New approaches to survival extrapolation with inference using piecewise deterministic Monte Carlo

Luke Hardcastle

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Statistical Science

University College London

December 5, 2025

I, Luke Hardcastle, 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 work.

Abstract

This thesis is primarily concerned with developing novel survival models that are able to extrapolate hazards beyond final event times, and the development of novel, efficient posterior sampling methods based on non-reversible processes.

Polyhazard models are a class of flexible parametric models for modelling survival over extended time horizons. Significant user input is required, however, in selecting the number of latent hazards to model, their distributions and the choice of which variables to associate with each hazard. The resulting set of models is too large to explore manually, limiting their practical usefulness. To address this we extend the standard polyhazard model through a prior structure allowing for joint inference of parameters and structural quantities.

The piecewise exponential model utilises a piecewise constant hazard function. We develop a novel extension to this model to allow for principled extrapolations, based on a two part prior: i) A discretisation of an underlying diffusion process, allowing prior information to inform extrapolations. ii) A Poisson point process prior for the set of knots, allowing this set to be extrapolated beyond final event times. Posterior inference in both cases is achieved using Markov Chain Monte Carlo methods based on Piecewise Deterministic Markov Processes. These processes have seen significant theoretical interest due to non-reversible dynamics allowing for efficient exploration of the state space, and tractable continuous trajectories that allow for efficient sampling from transdimensional posteriors. This thesis provides a literature review for the current state of these processes and makes several contributions to improving their implementation, including extending methods for generating the underlying Poisson process and extending the range of transdimensional posteriors they can be applied to. With respect to the latter, we develop theory that allows these processes to navigate posteriors comprised of mixtures of a manifold and the ambient space the manifold is embedded in.

Impact Statement

This thesis provides contributions both to the modelling of time-to-event data in the context of Health Technology Assessment and to the application of piecewise deterministic Monte Carlo methods. This thesis will result in at least two papers published in academic journals. The content of Chapter 4 is based on a paper that is currently in press at the *Annals of Applied Statistics*. The content of Chapter 5 is currently undergoing major revisions for *Bayesian Analysis*. The work of Section 6.1 is an early version of joint work with researchers at the Institute of Statistical Mathematics in Tokyo, and will be submitted for journal publication. The work in this thesis has been presented as several conferences and workshops, and in seminars to both clinical and computational academic, and industry-focused audiences.

Within academia, this thesis contributes to the continued development of advanced Bayesian methods for survival analysis. In particular, the work of this thesis highlights the benefits of developing bespoke survival models for survival extrapolation. Perhaps surprisingly, Bayesian methods are relatively under-employed in survival analysis, despite natural benefits in the context of missing data. Chapter 4 is concerned with extending polyhazard models. In addition to the methodological work in that chapter, we provide guidelines of prior specification that should be beneficial to analysts in both academia and industry.

Chapter 5 highlights the importance and advantages of employing Bayesian approaches in this context, and the range of information that can be incorporated to guide extrapolation of hazard functions. To increase the application of these methods in HTA analyses I am actively developing packages in R and julia to allow for easy implementation of these models. Further, a crucial point of these models is the incorporation of prior information. I am planning on working with groups in academia and industry to develop principled guidelines to elicit and incorporate this prior information. Further work will also rely on training and development in organisations outside of academia. Application of these models will allow for faster and more principled assessment of novel medical technologies and interventions by the NHS and other publicly funded healthcare systems, ultimately improving outcomes for patients and reduced financial cost.

The inferential methods in this thesis are based on piecewise deterministic Monte Carlo methods. These methods have seen a large amount of theoretical interest, but minimal uses in applied statistical analyses. This thesis contributes directly to this field by developing strong case studies that advocate for the continued development and application of these methods. Further, insights from employing these methods will help inform future theoretical and methodological research, and provide insights to those looking to employ these methods in future work.

UCL Research Paper Declaration Form: Referencing the Doctoral Candidate's Own Published Work(s)

1. **For a research manuscript that has already been published** (if not yet published, skip to section 2):

   (a) **Title of the manuscript:** Averaging polyhazard models using Piecewise deterministic Monte Carlo with applications to data with long-term survivors

   (b) **Provide the DOI or direct link to the published work:** `https://imstat.org/journals-and-publications/annals-of-applied-statistics/annals-of-applied-statistics-next-issues/`

   (c) **Publication name (e.g., journal or textbook):** Annals of Applied Statistics

   (d) **Publisher name (e.g., Elsevier, Oxford University Press):** Institute of Mathematical Statistics

   (e) **Date of publication:** TBD

   (f) **List all authors as they appear in the publication:** Luke Hardcastle, Samuel Livingstone, Gianluca Baio

   (g) **Was the work peer-reviewed?** Yes

   (h) **Do you retain copyright for the work?** No

   (i) **Was an earlier version uploaded to a preprint server (e.g., medRxiv, arXiv)?** If **Yes**, provide the DOI or direct link. If not applicable, leave blank. `https://arxiv.org/abs/2406.14182`

   If **No**, please seek publisher permission and check the box below:

   □ *I acknowledge permission from the publisher named in item 1d to include in this thesis portions of the publication cited in item 1c.*

2. **2. For a manuscript prepared for publication but not yet published** (if already published, skip to section 3):

   (a) **Current title of the manuscript:**

(b) **Has it been uploaded to a preprint server (e.g., medRxiv, arXiv)?**

If **Yes**, provide the DOI or direct link. If not applicable, leave blank.

(c) **Intended publication outlet (e.g., journal name):**

(d) **List all authors in the intended authorship order:**

(e) **Current stage of publication (e.g., in submission, under review):**

3. **3. For multi-authored work, please provide a contribution statement detailing each author's role** (if single-authored, skip to section 4): LH developed the project, developed of the model, sampling architecture, wrote the code, ran the experiments and applied examples. LH wrote the initial draft, edited further drafts, and incoporated reviewer feedback. SL and GB supported the development of the project and provided advice through supervisory meetings. SL and GB provided feedback on initial drafts on the work, and supported LH in incorporating reviewer feedback.

4. **4. In which chapter(s) of your thesis can this material be found?** Chapter 4

e-Signatures confirming accuracy of the above information

(This form should be co-signed by the supervisor/senior author unless the work is single-authored):

Candidate signature: Luke Hardcastle

Date: 05/09/2025

Supervisor/Senior Author signature (where appropriate): Samuel Livingstone, Gianluca Baio

Date: 05/09/2025

UCL Research Paper Declaration Form: Referencing the Doctoral Candidate's Own Published Work(s)

1. **For a research manuscript that has already been published** (if not yet published, skip to section 2):

   (a) **Title of the manuscript:**

   (b) **Provide the DOI or direct link to the published work:**

   (c) **Publication name (e.g., journal or textbook):**

   (d) **Publisher name (e.g., Elsevier, Oxford University Press):**

   (e) **Date of publication:**

   (f) **List all authors as they appear in the publication:** Luke Hardcastle, Samuel Livingstone, Gianluca Baio

   (g) **Was the work peer-reviewed?** Yes

   (h) **Do you retain copyright for the work?** No

   (i) **Was an earlier version uploaded to a preprint server (e.g., medRxiv, arXiv)?** If **Yes**, provide the DOI or direct link. If not applicable, leave blank. `https://arxiv.org/abs/2406.14182`

   If **No**, please seek publisher permission and check the box below:

   □ *I acknowledge permission from the publisher named in item 1d to include in this thesis portions of the publication cited in item 1c.*

2. **2. For a manuscript prepared for publication but not yet published** (if already published, skip to section 3):

   (a) **Current title of the manuscript:** Diffusion piecewise exponential models for survival extrapolation using Piecewise Deterministic Monte Carlo

   (b) **Has it been uploaded to a preprint server (e.g., medRxiv, arXiv)?** `https://arxiv.org/abs/2505.05932`

   (c) **Intended publication outlet (e.g., journal name):** Bayesian Analysis

(d) **List all authors in the intended authorship order:** Luke Hardcastle, Samuel Livingstone, Gianluca Baio

(e) **Current stage of publication (e.g., in submission, under review):** Initial decision received: Major revisions required.

3. **3. For multi-authored work, please provide a contribution statement detailing each author's role** (if single-authored, skip to section 4): LH developed the project, developed of the model, sampling architecture, wrote the code, ran the experiments and applied examples. LH wrote the initial draft and edited further drafts. SL and GB supported the development of the project and provided advice through supervisory meetings. SL and GB provided feedback on initial and final drafts on the work.

4. **4. In which chapter(s) of your thesis can this material be found?** Chapter 5

e-Signatures confirming accuracy of the above information

(This form should be co-signed by the supervisor/senior author unless the work is single-authored):

Candidate signature: Luke Hardcastle

Date: 05/09/2025

Supervisor/Senior Author signature (where appropriate): Samuel Livingstone, Gianluca Baio

Date: 05/09/2025

Acknowledgements

I would like to begin these acknowledgements by thanking my doctoral supervisors Sam Livingstone and Gianluca Baio. I feel deeply privileged to have had the opportunity to work with you over the past four years. Thank you for your advice, patience, and the opportunities you have offered me that have helped me grow both as a researcher and as a person.

I am grateful to everyone in the Department of Statistical Science at UCL for making my time as a PhD student as enjoyable as it has been. I would particularly like to thank everyone in the PhD room for your friendship and support. To be in the company of so many individuals who I know will go on to do great things has been incredibly inspiring.

I would like to thank Jim Griffin and Javier Rubio for the sitting on my upgrade panel and for the feedback they provided. I would also like to thank Terry Soo and Marina Lewis for their support in the roles of Graduate Tutor and Teaching and Learning Administrator for PhD students. Additionally, I am immensely grateful to all those I have had the chance to speak to and learn from at the workshops and conferences I have attended during the course of my studies.

I have been very lucky to have been able to take on several additional opportunities alongside my PhD. I would like to thank Parexel for their support during my internship. I would like to thank the Alan Turing Institute for supporting me through their enrichment scheme. I would like to thank the Isaac Newton Institute and particularly Codina Cotar for supporting me in a visit as part of the Stochastic systems for anomalous diffusion programme. Finally, I would like to thank Kengo Kamatani for sponsoring a visit to the Institute of Statistical Mathematics in Tokyo

as part of the Institute's international internship scheme, and Hirofumi Shiba for making my visit so enjoyable.

I would like to thank Alex Beskos and Chris Jackson for examining this thesis and for the enjoyable discussion we had during the viva.

Thank you to EPSRC (grant number EP/W523835/1) for funding this PhD.

I am grateful to my friends, in particular Tom, Theo, Robbie and Sam for being there for me throughout this journey and putting up with my unsolicited attempts to explain my research to them.

I am grateful to my parents, and my sisters, Saskia and Katya, for their constant support, especially during the final months when I have needed it most.

Finally, none of this would have been possible without my fiancée, Lauren, who gave me the strength and belief to continue when I thought I could not. Without you I could not have persevered over the past four years. This thesis is dedicated to you.

Contents

List of Figures

List of Tables

Chapter 1

Thesis overview

This thesis is primarily concerned with developing novel survival models that are able to extrapolate hazards beyond final event times, and the development of novel, efficient posterior sampling methods based on non-reversible Markov processes.

1.1 Survival extrapolation

Chapter 2 provides an introduction to survival analysis and a review of current methods for extrapolating hazard functions beyond final observation times. This problem is particularly relevant in Health Technology Assessment [HTA ; Latimer, 2013]. Healthcare systems, such as the UK's National Health Service, have the objective of maximising the health of the population given finite financial resources. HTA provides a decision-theoretic framework for analysing the cost-effectiveness of novel medical interventions in publicly funded healthcare systems, to ensure this objective is fulfilled.

In England, following the recommendations of the National Institute for Health and Care Excellence (NICE), expected life years (i.e mean survival) is commonly incorporated as the primary measure of benefit in these analyses. Estimation of this quantity requires survival, or equivalently hazard, curves to be estimated over a lifetime time horizon. This task is not trivial as data from both clinical trials and observational studies is often limited in follow up. Analysts are therefore tasked with inferring hazard curves on an extended interval given data from the initial period. There is therefore a focus on the development of methods that can extrapolate beyond final event times in a principled manner.

Chapter 4 is concerned with the extension of polyhazard models, a commonly used model for survival extrapolation [Berger and Sun, 1993, Demiris et al., 2015]. This model utilises an additive hazard formulation to extrapolate hazards beyond event times, informed by final observations. In practice, use of this model has been limited, however, by a challenging model selection problem that requires the analyst to specify the number of hazards, covariates associated with each hazard and the functional form of each hazard. The focus of Chapter 4 is the extension of the polyhazard model to infer these quantities within a Bayesian framework, with priors on each quantity. This allows analysts to fit a single model to the data, rather than having to fit a large set of candidate models that grows rapidly with the number of considered sub-hazards and covariates. The model is showcased on data arising from stroke survivors and kidney transplant patients.

Chapter 5 introduces a new prior structure for the piecewise exponential model as the discretisation of a latent diffusion process. In the context of survival extrapolation this allows for flexible, data-driven inference during the time period of the trial. Extrapolations are then informed by a pre-specified diffusion that encodes explicit prior beliefs about the long-term behaviour of the hazard. We outline extensions that incorporate non-proportional covariate effects, time-varying drifts and waning treatment effects. The model is showcased on data from colon cancer and Leukaemia patients.

Chapter 6 discusses future directions, focusing on the practical implementation of these models and the specification of alternative latent processes to those considered in Chapter 5.

1.2 Piecewise deterministic Monte Carlo

The primary inferential tool employed in this work is Markov Chain Monte Carlo (MCMC) methods based on Piecewise Deterministic Markov Processes [PDMPs ; Davis, 1993, Fearnhead et al., 2018]. MCMC methods, where samples of the posterior are generated by designing a Markov process with the posterior as its stationary distribution, are a well established tool for applied statisticians. Most popular MCMC methods rely on a reversibility condition for validity. This condition is typically easy to satisfy, but can introduce diffusive dynamics into sampling, limiting computational efficiency.

This has motivated the development of samplers that are non-reversible, replacing the diffusive dynamics of reversible samplers with ballistic exploration of the state space [Diaconis et al., 2000, Andrieu and Livingstone, 2021]. A class of processes that exhibit this behaviour are Piecewise Deterministic Markov Processes, where velocities drive piecewise deterministic exploration of the state space. In the context of Bayesian inference, there has been a large body of theoretical work studying these processes recently, however, there have been limited practical implementations. A key contribution of this thesis is the practical application of these methods.

Chapter 3 provides a review of Markov Chain Monte Carlo focused on Piecewise Deterministic Markov Processes. In particular we highlight the challenges associated with implementing these processes. Further we highlight an attractive property of these processes that, when they have tractable deterministic dynamics, they are able to move directly between nested models commonly found in Bayesian model averaging problems.

Chapter 4 applies these processes to the extended polyhazard model. We develop results that allow for incorporation of transdimensional birth-death processes alongside PDMPs and extend existing methods for implementing these processes.

Chapter 5 extends the transdimensional aspects of these processes to sample from the posterior of the piecewise exponential model when a Poisson point process prior is used for the number and location of knots in the sampler.

Chapter 6 introduces methodology that allows these processes to stick to an embedded manifold.

Chapter 2

Survival analysis for Health Technology Assessment

Understanding the benefit of medical interventions in terms of the amount of “life” gained is a foundational statistical problem. To the best of the author’s knowledge, the earliest attempt at the study of this problem was undertaken by Daniel Bernoulli [Bernoulli, 1766, Bernoulli and Blower, 2004], who advocated for the introduction of smallpox inoculation by developing a mathematical model to understand the number of life years gained given the eradication of the disease.

More broadly, time to event data (such as survival times given smallpox inoculation) are ubiquitous in many fields, perhaps most prominently in medical research but also engineering and reliability, insurance and financial risk modelling, and both social and environmental sciences. Following seminal contributions in the second half of the 20^{th} Century [Cox, 1972, Feigl and Zelen, 1965, Kaplan and Meier, 1958] the study of these data, commonly referred to as survival analysis, has become an established field of applied statistics [Ibrahim et al., 2001, Legrand, 2021].

Perhaps the defining feature of survival data is the presence of censoring, where a subset of observations are only partially observed. This commonly occurs in clinical trials and observational studies where individuals may not have experienced the event of interest by the end of the study.

This chapter provides an introduction to survival analysis primarily in the context of Health Technology Assessment [HTA; Latimer, 2011, Baio, 2013], where

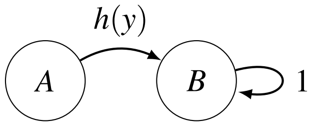

Figure 2.1: Visualisation of the data-generating process for survival models considered in this thesis, where A is the initial state (e.g being alive), B is the final absorbing state (e.g death), and $h(y)$ is the time-dependent hazard function, or equivalently the inhomogeneous rate of transitions from A to B.

the costs and benefits of novel medical interventions are analysed. Survival analysis is used to quantify the benefits within this framework primarily in terms of (quality-adjusted) life years gained, requiring estimation of expected survival. This contrasts with standard measures used in more traditional survival analysis, where measures of interest typically include median survival or hazard ratios.

2.1 Survival analysis

The standard data generating process for many survival models assumes that observations arise from a simple two-state continuous-time Markov chain, depicted in Figure 2.1, with an initial state, A, and a single final absorbing state, B. Transitions are determined by a time-dependent hazard function, $h(y)$, $y \in (0,\infty)$. Here, A and B can correspond to, for example, being alive and being dead or cancer having not progressed and cancer having progressed. The standard objective of survival analysis is the modelling of the time, Y, spent in state A and associated quantities.

Figure 2.1 encodes several assumptions associated with classical survival analysis. Primarily that events (i.e transitions) can only occur once and that there is a single transition of interest. Further, with the additional assumption that $\int_0^\infty h(y)dy = \infty$, that events always occur in finite time. Note that relaxing any of the above assumptions leads to several active areas of research in modern survival analysis, including the modelling of repeated events [Amorim and Cai, 2015], multi-state models [Jackson, 2011] and cure models [Amico and Van Keilegom, 2018].

Given the two-state Markov chain formulation above, we can define the probability density function of $Y \in \mathbb{R}_{>0}$ as $f(y)$, with corresponding cumulative density function $F(y) = \mathbb{P}(Y < y)$. More commonly, however, Y is analysed through the hazard and survival functions

$$h(y) := \lim_{\varepsilon \to 0} \frac{\mathbb{P}(Y \leq y + \varepsilon \mid Y > y)}{\varepsilon}, \quad S(y) := 1 - F(y). \tag{2.1}$$

Here, $h(y)$ can be interpreted as the instantaneous risk experienced by an individual, and $S(y)$ is simply the probability the transition has not occurred at time y. These quantities are directly linked to the probability density function and to each other as

$$f(y) = h(y)S(y), \quad S(y) = \exp\left(- \int_0^y h(u)du \right).$$

In short, specification of either $S(y)$ or $h(y)$ is sufficient to specify the entire data-generating process for Y. As such, hazard selection, i.e the process of deciding the form of $h(y)$, is equivalent to standard model selection. Finally, given a sequence of n independent observations for Y, $\{y_i\}_{i=1}^n$, and assuming that the hazard and survival functions are parametrised by a vector of parameters, θ, this allows us to specify a likelihood

$$\mathcal{L}(\theta; \{y_i\}_{i=1}^n) = \prod_{i=1}^n h_\theta(y_i) S_\theta(y_i).$$

2.1.1 Censoring

Throughout this thesis we will assume that data are partially right-censored due to, for example, random patient drop out or the end of clinical trials. This is accounted for in the data-generating process by observations arising according to

$$Y^O = \min\{Y, Y^C\},$$

where Y^O is the observed time and Y^C is the censoring time with corresponding probability density function $g(y)$ and cumulative density function $G(y)$. Specifically, this encodes the definition of right-censoring, that for all censored times $Y > Y^O$. In this work any references to censoring are referring to right-censoring; however, more generally, individuals may also be subject to left- or interval-censoring. Throughout we will assume that we know which event times are censored, i.e we observe the censoring indicator,

$$\Delta = \mathbb{1}(Y < Y^C),$$

realised as $\{\delta_i\}_{i=1}^n$. The data are therefore comprised of a tuple of event and censoring times, censoring indicators and, in some cases, covariates for individuals, $w \in \mathbb{R}^p$, summarised as $\mathcal{D} = \{y_i, \delta_i, w_i\}_{i=1}^n$. The resulting likelihood can then be written as

$$\begin{aligned}\mathcal{L}(\theta; \mathcal{D}) &= \prod_{i=1}^{n} [(1 - G(y_i)) f_\theta(y_i \mid w_i)]^{\delta_i} [S_\theta(y_i \mid w_i) g(y_i)]^{1-\delta_i}, \\ &= \prod_{i=1}^{n} h_\theta(y_i \mid w_i)^{\delta_i} S_\theta(y_i \mid w_i). \qquad (2.2)\end{aligned}$$

To move from the first to the second line we have crucially made the assumption that the censoring mechanism does not depend on θ, the parameters of the survival distribution [e.g., Legrand, 2021]. This is referred to as non-informative censoring and is a common assumption underpinning many survival models. In many of the examples of this thesis, survival times will be deterministically truncated at some time as a result of, e.g the end point of a clinical trial, in addition to the random censoring mechanism. In these cases we will refer to these observations being administratively censored, and denote this censoring time as y_+.

2.1.2 Survival extrapolation

Health Technology Assessment often requires the inference of expected survival over a lifetime time-horizon [Latimer, 2011, Baio, 2013], $(0, y_\infty)$ as a measure of the benefit received under a certain treatment,

$$\mathbb{E}[Y] = \int_0^{y_\infty} (1 - F(y)) \mathrm{d}y = \int_0^{y_\infty} S(y) \mathrm{d}y. \qquad (2.3)$$

A typical feature of data in this setting is that they are subject to a high degree of administrative censoring, often with $y_+ << y_\infty$. For example, Gibbons and Latimer [2024] estimate that since 2018 56% of NICE appraisals for cancer treatments have been conducted using immature survival data, where the majority of events occur after y_+. Re-writing (2.3) illustrates the difficulties this censoring can create,

$$\mathbb{E}[Y] = \int_0^{y_+} S(y)\mathrm{d}y + \int_{y_+}^{y_\infty} S(y)\mathrm{d}y, \tag{2.4}$$

as we now require a specification for $S(y)$ that can extrapolate beyond y_+ [Latimer, 2011]. In particular, as the rate of censoring increases, inferences of $\mathbb{E}[Y]$ will become increasingly sensitive to this extrapolation. This extrapolation excludes standard non-parametric approaches such as Kaplan-Meier estimators. Further, the problem cannot be circumvented by simply using an alternative estimand as a proxy for (2.4). For example, in the presence of high censoring rates median survival will often not be observed, and restricted mean survival estimates on the interval $(0, y_+)$ will be markedly different to those on $(0, y_\infty)$ as they ignore the non-negligible contribution from the second half of (2.4). We review current approaches to this problem in Section 2.2.

2.1.3 Bayesian survival analysis

This thesis is primarily concerned with Bayesian approaches to survival analysis. Within this paradigm, uncertainty about model parameters is represented through probability distributions. This is achieved by combining the survival likelihood (2.2) with a prior distribution, $\pi_0(\theta)$ for θ to derive a posterior distribution

$$\pi(\theta \mid \mathcal{D}) \propto \mathcal{L}(\theta; \mathcal{D})\pi_0(\theta).$$

This distribution is then used to make inferences about quantities of interest, in most cases considered in this thesis, $\mathbb{E}_\theta[Y]$. This perspective is particularly appealing in the context of survival extrapolation, as it allows for the principled incorporation of prior information into inferences either in the form of external data or expert opinion. We expand on this point in Section 2.2.3. Note, throughout this thesis we will occasionally drop the dependence on $\mathcal{D}$ in the posterior, denoting instead as $\pi(\theta)$.

2.2 Current approaches to survival extrapolation

For the remainder of this chapter we review current approaches to survival extrapolation in Health Technology Assessment.

2.2.1 Parametric models

Perhaps the most common approach to extrapolation is to assume Y is generated by a (typically two-parameter) parametric survival distribution [Latimer, 2013]. For example $Y \sim \text{Weibull}(\lambda, \gamma)$ gives the hazard function $h(y) = \lambda \gamma y^{\gamma - 1}$. The parametric form of the hazard function then allows extrapolation beyond y_+. This approach was reviewed in an influential NICE decision support unit report [Latimer, 2011], who recommend in the presence of incomplete survival data to compare several parametric survival models for the data based on information criteria and plausibility of extrapolations. A table of commonly used distributions is found in Table 1 of Baio [2020]. This approach is showcased in Appendix C.

This approach has several limitations:

1. Standard parametric models are only able to model a limited range of hazard behaviours. For example, none of the distributions in Table 1 of Baio [2020] can model a hazard with multiple points of inflection.

2. Extrapolations are based on the assumption the parametric model is correctly specified. This assumption is untestable when a large proportion of events occur in (y_+, y_∞).

3. As a direct consequence of the previous point, credible intervals for the hazard function on (y_+, y_∞) and therefore also the second half of (2.4) will shrink as more events are observed in $(0, y_+)$, despite no data being observed during the extrapolation period. If the model is misspecified, this will then underestimate uncertainty in the extrapolation period.

4. Further, any information incorporated into $\pi_0(\theta)$ [e.g Soikkeli et al., 2019, Palmer et al., 2023] to inform extrapolations will be overridden by the data as the number of events in $(0, y_+)$ increases.

The following sections review current methods for overcoming these limitations. The primary method for overcoming points 1 & 2 is to utilise a flexible parametric model for $(0, y_+)$. We review several approaches in this vein primarily following the review of Rutherford et al. [2020]. Flexible models can also prevent underestimation of uncertainty in the extrapolation period (point 3), but flexible assumptions do not automatically guarantee uncertainty is correctly calibrated. Points 3 & 4 are often overcome using external data and assumptions about the long-term form of the hazard, we review how these are often incorporated into standard modelling techniques in Section 2.2.3.

2.2.2 Flexible models

There are several choices of flexible parametric model available in the literature for survival extrapolation, allowing for data-driven inferences during the observation period. In some cases these models do not contain an automatic method for extrapolation, however, requiring the analyst to specify the extrapolation mechanism.

2.2.2.1 Polyhazard models

Polyhazard models are a class of flexible parametric models defined by additively combining hazards from one- or two-parameter survival distributions,

$$h(y) = \sum_{j=1}^{K} h_j(y).$$

This procedure results in hazard functions that are flexible and able to model a wide range of covariate effects. Originally developed for analysis of latent competing risks [Berger and Sun, 1993, Louzada-Neto, 1999], polyhazard models have become increasingly popular for modelling long-term survival required for Health Technology Assessment following the work of Demiris et al. [2015], who used a poly-Weibull model to analyse survival in transplant patients. Note they have also been used in the HTA literature to model data generating processes where the latent competing risks are given explicitly [Benaglia et al., 2015]. Polyhazard models can capture a wide range of hazard curves while retaining the interpretability and parsimony of simpler models. Further, due to the additive decomposition of the hazard function, later observations naturally have more influence on long-term survival. Recently Apsemidis and Demiris [2024] have also developed a piecewise version of the polyhazard model based on a single changepoint.

Polyhazard models are the primary focus of Chapter 4.

2.2.2.2 Piecewise models

A common approach for introducing flexibility into the hazard function is to define the hazard as a piecewise continuous function [Ibrahim et al., 2001, Feigl and Zelen, 1965, Fearnhead and Liu, 2011],

$$h(y) = \sum_{j=1}^{J} h_j(y)\mathbb{1}(y \in [s_{j-1}, s_j)).$$

where $\{s_j\}_{j=1}^J$ is a set of knots. The most common specification for $h_j(y)$ in this context is a constant, defining a piecewise exponential model. This could also be taken to be a linear or log-linear function, or taken as the hazard function from an existing parametric model, however we will focus on the piecewise constant case. These models naturally provide a flexible fit to the hazard assuming $\{s_j\}_{j=1}^J$ is sensibly chosen. Extrapolations are highly sensitive to model specification, however, with the analyst required to specify both how the the local hazard $h_j(y)$ evolves in time, and the location of knots in the extrapolation period.

The simplest way to define this extrapolation is through a random walk prior on the local (log-)hazards, and then manually placing knots in the extrapolation period [Che et al., 2023, Kearns et al., 2021]. The uncertainty associated with the resulting hazards, however, will increase indefinitely with the number of knots, often beyond a range of plausible long term hazard values. Further, this rate of increase is highly sensitive to the location of knots, increasing faster when more knots are specified and slower when fewer are specified.

An alternative perspective is provided by Cooney and White [2023a]. The piecewise exponential model is specified with independent gamma priors for the local constant hazards, and a Poisson prior for the number of knots. This prior allows for the location of s_J to be determined by the data. Hazards are then extrapolated as a constant after this point. This approach, advocated by Bagust and Beale [2014], has proved controversial due to the reliance on the model selected for the observation period, and the inability to inform long-term hazards with prior information [Latimer, 2014]. Further the independent priors for the local hazards reduce the flexibility of the hazard inferred during the observation period. This approach has also been extended to incorporate waning treatment effects [Cooney and White, 2023b].

Recently, Kearns et al. [2019, 2022] used piecewise models for extrapolation within the framework of dynamic survival models. These models incorporate a trend into the evolution of the hazard function, which can then be used to inform long-term extrapolations. These models retain the flexibility of standard piecewise models, however, for extrapolation they are reliant on the assumption that the long-term trend is correctly specified and can be inferred from the limited data in the observation period.

2.2.2.3 Spline models

A final class of approaches we consider model the hazard function using splines. In similar fashion to piecewise models, these approaches model the hazard function with a set of basis functions separated by knots. Similar considerations apply with their use, in that they are sensitive to placement of knots, in both the observation and extrapolation periods.

Spline-based extrapolation methods were introduced in Guyot et al. [2017], where Bayesian multi-parameter evidence synthesis was used to combine a restricted cubic spline model for the observed data with external information to guide long-term extrapolations. Restricted cubic splines can provide poor estimates in the context of hazard modelling, however, as they allow for negative hazards.

This approach was improved by Jackson [2023] who used M-splines as a model for $h(y)$. In contrast to restricted cubic splines, these guarantee that the hazard is positive. Extrapolation, without external data, is then based on placing a final knot in (y_+, y_∞). In the absence of additional information these methods will be sensitive to the placement of this knot. Further studies have validated the performance of this approach to fit data during the observation period [Timmins et al., 2025a], and the quality of extrapolations when external data is incorporated [Timmins et al., 2025b].

2.2.3 Incorporation of external information

When the observation period is short relative to the overall time-period of interest, $y_+ << y_\infty$, extrapolations can often be improved by the incorporation of additional information into survival models. In the Bayesian setting this information may be specified as a prior derived, for example, from expert opinion on the plausibility of long-term survival probabilities. Alternatively, it may be available as an external dataset from a population with some shared characteristics of the study population.

2.2.3.1 External data

Examples of external datasets include life-tables for the national population level, disease registries and previously conducted clinical trials [Bullement et al., 2024, Jackson et al., 2017]. Importantly, the similarity of the study and external population will inform how the external data are incorporated into the model.

These assumptions are reviewed in Jackson et al. [2017], where the hazard for the population of interest are related through one the following assumptions

$$h_S(y) = h_E(y), \quad y > y_*, \tag{2.5}$$

$$h_S(y) = \exp(\beta) h_E(y), \tag{2.6}$$

$$h_S(y) = h_E(y) + \gamma. \tag{2.7}$$

Here, (2.5) encodes a converging hazard assumption such that the study and external population hazards converge after some pre-specified time y_*; (2.6) encodes a long-term proportional hazard assumption between the study and external data populations; and (2.7) encodes a long-term additive difference between the external and study population hazards. Importantly each of these hazard forms are based on assumptions made by the analyst. These may be supported by data from the observation period, but an inherent feature of the survival extrapolation is that they are not able to be conclusively tested.

Principled incorporation of these assumptions with external data is undertaken through Bayesian multi-parameter evidence synthesis whereby external data, for example encoded through (2.5)-(2.7), is explicitly included in the likelihood function. The two spline-based approaches highlighted in Guyot et al. [2017], Jackson et al. [2017] are able to incorporate external data through these assumptions, and formally compared in Bullement et al. [2024]. Both approaches allow for the incorporation of data from both disease registries and background mortality.

In general, these assumptions are often incorporated into the analysis after the point of model fitting. This involves first fitting a model for $h_S(y)$, and then basing the analysis on, for example, the hazards $h(y) = h_S(y) + h_E(y)$, or $h(y) = \max\{h_S(y), h_E(y)\}$ [Andersson et al., 2013]. Note, there is no principled basis for either of these approaches. This is highlighted in van Oostrum et al. [2021], where the authors compare these approaches to methods incorporating external data directly into the likelihood and found superior performance in the latter.

A final method for incorporating external data is the blended survival approach [Che et al., 2023]. Two survival curves are inferred; The first, $S_O(y)$, for the observation period and based on study data, from a flexible, possibly non-parametric, survival model; The second, $S_E(y)$, for the extrapolation period is derived from external data, that represents the expected long-term survival of the study population. The two curves are then blended together as

$$S(y) = S_O(y)^{w(y)} \times S_E(y)^{(1-w(y))},$$
$$w(y) = F_B\left(\frac{y-a}{b-a}\right),$$

where F_B is the cumulative density function of a Beta(α, β) distribution. Here, a, b correspond to the limits of the interval over which the blending occurs, with α and β set to control the rate of blending between the two survival curves. Note, while we have introduced this method as an approach for incorporating external data, the authors also outline how $S_E(y)$ can be elicited via expert opinion.

2.2.3.2 Prior information

An alternative to direct incorporation of external data, is the use of a prior distribution to inform extrapolations. Derivation of this prior may be based on historical trial data. For example, Soikkeli et al. [2019], use historical data to derive a prior for the shape parameter of a Weibull distribution. Note, when standard parametric models are used for survival extrapolation, the influence of this prior information will decay as the number of events in the observation period increases, limiting the ability of this approach to influence extrapolations.

Cooney and White [2023c] propose an alternative method that incorporates expert opinion on the value of the survival function at a fixed time point, $y_* > y_+$, by eliciting a loss function for $S(y_*)$. This is then incorporated into the posterior by multiplicatively including the term

$$\pi(\theta \mid \mu_*, \sigma_*^2) \propto \exp\left(-\frac{1}{2\sigma_*^2}(S_\theta(y_*) - \mu_*)^2\right),$$

in addition to a standard prior for θ. Here, (μ_*, σ_*^2) are the expert's expected value of $S_\theta(y_*)$ and quantification of the corresponding uncertainty associated with this term. Theoretically the authors justify this approach within a generalised Bayes framework [Bissiri et al., 2016, Section 4.1]. This is a simpler approach than eliciting a prior distribution directly, as it does not require the elicited distribution to be transformed into an explicit density for θ. It does, however, suffer from the same drawbacks as when $S_\theta(y)$ is modelled with a parametric model, with data in the observation period dominating long-term inferences.

The above examples do not outline how prior information should be elicited. Prior elicitation is its own field within Bayesian statistics, with several general proposed frameworks. Recently, the Sheffield Elicitation Framework (SHELF) has been applied directly to the case of survival extrapolation for HTA [Gosling, 2017, Oakley et al., 2025] using standard parametric models [Cope et al., 2019] and in M-spline models [Jackson, 2023]. This provides a structured framework to elicit expert beliefs about long-term survival probabilities. These prior beliefs are then converted into synthetic datasets that can be incorporated into the analysis, as outlined earlier in this section.

Chapter 3

Piecewise Deterministic Monte Carlo

This chapter introduces and reviews recent advances in Markov Chain Monte Carlo methods [MCMC ; Brooks et al., 2011] primarily based on Piecewise Deterministic Markov Processes [PDMPs ; Davis, 1993]. This is an area of active research, and as such several of the works cited in this chapter have been published during the development of this thesis. The chapter begins with a review of the computational challenges presented by Bayesian inference before reviewing standard Markov Chain Monte Carlo approaches. Most state of the art approaches are based on a reversibility condition that introduces diffusivity into the dynamics of the sampler. This has motivated recent work on non-reversible processes that break this diffusivity by introducing velocities that can drive exploration of the state space. We review several of the recommended processes focusing in particular on their generation and transdimensional sampling.

3.1 Bayesian computation

Bayesian inference generates the posterior distribution, $\pi(\theta),\ \theta \in \Omega \subseteq \mathbb{R}^d$, resulting from the combination of the prior and likelihood. This distribution is then used to generate the quantities of interest to the analyst. For example, these may be marginal quantities of interest expressed as expectations of functions of θ,

$$\mathbb{E}_\pi[f(\theta)] = \int_\Omega f(\theta)\pi(\theta)\mathrm{d}\theta, \tag{3.1}$$

posterior predictive distributions

$$p(y_{n+1} \mid \mathcal{D}) = \int_{\Omega} p(y_{n+1} \mid \theta)\pi(\theta \mid \mathcal{D})\mathrm{d}\theta,$$

or marginal likelihoods as a measure of model evidence

$$p(y_{1:n}) = \int_{\Omega} \tilde{\pi}(\theta)\mathrm{d}\theta,$$

where $\tilde{\pi}(\theta) \propto \pi(\theta)$ is an unnormalised version of $\pi(\theta)$.

Outside of restricted cases, these integrals cannot be computed analytically due to the form of the posterior, or the high dimension of θ. One of the primary challenges associated with Bayesian inference, therefore, is the development of computational methods that can accurately approximate these quantities. Arguably the most popular and flexible of these methods is the Markov Chain Monte Carlo method [Brooks et al., 2011, Martin et al., 2024].

3.2 Markov Chain Monte Carlo

The results stated in this section, unless stated otherwise, can be found in [Brooks et al., 2011, Chapter 1]. Monte Carlo methods, in the context of Bayesian computation, use samples from the posterior distribution to approximate expectations (3.1) via the average

$$\mathbb{E}_{\pi}[f(\theta)] \approx N^{-1} \sum_{i=1}^{N} f(x_i), \quad x_1, \ldots, x_N \sim \pi(\cdot). \tag{3.2}$$

Practically this replaces the, now trivial, task of evaluating (3.1) with the task of generating suitably accurate samples from $\pi(\theta)$. Convergence of these estimates is then ensured by the Law of Large Numbers

$$N^{-1} \sum_{i=1}^{N} f(x_i) \to \mathbb{E}_{\pi}[f(\theta)], \quad N \to \infty.$$

This allows the approximation error to be reduced to an arbitrary degree by increasing the number of samples, although this result does not guarantee that the number of required samples can be generated in finite time. For guarantees of this form, we require the existence of central limit theorem

$$\sqrt{N}(N^{-1}\sum_{i=1}^{N} f(x_i) - \mu) \xrightarrow{d} \text{Normal}\left(0, \sigma^2\right), \tag{3.3}$$

where $\mu = \mathbb{E}_\pi[f(\theta)]$ and $\sigma^2 = \text{var}(f(\theta))$, and $\xrightarrow{d}$ denotes convergence in distribution. For independent sampling this holds if $\sigma^2 < \infty$.

Independent posterior sampling is a challenging task. Standard methods include (adaptive) rejection sampling [Gilks and Wild, 1992] and importance sampling [Kloek and Van Dijk, 1978]. They all, however, typically scale poorly with dimension, and are not necessarily applicable to generic target distributions. Note, for the remainder of this chapter we drop the distinction between the parameters of the model, denoted previously as θ, and the samples of the process, slightly abusing notation to denote both by x.

3.2.1 Markov Chain Monte Carlo and Metropolis-Hastings methods

Markov Chain Monte Carlo methods [Brooks et al., 2011] generate a sequence of dependent samples from $\pi(\cdot)$ by generating a Markov chain with $\pi(\cdot)$ as its stationary distribution. In particular, the Law of Large Numbers does not necessarily require samples to be independent and so (3.1) can be approximated by the ergodic average of the generated samples. Under certain conditions [e.g Roberts and Rosenthal, 1997], there also exists a Markov chain Central Limit theorem replacing σ^2 in (3.3) with

$$\sigma^2 = \text{var}(f(x_1)) + 2\sum_{k=1}^{\infty} \text{cov}(f(x_i), f(x_{i+k})),$$

where $x_1 \sim \pi(\cdot)$. This typically results in estimates with lower statistical efficiency when compared to independent sampling approaches, offset by significantly increased computational efficiency. Further, implementation of these methods requires no knowledge of the geometry of the posterior distribution beyond access to pointwise evaluations of $\log \pi$ and possibly its gradient. As such they have become the standard workhorse for Bayesian modelling, enhanced by a large suite of computational tools facilitating their use in applied statistics [Lunn et al., 2009, Stan Development Team, 2025, Fjelde et al., 2025].

The most common construction of a Markov chain that has π as its stationary distribution is based on constructing a chain that is *i)* $\pi-$irreducible, *ii)* aperiodic and *iii)* $\pi-$invariant. Full definitions and statements of relevant results are supplied in Appendix A. Here π-irreducibility and aperiodicity ensure that the chain can explore the entire posterior distribution and are often easy to verify. Arguably the simplest way of ensuring the chain is π-invariant is to ensure the chain is π-reversible, such that for all pairs $(x,x') \in \Omega \times \Omega$

$$\pi(x)p(x,x') = \pi(x')p(x',x),$$

where $p(x,x')$ is the one step ahead transition density of the chain conditional on the current state x.

Metropolis-Hastings methods [Metropolis et al., 1953, Hastings, 1970] ensure the above condition is met by taking a general Markov chain with transition kernel $q(x,x')$, and then coercing it to the correct stationary distribution through an acceptance step that either moves the adjusted chain to a new state or leaves it in its current position. More precisely, the transition kernel of the Metropolis-Hastings chain is given by

$$\begin{aligned} P(x_i, \mathrm{d}x_{i+1}) &= \alpha(x_i,x_{i+1})q(x_i,x_{i+1})\mathrm{d}x_{i+1} \\ &+ \int_\Omega (1-\alpha(x_i,x_{i+1}))q(x_i,x_{i+1})\mathrm{d}x_{i+1}\delta_0(\mathrm{d}x_{i+1}-x_i), \\ \alpha(x_i,x_{i+1}) &= \min\left\{1, \frac{\pi(x_{i+1})q(x_{i+1},x_i)}{\pi(x_i)q(x_i,x_{i+1})}\right\}. \end{aligned}$$

In the above the choice of q, referred to as the proposal distribution, will determine the efficiency of the chain. Common choices include independent proposal distributions [Tierney, 1994] and random walk proposals that centre q at x [Gelman et al., 1997]. More generally, proposal distributions can be improved by incorporating local information about the posterior via the gradient, $\nabla \log \pi(x)$, commonly incorporated by taking q as the one step transition kernel of a discretised π-stationary Langevin diffusion. Examples of this include the Metropolis Adjusted Langevin Algorithm [Roberts and Tweedie, 1996] and the more recently introduced Barker proposal [Livingstone and Zanella, 2022]. Full forms of these proposals are outlined in Appendix A.

Reversibility ensures that Metropolis-Hastings methods are widely applicable, however, it also results in processes that exhibit diffusive behaviour. This has motivated the development of methods that reduce this diffusive behaviour.

3.2.2 Kinetic sampling

A common solution to this problem is to consider kinetic sampling methods. In short these approaches augment the state space with velocities, $v \in \mathcal{V} \subseteq \mathbb{R}^d$, such that the resulting state, $z = (x, v) \in \mathbb{R}^d \times \mathcal{V}$, is driven through the state space by these velocities. In theory this reduces the diffusivity of reversible methods as the process retains memory of its trajectory through the state space.

The most common kinetic sampling methods are based on Hamiltonian dynamics [Neal, 2011], that target a joint stationary distribution defined by,

$$\pi(x, v) \propto \exp(-H(x, v)), \quad H(x, v) = U(x) + v^\top v.$$

where H is the Hamiltonian, U is the potential energy corresponding to the negative log-density of the desired target distribution, $\pi(x) \propto \exp(-U(x))$, and the remaining terms in H are referred to as the kinetic energy. The continuous-time evolution of (x_t, v_t) is then given by the system of differential equations

$$\frac{dx_t}{dt} = \frac{\partial H}{\partial v_t}, \quad \frac{dv_t}{dt} = -\frac{\partial H}{\partial x_t}. \tag{3.4}$$

In the context of sampling Hamiltonian dynamics are an attractive choice of proposal distribution as they are energy conserving. Explicitly, an initial (x_0, v_0) propagated for time t according to the above equations will have the property that $H(x_0, v_0) = H(x_t, v_t)$. An idealised sampler that utilises these dynamic, referred to in the literature as Randomised Hamiltonian Monte Carlo [Bou-Rabee and Sanz-Serna, 2017], is then defined by

1. Given (x_0, v_0), simulate $t^* \sim \text{Exponential}(\Lambda)$.

2. Generate (x_{t^*}, v_{t^*}) by evolving (x_0, v_0) for time t^* according to Hamiltonian dynamics, (3.4).

3. Take (x_{t^*}, v_{t^*}) to be the next samples in the chain, and refresh v from its stationary distribution $v \sim \text{Normal}(0, I)$.

The idealised version of this algorithm avoids a Metropolis correction due to the energy-preserving property of (3.4). In practice, however, for most target distributions (3.4) cannot be solved exactly and requires the use of a numerical integrator. The dynamics are therefore only approximately energy-conserving and require a Metropolis step to account for integrator error and ensure the correct distribution is targeted. The resulting algorithms, while highly efficient, require careful selection of tuning parameters or the use of adaptive MCMC methods [Hoffman and Gelman, 2014, Bou-Rabee et al., 2024].

An alternative perspective on Randomised Hamiltonian Monte Carlo is as an example of a Piecewise Deterministic Markov Process [Davis, 1993] as it is defined by: *i)* Generating a random event time. *ii)* Evolving (x, v) according to a deterministic flow until this event time. *iii)* Updating v according to some event dynamics. Computationally, for Randomised Hamiltonian Monte Carlo, steps *i)* and *iii)* are trivial with the computational cost of the method arising from generating the deterministic flow in step *ii)*. For the remainder of this chapter we will review a new class of MCMC methods based on PDMPs that retain the attractive properties of kinetic sampling methods, and utilise simple to compute deterministic dynamics removing the computational cost associated with *ii)*.

3.3 Piecewise Deterministic Markov Processes

We begin by formalising the definition of Piecewise Deterministic Markov Processes introduced in the previous section. PDMPs are defined by a state and velocity on the augmented space $(x, v) \in \Omega \times \mathcal{V}$. We denote in the following the j^{th} element of $z_t = (x_t, v_t)$ as $z_{t,j} = (x_{t,j}, v_{t,j})$. The processes are constructed via three components Davis [1993]:

1. **Deterministic dynamics:** Given by the system of ordinary differential equations
$$\frac{dz_{t,j}}{dt} = \Phi_j(z_t), \quad j = 1, \ldots, 2d$$
such that the process at time $t+s$, conditional on no event having occurred, is given deterministically by $z_{t+s} = \Psi(z_t, s)$, for known functions $\Phi_j(z_t)$ and $\Psi(z_t, s)$, where the latter is the flow map of the above ordinary differential equation.

2. **Event rate:** $\Lambda^E(z_t)$ a state dependent event rate under which events occur according to an inhomogeneous Poisson process. More precisely, given the current state of the sampler, z_t, this defines the next event time as
$$t^* = \inf\{s > 0 : \int_0^s \Lambda^E(z_{t+u})du = -\log V\}, \tag{3.5}$$
for $V \sim \text{Uniform}(0,1)$. We will usually suppress the dependence on z_t for the remainder of this thesis, denoting the event rate by $\Lambda^E(t)$.

3. **A transition kernel:** $q(\cdot \mid z_t)$, which determines the change in the velocity of the process occurring at each event.

For the remainder of this thesis, unless explicitly stated otherwise, the deterministic dynamics are defined by

$$\frac{dx_t}{dt} = v_t, \quad \frac{dv_t}{dt} = 0, \tag{3.6}$$

such that the state of the process evolves linearly with constant velocity. Further, all the examples of PDMPs discussed in this thesis will be designed to target from a stationary distribution given by

$$\pi^*(x, v) \propto \pi(x)\rho(v),$$

for a pre-defined stationary distribution for the velocities and a target posterior distribution $\pi(x)$. Samples from the posterior are recovered by simply marginalising out v. Under the above construction these processes are non-reversible, and will therefore typically result in faster mixing times and smaller asymptotic variances than reversible alternatives [Bierkens, 2016, Bierkens et al., 2019, Andrieu and Livingstone, 2021].

Note, in contrast to the MCMC methods defined in Section 3.2, samples are given by the piecewise continuous sample paths of the process rather than the value of the process at event times. In particular, the distribution of the state at event times is not π. In practice rather than computing the ergodic averages defined by

$$\mathbb{E}_{\pi}[f(x_t)] \approx T^{-1} \int_0^T f(x_t)dt,$$

it is computationally simpler to sample from the path of the process, either at fixed intervals or uniformly at random, and then compute averages using (3.2). Practically, these samples can be generated during a post-processing step, allowing only the skeleton points of the algorithm to be stored during sampling.

The primary challenge with generating these processes is the generation of event times that arise according to the inhomogeneous Poisson process with rate $\Lambda^E(t)$. This rate is additively comprised of two parts $\Lambda^E(t) = \Lambda^B(t) + \Lambda^R$, where $\Lambda^B(t)$ is an inhomogeneous bounce rate that depends on the local geometry of the posterior, and $\Lambda^R \geq 0$ is an homogeneous refreshment rate required to ensure irreducibility in some cases. Note, this is in contrast to randomised Hamiltonian Monte Carlo, where the event times are simple to generate and the computational cost arises from integrating the deterministic dynamics. Generating these event times is the focus of Section 3.5.

As an aside, the development of PDMPs for the purposes of sampling has primarily occurred in the context of statistical physics [Bernard et al., 2009, Michel et al., 2014] and are referred to as event chain Monte Carlo methods. Interestingly, this mirrors the original development of both the original MCMC method [Metropolis et al., 1953] and Hamiltonian Monte Carlo (referred to as Hybrid Monte Carlo) [Duane et al., 1987]. A recent review of connections between sampling in Bayesian

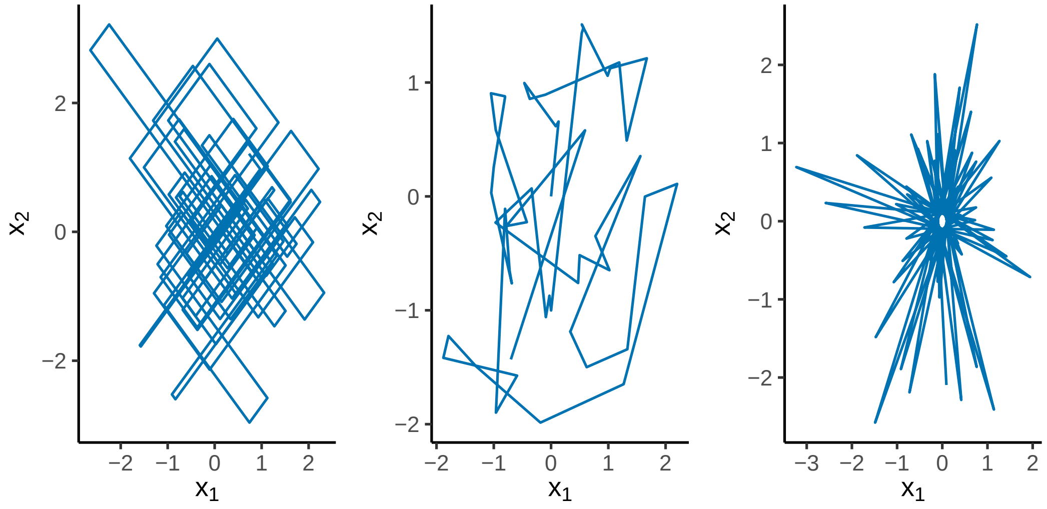

Figure 3.1: Sample paths of PDMPs targetting a two dimensional standard Gaussian distribution. (Left) The Zig-Zag sampler. (Centre) The Bouncy Particle sampler with $\Lambda^R = 1$. (Right) The Bouncy Particle sampler with $\Lambda^R = 0$.

inference and statistical physics can be found in Faulkner and Livingstone [2024].

3.4 Example processes

There are several examples of PDMPs that satisfy the definition introduced in the previous section. Here, we introduce the two processes that have seen the most interest in the Bayesian computation literature.

3.4.1 Zig-Zag sampler

The Zig-Zag sampler [Bierkens et al., 2019] utilises velocities with a uniform stationary distribution on $\mathcal{V} = \{-1,1\}^d$ and deterministic dynamics given by (3.6). The remaining dynamics of the process are then defined in coordinate-wise fashion with the event rate for the j^{th} coordinate given by

$$\Lambda_j^F(t) = \max\{0, v_{t,j}\partial_j U(x_{t,j})\},$$

and the corresponding event kernel flipping the associated velocity $v_{t,j} \mapsto -v_{t,j}$.

Intuitively, in the j^{th} coordinate, if the process is moving into areas of lower potential (equivalently higher posterior density) it continues uninterrupted. If, however, the converse is true, then $v_{t,j}$ flips with rate proportional to the rate of growth in the potential. The result is an almost-surely continuous (on x-space), piecewise deterministic process, whose sample paths produce a zig-zag pattern shown in Figure 3.1.

In practice, to avoid simulating d inhomogeneous Poisson processes, the next event time, t^*, can be generated by simulating $\Lambda^B(t) = \sum_{j=1}^d \Lambda_j^F(t)$. Once this event time is simulated, the coordinate to switch can then be chosen with probability proportional to $\Lambda_j^F(t^*)$. See Appendix A for more detail.

In general[1], the Zig-Zag sampler is irreducible with $\Lambda^R = 0$ [Bierkens et al., 2019]. This reduces the number of tuning parameters needed to implement the sampler in practice, and the diffusivity associated with refreshments.

3.4.2 Bouncy Particle Sampler

The Bouncy Particle sampler [Bouchard-Côté et al., 2018] takes $\mathcal{V} = \mathbb{R}^d$ or $\mathcal{V} = \mathbb{S}^{d-1}$ with respectively Gaussian or uniform invariant distribution, and linear deterministic dynamics. In contrast to the coordinate-wise updates of the Zig-Zag sampler, the Bouncy Particle sampler is defined by the global reflection rate

$$\Lambda^B(t) = \max\{0, \langle \nabla U(x_t), v_t \rangle\},$$

and transition kernel that updates velocities by reflecting them off the contours of the potential. This can be viewed as first computing the orthogonal decomposition of v_t with respect to the subspace spanned by ∇U and then flipping the component aligned with ∇U

$$v = v_\perp + v_{\nabla U}, \quad v \mapsto v_\perp - v_{\nabla U}. \tag{3.7}$$

The sample paths of the Bouncy Particles sampler with refreshment are shown in Figure 3.1. The intuition regarding when events occur is similar to the Zig-Zag sampler, however, rather than whether individual coordinates are moving into areas of higher potential, all coordinates are considered together, with reflection events

[1] For a counter-example consider a potential with square contours. The process is then irreducible as the the trajectory can only navigate the potential in clockwise or counter-clockwise fashion, depending on the initial conditions. This example was shown to us in a talk given by Prof. Gareth Roberts at a workshop at the University of Warwick in 2024.

only occurring when $\langle \nabla U(x_t), v_t \rangle > 0$. The event rate for the Bouncy Particle sampler is therefore often smaller than that of the Zig-Zag sampler due to the possibility of terms cancelling. Note, in one-dimension the two processes are identical.

Unlike the Zig-Zag sampler, the Bouncy Particle sampler typically requires $\Lambda^R > 0$ in order to be irreducible [Bouchard-Côté et al., 2018]. The behaviour of the process when $\Lambda^R = 0$ is shown on a Gaussian target distribution in Figure 3.1 where, without refreshments, the process is unable to reach a ball centred at the mode of the distribution. The efficiency of the process is highly sensitive to the choice of Λ^R. Scaling limit arguments suggest that optimal tuning of Λ^R results in 78.12% corresponding to refreshments [Bertazzi and Bierkens, 2022, Bierkens et al., 2022]. This introduces significant diffusivity into the process, limiting the benefit of the non-reversible dynamics.

Several authors have suggested generalising the dynamics of the Bouncy Particle sampler to incorporate randomness in the transition kernels [Wu and Robert, 2019, Michel et al., 2020]. The primary advantage of this is that it reduces the reliance on refreshments for irreducibility. In particular, in Chapter 5 we review the work of Michel et al. [2020] in this direction in more detail.

Initially it seems easy to conclude that the Bouncy Particle sampler is naturally more efficient than the Zig-Zag sampler as the velocities for each coordinate are updated at each event time. Note, however, that in high dimensions random vectors are close to orthogonal, and as such v will be increasingly dominated by $v_\perp$. Therefore at each bounce event the changes in v in a given coordinate will be small, requiring several events in order to flip the velocity in the fashion of the Zig-Zag sampler. An argument in favour of the Bouncy Particle sampler is in the context of anisotropic target distributions, the Bouncy Particle sampler has been shown to have preferable scaling behaviour in contrast to the Zig-Zag sampler [Bierkens et al., 2025].

3.4.3 Other processes

We briefly overview some alternative PDMP samplers that have not seen the same methodological interest that the Zig-Zag and Bouncy Particle samplers have.

3.4.3.1 The Boomerang sampler

The Boomerang sampler [Bierkens et al., 2020] replaces the linear dynamics of the above processes with the Hamiltonian dynamics defined with respect to a Gaussian reference measure

$$\frac{dx_t}{dt} = v_t, \quad \frac{dv_t}{dt} = -x_t.$$

The resulting deterministic trajectories then have an explicit solution, and events and reflections occur with the same rate and transitions as the Bouncy Particle sampler. The potential, however, is now defined relative to the reference measure, resulting in events that essentially correct the discrepancy between the posterior and the reference measure. This generalises an idea originally introduced in Vanetti et al. [2017].

3.4.3.2 The coordinate sampler

The coordinate sampler [Wu and Robert, 2020] takes the space of velocities to be $\mathcal{V} = \{\pm e_j, j = 1, \ldots, d\}$ where e_i are the canonical basis vectors of $\mathbb{R}^d$, with the motivation that often event times are simpler to generate when updating single coordinates at a time. The event rate is then taken to be $\Lambda^B(t) = \max\{0, \langle \nabla U(x_t), v_t \rangle\}$, and at event times a new velocity, $v^* \in \mathcal{V}$ is selected with probabilities proportional to $\max\{0, \langle \nabla U(x_t), v^* \rangle\}$.

3.4.3.3 Hamiltonianised PDMPs

Samplers that utilise Hamiltonian dynamics use momentum to explore the state space through kinetic energy that increases when potential energy is lost (3.4). In contrast, the previously discussed PDMP samplers only retain momentum in the sense that the velocities encode piecewise constant direction of movement.

This observation has motivated the development of “Hamiltonianised” versions of the Zig-Zag and Bouncy Particle samplers that build momentum as the potential energy reduces [Chin and Nishimura, 2024, Nishimura et al., 2025]. This is achieved for the Bouncy Particle sampler by replacing the generation of the next event time in (3.5), with

$$t^* = \inf_{t>0}\{s > 0 : \int_0^s v_{t+u}^\top \nabla U(x_{t+u}) \mathrm{d}u = l\}, \tag{3.8}$$

for $l \sim \text{Exponential}(1)$ that is re-drawn at each event time. As $v_{t+u}^\top \nabla U(x_{t+u})$ is negative when the potential is decreasing, this allows the processes to build momentum. The momentum introduced is an attractive theoretical property, however, it limits implementation as event times are no longer able to be generated via Poisson thinning (Section 3.5). Instead, (3.8) needs to be solved directly leaving the algorithm valid only for potentials with convex level sets.

3.5 Generating the process

For the PDMPs outlined in the previous section the computational cost associated with their implementation lies in the generation of event times given by the inhomogeneous Poisson process with rate $\Lambda^B(t)$. This is an active area of research that we review in this section.

3.5.1 Exact simulation and Poisson thinning

In simple cases (3.5) can be solved directly. For example, for standard Gaussian target distributions [Bouchard-Côté et al., 2018], t^* is directly computed as

$$t^* = \frac{1}{|v|^2}\begin{cases} -xv + \sqrt{-|v|^2 \log V}, & xv \leq 0, \\ -xv + \sqrt{(xv)^2 - |v|^2 \log V}, & xv > 0. \end{cases}$$

Such examples are the exception, however, and while this equation can be solved numerically [Bouchard-Côté et al., 2018, Pagani et al., 2024], these methods are computationally expensive due to the need for multiple evaluations of $U(x)$.

A more common approach to exactly simulate event times is to utilise Poisson thinning [Lewis and Shedler, 1979] (Appendix A). In short this requires upper bounding the event rate $\bar{\Lambda}^E(t) > \Lambda^E(t)$, with the event rate of a Poisson process for which (3.5) has an explicit solution. A candidate time is then generated using the upper bounding rate, $\bar{\Lambda}^E(t)$, with an event occurring at that time with probability $\Lambda^E(t)/\bar{\Lambda}^E(t)$. If the event is rejected, the process is repeated starting from the rejected candidate time.

Example bounds include constant upper bounds (applied to, for example, logistic regression models [e.g Bierkens et al., 2019]) and affine upper bounds that can be computed when the potential is Lipschitz with known Lipschitz constant. These bounds depend on the target distribution and their existence does not mean they will be efficient [e.g Section 6, Bertazzi et al., 2023], in that the ratio $\Lambda^E(t)/\bar{\Lambda}^E(t)$ may be small resulting in a large number of rejected events.

This has lead to several proposed methods that seek to find efficient methods for approximating $\Lambda^E(t)$ using local evaluations of ∇U. There are broadly two approaches to this problem. The first seeks to find a tractable rate $\bar{\Lambda}^E(t)$ that approximately upper bounds $\Lambda^E(t)$. The second seeks to directly approximate $\Lambda^E(t)$ with a tractable rate. In both cases the bias introduced by numerical errors and approximation is then controlled using tuning parameters. In the second, the bias can also be explicitly corrected for using a Metropolis correction. Four of these methods are visualised for $\Lambda^B(t) = t^4$ in Figure 3.2. This event rate is representative of the event rates found when sampling from posteriors with light tails. This commonly occurs in survival models due to exponential terms arising in the potential.

3.5.2 Approximating an upper bound

Methods for numerically generating an upper bound typically do this locally over an interval, $(0, t_{\max})$.

3.5.2.1 Automatic Zig-Zag

The automatic Zig-Zag method [Corbella et al., 2022] sets $\bar{\Lambda}^E(t)$ to a constant by upper bounding the event rate over the interval $(0, t_{\max})$. This is visualised in Figure 3.2 (A). The primary method for this is Brent's method [Brent, 1971]. To avoid repeated evaluations of ∇U, however, the authors introduce heuristics that check for monotonicity of the function on $(0, t_{\max})$. If these checks hold, the maximum at either end of the interval can be used in place of the maximum found by repeated iteration of Brent's method, saving significant computational cost. If the proposed time is greater than $t_{\max}$ the process is repeated on the next interval, $(t_{\max}, 2t_{\max})$ to ensure the upper bounds remain valid. We discuss this method further and extend it in Chapter 4.

Recently, Andral and Kamatani [2024] extended this approach by subdividing

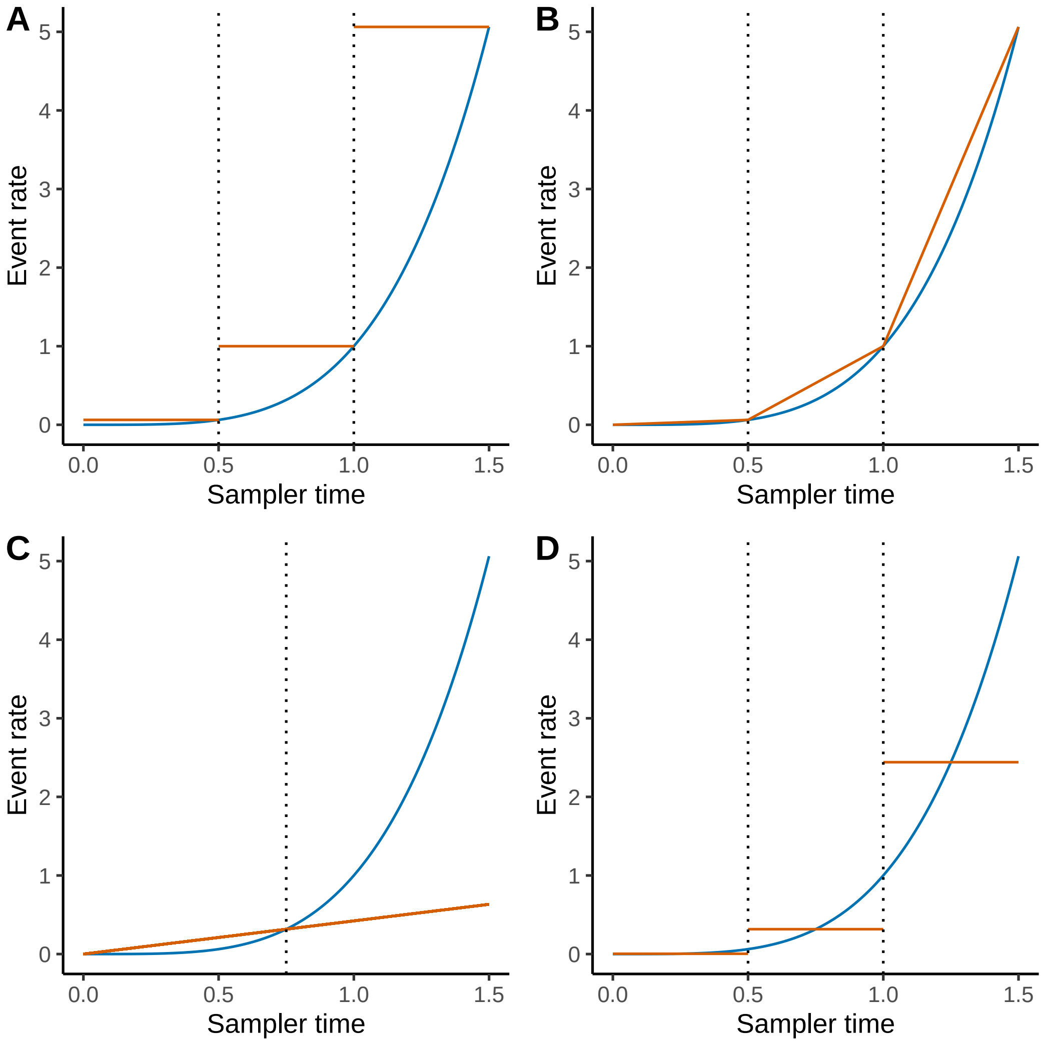

Figure 3.2: Visualisation of methods for upper bounding or approximating $\Lambda^B(t)$, for $\Lambda^B(t) = t^4$, assuming no events or refreshments occur. (Blue). The upper bound or approximation for $\Lambda^B(t)$ is shown in red. Points of evaluation of the event rate are shown with dotted lines. (A) The automatic Zig-Zag method, Section 3.5.2.1. (B) The concave-convex method, Section 3.5.2.2. (C) The linear interpolation method, Section 3.5.2.3. (D) Splitting schemes for approximating $\Lambda^B(t)$, Section 3.5.3.1.

$(0,t_{\max})$ into smaller intervals and then computing constant bounds on each interval. The additional computational cost is circumvented by evaluating ∇U at each time point in parallel.

3.5.2.2 Convex-concave bounding

Writing the event rate as $\Lambda^E(t) = \max\{0, l(t)\}$, $\Lambda^E(t)$ can be upper bounded analytically if $l(t)$ can be decomposed into convex and concave components, $l(t) = l_{\cup}(t) + l_{\cap}(t)$, respectively. Here $l_{\cup}(t)$ can be upper bounded using piecewise linear segments connecting points of l and $l_{\cap}(t)$ can be bounded by connecting the tangents at a set of evaluation points. The bounds are then combined to generate an upper bound for $\Lambda^E(t)$. This forms the basis of the concave-convex PDMP method [Sutton and Fearnhead, 2023].

This method is exact if this decomposition can be done analytically. Alternatively, $\Lambda^E(t)$ can be approximated, either by Taylor expansion or Lagrange polynomial interpolation on a fixed interval $[0,t_{\max})$. If the k^{th} derivative of l can be bounded, then a fixed offset can be added such that the method remains exact, otherwise this upper bound is only approximate and the method will bias the resulting samples. This method is visualised in Figure 3.2 (B). As the example event rate is convex, the method generates linear segments between evaluation points. For more complex target distributions upper bounds are unlikely to be as tight.

3.5.2.3 Linear interpolation

A final method for approximating upper bounds is introduced by Goan et al. [2023]. The method begins by evaluating $\Lambda^E(t)$ at the current state and some future time t_{init}. A piecewise linear upper bound is then constructed by interpolating between $\Lambda^E(t)$ at these two points. If the proposed event is rejected the next bound is then proposed by linearly interpolating between $\Lambda^E(t)$ at t^* and the rejected event time. This sequence is then repeated until an event time is accepted. To offset numerical bias this introduces, the authors suggested targeting a scaled Bouncy Particle sampler rate, $\tilde{\Lambda}^B(t) = \max\{0, \alpha\langle v_t, \nabla U(x_t)\rangle\}$, for $\alpha \geq 1$. Here larger values of α reduce sampling bias, at the cost of lower computational efficiency. The method is visualised in Figure 3.2 (C). In particular the method generates a poor bound without correction for the convex event rate. The performance would be improved with the suggested correction and on concave event rates [Figure 2, Goan et al., 2023].

3.5.2.4 Tuning parameters and bias reduction

All these methods are sensitive to the choice of tuning parameters. For the automatic Zig-Zag methods and concave-convex PDMP method the choice of bounding interval, $t_{\max}$, determines the efficiency of the method. As $t_{\max} \to 0$, the resulting rate becomes arbitrarily tight meaning that proposed events are accepted with probability close to 1 (meaning there is little to no thinning), but at the cost of having to compute $\bar{\Lambda}^E(t)$ over a large number of intervals before an event is observed. Alternatively, if $\Lambda^E(t)$ is unbounded, as $t_{\max} \to \infty$ we have $\Lambda^E(t)/\bar{\Lambda}^E(t) \to 0$ resulting in the need to compute many thinning events before an event is accepted. The computational efficiency of the sampler is therefore dependent on balancing the cost of constructing a tighter upper bound $\bar{\Lambda}^E(t)$ against that of rejecting too many proposed events when the bound is loose. Similar considerations hold for the choice of initial interval in the method of Goan et al. [2023].

3.5.3 Approximating the event rate

The alternative approach to generating PDMP dynamics is to directly approximate $\Lambda^E(t)$ with a tractable event rate.

3.5.3.1 Splitting schemes

This idea is first presented by Bertazzi et al. [2023] who introduce splitting schemes for PDMPs. Splitting schemes are a standard tool in the study of dynamical systems, whereby individual components of a system are simulated in turn, and has recently been developed for the PDMPs outlined in Section 3.4. Here the deterministic, reflection and refreshment rates are updated separately given a step size δ. This is outlined for the Bouncy Particle sampler in Algorithm 1, and the approximation to the event rate visualised in Figure 3.2. The computational advantage of this method is that by updating each component of the algorithm in turn, the inhomogeneous event rate is replaced by a fixed homogeneous event rate that can be simulated exactly.

This approach introduces a bias into the resulting posterior, that vanishes as

Algorithm 1 A single iteration of a splitting scheme for the bouncy particle sampler

```
Input state and velocity $(x_0, v_0)$ and step-size, $\delta$.
Set $v_1 = v_0$
With probability $(1 - \exp(-\Lambda^R \delta/2)$ refresh $v_1$.                ▷ Refreshment
Set $x_1 = x_0 + v_1 \delta/2$.                                         ▷ Deterministic dynamics
With probability $(1 - \exp(-\Lambda^E(x_1)\delta)$ update $v_1$ via (3.7).     ▷ Reflection
Set $x_1 = x_1 + v_1 \delta/2$.                                         ▷ Deterministic dynamics
With probability $(1 - \exp(-\Lambda^R \delta/2)$ refresh $v_1$.                ▷ Refreshment
Return $(x_1, v_1)$.
```

$\delta \to 0$. Alternatively a Metropolis correction can be used to fully correct this bias. Notably, this approach replaces the continuous time process with a discrete time approximation.

A similar proposal has recently been introduced by Chevallier et al. [2025], where the authors use PDMPs as a proposal distribution within a skew-reversible Metropolis framework. Proposals are generated by propagating the process forwards and backwards in time until a stopping criterion is reached in similar fashion to the No U-turn sampler [Hoffman and Gelman, 2014].

Generation of the event rate remains the main limitation for wider implementation of PDMP based samplers. In particular, non-reversible samplers will often out-perform reversible alternatives in terms of *statistical* efficiency, however recent work has suggested this benefit may be limited [Roberts and Rosenthal, 2025]. Fast generation of the event rate is therefore required, to ensure the computational cost of these processes is low enough that these benefits may be realised. We briefly note that a promising development in this direction is the use of unbiased sub-sampling techniques for PDMPs, allowing event times to be generated using only a single sample from the data [Bierkens et al., 2019, Agrawal et al., 2024].

3.6 Transdimensional sampling

The MCMC methods discussed so far have been focused on sampling from posteriors where the dimension of the state space is fixed. In this section we review methods for sampling from posteriors where the dimension of the state space is updated during sampling. This phenomenon commonly occurs in Bayesian models where a prior has been placed on a structural quantity, for example the choice of covariates to include in a linear predictor, or the number of components in a mixture model [Mitchell and Beauchamp, 1988, Richardson and Green, 1997]. The posterior is then defined as

$$\pi(x,k) \propto \pi(x \mid k)\pi(k), \quad k \in \mathcal{K},$$

where $\mathcal{K}$ is a set of model indicators, and $\pi(x \mid k)$ is the posterior conditional on the specification of the k^{th} model.

3.6.1 Reversible Jump MCMC

The standard approach to this problem is reversible jump MCMC [Green, 1995]. This generalises Metropolis-Hastings methods to cases where dimension of the posterior needs to be updated during sampling.

The core principle behind reversible jump MCMC is to separate the process of proposing candidate states into two parts. First generating some random innovation u with density $g(u)$, and then generating a proposal using the diffeomorphism $h : (x,u) \mapsto (x',u')$, with inverse h'. Here, x' is the candidate state, and u' is the innovation required in the reverse move from $x' \mapsto x$. The new state is then accepted with probability

$$\alpha(x,x') = \min\left\{1, \frac{\pi(x')g(u')}{\pi(x)g(u)}\left|\frac{\partial(x',u')}{\partial(x,u)}\right|\right\}, \tag{3.9}$$

where the ratio on the right hand side is referred to as the Metropolis-Hastings-Green ratio.

This framework can then be applied freely to the transdimensional case as long as h remains a diffeomorphism. This is achieved through a dimension matching condition, whereby if the dimension of x,u,x',u' are given by d,r,d',r', we set that we require $d+r = d'+r'$. If this does not hold, either h or h' would not be differentiable.

Further, multiple types of move can be included in the sampler. When the probability of making move m given state x is $j_m(x)$, the acceptance probability is then given by

$$\alpha(x,x') = \min\left\{1, \frac{\pi(x')j_m(x')g_m(u')}{\pi(x)j_m(x)g_m(u)}\left|\frac{\partial(x',u')}{\partial(x,u)}\right|\right\}. \tag{3.10}$$

This is commonly used to alternate between different across-model proposals, and between reversible jump moves that update k and fixed dimension Metropolis-Hastings steps that sample from $\pi(x \mid k)$.

3.6.1.1 Designing proposals

The efficiency of reversible jump MCMC methods is strongly dependent on carefully designed between model moves. In general, a common desiderata for these proposals is that they are constructed such that the proposed state has similar posterior support to the current state, ensuring high acceptance rates for the move and its reverse.

To ensure this most reversible jump schemes only consider local moves within model space e.g adding or removing a single variable in variable selection. Many popular methods then utilise some known aspect of the model structure. For example, the reversible jump scheme for Normal mixtures with an unknown number of components developed by Richardson and Green [1997], which involves split-merge and birth-death moves. Merge moves combine two components into a single component such that $\mu = \frac{1}{2}(\mu_1 + \mu_2)$, with the reverse split moves generating two components from one, and the generation of new variances being aided by moment-matching. Death moves remove a component with no allocated observations from the model, and birth moves introduce an empty component into the model.

This intuition is developed further by Brooks et al. [2003] who suggest designing transdimensional proposals around a centring point between two nested models. This point is defined as a subspace $\Gamma \subset \mathbb{R}^d$, such that the two likelihoods are equivalent. For example, in Bayesian variable selection this corresponds to a covariate equalling 0 [Mitchell and Beauchamp, 1988].

3.6.2 Transdimensional PDMP sampling

PDMPs are defined by deterministic continuous sample paths between event times. A consequence of this feature is that, when sampling from nested models where Γ

is a $d-1$-dimensional subspace, these processes will eventually intersect Γ. This feature has been utilised to design samplers that move into the lower-dimensional model at exactly this point, when the hyperplane is of the form $\Gamma = \{x \in \mathbb{R}^d : x_j = 0\}$ [Chevallier et al., 2023, Bierkens et al., 2023a]. This is commonly induced by the use of spike and slab priors,

$$\pi(dx) \propto \omega\pi(x) + (1-\omega)\delta_0(dx),$$

where $\delta_0(\cdot)$ is a Dirac mass at 0 representing the spike, and $\pi(x)$ is the slab, i.e the prior density conditional on not being in the model.

Given the posterior induced by these priors, PDMPs sample as

1. A standard PDMP sampler on $\mathbb{R}^d$.
2. When the process intersects with Γ the velocity in the j^{th} coordinate is set to 0.
3. The process continues as a standard PDMP on $\mathbb{R}^{d-1}$.
4. The process moves back to the higher dimensional space with rate Λ^S.

In designing the above process Chevallier et al. [2023] refresh v_j when the process returns to the higher-dimensional space, basing their construction on a reversibility condition. In contrast, the sampler of Bierkens et al. [2023a] resets v_j to its value when the sampler intersected Γ, retaining the non-reversible dynamics of the original PDMP. These differences are illustrated in the contrasting unsticking rates for the Bouncy Particle sampler with Gaussian velocities

$$\Lambda_1^S = \frac{\omega}{1-\omega}\pi(0)\frac{2}{\sqrt{2\pi}}, \quad \Lambda_2^S = \frac{\omega}{1-\omega}\pi(0)|v_j|,$$

with Λ_1^S corresponding to the rate in Chevallier et al. [2023] and Λ_2^S the rate in Bierkens et al. [2023a].

In Chapter 6 we further illustrate the differences between these approaches and extend both methods to the case when Γ is a general $d-1$-dimensional embedded manifold.

3.7 Bayesian computation for survival analysis

We conclude this chapter by highlighting current approaches to Bayesian computation for survival models, with a focus on those used in HTA. Many survival models used in HTA have fixed dimension, a small or moderate number of parameters, and small or moderate sample sizes. These models can therefore be fit using generic sampling approaches. The current state-of-the-art is the implementation of the No-U-Turn sampler in Stan [Hoffman and Gelman, 2014, Stan Development Team, 2025]. In R, parametric and spline survival models can be fit through the development branch of the `rstanarm` package [Brilleman et al., 2020]. Further, bespoke Stan implementations tailored to HTA are available for parametric models via the `survhe` package [Baio, 2020], and for M-splines in the `survextrap` package [Jackson, 2023]. In addition, Kearns et al. [2021] utilises a Stan implementation for dynamic survival models with a cure fraction. Alternative implementations include the use of Integrated Nested Laplace Approximation [Baio, 2020, Che et al., 2023] and Gibbs sampling via BUGS [Lunn et al., 2009, Demiris et al., 2015].

Transdimensional sampling is less commonly used in HTA survival models, in part due to limited off the shelf tools that can be used for posterior sampling. Cooney and White [2023a] implement a reversible jump sampler to average over the location and number of knots in a piecewise exponential model. Notably, they specify independent Gamma priors for the local hazards. This simplifies the posterior to independent, conjugate exponential-gamma models for each local hazard. Alternatively approaches to Bayesian model averaging in an HTA context have also implemented information criteria based approximations [Negrín et al., 2017].

Chapter 4

Averaging polyhazard models

The content of this chapter is based on the paper

> L. Hardcastle, S. Livingstone, and G. Baio. Averaging polyhazard models using Piecewise Deterministic Monte Carlo with applications to data with long-term survivors. *arXiv preprint arXiv:2406.14182*, 2024

in press at *Annals of Applied Statistics*.

4.1 Introduction

Polyhazard models, introduced in Section 2.2.2.1, are a class of flexible parametric models for time-to-event data, defined by additively combining hazards from simpler, typically one- or two-parameter survival distributions

$$h_Y(y) = \sum_{j=1}^{K} h_j(y),$$

where Y is a random variable representing a time-to-event outcome, and $h_Y(y), h_j(y)$ are hazard functions.

As reviewed in Chapter 2, standard methods for estimation of mean survival involve imposing parametric assumptions on Y, which given parameters θ, allows mean survival to be computed either analytically or through a simple numerical approximation, with or without censored observations. In many cases these models encode identical covariate assumptions as non-parametric alternatives (e.g proportional hazards or accelerated failure time) with the addition of a suitable parametric extrapolation mechanism. This broadly follows the recommendations of Latimer [2011] who proposes a set of two- or three-parameter survival distributions to be used for this purpose; given the leading role, globally, of NICE, these have become the gold-standard in HTA. While parsimonious, these distributions are typically restricted to hazards that are increasing, decreasing or unimodal and covariate effects restricted by assumptions of proportional hazards or odds. Further, these standard models infer the parameters dictating extrapolation from the whole sample, while in reality observations at the end of the trial are likely to contain more information about how survival can be expected to evolve in the long-term.

Polyhazard models can capture a much wider range of hazard curves while retaining the interpretability and parsimony of simpler models. Further, due to the additive decomposition of the hazard function, later observations naturally have more influence on long-term survival. This has resulted in an increased interest in applications to Health Technology Assessment [Demiris et al., 2015, Rutherford et al., 2020].

Despite theses advantages applications of polyhazard models have been limited due to: *i)* the lack of accessible computational tools and understanding of how prior specification affects inference; *ii)* a number of structural choices which, in the presence of even a small number of covariates, leads to a space of candidate models which is infeasibly large to explore manually.

This chapter addresses these issues via Bayesian model averaging, to facilitating wider application of polyhazard models. In Section 4.2 we extend the polyhazard model by accounting for uncertainty in structural choices through an extended prior specification leading to a Bayesian model averaging approach. In Section 4.3 we develop bespoke Markov Chain Monte Carlo (MCMC) methodology extending existing sampling methods based on Piecewise Deterministic Markov Processes [PDMPs; Fearnhead et al., 2018]. This allows for efficient generation of posterior samples, reducing the computational burden from fitting each individual polyhazard model to fitting a small set of models with high posterior mass. PDMP-based samplers have emerged as a promising new direction in Bayesian computation.

Their development has been hindered, however, by a limited understanding of their effectiveness in applied settings, a limitation this chapter begins to address. In Section 4.4 we study the extended model by re-analysing a digitised version of data first studied by Demiris et al. [2015]. Through this comparative analysis we show the effects of non-informative vs weakly informative priors in this setting and the importance of accounting for structural uncertainty. Following this we apply the extended polyhazard model to two complex data sets based on survival times in stroke survivors from the Copenhagen Stroke study (COST) [Jørgensen, 1996] and from kidney transplant patients [Chen et al., 2022b].

4.2 Polyhazard models

We maintain the notation and definitions introduced in Chapter 2. In particular throughout this chapter we assume the data take the form of $\mathcal{D} = (y_i, \delta_i, w_i)_{i=1}^{n}$, where y_i are, possibly right-censored, survival times, δ_i are event indicators, and $w_i \in \mathbb{R}^p$ are vectors of individual covariates.

4.2.1 Polyhazard model definition

Polyhazard models [Berger and Sun, 1993, Louzada-Neto, 1999] are constructed by combining multiple independent parametric hazards via the additive formulation

$$h_{D,\theta,\gamma}(y \mid w) = \sum_{k=1}^{K} h_{D_k,\gamma_k,\theta_k}(y \mid w). \tag{4.1}$$

Each subhazard corresponds to a proper hazard function from a known distribution $D_k \in \mathcal{H}$, where $\mathcal{H}$ is a set of candidate distributions (for the examples considered in Section 4.4, $\mathcal{H} = \{\text{Weibull}, \text{Log-logistic}\}$). Each θ_k is a vector of subhazard specific parameters composed of a shape parameter ν_k, and rate, scale or location parameter μ_k, such that $\theta_k = (\nu_k, \mu_k)$. For each hazard, covariate information is included in the location parameter via a log-link, such that

$$\mu_k(w, \gamma_k) = \exp\left(\beta_{k,0} + \sum_{j:\gamma_{kj}=1} w_j \beta_{k,j}\right), \tag{4.2}$$

where $\gamma_{kj} \in \{0,1\}$ indicates whether the j^{th} covariate is included in the k^{th} subhazard. In practice we will centre and normalise each element of w such that for a given hazard $\beta_{k,0}$ can be interpreted as the location parameter for the average individual in the sample. This information is collated as $D = (D_k)_{k=1}^{K}$, $\theta = (\theta_k)_{k=1}^{K}$, and $\gamma = (\gamma_k)_{k=1}^{K}$, such that the model is completely defined by the specification of (K, D, γ, θ).

We place no restriction on the combination of simpler hazard forms, neither requiring each subhazard to be from the same parametric family nor requiring each parametric family to be represented in (4.1). Similarly, γ_k, need not be identical across all subhazards.

In this chapter we will focus on polyhazard models where $\mathcal{H}$ contains the Weibull and log-logistic distributions with respective hazard functions

$$h_W(y) = \mu \nu y^{\nu-1}, \quad h_{LL}(y) = \frac{(\frac{\nu}{\mu})(\frac{y}{\mu})^{\nu-1}}{1+(\frac{y}{\mu})^{\nu}},$$

while noting that the methods presented naturally extend to other choices [see for example Louzada-Neto, 1999].

Combining hazard functions with different shapes results in flexible baseline hazards and covariate effects that are more flexible than those possible with simpler models. Various example hazard shapes generated by combining Weibull and log-logistic hazards are shown in Figure 4.1.

We briefly address two common misconceptions regarding the polyhazard model. *i)* The polyhazard model is *not* a mixture model. In contrast, each individual in the population is subject to risk from every subhazard (with intensity determined by relevant covariates), and there is no explicit weighting of subhazards within the population. *ii)* While the form of (4.1) is recognisable as the hazard for an individual subjected to independent, latent competing risks, we do not necessarily assume that the data were generated in this way. Rather, we utilise the form of (4.1) as a flexible modelling assumption.

Standard application of polyhazard models typically follows one of two approaches. In the first K, D and γ are fixed *a priori*, meaning inference is performed

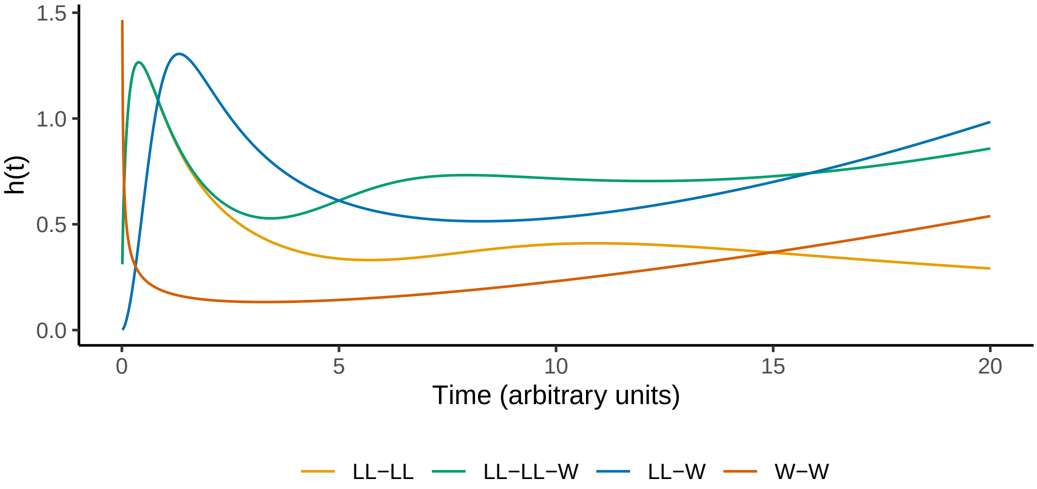

Figure 4.1: Example hazard shapes obtainable by the polyhazard model with combinations of log-logistic (LL) and Weibull (W) latent hazards.

on θ only. In HTA applications, for example, it has become common to only consider the bi-Weibull model (e.g Negrín et al. 2017). This often means that potentially viable candidate models are excluded from the analysis without justification. Alternatively, K, D and γ are reduced to a small set of possible values for which all models are fitted and compared *a posteriori*. Demiris et al. [2015] compare poly-Weibull models with K from 1 to 4 and γ based on the deviance and clinical plausibility, and Benaglia et al. [2015] compare the bi-Weibull and bi-Gompertz model based on visual fit.

Both these approaches rely on the set of candidate models being small enough to fit and interrogate individually, which is very restrictive, as for a fixed maximum number of subhazards, $K_{\max}$, the size of the set of candidate models is given by

$$\sum_{k=1}^{K_{\max}} 2^{pk} \binom{|\mathcal{H}|+k-1}{k}.$$

The result is a model space which is infeasible to explore manually for anything beyond very small $K_{\max}$, $\mathcal{H}$ and p. For the kidney transplant data analysed in Section 4.4.3, taking $K_{\max} = 3$, $|\mathcal{H}| = 2$ and $p = 13$, results in 274,928,246,784 candidate models.

4.2.2 Priors

We now introduce an extended specification of prior for the polyhazard model, which will incorporate uncertainty across each element of (K, D, γ, θ). This induces posterior model weights that can then be used for Bayesian model selection or averaging. The prior, denoted throughout by $\pi_0(\cdot)$, is specified as

$$\pi_0(K,D,\gamma,\theta,\phi) \propto \pi_0(\theta \mid K,D,\gamma,\phi)\pi_0(\gamma \mid K,\phi)\pi_0(\phi)\pi_0(D \mid K)\pi_0(K),$$

where $\phi = (\omega, \sigma_\beta)$ is a vector of hyperparameters to be defined.

First considering $\theta \mid K, D, \gamma, \phi$, we specify

$$\begin{gathered}\log(\nu_k) = \alpha_k \sim \text{Normal}(0, \sigma_\alpha), \quad k = 1, \ldots, K, \\ \beta_{k,0} \sim \text{Normal}(0, \sigma_{\beta_0}), \quad k = 1, \ldots, K.\end{gathered}$$

We place weakly informative priors on each $(\nu_k, \beta_{k,0})$ independent of distribution, in the first case following the reasoning of Demiris et al. [2015]. Crucially the specification of $(\sigma_\alpha, \sigma_{\beta_0})$ will depend both on the scale of the data (years in all the examples in Section 4.4) and the rate of censoring in the data [De Santis et al., 2001]. Specifically, as the rate of censoring increases tighter priors are required in order to regularise long-term hazards. Further justification for, and discussion of, this choice is provided in Section 4.4.1.1.

A contrasting approach is taken for the poly-Weibull model by Demiris et al. [2015] and Benaglia et al. [2015] who place a Uniform$(0,1)$ prior on ν_1. While justifiable for fixed (K,D), the effect of this prior on the posterior is unclear when (K,D) are also being inferred.

We note that in the above specification the same priors are utilised for all sub-hazard functions. When different parametric families are considered this will result in different implied priors for certain sub-hazard quantities (e.g median sub-hazard survival), however, *by design* this information is only weakly informative and therefore we expect the effect on posterior inferences of mean survival to be minimal. In addition, the universal priors used in our case require the specification of single hyperparameter, allowing for the impact of prior assumptions to be easily investigated, while this is not the case if different priors were used for different sub-hazard distributions.

For the remaining linear predictor terms in (4.2) we account for uncertainty in the effect of the covariates on the outcome through the specification of the spike-and-slab prior [Mitchell and Beauchamp, 1988]

$$\pi_0(d\beta_{k,j} \mid \phi) \propto (1-\omega)\delta_0(d\beta_{k,j}) + \omega\tilde{\pi}_0(\beta_{k,j} \mid \sigma_\beta), \tag{4.3}$$
$$k = 1,\dots,K, \quad j = 1,\dots,p,$$

where $\tilde{\pi}_0(\cdot \mid \sigma_\beta)$ is the density of a Normal distribution with mean 0 and δ_0 is a Dirac measure centred at 0. This formulation implies independent Bernoulli(ω) priors for each element of γ_k, resulting in

$$\pi_0(\gamma \mid K) \propto \omega^{\sum_{k,j}\gamma_{k,j}}(1-\omega)^{pK-\sum_{k,j}\gamma_{k,j}},$$

and we extend this to a hierarchical modelling setting through a conjugate Beta prior on ω,

$$\omega \sim \text{Beta}(a,b),$$

as recommended by Kohn et al. [2001]. This is a well established approach, which reduces the influence of prior specification in the context of Bayesian model averaging [Ley and Steel, 2009]. When applied to the COST and kidney transplant data we set $a = b = 4$. Further to this, in order to regularise the effect sizes observed in the linear predictors we utilise a horseshoe, half-Cauchy hyperprior on σ_β,

$$\sigma_\beta \sim \text{Cauchy}_{>0}(0,1),$$

designed to circumvent well known model misspecifcation issues arising from using a fixed σ_β [Polson and Scott, 2012].

Note that in the above formulation ω and σ_β are shared hyperparameters across subhazards encouraging sharing of information between hazards about expected effect sizes, which implies that the induced prior on $|\gamma|$ should be interpreted as a prior on the number of covariates across the model, rather than the number of covariates associated with each individual subhazard.

Each subhazard distribution, D_k, is drawn uniformly from the set of candidate distributions

$$D_k \mid K \sim \text{Uniform}(\mathcal{H}),$$

inducing a multinomial prior on D. If expert knowledge favours certain subhazards being present in the model this can be encoded at this stage.

Finally, prior belief about the number of hazards in the model is represented through a truncated Poisson prior

$$K \sim \text{Poisson}_{>0}(\xi),$$

for fixed ξ. We set $\xi = 2$ defining a weakly informative prior, encoding a soft preference for models with a smaller number of hazards. Any discrete distribution could be used as, for example, there may be expert knowledge which suggests a strong prior belief that $K > 2$, however in practice we find there is rarely justification for $K > 4$ (see e.g [Louzada-Neto, 1999, Demiris et al., 2015]). This is reflected in the choice of ξ which implies $\mathbb{P}(K > 4) = 0.061$ *a priori*. A full expression for the resulting posterior is provided in Appendix B.

4.3 Posterior sampling

The posterior induced by the prior formulation of Section 4.2 presents a challenging target distribution for many of the standard posterior sampling tools of Bayesian inference. Difficulties stem from the varying dimension of the parameter space and changing form of the likelihood due to the priors on (K, D, γ), as well as the geometry of the posterior when (K, D, γ) are fixed. Here, when the data are highly censored, the marginal posteriors of parameters for subhazards (which are influential later in the follow-up period) are often skewed due to partial information from censored observations. Further, subhazards can switch roles in the model. When these subhazards are from the same distribution, exchangeable prior information results in a symmetric, multimodal posterior with $K!$ modes. Role switching, however, can also occur when the subhazards have different distributions, inducing a non-symmetric, multimodal posterior. This is akin to the label switching problem in mixture models (e.g. Jasra et al. [2005]). An example is shown in Appendix B and we discuss this issue further in Section 4.3.5.1. In this Section we develop a bespoke sampling algorithm to handle these challenging posterior features.

Current approaches to posterior computation for fixed (K,D,γ) include a Gibbs sampler implemented in WinBUGS and a Stan implementation of the No-U-Turn Sampler, both for the poly-Weibull model [Demiris et al., 2015, Baio, 2020]. Neither of these approaches naturally extend to the transdimensional case. The former is also susceptible to high levels of auto-correlation, while both can struggle in the presence of multimodality.

The foundation of the method developed in this section is the Zig-Zag sampler [Bierkens et al., 2019] (Section 3.4), an example of a class of novel MCMC methods based on continuous-time Piecewise Deterministic Markov Processes [PDMPs; Fearnhead et al., 2018] (Section 3.3). As outlined in Chapter 3 these processes are non-reversible. As a result, and in contrast to more commonly used reversible MCMC methods, they often exhibit faster convergence and can use ballistic motion to help navigate the challenging geometry of the posterior [Diaconis et al., 2000, Andrieu and Livingstone, 2021]. Further, they are able to use their continuous, piecewise deterministic sample paths to directly sample from spike and slab distributions [Chevallier et al., 2023, Bierkens et al., 2023a] as defined by (4.3). These continuous time dynamics are combined with jump processes for updating (K,D,ϕ), allowing navigation of the full posterior. A summary of the algorithm is provided in Algorithm 2 with a fully detailed Algorithmic presentation provided in Appendix B.

Discrete time MCMC methods in transdimensional settings typically resort to the use of Gibbs sampling for within-model sampling, despite often superior mixing properties of gradient informed samplers. This is due to the sensitivity of these methods to the choice of step-size and mass matrix [Livingstone and Zanella, 2022].

In a large model space, these need to be tuned individually for each sub-model. In contrast, continuous time PDMP samplers require minimal tuning as the step size is replaced by constant velocity terms. A further advantage is the ability to perform transdimensional updates to γ without the specification of a proposal distribution or the need to evaluate likelihoods (Section 4.3.3), in contrast to reversible jump MCMC that requires careful tuning of proposal distributions to ensure modest acceptance rates, and likelihood evaluations at every step [Green, 1995]. Alternative model averaging approaches are possible using, for example, the Bayesian Information Criteria [Volinsky and Raftery, 2000]. These require each sub-model to be computed, however, dramatically increasing computational cost. Further, the approximation to the marginal likelihood is only asymptotically valid in the number of observed events, and is therefore likely to be inaccurate in the highly censored examples we consider. This is supported by recent empirical studies in the context of extrapolating survival curves [Bütepage et al., 2022]. For the remainder of this Section we will use $\pi(\cdot)$ to denote the posterior, conditional on any parameters not given as the argument.

Algorithm 2 Sampling algorithm

```
Initialise $(\theta, v, \gamma, \phi, K, D)$ at $t = 0$.
while $t < t_{\text{end}}$ do
    Sample next event time $t_e \sim$ Exponential$(\Lambda^b + \Lambda^d + \Lambda^s + \Lambda^h)$,.
    Sample $\pi(\theta, v, \gamma \mid \phi, K, D)$ until time $t + t_e$.  ▷ PDMP with sticky dynamics (4.3.1,4.3.3)
    Set $t \mapsto t + t_e$.
    Select event $i$ with probability proportional to $\Lambda^i$.
    if $i = h$ then
        Sample $\omega \sim \pi(\omega \mid \gamma)$  ▷ Gibbs step (4.3.2)
        Sample $\sigma \sim \pi(\sigma \mid \theta)$  ▷ Adaptive Metropolis-within-Gibbs step (4.3.2)
    end if
    if $i \in \{b, d, s\}$ then
        Perform move $i$ with probability $\Lambda^i(t)/\Lambda^i$.  ▷ Birth-death-swap process update for $(K, D)$ (4.3.4)
    end if
end while
```

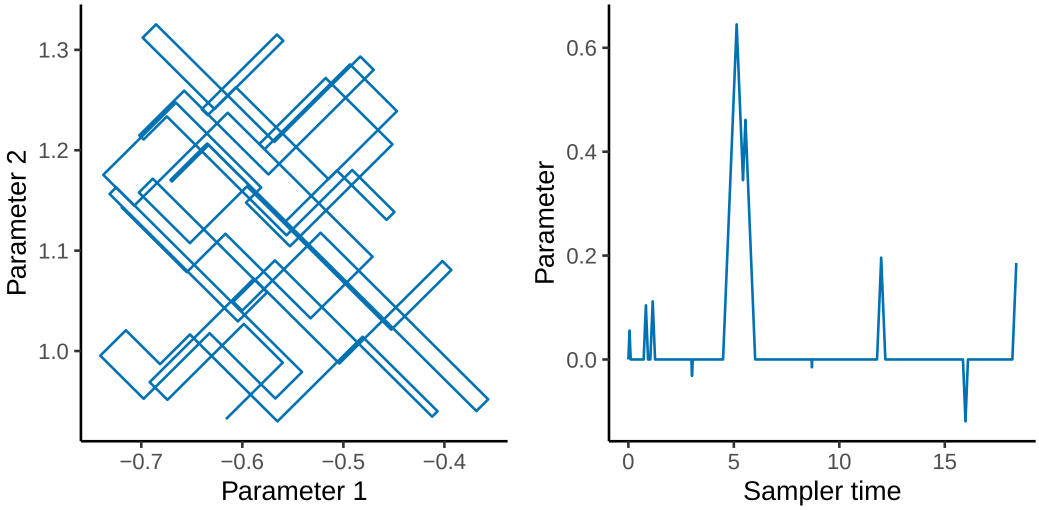

Figure 4.2: Trajectories from the Zig-Zag sampler (left) and variable selection Zig-Zag sampler (right) for arbitrary parameters.

4.3.1 Zig-Zag sampling

Zig-Zag sampling was introduced and reviewed in Chapter 3. In this chapter, the sampler is used to sample the parameters $\theta \in \mathbb{R}^{2K+|\gamma|}$ conditional on fixed (K, D, γ, ϕ) with the dynamics outlined in Section 3.4.

4.3.1.1 Generating the inhomogeneous Poisson process

The efficiency of the Zig-Zag sampler is crucially dependent on the cost of generating event times from an IHPP with rate $\Lambda^B(t)$. This is most commonly achieved via Poisson thinning [Lewis and Shedler, 1979], in which a proposed event time t^* is generated from a dominating Poisson process with rate $\bar{\Lambda}^B(t) > \Lambda^B(t)$, accepted with probability $\Lambda^B(t^*)/\bar{\Lambda}^B(t^*)$; if the proposed move is rejected, the process continues with the same dynamics from time t^*.

While it is possible to derive a tight upper bound analytically in some cases, we know of no such choice of $\bar{\Lambda}^B(t)$ that is suitable for polyhazard models. We therefore numerically bound $\Lambda^B(t)$ on the interval $[t_0, t_0 + t_{\max})$, via an extension of the Automatic Zig-Zag method of Corbella et al. [2022] (Section 3.5.2.1). In the Automatic Zig-Zag approach a constant upper bound for $\Lambda^B(t)$ is found using Brent's method on an interval with fixed length. Costly, repeated gradient evaluations are avoided by performing a monotonicity check after the first iteration, which if passed allows the evaluation of $\Lambda^B(t)$ at one end of the interval to be used as the bounding rate.

We make three modifications to this approach, summarised here with full details provided in Appendix B:

1. In the first iteration we check for monotonicity *and* local convexity. If local convexity holds we use a tighter linear bound.

2. We adaptively set the length of the bounding interval $t_{\max}$ using a modified version of the scheme suggested by Sutton and Fearnhead [2023] in a similar context.

3. We add a constant offset rate Λ_0 to $\bar{\Lambda}^B(t)$ to offset numerical errors and failures in the above checks.

The above modifications allow the sampler to adapt to the changing geometry and curvature of the target induced by the priors on (K,D). Further, if the bounding does fail, this is easily diagnosed by reporting instances when the upper bound is exceeded. These errors can then be investigated or the offset increased.

4.3.2 Updating hyperparameters

The hyperparameters (ω, σ_β) could be sampled directly by the Zig-Zag sampler, but strong posterior dependence between parameters and hyperparameters induced by the hyperprior structure would inhibit sampling efficiency. A more elegant solution is to follow the Gibbs Zig-Zag approach of Sachs et al. [2023], which allows traditional Gibbs updates to be interwoven into the Zig-Zag sampler at exponentially distributed intervals with rate Λ^H. In particular this allows ω to be updated by the closed form full conditional due to the Beta-Binomial prior formulation.

Full conditionals for σ_β are not available in closed form. However, sampling can be performed via adaptive random walk Metropolis steps. To avoid sampling difficulties resulting from the heavy-tails of the Cauchy distribution we utilise the re-parameterisation proposed by Betancourt [2018]

$$\sigma_\beta = z_1\sqrt{z_2}, \quad z_1 \sim \text{Normal}(0,1), \quad z_2 \sim \text{Inv-Gamma}(1/2, 1/2),$$

and determine the step-size and covariance matrix of the random walk Metropolis proposal adaptively using a Robbins-Monro style updating scheme as seen in Algorithm 4 of Andrieu and Thoms [2008].

4.3.3 Zig-Zag sampling for variable selection

Bayesian variable selection is a challenging problem even in standard parametric survival models. Current state-of-the-art approaches involve focusing sampling efforts on the marginal posterior for the variable inclusion indicator $\pi(\gamma)$, where $\gamma \in \{0,1\}^p$. Efficient exploration of the state space, however, requires efficient approximations of the marginal likelihood, which are typically not straightforward for polyhazard models [Liang et al., 2023]. Furthermore, simpler, uninformed schemes such as the add-delete-swap reversible jump sampler of Newcombe et al. [2017] are likely inhibited by poor acceptance rates.

An alternative approach, concurrently developed by Chevallier et al. [2023] and Bierkens et al. [2023a] was reviewed in Section 3.6.2, is to utilise the continuous sample paths of the Zig-Zag sampler to directly sample from the spike and slab posterior induced by (4.3). Here the process sticks to the hyperplane $\{\theta : \beta_{k,j} = 0\}$, corresponding to the spike, whenever it crosses it, by setting the corresponding velocity to 0 and then resetting the velocity after a waiting time, τ_β. Specifying τ_β as the first time of the homogeneous Poisson process

$$\Lambda^V_{k,j}(t) = \frac{\omega}{1-\omega}\tilde{\pi}_0(0 \mid \sigma_\beta),$$

preserves the correct target distribution. We note that the key point of this construction is that the rate of unsticking is given by the posterior ratio between the models with $\gamma_{kj} = 1$ and $\gamma_{kj} = 0$. Since this ratio is being evaluated at $\beta_{k,j} = 0$, the likelihood takes the same value for $\gamma_{kj} = 1$ and $\gamma_{kj} = 0$, and this ratio cancels to a ratio of priors resulting in a homogeneous Poisson process. This approach, therefore, has the dual advantage of being informed by the current state of the process and also being computationally efficient as updates to γ do not require any likelihood or gradient evaluations beyond those required for sampling θ. Example trajectories for this process are given in Figure 4.2 (right).

We extend the work of Chevallier et al. [2023], Bierkens et al. [2023a] by including a hyperprior structure on (ω, σ_β) as detailed in Section 4.2.2. Directly sampling (ω, σ_β) via the Zig-Zag sampler would result in unsticking times given by an inhomogeneous Poisson process requiring additional computational cost to generate. Alternatively by updating (ω, σ_β) with a continuous-time jump process as described in Section 4.3.2, the waiting times remain easy to generate as the first time of a Poisson process with piecewise constant rate.

4.3.4 Birth-death-swap processes

The final sampling ingredient is a birth-death-swap process which is able to update the number of hazards K and the vector of subhazard distributions D in continuous time. Births, deaths and swaps occur at rates given by $\Lambda^b(t)$, $\Lambda^d(t)$ and $\Lambda^s(t)$ respectively, with corresponding proposal distributions for new parameters given by $q_b(u)$, $q_d(u)$ and $q_s(u)$. We note that in addition to allowing exploration of the posterior for (K,D), these transdimensional updates also allow for traversal between modes for fixed (K,D).

4.3.4.1 Birth-death process

To define the birth-death process we require that a detailed balance condition is met

$$\Lambda^b(t)\pi(\theta,D,K)q_b(u) = \Lambda^d(t)\pi(\theta',D',K+1)q_d(u'). \tag{4.4}$$

In similar fashion to reversible jump MCMC [Green, 1995] and birth-death MCMC [Stephens, 2000], we also require that the transformation that maps $(\theta,u) \mapsto (\theta',u')$ is a bijection and that a dimension matching condition is met. To satisfy these conditions birth moves are defined by drawing parameters for a new hazard, u, from the prior conditional on ϕ and selecting the distribution of the new hazard uniformly at random. The reverse move then selects a hazard uniformly at random to remove from the model.

To satisfy (4.4), a simple way of specifying $\Lambda^b(t)$ is via a balancing function [e.g. Zanella, 2020] $g : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $g(a) = a \cdot g(1/a)$, and taking the

Metropolis–Hastings–Green ratio

$$a(t) := \frac{\pi(\theta'_t, v'_t, \phi, D', K+1) q_D(u')}{\pi(\theta_t, v_t, \phi, D, K) q_B(u)},$$

as its argument. The required death move is then defined similarly but with the argument $a(t)^{-1}$. The most commonly used example of this is the Metropolis balancing function, $g_M(a) = \min\{1, a\}$, which is the foundation of the Metropolis-Hastings algorithm. Extended theoretical justification of this approach and further discussion of the role of balancing functions is provided in Appendix B.

An alternative specification, which is the birth-death MCMC approach, is to take $\Lambda^b(t)$ constant and set $\Lambda^d(t) = a^{-1}$. This method fails in our setting as, in contrast to Stephens [2000], θ is being updated in continuous-time. The resulting ratio of posterior densities is then challenging to upper bound, which is needed to apply Poisson thinning. Note, however, that $g_M(a) \leq 1$ and therefore this birth rate is amenable to Poisson thinning. The specification of $g_M(a)$ holds up to a multiplicative constant, Λ^K, which can be used to control the intensity of transdimensional updates.

4.3.4.2 Swap moves

While the birth-death process is sufficient to sample from the correct target distribution, we find that posterior exploration can be significantly improved by the introduction of moves which swap subhazard distributions without updating K. These allow the sampler to move between models with the same number of hazards but different underlying distributions. The improvement in mixing is most noticeable when the posterior for K is concentrated but the posterior for $D \mid K$ is more diffuse, as it avoids the need for transitions through higher or lower order hazard models with low posterior mass.

We define our swap moves, $q_s(\cdot)$ between distributions based on the principle of median matching. Moment matching is a well established approach in defining reversible jump moves [Richardson and Green, 1997] but is not applicable here as moments for some survival distributions are not well defined (e.g the log-logistic distribution with $\nu < 1$).

We propose using median matching as a novel deterministic proposal in which the distribution of a subhazard is swapped from log-logistic to Weibull or vice versa. Considering the case without covariates first, the method keeps the shape parameters of the old and new hazards the same, and then transforms the location parameter to keep the medians the same, using the formula

$$\begin{gathered}\mathrm{Med}_{LL}(\nu,\mu) = \mu = \left(\frac{1}{\mu'}\right)^{\frac{1}{\nu}} (\log 2)^{1/\nu} = \mathrm{Med}_{W}(\nu,\mu'),\\ \implies \mu' = \mu^{-\nu}\log 2.\end{gathered}$$

When including standardised covariates, the interpretation of the above is that the subhazard median is preserved for the average individual. To include covariates in the transformation we apply the mapping $\beta_{LL} \mapsto -\beta_{LL} = \beta_W$. Intuitively it seems reasonable to expect the magnitude of the coefficient effects to be the same when altering the subhazard distribution. However, the interpretation of the effect is inverted, hence the switching of the sign. The median matching proposal can be placed into the balancing function framework outlined previously, although the Metropolis–Hastings–Green ratio now requires a Jacobian to account for the transformation.

Figure 4.3 shows trace plots of posterior model probabilities for samplers using solely the birth-death process; independent swaps; and median matching swaps; based on data containing 100 simulated survival times and a single binary covariate. Note that swap moves and birth-death moves have the same computational cost and, as the overall birth-death-swap rate was set to 10 in each case, the expected computational cost is identical for each sampler. Almost all the posterior mass is placed on models such that $K < 3$, but posterior mass is spread relatively evenly between these models. The median matching moves provide clearly superior convergence in comparison to the alternative processes, where slow convergence is observed for the log-logistic and Weibull models as posterior exploration between these models requires moving through higher order models. Acceptance rates for independent swaps and median match swaps were respectively 6.09% and 44.17%, showing the

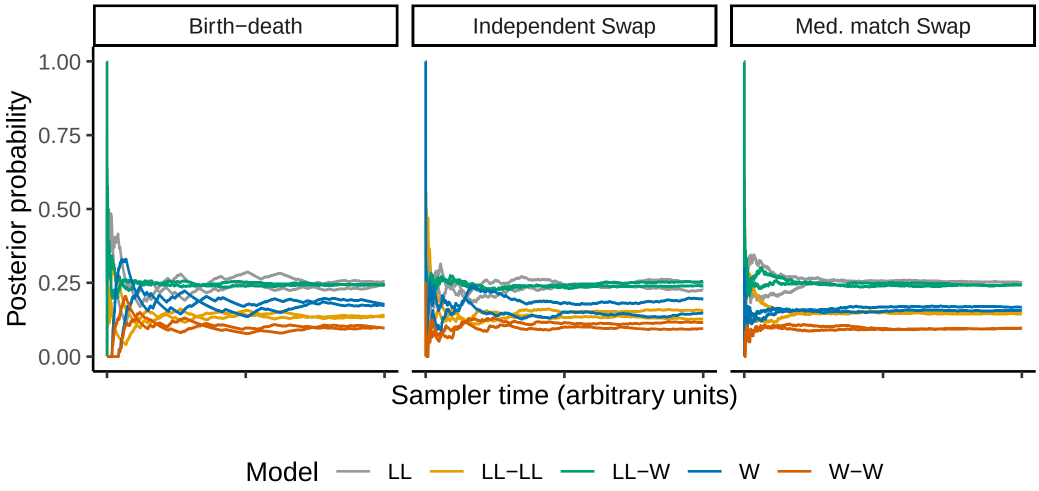

Figure 4.3: Experiment comparing the efficiency of the median matching swap moves to the birth-death moves and independent swap moves, on data simulated from a poly-log-normal-Weibull model with a single covariate. Coloured lines represent different subhazard combinations. Two chains for each method were produced running for 10,000 time units, with reversible jump moves occurring at the same rate. The median matching swap moves provide more stable and efficient mixing in comparison to the alternative methods.

clear superiority of the bespoke moves.

4.3.5 Practical implementation and computational cost

The methodology outlined in this Section requires the generation of multiple event times simultaneously. For computational efficiency this is done via the multinomial trick, whereby a single event time is generated with rate equal to the sum of rates and then a single event is chosen with probability proportional to its rate. Times until deterministic sticking events are also simply tracked and updated when necessary.

The majority of the computational cost for the method outlined in this Section arises from two areas, generating PDMP event times given by the inhomogeneous Poisson process and simulating the birth-death-swap process. Notably, updates to γ incur negligible computational cost, as the only requirements are computing sticking events (which is trivial given constant velocities) and computing unsticking times involving the simulation of exponential random variables, because of the use of the Gibbs Zig-Zag approach for updating hyperparameters. For the method outlined in Section 3.1.1, for each interval over which the event rate is generated we typically require two evaluations of $\nabla U(\theta)$ (as, in the absence of an event, evaluations can be saved between intervals), plus an additional gradient evaluation for each thinning event. We note that this cost should *not* be compared directly to the cost of a single Metropolis-Hastings step, as the trajectory between events in a PDMP typically corresponds to multiple equivalent discrete-time steps. This step is the most computationally costly. Fortunately, however, it is amenable to any methodological improvements in generating the event rate which is currently an area of active research. The birth-death process requires two evaluations of $U(\theta)$ per thinning step. Using the constant bounds derived here this cost is identical to reversible jump MCMC. While we do not believe it is possible to find tighter bounds in the case of polyhazard models, in alternative settings these may exist, meaning the resulting cost is always lower than the discrete time alternative.

An alternative approach is to perform the averaging procedure conditional on each potential K, given the small set of values considered. This will in most cases lead to a dramatically higher computational cost, due to the additional computational resources required for values of K with negligible posterior mass. In contrast, the sampler developed here focuses the computation on a much smaller set of viable models, while maintaining the weights required for averaging over K.

4.3.5.1 MCMC output

As stated previously the Zig-Zag sampler outputs piecewise continuous sample paths. This can be stored either as a skeleton of points which indicate updates to one of (v, K, D, γ), or as samples at exponential times. The effect of this and the rate of drawing samples is analogous to the role of thinning in discrete time MCMC.

For identifiability purposes we place an ordering constraint on the shape parameters of hazards with the same distribution as a post-processing step to sort the MCMC output. This is appropriate in this setting as: a) The quantity of interest, mean survival, is invariant to permutation, and so our inference should not suffer due to the re-labelling issue. b) Kozumi [2004] explored the use of loss functions in the poly-Weibull model and found that the resulting inferences were almost identical to the use of an ordering constraint. We therefore believe that alternative approaches would have little benefit, and that the ordering constraint is sufficient when examining individual subhazards during, for example, model checking.

To summarise, our approach utilises the sticky Zig-Zag sampler to update $\theta, \gamma \mid K, D, \phi$ in continuous-time using gradient information and non-reversible dynamics to ensure efficient exploration of the posterior. This sampler is combined with continuous-time jump processes for updating K, D, ϕ based on conjugate updates, adaptive Metropolis steps and a bespoke birth-death-swap process. The shared continuous-time framework allows for events to be efficiently generated via Poisson thinning and the multinomial trick.

Code for implementing the models developed in this chapter is available at `https://github.com/LkHardcastle/PolyhazardPaper`.

4.4 Real data case studies

In this section we apply the methodological extensions to polyhazard models proposed in the previous two Sections to three real world examples focusing on the effect of prior specification on computation and inference and the non-linear covariate effects produced by polyhazard models.

4.4.1 Lung transplant data

Demiris et al. [2015] used poly-Weibull models to calculate mean survival in lung transplant patients, focusing particularly on differences between patients who received single and double lung transplants. The data contain survival or censoring times of 338 patients, 173 (144 observed) of whom received single lung transplants and 165 (79 observed) of whom received double lung transplants. They focus their analysis on a set of ‘highly likely‘ variations of the poly-Weibull model, as assessed by the mean deviance, all of which indicate small differences in early survival but higher risk for single lung transplant patients in the long-term. This is due to a partial treatment effect, which increases the risk patients experience over a lifetime time horizon.

Although the original data are not publicly available we have constructed a similar dataset by digitising Figure 1 of Demiris et al. [2015]. This was done using

Model	Post. prob.	Mean survival DLT	Mean survival SLT	Difference
Avg. model	—	7.41 (5.26, 12.81)	4.59 (3.85, 5.53)	2.81 (0.01, 8.26)
Original W-W	—	8.78 (6.14, 13.7)	4.96 (4.32, 5.75)	3.83 (1.04, 8.72)
W-L	0.192	7.41 (5.30, 11.85)	4.50 (3.79, 5.34)	2.9 (0.76, 7.37)
W-W	0.010	7.64 (5.52, 11.37)	4.58 (3.74, 5.62)	3.06 (0.86, 6.77)
L-L	0.638	7.40 (5.24, 13.12)	4.61 (3.87, 5.57)	2.78 (0.00, 8.58)
W-W-L	0.016	7.47 (5.22, 11.57)	4.49 (3.75, 5.42)	2.98 (0.54, 7.07)
W-L-L	0.068	7.39 (5.30, 12.50)	4.56 (3.83, 5.49)	2.83 (0.47, 7.94)
L-L-L	0.070	7.37 (5.23, 12.73)	4.59 (3.87, 5.55)	2.77 (0.0, 8.23)

Table 4.1: Model summaries for the averaged (Avg.) model, original (Orig.) model, and sub-models with $> 1\%$ posterior mass. Posterior model probabilities are reported in the second column (Post. prob.). Mean survival estimates are shown for single (SLT) and double (DLT) lung transplant patients along with the expected difference in survival (and relevant 95% credible intervals). Estimates from the original bi-Weibull model are as reported in the original analysis.

the implementation of the method of Guyot et al. [2012] available via the `Survhe` `R` package. We re-analyse these data with the same objective using the extended polyhazard model. We set $\sigma_\alpha = 2$, $K_{\max} = 4$, and adjust the above prior structure by fixing $\sigma_\beta = 5$ and $\omega = 0.5$, which prevents (ω, σ_β) from being essentially nonidentifiable in the presence of a single covariate.

The number of candidate models in this scenario is 128, which, although possible to explore manually, would still be computationally expensive. Our approach has the dual advantage of saving computational cost by focusing on models with high posterior probability, and also providing posterior probabilities for each sub-model. Sampler trace plots are available in Appendix B.

Table 4.1 shows model summaries for the original bi-Weibull model chosen by [Demiris et al., 2015, original W-W], the averaged polyhazard model and all submodels with posterior probability greater than 1%. Notably the original bi-Weibull model receives 1% posterior probability, with the majority of the posterior mass focused on the bi-log-logistic model (63.8%), with reasonable mass on the Weibull-log-logistic model (19.2%) and 15.4% posterior probability shared between three of the three hazard models.

Mean survival estimates for single (SLT) and double (DLT) lung transplant patients are more conservative than those reported in the original analysis. In

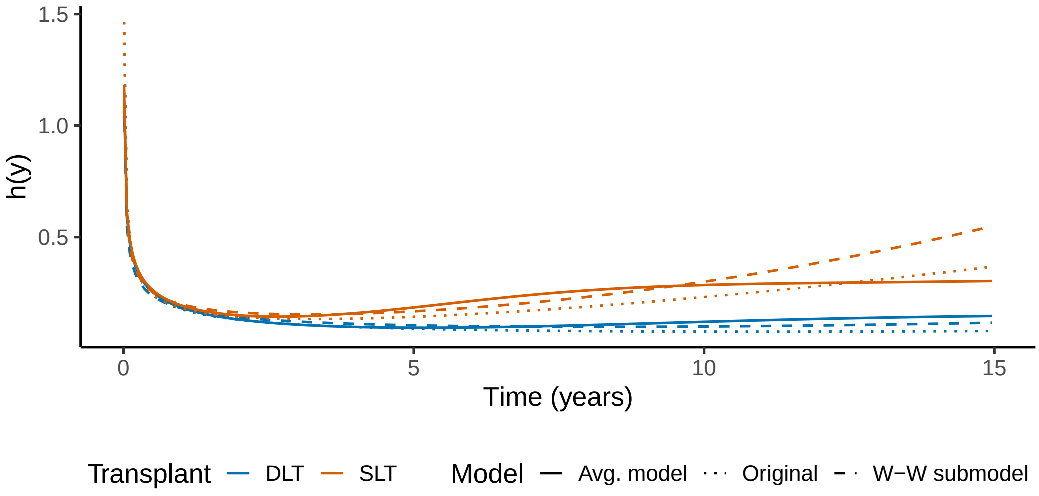

Figure 4.4: Hazards for different models fit to the lung transplant data. The hazards for the model from the original analysis (dash-dot), from the bi-Weibull model in our analysis (dashed) and from the overall hazard from our analysis (solid). These are plotted for DLT (blue) and SLT (red) patients.

particular, the credible interval for the difference in mean survival between the two groups is close to 0 under our analysis. As the reduction in DLT survival is larger than for SLT survival, the analysis using the averaged model reports a smaller difference in expected survival. Although this disparity is driven by a preference for the bi-log-logistic model, the estimates from the bi-Weibull sub-model also suggest more conservative survival estimates and a smaller difference in survival. These differences are discussed in Section 4.4.1.1. Negligible posterior mass was placed on the single hazard models, corroborating the results from the original analysis, which suggested that single hazard models were insufficient.

Figure 4.4 shows hazards for SLT and DLT patients from the overall model, the bi-Weibull model from our analysis and the bi-Weibull model from the original analysis. Notably all three models produce very similar results in the short-term and only differ noticeably after 3 years. This suggests the difference in results reported in Table 4.1 is due to differences in hazards for long-term survivors. Compared to the original analysis the hazard for SLT patients increases faster than in the original analysis after five years explaining the difference in the results reported in Table 4.1.

A key foundation of the original analysis is that the bathtub curve is commonly observed for transplant patients. This can be seen in our example where, although the bi-log-logistic model with highest posterior probability is not a bathtub curve, a decreasing-increasing pattern is observed over a typical patient lifetime, with the overall mixture of polyhazard models ensuring that as $y \to \infty$ we observe $h(y) \to \infty$.

4.4.1.1 Weakly informative priors

The two bi-Weibull models in Table 4.1 report different estimates of difference in mean survival between transplant types. While some of this difference arises from the data digitisation process, this is also due to the use of weakly informative rather than non-informative prior information.

Increasing the standard deviation of the prior for β, increases the posterior estimate for mean survival in both arms and the corresponding credible intervals. This is due to the increasing mass placed on extreme mean survival values by the increasingly non-informative prior. In a single hazard model this is not problematic as the likelihood provides sufficient regularisation of β_{10}. In a K hazard model, however, this behaviour results in the k^{th} subhazard having negligible influence on the likelihood and the model in effect reducing to a $K-1$ hazard model. This has the combined effect of hindering computation, whether via Gibbs sampling or using gradient-based samplers, and impairing the resulting inference. We note that this effect is independent of the prior for γ which has historically been the focus of identifiability in polyhazard models.

This undesirable behaviour can be excluded by the use of weakly informative priors for $\beta_{k,0}$, as outlined in Section 4.2.2. Although tighter than those used previously in the literature, we would argue that these priors are still weakly informative in that they are able to generate data and inferences well beyond the range of plausible values following similar arguments made in Gabry et al. [2019]. As such these priors should be robust to small changes in the choice of σ_{β_0}. We recommend conducting prior sensitivity analysis to ensure this regularisation is sufficient but not unnecessarily influential. In cases with a large number of candidate models, this can be focused on the small subset of models with high posterior probability to preserve computational efficiency.

4.4.2 COST data

We now apply the methodology to a more challenging example – data from the Copenhagen Stroke Study (COST), a prospective, cohort study of stroke survivors in Copenhagen starting in 1991 [Jørgensen, 1996]. The data contains survival times for stroke survivors with 13 relevant covariates. Previous works have used this study to investigate the long-term risks faced by stroke survivors. Kammersgaard et al. [2004] sought to understand the prognosis for very old patients (defined as age ≥ 85), conducting a subanalysis using Cox proportional hazards regression with very old age, stroke severity score and presence of atrial fibrilation as covariates. Andersen et al. [2005] investigated the association between sex and survival outcomes, fitting a Cox proportional hazards model to artificial 1-, 5- and 10- year data cuts to assess the changing effect of sex on survival in the short- and long-term. Similarly, Andersen and Olsen [2011] investigated the interaction between stroke severity, as defined by the stroke severity score, and other prognostic indicators.

In this setting, extrapolation using standard parametric models relies either on simplifying assumptions (e.g proportional hazards) or fitting separate models to each subgroup. Neither approach is ideal. Given the number of covariates, it is unreasonable to assume that proportional hazards hold for each subgroup. Furthermore, fitting separate models for each subgroup will increase uncertainty in extrapolations and provide a poor fit to the data due to small sample sizes. Polyhazard models adapt to smaller sample sizes via assuming homogeneous shape parameters in the sub-hazard functions between sub-groups. However, heterogeneous effects still arise from the model, due to differing covariate effects between sub-hazards, and by averaging across models.

A subset of the data containing survival times, event indicators and 13 covariates, including those discussed previously, for 518 patients is available via the `pec` `R` package [Mogensen et al., 2012]. A complete summary of the dataset is provided in Appendix B.

4.4.2.1 COST results

We fit the model using the full prior structure outlined in Section 4.2.2. Given the larger sample size and lower censoring rate posterior submodel probabilities are relatively concentrated, with 86.44% of the posterior mass given to the bi-log-logistic model, 6.47% to the tri-log-logistic model, 4.96% to the W-L-L model, and 1.71% to the Weibull-log-logistic model. All other models have less than 1% posterior mass.

An advantage of using polyhazard models is the ability to model covariate effects more flexibly than under standard assumptions of proportional hazards or accelerated failure times. This can be seen in Figure 4.5, where we plot the hazard ratios over time for atrial fibrillation, age, sex and stroke score. For continuous covariates these are defined as the hazard ratio between the observed 25% quantile and 75% quantile in the data with all other covariates set to 0, corresponding to their sample mean after standardisation. Notably the averaged model is able to capture a wide variety of flexible hazard ratios. These ratios are compared to the hazard ratios for the simpler Weibull and log-logistic models. Further, we also estimate hazard ratios using M-splines [Jackson, 2023] either combined with a proportional hazards assumption, or using a non-proportional hazards model with partially pooled effects. The parametric models are the established method for survival extrapolation following the initial recommendations of Latimer [2011].

The hazard ratio for age suggests older stroke sufferers have a higher risk of death, which decreases but remains notable for 10 years post-stroke. This aligns with the analysis of Kammersgaard et al. [2004]. The hazard ratio for sex corroborates the findings of Andersen et al. [2005] that women have higher survival than men, although it suggests that the difference in risk decreases in time after an initial peak. A similar pattern is observed for atrial fibrillation. Stroke severity (as measured by stroke score), shows that survivors of less severe strokes are at lower risk of death in the short-term, but that this difference in risk becomes less prevalent in the long-term. In each case the single Weibull and log-logistic hazard ratios are unable to match the increased flexibility of the polyhazard model.

The M-spline models are fit using Stan. Two chains of were generated each consisting of 2000 posterior draws, the first 1000 of which were discarded as burn-in. R-hat values for all parameters were < 1.01. For the M-spline models, the proportional hazards assumption estimates hazard ratios close to the Weibull model. The non-proportional M-spline model hazards appear to have, at least partially, over-fit the data as it exhibits multiple inflection points and sharp changes that are implausible given the study population and associated covariates. We note that over-fitting could also be established using information criteria as outlined in Section 5.4.1. An alternative specification of the M-spline model with fewer knots is presented in Appendix B. The resulting hazard ratios are close to constant, suggesting this over-fitting is due to the choice of knot location. We note that it may be possible to optimally place knots such that the hazard ratios are smooth and capture similar variability to the hazard ratios presented by the polyhazard model. This would be challenging, however, given the number covariates in the data, and the need to place knots manually.

An interesting feature is that the averaged polyhazard model estimates smaller effect sizes for atrial fibrilation and stroke score, at 2.5 and 7.5 years respectively, compared to the M-spline model. This is due to the sub-hazard functions considered in this work that are unable to incorporate sudden peaks in hazard ratios for single covariates. Alternative specifications of the non-proportional spline model (presented in Appendix B along with baseline hazards) suggest these spikes are likely due to over-fitting; however, in analyses where these spikes may be expected, the polyhazard model may not provide sufficient flexibility without the specification of additional sub-hazard forms.

Figure 4.6 plots hazards for each covariate group from the overall models (solid lines) and from the two hazards from the dominating bi-log-logistic model (dashed lines) for the same covariates. Interpreting the first hazard as the immediate post-stroke risk and the second as the longer-term risks, we can understand the influence of different covariates. In particular age increases both the immediate risk post-stroke and the long-term risk, while atrial fibrillation and being male has no immediate effect, but a noticeable long-term effect. Conversely, less severe strokes reduce risk

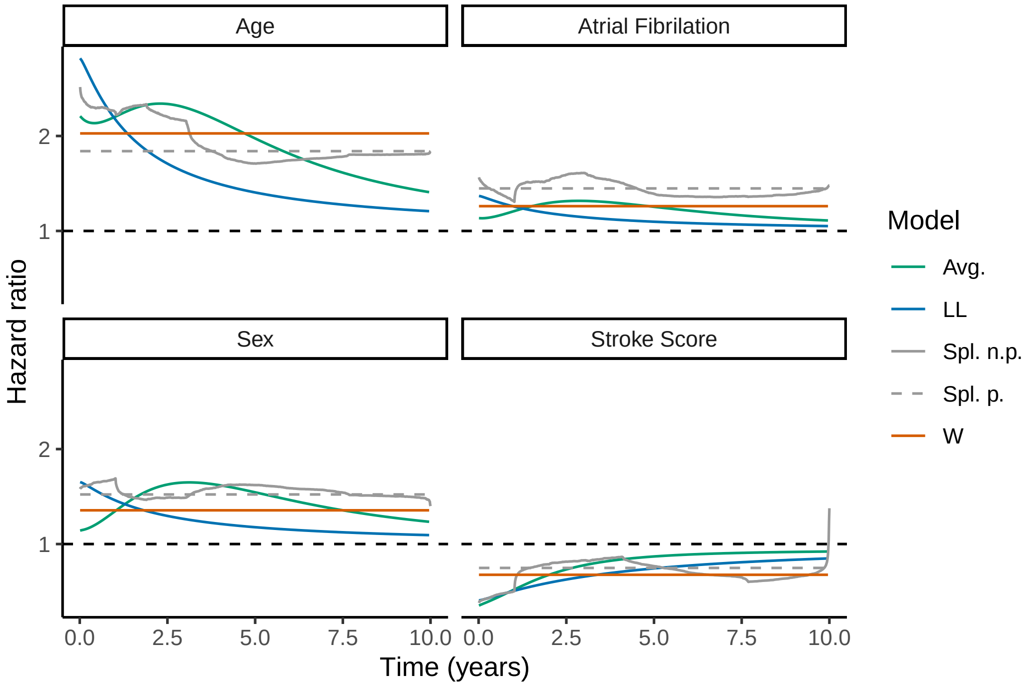

Figure 4.5: Posterior median hazard ratios (HRs) for atrial fibrilation, age, sex and stroke score from the COST dataset. The green line is the HR from the averaged polyhazard model (Avg.), the blue and orange lines are hazard ratios obtained from the simpler log-logistic (LL) and Weibull (W) models. The grey lines are HRs from the proportional (dashed, Spl. p.) and non-proportional (solid, Spl. n.p.) M-spline hazards model. A hazard ratio of 1 is indicated by a black dashed line on each plot.

in the short-term but have a less noticeable effect in the long-term. Figure 4.6 also contains estimates of mean survival and difference in mean survival. In each of the highlighted covariates the 95% credible interval for difference in mean survival does not contain 0, although for atrial fibrillation it coincides with the boundary of the interval, presenting clear evidence that the presence of atrial fibrillation, increasing age and being male lower survival, while less severe strokes improve survival.

4.4.3 Taiwan Kidney Transplant data

We apply our methodology to data on survival times of 3,562 Taiwanese patients following uncomplicated kidney transplantation with the primary objective of understanding the impact of waiting times on mean survival [Chen et al., 2022a]. The data were accessed via Dryad [Chen et al., 2022b]. The original analysis used hazard ratios provided by a Cox regression to understand the impact of transplant waiting

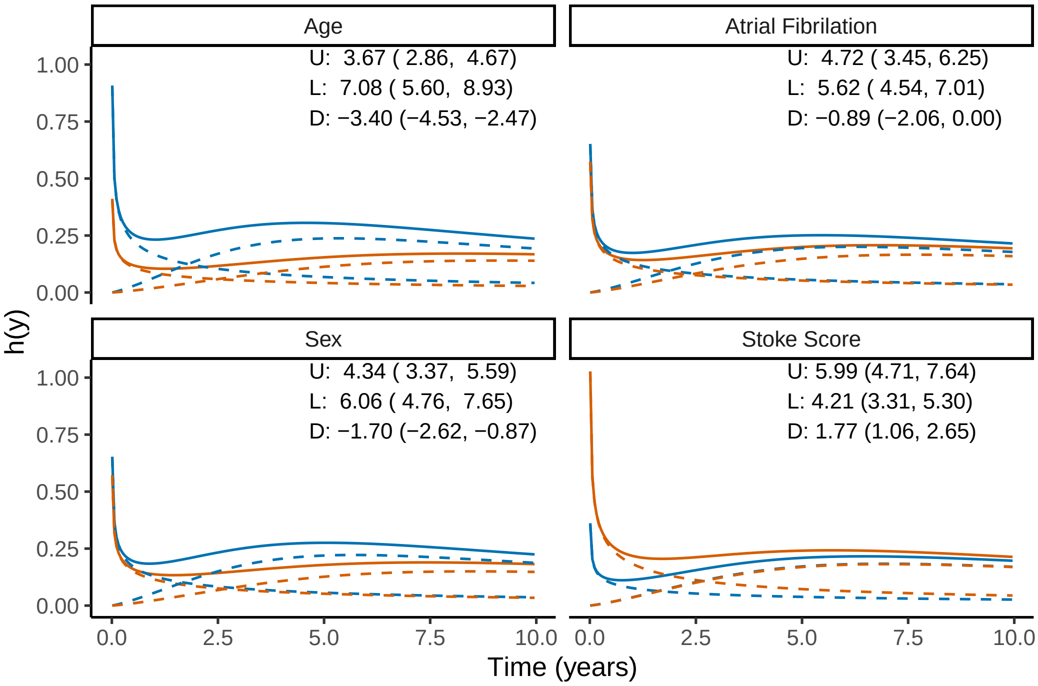

Figure 4.6: Hazards for different values of atrial fibrilation, age, sex and stroke score from the COST dataset. Arial Fibrilation and sex: 0 (orange), 1 (blue). Age and Stroke Score: Sample lower quartile (orange), sample upper quartile (blue). Other covariates are set to 0 representing the average patient. Estimates of mean survival are included in each plot for the blue hazard (U), orange hazard (L) and the difference in mean survival (D).

times on long-term survival. Patients were split into four groups based on wait times (<1 year, 1-3 years, 3-6 years, >6 years). Additional covariates in the data include age at time of transplantation (defined in 10 year blocks), sex, hypertension and Dyslipidemia. The primary challenge with the analysis of these data are the high censoring rates in all age and waiting time groups, with only the oldest patient group (71-80 years) reaching median survival with 41.18% censored, and censoring rates of 89.90% and 92.00% in the youngest two age groups.

Using the prior structure in Section 4.2.2 we fit the averaged model to this data. We make the modification of only considering models with $K < 4$ as, given the high censoring rates, it is unlikely that there is sufficient information in the data to define more than 3 hazards.

In addition, we use a slightly more informative $\text{Normal}(0,1)$ prior for the shape parameters as we otherwise encounter identifiability issues similar to those

Model	W-L	W-W	L-L	W-W-L	W-L-L	W-W-W	L-L-L
Post. prob.	0.321	0.322	0.054	0.140	0.064	0.085	0.010

Table 4.2: Posterior sub-model probabilities for the averaged model applied to the Taiwanese Kidney Transplant dataset restricted to models with posterior mass above 0.005.

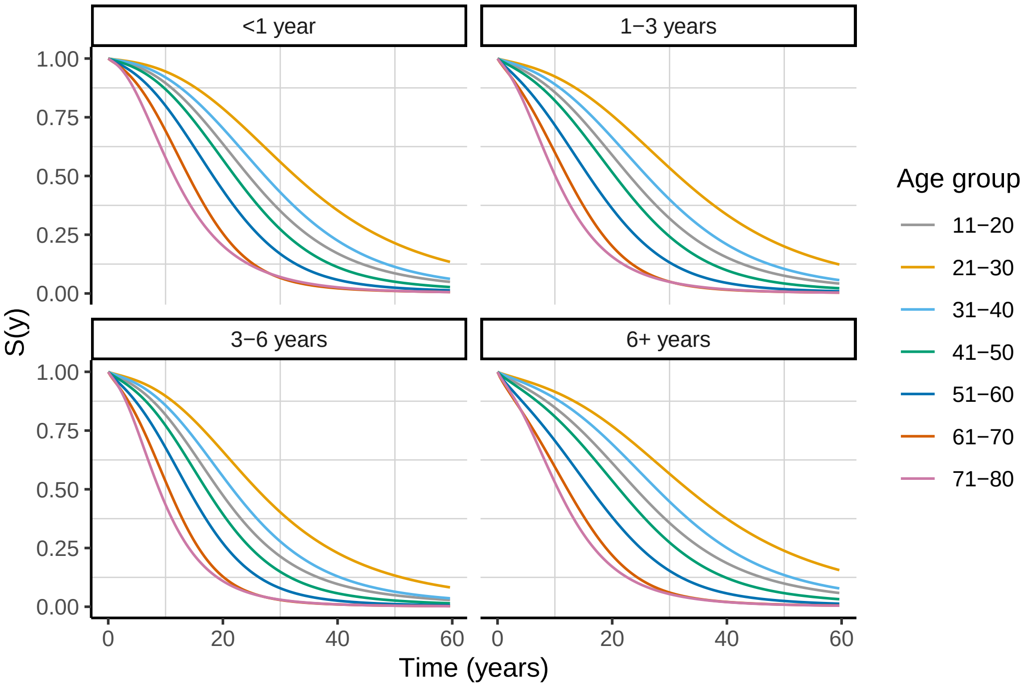

Figure 4.7: Mean survival curves from the averaged model for the Kidney transplant data set stratified by waiting time and age.

highlighted in Section 4.1.1 due to very high censoring rates in certain subgroups. This resulted in estimated survival curves that allowed for unrealistically long survival times. The sampler was run for 20,000 time units, with the rate of reversible jumps or Gibbs moves set to 20. This took 11.76 hours to run.

Posterior model probabilities are reported in Table 4.2. The majority of the posterior mass is shared between the bi-Weibull, Weibull-log-logistic and bi-Weibull-log-logistic models. The posterior is less concentrated than in the previous examples, due to the limited complete data in the sample.

Figure 4.7 shows survival curves for each waiting time group stratified by age. Each curve appears to reach 0 in a reasonable time frame. Of particular note is the apparently non-linear effect of age, with patients in the youngest age group (11-20) having worse survival than patients aged 21-40. This effect is not implausible due to

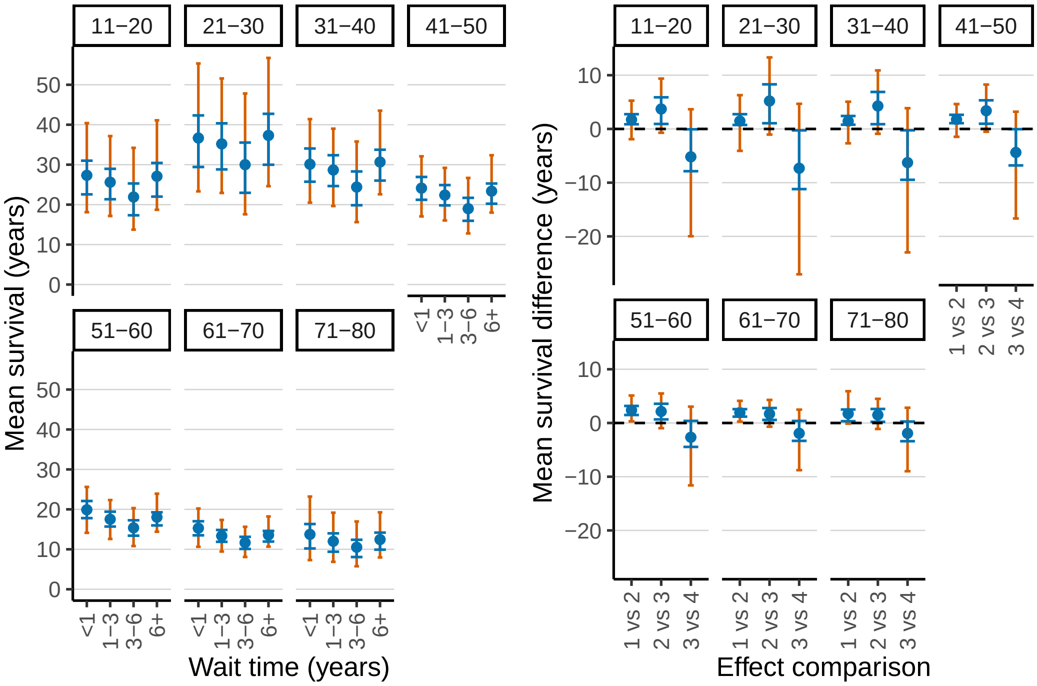

Figure 4.8: (Left) Posterior summaries for mean survival, stratified by age and waiting time. (Right) Posterior summaries for mean survival difference. Results are stratified by age and waiting times encoded as (1: <1 years, 2: 1-3 years, 3: 3-6 years, 4: 6+ years). A dashed line is used to indicate 0 difference. (Both) Posterior mean (solid blue dot), 50% credible interval (blue, larger, error bar), 95% credible intervals (orange, smaller error bar).

the differing reasons for requiring a kidney transplant in different age groups, which are possibly more likely to be due to genetic or hereditary conditions for younger patients, and more likely due to lifestyle factors in older patients. Further in all waiting time groups there are minimal differences in survival between patients in the oldest age groups.

To understand the effect of waiting times on mean survival, posterior estimates of mean survival stratified by age and waiting time group are presented in Figure 4.8 (Left), with posterior means, 75% and 95% credible intervals plotted. Similarly, the effect of moving reducing waiting time by one group is shown in Figure 4.8 (Right). The uncertainty associated with these estimates reduces with age in both cases as the number of censored observations decreases, except for the oldest age group which corresponds to only 17 patients in the sample, resulting in very high uncertainty. Similarly there is high uncertainty in each age group for mean survival in patients who waited more than 6 years for a transplant which propagates through to the estimates of difference in mean survival between patients who waited 6+ years and those who waited 3-6 years. From Figure 4.8 (Right) there is strong evidence to suggest that in the youngest age group and patients over 51 reducing waiting times from 1-3 years to <1 year improves mean survival and similarly reducing wait times from 3-6 years to 1-3 years for patients under 50 improves mean survival. In each age group the lack of information for patients with wait times over 6 years means there is high uncertainty related to the corresponding effect size.

4.5 Discussion

In this work we have developed an extended version of the polyhazard model, using an extended prior specification and novel posterior sampling methodology. This allows for the efficient application of polyhazard models to two motivating data sets for which previous approaches to model selection and computation would have been infeasible. Further, through the use of Bayesian model averaging, we limit the risk of survival extrapolation and mean survival inferences being affected by model misspecification when compared to selecting a single best model.

The findings from the analysis of the digitised lung transplant data from Demiris et al. [2015] suggest that non-informative priors are not appropriate in the polyhazard model setting as they place too much mass on unreasonably large mean survival values. This results in poor posterior estimates and identifiability issues not previously commented on in the literature.

The analysis of the COST dataset shows how the polyhazard model is able to translate epidemiological findings to a cost-effectiveness analysis in the presence of covariates. In particular our approach circumvents issues with current approaches, that either fit models for each subgroup or rely on strong covariate assumptions. The analysis of the kidney transplant data set shows that the extended polyhazard model is able to account for high censoring rates. In particular, being able to combine estimates from many plausible models provides more principled extrapolations in the presence of partial information.

The approach of this chapter is an addition to a number of methods which seek to provide more principled extrapolations by learning the parameters for extrapolation primarily from data towards the end of the observation period. Other examples include the use of M-splines [Jackson, 2023] and dynamic survival models [Kearns et al., 2022]. Compared to the M-spline models, our approach retains a degree of interpretability, and as we show in Section 4.4.2 it is also more stable in the presence of many covariates.

We note that the extended polyhazard model can be easily combined with several methods for improving extrapolations and integrating external information. In particular polyhazard models are the natural form for integrating external information, whether this relates to specific causes of death [Benaglia et al., 2015] or life table data for the wider population [van Oostrum et al., 2021]. Alternatively, the extended polyhazard model could be used to model the observed period and then combined with life-table data via the blended survival approach of Che et al. [2023]. Further simple adjustments to the model could also be made to combine it with other model averaging approaches to extrapolation. For example, the adjusted model averaging approach of Negrín et al. [2017] can be combined with our methods by adjusting posterior weights to account for optimistic and skeptical scenarios.

We briefly outline some obvious extensions to the model presented in Section 4.2. Alternative prior structures for sub-hazard parameters could be considered. For example, as suggested by a reviewer, these could be based on the right-tail behaviour of the sub-hazard functions. This could be particularly beneficial when strong prior information is available about the behaviour of the long-term hazard function. We have assumed that covariates enter the model through location parameters, in line with the recommendations of Latimer [2011]. A linear predictor could also be introduced for shape parameters. This would provide additional flexibility to the model; however, we expect that this would require stronger prior regularisation and increase the computational cost for limited additional flexibility.

We can naturally extend the model to include additional subhazard forms. Although there are many two-parameter survival distributions in the literature, selecting a small number of additional distributions should provide sufficient flexibility to model many datasets. In this context the swap moves from Section 4.3 could be extended to define pairwise transformations between different types of subhazards, or replaced with moment-matching moves where appropriate. Another novel extension would be to introduce the possibility of improper subhazards such that for the corresponding survivor functions

$$S_{k,\theta}(y) \to c > 0, \quad y \to \infty.$$

This would correspond to a cure model for that subhazard, but would need highly informative external information to ensure principled extrapolations. A final extension would be to introduce dependence between hazards, as explored by Tsai et al. [2013].

Finally, we believe we have made important contributions to the applications of PDMP samplers. While these samplers have seen several methodological and theoretical developments, they have seen limited practical application. We hope that their usage in this work can motivate their usage in other contexts. In particular, the bounding method developed in Section 4.3 is not model dependent so could be applied in other contexts, as could the extension of the Gibbs Zig-Zag approach to transdimensional updates. Moreover, we expect the birth-death process to be applicable in wider applications. A natural setting would be in mixture models with an unknown number of components. A further possibility would be the incorporation of global jump moves in fixed-dimensions to improve the mixing of the inherently local PDMP dynamics. In the context of PDMP samplers for variable selection, the combination of variable selection dynamics with the Gibbs Zig-Zag approach for updating hyperparameters efficiently is an important advancement, which can avoid the use of fixed spike and slab weights. Finally, the median matching heuristic developed for the swap moves may be useful in other contexts.

Chapter 5

Diffusion piecewise exponential models

The content of this chapter is based on the paper

> L. Hardcastle, S. Livingstone, and G. Baio. Diffusion piecewise exponential models for survival extrapolation using Piecewise Deterministic Monte Carlo. *arXiv preprint arXiv:2505.05932*, 2025

available on arxiv and currently undergoing the journal review process.

5.1 Introduction

This chapter develops a novel prior specification for the piecewise exponential model allowing for the principled inclusion of prior information to inform extrapolation of the hazard function beyond final event times y_+. Piecewise models are increasingly used to model survival in HTA analyses. This allows for flexible, data-driven inference of hazards during the observation period. In each of the examples discussed in Section 2.2, however, extrapolations are either driven by behaviour of the hazard function inferred during the observation period, external data included in the model, or both, and are often sensitive to modelling assumption, e.g the placement of knots beyond y_+.

In Section 2.2.3 we reviewed several recent advances in the incorporation of explicit prior information to inform extrapolations. We adopt two primary considerations for specifying this prior information: *i)* Assumptions about the form of this prior information should be minimal allowing the analyst maximal flexibility in its specification [Mikkola et al., 2023]. *ii)* The prior should be at least moderately informative during the extrapolation period. We argue, given the often sparse nature of data in these applications, that specification of an informative prior is the *only* way to ensure sensible inference in the extrapolation period.

5.1.1 Our contributions

We introduce the Diffusion Piecewise Exponential Model. The piecewise exponential model is defined by a piecewise constant log-hazard function,

$$\log h(y) = \sum_{j=1}^{J} \alpha_j \mathbb{1}\left(y \in (s_{j-1}, s_j]\right), \tag{5.1}$$

where $\{\alpha_j\}_{j=1}^J$ are a sequence of local log-hazards, and $\{s_j\}_{j=0}^J$ are a sequence of knot locations with $s_0 = 0$. Explicitly, our contributions are as follows.

In Section 5.2, we introduce a novel prior formulation for the sequences $\{\alpha_j\}_{j=1}^J$ and $\{s_j\}_{j=0}^J$, allowing for the principled combination of inferences for the observation period, primarily driven by the data, and inferences for the extrapolation period, primarily driven by prior information. This prior for $\{\alpha_j\}_{j=1}^J$ is given by the discretisation of a diffusion, with drift function used to encode strong prior information about the long-term behaviour of the hazard function. Notably, restrictions on the form of the drift are minimal allowing for a range of prior information to be encoded into the model. The prior for $\{s_j\}_{j=0}^J$ is given by a Poisson point process. This acts as a time change between the underlying diffusion and $\{\alpha_j\}_{j=1}^J$ allowing for intensity in the changes of the hazard during the extrapolation period to be informed by those observed on $(0, y_+)$.

In Section 5.3, we introduce a novel Markov Chain Monte Carlo (MCMC) sampling algorithm based on Piecewise Deterministic Markov Processes (PDMPs). In particular we make use of recent developments in defining and generating these processes to design an efficient sampler that requires minimal user tuning [Bertazzi et al., 2023, Michel et al., 2020]. Further, to handle the transdimensional posterior resulting from the prior on $\{s_j\}_{j=0}^J$, we extend recent results that use PDMPs to sample from posteriors induced by spike and slab priors [Bierkens et al., 2023a, Chevallier et al., 2023] to more general transdimensional posteriors.

In Section 5.4 we demonstrate the flexibility of the model and prior structure, and provide practical guidelines for its use via case studies corresponding to two clinical data sets. We conclude with a discussion in Section 5.5.

5.2 The Diffusion Piecewise Exponential Model

Throughout we adopt the notation and assumptions introduced in Chapter 2, and assume that we observe data, $\mathcal{D} = \{y_i, \delta_i, w_i\}_{i=1}^n$, consisting of n independent survival times, y_i, event indicators, δ_i and covariate vectors $w_i \in \mathbb{R}^p$.

5.2.1 Piecewise exponential models

Piecewise exponential models [Feigl and Zelen, 1965, Ibrahim et al., 2001] are constructed via a piecewise constant log-hazard function (5.1). Covariates can be incorporated into (5.1) by replacing α_j with $\eta_{ij} = \alpha_j + w_i^\top \beta_j$. We refer to α_j as the local baseline log-hazard and $\beta_j \in \mathbb{R}^p$ as a vector of local covariate effects, which can encode a local proportional hazards assumption.

To complete the model specification we require priors for $\{\alpha_j, \beta_j, s_j\}$. Computational convenience is a common motivation for prior selection, primarily through the use of independent, conjugate Gamma priors on $\exp(\alpha_j)$. Another common objective is some degree of smoothing between local hazards, by using either a random-walk prior on α_j [Fahrmeir and Lang, 2001], Markov-Poisson-Gamma priors [Lin et al., 2021] or priors incorporating local and global trend terms [Kearns et al., 2019]. A broader review of prior structures used for survival extrapolation is provided in Section 2.2. The prior we introduce in the following section will contain most of these prior structures as special cases, while providing a weakly informative prior during the observation period.

5.2.2 Discretised Diffusion Priors

To capture prior knowledge about the long-term behaviour of the hazard we assume that the *discrete*-time log-hazard process $\{\alpha_j\}_{j=1}^J$ can be described via a *continuous*-time stochastic process $(\check{\alpha}_{\check{y}})_{\check{y} \geq 0}$ with dynamics governed by the stochastic differential equation

$$\mathrm{d}\check{\alpha}_{\check{y}} = \mu(\check{\alpha}_{\check{y}})\mathrm{d}\check{y} + \mathrm{d}W_{\check{y}}, \quad \check{\alpha}_0 = a_0, \tag{5.2}$$

with drift $\mu(\check{\alpha}_{\check{y}})$, where $(W_{\check{y}})_{\check{y}\geq 0}$ is a standard Brownian motion [Oksendal, 2013]. The random variables $\alpha_1,...,\alpha_J$ are then defined through the relation $\alpha_j := \check{\alpha}_{j\sigma^2}$, where σ^2 is a step size defined later in this section. The primary motivation behind this prior is that information about the evolution of the hazard can be encoded into $\mu(\check{\alpha}_{\check{y}})$. During the observation period, where data are more abundant, this acts as a weakly informative prior with limited impact on the resulting inference. However, as observations become sparser and the hazard is extrapolated beyond y_+, this prior naturally becomes increasingly informative, allowing for long-term inferences to be driven by expert opinion encoded through $\mu(\check{\alpha}_{\check{y}})$.

Previous works have utilised diffusions as priors for hazard functions, including Aalen and Gjessing [2004] in which the hazard function is modelled as a squared Ornstein-Uhlenbeck process and Roberts and Sangali [2010] in which $\mu(\check{\alpha}_{\check{y}})$ is defined such that the resulting diffusion is a stochastic perturbation around a pre-specified hazard function. The challenges of working directly with diffusions are primarily computational. Diffusions of interest rarely have tractable solutions, and therefore need to be finely discretised, increasing computational cost. To combat this, our approach involves a hierarchical formulation in which the numerical discretisation is dictated by the knot locations $\{s_j\}_{j=1}^{J}$, which in turn are sampled from an underlying process, and a prior on the discretisation step-size. This allows for more parsimonious and computationally convenient hazard functions to be specified. More details are given in Section 5.2.3.

5.2.2.1 Example choices of $\mu(\check{\alpha}_{\check{y}})$

We briefly outline some example choices for $\mu(\check{\alpha}_{\check{y}})$, with their behaviour illustrated in Figure 5.1. A first trivial example is to set $\mu(\check{\alpha}_{\check{y}}) = 0$. The underlying diffusion is then a Brownian motion and the discretised version recovers the random walk prior [Fahrmeir and Lang, 2001]. In practice this corresponds to having no expert opinion about the long-term behaviour of the hazard, with credible intervals for the log-hazard increasing in width at a constant rate as $y \to y_\infty$. This assumption will often contradict available prior information, however, and can be improved upon in the following examples.

Stationary distributions: There will often be prior information available about a range of plausible values for the hazard function in the extrapolation period. In our framework this is encoded as a Langevin diffusion, such that $\mu(\breve{\alpha}_{\breve{y}}) = \nabla \log f_\psi(\breve{\alpha}_{\breve{y}})/2$, where $f_\psi(\breve{\alpha}_{\breve{y}})$ is the density of the required stationary distribution for the log-hazard, with parameters ψ. We consider log-Normal and Gamma (equivalently Normal and log-Gamma) stationary distributions for the hazard function (equivalently log-hazard function). The required drifts are then given by

$$\mu_{LN}(\breve{\alpha}_{\breve{y}}) = \frac{1}{\psi_2}(\breve{\alpha}_{\breve{y}} - \psi_1), \quad \mu_G(\breve{\alpha}_{\breve{y}}) = \psi_1 - \psi_2 \exp(\breve{\alpha}_{\breve{y}}). \tag{5.3}$$

Underlying hazards: The stochastic perturbation approach introduced by Roberts and Sangali [2010] can also be incorporated into our framework. In short we suppose that we have access to a known hazard function $h_0(y)$ that quantifies our belief about how the hazard function evolves in the extrapolation period derived, for example, from data from previous clinical trials. A suitable drift function can then be derived by viewing $h_0(y)$ as the solution to an autonomous ordinary differential equation,

$$\frac{dh_0(y)}{dy} = g(h_0(y)), \quad \mu_0(\breve{\alpha}_{\breve{y}}) = g(\breve{\alpha}_{\breve{y}}).$$

In Roberts and Sangali [2010], the absolute value of the diffusion is used to map the diffusion from $\mathbb{R}$ to $\mathbb{R}_{>0}$. In our case $g(\breve{\alpha}_{\breve{y}})$ requires a final change of variables to be transformed to a drift for the log-hazard. As a running example, we consider the case where $h_0(y)$ corresponds to a Gompertz hazard function. This is a natural choice as the Gompertz distribution is often used to model long-term survival in the general population [Thatcher, 1999], and therefore intuitively should provide sensible inferences for the extrapolation period. In our framework this ensures estimates of mean survival in the population of interest are consistent with those seen in the (usually healthier) general population. The corresponding stochastic

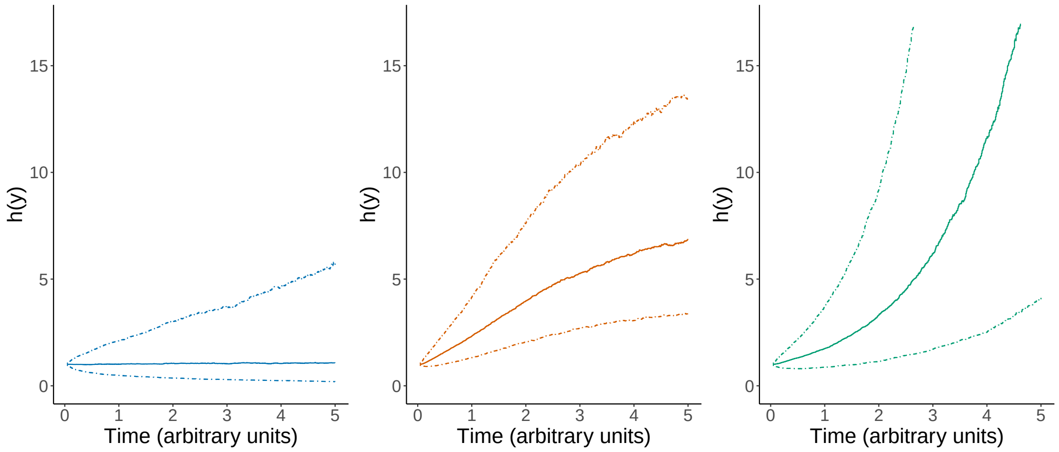

Figure 5.1: Prior simulations for $h(y)$ under different specifications for $\mu(\breve{\alpha}_{\breve{y}})$. (Left) Random Walk prior $\mu(\breve{\alpha}_{\breve{y}}) = 0$. (Centre) Gaussian Langevin prior (5.3). (Right) Gompertz prior dynamics (log-linear drift) (5.4).

differential equation has a linear drift

$$\mu(\breve{\alpha}_{\breve{y}}) = \psi, \tag{5.4}$$

where ψ is the scale parameter of the required Gompertz distribution. The derivation of this quantity is provided in Appendix C.

Time-varying drifts: The above examples have utilised time-homogeneous drift functions. This is, however, not a necessary requirement. In particular, expert opinion on the evolution of the hazard function will often evolve with time. A more flexible class of diffusions can therefore be defined with time-varying drifts $\mu(\breve{\alpha}_{\breve{y}}, y)$. We investigate this possibility further in Section 5.4.2.

5.2.2.2 Discretisation

As noted previously, stochastic differential equations rarely have analytic solutions and therefore implementation requires (5.2) to be discretised. The standard approach is the Euler-Maruyama discretisation [Platen and Bruti-Liberati, 2010]

$$\breve{\alpha}_{(j+1)\sigma^2} = \breve{\alpha}_{j\sigma^2} + \theta_{j+1}, \quad \theta_{j+1} \sim \text{Normal}(\sigma^2 \mu(\breve{\alpha}_{j\sigma^2}), \sigma^2). \tag{5.5}$$

where σ^2 is the step size, for $1 \leq j \leq J$. Note the slight abuse of notation, with $\{\breve{\alpha}_{j\sigma^2}\}_{j=1}^{J}$ now used to denote the discretised version of $(\breve{\alpha}_{\breve{y}})_{\breve{y}\geq 0}$. It is well established that (5.5) can be numerically unstable when $\mu(\breve{\alpha}_j)$ is not globally Lipschitz [Roberts and Tweedie, 1996]. In our application this condition is particularly restrictive and is not satisfied, for example, by the log-Gamma Langevin drift (5.3). More broadly, it is unrealistic to ask practitioners without a mathematical background to carefully check whether the drifts they elicit meet this condition before implementation, and an ideal generalisable prior would not rely on a Lipschitz drift.

To mitigate instabilities when considering non-Lipschitz drifts we utilise a recently introduced scheme based on skew-symmetric innovation densities [Iguchi et al., 2024],

$$\breve{\alpha}_{(j+1)\sigma^2} = \breve{\alpha}_{j\sigma^2} + \theta_{j+1}, \quad f_0(\theta_{j+1} \mid \breve{\alpha}_{j\sigma^2}) \propto \Big(1 + \tanh(\mu(\breve{\alpha}_{j\sigma^2})\theta_{j+1})\Big)\, \phi(\theta_{j+1} \mid \sigma^2). \tag{5.6}$$

Here $\phi(\cdot \mid \sigma^2)$ is the density of a Normal$(0,\sigma^2)$ random variable, and $1 + \tanh(\mu(\breve{\alpha}_{j\sigma^2})\theta_{j+1})$ is a skewing term corresponding to the cumulative distribution function of a logistic distribution evaluated at $\mu(\breve{\alpha}_{j\sigma^2})\theta_{j+1}$.[1] Similarly to the Euler-Maruyama method this approach introduces approximation error that vanishes as $\sigma \to 0$.

Intuitively, while the Euler-Maruyama method *shifts* θ_{j+1} in the direction of the drift, the skew-symmetric scheme *skews* θ_{j+1} in the direction of the drift. This difference is depicted in Figure 5.2 for fixed σ and increasing values of $\mu(\breve{\alpha}_{\breve{y}})$. In Iguchi et al. [2024] the authors show that (5.6) is more robust than (5.5), both to the choice of σ and to non-globally Lipschitz $\mu(\breve{\alpha}_{\breve{y}})$. In both of the above cases we initialise the process at $\breve{\alpha}_0 \sim$ Normal$(0,\sigma_0^2)$. In Section 5.3.5 we will also show that this approach is computationally advantageous, when combined with the prior for $\{s_j\}_{j=1}^{J}$ introduced in the following section.

To complete the specification of the above process, we place an exponential

[1] In fact, this construction is more general in that any CDF of a centred symmetric random variable is sufficient.

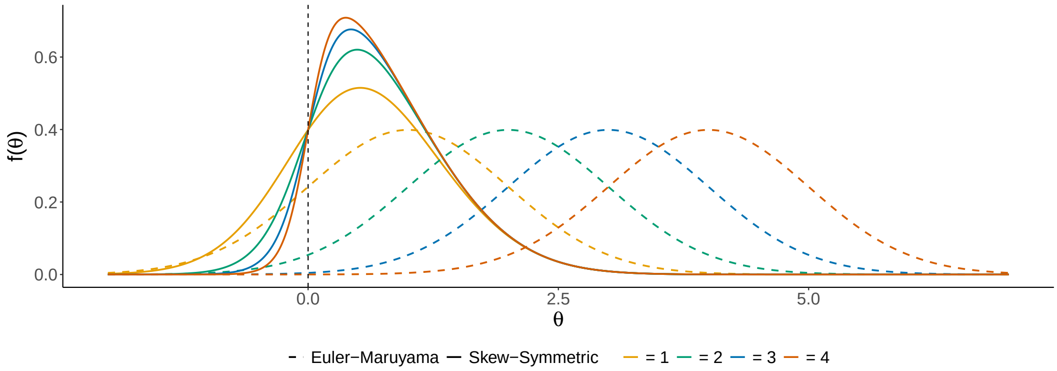

Figure 5.2: Density functions for the innovations θ under the Euler-Maruyama (dashed) and skew-symmetric (solid) schemes for increasing values of $\mu(\check{\alpha}_{\check{y}}) = 1,2,3,4$ and fixed σ^2.

prior on σ,

$$\sigma \sim \text{Exponential}(a),$$

corresponding to a penalised-complexity prior [Simpson et al., 2017]. This prior shrinks the innovation standard deviation towards 0, thus shrinking the overall hazard function towards a single constant value. In all the examples here we set the rate of the exponential prior to $a = 2$. Justification for this choice is provided in Appendix C, however we expect inferences to be generally unaffected for sensible choices of a.

5.2.3 A prior for knot locations

The specification of the model is completed with a prior for the knot locations, $\{s_j\}_{j=1}^J$. The standard approach is for $\{s_j\}_{j=1}^J$ to be fixed a priori, for example at set quantiles of observed event times or at regular intervals [Murray et al., 2016]. The resulting hazard, however, will be sensitive to this specification, particularly in the absence of data during extrapolation period.

We address these issues directly by assuming that $\{s_j\}_{j=1}^J$ arise from a Poisson Point Process with intensity γ on the interval $(0, y_\infty)$, denoted throughout as $\{s_j\}_{j=1}^J \sim \text{PPP}(\gamma, (0, y_\infty))$. This can be expressed equivalently as

$$J \sim \text{Poisson}(y_\infty \gamma), \quad \{s_j\}_{j=1}^J \overset{iid}{\sim} \text{Uniform}(0, y_\infty). \tag{5.7}$$

This prior (and variations) have been considered previously [Chapple et al., 2020, Demarqui et al., 2012]; however, this specification is commonly avoided due to the computational challenges it introduces.

Note that, in contrast to [e.g Roberts and Sangali, 2010], in the above construction the discretisation step size σ^2 is independent of the distance between knots $(s_j - s_{j-1})$. Because of this, the prior for $\log h(y)$ is in fact given by (5.2) through a random time-change defined by (5.7), such that a priori

$$\log h(y) = \alpha_j = \breve{\alpha}_{j\sigma^2}, \quad j = \min\{l : y < s_l\}.$$

This construction can be viewed as first simulating a numerical skeleton $\{\breve{\alpha}_{j\sigma^2}\}_{j=1}^{J}$, via (5.6) and then mapping this to the time-scale of interest, $(0, y_\infty)$, via (5.7).

Under this prior the number of knots, and therefore the flexibility of the hazard function, is directly controlled by γ. The diffusion speeds up when the data require a more volatile hazard and slows down when the hazard is less volatile, adapting to the data without being constrained by the prior. In terms of extrapolation, the advantage of this formulation is that γ determines the speed at which μ dominates the long-term hazard. Intuitively, if the hazard function is more volatile in the observation period we should expect the influence of the data to decay faster in the extrapolation period (with the prior taking over faster). Conversely, if the hazard is less volatile the data should remain informative for longer during the extrapolation period.

We consider two approaches for specifying γ. The first is to consider a set of models for a fixed number of values for γ. These models can then be compared using information criteria. A second, fully Bayesian approach places a prior $\gamma \sim \text{Gamma}(a, b)$, equivalent to a Negative Binomial prior on J. We compare these methods further in Section 5.4.1.

5.2.4 Incorporating covariates

We have focused so far on specification of a prior for the log-baseline hazard and corresponding knots. Both priors extend to the case when covariates are incorporated in the model.

For the underlying diffusion, it suffices to provide a specification for each β_j

process, independently of the diffusion for α_j. As β_j is a covariate effect, a natural process to specify is a Langevin diffusion with Gaussian stationary distribution.[2] Setting the mean to 0 implies the expected long-term treatment effect vanishes as $y \to \infty$. In Section 5.4.2 we show that $\mu(\beta_j, y)$ can be modified to incorporate a waning long-term treatment effect, a common and important assumption in many HTA analyses [Jackson et al., 2017]. Specifying a non-zero mean would imply a long-term proportional average treatment effect, but this would need to be supported by strong clinical opinion. Similarly for the prior for $\{s_j\}_{j=1}^J$, we define a set of knots $\{s_j^k\}$ independently of the set of knots for the baseline log-hazard.

5.3 Posterior sampling

The diffusion piecewise exponential model generates several challenges for commonly used Bayesian inference engines primarily associated with the prior on $\{s_j\}_{j=1}^J$. The resulting posterior is transdimensional for which the standard sampling approach is to use reversible jump MCMC [Green, 1995]. These samplers require the specification of a between-model proposal distribution that must be carefully tuned to achieve modest acceptance rates. This results in a noticeable increase in computational cost due to the additional likelihood evaluations required at each transdimensional step.

Note, in addition to the above, that the fixed $\{s_j\}_{j=1}^J$ model can still present sampling challenges. The potential function, $U(x) := -\log \pi(x)$ is non-Lipschitz, causing instability in gradient-based methods such as the Metropolis Adjusted Langevin Algorithm (MALA) [Roberts and Tweedie, 1996] and Hamiltonian Monte Carlo (HMC) [Livingstone et al., 2019]. Further, in the presence of high censoring rates (which is precisely the scenario we are considering), posteriors of survival models can exhibit high skew, again challenging standard Metropolis-Hastings methods [Hird et al., 2020].

To circumvent these issues we utilise sampling techniques based on continuous time Piecewise Deterministic Markov Processes (PDMPs) [Fearnhead et al., 2024].

[2] Note this is equivalent to specifying an Ornstein-Uhlenbeck prior for β_j.

These processes are non-reversible (e.g. Andrieu and Livingstone [2021]) and use ballistic motion and gradient information to efficiently explore the target distribution. Further, in contrast to MALA and HMC, they have constant velocity and require minimal tuning making them more robust to non-Lipschitz potentials. Recent works have also shown that they are able to sample from transdimensional posteriors induced by spike and slab priors without the need for additional likelihood evaluations or tuning of between-model proposals [Chevallier et al., 2023, Bierkens et al., 2023a]. The key contribution of this Section is to show how these results can be extended to more general transdimensional posteriors. A concise presentation of the algorithm is given in Appendix C. Code to implement these model is available at `https://github.com/LkHardcastle/DiffusionPiecewiseExponential.jl` (see also Section 6.2).

We briefly note that for posterior sampling we use a non-centred parameterisation of the model [Betancourt and Girolami, 2015]

$$\tilde{\theta}_0 = \alpha_0 \quad \tilde{\theta}_j = \sigma^{-1}\theta_j.$$

This avoids strong posterior dependence between α_j's and eliminates funnel-shaped geometry that can arise when simultaneously updating θ and σ (e.g. Betancourt and Girolami [2015]).

5.3.1 The Bouncy Particle Sampler and Forward event chain Monte Carlo

Piecewise Deterministic Monte Carlo methods have emerged as a promising class of non-reversible processes for posterior sampling in challenging Bayesian inference problems. In this work we use a variation of the bouncy particle sampler [Bouchard-Côté et al., 2018], known as Forward Event Chain Monte Carlo [Michel et al., 2020]. These methods were reviewed in Chapter 3.

Given the current sampler time, t, recall that the bouncy particle sampler is defined on an state-space augmented with velocities $z_t = (x_t, v_t) \in \mathbb{R}^d \times \mathbb{S}^{d-1}$, with $x = (\tilde{\theta}, \sigma)$ and $d = J(p+1)+1$. The continuous-time deterministic evolution of z_t

is given by the system of ordinary differential equations

$$\frac{dx_t}{dt} = v_t, \quad \frac{dv_t}{dt} = 0,$$

which results in v_t driving the linear evolution of x_t. This evolution is interrupted by a jump process, with jump times given by the inhomogeneous event rate, $\Lambda^E(t)$, and the transformation of z_t at these times given by the deterministic map Q. For the standard bouncy particle sampler these are defined as

$$\Lambda^E(t) = \max\{0, \langle v_t, \nabla U(x_t)\rangle\} + \Lambda^R,$$
$$Q : (x_t, v_t) \mapsto (x_t, v_t - 2v_t^{\nabla U}).$$

Recall from Section 3.4, $\Lambda^R \in \mathbb{R}_{\geq 0}$ is the refreshment rate, with $\Lambda^R > 0$ required to ensure the process is irreducible, and $v_t^{\nabla U}$ arises from the orthogonal decomposition of v_t with respect to $\nabla U(x_t)$, $v_t = v_t^{\nabla U} + v_t^{\perp}$, with $v_t^{\perp} \perp \nabla U(x_t)$. Events associated with the first term of $\Lambda^E(t)$ only occur when $\langle v, \nabla U(x_t)\rangle\ > 0$, i.e when the process is moving into areas of lower posterior mass, resulting in fast convergence towards areas of high posterior mass, meanwhile the map Q corresponds to a reflection of the velocity off the tangent to the potential.

A well-known drawback of the bouncy particle sampler is that Λ^R requires careful tuning, and that optimal values of Λ^R can result in approximately 78% of events being refreshments [Bertazzi and Bierkens, 2022]. This replaces the ballistic motion of the process with increasingly diffusive dynamics, inhibiting sampling efficiency. To remedy this issue the forward event chain method [Michel et al., 2020] replaces the the deterministic mapping Q with a jump kernel stochastically updating both of $(v_t^{\nabla U}, v_t^{\perp})$. This incorporates refreshment into reflections while ensuring the process targets the correct stationary distribution. In particular, $v_t^{\perp}$ is re-sampled as $\tilde{v}_t^{\perp}$ such that $\langle v_t^{\perp}, \tilde{v}_t^{\perp}\rangle \geq 0$, reducing the diffusivity as compared to full refreshments. Further, the update to $v_t^{\perp}$ need not occur at every event, but can be set to update at the first event after each time given by a homogeneous Poisson process with rate Λ^{R^*}.

To understand the robustness these alterations introduce we can consider the process when Λ^{R^*} is poorly tuned. In the case Λ^{R^*} is set to be too large the refreshment rate is capped by the rate at which reflections occur. Conversely, when Λ^{R^*} is too small, the stochastic updates of $v_t^{\nabla U}$ ameliorate the irreducibility issues observed in the original bouncy particle sampler. In contrast, poor tuning of the refreshment rate Λ^R in the bouncy particle sampler can significantly impact the resulting process. The forward event chain approach has seen uptake in the statistical physics literature, but we believe this to be the first application to an applied Bayesian statistics problem. The specific strategies from Michel et al. [2020] used in this work are outlined in Appendix C.

5.3.2 Generating the process

The deterministic dynamics and jump kernel of the bouncy particle sampler and forward event chain Monte Carlo are simple to generate, but the inhomogeneous Poisson process associated with $\Lambda^E(t)$ is typically more challenging, and an area of active research [Andral and Kamatani, 2024, Corbella et al., 2022, Sutton and Fearnhead, 2023]. The primary method we use to generate this event rate is the splitting schemes approach of [Bertazzi et al., 2023] (Section 3.5.3.1), which alternates between updating the deterministic and event rate processes over a given time step Δt. Without adjustment this scheme introduces a small approximation error into the posterior, which could in principle be corrected for using a non-reversible Metropolis–Hastings filter as described in Bertazzi et al. [2023], though we found this to be unnecessary here. Note that this replaces the continuous time sample paths of the original process with a discrete time approximation.

A second exact scheme we consider is to update σ using conditional Metropolis-within-Gibbs steps at exponential times in the sampler [Sachs et al., 2023]. For the random walk, Gaussian Langevin and Gompertz drifts, the potential for $\tilde{\theta}$ is then convex and the process can be generated exactly by determining event times using a line search [Bouchard-Côté et al., 2018, Example 1].

Of the two schemes we prefer the first. Both algorithms have tuning parameters that are easy to specify, although we find this is marginally easier to do in the former case. Further the sampling efficiency of the second method seems to be inhibited, both by reversible updates for σ, and conditional updating that struggles to explore the geometry of the posterior. Finally, the splitting method allows for a general drift function μ in the diffusion prior to be specified, which is an important goal of this work. Full algorithms for both methods are produced in Appendix C.

5.3.3 Spike and slab PDMPs

Transdimensional posteriors are often induced by priors that are mixtures of continuous and atomic components (commonly referred to as spike and slab priors) of the form

$$\pi_0(d\tilde{\theta}_j) \propto (1-\omega)\delta_0(d\tilde{\theta}_j) + \omega f_0(d\tilde{\theta}_j), \tag{5.8}$$

where δ_0 is a Dirac mass at 0, f_0 is a continuous density and $\omega \in (0,1)$. In all of our examples we set $\omega = 0.5$.

These posteriors can be sampled from directly using the forward event chain sampler, by moving from the continuous component to the atomic component at exactly the point when $\tilde{\theta}$ intersects the hyperplane $\{\tilde{\theta} : \tilde{\theta}_j = 0\}$ [Bierkens et al., 2023a, Chevallier et al., 2023]. Equivalently this can be seen as setting $v_j \mapsto 0$ at this point, with an appropriate renormalisation step when $v \in \mathbb{S}^{d-1}$. For forward event chain Monte Carlo, v_j is then refreshed after an exponential time, τ with

$$\tau \sim \text{Exponential}\left(\frac{\omega}{1-\omega} f_0(0)|\mathcal{J}|\right), \tag{5.9}$$

where $|\mathcal{J}|$ is the Jacobian associated with renormalising v. The remaining terms in this rate are given by a posterior ratio, between the model where $\tilde{\theta}_j$ is on the slab and $\tilde{\theta}_j$ is on the spike. Homogeneity of (5.9) arises due to the transdimensional updates occurring at a point where the likelihoods in both models are equivalent and therefore cancel, along with the majority of prior terms, simplifying this posterior ratio. When multiple components are considered simultaneously the next unsticking time is simply given by summing together the unsticking rates, with the component to update then selected uniformly at random.

The construction of (5.9) is remarkable as, in contrast to most reversible jump MCMC methods [Green, 1995], transdimensional updates do not require either the specification of tuning parameters or likelihood evaluations. Following the terminology of Bierkens et al. [2023a] we will refer to these dynamics as sticky PDMP dynamics from this point forward.

5.3.4 Sticky PDMPs for knot selection

While the computational efficiency and lack of tuning parameters in the above construction is appealing they have not yet been applied to transdimensional posteriors beyond those induced by spike and slab priors. We now show that the above dynamics can be extended to sampling the location of knots under a Poisson process prior following a two step procedure: *i)* Given a fixed set of candidate knots, use sticky PDMP dynamics to update which knots are active in the model and which are inactive. *ii)* Use a Gibbs step to update the set of candidate knots solely through updating the set of inactive knots.

5.3.4.1 Updating given fixed candidate knot locations

We begin by considering the simpler case in which a fixed a set of unique candidate knot locations $\{m_i\}_{i=1}^{M}$ with scaled innovation parameters $\tilde{\theta} \in \mathbb{R}^M$ are chosen. We will assume that this set is composed of a set of active knots $\{s_j\}_{j=1}^{J}$ such that $m_i \in \{s_j\}_{j=1}^{J}$ implies that $\tilde{\theta}_i \neq 0$ almost surely, and a set of inactive knots, $\{r_j\}_{j=1}^{M-J}$, such that $m_i \in \{r_j\}_{j=1}^{M-J}$ implies $\tilde{\theta}_i = 0$. Assuming a priori that $\mathbb{P}(m_i \in \{s_j\}_{j=1}^{J}) = \omega$ is equivalent to defining a spike and slab prior introduced in (5.8) independently for each $\tilde{\theta}_i$, with $f_0(d\tilde{\theta}_i)$ corresponding to prior density for $\tilde{\theta}_i$ induced by (5.6).

Sticky PDMP dynamics can then be directly applied without modification, with moves onto and off the spike updating membership of $\{s_j\}_{j=1}^{J}$ and $\{r_j\}_{j=1}^{M-J}$. When viewed in α-space, the resulting dynamics split and merge the trajectory of neighbouring α's in continuous time, showcasing a natural connection to split-merge reversible jump moves used in several settings [Brooks et al., 2003]. These dynamics are illustrated in Figure 5.3.

Note, these dynamics cannot be immediately extended to the model with a Poisson process prior, as the set of candidate knots is uncountable. Each element of $\{r_j\}$ has an associated Poisson clock with rate defined in (5.9), and therefore the

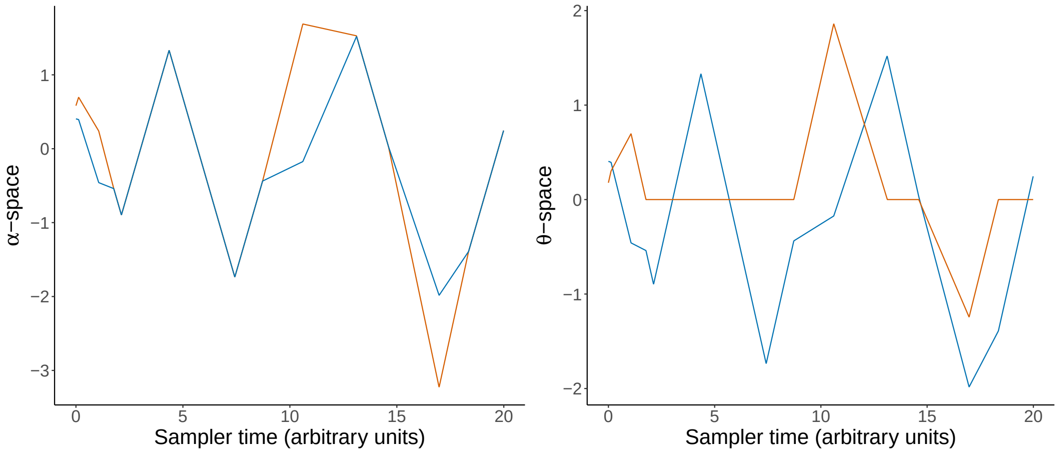

Figure 5.3: Trajectories for the PDMP sampler for knot selection viewed on α-space (left) and θ-space (right). (Blue) First coordinate, (Orange) second coordinate of both α and θ.

resulting combined unsticking rate will be infinite unless only a countable number of them are non-zero.

5.3.4.2 Updating the set of candidate knots

The second part of this procedure circumvents this explosivity, by initialising the sampler with a finite set of candidate knots that is then regularly updated via a Gibbs step. To define this update, we first let the intensity $\gamma := \omega\Gamma$ with ω defined in equation (5.8) and $\Gamma > 0$. This does not alter the prior introduced in Section 5.2.3. Under this specification the set of (now random) candidate knot locations $\{m_i\}_{i=1}^{M} \sim \text{PPP}(\Gamma, (0, y_+))$, and $\{s_j\}_{j=1}^{J}$ can be viewed as a thinned version of this process with thinning probability ω. As in the previous section, this is equivalent to defining a spike and slab prior (5.8) independently for each $\tilde{\theta}_i$.

A valid and computationally efficient Gibbs step then proceeds by re-sampling $\{r_j\}_{j=1}^{M-J} \sim \text{PPP}((1-\omega)\Gamma, (0, y_+))$. As these knots are inactive, updating their location does not alter the value of the likelihood and they can therefore be drawn directly from the prior. Conversely, if $\{m_i\}_{i=1}^{M}$ was updated this would require a Metropolis correction with corresponding likelihood evaluations. These Gibbs updates can occur at fixed or exponentially distributed times in the sampler [Sachs et al., 2023]. Further, if a hyperprior has been placed on γ, this can also be updated at these times. These steps are shown in full in the algorithms presented in Appendix C.

5.3.5 Mixing time of the process

The efficiency of the above process is dependent on the value of $f_0(0)$, i.e the continuous part of the prior for $\tilde{\theta}_j$ evaluated at 0. This can be seen through (5.9), as smaller values of $f_0(0)$ result in longer sticking times at 0, requiring the process to be run for longer to obtain the same estimates. We can in fact formalise this intuition to compare the mixing times of the process under different parameterisations of the underlying diffusion.

Proposition 1. *Given a fixed set of candidate knots, let τ_0^S (respectively τ_0^{EM}) be the recurrence time to the null model (i.e the model when all knots are inactive) under the skew-symmetric parameterisation (respectively the Euler-Maruyama parameterisation). Then*

$$\mathbb{E}[\tau_0^S] \leq \mathbb{E}[\tau_0^{EM}]. \tag{5.10}$$

Proof. Following [Bierkens et al., 2023a, Remark 2.4], as the process is invariant the expected recurrence time to the null model is inversely proportional to the expected occupation time in the null model,

$$\mathbb{E}[\tau_0] \propto \left(\frac{\omega}{1-\omega} f_0(0 \mid \tilde{\theta}, \sigma) |\mathcal{J}|\right)^{-M} \propto f_0(0 \mid \tilde{\theta}, \sigma)^{-M}. \tag{5.11}$$

Then note that under the skew-symmetric parameterisation, $f_0^S(0 \mid \tilde{\theta}, \sigma)$ is the density of a standard Normal distribution evaluated at 0, as the skewing term equals one when $\tilde{\theta}_j = 0$. Further under the Euler-Maruyama scheme the density, $f_0^{EM}(\cdot \mid \tilde{\theta}, \sigma)$ is that of a Normal$(\sigma^2 \mu(\alpha_{j-1}), 1)$ distribution. Therefore $f_0^{EM}(0 \mid \tilde{\theta}, \sigma) \leq f_0^S(0 \mid \tilde{\theta}, \sigma)$ and (5.10) follows directly. □

A direct consequence of Proposition 1 is that we can expect faster mixing times under the skew-symmetric parameterisation. We support this argument empirically by examining the performance of the sampler under each parameterisation for identical $\mu(\alpha_j)$. In particular we consider mean 0 Gaussian Langevin diffusions with standard deviations $\phi_2 = 2$ and $\phi_2 = 0.2$, for increasing values of fixed σ. Note that the resulting $\mu(\alpha_j)$ is Lipschitz and the approximation of the drift should be stable under both parameterisations.

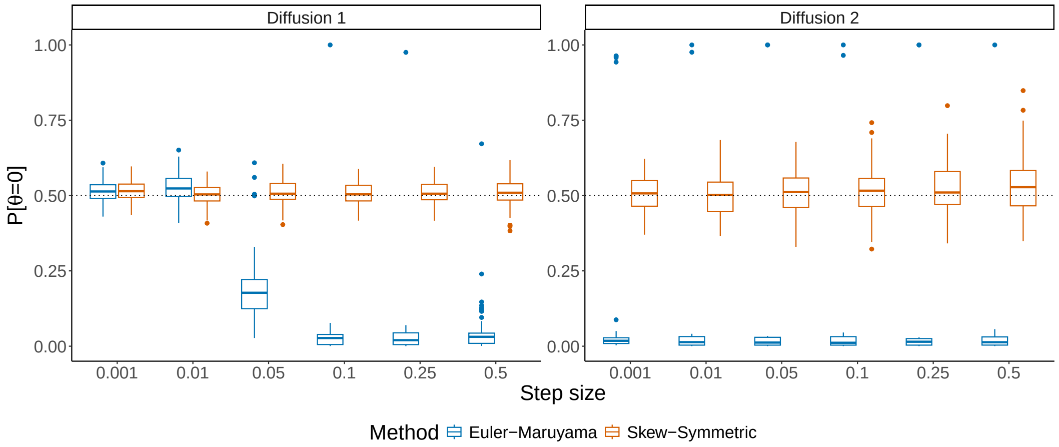

Figure 5.4: Comparison of efficiency of the PDMP sampler under the skew-symmetric and Euler-Maruyama parameterisations for different fixed values of σ. (Diffusion 1) $\mu(\alpha_j) = \alpha_j/2^2$, (Diffusion 2) $\mu(\alpha_j) = \alpha_j/0.2^2$. The dotted line indicates the true value $\omega = \mathbb{P}(\theta = 0) = 0.5$.

Figure 5.4 shows the resulting estimates of ω. For the case $\phi_2 = 2$, $\mu(\alpha)$ is relatively flat and so both parameterisations provide good sampling for small values of σ, however the Euler-Maruyama parameterisation becomes increasingly unstable as σ increases. For $\phi_2 = 0.2$ the Euler-Maruyama parameterisation remains stuck either on or off indicating noticeably slower mixing. There is larger variance in the estimates provided by the skew-symmetric parameterisation for larger values of σ, but the sampling is clearly improved compared to the Euler-Maruyama parameterisation.

5.3.6 Generating extrapolations

We note that the sampling methodology presented in this section has been designed to sample from parameters corresponding to the observed data period. Sampling parameters for the extrapolation period is easily handled using the skew-symmetric scheme directly along with posterior samples for $(\alpha_M, \sigma, \gamma)$. This direct sampling is more efficient than using the PDMP in the absence of data, and helps mitigates strong posterior dependencies that arise over extended time horizons.

Discretising the diffusion does introduce a first order bias that vanishes as $\sigma \to 0$. To reduce the bias in the extrapolation period, the (γ, σ) can be rescaled during this procedure. Full details are provided in Appendix C.

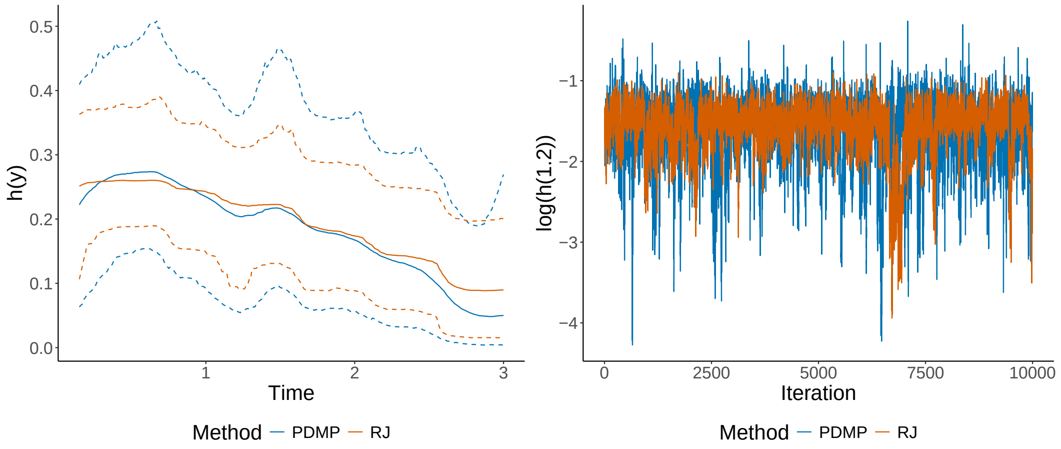

Figure 5.5: (Left) Inferred hazards under the reversible jump sampler (Orange) and the PDMP sampler (blue) with median hazards (solid) and 95% credible intervals (dashed) reported. (Right) Trace plots for $\log(h(1.2))$ under the two samplers. Note that the reversible jump sampler is struggling to fully explore the tails of the hazard function.

5.3.7 Comparison to reversible jump

To understand the efficiency of the developed methodology we compare the sampler to a comparable reversible jump scheme consisting of alternating an update for $\{s_j\}_{j=1}^J$ by either adding or removing a knot at each iteration with a random walk Metropolis update for $\tilde{\theta}$. We prefer the Random Walk to other choices of proposal kernel due to its robustness to tuning parameters that can be challenging to tune correctly within transdimensional sampling algorithms [Livingstone and Zanella, 2022].

We fit the diffusion piecewise exponential model to the Colon data set analysed in Section 5.4.1 using both the introduced PDMP sampler and the reversible jump sampler run for the same computational budget. Plots of the resulting hazard functions are shown in Figure 5.5, along with trace plots for $h(1.2)$. Notably, the reversible jump sampler is unable to sufficiently explore the tails of the posterior for the hazard function, and therefore underestimates posterior uncertainty.

It is natural to wonder how the design choices we have made affect the efficiency of the reversible jump sampler. In Appendix C we show results for alternative values of tuning parameter in the reversible jump proposal and full details of the algorithm. Further, we provide a comparison to the sampler introduced by Chapple et al. [2020]

for a similar model specification.

5.4 Applications

5.4.1 Colon Cancer data

Our first illustrative application is to a dataset consisting of survival times from 191 colon cancer patients, of whom 22 were censored before 3 years and 104 were administratively censored at 3 years. This data is available via the `survextrap` R package [Jackson, 2023]. To implement the model the practitioner is required to specify two quantities, the hyperprior (or fixed value) for γ, and the drift $\mu(\alpha_j)$.

5.4.1.1 Specifying γ

We consider both methods for the specification of γ highlighted in Section 5.2.2, namely *i)* Selecting an optimal value of γ based on information criteria. *ii)* Placing a hyperprior on γ.

Information criteria are commonly used when selecting a model for survival extrapolation [Baio, 2020]. These results must be combined with an assessment of the plausibility of extrapolated hazards, however, as information criteria only assess goodness-of-fit within the observation period, providing no guarantees for the quality of extrapolations. As a result, analysts are often faced with the choice of either selecting a model that fits the observed data poorly or a model that exhibits unrealistic long-term behaviour. We note that the use of a more flexible model is not an automatic remedy to this issue. If no additional information is provided to guarantee the quality of extrapolations, then the above scenario will always be a possibility.

The diffusion piecewise exponential model avoids this trade-off through the specification of μ, breaking the dependence between the model fit to the observation period and the limiting behaviour of the hazard function. As such information criteria can be used to select γ. The practical impact of this choice beyond y_+, is to control the rate at which the influence of the data in the observation period decays. Intuitively, if the observed hazard is more volatile, we can expect this influence to decay faster in comparison to a more stable hazard function.

In this work we use the leave-one-out information criteria estimated using Pareto-smoothed importance sampling [Vehtari et al., 2017] due to its stability properties compared to alternative criteria. The same criteria are used by Jackson [2023] to determine the number and location of knots when using M-splines. We believe that our approach is simpler, however, as it requires only the selection of a single parameter.

We find the approximation to the leave-one-out cross-validation score does not always sufficiently penalise overly complex models, resulting in implausibly shaped hazard functions. We therefore suggest that this score should be minimised, while also ensuring the shape of the hazard function remains plausible. The optimal value for the colon cancer data, using $\mu(\alpha_j) = 0$, is $\gamma = 3.5$. The full results of this procedure are available in Appendix C.

For the second approach, to allow for consistent comparisons with the above procedure we specify

$$\gamma \sim \text{Gamma}(3.5, 1).$$

As noted previously we can view this as a Negative Binomial prior. Previous applications of Negative Binomial priors in similar contexts have found they are less informative than Poisson priors for J [Sharef et al., 2010]. In practice we find that while the Negative Binomial prior is robust to the specification of the overdispersion parameter, in the sense that posterior inferences are minimally affected, the specified prior mean can still be influential. Therefore in practice modelling needs to be coupled with sensitivity analysis to understand the influence of this choice.

5.4.1.2 Specifying $\mu(\alpha_j)$

Specification of $\mu(\alpha_j)$ drives the behaviour of the hazard function during the extrapolation period, and should be elicited using expert opinion or external data on the long-term behaviour of the hazard. In particular it should *not* be selected using information criteria, as this only measures predictive ability during the observation period.

We consider various specifications of the time-homogeneous drifts outlined

Model	$\mathbb{E}[Y]$ on $(0, y_+)$	$\mathbb{E}[Y]$ on $(0, y_\infty)$
Random Walk (Poisson)	2.19 (2.01, 2.36)	4.73 (3.14, 6.09)
Random Walk (Neg. Binomial)	2.21 (2.02, 2.38)	4.67 (3.21, 6.06)
Log-Normal stationary (Poisson)	2.19 (1.99, 2.36)	3.80 (3.22, 4.49)
Log-Normal stationary (Neg. Binomial)	2.20 (1.99, 2.39)	3.97 (3.26, 4.82)
Gamma stationary (Poisson)	2.19 (2.01, 2.36)	4.31 (3.29, 5.52)
Gamma stationary (Neg. Binomial)	2.21 (2.01, 2.39)	4.39 (3.34, 5.57)
Gompertz (Poisson)	2.19 (2.02, 2.36)	4.43 (2.94, 5.89)
Gompertz (Neg. Binomial)	2.20 (2.01, 2.38)	4.44 (3.03, 5.84)
Log-normal parametric	2.18 (2.03, 2.32)	5.79 (4.71, 6.86)
Independent piecewise exponential	2.27 (2.11, 2.42)	5.37 (4.10, 7.34)
M-spline (final knot = 5)	2.25 (2.10, 2.40)	6.89 (4.65, 9.10)
M-spline (final knot = 10)	2.25 (2.10, 2.41)	6.57 (4.00, 8.81)
M-spline (final knot = 15)	2.26 (2.10, 2.40)	6.45 (3.65, 8.56)

Table 5.1: Mean survival estimates for the colon cancer data for the observation period and total window of interest with 95% credible intervals. (Top) Estimates under varying specifications of $\mu(\alpha_j)$ for both the Poisson and Negative Binomial priors. (Bottom) Estimates from the log-normal standard parametric model, an independent piecewise exponential model, and M-spline hazard model.

in Section 5.2.2. The use of Langevin diffusions with log-Gamma or Gaussian stationary distributions encodes an assumption that the expected hazard function will be constant as $y \to \infty$. To illustrate the method in the following examples we use Langevin diffusions with Normal$(\log(0.29), 0.4)$ and log-Gamma$(2,7)$ stationary distributions for the log-hazard, and the Gompertz diffusion (5.4) with $\psi = 0.3$. For each model generating two chains of 10,000 samples including burn-in took approximately 45 seconds. Examples of how to derive these prior drifts, full computational and modelling details are provided in Appendix C.

5.4.1.3 Results

Mean survival estimates for each specification of $\mu(\alpha_j)$ under the Poisson and Negative Binomial priors are presented in Table 5.1, with corresponding hazard functions in Figure 5.6. Posterior mean survival estimates for the observation period are almost identical under each specification of γ and $\mu(\alpha_j)$, with inferences driven by the observed data. Similarly, $\mu(\alpha_j)$ has minimal influence on the hazard functions in the observation period, although notably the Negative Binomial prior provides a smoother fit than the corresponding Poisson prior.

In contrast, mean survival estimates in the extrapolation period are highly reliant on the information encoded in $\mu(\alpha_j)$. In particular the credible intervals under the Random Walk and Gompertz drifts are larger than those under the Log-Normal and Gamma Langevin drifts. This difference in behaviour can also been seen in the hazard functions, where the credible intervals are noticeably larger under the former prior specifications. In general, although not for the random walk prior, the Negative Binomial specification results in larger estimates of mean survival. This is due to the smoother hazard function inferred for the observation period, slowing the speed of the underlying diffusion and the corresponding rate that the influence of the prior grows. Note that this behaviour is because the prior information encodes a typically higher hazard value than that observed at the end of the observation period. If the converse were true, then the Negative Binomial prior would result in more conservative estimates of mean survival. Finally, Figure 5.6 shows that the Gompertz drift results in large credible intervals (larger in fact than the random walk prior), suggesting this prior does not encode much information in the extrapolation period. This is due to the exponential form of the Gompertz hazard function. As such, extrapolations are highly sensitive to the hazard observed at the end of observation period. We explore improvements to this specification in Section 5.4.2.

5.4.1.4 Alternative approaches

To contextualise the inferences obtained under the diffusion piecewise exponential model, we consider three alternative methods: *i)* The standard approach of selecting a two-parameter parametric model using information criteria (in this case the log-Normal parametric model) [Latimer, 2011, Baio, 2020]. *ii)* The piecewise exponential model with independent priors, where the hazard at the end of the observation period is taken as the hazard for the extrapolation period [Cooney and White, 2023a]. *iii)* Modelling the hazard using M-splines [Jackson, 2023]. In particular, as extrapolations are based on the placement of a final knot on (y_+, y_∞), we consider inferences under three different knot locations. Full implementation details and additional analysis are provided in Appendix C. Mean survival estimates are reported in Table 5.1.

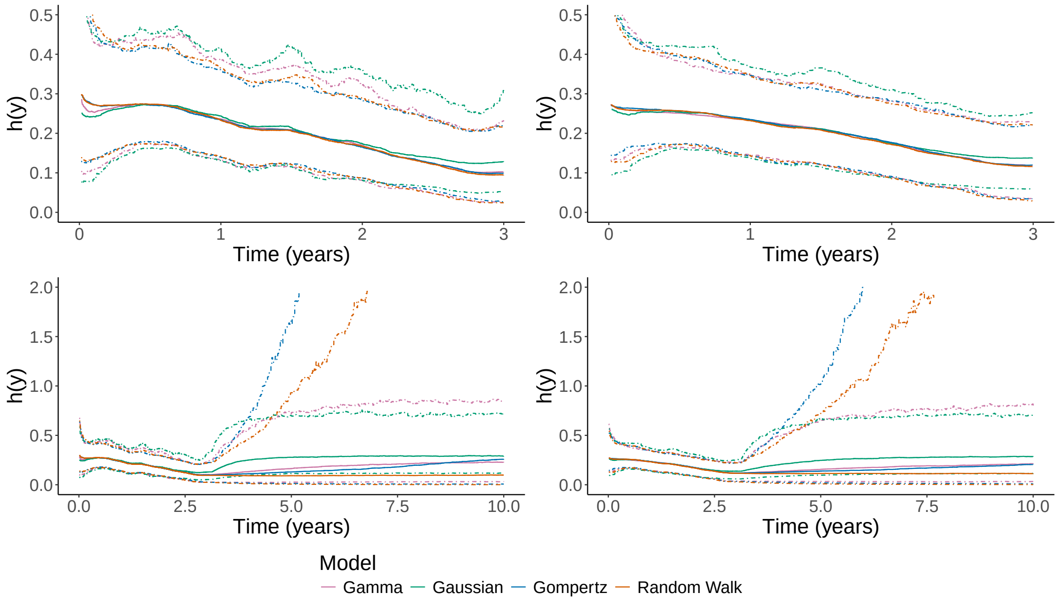

Figure 5.6: Hazard functions for the colon cancer data for observation period (top) and total period of interest (bottom) under the Poisson (left) and Negative Binomial (right) prior specifications. Median hazard values (solid) and corresponding 95% credible intervals (dashed) are reported for varying specifications of $\mu(\alpha_j)$.

In each case mean survival estimates for the observation period are close to those reported by the diffusion piecewise exponential model, although the spline and independent piecewise model provide slightly larger estimates of mean survival. We expect this to be due to the influence of $\mu(\alpha_j)$ at the end of the observation period when less data are available. Total mean survival estimates vary significantly between models. Note that the log-normal reports the smallest credible intervals, as the hazard in the extrapolation period inherits the parameter uncertainty from the observation period, and is therefore underestimating the uncertainty associated in total mean survival.

Both the independent piecewise model and the M-spline models report far higher values of mean survival in the extrapolation period. As the independent piecewise model extrapolates a constant hazard from the end of the observation period, this estimate is large with smaller credible intervals than the diffusion piecewise exponential model, as there is no additional uncertainty associated with the hazard as $y \to \infty$. Under the M-spline model, the uncertainty associated with the hazard grows until the final knot, after which a constant hazard is extrapolated. As evidenced in the estimates reported in Table 5.1 the placement of this final knot is highly influential, yet it is unclear how this knot should be placed beyond trial and error.

5.4.2 Time varying drifts

In the preceding Section we have only considered time homogeneous drift functions to guide extrapolations. As observed in Section 5.2.2, however, the prior structure can naturally be extended to incorporate time-varying drifts, $\mu(\alpha_j, y)$. This allows for a far more expressive range of expert information to be encoded into the prior.

5.4.2.1 Example time-varying drifts

For the log-baseline hazard we consider two time-varying drifts

$$\mu(\alpha_j, y) = \psi_1(y) - \psi_2(y)\exp(\alpha_j), \tag{5.12}$$

$$\mu(\alpha_j, y) = \frac{1}{\psi_2^2}(\alpha_j - \psi_1(y)), \tag{5.13}$$

constructed by adding time-varying parameters into the two Langevin drifts considered previously. In particular for the first drift $(\psi_1(y), \psi_2(y))$ are constructed such that they taper between the parameters of two different Gamma distributions on a finite given interval. As such this drift encodes a highly informative prior about the long-term hazard, but a weaker prior to be used for the observation period. The second drift allows the prior mean of the log-hazard to vary with time. In particular this allows for a pre-specified hazard function, for example elicited from previous clinical trials, to be used to guide long-term extrapolations. We note that this combining of the observed hazard with a pre-specified long-term hazard bears a strong resemblance to the blended survival approach of Che et al. [2023], albeit on the hazard rather than survival function.

A similar consideration can be taken when incorporating covariates directly in the model. Often in these cases, analysts will seek to encode a waning treatment effect assumption into extrapolations [Jackson et al., 2017]. This can be done explicitly within our framework as

$$\mu(\beta_j, y) = \frac{1}{\psi_2(y)^2}\beta_j, \tag{5.14}$$

shrinking the treatment effect to 0 as $y \to \infty$.

5.4.2.2 CLL-8 trial data

We apply the time-varying drifts to data from the CLL-8 trial [Williams et al., 2017], that investigated the effect of an immunotherapy treatment in combination with chemotherapy on survival in chronic lymphocytic leukemia (CLL) patients, compared to survival in patients who received chemotherapy alone. Here 810 patients were enrolled with 403 randomised to the treatment group and 407 to the control group, with only 11.5% of patients dying during the trial. Previous analysis has noted that there is expected to be a notable drop in $S(y)$ after 4 years [Che et al., 2023]. In particular we compare a Langevin diffusion prior with a fixed Gamma(10,10) distribution to one that converges to a Gamma(10,10) distribution in the extrapolation period, specified by (5.12). We also compare the baseline Gompertz drift to another centred around a given Gompertz hazard function (5.13). Generating two chains of 10,000 samples including burn-in took under 2 minutes for each model, except for the Gamma(10,10) model where poor prior specification hindered computation. Full prior specifications, computational details and further results are provided in Appendix C.

Survival curves for the above drifts are provided in Figure 5.7 for both the treatment and control arms along with corresponding posterior estimates of mean survival in Table 5.2. Expected mean survival is larger under each prior specification. Note that compared to the data in Section 5.4.1 events are rarer near the point of administrative censoring and therefore $\mu(\alpha_j, y)$ is more influential before y_+, as can be observed in Figure 5.7. This effect is particularly profound for the fixed Gamma(10,10) drift, in contrast (5.12) allows for the data to remain informative for longer before $\mu(\alpha_j, y)$ becomes influential. As can be seen in both trial arms, the Gompertz baseline prior is highly sensitive to the value of the survival function at y_+.

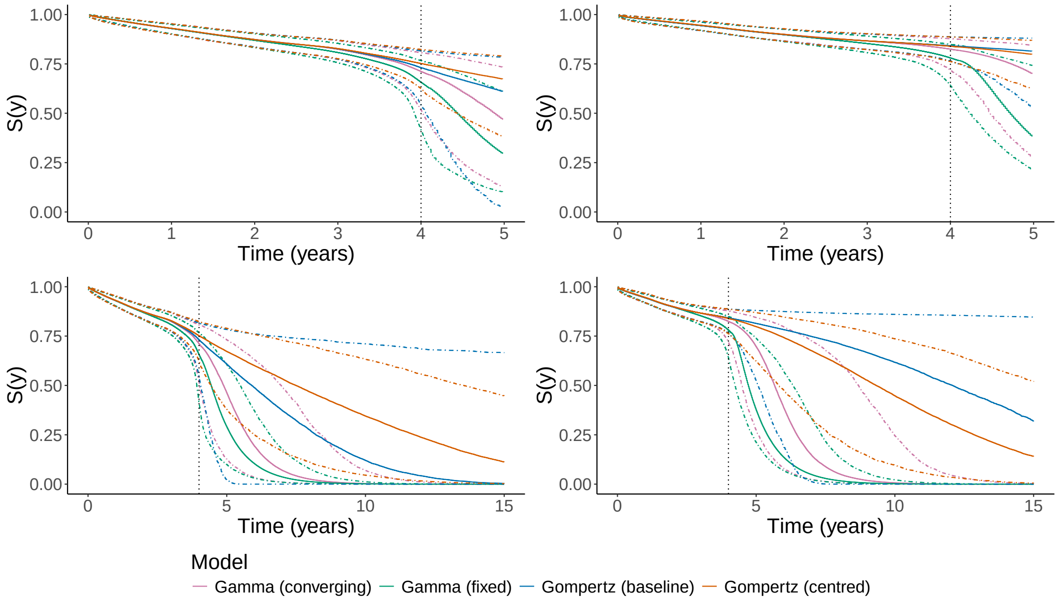

Figure 5.7: Survival curves for the diffusion piecewise exponential model for varying specifications of $\mu(\alpha_j)$ fit to the control (left) and treatment (right) arms of the CLL-8 trial data. Curves are plotted for the observation period (top) and extrapolation period (bottom), with $y_+ = 4$ denoted by the dotted line. Median values for $S(y)$ are given by the solid lines with 95% credible intervals indicated by the dashed lines.

Model	Trial arm	$\mathbb{E}[Y]$ on $(0, y_+)$	$\mathbb{E}[Y]$ on $(0, y_\infty)$
Gamma fixed	Control	3.51 (3.36, 3.64)	4.25 (3.70, 5.03)
Gamma fixed	Treatment	3.66 (3.53, 3.77)	4.67 (4.14, 5.81)
Gamma waning	Control	3.55 (3.40, 3.68)	4.71 (3.83, 6.25)
Gamma waning	Treatment	3.70 (3.58, 3.80)	5.54 (4.37, 7.95)
Gompertz Baseline	Control	3.56 (3.41, 3.69)	6.61 (3.72, 11.30)
Gompertz Baseline	Treatment	3.70 (3.58, 3.80)	9.78 (4.80, 13.11)
Gompertz centred	Control	3.58 (3.44, 3.71)	10.75 (8.35, 12.18)
Gompertz centred	Treatment	3.71 (3.60, 3.82)	12.17 (10.65, 13.08)

Table 5.2: Estimates for mean survival for the CLL-8 trial in the control and treatment arms when modelled independently under various prior assumptions. Expected mean survival and 95% credible intervals are reported for the observation period and the entire window of interest.

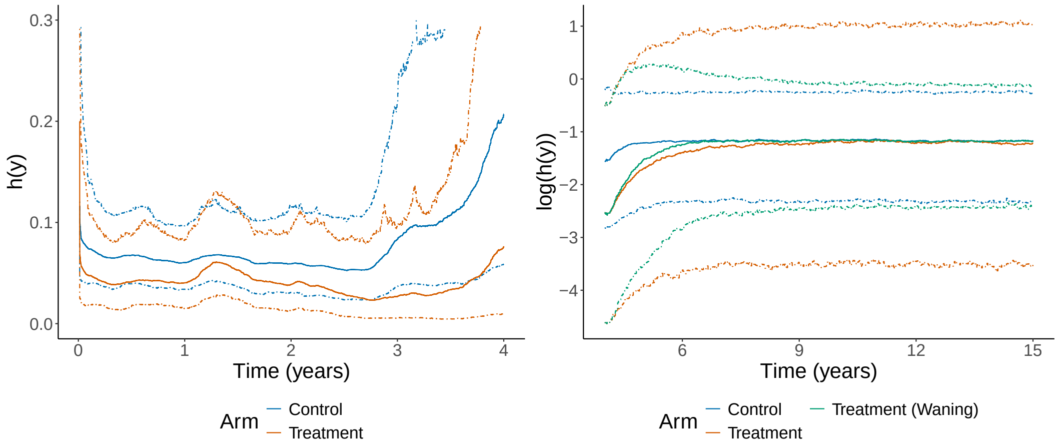

Figure 5.8: (Left) Hazard functions for the control and treatment arms with corresponding 95% credible intervals during the observation period. (Right) Log-hazard functions for the control and treatment arms (under both waning and non-waning assumptions) during the extrapolation period.

Treat. arm	$\mathbb{E}[Y]$ on $(0, y_+)$	$\mathbb{E}[Y]$ on $(0, y_\infty)$	$\mathbb{E}[Y_t] - \mathbb{E}[Y_c]$
Control	3.55 (3.41, 3.67)	5.62 (4.84, 6.60)	—
Treatment (fixed)	3.73 (3.61, 3.82)	6.34 (4.56, 9.02)	0.73 (-1.07, 3.17)
Treatment (waning)	3.73 (3.61, 3.82)	6.29 (4.90, 7.73)	0.68 (-0.64, 1.87)

Table 5.3: Estimates for mean survival and corresponding 95% credible intervals for the CLL-8 trial in the control and treatment arms (under both waning and non-waning assumptions) during the observation period, $(0, y_+)$, and the entire window of interest, $(0, y_\infty)$. The final column reports estimates of the difference between mean survival for the treatment and control groups.

As a result, minor differences in the data (as seen between the two trial arms) give rise to very different long-term survival estimates. In contrast, the centred hazard provides a far more controlled method for incorporating informative long-term information.

We conclude by investigating the model when covariates are directly incorporated rather than modelled independently. Here we use a Gamma$(5, 15)$ Langevin diffusion for the baseline log-hazard and compare a Normal$(0, 1)$ Langevin drift for $\mu(\beta_y)$, with (5.14) where the waning begins after y_+ resulting in identical inferences during the observation period. Hazards for the observation period and log-hazards for the extrapolation period are shown in Figure 5.8, with mean survival estimates provided in Table 5.3.

From the hazard functions, there is clear evidence of some non-proportionality in the observation period, and some weak evidence to suggest the treatment is beneficial compared to the control, that is corroborated by mean survival estimates. Examining the extrapolation period, both drifts for the treatment arm imply that the expected hazard should converge to the hazard for the control arm. For the fixed Langevin diffusion, uncertainty then arises from both the process for α_j and β_j. In contrast, the treatment hazard converges faster to the hazard of the control arm, and the associated credible intervals are far smaller. This is reflected in the estimates of the difference in mean survival where the waning assumption reduces the uncertainty in the estimates of difference in mean survival. We note that treatment effect waning is a strong and untestable assumption that in practice will require expert justification to be incorporated.

5.5 Discussion

In this work we have introduced the diffusion piecewise exponential model, a novel prior structure combining flexible modelling of the hazard function in the observation period with expert information in the extrapolation period within a principled Bayesian framework.

No model can automatically guarantee plausible extrapolations. The diffusion piecewise exponential model is no exception, with reasonable extrapolation relying on sensible specification of μ. Our approach has key advantages, however, compared to current state-of-the-art methods. First, as demonstrated through the variety of drifts used in Section 5.4, μ is able to incorporate a wide-range of prior information, with minimal restrictions on the form this should take. Second, specification of this prior information is only weakly informative during the observation period, becoming increasingly influential as the data become sparse. Finally, the assumptions encoded into this prior are explicit and easy to interrogate. This is a core part of the process of appraising the cost-effectiveness of novel medical interventions, and as such our model promotes improved decision making and analysis by both pharmaceutical companies and regulatory bodies.

In this chapter, we have focused on the process of incorporating prior information into long-term hazard extrapolations assuming that this information has already been elicited from subject-matter experts. Formalising the process of eliciting this information is the subject of future work, ensuring that processes exist that allow for information to be translated from expert opinion into principled prior information. Further work could also focus on implementing existing elicitation methods into this framework, for example the Sheffield elicitation framework [Gosling, 2017].

We have presented a wide range of possibilities for the specification of μ, but the examples considered here are by no means exhaustive. In the context of clinical trial data with two treatment arms, for example, dependence between each hazard could be introduced through μ rather than the local proportional hazard assumption incorporated in this work. We believe the design of drift functions that can capture an even wider range of expert information to be an exciting avenue for future research.

We have assumed a Poisson process prior for $\{s_j\}_{j=1}^{J}$ with homogeneous intensity. This assumption could be altered to incorporate a process with, for example, decreasing intensity if more volatility is expected at the start of the observation period. A particular strength of the sampling methods developed for this work is that changes to μ and γ in general do not require changes to the sampler. The only weak condition for the prior on $\{s_j\}_{j=1}^{J}$ is the existence of a dominating process that is simple to sample from, given the thinning procedure outlined in Section 5.3.4.

We have shown how efficient computational procedures for sampling from posteriors induced by spike and slab priors using PDMPs can be extended to more general transdimensional posteriors. The key feature of this construction was the identification of a hyperplane in θ-space such that the likelihoods of the simpler and more complex models were identical. In the reversible jump literature this is referred to as a centring point [Brooks et al., 2003] and is a common feature of many transdimensional posteriors. The sampling framework provided in Section 5.3 should therefore allow for the extension of sticky PDMP dynamics to a far wider range of transdimensional sampling problems.

Chapter 6

Ongoing and future work

In this chapter we outline ongoing and future work stemming from the research discussed in this thesis. Extensions to PDMP samplers are outlined in Section 6.1, generalising the sticky dynamics utilised in Chapters 4 and 5 to general surfaces. Chapter 5 introduced an underlying process as a prior for the hazard function based on the discretisation of a diffusion. In Section 6.2 we outline initial work towards the implementation of these methods into software packages. We conclude with a discussion of alternative processes that could be utilised as prior distributions for survival models.

6.1 Sticky manifold PDMP

The work in this section was partially undertaken during a visit to the Institute of Statistical Mathematics in Tokyo, Japan, and is a joint work with Professor Kengo Kamatani and Mr Hirofumi Shiba.

Chapter 5 discussed sampling from the posterior of the piecewise exponential model with a prior over the number and location of knots. The primary sampling tool used was sticky PDMP dynamics that allow PDMPs to sample from target distributions defined by a mixture of atomic and continuous components. Implementing this sampler required a re-parameterisation to sample on the space of (scaled) innovations between local hazards visualised in Figure 5.3. A natural question is whether this step can be avoided through sampling directly on the space of log-hazards. This requires the sampler to stick to an embedded $d-1$-dimensional hyperplane, rather than a single parameter sticking to zero under the construction of Chevallier et al. [2023], Bierkens et al. [2023a]. In this section we show that this sampling is possible, and provide results that show how to construct samplers that stick to general embedded surfaces.

More precisely, we aim to construct a d-dimensional PDMP with invariant measure that is jointly defined on Euclidean space and an embedded manifold, $\mathcal{M} \subset \mathbb{R}^d$

$$\pi^*(\mathrm{d}x, \mathrm{d}v) = (\omega \pi_1(x)\mathrm{d}x + (1-\omega)\pi_0(x)\mathrm{d}\mathcal{H}^{d-1}(x))\rho(\mathrm{d}v), \tag{6.1}$$

where π_1 is a density with respect to the d-dimensional Lebesgue measure (denoted $\mathrm{d}x$), π_0 is a density on $\mathcal{M}$ with respect to the Hausdorff measure (denoted $\mathrm{d}\mathcal{H}^{d-1}(x)$), and ρ is the density of the velocities. The construction of the process follows the construction in Bierkens et al. [2023a], comprising of

1. A PDMP defined on the ambient space, that targets π_1 as its invariant distribution. Throughout this section we will take this to be the Bouncy Particle sampler.

2. Deterministic sticking dynamics when the ambient process intersects $\mathcal{M}$, allowing movement from the ambient space to $\mathcal{M}$.

3. A PDMP on the constrained space, with π_0 as its invariant distribution, where π_0 will often be taken as the restriction of π_1 to $\mathcal{M}$. In particular, the deterministic dynamics of the constrained process are given by the geodesic flow on $\mathcal{M}$.

4. An unsticking rate and kernel that preserves the velocity of the process relative to $\mathcal{M}$ before the sticking event, allowing the process to move from $\mathcal{M}$ to the ambient space. This rate is given by

$$\Lambda^s(t) = \|u_\perp\| \frac{\omega \pi_1(x)\rho(v)}{(1-\omega)\pi_0(x)\rho(v)}, \tag{6.2}$$

where $u_\perp$ is the component of the velocity that is orthogonal to the tangent space when the process hits $\mathcal{M}$.

Note this is a generalisation of the processes designed by Bierkens et al. [2023a], where $\mathcal{M} = \{x \in \mathbb{R}^d : x_j = 0\}$. We will focus on two examples throughout this section:

Example 1 (Embedded hyperplane)**.**

$$\mathcal{M}_1 = \{x \in \mathbb{R}^2 : x_2 = \tan\theta x_1\}.$$

Taking $\theta = 0$ recovers the original sticky PDMP process, while taking $\theta = \pi/4$ results in the diagonal hyperplane occurring in the piecewise exponential model without reparametrisation. Similarly to Bierkens et al. [2023a] the process on $\mathcal{M}_1$ is simply a lower-dimensional version of the process defined in the ambient space PDMP.

Example 2 (Embedded hypersphere)**.**

$$\mathcal{M}_{2,R} = \{x \in \mathbb{R}^d : \|x\| = R\}.$$

Here the process on $\mathcal{M}_{2,R}$ has deterministic dynamics defined by the geodesic flow

$$(x_{t+s}, v_{t+s}) = \left(\sin\left(\frac{s\alpha}{R}\right)\frac{v_t R}{\alpha} + \cos\left(\frac{s\alpha}{R}\right)x_t, \cos\left(\frac{s\alpha}{R}\right)v_t - \sin\left(\frac{s\alpha}{R}\right)\frac{x_t\alpha}{R}\right).$$

Dynamics of PDMPs of this form have been introduced as the stereographic Bouncy Particle sampler [Yang et al., 2024, Bell et al., 2024]. Note, in the stereographic BPS the target measure is defined as a projection of the measure from $\mathbb{R}^d$ to $\mathbb{S}^d$, while we define the measure on the sphere as the restriction of the density in $\mathbb{R}^d$ to (a scaled version of) $\mathbb{S}^{d-1}$.

Theorem 1. *] Let $\mathcal{M}$ be a $d-1$ dimensional, two-sided manifold embedded in $\mathbb{R}^d$. Then the PDMP defined by steps 1), 2), 3) and 4) has* (6.1) *as its invariant distribution.*

6.1.1 Sketch proof of Theorem 1

The sketch proof of Theorem 1 generalises the approach of Bierkens et al. [2023a] to embedded manifolds. The main derivation is to show that the process is π^*-invariant. Bierkens et al. [2023a] show that sticky PDMP samplers are Harris recurrent, and that some skeleton of the chain is irreducible. If the process is also π-invariant then the chain is π-ergodic [Meyn and Tweedie, 1993, Theorem 6.1]. In this section we show π-invariance for the above processes. We believe that the proofs of Harris recurrence and irreducibility generalise to the manifold process, but have not formally shown this at this point.

Invariance of PDMPs is typically studied via the infinitesimal generator [Davis, 1993]

$$\mathcal{L}f(z) = \Phi(z) \cdot \nabla f(z) + \Lambda^B(z) \int q(z' \mid z)[f(z') - f(z)]dz',$$

for all functions f in a core of $\mathcal{L}$. Defining this core is typically non-trivial, and we omit this step during this sketch proof. The process is then π^*-invariant if

$$\int \mathcal{L}f \pi^*(dz) = 0, \tag{6.3}$$

for all f. The generator of the sticky manifold PDMP is given by

$$\mathcal{L}f(z) = \begin{cases} \Phi_1(z) \cdot \nabla f(z) + \Lambda_1^R(Q_R f - f) + \Lambda_1^E(z)(Q_E f - f), & z \in \mathbb{R}^d \times \mathcal{V}_1 \\ \Phi_0(z) \cdot \nabla f(z) + \Lambda_0^R(Q_R f - f) + \Lambda_0^E(z)(Q_E f - f) + \Lambda_0^S(z)(Tf - f) & z \in \mathcal{M} \times \mathcal{V}_0. \end{cases}$$

Here, the first line corresponds to the process in the ambient space, with, respectively, deterministic dynamics, reflection events and refreshment events. The second line corresponds to the process on $\mathcal{M}$ through either representation. The additional term is the unsticking rate, determining when the process leaves $\mathcal{M}$, with T denoting the transfer mapping that dictates the transformation of z at these events.

If the deterministic dynamics, reflection and refreshment components on both the ambient and embedded spaces are designed such that they are π^*-invariant, for π^* restricted to that space, then standard invariance results state that the corresponding terms in the generator cancel. Similar arguments are made in the processes outlined in Chevallier et al. [2023], Bierkens et al. [2023a]. The remaining terms are given by the unsticking rate on $\mathcal{M}$ and the boundary terms of the process in the ambient space, reducing (6.3) to

$$\begin{aligned}\int \mathcal{L}f \mathrm{d}\pi^* = &\int_{\mathcal{V}_1}\int_{\mathcal{M}^+\cup\mathcal{M}^-} \|u_\perp\| f \exp(-U(x))\mathcal{H}^{d-1}(\mathrm{d}x)\rho(\mathrm{d}v) \\ &+ \int_{\mathcal{V}_0}\int_{\mathcal{M}} \Lambda^S(z)(Tf-f)\mathcal{H}^{d-1}(\mathrm{d}x)\rho(\mathrm{d}v),\end{aligned} \tag{6.4}$$

where $\mathcal{H}^{d-1}$ is the Hausdorff measure and $\|u_\perp\|$ arises as a consequence of the divergence theorem [Bierkens et al., 2023b, Proposition 2.7]. To construct a non-reversible process on this space we need to assume that $\mathcal{M}$ is two-sided, such that we can divide $\mathcal{M}$ into two parts corresponding to its sides, $\mathcal{M}^-$ and $\mathcal{M}^+$. This is a generalisation of the construction in Bierkens et al. [2023a], where two copies of 0 are introduced as 0^- and 0^+. This is possible for the two examples introduced previously, but is not possible, for example, if $\mathcal{M}$ is a Mobius strip[1]. For the examples we consider, the ambient space, $\mathbb{R}^d$ is orientable, so this condition is equivalent to $\mathcal{M}$ being orientable.

Applying this to (6.4) gives

$$\begin{aligned}\int \mathcal{L}f\mathrm{d}\pi^* = &\\ &\int\int_{\mathcal{M}} (f(x^-,u_\perp)-f(x^+,u_\perp))\exp(-U(x))\|u_\perp\|\kappa(x)\mathcal{H}^{d-1}(\mathrm{d}x)\rho(du) \\ &+\int\int_{\mathcal{M}} -(f(x^+,-u_\perp)-f(x^-,-u_\perp))\exp(-U(x))\|u_\perp\|\kappa(x)\mathcal{H}^{d-1}(\mathrm{d}x)\rho(du) \\ &+\int\int_{\mathcal{M}} \lambda_{s,0}(z)(f(x^+,u_\perp)-f(x^-,u_\perp))\exp(-U(x))\mathcal{H}^{d-1}(\mathrm{d}x)\rho(du) \\ &+\int\int_{\mathcal{M}} \lambda_{s,0}(z)(f(x^-,-u_\perp)-f(x^+,-u_\perp))\exp(-U(x))\mathcal{H}^{d-1}(\mathrm{d}x)\rho(du).\end{aligned}$$

Matching terms gives the resulting rate as stated in (6.2).

[1] Note, the Mobius strip can be made two-sided, by cutting the manifold at two points, and then considering each surface separately.

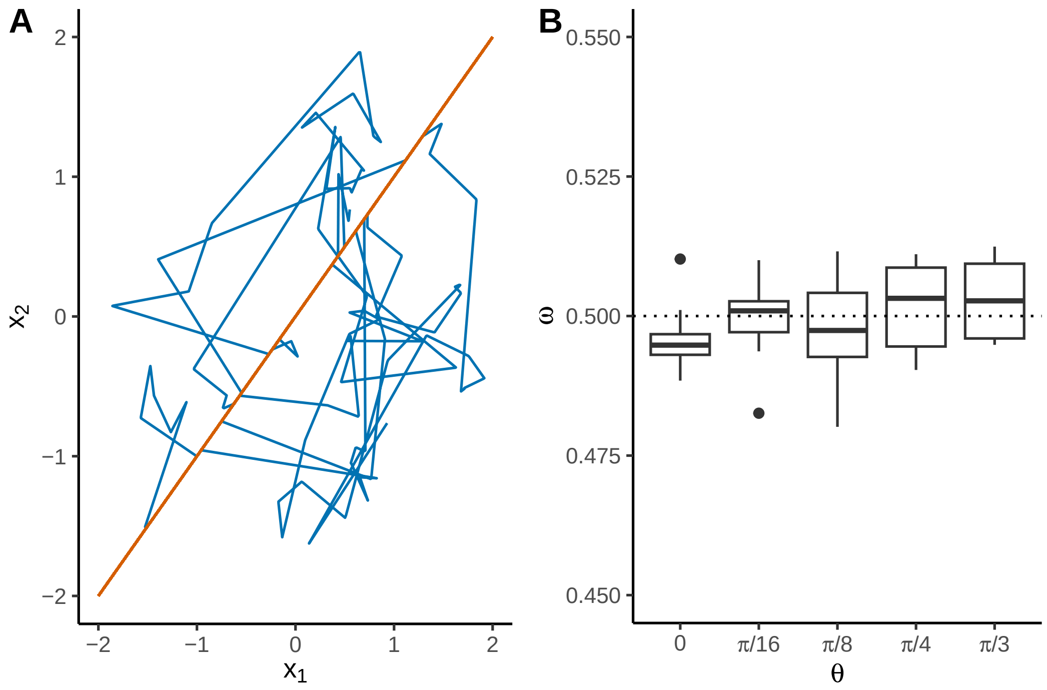

Figure 6.1: (A) Trajectories for the sticky bouncy particle sampler for Example 1, with $\mathcal{M}$ highlighted in red. (B) Box plots of estimates of ω based on 10 chains, for varying θ.

6.1.2 Examples

We outline the construction of these processes for the two introduced examples.

6.1.2.1 Example 1

Here we take $\pi(x)$ to be a standard two dimensional Gaussian with $\omega = 1/2$. The rate is then given by

$$\Lambda^S(z) = \frac{\|u_\perp\|}{\sqrt{2\pi}},$$

where $\sqrt{2\pi}$ arises as the ratio of the normalising constants between the two-dimensional Gaussian and the Gaussian restricted to $\mathcal{M}_1$. Note in particular when $\theta = 0$ this recovers the sticky bouncy particle sampler rate of Bierkens et al. [2023a], as $u_\perp$ is aligned with v_1. Figure 6.1 (A) shows the trajectory of this process for the $\theta = \pi/4$. Figure 6.1 (B) shows the estimated value of ω for varying values of θ, based on estimates from 10 separate chains. The process in the ambient space and on $\mathcal{M}_1$ can be computed exactly as outlined in Section 3.5.

6.1.2.2 Example 2

Here we take $\pi(x)$ to be a d-dimensional Gaussian. When restricted to $\mathcal{M}_{2,R}$ the resulting distribution is therefore uniform. The rate is given by

$$\Lambda^S(z) = \|u_\perp\| \frac{(2\pi)^{d/2} \exp(-\frac{1}{2}R^2)}{A_{d-1}(R)^{-1}},$$

where $A_d(R)$ is the area of the $d-1$-sphere with radius R.

Figure 6.2 shows the trajectories of this process for a sphere embedded in $\mathbb{R}^3$. Note the trajectories on the sphere are given by the geodesic flow, rather than the linear dynamics in the ambient space. Figure 6.3 shows the sample paths of $\|x\|$ for $d = 10$, and $R = \sqrt{5}$, (A), and $R = 1$, (B). For standard Gaussian distributions, mass concentrates around $\|x\| = \sqrt{d}$. This figure highlights the differences in sampler dynamics when the mass of the distribution is concentrated around $\mathcal{M}$ and when it is located away from $\mathcal{M}$. In the former case, sticking times are short to account for shorter excursions away from $\mathcal{M}$. In the latter, excursions away from $\mathcal{M}$ are longer, and this is compensated for by longer sticking times.

6.1.3 Future work

Future work will seek to formalise the proof of π-ergodicity presented for the sticky manifold process. In particular focusing on clarifying the conditions on $\mathcal{M}$. The processes here have been constructed using the geodesic flow on the manifold to define the lower dimensional PDMP. In most cases this flow is intractable. These flows can be approximated using numerical integrators [Ryckaert et al., 1977, Andersen, 1983], however this increases the computational cost of generating the process. Alternatively, $\mathcal{M}$ may be defined in terms of a coordinate chart, for example spherical coordinates can be used to represent $\mathcal{M}_{2,R}$, and a standard PDMP used on lower-dimensional Euclidean space. This is closer in spirit to the reversible jump PDMP introduced by [Chevallier et al., 2023], and would typically require the introduction of Jacobian terms into the rate to account for the corresponding transformation.

For more general choices of $\mathcal{M}$, it may not be possible to define a single

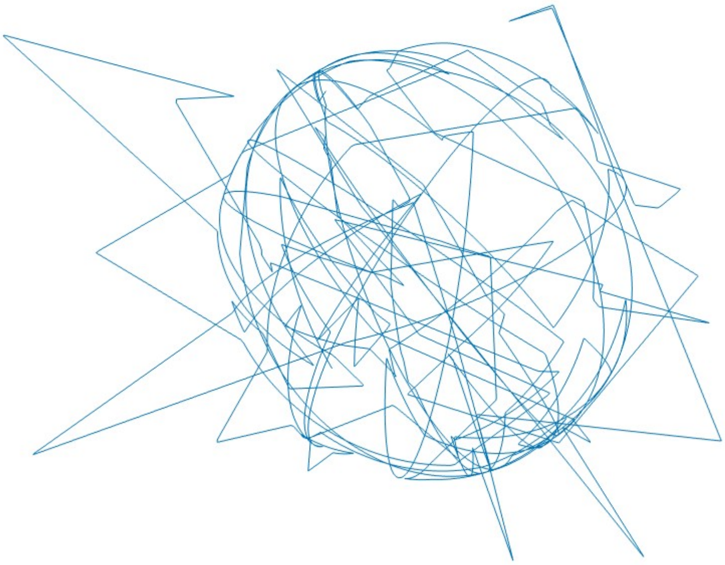

Figure 6.2: Trajectories of the sticky PDMP sampler for $\mathcal{M}_2 \subset \mathbb{R}^3$ with $R = 1$.

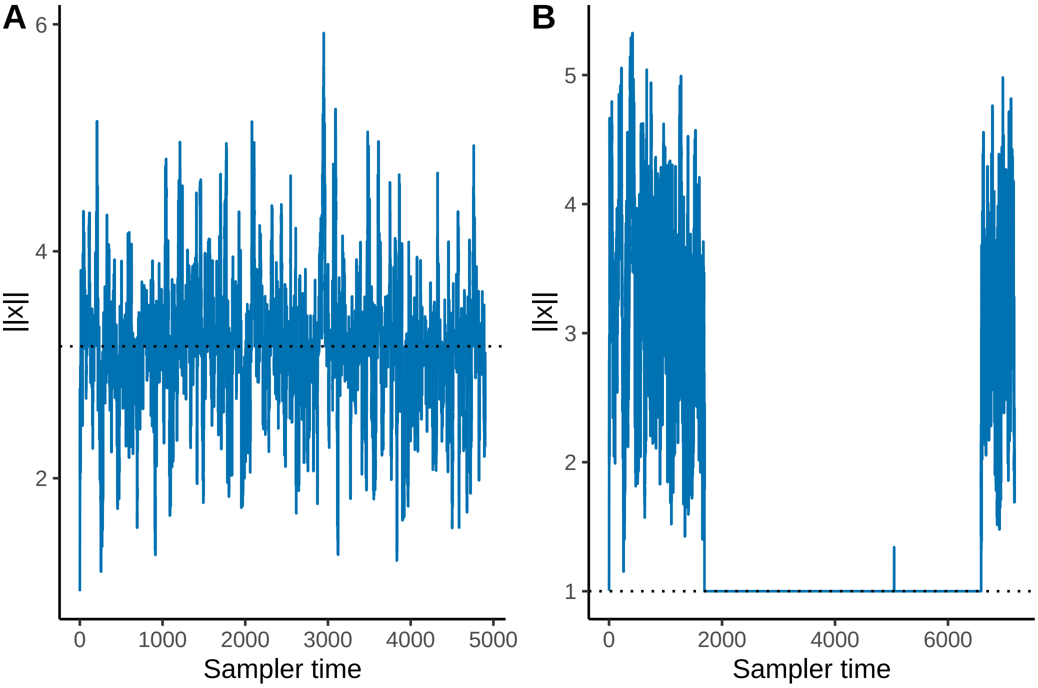

Figure 6.3: Trace plots of the $\|x\|$ for $\mathcal{M}_2 \subset \mathbb{R}^{10}$, with $R = \sqrt{5}$, (A), and $R = 1$, (B).

coordinate chart that spans the entire manifold. Instead an atlas, a countable union of coordinate charts, may be used, with the process transitioning between charts at predetermined boundaries.

The primary application of the above processes is likely to be in transdimensional sampling problems, for example the model introduced in Chapter 5. A similar process can be used to update the poly-Weibull model discussed in Chapter 4[2]. In particular, note that a bi-Weibull model whose shape parameters are equal, is equivalent to a single Weibull model with rate $\lambda_1 + \lambda_2$,

$$h(y) = \lambda_1 \gamma y^{\gamma-1} + \lambda_2 \gamma y^{\gamma-1} = (\lambda_1 + \lambda_2)\gamma y^{\gamma-1}$$

This is an example of a centring point [Brooks et al., 2003] (also Section 3.6). To construct the required process, the hyperplane example would need to be extended to account for the transformation $(\lambda_1, \lambda_2) \mapsto \lambda$. Alternatively, the lower order model could be treated as over-parametrised with both parameters sampled and jointly constrained by a prior.

6.2 Diffusion piecewise exponential package

The models and samplers developed in Chapter 4 and Chapter 5 were implemented in the julia programming language [Bezanson et al., 2017]. In particular, the diffusion piecewise exponential model developed in Chapter 5 is available via the `DiffusionPiecewiseExponential.jl` package that is in the early stages of development.

6.2.1 Model specification

An example model call is given by

```
    dpem_model = pem_fit(state0, data, priors,
                    settings, test_times, burn_in)
```

[2] This was, in fact, the primary motivation for this work!

this implements two chains of the PDMP sampler introduced in Chapter 5 to fit the diffusion piecewise exponential model. The primary terms in this call relating to the model are the `data` and `priors` objects. These are encoded in julia as structs. These are a composite type that allows multiple, related data fields of potentially differing type to be grouped together. The data struct in the above code is defined as

```julia
struct PEMData
    y::Vector{Float64}
    cens::Vector{Float64}
    covar::Matrix{Float64}
    grp::Vector{Int64}
    p::Int64
    n::Int64
    \delta::Matrix{Int64}
    W::Matrix{Float64}
    UQ::Matrix{Float64}
end
data = init_data(y, cens, covar, breaks)
```

In practice, the majority of these terms are not required to be specified by the analyst as they are precomputed quantities for efficient evaluation of the likelihood. Instead, the struct can be initialised by the `init_data` function, with analyst specifying survival times, a vector of event indicators, a matrix of covariates, and the initial specification of the grid of knots.

The `priors` struct is defined similarly. An example initialisation of this object is given by

```julia
priors = BasicPrior(..., PC(1.0, 2,...),
    FixedW([0.5]), ...
    CtsPois(7.0, 100.0, max(data.y)),
    [GaussLangevin(t -> log(0.29), t-> 0.4)],...)
```

In the above, additional terms to specify have been supressed by ellipses. In the full release of the package these are options that will be set automatically, unless specified. The remaining, highlighted terms respectively encode the prior for σ, the value of ω, the prior for $\{s_j\}_{j=1}^{J}$ and the underlying diffusion. Note in particular, each of these options is defined as its own struct. This allows the code to take advantage of julia's multiple dispatch functionality, reducing the amount of code that needs to be written.

The `CtsPois()` struct specifies the fixed Poisson prior for the set of knots, with a maximum knot value set to the maximum time in the data, $\Gamma = 7$ and the maximum number of knots truncated at 100. The Negative-Binomial prior can be similarly specified through the `CtsNB()` struct, with additional arguments for the hyperparameters of the Gamma hyperprior.

The diffusion struct, `GaussLangevin(t -> log(0.29), t-> 0.4)`, specifies a Gaussian stationary distribution for the baseline log-hazard with mean $\log(0.29)$ and standard deviation 0.4. In particular, the arguments for this struct are given by functions, set to constants in this example. This allows for the specification of the time varying drifts introduced in Section 5.4.2. Similar structs are defined for the alternative diffusions introduced in Chapter 5.

Currently specification of new diffusions requires specifications of functions for the drift and its derivative with respect to the standardised innovations $\tilde{\theta}_j$. In the full release of the package, users should ideally be able to specify the drift in terms of the local log-hazards, with automatic differentiation tools being used to generate the required derivatives [Revels et al., 2016].

6.2.2 Sampling and diagnostics

The above code implements the PDMP sampler introduced in Chapter 5. The transdimensional nature of the algorithm means individual parameters for the log-hazard cannot be monitored for convergence, as their definition shifts with the sampler. Instead we track convergence by monitoring the value of the hazard function at pre-specified time points, and the innovation standard deviation σ. Convergence for these values can then be monitored through Gelman-Rubin statistics and effective sample sizes [Gelman and Rubin, 1992, Fjelde et al., 2025], that are automatically calculated in the call to `pem_fit`.

Future work will also implement the Metropolis adjusted version of the PDMP algorithm. This requires the specification of a step-size parameter. Given the constant speed of the algorithm, however, the method should be robust to the choice of this tuning parameter. Practically a typical workflow would involve using the faster, unadjusted algorithm for exploratory model fitting, including the procedures for selecting γ outlined in Section 5.4.1. The adjusted version of the algorithm can then be used for the implementation of the final model.

Note, in particular, that the above sampler only implements the model for the observation period. As noted in Chapter 5, as extrapolations only depend on the data through the posterior of the log-hazard at the final event times, and posteriors for σ, Γ, extrapolations can be generated via direct implementation of the skew-symmetric discretisation scheme. This is implemented in the package as a separate function that can be called after the model has been fit.

6.2.3 R integration

The julia language is not typically used in most HTA workflows. This is typically done using a mixture of R and excel [Incerti et al., 2019]. To support practitioners using these methods, practitioners can currently implement these methods from R using the JuliaCall package. This allows R users to call julia from R, integrating with R workflows as

```r
julia_command("dpem_model = pem_fit(state0, data, priors,
                settings, test_times, burn_in)")
dpem_model_in_R = julia_eval("dpem_model")
```

The current implementation requires users to still write julia code. Future work will wrap these calls in an R package, matching existing syntax with established packages [Baio, 2020, Jackson, 2016] allowing practitioners to interact with these methods entirely within the R ecosystem. Future work will format the output of the model call, such that the results can interface with the HTA specific outputs provided by the `survHE` package [Baio, 2020].

6.3 Alternative processes as priors for hazards

The prior constructed in Chapter 5 is based on an approximation to an underlying diffusion process, with long-term behaviour that aligns with the expected long-term behaviour of the hazard function. In this section we explore alternative formulations of the underlying stochastic process.

6.3.1 Dense volatility matrix

The prior in the diffusion piecewise exponential model utilises a diagonal volatility matrix, with covariates encoded through a local proportional hazards assumption. An alternative specification, when the data contain multiple subgroups, is to directly model the log-hazard within each subgroup, and induce between-group dependence via a dense volatility matrix. Focusing on the case when the prior for the long-term log-hazard is encoded as a stationary distribution with density f, let $\alpha_y = (\alpha_{y,1}, \dots, \alpha_{y,k})$ denote the vector of log-hazards at time y for subgroups $1, \dots, k$. The underlying diffusion is then specified as

$$\mathrm{d}\alpha_y = \frac{1}{2} A \nabla \log f(\alpha) \mathrm{d}y + \sqrt{A} \mathrm{d}W_y,$$

for positive definite matrix A, and with W_y denoting a k-dimensional Brownian motion [Oksendal, 2013].

Here, A is a parameter in the model requiring a prior distribution. If there exists strong prior information about the structure of A this can be encoded at this stage. For example, this could encode spatial dependence, with the off-diagonal entries encoded as $A_{ij} = \rho^{|i-j|}$, or clustering between subgroups through a block-diagonal structure. The latter is particularly relevant in the context of basket trials and heterogeneous treatment groups, where analysts often seek to cluster subgroups as responsive and non-responsive [Chen et al., 2023, Hobbs et al., 2022, Lin et al., 2021].

6.3.2 Underdamped Langevin

A limitation of the previously utilised diffusions is that the prior for the innovations θ only depends on the current state of the hazard in relation to the underlying dynamics. As a result the innovation density cannot account for information about the recent trajectory of the hazard. In practice this results in sharp turning points in the hazard at the start of the extrapolation period when the expected trajectory of the hazard is moving away from areas of high density of the specified stationary distribution.

An immediate solution to this is to replace the over-damped Langevin dynamics encoded in the prior with under-damped Langevin dynamics

$$\begin{aligned} \mathrm{d}v_t &= -\gamma v_t \mathrm{d}t - \frac{1}{2} u \nabla f(\alpha_t) \mathrm{d}t + \sqrt{\gamma u} \mathrm{d}B_t, \\ \mathrm{d}\alpha_t &= v_t \mathrm{d}t, \end{aligned}$$

where α_t is the log-hazard with stationary distribution that has density proportional to $\exp(-f(x))$ and v_t are velocities with stationary distribution $\mathrm{Normal}(0, U)$.

The dynamics would need to be realised via splitting schemes. This approach would increase the number of parameters in the model, and sampling would likely be challenging due to the dependence between the state and velocities of the latent process. Further, a prior for γ would need to be defined which is not necessarily trivial.

6.3.3 PDMPs as a prior

An alternative to the above dynamics, that allows for momentum to be retained in the hazard function would be to use a Piecewise Deterministic Markov Process as a prior distribution for a log-linear hazard function. Stationary distributions could be encoded by the processes used for sampling outlined in Section 3.4. Further, the number of parameters in the model would be reduced due to the use of log-linear hazards.

To be more specific the model is defined by an initial log-hazard α_0, a set of velocities $\{v_j\}_{j=1}^J$ and a set of knots, $\{s_j\}_{j=1}^J$ corresponding to the event times of the PDMP, with $s_0 = 0$. Denoting the value of the log-hazard at time s_j as α_{s_j}, this results in the log-hazard function

$$\log h(y) = \alpha_{s_{j-1}} + v_j(y - s_{j-1}).$$

Note, in contrast to the underdamped Langevin prior, the deterministic dynamics of the PDMP reduce the number of parameters in the model, as the hazard is uniquely defined by the velocity and event times.

The primary sampling challenge associated with this prior is that the density of the between knot intervals will typically be intractable as it arises from an inhomogeneous Poisson process with rate $\Lambda^E(y)$. This challenge can be resolved if the event rate can be bounded above by a homogeneous Poisson process, $\bar{\Lambda}^E$. The location of candidate events, $\{m_l\}_{l=1}^L$ can then be generated via the homogeneous Poisson process with rate $\bar{\Lambda}^E$. Each event then corresponds to a knot, $\{s_j\}_{j=1}^J$ with probability $\Lambda^E(m_l)/\bar{\Lambda}^E$.

This representation makes the prior distribution tractable, but standard MCMC samplers may still struggle due to strong dependence encoded via the velocities and changes at event times. For example, removing s_k from $\{s_j\}_{j=1}^J$ results in an update of $h(y)$ for all $y > s_k$.

We finish by noting two additional challenges of implementation. The first is that, without refreshments, the process is unable to change direction when moving towards its stationary distribution. Applications, therefore, require a positive refreshment rate to allow the process to accurately model the hazard function. A second consideration is that typically one-dimensional stationary PDMPs have velocities constrained to $\{-v, v\}$ for fixed $v \in \mathbb{R}_+$. In applications, a prior could be placed on v, allowing the speed of the hazard to be learned from the process. Alternatively, processes with a wider range of velocity values could be incorporated such as the multi-directional Zig-Zag process [Vasdekis, 2021].

Chapter 7

General Conclusions

This thesis has focused on the joint development of novel Bayesian survival models for inferring long-term survival in the context of Health Technology Assessment, and the development and application of sampling algorithms based on Piecewise Deterministic Markov Processes. The current state of both these fields was reviewed in Chapter 2 and Chapter 3, respectively.

7.1 Contributions to survival analysis

Chapter 4 extended the existing polyhazard model, introducing a prior specification that allows for structural quantities to be inferred from observed data through Bayesian model averaging. This addresses a model selection problem that has hindered the further development and application of these models.

Chapter 5 introduces a novel prior structure for the piecewise exponential model based on a latent diffusion process for the log-hazard and a Poisson point process prior for the location of knots. This contributes to a growing literature on the use of time-varying parameter models for survival extrapolation [Kearns et al., 2019, Jackson, 2023].

A natural question is to consider which of the above approaches is preferable in certain situations. Both models provide a flexible fit to the observed data. The former then bases extrapolation on the hazard structure inferred during the observation period, while the latter informs extrapolations based on explicit prior information. The extrapolations from the diffusion piecewise exponential model are therefore likely to be more stable, when this prior information exists and can be encoded in the model. In contrast, where this prior information is unavailable, the polyhazard model is likely to provide superior extrapolations, as it is able to infer more structure from the observation period. Both models are able to incorporate covariates, however, we expect the diffusion piecewise exponential model to perform similarly to the non-proportional hazard M-spline model when the number of covariates is larger, suggesting the polyhazard model is preferable in these scenarios.

Both models developed in this thesis contribute to the increased use of flexible survival models in Health Technology Assessment. In particular, the extended polyhazard model presents a promising approach to incorporating flexible covariate effects directly in the model. This avoids the common practice of fitting separate models for each subgroup, which scales poorly with the number of covariates. Further, it allows for the structural uncertainty connected with these models to be directly accounted for in the analysis. The diffusion piecewise exponential model allows for a wide range of prior information to be incorporated to inform extrapolations. In particular, due to the flexibility in specifying the underlying drift, it separates the process of selecting a model for extrapolation from the specification of prior information, allowing for more principled inferences of mean survival in both the observation and extrapolation periods.

7.2 Contributions to posterior sampling methods

The sampling methods presented in this thesis are primarily based on Piecewise Deterministic Markov Processes. In contrast to the notable theoretical and methodological interest these processes have seen, applications in Bayesian modelling have been limited. Outside of the contributions of this thesis, we believe the primary applied work to be that of Koskela [2022].

In Chapter 4 we applied the Zig-Zag sampler to the transdimensional posterior arising from the extended polyhazard model. Generation of the event times required for sampling was achieved using a modified version of the Automatic Zig-Zag method [Corbella et al., 2022]. For non-regular, multi-modal geometries, as exhibited in polyhazard models, we believe this to be the current best choice for generating the event rate. There is a pressing need, however, for a more formal comparison of event rate generation methods. In particular, it is unlikely that there is a universally superior method out of those outlined in Section 3.5, and deeper understanding of the strengths and weaknesses of each would facilitate further application of PDMPs.

In Chapter 5 we utilised the forward event chain Monte Carlo method, discretised using splitting schemes, to sample from the diffusion piecewise exponential model posterior. An important feature of the forward event chain method is the robustness it exhibits to the choice of refreshment rate.

An important feature of PDMPs is the ability to efficiently sample from transdimensional posteriors when this is induced via a spike and slab prior [Chevallier et al., 2023, Bierkens et al., 2023a]. Chapter 4 extended these developments to incorporate continuous-time birth-death dynamics within PDMP samplers. Further, Chapter 5 showed how existing dynamics can be incorporated to sampling from more complex transdimensional posteriors via re-parameterisation and augmentation of the state space. Finally, Chapter 6 outlined initial work that extends sticky dynamics to more general embedded manifolds. Transdimensional sampling is a key strength of these samplers. For these to become commonly applied, however, there needs to be generally applicable methods for the generation of the event rate and more widely available software for their implementation.

Appendix A

Results on Markov processes

A.1 Markov chains and MCMC

A.1.1 Definitions for Section 3.2

Definition 1. *A Markov Chain* $\{x_i\}_{i=1}^{N}$ *with transition kernel* $p(x_i, x_{i+1})$ *is* $\pi-$*irreducible if for all* $A \in \mathcal{B}$*, and all* $x \in \Omega$*, there exists n such that*

$$\pi(A) > 0 \implies p^n(x, A) > 0$$

Definition 2. *A* π*-irreducible Markov chain is said to periodic if there exists a partition of the state space* $A_1, \ldots, A_p$*, with* $A_i \cap A_j =$ *for all* $i \neq j$*, and* $\cup_{j=1}^{p} A_j$ *if for* $x_i \in A_i$

$$p(x_i, A_j) = \begin{cases} 1, & j = i + k \mod k, \\ 0, & otherwise. \end{cases}$$

If the chain is not periodic it is aperiodic.

Definition 3. *A Markov chain is* π*-invariant if for all* $k \geq 1$,

$$x_i \sim \pi(\cdot) \implies x_{i+k} \sim \pi(\cdot).$$

A.2 Reversible Metropolis-Hastings proposal distributions

In Section 3.2 we referred to several choices of proposal kernel, $q(x_i, x')$ for Metropolis-Hastings MCMC methods. We explicitly state some forms of these proposals here.

1. **Independent:** The independent Metropolis-Hastings sampler [Tierney, 1994] takes q to be independent of the current state, e.g

$$q(x_i, x') \propto \exp\left(-\frac{1}{2}\|x'\|^2\right).$$

2. **Random walk:** The random walk proposal [Gelman et al., 1997] takes q as the density of a symmetric distribution centred at x_i, e.g,

$$q(x_i, x') \propto \exp\left(-\frac{1}{2\sigma^2}\|x' - x_i\|^2\right).$$

   By symmetry the acceptance probability simplifies to

$$\alpha(x_i, x') = \min\left\{1, \frac{\pi(x')}{\pi(x_i)}\right\}.$$

3. **Metropolis adjusted Langevin proposal:** MALA [Roberts and Tweedie, 1996] utilises the Euler-Maruyama discretisation of a Langevin diffusion as the proposal distribution

$$q(x_i, x') \propto \exp\left(-\frac{1}{2\sigma^2}\|x' - x_i + \frac{\sigma^2}{2}\nabla U(x_i)\|^2\right).$$

4. **Barker proposal:** The Barker proposal proposes a new point using the skew-symmetric proposal. Generation of a new point is as follows:

   (a) Draw $z \sim \text{MultivariateNormal}(0, \Sigma)$

(b) For each dimension, i, draw

$$v_i \sim \text{Rademacher}\left(\frac{1}{1+e^{z_i \partial_i U(x)}}\right).$$

(c) Propose the next iteration as

$$x_i = x' + v_i z_i.$$

Note this is equivalent to the skew-symmetric density of innovations reviewed in Chapter 5.

A.3 Poisson Point Processes

We briefly review important results from Poisson processes.

Definition 4. *Given integrable function $\Lambda : [0,\infty) \to [0,\infty)$, an integer valued counting process $\{N(t), t \geq 0\}$ is an inhomogeneous Poisson process with rate $\Lambda(t)$ if the following conditions hold:*

1. $N(0) = 0$;

2. $N(t)$ *has independent increments;*

3. *For any* $t \in [0,\infty)$

$$\begin{aligned}
\mathbb{P}(N(t+\delta) - N(t) = 0) &= 1 - \Lambda(t)\delta + o(\delta), \\
\mathbb{P}(N(t+\delta) - N(t) = 1) &= \Lambda(t)\delta + o(\delta), \\
\mathbb{P}(N(t+\delta) - N(t) \geq 2) &= o(\delta).
\end{aligned}$$

Given this definition, we can define the distribution of the number of events on a given interval $[t, t+s)$

$$N(t+s) - N(t) \sim \text{Poisson}\left(\int_t^{t+s} \Lambda(u)du\right).$$

Proposition 2. *(Thinning) Let* $\Lambda(t) \leq \bar{\Lambda}(t)$ *for* $t \geq 0$*, and let* $\{t_i\}_{i=1}^{\infty}$ *be the set of points of the Poisson process with rate* $\bar{\Lambda}(t)$*. For each i, delete* t_i *with probability* $1-\Lambda(t_i)/\bar{\Lambda}(t_i)$*. Then the remaining points form a Poisson process with intensity* $\Lambda(t)$*.*

Proposition 3. *(Superposition) Given two independent Poisson point processes* $\{N_1(t), t \geq 0\}, \{N_2(t), t \geq 0\}$ *with respective rates* $\Lambda_1(t), \Lambda_2(t)$*. Then* $N(t) = N_1(t) + N_2(t)$ *is a Poisson point process with intensity* $\Lambda_1(t) + \Lambda_2(t)$*.*

Appendix B

Appendix for Chapter 4

As stated in the main paper the likelihood for parametric survival models is given by

$$\mathcal{L}(\theta;\mathcal{D}) = \prod_{i=1}^{n} h_\theta(y_i)^{\delta_i} S_\theta(y_i).$$

Explicitly the hazard function for the polyhazard model are given by

$$h_{D,\theta,\gamma}(y \mid w) = \sum_{k=1}^{K} h_{D_k,\gamma_k,\theta_k}(y \mid w).$$

This also defines the survival function through the relation

$$\begin{aligned} S_{D,\theta,\gamma}(y \mid w) &= \prod_{k=1}^{K} S_{D_k,\gamma_k,\theta_k}(y \mid w) \\ &= \exp\left(-\sum \int_0^y h_{D,\theta,\gamma}(u \mid w) du\right). \end{aligned}$$

As stated in the paper the prior terms are given by

$$\pi_0(K,D,\gamma,\theta,\phi) \propto \pi_0(\theta \mid K,D,\gamma,\phi)\pi_0(\gamma \mid K,\phi)\pi_0(\phi)\pi_0(D \mid K)\pi_0(K).$$

Explicitly these are

$$\pi_0(\theta \mid K, D, \gamma, \phi) = \prod_{j:\gamma_j=1} \mathrm{N}(\beta_j \mid 0, \sigma_\beta^2) \prod_{k=1}^{K} \mathrm{N}(\beta_{k,0} \mid 0, \sigma_{\beta_0}^2),$$

$$\pi_0(\gamma \mid K, \phi) \propto \binom{Kp}{|\gamma|} \omega^{|\gamma|}(1-\omega)^{Kp-|\gamma|},$$

$$\pi_0(\sigma_\beta) \propto \frac{1}{1+\sigma_\beta^2}, \quad \sigma_\beta > 0,$$

$$\pi_0(\omega) \propto \omega^{a-1}(1-\omega)^{b-1},$$

$$\pi_0(D \mid K) \propto 1, \quad \pi_0(K) \propto \frac{\xi^K}{K!} 1\left(K \in \{1, \ldots, K_{\max}\}\right).$$

The resulting posterior is then

$$\pi(K, D, \gamma, \theta, \phi \mid \mathcal{D}) = \pi(K, D, \gamma, \theta, \phi) \propto \mathcal{L}(K, D, \gamma, \theta, \phi; \mathcal{D}) \pi_0(K, D, \gamma, \theta, \phi)$$

B.1 The sampler

Algorithm 3 Sampling algorithm

```
Initialise $(\theta, v, \gamma, \phi, K, D)$ at $t = 0$.
while $t < t_{\mathrm{end}}$ do
    Sample next event time $t_e \sim \mathrm{Exponential}(\Lambda^b + \Lambda^d + \Lambda^s + \Lambda^V + \Lambda^h)$,.
    Sample $\pi(\theta, v, \gamma \mid \phi, K, D)$ until time $t + t_e$.    ▷ PDMP with sticky components (Algorithm 4)
    Set $t \mapsto t + t_e$.
    Select event $i$ with probability proportional to $\Lambda^i$.
    if $i = h$ then
        Update hyperparameters    ▷ (Algorithm 5)
    end if
    if $i \in \{b, d, s\}$ then
        Perform move $i$ with probability $\Lambda^i(t)/\Lambda^i$.    ▷ Birth-death-swap process update for $(K, D)$
    end if
    if i = V then    ▷ Unsticking event
        $j \sim \mathrm{Uniform}(l : \gamma_l = 0)$.
        $v_j \sim \mathrm{Uniform}(\{-1, 1\})$.
        $\gamma_j \mapsto 1$.
    end if
end while
```

Algorithm 4 Automatic Zig-Zag variant

```
Given $(\theta, v, \gamma, \phi, K, D)$, and $t_e$ at time $t_0$.
Evaluate $\Lambda^B(t)$.                    ▷ Possibly retained from the previous iteration.
Find next sticking event $t_v$. Denote the coordinate by $j'$.
Set $t_{\max} = \min\{t^*, \Lambda^B(t)^{-1}, t_v\}$.
Evaluate $\Lambda^B(t+t_{\max}), \Lambda^B(t+t_{\max}/2)$.
if $\Lambda^B(t+t_b/2) < (\Lambda^B(t) + \Lambda^B(t+t_{\max}))/2$ then
    Set $\bar{\Lambda}^B(t) = \Lambda^B(t_0) + \frac{\Lambda^B(t_0+t_{\max})-\Lambda^B(t_0)}{t_{\max}} t$    ▷ Convexity check
else if $\min\{\Lambda^B(t), \Lambda^B(t+t_{\max})\} < \Lambda^B(t+t_{\max}/2) < \max\{\Lambda^B(t), \Lambda^B(t+t_{\max})\}$ then
    $\bar{\Lambda}^B(t) = \max\{\Lambda^B(t), \Lambda^B(t+t_{\max})\}$    ▷ Monotonicity check
else
    $\bar{\Lambda}^B(t) = \max\{\Lambda^B(s) : s \in (t, t_{\max})\}$    ▷ Brent's Method
end if
Generate $t'$ as the first time from an IHPP with rate $\bar{\Lambda}^B(t) + \Lambda_0$.
$t \mapsto t + \min\{t', t_{\max}\}$    ▷ Update time
$\theta \mapsto \theta + v \min\{t', t_{\max}\}$    ▷ Update state
if $t' < t_b$ then    ▷ Flip event
    With probability $1 - \Lambda^B(t')/\bar{\Lambda}^B(t')$ leave all velocities unchanged.
    With probability $\Lambda^B(t')/\bar{\Lambda}^B(t')$ select a single velocity to flip with probabilities proportional to $\Lambda_i^B(t')$.
end if
if $t' = t_v$ then
    Set $v_{j'} = 0$ and $\gamma_{j'} = 0$.    ▷ Sticking event.
end if
```

Algorithm 5 Hyperparameter updates

```
Given $(\theta, v, \gamma, \phi, K, D)$, and $\Sigma \in \mathbb{R}^{2\times 2}$
$\omega \sim \text{Beta}(a + |\gamma|, b + Kp - |\gamma|)$    ▷ Conjugate Gibbs update
Draw $u \sim \text{Normal}(0, \Sigma)$.    ▷ Metropolis-within-Gibbs
Set $(z_1', z_2') = (z_1 + u_1, z_2 + u_2)$
With probability $\min\left\{1, \frac{\pi(z_1', z_2')}{\pi(z_1, z_2)}\right\}$, set $(z_1, z_2) \mapsto (z_1', z_2')$.
Update $\Sigma$ using Algorithm 4 of Andrieu and Thoms [2008].
```

Concise summaries of the overall loop of the algorithm, the IHPP generation procedure and the hyperparameter update procedure are summarised in Algorithms 3, 4 & 5. We now provide additional details to the extensions developed for the automatic Zig-Zag method.

B.1.1 Extensions to the automatic Zig-Zag method

In the main paper we outlined the following three extensions to the automatic Zig-Zag method:

1. In the first iteration we check for monotonicity *and* local convexity. If local convexity we use a tighter linear bound.

2. We adaptively set the length of the bounding interval t_b using the scheme suggested by Sutton and Fearnhead [2023] in a similar context.

3. We add a constant offset rate Λ_0 to $\bar{\Lambda}^B(t)$ to offset numerical errors and failures in the above checks.

The full details of these extensions are summarised in the following. Firstly, we replace the first iteration of Brent's method with evaluations of $\Lambda^B(t)$ at $\{t_0, t_0 + t_{\max}/2, t_0 + t_{\max}\}$. We use these evaluations to check monotonicity *and* convexity. If both these checks are passed we then use the linear bound

$$\bar{\Lambda}^B(t) = \Lambda^B(t_0) + \frac{\Lambda^B(t_0 + t_{\max}) - \Lambda^B(t_0)}{t_{\max}} t, \quad t \in [t_0, t_0 + t_{\max}),$$

which is provably tighter than the constant choice. If monotonicity holds but convexity does not we use the relevant evaluation at the end of the interval as a constant upper bound, and if neither hold we resort to Brent's method. In both of the latter two instances the resulting bound is as in the Automatic Zig-Zag, but when it is applicable we have found that the linear bound can be much tighter than a constant choice, which can speed up the sampler significantly.

The second modification is to adaptively set the length of the bounding interval, as has previously been suggested in a similar context by Sutton and Fearnhead [2023], who recommend setting the length of the interval $t_{\max}$ to be the 80th percentile of observed inter-event times, t^*. We extend this approach to set $t_{\max} = \min\{t^*, \Lambda^B(t_0)^{-1}\}$, which uses information from both the history and current state of the chain. Intuitively, if the evaluation of the rate is high at the current state of the chain, a shorter interval is likely to be appropriate. This heuristic is regularised by t^* to avoid long intervals induced by small $\Lambda^B(t_0)$, which are likely to result in inefficient bounds. We note that in contrast to many adaptive MCMC schemes, this approach does not change the law of the process, and therefore we do not need to make considerations such as diminishing adaptation [Andrieu and Thoms, 2008].

The offset introduced in the final point is as an alternative to the use of smaller intervals recommended in Corbella et al. [2022].

The full method is outlined in Algorithm 4. We note that in practice rather than repeating the whole procedure after a single event time is simulated, in practice the bound can be re-used until the bounding interval is surpassed.

B.2 Birth-death-swap MCMC within Zig-Zag sampling

We present an argument for the validity of the transdimensional moves by extending the arguments presented in Sachs et al. [2023]. The main idea is to replace the Gibbs kernel in (5) of Sachs et al. [2023] with a reversible jump kernel [Green, 1995], including a Jacobian to account for the corresponding transformation (as is the case with the median-matching swap moves).

Without the underlying PDMP sampler this would correspond to birth-death MCMC [Stephens, 2000], although with an alternative specification of jumping rates, and would (inefficiently) provide valid posterior samples. Following arguments from Sachs et al. [2023], we can then superimpose the generators for this process, the Zig-Zag sampler for sampling (θ, γ) and the (Metropolis-within-)Gibbs updates for hyperparameters to construct a process with the correct target distribution.

B.2.1 On the role of balancing functions

In this work we use the Metropolis balancing function to define the birth-death-swap process for updating (K,D). Other choices of balancing function are available, however, for example the barker balancing function,

$$g_B(a) = \frac{a}{1+a}.$$

In the context of discrete time MCMC Peskun [1973] showed that the Metropolis balancing function dominates the Barker function in terms of variance of ergodic averages. When generating birth-death-swap rates via Poisson thinning as outlined in Section 3.4 we expect these results to still hold optimal.

An interesting prospect is raised, however, when considering whether this birth-death-swap process could be generated with more efficient Poisson thinning bounds. In this case the Metropolis balancing function may not be optimal and other balancing functions may be worth investigating.

B.3 Swap moves efficiency experiment details

To conduct the swap experiment in Section 3.4.2 data were generated as the following

$$\begin{aligned}
Y_1 &\sim \text{log-Normal}(0, 0.5),\\
Y_2 &\sim \text{Exponential}(1),\\
Y &= \min\{Y_1, Y_2\},\\
Y_{\text{Censored}} &\sim \text{Exponential}(0.5),\\
Y_{\text{Observed}} &= \min\{Y, Y_{\text{Censored}}\}.
\end{aligned}$$

A single binary covariate was also generated from Bernoulli random variables with $p = 0.5$.

Samplers were run for 10,000 time units generating approximately 10,000 samples with birth-death swap moves occurring with $\Lambda^S + \Lambda^{BD} = 10$. The sampler took on average 10 minutes to run. Birth, death and swap acceptance rates of 4.90%, 4.89% and 6.10% respectively this results in an across model update approximately every 2 time units.

B.4 Additional details for Section 4

This Section contains descriptions of the sampler settings used for the Lung transplant data, COST data and Kidney transplant data as well as trace plots. The digitised lung transplant data is provided online. The subset of data used for the COST analysis is available via the R `pec` package, and the kidney transplant data is available via Dryad as described in the main paper.

B.4.1 Lung transplant data

The extended polyhazard model was fit using the prior specification outlined in Section 2.3. except for the hyperprior specification, where fixed hyperparameters were used. The sampler was run for 10,000 time units, with samples taken with rate 4 and $\Lambda^{BD} + \Lambda^{S} = 10$. Trace plots for submodel posterior probabilities and a subset of parameters are shown in Figure B.1.

B.4.2 COST data

The extended polyhazard model was fit using the prior specification outlined in Section 2.3. The sampler was run for 50,000 time units, with samples taken with rate 5 and $\Lambda^{BD} = \Lambda^{S} = \Lambda^{H} = 3.33$. Convergence plots for submodel posterior probabilities along with trace plots for a subset of parameters are shown in Figure B.2 using three chains. Birth, death and swap acceptance rates were 4.36%, 4.32% and 1.99% respectively.

The M-spline models presented in the main text were fit using the default specification of the `R survextrap` package. As shown in Figure 5 (main text), the non-proportional hazards model significantly overfit. To attempt to reduce this effect a model fit with 4 knots instead of the default 10 was used. This results in under-fit hazard ratios seen in Figure B.3. The baseline hazard for the proportional hazard is also shown. The fit is very similar to the hazards in Figure 6 (main text), however with more pronounced peaks in the middle of the time period. This is discussed in Section 4.2.1.

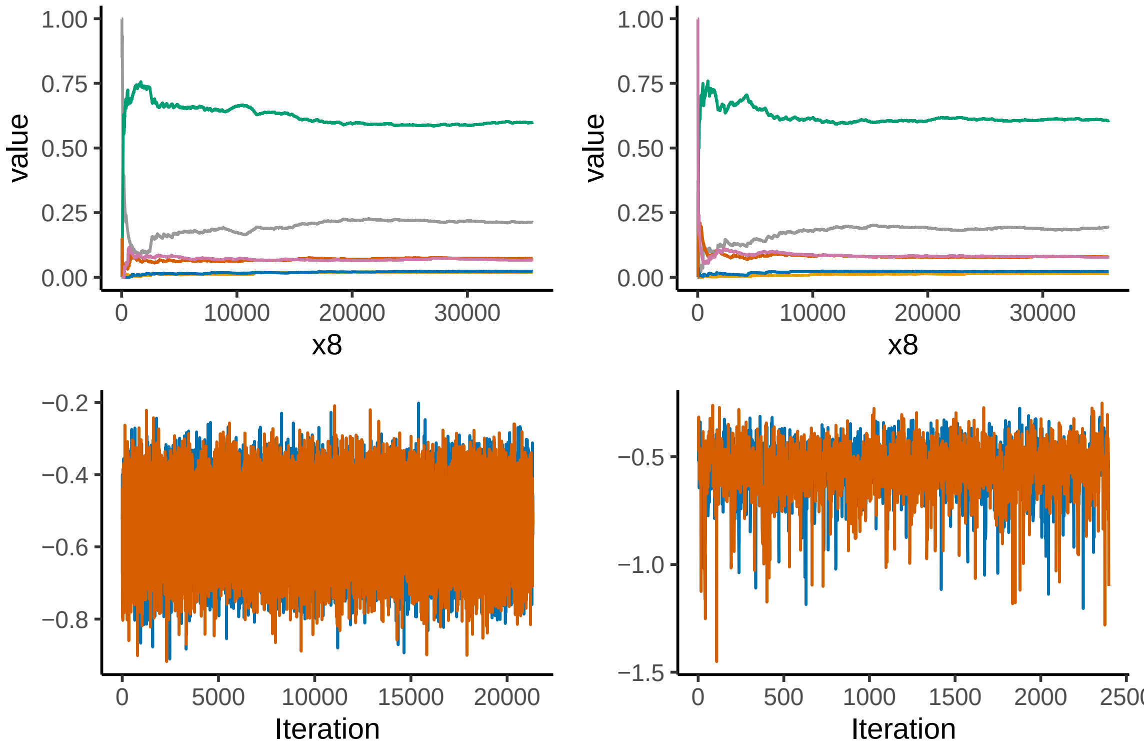

Figure B.1: Trace plots for the extended polyhazard model fit to the Lung transplant data for: (First row) Posterior sub-model probabilities from the first and second chain; (Second row) The α_1 from the bi- and tri-log-logistic models.

B.4.3 Kidney transplant data

The extended polyhazard model was fit using the prior specification outlined in Section 2.3. The sampler was run for 10,000 time units, with samples taken with rate 10 and $\Lambda^{BD} = \Lambda^{S} = \Lambda^{H} = 6.67$. Convergence plots for submodel posterior probabilities along with trace plots for a subset of parameters are shown in Figure B.5 using three chains.

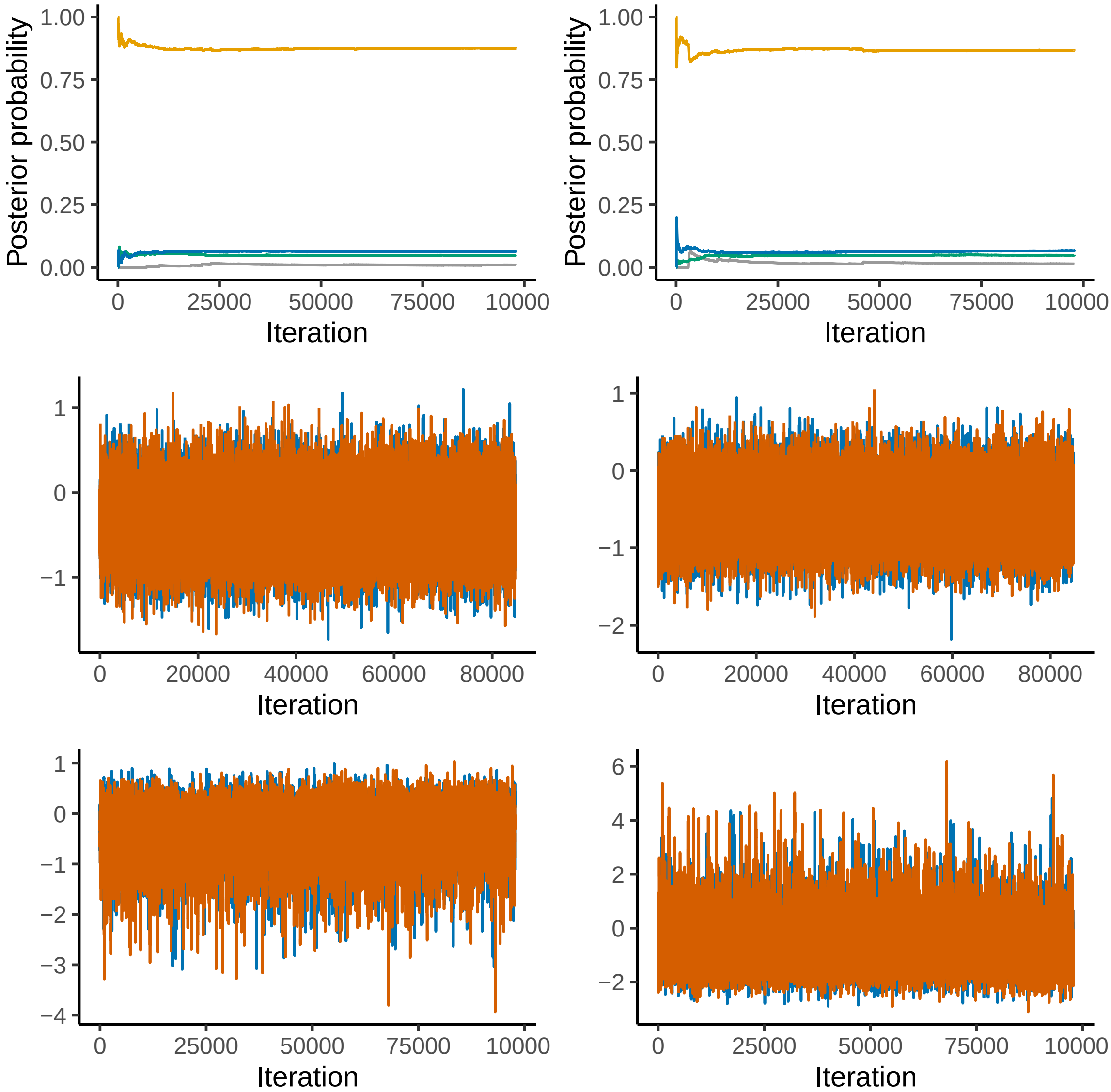

Figure B.2: Trace plots for the extended polyhazard model fit to the COST data for: (First row) Posterior sub-model probabilities from the first and second chain; (Second row) Coefficient effects from the bi-log-logistic model; (Third row) z_1 and z_2.

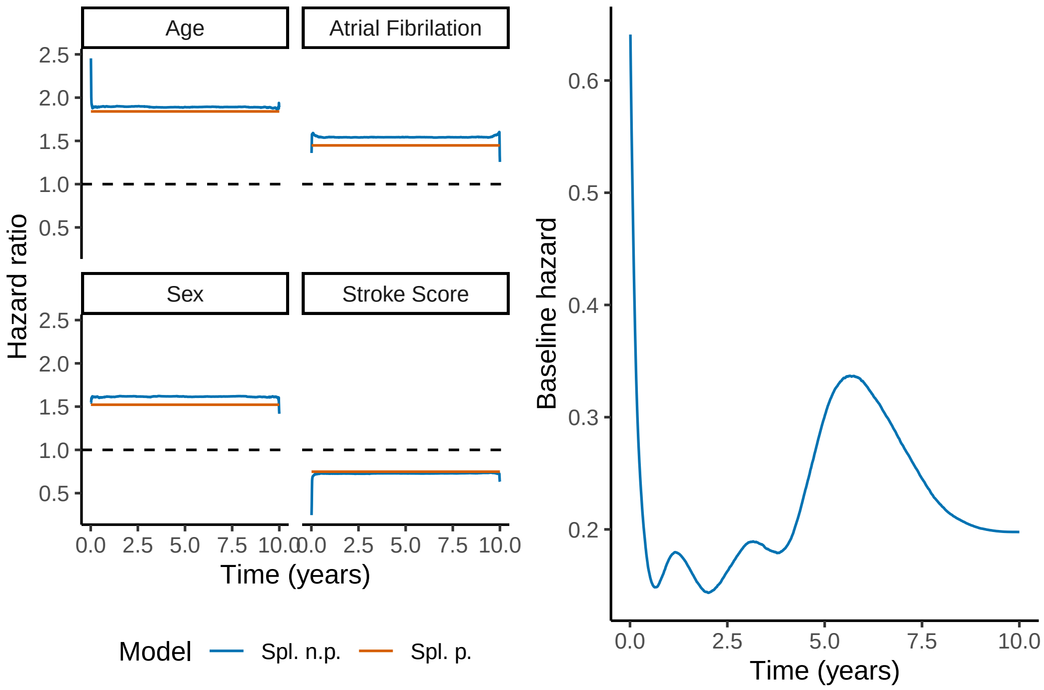

Figure B.3: (Left) Hazard ratios for the alternative specification of the M-spline model. (Right) Baseline hazards for the proportional hazard spline model.

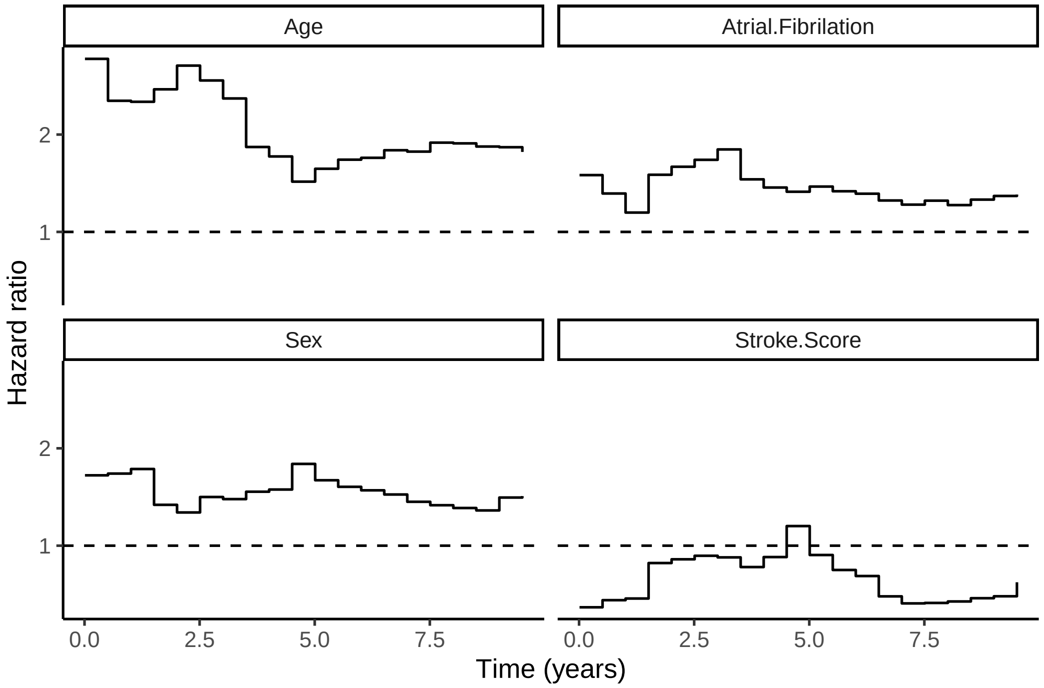

Figure B.4: Alternative presentation of the M-spline hazard ratios based on a coarser discretisation.

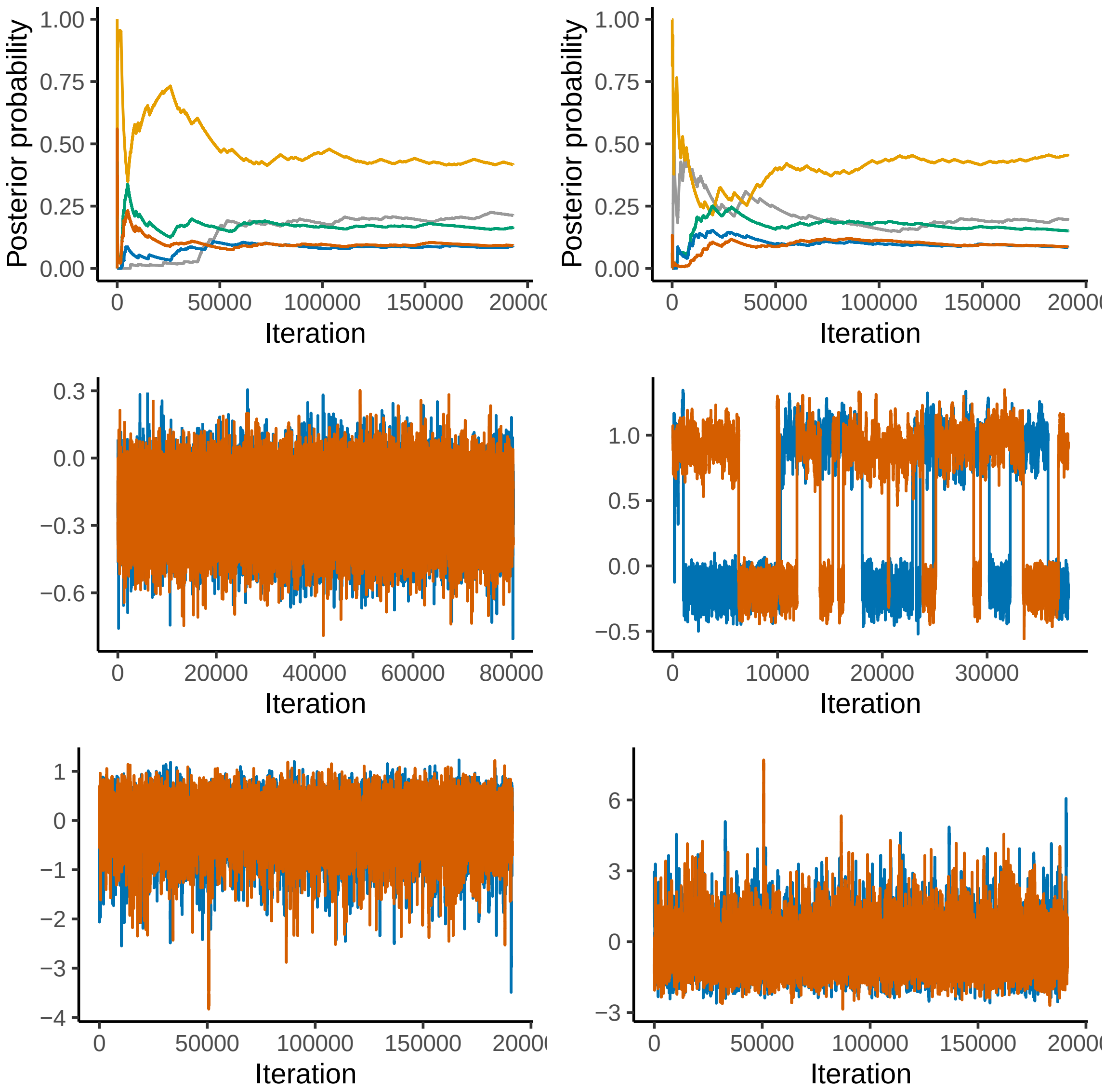

Figure B.5: Trace plots for the extended polyhazard model fit to the COST data for: (First row) Posterior sub-model probabilities from the first and second chain; (Second row) A coefficient effect from the bi-Weibull (left) and the Weibull shape parameter from the Weibull-log-logistic (right) models. Note the multi-modality in the model with hazards from different distributions; (Third row) z_1 and z_2.

Appendix C

Appendix for diffusion piecewise exponential models

C.1 Additional modelling details

C.1.1 Derivation of the Gompertz drift

The Gompertz hazard function is given by $h_0(y) = \psi_1 \exp(\psi_2 y)$. Following [Roberts and Sangali, 2010] we assume this is the solution to an autonomous differential equation

$$\begin{aligned}\frac{dh_0(y)}{dy} &= g(h_0(y)) \\ &= \psi_1 \psi_2 \exp(\psi_2 y) \\ &= \psi_2 h_0(y).\end{aligned}$$

This can then be transformed to the log-scale via a change of variables to arrive at the required drift

$$\begin{aligned}h_1(y) &= \log(h_0(y)), \\ \frac{dh_1(y)}{dh_0(y)} &= \frac{1}{h_0(y)}\psi_1 h_0(y) = \psi_1 = \mu(\alpha_y).\end{aligned}$$

C.1.2 Penalised-complexity prior derivation

We place a penalised complexity prior [Simpson et al., 2017] on the step size, σ, corresponding to the prior

$$\sigma \sim \text{Exponential}(a).$$

Following the reasoning presented in Simpson et al. [2017] we calibrate a through the probability

$$\mathbb{P}(\sigma > U) = \alpha.$$

Given the discretised diffusion prior presented in Section 2, this prior should place the majority of its mass < 1 to preserve the numerical stability of the skew-symmetric discretisation scheme [Iguchi et al., 2024]. Setting $a = 2$ gives

$$\mathbb{P}(\sigma > 1) = 0.135,$$

suggesting this is appropriately penalising σ. Following the reasoning of [Simpson et al., 2017] we expect this prior to be relatively insensitive to the specification of a.

C.2 Additional computational details

C.2.1 Algorithms

The core loop of the sampling algorithm consists of two components *i)* Generating the sticky PDMP dynamics for a fixed set of candidate knots $\{m_i\}_{i=1}^{M}$. *ii)* Updating the set of candidate knots and (if required) the hyperparameter γ. The former can be achieved using either Algorithm 6 or Algorithm 7. The latter is specified in Algorithm 8. Note that the method provided in Algorithm 6 is inexact without a Metropolis correction, with the induced bias vanishing as $\delta \to 0$ [Bertazzi et al., 2023]. This can be added after each loop of the algorithm. The results in the main paper are generated using the uncorrected version of Algorithm 6. The updates to $v^{\nabla U}, v^{\perp}$ are given by the positive p-orthogonal refresh forward event chain Monte Carlo method of [Michel et al., 2020].

Algorithm 6 Generating the PDMP via splitting schemes

```
Given step size $\Delta t$, current state $z_0 = (x_0, v_0)$ and current $b$
Simulate $U_1, U_2, U_3 \overset{iid}{\sim}$ Uniform$(0,1)$
if $U_1 < 1 - \exp(-\Lambda^{R^*} \Delta t/2)$ then
    Set $b = 1$
end if
Update $x_0 \mapsto x_{\Delta t/2}$,                    ▷ Sticky PDMP dynamics
if $U_2 < 1 - \exp(-\Lambda^{B}(\Delta t/2)\Delta t)$ then
    if $b = 0$ then
        Update $v^{\nabla U}$
    end if
    if $b = 1$ then
        Update $v^{\nabla U}, v^{\perp}$
        Set $b = 0$
    end if
end if
Update $x_{\Delta t/2} \mapsto x_{\Delta t}$.
if $U_3 < 1 - \exp(-\Lambda^{R^*} \Delta t/2)$ then
    Set $b = 1$
end if
```

Algorithm 7 Generating the PDMP using Gibbs updates and line search

```
Given current $z_0 = (x_0, v_0)$, $\sigma_0$.
Simulate $t_h \sim$ Exponential$(\lambda_h)$.
Update $z_0 \mapsto z_{t_h}$    ▷ Sticky PDMP dynamics via [Bouchard-Côté et al., 2018,
    Example 1]
Sample $\sigma$ from the full conditional $\pi(\sigma \mid z_{t_h})$.    ▷ Metropolis-within-Gibbs
```

Algorithm 8 Updating $\{m_i\}_{i=1}^{M}$ and Γ

```
Given $z_t, \Gamma_t, \{s_j\}_{j=1}^{J}, \{m_i\}_{i=1}^{M}$.
Update $z_t, \{s_j\}_{j=1}^{J}$                    ▷ Algorithm 6 or Algorithm 7
Update $K = M - J$, $K \sim$ Poisson$((1-\omega)\Gamma y_+)$
Update $\{r_j\}_{j=1}^{M-J} \overset{iid}{\sim}$ Uniform$(0, y_+)$
Update $\Gamma \sim$ Gamma$(J + \alpha, \omega/(\beta + 1))$        ▷ If hyperprior specified
```

C.2.2 Generating extrapolations

In Section 3.6 we highlight that extrapolations can: *i)* Be generated using the skew-symmetric scheme directly. This reduces the computational cost of the methods. *ii)* Be generated using a re-scaled consistent step size σ^*. This gives the practitioner more control over the computational cost of generating extrapolations, and the bias induced by using a discretisation scheme. Given σ^*, extrapolations can be generated using a set of times with inter-arrival times given by

$$\{\tau_i\}_{i=1}^N \overset{iid}{\sim} \text{Exponential}(\Gamma\omega(\sigma/\sigma^*)^2).$$

In the examples of Section 4 we use $\sigma^* = 0.1$.

C.2.3 Reversible jump MCMC

The reversible jump algorithm of Section 3.7 alternates between a random walk Metropolis update to $\tilde{\theta}$ and reversible jump moves which add and delete knots in the samplers [Green, 1995]. Knots are added as

1. Propose a new knot location $s_j^* \sim \text{Uniform}(s_1, y_+)$.
2. Propose a new value for the scaled innovation at that knot, $\tilde{\theta}_j^* \sim \text{Normal}(0, \sigma_{RJ}^2)$.
3. Accept the proposed knot and innovation with probability $\min\{1, A\}$ with

   $$A = \frac{\pi(\tilde{\theta}^*, s^*)/(J+1)}{\pi(\tilde{\theta}, s)q(\tilde{\theta}^*)}.$$

   Where $q(\cdot)$ is the proposal density, and the $J+1$ term arises as the probability of picking a knot to remove in the reverse move.

Knots are removed by selecting a knot and corresponding innovation to remove from the model. These moves are accepted with probability $\min\{1, A^{-1}\}$.

The reversible jump sampler in Section 3.7 was run using a step size of 0.05 for the Random Walk Metropolis kernel and 0.01 for the step size of the reversible jump proposal. The sampler was run for 1,000,000 iterations, with one iteration consisting

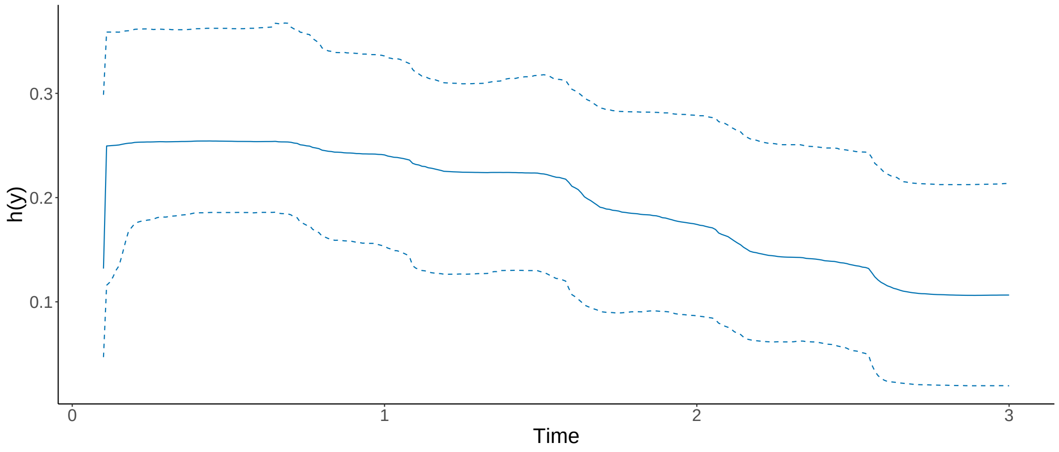

Figure C.1: Hazards under the reversible jump sampler with alternative reversible jump proposal parameter.

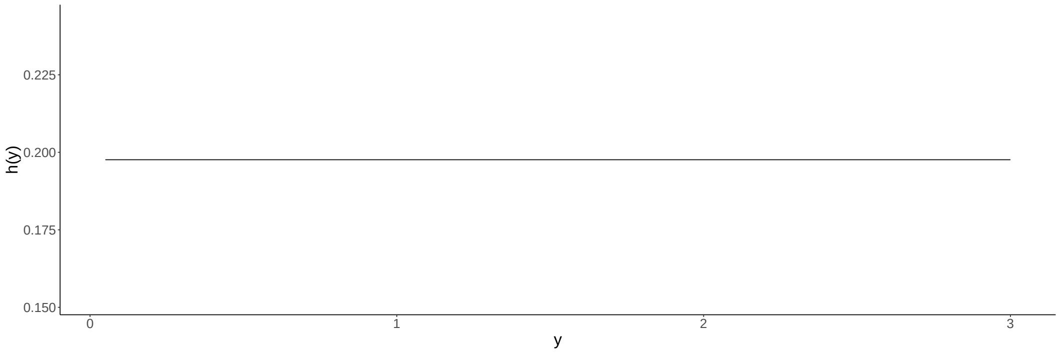

Figure C.2: The mean hazard under the piecewise exponential model and sampler of Chapple et al. [2020].

of a single Random Walk Metropolis and reversible jump step. The PDMP sampler was run for the same computational budget. Using a step size of 1 in the reversible jump proposal results in the hazards in Figure C.1.

We also sought to compare the sampler developed in Section 3 to an existing sampler. To this end the results of applying the piecewise exponential model of [Chapple et al., 2020] to the Colon cancer. The resulting hazard is plotted in Figure C.2. In this case both the reversible jump and within model sampling components have failed to explore the state space. This serves to illustrate the difficulty in designing and implementing these samplers.

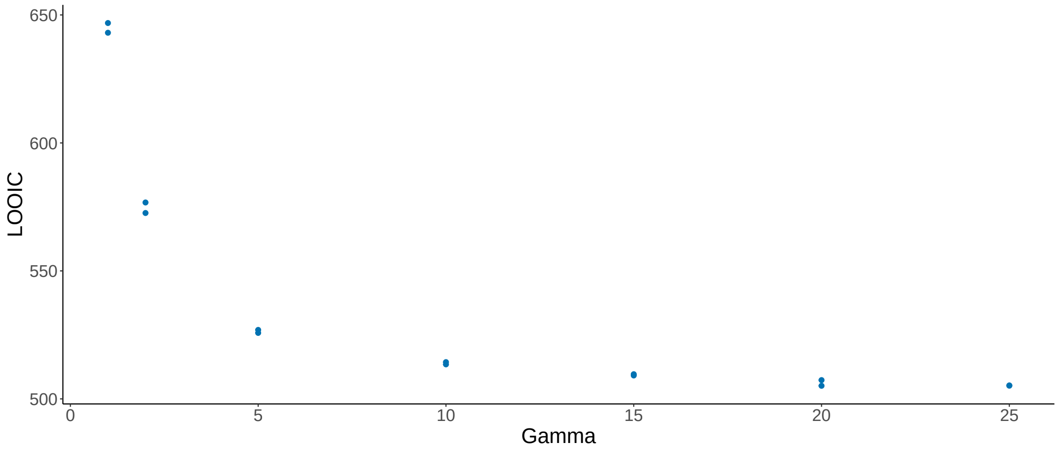

Figure C.3: LOOIC values for various values of Γ for the colon cancer data.

C.3 Additional details for example applications

All the models were implemented in Julia with code available at `https://github.com/LkHardcastle/PEM_extrap`.

C.3.1 Specification of γ

To find the optimal value of γ models were fit to the Colon cancer data using $\mu(\alpha_j) = 0$, for $\Gamma \in \{1,2,5,10,15,20,25\}$. LOOIC values were computed using Pareto-smoothed importance sampling [Vehtari et al., 2017]. These values are plotted in Figure C.3. While the LOOIC decreases as γ increases, the improvement begins to plateau between $\Gamma = 5$ and $\Gamma = 10$, indicating $\Gamma = 7$ as a good choice of hyperparameter.

C.3.2 Colon cancer data

The Colon cancer data were accessed via the `R survextrap` package. Each model was run for 2 chains of 10,000 iterations, 5,000 of which were burn-in, and where each iteration consists of a single Gibbs update for $\{r_k\}_{k=1}^{M-J}$ and 50 iterations of Algorithm 6 with $\Delta t = 0.01$ Convergence was assessed by examining trace plots, $\hat{R}$ values for the hazard at fixed time points and effective sample sizes. In this example, and all following examples, chains were run until $\hat{R} < 1.01$ for all time points (taken at intervals of 0.2 on $(0, y_+)$). Priors were as specified in Sections 2 and Section 4.1. The specific drift functions were derived as follows. The log-Gamma(2,7) stationary

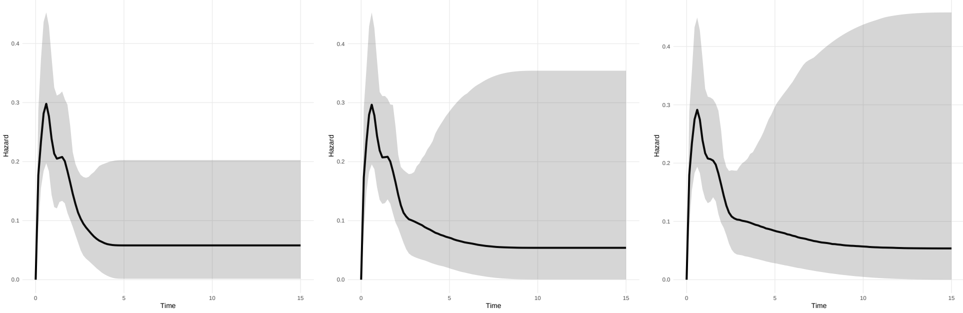

Figure C.4: Hazards for the M-spline hazard model fit to the Colon cancer data with the final knot placed at 5, 10 and 15 years [Jackson, 2023]. Note how extrapolations are strongly dependent on the placement of the final knot.

Model	AIC	$\mathbb{E}[Y](0, y_+)$	$\mathbb{E}[Y](0, y_\infty)$
Exponential	431.57	2.22 (2.06, 2.35)	4.71 (3.80, 5.63)
Gamma	433.56	2.21 (2.08, 2.35)	4.73 (3.73, 5.72)
Gompertz	428.20	2.15 (1.98, 2.30)	7.12 (4.70, 8.75)
Weibull	433.34	2.20 (2.05, 2.35)	4.89 (3.78, 5.96)
Log-logistic	428.24	2.18 (2.02, 2.32)	5.60 (4.57, 6.61)
Log-normal	422.19	2.18 (2.03, 2.32)	5.79 (4.71, 6.86)

Table C.1: Results for the standard parametric models fit to the Colon cancer data. Mean survival results and 95% confidence intervals are reported.

distribution was elicited by assuming a constant (exponential) hazard as $y \to y_\infty$. Using standard conjugacy results this can be elicited by assuming the observation of a individuals for b time until events was observed. This stationary prior implies 2 individuals observed for a total of 7 years in the limit. The Gaussian Langevin stationary distribution was then selected to approximately match the uncertainty intervals of this Gamma distribution. We note that these examples are purely illustrative and can likely be improved on in practice.

C.3.3 Comparators

Figure C.4 and Figure C.5 show hazard functions for the comparators in Section 4.1 [Baio, 2020, Cooney and White, 2023a, Jackson et al., 2017]. The results for the standard parametric models are presented in Table C.1. AIC is minimised for the log-normal model, and as such this is the model used as the comparator in the main manuscript.

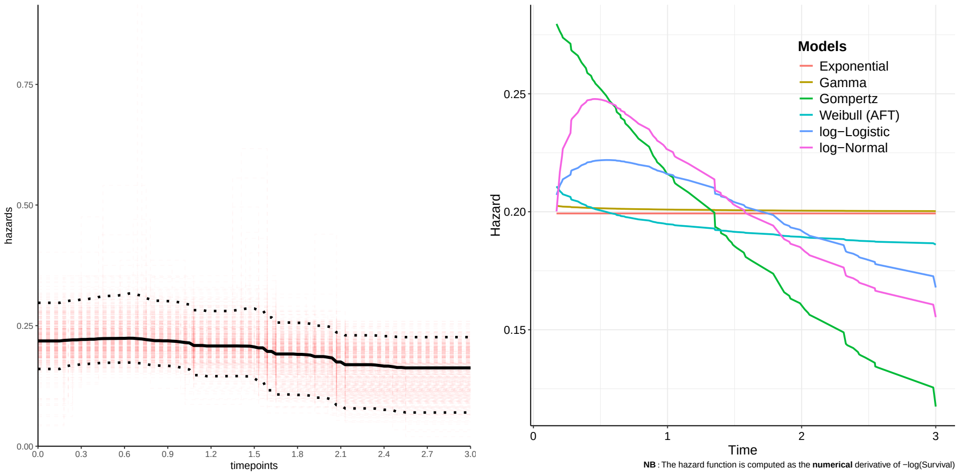

Figure C.5: (Left) Hazards for the independent piecewise exponential model fit to the Colon cancer data [Cooney and White, 2023a]. Note how the hazard is not as expressive as either the diffusion piecewise exponential model or the M-spline model. (Right) Hazards for the standard parametric models fit to the colon cancer data. Hazards are computed using numerical derivatives of $-\log(S(y))$ and as such appear non-smooth in the plot.

C.3.4 CLL-8 trial data

The same procedure as for the Colon cancer data was used to determine an optimal value of Γ. The results of this procedure are shown in Figure C.6. While the LOOIC is minimised for $\Gamma = 20$, the values begin to plateau around $\Gamma = 10$. Each model was run for 2 chains of 10,000 iterations, 5,000 of which were burn-in, and where each iteration consists of a single Gibbs update for $\{r_k\}_{k=1}^{M-J}$ and 50 iterations of Algorithm 6 with $\Delta t = 0.01$ Convergence was assessed by examining trace plots, $\hat{R}$ values for the hazard at fixed time points and effective sample sizes.

We outline the time-varying drift functions used in Section 4.2.

Gamma waning:

$$\mu(\alpha_y, y) = \psi_1(y) - \psi_2(y)\exp(\alpha_y),$$

$$\psi_1(y) = \psi_1 \max\{\min\{1, y/c\}, 1/c\}, \quad \psi_2(y) = \psi_2 \max\{\min\{1, y/c\}, 1/c\}.$$

Waning treatment effect:

$$\mu(\beta_y, y) = \frac{1}{\psi_2(y)^2}\beta_y, \quad \psi_2(y) = \max(1, (y/4)^2)^{-1}.$$

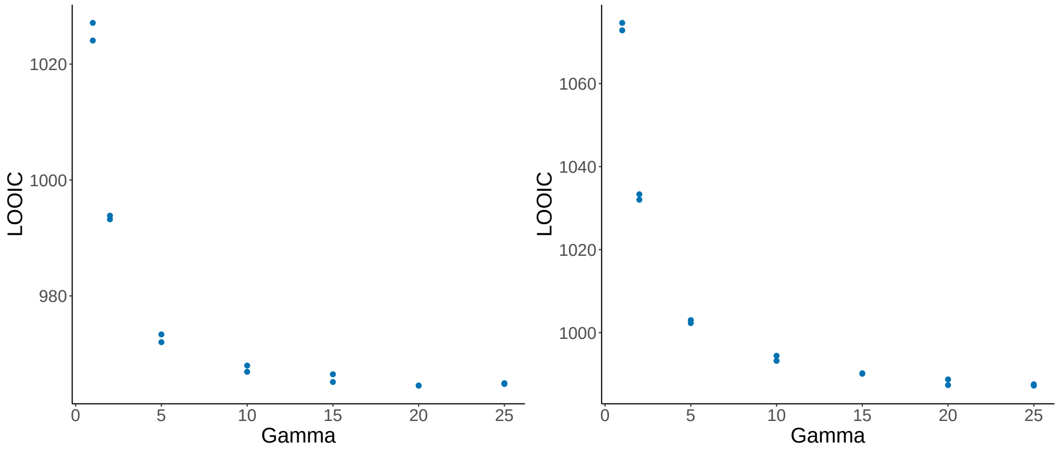

Figure C.6: LOOIC values for various values of Γ for the CLL-8 data.

Bibliography

O. O. Aalen and H. K. Gjessing. Survival models based on the ornstein-uhlenbeck process. *Lifetime data analysis*, 10:407–423, 2004.

S. Agrawal, J. Bierkens, and G. O. Roberts. Large sample scaling analysis of the zig-zag algorithm for bayesian inference. *arXiv preprint arXiv:2411.14983*, 2024.

M. Amico and I. Van Keilegom. Cure models in survival analysis. *Annual Review of Statistics and Its Application*, 5(1):311–342, 2018.

L. D. Amorim and J. Cai. Modelling recurrent events: a tutorial for analysis in epidemiology. *International journal of epidemiology*, 44(1):324–333, 2015.

H. Andersen. Rattle: a 'velocity'version of the shake calculations algorithm for molecular dynamics. *J. Comput. Phys*, 52:24–34, 1983.

K. K. Andersen and T. S. Olsen. One-month to 10-year survival in the copenhagen stroke study: interactions between stroke severity and other prognostic indicators. *Journal of Stroke and Cerebrovascular Diseases*, 20(2):117–123, 2011.

M. N. Andersen, K. K. Andersen, L. P. Kammersgaard, and T. S. Olsen. Sex differences in stroke survival: 10-year follow-up of the copenhagen stroke study cohort. *Journal of Stroke and Cerebrovascular Diseases*, 14(5):215–220, 2005.

T. M.-L. Andersson, P. W. Dickman, S. Eloranta, M. Lambe, and P. C. Lambert. Estimating the loss in expectation of life due to cancer using flexible parametric survival models. *Statistics in medicine*, 32(30):5286–5300, 2013.

C. Andral and K. Kamatani. Automated techniques for efficient sampling of piecewise-deterministic markov processes. *arXiv preprint arXiv:2408.03682*, 2024.

C. Andrieu and S. Livingstone. Peskun–tierney ordering for markovian monte carlo: beyond the reversible scenario. *The Annals of Statistics*, 49(4):1958–1981, 2021.

C. Andrieu and J. Thoms. A tutorial on adaptive mcmc. *Statistics and computing*, 18:343–373, 2008.

A. Apsemidis and N. Demiris. Stable survival extrapolation via transfer learning. *arXiv preprint arXiv:2409.16044*, 2024.

A. Bagust and S. Beale. Survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: An alternative approach. *Medical Decision Making*, 34(3):343–351, 2014. doi: 10.1177/0272989X13497998. PMID: 23901052.

G. Baio. *Bayesian methods in health economics*, volume 15. CRC Press Boca Raton (FL), 2013.

G. Baio. survhe: survival analysis for health economic evaluation and cost-effectiveness modeling. *Journal of Statistical Software*, 95:1–47, 2020.

C. Bell, G. O. Roberts, et al. Adaptive stereographic mcmc. *arXiv preprint arXiv:2408.11780*, 2024.

T. Benaglia, C. H. Jackson, and L. D. Sharples. Survival extrapolation in the presence of cause specific hazards. *Statistics in Medicine*, 34(5):796–811, 2015.

J. O. Berger and D. Sun. Bayesian analysis for the poly-weibull distribution. *Journal of the American Statistical Association*, 88(424):1412–1418, 1993.

E. P. Bernard, W. Krauth, and D. B. Wilson. Event-chain monte carlo algorithms for hard-sphere systems. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 80(5):056704, 2009.

D. Bernoulli. Essai d'une nouvelle analyse de la mor-talite causee par la petite verole. *Mémoires de mathématique et de physique présentés par divers savants à l'Académie Royale des Sciences*, 1766.

D. Bernoulli and S. Blower. An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. *Reviews in medical virology*, 14(5):275, 2004.

A. Bertazzi and J. Bierkens. Adaptive schemes for piecewise deterministic monte carlo algorithms. *Bernoulli*, 28(4):2404–2430, 2022.

A. Bertazzi, P. Dobson, and P. Monmarché. Piecewise deterministic sampling with splitting schemes. *arXiv preprint arXiv:2301.02537*, 2023.

M. Betancourt. Fitting the cauchy, 2018. URL `https://betanalpha.github.io/assets/case_studies/fitting_the_cauchy.html`.

M. Betancourt and M. Girolami. Hamiltonian monte carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79(30):2–4, 2015.

J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017. doi: 10.1137/141000671. URL `https://epubs.siam.org/doi/10.1137/141000671`.

J. Bierkens. Non-reversible metropolis-hastings. *Statistics and Computing*, 26(6): 1213–1228, 2016.

J. Bierkens, P. Fearnhead, and G. Roberts. The zig-zag process and super-efficient sampling for bayesian analysis of big data. *Annals of Statistics*, 47(3):1288–1320, 2019.

J. Bierkens, S. Grazzi, K. Kamatani, and G. Roberts. The boomerang sampler. In *International conference on machine learning*, pages 908–918. PMLR, 2020.

J. Bierkens, K. Kamatani, and G. O. Roberts. High-dimensional scaling limits of piecewise deterministic sampling algorithms. *The Annals of Applied Probability*, 32(5):3361–3407, 2022.

J. Bierkens, S. Grazzi, F. v. d. Meulen, and M. Schauer. Sticky pdmp samplers for sparse and local inference problems. *Statistics and Computing*, 33(1):8, 2023a.

J. Bierkens, S. Grazzi, G. Roberts, and M. Schauer. Methods and applications of pdmp samplers with boundary conditions. *arXiv preprint arXiv:2303.08023*, 2023b.

J. Bierkens, K. Kamatani, and G. O. Roberts. Scaling of piecewise deterministic monte carlo for anisotropic targets. *Bernoulli*, 31(3):2323–2350, 2025.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, 2016.

N. Bou-Rabee and J. M. Sanz-Serna. Randomized hamiltonian monte carlo. *The Annals of Applied Probability*, pages 2159–2194, 2017.

N. Bou-Rabee, B. Carpenter, and M. Marsden. Gist: Gibbs self-tuning for locally adaptive hamiltonian monte carlo. *arXiv preprint arXiv:2404.15253*, 2024.

A. Bouchard-Côté, S. J. Vollmer, and A. Doucet. The bouncy particle sampler: A nonreversible rejection-free markov chain monte carlo method. *Journal of the American Statistical Association*, 113(522):855–867, 2018.

R. P. Brent. An algorithm with guaranteed convergence for finding a zero of a function. *The computer journal*, 14(4):422–425, 1971.

S. L. Brilleman, E. M. Elci, J. B. Novik, and R. Wolfe. Bayesian survival analysis using the rstanarm r package. *arXiv preprint arXiv:2002.09633*, 2020.

S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

S. P. Brooks, P. Giudici, and G. O. Roberts. Efficient construction of reversible jump markov chain monte carlo proposal distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(1):3–39, 2003.

A. Bullement, M. Edmondson-Jones, P. Guyot, N. J. Welton, G. Baio, M. Stevenson, and N. R. Latimer. Mpes-r: multi-parameter evidence synthesis in r for survival extrapolation—a tutorial. *PharmacoEconomics*, 42(12):1317–1327, 2024.

G. Bütepage, C. Vitor, and P. Carlqvist. Msr71 performance of aic and bic for the extrapolation of survival data with different levels of censoring. *Value in Health*, 25(12):S363, 2022.

A. G. Chapple, T. Peak, and A. Hemal. A novel bayesian continuous piecewise linear log-hazard model, with estimation and inference via reversible jump markov chain monte carlo. *Statistics in medicine*, 39(12):1766–1780, 2020.

Z. Che, N. Green, and G. Baio. Blended survival curves: a new approach to extrapolation for time-to-event outcomes from clinical trials in health technology assessment. *Medical Decision Making*, 43(3):299–310, 2023.

H.-H. Chen, Y.-B. Chern, C.-Y. Hsu, P.-L. Tang, and C.-C. Lai. Kidney transplantation waiting times and risk of cardiovascular events and mortality: a retrospective observational cohort study in taiwan [dataset]. *Dryad*, 2022a.

H.-H. Chen, Y.-B. Chern, C.-Y. Hsu, P.-L. Tang, and C.-C. Lai. Kidney transplantation waiting times and risk of cardiovascular events and mortality: a retrospective observational cohort study in taiwan. *BMJ open*, 12(5):e058033, 2022b.

X. Chen, J. Zhang, L. Jiang, and F. Yan. Shotgun-2: A bayesian phase i/ii basket trial design to identify indication-specific optimal biological doses. *Statistical Methods in Medical Research*, 32(3):443–464, 2023.

A. Chevallier, P. Fearnhead, and M. Sutton. Reversible jump pdmp samplers for variable selection. *Journal of the American Statistical Association*, 118(544): 2915–2927, 2023.

A. Chevallier, S. Power, and M. Sutton. Towards practical pdmp sampling: Metropolis adjustments, locally adaptive step-sizes, and nuts-based time lengths. *arXiv preprint arXiv:2503.11479*, 2025.

A. Chin and A. Nishimura. Mcmc using bouncy hamiltonian dynamics: A unifying framework for hamiltonian monte carlo and piecewise deterministic markov process samplers. *arXiv preprint arXiv:2405.08290*, 2024.

P. Cooney and A. White. Extending beyond bagust and beale: Fully parametric piecewise exponential models for extrapolation of survival outcomes in health technology assessment. *Value in Health*, 26(10):1510–1517, 2023a. ISSN 1098-3015. doi: https://doi.org/10.1016/j.jval.2023.06.007.

P. Cooney and A. White. Extrapolation of relative treatment effects using change-point survival models. *arXiv preprint arXiv:2401.00568*, 2023b.

P. Cooney and A. White. Direct incorporation of expert opinion into parametric survival models to inform survival extrapolation. *Medical Decision Making*, 43 (3):325–336, 2023c.

S. Cope, D. Ayers, J. Zhang, K. Batt, and J. P. Jansen. Integrating expert opinion with clinical trial data to extrapolate long-term survival: a case study of car-t therapy for children and young adults with relapsed or refractory acute lymphoblastic leukemia. *BMC medical research methodology*, 19(1):182, 2019.

A. Corbella, S. E. Spencer, and G. O. Roberts. Automatic zig-zag sampling in practice. *Statistics and Computing*, 32(6):107, 2022.

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

M. H. Davis. *Markov models & optimization*. Routledge, 1993.

F. De Santis, J. Mortera, and A. Nardi. Jeffreys priors for survival models with censored data. *Journal of Statistical Planning and Inference*, 99(2):193–209, 2001.

F. N. Demarqui, R. H. Loschi, D. K. Dey, and E. A. Colosimo. A class of dynamic piecewise exponential models with random time grid. *Journal of Statistical Planning and Inference*, 142(3):728–742, 2012.

N. Demiris, D. Lunn, and L. D. Sharples. Survival extrapolation using the poly-weibull model. *Statistical Methods in Medical Research*, 24(2):287–301, 2015.

P. Diaconis, S. Holmes, and R. M. Neal. Analysis of a nonreversible markov chain sampler. *Annals of Applied Probability*, pages 726–752, 2000.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

L. Fahrmeir and S. Lang. Bayesian inference for generalized additive mixed models based on markov random field priors. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 50(2):201–220, 2001.

M. F. Faulkner and S. Livingstone. Sampling algorithms in statistical physics: a guide for statistics and machine learning. *Statistical Science*, 39(1):137–164, 2024.

P. Fearnhead and Z. Liu. Efficient bayesian analysis of multiple changepoint models with dependence across segments. *Statistics and Computing*, 21(2):217–229, 2011.

P. Fearnhead, J. Bierkens, M. Pollock, and G. O. Roberts. Piecewise deterministic markov processes for continuous-time monte carlo. *Statistical Science*, 33(3): 386–412, 2018.

P. Fearnhead, C. Nemeth, C. J. Oates, and C. Sherlock. Scalable monte carlo for bayesian learning. *arXiv preprint arXiv:2407.12751*, 2024.

P. Feigl and M. Zelen. Estimation of exponential survival probabilities with concomitant information. *Biometrics*, pages 826–838, 1965.

T. E. Fjelde, K. Xu, D. Widmann, M. Tarek, C. Pfiffer, M. Trapp, S. D. Axen, X. Sun, M. Hauru, P. Yong, W. Tebbutt, Z. Ghahramani, and H. Ge. Turing.jl: a general-purpose probabilistic programming language. *ACM Trans. Probab. Mach. Learn.*, Feb. 2025. doi: 10.1145/3711897.

J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman. Visualization in bayesian workflow. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(2):389–402, 2019.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

A. Gelman, W. R. Gilks, and G. O. Roberts. Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1): 110–120, 1997.

C. L. Gibbons and N. R. Latimer. Prevalence of immature survival data for anticancer drugs presented to the national institute for health and care excellence between 2018-2022. *Value in Health*, 2024.

W. R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):337–348, 1992.

E. Goan, D. Perrin, K. Mengersen, and C. Fookes. Piecewise deterministic markov processes for bayesian neural networks. In *Uncertainty in Artificial Intelligence*, pages 712–722. PMLR, 2023.

J. P. Gosling. Shelf: the sheffield elicitation framework. In *Elicitation: The science and art of structuring judgement*, pages 61–93. Springer, 2017.

P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

P. Guyot, A. Ades, M. J. Ouwens, and N. J. Welton. Enhanced secondary analysis of survival data: reconstructing the data from published kaplan-meier survival curves. *BMC medical research methodology*, 12:1–13, 2012.

P. Guyot, A. E. Ades, M. Beasley, B. Lueza, J.-P. Pignon, and N. J. Welton. Extrapolation of survival curves from cancer trials using external information. *Medical Decision Making*, 37(4):353–366, 2017.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

M. Hird, S. Livingstone, and G. Zanella. A fresh take on 'barker dynamics' for mcmc. In *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 169–184. Springer, 2020.

B. P. Hobbs, R. C. Pestana, E. C. Zabor, A. M. Kaizer, and D. S. Hong. Basket trials: review of current practice and innovations for future trials. *Journal of Clinical Oncology*, 40(30):3520–3528, 2022.

M. D. Hoffman and A. Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 11 2014. ISSN 15337928. URL `http://arxiv.org/abs/1111.4246`.

J. G. Ibrahim, M.-H. Chen, D. Sinha, J. Ibrahim, and M. Chen. *Bayesian survival analysis*, volume 2. Springer, 2001.

Y. Iguchi, S. Livingstone, N. Nüsken, G. Vasdekis, and R.-Y. Zhang. Skew-symmetric schemes for stochastic differential equations with non-lipschitz drift: an unadjusted barker algorithm. *arXiv preprint arXiv:2405.14373*, 2024.

D. Incerti, H. Thom, G. Baio, and J. P. Jansen. R you still using excel? the advantages of modern software tools for health technology assessment. *Value in Health*, 22 (5):575–579, 2019.

C. Jackson. Multi-state models for panel data: the msm package for r. *Journal of statistical software*, 38(1):1–28, 2011.

C. Jackson. flexsurv: a platform for parametric survival modeling in r. *Journal of statistical software*, 70:1–33, 2016.

C. Jackson, J. Stevens, S. Ren, N. Latimer, L. Bojke, A. Manca, and L. Sharples. Extrapolating survival from randomized trials using external data: A review of methods. *Medical Decision Making*, 37(4):377–390, 2017. doi: 10.1177/0272989X16639900. PMID: 27005519.

C. H. Jackson. survextrap: a package for flexible and transparent survival extrapolation. *BMC Medical Research Methodology*, 23(1):282, 2023.

A. Jasra, C. Holmes, and D. Stephens. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 20(1): 50–67, 2005.

H. S. Jørgensen. The copenhagen stroke study experience. *Journal of Stroke and Cerebrovascular Diseases*, 6(1):5–16, 1996.

L. P. Kammersgaard, H. Jørgensen, J. Reith, H. Nakayama, P. Pedersen, and T. S. Olsen. Short-and long-term prognosis for very old stroke patients. the copenhagen stroke study. *Age and ageing*, 33(2):149–154, 2004.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

B. Kearns, M. D. Stevenson, K. Triantafyllopoulos, and A. Manca. Generalized linear models for flexible parametric modeling of the hazard function. *Medical Decision Making*, 39(7):867–878, 2019.

B. Kearns, M. D. Stevenson, K. Triantafyllopoulos, and A. Manca. The extrapolation performance of survival models for data with a cure fraction: a simulation study. *Value in Health*, 24(11):1634–1642, 2021.

B. Kearns, M. D. Stevenson, K. Triantafyllopoulos, and A. Manca. Dynamic and flexible survival models for extrapolation of relative survival: A case study and simulation study. *Medical Decision Making*, 42(7):945–955, 2022.

T. Kloek and H. K. Van Dijk. Bayesian estimates of equation system parameters: an application of integration by monte carlo. *Econometrica: Journal of the Econometric Society*, pages 1–19, 1978.

R. Kohn, M. Smith, and D. Chan. Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, 11(4):313–322, 2001.

J. Koskela. Zig-zag sampling for discrete structures and nonreversible phylogenetic mcmc. *Journal of Computational and Graphical Statistics*, 31(3):684–694, 2022.

H. Kozumi. Posterior analysis of latent competing risk models by parallel tempering. *Computational statistics & data analysis*, 46(3):441–458, 2004.

N. Latimer. Nice dsu technical support document 14: survival analysis for economic evaluations alongside clinical trials-extrapolation with patient-level data. *Report by the Decision Support Unit*, 2011.

N. R. Latimer. Survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. *Medical Decision Making*, 33(6):743–754, 2013.

N. R. Latimer. Response to "survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: an alternative approach" by bagust and beale. *Medical Decision Making*, 34(3):279–282, 2014.

C. Legrand. *Advanced survival models*. Chapman and Hall/CRC, 2021.

P. W. Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval research logistics quarterly*, 26(3):403–413, 1979.

E. Ley and M. F. Steel. On the effect of prior assumptions in bayesian model averaging with applications to growth regression. *Journal of applied econometrics*, 24(4):651–674, 2009.

X. Liang, S. Livingstone, and J. Griffin. Adaptive mcmc for bayesian variable selection in generalised linear models and survival models. *Entropy*, 25(9):1310, 2023.

R. Lin, P. F. Thall, and Y. Yuan. Bags: A bayesian adaptive group sequential trial design with subgroup-specific survival comparisons. *Journal of the American Statistical Association*, 116(533):322–334, 2021.

S. Livingstone and G. Zanella. The barker proposal: combining robustness and efficiency in gradient-based mcmc. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):496–523, 2022.

S. Livingstone, M. F. Faulkner, and G. O. Roberts. Kinetic energy choice in hamiltonian/hybrid monte carlo. *Biometrika*, 106(2):303–319, 2019.

F. Louzada-Neto. Polyhazard models for lifetime data. *Biometrics*, 55(4):1281–1285, 1999.

D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. The bugs project: Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067, 2009.

G. M. Martin, D. T. Frazier, and C. P. Robert. Computing bayes: From then 'til now. *Statistical Science*, 39(1):3–19, 2024.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

S. P. Meyn and R. L. Tweedie. Stability of markovian processes ii: Continuous-time processes and sampled chains. *Advances in Applied Probability*, 25(3):487–517, 1993.

M. Michel, S. C. Kapfer, and W. Krauth. Generalized event-chain monte carlo: Constructing rejection-free global-balance algorithms from infinitesimal steps. *The Journal of chemical physics*, 140(5), 2014.

M. Michel, A. Durmus, and S. Sénécal. Forward event-chain monte carlo: Fast sampling by randomness control in irreversible markov chains. *Journal of Computational and Graphical Statistics*, 29(4):689–702, 2020. doi: 10.1080/10618600.2020.1750417.

P. Mikkola, O. A. Martin, S. Chandramouli, M. Hartmann, O. A. Pla, O. Thomas, H. Pesonen, J. Corander, A. Vehtari, S. Kaski, et al. Prior knowledge elicitation: The past, present, and future. *Bayesian Analysis*, 1(1):1–33, 2023.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032, 1988.

U. B. Mogensen, H. Ishwaran, and T. A. Gerds. Evaluating random forests for survival analysis using prediction error curves. *Journal of statistical software*, 50: 1–23, 2012.

T. A. Murray, B. P. Hobbs, D. J. Sargent, and B. P. Carlin. Flexible bayesian survival modeling with semiparametric time-dependent and shape-restricted covariate effects. *Bayesian analysis (Online)*, 11(2):381, 2016.

R. M. Neal. MCMC using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113–162, 6 2011. doi: 10.1201/b10905-6.

M. A. Negrín, J. Nam, and A. H. Briggs. Bayesian solutions for handling uncertainty in survival extrapolation. *Medical Decision Making*, 37(4):367–376, 2017.

P. J. Newcombe, H. Raza Ali, F. M. Blows, E. Provenzano, P. D. Pharoah, C. Caldas, and S. Richardson. Weibull regression with bayesian variable selection to identify prognostic tumour markers of breast cancer survival. *Statistical methods in medical research*, 26(1):414–436, 2017.

A. Nishimura, Z. Zhang, and M. A. Suchard. Zigzag path connects two monte carlo samplers: Hamiltonian counterpart to a piecewise deterministic markov process. *Journal of the American Statistical Association*, 120(550):1077–1089, 2025.

J. E. Oakley, S. Ren, J. E. Forsyth, J. P. Gosling, K. Wilson, N. Latimer, M. J. Rutherford, L. Uttley, and J. Fotheringham. Nice dsu technical support document 26: Expert elicitation for long-term survival outcomes. Technical Support Document 26, Decision Support Unit, National Institute for Health and Care Excellence (NICE), 2025.

B. Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

F. Pagani, A. Chevallier, S. Power, T. House, and S. Cotter. Nuzz: Numerical zig-zag for general models. *Statistics and Computing*, 34(1):61, 2024.

S. Palmer, Y. Lin, T. G. Martin, S. Jagannath, A. Jakubowiak, S. Z. Usmani, N. Buyukkaramikli, H. Phelps, R. Slowik, F. Pan, et al. Extrapolation of survival data using a bayesian approach: a case study leveraging external data from cilta-cel therapy in multiple myeloma. *Oncology and Therapy*, 11(3):313–326, 2023.

P. H. Peskun. Optimum monte-carlo sampling using markov chains. *Biometrika*, 60 (3):607–612, 1973.

E. Platen and N. Bruti-Liberati. *Numerical solution of stochastic differential equations with jumps in finance*, volume 64. Springer Science & Business Media, 2010.

N. G. Polson and J. G. Scott. On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012.

J. Revels, M. Lubin, and T. Papamarkou. Forward-mode automatic differentiation in julia. *arXiv preprint arXiv:1607.07892*, 2016.

S. Richardson and P. J. Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(4):731–792, 1997.

G. Roberts and J. Rosenthal. Geometric ergodicity and hybrid markov chains. *Electronic Communications in Probability*, 1997.

G. O. Roberts and J. S. Rosenthal. Quantifying the speed-up from non-reversibility in mcmc tempering algorithms. *arXiv preprint arXiv:2501.16506*, 2025.

G. O. Roberts and L. M. Sangali. Latent diffusion models for survival analysis. *Bernoulli*, 16(2):435 – 458, 2010. doi: 10.3150/09-BEJ217.

G. O. Roberts and R. L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(3):341–363, 1996.

M. J. Rutherford, P. C. Lambert, M. J. Sweeting, R. Pennington, M. J. Crowther, K. R. Abrams, et al. Nice dsu technical support document 21: flexible methods for survival analysis. *Decision Support Unit, ScHARR, University of Sheffield*, 2020.

J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of computational physics*, 23(3):327–341, 1977.

M. Sachs, D. Sen, J. Lu, and D. Dunson. Posterior computation with the gibbs zig-zag sampler. *Bayesian Analysis*, 18(3):909–927, 2023.

E. Sharef, R. L. Strawderman, D. Ruppert, M. Cowen, and L. Halasyamani. Bayesian adaptive b-spline estimation in proportional hazards frailty models. *Electronic Journal of Statistics*, 4:606–642, 2010.

D. Simpson, H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye. Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science*, 32(1):1 – 28, 2017. doi: 10.1214/16-STS576.

F. Soikkeli, M. Hashim, M. Ouwens, M. Postma, and B. Heeg. Extrapolating survival data using historical trial–based a priori distributions. *Value in Health*, 22(9): 1012–1017, 2019.

Stan Development Team. RStan: the R interface to Stan, 2025. URL `https://mc-stan.org/`. R package version 2.26.24.

M. Stephens. Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *Annals of statistics*, pages 40–74, 2000.

M. Sutton and P. Fearnhead. Concave-convex pdmp-based sampling. *Journal of Computational and Graphical Statistics*, 32(4):1425–1435, 2023.

A. R. Thatcher. The long-term pattern of adult mortality and the highest attained age. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1): 5–43, 1999.

L. Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

I. R. Timmins, F. Torabi, C. H. Jackson, P. C. Lambert, and M. J. Sweeting. Simulation-based assessment of a bayesian survival model with flexible baseline hazard and time-dependent effects. *arXiv preprint arXiv:2503.21388*, 2025a.

I. R. Timmins, F. Torabi, C. H. Jackson, P. C. Lambert, and M. J. Sweeting. A simulation and case study to evaluate the extrapolation performance of flexible bayesian survival models when incorporating real-world data. *arXiv preprint arXiv:2505.16835*, 2025b.

R. Tsai, L. K. Hotta, et al. Polyhazard models with dependent causes. *Brazilian Journal of Probability and Statistics*, 27(3):357, 2013.

I. van Oostrum, M. Ouwens, A. Remiro-Azócar, G. Baio, M. J. Postma, E. Buskens, and B. Heeg. Comparison of parametric survival extrapolation approaches incorporating general population mortality for adequate health technology assessment of new oncology drugs. *Value in Health*, 24(9):1294–1301, 2021.

P. Vanetti, A. Bouchard-Côté, G. Deligiannidis, and A. Doucet. Piecewise-deterministic markov chain monte carlo. *arXiv preprint arXiv:1707.05296*, 2017.

G. Vasdekis. *On zig-zag extensions and related ergodicity properties*. PhD thesis, University of Warwick, 2021.

A. Vehtari, A. Gelman, and J. Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27:1413–1432, 2017.

C. T. Volinsky and A. E. Raftery. Bayesian information criterion for censored survival models. *Biometrics*, 56(1):256–262, 2000.

C. Williams, J. D. Lewsey, D. F. Mackay, and A. H. Briggs. Estimation of survival probabilities for use in cost-effectiveness analyses: a comparison of a multi-state modeling survival analysis approach with partitioned survival and markov decision-analytic modeling. *Medical Decision Making*, 37(4):427–439, 2017.

C. Wu and C. Robert. Generalized bouncy particle sampler. *Statistics and Computing*, 2019.

C. Wu and C. P. Robert. Coordinate sampler: a non-reversible gibbs-like mcmc sampler. *Statistics and Computing*, 30(3):721–730, 2020.

J. Yang, K. Łatuszyński, and G. O. Roberts. Stereographic markov chain monte carlo. *The Annals of Statistics*, 52(6):2692–2713, 2024.

G. Zanella. Informed proposals for local mcmc in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865, 2020.