

Marta Tallarita¹ / Maria De Iorio² / Alessandra Guglielmi³ / James Malone-Lee⁴

Bayesian Autoregressive Frailty Models for Inference in Recurrent Events

¹ Department of Statistical Science, University College London, London, UK, E-mail: m.tallarita@ucl.ac.uk.
<https://orcid.org/0000-0002-7895-8302>.

² Department of Statistical Science, University College London, London, UK, E-mail: m.deiorio@ucl.ac.uk

³ Dipartimento di Matematica, Politecnico di Milano, Milan, Italy, E-mail: alessandra.guglielmi@polimi.it

⁴ Division of Medicine, University College London, London, UK, E-mail: james.malone-lee@ucl.ac.uk

Abstract:

We propose autoregressive Bayesian semi-parametric models for gap times between recurrent events. The aim is two-fold: inference on the effect of possibly time-varying covariates on the gap times and clustering of individuals based on the time trajectory of the recurrent event. Time-dependency between gap times is taken into account through the specification of an autoregressive component for the frailty parameters influencing the response at different times. The order of the autoregression may be assumed unknown and is an object of inference. We consider two alternative approaches to perform model selection under this scenario. Covariates may be easily included in the regression framework and censoring and missing data are easily accounted for. As the proposed methodologies lie within the class of Dirichlet process mixtures, posterior inference can be performed through efficient MCMC algorithms. We illustrate the approach through simulations and medical applications involving recurrent hospitalizations of cancer patients and successive urinary tract infections.

Keywords: autoregressive models, Dirichlet process mixtures, model selection

DOI: 10.1515/ijb-2018-0088

Received: August 28, 2018; **Revised:** July 18, 2019; **Accepted:** July 19, 2019

1 Introduction

Recurrent event processes generate events repeatedly over time and recurrent event data arise in many applications, for example in medicine (e. g. [1, 2]), economy (e. g. [3]) and technology (for example, [4]). Typical examples include recurrent infections, asthma attacks, hospitalizations, product repairs and machine failures. In particular, in this work, we are interested in settings where recurrent event processes are available from a large number of individuals, but with a small number of occurrences for each subject. Here we consider a benchmark dataset, in which we are interested in modelling the time between re-hospitalisation after surgery in colon cancer patients, and an application where the main clinical outcome is recurrence of urinary tract infections.

Typically, the focus is on modeling the rate of occurrence, accounting for the variation within and between individuals. Moreover, in applications, it is often of interest to assess the relationship between event occurrence and potential explanatory factors. The two main statistical approaches to perform inference on recurrent event data are (see [5]): (i) modelling the intensity function of the event counts $\{N(t), t \geq 0\}$, where $N(t)$ is the number of events between the time origin and time t ; (ii) modelling the whole sequence of gap times between successive realizations of the recurrent events.

The first approach is most suitable when individuals frequently experience the event of interest and the occurrence does not alter the process itself. The canonical framework for the analysis of event counts is the Poisson process. For a full review of this class of methods, see [6]. Wang et al. [7] suggest a subject-specific nonstationary Poisson process via a latent variable to model the occurrence of recurrent events. An alternative method was introduced by Thall [8] who considers a parametric mixed Poisson model which deals with covariates. Further, [9] propose a semiparametric model based on a non-homogeneous Poisson process. Finally, Pepe and Cai [10], Lawless and Nadeau [11] and Lawless et al. [12] propose semiparametric procedures for making inferences about the mean and rate functions of the counting process without the Poisson-type assumption.

The second approach based on gap times is more appropriate when the events are relatively infrequent, when, after an event, individual renewal takes place in some way, or when the focus of the analysis is the prediction of the time to the next event. The two dominant families of models in this framework are accelerated

Marta Tallarita is the corresponding author.

© 2019 Walter de Gruyter GmbH, Berlin/Boston.

failure time (AFT) and proportional hazards (PH) regressions. For example, Prentice et al. [13] propose a PH model for recurrent events, while Chang and Wang [14] propose a slightly different model by incorporating two kinds of covariates: some structural covariates (fixed) and some episode-specific covariates. Frailty models are discussed in McGilchrist and Aisbett [15] and Duchateau and Janssen [16]. However, in the frailty models time dependence is not modelled, but in many situations the occurrence of a past event may change the risk of a later event and so it is essential to incorporate and quantify how the occurrence of a past event may change the risk of a later event. Lin et al. [17] overcome this problem and propose a Bayesian recurrent event model to account for both subject-specific heterogeneity and event dependence. An alternative approach to frailty models consists in specifying a multivariate distribution for the gap times using copulas (see, for example, [18, 19]). This strategy requires specifying the marginal distribution of each gap time and then using a copula to introduce dependence.

1.1 Bayesian nonparametric framework

This paper lies within the gap times approach and develops a Bayesian semiparametric model for gap times between events. We assume that the joint distribution of the finite sequence of gap times for each individual is the product of the conditional distributions of each gap time, given the previous ones. Moreover, we specify a regression model for each of these conditional distributions to link the realization of each gap time to possibly time-varying covariates and previous gap times. To account for inter-subject variability, we introduce individual specific frailty parameters which we model flexibly using a Dirichlet process mixture prior as random effects distribution. Dirichlet process mixture (DPM) models [20, 21] are arguably the most common nonparametric Bayesian prior and have proved successful in many applications due to their flexibility and ease of computation. DPM models are mixtures of a parametric distribution where the mixing measure is the Dirichlet process (DP) introduced by Ferguson [22]. It is well known that the DP is almost surely discrete, and that if G is a $DP(M, G_0)$ with total mass parameter M and baseline distribution G_0 , then G admits the following representation [23]:

$$G(\cdot) = \sum_{h \geq 1} w_h \delta_{\theta_h}(\cdot)$$

where δ_θ is a point-mass at θ , the weights follow a stick-breaking process, $w_h = V_h \prod_{j < h} (1 - V_j)$, with $V_h \stackrel{\text{iid}}{\sim} \text{Beta}(1, M)$, and the atoms $\{\theta_h\}_{h \geq 1}$ are such that $\theta_h \stackrel{\text{iid}}{\sim} G_0$. As the discreteness of G is inappropriate in many applications, it is common to convolve a parametric kernel $k(y | \theta)$ with respect to G , obtaining a DPM:

$$H(y) = \int k(y | \theta) G(d\theta).$$

Due to the discreteness of G , the DPM prior induces clustering of the subjects in the sample based on the trajectory of the recurrent events over time, where the number K of clusters is unknown and learned from the data. Earliest examples of the use of a Bayesian nonparametric distribution for the random effects can be found in Müller and Rosner [24] and Kleinman and Ibrahim [25]. More recently, Pennell and Dunson [26] employ a Dirichlet process prior to build semiparametric dynamic frailty models for multiple event time data, allowing also the frailty parameter to change over time. Moreover, Paulon et al. [27] develop a joint model for gap times between two consecutive hospitalizations and survival time using the DP as random effect distribution, but do not explicitly account for the temporal dependence between gap times observed on the same individual.

In this work, we investigate different strategies to link gap times at time t with previous gap times. We start by assuming a standard Markov model where also the order of dependence p is unknown and is object of inference. We explore two different strategies to specify a prior distribution on p : one involves eliciting a prior directly on the space of all possible Markov models for $p \in \{0, 1, \dots, P\}$, while the other approach employs spike and slab base measures and it is in the spirit of stochastic search variable selection [28]. In conclusion, the main contribution of this work is to provide, within a Bayesian nonparametric framework, a model for gap times which accommodates for temporal dependence, performs variable selection on the order of dependence with two different approaches, allows for clustering of individuals and inter-subject variability, is able to handle different numbers of gap times per individual and missingness.

In Section 2 we introduce the model, while in Section 3 we explain how to perform inference on the order of dependence in the Markov structure. In Section 4 we investigate the performance of the proposed approach in simulations and compare the different strategies to model time dependency and to select the order p . Sections 5 and 6 present two medical applications involving recurrent hospitalizations and urinary tract infections, respectively. We conclude the paper in Section 7.

2 Autoregressive frailty models via Dirichlet process mixtures

We consider data on N individuals. We assume that $0 = T_{i0}$ corresponds to the start of the event process and that individual i is observed over the time interval $[0, \tau_i]$. If n_i events are observed at times $0 < T_{i1} < \dots < T_{in_i} < \tau_i$, let $W_{ij} = T_{ij} - T_{ij-1}$ for $j = 1, \dots, n_i$ denote the gap times between events of subject i and $W_{in_i+1} = \tau_i - T_{in_i}$. Note that if τ_i corresponds to an actual event, then W_{in_i+1} is observed. Otherwise τ_i becomes a censoring time and W_{in_i+1} is a censored observation. Therefore W_{ij} , $j = 1, \dots, n_i$, are the observed gap times for individual i with a possibly censoring time W_{in_i+1} . Let J be the maximum number of observed repeated events, i. e. $J = \max_{i=1, \dots, N}(n_i)$ and let $Y_{ij} = \log(W_{ij})$. We describe the joint distribution of $\mathbf{Y}_i := (Y_{i1}, \dots, Y_{in_i}, Y_{in_i+1})$ through the specification of the conditional laws $\mathcal{L}(Y_{ij} | \mathbf{x}_{ij}, Y_{i1}, \dots, Y_{ij-1})$, where \mathbf{x}_{ij} denotes the vector of possibly time-varying covariates for the i th individual. In particular, we assume that an observation at time j , for each subject i , $i = 1, \dots, N$ and $j = 1, \dots, n_i$, is distributed as follows

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_j + \alpha_{ij} + \sigma \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1) \quad (1)$$

where $\boldsymbol{\beta}_j$ is the vector of regression coefficients at time j common to all individuals. Covariates and regression parameters here have dimension q . Moreover, the random parameters α_{ij} 's represents time-specific frailties, for which we assume a nonparametric prior with a time-dependent modeling structure as described in Sections 2.1 and 2.2. Given the parameters in the model, the individual recurrent processes are assumed conditionally independent. Note that the number of recurrent events does not need to be the same for all individuals and that missing data are at least in principle easily accommodated in a Bayesian framework by assuming missingness at random.

The likelihood for all the sample is then given by:

$$L = \prod_{i=1}^N \left\{ \left(\prod_{j=1}^{n_i} f(y_{ij} | \mathbf{z}_{ij}, \boldsymbol{\beta}_j, \alpha_{ij}, \sigma) \right) S^{\nu_i}(y_{in_i+1} | \mathbf{z}_{in_i+1}, \boldsymbol{\beta}_{n_i+1}, \alpha_{in_i+1}, \sigma) \right. \\ \left. \times f^{1-\nu_i}(y_{in_i+1} | \mathbf{z}_{in_i+1}, \boldsymbol{\beta}_{n_i+1}, \alpha_{in_i+1}, \sigma) \right\}$$

where $\mathbf{z}_{ij} = (\mathbf{x}_{ij}, y_{i1}, \dots, y_{ij-1})$, f is the density of the gap times (in this case a Gaussian density), S denotes the survival function of the last (censored) gap times and ν_i is the censoring indicator equal 1 if the last observation is censored and 0 otherwise. We are assuming that the censoring mechanism is independent of the recurrent event process.

The vector \mathbf{x}_{ij} can contain both time-varying and fixed covariates and the effect of the covariates can be assumed to be constant over time if appropriate, i. e. $\boldsymbol{\beta}_j = \boldsymbol{\beta}$. The vector $\boldsymbol{\beta}_j$ does not include the intercept term, because of identifiability issues with α_{ij} . Moreover, the model can be generalised to include a subject specific and/or time specific observational variance σ^2 and/or a different distribution for the gap times.

2.1 Nonparametric AR(1)-type models

Following a similar modelling strategy to the one described in Di Lucca et al. [29], a straightforward way to introduce dependence among random effects at different times is to allow the distribution of α_{ij} to depend on some summary of the observations up to time $j-1$:

$$\alpha_{ij} | m_{i0}, m_{i1}, \tau \stackrel{\text{ind}}{\sim} \mathcal{N}(m_{i0} + m_{i1} g(Y_{i1}, \dots, Y_{ij-1}), \tau^2), \quad j = 1, \dots, n_i + 1 \quad (2)$$

$$(m_{i0}, m_{i1}) | G \stackrel{\text{iid}}{\sim} G, \quad G \sim DP(M, G_0). \quad (3)$$

When $j = 1$, the distribution of the random effect α_{i1} simplifies as the autoregressive term in (2) disappears and it reduces to the Normal distribution with mean m_{i0} .

We assume conditional independence among subjects, given the parameters, and that (m_{i0}, m_{i1}) are independent under the base measure G_0 , which becomes the product of a Normal density for m_{i0} and a Normal

density for the autoregressive coefficient m_{i1} . The prior specification is completed as follows:

$$\begin{aligned}\beta_j &\stackrel{\text{iid}}{\sim} N_q(0, \beta_0^2 I_q) \\ \sigma^2 &\sim \text{Inv-Gamma}(a_\sigma, b_\sigma) \\ \tau^2 &\sim \text{Inv-Gamma}(a_\tau, b_\tau) \\ M &\sim U(0, M_0) \\ G_0 &= N(0, \sigma_g^2) \times N(\mu_z, \sigma_z^2),\end{aligned}\tag{4}$$

assuming a priori independence among the different parameters.

When implementing the model in JAGS, we opt for marginal uniform priors on bounded interval for σ^2 and τ^2 as suggested by Gelman [30] as inverse-gamma distributions give numerical instability problems when hyperparameters are small. Moreover, Gelman [30] recommends uniform distributions (on bounded intervals) for scale parameters in hierarchical models as being weakly-informative, opposite to inverse-gamma priors for variance parameters, which are very sensitive to the choice of the hyperparameters themselves.

The choice of g in (2) is obviously crucial and depends on the context and the goals of the inference problem. Common choices in the literature are:

- $g(Y_{i1}, \dots, Y_{ij-1}) = Y_{ij-1}$, i. e. the random effect at time j depends only on the observation at time $j-1$, conditionally on the remaining parameters;
- $g(Y_{i1}, \dots, Y_{ij-1}) = (Y_{i1} + \dots + Y_{ij-1}) / (j-1)$, i. e. the conditional expected value of each α_{ij} depends on the sample mean of the observations up to time $j-1$;
- $g(Y_{i1}, \dots, Y_{ij-1}) = (Y_{i1} \times \dots \times Y_{ij-1})^{1/(j-1)}$, this is equivalent to the geometric mean of the observations up to time $j-1$.

Note that, when $g(Y_{i1}, \dots, Y_{ij-1}) = Y_{ij-1}$, then (2)–(3) imply that the random effects distribution at time j is a DPM of AR(1) processes, with dependence only on the gap time at time $j-1$.

2.2 Nonparametric AR(p) models

The model in Section 2.1 can be extended to include higher order dependence, by modifying (2) – (3) as follows:

$$\alpha_{ij} \mid m_{i0}, m_{i1}, \dots, m_{ip}, \tau \stackrel{\text{iid}}{\sim} \mathcal{N}(m_{i0} + \sum_{l=1}^p m_{il} Y_{ij-l}, \tau^2), \quad j = p+1, \dots, n_i+1\tag{5}$$

$$(m_{i0}, m_{i1}, \dots, m_{ip}) \mid G \stackrel{\text{iid}}{\sim} G, \quad G \sim DP(M, G_0)\tag{6}$$

$$G_0 = \mathcal{N}(0, \sigma_g^2) \times \underbrace{\mathcal{N}(\mu_z, \sigma_z^2) \times \dots \times \mathcal{N}(\mu_z, \sigma_z^2)}_{p \text{ times}}.\tag{7}$$

The distribution of α_{ij} for $j \leq p$, depends only on the available past observations as in any AR(p) model. Prior specification for the remaining parameters is the same as described in Section 2.1.

3 Estimating the order of dependence

In (5) we assume that the order of dependence on past observations is a fixed integer p . However, this parameter is often unknown in applications, and it needs to be estimated. A wealth of research focuses on Bayesian model selection (see, for example, [31, 32]). Here we concentrate on two approaches. The first one modifies the base measure of the DP by including a spike and slab distribution on the autoregressive coefficient, leading to the Spiked Dirichlet process prior introduced by Kim et al. [33]. The second one involves the direct specification of a prior on p , and then, conditional on p , we specify the prior distribution for the remaining parameters; in this case the dimension of the parameter vector $(m_{i0}, m_{i1}, \dots, m_{ip})$ changes according to p and consequently the dimension of the space where the Dirichlet process measure is defined.

3.1 Spike and slab base measure

Kim et al. [33] introduce the Spiked Dirichlet process prior in the context of regression. A key feature of their method is to employ the spike and slab distribution, i. e. a mixture of a point mass at 0 and a continuous distribution as centering distribution of the DP. This implies that, in a regression context, some coefficients have a positive probability of being equal to 0 and therefore they are not influential on the response of interest. Their technique is easily accommodated in our context by simply modifying G_0 in (7) as

$$\begin{aligned}
 G_0 &= \mathcal{N}(0, \sigma_g^2) \times \underbrace{\pi_1(\mu_z, \sigma_z^2) \times \cdots \times \pi_p(\mu_z, \sigma_z^2)}_{p \text{ times}} \\
 \pi_l(\mu_z, \sigma_z^2) &= (1 - \eta_l)\delta_0 + \eta_l \mathcal{N}(0, \sigma_z^2), \quad l = 1, \dots, p \\
 \eta_l &\stackrel{\text{iid}}{\sim} \text{Bernoulli}(c_l) \\
 c_l &\stackrel{\text{iid}}{\sim} \mathcal{U}(0, 1)
 \end{aligned} \tag{8}$$

where the introduction of hyperpriors on the weights of the mixture induces sparsity.

3.2 Prior on the Order of Dependence

Following Quintana and Müller [34], we specify a prior directly on the order p of the autoregressive process and then, conditioning on p , we set a Dirichlet Process prior of appropriate dimension for the parameters of the AR(p), i. e. the vector $(m_{i0}, m_{i1}, \dots, m_{ip})$. Let P be the maximum possible order. Then we can specify the following hierarchy:

$$\begin{aligned}
 \alpha_{ij} \mid p, m_{i0}, m_{i1}, \dots, m_{ip}, \tau &\stackrel{\text{iid}}{\sim} \mathcal{N}(m_{i0} + \sum_{l=1}^p m_{il} Y_{ij-l}, \tau^2), \quad j = p + 1, \dots, n_i + 1 \\
 (m_{i0}, m_{i1}, \dots, m_{ip}) \mid p, G_p &\stackrel{\text{iid}}{\sim} G_p \\
 G_p &\sim DP(M, G_{0p}) \\
 G_{0p} &= \mathcal{N}(0, \sigma_g^2) \times \underbrace{\mathcal{N}(\mu_z, \sigma_z^2) \times \cdots \times \mathcal{N}(\mu_z, \sigma_z^2)}_{p \text{ times}} \\
 p &\sim \text{Discrete Uniform on } \{0, 1, \dots, P\}.
 \end{aligned} \tag{9}$$

When $p = 0$, the process simplifies as the autoregressive term in (9) disappears and the base measure of the DP reduces to the Normal distribution for the intercept term.

An MCMC algorithm to perform posterior inference on the order of dependence is described in [34], who sample from the posterior distribution of p given the data. For the JAGS implementation we have opted for a pseudo-prior approach, see [35].

4 Simulated data

In order to check the performance of the class of models proposed in the previous sections, different simulation scenarios have been created. The simulated datasets discussed in this section, as well as the scenario in Section 3 of Supplementary Material, are generated from a mixture of Gaussian distributions. In Section 4 of Supplementary Material, results from a simulation scenario where data are generated from a mixture of Exponential distributions are presented. For all scenarios we generate data without censoring and datasets in which the subjects are right-censored with respect to their last gap time. Posterior inference for these examples, as well as for the real data applications in Section 5 and 6, can be performed through a standard Gibbs sampler algorithm, which we implement in JAGS [36], using a truncation-based algorithm for stick-breaking priors [37]. For all simulations, we run the algorithm for 251,000 iterations, discarding the first 1,000 iterations as burn-in and thinning every 50 iterations to reduce the autocorrelation of the Markov chain. The final sample size is then 5,000. The computational time for both simulation scenarios was close to three hours using a MacBook Pro with 2,4 GHz Intel Core i5 processor type. Unless otherwise stated, we check through standard diagnostics criteria such as those available in the R package CODA [38] that convergence of the chain is satisfactory for most of the parameters.

4.1 Simulation scenario 1

We consider a simulated dataset of $N = 300$ subjects, with n_i randomly generated from a Poisson distribution. In particular we select n_i in such a way that all subjects have at least three gap times: $n_i \sim \mathcal{P}(2.5) + 3$, for all i . One third of the observations are generated from

$$Y_{ij} \sim \mathcal{N}(0, (1.2)^2), \quad j = 1, \dots, n_i$$

while another third is generated from

$$Y_{i1} \sim \mathcal{N}(0, (1.5)^2), \quad Y_{i2}|Y_{i1} \sim \mathcal{N}(Y_{i1}, (1.5)^2) \\ Y_{ij}|Y_{ij-1}, Y_{ij-2} \sim \mathcal{N}(Y_{ij-1} + 0.7 \times Y_{ij-2}, (1.5)^2), \quad j = 3, \dots, n_i$$

and the last 100 observations are generated from

$$Y_{i1} \sim \mathcal{N}(0, (0.9)^2), \quad Y_{i2}|Y_{i1} \sim \mathcal{N}(Y_{i1}, (0.9)^2) \\ Y_{i3}|Y_{i2}, Y_{i1} \sim \mathcal{N}(Y_{i2} + 0.7 \times Y_{i1}, (0.9)^2) \\ Y_{ij}|Y_{ij-1}, Y_{ij-2}, Y_{ij-3} \sim \mathcal{N}(Y_{ij-1} + 0.7 \times Y_{ij-2} + 0.4 \times Y_{ij-3}, (0.9)^2), \quad j = 4, \dots, n_i.$$

In simulating the data, we assume independence across subjects. In this example, for ease of explanation, we do not include covariates. We introduce censoring according to the following procedure. Firstly, we set the percentage of the sample for which we have censoring. In this simulation we have used three different censoring rates: 10%, 25% and 50%. We consider at least two gap times for each individual and so for all patients i with censoring we generate a censoring time C_i from a Uniform distribution defined on the interval (T_{i2}, T_{in_i}) . The results obtained without censored data and with the different censoring rates are compared.

We fit the model (1), (5)–(6) with $p = 3$:

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_j + \alpha_{ij} + \sigma \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1) \\ \alpha_{ij} | m_{i0}, m_{i1}, \dots, m_{ip}, \tau \stackrel{\text{iid}}{\sim} \mathcal{N}(m_{i0} + \sum_{l=1}^3 m_{il} Y_{ij-l}, \tau^2), \quad j = 4, \dots, n_i + 1 \\ (m_{i0}, m_{i1}, \dots, m_{ip}) | G \stackrel{\text{iid}}{\sim} G, \quad G \sim DP(M, G_0)$$

where G_0 is given by the product of spike and slab distributions as defined in (8):

$$G_0 = \mathcal{N}(0, 10) \times \underbrace{\pi_1(0, 100) \times \dots \times \pi_3(0, 100)}_{3 \text{ times}} \\ \pi_l(0, 100) = (1 - \eta_l) \delta_0 + \eta_l \mathcal{N}(0, 100), \quad l = 1, 2, 3 \\ \eta_l \stackrel{\text{iid}}{\sim} \text{Bernoulli}(c_l) \\ c_l \stackrel{\text{iid}}{\sim} \mathcal{U}(0, 1).$$

In fitting the model we set:

$$\sigma \sim U(0, 10) \\ \tau \sim U(0, 10) \\ M_0 = 10.$$

Hyperparameters are chosen in order to specify vague marginal prior distributions. Figure 1 shows the predictive distributions of m_{i0} , m_{i1} , m_{i2} and m_{i3} obtained with the spike and slab base measures. Red, blue, green and grey lines correspond to estimates obtained using the dataset with 0%, 10%, 25% and 50% censoring rates, respectively. By visual inspection, it is clear that the marginal posterior predictive distributions of the m_{ij} 's agree with the true values used to generate the dataset. In fact, the predictive distribution of m_{i0} is concentrated around 0, while the predictive distributions of m_{i1} , m_{i2} and of m_{i3} are bimodal. Furthermore, the variance of the predictive distributions of each m_{ij} becomes larger with increasing censoring rate, because we have less information about the population parameters. The marginal posterior distributions of η_1 , η_2 , and η_3 in the prior (8), not reported here, concentrate most mass on 1.

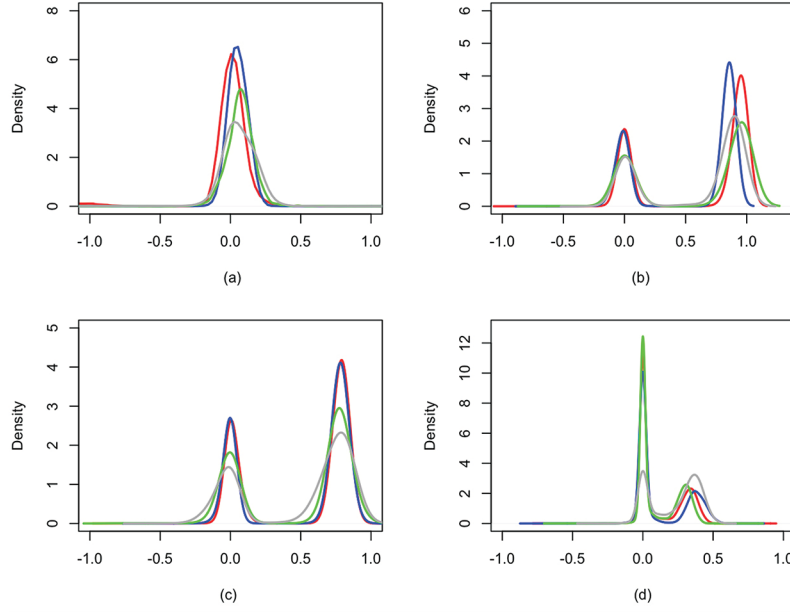


Figure 1: Simulation scenario 1: predictive marginal distributions of m_{i0} (a), m_{i1} (b), m_{i2} (c) and m_{i3} (d) under the spike and slab prior. Red, blue, green and grey lines correspond to estimates obtained using the dataset with 0%, 10%, 25% and 50% censoring rates, respectively.

Moreover, there is a clear difference between the posterior distribution of K , the number of distinct components in the mixture (5)–(6), using the data with and without censoring (results shown in Figure 2 of Supplementary Material). Using data without censoring, the configuration involving $K = 3$ clusters has the highest posterior probability: posterior inference on K is in agreement with the 3 components used to generate the data. Instead, using data with censoring, the number of clusters increases as the model needs to accommodate for the lack of information and the greater uncertainty in the distribution of the gap times. We also fit model (1), (9) to this dataset and similar results are obtained using the prior on the order of dependence. For further details, see Section 1 and Figure 1 of Supplementary Material.

4.2 Simulation Scenario 2

In this section we simulate a dataset of $N = 200$, with n_i randomly generated from a Poisson distribution. Similarly as in the previous example, we select $n_i \sim \mathcal{P}(2.5) + 3$ for all i , so that all subjects have at least three gap times. Half observations are generated independently from

$$Y_{i1} \sim N(0, 1.5^2), \quad Y_{i2}|Y_{i1} \sim N(0.9 \times Y_{i1}, 0.9^2) \\ Y_{ij}|Y_{ij-1}, Y_{ij-2} \sim N(0.9 \times Y_{ij-1} + 0.7 \times Y_{ij-2}, 0.9^2), \quad j = 3, \dots, n_i$$

while the other half is independently generated from

$$Y_{i1} \sim \mathcal{N}(0, 1.5^2), \quad Y_{i2}|Y_{i1} \sim \mathcal{N}(-0.9 \times Y_{i1}, 1.5^2) \\ Y_{ij}|Y_{ij-1}, Y_{ij-2} \sim \mathcal{N}(-0.9 \times Y_{ij-1} - 0.7 \times Y_{ij-2}, 1.5^2), \quad j = 3, \dots, n_i.$$

As in the previous example, covariates are not included and we introduce censoring using the same strategy. Also in this case we set the censoring rate equal to 10%, 25% and 50%. We consider at least two gap times for each individual and for all the patients with censoring we generate a censoring time C_i from a Uniform distribution defined on the interval (T_{i2}, T_{i10}) . The results obtained without censoring and with the different censoring rates are compared.

Also for this scenario, we fit both model (1), (5)–(6), (8) and model (1), (9). Here we report posterior inference of the latter model, with the prior on the order of dependence, where the maximum order of dependence is $P = 3$ and the prior hyperparameters (corresponding to a vague prior) are set as follows:

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_j + \alpha_{ij} + \sigma \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$$

$$\alpha_{ij} | p, m_{i0}, m_{i1}, \dots, m_{ip}, \tau \stackrel{\text{ind}}{\sim} N(m_{i0} + \sum_{l=1}^p m_{il} Y_{ij-l}, \tau^2), \quad j = p+1, \dots, n_i+1$$

$$(m_{i0}, m_{i1}, \dots, m_{ip}) | p, \tilde{G}_p \stackrel{\text{iid}}{\sim} \tilde{G}_p$$

$$\tilde{G}_p \sim DP(M, G_{0p})$$

$$G_{0p} = N(0, 10) \times N(0, 100) \times N(0, 100) \times N(0, 100)$$

$$p \sim \text{Discrete Uniform on } \{0, 1, 2, 3\}$$

$$\sigma \sim U(0, 10)$$

$$\tau \sim U(0, 10)$$

$$M \sim U(0, 5).$$

The mode of the marginal posterior distribution of p is 2 for all datasets without and with censoring, with corresponding posterior probability reported in the Table 1. In particular, the mass placed on $p=2$ decreases with increasing censoring rate.

Table 1: Simulation scenario 2: posterior distributions (masses) of p under the prior on the order of dependence for different rates of censoring.

Censoring rate	$p=0$	$p=1$	$p=2$	$p=3$
0 %	0.01	0.01	0.98	0.00
10 %	0.18	0.01	0.81	0.00
25 %	0.21	0.02	0.77	0.00
50 %	0.25	0.02	0.73	0.00

Conditional on $p=2$, Figure 2 reports the predictive distributions of m_{i0} , m_{i1} and m_{i2} . Red, blue, green and grey lines correspond to estimates obtained using the dataset with censoring rates equal to 0 %, 10 %, 25 % and 50 %, respectively.

More in detail, the predictive distribution of m_{i0} is concentrated around 0, while the predictive distributions of m_{i1} and of m_{i2} are bimodal with mode around $\{-0.9, 0.9\}$ and $\{-0.7, 0.7\}$, respectively. Also in this scenario, the variance of the predictive distributions of m_{ij} becomes larger with increasing censoring rate.

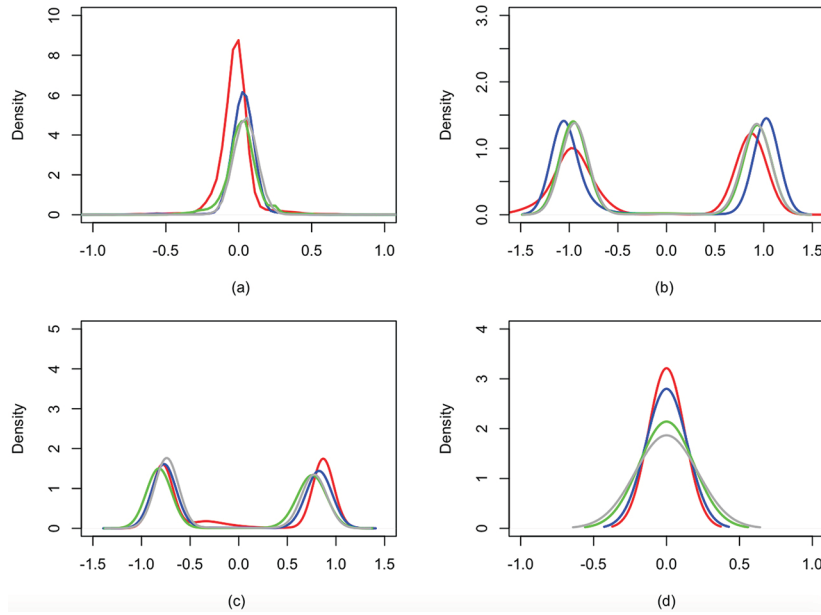


Figure 2: Simulation scenario 2: predictive marginal distributions of m_{i0} (a), m_{i1} (b), m_{i2} (c) and m_{i3} (d) under the prior on the order of dependence, conditioning on $p=2$. Red, blue, green and grey lines correspond to estimates obtained using the dataset with 0 %, 10 %, 25 % and 50 % censoring rates, respectively.

Finally, in the dataset without censoring time, conditioning on $p=2$, the posterior mode for the number K of clusters is 2 (results not shown), with associated posterior probability equal to 0.5, while for the datasets with censoring, the number of clusters increases because of the greater uncertainty on the distribution of gap times.

See Section 2 of Supplementary Material for the results obtained by fitting the model with the spike and slab base measure.

To evaluate the computational performance of the two approaches for model selection, we report in Table 2 summary statistics for the Effective Sample Size (ESS). ESS measures the amount by which autocorrelation within a MCMC chain increases uncertainty in estimates. The ESS of a parameter aims to estimate how many iid samples have the same estimation performance as the correlated samples found in the MCMC chain and it will be typically smaller than the final sample size. Hence large values of the ESS denote smaller autocorrelation in the chains; for further details, see [39]. Summary statistics are evaluated across the ESS of all estimated parameters. We report results for both simulations scenarios and four different censoring rates. The empirical distributions of the ESS in Table 2 seem comparable across different prior specifications and scenarios.

Table 2: ESS: summary statistics. We report 2.5 %, 50 %, 97.5 % empirical quantiles of the ESS evaluated for all estimated parameters. SS denotes the spike and slab base measure, while POD indicates the prior on the order of dependence.

Method	Censoring rate	Scenario 1	Scenario 2
SS	0 %	(37.45, 104.40, 245.09)	(38.96, 80.780, 207.70)
	10 %	(57.57, 144.70, 345.70)	(101.96, 120.78, 296.70)
	25 %	(43.50, 111.50, 241.80)	(167.96, 225.78, 370.70)
	50 %	(698.60, 1235.00, 1865.00)	(174.67, 253.82, 390.70)
POD	0 %	(41.52, 114.80, 265.09)	(143.30, 519.70, 573.00)
	10 %	(47.79, 141.01, 351.04)	(165.92, 290.60, 589.20)
	25 %	(53.50, 131.09, 298.09)	(821.90, 2125.00, 4521.00)
	50 %	(398.06, 1565.00, 1951.20)	(812.99, 2245.30, 4728.55)

4.3 Comparison with the time-dependent frailty model

In this subsection, we compare our model with the *time-dependent frailty* model of [40] on the same simulated data described in the previous section. This strategy not only accounts for dependence of recurrent failure times within observations for the same subject by introducing a subject-specific random frailty, but also for the order of events, which is a significant feature of dependence. In this setup, time independence is introduced in the model for the intensity function of the recurrent event process at time t :

$$\lambda_i(t | \beta, V_{i(t)}) = c_i(t) \lambda_0(t) \exp(V_{i(t)})$$

where the variable $c_i(t)$ is the censoring indicator equal to 1 if patient i is observed at time t and 0 otherwise. The function $\lambda_0(t)$ is an unspecified baseline intensity, while $V_{i(t)}$ is a subject-specific random frailty capturing the risk not accounted for by potential risk variables included in the analysis. The model can be extended to include time-dependent covariates. Following [40] we assume $\lambda_0(t) = \exp(\beta_0)$ and a first-order autocorrelated process prior for the frailties:

$$V_{i(t)} | V_{i(t-1)} = \phi V_{i(t-1)} + e_{i(t)}$$

where $V_{i(0)} \sim \mathcal{N}(0, \sigma_W^2)$ and $e_{i(t)}$ are i.i.d. random variables having a $\mathcal{N}(0, \sigma_W^2)$ distribution. The parameter ϕ is constrained to lie between -1 and 1 (a priori $\phi \sim \text{Translated Beta}(3, 3)$, i. e. with density proportional to $(\phi + 1)^2(1 - \phi)^2 \mathbb{1}_{(-1,1)}(y)$), and it measures the degree of serial correlation in the subject-specific frailty. The prior specification is completed as follows:

$$\begin{aligned} \beta_0 &\sim \mathcal{N}(0, 1000) \\ \sigma_W^2 &\sim \text{Inv-Gamma}(1, 1).y \end{aligned}$$

Also for this model posterior inference can be performed through a standard Gibbs sampler algorithm, which we implement in JAGS [36]. For each simulated scenarios, we run the algorithm for 251,000 iterations, discarding the first 1,000 iterations as burn-in and thinning every 50 iterations to reduce the autocorrelation of the Markov chain. The final sample size is 5,000 as before. Table 3 reports the posterior mean and standard deviation of the frailty correlation parameter ϕ , for the first simulation scenario, described in Section 4.1; Table 4 reports similar summaries of ϕ for the second simulation scenario (see Section 4.2).

Table 3: Simulation scenario 1: posterior mean and standard deviation of the frailty correlation parameter ϕ in [40].

Censoring rate	Mean	Standard Deviation
0	0.698	0.513
10 %	0.634	0.502
25 %	0.711	0.453
50 %	0.684	0.498

From Table 3 it is clear that Manda and Meyer's model is able to detect a significant positive correlation between recurrent events (as evident from the posterior means and standard deviations). In fact, in the first scenario, the data are generated so that observations at time j depend on observations at time $t - j$ in two clusters, while in the third one the observations are i.i.d.

Table 4: Simulation scenario 2: posterior mean and standard deviation of the frailty correlation parameter ϕ in [40].

Censoring rate	Mean	Standard Deviation
0	0.002	0.428
10 %	0.002	0.487
25 %	0.002	0.424
50 %	0.002	0.434

On the other hand, the posterior mean of ϕ in the second scenario is centred around 0. This shows the limitations of a parametric approach as observations are generated from two equal size groups, with the same order of dependence, but opposite effect. Therefore, it is difficult to capture the effect of the correlation between the recurrent event times. Moreover, from Figure 1 and Figure 2, it is clear that our model is more flexible. Both Figures show multimodal distributions for the autoregressive coefficients; this implies that our approach captures the clustering of observations and the cluster specific dependence structure. For example in Figure 1, the predictive distribution for m_{i3} has a mode in 0, i. e. in one cluster there is no dependence of third order, while the other mode is centred over 0.4, i. e. in the other cluster there is a positive dependence of the third order.

5 Hospitalization dataset

We fit model (1)–(3) to the *readmission* dataset in the R package *frailtypack* for all the possible choices of g described in Section 2.1. The dataset contains rehospitalization times (in days) after surgery in patients diagnosed with colorectal cancer. Data are available on $N = 403$ patients, for a total number of 861 recurrent events. In addition to gap times between successive rehospitalizations, the dataset contains information for each patient on the following covariates:

- *chemo*: variable indicating if the patient received chemotherapy.
- *sex*: gender of the patient.
- *dukes*: ordinal variable indicating the classification of the colorectal cancer. The baseline A–B denotes the invasion of the tumor through the bowel wall penetrating the muscle layer but not involving lymph nodes; the value C indicates the involvement of lymph nodes; the value D implies the presence of widespread metastases. Category D corresponds to the most severe type of cancer.
- *charlson*: Charlson comorbidity index. It measures ten-year mortality for a patient who may have a range of comorbidity conditions, and ranges within 3 classes, i. e. $\{0, 1 - 2, 3\}$. This is the only time-varying covariate.

The recurrent events in this study are readmission times (colorectal cancer patients may have several readmissions after first discharge). The origin of the time axis is the date of the surgical procedure for each patient and the recurrent events are subsequent rehospitalizations related to colorectal cancer. We exclude from the analysis patients with just one censored gap time, i. e. patients for which no further rehospitalization has been observed, and eight patients (approximately 2% of the subjects) with 7 or more events. We are then left with a dataset of $N = 197$ patients and a total number of 495 recurrent events. Table 5 reports the number of patients with exactly j gap times, for $j = 1, \dots, 6$. Moreover, 119 observations out of 197 are right-censored with respect to their last gap time. Since the proportion of censored data is considerably high, we need to take censoring into account.

Table 5: *Readmission* dataset: number of patients with exactly j gap times, $j = 1, \dots, J$.

j	1	2	3	4	5	6	TOT
n_j	30	96	36	18	9	8	197

Prior hyperparameters in (4) are set as follows:

$$\begin{aligned} \beta_0^2 &= 1,000 \\ \sigma &\sim \mathcal{U}(0, 10) \\ \tau &\sim \mathcal{U}(0, 10) \\ \sigma_g^2 &= 10, \quad \mu_z = 0, \quad \sigma_z^2 = 100 \\ M &= 1. \end{aligned}$$

5.1 Testing for the order of dependence

When testing the order of dependence, we first fit model (1), (5)–(6) and (8) with $p = 3$ (G_0 being the spike and slab distribution) and then model (1), but specifying a prior directly on model size as in (9) with $P = 3$.

Figure 3 reports the marginal posterior predictive marginal distributions of $m_{i,l}$, for $l = 0, 1, 2, 3$, obtained with the spike and slab base measure. Since the predictive distributions of m_{i0}, m_{i1}, m_{i2} are not concentrated around 0, unlike that of m_{i3} , we can conclude that the process best describing the *readmission* dataset has a dependency of the second order. This result is confirmed also using the approach described in Section 3.2. Indeed, the posterior distribution of p , displayed in Figure 9 of Supplementary Material, places most of its mass on 2.

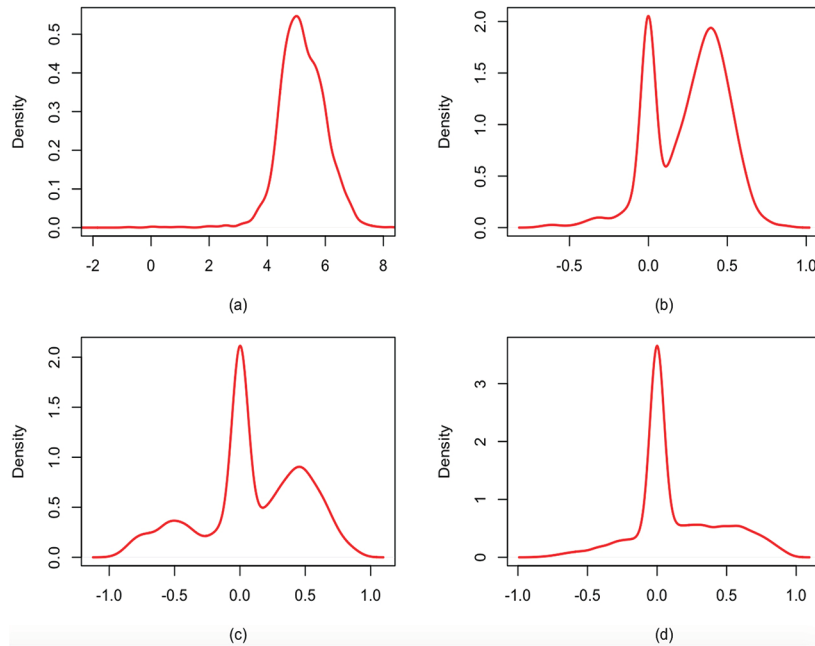


Figure 3: *Readmission* dataset: marginal posterior predictive distributions of m_{i0} (a), m_{i1} (b), m_{i2} (c) and m_{i3} (d) under the spike and slab prior.

5.2 Posterior analysis

We compare now the results of the nonparametric AR(2) model for the random effects α_{ij} 's as in (5)–(7), selected in the previous section, with models (2)–(3) built using different choices of g . In particular we consider two summary statistics: $g(Y_{i1}, \dots, Y_{ij-1}) = (Y_{i1} + \dots + Y_{ij-1}) / (j-1)$ and $g(Y_{i1}, \dots, Y_{ij-1}) = (Y_{i1} \times \dots \times Y_{ij-1})^{1/(j-1)}$. The goal is to understand if higher order temporal dependency can be approximated by an AR(1)-type process built on some appropriate function of past observations as described in (2). We analyze the posterior distribution of K , the number of components in the mixture (5)–(6) under different alternatives. In particular, the results show that the posterior modes of K are 2 or 3 with a probability of around 30% for the AR(1)-type models. On the

other hand, the AR(2) model, suggests the existence of 3, 4 or 5 groups (See Figure 10 of Supplementary Material for results).

In Figure 4 we present the marginal posterior predictive distributions of Y_j^{new} for a hypothetical new subject, for each time $j, j = 1, \dots, 6$, setting the values of the covariates to the empirical mode. From the figure, it is evident that the two AR(1)-type models produce very similar results. Obviously, for $j = 1$ and $j = 2$ the three distributions are almost identical, as the models are closer. For $j > 2$, it is clear that the posterior predictive distributions of Y_j^{new} have a larger variance and are more skewed under the AR(2) model. This experiment shows that it is not straightforward to approximate higher order dependency using summary statistics.

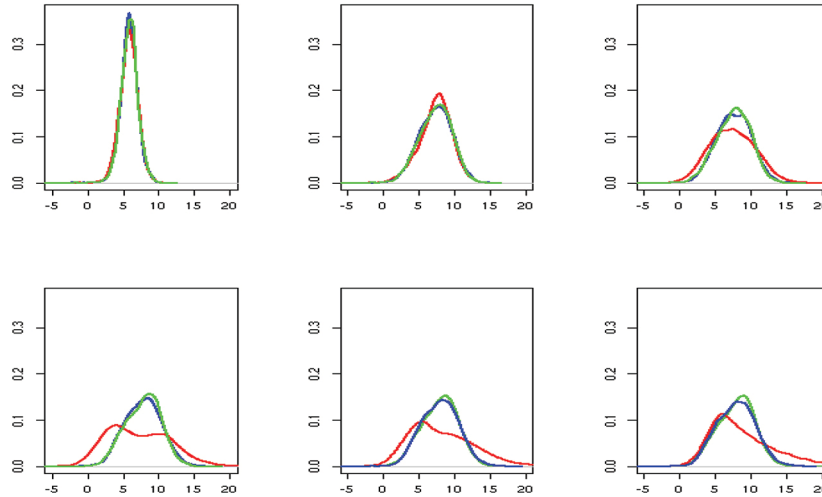


Figure 4: Readmission dataset: posterior predictive distributions of $Y_j^{new}, j = 1, \dots, 6$ for a hypothetical new subject, with covariates equal to the empirical mode of all covariates in the sample. The green and blue lines represent AR(1)-type models, with $g(Y_{i1}, \dots, Y_{ij-1}) = (Y_{i1} + \dots + Y_{ij-1}) / (j - 1)$ and $g(Y_{i1}, \dots, Y_{ij-1}) = (Y_{i1} \times \dots \times Y_{ij-1})^{1/(j-1)}$, respectively, and the red distribution indicates AR(2) model.

5.3 Posterior inference on the regression parameters

We now discuss the inference on the regression parameters in order to understand how covariates influence the recurrent event process. Although some covariates are fixed and do not vary over time, we still assume that their effect can be different in time and therefore we estimate a different vector of regression coefficients for all covariates in the model for each gap time $j, j = 1, \dots, 6$. Covariates *chemo* and *sex* are binary variables, while *dukes* and *charlson* are 3 levels categorical variables and we need to introduce 2 dummy variables for each of them in the model, with baseline set to A–B for *dukes* and to 0 for *charlson*. Therefore, the final covariate vector for individual i is given by $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}) = (\text{indicator for chemotherapy, indicator for female, indicator for dukes equal to C, indicator for dukes equal to D, indicator for charlson in 1–2, indicator for charlson equal to 3})$. The vector of regression parameters $\boldsymbol{\beta}_j = (\beta_{1j}, \beta_{2j}, \beta_{3j}, \beta_{4j}, \beta_{5j}, \beta_{6j})$ for each gap time $j, j = 1, \dots, J (= 6)$, is therefore 6-dimensional.

Figure 5 displays 95% credible intervals for the marginal posteriors of the regression parameter of each covariate for the first 5 gap times, i. e. of $\tilde{\boldsymbol{\beta}}_r := (\beta_{r1}, \beta_{r2}, \beta_{r3}, \beta_{r4}, \beta_{r5})$. In general there is no evident effect of chemotherapy on the outcome on any gap time. However, $\tilde{\boldsymbol{\beta}}_2$, which measures the effect of sex on the gap times, indicates that women have mainly larger gap times.

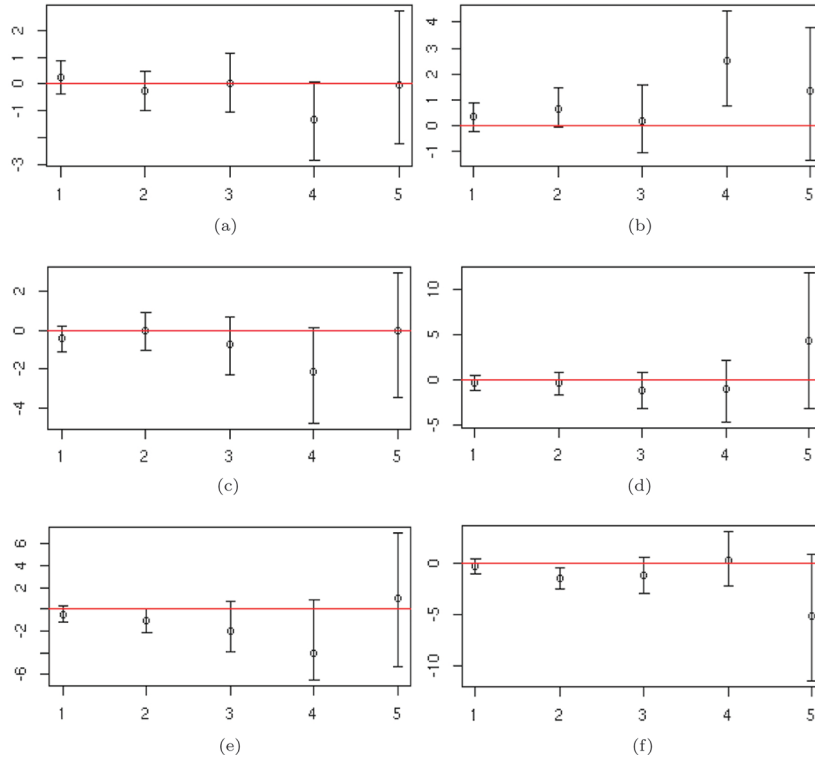


Figure 5: *Readmission* dataset: 95% posterior credible intervals for the regression parameters of each covariate across the first five gap times. (a) Regression coefficients of *chemo*. (b) Regression coefficients of *sex*. (c) Regression coefficients of *dukes = C*. (d) Regression coefficients of *dukes = D*. (e) Regression coefficients of *charlson = 1–2*. (f) Regression coefficients of *charlson = 3*.

The dummy variable relative to Dukes stage C of the tumour and the regression coefficients corresponding to Dukes stage D are never significant. Finally, posterior CI’s of the parameters corresponding to the Charlson index show that, in general, patients with index 0 will have larger gap times with respect to the one with index 3, in fact the medians of the posterior distribution are concentrated on negative values. Moreover, credible intervals are larger for the last gap times, as expected, since few individuals have a large number of events. The regression coefficients at time $j = 6$ are not shown as the credible intervals are not comparable with those of the previous times.

5.4 Comparison with existing methods

In this subsection, we compare our model with the frequentist *shared frailty* model of [41] as implemented in the R package *frailtypack* [42]. In this setup the dependence between recurrences is modelled through a frailty variable, such that events which occur several times within the same subject during the period of observation share an unobserved random effect, i. e. frailty. The hazard function for the j -th event ($j = 1, \dots, n_i + 1$) and i -th individual ($i = 1, \dots, N$), conditional on the frailty parameter ω_i , is

$$h_{ij}(t|\omega_i) = h_0(t)\omega_i e^{\beta^T x_{ij}}$$

$$\omega_i \stackrel{iid}{\sim} \Gamma\left(\frac{1}{\theta}, \frac{1}{\theta}\right)$$

where $h_0(t)$ is the baseline hazard function, β is the vector of the regression coefficients associated to the covariate vector x_{ij} . Table 6 reports the estimates of the regression coefficients and the associated p-value.

Table 6: *Readmission* dataset: estimates of the regression coefficients and associated p-values in the *shared frailty* model of [41].

	coef	p-value
chemo	−0.206	1.48e-01
sex	−0.537	1.06e-04

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

dukes C	0.297	6.42e-02
dukes D	1.056	5.82e-08
charlson 1–2	0.451	8.11e-02
charlson 3	0.410	2.74e-03

The estimate of the variance θ of the frailty term is 0.67 with standard error 0.14, so θ is significantly different from 0, meaning that there is heterogeneity between subjects. Differently from our approach, in this model the regression coefficients do not vary across time. The most significant variables are the indicator for *female*, indicator for *dukes* equal to D and indicator for *charlson* equal to 3. A negative coefficient implies a lower risk of rehospitalization, while a positive a higher risk. In our analysis we obtain that (i) the coefficient for indicator of *female* ($sex=1$) remains in general positive over time, implying longer times between rehospitalizations; (ii) the coefficient for *duke* equal to D is initially negative but becomes positive for the last gap times; (iii) the regression parameters corresponding to *charlson* index equal to 3 are mostly negative over time. These results show a good agreement between the two models. Finally, Section 6 of Supplementary Material reports a comparison of our model with the dynamic frailty models of [26], who also adopt a Bayesian nonparametric approach based on the DP.

6 Urinary tract infection dataset

We consider data on patients at risk of urinary tract infection (UTI). The best clinical marker of UTI available is pyuria, i. e. White Blood Cell count (WBC) μl^{-1} (1/microliter) ≥ 1 , detected by microscopy of a fresh unspun, unstained specimen of urine ([43, 44]). Let T_{i0} correspond to the first visit attendance at the *Lower Urinary Tract Service Clinic* (Whittington Hospital, London, UK) and let T_{ij} be the time of the j -th new infection for the patient i . Note that at time 0, all patients suffer of UTI. For each patient and at each visit the result of the microanalysis of a sample of urine has been recorded in terms of the WBC count. Presence of WBC in the urine (regardless of the quantity) indicates the presence of Urinary Tract Infection. We include in the analysis only female patients with at least two gap times, giving a total of $N = 306$ patients. The number of observations with exactly j gap times is displayed in Table 7.

Table 7: UTI dataset: number of patients with exactly j gap times, $j = 1, \dots, J$.

j	2	3	4	5	6	7	8	9	TOT
n_j	121	89	54	21	10	6	2	3	306

We note that 85 subjects out of 306 are right-censored with respect to their last gap time. Since the proportion of censored data is considerable, we have taken censoring into account and modified the likelihood appropriately. Figure 6 displays the recurrent events of two randomly selected patients, in which the last gap time of the patient in the left panel is observed, while that of the patient in the right panel is censored. Indeed, the first patient suffers of infection at her last visit, while the second patient has a WBC count equal to zero implying that a new infection will happen necessarily after her last visit.

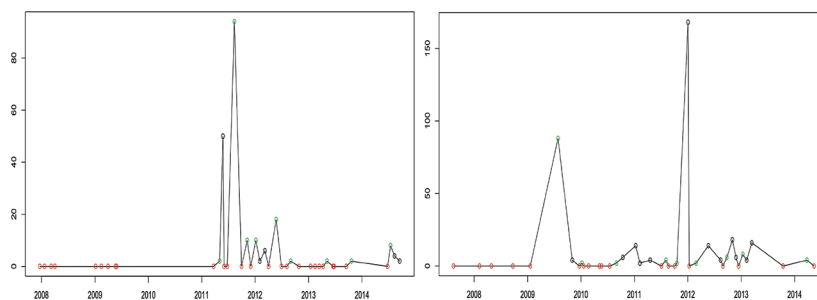


Figure 6: UTI dataset: recurrent events for two patients; the last gap time of the patient on the left is observed, while that of the patient on the right is censored. The vertical axis reports the value of WBC in the urine, red circles denote zero WBC at the visit while green circles denote WBC greater than 0.

We fit model (1), including for each patient a 5-dimensional vector of time-varying covariates: a continuous covariate representing the standardized age of the patient i during gap time j and four binary variables denoting the presence, during the j -th gap time, of urgency, pain, stress incontinence and voiding symptoms

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

(=1 if the symptom is present, 0 otherwise). Therefore, the final covariate vector for individual i is given by $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}) = (\text{age}, \text{indicator for urgency}, \text{indicator for incontinence}, \text{indicator for pain}, \text{indicator for voiding})$. Descriptive statistics of the covariates are given in Table 8.

Table 8: UTI dataset: descriptive statistics of the covariates.

Covariate	Mean	Standard Deviation
age	53.87	16.01
presence of urgency symptoms	0.56	0.50
presence of incontinence symptoms	0.21	0.41
presence of pain symptoms	0.47	0.50
presence of voiding symptoms	0.45	0.50

In the analysis we set the prior hyperparameters in (4) in order to specify vague prior distributions:

$$\begin{aligned} \beta_0^2 &= 1000 \\ \sigma &\sim \mathcal{U}(0, 10) \\ \tau &\sim \mathcal{U}(0, 10) \\ \sigma_g^2 &= 10, \quad \mu_z = 0, \quad \sigma_z^2 = 100. \\ M &= 1. \end{aligned}$$

6.1 Posterior inference

We perform inference on the order of dependence using both approaches introduced in Section 3. The marginal posterior predictive distributions of $m_{i,l}$, for $l = 0, 1, 2, 3$, obtained with the spike and slab base measure, are displayed in red in Figure 7. Panels (b), (c) and (d) show that the predictive distributions of $m_{i,1}, m_{i,2}, m_{i,3}$ are all concentrated around 0. In addition, specifying directly a prior on p , with $P = 3$, leads to a posterior distribution for the order of temporal dependence with mode in 0 (result shown in blue). These results indicate that there is no evidence of dependence between gap times, although the presence of a small cluster whose distribution is concentrated slight away from zero is detectable. This result is confirmed also by running the model for the three choices of function g described in Section 2.1, assuming an AR(1) structure. We obtain similar posterior predictive marginal distributions for m_{i0} and m_{i1} , as well as the same posterior inference for K . In particular, a posteriori, the marginal distribution of m_{i1} is concentrated around 0, supporting the hypothesis of independence between gap times.

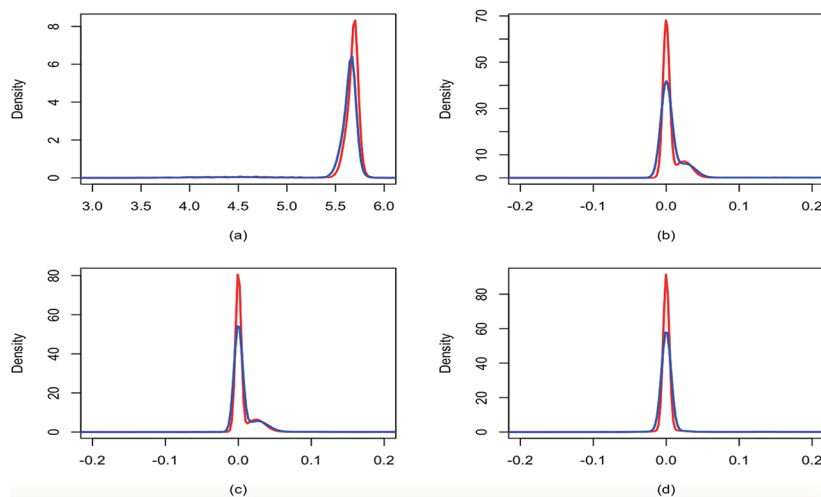


Figure 7: UTI dataset: marginal posterior predictive distributions of m_{i0} (a), m_{i1} (b), m_{i2} (c) and m_{i3} (d) obtained with the spike and slab base measure (red) and with the prior on the order of dependence (blue).

7 Conclusion

In this work we propose novel Bayesian nonparametric approaches for modelling gap times between recurrent events. Time-dependency is taken into account through the specification of an autoregressive model on the frailty parameters governing the distributions of the gap times. To allow for clustering of patients, overdispersion and outliers, we introduce Dirichlet process mixtures as random effects distribution. Both time-homogeneous and inhomogeneous covariates can be included in the model and we allow for their effect to vary through time.

The strategy we adopt is flexible and allows testing for the order of dependence among random effects at different times, that is a key feature of the nonparametric AR(p) model. We propose two different methods to test the order of dependence: spike and slab base measure and direct prior specification on the order of dependence. In the first case we can simply modify the base measure of the DP, whereas with the second technique, we elicit a prior on the order p of the autoregressive process and then, conditioning on p , we set a Dirichlet Process prior of appropriate dimension for the parameters of the AR(p) model.

We can introduce time-dependency between subject-specific gap times in different ways. The simplest and probably most natural way consists of assuming that the random effects at time $j-1, \dots, j-p$ influence the behaviour of the random effect at time j . We then investigate the possibility of approximating higher order of dependency using summary statistics of past observations. Our results show that the choice of summary statistics is crucial and not obvious and that the approximation worsens as the number of gap times increases. As such, this topic will be object of future research, possibly borrowing ideas from the Approximate Bayesian Computation literature.

In our model the random effect has a key clinical importance as it highlights the temporal evolution of the gap times. Our model allows for situations in which you can have independence among gap times as well as temporal dependence. This gives insight into the mechanisms of disease progression and the efficacy of medical interventions. In the cancer example, from Figure 3 it is evident that there is a time dependence between gap times, for some clusters of observations, up to order 2. For example, we find a cluster of individuals where there is a negative correlation with the second lag gap time. For another group this correlation is positive. These considerations have consequences on the expected length of time between events. On the other hand, there is no strong evidence of time dependency between gap times in the urinary infection example, which can impact therapeutic decisions.

The model strategy we propose can be extended to other fields of application; in particular it is straightforward to adapt the proposed approach to model multiple time series analysis [see [29, 45]. In fact, in this case, the data consist in N time series $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})$, where i denotes the time series and n_i is the number of observations for each series. The likelihood for each time series can be expressed as in (1) and temporal dependence may be introduced as in (2)–(3) with appropriate choice of the function $g(\cdot)$.

Finally although the main focus of the paper is on modelling the recurrent event process itself, the proposed model can also be used as building block in a hierarchy to describe the relationship between recurrent events and survival up to a terminating event. In recent years there has been a surge of interest in methods for analysing recurrent events data with risk of termination dependent on the recurrent process [46]. This can be achieved by specifying a joint distribution of the gap times and event (termination) time. Let \mathcal{T}_i denote the event time of interest for individual i . We can assume a survival regression model for \mathcal{T}_i , for example:

$$\mathcal{T}_i \mid \boldsymbol{\gamma}_i, \mathbf{x}_i^*, \rho_i \sim f(t_i \mid \mathbf{x}_i^* \boldsymbol{\gamma}_i, \rho_i)$$

where \mathbf{x}_i^* is a vector of covariates relevant for survival and $\boldsymbol{\gamma}_i$ is the vector of respective coefficients. We link the survival and recurrence processes through the parameter ρ_i by setting a joint nonparametric prior on ρ_i and (m_{i0}, m_{i1}) in (3):

$$(\rho_i, m_{i0}, m_{i1}) \mid G \stackrel{\text{iid}}{\sim} G, \quad G \sim DP(M, G_0)$$

with $G_0 = \mathcal{N}(0, \sigma_\rho^2) \times \mathcal{N}(0, \sigma_g^2) \times \mathcal{N}(\mu_z, \sigma_z^2)$. This latter extension is object of on-going investigation.

References

- [1] González JR, Fernandez E, Moreno V, Ribes J, Peris M, Navarro M, Cambray M, Borràs JM. Sex differences in hospital readmission among colorectal cancer patients. *J Epidemiol Commun Health.* 2005;59:506–11.
- [2] Kregge S, Gianì G, Meyer R, Otto T, Rubben H. A randomized multicenter trial of adjuvant therapy in superficial bladder cancer: transurethral resection only versus transurethral resection plus mitomycin c versus transurethral resection plus bacillus calmette-guerin. *J Urol.* 1996;156:962–66.

- [3] Ma Z, Krings AW. Multivariate survival analysis (i): shared frailty approaches to reliability and dependence modeling. In: 2008 IEEE Aerospace Conference, IEEE, 2008:1–21.
- [4] Callens M, Croux C. Poverty dynamics in europe: a multilevel recurrent discrete-time hazard analysis. *Int Sociol.* 2009;24:368–96.
- [5] Cook RJ, Lawless JF. The statistical analysis of recurrent events. New York: Springer, 2007.
- [6] Aalen O, Borgan O, Gjessing H. Survival and event history analysis: a process point of view. New York: Springer, 2008.
- [7] Wang MC, Qin J, Chiang CT. Analyzing recurrent event data with informative censoring. *J Am Stat Assoc.* 2001;96:1057–65.
- [8] Thall PF. Mixed poisson likelihood regression models for longitudinal interval count data. *Biometrics.* 1988;44:197–209.
- [9] Ma L, Sundaram R. Analysis of gap times based on panel count data with informative observation times and unknown start time. *J Am Stat Assoc.* 2016;113:294–305.
- [10] Pepe MS, Cai J. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *J Am Stat Assoc.* 1993;88:811–20.
- [11] Lawless JF, Nadeau C. Some simple robust methods for the analysis of recurrent events. *Technometrics.* 1995;37:158–68.
- [12] Lawless J, Nadeau C, Cook R. Analysis of mean and rate functions for recurrent events. In: *Proceedings of the First Seattle Symposium in Biostatistics*, Springer, 1997:37–49.
- [13] Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data. *Biometrika.* 1981;68:373–9.
- [14] Chang SH, Wang MC. Conditional regression analysis for recurrence time data. *J Am Stat Assoc.* 1999;94:1221–30.
- [15] McGilchrist C, Aisbett C. Regression with frailty in survival analysis. *Biometrics.* 1991;47:461–66.
- [16] Duchateau L, Janssen P. The frailty model. Springer Science & Business Media, 2007.
- [17] Lin LA, Luo S, Davis BR. Bayesian regression model for recurrent event data with event-varying covariate effects and event effect. *J Appl Stat.* 2018;45:1260–76.
- [18] Chatterjee M, Sen Roy S. A copula-based approach for estimating the survival functions of two alternating recurrent events. *J Stat Comput Simul.* 2018;88:3098–115.
- [19] Meyer R, Romeo JS. Bayesian semiparametric analysis of recurrent failure time data using copulas. *Biometrical J.* 2015;57:982–1001.
- [20] Antoniak CE. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Stat.* 1974;2:1152–74.
- [21] Lo A. On a class of Bayesian nonparametric estimates. I. density estimates. *Ann Stat.* 1984;12:351–57.
- [22] Ferguson TS. A Bayesian analysis of some nonparametric problems. *Ann Stat.* 1973;1:209–30.
- [23] Sethuraman J. A constructive definition of Dirichlet priors. *Stat Sin.* 1994;4:639–50.
- [24] Müller P, Rosner GL. A Bayesian population model with hierarchical mixture priors applied to blood count data. *J Am Stat Assoc.* 92:1279–92.
- [25] Kleinman KP, Ibrahim JG. A semiparametric Bayesian approach to the random effects model. *Biometrics.* 1998;54:921–38.
- [26] Pennell ML, Dunson DB. Bayesian semiparametric dynamic frailty models for multiple event time data. *Biometrics.* 2006;62:1044–52.
- [27] Paulon G, De Iorio M, Guglielmi A, and Ieva F. Joint modeling of recurrent events and survival: a Bayesian non-parametric approach, 2018. <https://dx.doi.org/10.1093/biostatistics/kxyo26>.
- [28] George E, McCulloch R. Variable selection via Gibbs sampling. *J Am Stat Assoc* 1993;88:881–89.
- [29] Di Lucca MA, Guglielmi A, Müller P, Quintana FA. A simple class of Bayesian nonparametric autoregression models. *Bayesian Anal.* 2013;8:63–88.
- [30] Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Anal.* 2006;1:515–34.
- [31] Clyde M, George EI. Model uncertainty. *Stat Sci.* 2004;19:81–94.
- [32] George EI, McCulloch RE. Approaches for Bayesian variable selection. *Stat Sin.* 1997;7:339–73.
- [33] Kim S, Dahl DB, Vannucci M. Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models. *Bayesian Anal* 2009;4:707–32.
- [34] Quintana FA, Müller P. Nonparametric Bayesian assessment of the order of dependence for binary sequences. *J Comput Graphical Stat.* 2012;13:213–31.
- [35] Carlin BP, Chib S. Bayesian model choice via markov chain monte carlo methods. *J Royal Stat Soc Ser B (Methodological).* 1995;57:473–84.
- [36] Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, 2003.
- [37] Ishwaran H, Zarepour M. Exact and approximate sum representations for the Dirichlet process. *Can J Stat.* 2002;30:269–83.
- [38] Plummer M, Best N, Cowles K, Vines K. CODA: Convergence diagnosis and output analysis for MCMC. *R News.* 2006;6:7–11.
- [39] Kass RE, Carlin BP, Gelman A, Neal RM. Markov chain Monte Carlo in practice: a roundtable discussion. *Am Stat.* 1998;52:93–100.
- [40] Manda SO, Meyer R. Bayesian inference for recurrent events data using time-dependent frailty. *Stat Med.* 2005;24:1263–74.
- [41] Rondeau V, Commenges D, Joly P. Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime Data Anal.* 2003;9:139–53.
- [42] Rondeau V, Mazroui Y, Gonzalez JR. frailtypack: an R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *J Stat Softw.* 2012;47:1–28.
- [43] Khasriya R, Khan S, Lunawat R, Bishara S, Bignal J, Malone-Lee M, Ishii H, O'Connor D, Kelsey M, Malone-Lee J. The inadequacy of urinary dipstick and microscopy as surrogate markers of urinary tract infection in urological outpatients with lower urinary tract symptoms without acute frequency and dysuria. *J Urol.* 2010;183:1843–7.
- [44] Kupelian AS, Horsley H, Khasriya R, Amussah RT, Badiani R, Courtney AM, Chandhoyke NS, Riaz U, Savlani K, Moledina M, Montes S, O'Connor D, Visavadia R, Kelsey M, Rohn JL, Malone-Lee J. Discrediting microscopic pyuria and leucocyte esterase as diagnostic surrogates for infection in patients with lower urinary tract symptoms: results from a clinical and laboratory evaluation. *BJU Int.* 2013;112:231–38.
- [45] Nieto-Barajas LE, Quintana FA. A Bayesian non-parametric dynamic AR model for multiple time series analysis. *J Time Ser Anal.* 2016;37:675–89.
- [46] Sinha D, Maiti T, Ibrahim JG, Ouyang B. Current methods for recurrent events data with dependent termination: a bayesian perspective. *J Am Stat Assoc.* 2008;103:866–78.

Supplementary Material: The Supplementary Materials file includes further posterior inference under our models for the different simulated scenarios referenced in Section 4 (see Sections 1–2 there) and two extra simulation scenarios (Sections 3–4 there). Further posterior inference on the *Readmission* dataset is included in Section 5 and 6. Sections 7 and 8 contains the JAGS files.

The online version of this article offers supplementary material (<https://doi.org/10.1515/ijb-2018-0088>).