$.

The recorded $n = 1832$ follow-up visits included patients reporting all four possible combinations of censoring for the two outcomes: $n_{00} = 1303$ patients did not experience symptoms and tested negative for the infection; $n_{10} = 121$ patients tested positive for the infection but were not experiencing any symptoms (asymptomatic infections); $n_{01} = 325$ patients tested negative for the infection but were experiencing symptoms (due to other causes); $n_{11} = 83$ patients tested positive for the infection and were also experiencing symptoms (symptomatic infections).

Figure 2 shows two univariate nonparametric maximum likelihood estimates (MLE) (Groeneboom and Wellner, 1992) for the distributions of time to infection I_i and time to symptoms S_i, stratified by two covariates (gender and intervention) under the assumption of independent censoring. Female participants seem to experience symptoms sooner than men. The flat region of survival probability in the middle of the range of the

observed data is typical for the nonparametric MLE and is clinically highly implausible. In Section 3 we show that the accumulation of probability mass toward the bounds of the observation range is a common issue when dealing with current status data under dependent censoring. Moreover, these nonparametric MLE estimates represent marginal effects and do not take into account any correlation that is expected between the time to infection and time to symptoms.

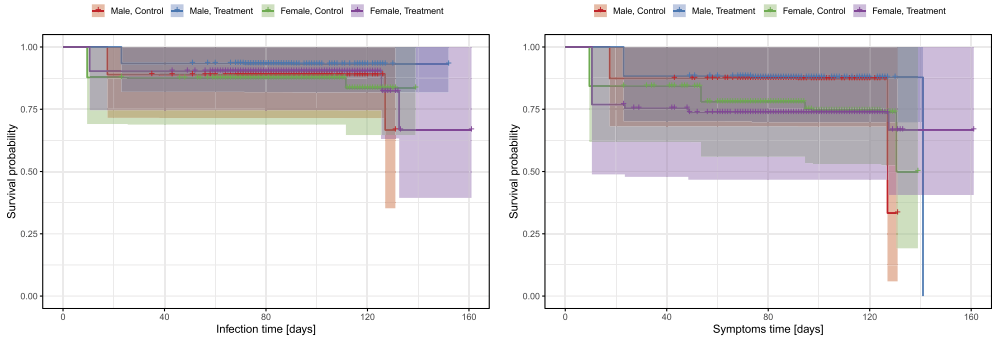


Figure 2: Nonparametric MLE for infection times (left panel) and time until symptoms (right panel), stratified by the binary covariates *gender* and *treatment* fixing *age* to the average age in the sample. Shaded areas represent pointwise 95% confidence intervals.

3 Univariate Survival Analysis for Current Status Data

We introduce a Bayesian nonparametric (BNP) modeling strategy for current status data, first in a simple univariate case. We show that the nonparametric MLE for current status data has an undesirable feature that makes it biologically uninterpretable when the independent censoring assumption is violated. More specifically, most of the probability mass is accumulated toward the extremes of the data range.

Let S_i represent the latent event time for patient i, Δ_i be a censoring indicator with $\Delta_i = 1$ if the event has been detected and $\Delta_i = 0$ otherwise, and let C_i denote the censoring time. That is, when $\Delta_i = 1$, then $S_i \leq C_i$ (left censored), otherwise $S_i > C_i$ (right censored). We want to infer the unknown density $f_S(s)$ based on only the observed censoring times and indicators $(C_i, \Delta_i), i = 1, \ldots, n$.

3.1 Limitations of the Maximum Likelihood Estimator

We show that under moderate sample sizes the nonparametric MLE does not provide meaningful estimates of the event time distribution for current status data under dependent censoring. Without loss of generality, we assume that the censoring times are ordered, $C_i \leq C_{i-1}$, and that $\Delta_1 = 1, \Delta_n = 0$. Define $P = \{\{i > 1 \text{ s.t. } \Delta_i = 1, \Delta_{i-1} = 0\} \cup \{1\}\}$ as the set of indices of left censored observations immediately following a

right censored observation, i.e. the set of indices of the pairs $(\Delta_{i-1}, \Delta_i) = (0, 1)$. Next, let $J = |P|$ and $\mathbf{C}^\star = (C_1^\star, \ldots, C_J^\star) = (C_i, i \in P)$ denote the corresponding censoring times. See Figure 3 for an illustration.

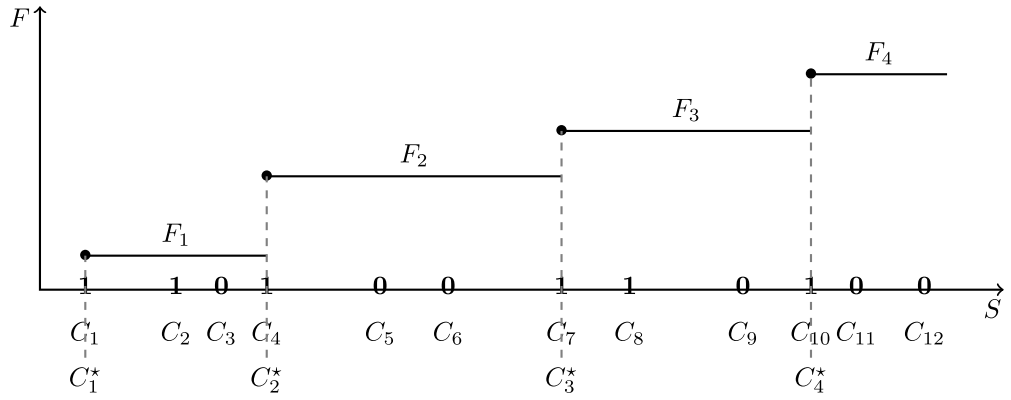


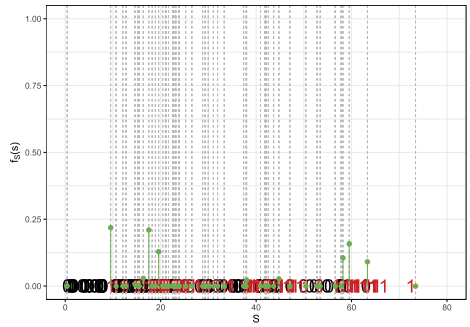
Figure 3: An example with $n = 12$ latent event times. The set of support points is $P = \{1, 4, 7, 10\}$. On the x-axis, 0 and 1 indicate the values of Δ_i.

Let $C_{J+1}^\star$ denote any point to the right of the last right censored observation. The times $\mathbf{C}^\star \cup \{C_{J+1}^\star\}$ are the only points where probability mass can accumulate under the nonparametric MLE. In other words, the support of a discrete nonparametric density estimate for the latent event times can have probability mass only at the left censoring times. More specifically, the support of the MLE is restricted to C_i's corresponding to (i) the left censored observation in every "01" pair, (ii) the first left censored observation, and (iii) any point to the right of the last right censored observation. To see this, write the unknown distribution $f_S(\cdot)$ of the latent times S_i as a discrete probability measure with atoms at the $C_j^\star$, i.e.

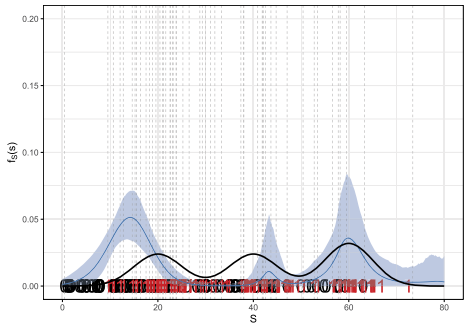
$$f_S(s) = \sum_{j=1}^{J+1} p_j \delta_{C_j^\star}. \tag{3.1}$$

We denote with $F_j = \sum_{k \le j} p_k$ the cumulative density function (c.d.f.) and with $\bar{F}_j = 1 - F_j$ the survival function at the supporting point $C_j^\star$. To see that the nonparametric MLE for $f_S(s)$ can only have support on the set $\mathbf{C}^\star$, assume that $f_S(s)$ were to include any additional probability mass p at $C_i \neq C_j^\star, j = 1, \ldots, J$. Let $j^\star = \max_j\{C_j^\star < C_i\}$ and $j' = \min_j\{C_j^\star > C_i\}$ denote the point mass in $\mathbf{C}^\star$ closest to C_i from the left and from the right, respectively. Then, if $\Delta_i = 1$ one could move the probability mass p to $C_{j^\star}^\star$, and if $\Delta_i = 0$ one could move the probability mass p to $C_{j'}^\star$. Either would leave the likelihood function unchanged.

Groeneboom and Wellner (1992) introduce a simple EM algorithm to estimate the unknown c.d.f for the latent times under the independent censoring assumption. Let $l_j = \#\{C_i \text{ s.t. } \Delta_i = 1, C_j^\star \le C_i < C_{j+1}^\star\}$ and $r_j = \#\{C_i \text{ s.t. } \Delta_i = 0, C_j^\star < C_i \le C_{j+1}^\star\}$ denote the runs of left and right censored observations, respectively. Let $\mathbf{Y} = \{(C_i, \Delta_i)\}_{i=1}^n$



(a) Green vertical pins represent the non-parametric MLE estimate for the point masses obtained via the EM algorithm.



(b) In blue, posterior mean (and shaded pointwise 95% credible intervals) for a simple mixture of $K = 3$ normal distributions. In black, the simulation truth.

Figure 4: Simulated data. Right and left censoring times are represented by black "0" and red "1", respectively, on the x-axis. Vertical dashed lines represent the possible support points for $f_S(s)$.

denote the data and $\mathbf{p} = \{p_j\}_{j=1}^{J+1}$ denote the parameters. The log-likelihood function under model (3.1) is

$$\begin{aligned}\ell(\mathbf{p}; \mathbf{Y}) &= \sum_{i=1}^{n} \{\delta_1(\Delta_i) \cdot \log F(C_i) + \delta_0(\Delta_i) \cdot \log \bar{F}(C_i)\} \\ &= \sum_{j=1}^{J} \{l_j \log F_j + r_j \log \bar{F}_j\}.\end{aligned}$$

If instead we knew the latent times $\mathbf{z} = \{S_i\}_{i=1}^{n}$, we could use the full data log-likelihood $\ell(\mathbf{p}, \mathbf{z}) = \sum_{j=1}^{J} n_j \log(p_j)$ where $n_j = \#\{S_i = C_j^\star\}$. The expectation of this full data log-likelihood with respect to $\mathbf{z}$ involves only $\mathbb{E}(n_j \mid \mathbf{p})$. This motivates an easy Expectation Maximization (EM) algorithm, illustrated in Algorithm 2 in the supplementary materials.

We illustrate the algorithm on simulated data with $n = 200$ latent times generated from a mixture of three normal distributions with weights $\boldsymbol{\pi} = (0.4, 0.2, 0.4)^\intercal$, locations $\boldsymbol{\mu} = (20, 40, 60)^\intercal$ and scale parameters $\boldsymbol{\sigma}^2 = (25, 25, 25)^\intercal$. The censoring times C_i were simulated according to model (3.2), defined below. As shown in Figure 4a, despite a large number of support points $\mathbf{C}^\star$, in this simulation study most of the probability mass under the unconstrained MLE accumulates close to the bounds of the range of the data. One might conjecture that the issue is caused by the excessively flexible nature of the unconstrained MLE. However, under dependent censoring even parametric models fail to capture the underlying distribution of the latent times. For comparison, we carried out inference using a mixture of $K = 3$ Gaussian distributions for the latent times S, matching the nature of the actual simulation truth. In Figure 4b, we show the posterior

mean for the unknown event time distribution under this model when fitted to the current status data in the simulation study. The posterior estimated distribution still allocates most probability mass toward the extremes of the data and misses the central peak, despite using an analysis model that matched the actual simulation truth. The same issues arise to either side of the range of the data when only either left censored observations or right censored observations are available.

3.2 A Bayesian Nonparametric Model

We introduce some assumptions to address the issues described in the previous section. In short, we regularize the model by (i) explicitly modeling the dependence between censoring times and latent event times, and (ii) introducing prior shrinkage with a flexible nonparametric Bayesian prior. We opt for a proper prior probability model to ensure that the posterior is proper. As a consequence, inference of any summary of interest is regularized and shrunk towards the corresponding prior summary.

Knowledge about dependent censoring allows us to gain some information on $f_S(\cdot)$ from the censoring times. For example, in the motivating case study it is expected that patients seek help shortly after they experience symptoms. This information can be incorporated in the model in many ways. For our specific application, we assume that the censoring times C_i's arise from a race between a return by schedule versus a return driven by the onset of symptoms, as

$$C_i \mid S_i, \lambda \overset{d}{=} \min\{S_i + \text{Exp}(\lambda); \text{Unif}(A, B)\}, \tag{3.2}$$

where $\overset{d}{=}$ denotes equality in distribution, A and B represent the range of the observation window, and $\text{Exp}(\lambda)$ and $\text{Unif}(A, B)$ refer to random variables with the respective distribution. In other words, the visit time to the hospital can either occur uniformly in the observation range (visit by protocol) or it can closely follow the symptoms onset (visit prompted by symptoms). The resulting distribution can be easily evaluated. Below we provide an explicit expression for the probability density function in the case $A \leq c \leq B$, $s \geq A$, which in our application was guaranteed by choosing $A = 0$.

Proposition 1. *The p.d.f. of the conditional distribution of censoring times given the event times is given by*

$$f_{C|S}(c \mid s) = \frac{\mathbb{1}\{c \leq s\}}{B - A} + \frac{\mathbb{1}\{c > s\}}{B - A} e^{-\lambda(c-s)}\{1 + \lambda(B - c)\}, \quad A \leq c \leq B, s \geq A.$$

A proof is provided in the supplementary materials. An alternative expression for the p.d.f. can be found for the case $s < A$.

The regularization induced by the dependent censoring mechanism in (3.2) yields more interpretable inference, but some issues remain. Figure 5 shows the nonparametric density estimate for such a model under dependent censoring, and it highlights that inference still fails to recover the simulation truth. Two important features that are missing from this model are prior smoothing for the distribution of the latent event times as well as borrowing of information within homogeneous patient subpopulations.

Motivated by the described limitations we specify a Bayesian nonparametric prior for the latent event times. Relaxing parametric assumptions allows for greater modeling flexibility, robustness against misspecification of a parametric statistical model and, as a result, more honest uncertainty assessment than under a parametric model. At the same time, prior smoothing and shrinkage result in more realistic and clinically meaningful estimates compared to a nonparametric MLE. In addition, a BNP model can allow to accommodate heterogeneous patient populations, for example using a Dirichlet process (DP) mixture model.

Let f denote the distribution of the variable of interest (in our case the event times). A DP mixture model assumes $f(y) = \int k(y \mid \theta)\, dH(\theta)$ with $H \sim DP(M, H_0)$, where DP indicates a DP prior with total mass α and base measure H_0. See, for example Müller et al. (2015, Chapter 2) for a review of the DP and DP mixtures. For later reference we note that $H = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}$ is a.s. discrete with $\theta_h \overset{\text{iid}}{\sim} H_0$, and a stick breaking prior (Sethuraman, 1994) for the weights $\pi_h = q_h \prod_{\ell<h}(1-q_\ell)$ with $q_h \overset{\text{iid}}{\sim}$ Beta$(1, M)$. Two natural choices of sampling models $k(y \mid \theta)$ for survival data are the log normal and the Weibull families. In applications with event times close to 0, it can be convenient to first log transform the data and then use normal kernels, i.e. use log normal kernels. In many instances, however, a mixture of normals may suffice (Lo, 1984) and is often preferred. Another attractive choice are gamma mixtures, especially with a view towards asymptotic results (Hanson, 2006; Poynor and Kottas, 2019). Although our implementation uses normal kernels for computational convenience, we recognize that distributions with support on the positive real line would be more natural and that very little would change in the model construction if gamma kernels are used. In our particular motivating case study, however, we did not observe any imputed negative event times over the $1,250$ Monte Carlo iterations (after burn-in and thinning) using normal kernels.

Finally, like most survival analysis approaches we treat the data as continuous random variables. This is a reasonable approximation even when the data is discrete (as in the case of dates) but with a resolution relatively small compared with time window of the experiment, as in the application motivating this article.

The BNP-CS model

The resulting model can be summarized as

$$\begin{aligned} C_i \mid S_i, \lambda &\overset{d}{=} \min\{S_i + \text{Exp}(\lambda); \text{Unif}(A, B)\} \\ S_i \mid H &\sim \int \text{N}(S_i \mid \mu, \sigma^2) dH(\mu, \sigma^2), \quad H \sim \text{DP}(M, H_0). \end{aligned} \tag{3.3}$$

The model is completed with base measure

$$H_0 = \text{N}(\mu_k \mid \mu_0, \sigma_k^2/\kappa_0) \times \text{IG}(\sigma_k^2 \mid a_\sigma, b_\sigma),$$

and priors $M \sim \text{Gamma}(a_M, b_M)$, $\lambda \sim \text{Gamma}(a_\lambda, b_\lambda)$. We refer to (3.3) as BNP for current status (BNP-CS) model, with the name implying that alternative BNP priors

other than the DPM (see, e.g. Müller et al., 2015) could be used if desired. Using the stick-breaking construction of the DP, the second line of model (3.3) can be rewritten as

$$S_i \mid \{\mu_k, \sigma_k^2, \pi_k\}_{k=1}^{+\infty} \sim \sum_{k=1}^{+\infty} \pi_k \mathrm{N}(S_i \mid \mu_k, \sigma_k^2)$$

with $(\mu_k, \sigma_k^2) \sim H_0$, i.i.d., and $\boldsymbol{\pi} \sim \mathrm{SB}(M)$, where $\mathrm{SB}(M)$ denotes the stick-breaking construction for the weights, with concentration parameter M. In our implementation, we also use priors on the hyperparameters $\mu_0, \kappa_0, b_\sigma$.

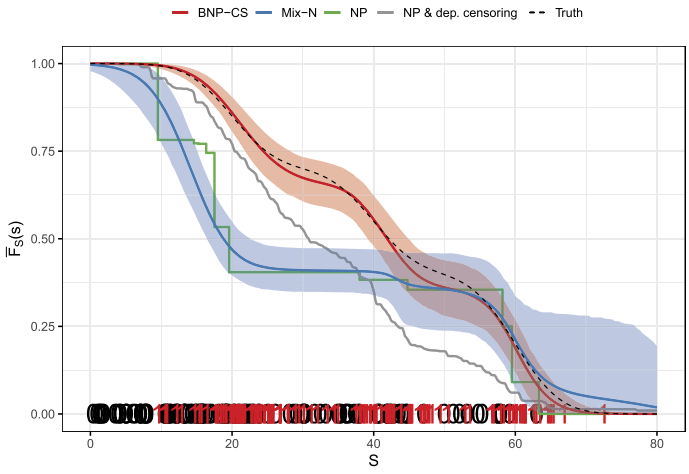


Figure 5: Simulated data: Right and left censoring times are represented by "0" and "1", respectively, on the x-axis. The green step function shows an estimate of the survival function under the nonparametric MLE using independent censoring. The gray step function shows an estimate of the survival function under the nonparametric MLE using dependent censoring. The blue curve shows an estimate of the survival function under a mixture of normals model (note that the simulation truth is in the same parametric family). The red curve shows an estimate of the survival function under the proposed model. Shaded areas represent pointwise 95% credible intervals for the estimated survival functions. The black dashed line represents the simulation truth.

Inference under the BNP-CS model for the same data used in the illustration of Section 3.1 recovers the underlying truth better than inference under the model with independent censoring. Figure 5 shows the survival function estimated under (i) an unconstrained nonparametric model, (ii) a nonparametric model under dependent censoring, (iii) a mixture of $K = 3$ normal distributions with independent censoring, and (iv) the proposed nonparametric model with dependent censoring. Although the model under (iii) matches the simulation truth, inference under models (i) – (iii) fails to recover a meaningful estimate, while inference under (iv) successfully exploits the information that is contained in the observed C_i.

4 Bivariate Survival Regression for Partially Ordered Current Status Data

4.1 A Bivariate Event Time Model

We now use the BNP-CS model (3.3) as a building block for bivariate outcomes. Beyond the already discussed dependence of S_i and C_i, we add some more structure based on prior knowledge of the underlying process. Without additional assumptions, the joint distribution for bivariate current status data is not likelihood identifiable, in general (Wang and Ding, 2000).

To see that this is the case, let $F_I = P(I_i \leq C_i) = P(\Delta_{I_i} = 1)$, $F_S = P(S_i \leq C_i) = P(\Delta_{S_i} = 1)$, and $F_{IS} = P(I_i \leq C_i, S_i \leq C_i) = P(\Delta_{I_i} = \Delta_{S_i} = 1)$. Note that F_I, etc. are functions of C_i. Assuming independent censoring, the joint likelihood function for bivariate current status data $\mathbf{Y} = (\Delta_{I_i}, \Delta_{S_i}, C_i, i = 1, \ldots, n)$ is then

$$\prod_i \left\{ F_{IS}^{\Delta_I \Delta_S} (F_I - F_{IS})^{\Delta_I (1-\Delta_S)} (F_S - F_{IS})^{(1-\Delta_I)\Delta_S} (1 - F_S - F_I + F_{IS})^{(1-\Delta_I)(1-\Delta_S)} \right\}, \tag{4.1}$$

where, for ease of notation, we suppressed the i index in Δ_I and Δ_S. Only the three univariate distributions F_I, F_S and F_{IS} are likelihood identifiable. In other words, (4.1) is invariant with respect to changing any other element of the joint distribution that does not change F_I, F_S or F_{IS}. For example, any general bivariate quantile $P(I < a, S < b)$ for $a \neq b$ would not be identifiable. In particular, this implies that nonparametric inference on the bivariate event time distribution is only possible with some additional assumptions. To achieve inference on the joint distribution of (I, S) we can either (i) estimate the joint distribution under parametric or semiparametric assumptions, or (ii) build the joint model from the two identifiable marginal distributions and a particular choice for their dependence structure. Our approach follows mainly the latter strategy.

In the application, we distinguish between symptoms that arise due to the infection and symptoms that arise due to other causes. In the former case, we assume a parametric model for the lag time $L = S - I$ between infection time I and onset of symptoms S. In the latter case, we assume independence between I and S. That is, we model the bivariate event time distribution $f_{IS}(I, S)$ as a mixture model in which one of the two components is subject to the order constraint $I < S$, i.e.

$$f_{IS}(I, S) = w f^{\star}_{IS}(I, S) + (1 - w) f'_{IS}(I, S), \tag{4.2}$$

where $f'_{IS}(I, S)$ is subject to $I < S$, whereas $f^{\star}_{IS}(I, S)$ is not. Therefore, $f^{\star}_{IS}(I, S)$ can be interpreted as the distribution of (I, S) for a patient with symptoms "due to other causes". Figure 6 shows the support of the two components of the mixture as well as the support for the latent times corresponding to the four possible censoring indicators, i.e. factors in (4.1).

We add two main assumptions to introduce more structure in (4.2), which will eventually facilitate inference: (i) under $f^{\star}_{IS}(I, S)$, the time to symptoms (due to other causes) and time to infection are independent; (ii) under $f'_{IS}(I, S)$, the latency time

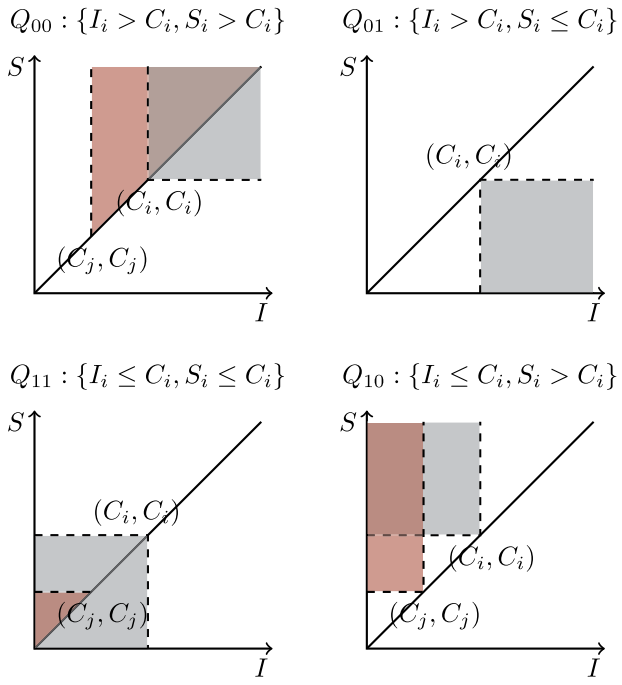


Figure 6: Support for the latent times $I > 0, S > 0$, corresponding to the four cases Q_{00}, Q_{01}, Q_{11}, and Q_{10}. The gray quadrants represent the support for the latent times corresponding to the observed censoring times (C_i, C_i) under $f^{\star}_{IS}(I, S)$. The red shaded areas represent the support for the latent times under $f'_{IS}(I, S)$.

$L = S - I$ and the time to infection are independent. These are reasonable model assumptions which are clinically plausible in the motivating application. Here L is the delay from the onset of illness to the development of symptoms. The assumed marginal $f_I(\cdot)$ on I is shared by both, $f^{\star}_{IS}$ and f'_{IS}. Thus, model (4.2) becomes

$$f_{IS}(I, S) = w f_I(I) f^{\star}_S(S) + (1 - w) f_I(I) f_L(S - I). \tag{4.3}$$

Finally, note that by introducing in (4.3) dependence between S and I, we implicitly also introduce dependence between I and C through (3.3), thus regularizing inference on both f_I and f_S. For later reference we note that sampling $(I_i, S_i) \sim f_{IS}$ can be equivalently written as a hierarchical model with latent indicators, say v_i, with $p(v_i = 1) = w$ and

$$(I_i, S_i \mid v_i) \sim \begin{cases} f_I(I) f^{\star}_S(S) & \text{if } v_i = 1 \\ f_I(I) f_L(S - I) & \text{if } v_i = 0. \end{cases} \tag{4.4}$$

The second component in (4.3) includes the constraint $I < S$ as a positivity constraint on the latency time $L > 0$. Recent approaches to deal with hard constraints use relaxation methods that replace the hard constraint with priors that penalize departures

outside of the constraint subspace (Duan et al., 2020). Alternatively, Patra and Dunson (2018) developed methodology that uses unconstrained inference and then projects the posterior draws onto the constrained subspace. In our model, assigning positive support to the reparametrized variable L automatically ensures the required order constraint $I < S$. In the following, we will use $L \mid \lambda_L \sim \text{Exp}(\lambda_L)$. However, if desired, any richer parametric family, e.g. a Gamma distribution, could be used. As a consequence, under $f'_{IS}(I, S) = f_I(I)f_L(S - I)$, time to symptoms and time to infection are dependent.

Let $M_{00}, M_{01}, M_{11}, M_{10}$ denote the likelihood factors corresponding to the four cases in Figure 6, i.e. the four factors in (4.1). Dropping the subject subscripts, let $F_I = F_I(C_i)$, $\bar{F}_I = 1 - F_I(C_i)$, and similarly for $F^\star_S$ and $\bar{F}^\star_S$. The structural assumptions allow us to replace the general bivariate quantiles arising from (4.2) by simple expressions that only use the univariate marginal distributions. Hence we get

$$\begin{aligned} M_{00} &= w\bar{F}_I\,\bar{F}^\star_S + (1-w)\bar{F}_I, \\ M_{01} &= w\bar{F}_I F^\star_S, \\ M_{11} &= wF_I F^\star_S + (1-w)\int_0^C f_I(I)F_L(C-I)dI, \\ M_{10} &= wF_I\bar{F}^\star_S + (1-w)\int_0^C f_I(I)\bar{F}_L(C-I)dI. \end{aligned} \tag{4.5}$$

Simulation experiments to assess successful estimation of the model parameters can be found in the supplementary materials.

Considering identifiability of $(F_I, F^\star_S, w, \lambda_L)$, we use a finite grid $(g_1, \ldots, g_G)$ over I and S as it is used for the posterior summary figures in Figure 7 and Figure 8. In what follows, we use a definition introduced in Basu (1983), as stated in Swartz et al. (2004).

Definition 1 (Identifiability)**.** *Let U be an observable random variable with distribution function F_θ and let F_θ belong to a family $\mathcal{F} = \{F_\theta : \theta \in \Omega\}$ of distribution functions indexed by a parameter θ. Here θ could be scalar or vector-valued. We say that θ is nonidentifiable by U if there is at least one pair $(\theta, \theta') \in \Omega^2, \theta \neq \theta'$, such that $F_\theta(u) = F_{\theta'}(u)$ for all u. In the contrary case we shall say θ is identifiable.*

Proposition 2. *Assume the model for the lag time L is such that an equation for the c.d.f., $F_L(x) = u$ for $u \in (0, 1)$ has a unique solution in the parameter λ_L. Assume that the prior support $\mathcal{S}_w$ for w is s.t. $\mathcal{S}_w \subseteq (0, 1)$ and the prior support for F_I, $F^\star_S$ implies for the discretized event times $F_I(g_k) < 1$, $F^\star_S(g_k) > 0$ for $k = 1, \ldots, G$. Then model (4.5) is identifiable.*

A proof is provided in the supplementary materials.

4.2 Bayesian Nonparametric Priors

The model is completed by introducing priors for the two unknown distributions, assuming nonparametric mixture models for both $f_I(I)$ and $f_S^\star(S)$,

$$\begin{aligned} f_I(I) &= \int \mathrm{N}(I \mid \boldsymbol{\theta}^{(I)}) dH^{(I)}(\boldsymbol{\theta}^{(I)}) = \sum_{k=1}^{+\infty} \pi_k^{(I)} \mathrm{N}(I \mid \mu_k^{(I)}, \sigma_k^{(I)2}), \\ f_S^\star(S) &= \int \mathrm{N}(I \mid \boldsymbol{\theta}^{(S)}) dH^{(S)}(\boldsymbol{\theta}^{(S)}) = \sum_{k=1}^{+\infty} \pi_k^{(S)} \mathrm{N}(S \mid \mu_k^{(S)}, \sigma_k^{(S)2}), \end{aligned} \tag{4.6}$$

where $\boldsymbol{\theta}^{(I)} = (\mu^{(I)}, \sigma^{(I)2})$ and $\boldsymbol{\theta}^{(S)} = (\mu^{(S)}, \sigma^{(S)2})$. Here $H^{(I)}(\cdot) = \sum_k \pi_k^{(I)} \delta_{\boldsymbol{\theta}^{(I)}}$, and similarly $H^{(S)}$, are the random mixing measures. The model is completed with a prior probability model on $H^{(I)}$ and $H^{(S)}$. Prior distributions on random probability measures are known as nonparametric Bayes (BNP) models.

Using a nonparametric prior on $H^{(I)}$ and $H^{(S)}$ the model becomes a mixture of normals with respect to the chosen random mixing measure. For example, in our implementation we assume a DP prior again, as in (3.3), now using two instances for f_I and $f_S^\star$. Alternatively, any other nonparametric Bayesian prior (e.g. James et al., 2009) could be used. The following result gives the marginal distributions implied by our construction.

Theorem 1. *The marginal distributions implied by model* (4.3) *with priors* (4.6) *are*

$$f_I(I) = \sum_{k=1}^{+\infty} \pi_k^{(I)} \mathrm{N}(I \mid \mu_k^{(I)}, \sigma_k^{(I)2}), \tag{4.7}$$

$$f_S(S) = w \sum_{k=1}^{+\infty} \pi_k^{(S)} \mathrm{N}(S \mid \mu_k^{(S)}, \sigma_k^{(S)2}) + (1-w) \sum_{k=1}^{+\infty} \pi_k^{(I)} \mathrm{EMG}(S \mid \mu_k^{(I)}, \sigma_k^{(I)2}, \lambda_L), \tag{4.8}$$

where $\mathrm{EMG}(\mu, \sigma^2, \lambda)$ *denotes the exponentially modified Gaussian distribution (Grushka, 1972).*

An easy proof is provided in the supplementary materials. Model (4.3) together with (4.6) and (3.3) for $p(C_i \mid S_i)$ defines the proposed **bivariate BNP-CS model** for current status data.

One of the reasons for the wide use of BNP mixtures like (4.6) is the induced prior on a random partition. Consider $I_i \sim f_I$, $i = 1, \ldots, n$. Under model (4.6) we can introduce latent indicators, say $r_i^{(I)}$, and write instead

$$p(I_i \mid r_i^{(I)} = k) = \mathrm{N}(\mu_k^{(I)}, \sigma_k^{(I)2}) \quad \text{and} \quad p(r_i^{(I)} = k) = \pi_k^{(I)}.$$

The $r_i^{(I)}$'s can be interpreted as cluster membership indicators. We see then how this formulation implicitly defines a probability model $p(\mathbf{r}^{(I)})$ on a partition $\mathbf{r}^{(I)} =$

$(r_1^{(I)}, \ldots, r_n^{(I)})$. Two observations are clustered together if they are assigned the same group-specific parameters $\boldsymbol{\theta}_k = (\mu_k, \sigma_k^2)$, where for brevity we now omit the superscript (I). Recall the indicators v_i in (4.4). Without loss of generality assume that $v_i = 1$ (symptoms due to other causes) for $i = 1, \ldots, n_1$, and $v_i = 0$ (symptoms due to disease), $i = n_1 + 1, \ldots, n$. Similar to $p(\mathbf{r}^{(I)})$ we get a random partition $p(\mathbf{r}_\star^{(S)})$ induced by sampling from $f_S^\star(\cdot)$ for patients $i = 1, \ldots, n_1$. For $i = n_1+1, \ldots, n$ we have $S_i = I_i + L_i$ with the infection times I_i subject to the already described partition $\mathbf{r}^{(I)}$, and L_i i.i.d. under the assumed parametric model for the lag times $L_i = S_i - I_i$. In words, under the proposed model, the clustering structures $\mathbf{r}^{(S)}$ and $\mathbf{r}_\star^{(S)}$ for symptoms due to infection and for symptoms due to other causes, respectively, are modeled separately and are independent. In fact, symptoms due to infection inherit the clustering structure $\mathbf{r}^{(I)}$, which is induced by the marginal distribution for the infection times.

In order to cluster grouped data, other approaches have been proposed (Teh et al., 2005; Rodriguez et al., 2008; Camerlenghi et al., 2019; Argiento et al., 2020). These strategies allow for the possibility of sharing atoms of the random probability measures across groups, thus borrowing information and yielding more precise inference. However, the random partition is not the main inference target here and we shall therefore not further explore such alternatives.

4.3 Regression on Covariates

We now add covariate effects in the proposed nonparametric model. In the context of model (4.6) this takes the form of replacing $H^{(I)}$ and $H^{(S)}$ by families of random probability measures (r.p.m). That is, we introduce a family $\{H_{\mathbf{x}}^{(I)}, \mathbf{x} \in \mathcal{X}\}$, and similarly for $H^{(S)}$. Here $\mathbf{x}$ are patient specific covariates, and we replace $H^{(I)}$ and $H^{(S)}$ by $H_{\mathbf{x}_i}^{(I)}$ and $H_{\mathbf{x}_i}^{(S)}$ for patient i in equation (4.6). Dropping for the moment the superscript for easier exposition, let $\mathcal{H} = \{H_{\mathbf{x}} = \sum_k \pi_{xk}\delta_{\mu_{xk}}, \mathbf{x} \in \mathcal{X}\}$ denote a family of r.p.m.'s indexed by $\mathbf{x}$. The most widely used class of priors on families like $\mathcal{H}$ are dependent DP (DDP) models (MacEachern, 1999). The DDP construction implies marginally for each $H_{\mathbf{x}}$ a DP prior, and allows for the desired dependence across $\mathbf{x}$. The definition of the marginal DP implies that the μ_{xk}'s are independent across k and that the weights have stick-breaking priors, but it does not restrict the distribution across $\mathbf{x}$. This is what the DDP construction exploits to borrow information across covariate values. The DDP induces dependence across $\mathbf{x}$ through the atoms μ_{xk} and/or the weights π_{xk} of the marginal r.p.m.'s.

In the Partner Notification study the predictors are $\mathbf{x}_i = \{\text{gender}, \text{arm}, \text{age}\} \in \{0;1\}^2 \times \mathbb{R}^+$, i.e. two binary and one continuous covariate. We use a simple analysis of variance (ANOVA) structure to induce dependence of μ_{xk} across $\mathbf{x}$ and common weights π_h. DDP models with ANOVA-type dependence across categorical factors are introduced as the ANOVA-DDP in De Iorio et al. (2004) and then extended to continuous covariates in De Iorio et al. (2009). The dependence structure of the random probability measures $H_{\mathbf{x}}$ is modeled by constructing the atoms as $\mu_{xk} = \delta_k + \alpha_k x_1 + \beta_k x_2 + \gamma_k x_3$. The interpretation of the linear model coefficients $\mathbf{m}_k = (\delta_k, \alpha_k, \beta_k, \gamma_k)^\intercal$ is exactly as in an ANOVA model, inducing the desired dependence of $H_{\mathbf{x}}$ across $\mathbf{x}$ by sharing, for

example, the same β_k for any two covariate vectors $\mathbf{x}$ and $\mathbf{x}'$ that share the same x_2. Finally, using a design vector $\mathbf{d}_i = (1, x_{i1}, x_{i2}, x_{i3})^\intercal$ to select the desired ANOVA effects we can write $\mu_{x_i k} = \mathbf{d}_i^\intercal \mathbf{m}_k$ to get $H_{\mathbf{x}_i} = \sum_{k=1}^{+\infty} \pi_k \delta_{\mathbf{d}^\intercal \mathbf{m}_k}$. Defining $\boldsymbol{\theta}_k = (\mathbf{m}_k, \sigma_k^2)^\intercal$ to allow for a mixture also with respect to the kernel variances, we can define one common mixing measure

$$H(\cdot) = \sum_{k=1}^{+\infty} \pi_k \delta_{\boldsymbol{\theta}_k},$$

and push the linear model into the mixture kernel in (4.6). We now add back the superscripts (I) and (S) for the two models $H^{(I)}$ and $H^{(S)}$ in (4.6). Using $H^{(I)}$ the marginal distribution $f_I(I_i \mid \mathbf{x}_i)$ can thus be rewritten equivalently as a DP mixture of linear models, now using a single mixing measure H for all $\mathbf{x}$ (linear dependent DDP, Jara et al., 2010)

$$f_I(I_i \mid \mathbf{x}_i) = \int \mathrm{N}(I_i \mid \mathbf{d}_i^\intercal \mathbf{m}^{(I)}, \sigma^{(I)2}) dH^{(I)}(\boldsymbol{\theta}^{(I)}) \quad \text{with} \quad H^{(I)} \sim \mathrm{DP}(M^{(I)}, H_0^{(I)}). \tag{4.9}$$

Another instance of the same model is used for the marginal distribution of symptoms due to other causes $f_S^\star(S_i \mid \mathbf{x}_i)$. The full model is

$$\begin{aligned} C_i \mid S_i, \lambda &\overset{d}{=} \min\{S_i + \mathrm{Exp}(\lambda); \mathrm{Unif}(A, B)\} \\ (S_i, I_i) \mid \boldsymbol{\theta}^{(S)}, \boldsymbol{\theta}^{(I)}, w, \lambda_L &\sim f_{IS}(I, S). \end{aligned}$$

using (4.9) for f_I and similarly for $f_S^\star$. The complete model now defines a **bivariate BNP-CS survival regression**. Using the stick-breaking representation, the DP priors on $H^{(I)}$ and $H^{(S)}$ can be written as follows. Using superscripts $E \in \{I, S\}$ to refer to the construction of f_I and $f_S^\star$ respectively, we have

$$\begin{aligned} &\{\mathbf{m}_k^{(E)}, \sigma_k^{(E)2}\}_{k=1}^{+\infty} \overset{\text{iid}}{\sim} H_0^{(E)} = \mathrm{N}(\mathbf{m}_k^{(E)} \mid \mathbf{m}_0^{(E)}, \Sigma_0^{(E)}) \times \mathrm{IG}(\sigma_k^{(E)2} \mid a_\sigma^{(E)}, b_\sigma^{(E)}) \\ &\boldsymbol{\pi}^{(E)} \mid M^{(E)} \sim \mathrm{SB}(M^{(E)}); \quad M^{(E)} \sim \mathrm{Ga}(a_M, b_M), \end{aligned}$$

and $\lambda \sim \mathrm{Ga}(a_\lambda, b_\lambda)$, $\lambda_L \sim \mathrm{Ga}(a_L, b_L)$, $w \sim \mathrm{Beta}(a_w, b_w)$. This completes the model construction. Hyperparameter choices are described in the supplementary materials.

For later reference we note that the random probability measures $H^{(I)}(\boldsymbol{\theta}^{(I)})$ and $H^{(S)}(\boldsymbol{\theta}^{(S)})$ that serve as the mixing measure in (4.9) are multivariate distributions for $\boldsymbol{\theta}^{(I)} = (\mathbf{m}^{(I)}, \sigma^{(I)2})^\intercal = (\delta^{(I)}, \alpha^{(I)}, \beta^{(I)}, \gamma^{(I)}, \sigma^{(I)2})^\intercal$, and similarly for $\boldsymbol{\theta}^{(S)}$. Let

$$H_\beta^{(I)} = \sum_{k=1}^{+\infty} \pi_k^{(I)} \delta_{\beta_k^{(I)}} \tag{4.10}$$

denote the implied univariate marginal for the ANOVA effect $\beta^{(I)}$. Analogous notation can be used for $H_\beta^{(S)}$ and any of the other ANOVA effects. We will later use inference on $H_\beta^{(E)}$, $E \in \{I, S\}$, to summarize inference on the treatment effect. Note that $H_\beta^{(E)}$, $E \in \{I, S\}$ should only be interpreted as summaries of a comparison of posterior inference on the distributions f_I and $f_S^\star$ across $\mathbf{x}$, as there is no notion of the mixtures $H_\beta^{(E)}$, $E \in \{I, S\}$ being identifiable.

4.4 Posterior Inference

To implement posterior inference under a Dirichlet process mixture model, the two main strategies are marginal (Escobar and West, 1995; MacEachern and Müller, 1998; Neal, 2000) and conditional MCMC posterior simulation, including the truncated Dirichlet process of (Ishwaran and James, 2001) and the slice sampler implementation proposed in Kalli et al. (2011). In our implementation, we employ the truncation algorithm of Ishwaran and James (2001). In particular, we rewrite the mixture model as a hierarchy by explicitly introducing the latent cluster membership variables $\mathbf{v}$, $\mathbf{r}^{(I)}$ and $\mathbf{r}_{\star}^{(S)}$. Moreover, we impute the latent symptoms and infection times from their corresponding full conditionals. We use efficient sampling for truncated normal distributions, originally proposed in Geweke (1991). This allows us to use standard algorithms for inference under a DPM.

The total masses $M^{(I)}$ and $M^{(S)}$ for the two random probability measures are included in the MCMC scheme and assigned Gamma priors, as recommended in Escobar and West (1995). Moreover, we put priors on the hyperparameters for the base measures $H_0^{(I)}$ and $H_0^{(S)}$. Additional details of the algorithm are deferred to the supplementary materials.

We carried out extensive simulation studies to verify the use of the proposed model with relevant sample sizes (details in the supplementary materials). Figure 7 shows the summary of one of these simulations, which was designed to closely mimic the setup of the Partner Notification Study. We simulate one binary and one continuous covariate. The underlying distributions for the infection times and for the symptom times are two mixtures of linear models, each one with different parameters. For more details on how the data were generated, we refer to the supplementary materials. Simulations show that our model can recover the underlying bivariate density for the two events when inference is conditioned on censoring times and censoring indicators only, using sample sizes as in the application. Figure 7 highlights that the underlying true bivariate density is recovered well by our method. In the supplementary materials, we show that our proposed method outperforms parametric and nonparametric alternatives in terms of goodness-of-fit.

5 Partner Notification Study - Results

We apply the proposed model for inference in the Partner Notification study described in Section 2. The primary inference goal is to understand the effect of covariates, in particular treatment assignment, on the joint distribution of the two latent times of interest. Furthermore, we are interested in assessing what factors drive time to reinfection and how time to symptoms onset of these cases can improve such estimation.

Inference under the proposed model includes the full joint distribution of latent times to symptoms and infection times. Figure 8 shows the posterior estimated distribution $f_{IS}(I, S)$ and the two components $f_{IS}^{\star}(I, S)$, $f_{IS}'(I, S)$ corresponding to a 'baseline' covariate combination (male, control group, median age). There is significant probability mass in the lower triangle ($S < I$) that is not concentrated around the 45° line but is

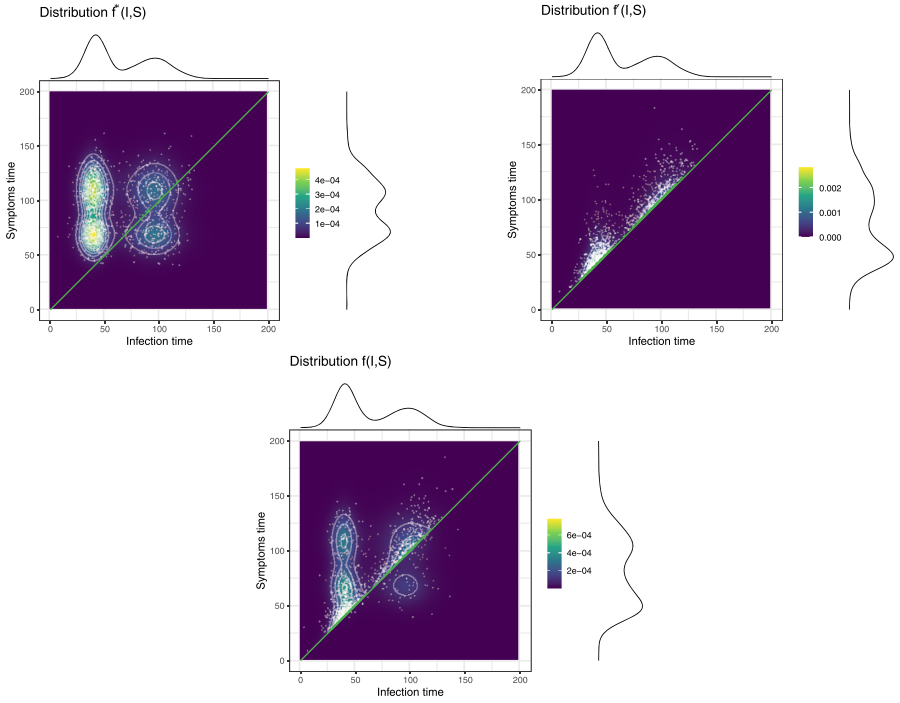


Figure 7: Results for simulated data: Posterior mean density estimate for $f^{\star}_{IS}$, f'_{IS} and f_{IS} corresponding to the baseline covariate levels. The green line is the 45° line $I = S$. The corresponding marginal distributions are shown on the top and right side of the density plot. The white points are a sample of the true latent times corresponding to the same covariate levels. Inference is only conditioned on Δ_I and Δ_S.

quite spread out. Instead, for the constrained component ($S > I$) the probability mass is concentrated very close to the 45° line. In other words, most of the inferred symptoms times due to infection concentrate in $I < S < I + 10$. This is simply reflecting that symptoms due to the infection follow shortly after the disease onset.

These results differ from the prior expectation implied by our choice of the hyperparameters. As one can see in the supplementary materials, the prior is much more diffuse than the posterior density estimate.

To show the estimated covariate effects, we could compare estimated survival functions for different combinations of the predictors. Alternatively, we can report posterior estimated marginal distributions of the ANOVA effects, for example $H_{\beta}^{(E)}$, $E \in \{I, S\}$ from (4.10). These are the univariate marginal distributions of the treatment effect in the DDP model, and concisely summarize the change of the bivariate survival distribution with respect to treatment versus control. The top center panel in Figure 9 shows the posterior estimated distributions $\mathbb{E}(H_{\beta}^{(I)} \mid \text{data})$, and similarly for other regression effects. Two significant effects can be detected. Importantly, treatment delays reinfec-

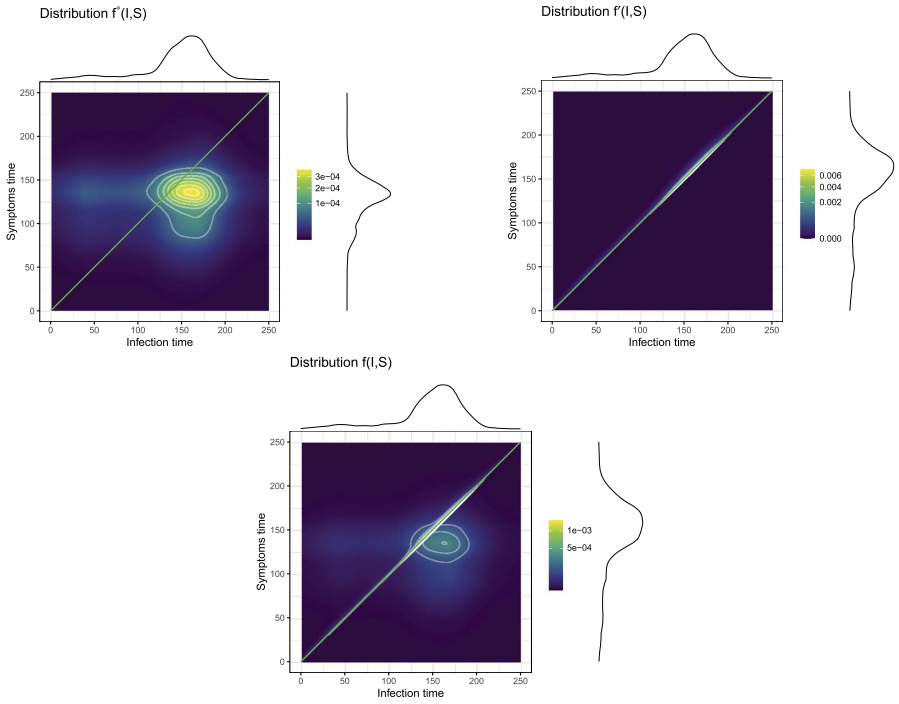


Figure 8: Results: Posterior mean density estimate for $f^{\star}_{IS}$, f'_{IS} and f_{IS} corresponding to the baseline covariate levels (male, control group, mean age). The green line corresponds to the 45° line, i.e. $I = S$. The corresponding marginal distributions are shown on the top and right side of the density plot.

tion times, confirming what was found in an earlier analysis in Sal y Rosas and Hughes (2011). Moreover, gender has an effect on the time to symptoms due to other causes, with women seeking hospital visits earlier, when the visit is prompted by symptoms. This might be simply due to the fact that women are more aware of their symptoms and are more inclined to hospital visits, suggesting that a health education campaign for men might improve their health outcome. Age has also been found to have a weak effect: younger individuals have shorter times to reinfection, possibly due to their more risky behavior.

Two parameters of the model, namely λ_L and λ, can give insights into how long it takes for participants to develop symptoms and to seek a visit to the hospital. In particular, the 95% credible interval for the exponential parameter λ is $[0.70, 1.42]$, suggesting that people seek a doctor visit, on average, one day after onset of symptoms. Moreover, the 95% credible interval for the exponential parameter λ_L is $[0.22, 0.80]$, which implies that patients develop symptoms due to infection, on average, 2.5 days after reinfection. Inference includes an estimate for the proportion of patients that experience symptoms due to the infection, in our notation $1 - w$. The posterior mean of such proportion is 0.18 (95% CI: [0.12, 0.24]). This is coherent with what we see empirically

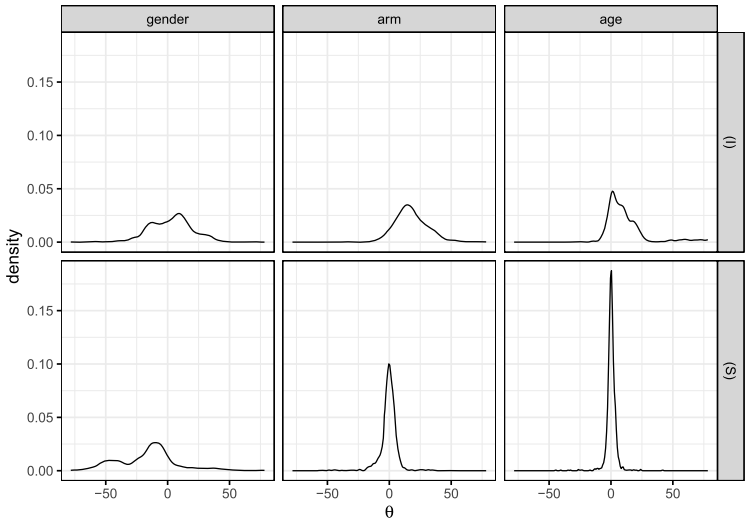


Figure 9: Results: Posterior means for the distributions H_α, H_β and H_γ associated to the regression coefficients α (left), β (middle) and γ (right) under f_I (top panels) and $f_S^\star$ (bottom panels). The figures show kernel density estimates based on MCMC posterior simulations.

in the data. There are more observed symptoms than observed infections, which implies that most of the symptoms should be attributed to other causes. This finding has important practical implications as it can help better planning for the treatment of patients.

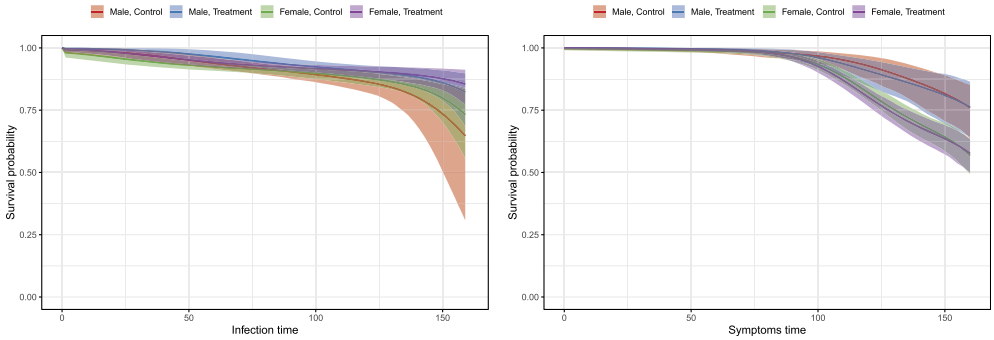


Figure 10: Results: Estimated survival curves under the proposed model for infection times (left panel) and times until symptoms (right panel) corresponding to the possible combinations of the binary covariates *gender* and *treatment* fixing the predictor *age* to the average age in the sample.

By way of comparison we carry out alternative inference, first using a model with flexible marginal distributions that assumes independence between the two events, and then using a model with parametric marginal distributions that instead implies dependence between the events. For the first strategy, we implement inference under two independent linear dependent Dirichlet process (LDDP) mixture of survival models for the marginal distributions of infection and symptoms times. This method is described in De Iorio et al. (2009) and implemented in the `DPpackage` (Jara et al., 2011). For a fair comparison, we used the same prior specifications for the shared parameters under the two models. The results are shown in Figure 11. Note the inappropriate posterior shrinkage of probability mass toward the extremes.

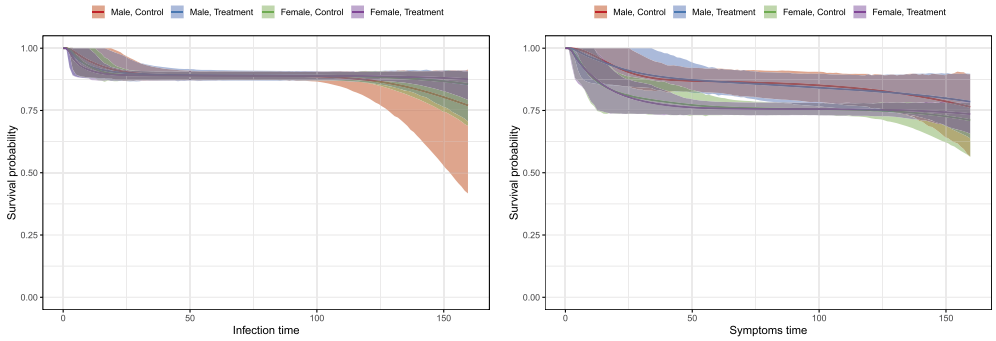


Figure 11: Results: LDDP estimated survival curves for infection times (left panel) and times until symptoms (right panel) corresponding to the possible combinations of the binary covariates *gender* and *treatment* fixing the predictor *age* to the average age in the sample.

The second comparison uses a bivariate Gumbel model (Gumbel, 1960). Two variables (I, S) have a Gumbel bivariate exponential distribution if their probability density function is

$$f(I,S) = \lambda_I e^{-\lambda_I I} \lambda_S e^{-\lambda_S S}[1 + \alpha\{1 - 2e^{-\lambda_I I}\}\{1 - 2e^{-\lambda_S S}\}],$$

where $-1 \leq \alpha \leq 1$ is a measure of dependence between the two variables. To include covariates, we generalize this model to a bivariate Gumbel regression by using $\log(\lambda_I) = \lambda_{I0} + X^\intercal\boldsymbol{\beta}, \log(\lambda_S) = \lambda_{S0} + X^\intercal\boldsymbol{\gamma}$. Under this model, both I and S have marginal exponential distributions with parameters λ_I and λ_S, respectively. Vague priors on all parameters were specified.

Some consistent results can be found across the models. For example, under the estimated models women have shorter time until symptoms as measured by the distribution for the corresponding regression coefficient in Figure 9 (or by the survival curves in Figure 10, right panel) and by the survival curves in Figure 11 (right) and Figure 12 (right). Unlike inference under the competing models, inference under the proposed bivariate model also shows a weak effect of the treatment on the infection time. Patients in the

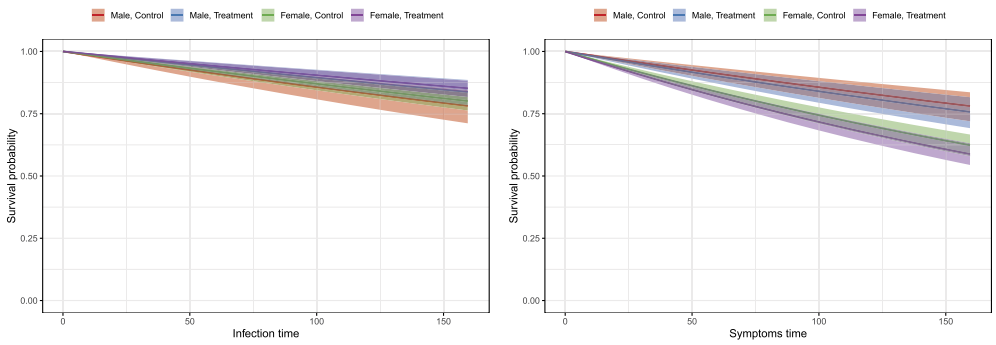


Figure 12: Results: Estimated survival curves under the bivariate Gumbel model for infection times (left panel) and times until symptoms (right panel) corresponding to the possible combinations of the binary covariates *gender* and *treatment* fixing the predictor *age* to the average age in the sample.

intervention group have a delayed re-infection time. The proposed model yields more interpretable results compared to the two independent LDDP models. In fact, under the LDDP models the probability mass accumulates toward the bounds of the observed censoring times, yielding a "flat" survival curve in the middle region (see Figure 11 compared to Figure 10), exactly where we expect events to happen. In fact, most right censored observations are imputed to the right of the rightmost censoring time, whereas most left censored observations are imputed to the left of the leftmost censoring time. This shows that the prior shrinkage alone does not suffice for regularization, and it is consistent with the observations of Section 3.1. On the other hand, the Exponential marginal distributions implied by the bivariate Gumbel model represent a very strict parametric assumption that seems not to fit well the data. The specification of such a simple parametric model also yields underestimation of uncertainty. The bivariate Gumbel model also implies a positive correlation between the events, as measured by the 95% credible interval for the parameter α is $[0.72, 0.99]$.

Additional simulations that illustrate the flexibility of our model compared to its competitors are shown in the supplementary materials.

6 Discussion

We proposed a novel Bayesian nonparametric bivariate survival regression model that is especially suited for current status data (BNP-CS regression). This research was motivated by the failure of available methods for such data formats. For example, we showed that under dependent censoring widely used nonparametric mixture priors lead to biologically uninterpretable results. Our model was built by incorporating simple structural dependence assumptions in a linear dependent Dirichlet process mixture of survival models. While the specific structural assumptions have natural interpretations in the motivating application, a similar inference framework remains valid in more gen-

erality. Any bivariate event times with a weak notion of ordering could be modeled very similarly, by introducing heterogeneity with an order constraint under one subpopulation and independence otherwise.

Applied to a recurrent infection study, the method provides novel insights into how symptoms-related hospital visits are affected by covariates. Notably, we were able to replicate previous results showing a significant effect of the intervention in the randomized controlled trial under consideration. In particular, patients in the intervention group have an improved outlook as measured by delayed reinfections. We also detect an effect of age, with young people having earlier reinfections, which might be due to more risky behaviours. Furthermore, we show that gender has a significant effect on the time until symptoms, but not on infection times. Our study shows that men seek hospital visits later compared to women, suggesting that investing in an awareness campaign could be beneficial.

The ideas presented in this article can be extended to different dependence structures. The present data called for a positive correlation between infection times and infection-related symptom times. A similar model specification can be used for negative correlations. Once the marginal models are flexibly specified, one could for example use copula models to construct a joint distribution with the desired dependence structure. A similar approach, but with positive correlations, could be used for general positively correlated event times when the assumptions used in this application are not available.

Supplementary Material

Supplementary Materials for Bayesian Nonparametric Bivariate Survival Regression for Current Status Data (DOI: 10.1214/22-BA1346SUPP; .pdf). Supplementary materials present additional details. These include proofs of the theorems, the choice of prior hyper-parameters and the MCMC scheme, convergence diagnostics and simulation studies. In separate files, the supplementary materials additionally include the `R` programs implementing the model developed in this article

References

Argiento, R., Cremaschi, A., and Vannucci, M. (2020). "Hierarchical Normalized Completely Random Measures to Cluster Grouped Data." *Journal of the American Statistical Association*, 115: 318–333. MR4078466. doi: https://doi.org/10.1080/01621459.2019.1594833. 15

Basu, A. P. (1983). "Identifiability." In Kotz, S. and Johnson, N. L. (eds.), *Encyclopedia of Statistical Sciences*, volume 4. Wiley Interscience. 13

Buckley, J. and James, I. (1979). "Linear regression with censored data." *Biometrika*, 66: 429–436. 2

Cai, B., Lin, X., and Wang, L. (2011). "Bayesian proportional hazards model for current status data with monotone splines." *Computational Statistics & Data Analysis*, 55: 2644–2651. MR2802342. doi: https://doi.org/10.1016/j.csda.2011.03.013. 2

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019). "Distribution theory for hierarchical processes." *The Annals of Statistics*, 47: 67–92. MR3909927. doi: https://doi.org/10.1214/17-AOS1678. 15

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). "BART: Bayesian additive regression trees." *The Annals of Applied Statistics*, 4: 266–298. MR2758172. doi: https://doi.org/10.1214/09-AOAS285. 2

Christensen, R. and Johnson, W. (1988). "Modelling accelerated failure time with a Dirichlet process." *Biometrika*, 75: 693–704. MR0995112. doi: https://doi.org/10.1093/biomet/75.4.693. 2

Cox, D. R. (1972). "Regression models and life-tables." *Journal of the Royal Statistical Society: Series B*, 34: 187–202. MR0341758. 2

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2009). "Bayesian nonparametric nonproportional hazards survival modeling." *Biometrics*, 65: 762–771. MR2649849. doi: https://doi.org/10.1111/j.1541-0420.2008.01166.x. 15, 21

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). "An ANOVA model for dependent random measures." *Journal of the American Statistical Association*, 99: 205–215. MR2054299. doi: https://doi.org/10.1198/016214504000000205. 3, 15

Duan, L. L., Young, A. L., Nishimura, A., and Dunson, D. B. (2020). "Bayesian Constraint Relaxation." *Biometrika*, 107: 191–204. MR4064148. doi: https://doi.org/10.1093/biomet/asz069. 13

Dunson, D. B. and Dinse, G. E. (2002). "Bayesian models for multivariate current status data with informative censoring." *Biometrics*, 58: 79–88. MR1891046. doi: https://doi.org/10.1111/j.0006-341X.2002.00079.x. 1, 2

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90: 577–588. MR1340510. 17

Geweke, J. (1991). "Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities." In *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, 571–578. 17

Golden, M. R., Whittington, W. L., Handsfield, H. H., Hughes, J. P., Stamm, W. E., Hogben, M., Clark, A., Malinski, C., Helmers, J. R., Thomas, K. K., and Holmes, K. K. (2005). "Effect of expedited treatment of sex partners on recurrent or persistent gonorrhea or chlamydial infection." *New England Journal of Medicine*, 352: 676–685. 2, 3

Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Birkhäuser Basel. MR1180321. doi: https://doi.org/10.1007/978-3-0348-8621-5. 1, 4, 6

Grushka, E. (1972). "Characterization of exponentially modified Gaussian peaks in chromatography." *Analytical Chemistry*, 44: 1733–1738. 14

Gumbel, E. J. (1960). "Bivariate exponential distributions." *Journal of the American Statistical Association*, 55: 698–707. MR0116403. 21

Hanson, T. and Johnson, W. O. (2002). "Modeling regression error with a mixture of Polya trees." *Journal of the American Statistical Association*, 97: 1020–1033. MR1951256. doi: https://doi.org/10.1198/016214502388618843. 2

Hanson, T. and Johnson, W. O. (2004). "A Bayesian semiparametric AFT model for interval-censored data." *Journal of Computational and Graphical Statistics*, 13: 341–361. MR2063989. doi: https://doi.org/10.1198/1061860043489. 2

Hanson, T. E. (2006). "Modeling censored lifetime data using a mixture of gammas baseline." *Bayesian Analysis*, 1: 575–594. MR2221289. doi: https://doi.org/10.1214/06-BA119. 9

Hjort, N. L. (1990). "Nonparametric Bayes estimators based on beta processes in models for life history data." *The Annals of Statistics*, 18: 1259–1294. MR1062708. doi: https://doi.org/10.1214/aos/1176347749. 2

Huang, J. (1996). "Efficient estimation for the proportional hazards model with interval censoring." *The Annals of Statistics*, 24: 540–568. MR1394975. doi: https://doi.org/10.1214/aos/1032894452. 2

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96: 161–173. MR1952729. doi: https://doi.org/10.1198/016214501750332758. 3, 17

James, L. F., Lijoi, A., and Prünster, I. (2009). "Posterior analysis for normalized random measures with independent increments." *Scandinavian Journal of Statistics*, 36: 76–97. MR2508332. doi: https://doi.org/10.1111/j.1467-9469.2008.00609.x. 14

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011). "DPpackage: Bayesian Semi- and Nonparametric Modeling in R." *Journal of Statistical Software*, 40: 1–30. MR3309338. doi: https://doi.org/10.1007/978-3-319-18968-0. 21

Jara, A., Lesaffre, E., De Iorio, M., and Quintana, F. (2010). "Bayesian semiparametric inference for multivariate doubly-interval-censored data." *The Annals of Applied Statistics*, 4: 2126–2149. MR2829950. doi: https://doi.org/10.1214/10-AOAS368. 16

Jewell, N. P. and Shiboski, S. C. (1990). "Statistical analysis of HIV infectivity based on partner studies." *Biometrics*, 46: 1133–1150. 1

Kalbfleisch, J. D. (1978). "Non-parametric Bayesian analysis of survival time data." *Journal of the Royal Statistical Society: Series B*, 40: 214–221. MR0517442. 2

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). "Slice sampling mixture models." *Statistics and Computing*, 21: 93–105. MR2746606. doi: https://doi.org/10.1007/s11222-009-9150-y. 17

Kottas, A. and Gelfand, A. E. (2001). "Bayesian semiparametric median regression modeling." *Journal of the American Statistical Association*, 96: 1458–1468. MR1946590. doi: https://doi.org/10.1198/016214501753382363. 2

Kuo, L. and Mallick, B. (1997). "Bayesian semiparametric inference for the accelerated failure-time model." *Canadian Journal of Statistics*, 25: 457–472. 2

Lo, A. Y. (1984). "On a class of Bayesian nonparametric estimates: I. Density estimates." *The Annals of Statistics*, 12: 351–357. MR0733519. doi: https://doi.org/10.1214/aos/1176346412. 9

MacEachern, S. N. (1999). "Dependent nonparametric processes." In *ASA proceedings of the section on Bayesian statistical science*, volume 1. 3, 15

MacEachern, S. N. and Müller, P. (1998). "Estimating mixture of Dirichlet process models." *Journal of Computational and Graphical Statistics*, 7: 223–238. 17

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian nonparametric data analysis*. Springer. MR3309338. doi: https://doi.org/10.1007/978-3-319-18968-0. 9, 10

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9: 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 17

Paulon, G., Müller, P., and Sal y Rosas, V. G. (2022). "Supplementary Materials for Bayesian Nonparametric Bivariate Survival Regression for Current Status Data" *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1346SUPP. 3

Patra, S. and Dunson, D. B. (2018). "Constrained Bayesian Inference through Posterior Projections." *arXiv preprint arXiv:1812.05741*. MR4035485. 13

Poynor, V. and Kottas, A. (2019). "Nonparametric Bayesian inference for mean residual life functions in survival analysis." *Biostatistics*, 20: 240–255. MR3922131. doi: https://doi.org/10.1093/biostatistics/kxx075. 9

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The nested Dirichlet process." *Journal of the American Statistical Association*, 103: 1131–1154. MR2528831. doi: https://doi.org/10.1198/016214508000000553. 15

Sal y Rosas, V. G. and Hughes, J. P. (2011). "Nonparametric and semiparametric analysis of current status data subject to outcome misclassification." *Statistical Communications in Infectious Diseases*, 3. MR2861479. doi: https://doi.org/10.2202/1948-4690.1032. 4, 19

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4: 639–650. MR1309433. 9

Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). "Nonparametric survival analysis using Bayesian additive regression trees (BART)." *Statistics in Medicine*, 35: 2741–2753. MR3513715. doi: https://doi.org/10.1002/sim.6893. 2

Swartz, T. B., Haitovsky, Y., Vexler, A., and Yang, T. Y. (2004). "Bayesian identifiability and misclassification in multinomial data." *Canadian Journal of Statistics*, 32: 285–302. MR2101757. doi: https://doi.org/10.2307/3315930. 13

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2005). "Sharing clusters among related groups: Hierarchical Dirichlet processes." In *Advances in Neural Information Processing Systems*, 1385–1392. MR2279480. doi: https://doi.org/10.1198/016214506000000302. 15

Wang, N., Wang, L., and McMahan, C. S. (2015). "Regression analysis of bivariate current status data under the Gamma-frailty proportional hazards model using the EM algorithm." *Computational Statistics & Data Analysis*, 83: 140–150. MR3281802. doi: https://doi.org/10.1016/j.csda.2014.10.013. 2

Wang, W. and Ding, A. A. (2000). "On assessing the association for bivariate current status data." *Biometrika*, 87: 879–893. MR1813981. doi: https://doi.org/10.1093/biomet/87.4.879. 1, 2, 11

Xue, H., Lam, K., and Li, G. (2004). "Sieve maximum likelihood estimator for semiparametric regression models with current status data." *Journal of the American Statistical Association*, 99: 346–356. MR2062821. doi: https://doi.org/10.1198/016214504000000313. 2