University College London

DOCTORAL THESIS

Non-parametric Bayes in biostatistics

Author:
Marta Tallarita

Supervisor:
Prof. Maria De Iorio

A thesis submitted in fulfillment of the requirements

for the degree of Doctor of Philosophy

in the

Department of Statistical Science

Faculty of Mathematical and Physical Sciences

Declaration of Authorship

I, Marta Tallarita confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: Marta Tallarita

Abstract

The main focus of this Phd project is the application of Bayesian models in Biostatistics. It has become indeed evident that healthcare management is in need for methods able to improve evidence-based practice.
The first problem we consider is modelling recurrent event and survival time. Recurrent event processes generate events repeatedly over time and they arise in many applications.Typically, the focus is on modelling 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. Although the first focus of our work is on modelling the recurrent event process itself, we also extend the proposed model as building block in a hierarchy to describe the relationship between recurrent events and survival up to a terminating event. This is achieved by specifying a joint distribution of the gap times and event (termination) time.
The second objective is to identify the most promising methods that can be applied in a network meta-analysis (NMA) across longitudinal time points, compare them and extend existing models in a B-spline setting. The network meta-analysis methods extend the standard meta-analysis methods, allowing pairwise comparison of all treatments in a network in the absence of head-to-head comparisons. We focus on the most recent methods suggested in the literature that incorporate multiple time points and allow indirect comparisons of treatment effects across different longitudinal studies. In particular, we compare the Mixed Treatment Comparison model (MTC) of Dakin et al. (2011), the Bayesian evidence synthesis techniques — integrated two-component prediction (BEST-ITP) developed by Ding and Fu (2013) and the more recent method based on fractional polynomials of Jansen et al. (2015). After a comparison of these methods, we develop some models within a B-spline framework.

Impact Statement

We assist nowadays to a period of profound public healthcare demand. To address this need, this research has informed strategies to prevent disease and promote health, together with facilitating high-quality management aimed at improving healthcare and reducing inequalities. Our strategy is to develop research models with high impact on health services, public health and prevention; we focus on the recurrent event and survival time but we also work with the network meta-analysis (NMA) across longitudinal time points.

On one hand, by modelling the whole sequence of waiting times between successive realizations of the recurrent events, we develop a Bayesian semiparametric model. This model allows us to investigate the order of the autoregression as well as to do 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. Moreover, in recent years there has been a surge of interest in methods for analysing recurrent events data with risk of termination, which may be dependent on the recurrent process and that is essential to account for. For instance, death terminates recurrent disease events, while frailty affects both recurrence and survival. Therefore, modelling recurrence and termination jointly can better capture dependency and improve interpretation. We propose a joint model for recurrence and survival processes. The strength of the association between recurrences and terminal event may then be interpreted in terms of patients' risk profiling, and a better understanding of how recurrences affect survival may lead to a more effective planning of healthcare resources.

Furthermore, since HTA is ultimately focused on base drugs approval, clinical protocols, guidelines formulation and decision-making, another important goal in the healthcare fold is the systematic reviews and pairwise meta-analyses of randomized controlled trials as well as the cost effectiveness of different interventions. Thus, we aim to provide a basic explanation of network meta-analysis conduction, highlighting its risks and benefits for evidence-based practice. Over the last years this statistical tool has matured as technique with models available for all types of raw data, producing differ-

ent pooled effect measures, using both Frequentist and Bayesian frameworks. However, the conduction, report and interpretation of network meta-analysis still poses multiple challenges that should be carefully considered, especially because this technique inherits all assumptions from pairwise meta-analysis but with increased complexity. Therefore, we provide a basic explanation of the most recent Bayesian methods suggested in the literature, including information on statistical methods evolution based on a B-spline.

To conclude, possible future research may be well developed on the arguments here analyzed, we hope nonetheless that this Phd project manages to provide, in the near future, healthcare management with solid support in its daily tasks.

Contents

List of Figures

List of Tables

Chapter 1

Bayesian inference

In this Chapter, we introduce the fundamentals of Bayesian data analysis and in particular we focus on Bayesian nonparametric methods, introducing the most popular priors (Dirichlet process). Dirichlet process mixture (DPM) models and the sampling strategies for posterior simulations are discussed.

1.1 The general framework

In this Section we briefly introduce the main basics of Bayesian inference. Let $(X_n)_{n\geq 1}$ be a realization of random variables taking values in a measurable space, referred to as sample space. We assume that, conditionally to an unknown quantity $\theta \in \Theta$, the observations are independent and identically distributed (i.i.d.) from a probability distribution $p(X|\theta)$. Θ is called the parameter space, while $p(X|\theta)$ is the sampling model (or model). The main aim of statistical inference is to collect information about θ. This may be achieved using different alternatives and in this work we focus on Bayesian inferential procedures. Under this approach the unknown quantity θ is treated as a random variable. To perform the inference, we need to specify a prior distribution $\pi(\theta)$ which is the assumed distribution for θ before running the experiment that generated the observations. More precisely, the model is described as:

$$\begin{aligned} X_1, \ldots, X_n|\theta &\overset{\text{iid}}{\sim} p(X|\theta) \\ \theta &\sim \pi(\theta) \end{aligned}$$

The prior distribution reflects the understanding, knowledge and belief about θ before analyzing the experiment that generated the data. A key question, in Bayesian analysis, is the choice of this distribution, as it should collect all information on the unknown parameter (see Gelman et al. (2008)). There exist various ways to solve this problem, nevertheless a widely used alternative is

to utilize a non-informative prior that essentially reflects a lack of strong and precisely quantified prior information. For more details, see Robert (2007). After seeing the data $X_1, \ldots, X_n$, it is possible to compute the posterior distribution for θ, i.e. the distribution of θ given the information contained in the data. Using Bayes theorem, it is given by:

$$p(\theta|X_1, \ldots, X_n) = \frac{p(X_1, \ldots, X_n|\theta)\pi(\theta)}{p(X_1, \ldots, X_n)}$$

where, $p(X_1, \ldots, X_n|\theta)$ is the likelihood of observing the result given the distribution for θ. Since the observations are independent and identically distributed, it can be computed as:

$$p(X_1, \ldots, X_n|\theta) = \prod_{i=1}^{n} p(X_i|\theta)$$

Instead, $p(X_1, \ldots, X_n)$ is the marginal likelihood and represents the evidence. This is the probability of data as determined by integrating across all possible values of θ:

$$p(X_1, \ldots, X_n) = \int_{\Theta} \prod_{i=1}^{n} p(X_i|\theta)\pi(\theta)d\theta$$

As in classical statistics, two approaches are distinguished in Bayesian statistics: parametric and nonparametric. In the former case we have a prior distribution defined on a finite dimensional space. Instead, in the latter case Θ is an infinite vector space. This means that parametric models fix the complexity of the model a priori, while nonparametric methods let the model increase in complexity as new observations become available. The most common models in Bayesian literature are the parametric models and they can be useful in many situations. However, in certain circumstances, constraining the analysis to a specific parametric form may be over-restrictive. Therefore, we would like to use a nonparametric model where we relax parametric assumptions in order to allow greater modelling flexibility and robustness against misspecification of a parametric statistical model. The part of Bayesian inference dealing with nonparametric models is referred to as BNP.

To define a nonparametric Bayesian model, we have to define a probability distribution (the prior) on an infinite-dimensional space. A distribution on an infinite-dimensional space Θ is a stochastic process with paths in Θ. Such distributions are typically harder to define than distributions on a finite space and in many cases we cannot explicitly write down a formula for the prior. There is a Bayesian literature on stochastic processes to priors which could be useful for practical problems and for which stochastic process posteriors

could be derived analytically. The most common choices include Gaussian processes (Williams and Rasmussen (2006)) and Completely Random Measures (CRM), like the well known Gamma process and Beta process (Kingman (1967) and Lijoi et al. (2010)). In this work we choose a Dirichlet process prior, which is obtainable normalising a Gamma process. A brief review about Dirichlet Process and its properties is presented in the following Sections.

1.2 Dirichlet Process

The Dirichlet process (DP) is one of the most important families of prior distribution in Bayesian nonparametric settings. It was firstly introduced by Ferguson (1973a), who used the Kolmogorov Consistency Theorem, gave the definition of a Dirichlet Process and described the Posterior Dirichlet Process. After, Blackwell (1973) used the de Finetti's Theorem to prove the existence of such a random probability measure and introduced the Blackwell-MacQueen urn scheme which satisfies the properties of Dirichlet Process. In 1994 Sethuraman (1994) provided an additional simple and direct way of constructing a DP by introducing the Stick-breaking construction. Finally another representation was provided by Aldous (1985) who introduced the Chinese Restaurant Process as an effective way to construct a Dirichlet Process.
The Dirichlet prior is easy to elicit, has a manageable posterior and various useful properties, such as the conjugacy and discreteness of the realizations. Before giving a formal definition, it is important to highlight that a realization from DP is neither a scalar nor a vector or a matrix, but is a probability measure.
It can be seen as an infinite dimensional generalization of the finite dimensional Dirichlet distribution (see Griffiths and Ghahramani (2011)):

Definition 1 (Dirichlet process) *Let us consider a measurable space (Θ, $\mathcal{A}$), a positive scalar α and a diffuse probability measure G_0 on (Θ, $\mathcal{A}$). We call DP with parameters α and G_0 the stochastic process whose realizations are random probability measures with the following property. Taking G to be a realization of a DP, we have that for every partition $\{S_1, \ldots, S_k\}$ of Θ the random vector $(G(S_1), G(S_2), \ldots, G(S_k))$ follows a Dirichlet distribution with parameters $(\alpha G_0(S_1), \alpha G_0(S_2), \ldots, \alpha G_0(S_k))$.*

This process is usually indicated with $G \sim DP(\alpha, G_0)$, where the parameter α is called a precision or total mass parameter and G_0 indicates the base measure of the DP (also called centre measure). This definition does not give a method to construct the DP, but it provides the properties for a process to

be a DP. Using the consistency conditions of Kolmogorov (1933), Ferguson (1973a) showed that this process exists. Moreover, another proof for the existence of the DP was stated by Blackwell (1973). The Dirichlet process has several important properties, including, but not limited to, the following.

1. Let be $G \sim DP(\alpha, G_0)$ on (Θ, $\mathcal{A}$). The support of the DP involves the set of all measures absolutely continuous with respect to G_0. Moreover, the support of each realization G is Θ, i.e. it is the same as G_0. If we consider the random quantity $G(A)$, it follows that $\mathbb{E}[G(A)] = G_0(A)$ for any measurable set A in $\mathcal{A}$, and thus we can say that G_0 is the prior expectation of G.

2. The Dirichlet process prior is a conjugate prior on Θ. Let consider the hierarchical model given by:

$$\begin{aligned} \theta_1, \ldots, \theta_n \mid G &\overset{ind}{\sim} G \\ G &\sim DP(\alpha, G_0), \end{aligned}$$

   then, the posterior distribution of $G \mid \theta_1, \ldots, \theta_n$ is still a DP. In particular,

$$G \mid \theta_1, \ldots, \theta_n \sim DP\left(\alpha + n, \frac{\alpha\, G_0 \, \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}\right),$$

   where δ_{θ_i} is the Dirac measure that places a unitary mass of probability in corrispondence of θ_i.

3. The Dirichlet process has discrete trajectories, even if G_0 is continuous. Given this property, the DP is an ideal process for mixture models and for the model based clustering. This charateristic is evident from the Sethuraman's constructive definition of DP, described in Section 1.2.2.

The different representations of the DP furnished by Blackwell et al. (1973), byAldous (1985) and by Sethuraman (1994) are described in the following sections.

1.2.1 Blackwell-MacQueen urn

Blackwell et al. (1973) described the construction of a Dirichlet prior distribution by a generalization of P*ó*lya's urn scheme, called Blackwell-Macqueen Urn (BMU). In the basic P*ó*lya urn model, the urn contains a certain number of red and green balls. One ball is drawn randomly from the urn and its color observed; it is then returned in the urn, and an additional ball of the same

color is added to the urn. This selection process is repeated. See Mahmoud (2008) for further details. In Blackwell-Macqueen urn model, the color of the balls is instead replaced with values from a probability measure G_0. The joint law of the sample $\boldsymbol{\theta}$ is fully specified by the following full conditionals:

$$\begin{aligned} \theta_1 &\sim G_0 \\ \theta_2 \mid \theta_1 &\sim \frac{\alpha}{\alpha+1}\, G_0 + \frac{1}{\alpha+1}\, \delta_{\theta_1} \\ &\dots \\ \theta_n \mid \theta_1, \dots, \theta_{n-1} &\sim \frac{\alpha}{\alpha+n-1}\, G_0 + \frac{1}{\alpha+n-1} \sum_{i=1}^{n-1} \delta_{\theta_i} \end{aligned} \tag{1.1}$$

Blackwell et al. (1973) showed that:

- if n tends to ∞ then the probability distribution in 1.1 converges almost surely to a discrete probability measure, called G;
- $G \sim DP(\alpha, G_0)$;
- if G is distributed as a $DP(\alpha, G_0)$ then $\theta_1, \dots, \theta_n \mid G \overset{\text{iid}}{\sim} G$.

Let us introduce the *De Finetti's Representation Theorem* (De Finetti (1931)) to understand the importance of the last item. In order to state de Finetti's Representation Theorem, it is necessary to define the concept of exchangeablity:

Definition 2 (Exchangeablity) *Let $\theta_1, \theta_2, \dots$ be a sequence of random elements taking value on Θ and let $A_1, \dots A_n$ be subsets of Θ where n is an arbitrary integer. This sequence is said to be infinitely exchangeable when, $\forall\ n \geq 1$ and for any finite permutation σ of $(1, 2, \dots, n)$, the random vectors $(\theta_1, \dots, \theta_n)$ and $(\theta_{\sigma(1)}, \dots, \theta_{\sigma(n)})$ have the same probability distribution. More formally we have that*

$$\textit{Pr}\left[\theta_1 \in A_1, \dots, \theta_n \in A_n\right] = \textit{Pr}\left[\theta_{\sigma(1)} \in A_1, \dots, \theta_{\sigma(n)} \in A_n\right]$$

An example of exchangeable sequence is $\theta_1, \theta_2, \dots$ generated using Equation 1.1. From Definition above, De Finetti showed that the sequence $\theta_1, \theta_2, \dots$ is exchangeable if and only if the joint distribution of any subsequence $(A_1, \dots, A_n)$ in Θ can be written as:

$$\text{Pr}\left[\theta_1 \in A_1, \dots, \theta_n \in A_n\right] = \int \prod_{i=1}^{n} G(A_i)\, dP(G), \tag{1.2}$$

where G is a random probability measure (directing random measure), distributed according P (De Finetti's mixing measure). From Equation 1.2, it is possible to conclude that, conditioning on G, the random sequence $\theta_1, \ldots, \theta_n$ are iid. So, De Finetti's theorem connects the concept of iid with the concept of exchangeablity.

Now, going back to the results provided by Blackwell et al. (1973) and considering G as the limit of distribution 1.1, then G is the directing random measure for the sequence obtained by the BMU and the law of G, i.e. P, is the DP. Moreover, the joint distribution of a sequence $\theta_1, \ldots, \theta_n$, from 1.1, is given by:

$$\text{Pr}\left[\theta_1 \in A_1, \ldots, \theta_n \in A_n\right] = \prod_{i=1}^{n} \frac{\alpha\, G_0(A_i) + \sum_{l<i}\, \delta_{\theta_l}(A_i)}{\alpha + i - 1},$$

with $\delta_{\theta_l}(A_i)$ equal to 1 if θ_l belongs to A_i. The Blackwell-Macqueen urn scheme underlies the discrete nature of the samples generated from a DP, as well as the induced partition on the random vector $\boldsymbol{\theta}$. This has two main implications: (1) the sequence $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ can be rewritten as the sequence of its unique values $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_k^*)$, with $k \leq n$, (2) the vector $\boldsymbol{s} = (s_1, \ldots, s_n)$ with $s_i \in \{1, ..., k\}$, which indicates at which component k is associated each $i \in \{1, ..., n\}$, defines a partition of the set $\{1, ..., n\}$.

The BMU is linked with another important process called Chinese Restaurant Process (CRP), which is obtained integrating out $\boldsymbol{\theta}^*$ from the BMU. This representation was provided by Aldous (1985), who introduced the Chinese Restaurant Process as an effective way to construct a Dirichlet Process. This scheme uses the following analogy: assume that there is a Chinese restaurant with infinite many tables. We start with an empty restaurant. As the first customer, $i = 1$ enter the restaurant he sits randomly to any of tables. We denote this table with $j = 1$. The second customer $i = 2$ entering the restaurant, has 2 choices: 1) sit on the occupied table ($j = 1$) 2) ask for a new table ($j = 2$). The probabilities for the two options are assumed to be proportional to the number of people sitting at $j = 1$, which we denote n_1 , and α respectively. If the second customer asked for a new table, $j = 2$, then the third costumer, $i = 3$, will have the choice of sitting at $j = 1$, $j = 2$ or at new table with probability proportional to n_1 , n_2 and α. By iterating the above mentioned mechanism, this method creates a partition of the customers based on which table they sit at. Considering now the $(n + 1)$-th customer enters the restaurant he will sit at one of the k (with $k \leq n$) tables already occupied with probabilities proportional to $n_1, \ldots, n_k$ or to a new table with probability

proportional to α. In other words:

$$\text{Pr}\left[i = n+1 \text{ assigned to } j \mid \rho_n\right] = \begin{cases} \frac{n_j}{\alpha+n} & \text{for } j = 1, \ldots, k \\ \frac{\alpha}{\alpha+n} & \text{for } j = k+1 \end{cases}, \tag{1.3}$$

where ρ_n is a partition of $N = \{1, \ldots, n\}$. Such partitions are exchangeable. In fact, the probability for two customers of sitting together is constant. From the 1.3, it is clear that the probability to be assigned to a specific table depends only on the number of costumers sitting at each of the tables and it is not determined neither by the names of the tables nor by the names of the customers. Under this process, the general formula of the probability of a generic partition of n items, is given by:

$$p(\rho_n) = \frac{\alpha^k}{\alpha_{(n)}} \prod_{j=1}^{k} (n_j - 1)! \tag{1.4}$$

where $\alpha_{(n)} = \alpha(\alpha+1)\ldots(\alpha+n-1)$. The Chinese Restaurant Process is strongly connected to P*ó*lya urn scheme and Dirichlet Process. The CRP is a way to specify a distribution over partitions (table assignments) of n points and can be used as a prior on the space of latent variable which determines the cluster assignments. The CRP is equivalent to P*ó*lya's urn scheme with only difference that it does not assign parameters to each table/cluster.
Finally, we report some properties based on the partition obtained sampling from DP:

- the expected number of ties a priori in the sequence $\theta_1, \ldots, \theta_k$ is given by:
$$\mathbb{E}[k] = \sum_{i=i}^{n} \frac{\alpha}{\alpha + i - 1},$$

- the distribution of the number of ties (k) in the sequence $\theta_1, \ldots, \theta_k$ is given by:
$$p(k) = |S_1(n,k)| n! \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \alpha^k,$$

where $S_1(n,k)$ is the Stirling number of the first kind.
So far we have introduced a representation of the DP samples, which furnish in details the effects of the discreteness of the DP samples. In the next Section we will present a constructive definition of the DP.

1.2.2 Sethuraman's construction

Sethuraman (1994) provided a constructive definition of the Dirichlet process. This construction, called *stick-breaking*, gives an idea of the structure of the process and provides an easy way to simulate its paths. See for more details Ishwaran and James (2001).
Let us consider a random discrete measure

$$G = \sum_{h=1}^{\infty} w_h \delta_{\theta_h}(\cdot) \tag{1.5}$$

where the weights $w_1, w_2, \ldots$ and the locations $\theta_1, \theta_2, \ldots$ in Θ are random and independently sampled. It follows that, if $\sum_{h=1}^{\infty} w_h = 1$ then G is a random probability measure associated with Θ. Stick-breaking process provides a way for generating $w_1, w_2, \ldots$ in 1.5, such that their sum equals to one. In particular, in this construction there is an independent sequence of r.v.s, $\{v_h\}_{h \geq 1}$ such that $v_h \overset{iid}{\sim} Beta(v_h | a_h, b_h)$. This is useful to define the weights:

$$\begin{aligned} w_1 &= v_1 \\ w_2 &= v_1 \times (1 - v_2) \\ &\vdots \\ w_h &= v_h \prod_{l<h} (1 - v_l). \end{aligned} \tag{1.6}$$

From a practical standpoint, let us imagine a stick of length equal to one. This process here described starts by breaking the stick at a random point. The broken part is saved, while the second part is broken randomly. This step continues iteratively. The locations are instead generated from a diffuse distribution G_0. Sethuraman (1994) proved that if $a_h = 1$ and $b_h = \alpha$, $\forall h = 1, 2, \ldots$ and, independently, $\theta_h \overset{iid}{\sim} G_0$, $\forall h = 1, 2, \ldots$, then G define a Dirichlet process with parameters (α, G_0). Jordan and Teh (2014) presented an important link between the stick-breaking definition of the DP and the BMU. In fact, they showed that the probability of belonging to the first cluster represented by the location θ_1 under 1.6, i.e. $w_1 \sim Beta(1, \alpha)$, corresponds to the one obtained from the BMU.

1.3 Dirichelet Process Mixture Models

As said in Section 1.2, the discreteness of the paths of the Dirichlet process makes DP models particularly effective to deal with clustering problems, con-

sidering a mixture of a continuous kernel densities. This approach has been widely studied by Escobar and West (1995a) and they define this model hierarchically as follows:

$$\begin{aligned} X_1, \dots X_n \mid \theta_1, \dots \theta_n &\overset{\text{ind}}{\sim} p(X_i \mid \theta_i) \\ \theta_1, \dots \theta_n \mid G &\overset{\text{iid}}{\sim} G \\ G &\sim DP(\alpha, G_0) \end{aligned} \tag{1.7}$$

Using this formulation (1.7), it is clear that assuming that $\theta_i,\ i = 1, \dots, n$ is a sample from a Dirichlet process, generates ties and the number of distinct values for θ_i, say k, is a random variable with a prior distribution and a posterior distribution. This implies that the latent parameters θ_i can equivalently be written in terms of its unique values $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_k^*)$ and of the assignment vector $s = (s_1, ..., s_n)$.

Thus, let us denote with $n = (n_1, \dots, n_k)$ the cluster sizes for a partition of n observations into clusters $1, \dots, k$. The prior distribution induced on a partition of the indexes $1, 2, \dots, n$ (called ρ_n) by the DP is:

$$p(\rho_n; n_1, \dots, n_k) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + \sum_{i=1}^{k} n_i)} \alpha^k \prod_{j=1}^{k} \Gamma(n_j)$$

where $k = 1, \dots, n$.

From this prior distribution, Antoniak (1974) showed that the marginal law of the prior number of clusters is

$$p(k) = |S_1(n, k)| \frac{\Gamma(\alpha)}{\Gamma(\alpha + \sum_{i=1}^{k} n_i)} \alpha^k$$

where $S_1(n, k)$ is the Stirling number of the first kind, which can be tabulated or computed by a software.
In order to highlight that 1.7 defines an infinite mixture model we can equivalently write the DPM as:

$$\begin{aligned} X_1, \dots, X_n \mid G &\overset{\text{iid}}{\sim} p\left(X \mid G\right) \\ p\left(X \mid G\right) &= \int p\left(X \mid \theta\right) dG(\theta) \\ G &\sim DP(\alpha, G_0) \end{aligned} \tag{1.8}$$

Moreover, recalling Sethuraman's construction, i.e. the discrete nature of the

DP sample, we can rewrite the sampling model as an infinite mixture model:

$$X_1, \dots X_n \mid G \overset{\text{iid}}{\sim} \sum_{h=1}^{\infty} w_h p(X|\theta_h) \tag{1.9}$$

An important point is that, using the P*ó*lya Urn scheme described in Section 1.2.1, it is possible to specify the conditional posterior distribution of θ_i for the DPM model as follows:

$$p(\theta_i|\boldsymbol{\theta}_{-i}, \boldsymbol{X}) \propto \sum_{l \neq i} p(X_i|\theta_i)\delta_{\theta_l}(\theta_i) + \alpha \int p(X_i|\theta) dG_0(\theta) g_0(\theta_i|X_i), \tag{1.10}$$

where $\boldsymbol{\theta}_{-i}$ is the vector of $\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n$ and $g_0(\theta_i|X_i)$ is given by:

$$g_0(\theta_i|X_i) = \frac{p(X_i|\theta_i)g_0(\theta_i)}{\int p(X_i|\theta) dG_0(\theta)}.$$

This equation can be considered as the posterior distribution of θ_i when s_i is different from all other components in $\boldsymbol{s}$.

1.3.1 Computational aspects of DPM models

Posterior inference in DPM models is often derived analytically using Markov Chain Monte Carlo (MCMC) algorithms (for an introductory review on MCMC methods see Andrieu et al. (2003)). These techniques allow to numerically evaluate posterior density, even in very complicated cases where, for example, when the likelihood is obtained through numerical simulations, the size of the parameter space is large and the posterior has a complex structure. In the case of models with DP priors, the algorithms are divided into two main classes, according to their computational strategy. The first group consists of schemes marginalising out the random probability measure G (collapsed Gibbs Samplers); while the second group consists of all the algorithms which impute the Dirichlet process and update it as a component of the Gibbs sampler (conditional methods). In the latter group, some algorithms exploit the stick breaking representation of G (blocked Gibbs Samplers), see Ishwaran and James (2001), whereas others rely on the technique of retrospective sampling, see Papaspiliopoulos and Roberts (2008). For a full and exhaustive discussion on the algorithms for the DP and on their implementation schemes, see Neal (2000a).

Chapter 2

Bayesian Autoregressive Frailty Models for Inference in Recurrent Events

In this Chapter we propose an autoregressive Bayesian semiparametric model 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 it 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. Since 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. In particular, in Section 2.1 the basic theory underlying recurrent event times is presented, as well as a review of existing Bayesian semiparametric approaches to gap times. In Section 2.2 we introduce our model, while in Section 2.3 we explain how to perform inference on the order of dependence in the Markov structure. In Section 2.4 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 2.5 and 2.6 present two medical applications involving recurrent hospitalizations and urinary tract infections, respectively. We conclude the Chapter in Section 2.7.

2.1 Recurrent events and gap times

Processes that generate events repeatedly over time are referred to as recurrent event processes and the data they provide are called recurrent event data. Data may be available from a large number of individuals, but with a small number of occurrences for each subject. These types of processes arise frequently in several settings, for example in medicine (Krege et al., 1996; González et al., 2005, e.g.), economy (e.g. Ma and Krings, 2008) and technology (for example, Callens and Croux, 2009). Typical examples include recurrent infections, asthma attacks, hospitalizations, product repairs, machine failures.

For a single recurrent event process, which is a point process, starting for simplicity at T_0 = 0, let $0 < T_1 < T_2 < ...$ denote the event times, where T_k is the time of the k-th event. The associated counting process $\{N(t), 0 \leq t\}$ records the cumulative number of events generated by the process; specifically, $N(t) = \sum_{k=1}^{\infty} I(T_k \leq t)$ is the number of events occurring over the time interval $[0, t]$. More generally, $N(s, t) = N(t) - N(s)$ represents the number of events occurring over the interval $(s, t]$. As defined here, counting processes are right-continuous, that is, $N(t) = N(t^+)$, where t^+ denote times that are infinitesimally larger than t. Typically, the focus is on modelling 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 Cook and Lawless (2007a)): (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 waiting times between successive realizations of the recurrent events.

The first approach uses results from stochastic calculus and point processes. Methods based on counts are often useful when individuals frequently experience the events of interest, and the events are “incidental" in the sense that their occurrence does not materially alter the process itself. The canonical framework for the analysis of event counts is the Poisson process. Models of this kind can be specified very generally by considering the probability distribution for the number of events in short intervals $[t, t+\triangle t)$. For events occurring in continuous time we make the mathematically convenient assumption that two events cannot occur simultaneously. Then, the intensity process $\lambda(t)$ is defined as the conditional probability that an event occurs in $[t, t + \triangle t)$, given all that has been observed prior to this interval, divided by

the length of the interval. More formally

$$\lambda(t \mid H(t)) = \lim_{\triangle t \to 0} \frac{P(\triangle N(t) = 1 \mid H(t))}{\triangle t} \tag{2.1}$$

where $\triangle N(t) = N(t + \triangle t^-) - N(t^-)$ denotes the number of events in the interval $[t, t + \triangle t)$, and $H(t) = \{N(s) : 0 \leq s < t\}$ denotes the history of the process at time t. For a full review of this class of methods, see Aalen et al. (2008). Wang et al. (2001) 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 (1988), who considers a parametric mixed Poisson model which deals with covariates. Further, Ma and Sundaram (2018) propose a semiparametric model based on a nonhomogeneous Poisson process. Finally, Pepe and Cai (1993), Lawless and Nadeau (1995) and Lawless et al. (1997) propose semiparametric procedures for making inferences about the mean and rate functions of the counting process without the Poisson-type assumption.

The second approach, which focuses on the sequence of gap times, is more appropriate when the recurrent events are relatively infrequent, when individual renewal takes place after an event, or when the goal is prediction of the time to the next event. Therefore, this approach is highly relevant for biomedical applications. For instance, major recurrent cardiac events for one patient are often rather infrequent from a statistical viewpoint. Also, healthcare planning can benefit from time-to-event predictions, especially if events require hospitalization. Nonetheless, there is less existing work on the second approach than on the first.

This work places itself within the second framework, as the events in our application are infrequent but measured on many individuals. We develop Bayesian semiparametric models for gap times between events. Following the notation and the general framework of Cook and Lawless (2007a), let us assume that T_0 = 0 corresponds at the start of each event process and that individual i is observed over the time interval $[0,\ \tau_i]$. Let n_i^* be the number of the events observed at times $0 < T_{i1} < \cdots < T_{in_i^*} \leq \tau_i$. Note that if τ_i corresponds to an actual event, then the last event is observed. Otherwise τ_i becomes a censoring time. We define n_i as the total number of the events; thus if the last event is observed $n_i = n_i^*$, while if the last event is censored $n_i = n_i^* + 1$. Let $W_{ij} = T_{ij} - T_{ij-1}$ for j=1, ...,$n_i - 1$ and $W_{in_i} = \tau_i - T_{in_i-1}$. These are the observed gap times for individual i with the final time being possibly censored. Representation of recurrent events for a generic observation i is shown in Figure 2.1. In the top panel the last event is observed, while in the bottom panel the last gap time is censored.

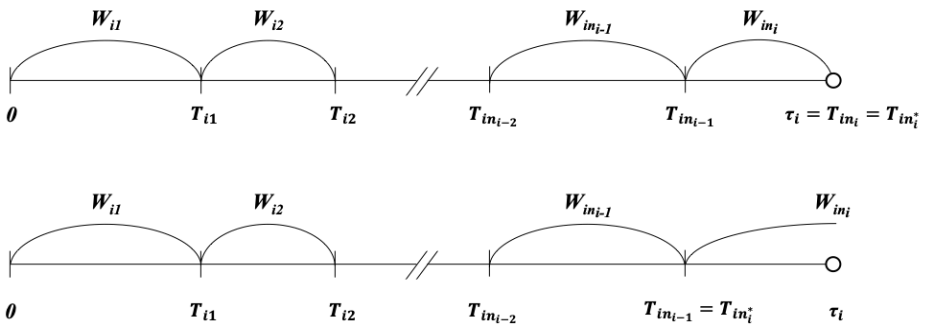


Figure 2.1: Representation of recurrent events for a generic obseravation i. In the top panel the last event is observed, while in the bottom panel the last gap time is censored.

In this scenario the joint distribution $(W_{i1}, ..., W_{in_i})$ is modelled through the specification of the conditional laws $\mathcal{L}(W_{ij}|\mathbf{x}_{ij}, W_{i1}, ...W_{ij-1})$, where by $\mathbf{x}_{ij}$ we denote a vector of possibly time-varying covariates at time j. In the following Subsection, we focus our attention on methods to model the whole sequence of waiting times between successive realizations of the recurrent events.

2.1.1 Renewal Processes

Renewal processes are the canonical models for waiting times and are defined as processes for which the gap times of each patient W_{ij}, $j = 1, \ldots, n_i$ are i.i.d, conditionally on covariates and parameters.
Let $H(t) = \{N(s) : 0 \leq s < t\}$ denote the history of the process at time t and let $h(t)$ be the hazard function, i.e. the instantaneous rate of failure at time t, defined by:

$$h(t) = \lim_{\triangle t \to 0} \frac{\mathbb{P}(t \leq T < t + \triangle t | T > t)}{\triangle t} = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

where $f(t)$ denotes the probability density function (p.d.f.) of T, $F(t) = P(T \leq t)$ is the cumulative distribution function of T and $S(t)$ indicates the survival function, defined as $S(t) = 1 - F(t) = P(T > t)$.
For renewal processes, the intensity function is equal to the hazard rate, i.e. $\lambda(t \mid H(t)) = h(t - T_{N(t^-)})$. The likelihood function from L independent individuals is of the form

$$\mathcal{L} = \prod_{i=1}^{L} \left[\prod_{j=1}^{n_i^*} f(W_{ij}|\mathbf{x}_{ij}) \right] S(W_{in_i^*+1}|\mathbf{x}_{in_i^*+1}) \tag{2.2}$$

where f and S denote the density and survivor functions for W_{ij} given $\mathbf{x}$. This is the familiar likelihood function for a random sample involving failure times W_{ij}, $j = 1, \ldots, n_i^*$ and right censoring times $W_{in_i^*+1}$ for $i = 1, \ldots, L$. If $W_{in_i^*+1}$ = 0, that is, if observation terminates after the n_i^*-th event, the term $S(W_{in_i^*+1}|\mathbf{x}_{in_i^*+1})$ in 2.2 disappears.
However, the strong assumption of independent gap times corresponds to the setting in which individuals are restored to the original physical state after each event. This assumption is often not realistic in applications. However, by extending renewal processes in various ways one can obtain more flexible models. These extensions are presented in the following Subsection.

2.1.2 Extensions and generalisations

In a more general case, when the assumption of independent gap times is unrealistic, models can be formulated through the sequence of conditional laws $\mathcal{L}(W_{ij}|\mathbf{x}_{ij}, W_{i1}, ...W_{ij-1})$, $j = 1, \ldots, n_i$. In this case, the cumulative distribution functions

$$F_j(W|\mathbf{x}_{ij}, W_i^{(j-i)}) = P(W_{ij} \leq w|\mathbf{x}_{ij}, W_i^{(j-1)}) \tag{2.3}$$

where $W_i^{(j-1)} = (W_{i1}, ..., W_{ij-1})'$, can change at each gap time. This format allows various types of dependence on previous event history to be considered, including elapsed time $W_{i1} + \cdots + W_{ij-1}$ up to $(j-1)^{st}$ event. For parametric models, the likelihood function from a set of L independent processes is an extensions of the Equation 2.2:

$$\mathcal{L} = \prod_{i=1}^{L} \left\{ \left[\prod_{j=1}^{n_i^*} f_j(W_{ij}|\mathbf{z}_{ij}) \right] S(W_{in_i^*+1}|\mathbf{z}_{i,n_i^*+1}) \right\} \tag{2.4}$$

where $\mathbf{z}_{ij}$ is a vector that models the dependence of W_{ij} on $\mathbf{x}_{ij}$ and $W_i^{(j-1)}$.
Prentice et al. (1981) propose a PH model for recurrent events. This semiparametric model is obtained by specifying the intensity function as one of the following

$$\lambda(t|N(t),\ X(t)) \ = \ \lambda_{0s}(t) \exp(\mathbf{x}(t)\beta_s) \tag{2.5}$$

$$\lambda(t|N(t),\ X(t)) \ = \ \lambda_{0s}(t \ - \ T_{n(t)}) \exp(\mathbf{x}(t)\beta_s) \tag{2.6}$$

where $T_{n(t)}$ is the time of the preceding event. These two choices correspond to the natural time scale for the baseline hazard function: one is the time t from the beginning of the study and the other is $t - T_{n(t)}$, the time elapsed since the immediately preceding event. Moreover, the index s allows the baseline hazard to be stratum-specific. Chang and Wang (1999) propose a slightly different model by incorporating two kinds of covariates: some structural covariates (fixed) and some episode-specific covariates. For example, in a study of schizophrenia, gender and marital status may have the same effect for different episodes, but the age of disease onset may have distinct effects over different episodes. Frailty models are discussed in McGilchrist and Aisbett (1991) and Duchateau and Janssen (2007). 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. (2018) 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, Chatterjee and Sen Roy, 2018; Meyer and Romeo, 2015). This strategy requires specifying the marginal distribution of each gap time and then using a copula to introduce dependence.

In this work, we investigate different strategies to link gap times at time t with 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. 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 $p \in \{0, 1, \ldots, P\}$, while the other approach employs spike and slab base measures and it is in the spirit of stochastic search variable selection (George and McCullogh, 1993). 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.

2.2 Autoregressive frailty models via Dirichlet process mixtures

Recalling Section 2.1, we consider data on L 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 $[0, \tau_i]$. n_i^* are the observed events at times $0 < T_{i1} < \cdots < T_{in_i^*} \leq \tau_i$ and n_i are the total number of events. The gap times between events of subject i are defined as $W_{ij} = T_{ij} - T_{ij-1}$ for $j = 1, \ldots, n_i - 1$ and $W_{in_i} = \tau_i - T_{in_i - 1}$. We recall that if τ_i corresponds to an actual event, then the last gap time is observed and $n_i = n_i^*$. Otherwise τ_i becomes a censoring time, W_{in_i} is a censored observation and $n_i = n_i^* + 1$. Therefore W_{ij}, $j = 1, \ldots, n_i - 1$, are the observed gap times for individual i with a possibly censored time W_{in_i}. Let J be the maximum number of observed repeated events, i.e. $J = \max_{i=1,\ldots,L}(n_i^*)$ and let $Y_{ij} = \log(W_{ij})$. We describe the joint distribution of $\boldsymbol{Y}_i := (Y_{i1}, \ldots, Y_{in_i})$ through the specification of the conditional laws $\mathcal{L}(Y_{ij} | \boldsymbol{x}_{ij}, Y_{i1}, \ldots, Y_{ij-1})$, where $\boldsymbol{x}_{ij}$ denotes the vector of possibly time-varying covariates for the ith individual. In particular, we assume that an observation at time j, for each subject i, $i = 1, \ldots, L$ and $j = 1, \ldots, n_i$, is distributed as follows

$$Y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_j + \alpha_{ij} + \sigma \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1) \tag{2.7}$$

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 modelling structure as described in Subsections 2.2.1 and 2.2.2. 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:

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

where $\boldsymbol{z}_{ij} = (\boldsymbol{x}_{ij}, y_{i1}, \ldots, y_{ij-1},)$, 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 $\boldsymbol{x}_{ij}$ 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.

2.2.1 Nonparametric AR(1)-type models

Following a similar modelling strategy to the one described in Di Lucca et al. (2013), 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 $j-1$:

$$\alpha_{ij} \mid m_{i0}, m_{i1}, \tau \overset{\text{ind}}{\sim} \mathcal{N}(m_{i0} + m_{i1}\, g(Y_{i1}, \ldots, Y_{ij-1}), \tau^2), \quad j = 1, \ldots, n_i \tag{2.8}$$

$$(m_{i0}, m_{i1}) \mid G \overset{\text{iid}}{\sim} G, \quad G \sim DP(M, G_0). \tag{2.9}$$

When $j = 1$, the distribution of the random effect α_{i1} simplifies as the autoregressive term in (2.8) 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:

$$\begin{aligned} \boldsymbol{\beta}_j &\overset{\text{iid}}{\sim} \mathcal{N}_q(0, \beta_0^2 I_q) \\ \sigma^2 &\sim \text{Inv-Gamma}(a_\sigma, b_\sigma) \\ \tau^2 &\sim \text{Inv-Gamma}(a_\tau, b_\tau) \\ M &\sim \mathcal{U}(0, M_0) \\ G_0 &= \mathcal{N}(0, \sigma_g^2) \times \mathcal{N}(\mu_z, \sigma_z^2), \end{aligned} \tag{2.10}$$

assuming a priori independence among the different parameters.

When implementing the model in JAGS, we opt for marginal uniform priors on a bounded interval for σ^2 and τ^2 as suggested by Gelman (2006) as

inverse-gamma distributions give numerical instability problems when hyperparameters are small. Moreover, Gelman (2006) recommends uniform distributions (on bounded intervals) for scale parameters in hierarchical models as being weakly-informative, as opposed to inverse-gamma priors for variance parameters, which are very sensitive to the choice of the hyperparameters themselves.

The choice of g in (2.8) is obviously crucial and depends on the context and the goals of the inference problem. Common choices in the literature are:

- $g(Y_{i1}, \ldots, Y_{ij-1}) = Y_{ij-1}$, i.e. the random effect at time j depends only on the observation at time $j-1$, conditionally on the remaining parameters;

- $g(Y_{i1}, \ldots, Y_{ij-1}) = (Y_{i1} + \cdots + Y_{ij-1})/(j-1)$, i.e. the conditional expected value of each α_{ij} depends on the sample mean of the observations up to time $j-1$;

- $g(Y_{i1}, \ldots, Y_{ij-1}) = (Y_{i1} \times \cdots \times Y_{ij-1})^{1/(j-1)}$, this is equivalent to the geometric mean of the observations up to time $j-1$.

Note that, when $g(Y_{i1}, \ldots, Y_{ij-1}) = Y_{ij-1}$, then (2.8)-(2.9) 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 $j-1$.

2.2.2 Nonparametric AR(p) models

The model in Subsection 2.2.1 can be extended to include higher order dependence, by modifying (2.8) -(2.9) as follows:

$$\alpha_{ij} \mid m_{i0}, m_{i1}, \ldots, m_{ip}, \tau \overset{\text{ind}}{\sim} \mathcal{N}(m_{i0} + \sum_{l=1}^{p} m_{il}\, Y_{ij-l}, \tau^2), \;\; j = p+1, \ldots, n_i \quad (2.11)$$

$$(m_{i0}, m_{i1}, \ldots, m_{ip}) \mid G \overset{\text{iid}}{\sim} G, \quad G \sim DP(M, G_0) \quad (2.12)$$

$$G_0 = \mathcal{N}(0, \sigma_g^2) \times \underbrace{\mathcal{N}(\mu_z, \sigma_z^2) \times \cdots \times \mathcal{N}(\mu_z, \sigma_z^2)}_{p \text{ times}}. \quad (2.13)$$

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 2.2.1.

2.3 Estimating the Order of Dependence

In (2.11) 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, George and McCulloch, 1997; Clyde and George, 2004). 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. (2009). 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 $(m_{i0}, m_{i1}, \ldots, m_{ip})$ changes according to p and consequently the dimension of the space where the Dirichlet process measure is defined.

2.3.1 Spike and Slab Base Measure

Kim et al. (2009) 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 (2.13) as

$$\begin{aligned} G_0 &= \mathcal{N}(0, \sigma_g^2) \times \underbrace{\pi_1(\mu_z, \sigma_z^2) \times \cdots \times \pi_p(\mu_z, \sigma_z^2)}_{p \text{ times}} \\ \pi_l(\mu_z, \sigma_z^2) &= (1-\eta_l)\delta_0 + \eta_l \mathcal{N}(0, \sigma_z^2),\ l = 1, \ldots, p \\ \eta_l &\overset{\text{ind}}{\sim} \text{Bernoulli}(c_l) \\ c_l &\overset{\text{iid}}{\sim} \mathcal{U}(0,1) \end{aligned} \tag{2.14}$$

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

2.3.2 Prior on the Order of Dependence

Following Quintana and Müller (2012), 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 $(m_{i0}, m_{i1}, \ldots, m_{ip})$. Let P be the maximum possible order. Then we

can specify the following hierarchy:

$$\begin{aligned}
\alpha_{ij} \mid p, m_{i0}, m_{i1}, \ldots, m_{ip}, \tau &\overset{\text{ind}}{\sim} \mathcal{N}(m_{i0} + \sum_{l=1}^{p} m_{il}\, Y_{ij-l}, \tau^2), \quad j = p+1, \ldots, n_i \\
(m_{i0}, m_{i1}, \ldots, m_{ip}) \mid p, \widetilde{G}_p &\overset{\text{iid}}{\sim} \widetilde{G}_p \\
\widetilde{G}_p &\sim DP(M, G_{0p}) \\
G_{0p} &= \mathcal{N}(0, \sigma_g^2) \times \underbrace{\mathcal{N}(\mu_z, \sigma_z^2) \times \cdots \times \mathcal{N}(\mu_z, \sigma_z^2)}_{p \text{ times}} \\
p &\sim \text{Discrete Uniform on } \{0, 1, \ldots, P\}.
\end{aligned} \tag{2.15}$$

When $p = 0$, the process simplifies as the autoregressive term in (2.15) 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 Quintana and Müller (2012), who sample from the posterior distribution of p given the data. For the JAGS implementation we have opted for a pseudo-prior approach, see Carlin and Chib (1995).

2.4 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 Subsections 2.4.1, 2.4.2 and 2.4.5 are generated from a mixture of Gaussian distributions, while in Subsection 2.4.6 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 2.5 and 2.6, can be performed through a standard Gibbs sampler algorithm, which we implement in JAGS (Plummer, 2003), using a truncation-based algorithm for stick-breaking priors (Ishwaran and Zarepour, 2002). 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 (Plummer et al., 2006) that convergence of the chain is satisfactory for most of the parameters.

2.4.1 Simulation scenario 1

We consider a simulated dataset of $L = 300$ 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

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

while another third is generated from

$$\begin{aligned} Y_{i1} \sim \mathcal{N}(0, (1.5)^2), \qquad & Y_{i2}|Y_{i1} \sim \mathcal{N}(Y_{i1}, (1.5)^2) \\ Y_{ij}|Y_{ij-1}, Y_{ij-2} \ \sim \ & \mathcal{N}(Y_{ij-1} + 0.7 \times Y_{ij-2}, (1.5)^2),\ j = 3, \ldots, n_i \end{aligned}$$

and the last 100 observations are generated from

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

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 (2.7), (2.11)-(2.12) with $p = 3$:

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

$$\alpha_{ij} \mid m_{i0}, m_{i1}, \ldots, m_{ip}, \tau \overset{\text{ind}}{\sim} \mathcal{N}(m_{i0} + \sum_{l=1}^{3} m_{il}\, Y_{ij-l}, \tau^2),\ \ j = 4, \ldots, n_i$$

$$(m_{i0}, m_{i1}, \ldots, m_{ip}) \mid G \overset{\text{iid}}{\sim} G, \quad G \sim DP(M, G_0)$$

where G_0 is given by the product of spike and slab distributions as defined in

(2.14):

$$\begin{aligned} G_0 &= \mathcal{N}(0,10) \times \underbrace{\pi_1(0,100) \times \cdots \times \pi_3(0,100)}_{3 \text{ times}} \\ \pi_l(0,100) &= (1-\eta_l)\delta_0 + \eta_l \mathcal{N}(0,100),\ l = 1,2,3 \\ \eta_l &\overset{\text{ind}}{\sim} \text{Bernoulli}(c_l) \\ c_l &\overset{\text{iid}}{\sim} \mathcal{U}(0,1). \end{aligned}$$

In fitting the model we set:

$$\begin{aligned} \sigma &\sim \mathcal{U}(0,10) \\ \tau &\sim \mathcal{U}(0,10) \\ M_0 &= 10. \end{aligned}$$

Hyperparameters are chosen in order to specify vague marginal prior distributions. Figure 2.2 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.

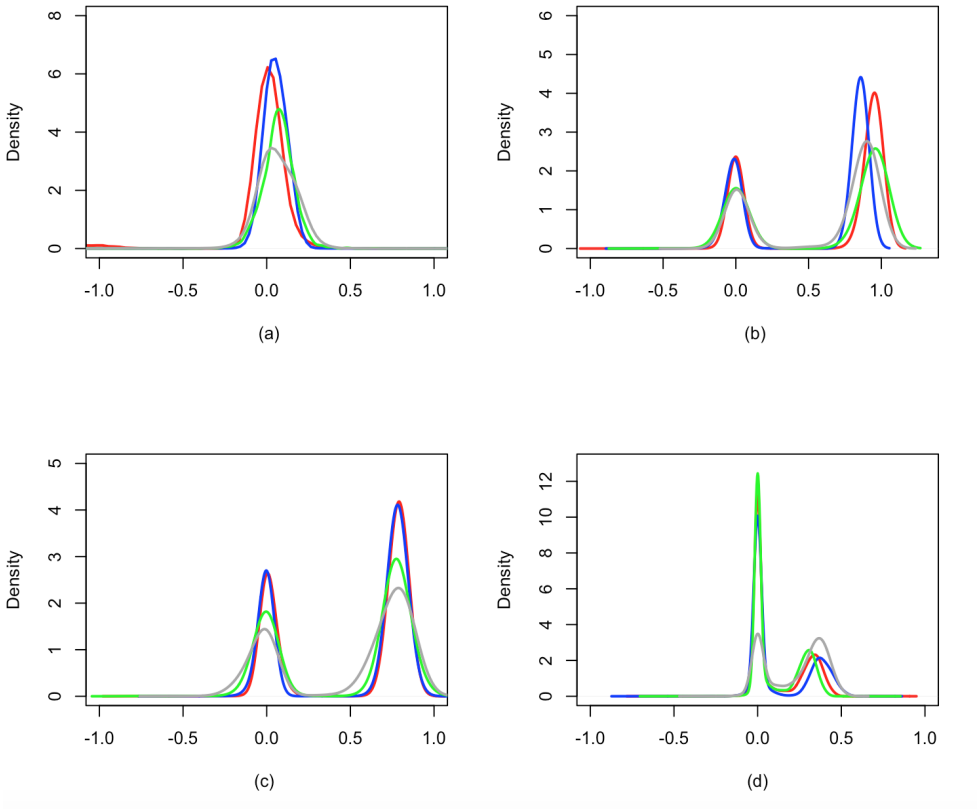


Figure 2.2: Simulation scenario 1: predictive marginal distributions of m_{i0} (a), m_{i1} (b), m_{i2} (c) and m_{i3} (d) under the spike and slab prior. Red, blue, green and grey lines correspond to estimates obtained using the dataset with 0%, 10%, 25% and 50% censoring rates, respectively.

The marginal posterior distributions of η_1, η_2, and η_3 in the prior (2.14), not reported here, concentrate most mass on 1. Moreover, there is a clear difference between the posterior distribution of K, the number of distinct components in the mixture (2.11)-(2.12), using the data with and without censoring (results shown in Figure 2.3). 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.

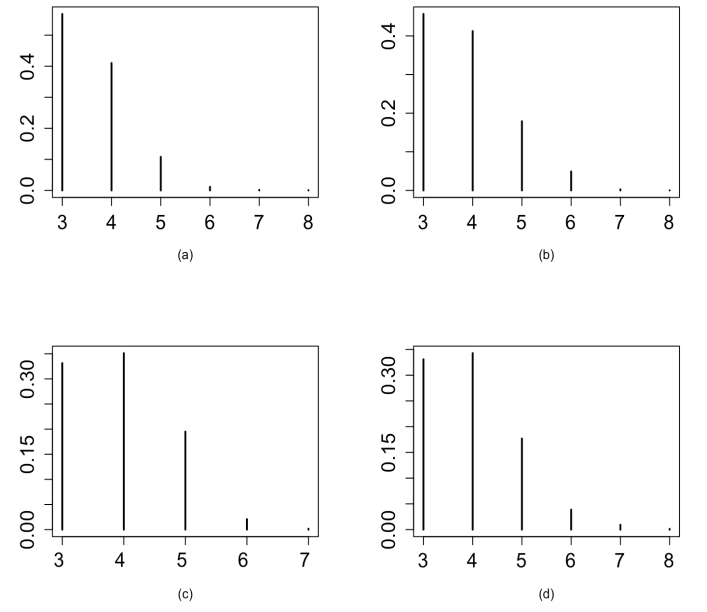


Figure 2.3: Simulation scenario 1: posterior distribution of K under the spike and slab prior. Panels (a), (b), (c), (d), correspond to estimates obtained using the dataset with 0%, 10%, 25% and 50% censoring rates, respectively.

We also fit model (2.7), (2.15) to this dataset. We assume here that the maximum order of dependence is $P = 3$. We fit the following model:

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

$$\begin{aligned}
\alpha_{ij} \mid p, m_{i0}, m_{i1}, \ldots, m_{ip}, \tau &\overset{\text{ind}}{\sim} \mathcal{N}(m_{i0} + \sum_{l=1}^{p} m_{il}\, Y_{ij-l}, \tau^2), \quad j = p+1, \ldots, n_i \\
(m_{i0}, m_{i1}, \ldots, m_{ip}) \mid p, \widetilde{G}_p &\overset{\text{iid}}{\sim} \widetilde{G}_p \\
\widetilde{G}_p &\sim DP(M, G_{0p}) \\
G_{0p} &= \mathcal{N}(0,10) \times \underbrace{\mathcal{N}(0,100) \times \cdots \times \mathcal{N}(0,100)}_{3 \text{ times}} \\
p &\sim \text{Discrete Uniform on } \{0,1,2,3\} \\
\sigma &\sim \mathcal{U}(0,10) \\
\tau &\sim \mathcal{U}(0,10) \\
M &\sim \mathcal{U}(0,5).
\end{aligned}$$

Similar results are obtained using the prior on the order of dependence. In fact, from Figure 2.4, that displays the marginal posterior predictive distributions of m_{i0}, m_{i1}, m_{i2} and m_{i3}, it is evident that these plots are very similar to those presented in Figure 2.2, displaying the predictive distributions of the

same parameters under the spike and slab prior.

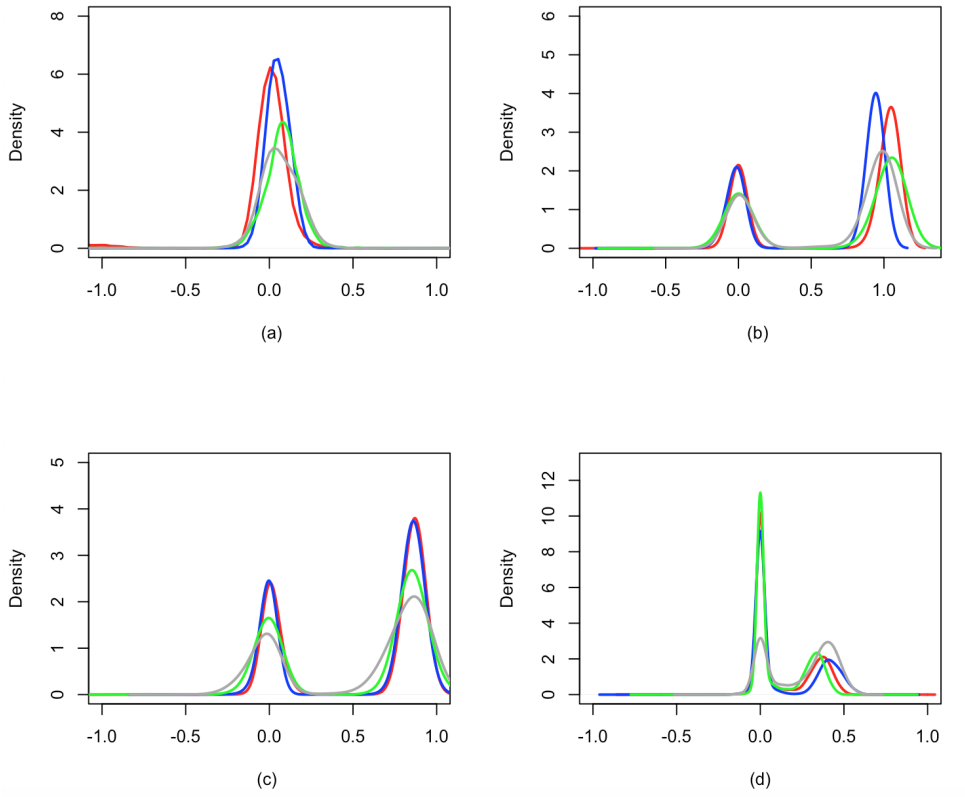


Figure 2.4: Simulation scenario 1: predictive marginal distributions of m_{i0} (a), m_{i1} (b), m_{i2} (c) and m_{i3} (d) obtained specifying a prior on the order of dependence. Red, blue, green and grey lines correspond to estimates obtained using the dataset with 0%, 10%, 25% and 50% censoring rates, respectively.

2.4.2 Simulation Scenario 2

In this Section we simulate a dataset of $L = 200$, 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

$$\begin{aligned} Y_{i1} \sim \mathcal{N}(0, 1.5^2), \qquad & Y_{i2}|Y_{i1} \sim \mathcal{N}(0.9 \times Y_{i1}, 0.9^2) \\ Y_{ij}|Y_{ij-1}, Y_{ij-2} \quad \sim \quad & \mathcal{N}(0.9 \times Y_{ij-1} + 0.7 \times Y_{ij-2}, 0.9^2), \ j = 3, \ldots, n_i \end{aligned}$$

while the other half is independently generated from

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

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 (2.7), (2.11)-(2.12), (2.14) and model (2.7), (2.15). Here we first 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:

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

$$\begin{aligned}
\alpha_{ij} \mid p, m_{i0}, m_{i1}, \ldots, m_{ip}, \tau &\overset{\text{ind}}{\sim} \mathcal{N}(m_{i0} + \sum_{l=1}^{p} m_{il}\, Y_{ij-l}, \tau^2), \quad j = p+1, \ldots, n_i \\
(m_{i0}, m_{i1}, \ldots, m_{ip}) \mid p, \widetilde{G}_p &\overset{\text{iid}}{\sim} \widetilde{G}_p \\
\widetilde{G}_p &\sim DP(M, G_{0p}) \\
G_{0p} &= \mathcal{N}(0,10) \times \mathcal{N}(0,100) \times \mathcal{N}(0,100) \times \mathcal{N}(0,100) \\
p &\sim \text{ Discrete Uniform on } \{0,1,2,3\} \\
\sigma &\sim \mathcal{U}(0,10) \\
\tau &\sim \mathcal{U}(0,10) \\
M &\sim \mathcal{U}(0,5).
\end{aligned}$$

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 2.1. In particular, the mass placed on $p = 2$ decreases with increasing censoring rate.

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

Table 2.1: Simulation scenario 2: posterior distributions (masses) of p under the prior on the order of dependence for different rates of censoring.

Conditional on $p = 2$, Figure 2.5 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.

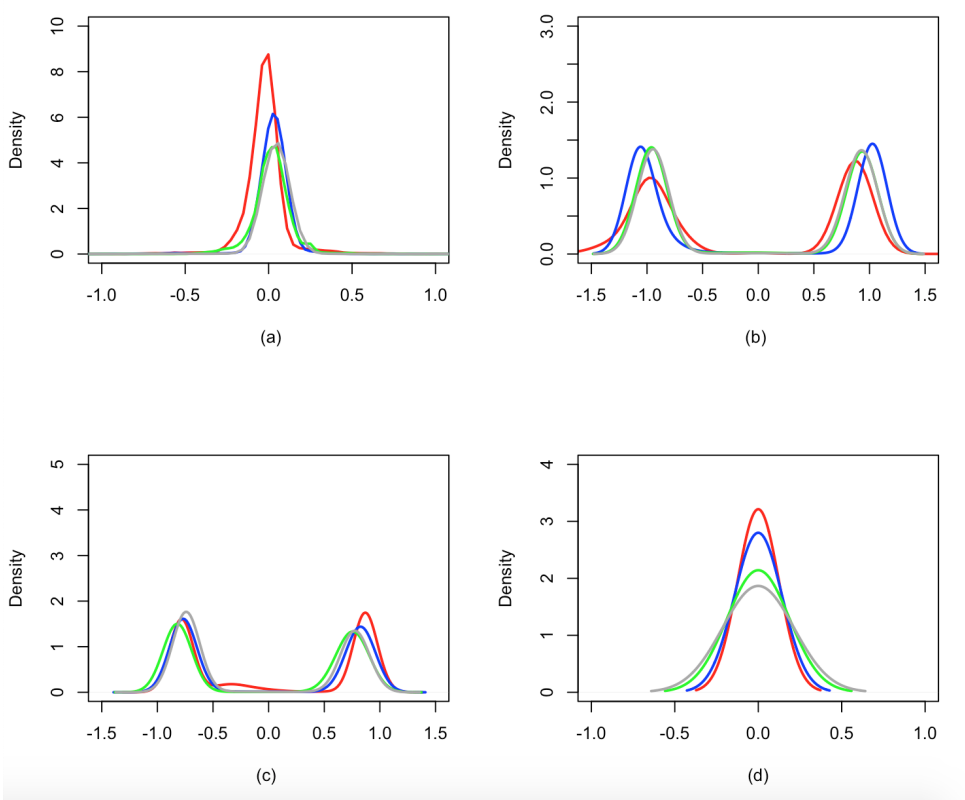


Figure 2.5: Simulation scenario 2: predictive marginal distributions of m_{i0} (a), m_{i1} (b), m_{i2} (c) and m_{i3} (d) under the prior on the order of dependence, conditioning on $p = 2$. Red, blue, green and grey lines correspond to estimates obtained using the dataset with 0%, 10%, 25% and 50% censoring 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.
Now, we briefly report the results obtained by fitting the model with the spike

and slab base measure. More specifically, we fix $p = 3$ and we fit the following model:

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

$$\alpha_{ij} \mid m_{i0}, m_{i1}, \ldots, m_{ip}, \tau \overset{\text{ind}}{\sim} \mathcal{N}(m_{i0} + \sum_{l=1}^{3} m_{il}\, Y_{ij-l}, \tau^2), \;\; j = 4, \ldots, n_i$$

$$(m_{i0}, m_{i1}, \ldots, m_{ip}) \mid G \overset{\text{iid}}{\sim} G, \quad G \sim DP(M, G_0)$$

where G_0 is given by the product of spike and slab distributions as defined in (2.14):

$$\begin{aligned} G_0 &= \mathcal{N}(0,10) \times \underbrace{\pi_1(0,100) \times \cdots \times \pi_3(0,100)}_{3 \text{ times}} \\ \pi_l(0,100) &= (1-\eta_l)\delta_0 + \eta_l \mathcal{N}(0,100), \; l = 1,2,3 \\ \eta_l &\overset{\text{ind}}{\sim} \text{Bernoulli}(c_l) \\ c_l &\overset{\text{iid}}{\sim} \mathcal{U}(0,1) \end{aligned}$$

In fitting the model we set:

$$\begin{aligned} \sigma &\sim \mathcal{U}(0,10) \\ \tau &\sim \mathcal{U}(0,10) \\ M_0 &= 10. \end{aligned}$$

Figure 2.6 displays the predictive marginal distributions of the m_{ij}'s. Once again, if we compare these results with those presented in Figure 2.5, it is clear that the predictive distributions of the m_{ij}'s under the spike and slab prior and the predictive distributions of the m_{ij}'s under the prior on the order of dependence are similar.

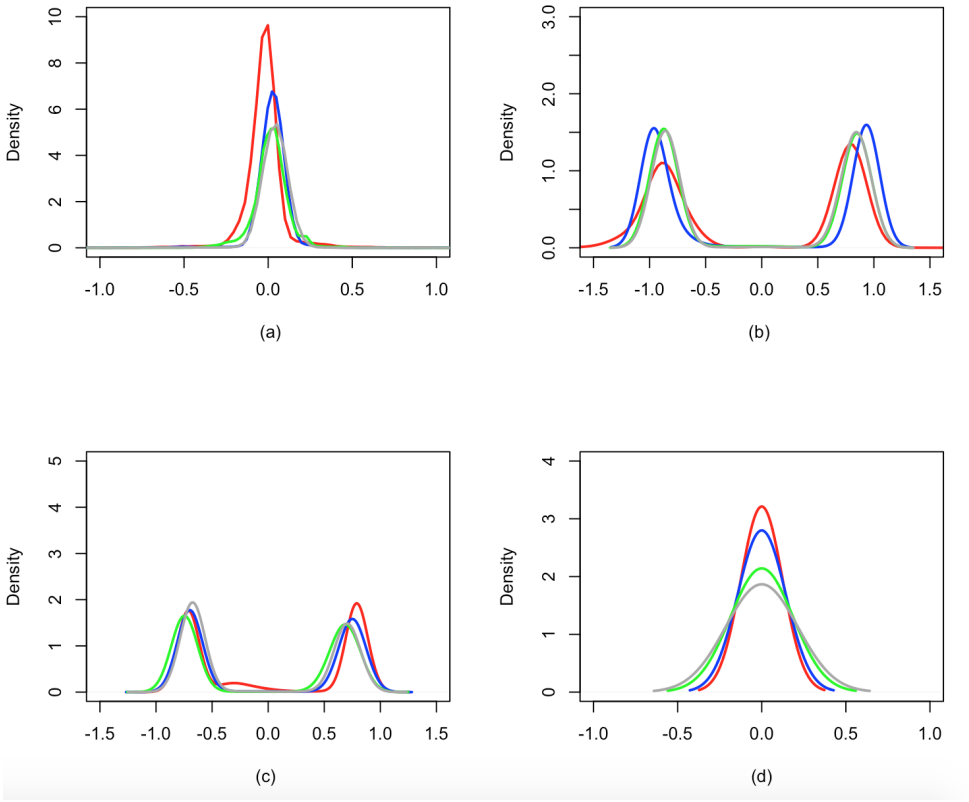


Figure 2.6: Simulation scenario 2: predictive marginal distributions of m_{i0} (a), m_{i1} (b), m_{i2} (c) and m_{i3} (d) obtained specifying a spike and slab prior. Red, blue, green and grey lines correspond to estimates obtained using the dataset with 0%, 10%, 25% and 50% censoring rates, respectively.

2.4.3 Effective Sample Size

To evaluate the computational performance of the two approaches for model selection, we report in Table 2.2 summary statistics for the Effective Sample Size (ESS). MCMC method produce samples that will typically be autocorrelated within a chain. This increases the uncertainty of the estimation of posterior quantities of interest. 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 Kass et al. (1998). Summary statistics are evaluated across the ESS of all estimated parameters. We report results for both simulation scenario 1 and simulation scenario 2 and four different censoring rates. The empirical

Method	Censoring Rate	Scenario 1	Scenario 2
SS	0%	(3678, 3821, 4001)	(3781, 3891, 4104)
	10%	(3599, 3734, 3999)	(3689, 3778, 3911)
	25%	(3602, 3758, 4024)	(3619, 3729, 3815)
	50%	(3548, 3629, 3875)	(3597, 3658, 3795)
POD	0%	(3812, 4017, 4109)	(3929, 4074, 4182)
	10%	(3728, 3982, 4099)	(3826, 3994, 4106)
	25%	(3702, 3858, 3907)	(3765, 3897, 3969)
	50%	(3698, 3795, 3951)	(3703, 3847, 3965)

Table 2.2: 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.

distributions of the ESS in Table 2.2 seem comparable across different prior specifications and scenarios.

2.4.4 Comparison with the Time-dependent Frailty Model

In this Subsection, we compare our model with the *time-dependent frailty* model of Manda and Meyer (2005) 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:

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

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 Manda and Meyer (2005) 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 $(\phi+1)^2(1-$

$\phi)^2\mathbb{I}_{(-1,1)}(y)$), and it measures the degree of serial correlation in the subject-specific frailty. The prior specification is completed as follows:

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

Also for this model posterior inference can be performed through a standard Gibbs sampler algorithm, which we implement in JAGS (Plummer, 2003). For each simulated scenario, 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 2.3 reports the posterior mean and standard deviation of the frailty correlation parameter ϕ, for the first simulation scenario, described in Subsection 2.4.1; Table 2.4 reports similar summaries of ϕ for the second simulation scenario (see Subsection 2.4.2).

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

Table 2.3: Simulation scenario 1: posterior mean and standard deviation of the frailty correlation parameter ϕ in Manda and Meyer (2005).

From Table 2.3 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..

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

Table 2.4: Simulation scenario 2: posterior mean and standard deviation of the frailty correlation parameter ϕ in Manda and Meyer (2005).

On the other hand, the posterior mean of ϕ in the second scenario is centered around 0. This shows the limitations of a parametric approach as obser-

vations 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 Figures 2.2 and 2.5, 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 2.2, 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.

2.4.5 Simulation scenario 3

In order to evaluate the performance of our model described above, we conduct a more complex simulation study, in which there are four groups of patients. In this case $L = 400$. Again, 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 fourth of the observations are generated from

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

while another fourth is generated from

$$\begin{aligned} Y_{i1} \sim \mathcal{N}(0, (1.5)^2), \qquad & Y_{i2}|Y_{i1} \sim \mathcal{N}(0.9 \times Y_{i1}, (1.5)^2) \\ Y_{ij}|Y_{ij-1}, Y_{ij-2} \ \sim \ & \mathcal{N}(0.9 \times Y_{ij-1} + 0.7 \times Y_{ij-2}, (1.5)^2),\ j = 3, \ldots, n_i \end{aligned}$$

other 100 observations are generated from

$$\begin{aligned} Y_{i1} \sim \mathcal{N}(0, (0.9)^2), \qquad & Y_{i2}|Y_{i1} \sim \mathcal{N}(Y_{i1}, (0.9)^2) \\ Y_{i3}|Y_{i2}, Y_{i1} \ \sim \ & \mathcal{N}(Y_{i2} + 0.7 \times Y_{i1}, (0.9)^2) \\ Y_{ij}|Y_{ij-1}, Y_{ij-2}, Y_{ij-3} \ \sim \ & \mathcal{N}(Y_{ij-1} + 0.7 \times Y_{ij-2} + 0.4 \times Y_{ij-3}, (0.9)^2),\ j = 4, \ldots, n_i \end{aligned}$$

and the last 100 observations are generated from

$$\begin{aligned} Y_{ij} \ \sim \ & \mathcal{N}(0, (0.9)^2),\ j = 1, \ldots, 3 \\ Y_{ij}|Y_{ij-1}, Y_{ij-2}, Y_{ij-3} \ \sim \ & \mathcal{N}(Y_{ij-1} + 0.7 \times Y_{ij-2} + 0.4 \times Y_{ij-3}, (0.9)^2),\ j = 4, \ldots, n_i \end{aligned}$$

We make the same assumptions as for scenarios 1 and 2, e.g. we assume independence across subjects and do not include covariates. We also introduce censoring and consider censoring rates 10%, 25% and 50%. We consider at

least two gap times for each individual and for all the patients i whose last gap time is censored 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 (2.7), (2.11)-(2.12), where we set p = 3 and hyperparameters as in the previous Section. Figure 2.7 shows the predictive distributions of m_{i0}, m_{i1}, m_{i2} and m_{i3} obtained with the spike and slab base measure. Red, blue, green and grey lines correspond to estimates obtained using the dataset with 0%, 10%, 25% and 50% censoring rates, respectively.

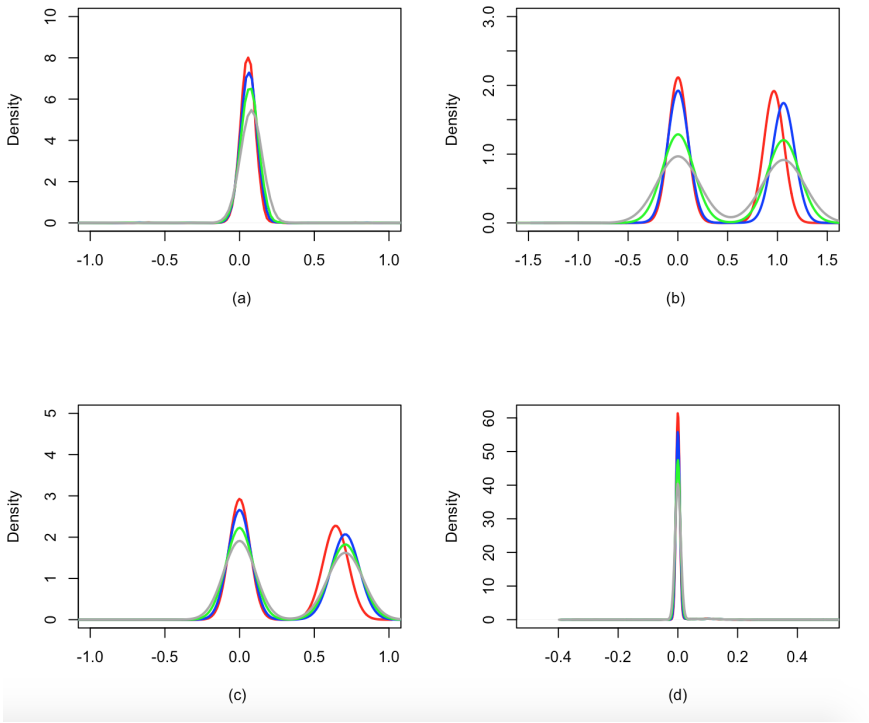


Figure 2.7: Simulation scenario 3: predictive marginal distributions of m_{i0} (a), m_{i1} (b), m_{i2} (c) and m_{i3} (d) under the the spike and slab prior. Red, blue, green and grey lines correspond to estimates obtained using the dataset with 0%, 10%, 25% and 50% censoring rates, respectively.

By visual inspection, it is clear that the results of the predictive distributions of 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 distribution of m_{ij} becomes larger with increasing censoring rate, because we have less information about the population parameters.
We also fit the model (2.7),(2.15) with maximum order of dependence $P = 3$ to this dataset and similar inference is obtained using the prior on the order

of dependence; see Figure 2.8 for predictive marginal distributions of m_{ij}'s. Moreover, Figure 2.9 shows the posterior distribution of K obtained specifying

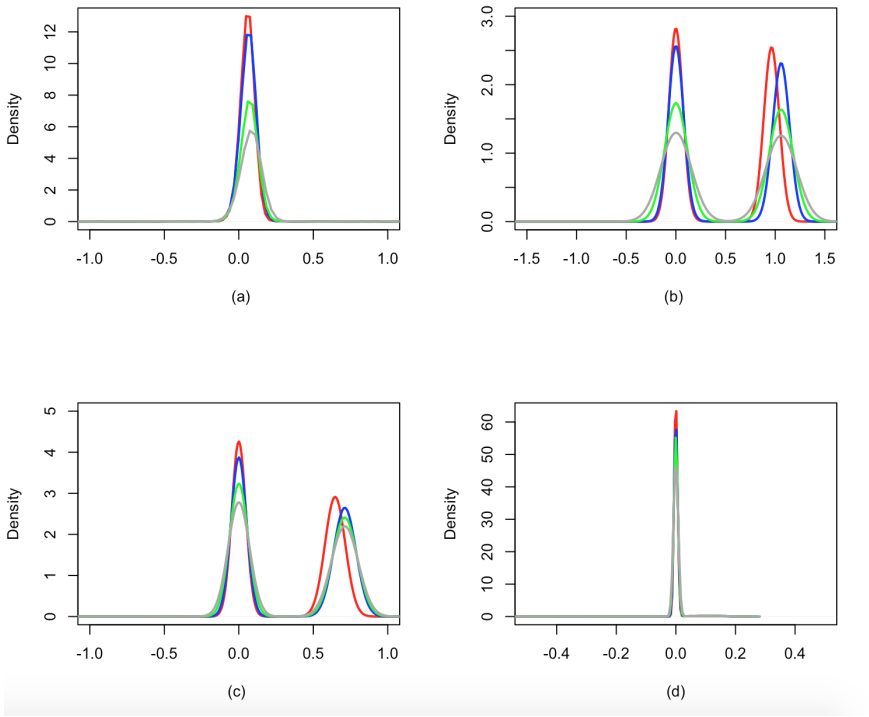


Figure 2.8: Simulation scenario 3: predictive marginal distributions of m_{i0} (a), m_{i1} (b), m_{i2} (c) and m_{i3} (d) obtained specifying a prior on the order of dependence. Red, blue, green and grey lines correspond to estimates obtained using the dataset with 0%, 10%, 25% and 50% censoring rates, respectively.

the prior on the order of dependence, the number of distinct components in the mixture (2.11)-(2.12), using the data with and without censoring. In all the panels, configurations with $K = 4$ clusters always have the highest posterior probability, and hence posterior inference on K is in agreement with the 4 components used to generate the data.

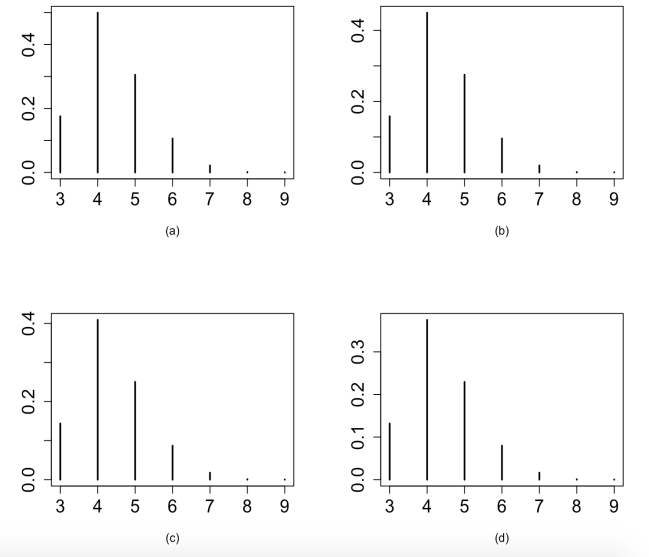


Figure 2.9: Simulation scenario 3: posterior distribution of K obtained using the prior on the order of dependence. The Figure (a), (b), (c), (d), correspond to estimates obtained using the dataset with 0%, 10%, 25% and 50% censoring rates, respectively.

2.4.6 Simulation scenario 4: a mixture of exponential distributions

Finally, we present the results obtained fitting out model to data that are generated from a mixture of Exponential distributions. The subjects are 200 ($L = 200$) and $n_i \sim \mathcal{P}(2.5) + 3$ for all i. The observations are generated from

$$W_{ij} \sim \mathcal{E}(\lambda_{ij}),\ j = 1, \ldots, n_i$$

where for half observations the parameters λ_{ij} are defined as

$$\log \lambda_{ij} = \log 0.5 + 0.7 \times W_{ij-1},\ j = 2, \ldots, n_i$$

and for the last 100 observations we have

$$\log \lambda_{ij} = \log 0.5,\ j = 1, \ldots, n_i.$$

The censoring mechanisms is the same as in the previous Sections. We fit model (2.7), (2.11)-(2.12) with p = 3 and model (2.7),(2.15) with maximum order of dependence $P = 3$. The hyperparameters are chosen in order to specify vague marginal prior distributions. The results obtained without censored data and with the different censoring rates are compared. Figures 2.10

and 2.11 show the marginal posterior predictive distributions of m_{ij}'s obtained with the spike and slab prior and the prior on the order of dependence, respectively. Red, blue, green and grey lines correspond to estimates obtained using the dataset with 0%, 10%, 25% and 50% censoring rates, respectively.

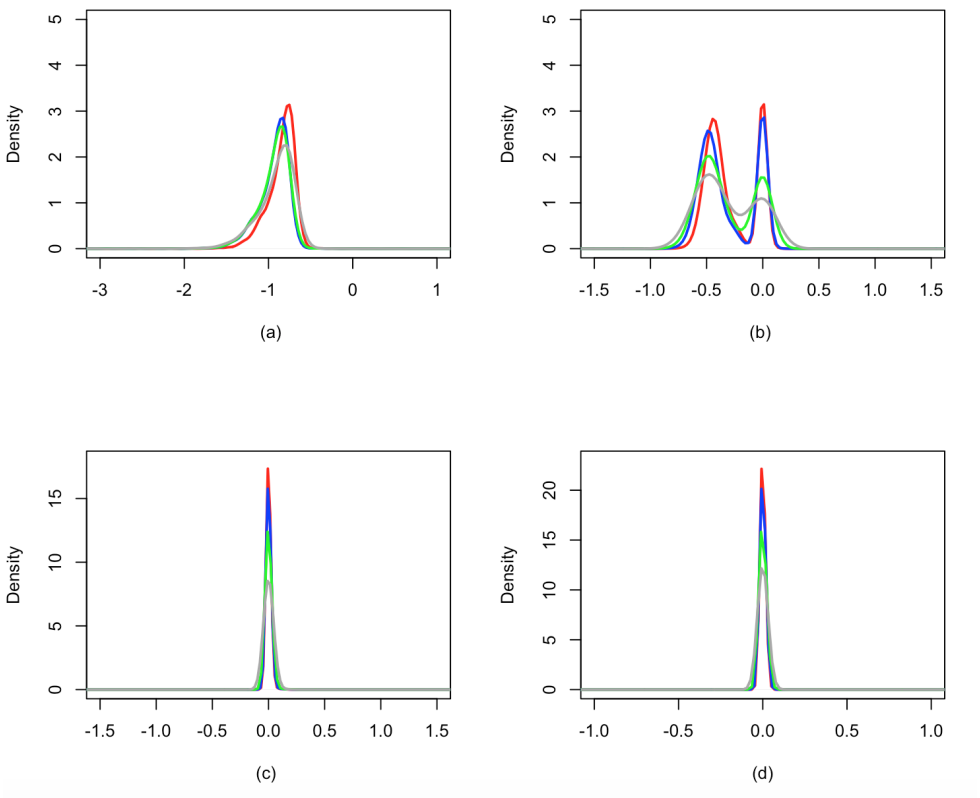


Figure 2.10: Simulation scenario 4: predictive marginal distributions of m_{i0} (a), m_{i1} (b), m_{i2} (c) and m_{i3} (d) obtained specifying the spike and slab prior base measure. Red, blue, green and grey lines correspond to estimates obtained using the dataset with 0%, 10%, 25% and 50% censoring rates, respectively.

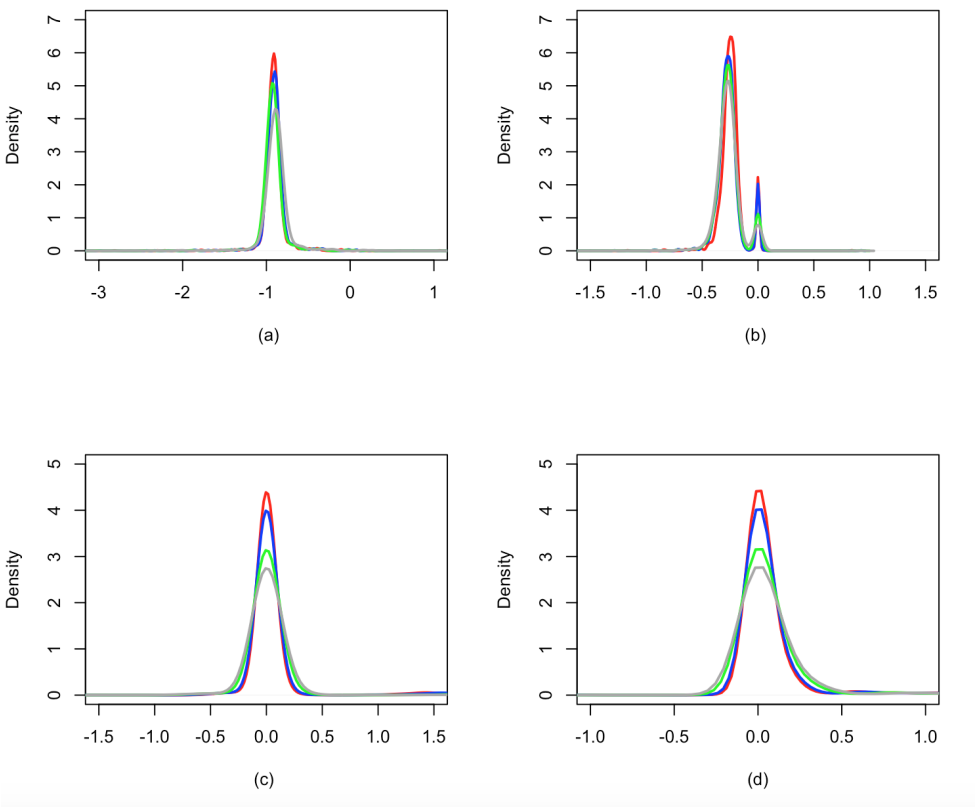


Figure 2.11: Simulation scenario 4: predictive marginal distributions of m_{i0} (a), m_{i1} (b), m_{i2} (c) and m_{i3} (d) obtained using the prior on the order of dependence. Red, blue, green and grey lines correspond to estimates obtained using the dataset with 0%, 10%, 25% and 50% censoring rates, respectively.

By visual inspection, it is clear that the predictive distributions of m_{ij}'s agree with the true values used to create the dataset. In fact, the predictive distributions of m_{i0}, m_{i2} and of m_{i3} are concentrated around 0, while the predictive distributions of m_{i1} is bimodal. Furthermore, the variance of the predictive distributions of m_{ij} become larger with increasing censoring rate because we have less information about the population.
Finally, in the dataset without censoring time, 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 data sets with censoring, the number of clusters increases because of the greater uncertainty on the distributions of gap times.

2.5 Hospitalization dataset

We fit model (2.7)-(2.9) 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 $L = 403$ 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 L = 197 patients and a total number of 495 recurrent events. Table 3.1 reports the number of patients with exactly j gap times, for $j = 1, \ldots, 6$. 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.

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

Table 2.5: *Readmission* dataset: number of patients with exactly j gap times, $j = 1, \ldots, J$.

Prior hyperparameters in (2.10) are set as follows:

$$\beta_0^2 = 1,000$$

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

2.5.1 Testing for the Order of Dependence

When testing the order of dependence, we first fit model (2.7), (2.11)-(2.12) and (2.14) with $p = 3$ (G_0 being the spike and slab distribution) and then model (2.7), but specifying a prior directly on model size as in (2.15) with $P = 3$. Figure 2.12 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.

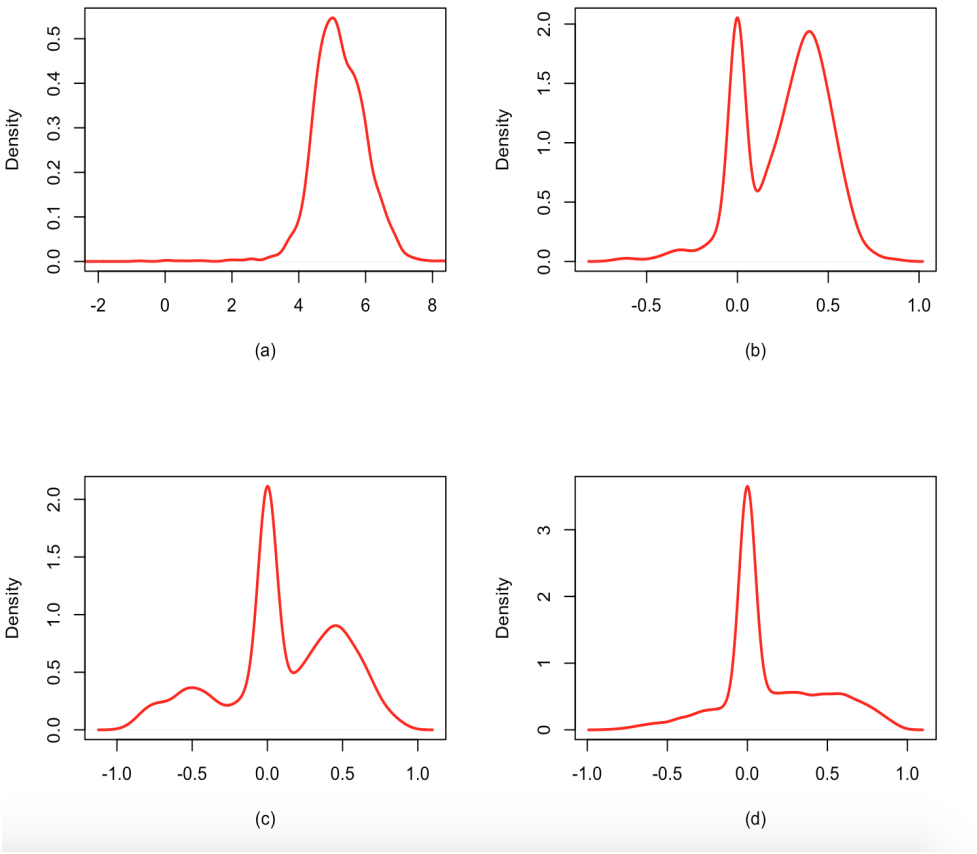


Figure 2.12: *Readmission* dataset: marginal posterior predictive distributions of m_{i0} (a), m_{i1} (b), m_{i2} (c) and m_{i3} (d) under the spike and slab prior.

This result is confirmed also using the approach described in Section 2.3.2.

Indeed, the posterior distribution of p, displayed in Figure 2.13, places most of its mass on 2.

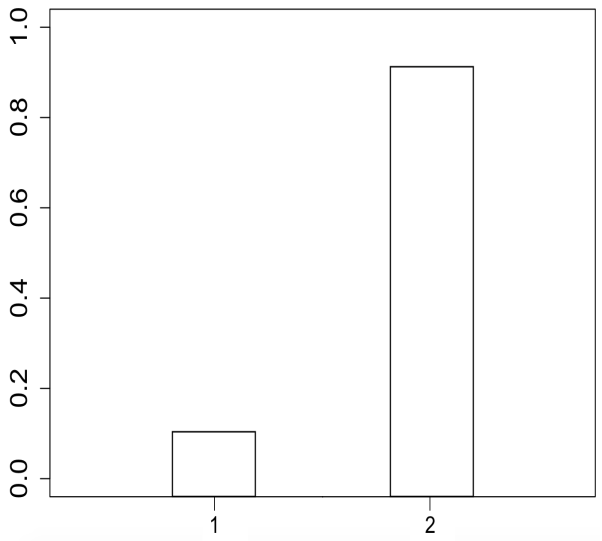
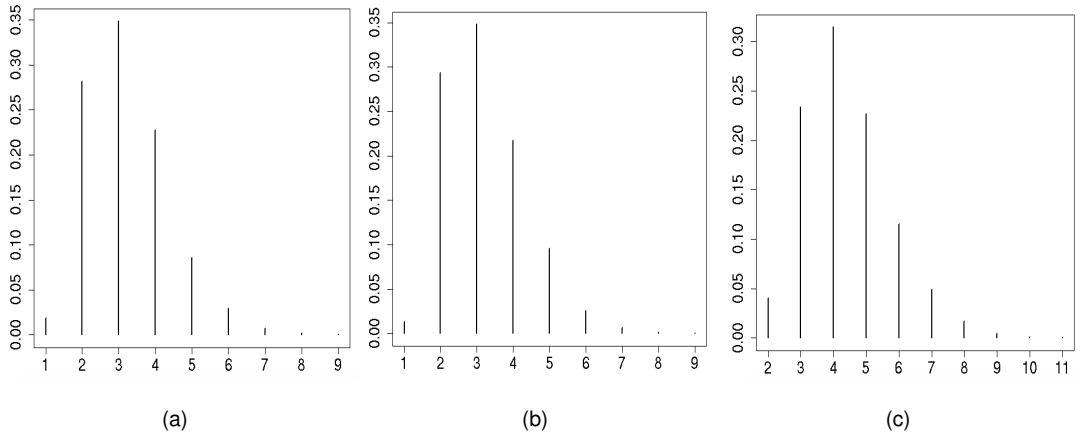


Figure 2.13: *Readmission* dataset: posterior distribution of p.

2.5.2 Posterior analysis

We compare now the results of the nonparametric AR(2) model for the random effects α_{ij}'s as in (2.11)-(2.13), selected in the previous Section, with models (2.8)-(2.9) built using different choices of g. In particular we consider two summary statistics: $g(Y_{i1}, \ldots, Y_{ij-1}) = (Y_{i1} + \cdots + Y_{ij-1})/(j-1)$ and $g(Y_{i1}, \ldots, Y_{ij-1}) = (Y_{i1} \times \cdots \times Y_{ij-1})^{1/(j-1)}$. 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 (2.8). We analyze the posterior distribution of K, the number of components in the mixture (2.11)-(2.12) 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 2.14).



(a) (b) (c)

Figure 2.14: *Readmission* dataset: posterior distribution of K, with $g(Y_{i1}, \ldots, Y_{ij-1}) = (Y_{i1} + \cdots + Y_{ij-1})/(j-1)$ (a) and $g(Y_{i1}, \ldots, Y_{ij-1}) = (Y_{i1} \times \cdots \times Y_{ij-1})^{1/(j-1)}$ (b). Panel (c) displays the posterior distribution of K using the AR(2) model.

In Figure 2.15 we present the marginal posterior predictive distributions of Y_j^{new} for a hypothetical new subject, for each time j, $j = 1, \ldots, 6$, 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.

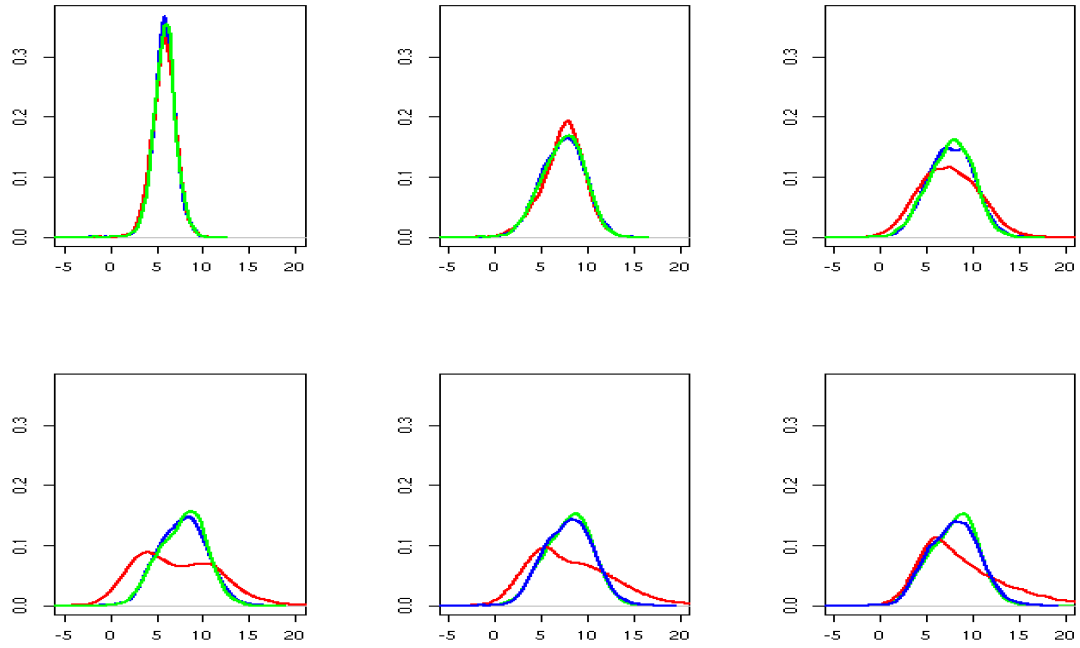


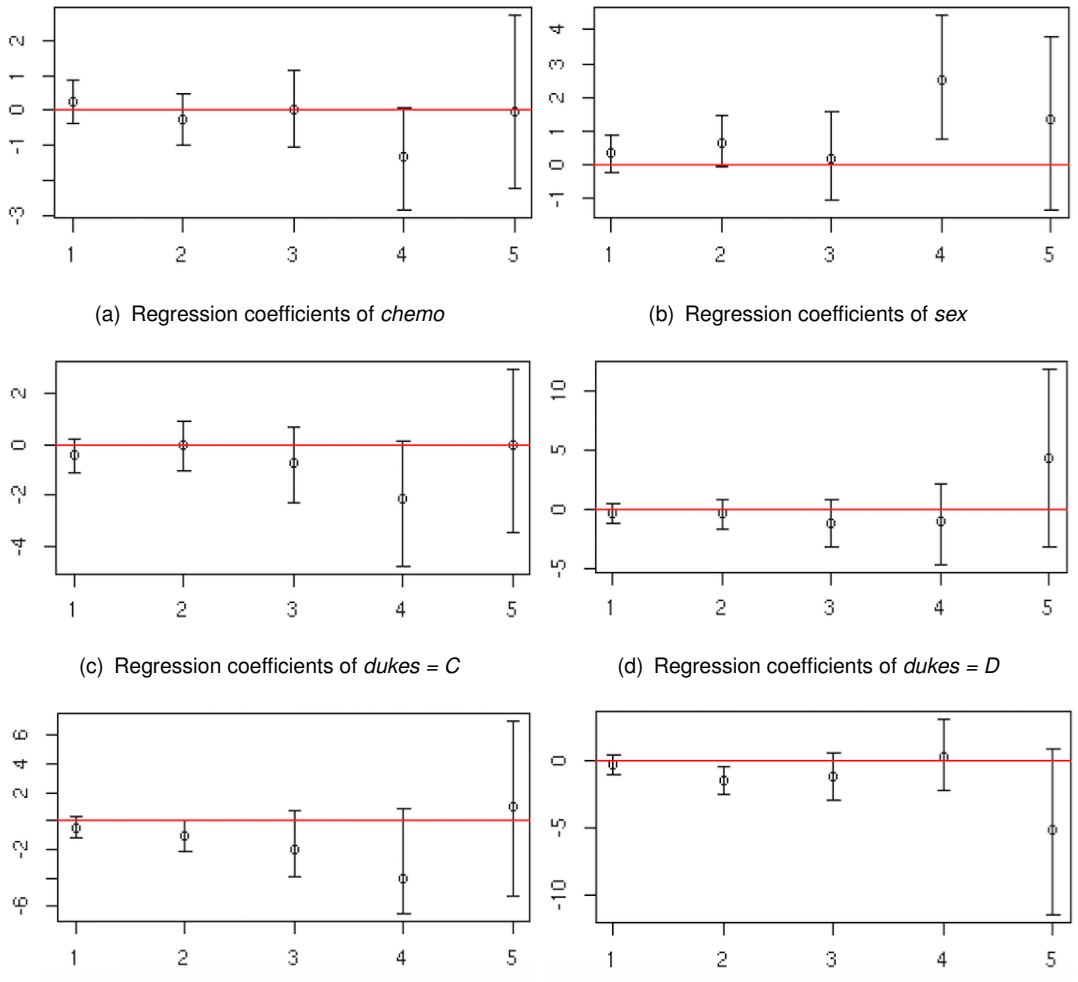
Figure 2.15: *Readmission* dataset: posterior predictive distributions of Y_j^{new}, $j = 1, \ldots, 6$ for a hypothetical new subject, with covariates equal to the empirical mode of all covariates in the sample. The green and blue lines represent AR(1)-type models, with $g(Y_{i1}, \ldots, Y_{ij-1}) = (Y_{i1} + \cdots + Y_{ij-1})/(j-1)$ and $g(Y_{i1}, \ldots, Y_{ij-1}) = (Y_{i1} \times \cdots \times Y_{ij-1})^{1/(j-1)}$, respectively, and the red distribution indicates AR(2) model.

2.5.3 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, $j = 1, \ldots, 6$. 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 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6})$ =(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, $j = 1, \ldots, J$ $(= 6)$, is therefore 6-dimensional.

Figure 2.16 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.



(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*

Figure 2.16: *Readmission* dataset: 95% posterior credible intervals for the regression parameters of each covariate across the first five gap times.

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.

2.5.4 Comparison with *shared frailty* model

In this Subsection, we compare our model with the frequentist *shared frailty* model of Rondeau et al. (2003) as implemented in the R package `frailtypack` (Rondeau et al., 2012a). 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 ($j = 1, \ldots, n_i$) and i-th individual ($i = 1, \ldots, L$), conditional on the frailty parameter ω_i, is

$$h_{ij}(t|\omega_i) = h_0(t)\omega_i e^{\boldsymbol{\beta}^T \boldsymbol{x_{ij}}}$$
$$\omega_i \overset{\text{iid}}{\sim} \Gamma\left(\frac{1}{\theta}, \frac{1}{\theta}\right)$$

where $h_0(t)$ is the baseline hazard function, $\boldsymbol{\beta}$ is the vector of the regression coefficients associated to the covariate vector $\boldsymbol{x}_{ij}$. Table 2.6 reports the estimates of the regression coefficients and the associated p-value.

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

Table 2.6: *Readmission* dataset: estimates of the regression coefficients and associated p-values in the *shared frailty* model of Rondeau et al. (2003).

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.

2.5.5 Comparison with Bayesian semiparametric dynamic frailty models

In this Section, we compare our model with the Bayesian semiparametric dynamic frailty models of Pennell and Dunson (2006) using the *readmission* dataset in the R package *frailtypack*. This model generalizes the shared frailty model to allow time-varying frailties and regression coefficients with a multiplicative parameterization to introduce autocorrelation and to smooth the time trajectories.

Differently from our approach, Pennell and Dunson (2006) nonparametrically model the intensity of a count process. In this set up the time axis is partitioned into M intervals, not necessarily equally spaced, and N_{im} denotes the number of events of patient i in interval m. The model has the following structure:

$$N_{im} \overset{\text{ind}}{\sim} \mathcal{P}(r_{im}\lambda_{im})$$

where r_{im} denote the time at risk and λ_{im} is the rate of event occurrence for subject i ($i = 1, \ldots, n$) in the interval m ($m = 1, \ldots, M$), given by

$$\lambda_{im} = \phi_i(\prod_{h=1}^{m} \phi_{ih})\lambda_{0m}\boldsymbol{\beta}^T X_{im}$$

where ϕ_i is a subject-specific shared frailty and ϕ_{ih} is the multiplicative innovation over time interval h. The q-dimensional vector $\boldsymbol{\beta}$ contains unknown parameters associated with the covariate vector X_{im}. The baseline hazard λ_{0m} is defined as follows:

$$\lambda_{0m} = \hat{\lambda}_{0m}\nu_0 \prod_{h=1}^{m} \nu_h,$$

where $\hat{\lambda}_{0m}$ is an initial estimate based on historical data. Following Pennell and Dunson (2006), we assume:

$$\nu_i \overset{\text{iid}}{\sim} \text{Gamma}(1, 1),\ i = 0, 1, \ldots, m$$

where ν_1 should be fixed for computation reasons ($\nu_1 = 1$). Moreover, to improve flexibility, Pennel and Dunson model the frailty distribution nonparametrically using an approach related to the Dirichlet process (DP) mixture models:

$$\phi_1 \mid G_1 \overset{\text{ind}}{\sim} G_1, \quad G_1 \sim DP(M_1, G_{01})$$

$$\phi_{im} \mid G_{m+1} \overset{\text{ind}}{\sim} G_{m+1}, \quad G_{m+1} \sim DP(M_2, G_{02})$$

The prior specification is completed as follows:

$$\begin{aligned} G_{01} &= \text{Gamma}(5, 0.5), \\ G_{02} &= \text{Gamma}(25, 0.5), \\ M_1 &= 5 \\ M_2 &= 5 \end{aligned}$$

Finally we set the prior for the vector of unknown regression parameters as follows:

$$\boldsymbol{\beta_m} \overset{\text{iid}}{\sim} \mathcal{N}_q(0, 1000 I_q)$$

Also for this model the posterior inference can be performed through a standard Gibbs sampler algorithm, which we implement in JAGS (Plummer, 2003), using a truncation-based algorithm for stick-breaking priors (Ishwaran and Zarepour, 2002). For each simulated scenario, 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 frailty model, that models event counts, takes a different approach from our model, that models gap times. Under our setup, comparing the models is difficult. However, changing our distributional assumptions in order to have similar modelling assumptions between these two models, both approaches should lead to the same inference. In the following analysis, we consider the predictive rate of event occurrence for a new subject in the interval m, when (i) the study period is divided in 6 intervals and when (ii) the study period is divided in 3 intervals. Moreover, we present inference on the regression parameters.

Posterior analysis

We partition the study period into 6 intervals ($M = 6$), with each interval having a length of 1 year, and into 3 intervals ($M = 3$) having a length of 2 year. Table 2.7 and 2.8 report the number of events (N_m) in each time window m, for $m = 1, \ldots, 6$ for the first partition, and for $m = 1, \ldots, 3$ for the second partition, respectively.

m	1	2	3	4	5	6
N_m	186	87	59	76	46	7

Table 2.7: Number of events in each time window m, for $m = 1, \ldots, 6$.

m	1	2	3
N_m	273	135	53

Table 2.8: Number of events in each time window m, for $m = 1, \ldots, 3$.

Figure 2.17 shows the predictive rate of event occurrence $\{\lambda_{n+1m}\}$ obtained using the two different partitions of the observation period ($M = 6$ is in red and $M = 3$ is in blue). Dashed lines denote 95% credible intervals (CI), while solid lines represent the posterior mean. By visual inspection, it is clear that the estimates for each time window are similar in both cases. Moreover, uncertainty of posterior estimates depends on the length of the time windows chosen. In fact, the credible interval became smaller for $M = 3$, because the number of events per windows increases.

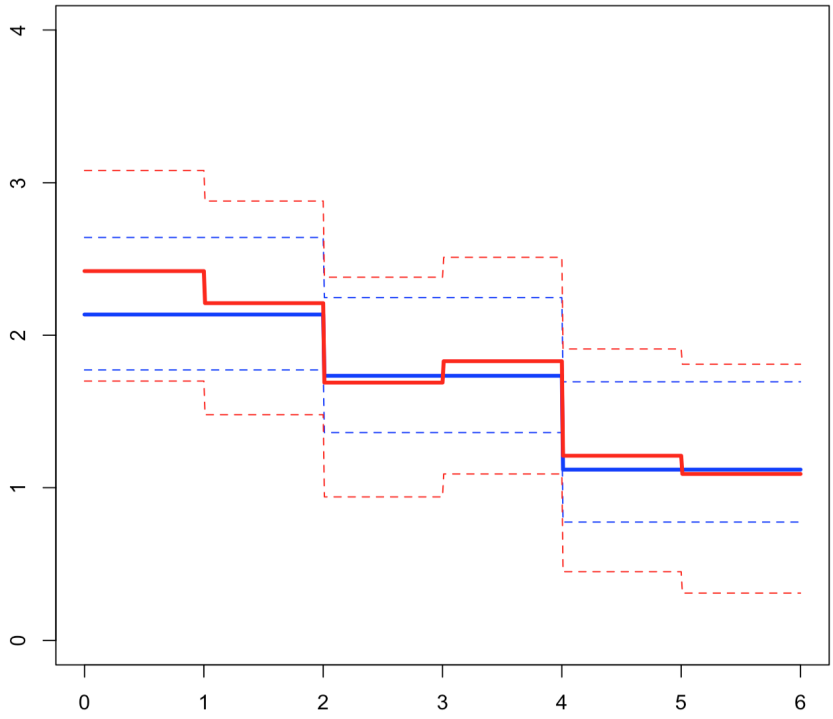
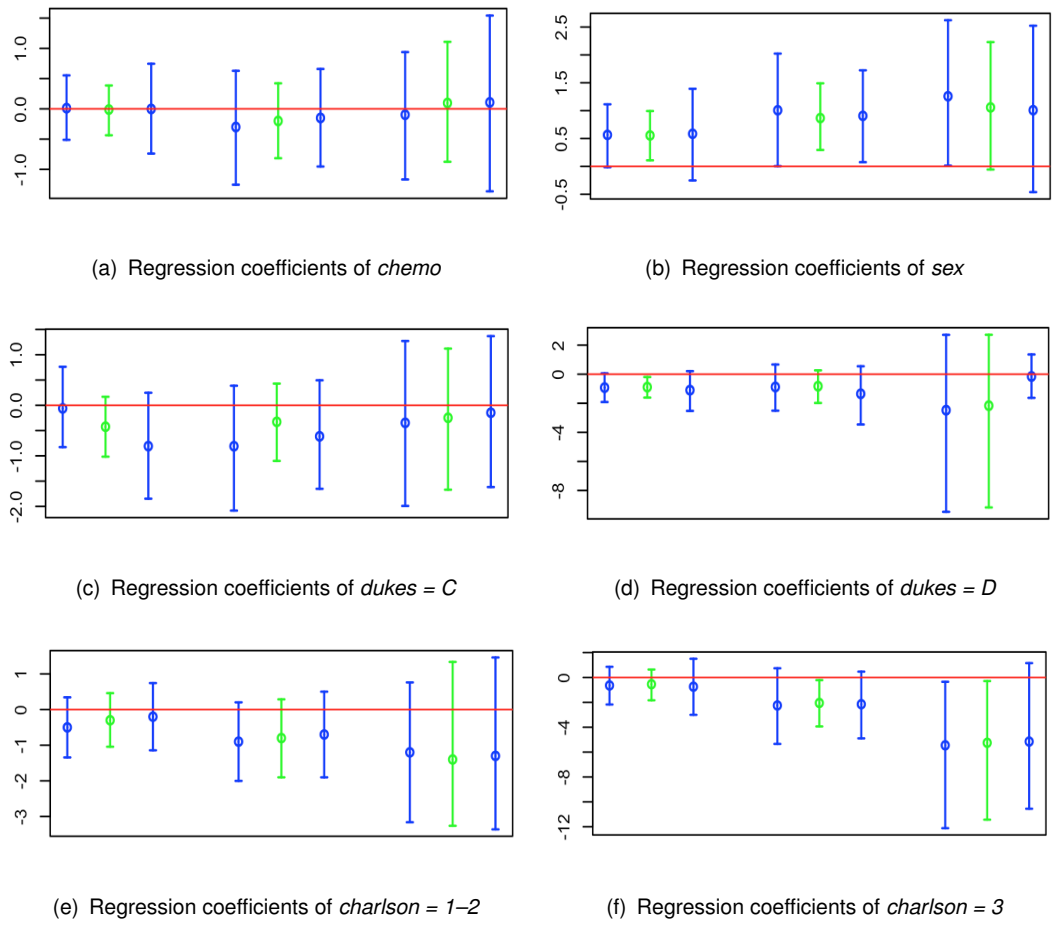


Figure 2.17: Posterior predictive rate of event occurrence $\{\lambda_{n+1m}\}$ for the two different partitions of the study period obtauned fitting the model by Pennell and Dunson (2006). Red lines indicate estimates obtained with $M = 6$, while blue lines indicate estimates obtained with $M = 3$. Dashed lines denote 95% credible intervals (CI), while solid lines represent the posterior mean.

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. Figure 2.18 displays 95% credible intervals for the marginal posterior of the regression parameter of each covariate. In particular the blue lines indicate the credible intervals when we partition the study period into 6 intervals, i.e. of $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}, \boldsymbol{\beta_4}, \boldsymbol{\beta_5}, \boldsymbol{\beta_6}$, while the green lines show the credible intervals when $M = 3$, i.e. of $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}$.



(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*

Figure 2.18: 95% posterior credible intervals of the regression parameters of each covariate across the 6 time windows ($M = 6$) and across the 3 time windows ($M = 3$), respectively, in blue and in green.

The inferential results are in agreement with the results presented in Subsection 2.5.3 in the manuscript. In fact, there is no evident effect of chemotherapy, while $\tilde{\boldsymbol{\beta}}_2$, which measures the effect of sex, is concentrated on positive values. Moreover, the regression coefficient of the indicator variable relative to the Dukes stage C has marginal posterior distribution centred around 0, while the medians of the posterior distribution of the coefficient of the indicator variable corresponding to Dukes stage D are concentrated, in particular in the first

time intervals, on negative values. Finally, the first dummy variables relative to the Charlson Index of the patient has a regression coefficient whose distribution is centred around 0, while the coefficient of the second dummy variable relative to the Charlson Index of the patient exhibits in general negative values over time.

2.6 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 (Khasriya et al. (2010); Kupelian et al. (2013)). 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 $L = 306$ patients. The number of observations with exactly j gap times is displayed in Table 2.9.

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

Table 2.9: *UTI* dataset: number of patients with exactly j gap times, $j = 1, \ldots, J$.

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 2.19 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.

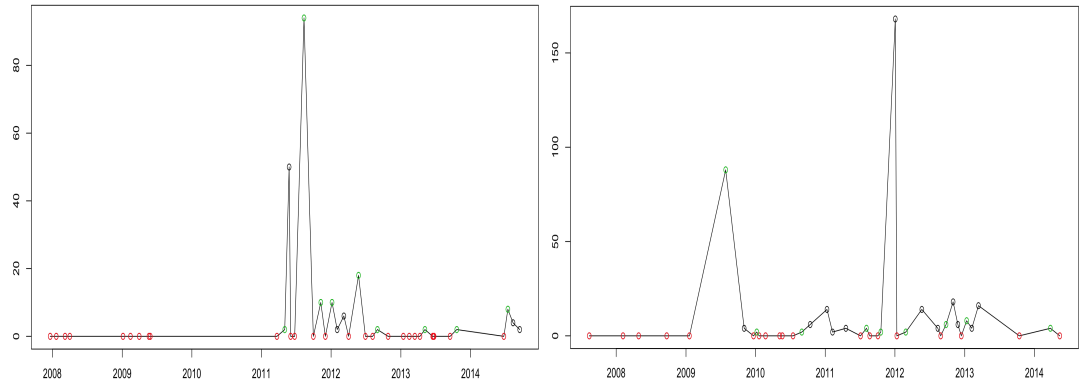


Figure 2.19: *UTI* dataset: recurrent events for two patients; the last gap time of the patient on the left is observed, while that of the patient on the right is censored. The vertical axis reports the value of WBC in the urine, red circles denote zero WBC at the visit while green circles denote WBC greater than 0.

We fit model (2.7), 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 (=1 if the symptom is present, 0 otherwise). Therefore, the final covariate vector for individual i is given by $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$ =(*age*, indicator for urgency, indicator for incontinence, indicator for pain, indicator for voiding). Descriptive statistics of the covariates are given in Table 2.10.

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

Table 2.10: *UTI* dataset: descriptive statistics of the covariates.

In the analysis we set the prior hyperparameters in (2.10) in order to specify vague prior distributions:

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

$$M = 1 .$$

2.6.1 Posterior Inference

We perform inference on the order of dependence using both approaches introduced in Section 2.3. 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 2.20. 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.

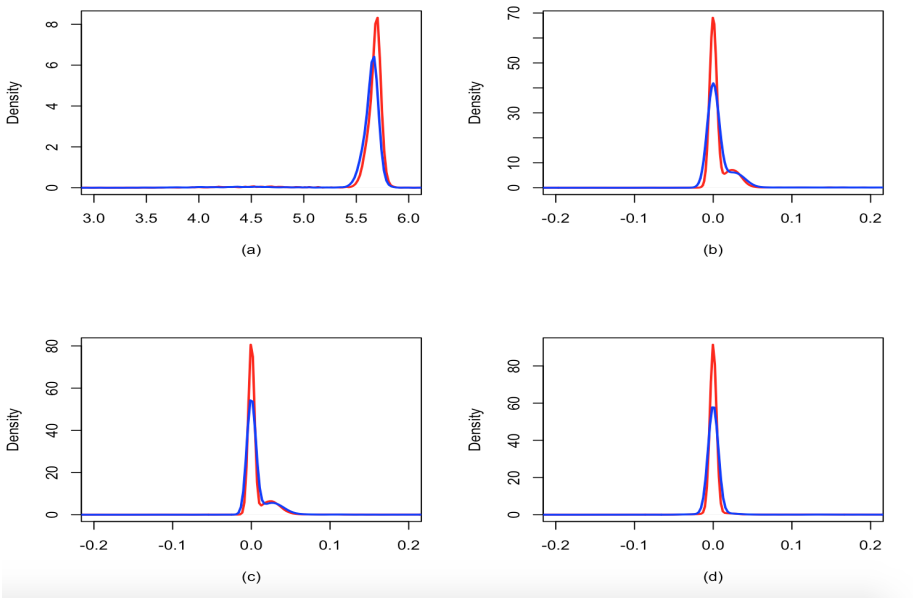


Figure 2.20: *UTI* dataset: marginal posterior predictive distributions of m_{i0} (a), m_{i1} (b), m_{i2} (c) and m_{i3} (d) obtained with the spike and slab base measure (red) and with the prior on the order of dependence (blue).

2.7 Conclusions

In this Chapter 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 a 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. The proposed 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 2.12 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 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 Nieto-Barajas and Quintana, 2016; Di Lucca et al., 2013). In fact, in this case, the data consist in L time series $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$, 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 (2.7) and temporal dependence may be introduced as in (2.8)-(2.9) with appropriate choice of the function $g(\cdot)$.

Since 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 (Sinha et al., 2008b), we extend the proposed model to describe the relationship between recurrent events and survival up to a terminating event. This target can be achieved by specifying a joint distribution of the gap times and event (termination) time. This latter extension is object of Chapter 3.

Chapter 3

Joint modelling of Recurrent Events and Survival: A Transdimensional Approach

Recurrent event times are often censored with dependence between the censoring time and the time between events. For instance, recurrent disease events are censored by the survival time while frailty might affect both disease recurrence and survival. As such, it is important to model the recurrent event process and the event time process jointly to better capture the dependency between them and improve interpretability of the results.

We propose a joint model for the time and the number of recurrent events, as well as for the terminal event.

Typically, the goal is modelling the rate of occurrence to account for variation within and between individuals and at the same time to understand how the recurrence process affects survival. Moreover, the relationship of both event occurrence and survival with potential explanatory factors is often of interest. An important factor in medical applications related to recurrent events and survival time is the overall frailty. Increased frailty is often associated with both increased disease recurrence and reduced survival. Frailty thus induces dependence between survival time and recurrent events. We introduce a model in which the number of recurrences before termination is a random variable of interest. Subsequently, conditionally on this number, we specify a joint distribution for recurrence and survival. This innovative conditional approach brings dependence between recurrence and survival. Posterior Inference is provided by a tailor-made Gibbs sampler involving reversible jump and slice sampling. We illustrate (show) our model on atrial fibrillation data and compare the performance with existing approaches.

In particular, in Section 3.1 the basic theory underlying survival analy-

sis is presented, as well as a review of existing Bayesian semiparametric approaches to dependence between recurrence and survival time. In Section 3.2 we introduce our model, we specify the prior and we explain how to perform posterior inference.Section 3.3 checks its performance on simulated data. Section 3.4 discusses an application to atrial fibrillation data. Section 3.5 compares our approach with existing ones. Section 3.6 concludes the Chapter.

3.1 Survival analysis

In this Section we consider the analysis of time-to-event data. We mention Christensen et al. (2011) for a comprehensive presentation of the main models and methods. Survival analysis is the term used to describe the analysis of time-to-event data in biological and medical contexts. Reliability analysis is often used for non-biological applications. Examples of this kind of data include: (i) the time until death after diagnosis with leukaemia; (ii) the time it takes to get sick after infection with a virus; (iii) the time until a machine breaks down after being installed. The most common goal of survival analysis is to compare survival prospects among different populations. Time-to-event data are distinguished by two features:

- they are positive, i.e. the random variable representing the survival times of individuals is a non-negative random variable. Moreover, time-to-event data are often skewed so we would need to take a log transformation before analyzing them;

- they are often censored (i.e. partially observed). We often know that a unit (a person, a machine) was operative (alive, working) up to a certain time but do not know exactly when it failed or would fail.

In this work, we consider time-to-event data as continuous random variables. As said in the previous Chapter, there are two main statistical approaches to inference on recurrent events: 1) modelling the intensity or hazard function of the event counts process and (2) modelling the sequence of times between recurrent events, known as gap times or waiting times (Cook and Lawless, 2007b). The first approach is most suitable when individuals frequently experience the recurrent event of interest and the occurrence does not alter the process itself. Here, we mention some examples that consider the dependence between recurrence and survival time. Liu et al. (2004), Rondeau et al. (2007), Ye et al. (2007), Huang et al. (2010), Sinha et al. (2008a) and Ouyang

et al. (2013) model the intensity of the recurrent events and the survival time. The latter two approaches propose Bayesian methods with an emphasis on modelling the risk of death and the risk of rejections for heart transplantation patients. Olesen and Parner (2006), Huang and Liu (2007), Yu and Liu (2011), Bao et al. (2012) and Liu et al. (2015) model the hazard function of the recurrent events and of the survival jointly, with the recurrent events and the survival being independent conditionally on the frailty parameters. Li et al. (2018, 2019) use a copula inside a Bayesian hazards model to allow for dependence between recurrence and survival conditionally on the frailty parameters. Yu et al. (2013) model the intensity of the recurrent events and the hazard function of survival jointly while assuming independent censoring before death.

The second approach, which focuses on the sequence of gap times, is more appropriate when the recurrent events are relatively infrequent, when individual renewal takes place after an event, or when the goal is prediction of the time to the next event. Therefore, this approach is highly relevant for biomedical applications. For instance, major recurrent cardiac events for one patient are often rather infrequent from a statistical viewpoint. Also, healthcare planning can benefit from time-to-event predictions, especially if events require hospitalization. Nonetheless, there is less existing work on the second approach than on the first.

This work places itself within the second framework, as the events in our application are infrequent but measured on many individuals. In this context, Paulon et al. (2018) proposes Bayesian non-parametric models for the gap times, with also considering dependence with survival time. We build on this previous work with some important differences. We propose an autoregressive process for gap times which has a constant mean as includes regression coefficients which would otherwise be hard to interpret. Unlike Paulon et al. (2018), we do not assume independence of gap times conditionally on random effects, and we have separate random effects for gap and survival times, enabling greater flexibility in capturing dependence or lack thereof between recurrence and survival.

The main novel methodological contribution of this work is how it incorporates inference on the number of recurrent events. Previous work (Olesen and Parner, 2006; Huang and Liu, 2007; Paulon et al., 2018) also considers censoring by a terminal event such as survival. In that work, a large number of recurrent events are assumed to exist for each individual with the recurrent event process defined also after the terminal event has occurred. Then, the gap times are censored either by the survival censoring time or by the actual

survival time. The contribution to the likelihood of the gap times after censoring is set equal to one such that the assumed large number of censored gap times does not affect the inference. On the other hand, we adopt a conditional approach which does not require the often unrealistic assumption of a large arbitrary number of recurrent events or the continuation of the recurrence process beyond the terminal event. More in detail, we explicitly model the number of recurrent events before the terminal event. Then, conditionally on the number of events, we specify a joint distribution for gap times and survival. To the best of our knowledge, this modelling strategy has never been employed in the context of joint modelling of survival and recurrence processes. This could be counter-intuitive as the number of events occurs after the gap times and is correlated with the gap times. However, in our model, the number of recurrent events could be censored if the terminal event is censored. As a consequence, the distribution of the unobserved gaptime is correlated to the observed gap time but we think this is a strength of the model as it allows borrowing information. The resulting posterior inference is computationally more challenging than the models proposed in the available literature and it is performed by the development of a Markov chain Monte Carlo algorithm which relies on various computational tools such as reversible jump Markov chain Monte Carlo (Green, 1995) and slice sampling (Neal, 2003). Our explicit modeling of the number of recurrent events not only yields a more intuitive model specification: it also allows capturing of the dependence structure between the recurrence and survival processes, which is important in medical applications and beyond. Subject-specific random effects describe the frailty by informing both the survival time and the dependence of subsequent gap times. The random effects are modelled flexibly with a Dirichlet process (DP, Ferguson, 1973b) prior as in the model of the previous Chapter. This choice allows for extra flexibility, variability between individual trajectories, overdispersion and clustering of observations, and overcomes the often too restrictive assumptions underlying a parametric random effects distribution. We introduce different random effects parameters, one for the recurrence process and one for the survival. We model these jointly using a DP prior, ensuring dependence between recurrence and survival. Additionally, we specify an autoregressive model for the gap times to capture the dependence between subsequent gap times. Following Paulon et al. (2018), in order to capture the strength of the relationship between gap and survival times, a possible extension of our model are joint models with shared parameters, commonplace in the literature for joint modelling of survival and longitudinal data. In this framework, by using frailty terms, it is possible to connect the likelihood for survival

and gap times. Therefore, it is possible to estimate the variance of the frailty and thus the degree of within-subject correlation of gap times, as well as the effect of disease recurrencens on survival.

3.2 Model

3.2.1 Notation

Here, where is possible, we use the same notation of Chapter 2. We consider data on L individuals. Let T_{i0} denote the start time of the recurrent event process for individual i. We assume $T_{i0} = 0$ for $i = 1, \ldots, L$. Let S_i denote the survival time for individual i since the start of the corresponding event process. Each individual i experiences N_i recurrent events over the time interval $(0,\, S_i]$. Let T_{ij} denote the jth event time for individual i. Then, the last event time T_{iN_i} is less than or equal to S_i.

Some event processes are right censored due to end of study, as in the application of Section 3.4, or loss to follow-up. We assume completely independent censoring. This contrasts with the survival time S_i which our model allows to depend on the event process. Let c_i denote the minimum of the censoring time and the survival time S_i for individual i, who is thus observed over the interval $(0,\, c_i]$. Let $n_i \leq N_i$ denote the number of events that are observed over the interval $(0,\, c_i]$. Either S_i or the censoring time is observed. If S_i is observed, then $N_i = n_i$ and $0 < T_{i1} < \cdots < T_{iN_i} \leq S_i = c_i$. If S_i is not observed, then $N_i \geq n_i$ and $S_i > c_i$ are unknown and object of inference. In this case, $0 < T_{i1} < \cdots < T_{in_i} \leq c_i < T_{i(n_i+1)} < \cdots < T_{iN_i} \leq S_i$. We define the log gap times as

$$Y_{ij} = \log(T_{ij} - T_{i(j-1)}), \tag{3.1}$$

for $j = 1, \ldots, N_i$. The q-dimensional vector $\boldsymbol{x}_i$ contains individual-specific covariates.

3.2.2 Likelihood specification

Firstly, we assume that the number of gap times N_i follows a Poisson distribution with rate parameter λ that is truncated by $N_i \geq 1$: $N_i \sim \mathbb{I}_{[1,\infty)}\,\text{Poisson}(\lambda)$, independently for $i = 1, \ldots, L$. Conditionally on N_i, we specify a joint model for the log gap times $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iN_i})^T$ and the survival time S_i. We define the joint density $p(\boldsymbol{Y}_i, S_i \mid —) = p(\boldsymbol{Y}_i \mid S_i, —)\, p(S_i \mid —)$ by specifying the conditional densities $p(\boldsymbol{Y}_i \mid S_i, —)$ and $p(S_i \mid —)$. This induces dependence among N_i, $\boldsymbol{Y}_i$ and S_i in a principled manner. We build on previous model (Chapter

2) by assuming that the gap times and survival times follow log-normal distributions where the pairs $(\boldsymbol{Y}_i, S_i)$ are mutually independent for $i = 1, \ldots, L$, conditionally on the random effects, number of recurrent events N_i and the other parameters in the model.

The random effects parameter for the gap times is the two-dimensional vector $\boldsymbol{m}_i$ which characterizes an autoregressive model for $\boldsymbol{Y}_i$:

$$p(\boldsymbol{Y}_i \mid \boldsymbol{\beta}, \boldsymbol{m}_i, N_i, S_i, \sigma^2, \boldsymbol{x}_i) \propto \mathcal{N}(Y_{i1} \mid \boldsymbol{x}_i^T\boldsymbol{\beta} + m_{i1},\, \sigma^2) \times \prod_{j=1}^{N_i} \mathcal{N}\{Y_{ij} \mid \boldsymbol{x}_i^T\boldsymbol{\beta} + m_{i1} + m_{i2}(Y_{i(j-1)} - \boldsymbol{x}_i^T\boldsymbol{\beta} - m_{i1}),\, \sigma^2\}, \quad (3.2)$$

for $T_{iN_i} \leq S_i$, where $T_{iN_i} = \sum_{j=1}^{N_i} e^{Y_{ij}}$ per (3.1) and the q-dimensional vector $\boldsymbol{\beta}$ consists of covariate effects on the gap times. This resembles the autoregressive model on the random effects in Equation (2.8) of Chapter 2. Two main differences are due to the fact that the the mean of Y_{ij} is the same for all j conditionally on the remaining parameters in our model and that Chapter 2 do not consider a survival process, which implies the truncation $T_{iN_i} \leq S_i$ in our work. The truncation results from our conditioning on N_i whereas existing literature (Paulon et al., 2018; Aalen and Husebye, 1991) specifies the likelihood as a joint distribution of the number of events n_i observed over the interval $(0,\, c_i]$ and their log gap times $Y_{i1}, \ldots, Y_{in_i}$. The regression coefficient $\boldsymbol{\beta}$ in (3.2) has the usual interpretation since the mean of Y_{ij} equals $\boldsymbol{x}_i^T\boldsymbol{\beta} + m_{i1}$ for all j.

The distribution of the log survival time is Gaussian:

$$\log(S_i) \sim \mathcal{N}(\boldsymbol{x}_i^T\boldsymbol{\gamma} + \delta_i,\, \eta^2), \quad (3.3)$$

independently for $i = 1, \ldots, L$, where the q-dimensional vector $\boldsymbol{\gamma}$ consists of covariate effects on the survival time and δ_i denotes a random effects parameter. Covariate effects can differ between gap and survival times, for instance if a therapy delays disease recurrence but does not prolong survival. Therefore, the model on the gap times in (3.2) and on the survival times in (3.3) have distinct regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively.

3.2.3 Prior specification

We specify a non-parametric prior for the random effects parameters $\boldsymbol{m}_i$ and δ_i in (3.2) and (3.3). In more detail, $(\boldsymbol{m}_i,\, \delta_i) \sim G$ independently for $i = 1, \ldots, L$, where $G \sim \text{DP}(M,\, G_0)$, with $M \sim \text{Gamma}(a_M,\, b_M)$ and base measure $G_0 = \mathcal{N}_2(\boldsymbol{0}_{2\times 1},\, \sigma_m^2\boldsymbol{I}_2) \times \mathcal{N}(0,\, \sigma_\delta^2)$. Finally, the prior distributions

on the remaining parameters are $\boldsymbol{\beta} \sim \mathcal{N}_q(\mathbf{0}_{q\times 1},\, \sigma^2_\beta\, \boldsymbol{I}_q)$, $\boldsymbol{\gamma} \sim \mathcal{N}_q(\mathbf{0}_{q\times 1},\, \sigma^2_\gamma\, \boldsymbol{I}_q)$, $\sigma^2 \sim$ Inv-Gamma$(\nu_{\sigma^2}/2,\, \nu_{\sigma^2}\, \sigma^2_0/2)$, $\eta^2 \sim$ Inv-Gamma$(\nu_{\eta^2}/2,\, \nu_{\eta^2}\, \eta^2_0/2)$ and $\lambda \sim$ Gamma$(a_\lambda,\, b_\lambda)$.

3.2.4 Posterior inference

Posterior inference is performed through a Gibbs sampler algorithm. This includes imputing N_i, $Y_{i(n_i+1)}, \ldots, Y_{iN_i}$ and S_i for each censored individual i. The Gibbs update for N_i and $\boldsymbol{Y}_i$ is transdimensional since the number of events N_i represents the dimensionality of the sequence of log gap times $\boldsymbol{Y}_i$. This requires devising a reversible jump sampler (Green, 1995) for N_i and $\boldsymbol{Y}_i$. The reversible-jump Markov chain Monte Carlo is an extension to standard Markov chain Monte Carlo methodology that allows simulation of the posterior distribution on spaces of varying dimensions.Thus, the simulation is possible even if the number of parameters in the model is not known. Most full conditional distributions are intractable due to the truncation $T_{iN_i} \leq S_i$, which for instance causes the normalization constant in (3.2) to depend on parameters of interest. We use slice sampling (Neal, 2003) to deal with this intractability. Slice sampling originates with the observation that to sample from a univariate distribution, we can sample points uniformly from the region under the curve of its density function and then look only at the horizontal coordinates of the sample points. A Markov chain that converges to this uniform distribution can be constructed by alternately sampling uniformly from the vertical interval defined by the density at the current point and from the union of intervals that constitutes the horizontal "slice" though the plot of the density function that this vertical position defines. To sample from a multivariate distribution, such single-variable slice sampling updates can be applied to each variable in turn. The normalization constant in (3.2) is also intractable. We therefore approximate it using the Fenton-Wilkinson method (Fenton, 1960). This method is based on the fact that the sum of lognormal distributions can be accurately approximated by a single lognormal. Algorithm 8 from Neal (2000b) is implemented to sample the DP parameters $(\boldsymbol{m}_i,\, \delta_i)$. Following Subsection details and derives the Markov chain Monte Carlo algorithms.

3.2.5 Gibbs sampler

This Subsection describes the Gibbs sampler summarized in Algorithm 1.

Algorithm 1 Gibbs sampler

For each iteration of the Gibbs sampler:

1. For $k = 1, \ldots, q$, update β_k and γ_k by slice sampling with (3.8) and (3.11).
2. For each censored individual i:
   (a) Update N_i and $Y_{i(n_i+1)}^{N_i}$ using the reversible jump sampler from Section 3.2.5.
   (b) For $j = n_i+1, \ldots, N_i$, sample Y_{ij} from (3.12) truncated by $T_{i(n_i+1)} > c_i$ and $T_{iN_i} \leq S_i$ using the inverse transformation method.
   (c) Sample S_i from (3.9) truncated by $S_i > \max(c_i, T_{iN_i})$ using the inverse transformation method.
3. Update $\boldsymbol{m}_i$ and δ_i for $i = 1, \ldots, L$ via Algorithm 8 from Neal (2000b) using slice sampling with (3.13–3.15).
4. Sample M from the distribution in Equation 13 from Escobar and West (1995b).
5. Update σ^2, η^2 and λ by slice sampling with (3.16–3.18).

Normalization constant in $p(\boldsymbol{Y}_i \mid \boldsymbol{\beta}, \boldsymbol{m}_i, N_i, S_i, \sigma^2, \boldsymbol{x}_i)$

The distribution in (3.2) is truncated by $T_{iN_i} \leq S_i$ such that the right-hand side of (3.2) is not normalized. The normalization constant depends on parameters of interest and is thus required to be able to sample from the full conditionals of these parameters. This differs from previous likelihood specifications with log-normally distributed gap times (Aalen and Husebye, 1991; Paulon et al., 2018) because $p(\boldsymbol{Y}_i \mid \boldsymbol{\beta}, \boldsymbol{m}_i, N_i, S_i, \sigma^2, \boldsymbol{x}_i)$ is conditional the number of events N_i. The conditioning on N_i requires normalizing the posterior distribution by a constant that depends on some of the parameters.

To derive the normalization constant in (3.2), note that

$$p(\boldsymbol{Y}_i \mid \boldsymbol{\beta}, \boldsymbol{m}_i, N_i, S_i, \sigma^2, \boldsymbol{x}_i) \propto \mathcal{N}_{N_i}(\boldsymbol{Y}_i \mid \boldsymbol{\mu}_{\boldsymbol{Y}_i}, \boldsymbol{\Sigma}_{\boldsymbol{Y}_i}), \quad T_{iN_i} \leq S_i, \tag{3.4}$$

where $\boldsymbol{\mu}_{\boldsymbol{Y}_i} = (\boldsymbol{x}_i^T\boldsymbol{\beta} + m_{i1})\,\mathbf{1}_{N_i \times 1}$ and the $N_i \times N_i$ matrix $\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}$ is defined by its tridiagonal inverse $\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}^{-1}$ with $(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}^{-1})_{jj} = (m_{i2}^2 + 1)/\sigma^2$ for $j = 1, \ldots, N_i - 1$,

$(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}^{-1})_{N_iN_i} = 1/\sigma^2$ and $(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}^{-1})_{j_1j_2} = -m_{i2}/\sigma^2$ for $|j_1 - j_2| = 1$.[1] Consider now the untruncated $\boldsymbol{Y}_i^* \sim \mathcal{N}_{N_i}(\boldsymbol{\mu}_{\boldsymbol{Y}_i}, \boldsymbol{\Sigma}_{\boldsymbol{Y}_i})$ and define $T_{iN_i}^* = \sum_{j=1}^{N_i} e^{Y_{ij}^*}$. Then, the normalization constant in (3.2) equals $\Pr(T_{iN_i}^* \leq S_i)$ where we drop the conditioning on $\boldsymbol{\beta}$, $\boldsymbol{m}_i$, N_i, S_i and σ^2 for notational convenience. $T_{iN_i}^*$ is the sum of log-normal random variables. The distribution of such sums and $\Pr(T_{iN_i}^* \leq S_i)$ have no closed-form expression (Asmussen et al., 2019), requiring us to resort to approximations.

It is infeasible to evaluate $\Pr(T_{iN_i}^* \leq S_i)$ by numerical integration using quadrature for the values of N_i that we encounter. Fortunately, there is a literature on approximating $\Pr(T_{iN_i}^* \leq S_i)$ (Botev et al., 2019, and references therein) which includes deterministic and Monte Carlo methods. As we aim to sample from the full conditionals of $\boldsymbol{\beta}$, $\boldsymbol{m}_i$, N_i, S_i and σ^2 as part of a Gibbs sampler, we need to evaluate the normalization constant $\Pr(T_{iN_i}^* \leq S_i)$ many times, requiring a fast approximation. We therefore choose the Fenton-Wilkinson method (Fenton, 1960) which approximates the distribution of $T_{iN_i}^*$ by a log-normal distribution with matched mean and variance. Asmussen et al. (2019) state that this approximation can be inaccurate for small N_i and when the elements in $\boldsymbol{Y}_i^*$ are dependent. However, their numerical results indicate good performance of the Fenton-Wilkinson method under these circumstances. Other fast approximations such as the saddle-point method from Asmussen et al. (2016) might be much more accurate, though they are also more complex than the Fenton-Wilkinson method.

For any matrix $\boldsymbol{A}$, denote the elementwise exponential by $e^{\boldsymbol{A}}$. Define $\operatorname{diag}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}) = \{(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i})_{11}, \ldots, (\boldsymbol{\Sigma}_{\boldsymbol{Y}_i})_{N_iN_i}\}^T$. For any vector $\boldsymbol{a}$, denote the outer product with itself by $\boldsymbol{a}^{2\otimes} = \boldsymbol{a}\boldsymbol{a}^T$. Then,

$$\begin{aligned}
\mathrm{E}(e^{\boldsymbol{Y}_i^*}) &= e^{\boldsymbol{\mu}_{\boldsymbol{Y}_i} + \operatorname{diag}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i})/2} &&= e^{\boldsymbol{x}_i^T\boldsymbol{\beta} + m_{i1}}\, e^{\operatorname{diag}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i})/2}, \\
\operatorname{Cov}(e^{\boldsymbol{Y}_i^*}) &= \mathrm{E}(e^{\boldsymbol{Y}_i^*})^{2\otimes} \circ (e^{\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}} - \mathbf{1}_{N_i\times N_i}) &&= e^{2(\boldsymbol{x}_i^T\boldsymbol{\beta} + m_{i1})}\, \{e^{\operatorname{diag}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i})/2}\}^{2\otimes} \circ (e^{\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}} - \mathbf{1}_{N_i\times N_i});
\end{aligned}$$

where '$\circ$' denotes the Hadamard product (Halliwell, 2015). Since $T_{iN_i}^* = \mathbf{1}_{1\times N_i}\, e^{\boldsymbol{Y}_i^*}$,

$$\begin{aligned}
\mathrm{E}(T_{iN_i}^*) &= \mathbf{1}_{1\times N_i}\, \mathrm{E}(e^{\boldsymbol{Y}_i^*}) = e^{\boldsymbol{x}_i^T\boldsymbol{\beta} + m_{i1}}\, \operatorname{sum}\{e^{\operatorname{diag}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i})/2}\}, \\
\operatorname{Var}(T_{iN_i}^*) &= \mathbf{1}_{1\times N_i}\, \operatorname{Cov}(e^{\boldsymbol{Y}_i^*})\, \mathbf{1}_{N_i\times 1} \\
&= e^{2(\boldsymbol{x}_i^T\boldsymbol{\beta} + m_{i1})}\, \operatorname{sum}[\{e^{\operatorname{diag}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i})/2}\}^{2\otimes} \circ (e^{\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}} - \mathbf{1}_{N_i\times N_i})];
\end{aligned} \tag{3.5}$$

where $\operatorname{sum}(\cdot)$ denotes the grand sum of a matrix which is the sum of all its elements. The Fenton-Wilkinson approximation to the distribution of $T_{iN_i}^*$ is a log-normal distribution with the same mean and variance. Let $\hat{T}_{iN_i}^*$ be distributed

[1] This implies $(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i})_{j_1j_2} = (m_{i2}^{|j_1-j_2|} - m_{i2}^{j_1+j_2})/(1-m_{i2}^2)$ for $j_1, j_2 = 1, \ldots, N_i$.

according to this log-normal distribution. Then, $\log(\hat{T}^*_{iN_i}) \sim \mathcal{N}[\log\{\mathrm{E}(T^*_{iN_i})\} - A,\, 2A]$ with

$$\begin{aligned} A &= \frac{1}{2}\log\left\{1 + \frac{\mathrm{Var}(T^*_{iN_i})}{\mathrm{E}(T^*_{iN_i})^2}\right\} \\ &= \frac{1}{2}\log(\mathrm{sum}[\{e^{\mathrm{diag}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i})/2}\}^{2\otimes} \circ e^{\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}}]) - \log[\mathrm{sum}\{e^{\mathrm{diag}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i})/2}\}] \\ &= \frac{\mathrm{LS}_2}{2} - \mathrm{LS}_1, \end{aligned}$$

where the second equality follows from (3.5) and $\mathrm{sum}\{e^{\mathrm{diag}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i})/2}\}^2 = \mathrm{sum}[\{e^{\mathrm{diag}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i})/2}\}^{2\otimes}]$, and $\mathrm{LS}_1 = \log[\mathrm{sum}\{e^{\mathrm{diag}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i})/2}\}]$ and $\mathrm{LS}_2 = \log(\mathrm{sum}[\{e^{\mathrm{diag}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_i})/2}\}^{2\otimes} \circ e^{\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}}])$ are introduced for notational convenience. Our approximation to the normalization constant is thus

$$\begin{aligned} \Pr(T^*_{iN_i} \leq S_i) \approx \Pr\{\log(\hat{T}^*_{iN_i}) \leq \log(S_i)\} &= \Phi\left[\frac{\log(S_i) - \log\{\mathrm{E}(T^*_{iN_i})\} + A}{\sqrt{2A}}\right] \\ &= \Phi\left\{\frac{\log(S_i) - \boldsymbol{x}_i^T\boldsymbol{\beta} - m_{i1} - 2\,\mathrm{LS}_1 + \mathrm{LS}_2/2}{\sqrt{\mathrm{LS}_2 - 2\,\mathrm{LS}_1}}\right\}, \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative density function of $\mathcal{N}(0,\, 1)$ and the second equality follows from (3.5). In the remainder of this appendix, we write $\Pr(T^*_{iN_i} \leq S_i)$ even though we use this approximation.

LS_1 and LS_2 only depend on $\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}$. Therefore, we only need to recompute LS_1 and LS_2 in the Gibbs sampler when $\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}$, which is a function of m_{i2}, σ^2 and N_i, changes.

Regression coefficients

The full conditional for $\boldsymbol{\beta}$ follows from the prior and the likelihood in Subsection 3.2.2 as

$$p(\boldsymbol{\beta} \mid —) \propto \mathcal{N}_q(\boldsymbol{\beta} \mid 0,\, \sigma^2_\beta \boldsymbol{I}_q) \prod_{i=1}^{L} p(\boldsymbol{Y}_i \mid \boldsymbol{\beta}, \boldsymbol{m}_i, N_i, S_i, \sigma^2, \boldsymbol{x}_i). \tag{3.6}$$

We can use the expression for $p(\boldsymbol{Y}_i \mid \boldsymbol{\beta}, \boldsymbol{m}_i, N_i, S_i, \sigma^2, \boldsymbol{x}_i)$ in (3.2) or (3.4) directly to evaluate (3.6), but that is computationally expensive as it involves a multitude of Gaussian density evaluations. Instead, we introduce $Y^{\boldsymbol{\beta}}_{i1} = Y_{i1} - m_{i1}$, $Y^{\boldsymbol{\beta}}_{ij} = \{Y_{ij} - m_{i1} - m_{i2}(Y_{i(j-1)} - m_{i1})\}/(1 - m_{i2})$ for $j = 2, \ldots, N_i$ and the $N_i \times N_i$ diagonal matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta},i}$ with $\mathrm{diag}(\boldsymbol{\Sigma}_{\boldsymbol{\beta},i}) = \sigma^2\,\{1,\, (1-m_{i2})^{-2},\, \ldots,\, (1-m_{i2})^{-2}\}^T$. Then, we can rewrite (3.2) as

$$p(\boldsymbol{Y}_i \mid \boldsymbol{\beta}, \boldsymbol{m}_i, N_i, S_i, \sigma^2, \boldsymbol{x}_i) \propto \mathcal{N}_{N_i}(\boldsymbol{Y}^{\boldsymbol{\beta}}_i \mid \boldsymbol{1}_{N_i \times 1}\, \boldsymbol{x}_i^T\boldsymbol{\beta},\, \boldsymbol{\Sigma}_{\boldsymbol{\beta},i}), \quad T_{iN_i} \leq S_i.$$

Inserting this expression into (3.6) yields a normal-normal model such that

$$p(\boldsymbol{\beta} \mid \text{—}) \propto \frac{\mathcal{N}_q(\boldsymbol{\beta} \mid \boldsymbol{\mu}^*_{\boldsymbol{\beta}},\, \boldsymbol{\Sigma}^*_{\boldsymbol{\beta}})}{\prod_{i=1}^{L} \Pr(T^*_{iN_i} \leq S_i)}, \tag{3.7}$$

where $\boldsymbol{\mu}^*_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^*_{\boldsymbol{\beta}} \sum_{i=1}^{L} \boldsymbol{x}_i \mathbf{1}_{1\times N_i} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta},i} \boldsymbol{Y}^{\boldsymbol{\beta}}_i$ and $\boldsymbol{\Sigma}^*_{\boldsymbol{\beta}} = \{\boldsymbol{I}_q/\sigma^2_\beta + \sum_{i=1}^{L} \mathrm{sum}(\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta},i})\, \boldsymbol{x}_i \boldsymbol{x}_i^T\}^{-1}$ with $\mathrm{sum}(\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta},i}) = \{1 + (N_i - 1)(1 - m_{i2})^2\}/\sigma^2$. Recalling the conditional distributions of a multivariate normal, we obtain

$$p(\beta_k \mid \text{—}) \propto \frac{\mathcal{N}(\beta_k \mid \mu^*_{\beta_k},\, \Sigma^*_{\beta_k})}{\prod_{i=1}^{L} \Pr(T^*_{iN_i} \leq S_i)}, \tag{3.8}$$

where $\mu^*_{\beta_k} = (\boldsymbol{\mu}^*_{\boldsymbol{\beta}})_k + (\boldsymbol{\Sigma}^*_{\boldsymbol{\beta}})_{k,-k} (\boldsymbol{\Sigma}^*_{\boldsymbol{\beta}})^{-1}_{-k,-k} \{\boldsymbol{\beta}_{-k} - (\boldsymbol{\mu}^*_{\boldsymbol{\beta}})_{-k}\}$ and $\Sigma^*_{\beta_k} = (\boldsymbol{\Sigma}^*_{\boldsymbol{\beta}})_{kk} - (\boldsymbol{\Sigma}^*_{\boldsymbol{\beta}})_{k,-k} (\boldsymbol{\Sigma}^*_{\boldsymbol{\beta}})^{-1}_{-k,-k} (\boldsymbol{\Sigma}^*_{\boldsymbol{\beta}})^T_{k,-k}$ with the $1 \times (q-1)$ row vector $(\boldsymbol{\Sigma}^*_{\boldsymbol{\beta}})_{k,-k}$ equal to the kth row of $\boldsymbol{\Sigma}^*_{\boldsymbol{\beta}}$ without its kth element, the $(q-1)\times(q-1)$ matrix $(\boldsymbol{\Sigma}^*_{\boldsymbol{\beta}})_{-k,-k}$ equal to $\boldsymbol{\Sigma}^*_{\boldsymbol{\beta}}$ without its kth row and kth column, and $\boldsymbol{a}_{-k}$ equal to the vector $\boldsymbol{a}$ without its kth element, for $k = 1, \ldots, q$. Now, the Gibbs update for $\boldsymbol{\beta}$ follows as slice sampling with (3.8) as target density for $k = 1, \ldots, q$.

For the other regression coefficient $\boldsymbol{\gamma}$, consider that $S_i \geq T_{iN_i}$. Therefore, (3.3) yields $\Pr(S_i \geq T_{iN_i} \mid \text{—}) = \Phi[\{\boldsymbol{x}_i^T\boldsymbol{\gamma} + \delta_i - \log(T_{iN_i})\}/\eta]$ and thus

$$p\{\log(S_i) \mid \text{—}\} = \frac{\mathcal{N}\{\log(S_i) \mid \boldsymbol{x}_i^T\boldsymbol{\gamma} + \delta_i,\, \eta^2\}}{\Phi[\{\boldsymbol{x}_i^T\boldsymbol{\gamma} + \delta_i - \log(T_{iN_i})\}/\eta]}, \quad S_i \geq T_{iN_i}. \tag{3.9}$$

The full conditional for $\boldsymbol{\gamma}$ then follows with the prior $\boldsymbol{\gamma} \sim \mathcal{N}_q(0,\, \sigma^2_\gamma\, \boldsymbol{I}_q)$ from Section 3.2.3 as

$$p(\boldsymbol{\gamma} \mid \text{—}) \propto \mathcal{N}_q(\boldsymbol{\gamma} \mid 0,\, \sigma^2_\gamma\, \boldsymbol{I}_q) \prod_{i=1}^{L} p\{\log(S_i) \mid \text{—}\} \propto \frac{\mathcal{N}_q(\boldsymbol{\gamma} \mid \boldsymbol{\mu}^*_{\boldsymbol{\gamma}},\, \boldsymbol{\Sigma}^*_{\boldsymbol{\gamma}})}{\prod_{i=1}^{L} \Phi[\{\boldsymbol{x}_i^T\boldsymbol{\gamma} + \delta_i - \log(T_{iN_i})\}/\eta]} \tag{3.10}$$

where $\boldsymbol{\mu}^*_{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}^*_{\boldsymbol{\gamma}}\, \boldsymbol{X}^T(\boldsymbol{U} - \boldsymbol{\delta})/\eta^2$ and $\boldsymbol{\Sigma}^*_{\boldsymbol{\gamma}} = (\boldsymbol{I}_q/\sigma^2_\gamma + \boldsymbol{X}^T\boldsymbol{X}/\eta^2)^{-1}$ with the $L \times q$ matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)^T$ and the L-dimensional vector $\boldsymbol{U} = \{\log(S_1), \ldots, \log(S_L)\}^T$. Analogously to (3.8), we obtain

$$p(\gamma_k \mid \text{—}) \propto \frac{\mathcal{N}(\gamma_k \mid \mu^*_{\gamma_k},\, \Sigma^*_{\gamma_k})}{\prod_{i=1}^{L} \Phi[\{\boldsymbol{x}_i^T\boldsymbol{\gamma} + \delta_i - \log(T_{iN_i})\}/\eta]}, \tag{3.11}$$

where $\mu^*_{\gamma_k} = (\boldsymbol{\mu}^*_{\boldsymbol{\gamma}})_k + (\boldsymbol{\Sigma}^*_{\boldsymbol{\gamma}})_{k,-k} (\boldsymbol{\Sigma}^*_{\boldsymbol{\gamma}})^{-1}_{-k,-k} \{\boldsymbol{\gamma}_{-k} - (\boldsymbol{\mu}^*_{\boldsymbol{\gamma}})_{-k}\}$ and $\Sigma^*_{\gamma_k} = (\boldsymbol{\Sigma}^*_{\boldsymbol{\gamma}})_{kk} - (\boldsymbol{\Sigma}^*_{\boldsymbol{\gamma}})_{k,-k} (\boldsymbol{\Sigma}^*_{\boldsymbol{\gamma}})^{-1}_{-k,-k} (\boldsymbol{\Sigma}^*_{\boldsymbol{\gamma}})^T_{k,-k}$. Similarly to $\boldsymbol{\beta}$, the Gibbs update for $\boldsymbol{\gamma}$ follows as slice sampling with (3.11) as target density for $k = 1, \ldots, q$.

Reversible jump sampler for N_i

If individual i is censored, then the number of events N_i is unknown and object of inference. Since N_i affects the dimensionality of $\boldsymbol{Y}_i$, we use a reversible jump sampler (Green, 1995; Waagepetersen and Sorensen, 2001) to update it. The sampler updates N_i and $Y_{i(n_i+1)}^{N_i} = (Y_{i(n_i+1)}, \ldots, Y_{iN_i})^T$ jointly. It is a Metropolis-Hastings algorithm on a state space of varying dimension. The state space is $\bigcup_{N_i=n_i}^{\infty} \mathbb{R}^{N_i-n_i}$ in our case.

The proposal distribution for N_i and $Y_{i(n_i+1)}^{N_i}$ is as follows. Since $N_i \geq n_i$, we sample $N_i \sim \mathbb{1}_{[n_i,\infty)}$ Poisson(λ) using the inverse transformation method. To complete the joint proposal, we only need to specify the proposal distribution of $Y_{i(n_i+1)}^{N_i}$ given N_i. We use $T_{in_i+1} \sim \mathcal{U}(c_i,\, S_i)$ and $T_{ij} \mid T_{ij-1} \sim \mathcal{U}(T_{ij-1},\, S_i)$ for $j = n_i + 2, \ldots, N_i$ as $T_{iN_i} \leq S_i$. We prefer this proposal over sampling along the lines of (3.2) as then the proposal density would involve an intractable normalization constant similarly to (3.2). Instead, we now have $\Pr(T_{in_i+1} \leq t) = (t - c_i)/(S_i - c_i)$ and $\Pr(T_{ij} \leq t \mid T_{ij-1}) = (t - T_{ij-1})/(S_i - T_{ij-1})$ for $j = n_i + 2, \ldots, N_i$. Inserting (3.1) shows $\Pr(Y_{i(n_i+1)} \leq y) = (e^y + T_{in_i} - c_i)/(S_i - c_i)$ and $\Pr(Y_{ij} \leq y \mid T_{ij-1}) = e^y/(S_i - T_{ij-1})$ for $j = n_i + 2, \ldots, N_i$. The proposal density is thus

$$f_{N_i}(Y_{i(n_i+1)}^{N_i}) = \begin{cases} \frac{e^{Y_{i(n_i+1)}} + T_{n_i} - c_i}{S_i - c_i} \prod_{j=n_i+2}^{N_i} \frac{e^{Y_{ij}}}{S_i - T_{ij-1}}, & N_i > n_i \\ 1, & N_i = n_i \end{cases}.$$

To derive the acceptance probability, we follow the notation in Waagepetersen and Sorensen (2001, Section 4) where proposed values are denoted by a prime ('). Specifically, the proposal distributions are written as $p_{N_iN_i'} \propto \mathbb{1}[N_i' \geq n_i]$ Poisson$(N_i' \mid \lambda)$ and $q_{N_iN_i'}(Y_{i(n_i+1)}^{N_i}, \cdot) = f_{N_i'}(\cdot)$. The target density follows from Sections 3.2.1, 3.2.2 and 3.2.5 as

$$\begin{aligned} \pi(N_i, Y_{i(n_i+1)}^{N_i}) \propto & \frac{\text{Poisson}(N_i \mid \lambda)}{\Pr\{\log(\hat{T}_{iN_i}^*) \leq \log(S_i)\}} \\ & \times \prod_{j=n_i+1}^{N_i} \mathcal{N}\{Y_{ij} \mid \boldsymbol{x}_i^T\boldsymbol{\beta} + m_{i1} + m_{i2}(Y_{i(j-1)} - \boldsymbol{x}_i^T\boldsymbol{\beta} - m_{i1}),\, \sigma^2\}, \end{aligned}$$

for $N_i \geq n_i$, $T_{i(n_i+1)} > c_i$ and $T_{iN_i} \leq S_i$. The dimension changing map can be written in the notation of Waagepetersen and Sorensen (2001, Section 4) as

$$g_{N_iN_i'}\{Y_{i(n_i+1)}^{N_i}, (Y')_{i(n_i+1)}^{N_i'}\} = \begin{pmatrix} g_{1N_iN_i'}\{Y_{i(n_i+1)}^{N_i}, (Y')_{i(n_i+1)}^{N_i'}\} \\ g_{2N_iN_i'}\{Y_{i(n_i+1)}^{N_i}, (Y')_{i(n_i+1)}^{N_i'}\} \end{pmatrix},$$

where $g_{1N_iN_i'}\{Y_{i(n_i+1)}^{N_i}, (Y')_{i(n_i+1)}^{N_i'}\} = (Y')_{i(n_i+1)}^{N_i'}$ and $g_{2N_iN_i'}\{Y_{i(n_i+1)}^{N_i}, (Y')_{i(n_i+1)}^{N_i'}\} = Y_{i(n_i+1)}^{N_i}$. The acceptance probability is then given by (Waagepetersen and Sorensen, 2001, Equation 19)

$$
\begin{aligned}
&a_{N_iN_i'}\{Y_{i(n_i+1)}^{N_i}, (Y')_{i(n_i+1)}^{N_i'}\} \\
&\quad = \min\left[1,\ \frac{\pi\{N_i', (Y')_{i(n_i+1)}^{N_i'}\}\, p_{N_i'N_i}\, q_{N_i'N_i}\{(Y')_{i(n_i+1)}^{N_i'}, Y_{i(n_i+1)}^{N_i}\}}{\pi(N_i, Y_{i(n_i+1)}^{N_i})\, p_{N_iN_i'}\, q_{N_iN_i'}\{Y_{i(n_i+1)}^{N_i}, (Y')_{i(n_i+1)}^{N_i'}\}}\ |J_{g_{N_iN_i'}}|\right],
\end{aligned}
$$

where $|J_{g_{N_iN_i'}}|$ denotes the determinant of the Jacobian of $g_{N_iN_i'}$. The elements of $J_{g_{N_iN_i'}}$ are all zero except for one entry in each row that equals one so that its determinant equals one. Additionally substituting the definitions of the various terms yields as acceptance probability

$$
a_{N_iN_i'}\{Y_{i(n_i+1)}^{N_i}, (Y')_{i(n_i+1)}^{N_i'}\} = \min\left[1,\ \frac{C_{N_i'}\{(Y')_{i(n_i+1)}^{N_i'}\}}{C_{N_i}(Y_{i(n_i+1)}^{N_i})} \times \frac{f_{N_i}(Y_{i(n_i+1)}^{N_i})}{f_{N_i'}\{(Y')_{i(n_i+1)}^{N_i'}\}}\right],
$$

where

$$
C_{N_i}(Y_{i(n_i+1)}^{N_i}) = \frac{\prod_{j=n_i+1}^{N_i} \mathcal{N}\{Y_{ij} \mid \boldsymbol{x}_i^T\boldsymbol{\beta} + m_{i1} + m_{i2}(Y_{i(j-1)} - \boldsymbol{x}_i^T\boldsymbol{\beta} - m_{i1}),\, \sigma^2\}}{\Pr\{\log(\hat{T}_{iN_i}^*) \le \log(S_i)\}}.
$$

This reversible jump sampler updates both N_i and $Y_{i(n_i+1)}^{N_i}$. Additionally, we update $Y_{i(n_i+1)}^{N_i}$ as described in the next Section to improve mixing of the Gibbs sampler in case the Metropolis-Hastings sampler in this Section rarely accepts the proposed samples.

Survival and log gap times

If individual i is censored, then the log gap times $Y_{i(n_i+1)}^{N_i}$ and the survival time S_i are imputed in the Gibbs sampler. The full conditional for the vector $Y_{i(n_i+1)}^{N_i}$ is hard to sample from due to the truncation $T_{iN_i} \le S_i$. Instead, we consider its elementwise full conditionals. By (3.2), for $j = 2, \ldots, N_i - 1$,

$$
\begin{aligned}
p(Y_{ij} \mid \text{—}) &\propto \mathcal{N}\left\{Y_{ij} \;\middle|\; \boldsymbol{x}_i^T\boldsymbol{\beta} + m_{i1} + \frac{m_{i2}}{1+m_{i2}^2}\,(Y_{i(j-1)} + Y_{i(j+1)} - 2\boldsymbol{x}_i^T\boldsymbol{\beta} - 2m_{i1}),\ \frac{\sigma^2}{1+m_{i2}^2}\right\}, \\
p(Y_{iN_i} \mid \text{—}) &\propto \mathcal{N}\{Y_{iN_i} \mid \boldsymbol{x}_i^T\boldsymbol{\beta} + m_{i1} + m_{i2}(Y_{i(N_i-1)} - \boldsymbol{x}_i^T\boldsymbol{\beta} - m_{i1}),\, \sigma^2\};
\end{aligned}
\tag{3.12}
$$

for $T_{i(n_i+1)} > c_i$ and $T_{iN_i} \le S_i$. Let $R_{ij} = S_i - T_{iN_i} + e^{Y_{ij}} = S_i - \sum_{j^* \ne j} e^{Y_{ij^*}}$. Then, $e^{Y_{ij}}$ is bounded from above by R_{ij} since $T_{iN_i} \le S_i$. Additionally, $e^{Y_{i(n_i+1)}}$ is bounded from below by $c_i - T_{in_i}$ since $T_{i(n_i+1)} > c_i$. Therefore, we can sample from (3.12) with these truncations for $j = n_i + 1, \ldots, N_i$ using the

inverse transform method.

We sample S_i from (3.9) truncated to $S_i > c_i$ using the inverse transform method.

Dirichlet process parameters

As detailed in Section 3.1, the discreteness of the DP induces clustering of the individuals. Denote the random effects in the hth cluster by $(\boldsymbol{m}_h^*, \delta_h^*)$ and the cluster that individual i belongs to by s_i. Then, $s_i = h$ if and only if $(\boldsymbol{m}_i, \delta_i) = (\boldsymbol{m}_h^*, \delta_h^*)$. To update $(\boldsymbol{m}_i, \delta_i)$ for $i = 1, \ldots, L$, we update the cluster allocations s_i and the cluster-specific parameters $(\boldsymbol{m}_h^*, \delta_h^*)$, using Algorithm 8 from Neal (2000b) with the algorithm-specific parameter $m = 2$. We choose this algorithm since independent sampling from the full conditional of $\boldsymbol{m}_h^*$, required for instance for Neal's Algorithm 2, is hard due to the intractability of the likelihood discussed in Subsection 3.2.5.

Neal's Algorithm 8 requires sampling of $(\boldsymbol{m}_h^*, \delta_h^*)$ that leaves its full conditional distribution invariant. We do this by first sampling m_{h1}^*, then m_{h2}^* and lastly δ_h^* such that their respective full conditionals remain invariant. A derivation analogous to the one for (3.7) yields the full conditional for m_{h1}^*. Specifically, we introduce $Y_{i1}^{m_{i1}} = Y_{i1} - \boldsymbol{x}_i^T\boldsymbol{\beta}$, $Y_{ij}^{m_{i1}} = \{Y_{ij} - \boldsymbol{x}_i^T\boldsymbol{\beta} - m_{i2}(Y_{i(j-1)} - \boldsymbol{x}_i^T\boldsymbol{\beta})\}/(1 - m_{i2})$ for $j = 2, \ldots, N_i$ and the $N_i \times N_i$ diagonal matrix $\boldsymbol{\Sigma}_{m_{i1}}$ with $\mathrm{diag}(\boldsymbol{\Sigma}_{m_{i1}}) = \sigma^2 \{1, (1 - m_{i2})^{-2}, \ldots, (1 - m_{i2})^{-2}\}^T$. Then,

$$p(m_{h1}^* \mid \text{—}) \propto \frac{\mathcal{N}(m_{h1}^* \mid \mu_{m_{h1}^*}^*, \Sigma_{m_{h1}^*}^*)}{\prod_{\{i|s_i=h\}} \Pr(T_{iN_i}^* \leq S_i)}, \tag{3.13}$$

where $\mu_{m_{h1}^*}^* = \Sigma_{m_{h1}^*}^* \sum_{\{i|s_i=h\}} \boldsymbol{1}_{1\times N_i} \boldsymbol{\Sigma}_{m_{i1}}^{-1} \boldsymbol{Y}_i^{m_{i1}}$ and $\Sigma_{m_{h1}^*}^* = 1/\{1/\sigma_m^2 + \sum_{\{i|s_i=h\}} \mathrm{sum}(\boldsymbol{\Sigma}_{m_{i1}}^{-1})\}$ with $\mathrm{sum}(\boldsymbol{\Sigma}_{m_{i1}}^{-1}) = \{1 + (N_i - 1)(1 - m_{i2})^2\}/\sigma^2$.

Similarly for m_{h2}^*, we introduce $Y_{i(j-1)}^{m_{i2}} = (Y_{ij} - \boldsymbol{x}_i^T\boldsymbol{\beta} - m_{i1})/(Y_{i(j-1)} - \boldsymbol{x}_i^T\boldsymbol{\beta} - m_{i1})$ for $j = 2, \ldots, N_i$ and the $(N_i - 1) \times (N_i - 1)$ diagonal matrix $\boldsymbol{\Sigma}_{m_{i2}}$ with $(\boldsymbol{\Sigma}_{m_{i2}})_{jj} = \sigma^2/(Y_{ij} - \boldsymbol{x}_i^T\boldsymbol{\beta} - m_{i1})^2$ for $j = 1, \ldots, N_i - 1$. Then,

$$p(m_{h2}^* \mid \text{—}) \propto \frac{\mathcal{N}(m_{h2}^* \mid \mu_{m_{h2}^*}^*, \Sigma_{m_{h2}^*}^*)}{\prod_{\{i|s_i=h\}} \Pr(T_{iN_i}^* \leq S_i)}, \tag{3.14}$$

where $\mu_{m_{h2}^*}^* = \Sigma_{m_{h2}^*}^* \sum_{\{i|s_i=h\}} \boldsymbol{1}_{1\times N_i} \boldsymbol{\Sigma}_{m_{i2}}^{-1} \boldsymbol{Y}_i^{m_{i2}}$ and $\Sigma_{m_{h2}^*}^* = 1/\{1/\sigma_m^2 + \sum_{\{i|s_i=h\}} \mathrm{sum}(\boldsymbol{\Sigma}_{m_{i2}}^{-1})\}$ with $\mathrm{sum}(\boldsymbol{\Sigma}_{m_{i2}}^{-1}) = \sum_{j=1}^{N_i-1} (Y_{ij} - \boldsymbol{x}_i^T\boldsymbol{\beta} - m_{i1})^2/\sigma^2$.

For δ_h^*, a derivation similar to the one for (3.10) yields

$$p(\delta_h^* \mid \text{—}) \propto \frac{\mathcal{N}(\delta_h^* \mid \mu_{\delta_h^*}^*, \Sigma_{\delta_h^*}^*)}{\prod_{\{i|s_i=h\}} \Phi[\{\boldsymbol{x}_i^T\boldsymbol{\gamma} + \delta_h^* - \log(T_{iN_i})\}/\eta]} \tag{3.15}$$

where $\mu^*_{\delta^*_h} = \Sigma^*_{\delta^*_h} \sum_{\{i|s_i=h\}} \{\log(S_i) - \boldsymbol{x}_i^T\boldsymbol{\gamma}\}/\eta^2$ and $\Sigma^*_{\delta^*_h} = 1/(1/\sigma^2_\delta + |\{i \mid s_i = h\}|/\eta^2)$. Now, the update of $(\boldsymbol{m}^*_h,\, \delta^*_h)$ follows as slice sampling with (3.13–3.15) as target density for $h = 1, \ldots, K$ where K is the number of clusters.

Equation 13 from Escobar and West (1995b) provides the Gibbs update for the DP concentration parameter M: First, draw $\eta_M \sim \text{Beta}(M+1,\, L)$. Then, with probability $1/[1 + L\,\{b_M - \log(\eta_M)\}/(a_M + K - 1)]$ where L denotes the number of individuals, draw $M \sim \text{Gamma}\{a_M + K,\, b_M - \log(\eta_M)\}$. Otherwise, draw $M \sim \text{Gamma}\{a_M + K - 1,\, b_M - \log(\eta_M)\}$.

Variance parameters and λ

Recalling the prior and likelihood for σ^2 from Section 3.2.3 and (3.2), respectively, we obtain

$$p(\sigma^2 \mid \text{—}) \propto \frac{\text{Inv-Gamma}\{\sigma^2 \mid (\nu_{\sigma^2} + \sum_{i=1}^L N_i)/2,\, (\nu_{\sigma^2}\,\sigma_0^2 + \sum_{i=1}^L \|\boldsymbol{Y}_i - \boldsymbol{\mu}_i^{\sigma^2}\|^2)/2\}}{\prod_{i=1}^L \Pr(T^*_{iN_i} \leq S_i)}, \tag{3.16}$$

where $\mu^{\sigma^2}_{i1} = \boldsymbol{x}_i^T\boldsymbol{\beta} + m_{i1}$ and $\mu^{\sigma^2}_{ij} = \boldsymbol{x}_i^T\boldsymbol{\beta} + m_{i1} + m_{i2}(Y_{i(j-1)} - \boldsymbol{x}_i^T\boldsymbol{\beta} - m_{i1})$ for $j = 2, \ldots, N_i$. Similarly for η^2, its prior from Section 3.2.3 and the likelihood in (3.9) yield

$$p(\eta^2 \mid \text{—}) \propto \frac{\text{Inv-Gamma}\{\eta^2 \mid (\nu_{\eta^2} + L)/2,\, (\nu_{\eta^2}\,\eta_0^2 + \sum_{i=1}^L \{\log(S_i) - \boldsymbol{x}_i^T\boldsymbol{\gamma} - \delta_i\}^2)/2\}}{\prod_{i=1}^L \Phi[\{\boldsymbol{x}_i^T\boldsymbol{\gamma} + \delta_i - \log(T_{iN_i})\}/\eta]}. \tag{3.17}$$

Now, the Gibbs updates for σ^2 and η^2 follow as slice sampling with (3.16) and (3.17) as target density, respectively.

Since $N_i \sim \mathbb{1}_{[1,\infty)}\,\text{Poisson}(\lambda)$ per Section 3.2.2, $\Pr(N_i = l \mid \lambda) = \text{Poisson}(l \mid \lambda)/(1 - e^{-\lambda})$ for $l = 1, 2, \ldots$, where $\text{Poisson}(\cdot \mid \lambda)$ is the probability mass function of a Poisson distribution with mean λ and the division by $1 - e^{-\lambda} = 1 - \text{Poisson}(0 \mid \lambda)$ ensures $\sum_{l=1}^\infty \Pr(N_i = l \mid \lambda) = 1$. Combining this likelihood with the prior $\lambda \sim \text{Gamma}(a_\lambda,\, b_\lambda)$ from Section 3.2.3 results in

$$p(\lambda \mid \text{—}) \propto \frac{\text{Gamma}(\lambda \mid a_\lambda + \sum_{i=1}^L N_i,\, b_\lambda + L)}{(1 - e^{-\lambda})^L}. \tag{3.18}$$

The Gibbs update for λ follows as slice sampling with (3.18) as target density.

3.3 Simulation study

We investigate the performance of our model and the Markov chain Monte Carlo algorithm via a simulation study. As in Chapter 2, we set the num-

ber of covariates q to zero for ease of explanation. We consider $L = 150$ individuals spread over 3 clusters of size 50 each. Then, we generate data according to the likelihood in Subsection 3.2.2 with $\lambda = 7$, $\sigma^2 = 1$, $\eta^2 = 1$, $\boldsymbol{m}_h^* = \{h,\, 0.8\,(h-2)\}^T$ and $\delta_h^* = h + 4$ where $\boldsymbol{m}_h^* = \boldsymbol{m}_i$ and $\delta_h^* = \delta_i$ if and only if individual i belongs to the hth cluster for $h = 1, 2, 3$. From these simulated data, we construct four different scenarios, namely where 0%, 50%, 80% and 90% of the individuals are censored. Similarly to Chapter 2, the censoring times c_i are sampled uniformly from the time interval $(T_{i1},\, S_i)$ between the first event recurrence and death.

We choose hyperparameters yielding uninformative prior distributions as detailed in the following Subsection. The base measure G_0 of the DP prior has high variance. A priori, σ^2 and η^2 have an expected value of one and a variance of 100. We run the Gibbs sampler for 200,000 iterations, discarding the first 20,000 as burn-in and thinning every 10 iterations, resulting in a final posterior sample size of 18,000.

Figures 3.1 through 3.4 summarize the results. The uncertainty quantification for the number of recurrent events N_i in Figure 3.1 is sensible with only 2 of the 330 credible intervals not covering the true N_i. Figures 3.2 and 3.3 show increased posterior uncertainty for higher levels of censoring. This increase in uncertainty is larger for δ_i and η^2, which relate to the survival time S_i, than for other parameters which relate to the recurrence process. A reason for this is that, for a censored individual i, some event times are observed while S_i is right-censored and thus not observed. The posterior mass on the actual number of clusters is marginally higher for the uncensored than for the censored data in Figure 3.4, although it must be noticed that posterior inference on K is robust across different level of censoring.

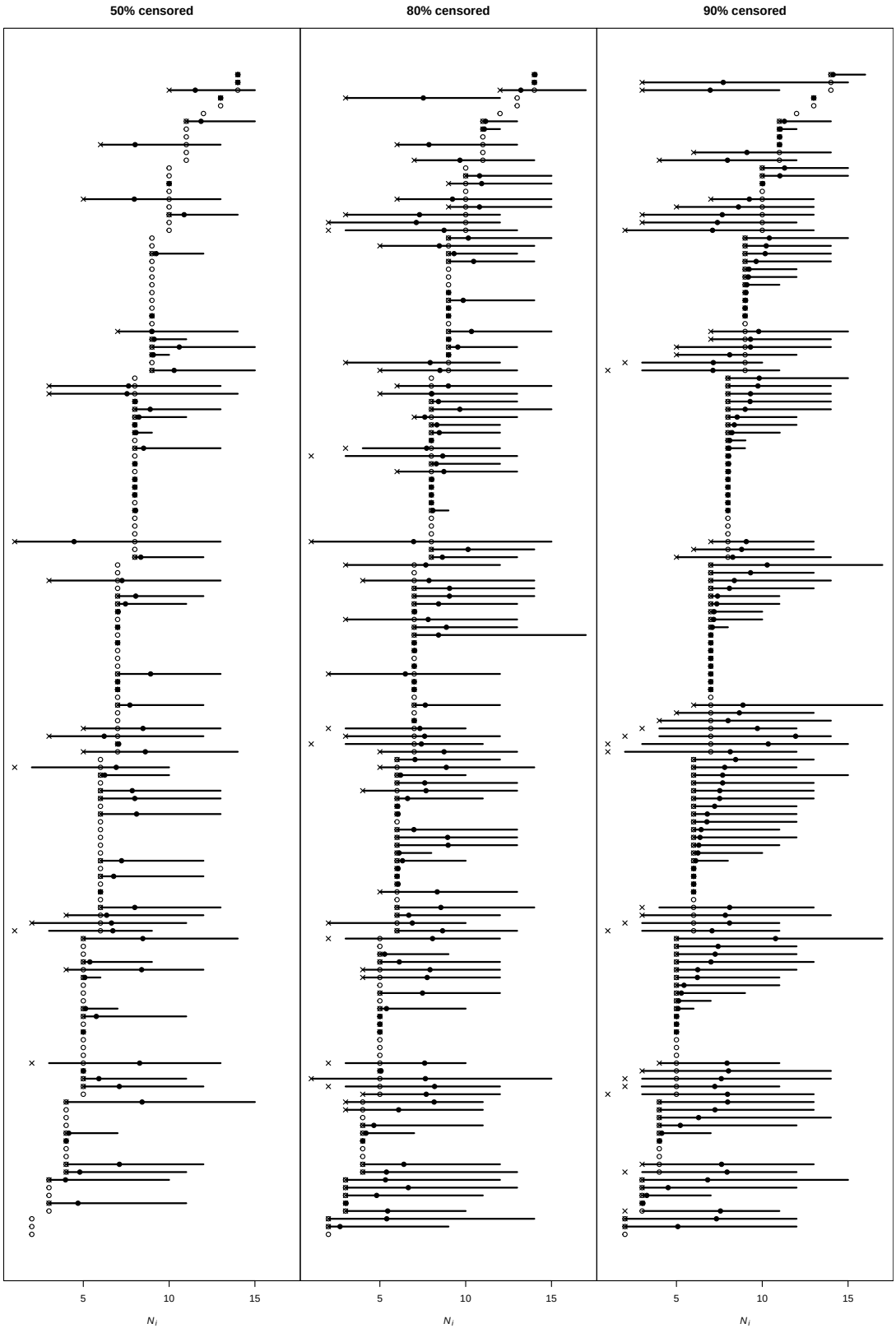


Figure 3.1: The number of gap times N_i (circle) and, if applicable, their posterior means (dot) and 95% posterior credible intervals (lines) for each individual from our model fit on the simulated data. For censored individuals, the number of observed gap times n_i is marked by '$\times$'.

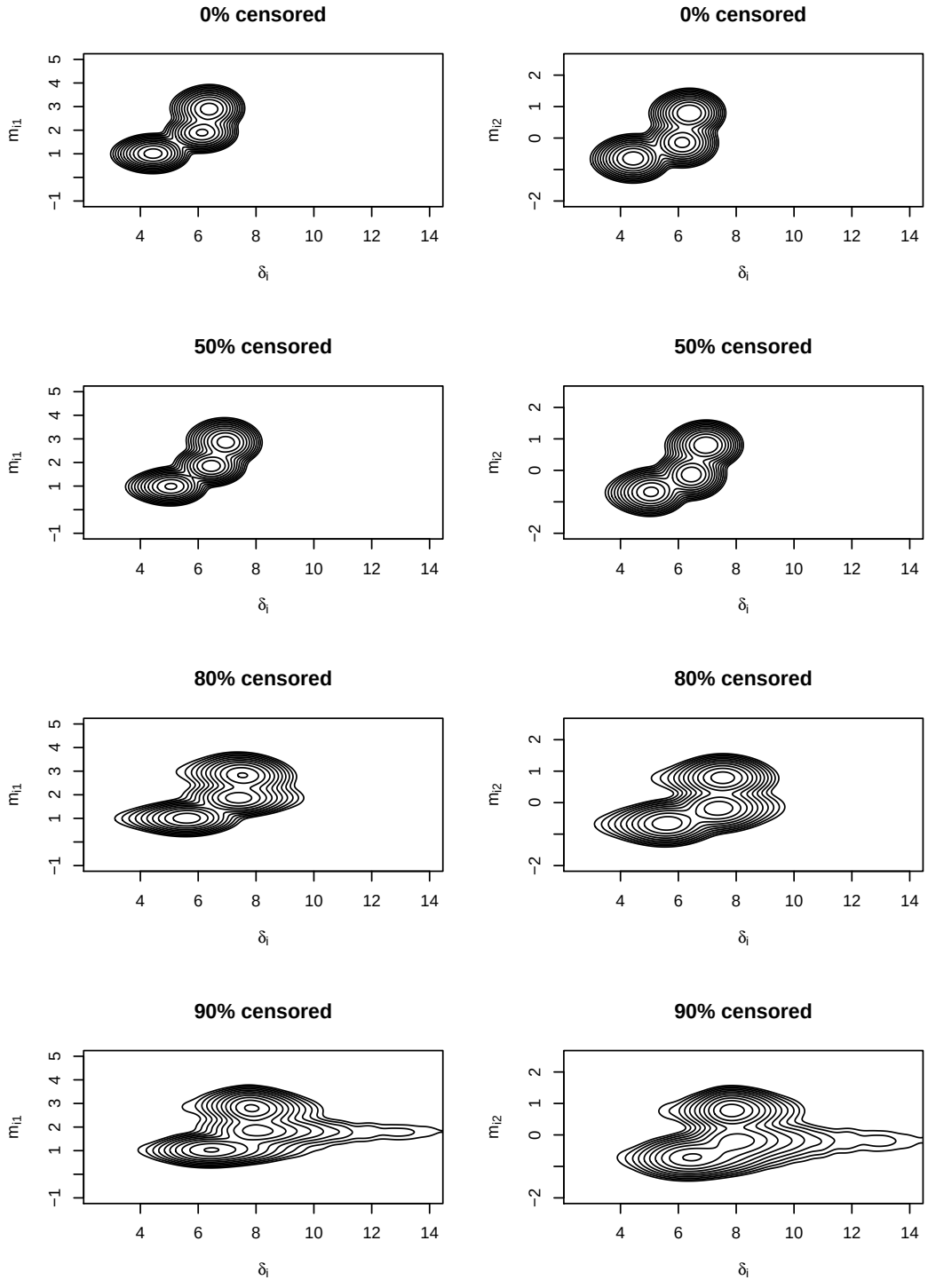


Figure 3.2: Contour plots of the log of the bivariate posterior predictive densities of $(m_{i1},\, \delta_i)$ (left) and $(m_{i2},\, \delta_i)$ (right) for a hypothetical new individual from the simulated data.

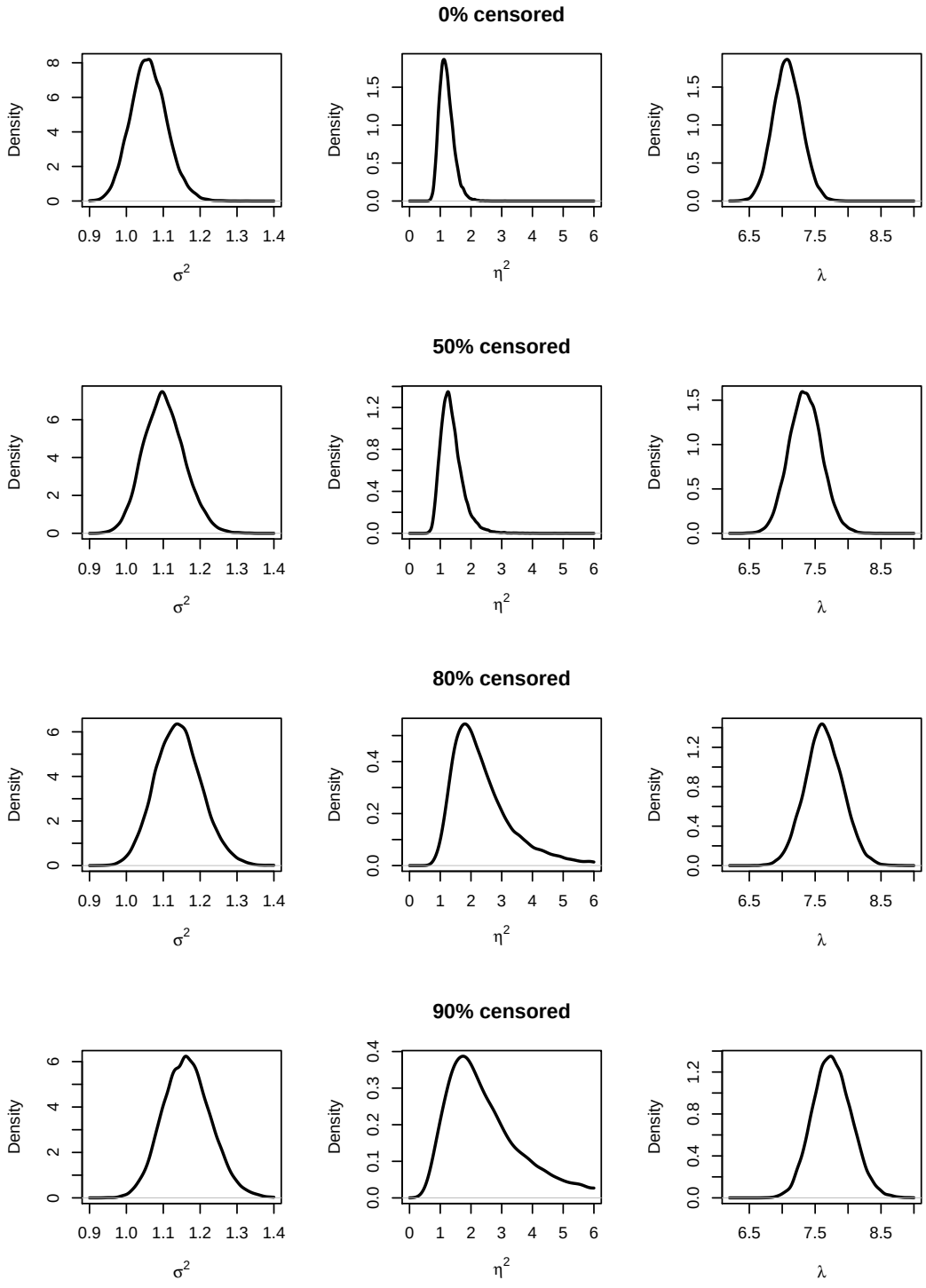


Figure 3.3: Posterior densities for σ^2, η^2 and λ from our model fit on the simulated data.

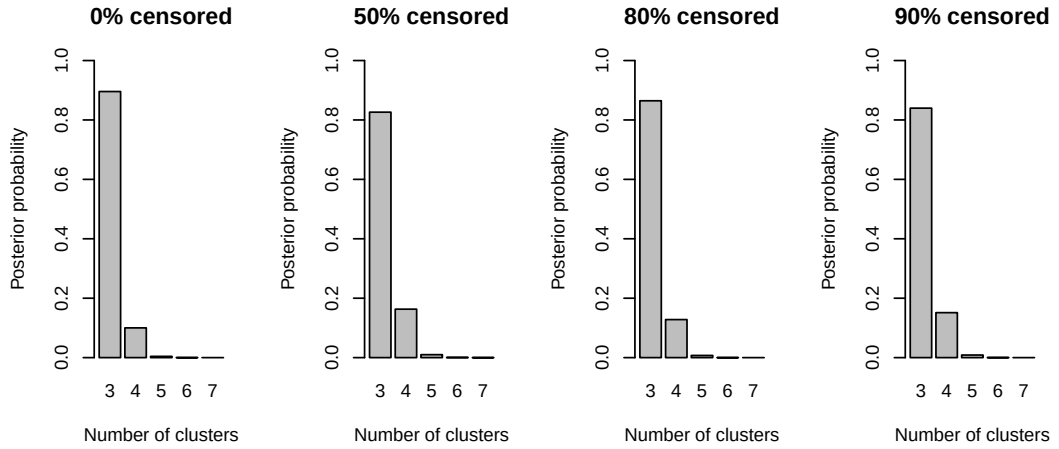


Figure 3.4: Posterior distribution of the number of clusters from our model fit on the simulated data.

3.3.1 Prior specification

For the application of our model to the simulation study in Section 3.3 and to the AF data in Section 3.4, the hyperparameters of the prior distributions in Subsection 3.2.3 are chosen as follows. The hyperparameters of the priors on σ^2 and η^2 are such that their prior means equal 1 and their prior variances equal 100. Specifically,

$$\begin{aligned}
\mathrm{E}(\sigma^2) &= \frac{\nu_1\,\sigma_0^2/2}{\nu_1/2-1} = \frac{\nu_1\,\sigma_0^2}{\nu_1-2} = 1,\\
\mathrm{Var}(\sigma^2) &= \frac{\nu_1^2\,\sigma_0^4/4}{(\nu_1/2-1)^2\,(\nu_1/2-2)} = \frac{2\,\nu_1^2\,\sigma_0^4}{(\nu_1-2)^2\,(\nu_1-4)} = 100,\\
\mathrm{E}(\eta^2) &= \frac{\nu_2\,\eta_0^2/2}{\nu_2/2-1} = \frac{\nu_2\eta_0^2}{\nu_2-2} = 1,\\
\mathrm{Var}(\eta^2) &= \frac{\nu_2^2\,\eta_0^4/4}{(\nu_2/2-1)^2\,(\nu_2/2-2)} = \frac{2\,\nu_2^2\,\eta_0^4}{(\nu_2-2)^2\,(\nu_2-4)} = 100;
\end{aligned}$$

yields $\nu_1 = 4.02$, $\sigma_0^2 = 2.02/4.02$, $\nu_2 = 4.02$ and $\eta_0^2 = 2.02/4.02$. Similarly, the prior variances of the elements of the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ equal $\sigma_\beta^2 = \sigma_\gamma^2 = 100$. The hyperparameters related to the base measure G_0 of the DP prior are chosen as non-informative with $\sigma_\delta^2 = \sigma_m^2 = 100$. We choose an uninformative prior on the DP concentration parameter M with $a_M = 2$ and $b_M = 1$. Finally, $a_\lambda = b_\lambda = 1$ specifies the prior for λ.

3.4 Application to atrial fibrillation data

3.4.1 Background

We apply our model to the data on atrial fibrillation (AF) described in Schroder et al. (2019b). AF is the most common serious cardiac arrhythmia with more than 33 million cases worldwide, increasing rapidly with 5 million new cases per year (Chung et al., 2020). It is characterized by an irregular and often high heart rate where the heart's upper chambers beat out of sync with its lower chambers. AF causes substantial morbidity and mortality, for instance due to heart failure and stroke. It places a high burden on health care systems, constituting 2.4% of the United Kingdom's National Health Service budget in 2000 (Thrall et al., 2006).

AF is often a chronic condition that requires repeated treatment. The goal of these treatments is to reduce the rate of AF episodes, as an increased number of episodes is associated with complications such as stroke (Munger et al., 2014). Thus, there is dependence between recurrence and survival, which our model is able to capture. Additionally, more frequent AF events are associated with an increase in the rate of episodes going forward (Wijffels et al., 1995). This points to temporal dependence in the AF recurrence process, as captured in our model by (3.2).

A variety of treatments exist. These include prophylactic anti-arrhythmic medication and cardioversion (Schroder et al., 2019b). Anti-arrhythmic medication aims to reduce the rate and duration of AF episodes. Cardioversion aims to restore the heart rhythm when it is abnormal, that is while someone is experiencing an AF episode. Cardioversion is either electrical, using direct currents, or pharmacologic. Electrical cardioversion takes place in a hospital. AF diagnosis usually requires an electrocardiogram (ECG).

The condition of AF can be categorized into three subtypes: paroxysmal, persistent and permanent (January et al., 2014). Episodes of paroxysmal AF terminate spontaneously without treatment. In contrast, persistent AF is when the episode only ends due to an intervention. Lastly, AF is permanent when the patient and clinician decide to no longer attempt to restore the heart rhythm. Additionally, AF episodes are either symptomatic or asymptomatic. Symptoms of AF include palpitations and chest pain.

As for many other chronic diseases, clinical interest lies in both the final outcome (death or survival time) and the dynamics of the process itself, since it determines the subsequent quality of the patient's life (Thrall et al., 2006). From an economic and healthcare planning perspective, there is great interest in reducing rehospitalization for AF. In fact, a better understanding of

both death and non-fatal clinical events could lead to improved prognosis and assessment of the impact and costs of AF by health providers. It is, therefore, of paramount importance to develop a comprehensive model for disease management, mortality and associated clinical event histories, which also accounts for the significant interindividual variability in disease course as it is typical of chronic diseases and biological events.

3.4.2 Data description and analysis

The data (Schroder et al., 2019a) consist of hospitalizations from January 1, 2008 to March 1, 2014 at the Department of Cardiology at University Hospital Copenhagen, Hvidovre, Denmark (Schroder et al., 2019b). The primary reason of all hospitalizations is symptomatic AF. Some include cardioversion treatment. AF is confirmed by ECG. We consider $L = 60$ patients that experience more than one hospitalization and thus at least one *re*hospitalization. This first hospitalization represents the origin of a patient's recurrence process such that $T_{i0} = 0$ for all i. Consequently, n_i represents the number of observed gap times between subsequent hospitalizations due to AF. Patients experience between 1 and 16 rehospitalizations each and $\sum_{i=1}^{L} n_i = 252$ in aggregate. Table 3.1 shows how they are distributed across patients. Gap times are defined as the difference between successive hospitalizations and, as such, capture both the length of stay in the hospital and the time between discharge and the next hospitalization.

n_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Frequency	14	10	9	4	3	8	4	3	0	2	1	0	1	0	0	1

Table 3.1: Frequency table of the number of observed gap times n_i in the AF data.

The main clinical outcome of interest is deterioration to permanent AF or death. We therefore define the survival time S_i as the time to permanent AF or death. If the terminal event is permanent AF, which is diagnosed during a rehospitalization, then the terminal event is also the last recurrent event such that $S_i = T_{iN_i}$. The survival times of 45 out of the 60 recurrence processes are censored due to the follow-up ending on March 1, 2014, resulting in unobserved total number of gap times N_i.

Patient characteristics are determined at the first hospitalization. They are 1) age; and the binary variables 2) gender; whether 3) AF is paroxysmal or persistent; and whether the patient has 4) hypertension; 5) heart disease; or 6) is on anti-arrhythmic medication. Here, heart disease includes heart

failure, heart valve disease and ischemic heart disease. These variables form the subject-specific 6-dimensional covariate vector $\boldsymbol{x}_i$, with $q = 6$. Being older or female, hypertension, and heart disease are known to be associated with more severe AF (January et al., 2014). Anti-arrhythmic medication aims to prevent and ameliorate the reoccurrence of AF. We standardize the age in $\boldsymbol{x}_i$. Table 3.2, and Figures 3.5 and 3.6 summarize the patient characteristics, and the gap and survival times.

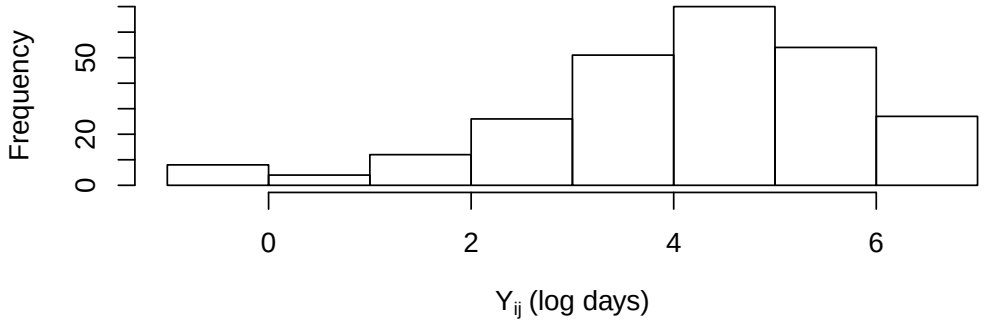


Figure 3.5: Histogram of the 252 observed log gap times Y_{ij} in the AF data.

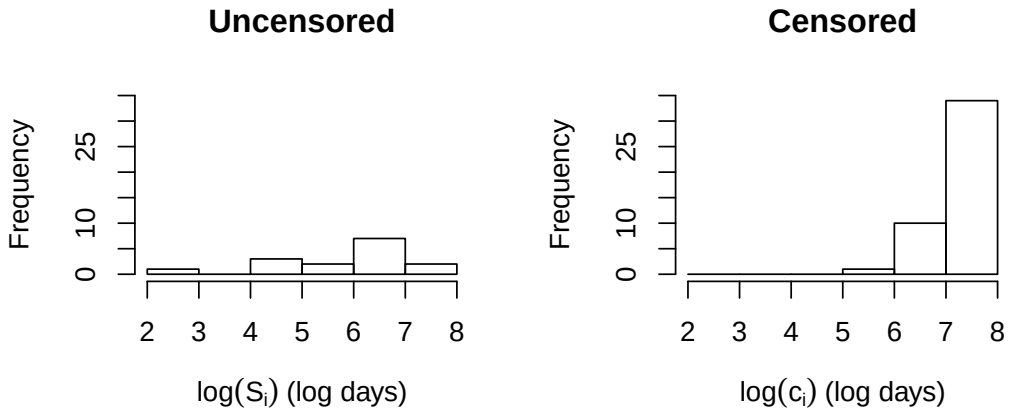


Figure 3.6: Histograms of the log of the 15 observed survival times S_i (left) and the 45 censoring times (right) in the AF data. If the survival time S_i is observed, then $c_i = S_i$.

We use the same priors, from Subsection 3.3.1 , and set-up of the Gibbs sampler as the simulation study in Section 3.3. Then, the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ have high prior variance.

3.4.3 Posterior inference on the number of recurrent events

Figure 3.7 summarizes the posterior distribution of the total number of gap times N_i for each patient.

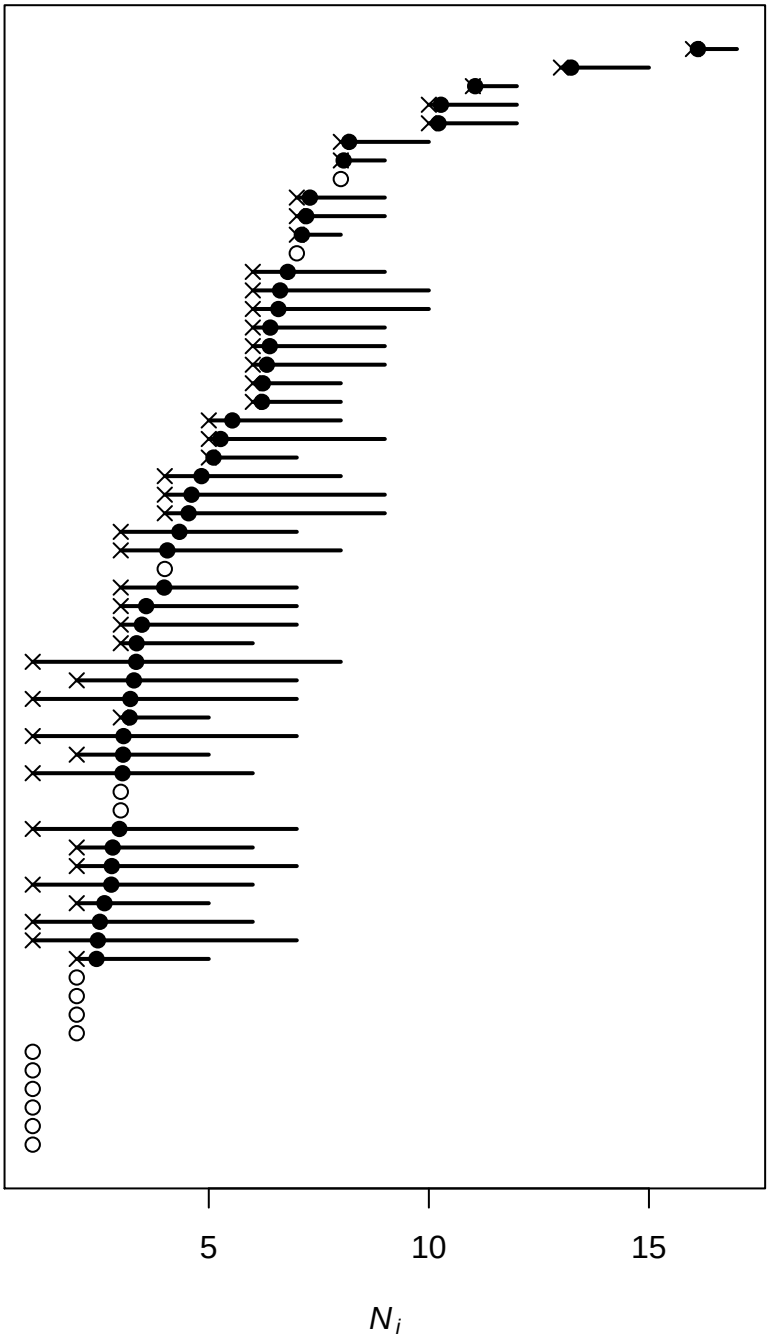


Figure 3.7: The number of gap times N_i (circle) if observed and otherwise their posterior means (dot) and 95% posterior credible intervals (lines) for each patient from our model fit on the AF data. For censored patients, the number of observed gap times n_i is marked by '$\times$'.

The posterior means for the censored N_i are generally in line with the observed N_i. The unobserved N_i are sometimes inferred to be larger than

the largest observed N_i. After all, the largest observed N_i equals 8 while the number of observed gap times n_i is 16 for one patient. This is expected since patients with longer survival times S_i are both more likely to have a higher number of gap times N_i and to be censored due to end of study. Our model flexibly captures N_i's uncertainty, which varies notably across censored patients. These findings highlight the importance of modelling N_i when the number of events is censored and, therefore, unknown.

3.4.4 Posterior inference on the regression coefficients

The regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ capture the covariate effects on the recurrent event and survival processes, respectively. Figure 3.8 shows no evident effect of any of the covariates as the corresponding 95% credible intervals include 0.

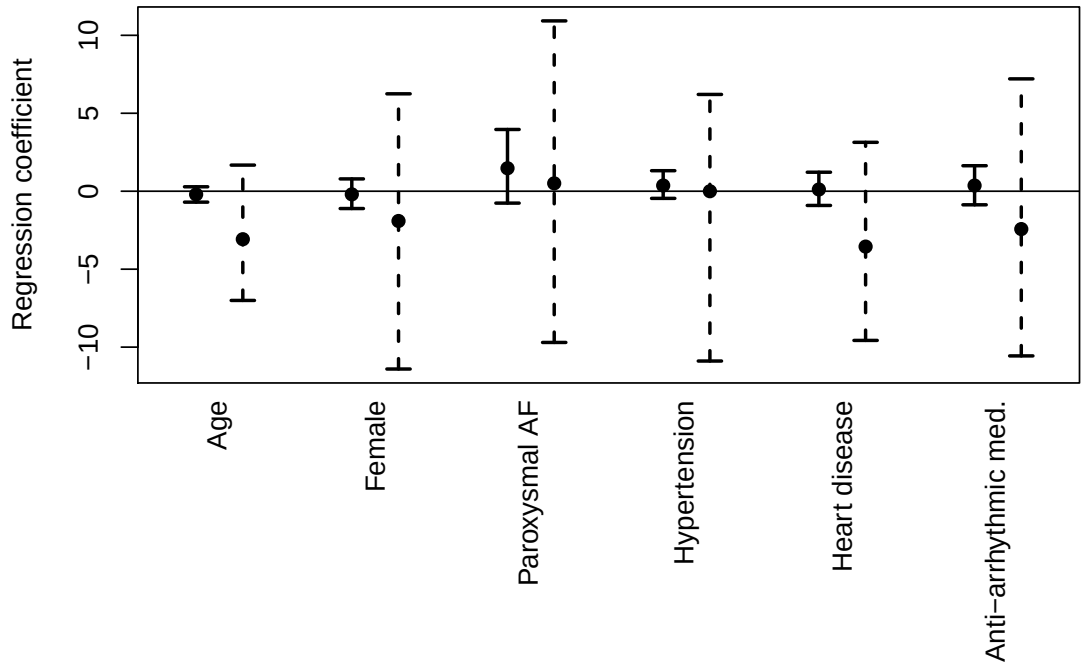


Figure 3.8: Posterior means (dot) and 95% marginal posterior credible intervals (lines) of the regression coefficients from our model fit on the AF data. The solid lines represent credible intervals for the regression coefficients $\boldsymbol{\beta}$ in (3.2) for the gap times model. The dashed lines correspond with the regression coefficients $\boldsymbol{\gamma}$ in (3.3) for the survival times.

This is in line with the analysis of these data described in Schroder et al. (2019b). The relatively small sample size of $L = 60$ might be the reason that we do not find strong effects, even though most of these covariates are risk factors for AF.

3.4.5 Posterior inference on the cluster allocation

As discussed in Section 3.1, the DP prior on $(\boldsymbol{m}_i, \delta_i)$ described in Subsection 3.2.3 allows for clustering of patients based on their recurrent event and survival profiles. The random effects parameters determine the clustering of patients and capture the dependence between the recurrence and survival processes. Indeed, the posterior predictive distribution of these parameters for a hypothetical new patient is multimodal as shown in Figure 3.9, indicating the presence of multiple patient subpopulations.

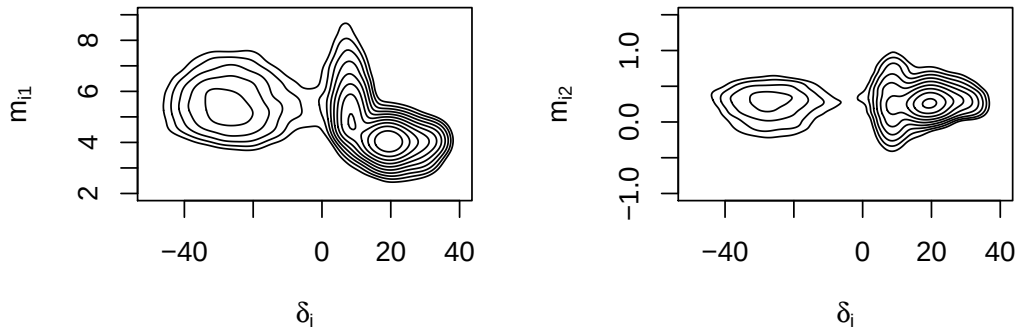


Figure 3.9: Contour plots of the log of the bivariate posterior predictive densities of (m_{i1}, δ_i) (left) and (m_{i2}, δ_i) (right) for a hypothetical new patient from the AF data.

The clustering depends on both gap time trajectories and survival outcomes thanks to the joint distribution on $\boldsymbol{m}_i$ and δ_i. In particular, Figure 3.9 reports the bivariate posterior marginals of (m_{i1}, δ_i) and (m_{i2}, δ_i), which are clearly trimodal. It is interesting to note that the modes for m_{i2} are centred around the same value, indicating a similar time dependence structure among gap times across clusters. This result is confirmed also by the posterior distribution of the number of clusters shown in Figure 3.10.

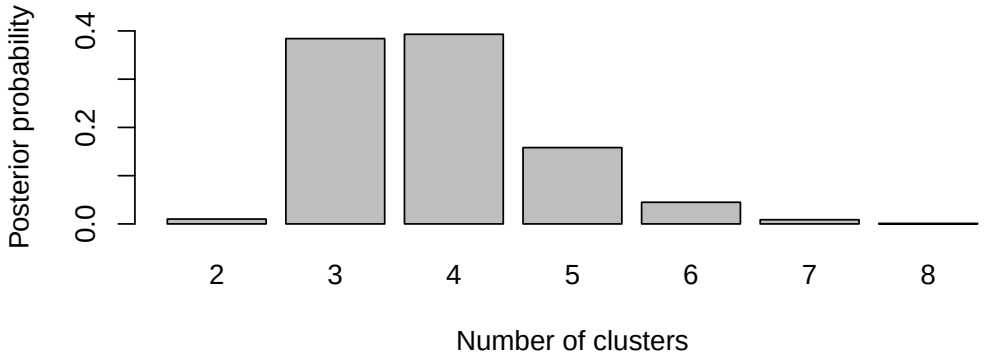


Figure 3.10: Posterior distribution of the number of clusters from our model fit on the AF data.

Posterior inference on the clustering structure of the patients is of clinical interest as it might guide more targeted therapies. Our Gibbs sampler provides posterior samples of the cluster allocation. Here, we report the cluster allocation that minimizes the posterior expectation of Binder's loss function (Binder, 1978) under equal misclassification costs, which is a common choice in the applied Bayesian non-parametrics literature (Lau and Green, 2007). See Appendix B of Argiento et al. (2014) for computational details. Briefly, Binder's loss function measures the difference for all possible pairs of individuals between the true probability of co-clustering and the estimated cluster allocation. In this context, the posterior estimate of the partition of the patients has 4 clusters, with 92% of the patients allocated to 2 clusters which are summarized in Table 3.2.
The larger cluster, Cluster 1, has longer gap and survival times than Cluster 2. Moreover, the Kaplan-Meier curves of each cluster in Figure 3.11 support the conclusion that Cluster 1 includes patients with longer survival times than Cluster 2. As shown in Table 3.2, Cluster 1 has a higher censoring rate than Cluster 2, as one might expect at longer survival times. The lower prevalence of hypertension and slightly lower age of Cluster 1 also confirm that it includes healthier subjects than Cluster 2.

	AF dataset	Cluster 1	Cluster 2
Number of patients	60	44	11
Proportion censored	75%	95%	0%
Average uncensored N_i	2.60 (2.20)	2.00 (0.00)	2.82 (2.52)
Average posterior mean of N_i (SD)	4.76 (3.08)	5.50 (3.06)	2.82 (2.52)
Average uncensored Y_{ij} (SD)	4.18 (1.60)	4.09 (1.60)	4.30 (1.51)
Average posterior mean of Y_{ij} (SD)	4.68 (1.42)	4.68 (1.44)	4.30 (1.51)
Average uncensored $\log(S_i)$ (SD)	5.82 (1.23)	6.76 (0.25)	5.52 (1.30)
Average posterior mean of $\log(S_i)$ (SD)	14.4 (6.15)	17.1 (4.50)	5.52 (1.30)
Average age (SD)	58.2 (11.4)	56.3 (11.8)	62.4 (7.26)
Proportion female	30%	32%	36%
Proportion with paroxysmal AF	7%	7%	9%
Proportion with hypertension	48%	45%	73%
Proportion with heart disease	20%	20%	18%
Proportion on anti-arrhythmic medication	15 %	14%	9%

Table 3.2: Summary statistics of the AF data and the posterior from our model. The two clusters are from a posterior estimate of the cluster allocation that minimizes the posterior expectation of Binder's loss function (Binder, 1978). The averages and standard deviations of posterior means are taken across patients and recurrent events. S_i is recorded in days and Y_{ij} in log days.

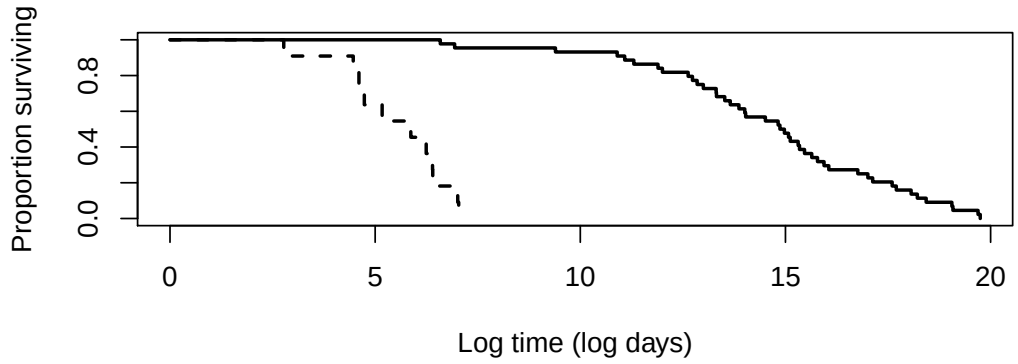


Figure 3.11: Kaplan-Meier survival estimates for the two largest clusters estimated by minimizing the expectation of Binder's loss function (Binder, 1978) under the posterior from our model on the AF data. The solid and dashed lines represent Cluster 1 and 2, respectively. The curves are based on the posterior means of $\log(S_i)$.

Figures 3.12 contains additional posterior results.

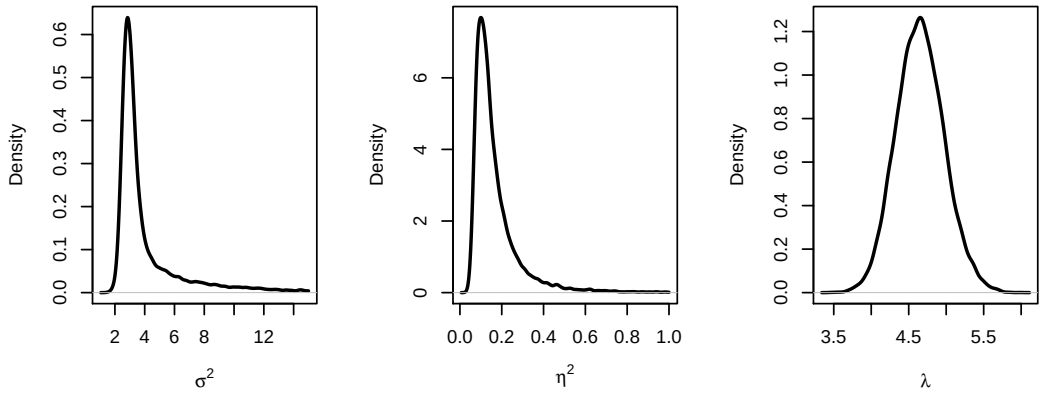


Figure 3.12: Posterior densities for σ^2, η^2 and λ from our model fit on the AF data.

3.5 Comparison with other models

3.5.1 Joint frailty model

We also compare our model with the joint frailty model by Rondeau et al. (2007) as implemented in the R package `frailtypack` (Rondeau et al., 2012b). The model estimates the hazard functions of rehospitalization and mortality jointly using two patient-specific frailty terms, namely u_i and v_i. The frailty term u_i captures the association between rehospitalization and mortality while v_i appears solely in the rehospitalization rate. Specifically, the hazard functions are $r_i(t \mid u_i, v_i, \boldsymbol{\beta}) = u_i\, v_i\, r_0(t)\, \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$ for rehospitalization and $\lambda_i(t|u_i, \boldsymbol{\gamma}) = u_i\, \lambda_0(t)\, \exp(\boldsymbol{x}_i^T \boldsymbol{\gamma})$ for mortality. Here, $r_0(t)$ and $\lambda_0(t)$ are baseline hazard functions, and $\boldsymbol{x}_i$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are defined as in Section 3.2. The random effects distributions are specified as follows: $v_i \sim$ Gamma$(1/\rho,\, 1/\rho)$ and $u_i \sim$ Gamma$(1/\epsilon,\, 1/\epsilon)$ independently for $i = 1, \ldots, L$.

To fit this model to the AF data, we drop hypertension from $\boldsymbol{x}_i$ due to convergence issues when using all covariates. The comparison of the joint frailty model results in Table 3.3 with our results in Figure 3.8 shows that both models obtain similar results and do not detect an association for most of the covariates. Exceptions are the effect of age and use of anti-arrhythmic medication on rehospitalization, and the effects of age and gender on survival. For these, the joint frailty model finds a statistically significant effect.

Finally, the estimate of ρ is 0.005 with a standard error of 0.0009. This suggests heterogeneity between patients that is not explained by the covariates. The estimate of ϵ is 0.032 with a standard error of 0.04. This implies that

the rate of rehospitalizations is positively associated with mortality. These results are in line with the posterior clustering results from our model in Table 3.2 where Cluster 1 is characterized by both the longest gap times and the longest survival times. Probably, given the small size of the dataset, including two random effects here might not be appropriate. This is reflected in the very small random effect variance estimates that were obtained when fitting the model.

Rehospitalization		
Covariate	Hazard ratio	p-value
Age	1.02	0.001
Female	0.95	0.71
Paroxysmal AF	0.76	0.40
Heart disease	1.04	0.82
Antiarrhythmic medication	0.63	0.02
Mortality		
Covariate	Hazard ratio	p-value
Age	1.28	0.02
Female	71.1	0.03
Paroxysmal AF	0.19	0.32
Heart disease	3.95	0.19
Antiarrhythmic medication	1.02	0.99

Table 3.3: Regression coefficients from the joint frailty model fit on the AF data.

3.5.2 Bayesian semi-parametric model from Paulon et al. (2018)

For a more direct comparison, we consider the method proposed by Paulon et al. (2018) as it models the gap and survival times jointly using Bayesian non-parametric priors. Paulon et al. (2018) assume that, conditionally on all parameters and random effects, the gap times are independent of both each other and the survival time. This contrasts with the temporal dependence between gap times in (3.2). Shared random effects induce dependence between different gap times of the same patient. Specifically, Paulon et al. (2018) assume $Y_{ij} \sim \mathcal{N}(\boldsymbol{x}_i^T\boldsymbol{\beta} + \alpha_i, \sigma_i^2)$ independently for $j = 1, \ldots, n_i + 1$

and $i = 1, \ldots, L$, and

$$\log(S_i) \sim \mathcal{N}(\boldsymbol{x}_i^T \boldsymbol{\gamma} + \psi\, \alpha_i,\, \eta^2), \tag{3.19}$$

independently for $i = 1, \ldots, L$, where $\boldsymbol{x}_i$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are defined as in Section 3.2, and α_i and σ_i are random effects. Paulon et al. (2018) do not model the total number of gap times N_i but assume that each patient has a censored $(n_i + 1)$th log gap time $Y_{i(n_i+1)}$. They also assume a priori independence among $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, ψ, η and $(\boldsymbol{\alpha}, \boldsymbol{\sigma}^2)$. The random effects $(\alpha_i, \sigma_i^2) \sim G$ independently for $i = 1 \ldots, L$ where $G \sim \text{DP}(M, G_0)$ with $M \sim \text{Uniform}(a_M, b_M)$ and $G_0 = \mathcal{N}(0, \alpha_0^2) \times \text{Inv-Gamma}(a_\sigma, b_\sigma)$. The priors on $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and η^2 are set as in Subsection 3.2.3. Finally, $\psi \sim \mathcal{N}(0, \psi_0^2)$.

In fitting this model to the AF data, we specify the same $\boldsymbol{x}_i$ and the same hyperparameters for the priors on $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and η^2 as in Subsection 3.4.2. Furthermore, we set $a_M = 0.3$, $b_M = 5$, $\alpha_0^2 = 100$, $a_\sigma = 2.01$, $b_\sigma = 1.01$ and $\psi_0^2 = 100$. This model yields conclusions that are consistent with those from our model. In particular, the posterior distributions on the coefficients in Figure 3.13 closely mimic our results in Figure 3.8.

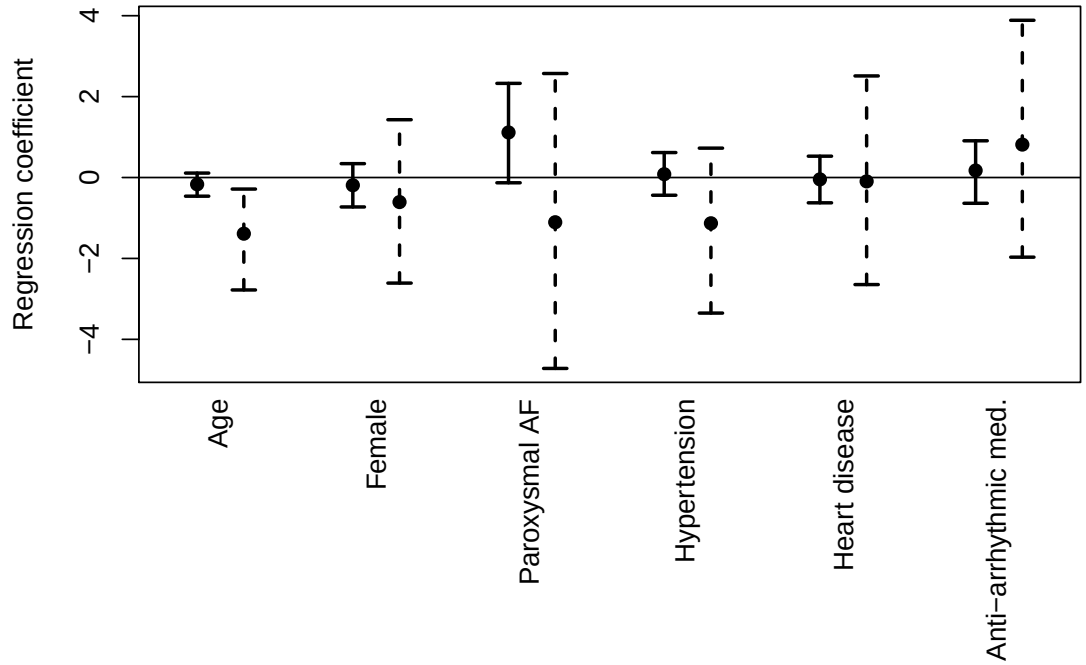


Figure 3.13: Posterior means (dot) and 95% marginal posterior credible intervals (lines) of the regression coefficients from the model in Paulon et al. (2018) fit on the AF data. The solid lines represent credible intervals for the regression coefficients $\boldsymbol{\beta}$ in the gap times model. The dashed lines correspond to the regression coefficients $\boldsymbol{\gamma}$ in (3.19) for the survival times.

Also, the posterior on ψ in (3.19) concentrates between 1.5 and 3.5 per

Figure 3.14.

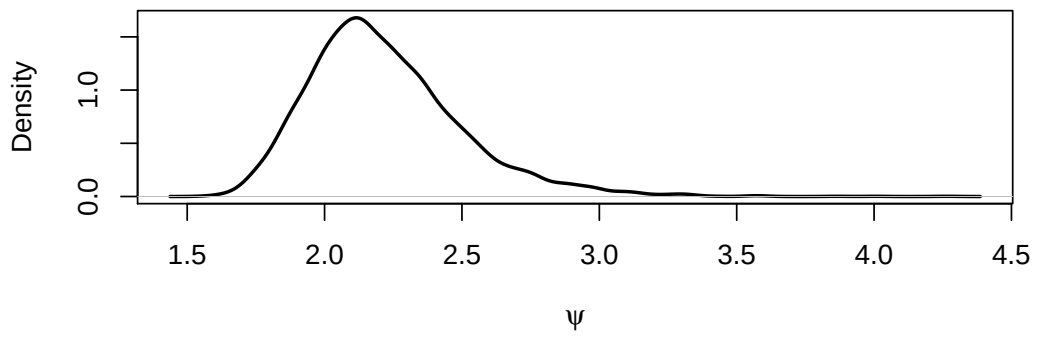


Figure 3.14: Posterior density for ψ in (3.19) from the model in Paulon et al. (2018) fit on the AF data.

This parameter captures the strength of the relationship between gap and survival times. Thus, the time between hospitalizations and survival have a positive association. This is consistent with the clustering results obtained from our model in Table 3.2. Lastly, the posterior on the number of clusters for this model and our model vary slightly, with a mode of 4 clusters for our model in Figure 3.10 of the supporting materials while Figure 3.15 has the mode at 3 for the model from Paulon et al. (2018).

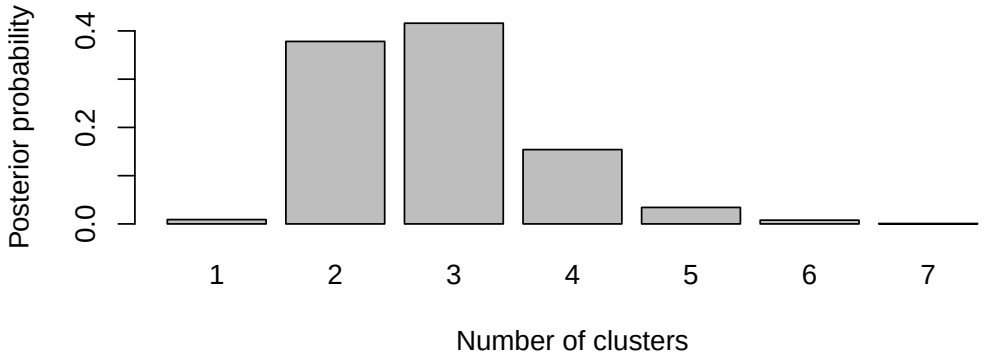


Figure 3.15: Posterior distribution of the number of clusters from the model from Paulon et al. (2018) fit on the AF data.

This is not surprising as our model introduces more structure: a temporal model for the gap time, as well as process-specific frailty terms which are jointly modelled non-parametrically. Moreover, the number of recurrent events is a random quantity and object of inference, which also informs the dependence between the two processes, in addition to the truncation $T_{iN_i} \leq S_i$ of (3.2). In contrast, Paulon et al. (2018) can capture such dependence using only ψ in (3.19), and the gap times are conditionally independent given ψ and the remaining parameters, with the total number of gap times per individual assumed arbitrarily large .

3.6 Conclusions

We have introduced a joint model on recurrence and survival that explicitly treats the number of recurrent events N_i before the terminal event as a random variable and object of inference. The explicit modelling of N_i as well as the specification of a joint distribution for the random effects of the recurrence and survival processes induces dependence between these processes. Moreover, temporal dependence among recurrent events is introduced through a first-order autoregressive process on the gap times. Extension to a more complex temporal structure is in principle straightforward. The model allows for estimating covariates effects on the recurrence and survival processes by introducing appropriate regression terms. Once again, extension to time-varying covariates is not problematic. The use of a nonparametric prior as random effects distribution allows for extra flexibility, patient heterogeneity and data-driven clustering of the patients.

A simulation study showed the effectiveness of our inference on the number of recurrent events N_i. Comparisons with the Cox proportional hazards model, the joint frailty model (Rondeau et al., 2007) and the Bayesian semi-parametric model from Paulon et al. (2018) yielded consistent results. Exceptions were some covariate effects which the Cox proportional hazards model and the joint frailty model detected but our model did not. This discrepancy might be the result of the fact that these models have fewer parameters and assume a single patient population while our model detected multiple sub-populations.

Chapter 4

A comparative review of network meta-analysis models in longitudinal randomized controlled trial

Network meta-analysis (NMA) is a well established technique used to synthetise a complex mass of evidence about a potentially large number of interventions (Borenstein et al. (2010), Higgins et al. (2009), Lipsey and Wilson (2001)). NMA is applied in several research areas, but arguably it plays a pivotal role in health technology assessment (HTA) (Welton et al. (2012)). This is the process by which bodies such as the National Institute for Health and Care Excellence (NICE) in the UK, the Pharmaceutical Benefits Advisory Committee in Australia or Canadian Agency for Drugs and Technologies in Health (CADTH) in Canada assess whether a new intervention provides value for money, or in other words, whether it is "cost-effective". While clinical studies are often primarily designed to assess the clinical effectiveness of a new intervention, possibly against placebo or standard of care, a full economic evaluation requires a comprehensive assessment against all possible relevant comparators. Thus, HTA is often based on a combination of individual- and aggregated-level data informing the model on a possibly large number of comparisons.

One feature that is often relevant in HTA is that individuals are observed at multiple time points throughout the follow up. From the statistical point of view, this implies that the repeated measurements will tend to be correlated and failure to account for this will lead to biased estimate of the underlying treatment effects. Modelling for NMA of repeated measures has been widely considered in the statistical and health economic literature, in recent

times. These methods have been applied in both frequentist and Bayesian frameworks. The Bayesian approach is able to address major statistical challenges associated with NMA, including the total number of trials in a network, the number of trials with more than two comparison arms (introducing an additional layer of complexity), heterogeneity (i.e., clinical, methodological, and statistical variability within direct and indirect comparisons, and across studies), inconsistency (i.e., discrepancy between direct and indirect comparisons), and potential bias that may influence effect estimates. Bucher et al. (1997) have presented an indirect method in meta-analysis that preserves the randomization of the originally assigned patient groups. However, this method allows one to compare two treatments only if a common comparator exists. Lumley (2002) have proposed a more flexible technique but this model can not handle multiple arm studies. An alternative method to avoid this problem is Olkin's linear regression method developed by Cooper et al. (2009a). Wandel et al. (2010) introduced a Bayesian repeated measures network meta-analysis model that allowed the synthesis of multiple time points. Finally, Lu et al. (2007) have proposed Bayesian hierarchical models for mixed treatment comparison using data sets with more than one follow-up time, and where different trials may report at different times. In particular, they focus on the number of individuals that recover over a specified time interval. By dividing the overall period of observation in intervals, they model the number of patients (still) at risk who are recovered over a specified time interval. In this way, they reconstruct the available data on each trial arm into a series of independent observations (over different time periods): each observation then consists of a separate reporting period, and the observations on each period follow a binomial distribution. The main parameter of interest becomes then the probability of recovering over a pre-specified time interval. Moreover, once a patient is recovered he cannot re-experience the event.

In this work, we focus on NMA techniques for longitudinal data, i.e., on the NMA of studies in which individuals are assessed at multiple time points throughout the follow-up period. Follow-up times may not coincide across studies. Repeated measurements for each individual tend to be correlated, which needs to be taken into account to avoid biased estimates of treatment effects. The three most recent methods suggested in the literature for NMA in the case of repeated measurements are the *Mixed Treatment Comparison* (MTC), developed by Dakin et al. (2011); the Bayesian evidence synthesis techniques — integrated two-component prediction (BEST-ITP) developed by Ding and Fu (2013); and the more recent method based on fractional polynomials by Jansen et al. (2015). Our first objective is to provide a comparative

review of these methods using a simulation study as well as real-world data, with a view of identifying specific characteristics that would make either more appropriate. After a comparison of these methods, in Chapter 5, we develop fractional polynomials models within a B-Spline framework.

We note here that all the models are originally specified under a Bayesian approach; this is mainly due to two reasons. Firstly, NMA is a process that inherently combines different (and possibly not completely homogeneous) sources of evidence by decomposing a complex problem in subcomponents which are modelled individually and then linked together via hyper-parameters in a probabilistically sound way. While it is possible to use frequentist methods for inference, a Bayesian approach is recognised as particularly effective as combining "modules", e.g. through the specification of joint prior distributions for some parameters of interest.

Secondly, HTA is ultimately focussed on decision-making, rather than just statistical inference. Again, a Bayesian approach is particularly helpful in this case, for example in terms of allowing a straightforward quantification of the impact of model and parameter uncertainty on the optimal decision, given current data (a process often termed "Probabilistic Sensitivity Analysis", which is usually mandatory in many jurisdictions. This process is required by many regulatory agencies (see Adalsteinsson and Toumi, 2013) and would simply translate in adding extra layers to the model hierarchy (see Berger, 2013; Baio and Dawid, 2011).

This Chapter is structured as follows. In Section 4.1 we briefly review the general set up and distributional assumptions for the three models; then in Section 4.2 we present a simulation study aimed at testing the performance of the methods under a range of scenarios in terms of the underlying treatment effect. In Section 4.3, Section 4.4 and Section 4.5 we illustrate the models using three real-world datasets, one collecting information on patients affected by chronic obstructive pulmonary disease (COPD), the other one investigating treatments for osteoarthritis(OA) of the knee and the last one involving diabetes patients. In Section 4.6 we summarise our conclusions and recommendations.

4.1 Models

In this Section we briefly review the three main models available in the literature to perform network meta-analysis of longitudinal studies. In this context the main objective is the evaluation over time of some suitably defined response, e.g. continuous outcomes, binary or count data. Let y_{sjt} denote the

response in study s, at time t in the jth treatment arm. We assume:

$$y_{sjt} \sim f\left(y_{ij}|\theta_{sjt}, \sigma^2_{sjt}\right) \tag{4.1}$$

where f is a probability density (usually parametric), $s = 1, \ldots, ; S, j = 1, \ldots, J$ and $t = 0, \ldots, T$. The main parameter of clinical interest is θ_{sjt}, which measures the treatment effect in study s of intervention j at time point t, on a suitable scale (e.g., logit in the case of binary variables or log when f is a Poisson distribution).

In this work we focus on continuous and binary responses as these are the most common outcomes in clinical trials. When the measurements y_{sjt} are continuous Equation (4.1) becomes:

$$y_{sjt} \sim \text{Normal}\left(\theta_{sjt}, \sigma^2_{sjt}\right),$$

while in case of a binary outcome the likelihood function becomes:

$$\begin{aligned} y_{sjt} &\sim \text{Binomial}\left(n_{sjt}, p_{sjt}\right), \\ g(p_{sjt}) &= \text{logit}(p_{sjt}) = \theta_{sjt}. \end{aligned}$$

Other alternatives include $g(\theta_{sjt}) = \ln(\theta_{sjt})$ if f is modeled as a Poisson distribution.

The fundamental difference among the three NMA methods is in the way in which the predicted (mean) outcome (θ_{sjt}) and the variance (σ^2_{sjt}), at time point t in arm j of trial s, are specified.

4.1.1 Mixed treatment comparison model

In the MTC model the mean outcome is specified in terms of a study- and time- specific baseline and an additive term describing the relative effect of each treatment. For simplicity, we set as the reference arm the treatment j = 1. The original specification develped by Dakin et al. (2011) presents 6 alternative versions of this general structure, where both the baseline and the relative effect may vary over time. Here, we consider a general formulation

$$\theta_{sjt} = \mu_{st} + \delta_{sj},$$

where, μ_{st} is the study- and time-specific effect pooled across treatment arms, while δ_{sj} is the study-specific arm deviation from the reference arm, with $\delta_{s1} := 0$. In this framework the parameters μ_{st} are assumed to be independent across time points in each study and the relative treatment effects δ_{sj} to be

constant over time.

In particular, μ_{st} is assumed to be normally distributed for each time-point and each study. Conversely, the parameters δ_{sj} are modelled as structured "random effects", as follows

$$\delta_{sj} \sim \text{Normal}\left(d_j - d_1, \sigma_\delta^2\right),$$

where d_j indicates the treatment effects of treatment j relative to reference treatment. Following Dakin et al. (2011), the parameters d_j are independently assigned vague Normal prior distribution centered on 0 and with large variance, whereas the between-studies standard deviation σ_δ^2 is given a uniform prior on a wide interval.

Finally, the variances for the observed data are modelled as

$$\sigma_{sjt}^2 = \left(\frac{\text{sd}_{sjt}}{\sqrt{n_{sjt}}}\right)^2 \qquad \text{for all } s, j, t,$$

where sd_{sjt} is an estimate of the standard deviation (possibly available in the literature) and n_{sjt} is the observed sample size for each treatment arm and time point in each study.

4.1.2 Bayesian evidence synthesis techniques - integrated two-component prediction model

Ding and Fu (2013) present two different models: one is based on fixed effects, while the other includes random effects. The general model with fixed effects specifies the mean and the variance of the observed outcome as

$$\begin{aligned} \theta_{sjt} &= (\phi_s + \delta_j)\left(\frac{1 - e^{p_j t}}{1 - e^{p_j T}}\right) \\ \sigma_{sjt}^2 &= \left[\frac{\text{sd}_{sjt}}{\sqrt{n_{sjt}}}\left(\frac{1 - e^{p_j t}}{1 - e^{p_j T}}\right)\right]^2, \end{aligned} \tag{4.2}$$

where the parameter δ_j indicates the j-th treatment mean effect at the end follow up period (time T). Here, the time period T is the same for all studies. It is straightforward to allow different time periods for different studies, by changing T to T_s. Moreover, the parameter p_j determines the shape of the j-th treatment effect over time. This parameter is assigned a uniform prior distribution defined on a negative interval. This implies that:

$$0 \leq \left(\frac{1 - e^{p_j t}}{1 - e^{p_j T}}\right) \leq 1.$$

More specifically, if p_j is large the rate of the change attributable to treatment j grows fast at the beginning and reaches a plateau quickly, whereas when p_j is close to 0, the trend is almost linear.

In the fixed effect specification, δ_j is assigned a vague Normal prior again centered on 0 and with large variance. The random effects version can be easily obtained by assuming a study-specific treatment effect δ_{sj} in Equation 4.3, e.g.,

$$\delta_{sj} \sim \text{Normal}\left(d_j - d_1, \sigma_\delta^2\right),$$

where the parameters d_j are assigned vague Normal priors and the between-studies standard deviation is given a wide uniform prior. Notice that in our simulations in Section 4.2 and examples in Section 4.3–4.5 we use the random effect model for a fairer comparison.

4.1.3 Fractional polynomial model

Jansen et al. (2015) propose an approach to network meta-analysis based on fractional polynomials (FPs), a family of flexible basis functions used to describe the relationships among variables, e.g., in a regression setting. This model assumes nonlinear dynamics of treatment effects over time.

A FP structure of order M for the mean outcome of the j-th treatment in study s is defined as

$$\theta_{sjt} = \begin{cases} \beta_{0sj} + \sum_{m=1}^{M} \beta_{msj} t^{p_m} & \text{if } p_1 \neq \cdots \neq p_M \\ \beta_{0sj} + \beta_{1sj} t^p + \sum_{m=2}^{M} \beta_{msj} t^p \left[\ln(t)\right]^{m-1}, & \text{if } M > 1, p_1 = \cdots = p_M = p \end{cases} \tag{4.3}$$

with $t^0 := \ln(t)$. Jansen et al. (2015) suggest selecting the power p_m from the set {-2, -1, -0.5, 0, 0.5, 1, 2, 3}, for $m = 1, \ldots, M$.

The vector $\boldsymbol{\beta_{sj}} = (\beta_{0sj}, \ldots, \beta_{Msj})^\intercal$ is modelled as

$$\begin{pmatrix} \beta_{0sj} \\ \vdots \\ \beta_{Msj} \end{pmatrix} = \begin{pmatrix} \mu_{0s} \\ \vdots \\ \mu_{Ms} \end{pmatrix} + \begin{pmatrix} \delta_{0sj} \\ \vdots \\ \delta_{Msj} \end{pmatrix},$$

where the vector $\boldsymbol{\mu}_s = (\mu_{0s}, \ldots, \mu_{Ms})^\intercal$ denotes the study specific mean, while the vector $\boldsymbol{\delta}_j = (\delta_{0sj}, \ldots, \delta_{Msj})^\intercal$ represents the study- and time-specific effects of treatment j relative to the reference treatment, arbitrarily coded as $j = 1$. The parameters δ_{msj} are modelled assuming a multivariate Normal distribution with the pooled estimates expressed in terms of the overall reference

treatment $j = 1$:

$$\begin{pmatrix} \delta_{0sj} \\ \vdots \\ \delta_{Msj} \end{pmatrix} \sim \text{Normal} \left(\begin{pmatrix} d_{0j} - d_{01} \\ \vdots \\ d_{Mj} - d_{M1} \end{pmatrix}, \Sigma \right), \tag{4.4}$$

where the vector $\boldsymbol{d}_j = (d_{0j}, \ldots, d_{Mj})^\intercal$ are assigned vague Normal priors, for $j > 1$, and $\boldsymbol{d}_1 = \mathbf{0}$ and with

$$\begin{pmatrix} \delta_{0s1} \\ \vdots \\ \delta_{Ms1} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \qquad \text{and} \qquad \Sigma = \begin{pmatrix} \sigma_0^2 & \cdots & \sigma_0 \sigma_M \lambda_{0M} \\ \vdots & \ddots & \vdots \\ \sigma_0 \sigma_M \lambda_{0M} & \cdots & \sigma_M^2 \end{pmatrix}.$$

The covariance matrix Σ captures between-study heterogeneity of the treatment effect parameters δ_{msj}, where $\sigma_m^2 = \text{Var}[\delta_{msj}]$ and λ_{lk} quantifies the correlation across the treatment effect parameters, with $m, l, k \in \{0, \ldots, M\}$.

Under a fixed-effects model, the formulation in (4.4) is replaced by

$$\begin{pmatrix} \delta_{0sj} \\ \vdots \\ \delta_{Msj} \end{pmatrix} = \begin{pmatrix} d_{0j} - d_{01} \\ \vdots \\ d_{Mj} - d_{M1} \end{pmatrix}$$

and as a result it is not necessary to estimate the between-study covariance matrix.

Finally, the variance of the main outcome

$$\sigma_{sjt}^2 = \left[\left(\frac{\text{sd}_{sjt}}{\sqrt{n_{sjt}}} \right) \Big/ \left(1 - \rho^2\right) \right]^2$$

is modelled as a function of the corresponding standard error adjusted by a factor $(1 - \rho^2)$, where ρ indicates the correlation coefficient between subsequent time points. Often, ρ is assumed to be known, but in a fully Bayesian setting it can be treated as unknown and object of inference, by specifying an appropriate prior distribution.

4.2 Simulation Study

In order to evaluate the performance of the three models described above, we conduct several simulations studies. We investigate different scenarios to highlight the main differences among the modelling strategies. Our goal is to reproduce different time-pattern for the main effect and investigate the ability

of each model to recover such structure. We consider treatment effects that over time: *a)* are linear; *b)* decrease logarithmically; *c)* are constant and *d)* non-monotonic treatment effect, corresponding to a situation in which a treatment could be first beneficial and then detrimental. Moreover, to provide fair comparisons, we also generate data based on the models by Dakin et al. (2011) and Ding and Fu (2013). Finally, we extend a continuous outcome for longitudinal studies to a binary outcomes (See Subsection 4.2.4). Firstly, we consider "closed" network, meaning that there is at least one study providing direct evidence for all possible pairwise comparisons. After that, in Subsection 4.2.5 we consider also a non-closed network to investigate if the nature of the network influence the conclusion of the comparison.
For each scenario, we randomly generate 50 datasets according to the following specifications. We then fit the NMA models on each dataset and average the results over the 50 simulated datasets. When generating the data we follow a similar simulation strategy as in Ding and Fu (2013).

For all scenarios with closed network, we simulate data from 3 hypothetical studies with 2 treatment arms. Figure 4.1 shows graphically the network of studies used for these scenarios; circles indicate treatments, while edges connecting them indicate the availability of direct evidence as provided by a study. In particular, in Figure 4.1 study 1 compares treatment A and B; study 2 has treatments A and C and study 3 matches B and C.

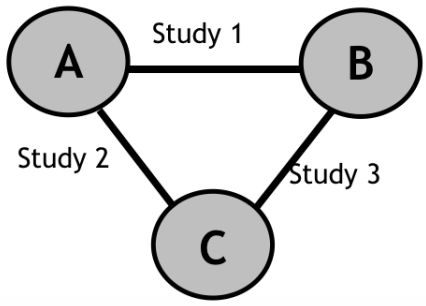


Figure 4.1: Network of studies used in the simulation scenarios 1–7. Treatments are represented as nodes and connected by an edge if a direct comparison study between them is available. The network is closed, with one study comparing all possible pairs of treatments.

The table below (4.1) describes the details of the simulations in terms of the time points at which follow up occurs as well as the number of individuals in each of the studies arms.

Study	Treatment	Time of Observation	Number of Subjects per arm (n_s)
Study 1	A and B	4, 8, 12, and 24 weeks	100
Study 2	A and C	4, 12, and 24 weeks	120
Study 3	B and C	4, 8, and 12 weeks	130

Table 4.1: Simulation parameters used in the simulation scenarios 1–7. For each study different treatments, times of observation and number of subjects per arm are specified.

For the first five scenarios (**Scenario 1–5**), to generate data from a meta-analysis of study, we first simulate individual level observations as $Z_{isjt} \sim$ Normal(θ_{sjt}, τ_s^2), for $i = 1, \ldots, n_s$, where n_s is the number of observation in study s. In addition, we set $\theta_{sjt} = \alpha_s + \gamma_{jt}$ with $\alpha_s \sim$ Normal$(0, 10)$, $\tau_1^2 = 1$, $\tau_2^2 = 2$ and $\tau_3^2 = 4$ and we show γ_{jt}, the true relationships between time and the change from baseline, in Figure 4.2. In the simulations, we assume that the variance of the outcome is constant over time and across treatments as our main interest is to investigate the performance of each model in capturing the time structure of the main outcome. Obviously this assumption may not be tenable in many real data applications as shown in Section 4.3, in Section 4.4 and in Section 4.5.
The main outcome of each study is reported as a sample mean $Y_{sjt} = \frac{\sum_{i=1}^{n_s} Z_{ijst}}{n_s}$, or $Y_{sjt} \sim$ Normal $(\theta_{sjt}, \sigma_s^2)$, with $\sigma_s^2 = \tau_s^2/n_s$. In other words, this setting amounts to including a set of independent study-effects α_s, while the component γ_{jt} in θ_{jst} represents the effect of the treatment over time.

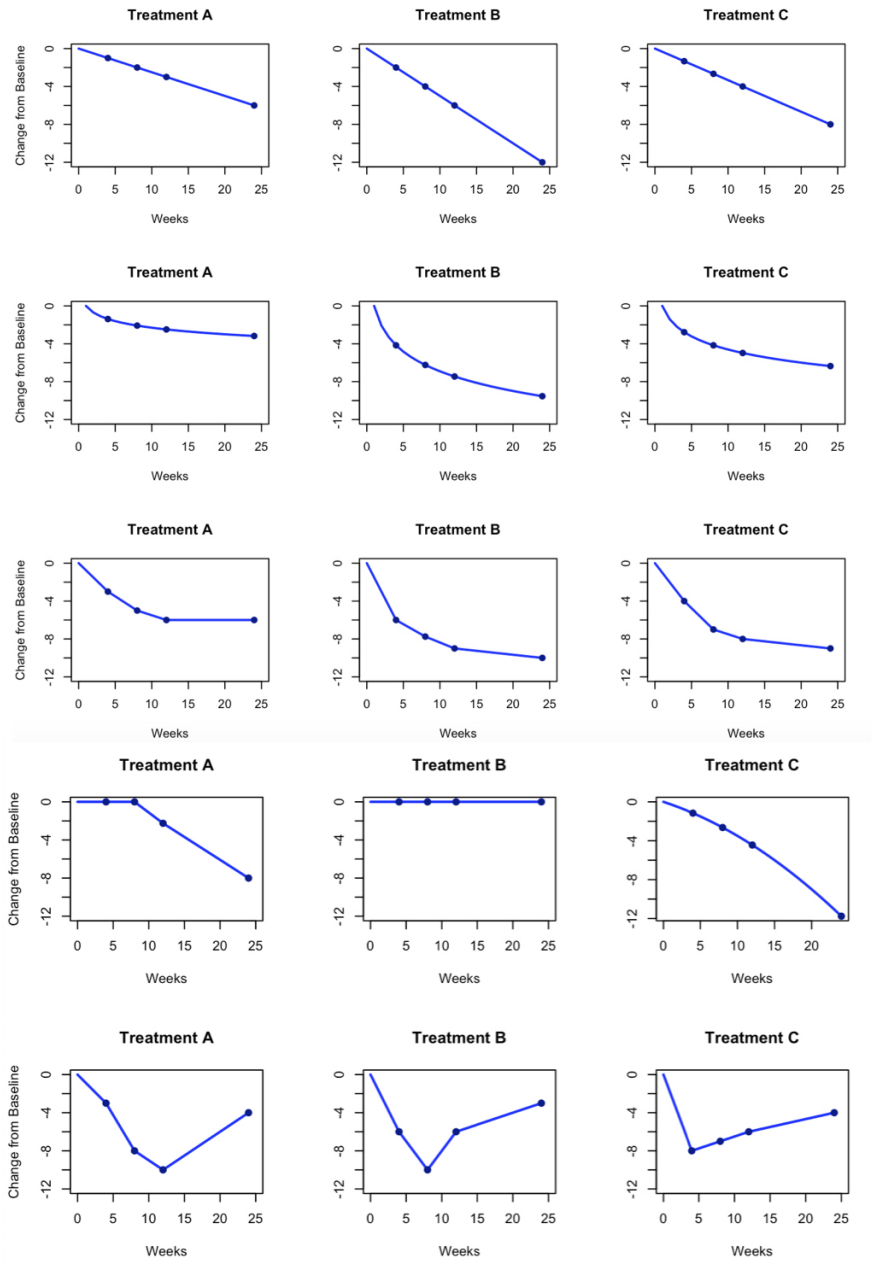


Figure 4.2: Temporal patterns for treatment effects of treatments A, B and C used in the simulations 1–5. Blue lines represent the treatment effect, γ_{jt}, over time. The columns refer to different treatments (A, B, C), while the rows represents the different simulation scenarios.

Through the different specifications of the curve γ_{jt} we create the various scenarios described at the beginning of this Section, according to the following scheme.

Scenario 1. The treatment effects are linear over the time with different slopes:

- Treatment A: $\gamma_{At} = -\frac{t}{4}$;

- Treatment B: $\gamma_{Bt} = -\frac{t}{2}$;
- Treatment C: $\gamma_{Ct} = -\frac{t}{3}$.

Scenario 2. We consider treatment effects with logarithmic decay over time:

- Treatment A: $\gamma_{At} = -\log(t+1)$;
- Treatment B: $\gamma_{Bt} = -3\log(t+1)$;
- Treatment C: $\gamma_{Ct} = -2\log(t+1)$;

Scenario 3. We assume a monotonic piecewise linear curves:

- Treatment A: $\gamma_{At} = \begin{cases} -\frac{3t}{4} & \text{if} \quad 0 \leq t \leq 4 \\ -\frac{t}{2} - 1 & \text{if} \quad 4 < t \leq 8 \\ -\frac{t}{4} - 3 & \text{if} \quad 8 < t \leq 12 \\ -6 & \text{if} \quad 12 < t \leq 24 \end{cases}$;
- Treatment B: $\gamma_{Bt} = \begin{cases} -\frac{3t}{2} & \text{if} \quad 0 \leq t \leq 4 \\ -\frac{10t}{23} - \frac{17}{4} & \text{if} \quad 4 < t \leq 8 \\ -\frac{5t}{16} - \frac{21}{4} & \text{if} \quad 8 < t \leq 12 \\ -\frac{t}{12} - 8 & \text{if} \quad 12 < t \leq 24 \end{cases}$;
- Treatment C: $\gamma_{Ct} = \begin{cases} -t & \text{if} \quad 0 \leq t \leq 4 \\ -\frac{3t}{4} - 1 & \text{if} \quad 4 < t \leq 8 \\ -\frac{t}{4} - 5 & \text{if} \quad 8 < t \leq 12 \\ -\frac{t}{12} - 7 & \text{if} \quad 12 < t \leq 24 \end{cases}$.

Scenario 4. We introduce also a treatment effect that is constant over the time (treatment B):

- Treatment A: $\gamma_{At} = \begin{cases} 0 & \text{if} \quad 0 \leq t \leq 8 \\ -\frac{t}{2} + 4 & \text{if} \quad 8 < t \leq 24 \end{cases}$;
- Treatment B: $\gamma_{Bt} = 0$;
- Treatment C: $\gamma_{Ct} = \frac{(-t^2 - 25t)}{100}$.

Scenario 5. We assume non-monotonic piecewise linear curves:

- Treatment A: $\gamma_{At} = \begin{cases} -\frac{3t}{4} & \text{if} \quad 0 \leq t \leq 4 \\ -\frac{5t}{2} + 2 & \text{if} \quad 4 < t \leq 8 \\ -\frac{t}{2} - 4 & \text{if} \quad 8 < t \leq 12 \\ \frac{t}{2} - 16 & \text{if} \quad 12 < t \leq 24 \end{cases}$;

- Treatment B: $\gamma_{Bt} = \begin{cases} -\frac{3t}{2} & \text{if} \quad 0 \leq t \leq 4 \\ -t-2 & \text{if} \quad 4 < t \leq 8 \\ t-18 & \text{if} \quad 8 < t \leq 12 \\ \frac{t}{4}-9 & \text{if} \quad 12 < t \leq 24 \end{cases}$;
- Treatment C: $\gamma_{Ct} = \begin{cases} -\frac{8t}{4} & \text{if} \quad 0 \leq t \leq 4 \\ \frac{t}{4}-9 & \text{if} \quad 4 < t \leq 12 \\ \frac{t}{6}-8 & \text{if} \quad 12 < t \leq 24 \end{cases}$.

For **Scenario 6**, we simulate data from the MTC model with the following parameters:

$$\begin{aligned} \mu_{st} &\sim \text{Normal}\,(-s, 1)\,, \\ \delta_{sj} &\sim \text{Normal}\,(0.5, 1)\,, \\ sd &= 1.2. \end{aligned}$$

Finally, in the last scenario, **Scenario 7**, we simulate data from the BEST-ITP model with the following parameters:

$$\begin{aligned} \phi_s &\sim \text{Normal}\,(-3, 1)\,, \\ \delta &= (0, -0.5, -1), \\ sd &= 1.2, \\ p_j &= (-0.1, -0.15, -0.15). \end{aligned}$$

Posterior inference for these examples, as well as for the real data applications in Section 4.3, in Section 4.4 and in Section 4.5, can be performed through a standard Gibbs sampler algorithm, which we implement in JAGS, see Plummer (2003). The first 5,000 iterations are discarded as 'burn-in' and the final sample size on which inference is based is 10,000 samples. We check convergence of the chain through the Gelman-Rubin potential scale reduction factor (Gelman and Rubin (1992)). The Gelman-Rubin diagnostic is evaluated by running multiple chains from different initial values and then comparing the estimated between-chains and within-chain variances for each model parameter. Large differences between these variances indicate that convergence has not been reached yet. In particular, Gelman and Rubin (1992) suggest that diagnostic values greater than 1.2 for any of the model parameters should indicate nonconvergence. In practice, a more stringent rule of diagnostic values < 1.1 is often used to declare convergence.

4.2.1 Models specification

For all the methods, we generally specify vague prior distributions for the parameters of interest. In the case of the MTC we follow the original proposal presented in Dakin et al. (2011) and assume

$$\mu_{st} \sim \text{Normal}\left(0, 10^4\right), \qquad d_{s_j} \sim \text{Normal}\left(0, 10^4\right) \qquad \text{and} \qquad \sigma_\delta \sim \text{Uniform}(0, 20).$$

Following Ding and Fu (2013), for the BEST-ITP we assume

$$\phi_s \sim \text{Normal}\left(0, 10^4\right) \qquad \text{and} \qquad p_j \sim \text{Uniform}(-10, 0).$$

Moreover, we follow the suggestion of the previous method (MTC model) to specify the random effect distribution as

$$d_j \sim \text{Normal}\left(0, 10^4\right) \qquad \text{and} \qquad \sigma_\delta \sim \text{Uniform}(0, 20).$$

Finally, regarding the FP approach, in the simulation study we investigate the performance of a first- and a second-order fractional polynomial model ($M = 1, 2$, respectively). In fact, in medical applications, FP1 and FP2 transformations are often used, with higher order transformations being used rarely. FP1 and FP2 functions allow representation of a wide range of non-linear relationships. A first-order fractional polynomial is obtained by describing the effect as a function of transformed time t in a linear model: FP of degree 1 functions are always monotonic, while FP degree 2 functions may be monotonic or unimodal. For greater detail, see Royston and Sauerbrei (2008).

When $M = 1$, Equation 4.3 reduces to

$$\theta_{sjt} = \beta_{0sj} + \beta_{1sj} t^p, \tag{4.5}$$

while for $M = 2$ it becomes

$$\theta_{sjt} = \begin{cases} \beta_{0sj} + \beta_{1sj} t^{p_1} + \beta_{2sj} t^{p_2} & \text{if } p_1 \neq p_2 \\ \beta_{0sj} + \beta_{1sj} t^p + \beta_{2sj} t^p \left[\ln(t)\right], & \text{if } p_1 = p_2 = p. \end{cases} \tag{4.6}$$

For consistency with the other models, we consider only specifications that include a random effect. In particular, when $M = 1$, there is only one between-study heterogeneity parameter, σ_0^2, related to the relative treatment effects for β_{0sj}. In our simulations, we consider another separate analysis where the between-study heterogeneity is assumed to affect treatment effects only in

terms of β_{1sj}. Moreover, if $M = 2$, we consider also the relative treatment effects for β_{2sj}. Here we assign the relevant parameters σ_m a uniform distribution on the intervall $(0, 10)$.

All the models are completed by specifying suitable prior distributions for the vectors $\boldsymbol{\mu}_s$ and $\boldsymbol{d}_j = (d_{0j}, \ldots, d_{Mj})^\intercal$. In particular, we define $\mathbf{0}$ as the zero vector of length $(M + 1)$ and model

$$\boldsymbol{\mu}_s \sim \text{Normal}\,(\mathbf{0}, \boldsymbol{T}_\mu) \qquad \text{and} \qquad \boldsymbol{d}_j \begin{cases} = \mathbf{0} & \text{for } j = 1 \\ \sim \text{Normal}\,(\mathbf{0}, \boldsymbol{T}_d) & \text{for } j > 1 \end{cases}.$$

The diagonal elements of the prior covariance matrices are set equal to 10^4, while the other elements are equal to zero. Thus, $\boldsymbol{T}_\mu = \boldsymbol{T}_d = 10^4\mathbb{I}$, with $\mathbb{I}$ denoting the identity matrix of appropriate dimension. Finally, we assume that the correlation ρ follows a Uniform distribution in the interval $(0, 0.95)$.

4.2.2 Selection of M and p_m in the FP model

Implementing the FP model requires selecting a value for M and fixing the power terms p or (p_1, p_2) in Equations (4.5) or (4.6), respectively. Following Jansen et al. (2015), we use the Deviance Information Criterion (DIC), originally proposed by Spiegelhalter et al. (2002) as an estimate of expected predictive error to guide the selection of the FP order and power terms, upon a set of possible choices. The DIC is defined as the sum of $\bar{D}$, the posterior mean of the model deviance and p_D, a posterior estimate of the effective number of parameters, acting as a penalty for more complex models in order to limit the effect of overfitting. Models with smaller DIC should be preferred. For each of the 50 replicates of each scenario, we compare the values of DIC of the different first-order fractional polynomial models (that is 8 different models, one for each value of p in our pre-specified range) and of the different second-order fractional polynomial models (35 models in total, given by all possible combinations of p_1 and p_2). For each scenario, we select M and corresponding power terms based on the DIC value. In Table 4.2 we report, as example, the power that achieves minimum DIC more often corresponding to the selected polynomial order M and the mean DIC evaluated over those models for the Scenario 3 and Scenario 5.

	Scenario 3		Scenario 5		
Random effects on	β_{0sj}	β_{1sj}	β_{0sj}	β_{1sj}	β_{2sj}
Power that achieves minimum DIC more often	$p=-1$	$p=-1$	$p_1=-2$, $p_2=0$	$p_1=-2$, $p_2=0$	$p_1=-2$, $p_2=0$
Mean of DIC	49.946	39.710	47.676	44.668	49.976
p_D	21.8	21.9	27.0	28.8	26.0

Table 4.2: DIC selection of the order M and of the power terms p_m in the FP model for Scenario 3 and Scenario 5. Shown are mean of DIC, mode of the power terms and posterior estimate of the effective number of parameters (p_D) obtained over 50 simulations for each scenario.

For scenario 3 the "best" models are associated with $M = 1$ and thus we only report the estimated values for the power term p; for scenario 5, the model fitting the data best among those we have tested is associated with $M = 2$ and thus we report the estimates for both p_1 and p_2. In each scenario, we also report the results upon varying the parameter on which we place the random effect structures. From the inspection of Table 4.2, it is possible to see that the DIC is consistently smaller for the models including a random effects on β_{1sj}.

4.2.3 Results

To compare the results obtained with the three different NMA methods we display estimated profiles of each treatment from each study obtained for each scenario. Figures 4.3 -4.9 show the estimated profiles obtained for the MTC, BEST-ITP and the appropriate FP models for all Scenarios. Red lines indicate the true values used to simulate the data, whereas magenta, green and blue lines indicate the estimates obtained using the MTC model, BEST-ITP model and fractional polynomials model, respectively.

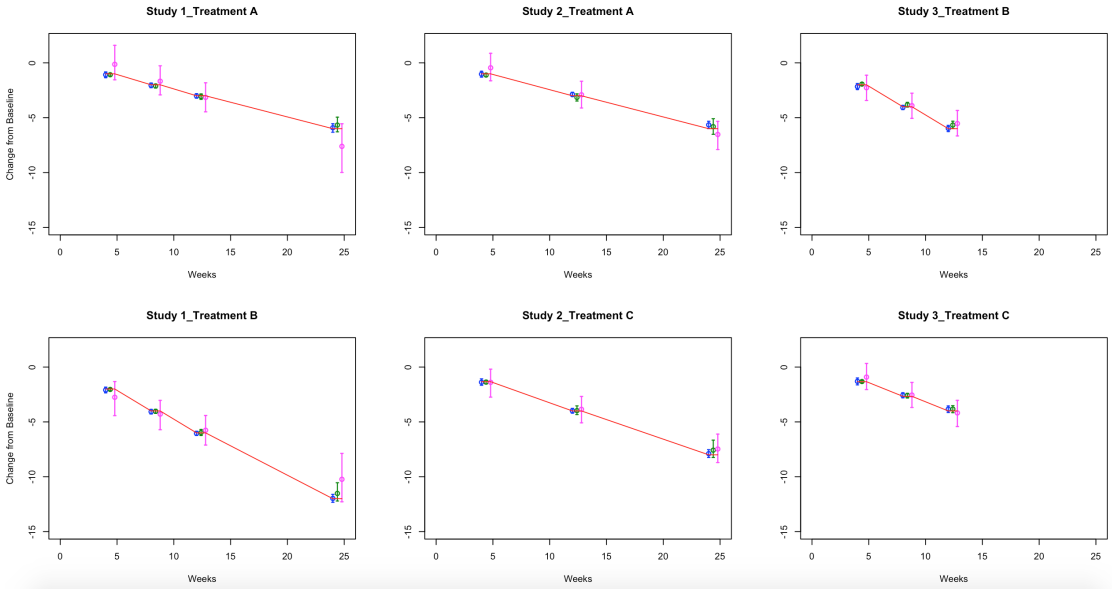


Figure 4.3: *Scenario 1*: posterior estimated profiles obtained. Red lines indicate the values used to simulate the data, whereas magenta lines, green lines and blue lines correspond to estimates obtained using the MTC model, BEST-ITP model and fractional polynomials model, respectively, along with 95% credible intervals.

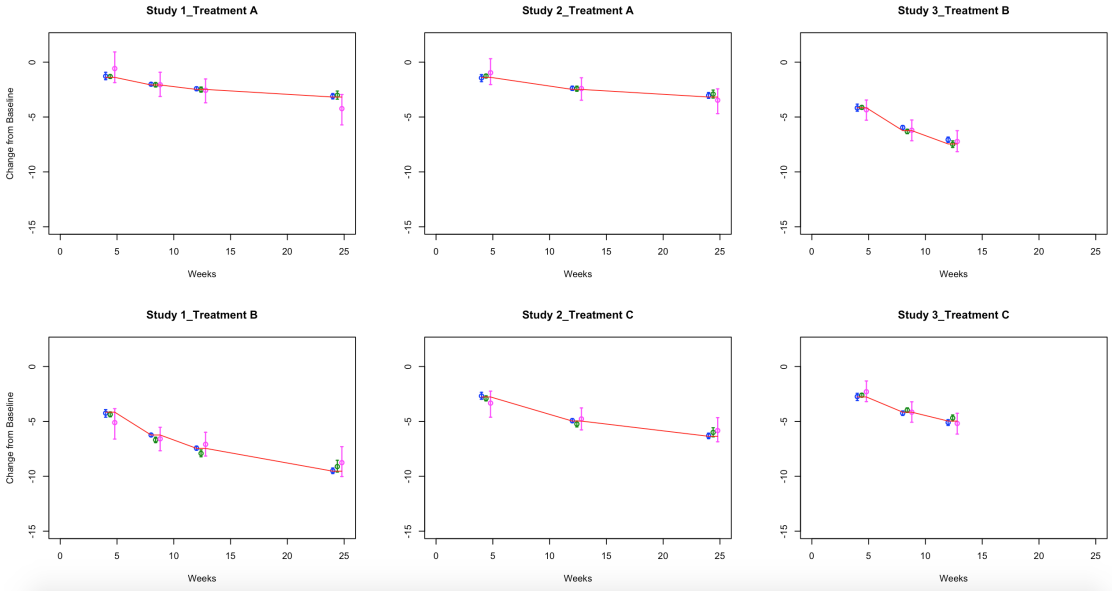


Figure 4.4: *Scenario 2*: posterior estimated profiles obtained. Red lines indicate the values used to simulate the data, whereas magenta lines, green lines and blue lines correspond to estimates obtained using the MTC model, BEST-ITP model and fractional polynomials model, respectively, along with 95% credible intervals.

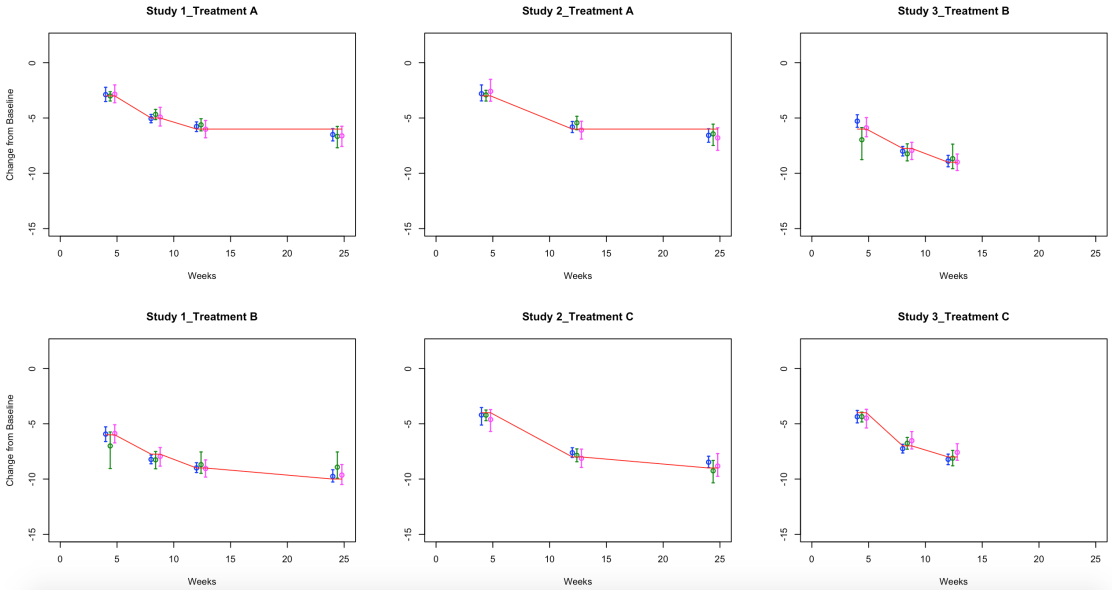


Figure 4.5: *Scenario 3*: posterior estimated profiles obtained. Red lines indicate the values used to simulate the data, whereas magenta lines, green lines and blue lines correspond to estimates obtained using the MTC model, BEST-ITP model and fractional polynomials model, respectively, along with 95% credible intervals.

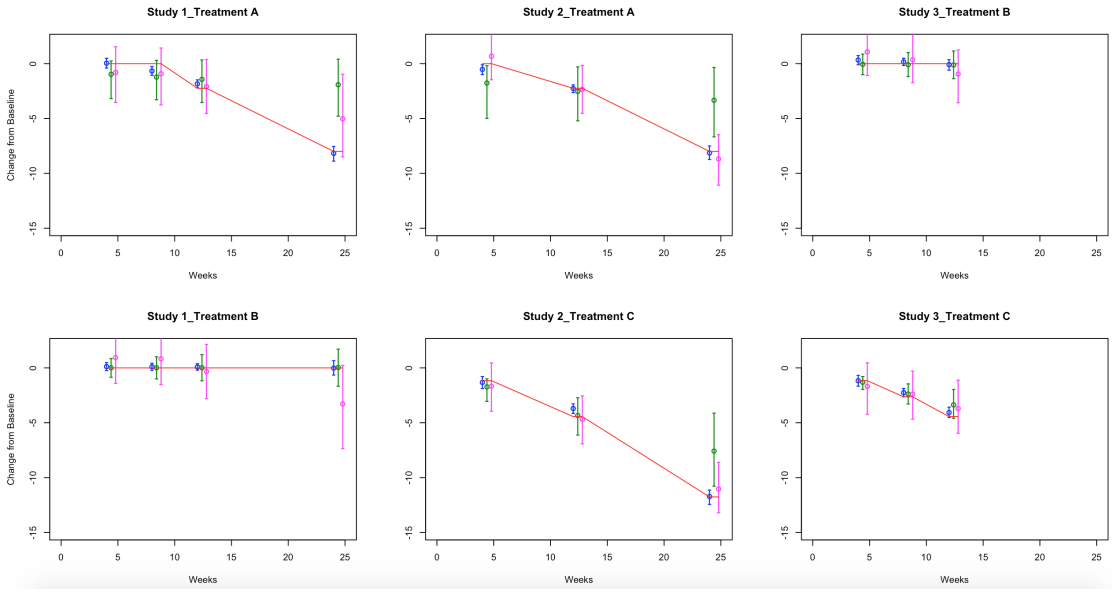


Figure 4.6: *Scenario 4*: posterior estimated profiles obtained. Red lines indicate the values used to simulate the data, whereas magenta lines, green lines and blue lines correspond to estimates obtained using the MTC model, BEST-ITP model and fractional polynomials model, respectively, along with 95% credible intervals.

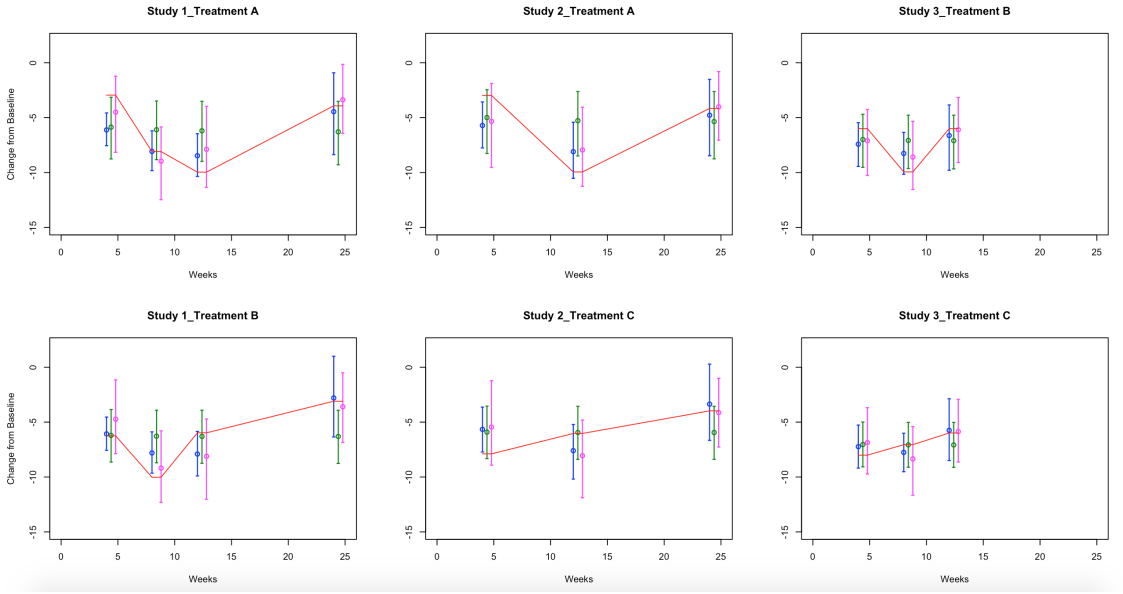


Figure 4.7: *Scenario 5*: posterior estimated profiles obtained. Red lines indicate the values used to simulate the data, whereas magenta lines, green lines and blue lines correspond to estimates obtained using the MTC model, BEST-ITP model and fractional polynomials model, respectively, along with 95% credible intervals.

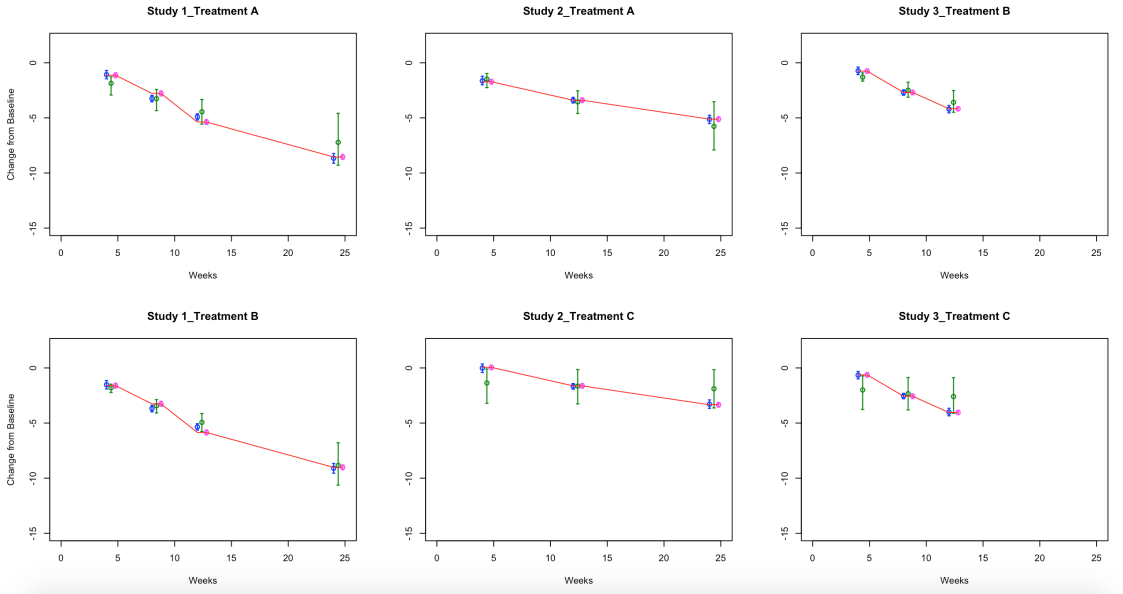


Figure 4.8: *Scenario 6*: posterior estimated profiles obtained. Red lines indicate the values used to simulate the data, whereas magenta lines, green lines and blue lines correspond to estimates obtained using the MTC model, BEST-ITP model and fractional polynomials model, respectively, along with 95% credible intervals.

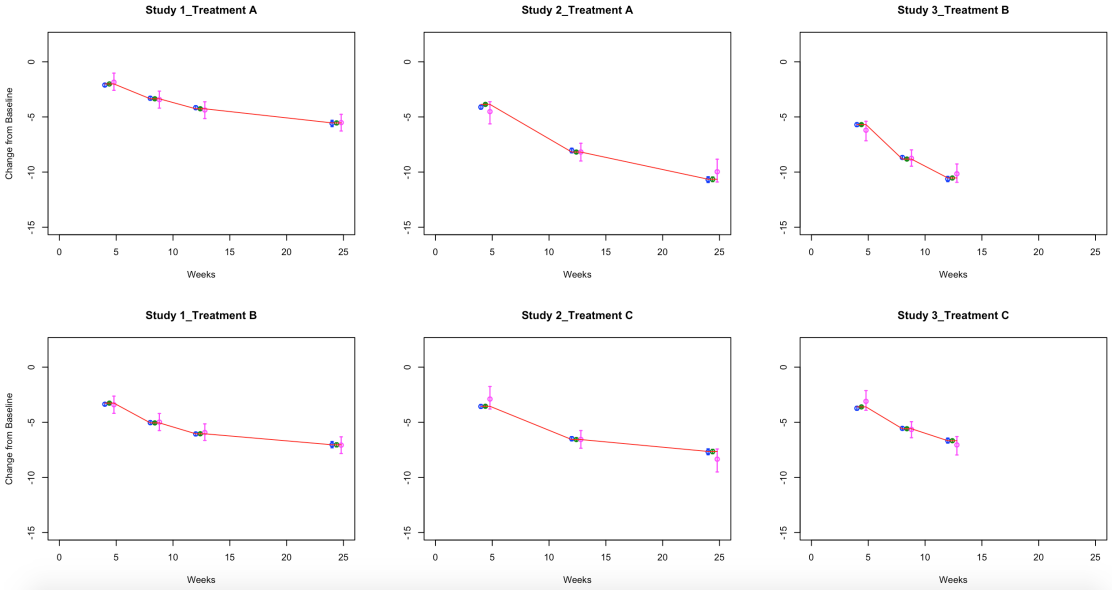


Figure 4.9: *Scenario 7*: posterior estimated profiles obtained. Red lines indicate the values used to simulate the data, whereas magenta lines, green lines and blue lines correspond to estimates obtained using the MTC model, BEST-ITP model and fractional polynomials model, respectively, along with 95% credible intervals.

Upon visual inspection, for all Scenarios it appears as though more accurate estimates are obtained with FP models. Moreover, it is evident that the credible intervals obtained from the MTC model are wider than those obtained from the FP and BEST-ITP counterparts. The only exception is for Scenario 6. This is what we expected since for scenario 6 we have generated data using the model by Dakin et al. (2011). Finally, the credible intervals corresponding to BEST-ITP model increase over the time for Figures 4.3-4.8, while in Figure 4.9 the credible intervals obtained from the BEST-ITP model remain close to real value over the time. This is not surprising since for scenario 7 we have generated data from the model by Ding and Fu (2013). Since the estimates obtained with MTC model, BEST-ITP model and FP model are not close to the real values in Figure 4.7, it is clear that all the models are less able to capture non-monotonic time effects.

To better quantify our comparison, we use the mean squared error (MSE) as a measure of goodness of fit. If $\hat{Y}_{sjt}^{(l)}$ denotes the predicted value of the response for study s, treatment j time t and at MCMC iteration l and Y_{sjt} is the "true", simulated values, the MSE is defined as

$$\text{MSE} = \frac{1}{R}\sum_{r=1}^{R}\left[\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{s=1}^{S}\left(Y_{sjt}-\hat{Y}_{sjt}^{(r)}\right)^2\right]$$

$$= \frac{1}{R}\sum_{r=1}^{R} \text{MSE}_r,$$

where $R = 10000$ is the number of saved MCMC iterations. Lower MSE is associated with better performance.
The MSE results for each model and for all the simulation scenarios are summarized in Table 4.3.

	MTC model	BEST-ITP model	FP model
Scenario 1	8.6004 (0.5524)	1.3959 (0.4903)	0.3505 (0.0388)
Scenario 2	4.8586 (0.5886)	1.1089 (0.3873)	0.3505 (0.0118)
Scenario 3	3.1465 (0.3951)	1.9744 (0.3398)	1.3069 (0.1823)
Scenario 4	27.6757 (3.9365)	81.6728 (8.5083)	1.5747 (0.1338)
Scenario 5	43.3330 (1.4238)	104.2221 (2.6848)	41.7533 (1.8354)
Scenario 6	0.0162 (0.3923)	6.1175 (0.6988)	0.9198 (0.2864)
Scenario 7	2.70591 (0.3365)	0.0125 (0.2658)	0.1701 (0.2898)

Table 4.3: Mean Square Error (MSE) comparison for MTC, BEST-ITP and FP network meta-analysis models. Displayed are the mean (standard deviation) of the respective MSEs averaged over 50 simulated data sets for each scenario. Scenarios that contain non-monotonic temporal behaviors appear to be challenging for the all methods.

We report the mean and the standard deviation of MSE across all 50 simulations. The mean squared error is lower for the FP models for all scenarios. Moreover, for the first three scenarios, the BEST-ITP model is preferable to the MTC model, while the BEST-ITP model is not able to capture the time constant effect in the fourth scenario, leading to a worse MSE. In general the values of MSE for the Scenario 5 are large, in agreement with the previous findings. Finally, in scenario 6 the smallest MSE is reached by the MTC model, while in the last scenario the best model is the BEST-ITP model. This is not unexpected since for scenario 6 we have generated the data from the model by Dakin et al. (2011) and for scenario 7 we have generated data from the model by Ding and Fu (2013).

4.2.4 Binary outcomes

Often in longitudinal studies the outcome of interest is discrete. As such it is important to investigate the performance of the above methods when a binary outcomes is available. Here, we report, as example, binary data correspond-

ing to scenario 3. They are simulated by simply taking a logit transformation of θ_{sjt}.

We fit the three network meta-analysis model, MTC model, BEST-ITP model and 1^{st} and 2^{nd} order fractional polynomial model (fixed and random effects models) model. The best fit (according to the DIC criterion) for the fractional polynomial model is obtained choosing a first order polynomial and treating β_0 as a fixed effect and β_1 as a random effect, with $p_1 = -2$. In Figure 4.10 we summarise the posterior distribution of p_{sjt} obtained from MTC model, BEST-ITP model and fractional polynomials. Red lines indicate the true values used to simulate the data, whereas magenta, green and blue lines indicate the estimates obtained using the MTC model, BEST-ITP model and fractional polynomials model, respectively.

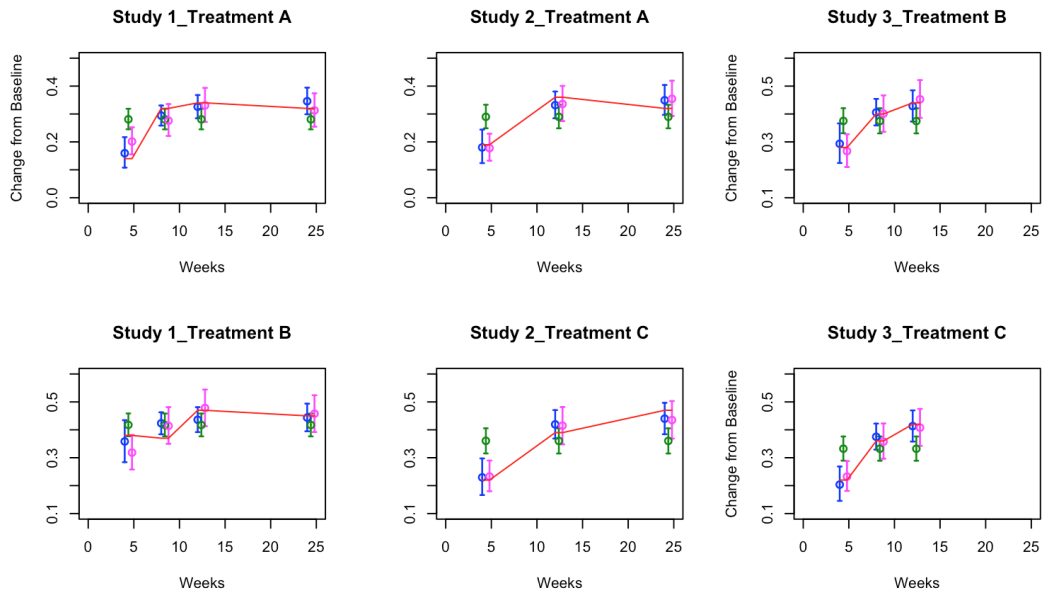


Figure 4.10: *Scenario 3: binary outcome*. Posterior distribution of p_{sjt}. Red lines indicate the values used to simulate the data, whereas the blue lines and the black lines correspond to the estimates obtained using the fractional polynomials model and the B-spline model, respectively, along with 95% credible intervals.

From the Figure 4.10 it is clear that the more accurate estimates are obtained with the MTC and FP models. Moreover, it is evident that the precision of the estimates given by BEST-ITP model increases over time. The values of the MSE confirm the results achieved by visual inspection, as the largest MSE is obtained with the BEST-ITP model (0.0897), while the smallest with the model of Jansen et al. (2015) (0.0060). These results are in agreement with the ones obtained in the continuous case.

4.2.5 Non-closed network

Finally, we consider a non-closed network to analyze whether or not a non-closed network could influence the behaviors of the three methods. In particular, we modify the simulation setting of the Scenario 3 adding a new study and a new treatment “D" (i.e. 4 hypothetical studies with 3 treatment arms), where the curve γ_{Dt} is given by: $\gamma_{Dt} = \begin{cases} -t & \text{if} \quad 0 \leq t \leq 4 \\ -\frac{3t}{4} - 1 & \text{if} \quad 4 < t \leq 8 \\ -\frac{t}{4} - 5 & \text{if} \quad 8 < t \leq 12 \\ -\frac{t}{12} - 7 & \text{if} \quad 12 < t \leq 24 \end{cases}$.

The network of studies used for this Scenario is displayed in Figure 4.11, this is a non-closed network where study 1 compares treatment A and B; study 2 has treatments A and C; study 3 matches B and C and study 4 analyze B and D.

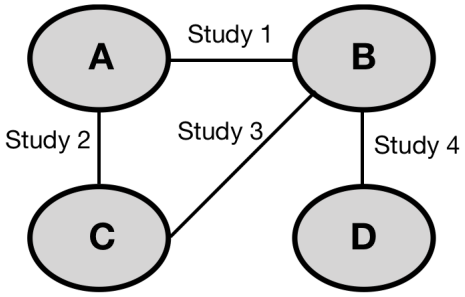


Figure 4.11: Network of studies used in the simulation scenario 8. Treatments are represented as nodes and connected by an edge if a direct comparison study between them is available. The network is non-closed, containing treatment pairs not directly compared by at least one study.

Table 4.4 describes the details of the simulations in terms of the time points at which follow up occurs as well as the number of individuals in each of the studies arms. Here we set the parameters τ_4^2 equal to 3.

Study	Treatment	Time of Observation	Number of Subjects per arm (n_s)
Study 1	A and B	4, 8, 12, and 24 weeks	100
Study 2	A and C	4, 12, and 24 weeks	120
Study 3	B and C	4, 8, and 12 weeks	130
Study 4	B and D	4, 8, 12, and 24 weeks	110

Table 4.4: Simulation parameters used in the simulation scenarios 8. For each study different treatments, times of observation and number of subjects per arm are specified.

In Figure 4.12 we summarise the estimated profiles obtained for the MTC, BEST-ITP and the appropriate FP models. Red lines indicate the true values used to simulate the data, whereas magenta, green and blue lines indicate the estimates obtained using the MTC model, BEST-ITP model and fractional polynomials model, respectively.

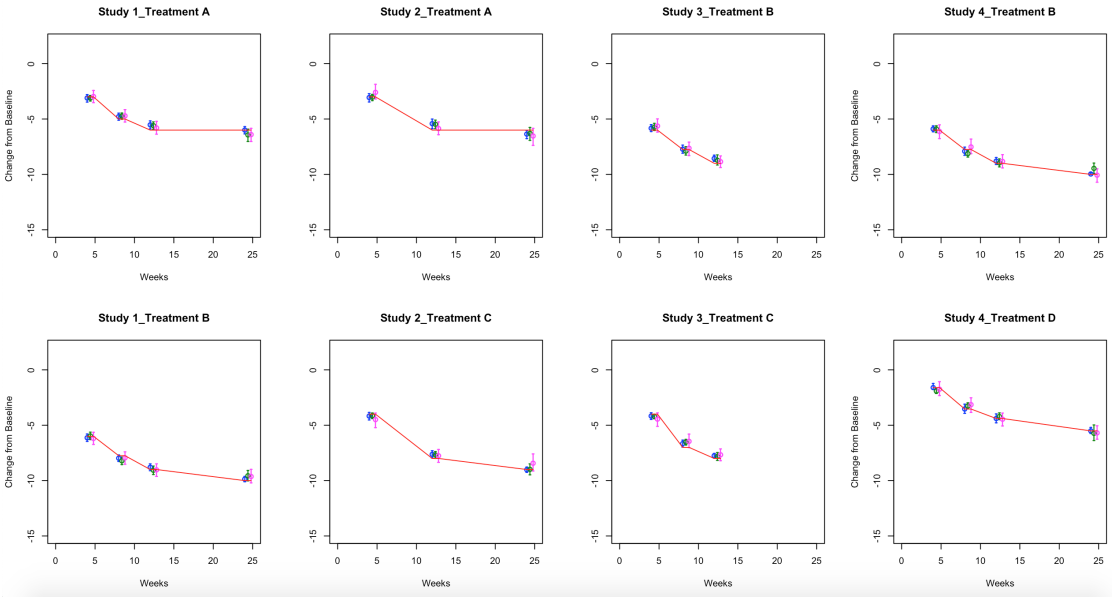


Figure 4.12: *Scenario 8: Non-closed network.* Posterior estimated profiles. lines indicate the values used to simulate the data, whereas magenta lines, green lines and blue lines correspond to estimates obtained using the MTC model, BEST-ITP model and fractional polynomials model, respectively, along with 95% credible intervals.

The main conclusions about model performance are not affected significantly by this new setup, as from Figure 4.10 it is clear that the most accurate estimates are obtained with the MTC and FP models. Moreover, it is evident that the precision of the estimates given by BEST-ITP model decrease over time. The values of the MSE confirm the results obtained in the closed case,

as the largest MSE is obtained with the BEST-ITP model (0.897), while the smallest with the model of Jansen et al. (2015) (0.090).

4.3 COPD Data

We consider here the dataset described in Karabis et al. (2013). The data are related to patients affected by chronic obstructive pulmonary disease (COPD). Subjects can receive a combination of three possible treatments: aclidinium 400 μg BID (AB400), tiotropium 18 μg QD (TIO18) or Placebo. The data are collected from a total of $S = 13$ studies: 3 studies compare aclidinium 400 μg with placebo, 8 compare tiotropium 18 μg with placebo and one compares aclidinium 400 μg and tiotropium 18 μg with placebo.
Figure 4.13 presents the network diagram, showing a total of 12 connections between the comparators. There is one closed loop, providing direct evidence. As all studies were placebo-controlled, placebo is used as the reference treatment.

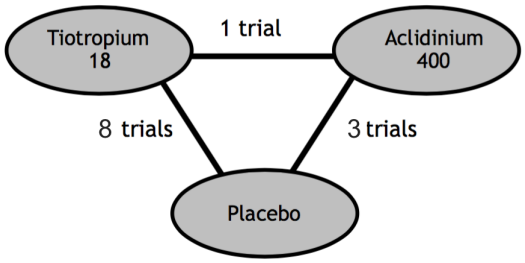


Figure 4.13: Networks of studies of the COPD data set. Edges indicate the number of studies providing direct comparison between adjacent treatments.

A brief overview of the study design and patient characteristics is presented in Table 4.3. For the studies identified that met the selection criteria, the change from baseline (CFB) and the standard error (SE) are reported. We have no information about the loss to follow-up, so we have assumed that the sample sizes n_{sjt} remain the same as at the start of the study. To account properly for loss to follow-up we would need to introduce a new layer in the hierarchy which is beyond the scope of this review. If a study did not report the number of patients (indicated in the right-most column of Table 4.3 as "NR"), then the sample size has been generated from an uniform distribution ranging from the minimum and the maximum number of patients across all studies. Moreover, where information about the standard errors is not reported, Dakin et al. (2011) suggest imputing the missing standard deviations

specifying the following prior distribution

$$\text{sd}_{sjt} \sim \text{Gamma}(\alpha_{se}, \beta_{se}), \quad \alpha_{sd} \sim \text{Uniform}(0, 10) \quad \text{and} \quad \beta_{sd} \sim \text{Uniform}(0, 10).$$

Study	Time point (weeks)	CFB aclidinium (SE)	CFB tiotropium (SE)	CFB placebo (SE)	Number of patients
LAS 39	0.14	0.07 (0.016)	0.02 (0.017)	-0.07 (0.022)	NR
	6.00	0.03 (0.018)	-0.01 (0.018)	-0.11 (0.024)	
ATTAIN	1.00	0.09 (0.012)	-	-0.03 (0.012)	269; 273
	4.00	0.08 (0.013)	-	-0.03 (0.013)	
	8.00	0.09 (0.014)	-	-0.03 (0.014)	
	12.00	0.06 (0.015)	-	-0.05 (0.015)	
	18.00	0.08 (0.015)	-	-0.06 (0.015)	
	24.00	0.06 (0.016)	-	-0.07 (0.016)	
ACCORD I	1.00	0.10 (0.012)	-	-0.01 (0.012)	190; 187
	4.00	0.11 (0.014)	-	-0.01 (0.014)	
	8.00	0.10 (0.013)	-	-0.01 (0.014)	
	12.00	0.10 (0.015)	-	-0.02 (0.015)	
ACCORD II	1.00	0.07 (0.013)	-	-0.03 (0.012)	178; 182
	4.00	0.06 (0.014)	-	-0.01 (0.014)	
	8.00	0.08 (0.016)	-	0.00 (0.016)	
	12.00	0.06 (0.016)	-	-0.01 (0.015)	
Tashkin 2008*	4.33	-	0.10 (0.013)	0.01 (0.013)	2987; 3006
	26.00	-	0.10 (0.013)	0.01 (0.010)	
	52.00	-	0.09 (0.013)	-0.01 (0.013)	
Verkindre 2006*	6.00	-	0.15 (0.060)	0.01 (0.055)	46; 54
	12.00	-	0.10 (0.060)	-0.01 (0.055)	
Brusasco 2003*	2.14	-	0.10 (NR)	-0.03 (NR)	402; 400
	24.14	-	0.09 (NR)	-0.05 (NR)	
Casaburi 2000	1.14	-	0.11 (0.010)	-0.01 (0.010)	279; 191
	7.14	-	0.11 (0.010)	-0.01 (0.010)	
	13.14	-	0.11 (0.010)	-0.04 (0.010)	
Donohue 2002*	2.14	-	0.12 (NR)	-0.02 (NR)	209; 201
	24.14	-	0.11 (NR)	-0.04 (NR)	
Moita 2008*	6.00	-	0.05 (NR)	-0.02 (NR)	147; 164
	12.00	-	0.07 (NR)	-0.01 (NR)	
Covelli 2005	8.00	-	0.17 (0.023)	0.003 (0.025)	100; 96
	12.00	-	0.19 (0.025)	0.001 (0.027)	
Donohue 2010*	2.00	-	0.14 (0.011)	0.01 (0.009)	420; 425
	12.00	-	0.15 (0.011)	-0.01 (0.010)	
	26.00	-	0.13 (0.011)	-0.03 (0.011)	
Casaburi 2002*	1.14	-	0.12 (0.010)	-0.01 (0.017)	550; 371
	13.14	-	0.12 (0.010)	-0.02 (0.017)	
	25.14	-	0.11 (0.010)	-0.04 (0.017)	
	49.14	-	0.11 (0.010)	-0.03 (0.017)	

Table 4.5: COPD data: mean change from baseline (CFB) in liters and corresponding SE at different time points.

We fit the MTC, BEST-ITP model and second order FP models (using both fixed and random effects models) to the COPD dataset and compare the results in the following.

4.3.1 Results

Firstly, we need to select the powers p_1 and p_2 for the FP model. Given that the outcome of interest is reported as change from baseline, the constant term β_{0sj} is always zero and can be ignored. The best fit according to the DIC criterion is obtained treating β_1 as fixed effect and β_2 as a random effect, with resulting estimated posterior means of $p_1 = -0.5$ and $p_2 = 0$.

We present estimated profiles obtained with the three methods in Figure 4.14. The solid lines indicate the results as mean differences in treatment effects over time for AB400 vs placebo, TIO18 vs placebo and AB400 vs TIO18 (red, blue and grey lines, respectively).

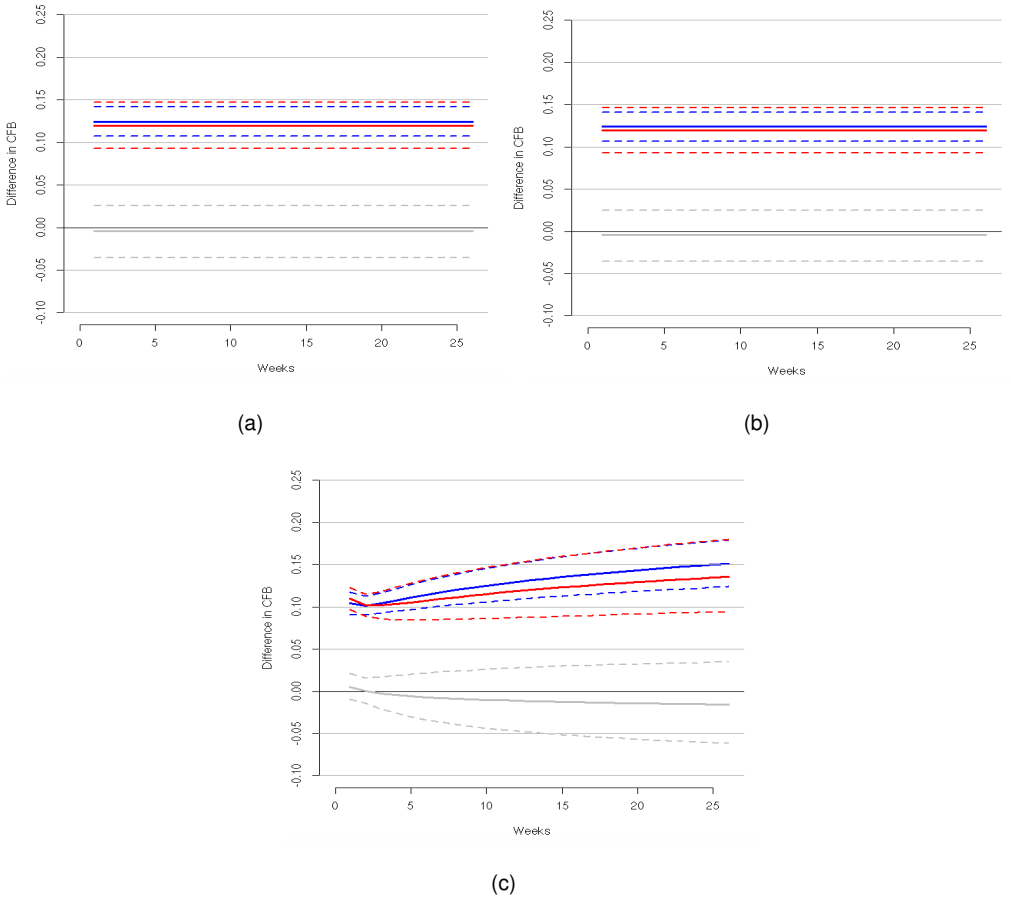


Figure 4.14: Shown are estimation results as mean differences in treatment effects over time for AB400 versus placebo (red), TIO18 versus placebo (blue) and AB400 versus TIO18 (grey) generated by by MTC model (a), BEST-ITP model (b) and Fractional polynomials (c). Dashed lines denote 95% credible intervals, while solid lines represent the posterior mean. These results are in agreement with the conclusions in Karabis et al. (2013).

It appears clear from the inspection of the graphs in Figure 4.14 that aclidinium has similar effect to tiotropium but it is better than placebo for the first

24 weeks. In fact, the grey lines, that represent the results as mean differences in treatment effects over time AB400 vs TIO18, are close to the value 0, whereas, the red lines and the blue lines, that indicate the results as mean differences in treatment effects over time for AB400 vs placebo and for TIO18 vs placebo have positive values. This results agree with the results provided by Karabis et al. (2013). Moreover, the posterior estimates obtained with fractional polynomials show a time pattern, by providing greater difference with increased time, while the results obtained with BEST-ITP model and MTC model are constant over time.

4.4 Osteoarthritis data

We also consider a second dataset described in Jansen et al. (2015). In this dataset there are 16 studies that investigate treatments for osteoarthritis (OA) of the knee based on different hyaluronan (HA)-based viscosupplements. OA is the most common chronic condition of the joints and it is a painful degenerative disease. In OA, the cartilage breaks down, causing pain, swelling and problems moving the joint. Due the effect of HA in synovial fluids, viscosupplementation with HA-based products is used to treat OA. Different treatments are considered: three, four, or five injections of HA with a molecular weight (MV) of 0.5 - 0.73 million Da (Hyalgan) (3 Hy-0.5-0.73; 4 Hy-0.5-0.73; 5 Hy-0.5-0.73); three injections of HA MW of 0.62-11.7 million Da (Supartz) (3 Hy-0.62-11.7); three injections of HA MW of 2.4-3.6 million Da (Euflexxa) (3 Hy-2.4-3.6); and three injections of Hylan GF-20 MW 6 million Da (Synvisc) (3 HyGF20). Table 4.6 and Figure 4.15 reports an overview of the study design and patient characteristics for the selected studies. Also for this dataset, we use the sample sizes at the initial step across the time points, because the sample sizes at other time points are not available.

Study	Time point (weeks)	CFB (SEM) 5 HY-0.5-0.73	CFB (SEM) 4 HY-0.5-0.73	CFB (SEM) 3 HY-0.5-0.73	CFB (SEM) 3 HYGF20	CFB (SEM) 3 HY-0.62-11.7	CFB (SEM) 3 HY-2.4-3.6	CFB (SEM) placebo	Number of patients
Altman (1998)	1	-15.0 (3.1)	-	-	-	-	-	-15.0 (3.1)	165;165
	2	-19.0 (3.0)	-	-	-	-	-	-20.0 (3.1)	
	3	-23.0 (2.9)	-	-	-	-	-	-21.0 (3.1)	
	4	-27.0 (2.9)	-	-	-	-	-	-27.0 (3.1)	
	5	-29.0 (2.9)	-	-	-	-	-	-30.0 (3.0)	
	9	-30.0 (6.1)	-	-	-	-	-	-32.0 (6.1)	
	12	-31.0 (3.0)	-	-	-	-	-	-31.0 (3.0)	
	16	-33.0 (2.9)	-	-	-	-	-	-33.0 (3.0)	
	21	-37.0 (2.8)	-	-	-	-	-	-34.0 (2.9)	
	26	-36.0 (2.8)	-	-	-	-	-	-31.0 (3.1)	
Carrabba et al. (1995)	2	-8.3 (4.0)	-	-6.0 (4.3)	-	-	-	-7.2 (3.3)	20;20;20
	4	-17.3 (4.2)	-	-14.6 (4.1)	-	-	-	-8.4 (3.3)	
	5	-21.5 (4.2)	-	-18.1 (4.6)	-	-	-	-8.0 (3.6)	
	8	-21.2 (3.9)	-	-16.3 (4.9)	-	-	-	-4.6 (3.8)	
Corrado et al. (1995)	5	-33.3 (7.5)	-	-	-	-	-	-20.7 (7.2)	20;20
	8	-39.0 (6.5)	-	-	-	-	-	-19.1 (6.7)	
Cubukcu et al. (2005)	1	-	-	-	-6.3 (2.0)	-	-	-9.0 (3.9)	30;10
	2	-	-	-	-18.3 (2.4)	-	-	-13.0 (3.7)	
	3	-	-	-	-24.3 (2.5)	-	-	-16.2 (2.8)	
	8	-	-	-	-31.0 (2.4)	-	-	-13.2 (3.1)	
Dougados et al. (1993)	2	-	-20.0 (3.0)	-	-	-	-	-15.0 (2.0)	55;55
	3	-	-31.0 (4.0)	-	-	-	-	-25.0 (3.0)	
	4	-	-23.0 (5.0)	-	-	-	-	-26.0 (3.0)	
	7	-	-35.5 (3.6)	-	-	-	-	-25.8 (2.9)	
	52	-	-38.9 (4.2)	-	-	-	-	-32.7 (3.9)	
Henderson et al. (1994)	1	-12.5 (2.3)	-	-	-	-	-	-16.1 (2.0)	45;46
Huskisson and Donnelly (1999)	5	-38.3 (4.1)	-	-	-	-	-	-21.3 (5.3)	50;50

	8	-33.5 (4.5)	-	-	-	-	-	-19.8 (5.3)	
	16	-32.8 (4.9)	-	-	-	-	-	-13.6 (5.5)	
	26	-26.4 (4.7)	-	-	-	-	-	-8.2 (5.3)	
Jubb et al. (2002)	4	-	-	-12.5 (1.9)	-	-	-	-11.0 (1.9)	204;204
	11	-	-	-10.0 (2.1)	-	-	-	-5.0 (1.9)	
	18	-	-	-9.0 (2.2)	-	-	-	-10.0 (1.9)	
Karlsson et al. (2002)	1	-	-	-	-7.0 (2.0)	-5.0 (2.0)	-	-7.0 (3.2)	70;70;70
	2	-	-	-	-16.0 (2.5)	-12.0 (2.6)	-	-11.0 (3.6)	
	3	-	-	-	-18.0 (2.9)	-20.0 (2.8)	-	-21.0 (4.0)	
	12	-	-	-	-22.0 (3.5)	-22.0 (3.2)	-	-19.0 (4.6)	
	20	-	-	-	-27.0 (3.5)	-21.0 (3.2)	-	-16.0 (3.8)	
	26	-	-	-	-20.0 (3.7)	-16.0 (3.8)	-	-21.0 (4.5)	
Scale et al. (1994)	1	-	-	-	-9.5 (5.8)	-	-	-2.5 (4.2)	40;40
	2	-	-	-	-32.0 (5.8)	-	-	-21.0 (4.2)	
	3	-	-	-	-44.0 (5.0)	-	-	-20.0 (4.2)	
	8	-	-	-	-52.0 (5.8)	-	-	-19.0 (4.2)	
	12	-	-	-	-53.0 (3.6)	-	-	-17.0 (3.6)	
Wobig et al. (1998)	1	-	-	-	-12.0 (4.6)	-	-	-5.0 (4.2)	57;60
	2	-	-	-	-26.0 (4.6)	-	-	-16.0 (4.2)	
	3	-	-	-	-39.0 (4.6)	-	-	-22.0 (4.2)	
	8	-	-	-	-46.0 (4.6)	-	-	-15.0 (4.2)	
	12	-	-	-	-46.0 (4.6)	-	-	-12.0 (4.2)	
Kirchner and Marshall (2006)	12	-	-	-	-28.7 (2.0)	-	-31.2 (2.0)	-	158;156
Stitik et al. (2007)	1	-5.6 (7.4)	-	-0.9 (5.2)	-	-	-	-	20;20
	2	-15.2 (8.0)	-	-5.7 (5.9)	-	-	-	-	
	5	-20.5 (8.0)	-	-13.2 (5.9)	-	-	-	-	
	12	-24.3 (8.0)	-	-23.2 (5.9)	-	-	-	-	
	26	-23.3 (8.0)	-	-21.8 (5.9)	-	-	-	-	
	36	-26.3 (8.0)	-	-10.4 (5.9)	-	-	-	-	
	52	-35.4 (8.0)	-	-16.1 (5.9)	-	-	-	-	
Wobig et al. (1999)	1	-	-	-	-8.0 (3.6)	-10.0 (4.5)	-	-	38;35
	2	-	-	-	-18.0 (3.6)	-20.0 (4.5)	-	-	

	3	-	-	-	-29.0 (4.2)	-30.0 (4.5)	-	-	
	8	-	-	-	-37.0 (4.2)	-30.0 (4.5)	-	-	
	12	-	-	-	-37.0 (4.2)	-30.0 (4.5)	-	-	
Altman et al. (2009)	1	-	-	-	-	-	-13.5 (1.0)	-10.0 (1.0)	293;295
	2	-	-	-	-	-	-20.0 (1.0)	-17.5 (1.0)	
	3	-	-	-	-	-	-24.0 (1.0)	-22.5 (1.0)	
	7	-	-	-	-	-	-25.0 (1.0)	-22.0 (1.0)	
	12	-	-	-	-	-	-25.5 (1.0)	-22.0 (1.0)	
	18	-	-	-	-	-	-27.5 (1.0)	-21.0 (1.0)	
	26	-	-	-	-	-	-25.0 (1.0)	-17.5 (1.0)	
Dickson et al. (2001)	12	-	-	-	-39.0 (4.0)	-	-26.0 (4.0)	-	NR

Table 4.6: OA dataset: mean CFB (Change from baseline) in liters and corresponding standard error (SE) at different time points.

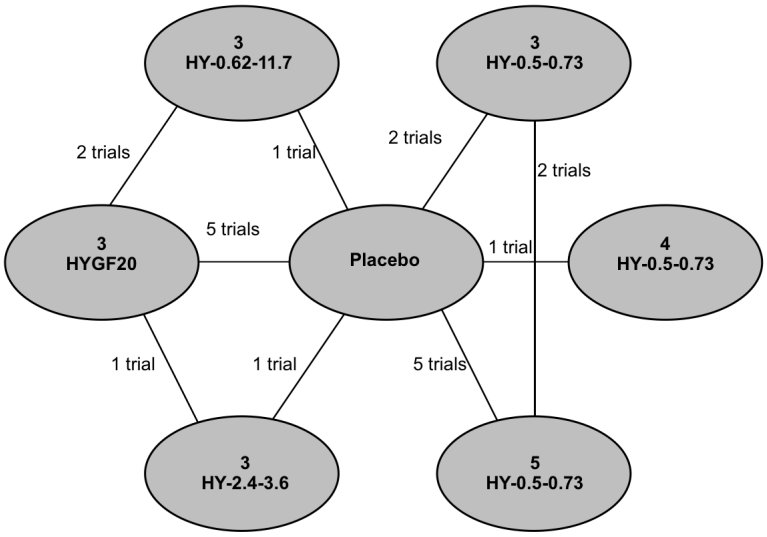


Figure 4.15: Networks of studies of the OA data set. Edges indicate the number of studies providing direct comparison between adjacent treatments.

4.4.1 Results

Also in this example, we apply the MTC model, BEST-ITP model and 1^{st} and 2^{nd} order fractional polynomial model (fixed and random effects models)(FP) models to perform a network meta-analysis of OA studies. In our analysis placebo is treated as reference treatment. The best fit for fractional polynomial model according to DIC is obtained using a second order polynomial and treating β_1 as a fixed effect and β_2 as a random effect, with p_1=0.5 and p_2=1.
In Figure 4.16 posterior estimated profiles obtained from the three models are shown. We report the posterior mean of the treatment effect relative to placebo over time for 5 HY-0.5-0.73 (magenta lines), 4 HY-0.5-0.73 (black lines), 3 HY-0.5-0.73 (red lines), 3 HYGF20 (green lines), 3 HY-0.62-11.7 (blue lines) and 3 HY-2.4-3.6 (light blue lines).

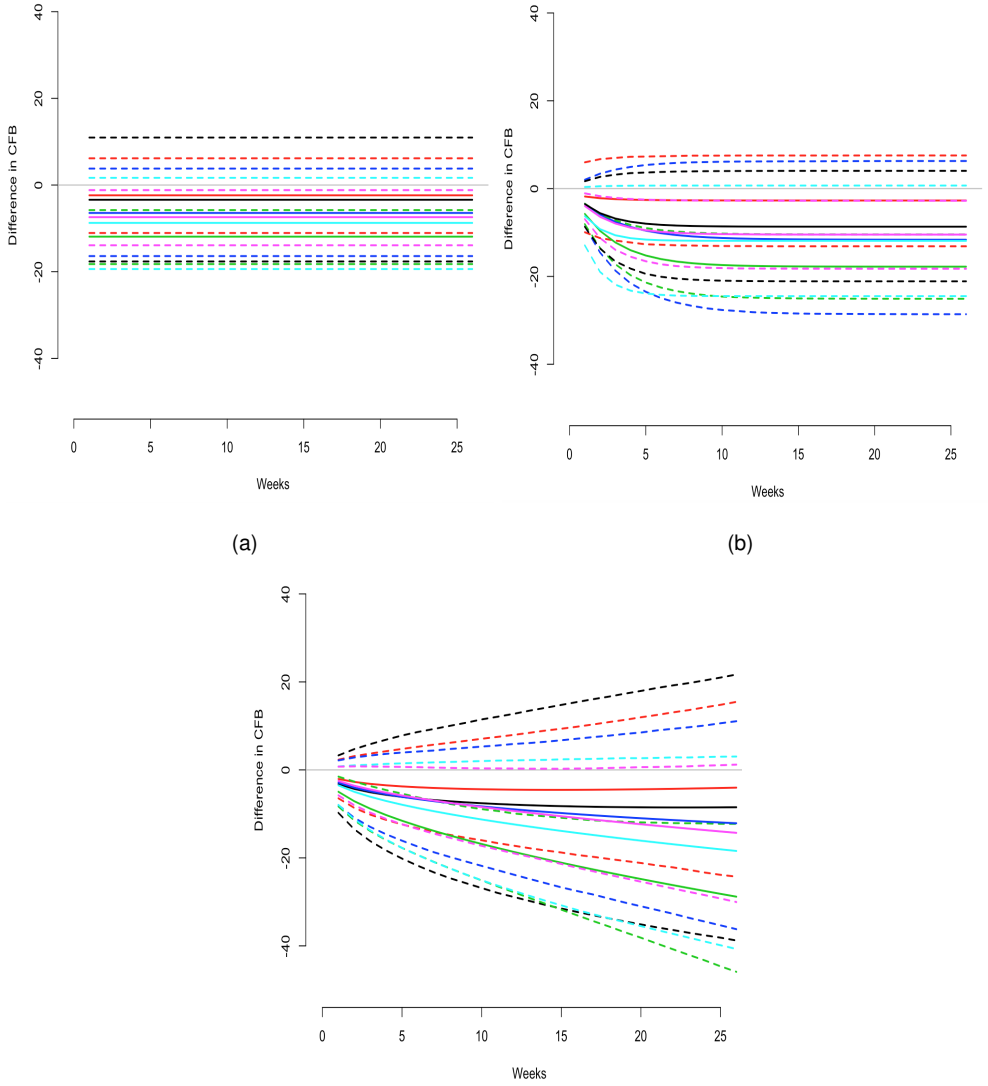


Figure 4.16: Results for the osteoarthritis (OA) data set. Shown are treatment effects relative to placebo over time for 5 HY-0.5-0.73 (magenta), 4 HY-0.5-0.73 (black), 3 HY-0.5-0.73 (red), 3 HYGF20 (green), 3 HY-0.62-11.7 (blue) and 3 HY-2.4-3.6 (light blue) estimated by MTC model (a), BEST-ITP model (b) and Fractional polynomials (c). Dashed lines denote 95% credible intervals, while solid lines represent the posterior means. All models suggest that the best results are obtained by 3 HYGF20, in agreement with the result obtained originally for this data with FPs by Jansen et al. (2015).

All treatments are better than placebo, consistently over the first 24 weeks. In particular, the best result are obtained with 3 HYGF20, in agreement with the result obtained originally for this data with FPs by Jansen et al. (2015). Moreover, the estimated treatment effects obtained from the fractional polynomials model change in time; for example for the first few weeks receiving 5 injections of HY-0.5-0.73 is better than 4 injections of the same drug, while

later on we find that the opposite is true. This type of change is not detected by the MTC model and the BEST-ITP model.

4.5 Diabetes Data

Finally, we consider a third dataset analysed in Ding and Fu (2013). Data are recorded on patients affected by type 2 diabetes. The authors consider 4 studies and 4 different treatments. All studies are randomized controlled, double-blind, parallel arm trials comparing two different oral anti-diabetic agents. The clinical effect of interest is hemoglobin A1c (HbA1c, %) reduction from baseline. Table 4.7 and Figure 4.17 report a summary of the study design and patient characteristics of the selected studies. Studies 1 and 2 both have a duration of 24 weeks and have measurements at 4, 8, 12, 16, and 24 weeks, whereas studies 3 and 4 have a shorter duration (16 weeks), with study 3 measuring the outcome variables at 8 and 16 weeks and study 4 measuring the outcome variables at 12 and 16 weeks. In the original paper (see Ding and Fu (2013)), the treatments are unknown and they are labeled as “Treatment1", “Treatment2", “Treatment3", “Treatment 4".

Study	Time point (weeks)	CFB (SE) Treatment1	CFB (SE) Treatment2	CFB (SE) Treatment3	CFB (SE) Treatment4	Number of patients
1	4	-	-0.467 (0.665)	0.259 (0.65)	-	525; 524
	8	-	0.925 (0.898)	0.652 (0.917)	-	496; 508
	12	-	1.292 (1.013)	1.024 (1.049)	-	502; 513
	16	-	1.52 (1.074)	1.309 (1.122)	-	509; 513
	24	-	1.782 (1.09)	1.578 (1.125)	-	515; 508
2	4	-	-	0.257 (0.562)	0.696 (0.54)	575; 561
	8	-	-	0.721 (0.766)	1.219 (0.817)	568; 561
	12	-	-	1.095 (0.938)	1.563 (0.975)	552; 546
	16	-	-	1.356 (1.034)	1.751 (1.056)	556; 553
	24	-	-	1.609 (1.043)	1.829 (1.101)	541; 548
3	8	0.185 (0.479)	-	0.807 (0.752)	-	95; 188
	16	0.198 (0.643)	-	1.395 (0.848)	-	88; 183
4	12	-	1.505 (1.103)	1.251 (1.102)	-	91; 93
	16	-	1.393 (1.134)	1.262 (1.298)	-	90; 98

Table 4.7: Type 2 Diabetes Dataset: mean CFB (Change from baseline) of haemoglobin and corresponding standard error at different time points for each of the 4 studies.

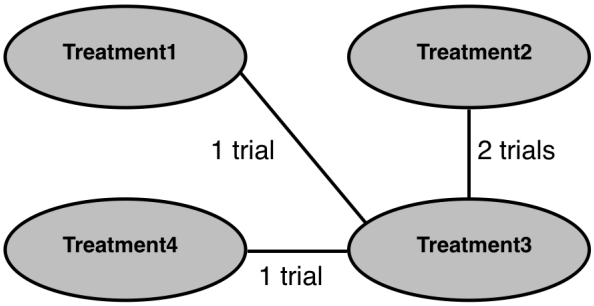


Figure 4.17: Networks of studies of the T2D data set. Edges indicate the number of studies providing direct comparison between adjacent treatments.

4.5.1 Results

We fit the three network meta-analysis model, MTC model, BEST-ITP model and 1^{st} and 2^{nd} order fractional polynomial model (fixed and random effects models)(FP) models to type 2 diabetes dataset. In our analysis Treatment 1 is set as reference treatment. The best fit (according to the DIC criterion) for the fractional polynomial model is obtained choosing a first order polynomial and treating β_0 as a fixed effect and β_1 as a random effect, with $p_1 = -2$. In the Figure 4.18 posterior estimated profiles obtained from MTC model, BEST-ITP model and fractional polynomials are shown. Results are reported as mean differences in treatment effect over time for Treatment 1 vs Treatment 2 (blue lines), Treatment 1 vs Treatment 3 (red lines), Treatment 1 vs Treatment 4 (green lines), Treatment 2 vs Treatment 3 (orange lines), Treatment 2 vs Treatment 4 (grey lines) and Treatment 3 vs Treatment 4 (magenta lines).
From Figure 4.18 we can conclude that, in general, Treatment 2, Treatment 3 and Treatment 4 are better than Treatment 1 and this is true over time. In particular, Treatment 3 leads to the smallest reduction in HbA1c, while Treatment 1 yields the largest reduction. These results are consistent with the ones described in the original paper. Moreover, the posterior estimates obtained with fractional polynomials and BEST-ITP models show that treatment effects changes over time, while the MTC model is not able to capture the change over time.

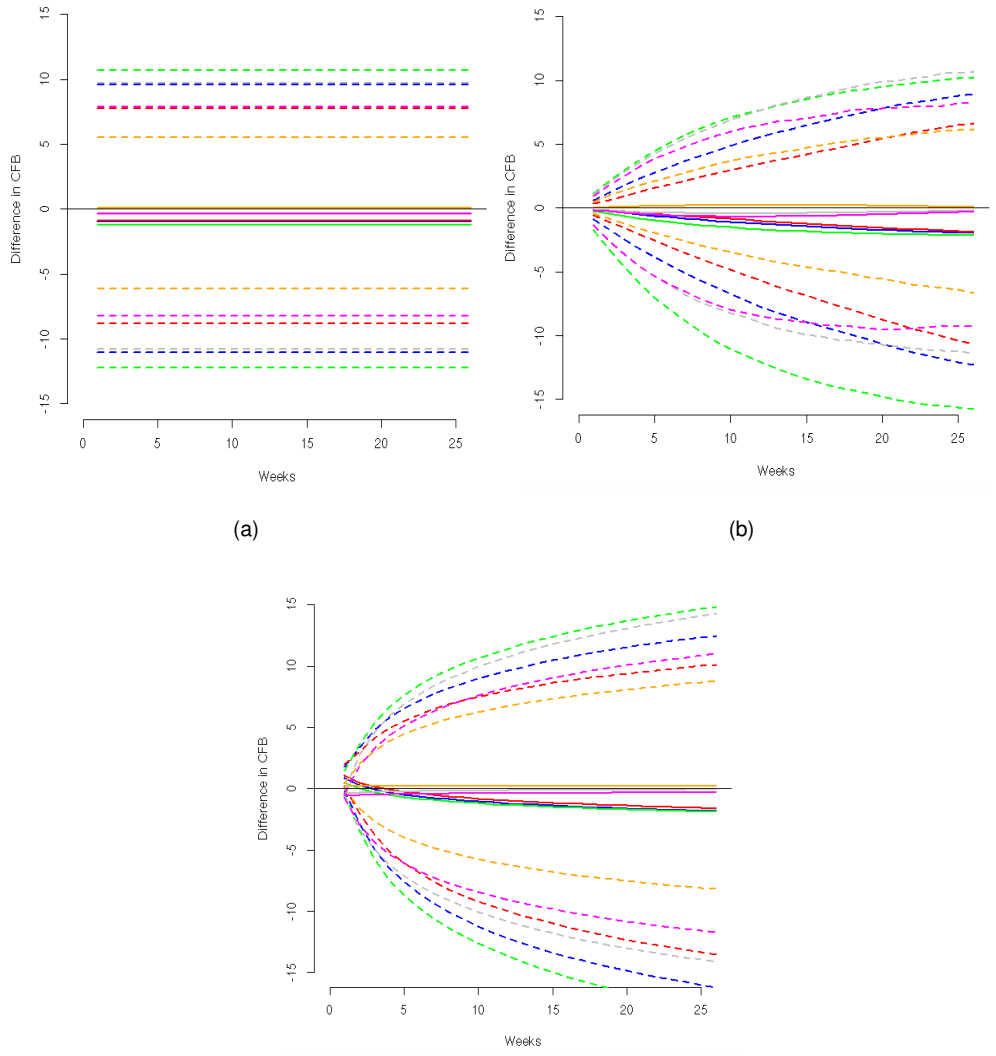


Figure 4.18: Results for the type 2 diabetes (T2D) data set. Shown are mean differences in treatment effects over time for treatments 1 versus 2 (blue), 1 versus 3 (red), 1 versus 4 (green), 2 versus 3 (orange), 2 versus 4 (grey) and 3 versus 4 (magenta) generated by the MTC model (panel a), BEST-ITP model (panel b) and Fractional polynomials (panel c). Dashed lines denote 95% credible intervals, while solid lines represent posterior means.

4.6 Conclusions

Driven by advances in science and technology, the number and types of available healthcare interventions is steadily increasing and clinicians are often faced with the challenge of choosing between a number of different treatment options. Relative treatment effects and cost effectiveness of different interventions need to be objectively and accurately assessed in clinical studies to bet-

ter inform healthcare decisions (Welton et al., 2012). It is generally accepted that randomized controlled trials provide the most rigorous and conclusive evidence on the relative effects of different interventions. For example, when comparing two treatments, A and B, the gold standard is a randomized control trial with two treatment arms. In practice, however, evidence from direct comparison trials may be limited and it is often impossible to have head-to-head comparisons for all relevant comparators of an intervention, making it necessary to resort to indirect comparisons (Song et al., 2009). For instance, direct comparison from two different studies on treatment A versus C, and B versus C, might be available and indirect methods exploit the common comparator C to provide an indirect comparison of treatment A versus B. In the context of systematic reviews, network meta-analysis (NMA) is a technique for comparing multiple treatments using both direct comparisons obtained within randomized clinical trials and indirect comparisons across trials having a common comparator, such as a placebo or standard treatment (Li et al., 2011; Borenstein et al., 2010; Higgins et al., 2009; Lipsey and Wilson, 2001). In this Chapter, we review the three most recent methods MTC, BEST-ITP and fractional polynomials models for Network Meta-analysis, highlighting the main features of the different methods. The model presented by Dakin et al. (2011) assumes random relative effects and allows for the relative treatment effects to vary over time, without temporal pattern restriction. Ding and Fu (2013) suggested a parametric model to describe a nonlinear relationship between outcome and time for each treatment, assuming a diminishing return time course of treatment responses. Finally, Jansen et al. (2015) suggested a more flexible approach where the relationship between outcome and time is a fractional polynomial. We evaluated the performance of the models in a a simulation study and on real data applications. Based on the results of our simulations, Dakin's model appears to be the most conservative in terms of estimation of the underlying effect-size, while fractional polynomial seems to offer the most flexible strategy, able to accommodate for different time patterns. The BEST-ITP model can not capture the constant course, i.e. an effect size that does not change over time. In general, it is challenging to capture non-monotonic temporal patterns, with the FP model leading to slightly more accurate estimates as shown in the simulation study.

In terms of computational time, there are no substantial differences in running times between the three different models, although fractional polynomials require extensive sensitivity analysis to select the optimal order and the power terms of the polynomials. Thus, to improve these problems, we extend in Chapter 5, the FP models in a B-spline settings.

Chapter 5

Evidence synthesis in network meta-analysis via Bayesian B-spline mapping

Following the comparison among MTC, BEST-ITP and fractional polynomials models in Chapter 4, FP methods turn out as the best of the three reviewed approaches. On the one hand, these models are the most flexible and can accommodate a variety of treatment effect patterns. On the other hand, they still present limitations both from a methodological and a computational point of view; they require intensive computations for the choice of parameters and they present modelling drawbacks (see Sections 5.2 and 5.7 below). In this Chapter, we propose an NMA method based on B-splines to model temporal behavior, which allows for the simultaneous analysis of outcomes at different time points, automatically accounting for correlation across time. We illustrate the model in simulations and on real data examples and compare its performance to the FP model. The proposed approach has many advantages in terms of model flexibility, computational burden and ease of specification. B-splines (and their extensions) are a natural competitor of FPs and as such they have been previously compared in the literature in different setups (see, among others, Strasak et al., 2011). This Chapter is organized as follows. In Section 5.1 we recall the model likelihood for the general NMA problem. In Section 5.2 we briefly review the FP approach and in Section 5.3 we introduce the proposed B-spline model and highlight its advantages. In Section 5.4 we specify the prior distribution for the parameters of the model and Section 5.5 we briefly describe the MCMC strategy. In Sections 5.6 and 5.7 we compare the the B-spline and FP models in simulations and real data applications, respectively. We conclude the Chapter in Section 5.8.

5.1 Network meta-analysis

Following the notation and the general framework of Section 4.1, we know that the main objective of NMA of longitudinal studies is the evaluation of a suitably defined response over time. The response y_{sjt} in study $s \in \{1, \ldots, S\}$, in treatment arm $j \in \{1, \ldots, J\}$ at time $t \in [0, T]$ is distributed as

$$y_{sjt} \sim f\left(y_{ij} | \theta_{sjt}, \sigma^2_{sjt}\right), \tag{5.1}$$

where f is a (usually parametric) probability density. The main parameter of clinical interest is θ_{sjt}, measuring the treatment effect of intervention j in study s at time t. As in the generalized linear model framework, a link function $g(\theta_{sjt})$ specifies the relationship between the main clinical effects θ_{sjt} and the response. For example, in the case of continuous outcomes we consider

$$y_{sjt} \sim \text{Normal}\left(\theta_{sjt}, \sigma^2_{sjt}\right), \tag{5.2}$$

with $g(\theta_{sjt}) = \theta_{sjt}$, while in the case of binary outcomes

$$\begin{aligned} y_{sjt} &\sim \text{Binomial}\left(n_{sjt}, p_{sjt}\right), \\ g(p_{sjt}) &= \text{logit}(p_{sjt}) = \theta_{sjt}. \end{aligned} \tag{5.3}$$

Other alternatives include $g(\theta_{sjt}) = \ln(\theta_{sjt})$ if f is modeled as a Poisson distribution. We focus on continuous and binary responses as these are the most common outcomes in clinical trials.

The fundamental difference between the NMA methods is the way the treatment effect θ_{sjt} is modelled.

In this Chapter, following the FPs model, we set the variances of the main outcomes in (5.2) equal to

$$\sigma^2_{sjt} = \frac{1}{(1-\rho^2)^2}\left(\frac{\text{sd}_{sjt}}{\sqrt{n_{sjt}}}\right)^2,$$

i.e., to the standard error (with sd_{sjt} an estimate of the standard deviation and n_{sjt} the sample size) for the corresponding study, treatment arm and time point, adjusted by a factor $\rho \in [0, 1)$ taking into account the within-study correlation between subsequent time points. While estimates for the correlation coefficient, here assumed constant over time, may be available from expert knowledge, in a fully Bayesian setting it is an object of inference and assigned an appropriate prior distribution. We assume $\rho \sim \text{Uniform}(0, 0.95)$.

5.2 Fractional polynomial model

As said in Subsection 4.1.3, in the FP regression framework, powers of the covariates of interest (in our case time), usually chosen from the set $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, are entered into the linear predictor (see Royston and Sauerbrei (2008) for a detailed account). Let us recall the approach to NMA proposed by Jansen et al. (2015), in which, given an order M and powers $p_1, \ldots, p_M \in S$, the mean outcome of the j-th treatment in study s at time $t > 0$:

$$\theta_{sjt} = \begin{cases} \beta_{0sj} + \sum_{m=1}^{M} \beta_{msj} t^{p_m}, & \text{if } p_1 \neq \cdots \neq p_M \\ \beta_{0sj} + \beta_{1sj} t^{p} + \sum_{m=2}^{M} \beta_{msj} t^{p} \left[\ln(t)\right]^{m-1}, & \text{if } M > 1 \text{ and } p_1 = \cdots = p_M = p, \end{cases} \tag{5.4}$$

where $t^0 := \ln(t)$. Thus, implementing the FP model requires selecting an order M and powers $p_1, \ldots, p_M$. In both the original work of Jansen et al. (2015) and the studies described in Chapter 4, the Deviance Information Criterion (DIC, Spiegelhalter et al., 2002) is used as an expected predictive error estimate to guide the selection of the FP order and powers. When using DIC, the aim is to select order and powers that minimize the sum of a goodness of fit term, given by the posterior mean of the model deviance, and a regularization term, given by the posterior estimate of the effective number of parameters. The latter is intended to penalize model complexity in order for the DIC to be a tool that prevents overfitting.

First and second order FPs allow representations of a considerable range of non-linear relationships, and higher orders are rarely used in medical applications. Even considering only the cases $M = 1$ and 2, using the DIC often presents a non-negligible computational task as it requires the comparison of the DIC values of 8 different first-order fractional polynomial models (one for each possible power from S) and of 35 different second-order fractional polynomial models (for all the possible combinations of powers from S). Besides the computational cost necessary to select the order and powers, the FP modelling approach lends itself to more fundamental criticism: (i) any choice of order and combination of powers imposes strong structural properties on the curve and therefore on the set of representable temporal patterns, which may not be supported by the data. Moreover, the selection of powers and order depends also on the prior specification of the other parameters in the model; (ii) FPs generally have singularities at zero or grow polynomially for

large arguments. Consequently, more complex FPs may only provide poor fit for values of t that are large or close to zero; (iii) the individual terms in an FP are functions over the entire time interval and, as discussed by Royston and Sauerbrei (2008), may be less responsive locally to perturbations of the response at a given point, but the fit at distant points may be affected considerably. To avoid such effects, basis functions with short support may be more appropriate in many situations (see the discussion in Royston and Altman, 1994); (iv) first-order FPs describe effects as a function of transformed time t in a linear model $\beta_{0sj} + \beta_{1sj}t^p$, in particular they are always monotonic. Higher order FPs become increasingly complex, with second-order FPs being either monotonic or unimodal. It may however be preferable to have a model which has the linear model always nested within more complex models; (v) although a second order FP as in (5.4) has three coefficients only (two in the case of modelling change from baseline, since the intercept is equal to zero and can be discarded), the choice powers increases model complexity, as many different FP models need to be fitted to find the one that best describes the temporal pattern of the treatment effect. Additionally, the use of DIC as model choice criterion effectively hinges the FP model to any potential shortcomings of the DIC which have been pointed out in the literature. We have indeed observed sub-optimal DIC-based selections in the real data applications, and therefore present a more detailed discussion in Section 5.7. To address the above issues we next propose to use B-spline basis functions to capture the treatment effect over time.

5.3 B-spline model

B-splines are a family of basis functions with many desirable theoretical and computational properties, making them widely used in function approximation and countless applications in statistics and engineering. For details we refer to de Boor (2001). Univariate (cardinal) B-splines can be defined inductively, starting from the B-spline of order 1 given by

$$B_1(t) := \begin{cases} 1 & \text{if } t \in [0,1], \\ 0 & \text{otherwise.} \end{cases}$$

B-splines of higher order $n \in \mathbb{N}$ are defined via consecutive convolutions by

$$B_n(t) := \int_{\mathbb{R}} B_{n-1}(t-s)B_1(t)\,ds,$$

i.e., with higher order they become increasingly smoother. For $n \geq 2$, the B-spline B_n is

(i) symmetric and positive valued, with finite support $[0, n]$,

(ii) $(n-2)$-times continuously differentiable,

(iii) polynomial of degree $n-1$ when restricted to any interval $[k, k+1]$, $k \in \mathbb{Z}$, and

(iv) satisfies $\sum_{k\in\mathbb{Z}} B_n(t-k) = 1$ for all $t \in \mathbb{R}$.

The integer translates $\{B_n(\cdot - k)\}_{k\in\mathbb{Z}}$ are thus a set of highly regular and well structured basis functions. They span the space of functions that are polynomial on each interval $[k, k+1]$, $k \in \mathbb{Z}$, and $(n-2)$-times continuously differentiable in every $k \in \mathbb{Z}$. Such piecewise polynomial functions with global smoothness restrictions are called splines. In the cardinal case the integers are called the knots and the above can be easily generalized to any uniform knot sequence $(hk)_{k\in\mathbb{Z}}$, where $h \in \mathbb{R}$, by considering the basis functions $\{B_n(t/h - k)\}_{k\in\mathbb{Z}}$ of translates of the appropriately dilated cardinal B-splines. In applications in which a finite time interval is considered, only those finitely many basis functions whose support intersects the interval are required.

A main advantage of B-spline basis functions comes from their locality due the finite support and their structure as translates of only one generating symmetric function. Choosing an order and adjacent knot distance does not impose any overall temporal behavior beyond smoothness. In fact, for sufficiently large order, the functions $\{B_n(2^n t - k)\}_{k\in\mathbb{Z}}$ can approximate any integrable function to arbitrary precision. This has to be contrasted with the FP model, in which the fractional monomials t^{p_m} may be considered as basis functions, and where choosing any combination of fractional monomials imposes rather strong geometrical restrictions on the representable functions. Furthermore, due to the translation structure of the B-spline bases, any part of the observation time-interval is treated equally, with no performance deterioration for small or large time arguments. The local support of the basis functions makes the B-spline expansions a local-influence model, whereas their overlapping support, the size of which goes hand in hand with the order of smoothness, acts as a regularizing mechanism that can facilitate stability. B-spline bases are also stable in the sense that small changes in coefficients do not perturb the function significantly (see Christensen, 2016). Small changes in a represented function will therefore reflect in small changes in its coefficient sequence and vice versa. Together with the locality, and the partition of unity property (iv), these characteristics contribute to a better interpretability

of the coefficients in a B-spline expansion as compared to the FP coefficients. Indeed, it is easy to understand changes in the function due to changes in the coefficients, as every coefficient corresponds to a particular interval of the domain of the function. On the other hand, FP are sums of different global functions and it is therefore harder to visualize the effect of changes in coefficients. Finally, B-spline bases have the property that, with increasing order and additional knots, the spaces of representable functions become strictly larger, containing all previously representable function. This nestedness is desirable since it ensures that simpler models are always embedded in more complex, a property not shared by the FP model.

Practical applications of B-splines are dominated by $n = 4$ as the order of choice, resulting in piecewise cubic polynomials with continuous curvature. In fact, given a number of data points, it is well known (e.g., Hastie et al. (2005)) that among all smooth regression functions, the unique solution to the objective of minimizing a weighted sum of squared approximation errors and the integral squared curvature as regularization is given by the cubic spline with knots at the data time-points. As such, in the remainder of this Chapter, we will restrict our attention to cubic B-splines.

The smoothness property of the spline functions at their knots makes them relatively insensitive to the precise locations of the knots. While observation times differ across studies, the B-spline basis functions, and in particular their knot positions, have to coincide across studies in order to guarantee the consistency assumption of the NMA in our modelling framework (see next Section). Since studies in NMA of longitudinal data typically have very few observation time points in the interval $[0, T]$, we choose $h = 3/T$, i.e., we consider the four uniformly spaced knots $0, T/3, 2T/3, T$, over the entire observation interval $[0, T]$. With this choice of 4 knots we prevent overfitting, given the limited number of time observations and the often regular behaviour of treatment effects over time in our application. However, in order to check this assumption, we conducted a sensitivity analysis for the number of knots ranging from 2 to 10. In fact, on the one hand, the number of knots should be sufficiently large to fit all essential spectral features of the desired curve and, on the other hand, a too large number of knots can result in serious over fitting of the data and may, therefore, fail to predict future observations reliably. The Deviance Information Criterion is used as an expected predictive error estimate to guide the selection of the the number of knots. In this setup, the aim is to select number of knots that minimize the sum of a goodness of fit term, given by the posterior mean of the model deviance, and a regularization term, given by the posterior estimate of the effective number of parameters. The

latter is intended to penalize model complexity in order for the DIC to be a tool that prevents overfitting. Based on our results, generally four knots provided the best choice (results not shown).

For cubic B-splines with four knots over the observation time interval, six of the generator translates have support intersecting $[0, T]$. We denote those by $B_0, \ldots, B_5$ and model the mean outcome of treatment j in study s at time $t > 0$ as cubic spline using the basis expansion

$$\theta_{sjt} = \sum_{k=0}^{5} \beta_{ksj} B_k(t). \tag{5.5}$$

As highlighted in the previous Section, it is not unreasonable to consider the B-spline model (5.5) as less complex than the FP model (5.4) despite having a greater number of coefficients, since all order and power parameter choices (for instance evaluated via the DIC) have to be additionally considered for (5.4), whereas we keep the model (5.5), i.e., the spline order and number of knots, fixed in all applications. The multi-resolution property of B-splines allows us to propose a default model choice, with good performance in most applications. This strategy is not possible for FPs, as they lack this property and impose global assumptions on the behavior of representable functions.

5.4 Prior specification

We follow the prior specification proposed for the FP approach by Jansen et al. (2015), which extends earlier approaches by Dakin et al. (2011) and Ding and Fu (2013). The regression coefficient vector $\boldsymbol{\beta_{sj}} = (\beta_{0sj}, \ldots, \beta_{Msj})^\intercal$ for both the FP model (5.4) and the B-spline model (5.5) is expressed as a sum of a study-specific random effect and a study-specific arm deviation from the reference treatment:

$$\boldsymbol{\beta_{sj}} = \boldsymbol{\mu}_s + \boldsymbol{\delta}_{sj}$$

where the study specific means $\boldsymbol{\mu}_s = (\mu_{0s}, \ldots, \mu_{Ms})^\intercal$ in our implementation are assigned a vague prior distribution

$$\boldsymbol{\mu}_s \sim \text{Normal}\left(\mathbf{0}, 10^4 \boldsymbol{I}\right),$$

with $\mathbf{0}$ denoting the zero vector and $\boldsymbol{I}$ the identity matrix of appropriate dimensions. The study specific treatment effects $\boldsymbol{\delta}_{sj} = (\delta_{0sj}, \ldots, \delta_{Msj})^\intercal$ in study s of treatment j relative to a reference treatment, indexed as $j = 1$, are modelled

as “structured random effects"

$$\begin{aligned} \boldsymbol{\delta}_{sj} &\sim \text{Normal}\left(\boldsymbol{d}_j - \boldsymbol{d}_1, \boldsymbol{\Sigma}\right), \\ \boldsymbol{d}_j &\sim \text{Normal}\left(\mathbf{0}, 10^4 \boldsymbol{I}\right), \end{aligned} \tag{5.6}$$

for $j > 1$, with $\boldsymbol{d}_1 = (d_{01}, \ldots, d_{M1})^\intercal := \mathbf{0}$ and $\boldsymbol{\delta}_{sj} = (\delta_{0s1}, \ldots, \delta_{Ms1})^\intercal := \mathbf{0}$. Clinical effects in (5.4) and (5.5) are therefore expressed as change from baseline (CFB) with respect to the reference treatment. The covariance matrix $\boldsymbol{\Sigma} = (\sigma_m \sigma_n \lambda_{mn})_{m,n=0,\ldots,M}$ captures heterogeneity between studies, where σ_m represents the variance in $(\delta_{msj})_{sj}$ and the covariances λ_{mn} quantify the correlation between these treatment effect parameters.

When assuming a (partially) fixed-effects model, (5.6) (or certain components of it) is replaced by $\boldsymbol{\delta}_{sj} = \boldsymbol{d}_j - \boldsymbol{d}_1$ and it is not necessary to estimate (the respective) between-study covariances. In this work however, in the B-spline case we only consider models including random effects per study, allowing for heterogeneity between studies, but fixed effects for $\boldsymbol{\delta}_{sj}$. For the FP model, as suggested by Jansen et al. (2015), we impose a random effect only on one of the components $\boldsymbol{\delta}_{sj}$. In particular, when $M = 1$, there is only one between-study heterogeneity parameter, related to the relative treatment effects for β_{1sj}. Here we assign the relevant parameters σ_m a uniform distribution on the intervall $(0, 10)$. Moreover, if $M = 2$, we consider also the relative treatment effects for β_{2sj}.

Note that for the Bayesian NMA model described here, i.e., for (5.6) to facilitate consistent comparisons of treatments that are not directly compared by the same study with those that are, the implicit assumption regarding the relative treatment effects is that $d_{j_1} - d_{j_2} = (d_{j_1} - d_1) - (d_{j_2} - d_1)$. Inconsistency can occur if there are systematic differences in relative treatment effect modifiers between different direct comparisons. For the consistency assumption to transfer through (5.5) and (5.4), the same B-splines (i.e., order and knots) and the same FPs (i.e., order and powers) have to be used across all studies and treatment arms. Finally, note that the described model does not account for correlations stemming from trials with more than two treatment arms but can be easily extended to consider them (see Cooper et al., 2009b).

5.5 MCMC algorithm

Posterior inference for all simulated examples and real data applications is performed via standard Gibbs sampling, implemented in JAGS (see Plummer, 2003). The first 5,000 iterations are discarded as ’burn-in’ and the final

sample size on which inference is based is 10,000 samples. Convergence of the chain is checked through the Gelman-Rubin potential scale reduction factor (see Gelman and Rubin, 1992). The Gelman-Rubin diagnostic is evaluated by running multiple chains from different initial values and comparing the estimated between-chains and within-chain variances for each model parameter. Large differences between these variances indicate that convergence has not been reached yet.

5.6 Simulation Study

We replicate the simulation study of Chapter 4 to compare the B-spline model with the FP approach. In the previous Chapter, we design various simulation scenarios for continuous and binary responses that involve widely differing temporal patterns of treatment effects. Moreover, we consider data sets simulated from the models by Dakin et al. (2011) and Ding and Fu (2013). For each scenario of the simulation study, we fit the B-spline and FP models to 50 randomly generated data sets and report results as averages. Following Section 4.2 we consider a closed network, see Figure 5.1 (left panel), as well as non-closed network, see Figure 5.1 (right panel). In particular, study 1 directly compares treatments A and B with follow-up at four time points (weeks 4, 8, 12 and 24). Study 2 compares treatments A and C at three time points (weeks 4, 12, 24), while study 3 compares treatments B and C at three time points (weeks 4, 8, 12). Finally, study 4 compares treatments B and D at four time points (weeks 4, 8, 12, 24).

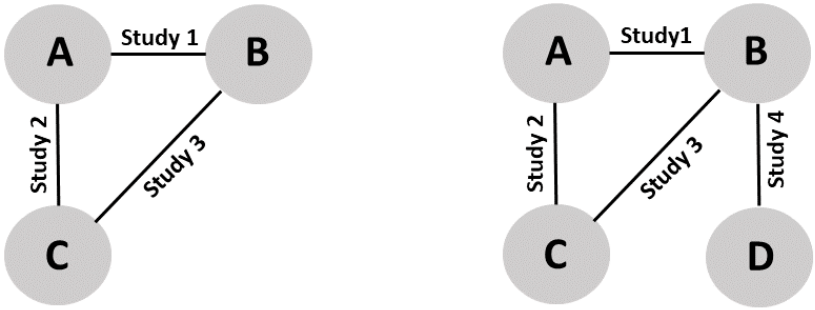


Figure 5.1: Network of studies used in the simulation scenarios. Treatments are represented as nodes and connected by an edge if a direct comparison study between them is available. The left network is closed, with one study comparing all possible pairs of treatments, while in the right network it is extended to become non-closed, containing treatment pairs not directly compared by at least one study.

Let us summarise how the data sets are generated. For **Scenario 1–5**, the main outcomes of each study (s), treatment (j) and time point (t) are reported as

$$Y_{sjt} \sim \text{Normal } \left(\gamma_{jt} + \alpha_s, \tau_s^2/n_s\right),$$

where $\alpha_s \sim$ Normal$(0, 10)$, the number of subjects per treatment arm in the different studies are chosen as $n_1 = 100, n_2 = 120$ and $n_3 = 130$ and as variances we set $\tau_1^2 = 1$, $\tau_2^2 = 2$ and $\tau_3^2 = 4$. Note that the variance is assumed as constant over time and across treatment arms. (This assumption may not hold in practice; see the real data applications below.) The simulated temporal patterns γ_{At}, γ_{Bt} and γ_{Ct} in the scenarios are (i) linear (**Scenario 1**), (ii) logarithmic (**Scenario 2**), (iii) piecewise linear monotonic (**Scenario 3**), (iv) a mix of the above with one of the treatment effects being constant (**Scenario 4**), and (iii) piecewise linear non-monotonic (**Scenario 5**). A description of the simulation study setup is reported in Table 5.1.

	γ_{At}	γ_{Bt}	γ_{Ct}
Linear	$-t/4$	$-t/2$	$-t/3$
Logarithmic	$-\log(t+1)$	$-3\log(t+1)$	$-2\log(t+1)$
Piecewise linear monotonic	$\begin{cases} -\frac{3t}{4}, & t \in [0,4] \\ -\frac{t}{2}-1, & t \in (4,8] \\ -\frac{t}{4}-3, & t \in (8,12] \\ -6, & t \in (12,24] \end{cases}$	$\begin{cases} -\frac{3t}{2}, & t \in [0,4] \\ -\frac{10t}{23}-\frac{17}{4}, & t \in (4,8] \\ -\frac{5t}{16}-\frac{21}{4}, & t \in (8,12] \\ -\frac{t}{12}-8, & t \in (12,24] \end{cases}$	$\begin{cases} -t, & t \in [0,4] \\ -\frac{3t}{4}-1, & t \in (4,8] \\ -\frac{t}{4}-5, & t \in (8,12] \\ -\frac{t}{12}-7, & t \in (12,24] \end{cases}$
Mixed	$\begin{cases} 0, & t \in [0,8] \\ -\frac{t}{2}+4, & t \in (8,24] \end{cases}$	0	$(-t^2-25t)/100$
Piecewise linear non-monotonic	$\begin{cases} -\frac{3t}{4}, & t \in [0,4] \\ -\frac{5t}{2}+2, & t \in (4,8] \\ -\frac{t}{2}-4, & t \in (8,12] \\ \frac{t}{2}-16, & t \in (12,24] \end{cases}$	$\begin{cases} -\frac{3t}{2}, & t \in [0,4] \\ -t-2, & t \in (4,8] \\ t-18, & t \in (8,12] \\ \frac{t}{4}-9, & t \in (12,24] \end{cases}$	$\begin{cases} -2t, t \in [0,4] \\ \frac{t}{4}-9, t \in (4,12] \\ \frac{t}{6}-8, t \in (12,24] \end{cases}$

Table 5.1: Temporal patterns for treatment effects of treatments A,B and C used in the simulations 1–5. The mixed scenario contains a constant treatment effect for treatment B and a quadratic effect for treatment C.

The main outcomes for the **Scenario 8** (non-closed network) are generated in the same way. The additional treatment D is given the temporal effect pattern

$$\gamma_{Dt} = \begin{cases} -t, & t \in [0,4] \\ -\frac{3t}{4}-1, & t \in (4,8] \\ -\frac{t}{4}-5, & t \in (8,12] \\ -\frac{t}{12}-7, & t \in (12,24] \end{cases}.$$

The number of observations is chosen as $n_4 = 110$, and as variance we assume $\tau_4^2 = 3$.

While, in **Scenario 6** the temporal patterns have been generated according to the mixed treatment comparison (MTC) model of Dakin et al. (2011). Mean outcomes are modelled as

$$\theta_{sjt} = \mu_{st} + \delta_{sj},$$

the sum of a study- and time-specific effect across treatment arm and a study-specific arm deviation, where

$$\mu_{st} \sim \text{Normal } (-s, 1),$$
$$\delta_{sj} \sim \text{Normal } (0.5, 1),$$

i.e., in this model relative treatment effects are assumed to be constant over

time. When generating the data we let $\sigma_{sjt} = 1.2/\sqrt{n_s}$, i.e., we assume no within-study correlation between subsequent time points.

Finally, in **Scenario 7** the temporal patterns have been generated from the Bayesian evidence synthesis techniques – integrated two-component prediction (BEST-ITP) model of Ding and Fu (2013). Mean outcomes are modelled as

$$\theta_{sjt} = (\phi_s + \delta_j)\left(\frac{1 - e^{-p_j t}}{1 - e^{-p_j T}}\right),$$

where

$$\phi_s \sim \text{Normal}(-3, 1),$$

and $\delta_1 = 0$, $\delta_2 = 0.5$ and $\delta_3 = 1$ are the treatment effects at the end follow-up time T, to which they increase according the shape given by the exponential function factor, in which we let $p_1 = -0.1$, $p_2 = -0.15$ and $p_3 = -0.15$.

For a detailed description of the different scenarios, see Section 4.2.

In Figures 5.2- 5.9 we present, for continuous outcomes in **Scenario 1**– **Scenario 8**, a comparison of the estimated profiles obtained using the B-spline model and the FP model, while in Figures 5.10- 5.11 we present a comparison of the estimated profiles obtained using the B- spline model and the FP model for binary outcomes in **Scenario 3** and **Scenario 5** respectively. We remind that binary outcomes are simulated by taking the inverse logit transformation of θ_{sjt} and then generating individual level outcomes from a Bernoulli distribution (see Equation (5.3)). Red lines indicate the true values used to simulate the data, whereas the blue lines and the black lines indicate the estimates obtained using the fractional polynomials model and the B-spline model, respectively.

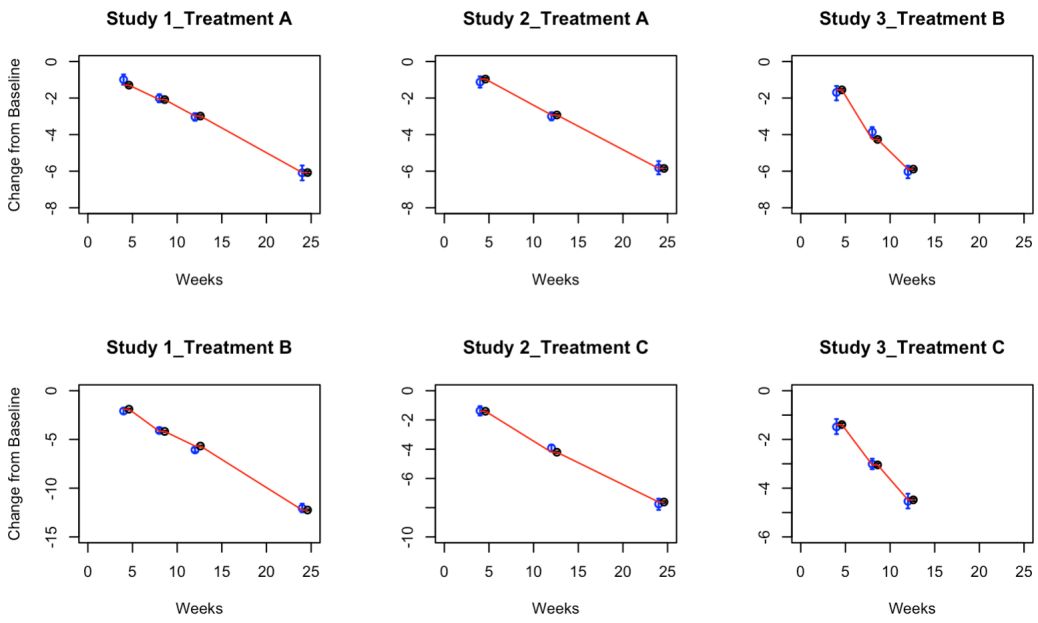


Figure 5.2: *Scenario 1*: posterior estimated profiles obtained. Red lines indicate the values used to simulate the data, whereas the blue lines and the black lines correspond to the estimates obtained using the fractional polynomials model and the B-spline model, respectively, along with 95% credible intervals.

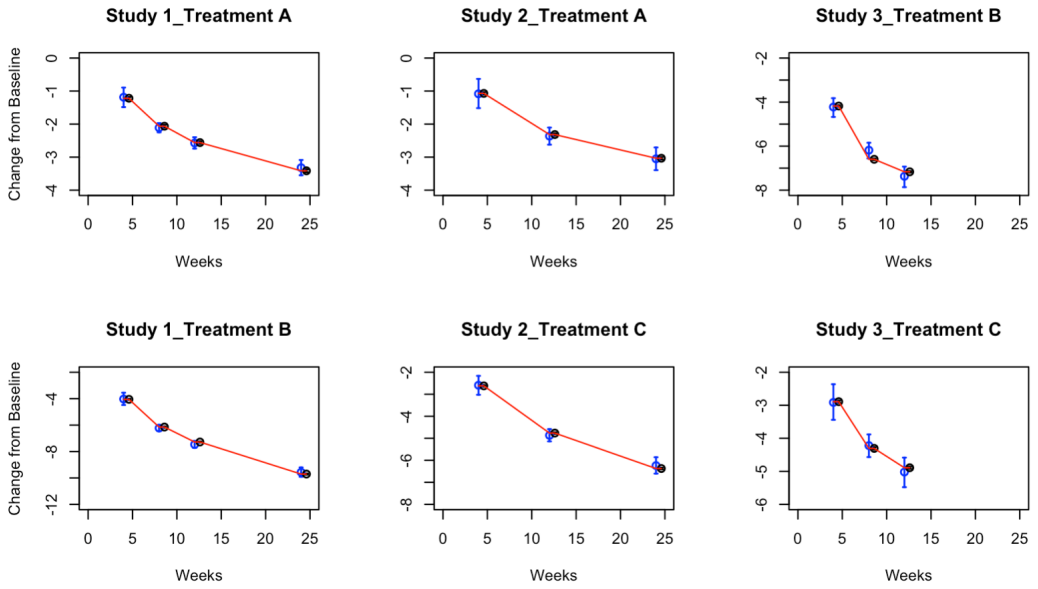


Figure 5.3: *Scenario 2*: posterior estimated profiles obtained. Red lines indicate the values used to simulate the data, whereas the blue lines and the black lines correspond to the estimates obtained using the fractional polynomials model and the B-spline model, respectively, along with 95% credible intervals.

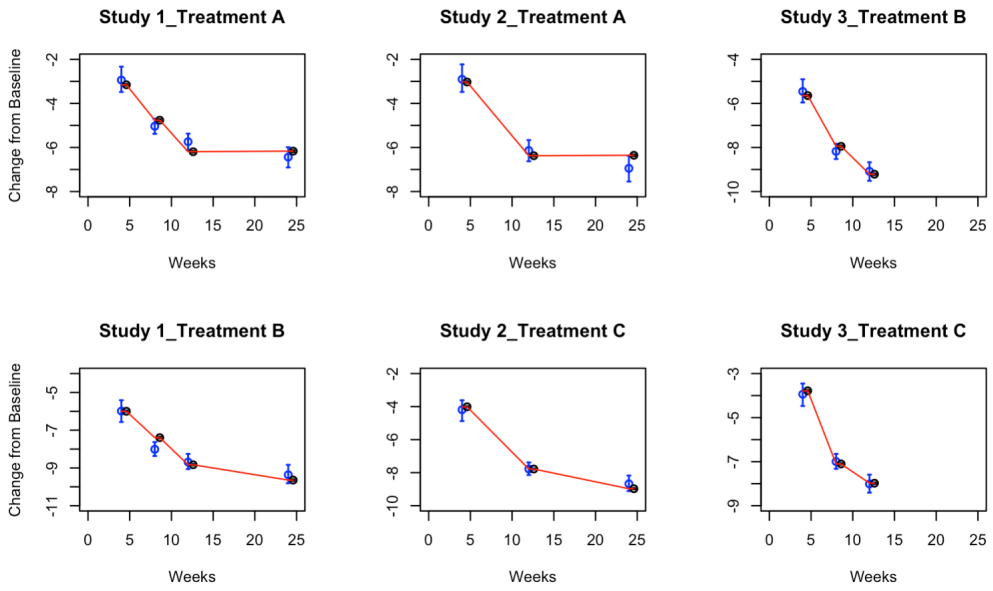


Figure 5.4: *Scenario 3*: posterior estimated profiles obtained. Red lines indicate the values used to simulate the data, whereas the blue lines and the black lines correspond to the estimates obtained using the fractional polynomials model and the B-spline model, respectively, along with 95% credible intervals.

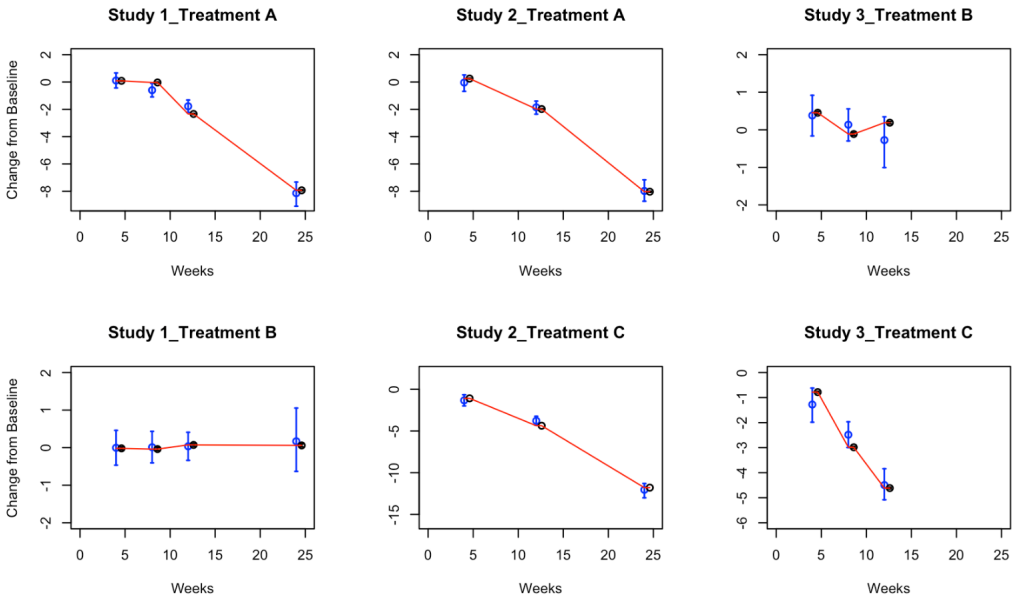


Figure 5.5: *Scenario 4*: posterior estimated profiles obtained. Red lines indicate the values used to simulate the data, whereas the blue lines and the black lines correspond to the estimates obtained using the fractional polynomials model and the B-spline model, respectively, along with 95% credible intervals.

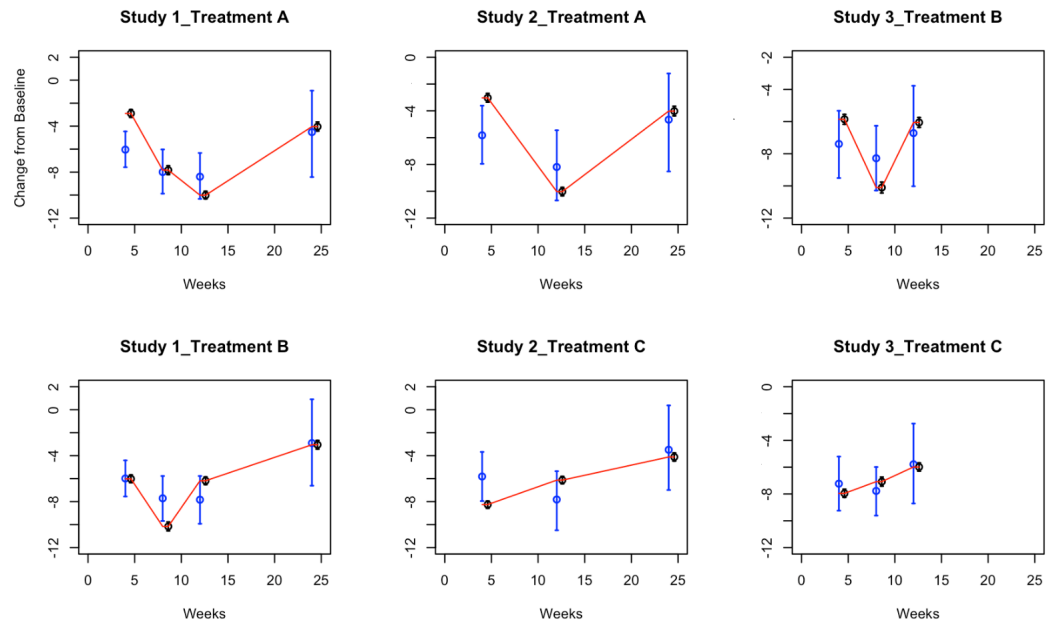


Figure 5.6: *Scenario 5*: posterior estimated profiles obtained. Red lines indicate the values used to simulate the data, whereas the blue lines and the black lines correspond to the estimates obtained using the fractional polynomials model and the B-spline model, respectively, along with 95% credible intervals.

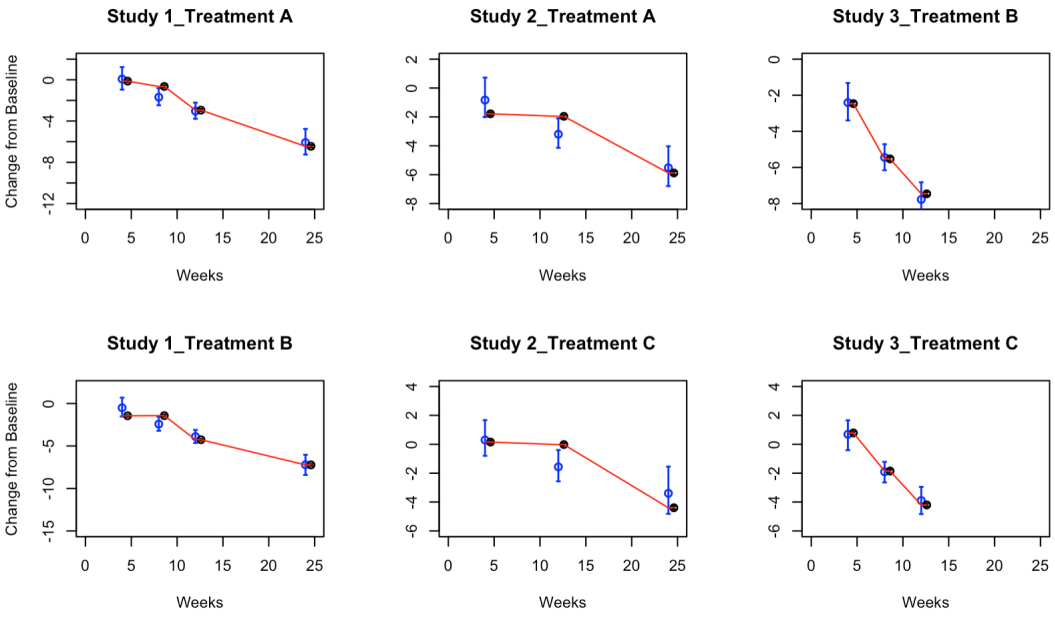


Figure 5.7: *Scenario 6*: posterior estimated profiles obtained. Red lines indicate the values used to simulate the data, whereas the blue lines and the black lines correspond to the estimates obtained using the fractional polynomials model and the B-spline model, respectively, along with 95% credible intervals.

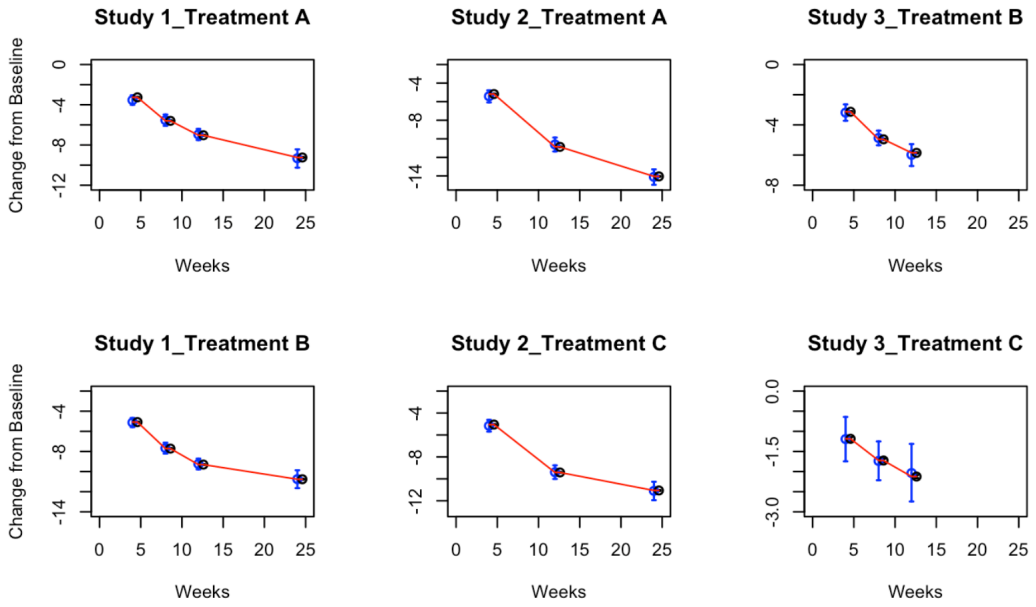


Figure 5.8: *Scenario 7*: posterior estimated profiles obtained. Red lines indicate the values used to simulate the data, whereas the blue lines and the black lines correspond to the estimates obtained using the fractional polynomials model and the B-spline model, respectively, along with 95% credible intervals.

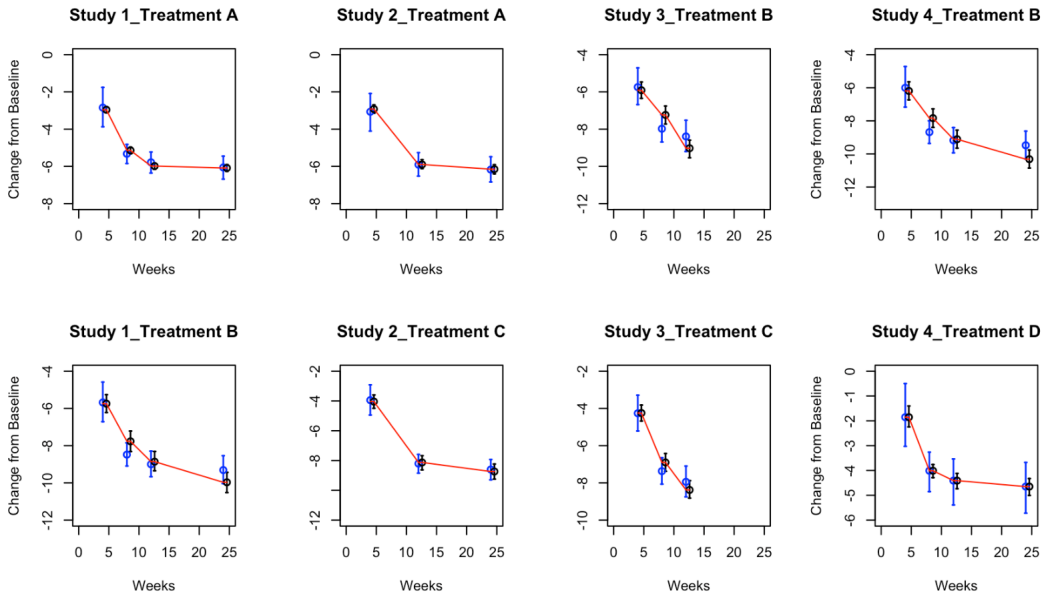


Figure 5.9: *Scenario 8: Non-closed network.* Posterior estimated profiles. Red lines indicate the values used to simulate the data, whereas the blue lines and the black lines indicate the estimates obtained using the fractional polynomials model and the B-spline model, respectively, along with 95% credible intervals.

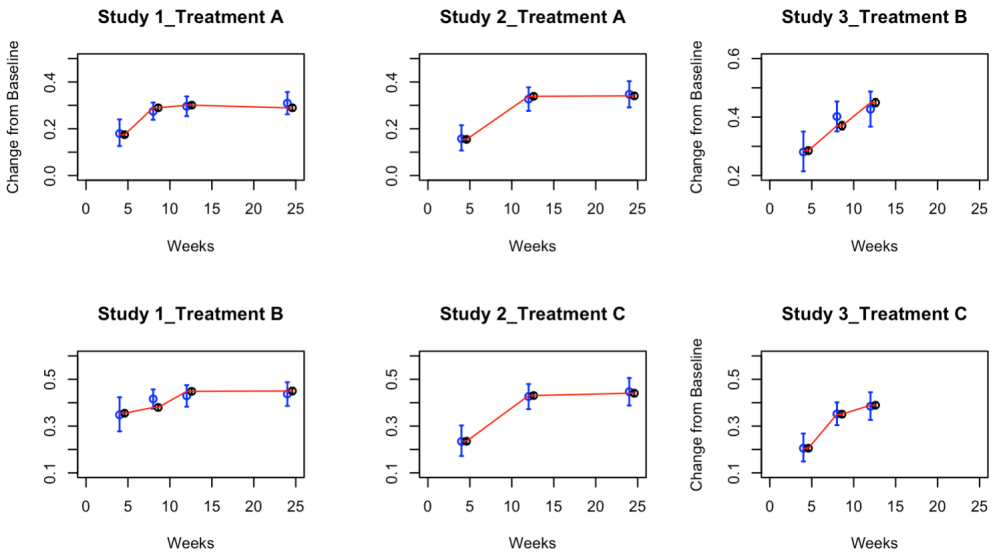


Figure 5.10: *Scenario 3: binary outcome*. Posterior distribution of p_{sjt}. Red lines indicate the values used to simulate the data, whereas the blue lines and the black lines correspond to the estimates obtained using the fractional polynomials model and the B-spline model, respectively, along with 95% credible intervals.

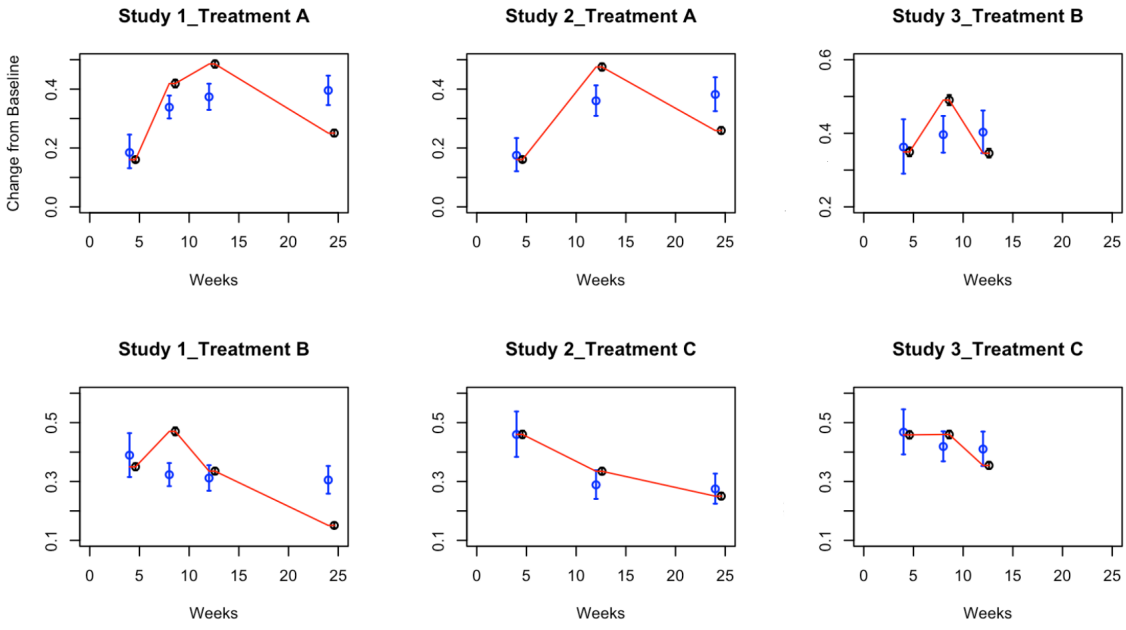


Figure 5.11: *Scenario 5: binary outcome*. Posterior distribution of p_{sjt}. Red lines indicate the values used to simulate the data, whereas the blue lines and the black lines correspond to the estimates obtained using the fractional polynomials model and the B-spline model, respectively.

These simulation results demonstrate the improvement of estimation of the treatment effects when utilizing B-spline model in the different scenarios. In fact, the 95% credible intervals obtained with the B-splines are narrow and always cover the true value used to generate the data. In particular, this is evident in **Scenario 5**. From the Figures 5.6 and 5.11, it is clear that the B-spline model is able to accurately estimate non-monotonic time effects in both the continuous and binary cases, which the FP model fails to capture. In Figure 5.12, we show estimated differences in treatment effects for one of the 50 simulation replicates in the continuous case. It is evident that the B-spline model is able to capture the non-monotonic curve shape for difference in treatment, while the FP model, flattens out the difference in treatment with consequent increase of uncertainty.

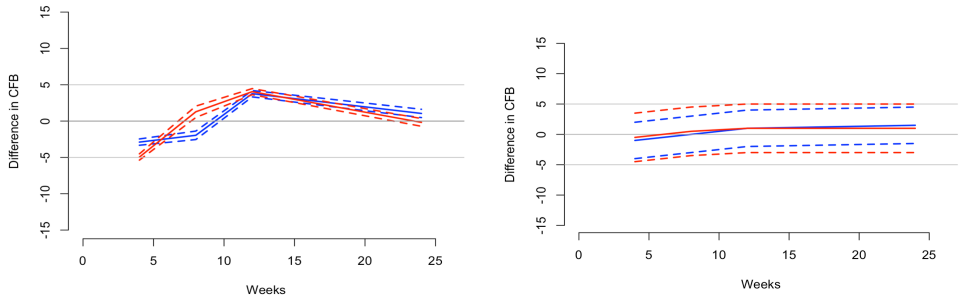


Figure 5.12: Difference in treatments for the non-monotonic simulation scenario. Treatments for A versus B (blue), and treatments A versus C (red) are shown in the left panel for the B-spline model and in the right panel for the FP model.

To conclude this Section, we report for an overall quantitative comparison of goodness of fit the mean squared error (MSE) evaluated for each simulation scenario across all replicates. The MSE is defined as

$$\text{MSE} = \frac{1}{R}\sum_{r=1}^{R}\sum_{s,j,t}\left(Y_{sjt} - \hat{Y}_{sjt}^{(r)}\right)^2,$$

where $\hat{Y}_{sjt}^{(r)}$ denotes the predicted value of the response for study s, treatment j, time t and MCMC iteration r. The number of saved MCMC iterations is $R = 10,000$. For all simulation scenarios the MSEs for the B-spline model and the FP model are summarized in Table 5.2. The table reports the means (and standard deviations) of the MSEs across the 50 simulations, quantitatively showing an increase in estimation accuracy for the B-spline model as compared to the FP model in all scenarios, and particularly in the non-monotonic scenario simulated (**Scenario 5**).

Scenario	B-spline model	FP model
Linear	$0.001954\ (0.0213)$	$0.3505\ (0.0388)$
Logarithmic	$0.001454\ (0.0111)$	$0.3505\ (0.0118)$
Piecewise linear monotonic	$0.001121\ (0.0923)$	$1.3069\ (0.1823)$
Mixed	$0.003709\ (0.0313)$	$1.5747\ (0.1338)$
Non-monotonic	$0.001980\ (0.0389)$	$41.7533\ (1.8354)$
MTC	$0.037441\ (0.0483)$	$0.9198\ (0.2864)$
BEST-ITC	$0.03971\ (0.0503)$	$0.1701\ (0.2898)$
Piecewise linear monotonic (binary)	$0.003491\ (0.0118)$	$0.0897\ (0.3698)$
Non-monotonic (binary)	$0.000401\ (0.0019)$	$0.9748\ (0.2976)$
Piecewise linear (non-closed network)	$0.0172261\ (0.0173)$	$0.0903\ (0.3854)$

Table 5.2: Mean Square Error (MSE) comparison for B-spline and fractional polynomial (FP) network meta-analysis models. Displayed are the mean (standard deviation) of the respective MSEs averaged over 50 simulated data sets for each scenario. Unless stated otherwise, considered outcomes are continuous. Scenarios that contain non-monotonic temporal behaviors appear to be challenging for the FP method.

From the results of this Section, it is apparent that the B-spline model offers a flexible strategy that is able to accommodate different time patterns and correlation structures among time points without the computational burden of the FP model, while still leading to better results. Cubic splines with four uniformly spaced knots appear to be able to provide accurate estimates and seem a natural choice, since it is common in NMA that few time measurements are available per treatment arm and study. This will be further underlined by the real data experiments conducted in the next Section.

5.7 Real data application

We consider five real NMA data sets. The description of the first three data sets, besides in the original references below, can also be found in Chapter 4. We provide the networks of studies for all data sets in Figures 5.13 and 5.14.

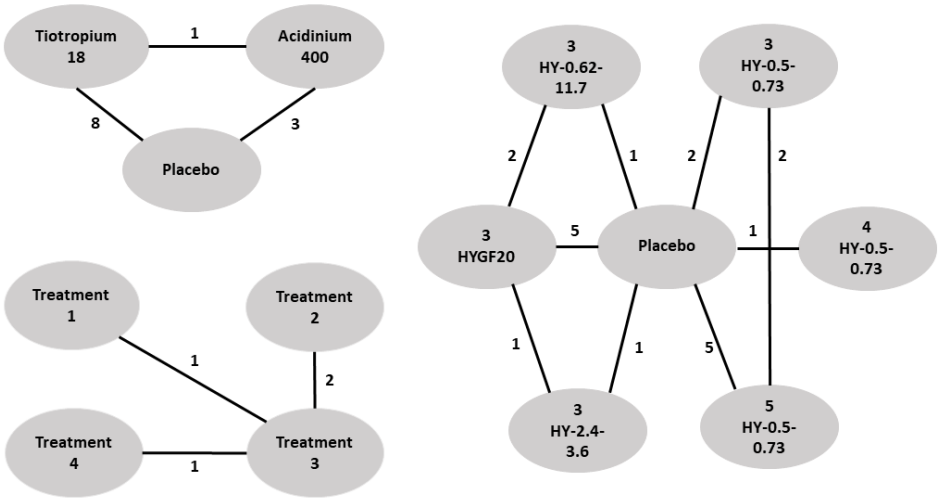


Figure 5.13: Networks of studies of the COPD data set (top left), the T2D data set (bottom left) and the OA data set (right). Edges indicate the number of studies providing direct comparison between adjacent treatments.

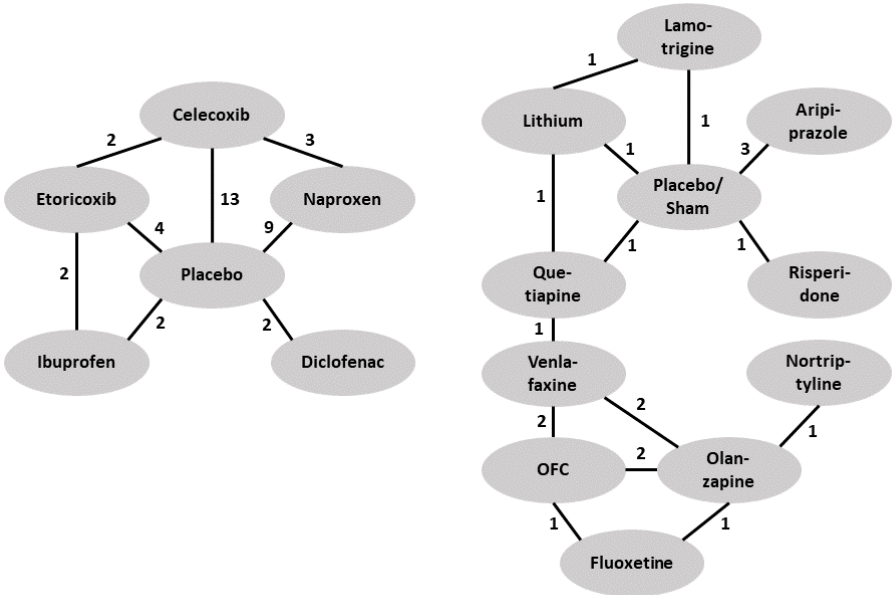


Figure 5.14: Networks of studies of the CMP data set (left) and the MMD data set (right). Edges indicate the number of studies providing direct comparison between adjacent treatments.

Main features of all data sets are summarized in Table 5.3.

	COPD	OA	T2D	CMP	MDD
Number of studies	13	16	4	27	13
Number of treatments	3	7	4	6	10
Number of follow ups	$2-6$	$1-10$	$2-5$	2	$3-4$
Number of patients per study	$46-3006$	$20-295$	$88-575$	$12-6769$	$24-1147$
Network type	closed	non-closed	non-closed	non-closed	non-closed

Table 5.3: Summary of main features of the chronic obstructive pulmonary disease (COPD), osteroarthritis (OA), type 2 diabetes (T2D), chronic muskuloskeletal pain (CMP), and treatment resistant major depressive disorder (MDD) data sets.

(i) The first data set contains information of 13 studies on patients affected by chronic obstructive pulmonary disease (COPD) and is described in Karabis et al. (2013).

(ii) The second data set, described in Jansen et al. (2015), contains a systematic review of six studies that investigate treatments for osteoarthritis (OA) of the knee.

(iii) The third data set, analyzed in Ding and Fu (2013), provides information of four studies on patients affected by type 2 diabetes (T2D).

(iv) The fourth data set is part of a large collection of studies on benefits and risks of drugs for treating chronic musculoskeletal pain (CMP) in patients with osteoartrithis or rheumatoid artrithis. The data is collected in van Walsem et al. (2015), where the authors conduct a Bayesian NMA. Here we concentrate on pain relief as the outcome of interest, measured by visual analogue scale (VAS), and only include studies with two follow-ups. Treatments are Diclofenac, Naproxen, Ibuprofen, Celecoxib, and Etoricoxib, which are compared to placebo. Outcomes are reported at 6 and 12 weeks (within 2-week range).

(v) The fifth data set is taken from a comparative review of efficacy and tolerability of interventions for treatment resistant major depressive disorder (MMD) by Papadimitropoulou et al. (2017). Interventions are Aripiprazole, Fluoxetine, Lamotrigine, Lithium, Nortriptyline, Olanzapine, Olanzapine/Fluoxetine combination (OFC), Quetiapine, Risperidone, Venlafaxine, which are compared to placebo/sham. The main outcome as CFB is measured on the Montgomery-Asberg Depression Rating Scale (MADRS). Outcomes are reported at $4, 6$ or 8 weeks.

For the data sets for which no information about the loss to follow-up is available, we assume that the sample sizes remain the same as at the start of the

study. For studies within the NMAs that do not report the number of patients, we generate the sample size from a uniform distribution ranging from the minimum to the maximum number of patients across all studies of the respective NMA. Finally, following Dakin et al. (2011), we impute any missing standard deviation information specifying the prior distributions

$$\begin{aligned} \text{sd}_{sjt} &\sim \text{Gamma}(\alpha_1, \alpha_2), \\ \alpha_i &\sim \text{Uniform}(0, 10). \end{aligned}$$

The results of the real data applications confirm those of the simulation study. For the FP fit we considered first and second order FPs. The estimated profiles using the B-spline (left panel) and FP (right panel) models are presented in Figures 5.15-5.19.

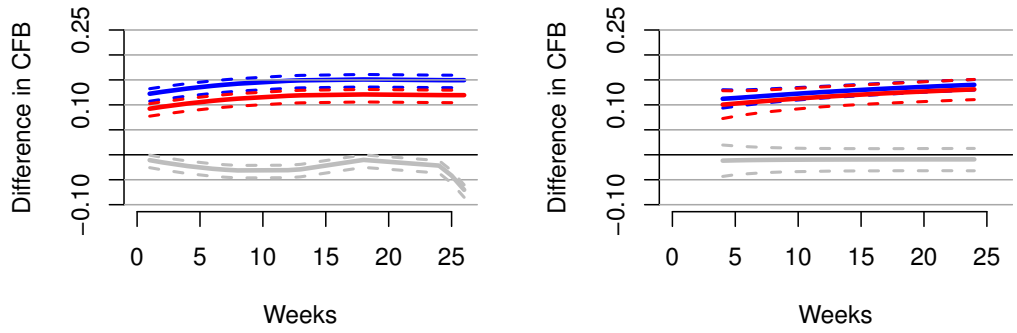


Figure 5.15: Results for the chronic obstructive pulmonary disease (COPD) data set. Shown are estimation results as mean differences in treatment effects over time for AB400 versus placebo (red), TIO18 versus placebo (blue) and AB400 versus TIO18 (grey) generated by the B-spline model (left panel) and the FP model (right panel). Dashed lines denote 95% credible intervals, while solid lines represent the posterior mean. These results are in agreement with the conclusions in Karabis et al. (2013).

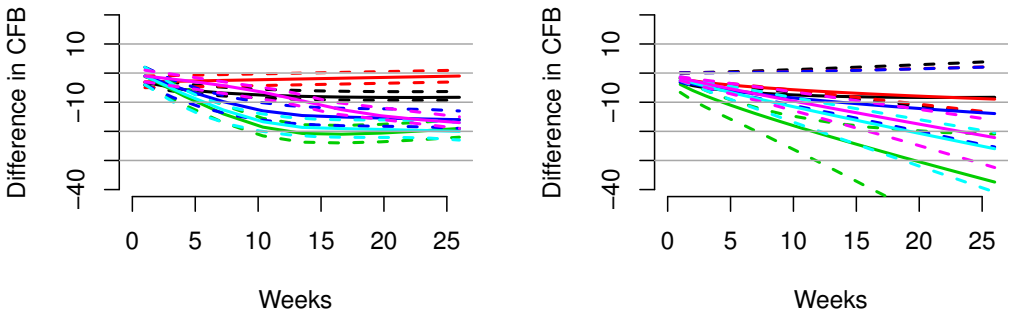


Figure 5.16: Results for the osteoarthritis (OA) data set. Shown are treatment effects relative to placebo over time for 5 HY-0.5-0.73 (magenta), 4 HY-0.5-0.73 (black), 3 HY-0.5-0.73 (red), 3 HYGF20 (green), 3 HY-0.62-11.7 (blue) and 3 HY-2.4-3.6 (light blue) estimated by the B-spline model (left panel) and and the FP model (right panel). Dashed lines denote 95% credible intervals, while solid lines represent the posterior means. Both models suggest that the best results are obtained by 3 HYGF20, in agreement with the result obtained originally for this data with FPs by Jansen et al. (2015).

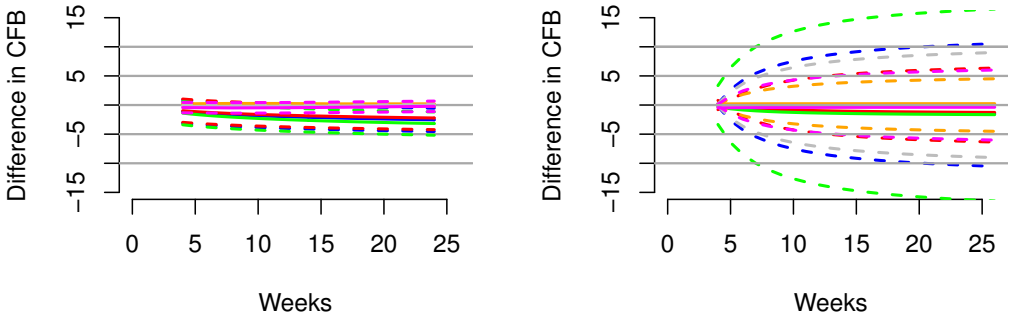


Figure 5.17: Results for the type 2 diabetes (T2D) data set. Shown are mean differences in treatment effects over time for treatments 1 versus 2 (blue), 1 versus 3 (red), 1 versus 4 (green), 2 versus 3 (orange), 2 versus 4 (grey) and 3 versus 4 (magenta) generated by the B-spline model (left panel) and FP model (right panel). Dashed lines denote 95% credible intervals, while solid lines represent posterior means.

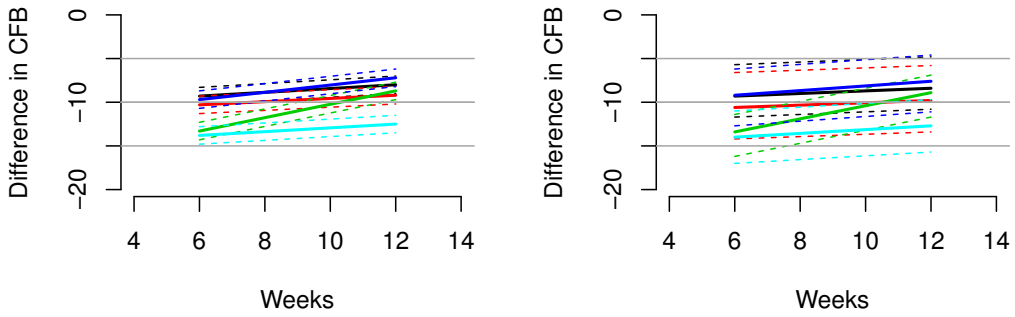


Figure 5.18: Results for chronic musculoskeletal pain (CMP) data set. Shown are treatment effects relative to placebo over time for Celecoxib (black), Naproxen (red), Etoricoxib (green), Ibuprofen (blue), and Diclofenac (light-blue) estimated using the B-Spline model (left panel) and the FP model (right panel). Dashed lines denote 95% credible intervals, while solid lines represent posterior means. Both figures are in agreement with the result of van Walsem et al. (2015) that Diclofenac is the most effective.

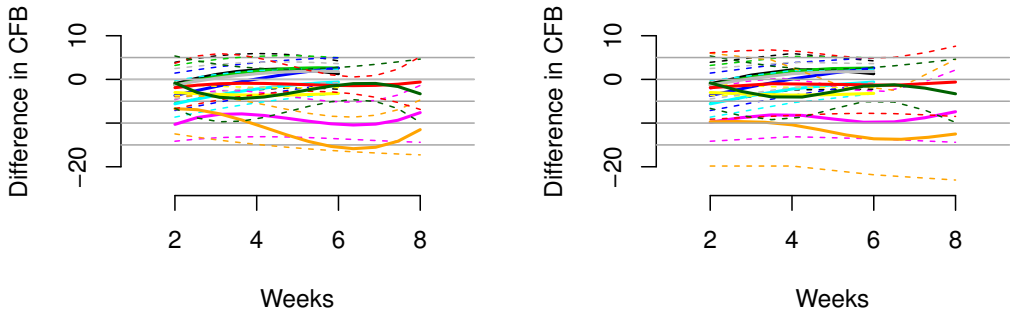


Figure 5.19: Results for the major depressive disorder (MMD) data set. Shown are treatment effects relative to placebo over time for Nortriptyline (black), Lithium (red), Fluoxetine (green), Olanzapine (blue), OFC (ligth-blue), Risperidone (magenta), Aripiprazole (yellow), Venlafaxine (grey), Quetiapine (orange) and Lamotrigine (dark green) estimated using the B-Spline model (left panel) and the FP model. Dashed lines denote 95% credible intervals, while solid lines represent posterior means. The results are in qualitative agreement with Papadimitropoulou et al. (2017).

B-splines provide narrower credible intervals for the posterior mean estimates, while still being able to detect slight changes over time. As a rule of thumb, if the 95% credible interval of the posterior distribution of the difference between two treatments does not cover zero, then we can conclude there is evidence in support of different effectiveness between them. In the results for the OA data it is evident that the B-spline model is able to detect non-monotonic time patterns, while the FP model forces a monotonic trend given the parameter choice which is dictated by the DIC. Furthermore, for the T2D and the CMP examples, the shape and size of the credible intervals obtained with the FP model is due to the choice (by DIC) of a first order polynomial with power $p = -2$. Any choice of FP parameters generally imposes structure on the temporal pattern, which might not always be supported by the data. In particular, this is the case for the choice of the function β/t^2 when, as for the CMP data, only two data points per study and treatment arm are available. In Table 5.4 we report the specific FP selected by DIC for each data set.

Data set	M	p_m	RE/FE
COPD	2	p_1 = -0.5, p_2 = 0	RE on β_2
OA	2	p_1 = 0.5, p_2 = 1	RE on β_2
T2D	1	p_1 = -2	RE on β_1
CMP	1	p_1 = -1	FE
MMD	2	p_1 = 0.5, p_2 = 0	FE

Table 5.4: Real data sets: DIC selection of the order M and of the power terms p_m in the FP model. RE refers to random effect, while FE to fixed effects.

These examples confirm that the DIC might not be the optimal model choice criterion when confronted with non-linear effects. Indeed, while the DIC has been shown to be an approximation to a penalized loss function based on the deviance with a penalty derived from a cross-validation argument, it has been warned that this approximation is in fact only valid when the effective number of parameters p_{eff} in the model is considerably smaller than the number of independent observations n (see Plummer, 2008). Since the latter assumption is usually not satisfied in the case of NMAs the DIC can tend to under-penalize more complex models. The poor empirical performance of DIC compared to cross-validation when the assumption $p_{\text{eff}} \ll n$ is violated has also been highlighted by Vehtari and Lampinen (2002).

5.8 Conclusions

In this Chapter we propose a random effect model for NMA of repeated measurements based on B-splines. The model is able to accommodate a large class of temporal treatment effect patterns, allowing for direct and indirect comparisons of widely varying shapes of longitudinal profiles. We argue for a fixed choice of the order and automatic selection of uniform knots of the B-spline. The model is not restricted to continuous or binary outcomes, but can be extended to any type of response variable by specifying an appropriate link function. We show that the B-spline model overcomes the methodological and computational limitations of the FP approach, which, beyond requiring extensive sensitivity analysis for each new scenario, might also suffer from potential shortcomings of the DIC. In detail, investigation of model choice criteria for FP selection is needed, but it is beyond the scope of this work. One of the main consequences of a sub-optimal choice of the polynomial can be observed in the uncertainty quantification as represented by the credible intervals for the FP model. The B-spline model is useful in understanding treatment effects as well as between-study variability. Moreover, this model naturally allows for different numbers of observations per study, different times of observations, different sample sizes across studies, as well as missing data. Finally, there are arguments for better interpretability of its coefficients as compared to the FP model. The concerns of Jansen et al. (2015), about splines with fixed number of uniformly spaced knots being too restrictive in possible curve shapes, are not supported by our results in simulations and real data applications, which highlight, on the contrary, the flexibility and efficiency of the B-spline approach.

Chapter 6

Final remarks

We conclude this work summarizing the main contributions of this research. For each of the previous chapters, we focus on the used modelling strategy and we review the motivating applications and results. In addition, we present open research questions and possible extension to the models.
This work focuses on the application of Bayesian models in Biostatistics. The aim is two fold: i) recurrent event and survival time, ii) network meta-analysis. In both cases, the health care industry must develop strategies aimed at containing the economic burden of the care, without compromising its effectiveness.

On the one hand, it is essential to recognize the clinical factors that contribute to the disease progression and the efficacy of medical interventions. Therefore, models that allow for situations in which you can have independence among gap times as well as temporal dependence are necessary. Moreover, since most analyses focus only on the primary clinical outcomes (such as death or time to first re-hospitalization), they ignore by construction the information contained in the recurrent event process and the relationship between multiple hospital admissions and patient's survival. Hence, it is essential to specify a model with a joint distribution of the length and recurrence of hospitalizations and event (termination) time in order to overcome the above mentioned limitation.

On the other hand, since another important goal in the healthcare fold is the relative treatments effects and cost effectiveness of different interventions and since HTA is ultimately focused on decision-making of new interventions, it is fundamental to gain a deeper understanding of the NMA techniques. Our first objective is to provide a comparative review of the most recent approach for longitudinal data using a simulation study as well as real-world data. We aim to identify specific characteristics that would make the model more appropriate in each scenario. After that, we propose an NMA method based

on B-splines to model temporal behavior, which allows for the simultaneous analysis of outcomes at different time points, automatically accounting for correlation across time.

In Chapter 2 we introduce the first contribution of this work, proposing a new Bayesian nonparametric model for recurrent event data. The literature on this field is based on two different approaches. The first method models the intensity or hazard function of the event counts process, while the second one performs inference modelling the sequence of times between recurrent events, known as gap times or waiting times. The latter approach is more appropriate when the events are relatively infrequent, when individual renewal takes place after an event, or when the goal is prediction of the time to the next event. Since the events in our application are infrequent but measured on many individuals, we develop a Bayesian semiparametric models for gap times between events based on this approach. In particular, time-dependency among waiting times is taken into account through an autoregressive model, whose parameters are a sample from DP, which is the main stochastic process over discrete probability measures used in BNP. Therefore, a clustering structure among the patients is allowed. We adopt a flexible strategy and we focus on the test of the order of dependence among random effects at different times, using two different approach: spike and slab base measure and direct prior specification on the order of dependence.
After that, we analyze the posterior marginal distribution of the latent variables of DP and the posterior of the number of clusters, i.e. the number of groups that the latent variables has done. Both fixed and time-dependent covariates may be included in this framework. This model is useful for health care management, whose interest is the prediction of the next hospitalization in order to plan the resources in the most appropriate way. Moreover, it can be extended to model multiple time series analyses in other fields, like economy and technology.

Albeit already useful for hospital planning, the introduction of the eventual death of the subject would allow to use this model for medical purposes too. In order to do so, we propose in Chapter 3 a nonparametric method that models the joint distribution of the gap times and the survival time. We explicitly model the number of events and, conditional on the number of events, we specify a joint distribution for gap times and survival. This choice captures the dependence between the recurrence and survival processes, which is an important feature in medical applications and beyond. Subject-specific random effects capture the frailty by informing both the survival time and the dependence of subsequent gap times. These random effects are modeled flexibly

with a DP. The discreteness of the DP induces clustering of the subjects in the sample, based on the unique values of the random effects, where the number of clusters is learned from the data. This choice allows for extra flexibility, variability between individual trajectories, overdispersion and clustering of the observations, and overcomes the often too restrictive assumptions underlying a parametric distribution. We use different random-effects parameters in the distribution of the recurrence and in the distribution of the survival. This still induces dependence between recurrence and survival since both parameters share the same DP cluster. Additionally, the way we model the number of recurrent events also induces dependence between recurrence and survival.

In Chapter 4 we introduce the second main argument of this work: network meta analysis (NMA). This technique compares multiple treatments using both direct comparisons, obtained within randomized clinical trials, and indirect comparisons across trials having a common comparator, such as a placebo or standard treatment. The three most promising methods found in literature are: Mixed Treatment Comparison (MTC), developed by Dakin et al. (2011); the Bayesian evidence synthesis techniques — integrated two-component prediction (BEST-ITP) developed by Ding and Fu (2013); and the more recent method based on fractional polynomials by Jansen et al. (2015). The model presented by Dakin et al. (2011) assumes random relative effects and allows for the relative treatment effects to vary over time, without temporal pattern restriction. Ding and Fu (2013) suggested a parametric model to describe a nonlinear relationship between outcome and time for each treatment, assuming a diminishing return time course of treatment responses. Finally, Jansen et al. (2015) suggested a more flexible approach where the relationship between outcome and time is a fractional polynomial. Based on the results of our comparisons, FP methods turn out as the best of the three reviewed approaches. These models are undoubtedly the most flexible and they can accommodate a variety of treatment effect patterns. On the other hand, they still present limitations both from a methodological and a computational point of view; they require intensive computations for the choice of parameters and present modelling drawbacks.

To overcome this limitation, we propose, in Chapter 5, an alternative method using cubic splines, i.e. smooth piecewise polynomial functions to describe outcomes over time. Moreover, although a wide range of curve shapes can be obtained with (second order) fractional polynomials, splines can be more flexible, depending upon the number of nodes. Among the various methods available for constructing polynomial functions, we selected the B-spline method because it provides some useful properties. The continuity of B-

spline's derivatives is guaranteed till the order $n - \mu$ (being n the order of the polynomial function and μ the multiplicity of the knots), thus, by choosing a cubic B-spline with a set of distinguished knots (μ = 1) enables us, to guarantee the B-spline function being C^2 . It is straightforward to understand how this property comes to our advantage in addressing the limitation and hurdle of the Fractional Polynomial model we described here above. The use of the cubic B-splines enables us, by keeping the order of the polynomial function low, to guarantee that the fitting curve is smooth (first order degree function is continuous) and it has the same curvature on the junction points of subsequent subintervals. We finally evaluate the performance of the B-Spline model and we compare our model with the FP model through simulations and real data. It is evident that the B-spline model offers a flexible strategy that is able to accommodate different time patterns and correlation structures among time points without the computational burden of the FP model, while still leading to better results. Moreover, cubic splines with four uniformly spaced knots appear to be able to provide accurate estimates and seem a natural choice, since it is common in NMA that few time measurements are available per treatment arm and study.

To conclude, even though possible future research may be developed on the arguments here analyzed, we hope this Phd project did manage, all-in-all, to provide healthcare management with solid support in its day by day tasks.

Bibliography

Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and event history analysis: a process point of view*. Springer, New York.

Aalen, O. O. and Husebye, E. (1991). "Statistical analysis of repeated events forming renewal processes." *Statistics in Medicine*, 10(8): 1227–1240.

Adalsteinsson, E. and Toumi, M. (2013). "Benefits of probabilistic sensitivity analysis–a review of NICE decisions." *Journal of market access & health policy*, 1(1): 21240.

Aldous, D. J. (1985). "Exchangeability and related topics." In *École d'Été de Probabilités de Saint-Flour XIII*, 1–198. Springer.

Altman, R. (1998). "Intra-articular sodium hyaluronate (Hyalgan®) in." *J Rheumatol*, 25: 2203–12.

Altman, R., Rosen, J., Bloch, D., Hatoum, H., and Korner, P. (2009). "A double-blind, randomized, saline-controlled study of the efficacy and safety of EUFLEXXA® for treatment of painful osteoarthritis of the knee, with an open-label safety extension (The FLEXX Trial)." In *Seminars in arthritis and rheumatism*, volume 39, 1–9. Elsevier.

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. (2003). "An introduction to MCMC for machine learning." *Machine learning*, 50(1-2): 5–43.

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *The Annals of Statistics*, 2: 1152–1174.

Argiento, R., Cremaschi, A., and Guglielmi, A. (2014). "A "density-based" algorithm for cluster analysis using species sampling Gaussian mixture models." *Journal of Computational and Graphical Statistics*, 23(4): 1126–1142.

Asmussen, S., Goffard, P.-O., and Laub, P. J. (2019). "Orthonormal Polynomial Expansions and Lognormal Sum Densities." In *Risk and Stochastics*, 127–150. World Scientific.

Asmussen, S., Jensen, J. L., and Rojas-Nandayapa, L. (2016). "Exponential Family Techniques for the Lognormal Left Tail." *Scandinavian Journal of Statistics*, 43(3): 774–787.

Baio, G. and Dawid, A. (2011). "Probabilistic Sensitivity Analysis in Health Economics." *Statistical Methods in Medical Research*.
URL http://www.ncbi.nlm.nih.gov/pubmed/21930515

Bao, Y., Dai, H., Wang, T., and Chuang, S.-K. (2012). "A joint modelling approach for clustered recurrent events and death events." *Journal of Applied Statistics*, 40(1): 123–140.
URL https://doi.org/10.1080/02664763.2012.735225

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Binder, D. A. (1978). "Bayesian cluster analysis." *Biometrika*, 65(1): 31–38.

Blackwell, D. (1973). "Discreteness of Ferguson selections." *The Annals of Statistics*, 1(2): 356–358.

Blackwell, D., MacQueen, J. B., et al. (1973). "Ferguson distributions via Pólya urn schemes." *The annals of statistics*, 1(2): 353–355.

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2010). "A basic introduction to fixed-effect and random-effects models for meta-analysis." *Research synthesis methods*, 1(2): 97–111.

Botev, Z. I., Salomone, R., and Mackinlay, D. (2019). "Fast and accurate computation of the distribution of sums of dependent log-normals." *Annals of Operations Research*, 280(1-2): 19–46.

Bucher, H. C., Guyatt, G. H., Griffith, L. E., and Walter, S. D. (1997). "The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials." *Journal of Clinical Epidemiology*, 50(6): 683 – 691.
URL http://www.sciencedirect.com/science/article/pii/S0895435697000498

Callens, M. and Croux, C. (2009). "Poverty dynamics in Europe: A multilevel recurrent discrete-time hazard analysis." *International Sociology*, 24(3): 368–396.

Carlin, B. P. and Chib, S. (1995). "Bayesian model choice via Markov chain Monte Carlo methods." *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3): 473–484.

Carrabba, M., Paresce, E., Angelini, M., Re, K., Torchiana, E., and Perbellini, A. (1995). "The safety and efficacy of different dose schedules of hyaluronic acid in the treatment of painful osteoarthritis of the knee with joint effusion." *European journal of rheumatology and inflammation*, 15(1): 25–31.

Chang, S.-H. and Wang, M.-C. (1999). "Conditional regression analysis for recurrence time data." *Journal of the American Statistical Association*, 94: 1221–1230.

Chatterjee, M. and Sen Roy, S. (2018). "A copula-based approach for estimating the survival functions of two alternating recurrent events." *Journal of Statistical Computation and Simulation*, 88(16): 3098–3115.

Christensen, O. (2016). *An introduction to frames and Riesz bases*. Springer-Verlag.

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. E. (2011). *Bayesian ideas and data analysis: an introduction for scientists and statisticians*. CRC Press.

Chung, M. K., Eckhardt, L. L., Chen, L. Y., Ahmed, H. M., Gopinathannair, R., Joglar, J. A., Noseworthy, P. A., Pack, Q. R., Sanders, P., and and, K. M. T. (2020). "Lifestyle and risk factor modification for reduction of atrial fibrillation: A scientific statement from the American Heart Association." *Circulation*, 141(16).

Clyde, M. and George, E. I. (2004). "Model uncertainty." *Statistical Science*, 19: 81–94.

Cook, R. J. and Lawless, J. F. (2007a). *The statistical analysis of recurrent events*. Springer, New York.

— (2007b). *The Statistical Analysis of Recurrent Events*. Springer, New York. URL https://doi.org/10.1007/978-0-387-69810-6

Cooper, H., Hedges, L. V., and Valentine, J. C. (2009a). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

Cooper, N. J., Sutton, A. J., Morris, D., Ades, A., and Welton, N. J. (2009b). "Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation." *Statistics in medicine*, 28(14): 1861–1881.

Corrado, E., Peluso, G., Gigliotti, S., De Durante, C., Palmieri, D., Savoia, M., Oriani, G., and Tajana, G. (1995). "The effects of intra-articular administration of hyaluronic acid on osteoarthritis of the knee: a clinical study with immunological and biochemical evaluations." *European journal of rheumatology and inflammation*, 15(1): 47–56.

Cubukcu, D., Ardıç, F., Karabulut, N., and Topuz, O. (2005). "Hylan GF 20 efficacy on articular cartilage quality in patients with knee osteoarthritis: clinical and MRI assessment." *Clinical Rheumatol*, 24(4): 336–341.

Dakin, H. A., Welton, N. J., Ades, A., Collins, S., Orme, M., and Kelly, S. (2011). "Mixed treatment comparison of repeated measurements of a continuous endpoint: an example using topical treatments for primary open-angle glaucoma and ocular hypertension." *Statistics in medicine*, 30(20): 2511–2535.

de Boor, C. (2001). *A practical guide to splines*. Appied Mathematical Sciences. Springer-Verlag.

De Finetti, B. (1931). "Sul significato soggettivo della probabilita." *Fundamenta mathematicae*, 17(1): 298–329.

Di Lucca, M. A., Guglielmi, A., Müller, P., Quintana, F. A., et al. (2013). "A simple class of Bayesian nonparametric autoregression models." *Bayesian Analysis*, 8: 63–88.

Dickson, D., Hosie, G., and English, J. (2001). "A double-blind, placebo-controlled comparison of hylan GF 20 against diclofenac in knee osteoarthritis." *JOURNAL OF DRUG ASSESSMENT*, 4(3): 179–190.

Ding, Y. and Fu, H. (2013). "Bayesian indirect and mixed treatment comparisons across longitudinal time points." *Statistics in medicine*, 32(15): 2613–2628.

Dougados, M., Nguyen, M., Listrat, V., and Amor, B. (1993). "High molecular weight sodium hyaluronate (hyalectin) in osteoarthritis of the knee: a 1 year placebo-controlled trial." *Osteoarthritis and cartilage*, 1(2): 97–103.

Duchateau, L. and Janssen, P. (2007). *The frailty model*. Springer Science & Business Media.

Escobar, M. D. and West, M. (1995a). "Bayesian density estimation and inference using mixtures." *Journal of the american statistical association*, 90(430): 577–588.

— (1995b). "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association*, 90(430): 577–588.

Fenton, L. (1960). "The sum of log-normal probability distributions in scatter transmission systems." *IEEE Transactions on Communications*, 8(1): 57–67.

Ferguson, T. S. (1973a). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1: 209–230.

— (1973b). "A Bayesian Analysis of Some Nonparametric Problems." *The Annals of Statistics*, 1(2): 209–230.

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515–534.

Gelman, A. and Rubin, D. (1992). "Inference from iterative simulation using multiple sequences." *Statistical science*, 457–472.

Gelman, A. et al. (2008). "Objections to Bayesian statistics." *Bayesian Analysis*, 3(3): 445–449.

George, E. and McCullogh, R. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88: 881–889.

George, E. I. and McCulloch, R. E. (1997). "Approaches for Bayesian variable selection." *Statistica Sinica*, 7: 339–373.

González, J. R., Fernandez, E., Moreno, V., Ribes, J., Peris, M., Navarro, M., Cambray, M., and Borràs, J. M. (2005). "Sex differences in hospital readmission among colorectal cancer patients." *Journal of Epidemiology & Community Health*, 59(6): 506–511.

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82(4): 711–732.
URL https://doi.org/10.1093/biomet/82.4.711

Griffiths, T. L. and Ghahramani, Z. (2011). "The indian buffet process: An introduction and review." *Journal of Machine Learning Research*, 12(Apr): 1185–1224.

Halliwell, L. J. (2015). "The Lognormal Random Multivariate." *Casualty Actuarial Society E-Forum*, Spring 2015: 1–5.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). "The elements of statistical learning: data mining, inference and prediction." *The Mathematical Intelligencer*, 27(2): 83–85.

Henderson, E., Smith, E., Pegley, F., and Blake, D. (1994). "Intra-articular injections of 750 kD hyaluronan in the treatment of osteoarthritis: a randomised single centre double-blind placebo-controlled trial of 91 patients demonstrating lack of efficacy." *Annals of the rheumatic diseases*, 53(8): 529–534.

Higgins, J., Thompson, S. G., and Spiegelhalter, D. J. (2009). "A re-evaluation of random-effects meta-analysis." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1): 137–159.

Huang, C.-Y., Qin, J., and Wang, M.-C. (2010). "Semiparametric Analysis for Recurrent Event Data with Time-Dependent Covariates and Informative Censoring." *Biometrics*, 66(1): 39–49.
URL https://doi.org/10.1111/j.1541-0420.2009.01266.x

Huang, X. and Liu, L. (2007). "A Joint Frailty Model for Survival and Gap Times Between Recurrent Events." *Biometrics*, 63(2): 389–397.
URL https://doi.org/10.1111/j.1541-0420.2006.00719.x

Huskisson, E. and Donnelly, S. (1999). "Hyaluronic acid in the treatment of osteoarthritis of the knee." *Rheumatology*, 38(7): 602–607.

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96(453): 161–173.

Ishwaran, H. and Zarepour, M. (2002). "Exact and approximate sum representations for the Dirichlet process." *Can J Statistics*, 30: 269–283.

Jansen, J., Vieira, M., and Cope, S. (2015). "Network meta-analysis of longitudinal data using fractional polynomials." *Statistics in medicine*, 34(15): 2294–2311.

January, C. T., Wann, L. S., Alpert, J. S., Calkins, H., Cigarroa, J. E., Cleveland, J. C., Conti, J. B., Ellinor, P. T., Ezekowitz, M. D., Field, M. E., Murray, K. T., Sacco, R. L., Stevenson, W. G., Tchou, P. J., Tracy, C. M., and Yancy, C. W. (2014). "2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation." *Circulation*, 130(23).

Jordan, M. I. and Teh, Y. W. (2014). "A gentle introduction to the dirichlet process, the beta process, and bayesian nonparametrics." *Dept. Statistics, UC Berkeley*.

Jubb, R., Piva, S., Beinat, L., Dacre, J., and Gishen, P. (2002). "A one-year, randomised, placebo (saline) controlled clinical trial of 500-730 kDa sodium hyaluronate (Hyalgan) on the radiological change in osteoarthritis of the knee." *International journal of clinical practice*, 57(6): 467–474.

Karabis, A., Lindner, L., Mocarski, M., Huisman, E., and Greening, A. (2013). "Comparative efficacy of aclidinium versus glycopyrronium and tiotropium, as maintenance treatment of moderate to severe COPD patients: a systematic review and network meta-analysis." *International Journal of Chronic Obstructive Pulmonary Disease*, 8: 405–423.

Karlsson, J., Sjögren, L., and Lohmander, L. (2002). "Comparison of two hyaluronan drugs and placebo in patients with knee osteoarthritis. A controlled, randomized, double-blind, parallel-design multicentre study." *Rheumatology*, 41(11): 1240–1248.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). "Markov chain Monte Carlo in practice: a roundtable discussion." *The American Statistician*, 52(2): 93–100.

Khasriya, R., Khan, S., Lunawat, R., Bishara, S., Bignal, J., Malone-Lee, M., Ishii, H., O'Connor, D., Kelsey, M., and Malone-Lee, J. (2010). "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." *The Journal of Urology*, 183: 1843–1847.

Kim, S., Dahl, D. B., and Vannucci, M. (2009). "Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models." *Bayesian Analysis*, 4: 707–732.

Kingman, J. (1967). "Completely random measures." *Pacific Journal of Mathematics*, 21(1): 59–78.

Kirchner, M. and Marshall, D. (2006). "A double-blind randomized controlled trial comparing alternate forms of high molecular weight hyaluronan for the treatment of osteoarthritis of the knee." *Osteoarthritis and cartilage*, 14(2): 154–162.

Kolmogorov, A. (1933). "Foundations of the Theory of Probability, trans." *Nathan Morrison (1956). Chelsea, New York*.

Krege, S., Giani, G., Meyer, R., Otto, T., and Rubben, H. (1996). "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." *The Journal of urology*, 156(3): 962–966.

Kupelian, A. S., Horsley, H., Khasriya, R., Amussah, R. T., Badiani, R., Courtney, A. M., Chandhyoke, N. S., Riaz, U., Savlani, K., Moledina, M., Montes, S., O'Connor, D., Visavadia, R., Kelsey, M., Rohn, J. L., and Malone-Lee, J. (2013). "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 International*, 112: 231–238.

Lau, J. W. and Green, P. J. (2007). "Bayesian model-based clustering procedures." *Journal of Computational and Graphical Statistics*, 16(3): 526–558.

Lawless, J., Nadeau, C., and Cook, R. (1997). "Analysis of mean and rate functions for recurrent events." In *Proceedings of the First Seattle Symposium in Biostatistics*, 37–49. Springer.

Lawless, J. F. and Nadeau, C. (1995). "Some simple robust methods for the analysis of recurrent events." *Technometrics*, 37(2): 158–168.

Li, T., Puhan, M. A., Vedula, S. S., Singh, S., and Dickersin, K. (2011). "Network meta-analysis-highly attractive but more methodological research is needed." *BMC medicine*, 9(1): 79.

Li, Z., Chinchilli, V. M., and Wang, M. (2018). "A Bayesian joint model of recurrent events and a terminal event." *Biometrical Journal*, 61(1): 187–202.
URL https://doi.org/10.1002/bimj.201700326

— (2019). "A time-varying Bayesian joint hierarchical copula model for analysing recurrent events and a terminal event: An application to the Cardiovascular Health Study." *Journal of the Royal Statistical Society: Series*

C (Applied Statistics). Advance online publication.
URL https://doi.org/10.1111/rssc.12382

Lijoi, A., Prünster, I., et al. (2010). "Models beyond the Dirichlet process." *Bayesian nonparametrics*, 28(80): 342.

Lin, L.-A., Luo, S., and Davis, B. R. (2018). "Bayesian regression model for recurrent event data with event-varying covariate effects and event effect." *Journal of applied statistics*, 45(7): 1260–1276.

Lipsey, M. W. and Wilson, D. B. (2001). *Practical meta-analysis.*. Sage Publications, Inc.

Liu, L., Huang, X., Yaroshinsky, A., and Cormier, J. N. (2015). "Joint frailty models for zero-inflated recurrent events in the presence of a terminal event." *Biometrics*, 72(1): 204–214.
URL https://doi.org/10.1111/biom.12376

Liu, L., Wolfe, R. A., and Huang, X. (2004). "Shared Frailty Models for Recurrent Events and a Terminal Event." *Biometrics*, 60(3): 747–756.
URL https://doi.org/10.1111/j.0006-341x.2004.00225.x

Lu, G., Ades, A., Sutton, A., Cooper, N., Briggs, A., and Caldwell, D. (2007). "Meta-analysis of mixed treatment comparisons at multiple follow-up times." *Statistics in Medicine*, 26(20): 3681–3699.

Lumley, T. (2002). "Network meta-analysis for indirect treatment comparisons." *Statistics in medicine*, 21(16): 2313–2324.

Ma, L. and Sundaram, R. (2018). "Analysis of gap times based on panel count data with informative observation times and unknown start time." *Journal of the American Statistical Association*, 113(521): 294–305.

Ma, Z. and Krings, A. W. (2008). "Multivariate survival analysis (I): shared frailty approaches to reliability and dependence modeling." In *2008 IEEE Aerospace Conference*, 1–21. IEEE.

Mahmoud, H. (2008). *Pólya urn models*. Chapman and Hall/CRC.

Manda, S. O. and Meyer, R. (2005). "Bayesian inference for recurrent events data using time-dependent frailty." *Statistics in medicine*, 24(8): 1263–1274.

McGilchrist, C. and Aisbett, C. (1991). "Regression with frailty in survival analysis." *Biometrics*, 47(2): 461–466.

Meyer, R. and Romeo, J. S. (2015). "Bayesian semiparametric analysis of recurrent failure time data using copulas." *Biometrical Journal*, 57(6): 982–1001.

Munger, T. M., Wu, L.-Q., and Shen, W. K. (2014). "Atrial fibrillation." *Journal of Biomedical Research*, 28(1): 1–17.

Neal, R. M. (2000a). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of computational and graphical statistics*, 9(2): 249–265.

— (2000b). "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265.
URL https://doi.org/10.1080/10618600.2000.10474879

— (2003). "Slice sampling." *The Annals of Statistics*, 31(3): 705–767.

Nieto-Barajas, L. E. and Quintana, F. A. (2016). "A Bayesian Non-Parametric Dynamic AR Model for Multiple Time Series Analysis." *Journal of Time Series Analysis*, 37: 675–689.

Olesen, A. V. and Parner, E. T. (2006). "Correcting for selection using frailty models." *Statistics in Medicine*, 25(10): 1672–1684.
URL https://doi.org/10.1002/sim.2298

Ouyang, B., Sinha, D., Slate, E. H., and Bakel, A. B. V. (2013). "Bayesian analysis of recurrent event with dependent termination: An application to a heart transplant study." *Statistics in Medicine*, 32(15): 2629–2642.
URL https://doi.org/10.1002/sim.5717

Papadimitropoulou, K., Vossen, C., Karabis, A., Donatti, C., and Kubitz, N. (2017). "Comparative efficacy and tolerability of pharmacological and somatic interventions in adult patients with treatment-resistant depression: a systematic review and network meta-analysis." *Current medical research and opinion*, 33(4): 701–711.

Papaspiliopoulos, O. and Roberts, G. O. (2008). "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models." *Biometrika*, 95(1): 169–186.

Paulon, G., Iorio, M. D., Guglielmi, A., and Ieva, F. (2018). "Joint modeling of recurrent events and survival: A Bayesian non-parametric approach." *Biostatistics*.
URL https://doi.org/10.1093/biostatistics/kxy026

Pennell, M. L. and Dunson, D. B. (2006). "Bayesian semiparametric dynamic frailty models for multiple event time data." *Biometrics*, 62: 1044–1052.

Pepe, M. S. and Cai, J. (1993). "Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates." *Journal of the American statistical Association*, 88(423): 811–820.

Plummer, M. (2003). "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling."

— (2008). "Penalized loss functions for Bayesian model comparison." *Biostatistics*, 9(3): 523–539.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News*, 6: 7–11.

Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981). "On the regression analysis of multivariate failure time data." *Biometrika*, 68: 373–379.

Quintana, F. A. and Müller, P. (2012). "Nonparametric Bayesian assessment of the order of dependence for binary sequences." *Journal of Computational and Graphical Statistics*, 13: 213–231.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Rondeau, V., Commenges, D., and Joly, P. (2003). "Maximum penalized likelihood estimation in a gamma-frailty model." *Lifetime data analysis*, 9(2): 139–153.

Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H., Brouste, V., and Soubeyran, P. (2007). "Joint frailty models for recurring events and death using maximum penalized likelihood estimation: Application on cancer events." *Biostatistics*, 8(4): 708–721.
URL https://doi.org/10.1093/biostatistics/kxl043

Rondeau, V., Mazroui, Y., and Gonzalez, J. R. (2012a). "frailtypack: an R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation." *J Stat Softw*, 47(4): 1–28.

— (2012b). "frailtypack: An R package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation or Parametrical Estimation." *Journal of Statistical Software*, 47(4): 1–28.
URL https://doi.org/10.18637/jss.v047.i04

Royston, P. and Altman, D. G. (1994). "Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(3): 429–453.

Royston, P. and Sauerbrei, W. (2008). *Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*. John Wiley & Sons.

Scale, D., Wobig, M., and Wolpert, W. (1994). "Viscosupplementation of osteoarthritic knees with hylan: a treatment schedule study." *Current Therapeutic Research*, 55(3): 220–232.

Schroder, J., Bouaziz, O., Agner, B. R., Martinussen, T., Madsen, P. L., Li, D., and Dixen, U. (2019a). "Full anonymized dataset used for statistical analysis." *figshare, doi:10.1371/journal.pone.0217983.s001*.

— (2019b). "Recurrent event survival analysis predicts future risk of hospitalization in patients with paroxysmal and persistent atrial fibrillation." *PLOS ONE*, 14(6): e0217983.

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

Sinha, D., Maiti, T., Ibrahim, J. G., and Ouyang, B. (2008a). "Current Methods for Recurrent Events Data With Dependent Termination." *Journal of the American Statistical Association*, 103(482): 866–878.
URL https://doi.org/10.1198/016214508000000201

— (2008b). "Current methods for recurrent events data with dependent termination: a Bayesian perspective." *Journal of the American Statistical Association*, 103(482): 866–878.

Song, F., Loke, Y. K., Walsh, T., Glenny, A.-M., Eastwood, A. J., and Altman, D. G. (2009). "Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews." *BMJ*, 338: b1147.

Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4): 583–639.

Stitik, T., Blacksin, M., Stiskal, D., Kim, J., Foye, P., Schoenherr, L., Choi, E., Chen, B., Saunders, H., and Nadler, S. (2007). "Efficacy and safety of

hyaluronan treatment in combination therapy with home exercise for knee osteoarthritis pain." *Archives of physical medicine and rehabilitation*, 88(2): 135–141.

Strasak, A. M., Umlauf, N., Pfeiffer, R. M., and Lang, S. (2011). "Comparing penalized splines and fractional polynomials for flexible modelling of the effects of continuous predictor variables." *Computational statistics and data analysis*, 55(4): 1540–1551.

Thall, P. F. (1988). "Mixed Poisson likelihood regression models for longitudinal interval count data." *Biometrics*, 197–209.

Thrall, G., Lane, D., Carroll, D., and Lip, G. Y. (2006). "Quality of life in patients with atrial fibrillation: A systematic review." *The American Journal of Medicine*, 119(5): 448.e1–448.e19.

van Walsem, A., Pandhi, S., Nixon, R. M., Guyot, P., Karabis, A., and Moore, R. A. (2015). "Relative benefit-risk comparing diclofenac to other traditional non-steroidal anti-inflammatory drugs and cyclooxygenase-2 inhibitors in patients with osteoarthritis or rheumatoid arthritis: a network meta-analysis." *Arthritis research therapy*, 17(1): 66.

Vehtari, A. and Lampinen, J. (2002). "Bayesian model assessment and comparison using cross-validation predictive densities." *Neural computation*, 14(10): 2439–2468.

Waagepetersen, R. and Sorensen, D. (2001). "A Tutorial on Reversible Jump MCMC with a View toward Applications in QTL-mapping." *International Statistical Review*, 69(1): 49–61.
URL https://doi.org/10.1111/j.1751-5823.2001.tb00479.x

Wandel, S., Jüni, P., Tendal, B., Nüesch, E., Villiger, P. M., Welton, N. J., Reichenbach, S., and Trelle, S. (2010). "Effects of glucosamine, chondroitin, or placebo in patients with osteoarthritis of hip or knee: network meta-analysis." *Bmj*, 341: c4675.

Wang, M.-C., Qin, J., and Chiang, C.-T. (2001). "Analyzing recurrent event data with informative censoring." *Journal of the American Statistical Association*, 96(455): 1057–1065.

Welton, N. J., Sutton, A. J., Cooper, N. J., Abrams, K. R., and Ades, A. E. (2012). *Evidence Synthesis for Decision Making in Healthcare*. John Wiley & Sons, Ltd.

Wijffels, M. C., Kirchhof, C. J., Dorland, R., and Allessie, M. A. (1995). "Atrial fibrillation begets atrial fibrillation." *Circulation*, 92(7): 1954–1968.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Wobig, M., Bach, G., Beks, P., Dickhut, A., Runzheimer, J., Schwieger, G., Vetter, G., and Balazs, E. (1999). "The role of elastoviscosity in the efficacy of viscosupplementation for osteoarthritis of the knee: a comparison of hylan GF 20 and a lower-molecular-weight hyaluronan." *Clinical therapeutics*, 21(9): 1549–1562.

Wobig, M., Dickhut, R., Aand Maier, and Vetter, G. (1998). "Viscosupplementation with hylan GF 20: a 26-week controlled trial of efficacy and safety in the osteoarthritic knee." *Clinical therapeutics*, 20(3): 410–423.

Ye, Y., Kalbfleisch, J. D., and Schaubel, D. E. (2007). "Semiparametric Analysis of Correlated Recurrent and Terminal Events." *Biometrics*, 63(1): 78–87.
URL https://doi.org/10.1111/j.1541-0420.2006.00677.x

Yu, Z. and Liu, L. (2011). "A joint model of recurrent events and a terminal event with a nonparametric covariate function." *Statistics in Medicine*, 30(22): 2683–2695.
URL https://doi.org/10.1002/sim.4297

Yu, Z., Liu, L., Bravata, D. M., and Williams, L. S. (2013). "Joint model of recurrent events and a terminal event with time-varying coefficients." *Biometrical Journal*, 56(2): 183–197.
URL https://doi.org/10.1002/bimj.201200160