Bayesian inference and prediction for the inhomogeneous Poisson process, and a robust competitor to Hamiltonian Monte Carlo

Szymon Urbas, B.Sc.(Hons.), M.Res

Submitted for the degree of Doctor of Philosophy at

Lancaster University.

September 2022

STOR-i
excellence with impact

Abstract

Modern Bayesian statistical methods utilise a plethora of mathematical and computational techniques to tackle real-world problems. The probabilistic formulation of a given model and its parameters can be used for making reliable predictions about future observations whilst accounting for parameter uncertainty. This, however, comes at a cost where the posterior distribution of the model parameters is not always tractable. To overcome intractability, computationally-intensive methods for Bayesian inference are employed; often, these need to be picked on a case-by-case basis, with algorithms themselves requiring careful tuning of the respective tuning parameters. This thesis explores three separate problems in Bayesian modelling and inference. In all three parts, the proposed solutions are based on computational methods relying on the simulation of particular stochastic processes.

The first contribution tackles the problem of modelling and predicting recruitment to clinical trials. Phase III trials usually involve large recruitment drives across hundreds of centres to enrol a target number of patients in a prespecified period. It is crucial to accurately monitor the recruitment process at interim points of a study. Early detection of poor recruitment can inform trial organisers when making changes

to the protocol. We introduce a new flexible hierarchical model for multi-centre recruitment based on the inhomogeneous Poisson process. The simple, yet flexible, form of the model allows for efficient and user-friendly inference carried out in the Bayesian paradigm. The proposed framework outperforms state-of-the-art methods for recruitment prediction and is robust to model misspecifications.

The second contribution is in the area of exact Cox process inference. A Cox process is a Poisson point process where the intensity itself is stochastic. Bayesian inference for Cox process models is an inherently difficult problem as the likelihood function of the intensity given observed data is not available in a closed form; this is known as a doubly-intractable problem. We introduce a novel unbiased estimator of the Cox process likelihood for processes with bounded intensity. The estimator can be implemented inside a random-weight particle filter to carry out exact inference (in a Monte Carlo sense) on the intensity function. In addition, it allows for efficient inference of the model hyperparameters through the pseudo-marginal Metropolis-Hastings algorithm.

The third contribution is a competitor to the Hamiltonian Monte Carlo (HMC) algorithm. HMC is commonly used for problems in computational physics and Bayesian inference. It is a gradient-based algorithm, ideal for sampling from complex and high-dimensional targets. The main caveat, however, is the algorithm's sensitivity to one of the tuning parameters. In this thesis, we introduce a new algorithm, the *Apogee to Apogee Path Sampler*, which is based on HMC and thus benefits from many of its desirable properties. We show it has comparable efficiency to HMC but is much more robust to the misspecification of its tuning parameters.

Acknowledgements

First and foremost, I would like to thank Professor Chris Sherlock for his tremendous patience and wisdom in guiding me through the PhD, and helping me become a better researcher. It was an incredible privilege to have worked with him on such a variety of problems.

I would also like to thank Dr Paul Metcalfe for organising the project with AstraZeneca in the first place, and for all the helpful comments and suggestions on the industry part of the project. Over at AstraZeneca, I would like to thank Dr Aris Perperoglou for his valuable insights, as well as for organising my internship. It was a great opportunity to get some hands-on experience.

My PhD was supported by the STOR-i Centre for Doctoral Training as funded by EPSRC. I am grateful to AstraZeneca for providing additional funding.

I am extremely grateful to everyone at STOR-i who made both the MRes and the PhD possible. In particular, Jon, Kevin and Idris for their ongoing commitment to running the centre and ensuring everyone comes out of the programme with a wide range of skill.

My time at STOR-i would not be the same without the people in my cohort

(in alphabetical order): Alan, Anja, Harry, Holly, Jess, Jordan, Livia, Matt, Mike, Mirjam, Nicola, Srsthi and Tom. I could not have asked for a better combination of people to spend time with.

I would like to express my gratitude to my viva examiners, Prof Andrew Titman and Prof Alexandros Beskos, for taking the time to read this thesis, as well as for their valuable comments, insights and suggestions.

Finally, I would like to thank Lady for her endless encouragement over all these years. I cannot overstate how crucial her support was in me finishing this.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

A version of Chapter 3 has been published as *Interim recruitment prediction for multi-center clinical trials* by Szymon Urbas, Chris Sherlock and Paul Metcalfe in Biostatistics.

Chapter 4 is original work under the supervision of Prof Chris Sherlock, and is yet to be submitted for publication.

A version of Chapter 5 has been submitted for publication and is available as a preprint *The Apogee to Apogee Path Sampler* by Chris Sherlock, Szymon Urbas and Matthew Ludkin. The publication is currently under revision.

Szymon Urbas

Contents

List of Figures

List of Tables

List of Abbreviations

a.s.	almost surely
AAPS	the Apogee to Apogee Path Sampler
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BST	bootstrapped test
CI	credible interval
ESS	effective sample size
GP	Gaussian process
HMC	Hamiltonian Monte Carlo
HPP	homogeneous Poisson process
i.i.d.	independent and identically distributed
LRT	likelihood-ratio test
MCMC	Markov chain Monte Carlo
MH	Metropolis-Hastings
MLE	maximum liklihood estimator
NUTS	the no-U-turn sampler

OU	Ornstein-Uhlenbeck
PE	Poisson estimator
PF	particle filter
PMMH	pseudo-marginal Metropolis-Hastings
RBTE	Rao-Blackwellised Thinning Estimator
RWM	random-walk Metropolis
RWPF	random-weight particle filter
SDE	stochastic differential equation
SGCP	sigmoidal Gaussian Cox process
SIR	sampling importance resampling
SMC	sequential Monte Carlo
SSM	state-space model

Chapter 1

Introduction

Modern problems in Bayesian statistics come in many forms. In this thesis, I explore three problems in Bayesian statistics ranging from the use of Bayesian modelling in clinical trial recruitment prediction, to exact Bayesian inference for a challenging class of point process models, to even general methods for sampling from intractable Bayesian posteriors. Each of the three projects is contained within a separate chapter in the main body of this thesis; as such, the chapters can be read in any order.

1. The first project deals with the problem of predicting recruitment to clinical trials. The focus is on developing a hierarchical model based on the inhomogeneous temporal Poisson process. (Chapter 3)

2. The second project focuses on the methodology for exact Bayesian posterior inference of the temporal Cox process. This is carried out with the aid of a novel Cox process likelihood estimator. (Chapter 4)

3. The third project deals with the problem of tuning the Hamiltonian Monte Carlo

algorithm which is often used for high-dimensional Bayesian inference problems. (Chapter 5)

The direction of the second project was in part inspired by the work carried out during the first project — the use of Cox processes to model recruitment is a suggested extension at the end of Chapter 3, and Chapter 4 addresses the immediate concerns with inference for such models, without being a direct sequel. In both chapters, the proposed methods rely on the use of stochastic processes. The models used throughout those two chapters involve relatively few hyperparameters which allows more straightforward posterior inference methods, such as importance sampling or the random walk Metropolis algorithm. Some of the suggested extensions involve the use of models involving a larger number of parameters. Carrying out efficient Bayesian inference of a large number of parameters can be a challenging task as inference algorithms can be particularly sensitive to the choice of tuning parameters. The work in Chapter 5 addresses this exact issue by introducing a novel algorithm that itself is robust to tuning misspecification. The general ideas introduced and developed in the chapter could be incorporated into the methods introduced in the previous chapters. More generally, all three projects can be put under the general umbrella of computationally-intensive Bayesian methods with easy-to-follow guidelines for the end user.

1.1 Contributions and thesis outline

Chapter 2: Background material

This chapter contains a literature review of the key topics referred to in the thesis. It discusses the relevant operational aspects of clinical trials which serves as a context for the statistical model introduced in Chapter 3. It covers required background material on stochastic processes, Bayesian inference and Monte Carlo methods which are used throughout the thesis.

Chapter 3: Interim recruitment prediction for multi-centre clinical trials

In this chapter, we introduce a general framework for monitoring, modelling, and predicting the recruitment to multi-centre clinical trials. The work is motivated by overly optimistic and narrow prediction intervals produced by existing time-homogeneous models for multi-centre recruitment. We first present two tests for the detection of decay in recruitment rates, together with a power study. We then introduce a model based on the inhomogeneous Poisson process with monotonically decaying intensity, motivated by recruitment trends observed in oncology trials. The general form of the model permits adaptation to any parametric curve-shape. A general method for constructing sensible parameter priors is provided and Bayesian model averaging is used for making predictions which account for the uncertainty in both the parameters and the model. The validity of the method and its robustness to misspecification are tested using synthetic datasets. The new methodology is then applied to oncology trial data, where we make interim accrual predictions, comparing them to those ob-

tained by existing methods, and indicate where unexpected changes in the accrual pattern occur.

Chapter 4: Exact sequential inference for bounded-intensity Cox processes

Temporal point processes are widely used to model phenomena in areas such as finance, biology or geology. One general model is the sigmoidal Gaussian Cox process (SGCP), first introduced in Adams et al. (2009), where a transformation of the intensity function governing data-generation is assigned a Gaussian process prior. This introduces a great deal of modelling flexibility; the setup does not require any prespecified parametric forms for the intensity shape. However, the resulting posterior is doubly-intractable, as the likelihood function has no closed form, and so various data augmentation schemes are needed to perform exact Bayesian inference on the intensity. The Markov chain Monte Carlo algorithms used for inference involve steps on spaces with varying dimensions which can be slow and difficult to tune.

We introduce a novel unbiased estimator of the Cox process likelihood, which we call the *Rao-Blackwellised Thinning Estimator*. It is constructed by simulating the thinning procedure conditional on the observed point process data. The proposed estimator is a generalisation of the existing *Poisson estimator* (Wagner, 1989; Beskos et al., 2006) in the context of bounded-intensity Cox process inference. The estimator is used within a random-weight particle filter (RWPF) algorithm to perform posterior inference on the intensity subject to a diffusion process prior. The inference is exact, with no need for data augmentation. Numerical experiments show that, when used within a RWPF, the proposed estimator outperforms the Poisson estimator in terms

of estimator variance for a given computational budget. Additionally, we employ a pseudo-marginal Metropolis-Hastings algorithm to sample from the marginal posterior of the model hyperparameters, bypassing the high posterior correlation between intensity shape and the parameters.

Chapter 5: The Apogee to Apogee Path Sampler

Hamiltonian Monte Carlo (HMC) is a general-purpose Markov chain Monte Carlo (MCMC) algorithm for sampling from complex, high-dimensional posterior distributions on $\mathbb{R}^d$ when the log-posterior gradients are available. It treats the negative log-posterior density as a potential surface, $U(x) = -\log \pi(x)$, and introduces a fictitious momentum variable. A proposal is constructed by applying the *leapfrog* integrator to simulate the Hamiltonian dynamics for a time $\mathcal{T}$, given an initial position-momentum pair; the proposals then undergo an appropriate accept-reject step. The strengths of HMC are the potentially high acceptance rate ($\approx 65.1\%$) for distant proposals, and the scaling of the efficiency in dimension d being $\mathcal{O}(d^{1/4})$. A major drawback of the algorithm is its sensitivity to the choice of the (continuous) integration-time tuning parameter $\mathcal{T}$, where slight misspecification can lead to drastic drops in efficiency.

In the chapter, we introduce a new MCMC algorithm based on HMC, which we call the *Apogee to Apogee Path Sampler* (AAPS). The algorithm constructs a set of points along the path of the approximate Hamiltonian dynamics, both forward and backward in time. The points are enclosed by local maxima in the potential along the path, or *apogees*. Due to the reversibility of the leapfrog integrator, this construction is invariant to the initial position-momentum pair from the constructed path. The

tuning parameter is the (discrete) number of apogees, K, contained within the path. We introduce a generic methodology for proposing an element from the path and then accepting or rejecting it, which produces samples from the target distribution. Numerical experiments in the chapter illustrate how AAPS achieves similar efficiency to HMC on a wide variety of target distributions, but it is much more robust to the misspecification of its equivalent tuning parameter.

Chapter 6: Conclusions

This chapter concludes the work carried out over the course of this thesis. It provides the final connections between the three separate pieces of work by providing suggestions for future research directions based on the intersections of the different problem areas.

Chapter 2

Background material

This chapter serves as a review of the relevant literature for the work contained within this thesis. Section 2.1 outlines the key aspects of clinical trials which serves as context for the work carried out in Chapter 3. Section 2.2 provides background on stochastic processes which underpin the statistical models and methodology developed within all chapters of the thesis. Similarly, all chapters involve Monte Carlo methods in the context of Bayesian inference; Section 2.4 provides the required background on those. Finally, Section 2.5 outlines a general state-space model formulation and the Bayesian filtering problem which are encountered in Chapter 4 of the thesis; relevant details on inference using a particle filter are also provided.

2.1 Clinical trials

Novel developments in medical treatment must be proven to be both safe and efficacious against a prespecified condition before being administered to the general

population. The gold standard for showing this is through *clinical trials* (see, for example, Pocock, 1983; Friedman et al., 1998). These are a series of prospective studies comparing the safety and efficacy of new treatments (drugs or procedures) against a control in humans. The prospective part of the study refers to a fixed experimental design finalised before any experimentation takes place, and the state of the volunteering participants is followed forward in time. The analysis of the experiment results involves pre-specified statistical tests which, for the benefit of the trial organiser, must have sufficiently high statistical power, usually at least 80% (e.g., Sakpal, 2010). This usually requires the studies to involve large numbers of participants from many different demographics. The new experimental treatments are rarely risk-free and so to ensure patient safety, the study is split up into a number of *phases*. Each phase aims to satisfy different objectives, and as the objectives increase in complexity the number of required participants also increases. The trials can generally be divided up into:

- **Phase 0:** If the new treatment is a novel drug, very low doses of it are administered to a small group of patients (usually ≤ 20). The aim is to determine the pharmacokinetic properties of the drug; that is, how it behaves once inside the human body.

- **Phase I:** The treatment is tested on 20-100 volunteers. The individuals can have different, but similar conditions (say, different types of cancer) and the goal of the phase is to identify common side effects and determine safe dosage levels.

- **Phase II:** The treatment is tested on 100-300 patients with similar diseases

(usually one or two types). This aims to establish the proposed therapeutic effect of the treatment; specifically, what type of disease it targets, and learn more about the side effects.

- **Phase III:** This phase involves large randomised control studies in order to determine the efficacy of the new treatment compared to the current best option. To ensure the comparisons use tests with high statistical power, large numbers (300-3000) of patients with the specific disease across many different countries are recruited into the trial.

- **Phase IV:** Once the treatment is open to the public, confirmatory trials are carried out to identify potential long-term or rare side effects.

These specific recruitment numbers above were taken from Cancer Research UK (2022) but can differ based on the therapeutic area; cardiovascular trials typically recruit a lot more volunteers than oncology. Trials sometimes span more than one phase; for example, Phase II/III trials involve a seamless transition between the second and the third phases.

The trials are a tremendous undertaking both in the time scale and the costs; it can take between 10 to 15 years for a treatment to move from initial discovery to approval and licencing, and the costs can vary from around $500 million to over $2 billion, for a given therapy or developing firm (Adams and Brantner, 2006). It is crucial that the new treatments are examined fairly and the results are reported with sufficiently high certainty, adhering to specific guidelines from the overseeing agencies (European Medicine Agency in Europe, and Food and Drug Administration in the

United States). In an independent review, Getz and Lamberti (2013) reports that 48% of examined trials failed to meet recruitment targets in the allotted time which led to numerous extensions of trial timelines.

The focus of the work in Chapter 3 of this thesis is to address the issues of monitoring the process of recruiting patients during Phase III. The recruitment is carried out across hundreds of *centres* which can be hospitals or clinics; occasionally those are referred to as *sites*. As such, these trials are commonly referred to as *multi-centre* trials. To speed up the recruitment, a trial organiser may bring more centres into the study and this in itself is a very costly process; initiating a new recruitment centre involves employing and training new staff, advertising, and adjusting the existing drug-supply chains. It is therefore crucial that under-recruitment can be clearly identified at interim points of a study.

The source of the problem of overpredicting recruitment rates stems from the overly optimistic initial estimates for the pool of suitable patients for the trial. This phenomenon was first noted in Lasagna (1979) and has since been referred to as "Lasagna's Law" in literature,

> *"The incidence of patient availability sharply decreases when a clinical trial begins and returns to its original level as soon as the trial is completed"*(Nahler, 2009).

Statistical models for monitoring the recruitment process additionally make very strong assumptions of time-homogeneity in the recruitment rates; that is, the average monthly recruitment rate should not vary from one month to another. Figure 2.1.1 (as taken from Anisimov, 2009) gives an example of the accrual (cumulative recruitment) in a small study along with interim predictions from three periods. The

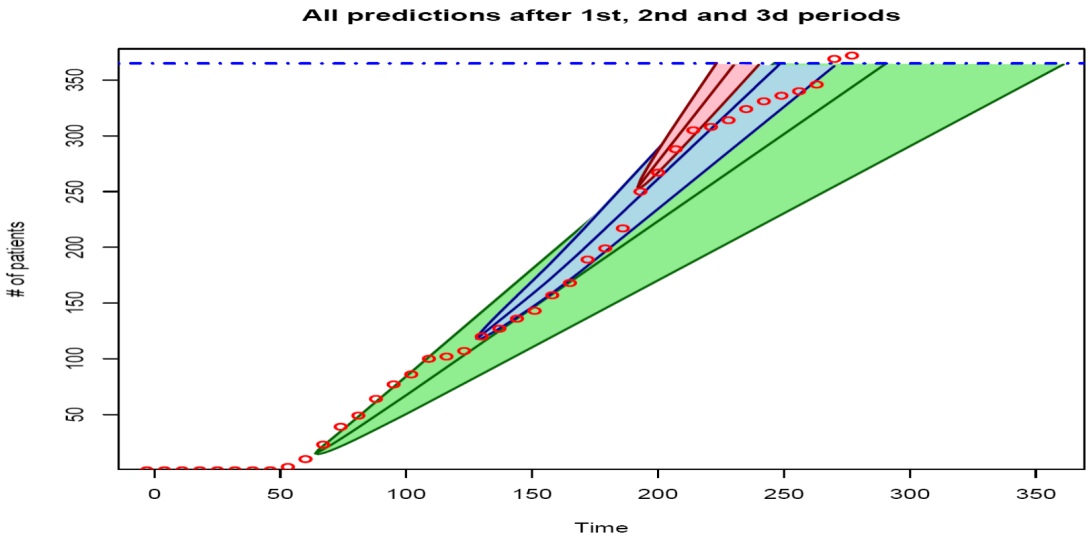


Figure 2.1.1: Example of patient accrual to a small clinical trial. Accrual numbers are indicated by "∘" and coloured bands provide 90% pointwise prediction intervals; point predictions are given by the middle lines. The horizontal dashed line is the target number of 366 recruitments. The figure was taken from Anisimov (2009).

initial plan of the study was to recruit 366 patients across 75 centres. However, due to poor initial recruitment estimates, a total of 119 centres needed to be initiated. Furthermore, the model predictions were getting increasingly worse indicating fundamental model misspecifcations; the specific model used is from Anisimov and Fedorov (2007) and is discussed in detail in Chapter 3.

In Chapter 3, we introduce a novel modelling framework for multi-centre clinical trials which can be used for monitoring trial recruitment by making predictions forward in time during interim analyses. Several relevant existing methods are outlined at the beginning of Chapter 3. A much more detailed review of the history of recruitment prediction methodology can be found in the MRes dissertation, Urbas (2018).[1]

[1]Available at https://www.lancaster.ac.uk/~urbas/SzymonUrbas_PhDProp_revised.pdf

2.2 Stochastic processes

The models and methodology developed in this thesis require a certain amount of background on stochastic processes. Here, we outline the key relevant concepts without delving into rigorous measure theory; for a more fundamental introduction see, for example, Doob (1953); Ross et al. (1996); Del Moral and Penev (2017).

A stochastic process $\{X_t,\, t \in \mathcal{T}\}$ is a collection of random variables on a common probability space $(\mathsf{X}, \mathcal{A}_{\mathsf{X}}, \mathbb{P})$, where $\mathcal{T}$ is a time domain which can be either discrete or continuous. We let $\mathcal{F}_t$ denote the σ-algebra generated by $\{X_s : 0 \leq s \leq t\}$. The process X_t is Markov; if conditional on all the information available up to and including the present state, the future of the process solely depends on the present state and is independent of the previous history. Formally, for all $A \in \mathcal{A}_{\mathsf{X}}$, $s \in [0, t]$,

$$\mathbb{P}\left(X_t \in A | \mathcal{F}_s\right) = \mathbb{P}\left(X_t \in A | X_s\right).$$

Markov processes on discrete time domains are often referred to as *Markov chains*.

2.2.1 Poisson processes

The inhomogeneous Poisson process underpins the models and methodology introduced in Chapters 3 and 5. In this thesis, we solely focus on the temporal Poisson process, that is, $\mathcal{T} = [0, \infty)$.

Definition 2.2.1 (Poisson Process)**.** *A Poisson process* $\{N_t,\, t \geq 0\}$ *with intensity* $\lambda(t) > 0$*, satisfies the following*

(i) $N_0 = 0$ *almost surely;*

(ii) for any $t > s \geq 0$, $(N_t - N_s) \sim \mathsf{Pois}\left(\int_s^t \lambda(u)\, \mathsf{d}u\right)$;

(iii) for any $v > u \geq t > s \geq 0$, $(N_v - N_u)$ *is independent of* $(N_t - N_s)$;

(iv) N_t *is right-continuous with left limits.*

As a consequence of condition (iii), the Poisson process is Markov. We say the process is homogeneous if $\lambda(t) \equiv \lambda_0$, for $\lambda_0 > 0$. The process is discrete-stable, that is, for any two Poisson processes $N_t^{(1)}$ and $N_t^{(2)}$, with respective intensities λ_1 and λ_2, the sum $N_t^{(1)} + N_t^{(2)}$ is itself a Poisson process with intensity $\lambda_1 + \lambda_2$. A realisation (sample path) of the process N_t is usually described through a time sequence of *arrivals* or *event times* $T_1, T_2, \ldots$, where $T_i = \mathsf{inf}\{t : N_t \geq i\}$, with the convention $T_0 = 0$. Letting Y_i be the length of the "gap" between T_{i-1} and T_i, $i = 1, 2, \ldots$, we can derive the inter-arrival distribution,

$$\begin{aligned}
\mathbb{P}(Y_i \leq y) &= 1 - \mathbb{P}(Y_i < y) \\
&= 1 - \mathbb{P}\left(N_{T_{i-1}+y} - N_{T_{i-1}} = 0\right) \\
&= 1 - \mathsf{exp}\left(-\int_{T_{i-1}}^{T_{i-1}+y} \lambda(u)\, \mathsf{d}u\right),
\end{aligned}$$

and so the density of Y_i is

$$f_{Y_i}(y) = \lambda(T_{i-1} + y)\mathsf{exp}\left(-\int_{T_{i-1}}^{T_{i-1}+y} \lambda(u)\, \mathsf{d}u\right), \quad y > 0.$$

If the process is homogeneous with rate λ_0 then $Y_i \overset{\mathsf{iid}}{\sim} \mathsf{Exp}(\lambda_0)$ for all i and their sum has a gamma distribution. In this thesis, we will be particularly interested in sampling the process N_t. The simplest exact method of sampling paths of N_t is through the inversion method (e.g., Çinlar, 1975).

Theorem 2.2.2. *Let* $\Lambda_t = \int_0^t \lambda(u)\,\mathrm{d}u,\ t \geq 0$ *be the integrated intensity function. The random variables* $T_1, T_2, \ldots$ *are event times corresponding to a Poisson process with intensity* $\lambda(t)$ *if and only if* $\Lambda_{T_1}, \Lambda_{T_2}, \ldots$ *are event times corresponding to a homogeneous Poisson process with unit rate.*

The consequence of the theorem is that the event times can be obtained by simulating $\mathsf{Exp}(1)$ random variables and applying the inverse transform Λ^{-1}, using numerical methods if the inverse not available in closed form.

To simulate the process on some closed interval $[0, \tau]$ we can use the *thinning* procedure (Lewis and Shedler, 1979). The procedure does not require the inversion or even evaluation of Λ_t, instead, it simply requires an upper bound on the intensity λ in the interval. A homogeneous process with rate λ^* is first simulated, where $\lambda^* \geq \lambda(t),\ t \in [0, \tau]$. It is composed of R points $\tilde{t}_1, \ldots \tilde{t}_R$. Each point is rejected (thinned) with probability $1 - \frac{\lambda(\tilde{t}_i)}{\lambda^*},\ i = 1, \ldots, R$, and the set of the remaining points forms a realisation of the desired process; Algorithm 1.1 details the procedure.

1.1 Thinning

1. **Input:** intensity function $\lambda(t)$; dominating rate λ^*; observation interval $[0, \tau]$
2. **Output:** realisation of a Poisson process with intensity $\lambda(t)$ on $[0, \tau]$, $(n, \boldsymbol{t})$
3. Sample $R \sim \mathsf{Pois}(\lambda^* \tau)$;
4. Sample locations $\tilde{t}_i \sim \mathsf{Unif}[0, \tau],\ i = 1, \ldots, R$;
5. Reject each point with probability $1 - \frac{\lambda(\tilde{t}_i)}{\lambda^*},\ i = 1, \ldots, R$;
6. Set $\boldsymbol{t}$ to be a vector of the remaining, sorted points, and n the vector length
7. **return** $(n, \boldsymbol{t})$;

The likelihood function of λ given data $(n, \boldsymbol{t})$ produced by N_t is

$$p(n, \boldsymbol{t}|\lambda) = \exp\left(-\int_0^\tau \lambda(s)\, \mathrm{d}s\right) \prod_{i=1}^{n} \lambda(t_i).$$

2.2.2 Gaussian processes

Definition 2.2.3 (Gaussian Process)**.** *For some domain $\mathcal{T}$, $X_t \in \mathbb{R}^d$, $t \in \mathcal{T}$ is a d-dimensional Gaussian process, if for all $\{t_1, \ldots, t_k\} \subseteq \mathcal{T}$ the random vector $\boldsymbol{Z} = (X_{t_1}, \ldots, X_{t_k})^\top$ has a multivariate normal distribution. That means there exist a vector $\boldsymbol{M} \in \mathbb{R}^{dk}$ and a non-negative definite matrix $C = [c_{jm}] \in \mathbb{R}^{dk \times dk}$ such that the characteristic function of $\boldsymbol{Z}$ is*

$$\mathbb{E}\left[\exp\left(i\boldsymbol{u}^\top \boldsymbol{Z}\right)\right] = \exp\left(-\frac{1}{2}\sum_{j,m} u_j c_{jm} u_m + i\sum_j u_j M_j\right)$$

for all $\boldsymbol{u} = (u_1, \ldots, u_{dk})^\top \in \mathbb{R}^{dk}$, where $i = \sqrt{-1}$.

Similarly as in the previous section, we are only interested in Gaussian processes on the temporal domain, *i.e.*, $\mathcal{T} = [0, \infty)$; the domain could also be $\mathbb{R}^2$ for spatial process modelling (see, for example, Chiles and Delfiner, 2009). In general, temporal Gaussian processes are not Markov. The mean vector $\boldsymbol{M}$ in the above definition is constructed from some mean function $\mu(t)$, $t \in \mathcal{T}$ and the covariance matrix C is constructed using some covariance function ρ. The process is said to be stationary if the covariance is only a function of the temporal distance between the two points, that is

$$\mathrm{Cov}\left[X_s, X_r\right] = \rho(|r - s|), \quad s, r \in \mathcal{T}.$$

The covariance function ρ must itself be positive definite.

Definition 2.2.4. *The function $f : \mathbb{R} \to \mathbb{C}$ is positive definite if for any $x_1, \ldots, x_n \in \mathbb{R}$, the $n \times n$ matrix*

$$A = [a_{jk}] \quad a_{jk} = f(x_k - x_j)$$

is positive semidefinite.

We use

$$X_t \sim \mathsf{SGP}(\mu, \rho)$$

to denote that X_t is a stationary Gaussian process with mean function μ and stationary covariance function ρ.

There are several ways of showing a given function is positive definite. In this thesis, we employ a version of Bochner's Theorem (Bochner, 1933; Loomis, 1953).

Theorem 2.2.5 (Bochner's Theorem)**.** *The function $\rho : [0, \infty) \to [0, \infty)$ is positive definite if and only if its Fourier transform $\mathsf{F}\{\rho\}$ defines a positive measure on $\mathbb{R}$.*

The above theorem serves as a quick way of verifying a given function is positive definite; for example, take $\rho_1(t) = \mathsf{exp}\left(-t^2/2\right)$ and $\rho_2(t) = \cos(t)$, then

$$\begin{aligned}
\mathsf{F}\left\{\rho_1(t)\right\}(\omega) &= \mathsf{F}\left\{\mathsf{exp}\left(-t^2/2\right)\right\}(\omega) = \mathsf{exp}\left(-\omega^2/2\right), \\
\mathsf{F}\left\{\rho_2(t)\right\}(\omega) &= \mathsf{F}\left\{\cos(t)\right\}(\omega) = \sqrt{\frac{\pi}{2}}\delta_1(\omega) + \sqrt{\frac{\pi}{2}}\delta_{-1}(\omega),
\end{aligned}$$

so the functions are positive definite.

In Chapter 5, we will study a certain process which is an average of n stationary (non-GP) components, and in particular, we will want to show it converges to a stationary GP as $n \to \infty$. It is sufficient to show that the random vector of its values at a finite number of points converges in distribution to a multivariate Gaussian, and this

can be done through the Lindeberg-Feller Central Limit Theorem (Feller, 1971). Consider a triangular array of d-dimensional random variables $\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}, \ldots, \boldsymbol{X}_{nn}$, $n \geq 1$,

$$\begin{array}{l} \boldsymbol{X}_{11} \\ \boldsymbol{X}_{21}\ \boldsymbol{X}_{22} \\ \boldsymbol{X}_{31}\ \boldsymbol{X}_{32}\ \boldsymbol{X}_{33} \\ \vdots \end{array} \quad ,$$

and suppose that for each n, $\boldsymbol{X}_{n1}, \ldots, \boldsymbol{X}_{nn}$ are independent, $\mathbb{E}[\boldsymbol{X}_{ni}] = \boldsymbol{\mu}_{ni}$ and $\mathbb{V}[\boldsymbol{X}_{ni}]$ are positive definite for all $i = 1, \ldots, n$. Let

$$\begin{aligned} \boldsymbol{Y}_{ni} &= \boldsymbol{X}_{ni} - \boldsymbol{\mu}_{ni}, \\ \boldsymbol{T}_n &= \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Y}_{ni}. \end{aligned} \tag{2.2.1}$$

The *Lindeberg condition* for triangular arrays states that for all $\varepsilon > 0$,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\|\boldsymbol{Y}_{ni}\|^2 \mathbb{I}_{\left\{\|\boldsymbol{Y}_{ni}\| \geq \varepsilon \sqrt{n}\right\}} \right] = 0. \tag{2.2.2}$$

Theorem 2.2.6 (Lindeberg-Feller Central Limit Theorem)**.** *Let $\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}, \ldots, \boldsymbol{X}_{nn}$, $n \geq 1$ be a triangular array of d-dimensional random variable and let $\boldsymbol{Y}_{ni}$, $\boldsymbol{T}_n$ be defined as in (2.2.1). If*

$$\lim_{n \to \infty} \mathbb{V}[\boldsymbol{T}_n] = V$$

is positive definite and the Lindeberg condition (2.2.2) holds, then

$$\sqrt{n} \boldsymbol{T}_n \xrightarrow{\mathcal{D}} \mathsf{N}(\boldsymbol{0}, V) \quad as \quad n \to \infty.$$

2.2.3 Diffusions

In this section we introduce only the main process of constructing a diffusion; more detailed introductions appear in, for example, Øksendal (2003); Kloeden and Platen (1992). A diffusion process (often called an Itô process) is a Markov process which solves a particular stochastic differential equation (SDE). A stochastic differential equation is a differential equation where at least one term introduces noise through a stochastic process. The construction of diffusive processes is based on the classic continuous-time Wiener process.

Definition 2.2.7 (Wiener Process)**.** *A Wiener process* $\{W_t,\, t \geq 0\}$ *satisfies the following properties:*

- $W_0 = 0$ *almost surely;*
- *for any* $t > s \geq 0$, $W_t - W_s \sim \mathsf{N}(0, t-s)$*;*
- *for any* $t > s \geq u \geq 0$, $W_t - W_s$ *is independent of* W_u*;*
- *samples paths of* W_t *are continuous in* t *(almost surely).*

A d-dimensional Wiener process is a vector of d one-dimensional independent Wiener processes. Here, we focus on a one-dimensional diffusion. We use the following notation to denote discrete increments of the Wiener process:

$$\Delta W_t := \; W_{t+\Delta t} - W_t \sim \mathsf{N}(0, \Delta t).$$

Suppose we want to construct a process X_t with its dynamics described by a deterministic *drift* component $\alpha(X_t, t)$, and some noise which can depend on (X_t, t). In

a time-discretised setting, the noise can be introduced through an increment of a Wiener process scaled by some diffusivity function $\sigma(X_t, t)$ so that the variance of the process in an increment Δt is equal to $\sigma(X_t, t)^2 \Delta t$. The discrete-time evolution of X_t is described through the increment

$$\Delta X_t := \; X_{t+\Delta t} - X_t = \mu(X_t, t)\Delta t + \sigma(X_t, t)\Delta W_t, \tag{2.2.3}$$

where the increment has a normal distribution

$$\Delta X_t \sim \mathsf{N}(\alpha(X_t, t)\Delta t, \sigma(X_t, t)^2 \Delta t).$$

Taking the sum of discrete-time increments the process at some time T is

$$X_T = X_0 + \sum_{j=0}^{\lfloor T/\Delta t \rfloor} \mu(X_j, j)\Delta t + \sum_{j=0}^{\lfloor T/\Delta t \rfloor} \sigma(X_j, j)\Delta W_j.$$

Using the usual integration notation when taking the limit $\Delta t \downarrow 0$, we arrive at

$$X_T = X_0 + \int_{s=0}^{T} \alpha(X_s, s)\ \mathsf{d}s + \int_{s=0}^{T} \sigma(X_s, s)\ \mathsf{d}W_s, \tag{2.2.4}$$

where $\int_{s=0}^{t} \sigma(X_s, s)\ \mathsf{d}W_s$ is the Itô integral. The exact details of the Itô integral are outside the scope of this thesis (see, for example, Kloeden and Platen, 1992; Øksendal, 2003); however, we make use of the following two properties which hold for any deterministic function $f(t)$:

$$\mathbb{E}\left[\int_0^T f(t)\ \mathsf{d}W_t\right] = 0, \quad \text{and} \quad \mathbb{E}\left[\left(\int_0^T f(t)\ \mathsf{d}W_t\right)^2\right] = \int_0^T f(t)^2\ \mathsf{d}t, \tag{2.2.5}$$

where the second equality is known as *Itô isometry*. The integral is a limit of a linear

combination of Gaussian increments, and so it is itself Gaussian, that is,

$$\int_0^T f(t)\ \mathsf{d}W_t \sim \mathsf{N}\left(0, \int_0^T f(t)^2\ \mathsf{d}t\right).$$

An important property of the diffusion X_t is that it is Markov (e.g., Øksendal, 2003). The integrated process form (2.2.4) can be represented using a stochastic differential equation

$$\mathsf{d}X_t = \alpha(X_t, t)\ \mathsf{d}t + \sigma(X_t, t)\ \mathsf{d}W_t.$$

This itself is strictly a shorthand notation. The above only describes a one-dimensional diffusion, however, a similar sequence of steps is used to construct a d-dimensional process; now μ is a vector-valued function, $\boldsymbol{W}_t$ is a p-dimensional Wiener process and σ is a $d \times p$ matrix.

Ornstein-Uhlenbeck process

In Chapter 4, we will examine a particular diffusion called the Ornstein-Uhlenbeck (OU) process (e.g., Stroock, 2003); the work in the chapter involves the derivation of a d-dimensional process. For illustrative purposes, we outline the derivation of a one-dimensional version of the (zero-mean) OU process.

The one-dimensional OU process X_t, $t \geq 0$ satisfies the following stochastic differential equation

$$\mathsf{d}X_t = -\theta X_t\ \mathsf{d}t + \sigma\ \mathsf{d}W_t, \quad t > 0,$$

$$X_0 \sim \mathsf{N}(\mu_0, \nu_0).$$

Using a transformation $Y_t = \mathsf{e}^{\theta t} X_t$ (noting $Y_0 = X_0$) the SDE describing the evolution

of Y_t is

$$\begin{aligned} \mathsf{d}Y_t &= \mathsf{d}(\mathsf{e}^{\theta t} X_t) \\ &= \theta \mathsf{e}^{\theta t} X_t \, \mathsf{d}t + \mathsf{e}^{\theta t} \, \mathsf{d}X_t \\ &= \mathsf{e}^{\theta t} \left(\theta X_t \, \mathsf{d}t - \theta X_t \, \mathsf{d}t + \sigma \, \mathsf{d}W_t \right) \\ &= \sigma \mathsf{e}^{\theta t} \, \mathsf{d}W_t. \end{aligned}$$

The increment from time 0 to T is given by

$$Y_T - Y_0 = \int_0^T \mathsf{d}Y_t = \int_0^T \sigma \mathsf{e}^{\theta t} \, \mathsf{d}W_t.$$

From (2.2.5), we get that

$$\mathbb{E}[Y_T - Y_0] = 0 \quad \text{and} \quad \mathbb{V}\left[(Y_T - Y_0)^2\right] = \frac{\sigma^2}{2\theta} \left(\mathsf{e}^{2\theta T} - 1 \right)$$

with the difference itself having a Gaussian distribution. The transformed process at time T is $Y_T = (Y_T - Y_0) + Y_0$, where the two Gaussian terms are independent from the diffusion being Markov. Applying the inverse transform to obtain the original process, $X_t = \mathsf{e}^{-\theta t} Y_t$, we find that

$$X_T \sim \mathsf{N}\left(\mathsf{e}^{-\theta T} \mu_0, \frac{\sigma^2}{2\theta} \left(1 - \mathsf{e}^{-2\theta t} \right) + \mathsf{e}^{-2\theta t} \nu_0 \right).$$

The limiting (stationary) distribution of the process as $T \to \infty$, is $\mathsf{N}(0, \sigma^2/2\theta)$. Figure 2.2.1 shows several sample paths of the process.

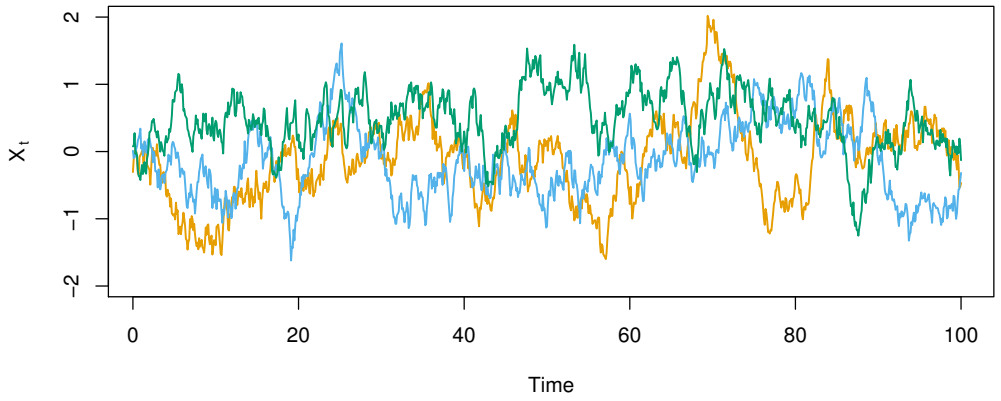


Figure 2.2.1: Sample paths of the Ornstein-Uhlenbeck process, $\theta = 0.1$, $\sigma = 0.4$; paths we initialised from the stationary distribution.

2.3 Bayesian inference problem

Given a set of data which are observations of real-world phenomena, a statistical model aims to at least approximately describe the data-generating process. Under the assumed model parametrised by an unknown set of parameters $\theta \in \Theta$, the data-generating mechanism for $\boldsymbol{Y}$ is fully encapsulated by the conditional density or mass function $p(\boldsymbol{y}|\theta)$. It is often referred to as the *likelihood function*, with notation

$$L(\theta; \boldsymbol{y}) = p(\boldsymbol{y}|\theta).$$

Through this likelihood, data $\boldsymbol{Y} = \boldsymbol{y}$ are then used to infer the unknown θ.

Statistical inference in the Bayesian paradigm involves assuming that the unknown parameter θ is itself a random variable with prespecified *prior* distribution π_0. This formulation allows for the construction of a *posterior* distribution for θ, based on

observed data $\boldsymbol{y}$ by applying Bayes' Theorem,

$$\pi(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta)\pi_0(\theta)}{\int_\Theta p(\boldsymbol{y}|\theta)\pi_0(\theta)\ \mathsf{d}\theta}. \tag{2.3.1}$$

The integral in the denominator of (2.3.1) is known as the marginal likelihood of the data. In general, it is not tractable, meaning that we can only evaluate $\pi(\theta|\boldsymbol{y})$ up to a constant of proportionality. However, in most applications, practitioners only care about quantities produced by expectations of functions over the posterior; for example,

- expectation: $\mathbb{E}[\theta]$;
- variance: $\mathbb{V}[\theta] = \mathbb{E}\left[(\theta - \mathbb{E}[\theta])(\theta - \mathbb{E}[\theta])^\top\right]$;
- probabilities: $\mathbb{P}\left(\theta \in A\right) = \mathbb{E}\left[\mathbb{I}_{\{\theta \in A\}}\right]$, $A \subset \Theta$.

More generally, we are interested in computing

$$\mathbb{E}_\pi\left[h(\theta)\right], \tag{2.3.2}$$

where h is a π-integrable function, but the exact form of the integral is not tractable.

2.4 Monte Carlo methods

In this section, we examine some of the methods of estimating integrals with respect to π of the type (2.3.2). This is carried out by producing finite Monte Carlo samples of a random variable $X \in \mathsf{X}$. For the purposes of this thesis, we assume π admits a density with respect to the Lebesgue measure, which we denote by $\pi(x)$.

Suppose we have a finite sample of independent, identically distributed (i.i.d.) random variables $\{x_i\}_{i=1}^n$. The i.i.d. sample could be produced by inversion of the distribution function (see, for example, Devroye, 1986). We denote the empirical distribution of this sample by $\hat{\pi}_{\mathsf{IID}}$, and the corresponding density is

$$\pi(x) \approx \hat{\pi}_{\mathsf{IID}}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x),$$

where δ_Y is a Dirac mass centered at Y. Thus the expectation (2.3.2) can be approximated by

$$\mathbb{E}_\pi\left[h(X)\right] \approx \mathbb{E}_{\hat{\pi}_{\mathsf{IID}}}\left[h(X)\right] = \frac{1}{n}\sum_{i=1}^{n} h(x_i).$$

Indeed it follows that the estimate is unbiased,

$$\mathbb{E}_\pi\left[\frac{1}{n}\sum_{i=1}^{n} h(X_i)\right] = \frac{n}{n}\mathbb{E}_\pi\left[h(X)\right],$$

from linearity of expectation. Additionally, by the strong law of large numbers

$$\mathbb{E}_{\hat{\pi}_{\mathsf{IID}}}\left[h(X)\right] \overset{\text{a.s.}}{\to} \mathbb{E}_\pi\left[h(X)\right] \quad \text{as} \quad n \to \infty.$$

Since the samples are i.i.d. the variance of the resulting estimator is given by

$$\mathbb{V}_\pi\left[\frac{1}{n}\sum_{i=1}^{n} h(X_i)\right] = \frac{\mathbb{V}_\pi[h(X)]}{n}.$$

If $\mathbb{V}_\pi[h(X)] < \infty$ and we can obtain the following central limit theorem (CLT) (e.g., Robert and Casella, 2005)

$$\sqrt{n}\left(\mathbb{E}_{\hat{\pi}_{\mathsf{IID}}}\left[h(X)\right] - \mathbb{E}_\pi\left[h(X)\right]\right) \overset{\mathcal{D}}{\to} \mathsf{N}\left(0, \mathbb{V}_\pi[h(X)]\right), \tag{2.4.1}$$

where $\overset{\mathcal{D}}{\to}$ denotes convergence in distribution. This indicates that the accuracy of the

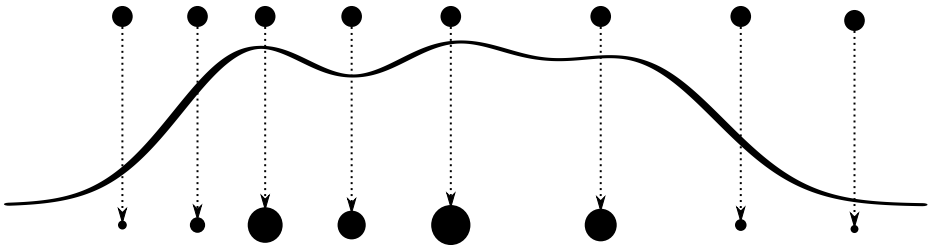


Figure 2.4.1: Illustration of importance sampling. First, samples from a tractable proposal distribution are produced. Then, each proposal is weighted relative to the density of interest. Resampling with probabilities proportional to the sample weights can be carried out at the end.

Monte Carlo approximation is independent of the dimension of X and gets better as we increase the sample size n; the rate of convergence is $\mathcal{O}(n^{-1/2})$.

2.4.1 Importance sampling

Often, we are unable to sample from π directly. One alternative is to use *importance sampling* which utilises a proposal distribution q. The proposal distribution is chosen such that it can be sampled from directly, and its support covers the support of π. Proposal q is used to produce the set of i.i.d. samples $\{x_i\}_{i=1}^n$. Defining weights for each element as

$$w(x_i) := \pi(x_i)/q(x_i), \quad i = 1, \ldots, n,$$

we have the approximate distribution $\hat{\pi}_{\mathsf{IS}}$ with

$$\hat{\pi}_{\mathsf{IS}}(x) = \frac{1}{n} \sum_{i=1}^{n} w(x_i) \delta_{x_i}(x).$$

Figure 2.4.1 illustrates the weighting procedure. Using $\hat{\pi}_{\mathsf{IS}}$ the expectation estimator has the form

$$\mathbb{E}_{\hat{\pi}_{\mathsf{IS}}}[h(X)] = \frac{1}{n} \sum_{i=1}^{n} w(x_i) h(x_i). \tag{2.4.2}$$

The unbiasedness of $\mathbb{E}_{\hat{\pi}_{\mathsf{IS}}}[h(X)]$ estimates follows from the fact that

$$\mathbb{E}_q[w(X)h(X)] = \int_{\mathsf{X}} w(x)h(x)q(x)\,\mathsf{d}x = \int_{\mathsf{X}} \frac{\pi(x)}{q(x)}h(x)q(x)\,\mathsf{d}x = \mathbb{E}_\pi[h(X)].$$

Similarly to the i.i.d. sampling estimator, The importance sampling estimator (2.4.2) will converge to $\mathbb{E}_\pi[h(X)]$ (almost surely) as $n \to \infty$ due to the strong law of large numbers.

For the estimates to have finite variance now require $\mathbb{V}_q[w(X)h(X)] < \infty$, in addition to $\mathbb{V}_\pi[h(X)] < \infty$. This can be satisfied by ensuring that q has heavier tails than π, that is,

$$\sup_{x\in\mathsf{X}} \frac{\pi(x)}{q(x)} < \infty.$$

To see this, suppose $\sup_{x\in\mathsf{X}} w(x) < K < \infty$, then

$$\mathbb{E}_q\left[w(X)^2h(X)^2\right] < K\mathbb{E}_q\left[w(X)h(X)^2\right] = \mathbb{E}_\pi\left[h(X)^2\right],$$

which is finite. The variance of the importance sampling estimator for a given h heavily depends on the choice of q. As the samples coming from q are i.i.d., we can focus on the variance of a single term in the summation (2.4.2);

$$\begin{aligned}\mathbb{V}_q[w(X)h(X)] &= \mathbb{E}_q\left[w(X)^2h(X)^2\right] - \mathbb{E}_q\left[w(X)h(X)\right]^2 \\ &= \mathbb{E}_q\left[w(X)^2h(X)^2\right] - \mathbb{E}_\pi\left[h(X)\right]^2.\end{aligned}$$

The second term is free of q so the estimator variance solely depends on the second moment of $w(X)h(X)$, $X \sim q$. Applying Jensen's inequality to the square of

$w(X)h(X)$ we have,

$$\mathbb{E}_q\left[w(X)^2h(X)^2\right] \geq \left(\mathbb{E}_q\left[w(X)|h(X)|\right]\right)^2 = \mathbb{E}_\pi\left[|h(X)|\right]^2. \tag{2.4.3}$$

This lower bound is achieved by using the *optimal proposal*

$$q(x) \propto |h(x)|\pi(x). \tag{2.4.4}$$

This result appears in Theorem 3.12 of Robert and Casella (2005). The optimal proposal requires sampling with density proportional to $|h(x)|\pi(x)$ which is infeasible in most scenarios. For Bayesian inference in Chapter 3, the same sample needs to be used for estimating expectations of several different h functions. It is sensible for those sort of situations to choose a proposal q which is similar to π in shape; ideally, the weights $w(x_i)$ would be approximately equal, so that $\{x_i\}_{i=1}^n$ is close to an i.i.d. sample from π. In Chapter 3, we employ an importance sampler where the proposal has the same mode as π and a similar covariance structure.

Direct pointwise evaluations of $\pi(x)$ are not possible in most Bayesian inference problems. Instead, $\pi(x) = f(x)/Z_f$, where $f(x)$ can be evaluated but its integral, $Z_f = \int_{\mathsf{X}} f(x)\ \mathsf{d}x$, has no closed form. In this situation, we have $w(x_i) = f(x_i)/q(x_i)$ and $\tilde{w}$ are the normalised weights,

$$\tilde{w}(x_i) = \frac{w(x_i)}{\sum_{i=1}^n w(x_j)}.$$

The resulting *normalised importance sampling* estimator is then

$$\mathbb{E}_{\hat{\pi}_{\mathsf{nIS}}}\left[h(X)\right] = \sum_{i=1}^n \tilde{w}(x_i)h(x_i) = \frac{\sum_{i=1}^n w(x_i)h(x_i)}{\sum_{i=1}^n w(x_i)}.$$

The motivation behind the normalisation step is that (setting $h(x) \equiv 1$)

$$Z_f = \mathbb{E}_q \left[f(x)/q(x) \right] \approx \frac{1}{n} \sum_{j=1}^{n} w(x_j).$$

The estimator is no longer unbiased,

$$\mathbb{E}_q \left[\tilde{w}(X) h(X) \right] \neq \mathbb{E}_\pi \left[h(X) \right],$$

but its bias decreases with increasing n. The estimator almost surely converges to the required $\mathbb{E}_\pi \left[h(X) \right]$ as $n \to \infty$, due to the strong law of large numbers (see, for example, Section 3.3.2 of Robert and Casella, 2005).

In Bayesian inference where $f(x)$ is the joint density of the parameter and the data, the average weight is an unbiased estimator of the marginal likelihood (provided it includes all multiplicative constants); see Expression (2.3.1).

Effective sample size

Since the quality of the estimates (in terms of variance) depends on the choice of proposal q, this motivates the need for a diagnostic tool of a given importance sampling scheme. This is done by quantifying the number of i.i.d. samples from π needed to obtain the same variance as the estimator using $\hat{\pi}$ made up of n samples (Kong, 1992). This is known as the *effective sample size* (ESS) and is the ratio

$$\mathsf{ESS}_{\hat{\pi}}[h(X)] := \frac{n \mathbb{V}_\pi[h(X)]}{\mathbb{V}_{\hat{\pi}}[h(X)]}. \tag{2.4.5}$$

In practice, ESS cannot be evaluated exactly, however, it can be approximated

(e.g., Kong et al., 1994; Doucet et al., 2001),

$$\mathsf{ESS}_{\hat{\pi}}[h(X)] \approx \frac{\left(\sum_{i=1}^{n} w(x_j)\right)^2}{\sum_{i=1}^{n} w(x_j)^2};$$

this is also the inverse of the second empirical moment of the normalised weights.

2.4.2 Markov chain Monte Carlo

In this section, we outline an alternative set of approaches to the i.i.d. sampling methods discussed thus far. Instead of independently producing samples from π (or some proposal distribution), a discrete-time Markov chain $\{X_i,\, i \geq 0\}$ is constructed in such a way that its stationary (and limiting) distribution is the target π. Notably, this implies that Monte Carlo samples are no longer i.i.d. but instead they form a correlated sample with the marginal distribution of X_i approaching π as $i \to \infty$. Algorithms which simulate such chains are collectively referred to as Markov chain Monte Carlo (MCMC).

As the process is Markov, the distribution of X_i given all the previous parts of the chain is fully encapsulated by the most recent value $X_{i-1} = x_{i-1}$. Formally,

$$p(x_i|x_{1:i-1}) = p(x_i|x_{i-1}) = K(x_{i-1}, x_i),$$

which defines a Markov kernel. The transition kernel K is homogeneous so that it is free of the particular position in the chain, i. The kernel is said to be π-invariant if $X_i \sim \pi \Rightarrow X_{i-1} \sim \pi,\ i > 0$. The Markov chain is stationary if the distribution of its

state does not change from one time point to the next, which formally is written as

$$\pi(x^*) = \int_{\mathsf{X}} K(x, x^*)\pi(x)\, \mathsf{d}x \quad x^* \in \mathsf{X}.$$

Definition 2.4.1 (Reversibility)**.** *Let* $\{X_i : i \geq 0\}$ *be a Markov chain with some transition kernel* $K(x, x^*)$*. The chain is reversible with respect to* π *if the transition kernel satisfies*

$$K(x, x^*)\pi(x) = K(x^*, x)\pi(x^*). \tag{2.4.6}$$

Reversibility, as defined above, is also known as *detailed balance*. It implies that, at stationarity, the probability of being at x and moving $x \to x^*$ must be equal to being at x^* and moving $x^* \to x$. If the Markov chain satisfies the detailed balance condition with respect to π, then we have

$$\int_{\mathsf{X}} K(x, x^*)\pi(x)\, \mathsf{d}x = \int_{\mathsf{X}} K(x^*, x)\pi(x^*)\, \mathsf{d}x = \pi(x^*) \int_{\mathsf{X}} K(x^*, x)\, \mathsf{d}x = \pi(x^*),$$

where the last equality follows from the fact that $K(x^*, x) = p(x|x^*)$ is a (normalised) conditional density. Detailed balance with respect to π implies that π is a stationary distribution of the chain.

Provided the chain satisfies two sufficient conditions, it will converge to its limiting distribution which is equal to the stationary π. The first is *aperiodicity*, which implies the time length of the passage between groups of states follows no deterministic pattern; an example of a periodic chain would be one where the return time to some state $x^\dagger$ is always a multiple of integer $k \geq 2$. For the algorithms appearing in this thesis, the condition can be satisfied if there is a non-zero probability of the event $X_t = X_{t-1}$. The second condition is *Harris recurrence* which ensures that the limiting

behaviour is the same for every starting position of the chain X_0, that is, there is a unique limiting distribution.

Any chain which satisfies those two conditions will eventually produce samples from the stationary target distribution π, and, by the ergodic theorem (e.g., Theorem 6.63 of Robert and Casella, 2005), the sample mean of some π-integrable function $h(X)$ will converge to $\mathbb{E}_\pi[h(X)]$ almost surely.

In practice, the chain may start from an initial value in the tails of π. In this scenario, the process will gradually make its way toward the main mass of the distribution. This transient part of some length $m > 0$ is called the *burn-in* and is discarded to reduce the bias in the estimators. The estimator of $\mathbb{E}_\pi[h(X)]$ is constructed from the empirical average of a finite part of the chain at stationarity,

$$\hat{h}_n = \frac{1}{n} \sum_{i=m+1}^{n+m} h(X_i). \tag{2.4.7}$$

The process of discarding the burn-in is not the focus of the work in this thesis so, without loss of generality, we assume $m = 0$. The asymptotic variance of the estimator $\hat{h}_n$ is defined as

$$\begin{aligned} \sigma_h^2 := & \lim_{n\to\infty} n\mathbb{V}\left[\hat{h}_n\right] \\ & = \lim_{n\to\infty} n\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} h(X_i)\right] \\ & = \mathbb{V}\left[h(X_0)\right] + 2\sum_{i=1}^{\infty} \mathsf{Cov}\left[h(X_0), h(X_i)\right], \end{aligned}$$

provided that the above summation exists (e.g., Geyer, 1992). Similar to (2.4.5), σ_h^2 motivates a way of quantifying the inefficiency of a MCMC algorithm as the *integrated*

autocorrelation time,

$$I_{\mathsf{ACT}}(h) := \frac{\sigma_h^2}{\mathbb{V}_\pi\left[h(X_0)\right]} = 1 + 2\sum_{i=1}^{\infty} \mathsf{Corr}\left[h(X_0), h(X_i)\right].$$

This can then be used to define the effective sample size (ESS)

$$\mathsf{ESS}\left(\hat{h}_n\right) := \frac{n}{I_{\mathsf{ACT}}(h)}.$$

If the chain were to produce i.i.d. samples then $I_{\mathsf{ACT}}(h) = 1$ and $\mathsf{ESS}\left(\hat{h}_n\right) = n$. Naturally, I_{ACT} cannot be evaluated exactly, but there are methods of estimating it. In this thesis, we employ the `R` package `coda` (Plummer et al., 2006) for calculating ESS through spectral density estimates of a given chain (Heidelberger and Welch, 1981).

Chapter 5 is focused on methodology relating to a particular MCMC algorithm known as Hamiltonian Monte Carlo (HMC) (see, for example, Neal, 2011). It requires the density of π to be in the form $\pi(x) = \mathsf{e}^{-U(x)}$, where $x \in \mathbb{R}^d$ and U is a differentiable potential surface. HMC introduces an auxiliary momentum variable $p \in \mathbb{R}^d$ which has a multivariate Gaussian distribution. Proposals are generated by propagating Hamiltonian dynamics conditional on a given (x, p), which then undergo an accept-reject step to produce samples from the target π. The exact details of the algorithm and a review of relevant literature are at the beginning of the Chapter 5.

Metropolis-Hastings

A commonly used general class of MCMC algorithms is the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970). An iteration of the algorithm is

composed of two steps: (1) proposal mechanism of generating new state X' given current state $X = x$, described through a distribution q with density $q(x'|x)$, and (2) Metropolis-Hastings acceptance probability

$$\alpha(x, x') = 1 \wedge \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)}. \tag{2.4.8}$$

With a probability $\alpha(x, x')$, the next value of the chain becomes x'; otherwise, it remains unchanged. The MH update only requires the evaluation of some unnormalised version of π, which is the main reason for its appeal in Bayesian inference problems. It can be shown that the above procedure defines a transition kernel which satisfies the detailed balance with respect to π. The Metropolis-Hastings kernel has the form

$$K_{\mathsf{MH}}(x, x^*) = q(x^*|x)\alpha(x, x^*) + r(x)\delta_x(x^*),$$

where $r(x) = 1 - \int_{\mathsf{X}} q(x'|x)\alpha(x, x')\, \mathsf{d}x'$ is the probability of rejecting a proposal for a given $X = x$. The kernel K_{MH} accounts for both the accept and reject scenarios of the MH update. To show the kernel satisfies the detailed balance with respect to π, note that

$$\begin{aligned} K_{\mathsf{MH}}(x, x^*)\pi(x) &= \pi(x)q(x^*|x)\alpha(x, x^*) + \pi(x)r(x)\delta_x(x^*) \\ &= [\pi(x)q(x^*|x) \wedge \pi(x^*)q(x|x^*)] + \pi(x)r(x)\delta_x(x^*) \end{aligned}$$

which is invariant to $x \leftrightarrow x^*$. Consequently, it admits π as its stationary distribution.

The accept-reject step results in the chain using MH updates being aperiodic. To show the chain is Harris recurrent we use the fact that *irreducible* MH chains imply Harris recurrence, as per Lemma 7.3 of Robert and Casella (2005). The chain is said

to be irreducible if it can reach any subset of the support from any starting point, given enough time. This can be satisfied if

$$q(x^*|x)\alpha(x,x') > 0, \quad \forall (x,x^*) \in \mathsf{X} \times \mathsf{X},$$

which will be the case for all algorithms appearing in this thesis. Thus, a chain produced by the Metropolis-Hastings algorithm will converge to the target distribution π and

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \overset{a.s.}{\to} \mathbb{E}_\pi\left[h(X)\right] \quad \text{as} \quad n \to \infty.$$

Pseudo-marginal Metropolis-Hastings

It can happen that even an unnormalised version of $\pi(x)$ may not be available for exact evaluation. Take for example the following model consisting of parameters of interest θ, latent variable Z and observable data Y. The variables follow

$$Y|(Z=z) \sim f(\,\cdot\,|z,\theta) \quad \text{and} \quad Z \sim g(\,\cdot\,|\theta),$$

and the marginal likelihood function is given by

$$L(\theta;y) = p(y|\theta) = \int f(y|z,\theta)g(z|\theta)\,\mathsf{d}z, \tag{2.4.9}$$

which, apart from special cases, is not tractable. As discussed in Section 2.4.1, importance sampling can be used to produce an unbiased estimates of (2.4.9), and since $\pi(\theta|y) \propto L(\theta;y)\pi_0(\theta)$, we can obtain non-negative unbiased estimates of target density π, up to a multiplicative. Remarkably, if one were to replace $\pi(x')$ and $\pi(x)$ in the standard MH acceptance probability with an unbiased estimate $\hat{\pi}(x')$ and the es-

timate $\hat{\pi}(x)$ from the previous iteration, the resulting algorithm would still target the correct stationary distribution (Beaumont, 2003; Andrieu and Roberts, 2009). Such an algorithm is referred to as *pseudo-marginal Metropolis-Hastings*.

Returning to the general notation used throughout this section, we are interested in sampling a random variable $X \sim \pi$. Suppose $\pi(x)$ is intractable but instead we have access to $\hat{\pi}(x;\nu)$ which is a non-negative unbiased estimator of $\pi(x)$ (up to a multiplicative constant), where $\nu \sim \mu(\cdot\,|x)$ is a vector of all the random variables used to generate the estimate. We follow similar steps as in the standard MH algorithm: given current position (x,ν) with estimate $\hat{\pi}(x;\nu)$ we propose $x' \sim q(\cdot\,|x)$ and sample $\nu' \sim \mu(\cdot\,|x')$ to construct $\hat{\pi}(x';\nu')$. The proposal is then accepted with probability

$$\alpha\left((x,\nu),(x',\nu')\right) = 1 \wedge \frac{\hat{\pi}(x';\nu')q(x|x')}{\hat{\pi}(x;\nu)q(x'|x)},$$

otherwise the chain remains at (x,ν). To see why the above should have the correct invariant distribution, consider the extended target

$$\tilde{\pi}(x,\nu) = \hat{\pi}(x;\nu)\mu(\nu|x)$$

and note that

$$\alpha\left((x,\nu),(x',\nu')\right) = 1 \wedge \frac{\hat{\pi}(x';\nu')q(x|x')}{\hat{\pi}(x;\nu)q(x'|x)} = 1 \wedge \frac{\tilde{\pi}(x',\nu')q(x|x')\mu(\nu|x)}{\tilde{\pi}(x,\nu)q(x'|x)\mu(\nu'|x')}.$$

The above ratio is invariant to $(x,\nu) \leftrightarrow (x',\nu')$ and so it targets the extended $\tilde{\pi}$, with the marginal for x being the desired distribution,

$$\int \tilde{\pi}(x,\nu)\,\mathsf{d}\nu = \int \hat{\pi}(x;\nu)\mu(\nu|x)\,\mathsf{d}\nu = \pi(x).$$

Random-walk Metropolis

The classic choice of proposal distribution q for the Metropolis-Hastings algorithm is to propose $X'|(X = x) \sim \mathsf{N}(x, \lambda\Sigma)$, where $\lambda > 0$ is the (tuning) scaling parameter and Σ is a positive definite matrix (Metropolis et al., 1953). A naive choice for the proposal covariance would be $\Sigma = I_d$, but more optimally it would be chosen to reflect the covariance matrix of the target π (Roberts and Rosenthal, 2001). The algorithm is commonly known as random-walk Metropolis (RWM). The proposal mechanism is symmetric, i.e., $q(x'|x) = q(x|x')$, which simplifies the resulting MH acceptance probability,

$$\alpha(x, x') = 1 \wedge \frac{\pi(x')}{\pi(x)}.$$

Figure 2.4.2 illustrates the RWM algorithm on a toy example. The algorithm was purposefully initialised in the tails of the distribution to result is a long burn-in period, after which we say the chain has converged to the stationary distribution; the burn-in would be discarded when constructing estimators of the form (2.4.7) to reduce any potential bias.

The optimal performance of the algorithm is of wide interest both in theory and practice. The general recommendation for RWM is that the scaling λ should be chosen to achieve an acceptance rate of 23.4%. This corresponds to $\lambda \propto 2.38/\sqrt{d}$ (Roberts and Rosenthal, 2001), with equality when the target is Gaussian. The results are based on the limits as the dimension d increases but the target acceptance rate seems to hold reasonably well in dimensions as low as 5.

In pseudo-marginal Metropolis-Hastings, this becomes less straightforward as noisy

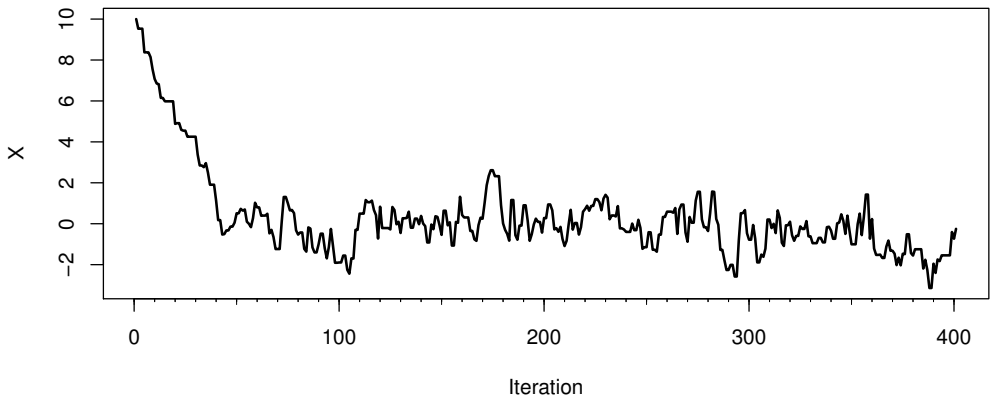


Figure 2.4.2: Traceplot of a chain produced by the random-walk Metropolis algorithm, on a standard normal target. The proposal distribution is a zero-mean Gaussian with a standard deviation of 0.75. The initial $50 - 70$ iterations would be considered as burn-in.

estimates of π are used. It can be shown that the efficiency of the PMMH algorithm will be strictly lower than that of its MH counterpart with direct evaluations of π (Andrieu and Vihola, 2016). Various works such as Pitt et al. (2012), Sherlock et al. (2015), Doucet et al. (2015) and Nemeth et al. (2016), suggest first tuning the number of random variables used in $\hat{\pi}$ estimates so that at $\bar{x} = \mathbb{E}_{\pi}[X]$ the variance of the log-estimates $\log \hat{\pi}(\bar{x})$ is somewhere in the region of $0.85 - 3.28$. Then a suitable scaling λ will result in acceptance rate for low dimensional targets to be $\approx 14\%$, with the optimal rate as $d \to \infty$ being $\approx 7\%$ (Sherlock et al., 2015).

2.5 Bayesian filtering problem

State-space models are a particular class of models also referred to as hidden Markov models. They are made up of a latent (unobservable) Markov process $\{X_t\}_{t \geq 0}$, initi-

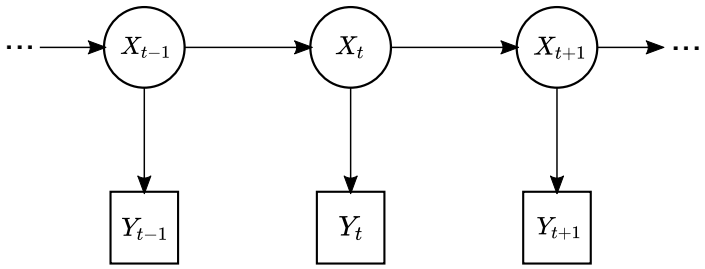


Figure 2.5.1: Illustration of the individual transitions of the variables in a simple hidden Markov model $\{X_t, Y_t\}_{t\geq 0}$. Rectangular nodes indicate observable data and circular nodes represent latent (unobservable) variables.

ated from $X_0 \sim \mu$ and obeying transition densities

$$p(x_t|x_{0:t-1}) = f(x_t|x_{t-1}), \quad t = 1, 2, \ldots.$$

Instead of observing X directly, we only have some noisy realisations $\{y_t\}_{t\geq 0}$ produced by some

$$p(y_t|x_{0:t}, y_{0:t-1}) = g(y_t, x_t), \quad t = 1, 2, \ldots.$$

Figure 2.5.1 gives a pictorial representation of the dependencies in the model.

We wish to infer $X_{1:T} = \{X_t\}_{t=0}^T$ based on a realisation $y_{0:T} = \{y_t\}_{t=0}^T$. The posterior for the latent process, conditional on the parameters θ is

$$\pi(x_{1:T}|y_{1:T}, \theta) \propto \pi(x_{1:T}, y_{1:T}|\theta) = \mu(x_0|\theta) \prod_{t=1}^{T} f(x_t|x_{t-1}, \theta) \prod_{t=0}^{T} g(y_t|x_t, \theta).$$

In the Bayesian filtering problem, we are interested in the posterior distribution of the latent process at times $t \leq 0$, in contrast to the whole joint posterior. The *filtering* density $\pi(x_t|y_{0:t}, \theta)$ can be obtained through Bayes' Theorem,

$$\pi(x_t|y_{0:t}, \theta) = \frac{p(x_t, y_t|y_{0:t-1}, \theta)}{p(y_t|y_{0:t-1}, \theta)},$$

where the joint density on the numerator follows a recursive relation

$$\begin{aligned} p(x_t, y_t|y_{0:t-1}, \theta) &= \int p(x_t, y_t, x_{t-1}|y_{0:t-1}) \, \mathsf{d}x_{t-1} \\ &= \int \pi(x_{t-1}|y_{0:t-1}, \theta) p(x_t|x_{t-1}, y_{0:t-1}, \theta) p(y_t|x_t, x_{t-1}, y_{0:t-1}, \theta) \, \mathsf{d}x_{t-1} \\ &= \int \pi(x_{t-1}|y_{0:t-1}, \theta) f(x_t|x_{t-1}, \theta) g(y_t|x_t, \theta) \, \mathsf{d}x_{t-1}, \end{aligned} \tag{2.5.1}$$

and $p(y_t|y_{0:t-1}, \theta)$ is the marginal of the above joint density. The above filtering densities are intractable apart from two special cases: (i) discrete state spaces resulting in the forward-backward algorithm (e.g., Rabiner, 1989), and (ii) Gaussian transition and observation densities resulting in a Kalman filter (Kalman, 1960). Neither of the two special cases appear in this thesis. The remainder of this section outlines a Monte Carlo method for approximating the filtering distribution.

2.5.1 Particle filters

In a particle filter, we approximate the (continuous) filtering density $\pi(x_{t-1}|y_{0:t-1}, \theta)$ using an approximate discrete distribution $\hat{\pi}$ constructed from a weighted sample of particles $\{x_{t-1}^{(b)}, \tilde{w}_{t-1}^{(b)}\}$,

$$\hat{\pi}(x|y_{0:t-1}, \theta) = \sum_{b=1}^{B} \tilde{w}_{t-1}^{(b)} \delta_{x_{t-1}^{(b)}}(x),$$

where the weights are normalised, that is, $\sum_{b=1}^{B} \tilde{w}_{t-1}^{(b)} = 1$. Substituting the approximation into the recursion relation (2.5.1) produces the approximate filtering density

$$\hat{\pi}(x_t|y_{0:t}, \theta) \propto \sum_{b=1}^{B} \tilde{w}_{t-1}^{(b)} f(x_t|x_{t-1}^{(b)}, \theta) g(y_t|x_t, \theta).$$

The sampling, importance resampling (SIR) particle filter (Gordon et al., 1993) approximates the filtering density at time t by resampling the pool of particles at the previous time-step $t-1$ according to their normalised weights $\tilde{w}_{t-1}^{(b)}$. The new particle approximation is constructed by the propagation mechanism $x_t^{(b)} \sim f(\cdot|x_{t-1}^{(b)}, \theta)$. Each new particle is then assigned an importance weight

$$w_t^{(b)} = \frac{g(y_t|x_t^{(b)}, \theta) f(x_t^{(b)}|x_{t-1}^{(b)}, \theta)}{f(x_t^{(b)}|x_{t-1}^{(b)}, \theta)} = g(y_t|x_t^{(b)}, \theta)$$

and the weights are normalised,

$$\tilde{w}_t^{(b)} = \frac{w_t^{(b)}}{\sum_{i=1}^{B} w_t^{(i)}}.$$

The propagation of the particles could in fact be replaced by some other proposal density $q(x_t|x_{t-1}^{(b)}, y_t, \theta)$. An *auxillary particle filter* (Pitt and Shephard, 1999) attempts to approximate the optimal proposal distribution q which can greatly improve the efficiency of the filter with respect to the effective sample size at a given iteration; this is similar to the optimal proposal for importance sampling (2.4.4). The construction of the approximately optimal proposal generally depends on the specific state-space model.

The resampling step introduces Monte Carlo noise since the pool of resampled particles is less diverse. One could omit the resampling altogether, instead updating the weights at each iteration of the filter

$$\dot{w}_t^{(b)} = g\left(y_t|x_t^{(b)}, \theta\right) \times \dot{w}_{t-1}^{(b)}.$$

This algorithm is referred to as *sequential importance sampling* (SIS) (e.g. Liu and

Chen, 1998; Arulampalam et al., 2002). Excluding the resampling step, however, leads to a phenomenon called *particle degeneracy*, where the variance of the importance weights increases at each iteration (Kong et al., 1994). The resampling could instead be carried out only when some criterion is satisfied; for example, if the effective sample size of the particles drops below 50% of the number of particles. In Chapter 4, we introduce a particle filter where the particle weights inherently have a very large variance and so to limit any potential degeneracy we will require resampling at every stage of the algorithm.

As a byproduct of the SIR, the product of unnormalised weight averages can be used to estimate the marginal likelihood,

$$\hat{p}(y_{0:T}|\theta) = \prod_{t=0}^{T} \frac{1}{n} \sum_{b=1}^{B} w_t^{(b)}.$$

In fact, the estimator is unbiased, that is,

$$\mathbb{E}\left[\prod_{t=0}^{T} \frac{1}{n} \sum_{b=1}^{B} w_t^{(b)}\right] = p(y_{0:T}|\theta);$$

see Theorem 1 of Pitt et al. (2012) or Proposition 7.4.1 of Del Moral (2004) for proofs of this result. As a result, the estimator can be used for exact posterior inference on θ through the pseudo-marginal Metropolis-Hastings algorithm, as outlined in Section 2.4.2.

Chapter 3

Interim recruitment prediction for multi-centre clinical trials

3.1 Introduction

Efficiently recruiting patients to clinical trials is a critical factor in running the trials and hence delivering new medicines to patients as quickly as possible. Late-stage clinical trials are commonly run across many sites, and successfully managing and running trials and subsequent processes requires accurate forecasts of trial recruitment.

Early recruitment rates can be high, for example, because patients with the required condition are already available, and rates can then drop once these patients have been recruited. Deterministic approaches and ad hoc techniques may yield simplified and, often, overly optimistic recruitment timelines, a phenomenon thus dubbed *Lasagna's Law* (Lasagna, 1979). For example, 48% of centres studied by Getz and Lamberti (2013) failed to enrol the required number of patients in the time originally

allocated, leading to extensions of the recruitment timelines and the need to bring more centres into the study, which itself is a costly process. The timelines are usually pushed to nearly twice the originally proposed plan. The most frequent reason for trial discontinuation appears to be poor recruitment; out of 253 discontinued trials studied in Kasenda et al. (2014), 101 were terminated due to under-recruitment.

This motivates the need for robust statistical methods for modelling and predicting the recruitment to clinical trials at site-level. Early detections of possible centre underperformance may allow practitioners to swiftly intervene in the operations. It can also provide realistic timelines for the completion of different stages of the trials.

In this work, we introduce a novel flexible framework for effectively modelling and predicting patient recruitment. We will focus on the oncology therapeutic area as it is known for sparse enrolments whose patterns are not sufficiently captured by the state-of-the-art methods (Anisimov and Fedorov, 2007; Lan et al., 2019). Our framework utilises time-varying recruitment rates whilst also permitting variation between recruitment centres. Inference is based on the set of known centre initiation times to date, whilst the prediction is conditional on a set of future initiation times. Past initiation times are known, but typically, whilst there is a plan for future initiation times along with potential contingencies, the actual times are not known precisely in advance. The proposed methodology can be used with user-specified initiation schedules to facilitate the choice between different initiation-time scenarios, or it can be combined with a centre-initiation model. Predictions of future recruitment incorporate parameter and model uncertainty, which is essential when data are limited.

Existing methods for predicting recruitment to clinical trials are overviewed in

Section 3.2. Section 3.3 outlines methods for detecting recruitment rate decay in the multi-centre recruitment setting along with result of a Monte Carlo power study. Section 3.4 introduces the flexible modelling framework and Section 3.5 presents a general method for choosing sensible Bayesian parameter priors, along with an appropriate posterior sampling method and diagnostics. A simulation study is presented in Section 3.6, illustrating the fitting of the model, model validation and forecasting recruitment using Bayesian model-averaging. In Section 3.7 the model is fitted to an oncology dataset, and this is followed by a discussion in Section 3.8.

3.2 Existing methods

The first statistical modelling framework for clinical trial recruitment was introduced in Lee (1983), where the recruitment was assumed to be a constant-rate Poisson process, leading to tractable inference based on interim data. Williford et al. (1987) built on the model by considering Bayesian inference with conjugate priors. Gajewski et al. (2008) and Jiang et al. (2015) further explored the effects various prior densities can have on predictions. Time-inhomogeneous accrual was first considered in Piantadosi and Patterson (1987), where the aggregated accrual across all sites was modelled as an inhomogeneous Poisson process with intensity $\lambda(t) = \zeta(1 - \mathsf{exp}(-\kappa t))$, $\zeta, \kappa > 0$. Zhang and Long (2010) took a non-parametric approach, using B-splines to model the trends in accrual and using the intensity value at the census time for predictions. Tang et al. (2012) proposed a Poisson model with a piece-wise linear intensity which captured aspects of recruitment such as slow initial recruitment and a spike in re-

cruitment close to the end of the trial. For a more thorough review of these as well as other methods see Heitjan et al. (2015). Accrual-only modelling methods do not consider the effect that initiating new centres can have on recruitment trends. For that reason, we shall focus on methods which can take advantage of centre-specific recruitment data.

Anisimov and Fedorov (2007) introduced the Poisson-gamma (PG) model of recruitment in a multi-centre setting, with the main appeal being the use of random effects for the recruitment rates of centres, providing a tractable, data-driven prior predictive distribution for recruitment in yet-unopened centres. The model consists of C centres, each recruiting N_c patients over τ_c days, $c = 1, \ldots, C$. The framework makes the following distributional assumptions,

$$\begin{aligned} \lambda_c &\sim \mathsf{Gamma}\left(\alpha, \alpha/\phi\right), \\ N_c|\lambda_c &\sim \mathsf{Pois}\left(\lambda_c \tau_c\right), \end{aligned} \qquad c = 1, \ldots, C. \tag{3.2.1}$$

The random effect λ_c is the *recruitment rate* for centre c. The rates, and thus the centre recruitments, are assumed to be independent conditional on α and ϕ. There are, however, several caveats with the approach taken. The paper advocates using the Empirical Bayes approach, that is, maximum likelihood estimation for the hierarchical parameters (α, ϕ) followed by re-estimation of the distribution of random effect λ_c given α, ϕ and n_c, for each centre. A method for obtaining the uncertainty in the hierarchical (α, ϕ) parameters is provided, but this uncertainty is not accounted for when making predictions, leading to overly confident prediction intervals. However, the main issue which could result from employing the model arises from the strong

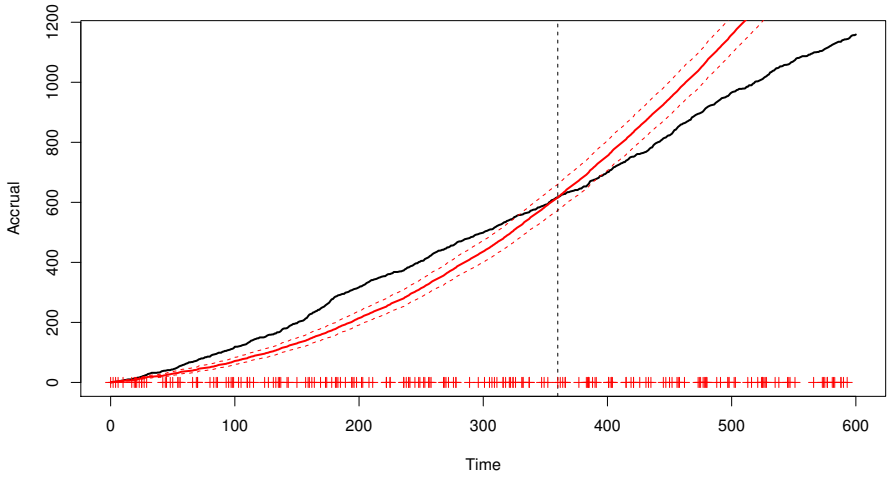


Figure 3.2.1: Accrual (black, solid) with the predictive mean (red, solid) and 95% prediction bands (red,dashed), based on the PG model ((3.2.1)) with the census time marked by the vertical, dashed line.

assumption of time-homogeneity of centre recruitments, which can lead to underestimations of the time to completion.

Figure 3.2.1 shows the accrual in a simulated trial where the rates gradually decay with time as well as the predictive distribution of the PG model fitted at a census time of three-fifths of the total length of the study; the initiation day for each centre is marked. The accrual appears to follow a straight line which could initially suggest using a time-homogeneous model. However, new centres are constantly being initatated so that a constant recruitment rate for each centre leads to an upward arching trend in accrual. This is encapsulated by the fitted predictive. Here the accrual is initially badly underestimated and then grossly overestimated after the census time. The apparent "matching" at the census time is due to predictions using re-estimated random-effect distributions.

Lan et al. (2019) describes the first multi-centre recruitment model in which the

rates decrease over time. The model assumes inhomogeneous Poisson for arrivals centre c with an intensity of the form

$$\lambda_c(t) = \begin{cases} \lambda_c^o, & t < t_o \\ \\ \lambda_c^o \mathsf{exp}(-\theta(t - t_o)), & t \geq t_o \end{cases},$$

where λ_c^o is a gamma random effect, as in ((3.2.1)), and t_o a user-specified parameter and is not estimated as part of the inference. By enforcing the specific intensity-form, the possibilities of time-homogeneous recruitments or even intensity decays with heavier tails are excluded. A more systematic alternative is to start by testing the time-homogeneity assumption.

3.3 Detecting time-inhomogeneity

Given series of daily centre recruitment counts over the recruitment period of τ_c days, $\{N_c(t)\}_{t=1}^{\tau_c}$, $c = 1, \ldots, C$, we can test the hypothesis of time-homogeneity. To detect a decay in the rate, we only need to use the sums $X_1^{(c)} = \sum_{t=1}^{\tau_c/2} N_c(t)$ and $X_2^{(c)} = \sum_{t=\tau_c/2+1}^{\tau_c} N_c(t)$ $(c = 1, \ldots, C)$, whose expectations we denote by $\mu_1^{(c)}$ and $\mu_2^{(c)}$ respectively. Detecting time-inhomogeneity in a single centre can be difficult as the infrequent counts will lead to low powers of tests (Krishnamoorthy and Thomson, 2004) (see also Tables 3.3.1 and 3.3.2). Thus we combine the recruitments across all centres leading to two counts: $X_1 = \sum_{c=1}^{C} X_1^{(c)}$ and $X_2 = \sum_{c=1}^{C} X_2^{(c)}$, and we choose our hypotheses to be

$$\mathrm{H}_0 : \sum_{c=1}^{C} \mu_1^{(c)} = \sum_{c=1}^{C} \mu_2^{(c)} \quad \text{vs} \quad \mathrm{H}_1 : \sum_{c=1}^{C} \mu_1^{(c)} > \sum_{c=1}^{C} \mu_2^{(c)}.$$

The tests are one-sided as we are only interested in recruitment which decays over time. We consider tests with respect to the following assumptions:

Assumption 1: *For each centre $c = 1, \ldots, C$, the counts in the first and second halves of that centre's recruitment period are independent and have the same distribution, $X_1^{(c)} \overset{\mathcal{D}}{=} X_2^{(c)}$, with expectation $\mu_1^{(c)}$. Furthermore, the recruitments at each centre are independent of each other.*

Assumption 2: *The patients arrive according to a Poisson process such that $X_1^{(c)}, X_2^{(c)} \sim \mathsf{Pois}\left(\mu_1^{(c)}\right)$, for some $\mu_1^{(c)}$, $c = 1, \ldots, C$.*

Assumption 1 implies that X_1 and X_2 must have the same distributions, with respective expectations $\mu_1 = \sum_{c=1}^{C} \mu_1^{(c)}$ and $\mu_2 = \sum_{c=1}^{C} \mu_2^{(c)}$ being equal. Assumption 2 further implies that the distributions must be Poisson. Figure 3.3.1 shows the construction of the quantities X_1 and X_2 by aligning the centres of the recruiting periods. The splitting of the series halfway is arbitrary, though splitting it in half (or at least close to this) would theoretically yield the highest power. It assumes that the τ_c are even. However, centres recruiting over odd numbers of days can still be used by removing the middle day observation. This reduces the power of the tests, though the reduction is negligible.

Gu et al. (2008) offer a detailed Monte Carlo study of the different methods used for testing for a difference in means of two Poisson variables. Here, we focus on the ones most applicable to the clinical-trial recruitment setting, bearing in mind statistical power and robustness. We identified two methods: the non-parametric bootstrapped test (BST), which is powerful yet robust, and the Poisson likelihood-ratio test (LRT), which makes stronger distribution assumptions to achieve an even

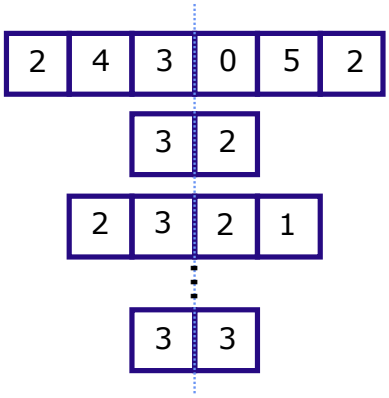


Figure 3.3.1: Count series are all centred and the sum of all the first halves is compared to the sum of second halves.

higher power. The BST only assumes that the counts in each day are independent and identically distributed (Assumption 1). With this assumption, resampling within each centre with replacement, from the original data would still produce a valid sample from the assumed distribution under H_0. A large number of bootstrap samples is used to simulate the distribution of the difference in two means, which is then used to test the hypothesis. Appendix A.1 details the sampling procedure for obtaining the distribution and the p-value.

For the LRT, we require Assumption 2, which is already an underlying assumption for the model in Anisimov and Fedorov (2007). Upon aggregation, the two sums follow Poisson distributions, that is, $X_1 \sim \mathsf{Pois}(\mu_1)$ and $X_2 \sim \mathsf{Pois}(\mu_2)$. The likelihood under the null model ($\mu_1 = \mu_2$) is compared to the likelihood under the alternative two-mean model ($\mu_1 > \mu_2$). Here, the likelihood function is

$$L(\mu_1, \mu_2; x_1, x_2) = \frac{\mu_1^{x_1}\mathsf{exp}(-\mu_1)}{x_1!}\frac{\mu_2^{x_2}\mathsf{exp}(-\mu_2)}{x_2!}, \quad \mu_1, \mu_2 > 0.$$

We let

$$T_L(x_1, x_2) = \begin{cases} 2[\log L(\hat{\mu}_1, \hat{\mu}_2; x_1, x_2) - \log L(\hat{\mu}, \hat{\mu}; x_1, x_2)], & \hat{\mu}_1 > \hat{\mu}_2 \\ 0, & \hat{\mu}_1 \leq \hat{\mu}_2 \end{cases},$$

where $\hat{\mu}$ is the MLE under the null, and $\hat{\mu}_1$ and $\hat{\mu}_1$ are the MLEs under the alternative hypothesis. Under the null, we would expect the test statistic $T_L(X_1, X_2)$ to asymptotically be zero half the time with the other half following a χ^2_1 distribution (Robertson et al., 1988). When using the LRT, the simulated significance levels can differ from the pre-specified level when μ values are low. This is due to using the asymptotic χ^2 distribution when calculating the p-value (Gu et al., 2008).

The performance of the two tests was assessed by carrying out a Monte Carlo study. Test powers were estimated using Poisson data with different expectations and ratios, $R = \mu_2/\mu_1$. For the LRT power estimates, 5×10^6 samples were used as the test itself is very computationally cheap. For the BST, 5×10^4 samples were used, with each test using a bootstrapped distribution of size 10^3. Tables 3.3.1 and 3.3.2 show the results of the study. The biggest difference in powers occurs for lower expectations, with the LRT outperforming BST. It must be noted, however, that the BST only requires the data to be i.i.d. within each centre and thus is robust to violations of the Poisson assumption; if the counts within each centre are overdispersed, for example, it does not affect the Type I error.

To exemplify the usefulness of this test, we can consider an interim likelihood ratio test where the expected number of enrolments is 170. This corresponds to $\mathbb{E}[X_1] = 100$ and $R = 0.7$, for example, and results in a statistical power of approximately 0.75.

$\mathbb{E}[X_1]$	$R = 1$	$R = 0.9$	$R = 0.8$	$R = 0.7$	$R = 0.6$	$R = 0.5$
5	0.06	0.08	0.11	0.15	0.20	0.27
10	0.05	0.08	0.12	0.18	0.26	0.37
20	0.05	0.09	0.17	0.27	0.41	0.58
50	0.05	0.13	0.28	0.50	0.73	0.90
100	0.05	0.18	0.44	0.75	0.94	0.99
200	0.05	0.27	0.68	0.95	1.00	1.00

Table 3.3.1: Power for likelihood-ratio test

$\mathbb{E}[X_1]$	$R = 1$	$R = 0.9$	$R = 0.8$	$R = 0.7$	$R = 0.6$	$R = 0.5$
5	0.04	0.06	0.08	0.11	0.14	0.18
10	0.05	0.08	0.12	0.16	0.24	0.33
20	0.05	0.10	0.16	0.25	0.39	0.57
50	0.05	0.14	0.28	0.48	0.70	0.88
100	0.05	0.18	0.42	0.74	0.93	0.99
200	0.05	0.28	0.67	0.94	1.00	1.00

Table 3.3.2: Power for non-parametric bootstrapped test

Considering many trials require an upward of 500 enrolments, informed decisions can be made relatively early on in the trial.

3.4 Proposed model

We consider a scenario of C centres recruiting patients, with each centre c being initiated for τ_c days. The number recruited by centre c on day t shall be denoted by $N_c^{(t)}$. We propose the following modelling framework for the multi-centre clinical-trial recruitment, based on the inhomogeneous Poisson process,

$$\lambda_c^o \sim \mathsf{Gamma}\left(\alpha, \frac{\alpha}{\phi}\right), \qquad c = 1, \ldots, C,$$

$$N_c^{(t)} \sim \mathsf{Pois}\left(\lambda_c^o \int_{t-1}^{t} g(s;\theta)\, \mathsf{d}s\right), \qquad t = 1, \ldots, \tau_c,$$

where g is a non-negative function which dictates the curve-shape of the intensity and θ is a parameter (or parameter vector) associated with the functional form. We use the (α, ϕ) parametrisation for the hierarchical gamma distribution as it leads to orthogonality of α and ϕ in the Poisson-gamma model (Huzurbazar, 1950). *A priori*, $\mathbb{E}[\lambda_c] = \phi$ and $\mathbb{V}[\lambda_c] = \phi^2/\alpha$. For notational simplicity, we define $G(t;\theta) = \int_0^t g(s;\theta)\,\mathsf{d}s$. The likelihood contribution from centre c is

$$\begin{aligned}\mathbb{P}(\mathbf{N}_c = \mathbf{n}_c|\lambda_c^o, \theta, \tau_c) &= \prod_{t=1}^{\tau_c} \mathbb{P}(N_c^{(t)} = n_c^{(t)}|\lambda_c^o, \theta) \\ &= \exp(-\lambda_c^o G(\tau_c;\theta))(\lambda_c^o)^{n_c^{(\cdot)}} \prod_{t=1}^{\tau_c} \frac{[G(t;\theta) - G(t-1;\theta)]^{n_c^{(t)}}}{n_c^{(t)}!},\end{aligned}$$

where $n_c^{(\cdot)} = \sum_{t=1}^{\tau_c} n_c^{(t)}$. Marginalising over the random-effect component gives

$$\mathbb{P}(\mathbf{N}_c = \mathbf{n}_c|\alpha, \phi, \theta, \tau_c) = \frac{(\alpha/\phi)^\alpha \Gamma\left(\alpha + n_c^{(\cdot)}\right)}{\Gamma(\alpha)[G(\tau_c;\theta) + \alpha/\phi]^{\left(\alpha + n_c^{(\cdot)}\right)}} \prod_{t=1}^{\tau_c} \frac{[G(t;\theta) - G(t-1;\theta)]^{n_c^{(t)}}}{n_c^{(t)}!},$$

whence the full likelihood of the model given the recruitment data is:

$$\begin{aligned}L(\alpha, \phi, \theta; \mathbf{n}, \boldsymbol{\tau}) &= \prod_{c=1}^{C} \mathbb{P}(\mathbf{N}_c = \mathbf{n}_c|\alpha, \phi, \boldsymbol{\tau}) \\ &= \frac{(\alpha/\phi)^{C\alpha}}{\Gamma(\alpha)^C} \prod_{c=1}^{C} \frac{\Gamma\left(\alpha + n_c^{(\cdot)}\right)}{[G(\tau_c;\theta) + \alpha/\phi]^{\left(\alpha + n_c^{(\cdot)}\right)}} \prod_{t=1}^{\tau_c} \frac{[G(t;\theta) - G(t-1;\theta)]^{n_c^{(t)}}}{n_c^{(t)}!}.\end{aligned} \tag{3.4.1}$$

If all the centres had been recruiting for the same amount of time, that is, $\tau_c \equiv \tau$ $\forall c$, then by fixing the integral of $g(t;\theta)$ over τ days we could introduce parameter orthogonality between (α, ϕ) and θ by imposing the normalisation: $\int_0^\tau g(t;\theta)\,\mathsf{d}t = \tau$. This generalises the homogeneous model with $g(t;\theta) = 1$ and leads to the following

factorisable likelihood,

$$L(\alpha, \phi, \theta; \mathbf{n}, \boldsymbol{\tau}) = \frac{(\alpha/\phi)^{C\alpha}}{\Gamma(\alpha)^C(\tau + \alpha/\phi)^{(C\alpha+n_\Sigma)}} \prod_{c=1}^{C} \Gamma\left(\alpha + n_c^{(\cdot)}\right) \prod_{t=1}^{\tau_c} \frac{[G(t;\theta) - G(t-1;\theta)]^{n_c^{(t)}}}{n_c^{(t)}!}$$

$$= L(\alpha, \phi; \mathbf{n}, \tau) L(\theta; \mathbf{n}, \boldsymbol{\tau}), \tag{3.4.2}$$

where $n_\Sigma = \sum_{c=1}^{C} n_c^{(\cdot)}$.

The factorisation means that now the θ parameter describes the shape of the intensity only, and α and ϕ describe the distribution of the magnitude of the integrated intensity, leading to a more interpretable model.

Even when centres are not all recruiting for the same length of time, we choose to impose a similar normalisation using some representative τ, here $\frac{1}{C}\sum_{c=1}^{C} \tau_c$. As demonstrated empirically in Section 3.6, the condition leads to approximate orthogonality even when the centres are initiated uniformly throughout the study.

3.4.1 Intensity curve-shape

In this work, we will restrict our choice of curve-shape g to parametric forms. The functional form of g is arbitrary and the best choices may depend on the context of the problem. When working with oncology datasets, for each centre we observe low-frequency counts which seem to become even less frequent over time but with varying tail behaviours. For this reason, we chose the following curve-shape

$$g_\kappa(t;\theta) \propto \left(1 + \frac{\theta t}{\kappa}\right)^{-\kappa}, \quad t \geq 0, \quad \theta, \kappa > 0. \tag{3.4.3}$$

The proportionality is used as multiplying g_κ by some positive constant and dividing ϕ by the same constant leads to the same model. The limit as $\kappa \to 0$ recovers the

standard PG model ((3.2.1)); and letting $\kappa \to \infty$, we obtain an exponential tail. The full (normalised) forms are then

$$g_0(t) \equiv 1, \tag{3.4.4}$$

$$g_1(t;\theta) = \frac{\theta(1+\theta t)^{-1}}{\mathsf{log}(1+\theta\tau)}\tau, \tag{3.4.5}$$

$$g_\kappa(t;\theta) = \frac{\theta(1-\kappa)(1+\theta t/\kappa)^{-\kappa}}{\kappa(1+\theta\tau/\kappa)^{1-\kappa}-\kappa}\tau, \quad \kappa \notin \{0,1,\infty\}, \tag{3.4.6}$$

$$g_\infty(t;\theta) = \frac{\theta\mathsf{exp}(-\theta t)}{1-\mathsf{exp}(-\theta\tau)}\tau. \tag{3.4.7}$$

The associated integrated forms, $G_\kappa(t;\theta)$ are provided in Appendix A.2.

The flexibility of the model, however, can result in potential identifiability issues. Inference methods, such as maximum likelihood, can run into numerical instabilities when $\kappa >> 1 > \theta$ or $\kappa < 1 << \theta$ (see Appendix A.2 for details). For this reason, we recommend restricting the choice of κ to a discrete set of values; in this work, we use $\{0, 0.5, 1, 2, \infty\}$. This will be elaborated on in Section 3.5.3.

3.5 Inference, diagnostics and predictions

We aim to construct a framework which can provide reliable predictions whilst capturing uncertainty in the estimated parameters and in the underlying model itself. We employ the Bayesian paradigm since it naturally incorporates the distribution of the random effects, λ_c, with the uncertainty in the model and the parameter values. However, we note that in some scenarios frequentist methods may be preferred and give a brief outline of how one may employ them in Appendix A.3.

Given a parametric statistical model, the Bayesian paradigm starts from a prior

distribution for the parameters, here denoted $\pi_0(\alpha, \phi, \theta)$ and updates this according to some data, y, to provide a posterior distribution, here denoted by $\pi(\alpha, \phi, \theta|y)$. When multiple parametric models, M_k, $k = 1, \ldots, K$, are being considered, the posterior probability for model k, here denoted by $\pi_p(M_k|y)$, may also be calculated. For the models under consideration for trial-recruitment data, neither the posterior model probabilities nor the posteriors for the parameters for any particular model are tractable, and so we employ importance sampling to obtain Monte Carlo samples $(\alpha_m, \phi_m, \theta_m)$, $m = 1, \ldots, M$ from the posterior distribution for any given model, as well as an estimate of $\pi(M_k)$, $k = 1, \ldots, K$. Chapter 2 of this thesis provides further details of this method, as well as of effective sample size (ESS), a diagnostic which indicates the reliability of the Monte Carlo estimates; see also Robert and Casella (2013) or Doucet et al. (2013).

In Sections 3.6 and 3.7, we carry out inference on $\tilde{\alpha} = \log \alpha$, $\tilde{\phi} = \log \phi$ and $\tilde{\theta} = \log \theta$ since analyses of trial data showed the likelihood in the log-parameters to be more symmetric about the mode, which can make sampling more efficient. For the importance sampling proposal distribution, we use a multivariate t-distribution on 4 degrees of freedom, with the same mode as the posterior and the shape matrix equal to the inverse Hessian at the posterior mode.

3.5.1 Prior choices

We base our prior specification on a maximum likelihood meta-analysis of 20 oncology clinical trial recruitment datasets. The trials studied were for seven different types of cancers: ovarian, prostate, breast, small and non-small lung, bladder and pancre-

atic. The number of centres ranged from 58 to 244 with a median of 140 and total enrolments ranged from 245 to 4391 with a median of 1035. In all cases, the parameter estimators were close to orthogonal justifying the use of independent priors: $\pi_0\left(\tilde{\alpha}, \tilde{\phi}, \tilde{\theta}\right) = \pi_0\left(\tilde{\alpha}\right)\pi_0\left(\tilde{\phi}\right)\pi_0\left(\tilde{\theta}\right)$.

We found that the α parameter does not change much from one study to another. The weakly informative prior $\tilde{\alpha} \sim \mathsf{N}(0.2, 2^2)$ sufficiently reflects the distribution of the estimated values.

The ϕ parameter estimates varied by orders of magnitude between studies. The parameter reflects the mean centre recruitment and is well identified by the data; it depends upon the catchment region, type of indication and protocol, for example. For this reason, we advocate using a vague prior unless reliable expert knowledge is available. In our analyses, we used the uninformative, proper prior $\tilde{\phi} \sim \mathsf{Unif}(-8, 8)$.

The difference between the homogeneous model (3.4.4) and the inhomogeneous models (3.4.5), (3.4.6), (3.4.7) is the presence of the curve-shape parameter θ. Lindley's paradox (Lindley, 1957) warns that assigning θ a vague prior can lower the posterior probabilities of the models that use θ, compared to the model with $\kappa = 0$ which does not use θ. To avoid the paradox we set an informative but sensible prior by considering the drop off in intensity after some time, t_0. We let $R_\kappa = g_\kappa(t_0; \theta)/g_\kappa(0; \theta)$ and set $R_\kappa \sim \mathsf{Beta}(a, b)$ *a priori*, with $a = b = 1.1$ to indicate a lack of information, excepting that this is not a constant intensity model, since this is covered by $\kappa = 0$, and that we do not expect a 100% drop off after a time of t_0 (expert opinion); here we take $t_0 = 4$ months. As R_κ is a monotonic function of θ, we can use a density transform to derive the corresponding prior for θ. If prior information is abundant, be

it in the form of historical data or expert knowledge, the beta distribution parameters can be adjusted to reflect this. Given (3.4.3), the resulting prior density for $\tilde{\theta}$ is given in Appendix A.4.

3.5.2 Predictive distribution

There are two complementary scenarios for which predictions might be required: the distribution of future recruitments within a set time interval, and the distribution of time until the target number of recruitments is reached. In this section, we focus on the former; details of the latter appear in Appendix A.5.

Suppose we are interested in sampling the recruitment, denoted N_c^+, at some day t^+ by centre c. Given samples from the parameter posteriors, we can sample exactly from the posterior predictive for N_c^+ by exploiting the Poisson-gamma conjugacy of the random-effect distribution. The posterior distribution for the λ_c^o random effect for centre c is

$$\lambda_c^o|\alpha,\phi,\theta,\mathbf{n}_c,\tau_c \sim \mathsf{Gamma}\left(\alpha+n_c^{(\cdot)},\alpha/\phi+G(\tau_c;\theta)\right)=\mathsf{Gamma}\left(\alpha_c^*,\frac{\alpha_c^*}{\phi_c^*}\right), \quad (3.5.1)$$

where $\alpha_c^*=\alpha+n^{(\cdot)}$ and $\phi_c^*=\phi\times\left(\frac{\alpha+n_c^{(\cdot)}}{\alpha+\phi G(\tau_c;\theta)}\right)$. The predictive distribution for N_c^+ conditional on the random effect is:

$$N_c^+|\lambda_c^o,\theta\sim\mathsf{Pois}\left(\lambda_c^o\int_{t^+-1}^{t^+}g(s;\theta)\,\mathsf{d}s\right)=\mathsf{Pois}\left(\lambda_c^oG_\theta^+\right), \quad (3.5.2)$$

where $G_\theta^+=\int_{t^+-1}^{t^+}g(s;\theta)\,\mathsf{d}s$.

Marginalising over the random effect posterior, we arrive at the negative binomial

distribution:

$$\mathbb{P}(N_c^+ = n|\alpha_c^*, \phi_c^*) = \frac{\Gamma(\alpha_c^* + n)}{\Gamma(\alpha_c^*)n!}\left(\frac{\alpha_c^*}{\alpha_c^* + \phi_c^* G_\theta^+}\right)^{\alpha_c^*}\left(\frac{\phi_c^* G_\theta^+}{\alpha_c^* + \phi_c^* G_\theta^+}\right)^n, \quad n \in \mathbb{N}. \quad (3.5.3)$$

The prediction interval t^+ does not need to be a day and could instead be a week or a month, depending on the context of the application. To obtain the full marginal predictive, we sample the recruitments conditional on parameters sampled from the posterior. For as yet unopened centres, we set $n_c^{(\cdot)} = \tau_c = 0$. For each triplet (or couplet, if $\kappa = 0$) of parameters sampled from the posterior, we sample N_c^+, $c = 1, \ldots, C$, and sum them to obtain a sample from $N^+|\alpha, \phi, \theta$. The collection of these sums is a sample from the posterior predictive distribution for the model.

If simulations for multiple distinct time periods are required for a given centre, c, as needed for the accrual curve for example, then we first sample λ_c^o from its posterior (3.5.1). We then simulate the Poisson counts for the individual time periods, which are conditionally independent given λ_c^o, from (3.5.2).

3.5.3 Model averaging

When predicting the enrolments using a fitted model, we implicitly assume that a single model best reflects reality; however, prediction methods should consider the uncertainty in the models used for inference. We shall, therefore, use model averaging for making predictions, that is, take a weighted average of predictions made by each model. Working in the Bayesian paradigm provides us with an intuitive choice for

weights in the form of marginal likelihoods of the models.

$$\mathbb{P}(N^+ = n^+|\mathbf{n}, \boldsymbol{\tau}) = \sum_{k=1}^{K} \mathbb{P}(N^+ = n^+|\mathbf{n}, \boldsymbol{\tau}, M_k)\pi_p(M_k|\mathbf{n}, \boldsymbol{\tau}),$$

where $\pi_p(M_k|\mathbf{n}, \boldsymbol{\tau}) \propto \pi(\mathbf{n}|\boldsymbol{\tau}, M_k)\pi_0(M_k)$, $k = 1, \ldots, K$, with $\pi_0(M_k)$ being prior model probabilities. The averaging framework fits in with the restriction of the shape parameter κ to a discrete space. Each κ value generates an inhomogeneous Poisson-gamma model with the tail behaviour of the associated intensity shape. This includes the null ($\kappa = 0$) model as in Anisimov and Fedorov (2007). In this work we set all prior model probabilities equal.

3.5.4 Model validation

Before making any statements in regards to the future recruitments, we should validate that the fitted model does indeed capture the true data-generating process sufficiently well. Since the true process is unknown, we compare the observed data to the modal model (the model with the highest posterior probability) fixed at posterior parameter means $(\hat{\alpha}, \hat{\phi}, \hat{\theta})$.

Firstly, we wish to assess that the chosen hierarchical structure is reflected in the data. The distribution of posterior means of the individual random effects should approximately follow the hierarchical $\mathsf{Gamma}(\hat{\alpha}, \hat{\alpha}/\hat{\phi})$ distribution. A QQ-plot can be used to visually compare the distributions. If deemed sufficiently similar, using the distribution for generating predictions for yet-unopened centres is appropriate. If the distributions are noticeably different, particularly if the true distribution is multimodal, any interim predictions for yet-unopened centres could (but need not;

see robustness study in Section 3.6) be inaccurate.

According to the model, the counts in any initial period $[0, t']$ (such as the first month) of each centre's recruitment period, follow a negative binomial distribution with shape parameter α and success probability $\phi G(t';\theta)/(\alpha + \phi G(t';\theta))$, similar to that given in (3.5.3) but using α and ϕ in place of α_c^* and ϕ_c^*. As the true parameters are unknown, we compare it to the distribution fixed at point-estimates $(\hat{\alpha}, \hat{\phi}, \hat{\theta})$. The diagnostic indicates if the combination of the gamma random effects and the modal decay model captures the behaviour over the initial period after centre initiation. Again, a QQ-plot can be used for comparing the theoretical distribution to the observation, giving an indication if the fitted model under- or overestimates initial recruitment. The initial period, $[0, t']$, should be long enough that the true recruitment decay should be apparent. However, since only centres that have been recruiting for a period of at least t' can be used for the diagnostic, to ensure a reasonable power, t' should be short enough that a large number of sites have been recruiting for this duration. In this work, we set $t' = 60$ (2 months).

3.6 Simulation results

We demonstrate our flexible framework through a simulation study, using simulated data sets to illustrate model fit and prediction and to highlight the effect model misspecification can have on predictions. In practice, patterns in centre initiation times can vary greatly between trials. For presenting the methodology, we consider an initiation schedule similar to that observed in a typical trial. We test the robustness

of the method to model misspecification using a uniform initiation schedule, with another type of schedule examined in Appendix A.6.

Our historical data set do not include the initiation times of the centres, so instead, to accurately reflect the historical data used in the meta-analysis and what is often available to researchers, we take the first recruitment time of a centre as its initiation time and adjust the models to include a single deterministic recruitment at the initiation time of each centre followed by stochastic recruitment as described in Section 3.4.

We simulate a study over a course of 600 days, with 200 centres. The parameters used for simulations were $\alpha = 1.4$, $\phi = 0.01$, $\kappa = 2.7$ and $\theta = 0.02$. The inference is carried out on data observed in the first 360 days. As motivated in Section 3.1, we condition the inference on a set of known initiation times, chosen by the practitioner; these could subsequently be varied to investigate the impact of different schedules or initiation models. We consider a set of models with flexible tails (Section 3.4.1) allowing $\kappa \in \{0, 0.5, 1, 2, \infty\}$, thus including the null model (Anisimov and Fedorov, 2007). The "normalisation" of the curve-shapes was imposed at $\bar{\tau} = \frac{1}{C}\sum_{c=1}^{C} \tau_c$. We purposely simulated using a κ value outside of those considered in our models to illustrate the flexibility of the framework. For Bayesian inference, we used parameter and model priors outlined in Sections 3.5.1 and 3.5.3 respectively. Based on the model fitted to the data at the census day 360, we wish to predict the daily accrual until day 600.

Performing the LRT and BST from Section 3.3, we find the p-values of both tests to be < 0.001. Table 3.6.1 provides the fits for the five models. The effective

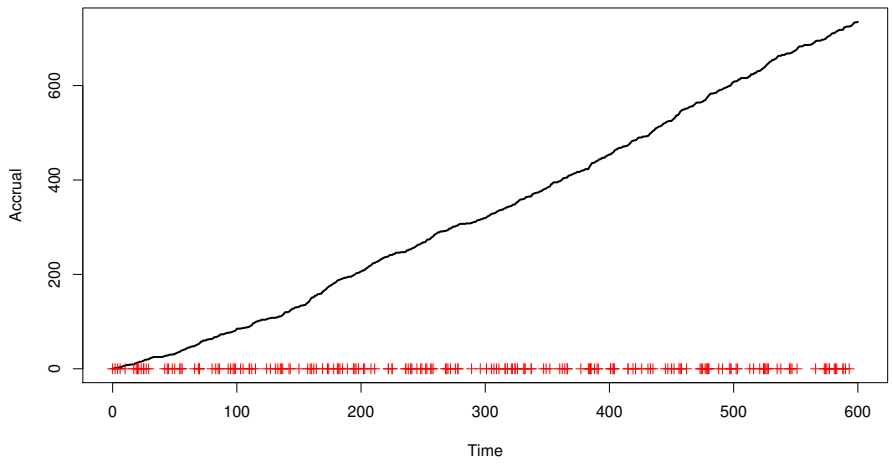


Figure 3.6.1: Accrual plot with the centre opening times marked by "+" symbols on the x-axis.

κ	α	ϕ	θ	$\pi(\mathcal{M}_k\|\boldsymbol{n})$	ESS
0	$1.141\ (0.771, 1.672)$	$0.013\ (0.011, 0.017)$	$--$	3.49×10^{-25}	9006
0.5	$1.167\ (0.759, 1.745)$	$0.013\ (0.010, 0.016)$	$0.143\ (0.044, 0.441)$	5.51×10^{-4}	8519
1	$1.144\ (0.742, 1.744)$	$0.013\ (0.011, 0.016)$	$0.033\ (0.021, 0.049)$	2.21×10^{-1}	8665
2	$1.142\ (0.728, 1.644)$	$0.014\ (0.011, 0.016)$	$0.017\ (0.012, 0.023)$	6.58×10^{-1}	8564
∞	$1.122\ (0.718, 1.645)$	$0.014\ (0.011, 0.017)$	$0.009\ (0.007, 0.011)$	1.20×10^{-1}	8610

Table 3.6.1: Posterior means and 95% credible intervals, posterior model probabilities and effective sample sizes, obtained using 10^4 importance samples for each model.

samples sizes are high, which means that each of the model posteriors is represented well by its respective sample and that the marginal likelihood estimates are accurate. If the ESS values had been low, we would have retried using more samples in the importance sampler. We see that model corresponding to $\kappa = \infty$ has the highest posterior probability. A trellis plot of the posteriors for $(\tilde{\alpha}, \tilde{\phi}, \tilde{\theta})$ from the modal model (see Appendix A.6) confirms at least approximate pairwise orthogonality between the parameters, as anticipated from Sections 3.4 and 3.5.1. QQ-plots for the modal model comparing the hierarchical gamma distribution to the posterior means of the random effects, and comparing the observed recruitments over the first two months of each centre's recruiting period to the model's negative binomial distribution both show approximate straight lines with unit gradient and are provided in the Appendix A.6.

Figure 3.6.2 shows the accrual forecast from the census time $\tau = 360$ up to the horizon $\tau^H = 600$, superimposed onto the true accrual plot. The forecast is based on the Bayesian model-averaged posterior predictive distribution. The true accrual is contained within the 95% predictive intervals.

Figures 3.6.3a and 3.6.3b use an earlier census time ($\tau = 240$) to illustrate the issues that can arise when making predictions using maximum likelihood estimation and model selection. The inference was carried out with the same set of candidate models, and predictions were obtained by simulating from the best model ($\kappa = \infty$, chosen using AIC) with parameters fixed at the MLEs. As shown in the plots, not accounting for parameter and model uncertainty may lead to overly confident and biased predictions. Simulations with $\tau = 360$ (see Appendix A.6) still showed bias due to the choice of a single model, although the contrast with Figure 3.6.2 in terms

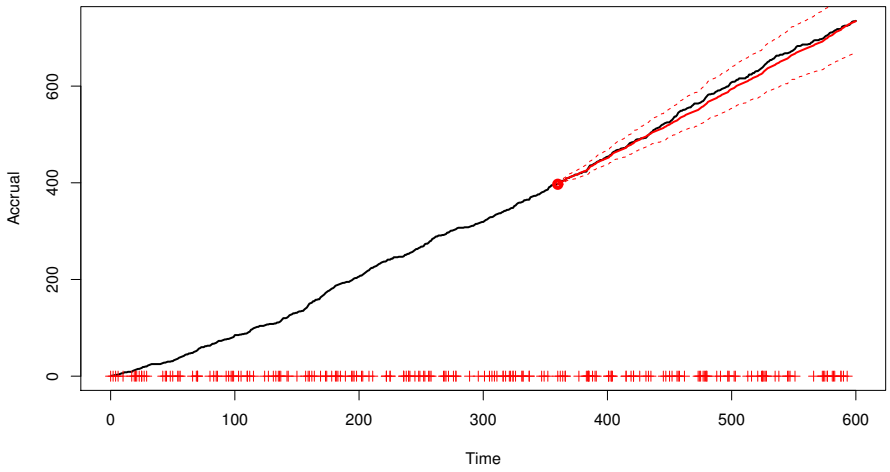


Figure 3.6.2: Accrual with Bayesian model-averaged forecast predictive mean (solid, red) and 95% prediction bands (red, dashed). Prediction bands are based on the 2.5% and 97.5% quantiles. The forecast begins from a point marked by the red dot and the "+" symbols on the x-axis indicate centre opening times.

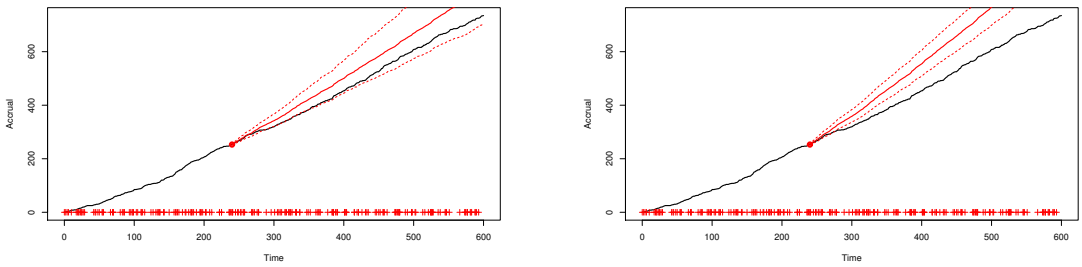
of prediction interval width was less marked.

We repeated the analysis with a different distribution of initiation times, making the centre initiations "clump" roughly every two months. The resulting forecast predictive distribution can be seen in Figure 3.6.4; performance appears to be robust to the type of initiation schedule.

To further test the robustness of the framework, we first consider the random effects λ_c^o now being generated from a mixture of two gamma distributions

$$\lambda_c^o|\alpha,\phi_1,\phi_2 \sim \frac{1}{2}\mathsf{Gamma}\left(\alpha,\frac{\alpha}{\phi_1}\right)+\frac{1}{2}\mathsf{Gamma}\left(\alpha,\frac{\alpha}{\phi_2}\right).$$

We considered data generated using the same α value and curve-shape as before, but now with centre initiation times uniformly sampled on the interval. The ratio of gamma expectations was fixed such that $\phi_2 = 10\phi_1$, and the random effect expecta-



(a) Bayesian model averaging (b) MLE and model selection

Figure 3.6.3: Comparison of accrual predictions produced by two methods; accruals (black, solid) with predictive means (red, solid) and 95% prediciton bands (red, dashed). Prediction bands are based on the 2.5% and 97.5% quantiles. The "+" symbols on the x-axis indicate centre opening times.

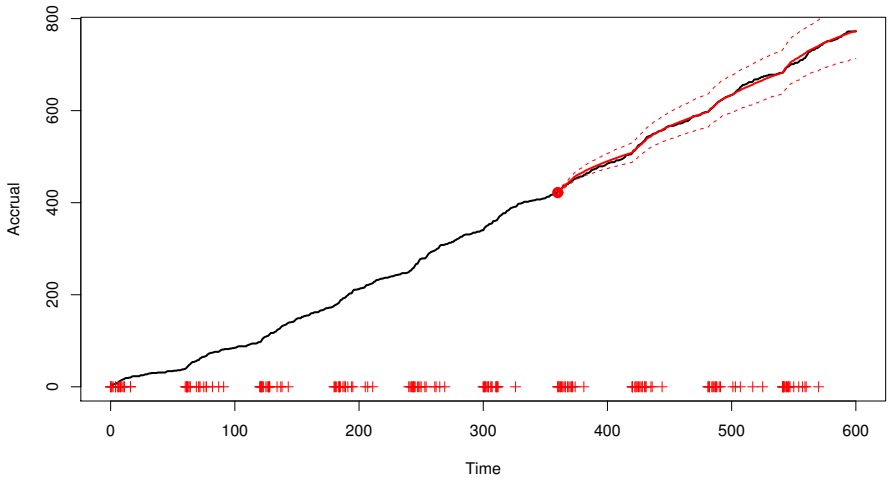
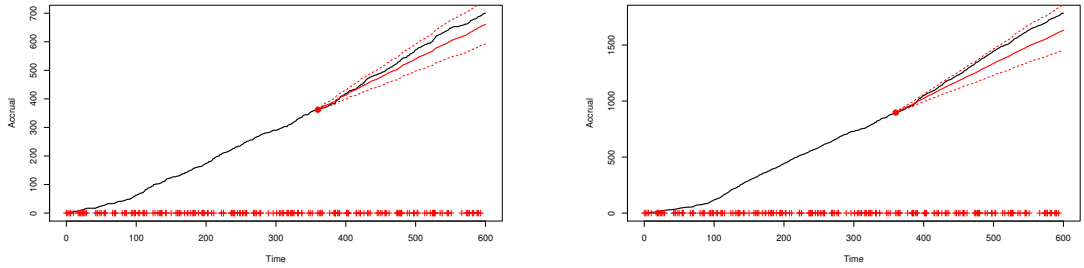


Figure 3.6.4: Accrual with forecast predictive mean (solid, red) and 95% prediction bands (red, dashed). Prediction bands are based on the 2.5% and 97.5% quantiles. The forecast begins from a point marked by the red dot and the "+" symbols on the x-axis indicate centre opening times.

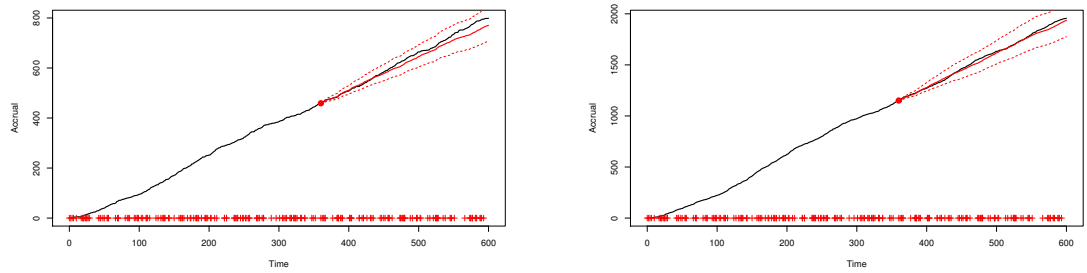


(a) Uniform openings, $\mathbb{E}[\lambda_c^o] = 0.01$ (b) Uniform openings, $\mathbb{E}[\lambda_c^o] = 0.03$

Figure 3.6.5: Accruals (black, solid) with predictive means (red, solid) and 95% prediciton bands (red, dashed) when the true random-effect distribution is a mixture. Prediction bands are based on the 2.5% and 97.5% quantiles. The "+" symbols on the x-axis indicate centre opening times.

tion, $\mathbb{E}[\lambda_c^o] = (\phi_1 + \phi_2)/2$, was set to 0.01 and then 0.03. Figures 3.6.5a and 3.6.5b show example forecasts for accruals with the two different expectations. The more data, that is, the larger $\mathbb{E}[\lambda_c^o]$, the more apparent the discrepancy in the random-effect distribution, and the concomitant predictions, becomes. This is visible in the clearly non-linear diagnostic QQ-plots, and the plotted forecasts (see Appendix A.6). The robustness of predictions comes from the fact that the random effects for initiated centres use re-estimated data-driven distributions, reducing the importance of the random-effect prior; thus the main source of forecasting error comes from the incorrect random-effect prior for new centres. Similar plots for the "clumped" initiation schedule, provided in Appendix A.6, show the same pattern. This mixture distribution of random effects represents the (extreme) scenario where roughly half of the centres recruit the vast majority of patients, with the remaining sites recruiting little to none each. When the ratio of the two means is closer to 1, the model still produces reliable predictions.

We also consider the effect of curve-shape misspecification on predictions, gener-



(a) Uniform openings, $\mathbb{E}[\lambda_c^o] = 0.01$ (b) Uniform openings, $\mathbb{E}[\lambda_c^o] = 0.03$

Figure 3.6.6: Accruals (black, solid) with predictive means (red, solid) and 95% prediciton bands (red, dashed) when the true intensity shape is Weibull, for two different values of $\mathbb{E}[\lambda_c^o]$. Prediction bands are based on the 2.5% and 97.5% quantiles. The "+" symbols on the x-axis indicate centre opening times.

ating data using an intensity proportional to the Weibull density function

$$g_W(t;\theta,k) = \frac{\frac{k}{\theta}\left(\frac{t}{\theta}\right)^{k-1}\exp\left(-(t/\theta)^k\right)}{1-\exp\left(-(\tau/\theta)^k\right)}\tau, \quad \text{so} \quad G_W(t;\theta,k) = \frac{1-\exp\left(-(t/\theta)^k\right)}{1-\exp\left(-(\tau/\theta)^k\right)}\tau,$$

where θ, $k > 0$. We simulated accrual datasets using the Weibull shape with $\theta = 30$ and $k = 1.5$, resulting in the highest recruitment rates occurring two weeks after centre initiation. The random-effect distribution used $\alpha = 1.4$ and two different values ϕ were used: 0.01 and 0.03; Figures 3.6.6a and 3.6.6b show example forecasts. For lower overall recruitment levels, the model still predicts future accrual well. Forecast inaccuracies due to model misspecifiation become more apparent when larger recruitment rates are used. The same pattern is observed when centre initiation times are clumped (see Appendix A.6).

3.7 Data results

We fitted the same set of models to a recruitment dataset of a prostate-cancer clinical trial. The recruitment was carried out across 244 sites. The accrual is presented as the proportion of the total number enrolled. Similarly, time is given as the proportion of the total recruiting period. Figures 3.7.1 and 3.7.2 show the diagnostic QQ-plots for the model fitted to data available at time 0.4. They indicate that there is sufficient concordance between the assumed model and observed enrolment giving validity to potential predictions. Figure 3.7.3 shows the accrual along with forecasts from four different census times. The predictive bands become narrower and parameter uncertainty decreases at each census as more data become available for inference. After the third census, there is an unexpected jump in accrual followed by a drop around the fourth census time, suggesting a global external factor, such as a change in the protocol. Table 3.7.1 shows p-values of the LRT and BST. Initially, when the accrual is still only a small proportion of the total, it is hard to detect the time-inhomogeneity. At later census points, the test outcomes indicate that the rates are not constant.

We compare the proposed framework to the standard homogeneous PG model (3.2.1) as well as a homogeneous Poisson process (HPP) model fitted only to the accrual. We used the same priors as outlined in Section 3.5.1 for fitting the PG model, and the HPP rate estimate was obtained using maximum likelihood. The methods were compared in terms of the predicted completion time of the recruitment for the study with the sampling details outlined in Appendix A.5. Forecast completion time from 6 different census points and can be seen in Figure 3.7.4; the first HPP

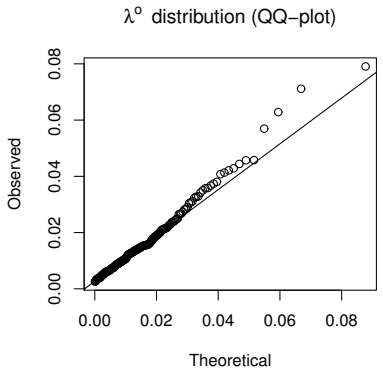


Figure 3.7.1: Re-estimated λ_c^o expectations compared to Gamma$\left(\hat{\alpha}, \hat{\alpha}/\hat{\phi}\right)$ distribution.

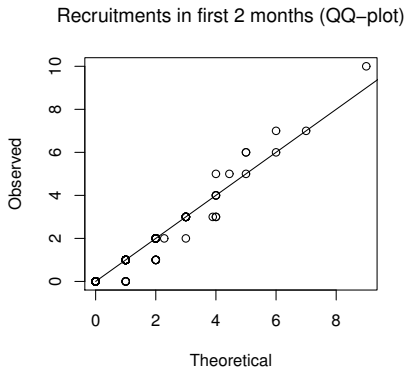


Figure 3.7.2: Observed recruitments compared to the theoretical negative binomial distribution.

predictions were centred at 3.67 and 1.84 which were outside the plot's range. The proposed framework produces better point predictions, especially at earlier interim analyses, and more closely represents the true uncertainty. The HPP predictions near the end of the trial are very accurate. At this point, the majority of the centres having already been initiated and have been recruiting for a long period of time. As a result, the total recruitment rates are not changing by much, with the slight decreasing trend offset by the occasional initiation of a new centre. This is a coincidence; if the decay rate had been sharper or shallower, or if fewer or more centres had been initiated then the naive overall Poisson process model would not have fitted as well. The underprediction of the completion time by the proposed model at the census time of $t = 0.71$ is likely a result of the unexpected surge in recruitment at around that time. The surge is examined in more detail in Appendix A.6.

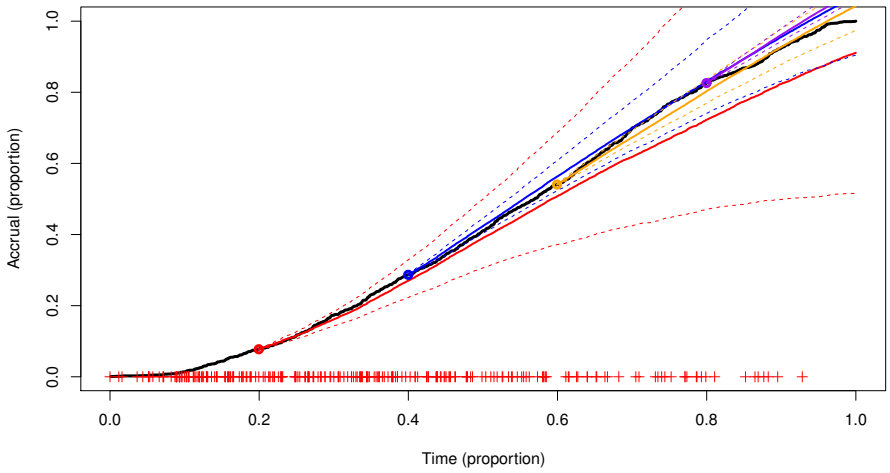


Figure 3.7.3: Accrual (black, solid) for an oncology study; coloured solid lines are mean predictions from census times, dashed lines are the 95% prediction bands, and the "+" symbols indicate opening times of centres.

Census time	BST p-value	LRT p-value
1	0.196	0.226
2	0.012	0.021
3	< 0.001	< 0.001
4	< 0.001	< 0.001

Table 3.7.1: Decay in rate test p-values and the forecasting p-values at four census times.

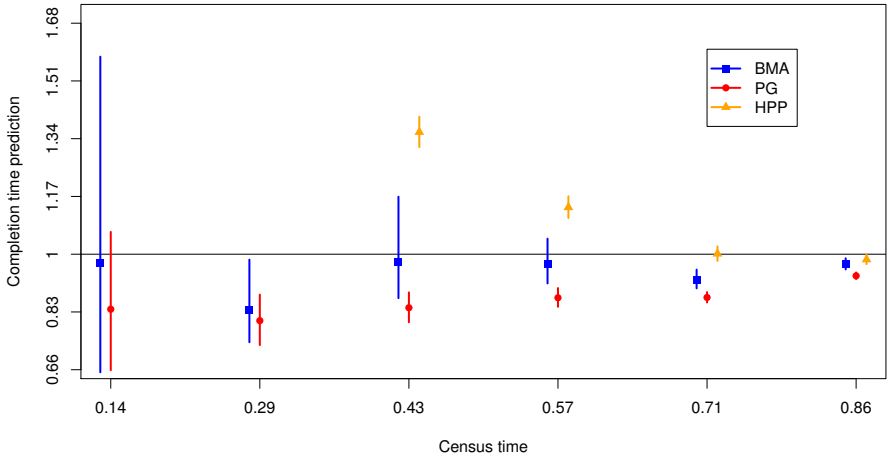


Figure 3.7.4: Predictive distributions for time needed to make the final recruitment in the data example in Section 7, as forecast by three different modelling frameworks: Bayesian model averaging (BMA), time-homogeneous Poisson-gamma (PG) and homogeneous Poisson process fit to accrual only (HPP). The horizontal line represents the true completion time and the prediction positions of the x-axis were off-set by 0.01 for clarity.

3.8 Discussion and further work

We have introduced a general, flexible framework for modelling and predicting recruitment to clinical trials. We suggest two tests for detecting decay in recruitment rates; comparing them both with respect to power and robustness. The particular form of the test statistic allows for a single, simple trial-level test. Alternative forms, such as splitting according to a global time, would either require a test for each centre, massively reducing the power, or estimates of all of the individual centre intensities which would introduces several layers of additional complexity because of the hierarchical connection between the centre intensities. If it were believed *a priori* that a particular global period would be unrepresentative then this time span, and the concomitant recruitment, could simply be removed, albeit at the cost of lower power.

The parametric curve-shape forms chosen for the intensity were based on the features encountered in oncology trials. We found that the model was still robust to moderate model misspecifications in the distribution of the random effect and intensity shape. Other therapeutic areas such as pulmonary or cardio-vascular diseases experience more frequent recruitments and different curve-shapes may be appropriate. As shown in Section 3.6, model misspecification becomes more of a problem at larger enrollment rates. However, with increased frequency, pattern changes in the early months of a centre are easier to identify. Using more complex parametric forms, such as Weibull or generalised gamma shape, could lead to more accurate predictions. Alternatively, if covariate information is available, say $\mathbf{x}_c$ for each centre, the following intensity form motivated by hazard models from survival analysis could be used: $\lambda_c(t) = \lambda_c^o \mathsf{exp}\left(\beta^\top \mathbf{x}_c\right) g\left(t; \mathsf{exp}\left(\eta^\top \mathbf{x}_c\right)\right)$, where λ_c^o are now random effects coming from a $\mathsf{Gamma}(\alpha, \alpha)$ distribution and β and η are vectors of unknown parameters.

As seen in the data example in Section 3.7 there can be external factors modulating the overall accrual. This could potentially be modelled via a short-term, constant global intensity modifier, which would maintain tractability. The framework is not constrained to parametric forms; non-parametric intensity models, such as those using B-splines (for example, Morgan et al. (2019)) or Gaussian processes (for example, Adams et al. (2009)), could be used instead. This, however, could greatly complicate the Bayesian inference and make the intensity extrapolation problem more difficult.

For the curve-shape parameter prior construction, our choice of the quantity of interest R_κ was motivated by simplicity of the form; one could just as well have used $\frac{G_\kappa(t_0/2;\theta)}{G_\kappa(t_0;\theta)}$, albeit with more algebraic manipulations. The general method was aimed

at models with monotonically decreasing intensities. If curve-shapes such as Weibull are considered then constructing sensible priors will be more complicated.

In presenting the method, we condition the inference and prediction on known initiation schedules for the centres. Incorporating stochastic centre initiation models, such as those in Anisimov (2009) and Lan et al. (2019), into the Monte Carlo prediction framework is straightforward, but would complicate the presentation of our methodology without adding novelty. In Appendix A.7, we demonstrate how recruitment can be predicted using our methodology when there is uncertainty in the initiation schedule. For illustration, we imagine a Weibull-distributed delay to each centre's initiation, but any other initiation model could be incorporated in a similar manner. We stress that full prediction intervals should take this uncertainty into account.

In this work, we focus on patient recruitment regardless of the numbers of dropouts observed. In practice, screening failure and patient withdrawal are both prevalent in clinical trials. Assuming the dropouts are independent of the recruitment process, existing survival analysis techniques such as Cox's proportional hazard model (Cox, 1972) or accelerated failure time frailty model (Wei, 1992) could be used in combination with the recruitment model to produce distributions of the numbers of patients in the system at a given time. Such knowledge would be useful to the practitioners and operational researchers in charge of drug-supply chains for the centres.

Anisimov and Fedorov (2007) introduced a method for determining the number of additional centres needed to be initiated for the study to finish on time. With minimal adaptation, the same method can also be used with our model. However, since the

method assumes that all new centres are initiated immediately, it may not apply in all practical scenarios. We would advocate a simulation-based approach, where forecasts based on different centre initiation schedules are compared. As different operational costs can be associated with different schedules, this would become a resource-constrained optimisation problem.

3.9 Software

Software in the form of `R` code is available at https://github.com/SzymonUrbas/ct-recuitment-prediction.

Chapter 4

Exact sequential inference for bounded-intensity Cox processes

4.1 Background

Temporal Poisson process models arise in many areas such as geology (Shlien and Nafi Toksöz, 1970), finance (Cariboni and Schoutens, 2009; Li and Godsill, 2021) or ecology (Warton and Shepherd, 2010), where events such as earthquakes or stock crashes occur in time with some underlying pattern. In many situations, the data are not time-homogeneous, with points exhibiting various time-dependent trends dictated by the *intensity function* $\lambda(t),\ t > 0$. Under the Poisson process model assumption, we have a *time sequence* of n *ordered* points, $\boldsymbol{t} = (t_1, \ldots, t_n)$, observed in $[0, \mathcal{T}]$ with the likelihood of the intensity

$$L(\lambda(\cdot); n, \boldsymbol{t}) = \mathsf{exp}\left(-\int_0^{\mathcal{T}} \lambda(s)\ \mathsf{d}s\right) \prod_{i=1}^{n} \lambda(t_i). \tag{4.1.1}$$

A major drawback of implementing inhomogeneous Poisson process models for general purposes is the need to prespecify the parametric form of λ which requires domain-specific knowledge; one such example appears in Chapter 3 of this thesis. Some approaches aim to bypass this issue by using a basis of spline functions to nonparametrically estimate the functional form (see, for example, Weinberg et al., 2007; Morgan et al., 2019). The resulting inference, however, can be sensitive to the choice of basis and smoothness constraints. Moreover, predictions, whether forward in time or over a censored region, may not fully capture the uncertainty.

A Cox process assumes a hierarchical structure where $\lambda(t)$ is itself a stochastic process, adding a great deal of modelling flexibility (Cox, 1955). This comes at a great cost as, apart from special cases, the exponential-functional term in (4.1.1) is not available in a closed form, which leads to the intractability of any direct likelihood-based inference. This is known as a doubly-intractable problem where even the likelihood function cannot be evaluated, let alone the posterior density. There is a multitude of formulations for Cox process models with different stochastic process priors (see, for example, Cox, 1955; Møller et al., 1998; Fearnhead et al., 2008; Lloyd et al., 2015) In this chapter, we only consider models where the resulting intensity $\lambda(t)$ is bounded (but the bound itself could be unknown). The physical interpretation is that there is some inherent limit to how frequently events can occur in a given setting. We also focus our attention on inference methods which allow for exact Bayesian posterior inference; the exactness is in the Monte Carlo sense.

Adams et al. (2009) introduces the sigmoidal Gaussian Cox process (SGCP) by

letting

$$\lambda(t) = \lambda_0 F(X_t), \tag{4.1.2}$$

where $\lambda_0 > 0$, $F : \mathbb{R} \to (0, 1)$ is an inverse-logit transform, and X_s is a univariate Gaussian process (GP) with a squared-exponential kernel. The main contribution of the article was a very specific data-augmentation scheme which can be used for any bounded monotonic F. The setup considers a set of $\widetilde{N}$ latent *thinned* points $\boldsymbol{r}$ which marginally form a realisation of a Cox process with an intensity $\lambda_0 (1 - F(X_t))$. This is directly motivated by the thinning procedure where such points would be discarded in order to obtain the observed process. The joint likelihood under the data-augmented model becomes

$$\begin{aligned} L_{\mathsf{DA}}(X_t, \lambda_0; n, \boldsymbol{t}, \widetilde{N}, \boldsymbol{r}) = \; & \exp\left(-\lambda_0 \int_0^{\mathcal{T}} F(X_s)\, \mathsf{d}s\right) \prod_{i=1}^{n} \lambda_0 F(X_{t_i}) \\ & \times \exp\left(-\lambda_0 \int_0^{\mathcal{T}} 1 - F(X_s)\, \mathsf{d}s\right) \prod_{i=1}^{\widetilde{N}} \lambda_0 (1 - F(X_{r_i})) \\ = & {\lambda_0}^{n+\widetilde{N}} \mathsf{e}^{-\lambda_0 \mathcal{T}} \prod_{i=1}^{n} F(X_{t_i}) \prod_{i=1}^{\widetilde{N}} (1 - F(X_{r_i})). \end{aligned}$$

Tractable, exact inference now involves a dimension-changing reversible-jump Markov chain Monte Carlo (Green, 1995). The main computational bottleneck is that, at each iteration of the algorithm, one is required to invert the GP covariance matrix which is $\mathcal{O}\left((n + \widetilde{N})^3\right)$, where $\widetilde{N}$ is random. Additionally, if the GP hyperparameters are not specified and thus included in the inference, the posterior correlation between the parameters and the functional form X_t can result in poorly mixing chains.

Teh and Rao (2011) further extends the data-augmentation approach to a more general class of renewal processes modulated by a GP. The work also proposes an al-

ternative MCMC algorithm composed of a Gibbs sampler for the augmented dataset $(\widetilde{N}, \boldsymbol{r})$ conditional on the current position of the chain. The latent variables are sampled from their marginal distribution, hence bypassing the reversible-jump updates. Numerical experiments show how this is an improvement in both computational speed and statistical efficiency. Similar Gibbs algorithms appear in Rao et al. (2017) and Gonçalves and Gamerman (2018); however, the latter points out that the intensity posterior seems to concentrate around different values depending on the inference scheme. Recent work of Alie et al. (2022) argues that the Gibbs samplers incorrectly assume independence of $\boldsymbol{t}$ and $\boldsymbol{r}$ conditional on X_t, which results in the sampler targetting a joint posterior density with respect to a different measure.

Li and Godsill (2021) (the sequel to Li and Godsill (2018)) replace the generic GP prior with a 2-dimensional Langevin diffusion prior, that is, X_t is now a diffusive process which solves a specific stochastic differential equation. The joint distribution of the latent process is still Gaussian but now the process is Markovian, which allows for cheap conditional simulation and joint density evaluations. The paper presents a number of inference schemes for this particular modelling setup, with the main being a novel sequential Markov chain Monte Carlo algorithm which infers the intensity function over subintervals of the domain. The iterations of the algorithm alternate between the reversible-jump moves and local Metropolis-within-Gibbs refinements; the Gibbs steps are akin to those in Teh and Rao (2011). The work also outlines a random-weight particle filter (Fearnhead et al., 2008) where the particle weights are estimated through the data-augmented likelihood L_{DA}. It appears that the resulting weights have a very large variance as the particle approximation of the intensity

posterior is very poor compared to other approaches. Throughout the article, the hyperparameters of the X_t process are fixed, and the proposed methods do not easily lend themselves into hyperparameter inference; the sequential MCMC algorithm does not produce unbiased estimates of the marginal likelihood needed for particle MCMC.

Gonçalves et al. (2020) outline a more general modelling framework where the GP prior is replaced with a generic diffusive process on $\mathbb{R}$, subject to conditions, and F can be replaced by any function mapping to $\mathbb{R}_+$. Inference is now based on the exact algorithm for sampling diffusions (Beskos et al., 2006). The main advantage of the methodology is its applicability to a large class of problems, though it still involves MCMC steps on parameter spaces of varying dimensions.

Some approaches aim to circumvent the computation overhead of exact inference by using a discrete-time approximation for the integral term (see, for example, Lechnerová et al., 2008; Ng and Murphy, 2019). Gunter et al. (2014) reduces the complexity of the inference by considering induced points for the Gaussian process regression and Lloyd et al. (2015) applies a full variational inference framework to further speed up model-fitting. All those schemes introduce some bias that, given current methods, cannot be easily quantified.

In this chapter, we introduce a novel extension of the Poisson Estimator (Beskos et al., 2006) to provide an unbiased estimate of the likelihood (4.2.1) which is then used within a particle MCMC scheme to allow for exact posterior inference on both the stochastic intensity function and the underlying hyperparameters. Section 4.2 describes the particular model employed in this chapter, and Section 4.3 outlines a generic inference scheme based on a random-weight particle filter which is then used to

produce posterior estimates for a pseudo-marginal Metropolis-Hastings algorithm. In Section 4.4, we introduce the novel unbiased estimator of the Cox process likelihood, which we call the Rao-Blackwellised Thinning Estimator, and discuss its variance and the computational cost; general guidelines are given for random-weight particle filter implementations. Section 4.5 illustrates the proposed estimator in practice on a number of synthetic and real data examples, and Section 4.6 contains the discussion along with suggestions for further work.

Notation

Throughout the chapter, we consider a multi-dimensional temporal stochastic process denoted by $\boldsymbol{X}_t,\ t \geq 0$, and its first one-dimensional component by X_t. We use $\boldsymbol{X}_{t_i}$ and X_{t_i}, to denote the respective realisations at times t_i, $i = 1, \ldots, n$. We let $\boldsymbol{X}_{(a,b)}$ be the process contained to the open interval $(a, b),\ b > a \geq 0$, with similar definitions for half-open and closed intervals. For latent variable vectors of realisations at $\boldsymbol{t}$, we use a shorthand notation $\boldsymbol{X}_{\boldsymbol{t}} = (\boldsymbol{X}_{t_1}, \ldots, \boldsymbol{X}_{t_n})$, with a similar definition for $X_{\boldsymbol{t}}$. To avoid confusion when indexing a sequence of vectors, we use subscripts in parentheses; for example, $\boldsymbol{v}_{(k)} \in \mathbb{R}^d,\ k = 0, 1, \ldots$.

We let $Y \sim \mathsf{TruncPois}(\mu; k)$ denote a $(k-1)$-truncated $\mathsf{Pois}(\mu)$ random variable with a mass function

$$\mathbb{P}(Y = y) = \frac{\mu^y \mathsf{e}^{-\mu}}{S_{\mathsf{P}}(k-1; \mu) y!}, \quad y = k, k+1, \ldots, \tag{4.1.3}$$

where $S_{\mathsf{P}}(\,\cdot\,; \mu)$ is the survival function of a standard Poisson random variable with expectation μ. Properties of the distribution can be found in Appendix B.1.

4.2 Model setup

We shall use the work of Li and Godsill (2021) as a starting point for constructing a new model and the corresponding sequential inference scheme, albeit with a number of modifications. Recall that the data are an ordered time-sequence $(n, \boldsymbol{t})$, observed in $[0, \mathcal{T}]$. The points are a realisation of the process conditional on an unobserved realisation of the stochastic intensity $\lambda(\cdot)$ which is assumed to have the form (4.2.1). We set F to be the inverse-logit transform, $F(x) = (1 - \mathsf{e}^{-x})^{-1}$, and X_t to be a single component of a multi-dimensional Ornstein-Uhlenbeck (OU) diffusion process parametrised by a vector $\boldsymbol{\psi}$, with full details given in the next section. We refer to the resulting model as the OU-process-driven sigmoidal Cox process (OU-SCP). The methods, as introduced in this chapter, are fully applicable to any bounded monotonic transformation F and any diffusion X_t provided it can be simulated exactly. We assume that the data are generated as follows:

- $\boldsymbol{X}_t$ is a single realisation of an OU process initialised from its stationary distribution, where the process is described by a stochastic differential equation (SDE) parameterised by $\boldsymbol{\psi}$;
- N_t is a Poisson process with an intensity function $\lambda(t) = \lambda_0 F(X_t)$;
- $(n, \boldsymbol{t})$ is a time sequence obtained from a realisation of N_t observed in $[0, \mathcal{T}]$.

The likelihood under this setup is

$$L(\boldsymbol{X}_t, \boldsymbol{\psi}, \lambda_0; n, \boldsymbol{t}) = {\lambda_0}^n \prod_{i=1}^{n} F(X_{t_i}) \mathsf{exp}\left(-\lambda_0 \int_0^{\mathcal{T}} F(X_s)\, \mathsf{d}s\right). \tag{4.2.1}$$

Given an observation $(n, \boldsymbol{t})$, we wish to perform Bayesian inference and sample from the posterior

$$\pi(\boldsymbol{X_t}, \boldsymbol{\psi}, \lambda_0 | n, \boldsymbol{t}) \propto L(\boldsymbol{X_t}, \boldsymbol{\psi}, \lambda_0; n, \boldsymbol{t}) p_{\mathsf{OU}}(\boldsymbol{X_t} | \boldsymbol{\psi}) \pi_0(\boldsymbol{\psi}, \lambda_0), \tag{4.2.2}$$

where p_{OU} is the joint prior density of the stochastic process evaluated at $\boldsymbol{t}$ and $\pi_0(\boldsymbol{\psi}, \lambda_0)$ is the joint hyperparameter prior.

4.2.1 Ornstein-Uhlenbeck process prior

We consider a similar diffusion state-space prior on X_t as in Li and Godsill (2021); relevant properties of diffusions can be found in Section 2.2.3 of Chapter 2. *A priori*, we assume the unknown curve X_t is a component of a realisation of an Ornstein-Uhlenbeck process which solves the stochastic differential equation

$$\mathsf{d}\boldsymbol{X}_t = -A\boldsymbol{X}_t \, \mathsf{d}t + \boldsymbol{h} \, \mathsf{d}W_t, \quad t > 0, \tag{4.2.3}$$

$$\boldsymbol{X}_0 \sim \mathsf{N}(\boldsymbol{\mu}_0, \Sigma_0), \tag{4.2.4}$$

where A is a positive definite drift matrix, h is a diffusivity vector term and W_s is a 1-dimensional Wiener process. For A and h, we take

$$A = \begin{bmatrix} \theta_1 & -1 \\ 0 & \theta_2 \end{bmatrix} \text{ and } \boldsymbol{h} = \begin{bmatrix} 0 \\ \sigma \end{bmatrix}, \tag{4.2.5}$$

where $\theta_1,\ \theta_2,\ \sigma > 0$. A scenario where $\theta_1 = \theta_2$ produces a valid model but, as we let the parameters be random for Bayesian inference, this occurs probability 0.

Proposition 4.2.1. *The solution to the SDE (4.2.3) subject to initial condition*

(4.2.4) *is a Gaussian process with the marginal distribution at time* t *being* $\boldsymbol{X}_t \sim \mathsf{N}\left(G_t\boldsymbol{\mu}_0, G_t\left(\Psi_t + \Sigma_0\right)G_t^\top\right)$, *where* $G_t = \mathsf{exp}(-At)$ *and* $\Psi_t = \int_0^t G_s^{-1}\boldsymbol{h}\boldsymbol{h}^\top(G_s^{-1})^\top \,\mathsf{d}s$.

If A and h are as in (4.2.5) then

$$G_t = \mathsf{exp}(-At) = \begin{bmatrix} \mathsf{e}^{-\theta_1 t} & \frac{\mathsf{e}^{-\theta_2 t} - \mathsf{e}^{-\theta_1 t}}{\theta_2 - \theta_1} \\ 0 & \mathsf{e}^{-\theta_2 t} \end{bmatrix},$$

and

$$\Psi_t = \begin{bmatrix} \left(\frac{\sigma}{\theta_2-\theta_1}\right)^2 \left(\frac{\mathsf{e}^{2\theta_2 t}}{2\theta_2} + \frac{\mathsf{e}^{2\theta_1 t}}{2\theta_1} - \frac{2\mathsf{e}^{(\theta_1+\theta_2)t}}{\theta_1+\theta_2} - \frac{\theta_1^2 - 2\theta_1\theta_2 + \theta_2^2}{2\theta_1\theta_2(\theta_2+\theta_1)}\right) & \frac{\sigma^2}{\theta_2-\theta_1}\left(\frac{1}{\theta_1+\theta_2} - \frac{1}{2\theta_2} - \frac{\mathsf{e}^{(\theta_1+\theta_2)t}}{(\theta_1+\theta_2)} + \frac{\mathsf{e}^{2\theta_2 t}}{2\theta_2}\right) \\ \frac{\sigma^2}{\theta_2-\theta_1}\left(\frac{1}{\theta_1+\theta_2} - \frac{1}{2\theta_2} - \frac{\mathsf{e}^{(\theta_1+\theta_2)t}}{(\theta_1+\theta_2)} + \frac{\mathsf{e}^{2\theta_2 t}}{2\theta_2}\right) & \frac{\sigma^2}{2\theta_2}\left(\mathsf{e}^{2\theta_2 t} - 1\right) \end{bmatrix},$$

where Ψ_0 is a matrix of zeros. The proof of Proposition 4.2.1 and the derivation of the above matrices are given in Appendix B.2. The key step of the SDE solution is the substitution $\boldsymbol{X}_t = G_t\boldsymbol{Y}_t$, where $\boldsymbol{Y}_t$ is a Markov process experiencing no drift which leads to $\boldsymbol{Y}_t \sim \mathsf{N}(\boldsymbol{\mu}_0, \Psi_t + \Sigma_0)$. Since A is positive definite, $\boldsymbol{X}_t$ is mean-reverting and we can obtain the limiting stationary distribution

$$\lim_{t\to\infty} \boldsymbol{X}_t \sim \mathsf{N}\left(\mathbf{0}, \begin{bmatrix} \frac{\sigma^2}{2\theta_1\theta_2(\theta_1+\theta_2)} & \frac{\sigma^2}{2\theta_2(\theta_1+\theta_2)} \\ \frac{\sigma^2}{2\theta_2(\theta_1+\theta_2)} & \frac{\sigma^2}{2\theta_2} \end{bmatrix}\right). \tag{4.2.6}$$

In contrast to state-space priors based on a 1-dimensional OU process (Lechnerová et al., 2008; Gonçalves et al., 2020), the first component of $\boldsymbol{X}_t$ is differentiable which can make for a more intuitive model depending on the setting. For particle filter implementation, we need to quantify the relaxation time of the process $\boldsymbol{X}_t$, and in particular its first component. To do this, we examine the cross-covariance matrix at

times 0 and t when the process is initialised from stationarity,

$$\begin{aligned}\mathsf{Cov}\left[\boldsymbol{X}_0, \boldsymbol{X}_t\right] &= \mathsf{Cov}\left[\boldsymbol{Y}_0, G_t \boldsymbol{Y}_t\right] \\ &= \mathsf{Cov}\left[\boldsymbol{Y}_0, \boldsymbol{Y}_0 + (\boldsymbol{Y}_t - \boldsymbol{Y}_0)\right] G_t^\top \\ &= \mathbb{V}[\boldsymbol{Y}_0] G_t^\top,\end{aligned}$$

where the third equality follows from the $\boldsymbol{Y}_t$ process being Markov. The auto-correlation of the first component, $X_{1,t}$, is obtained by evaluating

$$\mathsf{Corr}\left[X_{1,0}, X_{1,t}\right] = \frac{\left[\mathbb{V}[\boldsymbol{Y}_0] G_t^\top\right]_{1,1}}{\left[\mathbb{V}[\boldsymbol{Y}_0]\right]_{1,1}} = \frac{\theta_1 \mathsf{e}^{-\theta_2 t} - \theta_2 \mathsf{e}^{-\theta_1 t}}{\theta_1 - \theta_2}. \tag{4.2.7}$$

4.3 Sequential Inference

We wish to avoid the costly computations and possible poor mixing of chains associated with data-augmentation schemes without compromising on the exactness of inference. To do so, we devise a pseudo-marginal scheme which uses sequential Monte Carlo (SMC) to integrate over the $\boldsymbol{X}_t$ process, without introducing any numerical errors. We begin by dividing up the observation window, $[0, \mathcal{T}]$, using a partition

$$\boldsymbol{\tau} := \quad (0 = \tau_0 < \tau_1 < \ldots < \tau_{M-1} < \tau_M = \mathcal{T}),$$

into $M + 1$ subintervals

$$\{0\}, (0, \tau_1], \ldots, (\tau_{M-2}, \tau_{M-1}], (\tau_{M-1}, \mathcal{T}]$$

with $\Delta_m = \tau_m - \tau_{m-1}$, $m = 1, \ldots, M$; the corresponding portions of the data are $\left(n_m, \boldsymbol{t}_{(m)}\right)$, where

$$\boldsymbol{t}_{(m)} = \{t_i : \tau_{m-1} \leq t_i < \tau_m\} \quad \text{and} \quad n_m = |\boldsymbol{t}_{(m)}|,$$

with $n_0 = 0$ and $\boldsymbol{t}_{(0)} = t_0 = 0$. To simplify the notation we use $y_{(m)} = \left(n_m, \boldsymbol{t}_{(m)}\right)$ and $y_{(0:m)} = \left(y_{(0)}, \ldots, y_{(m)}\right)$. For the time being $\boldsymbol{\tau}$ is a fixed, non-uniform partition; Section 4.4.3 will discuss specific guidelines for choosing $\boldsymbol{\tau}$ for a given combination of data and parameters.

We first examine the ideal filter, where the integral term in the likelihood (4.2.1) can be evaluated exactly; the below procedure mirrors the general filtering algorithm outlined in Section 2.5.1 of Chapter 2. The process is initialised from its stationary distribution, that is, $\pi(\boldsymbol{X}_0|\boldsymbol{\psi})$ as given by (4.2.6). Due to the Markovian property of both the OU process and the Poisson process, we observe that

$$p\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]}|\boldsymbol{X}_{(\tau_{m-2},\tau_{m-1}]}, y_{(0:m-1)}, \boldsymbol{\psi}, \lambda_0\right) = g\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]}|\boldsymbol{X}_{\tau_{m-1}}, \boldsymbol{\psi}\right), \tag{4.3.1}$$

where g is the density of the continuous-time conditional propagation of $\boldsymbol{X}_t$ (Proposition 4.2.1); and

$$\begin{aligned} p\left(y_{(m)}|\boldsymbol{X}_{[0,\tau_m]}, y_{(0:m-1)}, \boldsymbol{\psi}, \lambda_0\right) &= p\left(y_{(m)}|\boldsymbol{X}_{(\tau_{m-1},\tau_m]}, \boldsymbol{\psi}, \lambda_0\right) \\ &= L\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]}, \lambda_0; y_{(m)}\right), \end{aligned}$$

which is the Cox process likelihood. The filtering density of the process conditional

on the observed data follows a recursion relation

$$\pi\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]}|y_{(0:m)},\boldsymbol{\psi},\lambda_0\right)=\frac{p\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]},y_{(m)}|y_{(0:m-1)},\boldsymbol{\psi},\lambda_0\right)}{p\left(y_{(m)}|y_{(0:m-1)},\boldsymbol{\psi},\lambda_0\right)}, \tag{4.3.2}$$

where

$$\begin{aligned}
p\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]},y_{(m)}|y_{(0:m-1)},\boldsymbol{\psi},\lambda_0\right) &= \int p\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]},y_{(m)},\boldsymbol{X}_{(\tau_{m-2},\tau_{m-1}]}|y_{(0:m-1)},\boldsymbol{\psi},\lambda_0\right)\ \mathsf{d}\boldsymbol{X}_{(\tau_{m-2},\tau_{m-1}]}\\
&= \int \pi\left(\boldsymbol{X}_{(\tau_{m-2},\tau_{m-1}]}|y_{(0:m-1)},\boldsymbol{\psi},\lambda_0\right)\\
&\quad\times p\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]}|\boldsymbol{X}_{(\tau_{m-2},\tau_{m-1}]},y_{(0:m-1)},\boldsymbol{\psi},\lambda_0\right)\\
&\quad\times p\left(n_m,\boldsymbol{t}_{(m)}|\boldsymbol{X}_{(\tau_{m-1},\tau_m]},\boldsymbol{X}_{(\tau_{m-2},\tau_{m-1}]},y_{(0:m-1)},\boldsymbol{\psi},\lambda_0\right)\ \mathsf{d}\boldsymbol{X}_{(\tau_{m-2},\tau_{m-1}]}\\
&= \int \pi\left(\boldsymbol{X}_{(\tau_{m-2},\tau_{m-1}]}|y_{(0:m-1)},\boldsymbol{\psi},\lambda_0\right)\\
&\quad\times g\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]}|\boldsymbol{X}_{\tau_{m-1}},\boldsymbol{\psi}\right)\\
&\quad\times L\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]},\boldsymbol{\psi},\lambda_0;y_{(m)}\right)\ \mathsf{d}\boldsymbol{X}_{(\tau_{m-2},\tau_{m-1}]},
\end{aligned}$$

and $p\left(y_{(m)}|y_{(0:m-1)},\boldsymbol{\psi},\lambda_0\right)$ is the above joint density integrated over $\boldsymbol{X}_{(\tau_{m-1},\tau_m]}$. Even if L could be evaluated, the resulting conditional density $\pi\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]}|y_{(0:m)},\boldsymbol{\psi},\lambda_0\right)$ would be intractable. To overcome this intractability, a particle filter can be employed to approximate $\pi\left(\boldsymbol{X}_{(\tau_{m-2},\tau_{m-1}]}|y_{(0:m-1)},\boldsymbol{\psi},\lambda_0\right)$ at iteration $m-1$ by using discrete distribution $\hat{\pi}$ constructed from a finite weighted sample $\{\boldsymbol{X}^{(b)}_{(\tau_{m-2},\tau_{m-1}]},\widetilde{w}^{(b)}_{m-1}\}_{b=1}^{B}$,

$$\hat{\pi}(\boldsymbol{X}_{(\tau_{m-2},\tau_{m-1}]}|y_{(0:m-1)},\boldsymbol{\psi},\lambda_0)=\sum_{b=1}^{B}\widetilde{w}^{(b)}_{m-1}\delta_{\boldsymbol{X}^{(b)}_{(\tau_{m-2},\tau_{m-1}]}}(\boldsymbol{X}_{(\tau_{m-2},\tau_{m-1}]}), \tag{4.3.3}$$

where $\sum_{b=1}^{B}\widetilde{w}^{(b)}_{m-1}=1$. Substituting the approximation into the recursion relation

leads to

$$\hat{\pi}\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]}|y_{(0:m)},\boldsymbol{\psi},\lambda_0\right) \propto \sum_{b=1}^{B} \widetilde{w}_m^{(b)} g\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]}|\boldsymbol{X}_{\tau_{m-1}}^{(b)},\boldsymbol{\psi}\right) L\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]},\lambda_0;y_{(m)}\right). \tag{4.3.4}$$

The sampling, importance resampling (SIR) particle filter (Gordon et al., 1993) produces $\hat{\pi}$ approximations at iteration m through importance sampling. The proposal mechanism at m starts with resampling $\{\boldsymbol{X}_{(\tau_{m-2},\tau_{m-1}]}^{(b)}\}_{b=1}^{B}$ according to the weights $\widetilde{w}_{m-1}^{(b)}$, and each particle is propagated via the conditional prior $g\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]}^{(b)}|\boldsymbol{X}_{\tau_{m-1}}^{(b)},\boldsymbol{\psi}\right)$. Proposing samples from the prior results in the new importance weights only depending on the likelihood given the data in the new subinterval,

$$w_m^{(b)} = L\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]}^{(b)},\lambda_0;y_{(m)}\right).$$

Those weights can then be normalised

$$\widetilde{w}_m^{(b)} = \frac{w_m^{(b)}}{\sum_{i=1}^{B} w_m^{(i)}}, \quad b = 1,\ldots,B,$$

and used to construct particle approximation to the filtering density at iteration m. This exact particle filter cannot be used in practice as it depends on exact evaluations of L; however, it serves as a foundation for the approach that follows.

Suppose we instead have access to some function that can produce non-negative estimates of $L\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]},\lambda_0;y_{(m)}\right)$ at each subinterval, which we call $\mathcal{L}$. We assume that the function produces those estimates through the simulation of random points $\tilde{\boldsymbol{t}}_{(m)}$ on $(\tau_{m-1},\tau_m]$, through a conditional distribution $f_{\mathcal{L}}\left(\cdot\,|y_{(m)},\lambda_0,\Delta_m\right)$; more generally, $\mathcal{L}$ could involve the simulation of any kind of auxiliary random variables. For producing the estimate using $\mathcal{L}$, $\boldsymbol{X}_{(\tau_{m-1},\tau_m]}$ only needs to be sampled at $\left(\boldsymbol{t}_{(m)},\tilde{\boldsymbol{t}}_{(m)}\right)$,

conditional on the discrete sampling locations in the previous subintervals. If the estimates produced by $\mathcal{L}$ are unbiased then

$$L\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]},\lambda_0;y_{(m)}\right)=\int\mathcal{L}\left(\boldsymbol{X}_{\boldsymbol{t}_{(m)}},\boldsymbol{X}_{\tilde{\boldsymbol{t}}_{(m)}},\lambda_0;y_{(m)};\tilde{\boldsymbol{t}}_{(m)}\right)f_{\mathcal{L}}\left(\tilde{\boldsymbol{t}}_{(m)}|y_{(m)},\lambda_0,\Delta_m\right)\,\mathsf{d}\tilde{\boldsymbol{t}}.$$

The continuous $\boldsymbol{X}_{(\tau_{m-1},\tau_m]}$ only interacts with the data through the discrete collection $(\boldsymbol{X}_{\boldsymbol{t}_{(m)}},\boldsymbol{X}_{\tilde{\boldsymbol{t}}_{(m)}})$ needed to estimate the likelihood. Sampling at the endpoint of the preceeding subinterval, τ_{m-1}, to obtain the transition density (4.3.1) is not needed and could in fact introduce unnecessary Monte Carlo noise. It is sufficient to condition the diffusion propagation only on the most recent sampling location of the process from the previous intervals, $t^*_{m-1}=\mathsf{max}\left\{\boldsymbol{t}_{(0:m-1)},\tilde{\boldsymbol{t}}_{(0:m-1)}\right\}$;

$$p\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]}|\boldsymbol{X}_{t^*_{m-1}},\boldsymbol{\psi}\right)=\int g\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]}|\boldsymbol{X}_{\tau_{m-1}},\boldsymbol{\psi}\right)g\left(\boldsymbol{X}_{\tau_{m-1}}|\boldsymbol{X}_{t^*_{m-1}},\boldsymbol{\psi}\right)\,\mathsf{d}\boldsymbol{X}_{\tau_{m-1}}.$$

This is still a valid way of proposing diffusion paths, whilst reducing the variation which would otherwise come with sampling $\boldsymbol{X}_{\tau_{m-1}}|\boldsymbol{X}_{t^*_{m-1}}$ followed by $\boldsymbol{X}_{(\tau_{m-1},\tau_m]}|\boldsymbol{X}_{t^*_{m-1}}$. It is helpful to define

$$\boldsymbol{X}_{(m)}=\left(\left(t^*_{m-1},\boldsymbol{X}_{t^*_{m-1}}\right),\boldsymbol{X}_{\boldsymbol{t}_{(m)}},\boldsymbol{X}_{\tilde{\boldsymbol{t}}_{(m)}}\right), \tag{4.3.5}$$

the vector of $\boldsymbol{X}_t$ at all required the sampling locations in the m^{th} subinterval, $(\tau_{m-1},\tau_m]$, along with the starting position; $\boldsymbol{X}_{(0)}$ is the initial position of the diffusion, and we use $\boldsymbol{X}_{(0:m)}$ to denote the union over the first m subintervals.

Using $\mathcal{L}$ instead of L in the relation (4.3.4), allows for the construction of a SIR *random-weight particle filter* (RWPF) (Fearnhead et al., 2008). The RWPF is in fact an auxiliary particle filter on the extended space of $(\boldsymbol{X}_t,\tilde{\boldsymbol{t}})$ and so it approximates

the joint filtering density (see Section 3.2 of Fearnhead et al., 2008). Discarding the auxiliary variables $\tilde{\boldsymbol{t}}$ and only retaining $\boldsymbol{X}_t$ realisations produces approximate samples from the desired marginal distributions (4.3.2). We will now outline the steps involved in the proposed particle filter.

At time 0, we initialise the filter by sampling $\boldsymbol{X}_0^{(b)},\ b = 1, \ldots, B$, from the stationary distribution (4.2.6) and setting it as the beginning of the sample path for each of the B particles, $\boldsymbol{X}_{(0)}^{(b)} = \boldsymbol{X}_0^{(b)}$.

At the m^{th} iteration, $m = 1, \ldots, M$, each particle from the previous iteration is propagated forward in time using the conditional prior distribution in Proposition (4.2.1). This produces $\boldsymbol{X}_{(m)}^{(b)}$ given the most recent realisation of each path $\boldsymbol{X}_{t_{m-1}^{*(b)}}^{(b)}$ (Expression (4.3.5)); the sampling takes place at $\boldsymbol{t}_{(m)}$, and $\tilde{\boldsymbol{t}}_{(m)}^{(b)} \sim f_{\mathcal{L}}\left(\cdot\,|y_{(m)}, \lambda_0, \Delta_m\right)$. Each particle is then weighted by a random weight

$$W_m^{(b)} = \mathcal{L}\left(\boldsymbol{X}_{(m)}^{(b)}, \lambda_0; y_{(m)}; \tilde{\boldsymbol{t}}_{(m)}^{(b)}\right).$$

The normalised weights

$$\widetilde{W}_m^{(b)} = \frac{W_m^{(b)}}{\sum_{i=1}^{B} W_m^{(i)}}$$

are then used to construct an empirical filtering distribution,

$$\pi\left(\boldsymbol{X}_{(\tau_{m-1},\tau_m]}|y_{(0:m)}, \boldsymbol{\psi}, \lambda_0\right) \approx \sum_{b=1}^{B} \widetilde{W}_{(m)}^{(b)} \mathcal{G}_m(\boldsymbol{X}_{(\tau_{m-1},\tau_m]}; \boldsymbol{X}_{(m)}^{(b)}, \boldsymbol{\psi}), \tag{4.3.6}$$

where $\mathcal{G}_m(\,\cdot\,; \boldsymbol{X}_{(m)}^{(b)}, \boldsymbol{\psi})$ is the density of the continuous $\boldsymbol{X}_t$ process in the m^{th} interval, conditional on the finite $\boldsymbol{X}_{(m)}^{(b)}$. Densities $\mathcal{G}_m$ are combination of Dirac masses at $\left(\boldsymbol{X}_{\boldsymbol{t}_{(m)}^{(b)}}, \boldsymbol{X}_{\tilde{\boldsymbol{t}}_{(m)}^{(b)}}\right)^{(b)}$ and conditional densities for the diffusion bridges between neighbouring sampling locations (see, for example, Lemma 3 of Bladt et al., 2016). This

can be thought of as Rao-Blackwellisation over all the OU process paths which pass through $\boldsymbol{X}_{(m)}^{(b)}$. We only store the discrete collection $\boldsymbol{X}_{(m)}^{(b)}$ but, in principle, one could keep track of all the conditional diffusion bridge densities.

As the weighting procedure used in the scheme is only based on estimates of the true likelihood, the discrepancy between the individual particle weights can be large due to the presence of the multiplicative noise. Furthermore, the usual method of estimating the effective sample size at each subinterval through

$$\mathsf{ESS}_m \approx \left(\sum_{b=1}^{B} \left(\widetilde{W}_m^{(b)} \right)^2 \right)^{-1}$$

can be unreliable due to the same noise. To mitigate the potentially rapid particle degeneracy, we employ resampling at every iteration of the filter; residual resampling (e.g., Beadle and Djuric, 1997; Liu and Chen, 1998) is used as it introduces less Monte Carlo noise than standard multinomial resampling and is straightforward to implement (e.g., Douc and Cappé, 2005). The full SIR random-weight particle filter is outlined in Algorithm 4.1.

The above procedure can be used to approximate the posterior distribution of the intensity over the whole observation interval $[0, \mathcal{T}]$, conditional on $(\boldsymbol{\psi}, \lambda_0)$. In addition, taking the average of the unnormalised weights at each iteration and then the product of those averages gives an estimate of the full likelihood of the parameters given the data,

$$L(\lambda_0, \psi; n, \boldsymbol{t}) \approx \widehat{L}_{\mathsf{PM}}(\lambda_0, \psi; n, \boldsymbol{t}) = \prod_{m=1}^{M} \frac{1}{B} \sum_{b=1}^{B} W_m^{(b)}. \tag{4.3.7}$$

This estimator is unbiased (Pitt et al., 2012) and can be used for inference of the pa-

rameters through the pseudo-marginal Metropolis-Hastings (PMMH) algorithm with

$$\pi(\lambda_0, \psi|n, \boldsymbol{t}) \approx \hat{\pi}_{\mathsf{PM}}(\lambda_0, \psi|n, \boldsymbol{t}) \propto \pi_0(\lambda_0, \psi)\widehat{L}_{\mathsf{PM}}(\lambda_0, \psi; n, \boldsymbol{t}).$$

Provided that $\mathcal{L}$ produces unbiased estimates, $\hat{\pi}_{\mathsf{PM}}$ will be an unbiased estimator (up to a normalising constant) and the resulting MCMC algorithm will satisfy the detailed balance with respect to the original posterior π (Expression (4.2.2)) (Beaumont, 2003; Andrieu and Roberts, 2009); see Chapter 2 Section 2.4.2 for details. Retaining a path of $\boldsymbol{X}_{(0:m)}$ at each iteration produces samples from the intensity posterior marginalised over the hyperparameters. Algorithm 4.2 outlines the PMMH update.

The efficiency of any inference scheme for $\boldsymbol{X}_t$ and $(\boldsymbol{\psi}, \lambda_0)$ will depend on the variance of the estimates produced by $\mathcal{L}$. In the next section, we introduce a novel method for unbiased estimation of the Cox process likelihood function.

4.1 Random-weight Particle Filter:

1. **Input:** data $(n, \boldsymbol{t})$; partition $\boldsymbol{\tau}$; dominating rate λ_0; hyperparameters $\boldsymbol{\psi}$; likelihood estimator function $\mathcal{L}$ with conditional joint density of random points f; number of particles B
2. **Output:** Samples from the posterior filtering distribution $\left\{\boldsymbol{X}_{(1:m)}^{(b)}\right\}_{b=1}^{B}$; estimate of pseudo-marginal likelihood $\widehat{L}_{\mathsf{PM}}(\boldsymbol{\psi}, \lambda_0; n, \boldsymbol{t})$
3. **Initialise:** Sample $\boldsymbol{X}_{(0)}^{(b)} = \boldsymbol{X}_0^{(b)}$ from stationary distribution (4.2.6) for $b = 1, \ldots, B$; set $\widehat{L}_{\mathsf{PM}}(\boldsymbol{\psi}; n, \boldsymbol{t}) = 1$;
4. **for** m **in** $1, \ldots, M$ **do**
   (a) **for** b **in** $1, \ldots, B$ **do**
      i. Sample $\tilde{\boldsymbol{t}}_{(m)}^{(b)} \sim f_{\mathcal{L}}\left(\cdot\,|n_m, \boldsymbol{t}_{(m)}, \lambda_0, \Delta_m\right)$;
      ii. Sample $\left(\boldsymbol{X}_{\boldsymbol{t}_{(m)}}, \boldsymbol{X}_{\tilde{\boldsymbol{t}}_{(m)}^{(b)}}\right)^{(b)}$ given $\boldsymbol{X}_{(m-1)}^{(b)}$ via the conditional prior distribution in Proposition 4.2.1, and construct $\boldsymbol{X}_{(m)}^{(b)}$;
      iii. Estimate particle weights $W_{(m)}^{(b)} = \mathcal{L}\left(\boldsymbol{X}_{(m)}^{(b)}, \lambda_0; n_m, \boldsymbol{t}_{(m)}; \tilde{\boldsymbol{t}}_{(m)}^{(b)}\right)$;
   (b) Update pseudo-marginal likelihood $\widehat{L}_{\mathsf{PM}}(\boldsymbol{\psi}, \lambda_0; n, \boldsymbol{t}) \leftarrow \widehat{L}_{\mathsf{PM}}(\boldsymbol{\psi}, \lambda_0; n, \boldsymbol{t}) \times \frac{1}{B}\sum_{b=1}^{B} W_{(m)}^{(b)}$;
   (c) Resample $\left\{\boldsymbol{X}_{(m)}^{(b)}\right\}_{b=1}^{B}$ according to normalised weights $\widetilde{W}_{(m)}^{(b)}$ using residual resampling;

5. **return** $\left(\left\{\boldsymbol{X}_{(1:m)}^{(b)}\right\}_{b=1}^{B},\ \widehat{L}_{\mathsf{PM}}(\boldsymbol{\psi}, \lambda_0; n, \boldsymbol{t})\right)$;

4.2 Pseudo-marginal Metropolis-Hastings:

1. **Input:** unbiased (up to a fixed normalising constant) estimator $\hat{\pi}_{\mathsf{PM}}$ of π; current value $z_0 \sim \pi$ with $\hat{\pi}_{\mathsf{PM}}(z_0)$; proposal distribution q
2. **Output:** new value $z \sim \pi$; estimate $\hat{\pi}_{\mathsf{PM}}(z)$
3. Propose $z' \sim q(\,\cdot\,|z_0)$;
4. Estimate $\pi(z')$ via $\hat{\pi}_{\mathsf{PM}}(z')$;
5. With probability
$$\alpha(z_0, z') = 1 \wedge \frac{\hat{\pi}_{\mathsf{PM}}(z')q(z_0|z')}{\hat{\pi}_{\mathsf{PM}}(z)q(z'|z_0)}$$
set $z \leftarrow z'$ and $\hat{\pi}_{\mathsf{PM}}(z) \leftarrow \hat{\pi}_{\mathsf{PM}}(z')$; else set $z \leftarrow z_0$ and $\hat{\pi}_{\mathsf{PM}}(z) \leftarrow \hat{\pi}_{\mathsf{PM}}(z_0)$;
6. **return** $(z, \hat{\pi}_{\mathsf{PM}}(z))$;

4.4 Thinning estimator

In this section, we focus on the problem of estimating the Cox likelihood function. To do this we employ ideas from the thinning procedure for simulating realisations of an inhomogeneous Poisson process with a bounded intensity (Lewis and Shedler, 1979). A homogeneous process with rate λ^* is first sampled, $(R, \tilde{\boldsymbol{t}})$, where $\lambda^* \geq \lambda(t),\ t \in [0, \mathcal{T}]$. Each point is rejected (thinned) with probability $1 - \frac{\lambda(\tilde{t}_i)}{\lambda^*},\ i = 1, \ldots, R$, and the set of the remaining points forms a realisation of the desired process; Algorithm 4.3 details the procedure. Simulation through thinning is the basis for our methodology.

4.3 Thinning:

1. **Input:** intensity function $\lambda(t)$; dominating rate λ^*; observation interval $[0, \mathcal{T}]$
2. **Output:** realisation of a Poisson process with intensity $\lambda(t)$ on $[0, \mathcal{T}]$, $(n, \boldsymbol{t})$
3. Sample $R \sim \mathsf{Pois}(\lambda^* \mathcal{T})$;
4. Sample locations $\tilde{t}_i \sim \mathsf{Unif}[0, \mathcal{T}],\ i = 1, \ldots, R$;
5. Reject each point with probability $1 - \frac{\lambda(\tilde{t}_i)}{\lambda^*},\ i = 1, \ldots, R$;
6. Set $\boldsymbol{t}$ to be a vector of the remaining, sorted points, and n the vector length
7. **return** $(n, \boldsymbol{t})$;

For the clarity of presentation in this section, we assume fixed hyperparameters $\boldsymbol{\psi}$, and we introduce the methodology for a single-interval filter, that is, inference with a partition $\boldsymbol{\tau} := \ (0 = \tau_0 < \tau_1 = \mathcal{T})$ and $\Delta = \tau_1 - \tau_0 = \mathcal{T}$. Since the Poisson process is Markov, the observations in one subinterval are independent of the observations in the other. As a result, the method still holds for partitions with multiple subintervals, where the likelihood is estimated sequentially for each portion of the partition, conditional on the $\boldsymbol{X}_t$ process from the subinterval before it. Taking the product over all those estimates gives the desired likelihood over the whole dataset.

The standard method of obtaining unbiased estimates of the likelihood (4.2.1) is to use the *Poisson Estimator* (PE) for the exponential-functional term (Wagner, 1989; Beskos et al., 2006). We note that Fearnhead et al. (2008) offers an extension called the Generalised Poisson Estimator, where the method is modified to work on unbounded forms of F for general partially-observed diffusion problems. However, in the bounded-intensity Cox process setting this generalisation does not significantly improve upon PE, thus we omit it from the investigations. The Cox process likelihood

estimator using PE is

$$\mathcal{L}_{\mathsf{Pois}}\left(\boldsymbol{X}_t, \lambda_0; n, \boldsymbol{t}; R, \boldsymbol{u}\right) = \prod_{i=1}^{n} \lambda_0 F(X_{t_i}) \cdot \prod_{j=1}^{R} \overline{F}(X_{u_j}), \tag{4.4.1}$$

$$R \sim \mathsf{Pois}\left(\lambda_0 \mathcal{T}\right),$$

$$u_j \overset{\mathsf{iid}}{\sim} \mathsf{Unif}\left(0, \mathcal{T}\right), \quad j = 1, \ldots, R,$$

where $\overline{F}(x) = 1 - F(x)$. Indeed, taking the expectation of the second product over $(R, \boldsymbol{u})$ gives the required exponential-integral term,

$$\begin{aligned}
\mathbb{E}\left[\prod_{j=1}^{R}(1 - F(X_{u_i}))\right] &= \mathbb{E}_R\left[\mathbb{E}_{\boldsymbol{u}}\left[\prod_{j=1}^{R}(1 - F(X_{u_i}))\middle| R\right]\right] \\
&= \mathbb{E}_R\left[\left(\int_0^{\mathcal{T}} \frac{1 - F(X_s)}{\mathcal{T}}\, \mathsf{d}s\right)^R\right] \\
&= \sum_{k=0}^{\infty} \frac{\mathsf{e}^{-\lambda_0 \mathcal{T}}(\lambda_0 \mathcal{T})^k \left(\int_0^{\mathcal{T}} \frac{(1-F(X_s))}{\mathcal{T}}\, \mathsf{d}s\right)^k}{k!} \\
&= \mathsf{exp}\left(-\lambda_0 \mathcal{T} + \lambda_0 \int_0^{\mathcal{T}} 1 - F(X_u)\, \mathsf{d}s\right) \\
&= \mathsf{exp}\left(-\lambda_0 \int_0^{\mathcal{T}} F(X_s)\, \mathsf{d}s\right).
\end{aligned}$$

This procedure actually estimates the probability of observing no arrivals in $[0, \mathcal{T}]$, that is, the probability of thinning every point of a dominating process with the rate $\lambda^* = \lambda_0$. The stochastic part of the estimating function $\mathcal{L}_{\mathsf{Pois}}$ is constructed regardless of the observed points in the interval. In the following, we extend the Poisson Estimator methodology to now simulate the auxiliary random variables conditional on the observed data, thus estimating the probability of observing the particular dataset through thinning.

Let us suppose that our data, $(n, \boldsymbol{t})$, were indeed obtained by thinning a realisation

of a dominating process with constant rate $\lambda^* = \lambda_0$ on $[0, \mathcal{T}]$, denoted by $(R, \tilde{\boldsymbol{t}})$. Clearly, to observe $(n, \boldsymbol{t})$ we first require $R \geq n$, so $R|(N = n) \sim \mathsf{TruncPois}(\lambda_0\mathcal{T}; n)$. Now, via the thinning process, n of those R points were accepted and $R - n$ were rejected. The accepted points are the observed data and the rejected ones could be any combination of the simulated events, with $\binom{R}{n}$ possible choices. Let $\boldsymbol{J}$ be a binary vector of length R, such that a value of 0 at position j indicates the thinning of the point $\tilde{t}_j$ and construct set $\mathcal{U}_{\boldsymbol{t}} = \{\boldsymbol{J} : \|\boldsymbol{J}\|_1 = n\}$, *i.e.*, the set of thinning choices that lead to exactly n unthinned points. The size of the set is $\binom{R}{n}$ and so it implicitly depends on $\tilde{\boldsymbol{t}}$.

We propose the following method for estimating $L(\boldsymbol{X}_t, \lambda_0; n, \boldsymbol{t})$ using a (naive) function $\mathcal{L}_0$:

1. Sample $R \sim \mathsf{Pois}(\lambda_0\mathcal{T})$, if $R < n$ return $\mathcal{L}_0\left(\boldsymbol{X}_t; n, \boldsymbol{t}; R\right) = 0$; else

2. For each $j = 1, \ldots, R$, sample locations $\tilde{t}_j \overset{\mathsf{iid}}{\sim} \mathsf{Unif}[0, \mathcal{T}]$ and relabel to fix increasing order $\tilde{t}_1 < \tilde{t}_2 < \ldots < \tilde{t}_R$;

3. Sample stochastic process $\boldsymbol{X}_t|\boldsymbol{\psi}$ from its prior distribution and evaluate $\left(X_{t_1}, \ldots, X_{t_n}, X_{\tilde{t}_1}, \ldots, X_{\tilde{t}_R}\right)$

4. Return

$$\mathcal{L}_0\left(\boldsymbol{X}_t; n, \boldsymbol{t}; R, \tilde{\boldsymbol{t}}\right) = \prod_{i=1}^{n} F(X_{t_i}) \cdot \sum_{\boldsymbol{J} \in \mathcal{U}_{\boldsymbol{t}}} \prod_{j=1}^{R} \overline{F}(X_{\tilde{t}_j})^{1-J_j}. \tag{4.4.2}$$

The combination of Step 1 and the conditional estimator of the likelihood (4.4.2)

implies that the overall estimator is

$$\mathcal{L}_0\left(\boldsymbol{X}_t; n, \boldsymbol{t}; R, \tilde{\boldsymbol{t}}\right) = \mathbb{I}_{\{R\geq n\}} \prod_{i=1}^{n} F(X_{t_i}) \cdot \sum_{\boldsymbol{J}\in\mathcal{U}_{\boldsymbol{t}}} \prod_{j=1}^{R} \bar{F}(X_{\tilde{t}_j})^{1-J_j}.$$

An estimate given by $\mathcal{L}_0$ is not unbiased but its expectation is proportional to $L(\boldsymbol{X}_t; n, \boldsymbol{t})$; $\mathcal{L}_0$ it can be modified to obtain a function producing unbiased estimates. Proposition 4.4.1 gives the full unbiased form of the modified estimator.

Proposition 4.4.1. *The modified estimator*

$$\mathcal{L}_1\left(\boldsymbol{X}_t; n, \boldsymbol{t}; R, \tilde{\boldsymbol{t}}\right) = \mathbb{I}_{\{R\geq n\}} \cdot \frac{n!}{\mathcal{T}^n} \prod_{i=1}^{n} F(X_{t_k}) \cdot \sum_{\boldsymbol{J}\in\mathcal{U}_{\boldsymbol{t}}} \prod_{j=1}^{R} \bar{F}(X_{\tilde{t}_j})^{1-J_j} \tag{4.4.3}$$

is an unbiased estimator of (4.2.1).

Proof.

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}_1(\boldsymbol{X}_t; n, \boldsymbol{t}; R, \tilde{\boldsymbol{t}})\right] &= \mathbb{E}\left[\mathbb{I}_{\{R\geq n\}} \cdot \frac{n!}{\mathcal{T}^n} \prod_{k=1}^{n} F(X_{t_k}) \cdot \sum_{\boldsymbol{J}\in\mathcal{U}_{\boldsymbol{t}}} \prod_{j=1}^{R} \bar{F}(X_{\tilde{t}_j})^{1-J_j}\right] \\
&= \sum_{r=n}^{\infty} \mathbb{P}\left(R=r\right) \cdot \frac{n!}{\mathcal{T}^n} \prod_{k=1}^{n} F(X_{t_k}) \cdot \sum_{\boldsymbol{J}\in\mathcal{U}_{\boldsymbol{t}}} \mathbb{E}\left[\prod_{j=1}^{r} \bar{F}(X_{s_j})^{1-J_j}\right] \\
&= \sum_{r=n}^{\infty} \mathbb{P}\left(R=r\right) \cdot \frac{n!}{\mathcal{T}^n} \prod_{k=1}^{n} F(X_{t_k}) \cdot \sum_{\boldsymbol{J}\in\mathcal{U}_{\boldsymbol{t}}} \int_{[0,\mathcal{T}]^{r-n}} \frac{\prod_{j=1}^{r} \bar{F}(X_{s_j})^{1-J_j}}{\mathcal{T}^{r-n}} \,\mathsf{d}\boldsymbol{s} \\
&= \sum_{r=n}^{\infty} \frac{\mathsf{e}^{-\lambda_0\mathcal{T}}\left(\lambda_0\mathcal{T}\right)^r}{r!} \cdot \frac{n!}{\mathcal{T}^n} \prod_{k=1}^{n} F(X_{t_k}) \cdot \sum_{\boldsymbol{J}\in U_{\boldsymbol{t}}} \frac{\left(\int_0^{\mathcal{T}} \bar{F}(X_s) \,\mathsf{d}s\right)^{r-n}}{\mathcal{T}^{r-n}} \\
&= {\lambda_0}^n \prod_{k=1}^{n} F(X_{t_k}) \cdot \mathsf{e}^{-\lambda_0\mathcal{T}} \sum_{r=n}^{\infty} \frac{n!}{r!} \binom{r}{n} \left(\int_0^{\mathcal{T}} \lambda_0 \bar{F}(X_s) \,\mathsf{d}r\right)^{r-n} \\
&= {\lambda_0}^n \prod_{k=1}^{n} F(X_{t_k}) \cdot \mathsf{e}^{-\lambda_0\mathcal{T}} \sum_{r=n}^{\infty} \frac{1}{(r-n)!} \left(\int_0^{\mathcal{T}} \lambda_0 \bar{F}(X_s) \,\mathsf{d}s\right)^{r-n} \\
&= {\lambda_0}^n \prod_{k=1}^{n} F(X_{t_k}) \cdot \mathsf{exp}\left(-\lambda_0\mathcal{T} + \lambda_0 \int_0^{\mathcal{T}} 1 - F(X_s) \,\mathsf{d}s\right) \\
&= {\lambda_0}^n \prod_{k=1}^{n} F(X_{t_k}) \cdot \mathsf{exp}\left(-\lambda_0 \int_0^{\mathcal{T}} F(X_s) \,\mathsf{d}s\right)
\end{aligned}$$

□

The form (4.4.3) is free of λ_0 which comes in through the accept-reject part of the indicator function. The $\mathcal{T}^{-n}$ term arises from the likelihood of uniformly sampling the n observed points on $[0, \mathcal{T}]$, and $n!$ arises from the data appearing in increasing order. The rejection of R means that the estimator can be exactly 0 which, when applied in a PF, would very quickly lead to particle degeneracy if $n > \lambda_0/\mathcal{T}$. For this reason, we marginalise over the number of the dominating points to obtain

$$\mathcal{L}_{\mathsf{Thin}}\left(\boldsymbol{X}_t, \lambda_0; n, \boldsymbol{t}; \widetilde{R}, \tilde{\boldsymbol{t}}\right) = \frac{S_{\mathsf{P}}(n-1; \lambda_0\mathcal{T})n!}{\mathcal{T}^n} \prod_{i=1}^{n} F(X_{t_i}) \cdot \sum_{\boldsymbol{J}\in\mathcal{U}_{\boldsymbol{t}}} \prod_{j=1}^{\widetilde{R}} \bar{F}(X_{\tilde{t}_j})^{1-J_j}, \quad (4.4.4)$$

where now $\widetilde{R} \sim \mathsf{TruncPois}(\lambda_0\mathcal{T}; n)$, and $S_{\mathsf{P}}(\,\cdot\,; \mu)$ denotes the survival function of a Poisson random variable with expectation μ. The estimator remains unbiased by the Rao-Blackwell Theorem. From this point onwards, we refer to (4.4.4) as the *Rao-Blackwellised Thinning Estimator* (RBTE). The Rao-Blackwellisation comes into it in two ways – first, we sum over all possibilities of that result in n unthinned points, given current R and $\boldsymbol{X}_t$; and second, we use a truncated Poisson random variable.

Depending on a given realisation of R it might be more efficient to compute the estimate using a sum of harmonic products, that is

$$\sum_{\boldsymbol{J}\in\mathcal{U}_{\boldsymbol{t}}} \prod_{j=1}^{\widetilde{R}} \bar{F}(X_{\tilde{t}_j})^{1-J_j} = \prod_{k=1}^{\widetilde{R}} \bar{F}(X_{\tilde{t}_k}) \sum_{\boldsymbol{J}\in\mathcal{U}_{\boldsymbol{t}}} \prod_{j=1}^{\widetilde{R}} \left(\frac{1}{\bar{F}(X_{\tilde{t}_j})}\right)^{J_j}$$

This is more efficient when $n < 2\widetilde{R}$. It is notable that when $n = 0$, RBTE reduces to the Poisson Estimator (4.4.1). The generalisation, however, is only valid in the context of Cox process models with bounded intensity.

The remainder of this section is devoted to examining the computational cost and

variance of the estimator, particularly for a given partition $\boldsymbol{\tau}$.

4.4.1 Computational cost

The most expensive parts of the estimation are the simulation of the $\boldsymbol{X}_t$ process and iteration over the elements of $\mathcal{U}_{\boldsymbol{t}}$. Assuming the simulation costs would be similar for any inference scheme, we focus on the latter. Given that $\widetilde{R} \sim \mathsf{TruncPois}(\lambda_0\mathcal{T}; n)$, we can calculate the expected number of the terms in the sum in (4.4.4),

$$\begin{aligned}
\mathbb{E}\left[\binom{\widetilde{R}}{n}\right] &= \mathbb{E}\left[\frac{\widetilde{R}!}{(\widetilde{R}-n)!n!}\right] \\
&= \frac{1}{S_{\mathsf{P}}(n-1;\lambda_0\mathcal{T})}\sum_{r=n}^{\infty}\frac{r!}{(r-n)!n!}\cdot\frac{\mathsf{e}^{-\lambda_0\mathcal{T}}(\lambda_0\mathcal{T})^r}{r!} \\
&= \frac{\mathsf{e}^{-\lambda_0\mathcal{T}}(\lambda_0\mathcal{T})^n}{S_{\mathsf{P}}(n-1;\lambda_0\mathcal{T})n!}\sum_{k=n}^{\infty}\frac{(\lambda_0\mathcal{T})^{r-n}}{(r-n)!} \\
&= \frac{(\lambda_0\mathcal{T})^n}{S_{\mathsf{P}}(n-1;\lambda_0\mathcal{T})n!}.
\end{aligned} \tag{4.4.5}$$

At first, the fact that the cost is exponential in n might suggest the impracticality of RBTE; however, carefully subdividing the observation window using a partition $\boldsymbol{\tau}$ for a given λ_0 can stabilise the cost by spreading it across multiple subintervals. For practical purposes, a suitable choice is one that results in the total estimation costs not exceeding the cost of simulating the diffusion $\boldsymbol{X}_t$. It is still possible that if a large R is sampled, the particular iteration's cost can become prohibitively large. This issue can be overcome by employing an unbiased Monte Carlo estimate of the sum; detailed guidelines are discussed in Section 4.4.3.

4.4.2 Estimator variance

For inference using pseudo-marginal Metropolis-Hastings (Section 4.3), we wish to control the variance of the logarithm of the marginal posterior density $\mathbb{V}[\log \hat{\pi}_{\mathsf{PM}}(\,\cdot\,|n, \boldsymbol{t})]$. In a random-weight particle filter, this term will be influenced by (1) the variance of the likelihood estimates obtained through RBTE for each particle in each subinterval, and (2) the quality of the particle filter with respect to particle degeneracy. The former requires us to establish the behaviour of $\mathbb{V}[\mathcal{L}_{\mathsf{Thin}}(\,\cdot\,;\,\cdot\,;\,\cdot\,)]$ for given $n, \lambda_0, \mathcal{T}$ and then examining how it extends to $\mathbb{V}[\log \widehat{L}_{\mathsf{PM}}(\,\cdot\,;\,\cdot\,)]$ (Expression (4.3.7)); is it better to have few long subintervals or many short ones? The latter is less straightforward as it depends on how well the diffusion simulations can explore the regions of high posterior mass, how often the particles are resampled, and the variance of the particle weight estimates; we address this through a simulation study later in this section.

Let us first begin by establishing that the variance of the estimates produced by RBTE is finite. By construction, the link function F is bounded above by 1 which means that

$$\begin{aligned}\mathcal{L}_{\mathsf{Thin}}\left(\boldsymbol{X}_t, \lambda_0; n, \boldsymbol{t}; \widetilde{R}, \tilde{\boldsymbol{t}}\right) &= \frac{S_{\mathsf{P}}(n-1; \lambda_0\mathcal{T})n!}{\mathcal{T}^n} \prod_{i=1}^{n} F(X_{t_i}) \cdot \sum_{\boldsymbol{J}\in\mathcal{U}_{\boldsymbol{t}}} \prod_{j=1}^{\widetilde{R}} \bar{F}(X_{\tilde{t}_j})^{1-J_j} \\ &\leq \frac{S_{\mathsf{P}}(n-1; \lambda_0\mathcal{T})n!}{\mathcal{T}^n} \prod_{i=1}^{n} F(X_{t_i}) \cdot \binom{\widetilde{R}}{n}.\end{aligned}$$

Since $\mathcal{L}_{\mathsf{Thin}}\left(\boldsymbol{X}_t, \lambda_0; n, \boldsymbol{t}; \widetilde{R}, \tilde{\boldsymbol{t}}\right)$ is non-negative and $\binom{\widetilde{R}}{n}$ has a finite second moment (see Appendix B.1), the variance of estimates produced by RBTE is finite. The full expression for $\mathbb{V}\left[\mathcal{L}_{\mathsf{Thin}}(\,\cdot\,;\,\cdot\,;\,\cdot\,)\right]$ is intractable, but, with some assumptions, we can obtain an approximate form for the stochastic part of the estimator and get some

heuristics for its behaviour on the log-scale.

Consider the simple scenario of some small subinterval of $[0, \mathcal{T}]$ of length Δ in which we observe $n > 0$ arrivals. This is a slight abuse of notation; strictly, it would be n_m arrivals as in Section 4.3 but the m subscript is omitted to aid in clarity. Let us define the quantity

$$\omega = \sum_{\boldsymbol{J} \in U_{\boldsymbol{t}}} \prod_{i=1}^{\widetilde{R}} \left(1 - F(X_{\tilde{t}_i})\right)^{1-J_i},$$

which is the stochastic part of $\mathcal{L}_{\mathsf{Thin}}$. For the following analyses, we use an approximation for X_t, thus removing the additional source of randomness that comes with sampling the process. If Δ is small enough that the intensity only meanders around some mean value, but long enough that $n < \lambda_0 \Delta$, we can use a method-of-moments estimator $F(X_{\tilde{t}_i}) \approx \hat{F} = \frac{n}{\lambda_0 \Delta}$, $i = 1, \ldots, \widetilde{R}$. This is likely to hold since the X_t is continuous and differentiable. Substituting the approximation into the quantity ω, we have

$$\begin{aligned}
\omega &= \sum_{\boldsymbol{J} \in U_{\boldsymbol{t}}} \prod_{i=1}^{\widetilde{R}} \left(1 - \hat{F}\right)^{1-J_i} \\
&= \sum_{\boldsymbol{J} \in U_{\boldsymbol{t}}} \left(1 - \hat{F}\right)^{\widetilde{R}-n} \\
&= \binom{\widetilde{R}}{n} \left(1 - \hat{F}\right)^{\widetilde{R}-n}.
\end{aligned}$$

On the log-scale, we have $\mathbb{V}[\log \omega] = \mathbb{V}\left[\log \binom{\widetilde{R}}{n} + \widetilde{R} \log\left(1 - \hat{F}\right)\right]$. For $n > 0$, $\log \binom{\widetilde{R}}{n}$ is an increasing function in $\widetilde{R}$; take any integer $r \geq n > 0$,

$$\log \binom{r+1}{n} - \log \binom{r}{n} = \log\left(\frac{n!(r+1)!(r-n)!}{n!r!(r+1-n)!}\right) = \log\left(\frac{r+1}{r+1-n}\right) > 0.$$

This implies that $\mathsf{Cov}\left[\log \binom{\widetilde{R}}{n}, \widetilde{R}\right] > 0$ (see, for example, Cuadras, 2002). Since

$\hat{F} \in (0,1)$, we have

$$\log\left(1-\hat{F}\right) < 0 \quad \Rightarrow \quad \mathrm{Cov}\left[\log\binom{\widetilde{R}}{n}, \widetilde{R}\log(1-\hat{F})\right] < 0,$$

which means that $\mathbb{V}[\log\omega] < \mathbb{V}\left[\log\binom{\widetilde{R}}{n}\right] + \mathbb{V}\left[\widetilde{R}\log\left(1-\hat{F}\right)\right]$. Figure 4.4.1a illustrates the ratio

$$\frac{\mathbb{V}\left[\widetilde{R}\log\left(1-\hat{F}\right)\right]}{\mathbb{V}\left[\log\binom{\widetilde{R}}{n}\right]} \tag{4.4.6}$$

for different choices of n and $\lambda_0\Delta$, and Figure 4.4.1b gives the $\log_{10}$-cost (Expression (4.4.5)) associated with each scenario. The two figures suggest that in a regime where $\lambda_0\Delta$ is sufficiently small to allow feasible estimation of the likelihood, the variance of the linear term dominates, that is,

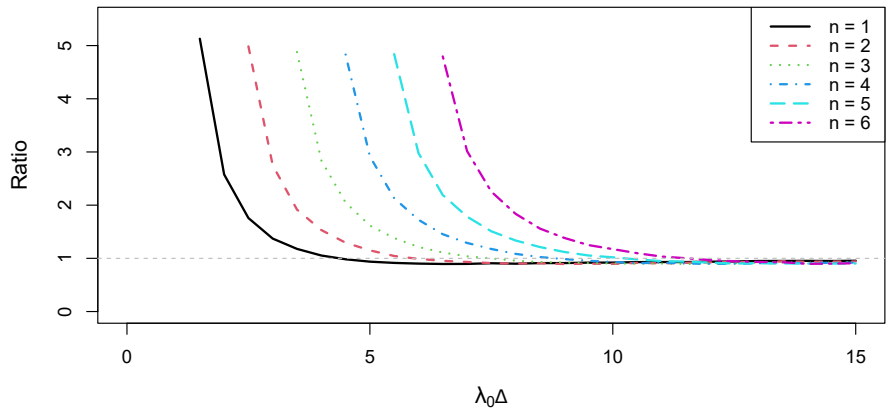
$$\mathbb{V}[\log\omega] \approx \log^2\left(1-\hat{F}\right)\mathbb{V}\left[\widetilde{R}\right], \tag{4.4.7}$$

where $\log^2(\cdot)$ is a shorthand notation for $[\log(\cdot)]^2$. As the variance of a Poisson random variable decreases with truncation, this seems to suggest that $\mathbb{V}[\log\omega]$ should decrease with increasing n, for λ_0 and Δ fixed (see Appendix B.1 for details).

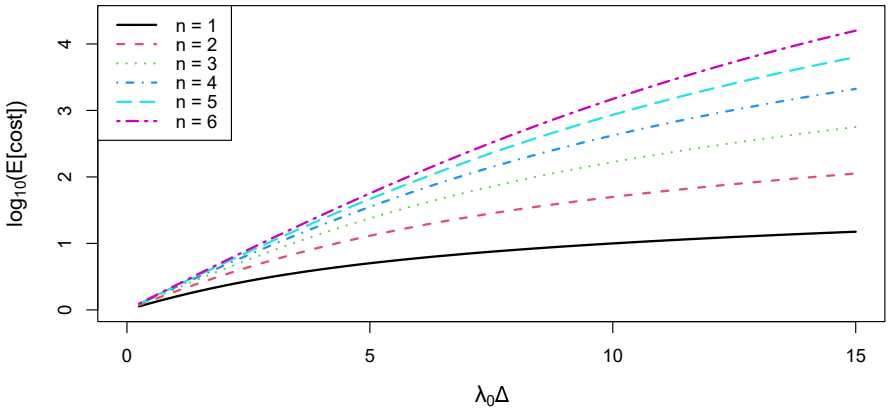
We now turn our attention to subintervals resulting from some general partition of $[0,\mathcal{T}]$,

$$\boldsymbol{\tau} := \quad (0 = \tau_0 < \tau_1 < \ldots < \tau_{M-1} < \tau_M = \mathcal{T}),$$

with $\Delta_m = \tau_m - \tau_{m-1}$ and respective numbers of counts n_m, $m = 1,\ldots,M$. For a single-particle scheme ($B = 1$), the likelihood is estimated based on the data observed in each subinterval, and the estimate of the marginal likelihood over the whole of $[0,\mathcal{T}]$



(a) Ratio of variance components of $\mathcal{L}_{\mathsf{Thin}}$, $\mathbb{V}\left[R \log\left(1-\hat{F}\right)\right] / \mathbb{V}\left[\log\binom{R}{n}\right]$.



(b) Expected computational cost.

Figure 4.4.1: Behaviour of variance components and cost of estimator $\mathcal{L}_{\mathsf{Thin}}$. The expected cost of computing is assumed to be dominated by $\mathbb{E}\left[|\mathcal{U}_t|\right]$ (Expression (4.4.5)).

is the product

$$\widehat{L}_{\mathsf{PM}} = \prod_{m=1}^{M} \binom{\widetilde{R}_m}{n_m} \left(1 - \frac{n_m}{\lambda_0 \Delta_m}\right)^{\widetilde{R}_m - n_m}, \tag{4.4.8}$$

where $\widetilde{R}_m \sim \mathsf{TruncPois}\left(\lambda_0 \Delta_m; n_m\right)$, $m = 1, \ldots, M$. Since the Poisson process is Markov and the estimation mechanism only depends on data in the given subinterval, the estimator components are independent, and so, using the variance approximation (4.4.7),

$$\mathbb{V}[\mathsf{log}\widehat{L}_{\mathsf{PM}}] \approx \sum_{m=1}^{M} \mathsf{log}^2 \left(1 - \frac{n_m}{\lambda_0 \Delta_m}\right) \mathbb{V}\left[\widetilde{R}_m\right]. \tag{4.4.9}$$

To examine the effect of splitting or merging subintervals, consider two adjacent equal-width subintervals, that is $\Delta_1 = \Delta_2 = \Delta$, with respective counts n_1 and n_2. The contribution to the variance (4.4.9) is composed of squared-logarithm leading coefficients and variances of truncated Poisson variates. The leading coefficients of the contributing term have the form $f(z) = \mathsf{log}^2(1 - z)$, restricted to $z \in (0, 1)$. The derivatives of the function are

$$f'(z) = -\frac{2\mathsf{log}(1 - z)}{1 - z}, \tag{4.4.10}$$

$$f''(z) = -\frac{2(\mathsf{log}(1 - z) - 1)}{(1 - z)^2}. \tag{4.4.11}$$

From (4.4.10) we see that f is increasing in z on $(0, 1)$ and additionally from (4.4.11) we know that f is convex. In this context, assuming that the method-of-moments

approximation still holds, the leading coefficients follow the inequality

$$\begin{aligned}\mathsf{log}^2\left(1-\frac{n_1+n_2}{\lambda_0(\Delta_1+\Delta_2)}\right) &= \mathsf{log}^2\left(1-\frac{n_1+n_2}{2\lambda_0\Delta}\right)\\ &= \mathsf{log}^2\left(1-\frac{1}{\lambda_0\Delta}\left(\frac{n_1+n_2}{2}\right)\right)\\ &\leq \frac{1}{2}\left[\mathsf{log}^2\left(1-\frac{n_1}{\lambda_0\Delta}\right)+\mathsf{log}^2\left(1-\frac{n_2}{\lambda_0\Delta}\right)\right],\end{aligned}$$

with the last line due to Jensen's inequality. However, a similar relationship cannot be established for $\mathbb{V}\left[\widetilde{R}_1\right]$, $\mathbb{V}\left[\widetilde{R}_2\right]$ and $\mathbb{V}\left[\widetilde{R}_*\right]$ where

$$\begin{aligned}\widetilde{R}_1 &\sim \mathsf{TruncPois}\left(\lambda_0\Delta; n_1\right),\\ \widetilde{R}_2 &\sim \mathsf{TruncPois}\left(\lambda_0\Delta; n_2\right),\\ \widetilde{R}_* &\sim \mathsf{TruncPois}\left(2\lambda_0\Delta; n_1+n_1\right).\end{aligned}$$

Using the variance for a truncated Poisson random variable, as derived in Appendix B.1, we can directly calculate the relative effect of merging the two adjacent subintervals on the overall variance,

$$\mathsf{RATIO} = \frac{\mathsf{log}^2\left(1-\frac{n_1+n_2}{2\lambda_0\Delta}\right)\mathbb{V}\left[\widetilde{R}_*\right]}{\mathsf{log}^2\left(1-\frac{n_1}{\lambda_0\Delta}\right)\mathbb{V}\left[\widetilde{R}_1\right]+\mathsf{log}^2\left(1-\frac{n_2}{\lambda_0\Delta}\right)\mathbb{V}\left[\widetilde{R}_2\right]}. \tag{4.4.12}$$

The ratio is illustrated under three example scenarios:

(i) $\lambda_0 = 1$, $\Delta = 2.2$ and $n_1 = n_2 = 2 \;\Rightarrow\; \mathsf{RATIO} \approx 0.93$;

(ii) $\lambda_0 = 1$, $\Delta = 5$ and $n_1 = n_2 = 2 \;\Rightarrow\; \mathsf{RATIO} \approx 1.07$;

(iii) $\lambda_0 = 1$, $\Delta = 5$ and $n_1 = 3$, $n_2 = 1 \;\Rightarrow\; \mathsf{RATIO} \approx 0.72$.

The approximate variance ratio suggests that the variance does not always decrease with the merging of adjacent subintervals. However, scenario (ii) corresponds to

$\lambda_0\Delta$ being too large for the linear approximation to hold (see Figure 4.4.1a). This highlights the limitation of the approximation.

Increasing the number of particles (to B) further complicates the analyses. Assuming we resample at each subinterval $m = 1, \ldots, M$ and take the average of the likelihood estimates $\overline{\omega}_m = \sum_{b=1}^{B} \omega_m^{(b)}$, we obtain the pseudo marginal likelihood, up to a multiplicative constant,

$$\widehat{L}_{\mathsf{PM}} = \prod_{m=1}^{M} \overline{\omega}_m.$$

The variance of the log-likelihood estimate for the whole dataset in $[0, \mathcal{T}]$ resulting from the random weights and conditional on the resampling choices has the form

$$\begin{aligned}
\mathbb{V}[\mathsf{log}\widehat{L}_{\mathsf{PM}}] &= \mathbb{V}\left[\sum_{m=1}^{M} \mathsf{log}\overline{\omega}_m\right] \\
&= \sum_{m=1}^{M} \mathbb{V}\left[\mathsf{log}\,\overline{\omega}_m\right] \\
&= \sum_{m=1}^{M} \mathbb{V}\left[\mathsf{log}\,\left(\sum_{b=1}^{B} \omega_m^{(b)}\right)\right],
\end{aligned}$$

where variances in the last line are intractable. We were unable to accurately approximate the $\mathbb{V}\left[\mathsf{log}\,\overline{\omega}_m\right]$ terms and so for the remainder of this section, we employ Monte Carlo integration.

Simulation study

Once again using the approximation $\hat{F}_m = \frac{n_m}{\lambda_0 \Delta_m}$, we illustrate the effect of concatenating two subintervals into one large one. Specifically, we focus on the ratio

$$\frac{\mathbb{V}_{\widetilde{\boldsymbol{R}}_*}\left[\log \sum_{b=1}^{B} \binom{\widetilde{R}_*^{(b)}}{n_1+n_2} \left(1 - \frac{n_1+n_2}{\lambda_0(\Delta_1+\Delta_2)}\right)^{\widetilde{R}_*^{(b)}}\right]}{\mathbb{V}_{\widetilde{\boldsymbol{R}}_1}\left[\log \sum_{b=1}^{B} \binom{\widetilde{R}_1^{(b)}}{n_1} \left(1 - \frac{n_1}{\lambda_0 \Delta_1}\right)^{\widetilde{R}_1^{(b)}}\right] + \mathbb{V}_{\widetilde{\boldsymbol{R}}_2}\left[\log \sum_{b=1}^{B} \binom{\widetilde{R}_2^{(b)}}{n_2} \left(1 - \frac{n_2}{\lambda_0 \Delta_2}\right)^{\widetilde{R}_2^{(b)}}\right]} \tag{4.4.13}$$

where

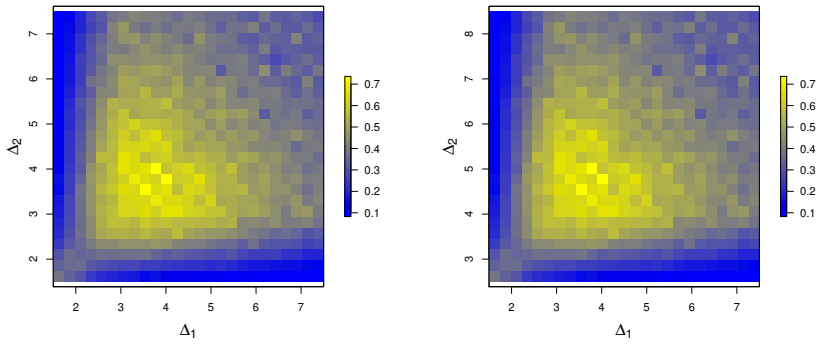
$$\widetilde{R}_1^{(b)} \overset{\text{iid}}{\sim} \mathsf{TruncPois}\left(\lambda_0 \Delta_1; n_1\right),$$

$$\widetilde{R}_2^{(b)} \overset{\text{iid}}{\sim} \mathsf{TruncPois}\left(\lambda_0 \Delta_2; n_2\right),$$

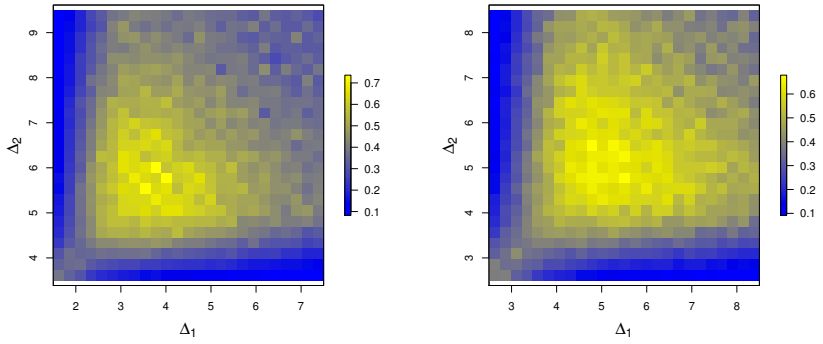
$$\widetilde{R}_*^{(b)} \overset{\text{iid}}{\sim} \mathsf{TruncPois}\left(\lambda_0 (\Delta_1 + \Delta_2); n_1 + n_1\right);$$

the random variables are sampled independently for all $b = 1, \ldots, B$. Without loss of generality, we fix $\lambda_0 = 1$ and examine the ratio when varying Δ for given n values. For the method-of-moments estimator to be valid, we only consider $\Delta > n/\lambda_0$. Figures 4.4.2 and 4.4.3 show the ratios resulting from using $B = 1$ and $B = 100$ particles; we found further increasing B did not affect the results. The results are based on means of $5,000$ Monte Carlo samples per pixel of the raster plot. The figures show how there is a clear and substantial reduction in variance resulting from merging subintervals.

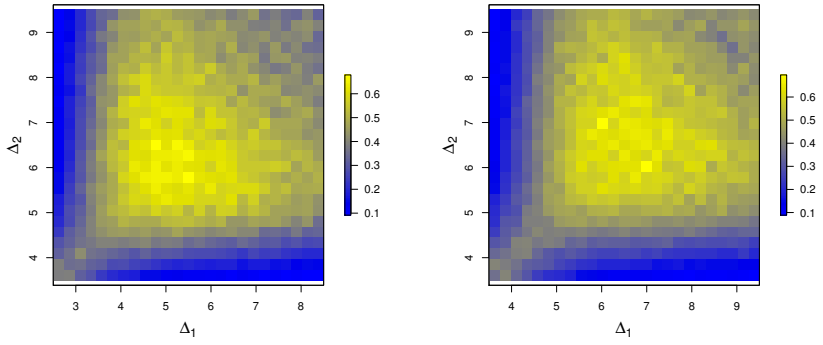
The previous analyses assumed a deterministic intensity approximation. For the purposes of the inference scheme outlined in Section 4.3, we require the behaviour of the estimator in nonparametric intensity estimation when used within a SIR random-weight particle filter. The variance will likely depend on the length of the subintervals



(a) $n_1 = 1, n_2 = 1$ (b) $n_1 = 1, n_2 = 2$

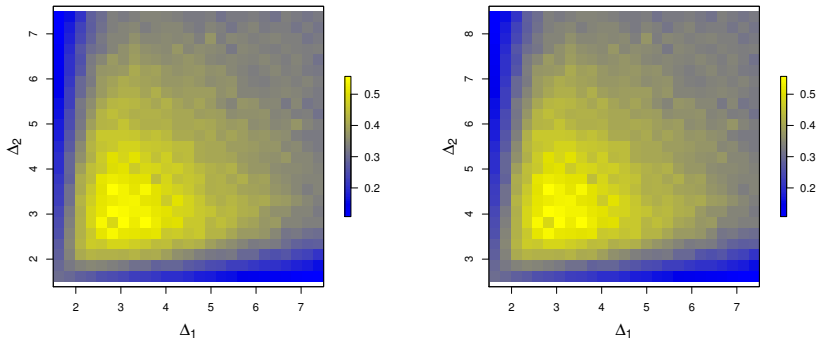


(c) $n_1 = 1, n_2 = 3$ (d) $n_1 = 2, n_2 = 2$



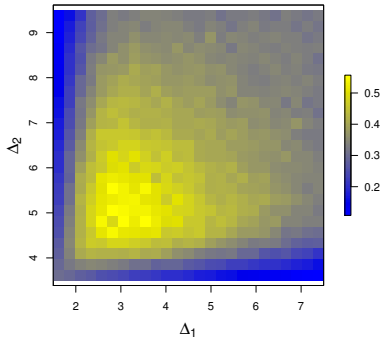
(e) $n_1 = 2, n_2 = 3$ (f) $n_1 = 3, n_2 = 3$

Figure 4.4.2: Monte Carlo estimate of the relative effect on the variance of the estimator, $\mathcal{L}_{\mathsf{Thin}}$, under merging subintervals of lengths Δ_1 and Δ_2, with respective counts n_1 and n_2, as given in the ratio (4.4.13); number of particles is $B = 1$.

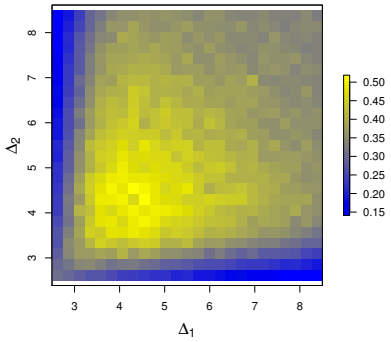


(a) $n_1 = 1, n_2 = 1$

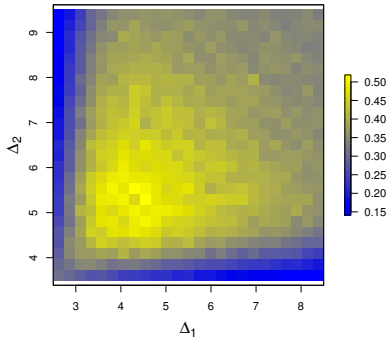
(b) $n_1 = 1, n_2 = 2$



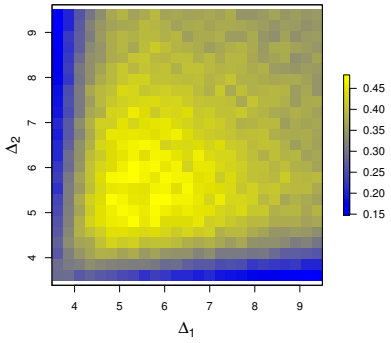
(c) $n_1 = 1, n_2 = 3$



(d) $n_1 = 2, n_2 = 2$



(e) $n_1 = 2, n_2 = 3$



(f) $n_1 = 3, n_2 = 3$

Figure 4.4.3: Monte Carlo estimate of the relative effect on the variance of the estimator, $\mathcal{L}_{\mathsf{Thin}}$, under merging subintervals of lengths Δ_1 and Δ_2, with respective counts n_1 and n_2, as given in the ratio (4.4.13); number of particles is $B = 100$.

Setting	θ_1	θ_2	σ	$\mathsf{IQ}(F(X))$	Δ_{max}
1	5.00	0.2	5.27	$(0.25, 0.75)$	8.25
2	5.00	0.2	3.72	$(0.31, 0.69)$	8.25
3	0.15	2.0	1.85	$(0.25, 0.75)$	11.25
4	0.15	2.0	1.31	$(0.31, 0.69)$	11.25
5	0.22	0.2	0.31	$(0.25, 0.75)$	14.29
6	0.22	0.2	0.22	$(0.31, 0.69)$	14.29

Table 4.4.1: Hyperparameter settings for multivariate OU process priors satisfying (4.2.3).

as compared to the correlation length-scale of the OU process $\boldsymbol{X}_t$, and the resampling procedure after each time step.

We test several scenarios under different hyperparameter settings given in Table 4.4.1. The table provides the inter-quartile range of the intensity prior at stationarity; we chose vague and weakly informative versions. Without loss of generality, the dominating rate λ_0 is fixed at 1 and we let Δ_{max} denote the time-lag at which the first component of $\boldsymbol{X}_t$ (initialised from stationarity) has an autocorrelation of less than 0.5, that is, $\Delta_{\mathsf{max}} = \mathsf{inf}\{s : \mathsf{Corr}\,[X_{1,t}, X_{1,t+s}] < 0.5\}$ where the correlation function is given by (4.2.7). For the experiments, we do not consider observation windows which are greater than $2\Delta_{\mathsf{max}}$, otherwise, the proposed OU process paths can have a very large variance by end of the window. In each experiment, we investigate the effect of merging subintervals under different parameter configurations, the number of observations in each subinterval and the total window lengths in terms of respective Δ_{max} values. The datasets are constructed as follows:

1. one hyperparameter configuration is picked from Table 4.4.1;

2. total window length $\mathcal{T} = 2\Delta$ is chosen to be one of $2\Delta_{\mathsf{max}}$, Δ_{max}, $\Delta_{\mathsf{max}}/2$;

3. a unique combination of (n_1, n_2) is chosen, with a constraint that $n_1 + n_2 \leq 6$ to avoid scenarios with prohibitive computational cost;

4. for each unique choice of $\mathcal{T}$ and (n_1, n_2), 5 different datasets are constructed using different random seeds; for each dataset, the observation window is bisected, and the respective subinterval arrivals are sampled in a stratified manner, *i.e.*,

$$t_k = (k - 1 + U_k)\Delta/n_1, \quad U_k \overset{\text{iid}}{\sim} \mathsf{Unif}(0, 1), \quad k = 1, \ldots, n_1,$$

and

$$t_l = \Delta + (l - 1 + U_l)\Delta/n_2, \quad U_l \overset{\text{iid}}{\sim} \mathsf{Unif}(0, 1), \quad l = 1, \ldots, n_2.$$

To reduce Monte Carlo error in the experiment results, we first set out to identify an appropriate number of particles B needed for two random-weight particle filters (Algorithm 4.1), one on separated subintervals and one on merged. To do this we considered the asymptotic variance, σ^2_{PM}, of the pseudo-marginal log-likelihood estimates (4.3.7) as produced by $\mathcal{L}_{\mathsf{Thin}}$ for the whole dataset $(n, \boldsymbol{t})$,

$$\lim_{B\to\infty} B \times \mathbb{V}\left[\log \widehat{L}_{\mathsf{PM}}(\boldsymbol{\psi}, \lambda_0; n, \boldsymbol{t})\right] = \sigma^2_{\mathsf{PM}}.$$

The asymptotic variance was identified through an initial set of experiemnts and B was chosen such that the approximation $B \times \mathbb{V}[\log \widehat{L}_{\mathsf{PM}}(\boldsymbol{\psi}, \lambda_0; n, \boldsymbol{t})] \approx \sigma^2_{\mathsf{PM}}$ was accurate for the two types of partitions. This ensured that the results would reflect the effect of the partition choice; preliminary experiments showed that $B = 1000$ safely satisfied this condition. For the "split" case, the particle filter employed residual resampling after the first subinterval; we found this to overall reduce the variance of the pseudo-

marginal log-likelihood estimates for a given B.

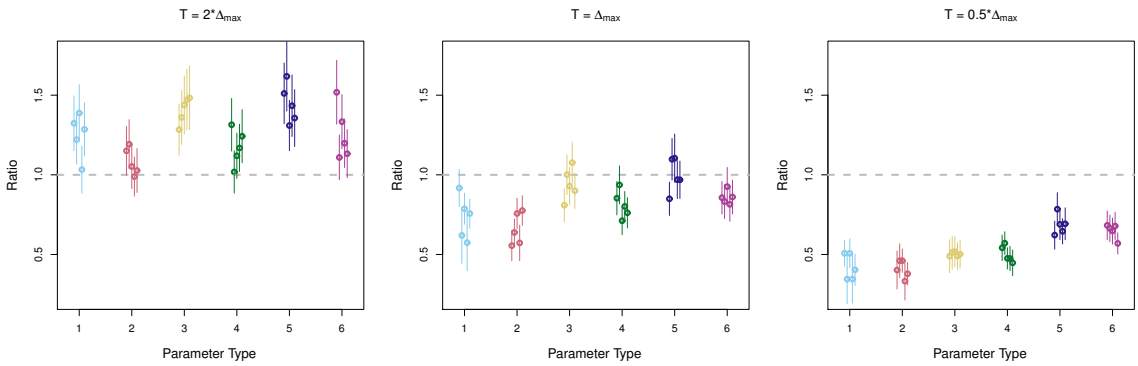
For each synthetic dataset, we ran the two RWPF algorithms 1000 times each, with fixed hyperparameters of $\boldsymbol{X}_t$ corresponding to each Δ_{max}. Monte Carlo estimates of $\mathbb{V}[\log\widehat{L}_{\mathsf{PM}}(\boldsymbol{\psi}, \lambda_0; n, \boldsymbol{t})]$ were obtained and we used them to estimate the relative change in the variance that comes with merging the subintervals. Additionally, nonparametric bootstrapping was used to estimate standard errors for these ratio estimates.

Figure 4.4.4 shows the results for the $(n_1 = 0, n_2 = 1)$ and $(n_1 = 1, n_2 = 0)$ settings, and the results for the remaining combinations can be found in Appendix B.2.1 (Figures B.2.1 - B.2.5). The results show how, in most cases, merging the subintervals is beneficial up to $\Delta = \Delta_{\mathsf{max}}$ (at least up to a total of 6 observed counts). Increasing Δ further actually impairs the estimation. This is likely due to the filter otherwise benefitting from the resampling step. For those $\Delta = 2\Delta_{\mathsf{max}}$ examples, the joint prior for the diffusion was likely too vague compared to the likelihood near the end of the interval and suffered too much from particle degeneracy.

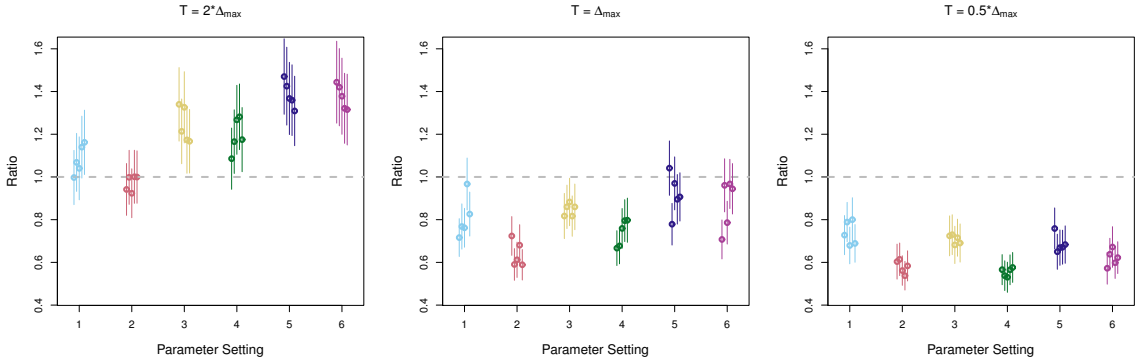
4.4.3 Window partition guidelines

Based on discussed aspects of the Rao-Blackwellised Thinning Estimator, we now outline some generic guidelines for picking the partition $\boldsymbol{\tau}$ based on observed data $(n, \boldsymbol{t})$, and parameters $(\lambda_0, \theta_1, \theta_2, \sigma)$. If used within a particle MCMC scheme, this can be carried out at every iteration. We aim to construct the subintervals according to the following:

(i) Identify a maximum subinterval length Δ_{max} which avoids particle degeneracy



(a) $n_1 = 0,\ n_2 = 1$



(b) $n_1 = 1,\ n_2 = 0$

Figure 4.4.4: Effect of merging adjacent equal-width subintervals on the variance contribution to the pseudo-marginal log-likelihood estimate $\log \widehat{L}_{\mathsf{PM}}$. The ratio represents the relative variance of estimates produced by a random-weight particle filter run on merged subintervals, compared to a filter on split subintervals. Respective Monte Carlo means are denoted by "○", and "|" is used to represent intervals within 2 standard errors from the mean. Subinterval widths are denoted by Δ with $\Delta_{\mathsf{max}} = \inf\{s : \mathsf{Corr}\left[X_{1,t}, X_{1,t+s}\right] < 0.5\}$ and parameter settings are given in Table 4.4.1.

due to poor proposals. For the SIR-RWPF, we set it to be $\inf\{s : \mathsf{Corr}\left[X_{1,t}, X_{1,t+s}\right] < 0.5\}$.

(ii) If a given subinterval contains zero observations, then there is no benefit to further subdividing it provided (i) is satisfied. The variance will likely remain similar but the particle filter cost will be higher due to additional resampling.

(iii) To keep the estimation costs down, we impose a hard constraint on the maximum

number of observations per subinterval, denoted by n^*.

(iv) As demonstrated by experiments, generally including fewer subintervals of longer length results in a lower variance.

If the number of observations in a time-length Δ_{max} exceeds n^*, we (arbitrarily) end a given subinterval at the midpoint of the $n^{*\text{th}}$ and $(n^*+1)^{\text{th}}$ observations. Algorithm 4.4 outlines how the above guidelines are implemented in our framework.

The cost of solving the non-linear equation for Δ_{max} is negligible compared to the simulation costs of the PF. The main reason behind guideline (i) is the use of a SIR particle filter – if the proposal paths were to be constructed differently, longer Δ_{max} could be feasible. The main diagnostic for this tuning can be the effective sample size (ESS) estimate of the particles at the resampling step; if the ESS is too low, shortening the corresponding subintervals can alleviate some of the degeneracy.

The cost of simulating the SDE solution will of course depend on the particular choice of the SDE and the method of simulation. In this work, we only consider an OU process which is sampled directly with no numerical or Monte Carlo approximations. In our experiments, we found that using (iii) $n^* = 3, 4, 5$ results in stable behaviour of $\mathsf{log}\mathbb{V}\left[\mathsf{log}\widehat{L}_{\mathsf{PM}}(\boldsymbol{\psi}, \lambda_0; n, \boldsymbol{t})\right]$ estimates and the corresponding computational cost not exceeding the OU process simulation cost. We suspect a similar guideline will follow for other types of diffusion.

It is also possible that even with moderate $\lambda_0\Delta$ and n^* we may sample $\widetilde{R}$ large enough to increase the cost of single weight estimation by an order of magnitude. This, however, can be circumvented a choosing an upper limit on $|\mathcal{U}_{\boldsymbol{t}}| = \binom{\widetilde{R}}{n}$ and,

once the limit is exceeded, approximating the full summation term by an average of N_{MC} Monte Carlo samples. Random binary vectors $\tilde{\boldsymbol{J}}^{(k)}$ of length $\widetilde{R}$, $k = 1, \ldots, N_{\mathsf{MC}}$, are constructed by sampling, without replacement, n indices and setting them to 1, with the remainder set to 0, independently for each vector. This results in uniform sampling over all elements of $\mathcal{U}_{\boldsymbol{t}}$ and gives an approximation

$$\sum_{\boldsymbol{J}\in\mathcal{U}_{\boldsymbol{t}}}\prod_{j=1}^{\widetilde{R}}\bar{F}(X_{\tilde{t}_j})^{1-J_j} \approx \frac{|\mathcal{U}_{\boldsymbol{t}}|}{N_{\mathsf{MC}}}\sum_{k=1}^{N_{\mathsf{MC}}}\prod_{j=1}^{\widetilde{R}}\bar{F}(X_{\tilde{t}_j})^{1-\tilde{J}_j^{(k)}}.$$

Taking the expectation of a single term of the Monte Carlo sum gives

$$\mathbb{E}\left[\prod_{j=1}^{\widetilde{R}}\bar{F}(X_{\tilde{t}_j})^{1-\tilde{J}_j}\right] = \frac{1}{|\mathcal{U}_{\boldsymbol{t}}|}\prod_{j=1}^{\widetilde{R}}\bar{F}(X_{\tilde{t}_j})^{1-J_j}$$

and so, by linearity of expectation, the Monte Carlo average is then unbiased. In a `C++` implementation, we found that choosing the upper limit $|\mathcal{U}_{\boldsymbol{t}}|$ to be 1000 and setting $N_{\mathsf{MC}} = 100$ avoids the occasional costly iteration without compromising the estimator efficiency.

4.4 Observation window partition:

1. **Input:** Observation window $[0, \mathcal{T}]$; data $(n, \boldsymbol{t})$, hyperparameters $(\theta_1, \theta_2, \sigma)$; target autocorrelation ρ; maximum arrivals per subinterval $n*$;
2. **Output:** M-subinterval partition $\boldsymbol{\tau} := (0 = \tau_0, \tau_1, \ldots, \tau_{M-1}, \tau_M = \mathcal{T})$;
3. **Initialise:** $\boldsymbol{\tau} = \tau_0 = 0,\ \Delta_{\mathsf{max}} = 0,\ \bar{\tau} = 0,\ \bar{N} = 0,\ \bar{n} = 0,\ M = 0$;
4. Solve $\mathsf{Corr}\,[X_{1,0}, X_{1,s}] = \rho$ for s by evaluating (4.2.7) and set $\Delta_{\mathsf{max}} \leftarrow s$;
5. **while** $\mathsf{max}\,\boldsymbol{\tau} < \mathcal{T}$ **do**
   (a) $\bar{\tau} \leftarrow \tau_M + \Delta_{\mathsf{max}}$;
   (b) $\bar{n} \leftarrow \sum_{i=1}^{n} \mathbb{I}_{\{t_i < \bar{\tau}\}} - \bar{N}$;
   (c) **if** $\bar{n} \leq n^*$ **do**
      i. $\tau_{M+1} \leftarrow \bar{\tau}$;
      ii. $\bar{N} \leftarrow \bar{N} + \bar{n}$;
   (d) **else do**

i. $\tau_{M+1} \leftarrow \frac{1}{2}\left(t_{\bar{N}+n^*} + t_{\bar{N}+n^*+1}\right)$;
ii. $\bar{N} \leftarrow \bar{N} + n^*$;

(e) Append $\boldsymbol{\tau}$ with $\mathsf{min}\{\tau_{M+1}, \mathcal{T}\}$;

(f) $M \leftarrow M + 1$;

6. **return** $\boldsymbol{\tau}$;

4.5 Numerical experiments

In this section, we illustrate the performance of the Rao-Blackwellised Thinning Estimator when applied in the Bayesian inference scheme outlined in Section 4.3. We use the framework to sample from the posterior of the OU process $\boldsymbol{X}_t$, hyperparameters $\boldsymbol{\psi} = (\theta_1, \theta_2, \sigma)$ governing the SDE given by (4.2.3) and (4.2.5), and the dominating rate λ_0. We assume the following hyperparameter priors:

$$\lambda_0 \sim \mathsf{Exp}(0.4),$$

$$\theta_1 \sim \mathsf{Gamma}(1.1, 2.2),$$

$$\theta_2 \sim \mathsf{Gamma}(1.1, 0.9),$$

$$\sigma \sim \mathsf{Gamma}(1.1, 0.9),$$

and we initialise the $\boldsymbol{X}_t$ process from its stationary distribution (4.2.6) conditional on $\boldsymbol{\psi}$. Based on initial experiments, we found the above priors able to capture a wide range of intensity behaviours. It may be more appropriate to use context-specific prior distributions depending on a given problem; this is further discussed in Section

4.6. The posterior inference of the parameters $(\boldsymbol{\psi}, \lambda_0)$ is carried out using the pseudo-marginal scheme outlined in Section 4.3, where the logarithms of $(\boldsymbol{\psi}, \lambda_0)$ are jointly updated through a Metropolis-Hastings step with Gaussian random walk proposals. The proposals are appropriately tuned with target $10-15\%$ acceptance rate (Sherlock et al., 2015). The covariance of the proposals was initially set to be $2.56/\sqrt{d} \times \hat{\Sigma}$, where $\hat{\Sigma}$ was a posterior covariance estimate from an initial tuning run, and the step size scale was gradually varied to achieve the required acceptance rate. For tuning the number of particles, B, we refer to Pitt et al. (2012), Sherlock et al. (2015), Doucet et al. (2015) and Nemeth et al. (2016), all suggesting a constant B such that the log-posterior estimator variance at the posterior mean, $\mathbb{V}\left[\log \hat{\pi}_{\mathsf{PM}}(\bar{\lambda}_0, \bar{\boldsymbol{\psi}} | n, \boldsymbol{t})\right]$, is somewhere in the range $0.85 - 3.28$; where computationally feasible, we adopt a conservative choice of 1. The full algorithm was implemented in `C++`.

4.5.1 Synthetic examples

The framework is first illustrated on synthetic datasets generated using 4 different intensity functions:

1. $\lambda_1(t) = 2\exp(-t/15) + \exp\left(-(t-25)^2/100\right)$, $\mathcal{T} = 50, n = 49$;
2. $\lambda_2(t) = 4F(X_t)$, where X_t simulated using $\theta = 0.1, \theta_2 = 0.6, \sigma = 0.5$, $\mathcal{T} = 50, n = 95$;
3. $\lambda_3(t) = 1.5 + \sin(t/5)$, $\mathcal{T} = 100, n = 151$;
4. $\lambda_4(t)$ is piecewise linear, as illustrated in Figure 4.5.1 , $\mathcal{T} = 100, n = 224$.

Method	Error	λ_1	λ_2	λ_3	λ_4
OU-SCP	L_1	0.06	0.31	0.19	0.12
	L_∞	0.58	2.81	0.91	0.73
KDE	L_1	0.06	0.26	0.30	0.13
	L_∞	0.69	3.25	1.09	0.78

Table 4.5.1: Intensity estimate errors for the OU-process-driven sigmoidal Cox process posterior mean and the kernel density estimate.

Posterior sampling given each dataset was carried out through the pseudo-marginal Metropolis-Hastings algorithm. Each chain was run for long enough to obtain $10,000$ samples after discarding the initial burn-in. The $\boldsymbol{X}_t$ process samples were obtained by taking a single particle path from each random-weight particle filter iteration. At each PMMH iteration, the partition $\boldsymbol{\tau}$ is chosen through Algorithm 4.4; we set $n^* = 4$ for $\lambda_1, \lambda_3, \lambda_4$, and $n^* = 3$ for λ_2. The second dataset contained a long gap with no observed events which resulted in large drops in the particle effective sample size – increasing n^* helped to mitigate some of that. Figures 4.5.1 show the results of the OU-SCP intensity inference, along with that obtained via a standard kernel density estimation (KDE) method (Diggle, 1985) with a correction for a bounded domain. We determine the model accuracy through the L_1- and $L_\infty-$errors between the posterior means and the truth, that is, $\left\| \hat{\lambda} - \lambda_{\text{True}} \right\|_p$, $p = 1, \infty$, where $\hat{\lambda}$ are the respective point estimates. The results for OU-SCP and KDE are given in Table 4.5.1.

The previous inferences used RBTE for likelihood estimation. We now compare the efficiency of the proposed estimator compared to the Poisson estimator (4.1.1). For each of the four datasets, we examine CPU time of a single random-weight particle filter run at the respective posterior means and identify the time needed to satisfy $\mathbb{V}\left[\log \hat{\pi}_{\mathsf{PM}}(\bar{\lambda_0}, \bar{\boldsymbol{\psi}} | n, \boldsymbol{t})\right] < 1$. We do this by running SIR-RWPF using each estimator

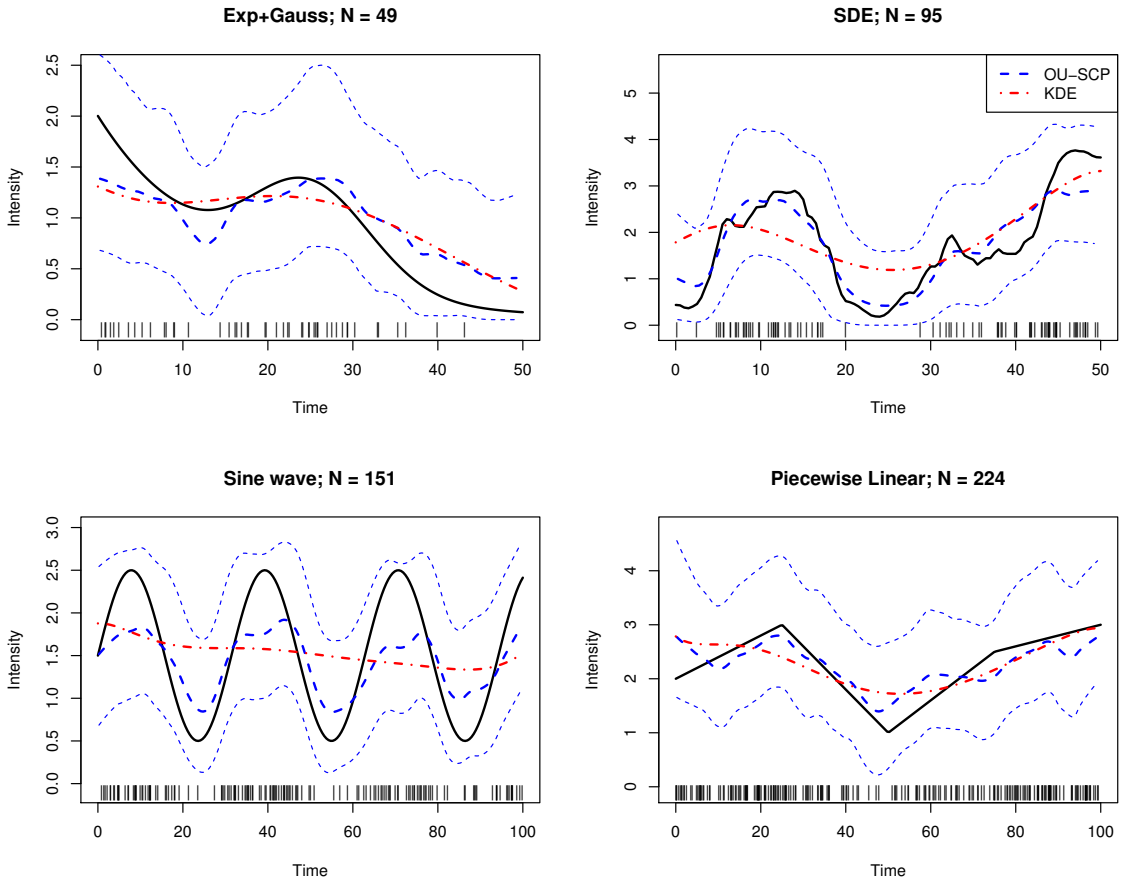


Figure 4.5.1: OU-SCP posterior intensity mean with pointwise 90% credible intervals, and kernel density intensity estimates for four synthetic examples.

and varying the number of particles used; each filter is run $2,500$ times and standard errors are estimated via nonparametric bootstrap. Figure 4.5.2 shows the Monte Carlo estimates the variances and the corresponding computational time per iteration, and Table 4.5.2 provides numerical details of the results. The final number of particles, B, was chosen such that the Monte Carlo mean was at least 2 standard errors below 1. The results show consistent improvement in using RBTE over the Poisson Estimator.

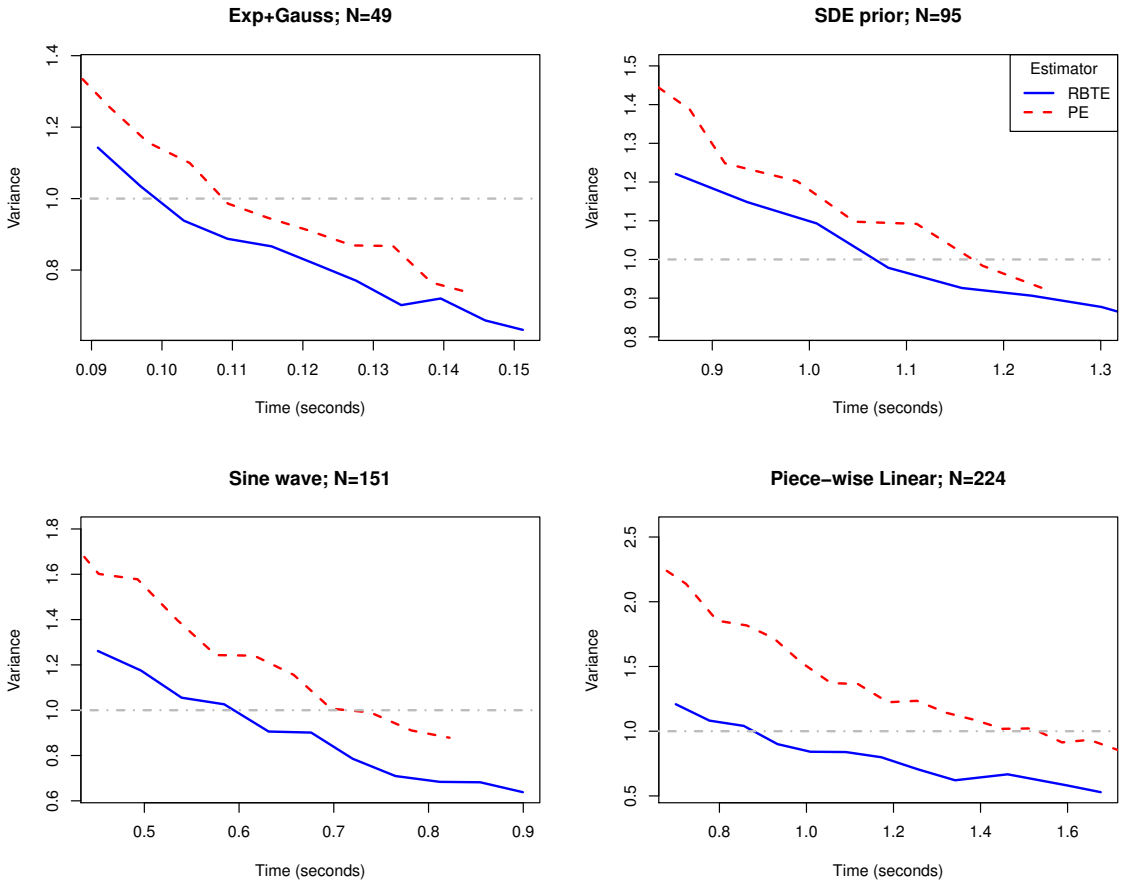


Figure 4.5.2: CPU time (seconds) comparison for random-weight particle filters using the Rao-Blackwellised Thinning Estimator $\mathcal{L}_{\mathsf{Thin}}$ and Poisson Estimator $\mathcal{L}_{\mathsf{Pois}}$.

4.5.2 Coal-mining disasters

We first illustrate the methodology on a classic coal mine disaster dataset (Jarrett, 1979). It represents the period from the 15$^{\text{th}}$ of March 1875 to the 22$^{\text{nd}}$ of March 1962, with events defined as coal mine explosions which killed at least 10 miners in Britain; the data contains 191 such events. In this example, we define the time units in years. This is a moderately-sized dataset and we found that the inference algorithm performed well with 120 particles. Figure 4.5.3a shows the posterior distribution of the intensity mixed over the hyperparameter posteriors. Chain traceplots and

Intensity	n	$\mathcal{T}$	Particles needed		Comp. Time (s)		Relative time
			$\mathcal{L}_{\mathsf{Thin}}$	$\mathcal{L}_{\mathsf{Pois}}$	$\mathcal{L}_{\mathsf{Thin}}$	$\mathcal{L}_{\mathsf{Pois}}$	
λ_1	49	50	34	42	0.103	0.121	0.85
λ_2	95	50	160	190	1.127	1.239	0.93
λ_3	151	100	70	95	0.631	0.782	0.81
λ_4	224	100	60	95	0.934	1.587	0.59

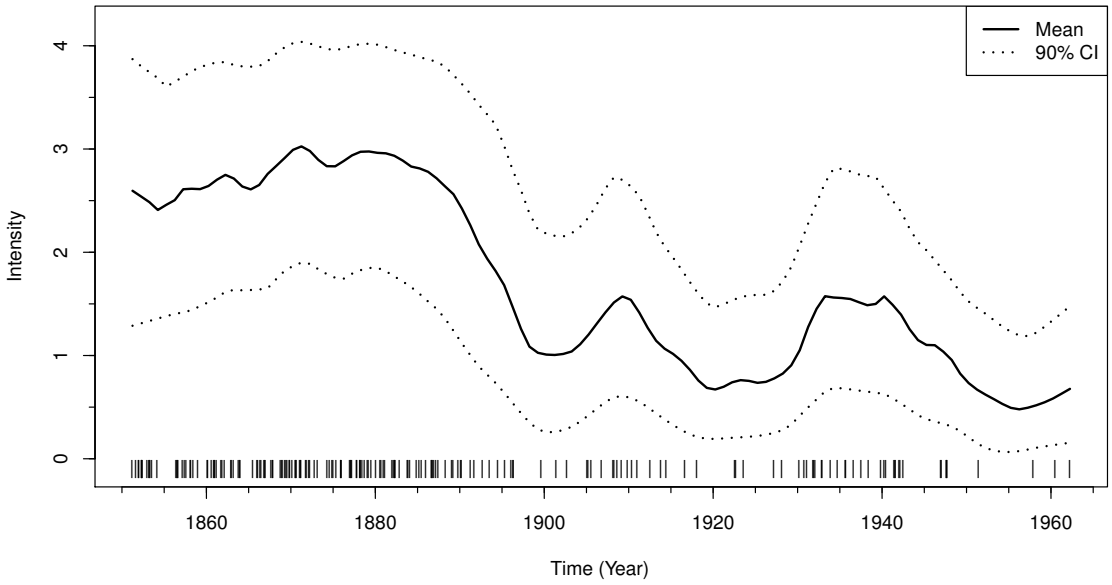
Table 4.5.2: Computational time comparison for SIR-RWPF schemes using $\mathcal{L}_{\mathsf{Thin}}$ and $\mathcal{L}_{\mathsf{Pois}}$. The number of particle B was chosen such that the Monte Carlo mean of the pseudo-marginal log-likelihood variance was at least 2 standard errors below 1. Intensity shapes are given at the start of Section 4.5.1.

autocorrelation plots are given in Appendix B.2.2.

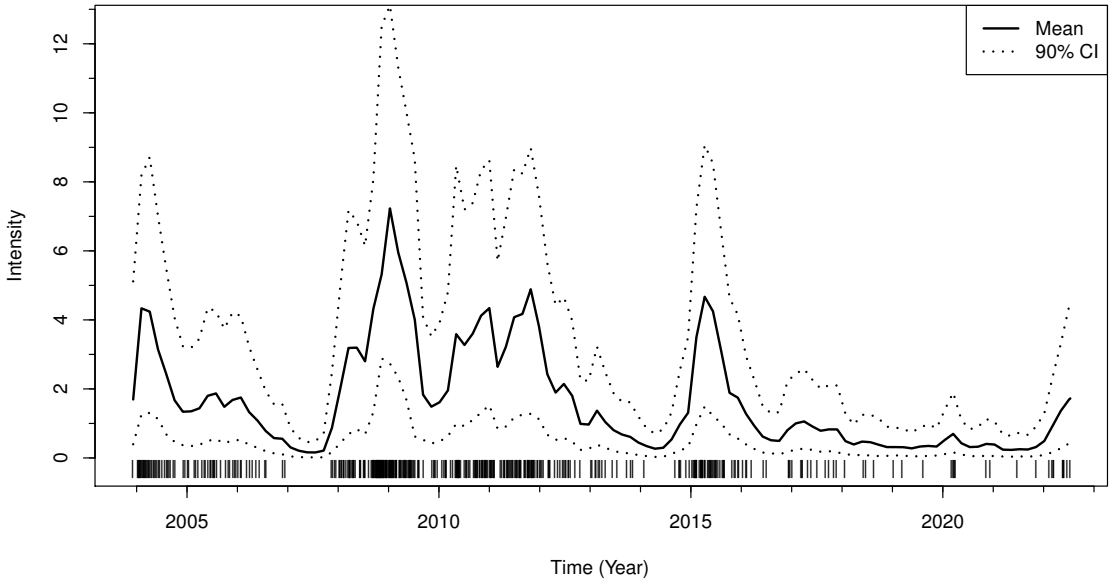
4.5.3 USD-EUR exchange rate

We apply the proposed methodology to a much larger dataset. We consider a time series of the exchange rate between the US Dollar and the Euro, from the 1$^{\text{st}}$ of December 2003 to the 12$^{\text{th}}$ of July 2022 (4857 days).[1] As events, we define instances where the daily exchange rate changed by more than 1%; resulting in 427 events. The inference was performed with a single time unit being a period of 30 days (month). For this large dataset, we found that the algorithm ran sufficiently well with $B = 350$ particles resulting in $\mathbb{V}[\hat{\pi}_{\mathsf{PM}}(\bar{\lambda_0}, \bar{\boldsymbol{\psi}}|n, \boldsymbol{t})] = 1.4$; further increasing the number of particles only produced marginal improvements. The resulting intensity posterior is given in Figure 4.5.3b. Chain traceplots and autocorrelation plots are given in Appendix B.2.2.

[1] The open data were obtained from www.finance.yahoo.com



(a) Coal-mining disasters



(b) USD-EUR exchange rate

Figure 4.5.3: Posterior intensity mean and pointwise 90% credible intervals for two datasets.

4.6 Discussion and further work

We have introduced a novel estimator of the Cox process likelihood functions, the Rao-Blackwellised Thinning Estimator, and devised a sequential Monte Carlo scheme to produce an approximate intensity posterior conditional on observed time-sequence data and latent process hyperparameters. As a byproduct of the SMC algorithm, we obtain a pseudo-marginal likelihood of the hyperparameters and use it in a particle MCMC algorithm. We provide general, setting-agnostic guidelines for partitioning the data, informed by some heuristics and a Monte Carlo study. We applied the method to various examples and show how RBTE outperforms the standard Poisson Estimator in the context of sigmoidal Cox process inference. Based on the results in Section 4.5, the number of particles needed for the random-weight particle filter using RBTE seems to scale better in the size of the data than one using Poisson estimator.

For fixed parameters, the random-weight particle filter is a black-box algorithm, only requiring the end user to specify the number of particles. This can be used in conjunction with statistical software packages such as `PyMC` (Salvatier et al., 2016) to carry sampling on the parameter space. Additionally, parts of the algorithm naturally lend themselves to a parallel infrastructure implementation; for example, groups of particles could be proposed and have their weights estimated on separate cores with the cores only communicating during the resampling step – this may be useful for very large datasets requiring many particles.

There is an inherent problem with identifying the hyperparameters of a latent diffusion using point process data, as noted in Gonçalves and Gamerman (2018).

Gonçalves et al. (2020) suggests that whilst the drift terms governing the diffusion process can be identified from point process data, the diffusivity parameter should be fixed depending on the scale of the problem. In our proposed model, this was the σ parameter which, as we found, could still be identifiable in the examples considered. However, during an initial exploratory investigation on synthetic datasets, we found that the intensity posterior would occasionally concentrate around the mean value of $\frac{1}{2}\lambda_0$ even if the true intensity varied throughout the interval. This coincided with a poor identification of the autocorrelation of the $\boldsymbol{X}_t$ process; time-lags longer than the window length itself were needed to achieve $\mathsf{Corr}\left[X_{1,0}, X_{1,t}\right] < 0.05$. This could be bypassed by constraining the prior on (θ_1, θ_2) such that the resulting correlation never exceeds some value, say $\frac{1}{3}\mathcal{T}$.

Further investigation is required to compare the PMMH scheme employed in this work to the MCMC algorithms based on data augmentation. Ideally, the proposed framework will be compared to that in Adams et al. (2009) and Gonçalves et al. (2020) as the two algorithms are guaranteed to converge to the correct respective posteriors. In this chapter, the number of particles for the PMMH algorithm was guided by a somewhat conservative condition $\mathbb{V}\left[\hat{\pi}_{\mathsf{PM}}(\bar{\lambda_0}, \bar{\boldsymbol{\psi}}|n, \boldsymbol{t})\right] = 1$; larger variances could result in a better overall efficiency of the inference (in terms of effective sample size per second).

A natural extension of the Rao-Blackwellised Thinning Estimator method is the generalisation of the estimator form to work in more general partially-observed diffusion problems. It seems sensible that RBTE could be used in any problem where the Poisson Estimator is used, such as the exact algorithm Beskos et al. (2006) or debi-

asing schemes for time-discretised approximations Jin et al. (2022). This, however, is not trivial as using an unbounded link function F can lead to infinite variances for the Poisson Estimator. Further investigation would likely involve extensions based on the methodology in Fearnhead et al. (2008) or Fearnhead et al. (2010).

Throughout this chapter, we focused solely on Bayesian inference but, in principle, the unbiased estimator could be used for maximum likelihood inference by identifying $\mathsf{argmax}\, L(\boldsymbol{\psi}, \lambda_0; n, \boldsymbol{t})$. This would require the use of a bias correction procedure for the pseudo-marginal log-likelihood (Andrieu et al., 2004) and a suitable optimisation algorithm for stochastic functions (for example, Kingma and Ba, 2015).

The introduced methodology could be modified to work under other data settings. In this chapter, we only considered point process data with exact observations but in many contexts, such as clinical trial enrolment, data undergo interval-censoring with only counts per interval being reported. This complicates the inference as now the integrated intensity term appears twice in the likelihood function. The likelihood for a single interval in this case would be

$$\mathcal{L}\big(\lambda(\cdot); n\big) = \frac{1}{n!}\mathsf{exp}\left(-\int_0^{\mathcal{T}} \lambda(s)\ \mathsf{d}s\right) \cdot \left[\int_0^{\mathcal{T}} \lambda(s)\ \mathsf{d}s\right]^n.$$

Gonçalves et al. (2020) suggest bypassing the issue by using the Gibbs step where each point is sampled uniformly (and independently) within the interval and the uncensored likelihood (4.2.1) is used. Efficiently scaling up the inference to datasets composed of multiple time sequences is not trivial. If the realisations were conditional on the hyperparameters of $\boldsymbol{X}_t$ then the pseudo-marginal log-likelihood estimate, $\widehat{L}_{\mathsf{PM}}$, would be composed of multiple estimates coming from independent particle filters.

Controlling its variance would require increasing the number of particles in each filter. Using parallel infrastructure could alleviate some of this computational burden as particle filters for each time sequence could be run on a separate core. If on the other hand the time sequences shared a common data-generating intensity, the processes could be superposed, essentially multiplying the intensity function by a constant. An appropriate modification to the estimator form could still allow exact inference, albeit at potentially considerable computational costs.

Chapter 5

The Apogee to Apogee Path Sampler

5.1 Hamiltonian Monte Carlo

We wish to draw samples from target distribution π on $\mathsf{X} = \mathbb{R}^d$, where the density can be written in the form $\pi(x) = \mathsf{e}^{-U(x)}$. To do so we utilise Hamiltonian Monte Carlo (HMC) (see, for example, Neal, 2011), which was originally referred to as hybrid Monte Carlo (Duane et al., 1987). The support is extended to now include auxiliary momentum variable $p \in \mathsf{P} = \mathbb{R}^d$ and we denote the density function for the *phase state* $z = (x, p)$ on $\mathsf{Z} = \mathsf{X} \times \mathsf{P}$ as $\tilde{\pi}$.

We use U as the potential energy function of a particle at position x and define K to be the kinetic energy function for an auxiliary momentum variable $p \in \mathbb{R}^d$. The Hamiltonian (total energy) of a particle at x and with momentum p is $\mathcal{H}((x, p)) = U(x) + K(p) = \log \tilde{\pi}(z)$. Given an initial state $z_0 = (x_0, p_0)$, a trajectory along

the contours of $\mathcal{H}$ (which preserves the total energy (Hairer et al., 2003)) can be constructed by solving the system of ordinary differential equations

$$\begin{aligned} \frac{\mathsf{d}x}{\mathsf{d}t} &= \nabla_p K(p) \\ \frac{\mathsf{d}p}{\mathsf{d}t} &= -\nabla_x U(x), \end{aligned} \tag{5.1.1}$$

with t denoting the ficticious time variable used to parametrise the trajectory. In most scenarios, $p \sim \mathsf{N}(0, M)$ independently of x, where M is the positive definite the mass matrix. This choice arises from mirroring Newtonian dynamics on the potential U with $K(p) = \frac{1}{2}p^\top M^{-1}p$ and $\nabla_p K(p) = M^{-1}p$. In general, the system (5.1.1) cannot be solved analytically and so numerical approximations must be employed. One such method is the *leapfrog* integrator (e.g., Hairer et al., 2003). Letting ε be the step-size, the transition $\mathsf{LeapFrog} : (x_0, p_0) \mapsto (x_1, p_1)$ follows the position variant of the integrator,

$$\begin{aligned} p_{0.5} &= p_0 - \frac{\varepsilon}{2}\nabla_x U(x_0), \\ x_1 &= x_0 + \varepsilon \nabla_p K(p_1), \\ p_1 &= p_{0.5} - \frac{\varepsilon}{2}\nabla_x U(x_1). \end{aligned}$$

The integrator applies three shear transformations to (x, p) and so the resulting Jacobian matrix has its determinant equal to one. A particularly useful feature of the leapfrog algorithm is the reversibility due to its symmetry. If we were to instead start at $z_1^- = (x_1, -p_1)$ then with a single iteration of the algorithm above we would arrive at $(x_0, -p_0) = z_0^-$, that is, $\mathsf{LeapFrog}((x_1, -p_1); \varepsilon) = (x_0, -p_0)$. This is also equivalent to using a negative step-size ε which will advance backwards in time,

$\mathsf{LeapFrog}((x_1, p_1); -\varepsilon) = (x_0, p_0)$.

In HMC, at each iteration, a new momentum p_0 is sampled from its marginal Gaussian distribution via a Gibbs move to give z_0 and then the leapfrog integrator is applied L times to produce a state z_L. To satisfy the detailed balance condition, the momentum at the endpoint is then negated to give a proposal $z' = z_L^-$. As the proposal mechanism is deterministic (conditional on z_0) we obtain a symmetric transition function q, that is, $q(z_0|z') = q(z'|z_0) = 1$, with $q(z^*|z_0) = 0$ for any $z^* \neq z'$. In Section 5.5 a "blurred" version of HMC is used (Mackenze, 1989). In it, the integration time $\mathcal{T} = \varepsilon L$ is jittered at each iteration using i.i.d. noise, $\mathcal{T} \sim \mathsf{Unif}(0.8\mathcal{T}^*, 1.2\mathcal{T}^*)$ with $\mathcal{T}^*$ fixed. In this case, the symmetry of q is still preserved since the jitter is sampled independently of z_0. In both standard and blurred HMC, the proposal z' is accepted with probability

$$\alpha(z, z') = 1 \wedge \mathsf{exp}\left(\mathcal{H}(z) - \mathcal{H}(z')\right),$$

which preserves detailed balance with respect to $\tilde{\pi}$ (Neal, 2011). This in fact is a Metropolis-Hastings acceptance probability. Discarding the auxiliary momentum p and only retaining the new x gives a sample from the marginal π. The standard HMC update is summarised in Algorithm 5.1. The appeal of HMC is its potential to give distant proposals which are then accepted with a high MH acceptance probability. The caveat, however, is that two parameters need to be tuned to ensure optimal performance, ε and L, or equivalently ε and $\mathcal{T}$.

5.1 Hamiltonian Monte Carlo:

1. **Input:** target distribution π; position $x_0 \sim \pi$; step-size ε; number of leapfrog steps L

2. **Output:** new position $x \sim \pi$
3. **Initialise:** sample momentum $p \sim \mathsf{N}(0, M)$ and set $z_0 = (x_0, p_0)$;
4. $z_L \leftarrow \mathsf{LeapFrog}(z; \varepsilon, L)$;
5. Set $z' = z_L^-$ and compute $\mathcal{H}(z')$;
6. Sample $z \in \mathcal{S}_0^-$ w.p. $\propto \omega(z'; z_0)$;
7. With probability
$$\alpha(z_0, z') = 1 \wedge \exp\left(\mathcal{H}(z) - \mathcal{H}(z')\right)$$
set $z \leftarrow z'$, else set $z \leftarrow z_0$;
8. **return** x;

The leapfrog integrator is symplectic, which is a much stronger property than simply preserving the volume. The symplecticness of the scheme ensures that the global error of $\mathcal{H}(z)$ along the integrated trajectory is $\mathcal{O}(\varepsilon^2)$, even after multiple leapfrog steps (e.g., Leimkuhler and Reich, 2005; Hairer et al., 2006). As the change in the energy does not strongly depend on the number of iterations L, the MH acceptance probability is not sensitive to the particular choice of integration time. Consequently, ε can be tuned independently of the integration time to produce sufficiently high acceptance rates.

There are guidelines for tuning ε conditional on a fixed $\mathcal{T}$ in the literature. For example, Beskos et al. (2013) suggests that, under mild conditions and as the dimension approaches infinity, the optimal acceptance rate for HMC is 0.651. Setting ε to give that rate gives the optimal discretisation of the path whilst controlling the Hamiltonian discrepancy. However, this is not strictly true as, in practice, the limit is approached slowly and optimal rates can be higher (see, for example, Girolami and

Calderhead, 2011). This is also demonstrated in Section 5.5.

As a result of this general guideline, adaptive MCMC schemes (see, for example, Andrieu and Thoms, 2008) can be devised to optimise ε with respect to the acceptance rate during the run of the algorithm. This can be done using the primal-dual averaging method (Nesterov, 2009), as done, for example, in Hoffman and Gelman (2014), Hoffman et al. (2021) and Mongwe et al. (2021), and is widely used in the `Stan` package (Stan Development Team, 2020).

On the other hand, tuning the integration time $\mathcal{T}$ (or equivalently L for a given ε) is less straightforward as the optimum is highly problem-dependent. Considering the example in Figure 5.1.1, low $\mathcal{T}$ would give proposals close to the original point, whereas a very large $\mathcal{T}$ would give a trajectory that wraps back onto itself, both cases resulting in chains with a higher autocorrelation. An optimal choice of $\mathcal{T}$ is one which maximises the overall displacement from the original point. In particular, if $\mathcal{T}$ is longer than optimal, the trajectory will start coming back towards z_0 resulting in more correlated proposals z' at a higher cost. Figure 5.1.2 illustrates the efficiency of both blurred and unblurred HMC on an asymmetric banana-shaped target based on the regularised Rosenbrock-type function (see Section 5.5.1). The sharp drops as L increases past the optimum for a given choice of ε indicate the costly trajectories overshooting the optimal proposals. The figure additionally highlights the potential multimodality of optimal integration time for HMC caused by the trajectories wrapping back multiple times. The sensitivity to $\mathcal{T}$ can be slightly mitigated by using the blurred version of the algorithm.

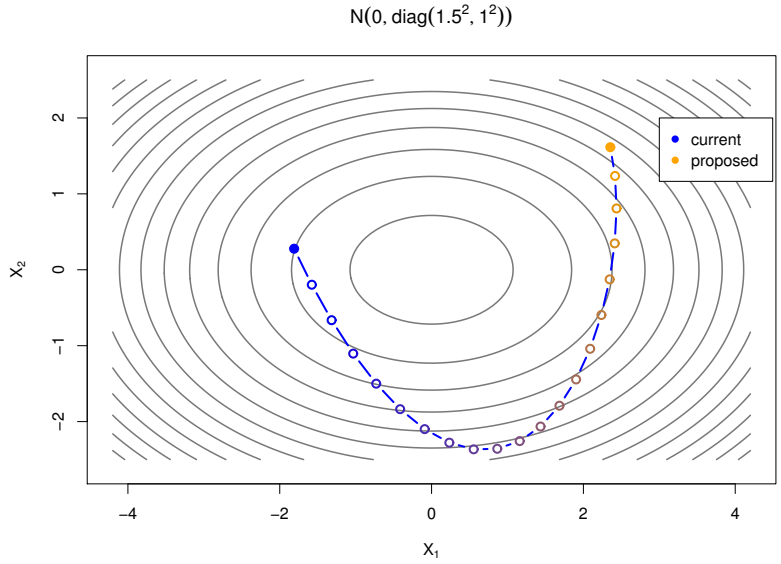


Figure 5.1.1: Approximate Hamiltonian dynamics using the leapfrog integrator on a Gaussian target with diagonal covariance matrix. The component standard deviations are 1.5 and 1. The trajectory starts at blue and ends at orange.

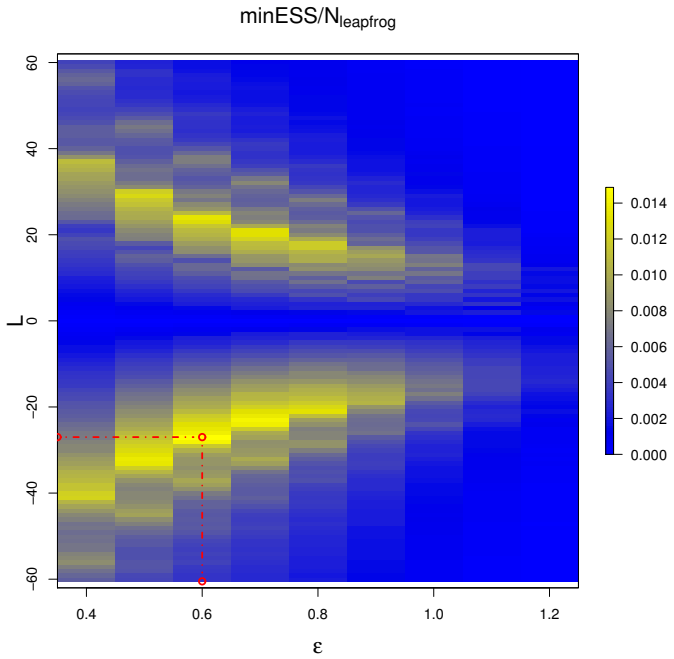


Figure 5.1.2: Performance of HMC on a 40-dimensional banana-shaped target (Section 5.5.1). Negative values of L refer to a blurred version of the algorithm. Performance is measured as the componentwise minimum ESS per gradient evaluation. Red, dashed lines give the global maximum.

5.1.1 The no-U-turn sampler

Hoffman and Gelman (2014) introduces the no-U-turn sampler (NUTS) which aims to eliminate the need for specifying $\mathcal{T}$ by simulating the dynamics backwards and forwards in time with respect to the current state and considering the U-turn condition,

$$\langle x^+ - x^-, M^{-1}p^- \rangle < 0 \quad \text{or} \quad \langle x^+ - x^-, M^{-1}p^+ \rangle < 0, \tag{5.1.2}$$

where (x^-, p^-) and (x^+, p^+) are the trajectory endpoints and $\langle \cdot, \cdot \rangle$ denotes the inner product. This criterion is used to terminate the dynamics if the endpoints are pointing towards each other. The motivation is that any further integration would likely be wasteful.

At first, a single step is taken from z_0. Then a direction is picked, forward or backward with equal probability, and two steps are taken from the respective endpoints with positive or negative ε. The procedure follows, such that at the j^{th} iteration 2^{j-1} additional points are added onto the trajectory. The procedure terminates once the condition (5.1.2) is met, at which point the last iteration is discarded. The algorithm aims to create a set of points $\mathcal{E}$ which could be constructed from any $z \in \mathcal{E}$ with the same probability. So if a U-turn is observed at iteration k then $\mathcal{E}$ can only be composed of all the points up to iteration $k-1$ and $\mathbb{P}(\mathcal{E}|z) \propto (1/2)^j,\ z \in \mathcal{E}$.

Given the set $\mathcal{E}$, a subset $\mathcal{C}$ is chosen so as to contain all states $z \in \mathcal{E}$ to which the chain can transition without violating detailed balance with respect to $\tilde{\pi}$. This is done using a slice sampling procedure where an auxiliary independent random variable $\upsilon \sim \mathsf{Unif}(0, 1)$ is introduced to define a "slice" (Neal, 2003). At each iteration, a new

realisation of U is drawn and $\mathcal{C}$ is formed by discarding all z^* where $\mathcal{H}(z) - \mathcal{H}(z^*) < \log v$ (which is equivalent to $\tilde{\pi}(z^*) < u\tilde{\pi}(z)$), that is,

$$\mathcal{C} = \{z^* \in \mathcal{E} \ : \ \mathcal{H}(z) - \mathcal{H}(z^*) \geq \log v\}.$$

Sampling z' uniformly from $\mathcal{C}$ is a valid Gibbs step which leaves the uniform distribution on $\mathcal{C}$ invariant. This combined with the resampling of u results in a Gibbs sampler which targets a mass function $\propto \tilde{\pi}(z^*)$ on $\mathcal{E}$.

The NUTS algorithm is implemented recursively where the double-backing procedure builds a balanced binary tree and the new z' are sampled sequentially to remove the need for storing all of $\mathcal{C}$. The average depth of the tree is very problem-dependent and so the recursive procedure can be costly depending on how much stack memory is assigned to the inference software when compiling. Furthermore, no parts of the algorithm can be parallelised as (5.1.2) requires the endpoints of the path at all times. The last iteration of the double-backing procedure is discarded meaning that nearly half the computational effort of the MCMC step is spent on verifying the U-turn condition. Additionally, it is possible the optimal states which maximise the squared jumping distance from z_0 are included in that last iteration right before being discarded.

5.1.2 Other algorithms

Betancourt (2016) suggests modification to NUTS where the new state is obtained using multinomial sampling with a bias towards the second half of the steps in the trajectory set $\mathcal{E}$ instead of slice sampling; this is used in the current version of `Stan` (Stan Development Team, 2020).

Wu et al. (2019) addresses the issue of wasteful gradient evaluations by outlining a procedure used to learn a suitable empirical distribution of the number of leapfrog steps, L, during the warm-up run of the HMC algorithm. At each iteration of the warm-up, empirical HMC, as the authors call it, runs the standard leapfrog integrator for L_0 steps and identifies the minimum number of steps L needed to satisfy the U-turn condition, which could require additional leapfrog steps beyond L_0. The minimum L is referred to as the *longest batch* and is stored to produce a marginal distribution of the global leapfrog scale. This empirical distribution is then used to draw suitable values of L during the main run of the algorithm. In the numerical experiments, the procedure is allocated 10% of MCMC iterations after tuning ε.

Wang et al. (2013) introduces an adaptive scheme that uses Bayesian optimisation to find ε and L $(= \lceil \mathcal{T}/\varepsilon \rceil)$ which seeks to maximise the cost-adjusted expected squared jumping distance $\mathbb{E}\left[\|X' - X\|^2\right]/\sqrt{L}$. The acquisition function used in the optimisation was a variant of the Upper Confidence Bound.

Hoffman et al. (2021) introduces a gradient-based adaptive scheme that utilises multiple parallel chains of HMC to tune for both ε and $\mathcal{T}$ at the same time. The tuning of $\mathcal{T}$ is done using the "Change in the Estimator of the Expected Square" (ChEES) statistic

$$\text{ChEES} = \frac{1}{4}\mathbb{E}\left[\left(\|X' - \mu\|^2 - \|X - \mu\|^2\right)^2\right],$$

where $X \sim \pi,\ P \sim \mathsf{N}(0, M),\ (X', P') = \mathsf{LeapFrog}((X, P); \varepsilon L)$ and $\mu = \mathbb{E}[X]$. The authors note that the criterion (i) is invariant to shifts and rotations, (ii) places greater weight on exploring large-variance directions, and (iii) is focused on second-moment

estimation, which is something NUTS can struggle with. The full algorithm maximises ChEES using estimates of its gradient with respect to $\mathcal{T}$ with the scale of gradient steps being adjusted by ADAM (adaptive moment estimation, Kingma and Ba, 2015). Based on the presented results, the scheme was substatially more efficient than NUTS and, in most cases, outperformed adaptive tuning with respect to standard expected squared jumping distance (ESJD); however, it was still outperformed by HMC with its trajectory length chosen via a grid search.

This work introduces a competitive variant of the HMC algorithm where the trajectory length termination is directly related to the density of interest π.

5.1.3 Apogee definition

As the particle traverses the potential surface U, there will come a point where it changes direction from "uphill" to "downhill" with respect to the potential. The point of this change in direction is called an *apogee* and is defined as the point at which $\langle M^{-1}p, \nabla U(x)\rangle$ changes sign from positive to negative. We incorporate this apogee condition into the leapfrog integrator by checking for

$$\langle M^{-1}p_0, \nabla U(x_0)\rangle > 0 \quad \text{and} \quad \langle M^{-1}p_1, \nabla U(x_1)\rangle < 0. \tag{5.1.3}$$

Substituting in (5.1.1), we find that

$$\langle M^{-1}p, \nabla U(x)\rangle = \left\langle \frac{\mathsf{d}x}{\mathsf{d}t}, \nabla U(x) \right\rangle = \frac{\mathsf{d}U}{\mathsf{d}t},$$

which means that (5.1.3) indicates a local maximum in the trajectory along the potential surface U. Figure 5.1.3 shows how the condition (5.1.3) can be used to split

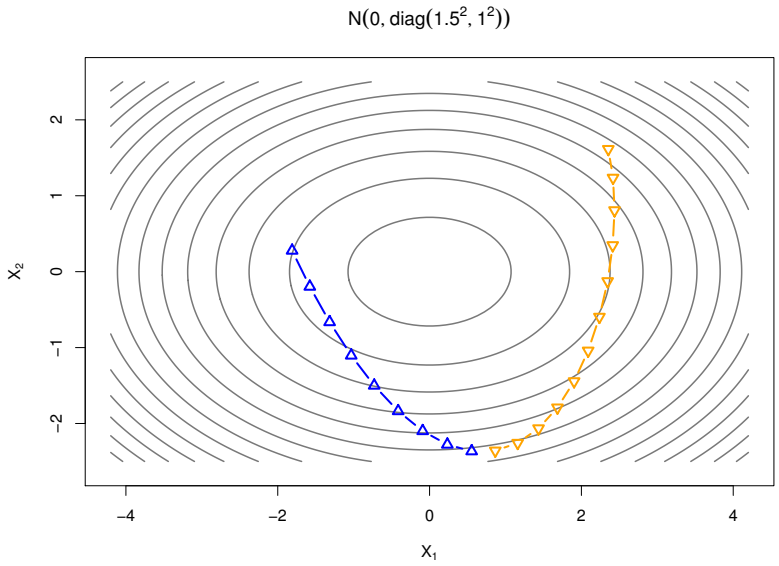


Figure 5.1.3: Approximate Hamiltonian dynamics on a Gaussian target. States before and after the apogee are indicated by "△" and "▽" respectively.

the trajectory. It is important to note that the condition indicates the existence of an apogee between states z_0 and z_1, as opposed to locating the apogee itself.

In Section 5.2, we outline how apogees can be used to define endpoints of HMC trajectories, thus eliminating the need to specify an arbitrary numeric value for the integration time. A sampling algorithm based on this criterion is introduced and we show how it satisfies detailed balance with respect to $\tilde{\pi}$. Section 5.3 provides general tuning guidelines for the algorithm. Section 5.4 examines the behaviour of the algorithm on a multivariate Gaussian target, providing theoretical results for the number of apogees passed along a Hamiltonian trajectory for a given integration time. Section 5.5 contains a numerical experiments comparing the proposed algorithm to HMC and NUTS, and Section 5.6 provides a discussion along with suggestions for future work.

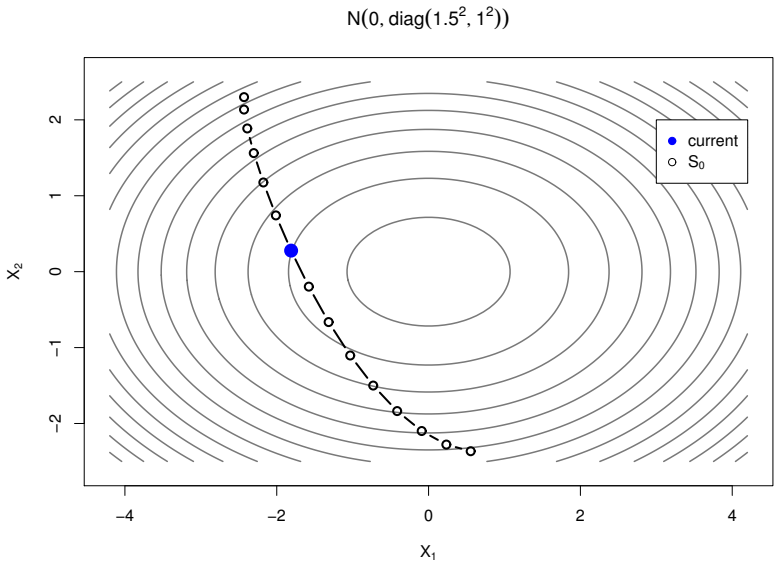


Figure 5.2.1: Segment $\mathcal{S}_0$ constructed from z_0 indicated by the blue dot.

5.2 AAPS

In this section, we introduce the Apogee to Apogee Path Sampler (AAPS) for producing samples from the extended target $\tilde{\pi}$. We begin by first outlining a simple variant of the algorithm, AAPS_0, and extending it to a much more efficient version, AAPS_K.

5.2.1 AAPS_0

With the condition (5.1.3), we construct an *apogee to apogee path* by simulating the Hamiltonian dynamics forward and backwards in time from the state z_0 until an apogee is detected at either end. This is given in Algorithm 5.2; the algorithm includes an additional stability condition discussed later in Section 5.2.4. We define the set of states produced in this way as a *segment*.

5.2 Apogee-to-apogee segment construction:

1. **Input:** target distribution π; initial point $z_0 = (x_0, p_0)$; step-size ε; maximum Hamiltonian

error Δ

2. **Output:** apogee-to-apogee segment $\mathcal{S}_0$
3. **Initialise:** sample momentum $p_0 \sim \mathsf{N}(0, M)$; initialise segment $\mathcal{S}_0 = z_0$; calculate $D_0 = \langle p_0, \nabla U(x_0) \rangle$; set $z = z_0,\ D = D_0,\ \texttt{no\_apogee} = 1$; $H_{\min} = \mathcal{H}(z_0)$; $H_{\max} = \mathcal{H}(z_0)$
4. **while** `no_apogee` **do**
   (a) $z_{\text{new}} = \mathsf{LeapFrog}(z; \varepsilon)$; $H_{\min} \leftarrow \mathsf{min}\{\mathcal{H}(z_{\text{new}}), H_{\min}\}$; $H_{\max} \leftarrow \mathsf{max}\{\mathcal{H}(z_{\text{new}}), H_{\max}\}$;
   (b) **if** $H_{\max} - H_{\min} < \Delta$ **do**
      i. $D_{\text{new}} = \langle p_{\text{new}}, \nabla U(x_{\text{new}}) \rangle$;
      ii. **if** $D > 0$ & $D_{\text{new}} < 0$ **do**
         A. `no_apogee` $\leftarrow 0$;
      iii. **else do**
         A. $\mathcal{S}_0 \leftarrow \mathcal{S}_0 \cup z_{\text{new}}^-$;
         B. $z \leftarrow z_{\text{new}}$, $D \leftarrow D_{\text{new}}$;
   (c) **else**
      i. **return** $\mathcal{S}_0 = z_0$;
5. $z \leftarrow z_0^-$, $D \leftarrow -D_0$, `no_apogee` $\leftarrow 1$;
6. Repeat step 4;
7. **return** $\mathcal{S}_0$;

Letting $\mathcal{S}_0(z_0)$ denote the segment produced from the point z_0, it is clear to see from Figure 5.2.1 that for any $z \in \mathcal{S}_0(z_0)$, $\mathcal{S}_0(z) = \mathcal{S}_0(z_0) =: \mathcal{S}_0$; the probability of constructing the segment is $\mathbb{P}(\mathcal{S}_0|z) = \mathbb{I}_{\{z \in \mathcal{S}_0\}}$. Using the segment, we then construct a proposal mechanism for a new point z' which leaves the target distribution $\tilde{\pi}$ invariant. Here, it is helpful to define $\mathcal{S}_0^-$ which is made up of the same states as $\mathcal{S}_0$ but with negated momenta, that is, $\mathcal{S}_0^- = \{z^- : z \in \mathcal{S}_0\}$. Given the segment $\mathcal{S}_0$, we can propose a point from $\mathcal{S}_0^-$ by weighting all its elements by some function $\omega : \mathbb{R}^{4d} \to \mathbb{R}^+$. The chosen weight is invariant to sign change of the momenta and can depend on z_0. A

new state is proposed by sampling from $\mathcal{S}_0^-$ with probability $\propto \omega(z'; z_0)$. The full transition function from z_0 to z' then becomes

$$\begin{aligned} q(z'|z_0) &= \mathbb{P}\left(\mathcal{S}_0|z_0\right)\mathbb{P}\left(z'|z_0, \mathcal{S}_0, \omega\right) \\ &= \frac{\omega(z'; z_0)}{\sum_{z\in\mathcal{S}_0^-}\omega(z; z_0)} \cdot \mathbb{I}_{\{z_0\in\mathcal{S}_0\}} \cdot \mathbb{I}_{\left\{z'\in\mathcal{S}_0^-\right\}}. \end{aligned}$$

Algorithm 5.3 outlines a full MCMC update which we call AAPS_0. The validity of the algorithm is shown in Section 5.2.3.

5.3 AAPS_0**:**

1. **Input:** target distribution π; position $x_0 \sim \pi$; step-size ε; weight function $\omega(\cdot;\cdot)$
2. **Output:** new position $x \sim \pi$
3. **Initialise:** sample momentum $p \sim \mathsf{N}(0, M)$ and set $z_0 = (x_0, p_0)$;
4. Construct path $\mathcal{S}_0$ from z_0 (Algorithm 5.2), flip state momenta to obtain $\mathcal{S}_0^-$ and weight each element using ω;
5. Sample $z \in \mathcal{S}_0^-$ w.p. $\propto \omega(z'; z_0)$;
6. With probability
$$\alpha(z_0, z') = 1 \wedge \frac{\tilde{\pi}(z')\omega(z_0; z')\sum_{z''\in\mathcal{S}_0^-}\omega(z''; z_0)}{\tilde{\pi}(z)\omega(z'; z_0)\sum_{z''\in\mathcal{S}_0}\omega(z''; z')}$$
set $z \leftarrow z'$, else set $z \leftarrow z_0$;
7. **return** x;

5.2.2 AAPS_K

For many distributions, a single-segment path might not be enough to cover a substantial distance from the initial position x_0. We can overcome this potential issue by constructing a path made up of multiple segments, which is the basis for AAPS_K. One way to create a longer path is by positioning $K + 1$ contiguous segments such

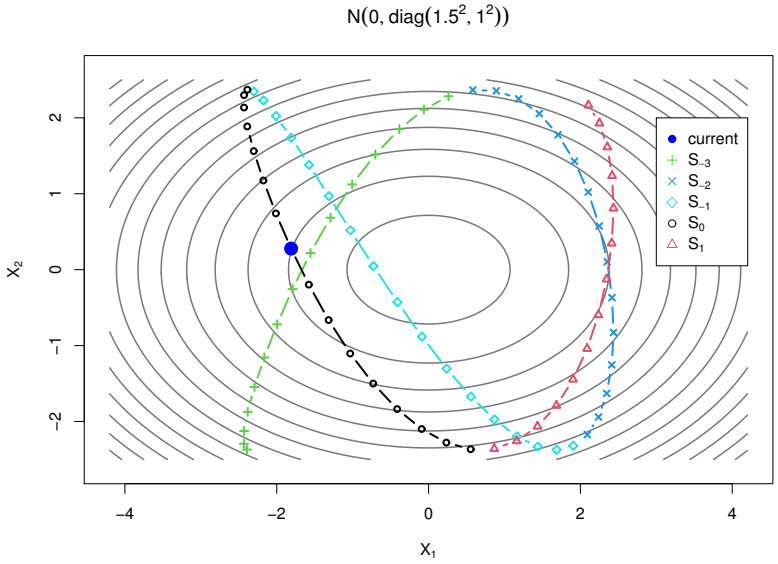


Figure 5.2.2: Path $\mathcal{P} = \mathcal{S}_{-3:1}$ constructed using four apogee-to-apogee segments.

that $\mathcal{S}_0$ is equally likely to be the 1^{st} , 2^{nd} , $\ldots$, $(K+1)^{\text{th}}$ segment. We denote the concatenation of the segments by $\mathcal{S}_{a:b}(z_0)$, where $b-a = K$ and $a \leq 0 \leq b$. The path $\mathcal{S}_{a:b}(z_0)$ can be written as $\mathcal{S}_{a^*:b^*}(z^*)$ (where $b^* - a^* = K$ and $a^* \leq 0 \leq b^*$) for any $z^* \in \mathcal{S}_{a:b}(z_0)$ and as a result

$$\mathbb{P}\left(\mathcal{S}_{a^*:b^*}(z^*)|z^*, K\right) = \frac{1}{K+1}\mathbb{I}_{\{z^* \in \mathcal{S}_{a^*:b^*}\}}.$$

To simplify the notation we let $\mathcal{P} = \mathcal{S}_{a:b}$ and $\mathcal{P}^- = \mathcal{S}^-_{a:b}$. With the same setup as before where a new state is proposed according to some weighting function ω, the acceptance probability is

$$\begin{aligned}
\alpha(z_0, z') &= 1 \wedge \frac{\tilde{\pi}(z')q(z_0|z')}{\tilde{\pi}(z_0)q(z'|z_0)} \\
&= 1 \wedge \frac{\tilde{\pi}(z')\mathbb{P}\left(\mathcal{P}^-(z')|z', K\right)\mathbb{P}\left(z_0|z', \omega, \mathcal{P}^-\right)}{\tilde{\pi}(z_0)\mathbb{P}\left(\mathcal{P}(z_0)|z_0, K\right)\mathbb{P}\left(z'|z_0, \omega, \mathcal{P}\right)} \\
&= 1 \wedge \frac{\tilde{\pi}(z')\omega(z_0; z')\sum_{z^* \in \mathcal{P}^-}\omega(z^*; z_0)}{\tilde{\pi}(z)\omega(z'; z_0)\sum_{z^* \in \mathcal{P}}\omega(z^*; z')},
\end{aligned}$$

where probability of constructing $\mathcal{P}$ cancels out in the MH ratio. The path-building procedure is described in Algorithm 5.4, and Algorithm 5.5 outlines the full apogee to apogee path sampler AAPS_K. Setting $K = 0$ results in AAPS_0.

5.4 Apogee-to-apogee path construction:

1. **Input:** target distribution π; initial point $z_0 = (x_0, p_0)$; step-size ε; number of internal apogees K;maximum Hamiltonian error Δ
2. **Output:** apogee-to-apogee path $\mathcal{P}$
3. **Initialise:** sample momentum $p_0 \sim \mathsf{N}(0, M)$; sample $b \sim \mathsf{Unif}(\{0, 1, \ldots, K\})$ and set $a = b - K$; initialise path $\mathcal{P} = z_0$; calculate $D_0 = \langle p_0, \nabla U(x_0)\rangle$; set $z = z_0$, $D = D_0$, `apogees_left` $= b + 1$; $H_{\min} = \mathcal{H}(z_0)$; $H_{\max} = \mathcal{H}(z_0)$
4. **while** `apogees_left` > 0 **do**
   (a) $z_{\text{new}} = \mathsf{LeapFrog}(z; \varepsilon)$;$H_{\min} \leftarrow \mathsf{min}\{\mathcal{H}(z_{\text{new}}), H_{\min}\}$; $H_{\max} \leftarrow \mathsf{max}\{\mathcal{H}(z_{\text{new}}), H_{\max}\}$;
   (b) **if** $H_{\max} - H_{\min} < \Delta$ **do**
      i. $D_{\text{new}} = \langle p_{\text{new}}, \nabla U(x_{\text{new}})\rangle$;
      ii. **if** $D > 0$ & $D_{\text{new}} < 0$ **do**
         A. `apogees_left` $\leftarrow$ `apogees_left` $- 1$;
      iii. **if** `apogees_left` > 0 **do**
         A. $\mathcal{P} \leftarrow \mathcal{P} \cup z_{\text{new}}^{-}$;
         B. $z \leftarrow z_{\text{new}}$, $D \leftarrow D_{\text{new}}$;
   (c) **else**
      i. **return** $\mathcal{P} = z_0$;
5. $z \leftarrow z_0^{-}$, $D \leftarrow -D_0$, `apogees_left` $\leftarrow |a| + 1$;
6. Repeat step 4;
7. **return** $\mathcal{P}$;

5.5 AAPS_K:

1. **Input:** target distribution π; position $x_0 \sim \pi$; step-size ε; weight function $\omega(\cdot;\cdot)$
2. **Output:** new position $x \sim \pi$
3. **Initialise:** sample momentum $p \sim \mathsf{N}(0, M)$ and set $z_0 = (x_0, p_0)$;
4. Construct path $\mathcal{P}$ from z_0 (Algorithm 5.4), flip state momenta to obtain $\mathcal{P}^{-}$ and weight each element using ω;

5. Sample $z' \in \mathcal{P}$ w.p. $\propto \omega(z'; z_0)$;
6. With probability
$$\alpha(z_0, z') = 1 \wedge \frac{\tilde{\pi}(z')\omega(z_0; z') \sum_{z'' \in \mathcal{P}^-} \omega(z''; z_0)}{\tilde{\pi}(z)\omega(z'; z_0) \sum_{z'' \in \mathcal{P}} \omega(z''; z')}$$
set $z \leftarrow z'$, else set $z \leftarrow z_0$;
7. **return** x;

5.2.3 Validity of AAPS

Here, we show that AAPS_K satisfies detailed balance with respect to $\tilde{\pi}$ and so the marginal of x will target the desired distribution π.

Proposition 5.2.1. *The* AAPS_K *algorithm satisfies detailed balance with respect to the canonical distribution* $\tilde{\pi}$*.*

Proof. The momentum refreshment step samples directly from $\mathsf{N}(0, M)$ which is the marginal distribution of p, satisfying detailed balance with respect to it.

At stationarity, we have

$$\tilde{\pi}(z_0) \times \mathbb{P}\left(\mathcal{P}|z_0\right) \times \mathbb{P}\left(\text{propose } z'|\mathcal{P}, z_0\right) \times \mathbb{P}\left(\text{accept } z'|\mathcal{P}, z_0, \text{proposed } z'\right)$$

which is

$$\tilde{\pi}(z_0) \times \frac{\mathbb{I}_{\{z_0 \in \mathcal{P}\}}}{K+1} \times \frac{\omega(z'; z_0)\mathbb{I}_{\{z' \in \mathcal{P}^-\}}}{\sum_{z^* \in \mathcal{P}^-} \omega(z^*; z_0)} \times 1 \wedge \frac{\tilde{\pi}(z')\omega(z_0; z') \sum_{z^* \in \mathcal{P}^-} \omega(z^*; z_0)}{\tilde{\pi}(z)\omega(z'; z_0) \sum_{z^* \in \mathcal{P}} \omega(z^*; z')},$$

and this in turn equates to

$$\frac{\mathbb{I}_{\{z_0 \in \mathcal{P}\}} \cdot \mathbb{I}_{\{z' \in \mathcal{P}^-\}}}{K+1} \times \left\{ \frac{\tilde{\pi}(z_0)\omega(z_0; z')}{\sum_{z^* \in \mathcal{P}} \omega(z^*; z')} \wedge \frac{\tilde{\pi}(z')\omega(z'; z_0)}{\sum_{z^* \in \mathcal{P}^-} \omega(z^*; z_0)} \right\}.$$

The final expression is invariant to $(z_0, \mathcal{P}) \leftrightarrow (z', \mathcal{P}^-)$. □

5.2.4 Stability of path construction

The trajectory may enter parts of the state-space Z where the numerical solution to (5.1.1) is unstable with the error of $\mathcal{H}$ increasing without bound. There are two main consequences of this: (i) when building the path $\mathcal{P}$ apogees may be incorrectly omitted, taking much longer to terminate the dynamics, and (ii) during the accept-reject step, ideal proposals are more likely to be rejected, thus wasting the effort taken to build the path. The NUTS algorithm suffers from the same problem and so we introduce a similar stability condition to that in Hoffman and Gelman (2014),

$$\mathsf{max}_{z \in \mathcal{P}} \mathcal{H}(z) - \mathsf{min}_{z' \in \mathcal{P}} \mathcal{H}(z') < \Delta, \tag{5.2.1}$$

where $\Delta > 0$ is a maximum energy discrepancy which prespecified by the user. If at any point the range of $\mathcal{H}$ along $\mathcal{P}$ exceeds Δ, the path is terminated returning z_0. It is straightforward to see that (5.2.1) is invariant to the starting position on the path. For the experiments in Section 5.5, we found that even a large value $\Delta = 1000$ is sufficient. Algorithms 5.2 and 5.4 explicitly check for the stability condition (5.2.1).

5.2.5 AAPS on bimodal targets

In $d = 1$, it is straightforward to see that AAPS_0 is reducible on multimodal targets; provided a stable ε is used, the dynamics will never cross the cusp between two modes. However, even in $d = 2$ this is no longer the case. Figure 5.2.3 shows Hamiltonian

dynamics on two bimodal target distribution

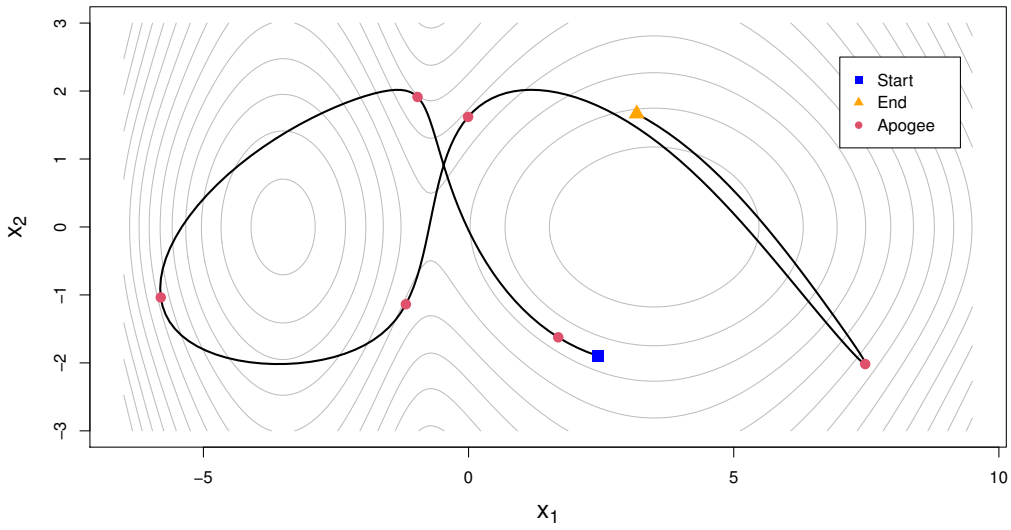
$$X \sim \frac{1}{2}\mathsf{N}\left((-a,0)^\top, I_d\right) + \frac{1}{2}\mathsf{N}\left((a,0)^\top, \sigma^2 I_d\right), \tag{5.2.2}$$

where $a = 3.5$, $\sigma = 2$ and $d = 2, 40$ (rotated variants of those used in Pompe et al., 2020). In both examples, the same initial positions and momenta were used for $(z_1, z_2)^\top$, with the remaining components in the $d = 40$ case being initialised from the second mixing distribution. The step-size ε was chosen sufficiently small so the numerically integrated path closely approximates the true solution to (5.1.1). From the figures it is clear that, even in two dimensions, a single apogee-to-apogee segment permits movement between the modes. The mode-hopping becomes more common as the dimension increases since the first component's relative contribution to the dot-product is smaller; the apogee for the whole trajectory may not be anywhere close to the saddle point. In Section 5.5.2, we compare AAPS, HMC and NUTS on a selection of bimodal targets of the form (5.2.2).

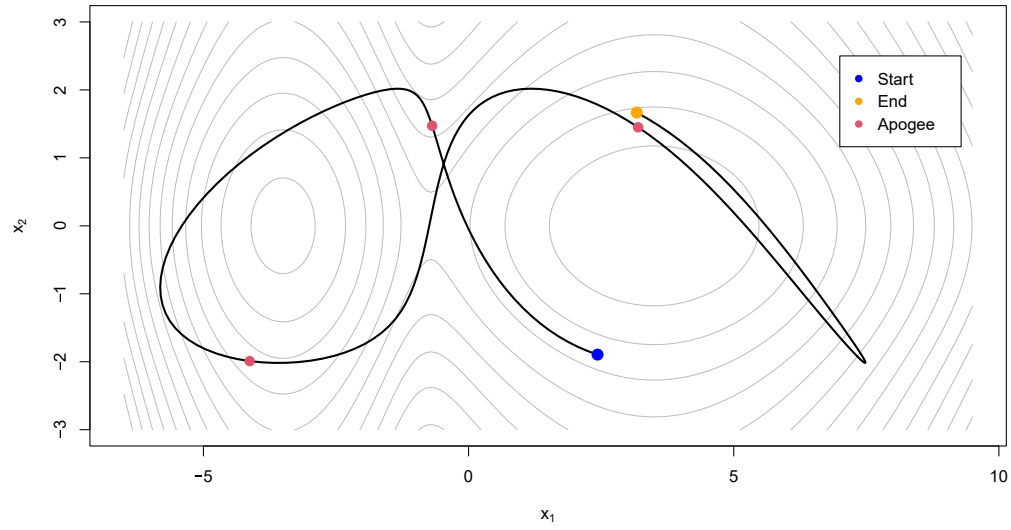
5.2.6 Sampling from the path

Because of the general formulation of AAPS, there are multiple valid ways of sampling points from the path $\mathcal{P}^-$. Below we outline choices for the weighting function $\omega(\cdot; z)$ considered in this work along with the respective acceptance probabilities $\alpha(z_0, z')$. All the weights are invariant to the signs of momenta p so, for notational clarity, we drop the superscript and simply write $\mathcal{P}$.

1. $\omega(z'; z) = \tilde{\pi}(z')$ and $\alpha(z_0, z') = 1$



(a) $d = 2$



(b) $d = 40$

Figure 5.2.3: Hamiltonian dynamics on bimodal targets with $d = 2$ (Top) and $d = 40$ (Bottom). Initial positions and momenta of z_1 and z_2 are fixed in both examples, with the remaining components of the 40-dimensional example initialised from the stationary distribution. The bottom panel shows the projection of the trajectory onto the first two dimensions.

2. $\omega(z';z) = \|x' - x\|^2$ and $\alpha(z_0, z') = \frac{\tilde{\pi}(z')\sum_{z\in\mathcal{P}}\|x-x_0\|^2}{\tilde{\pi}(z_0)\sum_{z\in\mathcal{P}}\|x-x'\|^2}$

3. $\omega(z';z) = \tilde{\pi}(z')\|x' - x\|^2$ and $\alpha(z_0, z') = \frac{\sum_{z\in\mathcal{P}}\tilde{\pi}(z)\|x-x_0\|^2}{\sum_{z\in\mathcal{P}}\tilde{\pi}(z)\|x-x'\|^2}$

4. $\omega(z';z) = \|x' - x\|$ and $\alpha(z_0, z') = \frac{\tilde{\pi}(z')\sum_{z\in\mathcal{P}}\|x-x_0\|}{\tilde{\pi}(z_0)\sum_{z\in\mathcal{P}}\|x-x'\|}$

5. $\omega(z';z) = \tilde{\pi}(z')\|x' - x\|$ and $\alpha(z_0, z') = \frac{\sum_{z\in\mathcal{P}}\tilde{\pi}(z)\|x-x_0\|^2}{\sum_{z\in\mathcal{P}}\tilde{\pi}(z)\|x-x'\|^2}$

6. $\omega(z';z) = \tilde{\pi}(z')\mathbb{I}_{\{z'\in H(z)\}}$, where $H(z)$ is defined in Section 5.2.6; this also results in $\alpha(z_0, z') = 1$

We wish to construct an algorithm that proposes distant proposal states which are then accepted with a high probability. By using weights that in some way incorporate the distance from the initial state z_0, we skew the sampling towards the elements of $\mathcal{P}$ which maximise the jumping distance, and this in turn decreases the autocorrelation of the chain. Weight 6 is a special case where the state is proposed from the "other half" of the path with respect to the cumulative sums of $\tilde{\pi}$ from one endpoint to the other.

The weights modulated by $\tilde{\pi}$ have the property of the ratio $\tilde{\pi}(z')/\tilde{\pi}(z)$ not appearing in the acceptance probability. Intuitively, we would expect this to result in higher acceptance rates even for large for stepsizes ε. Typically, the calculation of the sum $\sum_{z\in\mathcal{P}}\omega(z;z')$ for a lot of choices of ω requires the storage of the entire path which is made up of an unknown (and possibly large) number of variables. However, sampling according to weights 1, 2 and 3 can be done sequentially at a fixed memory cost. A numerical comparison of the sampling schemes is given a the end of this section.

Sequential sampling at fixed memory cost

This is a general method of sampling a proposal state Z' from the path

$$\mathcal{P} = \{z_{-B}, \ldots, z_{-1}, z_0, z_1, \ldots, z_F\},$$

where $-B$ and F are the indices of the endpoints in the "backwards" and "forward" parts of the trajectory respectively; the state z_0 is arbitrarily assigned to the front. The aim is to pick a proposal point without needing to store the entire path. If each point z_i is to be picked with probability proportional to $w_i = \omega(z'; z_0)$, then we do the following:

1. **Initialise:** Set $Z' = z_0$ and $Z^* = z_{-1}$;
2. **For** $k = 1, \ldots, F$**:** Set $Z' = z_k$ with probability $\dfrac{w_k}{\sum_{i=0}^{k} w_i}$;
3. **For** $k = -2, \ldots, -B$**:** Set $Z^* = z_k$ with probability $\dfrac{w_k}{\sum_{i=k}^{-1} w_i}$;
4. Set $Z' = Z_*$ with probability $\frac{\sum_{i=0}^{F} w_i}{\sum_{j=0}^{F} w_j + \sum_{k=-B}^{-1} w_k}$;
5. **Return** Z';

Proposition 5.2.2.

$$\mathbb{P}\left(Z' = z_k\right) = \frac{w_k}{\sum_{i=-B}^{F} w_i}.$$

Proof. Suppose $k \geq 0$. The probability of picking z_k is independent of the previously

picked or rejected points z_i, $i < k$. Letting $z_{a:b} = \{z_a, z_{a+1}, \ldots, z_{b-1}, z_b\}$, we have

$$\begin{aligned}
\mathbb{P}(Z' = z_k) &= \mathbb{P}((z_k \text{ picked}) \cap (z_l,\, l \neq k \underline{\text{ not}} \text{ picked})) \\
&= \mathbb{P}(Z' \in z_{0:F})\, \mathbb{P}\left(Z' \notin z_{0:(k-1)} \cup z_{(k+1):F} | Z' \in z_{0:F}\right) \\
&= \frac{\sum_{i=0}^{F} w_i}{\sum_{j=-B}^{F} w_j} \left[\frac{w_k}{\sum_{i=0}^{k} w_i} \prod_{l=k+1}^{F} \left(1 - \frac{w_l}{\sum_{i=0}^{l} w_i}\right)\right] \\
&= \frac{\sum_{i=0}^{F} w_i}{\sum_{j=-B}^{F} w_j} \left[\frac{w_k}{\sum_{i=0}^{k} w_i} \prod_{l=k+1}^{F} \frac{\sum_{j=0}^{l-1} w_j}{\sum_{i=0}^{l} w_i}\right] \\
&= \frac{\sum_{i=0}^{F} w_i}{\sum_{j=-B}^{F} w_j} \cdot \frac{w_k}{\sum_{i=0}^{F} w_i} \\
&= \frac{w_k}{\sum_{i=-B}^{F} w_i}.
\end{aligned}$$

Due to the symmetry of the problem, a similar argument is used to prove the case for $k < 0$.

□

Steps 2 and 3 are fully independent of each other and so they could potentially be carried out in parallel reducing the overall computational burden.

We now require a way of computing the acceptance probability. Any choice where the weight is only a function of the state on the path is suitable, $w_i = \omega(z_i; z) = f(z_i)$; for example, $w_i = \tilde{\pi}(z_i)$ (the canonical density). This holds as the same weights are used when computing the denominators for the Metropolis-Hastings ratio. However, some desirable choices such as weights proportional to a metric computed from the state z_0, for example, l_1 or l_2, are unsuitable; given a proposal z' we need to compute the sum of the distances to all other points on the path to obtain the denominator which defeats the purpose of the fixed-memory approach.

Another suitable choice for the weight is a function squared jumping distance of some function $g(z)$, $w_i \propto \|g(z_i) - g(z_0)\|_2^2$. In the simplest form, if $w_i = \|x_i - x_0\|^2$ ($g(x) \equiv x$) and x is 1-dimensional then

$$\sum_{i=-B}^{F} \|x_i - x_0\|_2^2 = \left(\sum_{i=-B}^{F} x_i^2\right) - 2x_0 \left(\sum_{i=-B}^{F} x_i\right) - x_0^2(F+B),$$

meaning that we only need to store the first two empirical moments of the positions along $\mathcal{P}$. Similarly, if $w_i = \tilde{\pi}(z_i)\|x_i - x_0\|^2$ then

$$\sum_{i=-B}^{F} \tilde{\pi}(z_i)\|x_i - x_0\|_2^2 = \left(\sum_{i=-B}^{F} \tilde{\pi}(z_i)x_i^2\right) - 2x_0 \left(\sum_{i=-B}^{F} \tilde{\pi}(z_i)x_i\right) - x_0^2 \sum_{i=-B}^{F} \tilde{\pi}(z_i).$$

Extending this to a d-dimensional space, we only need to store at most $3 \times d$ additional quantities.

Sampling Scheme 6 – Sampling from the opposite half of the path

For clarity, we relabel the points on the path $\mathcal{P} = \{z_{-B}, \ldots, z_F\}$ to be $\{z_1, z_2, ..., z_n\}$, where $n = B + F + 1$. We let $S_k = \sum_{i=1}^{k} \tilde{\pi}(z_i)$, $k = 1, \ldots, n$, and $h = \mathsf{min}\{k : S_k \geq \frac{1}{2}S_n\}$. The index h is the point where the cumulative sums exceed half of the total sum. We define $H_B = \{z_i : i < h\}$ and $H_F = \{z_i : i > h\}$ which are the sets of point "behind" and "in front of" z_h. The proposal z' is obtained for three possible cases:

Case 1, $z_0 \in H_B$: Sample $z' \in H_F \cup z_h$

$$\mathbb{P}(z' = z_i) = \begin{cases} \frac{\tilde{\pi}(z_i)}{S_n/2} = \frac{2\tilde{\pi}(z_i)}{S_n} & z_i \in H_F \\ 1 - \frac{S_n - S_h}{S_n/2} = \frac{2S_h - S_n}{S_n} & z_i = z_h \end{cases}.$$

Case 2, $z_0 \in H_F$: Sample $z' \in H_B \cup z_h$

$$\mathbb{P}(z' = z_i) = \begin{cases} \frac{\tilde{\pi}(z_i)}{S_n/2} = \frac{2\tilde{\pi}(z_i)}{S_n} & z_i \in H_B \\ 1 - \frac{S_{h-1}}{S_h/2} = \frac{S_n - 2S_h + 2\tilde{\pi}(z_h)}{S_n} & z_i = z_h \end{cases}.$$

Case 3, $z_0 = z_h$: Assign z_0 to H_B or H_F with the respective probabilities being

$$\mathbb{P}(z_0 \to H_B) = \frac{S_n - 2S_h + 2\tilde{\pi}(z_h)}{2\tilde{\pi}(z_h)}, \tag{5.2.3}$$

$$\mathbb{P}(z_0 \to H_F) = \frac{2S_h - S_n}{2\tilde{\pi}(z_h)}, \tag{5.2.4}$$

and sample from the other half according to the rules above. These specific assignment probabilities ensure that AAPS using Scheme 6 preserves the canonical distribution $\tilde{\pi}$ without the need for an accept-reject step.

Proposition 5.2.3. *Assignment probabilities* (5.2.3) *and* (5.2.4) *lead to a transition kernel which satisfies detailed balance with respect to* $\tilde{\pi}$.

Proof. Suppose $z_0 \in H_B$ and we propose $z' = z_h$. To satisfy detailed balance we require that $\tilde{\pi}(z_0)q(z_h|z_0) = \tilde{\pi}(z_h)q(z_0|z_h)$. Factoring out the indicator functions from the transition terms, we have

$$\begin{aligned} \tilde{\pi}(z_0)\left(\frac{2S_h - S_n}{S_n}\right) &= \tilde{\pi}(z_h)\mathbb{P}(z_0 \to H_F)\left(\frac{2\tilde{\pi}(z_0)}{S_n}\right) \\ \Rightarrow \mathbb{P}(z_0 \to H_F) &= \frac{2S_h - S_n}{2\tilde{\pi}(z_h)}. \end{aligned}$$

Similar steps follow for when $z_0 \in H_F$ and the resulting probabilities add up to 1. □

Sampling weights comparison

We compare the different weighting schemes when used within AAPS with respect to two quantities: the efficiency, defined as the component-wise minimum effective sam-

ple size divided by the total number of leapfrog steps taken; and the true acceptance rate, that is, the emprically estimated probability of moving from z_0 during an iteration of the algorithm. Figure 5.2.4 shows the results of the comparison on a particular target being π_G^H with $d = 40,\ \xi = 20$; see Section 5.5.1 for details. The step-size ε was chosen to be 1.0 which was smaller than the optimum of 1.4; this was to minimise the chance of skipping apogees which could affect the results. Modulating the weights by $\tilde{\pi}(z)$ resulted in higher acceptance rates, with a lower sensitivity to the choice of ε, compared to the unmodulated variants and led to an overall increase in the efficiency. The relative performance of the sampling schemes persisted for other choices of fixed ε and K, and experiments using different targets. Scheme 6 which forces proposals from the second half of the cumulative sum of $\tilde{\pi}(z)$ performed the best, with Schemes 3 and 5 reasonably close to it; both of those schemes incorporated some notion of jumping distance modulated by $\tilde{\pi}(z)$. However, Scheme 6 required the storage of the entire path $\mathcal{P}$ to propose z' at each iteration, and, similarly, Schemes 4 and 5 needed the whole path stored for the calculation of the acceptance probabilities α. The $\mathcal{O}(K)$ memory overhead could make these unsuitable for a general-purpose inference tool. On the other hand, Schemes 1, 2 and 3 can be fully implemented with a fixed $\mathcal{O}(1)$ memory cost. In practice, we found this to improve the efficiency as measured by the ESS per CPU second by a factor of between 2 and 5 on a range of toy targets described in Section 5.5 and a stochastic volatility model posterior (Section 5.5.3). The greatest improvements were observed when both the ratio between the smallest and largest component scales, and the dimension of the target were large (> 40 and > 100, respectively).

Scheme 3, $\omega(z; z_0) = \tilde{\pi}(z)\|x - x_0\|_2^2$, strikes the balance between sampling efficiency and problem-agnostic implementation, and is applied throughout the remainder of this work.

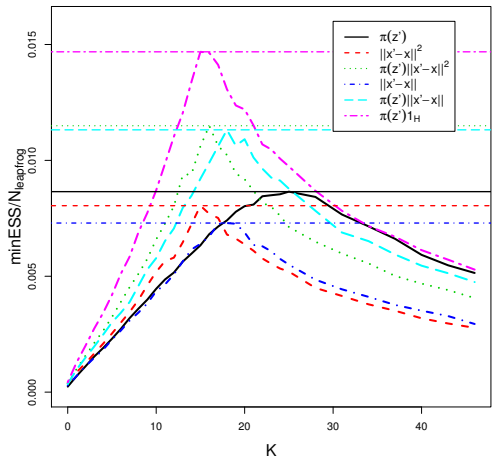
5.3 Tuning

AAPS, as introduced in this work, uses the leapfrog integrator and as a result benefits from its symplecticness. Similar to standard HMC, the variation in the Hamiltonian along $\mathcal{P}$ is only strongly affected by the choice of ε and not K. Figure 5.3.1 shows how HMC and AAPS_K differ in efficiency with respect to their tuning parameters; the efficiency is measured as the minimum ESS over all target components, divided by the total number of leapfrog steps taken. The HMC plot axes are $\varepsilon - L$ as, generally, a practitioner would first tune the step-size and then the integration time in integer multiples of ε. It is evident that AAPS_K is more robust to the variation of the number of segments K from its optimum than HMC is to the integration time $\mathcal{T}$. Additional tuning plots are provided in the Appendix C.1.

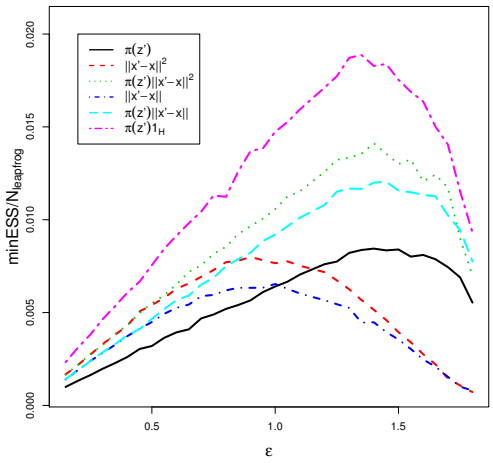
In this section, we discuss the tuning parameters of AAPS_K and outline some general strategies for tuning them.

5.3.1 Step-size tuning

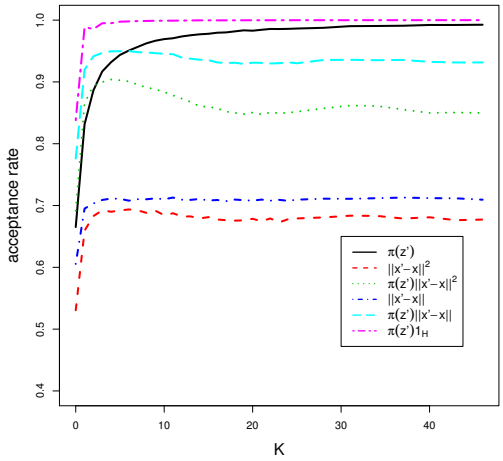
The efficiency of the algorithm will depend on the choice of the step-size ε. If ε is very small then the simulated Hamiltonian dynamics will be close to the true continuous solutions to (5.1.1). This, however, comes at a great computational cost associated



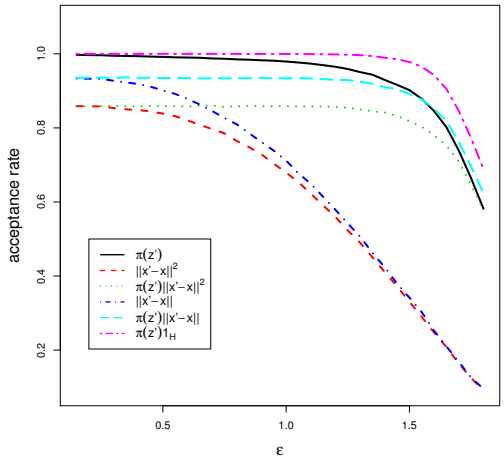
(a) K-efficiency; $\varepsilon = 1.0$



(b) ε-efficiency; $K = 15$

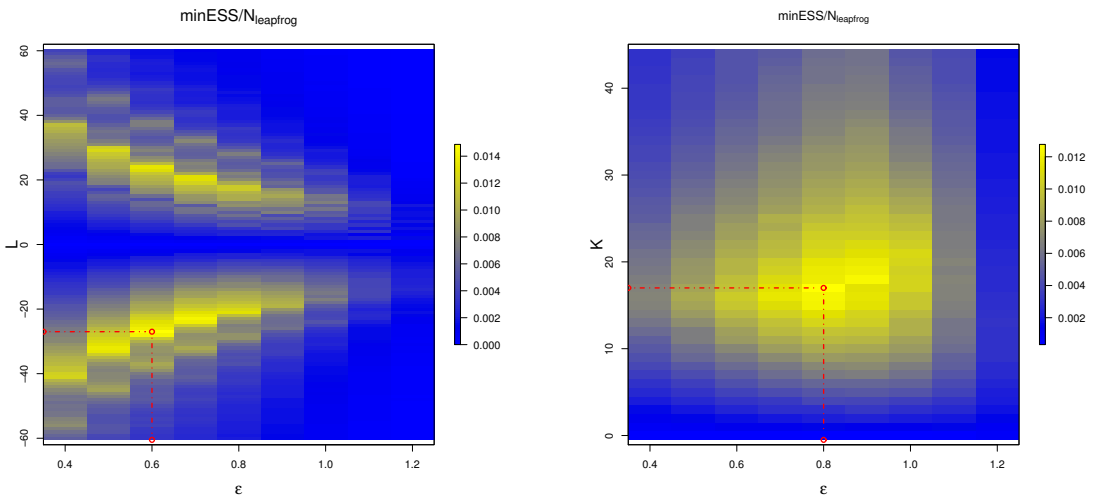


(c) K-acceptance rate; $\varepsilon = 1.0$



(d) ε-acceptance rate; $K = 15$

Figure 5.2.4: Comparison of sampling weights ω. Horizontal lines in (a) help to indicate the respective maxima of the weighting schemes. The acceptance rate given in the bottom plots is computed as the estimated probability of leaving z_0 at a given iteration.



(a) Rosenbrock – HMC (b) Rosenbrock – AAPS

Figure 5.3.1: Efficiency w.r.t. tuning parameters for HMC and AAPS on a Rosebrock-type product target (Section 5.5.1) with $d = 40$, and $\mathsf{range}(\beta) = (1, 100)$). HMC with negative L corresponds to blurred integration time.

with a great number of gradient evaluations, which is generally the most costly element of inference schemes. On the other hand, if ε is too large, the numerical solution will be unstable (see, for example, Hairer et al., 2003), resulting in an unbounded error in the Hamiltonian. Over a range of experiments carried out in Section 5.5, AAPS with sampling weights modulated by $\tilde{\pi}$ was able to achieve optimal performance at acceptance rates between $75 - 87\%$. Due to the simplecticness of the integrator, the acceptance rate was influenced mostly by the choice of ε, with little effect coming from the choice of K, provided $K > 0$; see Figure 5.2.4 for an example of this. In AAPS_0, the true acceptance rate can be quite variable if there are only a few points in the segment. The recommended general tuning procedure for ε in AAPS_K is then:

1. Fix $K = 0$ (AAPS_0) and identify a maximum stable ε for which the stability condition (5.2.1) holds over multiple iterations;

2. Increase K so that the average size of path, $|\mathcal{P}|$,does not change much between iterations.

3. Gradually reduce the step-size until the acceptance rate is within $75 - 87\%$.

Depending on the problem, acceptance rates lower than 75% may result in more efficient mixing. For example, in the stochastic volatility problem in Section 5.5, ε values larger than the optimal, with correspodning acceptance rates $< 75\%$, resulted in chains which were still suitably efficient.

5.3.2 Tuning the number of segments

Here, we introduce a useful diagnostic for determining a suitable choice for the K parameter of AAPS_K.

At each iteration, for given z_0, AAPS_K samples a the initial segment index c uniformly from $\{0, \ldots, K\}$ and proposes a state z' from $\mathcal{P} = S_{-c:(K-c)}$, which is composed of $K+1$ segments, with a probability proportional to our chosen weight function $\omega(z'; z) = \tilde{\pi}(z')\|x' - x_0\|^2$. It will be helpful in the sequel to consider the case where the segment index j of the proposal was, instead, sampled uniformly on $\{-c, \ldots, K-c\}$. This is equivalent to independently and uniformly sampling some values c and e from $\{0, \ldots, K\}$, and setting $j = e - c$. Thus the marginal distribution for j can be described by the mass function

$$\mathbb{P}(J = j) = \begin{cases} \frac{K+1-|j|}{(K+1)^2} & j = -K, \ldots, K, \\ 0 & \text{otherwise.} \end{cases}$$

In deciding on a suitable K, we first perform a relatively short run of AAPS_K for some large $K = K^*$. Each iteration, we note the segment number $j \in \{-c, \ldots, K^* - c\}$ from which the proposal arose. By the symmetry of sampling c and momentum p, the distribution of j is symmetric so we track $k = |j|$, and keep a running total of the number of times, $n_{K^*}(k)$, that there has been a proposal from a segment with an index whose absolute value is $k \in \{0, \ldots, K^*\}$. If the weights had been irrelevant and all segments contained the same number of points then

$$p_K(k) := \mathbb{P}\left(|j| = k | \text{proposed segment sampled uniformly, } K\right) = \begin{cases} \frac{1}{K+1} & k = 0, \\ 2\frac{K+1-k}{(K+1)^2} & k = 1, \ldots, K^*, \\ 0 & \text{otherwise.} \end{cases}$$

The method for placing the set of segments around the initial segment, as described in Section 5.2.2, inherently causes lower values of k to appear more frequently. To account for this, we adjust the running totals

$$m_{K^*}(k) = \frac{n_{K^*}(k)}{p_{K^*}(k)}$$

and choose the optimal number of additional apogees, $\hat{K}$, as

$$\hat{K} = \underset{k=0,\ldots,K^*}{\mathsf{argmax}} \; m_{K^*}(k).$$

The tuning procedure for the number of segments is then

1. Fix a stable ε and pick a large $K = K^*$.

2. Run for a sufficiently large number of iterations so that values $m_{K^*}(k)$ do not change much with iterations.

3. If the diagnostic suggests optimal $\hat{K}$ to be close to K^*, repeat the run with a larger number of segments.

4. If $\hat{K} << K^*$ then proceed with the $K = \hat{K}$.

Appendix C.2 gives a heuristic explanation for the diagnostic; the work was carried out by Prof Chris Sherlock. Figure 5.3.2 compares the optimal K suggested by the proposed diagnostic to the *true* optimal value identified via a grid-search on (K, ϵ). The comparisons are given for skewed-Gaussian product targets described in Section 5.5.1, with varying maximum ratios of the component scales. The plots show good agreement between the proposed diagnostic and the empirical estimate of the truth. Similar results were found for other types of targets investigated in Section 5.5.1.

5.4 Theory

5.4.1 Gaussian product target

We consider a d-dimensional Gaussian target with a diagonal covariance matrix; the potential is

$$U(x) = \sum_{i=1}^{d} \frac{1}{2}\nu_i x_i^2 + \text{constant}, \tag{5.4.1}$$

where $\nu_i > 0,\ i = 1, \ldots, d$, are squared inverse-scales of each component. We let $\pi_\nu(x)$ denote the density of a univariate zero-mean Gaussian random variable with the variance equal to ν^{-1}, and π_ν be the corresponding measure. Note that $\sqrt{\nu_i}x_i \sim \pi_1 \equiv \mathsf{N}(0, 1)$ for all $i = 1, \ldots, d$.

The deterministic map $\phi : (x_0, p_0) \mapsto (x_t, p_t)$ defined by the solution to (5.1.1) has a

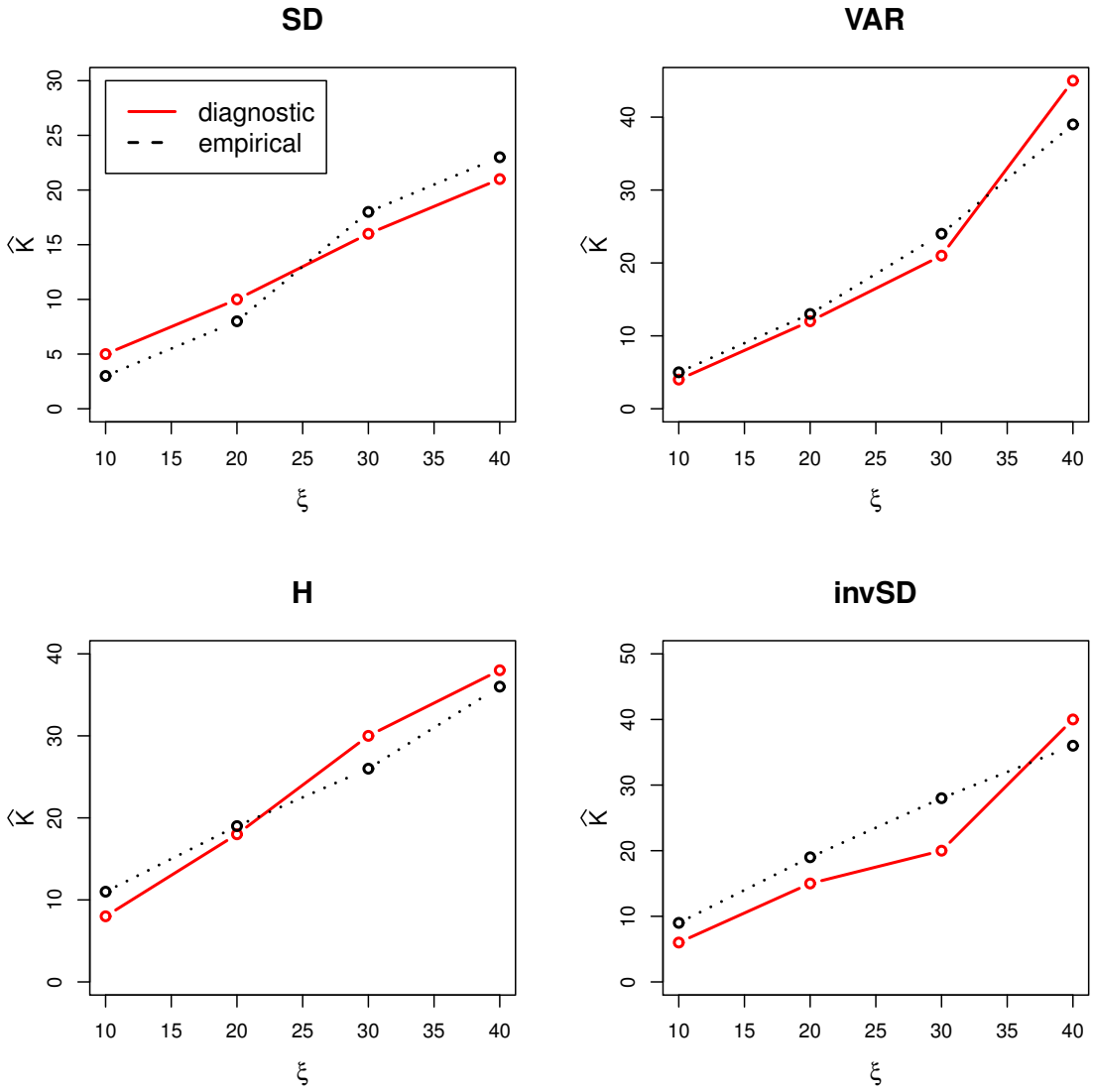


Figure 5.3.2: Comparison of the optimal choice of K as given by the diagnostic (5.3.2) and grid-search. The target distribution is a 40-dimensional Skewed-Gaussian with varying scales-ratio ξ, and scales-spacing linear in standard deviation (SD), variance (VAR), Hessian (H) and inverse standard deviation (invSD).

unit Jacobian and so the density of (X_t, P_t) is $f(x_t, p_t) = \tilde{\pi}_\nu(\phi(x_t, p_t; -t)) = \tilde{\pi}_\nu(x_0, p_0)$. The conservation of energy under the Hamiltonian dynamics is equivalent to

$$\frac{\mathsf{d}}{\mathsf{d}t}\tilde{\pi}_\nu(x_t, p_t) = 0,$$

so $f \equiv \tilde{\pi}_\nu$ and (X_t, P_t) is stationary.

We wish to analyse the behaviour of the Hamiltonian dynamics with respect to the apogees. Without loss of generality, we assume a unit mass, that is, $P \equiv \mathsf{N}(0, I_d)$. Define the scaled dot-product at time t in d dimensions, subject to initial phase-state (x_0, p_0), as

$$D^{(d)}\left(t; x_0^{(d)}, p_0^{(d)}\right) = \frac{1}{\sqrt{d}}\left\langle \nabla U_x\left(x_t^{(d)}\right), p_t^{(d)}\right\rangle. \tag{5.4.2}$$

In the results that follow, it is essential that we first solve the ODE system (5.1.1). As the target components are independent, the trajectories in the phase-space will evolve separately and so we can focus on an individual component, leading to the following Initial Value Problem

$$\frac{\mathsf{d}p(t)}{\mathsf{d}t} = -\sqrt{\nu}x(t), \quad \frac{\mathsf{d}x(t)}{\mathsf{d}t} = p(t); \quad x(0) = x_0, \quad \frac{\mathsf{d}x(0)}{\mathsf{d}t} = p_0. \tag{5.4.3}$$

Combining the two equation we find that $\frac{\mathsf{d}^2x(t)}{\mathsf{d}t^2} = -x(t)$, which is the simple harmonic oscillator equation, with the solution

$$\begin{aligned} x(t) &= x_0 \cos\sqrt{\nu}t + \frac{p_0}{\sqrt{\nu}}\sin\sqrt{\nu}t, \\ p(t) &= -\nu x_0 \sin\sqrt{\nu}t + p_0\cos\sqrt{\nu}t. \end{aligned} \tag{5.4.4}$$

Theorem 5.4.1. *Let the potential be defined as in* (5.4.1) *and let* μ *be a distribution*

with support on $\mathbb{R}^*$ *with*

$$\mathop{\mathbb{E}}_{\nu\sim\mu}\left[\nu^{1+\delta/2}\right] < \infty \tag{5.4.5}$$

for some $\delta > 0$*, and let* $\nu_i \overset{\text{iid}}{\sim} \mu$*. Let* $D^{(d)}$ *be the scaled dot-product defined in* (5.4.2)*, and let* $(X_0, P_0) \sim \tilde{\pi}$*; then, as* $d \to \infty$*,*

$$D^{(d)} \to \tilde{D} \sim \mathsf{SGP}(0, V), \tag{5.4.6}$$

where $Y \sim \mathsf{SGP}(b, V^*)$ *denotes that* Y *is a one-dimensional stationary Gaussian process with an expectation of* b *and a covariance function of* $\mathsf{Cov}\left[Y_t, Y_{t'}\right] = V^*(t'-t)$*, and*

$$V(t-s) = \mathop{\mathbb{E}}_{\nu\sim\mu}\left[\nu\cos 2\sqrt{\nu}(t-s)\right].$$

Proof. The dynamics are initialised randomly from the canonical distribution $\tilde{\pi}_\nu$, that is, $X_0 \sim \pi_\nu$ and $P_0 \sim \rho_1$. Consider the dot-product contribution of a single component

$$D(t; X_0, P_0, \nu) = \left\langle \frac{\mathsf{d}U}{\mathsf{d}x}(X_t), P_t \right\rangle = \nu X_t P_t.$$

Substituting in the solution (5.4.4), we get that

$$\begin{aligned} D(t; X_0, P_0, \nu) &= \nu\left(x_0\cos\sqrt{\nu}t + \frac{p_0}{\sqrt{\nu}}\sin\sqrt{\nu}t\right)\left(\nu x_0\sin\sqrt{\nu}t + p_0\cos\sqrt{\nu}t\right) \\ &= \sqrt{\nu}\left[\frac{1}{2}\left(P_0^2 - \nu X_0^2\right)\sin 2\sqrt{\nu}t + \sqrt{\nu}X_0P_0\cos 2\sqrt{\nu}t\right]. \end{aligned}$$

Due to stationarity of (X_t, P_t), $D(0; X_0, P_0, \nu) \overset{\mathcal{D}}{=} D(t; X_0, P_0, \nu)$ and so we have

$$\mathbb{E}_{(X_0,P_0)\sim\tilde{\pi}_\nu}\left[D(t; X_0, P_0, \nu)\right] = \mathbb{E}_{(X_0,P_0)\sim\tilde{\pi}_\nu}\left[D(0; X_0, P_0, \nu)\right] = \mathbb{E}_{(X_0,P_0)\sim\tilde{\pi}_\nu}\left[\nu X_0 P_0\right] = 0.$$

Now, let us consider the covariance of the dot-product at two different times $t, s > 0$, with $t > s$. Define $V(s,t;\nu) := \mathsf{Cov}_{(X_0,P_0)\sim\tilde{\pi}_\nu}\left[D(s; X_0, P_0, \nu), D(t; X_0, P_0, \nu)\right]$.

Since D is zero-mean,

$$\begin{aligned}
V(s,t;\nu) &= \mathbb{E}\left[D(s;X_0,P_0,\nu)D(t;X_0,P_0,\nu)\right] \\
&= \mathbb{E}\left[\nu^2 X_s P_s X_t P_t\right] \\
&= \nu\mathbb{E}\left[\nu X_0 P_0 X_{t-s} P_{t-s}\right] \\
&= \nu\mathbb{E}\left[\sqrt{\nu}X_0P_0\left(\frac{1}{2}(P_0^2-\nu X_0^2)\sin 2\sqrt{\nu}(t-s)+\sqrt{\nu}X_0P_0\cos 2\sqrt{\nu}(t-s)\right)\right] \\
&= \nu\mathbb{E}\left[\frac{1}{2}\sqrt{\nu}X_0P_0\left(P_0^2-\nu X_0^2\right)\sin 2\sqrt{\nu}(t-s)+\nu X_0^2P_0^2\cos 2\sqrt{\nu}(t-s)\right] \\
&= \nu\cos 2\sqrt{\nu}(t-s),
\end{aligned}$$

where the expectations are taken over $(X_0, P_0) \sim \tilde{\pi}_\nu$, noting that $\sqrt{\nu}X \sim \pi_1$. The Fourier transform of $V(s,t;\nu)$ is a positive measure on $\mathbb{R}$ which, by Bochner's theorem, means that the function itself is positive definite. Combining all components, by the strong law of large numbers,

$$\begin{aligned}
V^{(d)}(s,t) &= \mathsf{Cov}_{(X_0,P_0)\sim\tilde{\pi}_\nu}\left[D^{(d)}\left(s;X_0^{(d)},P_0^{(d)}\right),D^{(d)}\left(t;X_0^{(d)},P_0^{(d)}\right)\right] \\
&= \frac{1}{d}\sum_{i=1}^{d}\nu_i\cos 2\sqrt{\nu_i}(t-s) \\
&\overset{a.s.}{\to} \mathop{\mathbb{E}}_{\nu\sim\mu}\left[\nu\cos 2\sqrt{\nu}(t-s)\right] = V(t-s),\ \text{as } d\to\infty.
\end{aligned}$$

All $V^{(d)}$ are positive definite as they are sums of positive definite functions. Furthermore, letting F denote the Fourier transform operator and $f_1(r) = \cos 2\sqrt{\nu}r$, we find that

$$\mathsf{F}\{V\}(\omega) = \mathop{\mathbb{E}}_{\nu\sim\mu}\left[\nu\mathsf{F}\{f_1\}(\omega)\right] = \mathop{\mathbb{E}}_{\nu\sim\mu}\left[\nu\sqrt{\pi/2}\left(\mathbb{I}_{\{\omega=2\sqrt{\nu}\}}+\mathbb{I}_{\{\omega=-2\sqrt{\nu}\}}\right)\right].$$

Since $\nu > 0$, $\mathsf{F}\{V\}$ is a positive measure on $\mathbb{R}$ meaning that V is a positive definite

function.

Now, take any finite set of time-points $\{t_1, \ldots, t_n\},\ n \in \mathbb{N}$ and let

$$D_{1:n} = \left(D(t_1; X_0, P_0, \nu), \ldots, D(t_n; X_0, P_0, \nu)\right)^\top.$$

For any $\delta > 0$ such that $\mathbb{E}_\mu[\nu^{1+\delta/2}] < \infty$,

$$\begin{aligned}
\|D_{1:n}\|^{2+\delta} &= \left(\sum_{j=1}^{n} |D(t_j; X_0, P_0, \nu)|^2\right)^{1+\delta/2} \\
&= n^{1+\delta/2}\left(\frac{1}{n}\sum_{j=1}^{n} |D(t_j; X_0, P_0, \nu)|^2\right)^{1+\delta/2} \\
&\leq n^{1+\delta/2}\left(\frac{1}{n}\sum_{j=1}^{n} |D(t_j; X_0, P_0, \nu)|^{2+\delta}\right) \\
&= n^{\delta/2}\sum_{j=1}^{n} |D(t_j; X_0, P_0, \nu)|^{2+\delta},
\end{aligned}$$

with the third line being a consequence of Jensen's inequality with respect to the quadratic function. Taking the expectation over $(X_0, P_0) \sim \tilde{\pi}_\nu$, we have

$$\begin{aligned}
\mathbb{E}_{(X_0,P_0)\sim\tilde{\pi}_\nu}\left[\|D_{1:n}\|^{2+\delta}|\nu\right] &\leq n^{\delta/2}\sum_{j=1}^{n} \mathbb{E}_{(X_0,P_0)\sim\tilde{\pi}_\nu}\left[|D(t_j; X_0, P_0, \nu)|^{2+\delta}|\nu\right] \\
&= n^{1+\delta/2}\nu^{1+\delta/2}\mathbb{E}_{(X_0,P_0)\sim\tilde{\pi}_\nu}\left[|\sqrt{\nu}X_0P_0|^{2+\delta}|\nu\right] \\
&= n^{1+\delta/2}\nu^{1+\delta/2}\mathbb{E}_{\pi_1}\left[|X_0^*|^{2+\delta}\right]\mathbb{E}_{\rho_1}\left[|P_0|^{2+\delta}\right] \\
&= n^{1+\delta/2}\nu^{1+\delta/2}m(2+\delta)^2 < \infty, \qquad (5.4.7)
\end{aligned}$$

where $m(a) = \mathbb{E}[|Z|^a],\ Z \sim \mathsf{N}(0, 1)$.

As $\|D_{1:n}\| > 0$, we have (for some $a > 0$)

$$\begin{aligned}\mathbb{E}\left[\|D_{1:n}\|^{2+\delta}\right] &\geq \mathbb{E}\left[\|D_{1:n}\|^{2+\delta}\mathbb{I}_{\{\|D_{1:n}\|\geq a\}}\right] \\ &\geq a^{\delta}\mathbb{E}\left[\|D_{1:n}\|^{2}\mathbb{I}_{\{\|D_{1:n}\|\geq a\}}\right].\end{aligned}$$

Using the above inequality with $a = \varepsilon\sqrt{d},\ \varepsilon > 0$, in combination with (5.4.7), we arrive at the Lindeberg condition

$$\frac{1}{d}\sum_{i=1}^{d}\mathbb{E}\left[\left\|D_{i,1:n}^{(d)}\right\|^{2}\mathbb{I}_{\left\{\left\|D_{i,1:n}^{(d)}\right\|\geq\varepsilon\sqrt{d}\right\}}\right] \leq \frac{1}{d}\sum_{i=1}^{d}\frac{n^{1+\delta/2}}{\varepsilon^{\delta}d^{\delta/2}}\nu_i^{1+\delta/2}m(2+\delta)^2 \to 0 \quad \text{as} \quad d\to\infty.$$

Since V is positive definite, by the Lindeberg-Feller Central Limit Theorem the n-dimensional distribution tends towards a multivariate Gaussian with a covariance given according to V which proves the result. □

Corollary 5.4.2. *The expected number of apogees of $\tilde{D}$ over an integration time T is*

$$N(T) = \frac{T\ \mathbb{E}[\nu]}{\pi} \times \sqrt{\frac{\mathbb{E}[\nu^2]}{\mathbb{E}[\nu]^2}}. \tag{5.4.8}$$

Proof. Recall the covariance function of $\tilde{D}$,

$$V(t) = \mathop{\mathbb{E}}_{\nu\sim\mu}\left[\nu\cos 2\sqrt{\nu}t\right]$$

with the variance of the process being $V(0) = \mathbb{E}[\nu]$, with the integration over μ being implicit. The process, $D_* = \tilde{D}/\sqrt{V(0)}$ is a stationary Gaussian process with unit variance. Ylvisaker (1965) shows that the expected number of zeroes over a unit interval of a stationary GP with unit variance and covariance function C is:

$\frac{1}{\pi}\sqrt{-C''(0)}$. The second derivative of the covariance function of D_* evaluated at 0 is

$$V_*''(t) = -4\frac{\mathbb{E}[\nu^2 \cos 2\sqrt{\nu}t]}{\mathbb{E}[\nu]}$$

with $V_*''(0) = -4\mathbb{E}[\nu^2]/\mathbb{E}[\nu]$. The expected number of zeroes of D_* is the same as that of $\tilde{D}$. Thus, the expected number of apogees over a time T is

$$\frac{T}{2\pi}\sqrt{-V_*''(0)} = \frac{T}{2\pi}\sqrt{4\frac{\mathbb{E}[\nu^2]}{\mathbb{E}[\nu]}} = \frac{T\ \mathbb{E}[\nu]}{\pi} \times \sqrt{\frac{\mathbb{E}[\nu^2]}{\mathbb{E}[\nu]^2}}$$

as required. □

Given that the marginal distribution of the momentum variable, P_0, is set to be $\mathsf{N}(0, I)$ and an identity mass matrix is used in defining the Hamiltonian dynamics, the first term of the product (5.4.8) relates the integration time to the average length scale of the target density as ν is the inverse variance. The second terms of the product shows how $N(T)$, for a fixed T, increases with the variability in the squared inverse scales of the target components. Thus, the tuning parameter K simply relates to the properties intrinsic to the target density, being relatively unaffected by a uniform redefinition of the length scales of the target or changing the choice of ε. This is in contrast to L in HMC where the performance of the algorithm with change with such redefinition. This gives another reason for the robustness of AAPS with respect to its tuning parameters.

5.4.2 General product target

Theorem 5.4.4 and Corollary 5.4.5 were proved by Prof Chris Sherlock and the proofs can be found in Sherlock et al. (2021).

Theorem 5.4.1 can be extended to a general d-dimensional product target with a potential

$$U^{(d)}(x) = \sum_{i=1}^{d} g\left(\sqrt{\nu^{(d)} x_i^{(d)}}\right) + \text{constant}, \tag{5.4.9}$$

for some $g : \mathbb{R} \to \mathbb{R}$, where $\int_{\mathbb{R}^d} \mathsf{exp}\left(-U^{(d)}x\right) \mathsf{d}x = 1$.

Assumption 5.4.3. *We assume that $g \in C^1$, that there is a $\delta > 0$ such that*

$$\mathbb{E}_{X\sim\pi_1}\left[|g'(X)|^{2+\delta}\right] < \infty, \tag{5.4.10}$$

and that for each $y_0 \in \mathbb{R}$, there is a unique, non-explosive solution $a(t; y_0, y_0')$ to the initial value problem:

$$\frac{\mathsf{d}^2 y}{\mathsf{d}t^2} = -g'(y);\ y(0) = y_0,\ y'(0) = y_0'. \tag{5.4.11}$$

Theorem 5.4.4. *Let the potential be defined as in (5.4.9) and where g satisfies the assumptions around (5.4.10) and (5.4.11). Further, let μ be a distribution with support on $\mathbb{R}^+$ with*

$$\mathop{\mathbb{E}}_{\nu\sim\mu}\left[\nu^{1+\delta/2}\right] < \infty \tag{5.4.12}$$

for some $\delta > 0$, and let $\nu_i \overset{iid}{\sim} \mu$. Define

$$V(t) = \mathbb{E}\left[\nu\ g'(X)P\ g'(\sqrt{\nu}X_t; X, P)\ a'(\sqrt{\nu}t; X, P)\right] \tag{5.4.13}$$

where the expectation is over the independent variables $X \sim \pi_1$, $P \sim \rho_1$ and $\nu \sim \mu$, and assume that for any finite sequence of n distinct times $(t_1, \ldots, t_n)$, the $n \times n$ matrix Σ with $\Sigma_{i,j} = V(t_j - t_i)$ is positive definite. Let $D^{(d)}$ be the scaled dot-product

defined in (5.4.2)*, and let* $(X_0, P_0) \sim \tilde{\pi}$*; then, as* $d \to \infty$

$$D^{(d)} \to \tilde{D} \sim \mathsf{SGP}(0, V),$$

where $Y \sim \mathsf{SGP}(b, V)$ *denotes that* Y *is a one-dimensional stationary Gaussian process with an expectation of* b *and a covariance function of* $\mathsf{Cov}\,[Y_t, Y_{t'}] = V(t' - t)$.

Corollary 5.4.5. *For a potential of the form* (5.4.1)*, the expected number of apogees of* $\tilde{D}$ *over an integration time* T *is*

$$N(t) \propto T\ \mathbb{E}[\nu] \times \sqrt{\frac{\mathbb{E}[\nu^2]}{\mathbb{E}[\nu]^2}}.$$

5.5 Numerical experiments

The optimal performance of HMC is invariant to the choice of a linear transformation of the original variables, that is, an algorithm sampling x using a mass matrix M is equivalent to sampling Ax using a mass matrix $(A^\top)^{-1} M A^{-1}$. Thus, without a loss of generality, we use $M = I_d$ throughout the analyses. Each set of experiments is run for 10^5 iterations, initiated from the corresponding stationary distributions. The efficiency, unless specified otherwise, is given as the minimum ESS over all target components divided by the number of gradient evaluations (or equivalently leapfrog steps). All algorithms were implemented in `C++`.

5.5.1 Toy product targets

We evaluate the algorithms on a selection of d-dimensional centralised product targets where each component x_i has a different scaling parameter $\sigma_i > 0$:

- Gaussian:

$$\pi_G(x) = \prod_{i=1}^{d} \frac{1}{\sigma_i\sqrt{2\pi}} \mathsf{e}^{-\frac{x_i^2}{2\sigma_i}};$$

- Logistic:

$$\pi_L(x) = \prod_{i=1}^{d} \frac{\mathsf{e}^{-x_i/\sigma_i}}{\sigma_i(1+\mathsf{e}^{-x_i/\sigma_i})^2};$$

- Skew-Gaussian:

$$\pi_{SG}(x) = \prod_{i=1}^{d} \frac{2}{\sigma_i\sqrt{2\pi}} \mathsf{e}^{-\frac{x_i^2}{2\sigma_i}} \Phi\left(\frac{\alpha x_i}{\sigma_i}\right),$$

  where Φ is the distribution function of a standard normal random variable and $\alpha \in \mathbb{R}$. We set $\alpha = 3$.

In all three targets, we have $\mathbb{V}[X_i] \propto \sigma_i^2$, with equality in the Gaussian case. As a result, the square-root of the condition number of the covariance matrix in each case is dictated by the largest ratio of component scales; we denote this by

$$\xi = \max_{1\le i,j\le d} \sqrt{\sigma_i/\sigma_j}.$$

We consider four different kinds of jittered, linear spacings for scales σ_i. We consider the component variances (σ^2) linear in component indexing, as well as the standard deviation (σ), the inverse standard deviations (σ^{-1}), and the diagonal elements of the Hessian (σ^{-2}). Each set of scales is constructed using a d-dimensional vector $\boldsymbol{\gamma}$ such that $\gamma_1 = 0$ and $\gamma_d = 1$, and for $i = 2, \ldots, d-1$, $\gamma_i = (i + 1 + \upsilon_i)/(d-1)$ where υ_i are independent $\mathsf{Unif}(-0.5, 0.5)$ variables. For each of the four different spacings, the scales are:

- SD: $\sigma_i = (\xi - 1)\gamma_i + 1$;

- VAR: $\sigma_i^2 = (\xi^2 - 1)\gamma_i + 1$;
- H: $1/\sigma_i^2 = (1 - 1/\xi^2)\gamma_i + 1/\xi^2$;
- invSD: $1/\sigma_i = (1 - 1/\xi)\gamma_i + 1/\xi$

for $i = 1, \ldots, d$. In all cases, $\sigma_1 = 1$ and $\sigma_d = \xi$ to ensure stability in ε across experiments. The jittering prevents any scales, other than σ_1 and σ_d, from being integer multiples of each other; such scenario, which is unlikely to occur in practice, would result in very odd behaviour of the optimal integration time for HMC and could confound the results.

We also include a particularly difficult target, denoted by π_G^{RN}, which was used in an online comparison between HMC and NUTS.[1] The target is 30-dimensional and has $\xi = 110$.

Modified Rosenbrock

We additionally include a *banana*-shaped target which is a popular benchmark for MCMC algorithms (e.g., Hogg and Foreman-Mackey, 2018; Martin et al., 2012; Hoffman et al., 2021; Pagani et al., 2021) with its log-density based on the Rosenbrock function $f(x, y) = -\beta(x^2 - y)^2 - (x - \alpha)^2$, $\beta > 0$ and $\alpha \in \mathbb{R}$. We modify the original function as HMC is not an appropriate method for sampling from a light-tailed target; specifically, HMC is stable when the tails do not increase faster than quadratically (Livingstone et al., 2019), but here they would increase quartically. The modified

[1] https://radfordneal.wordpress.com/2012/01/27/evaluation-of-nuts-more-comments-on-the-paper-by-hoffman-and-gelman/

Rosenbrock-based target has the form

$$\log \pi_{MR}(x, y) = -\frac{1}{2}\left(\frac{cx^2}{1 + c^2x^2/2} - y\right)^2 - c^2(x - \alpha/c)^2 \tag{5.5.1}$$

up to an additive constant, where $\alpha > 0$ and $c = (2\beta)^{-\frac{1}{2}} > 0$. This is equivalent to

$$\begin{aligned} X &\sim \mathsf{N}(\alpha/c, c^{-2}/2), \\ Y|X = x &\sim \mathsf{N}\left(\frac{cx^2}{1 + c^2x^2/2}, 1\right). \end{aligned}$$

The partial derivatives of the log-density are

$$\begin{aligned} \partial_x \log \pi_{MR}(x, y) &= -\frac{8cx}{(c^2x^2 + 2)^2}\left(\frac{cx^2}{1 + c^2x^2/2} - y\right) - 2c(cx - \alpha), \\ \partial_y \log \pi_{MR}(x, y) &= \left(\frac{cx^2}{1 + c^2x^2/2} - y\right). \end{aligned}$$

The partial derivative with respect to x is now $\mathcal{O}(x)$ making it stable for HMC. In the numerical experiments, for a d-dimensional target (d even), we consider a product of $d/2$ bivariate densities based on (5.5.1) with the β parameters linearly distributed between 1 and 100; that is, $\beta_i = 1 + \frac{99(i-1)}{d/2-1}$, for $i = 1, \ldots, d/2$.

Comparison

Table 5.5.1 shows the efficiencies of standard HMC, blurred HMC and the no U-turn sampler relative to AAPS_K. For some targets, some of the other algorithms are slightly more efficient than AAPS_K, and for others, they are slightly less efficient. However, across the wide range of targets, no algorithm is 1.7 or more times as efficient as AAPS_K. Table 5.5.2 gives the respective acceptance rates of the algorithms fixed at optimal tuning parameters. For the given targets, the HMC acceptance rates are far

Target	ξ	d	AAPS_K	HMC	HMC-bl	NUTS
π_G^{SD}	20	40	1.000	0.722	0.718	1.182
π_G^{VAR}	20	40	1.000	1.016	1.091	1.461
π_G^{H}	20	40	1.000	0.162	0.644	0.392
π_G^{invSD}	20	40	1.000	0.162	0.461	0.460
π_{SG}^{VAR}	20	40	1.000	0.990	1.224	1.403
π_L^{VAR}	20	40	1.000	1.135	1.488	1.677
π_G^{VAR}	20	100	1.000	0.657	1.020	1.378
π_G^{VAR}	40	40	1.000	1.190	1.346	1.645
π_{MR}	-	20	1.000	1.647	1.582	0.728
π_{MR}	-	40	1.000	1.045	1.166	0.873
π_G^{RN}	110	30	1.000	*0.019	1.206	0.306

Table 5.5.1: Efficiency of algorithms measured minimum ESS across all d components per number of gradient evaluations. The values are scaled w.r.t. AAPS_K
*The tuning surface of HMC was extremely uneven so we could not determine the optimal tuning parameters with any degree of certainty.

from the recommended 65.1% and, similarly, NUTS is far from the default setting of 80% as it appears in `Stan`.

5.5.2 Bimodal distributions

In Section 5.2.5, we demonstrated that AAPS_0 is irreducible on bimodal targets with dimension $d > 1$. We now compare the efficiency of AAPS_K to HMC-bl and NUTS on a selection of targets of the form (5.2.2), with a and σ varied. Table 5.5.3 shows the results with the efficiency defined as the effective sample size of the first component per number of gradient evaluations. The a and σ values were chosen such that the 1^{st} and 4^{th} rows of the table correspond to a very slight separation of modes and rows 3 and 6 correspond to very substantial separation. Ideal HMC-bl, tuned on a fine (ε, L) grid, outperforms the other two algorithms quite substantially. Comparing the first

Target	ξ	d	AAPS$_K$	HMC	HMC-bl	NUTS
π_G^{SD}	20	40	86.6	90.7	78.1	99.1
π_G^{VAR}	20	40	84.9	91.6	88.4	99.8
π_G^{H}	20	40	83.7	63.0	71.7	82.8
π_G^{invSD}	20	40	83.8	70.1	72.4	91.2
π_{SG}^{VAR}	20	40	84.6	93.2	90.6	94.3
π_L^{VAR}	20	40	74.7	73.9	77.1	96.3
π_G^{VAR}	20	100	84.5	78.6	75.7	96.4
π_G^{VAR}	40	40	84.4	97.5	77.1	96.8
π_{MR}	-	20	82.8	78.9	84.8	89.9
π_{MR}	-	40	84.7	80.7	80.0	89.7
π_G^{RN}	110	30	83.7	*62.0	66.7	99.4

Table 5.5.2: Acceptance rates (%) at optimal algorithm settings. *The acceptance rate of HMC on π_G^{RN} was extremely sensitive to $\mathcal{T}$.

Target	HMC-bl	AAPS$_K$	NUTS
$a = 1.5, \sigma = 1$	8242.29	4325.50	6108.39
$a = 2.5, \sigma = 1$	738.37	383.78	654.23
$a = 3.5, \sigma = 1$	86.30	31.96	48.94
$a = 7, \sigma = 10$	347.35	231.37	209.38
$a = 10, \sigma = 10$	158.21	119.56	82.33
$a = 15, \sigma = 10$	37.20	20.91	14.47

Table 5.5.3: Algorithm performances on 40-dimensional bimodal targets of the form $\frac{1}{2}\mathsf{N}\left((-a, 0, \ldots, 0)^\top, I_{40}\right) + \frac{1}{2}\mathsf{N}\left((a, 0, \ldots, 0)^\top, \sigma^2 I_{40}\right)$. The efficiency is given as $\text{ESS}(X_1)/N_{\text{leapfrog}} \times 10^5$.

three rows of the tables to the last three, AAPS$_K$ decreases the least in performance.

5.5.3 Stochastic volatility

For the final comparison example, we use a stochastic volatility model as used in Girolami and Calderhead (2011) and Wu et al. (2019). Consider sequence of T latent volatility variables $X_{1:T}$ taking the form of an auto-regressive AR(1) process, with the

following distributional assumptions:

$$t = 1 :$$

$$X_1 \sim \mathsf{N}\left(0, \frac{\sigma^2}{1-\phi^2}\right),$$

$$t = 2, \ldots, T :$$

$$X_t | X_{t-1} = x_{t-1} \sim \mathsf{N}(\phi x_{t-1}, \sigma^2),$$

where $\phi \in (0, 1)$. We observe a sequence $Y_{1:T}$ generated by the latent volatilities, which are distributed as

$$Y_t | X_t = x_t \sim \mathsf{N}(0, \kappa^2 \mathsf{exp}(x_t)), \quad t = 1, \ldots, T.$$

We wish perform inference on $(\phi, \kappa, \sigma^2, X_{1:T})$ given observations $y_{1:T}$, with the parameter priors $\pi_0(\kappa) \propto 1/\kappa$, $\sigma^2 \sim \mathsf{inv}\text{-}\chi^2(10, 0.05)$ and $\frac{1}{2}(1+\phi) \sim \mathsf{Beta}(20, 1.5)$. For inference, each parameter is transformed to be on the whole of the real line,

$$\alpha = 2\mathsf{tanh}^{-1}(\phi), \quad \beta = \mathsf{log}(\kappa), \quad \text{and } \gamma = \mathsf{log}(\sigma^2).$$

The full model posterior as well as gradients of the potential are given in Appendix C.3. The data $Y_{1:T}$ are generated using true parameters $\phi = 0.98$, $\kappa = 0.65$ and $\sigma = 0.15$, with $T = 1000$, and can be seen in Figure 5.5.1 along with the true $X_{1:T}$.

In this example, we only consider the blurred version of HMC as for the version without the blurring it was substantially more difficult to identify a stable and optimal tuning; in real-world applications practitioners would be using inference packages such as `Stan` or `PyMC` (Salvatier et al., 2016) which perform the blurring by default. In the results, we include $\mathsf{AAPS}^{\mathsf{d}}_K$ which is the AAPS_K algorithm with K tuned only using

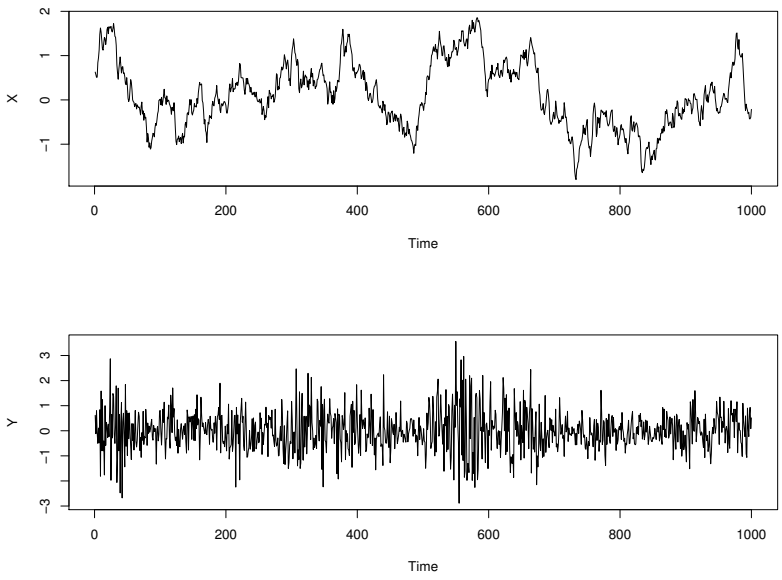


Figure 5.5.1: Synthetic dataset based on the stochastic volatility model. Top: Latent volatilites $X_{1:1000}$. Bottom: Observed returns $Y_{1:1000}$.

the diagnostic from Section 5.3.2. Tables 5.5.4 and 5.5.5 show the performance of the algorithms on the stochastic volatility example. Each cell is given as "mean (sd)" over 10 chains of 10^5 iterations where all chains are initialised from stationarity using a posterior sample obtained from the endpoint of a NUTS chain of length $250,000$.

It took approximately 53 minutes to run a single chain of HMC-bl; the relative times for AAPS_0, $\mathsf{AAPS}_K^{\mathsf{d}}$, AAPS_K and NUTS were 0.39, 2.50, 2.91 and 4.41 respectively.

Efficiency: ESS/$N_{\text{leapfrog}} \times 10^5$						
Algorithm	ESS(α)	ESS(β)	ESS(γ)	$\mathsf{min}_t\text{ESS}(X_t)$	$\text{min}_{\text{all}}\text{ESS}$	ESS($\log \pi$)
AAPS_K	32.29 (3.97)	14.53 (1.35)	14.43 (0.50)	33.60 (4.42)	13.86 (0.71)	14.11 (0.50)
$\mathsf{AAPS}^{\mathsf{d}}_K$	33.66 (6.00)	12.68 (1.61)	17.25 (0.74)	29.86 (5.41)	12.68 (1.61)	16.90 (0.58)
HMC-bl	33.96 (5.83)	15.16 (1.82)	16.41 (1.03)	36.12 (6.52)	14.71 (1.53)	15.97 (0.92)
NUTS	23.48 (8.20)	18.07 (4.22)	10.71 (2.74)	26.08 (14.7)	10.62 (2.93)	10.40 (2.68)
AAPS_0	89.92 (21.0)	1.78 (0.62)	42.57 (3.47)	29.47 (3.51)	1.78 (0.62)	43.02 (5.32)
Efficiency: ESS/Time						
Algorithm	ESS(α)	ESS(β)	ESS(γ)	$\mathsf{min}_t\text{ESS}(X_t)$	$\text{min}_{\text{all}}\text{ESS}$	ESS($\log \pi$)
AAPS_K	31.45 (3.95)	14.15 (1.39)	14.04 (0.47)	32.73 (4.45)	13.50 (0.76)	13.74 (0.44)
$\mathsf{AAPS}^{\mathsf{d}}_K$	32.85 (5.95)	12.37 (1.58)	16.83 (0.79)	29.14 (5.32)	12.37 (1.58)	16.49 (0.61)
HMC-bl	55.15 (9.51)	24.62 (2.94)	26.66 (1.72)	58.65 (10.6)	23.90 (2.50)	25.93 (1.55)
NUTS	15.84 (5.49)	12.19 (2.81)	7.22 (1.83)	17.56 (9.87)	7.16 (1.96)	7.02 (1.78)
AAPS_0	76.54 (18.05)	1.51 (0.53)	36.23 (3.07)	25.08 (3.02)	1.51 (0.53)	36.61 (4.55)

Table 5.5.4: Efficiencies (mean (sd)) of AAPS, HMC and NUTS, run for 10 chains each, on 1003-dimensional stochastic volatility model problem. N_{leapfrog} is the total number of gradient evaluations and time is given in hundreds of seconds. $\mathsf{AAPS}^{\mathsf{d}}_K$ is tuned using the diagnostic given in Section 5.3.2 and remaining algorithms were tuned via a grid search. The second column from the right shows the lowest efficiency over all of $\theta = (\alpha, \beta, \gamma, X_{1:T})$, averaged over 10 chains for each algorithm.

Efficiency: ESS/N_{leapfrog} (scaled)						
Algorithm	ESS(α)	ESS(β)	ESS(γ)	$\min_t$ESS(X_t)	$\min_{\text{all}}$ESS	ESS($\log\pi$)
AAPS_K	1.00 (0.12)	1.00 (0.09)	1.00 (0.03)	1.00 (0.13)	1.00 (0.05)	1.00 (0.04)
$\mathsf{AAPS}^{\mathsf{d}}_K$	1.04 (0.19)	0.87 (0.11)	1.20 (0.05)	0.89 (0.16)	0.91 (0.12)	1.20 (0.04)
HMC-bl	1.05 (0.18)	1.04 (0.13)	1.14 (0.07)	1.07 (0.19)	1.06 (0.11)	1.13 (0.06)
NUTS	0.73 (0.25)	1.24 (0.29)	0.74 (0.19)	0.78 (0.44)	0.76 (0.21)	0.74 (0.19)
AAPS_0	2.79 (0.65)	0.12 (0.04)	2.95 (0.24)	0.88 (0.10)	0.13 (0.04)	3.05 (0.38)
Efficiency: ESS/Time (scaled)						
Algorithm	ESS(α)	ESS(β)	ESS(γ)	$\min_t$ESS(X_t)	$\min_{\text{all}}$ESS	ESS($\log\pi$)
AAPS_K	1.00 (0.13)	1.00 (0.10)	1.00 (0.03)	1.00 (0.14)	1.00 (0.06)	1.00 (0.03)
$\mathsf{AAPS}^{\mathsf{d}}_K$	1.04 (0.19)	0.87 (0.11)	1.20 (0.05)	0.89 (0.16)	0.91 (0.12)	1.20 (0.04)
HMC-bl	1.75 (0.30)	1.74 (0.21)	1.90 (0.12)	1.79 (0.32)	1.77 (0.18)	1.89 (0.11)
NUTS	0.50 (0.17)	0.86 (0.20)	0.51 (0.13)	0.54 (0.30)	0.53 (0.15)	0.51 (0.13)
AAPS_0	2.43 (0.57)	0.10 (0.04)	2.58 (0.22)	0.77 (0.09)	0.11 (0.04)	2.67 (0.33)

Table 5.5.5: Efficiencies (mean (sd)) of AAPS, HMC and NUTS on 1003-dimensional stochastic volatility model problem. All values were scaled such that the eficiency of AAPS_K is 1. $\mathsf{AAPS}^{\mathsf{d}}_K$ is tuned using the diagnostic given in Section 5.3.2 and remaining algorithms were tuned via a grid search. The second column from the right shows the lowest efficiency over all of $\theta = (\alpha, \beta, \gamma, X_{1:T})$, averaged over 10 chains for each algorithm.

5.6 Discussion and further work

We introduced the general framework of the *Apogee to Apogee Path Sampler* which has similar efficiency to Hamiltonian Monte Carlo but is a lot easier to tune. The key feature of AAPS is the construction of a path, $\mathcal{P}$ made up of a fixed number of *segments*, with the construction being invariant to the choice of the initial phase-state $z \in \mathcal{P}$. From the path, AAPS proposes a state with respect to some generic predetermined weight function and then uses an accept-reject step to satisfy the detailed balance condition, thus ensuring the sampler targets the intended distribution.

Unlike the no-U-turn sampler algorithm, AAPS does not require recursive computation and parts of the algorithm can be parallelised; specifically, the backwards and forward integration from the current state z_0 can be carried out separately on two cores, and this can be combined with the fixed-memory proposal and acceptance procedure outline in Section 5.2.6.

In Section 5.2.6, we investigated six different mechanisms for proposing a state from the path, and for the numerical experiments in Section 5.5, we chose to sample a state from the path with a probability proportional to the proposal's squared displacement in the position variable with respect to the initial state. This could be achieved at an $\mathcal{O}(1)$ memory cost. As discussed in Section 5.2.6, a proposal mechanism using a squared jumping distance of any arbitrary function of interest could be implemented at a fixed cost to the memory. We could, for example, use a weighting motivated by the ChEES diagnostic of Hoffman et al. (2021); that is, $\omega(z'; z) \propto (\|x' - \mu\|^2 - \|x - \mu\|^2)^2$ for some central point μ.

In the work, we constructed the path by choosing the position of the initial segment, $\mathcal{S}_0$, uniformly at random from the $K+1$ segments. This is not the only choice that can preserve the detailed balance with respect to the intended target. For instance, the current segment could be fixed at segment 0 and we would only propose from segment K. This choice bears a resemblance to the window scheme in Neal (2011). However, in early experiments, we found that this had a negative impact on the robustness of the efficiency to the choice of K.

Because of the simplicity of AAPS, there are many ways in which the algorithm can be extended. For example, if the states along the path are stored, then a delayed rejection step can improve the efficiency by increasing the acceptance probabilities. Using the gradient evaluations from the leapfrog integrator, an approximation $\tilde{\pi}^* \approx \tilde{\pi}$ could be used in constructing the weight $\omega(z'; z) \propto \tilde{\pi}^*(z')$. Provided the approximation is agnostic of the initial state, the resulting Markov Chain would satisfy the detailed balance condition.

The algorithm, as introduced, is reversible due to the accept-reject step and, as a result, it is possible for costly, distant proposals to still be rejected, wasting the computational effort. A non-reversible version of the algorithm could set $K = 1$ and, instead of completely refreshing the momentum at each iteration, the momentum at the start of a new step could be a Crank-Nicolson perturbation of the momentum at the end of the previous step as in Horowitz (1991).

We focused on constructing a path of states which relies on the notion of apogees, the points where the trajectory of the Hamiltonian dynamics moves from travelling uphill to downhill with respect to the potential surface. A related concept is the

perigee which is the point of swapping from moving downhill to uphill with respect to the potential surface U. Future extensions could involve incorporating perigees into the path construction; although, the exact implementation is not obvious.

Chapter 6

Conclusions

This thesis tackled three different problems in Bayesian modelling and inference. In all three, the methods introduced as solutions to those problems employ the simulation of stochastic processes. Additionally, they all emphasise Bayesian inference through the use of computational methods with easy-to-follow tuning guidelines.

The first contribution was a novel, flexible framework for modelling and prediction of patient recruitment to Phase III clinical trials. The introduced model was based on the time-inhomogeneous Poisson process, and the efficient inference was carried out in the Bayesian paradigm; sensible choices for parameter priors were provided to encourage widespread use. The model was superior to time-homogeneous industry-standard models in terms of prediction.

The main contribution of the work in Chapter 4 was the development of a novel Cox process likelihood estimator, along with an analytical and empirical demonstration of its properties. The proposed Rao-Blackwellised Thinning Estimator was then applied in a random-weight particle filter for Cox process inference, with a diffusion process

prior. We demonstrate how the proposed estimation method is more efficient, in a sense of estimator precision for a given computational budget, than the widely-used Poisson Estimator when applied in a particle filter.

The final contribution of this thesis was the Apogee to Apogee Path Sampler algorithm for sampling from an unnormalised distribution on $\mathbb{R}^d$. The algorithm is based on the commonly used Hamiltonian Monte Carlo; however, the particular construction of the proposals mitigates the high sensitivity of the tuning parameters. Specifically, empirical evidence shows how sampling from a path constructed by integrating the Hamiltonian dynamics forwards and backwards in segments separated by apogees leads to an improvement in tuning misspecification robustness over the seemingly arbitrary integration time T as in standard HMC. The algorithm, as introduced, is competitive in efficiency and allows for many further modifications.

We now outline some suggestions for future research directions, particularly in the context of the whole thesis.

The base model introduced in Chapter 3 had a simple form to allow for widespread use of the methodology in the area of clinical trial operations. An immediate extension would be to modify the basic form of the intensity function. Generally, there are a number of covariates available for each recruiting centre. Using $\mathbf{x}_c$ to denote this vector of covariates, the intensity model could take the form

$$\lambda_c(t) = \lambda_c^o \mathsf{exp}\left(\boldsymbol{\beta}^\top \mathbf{x}_c\right) g\left(t; \mathsf{exp}\left(\boldsymbol{\eta}^\top \mathbf{x}_c\right)\right),$$

where now $\lambda_c^o \sim \mathsf{Gamma}(\alpha, \alpha)$, and $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are vectors of unknown parameters. Alternatively, the form of the functional shape g could be modelled via splines, similar

to the model appearing in (Morgan et al., 2019). In particular, choosing B splines with specific parameter constraints can ensure that the intensity function is always non-negative.

In either of the above suggestions, the resulting model would involve a larger number of hyperparameters. Importance sampling, as recommended in Chapter 3, might not be enough to deal with this larger parameter space (see, for example, Section 3.3 of Chopin, 2004). In fact, the Apogee to Apogee Path Sampler introduced in Chapter 5 could help with this by allowing for efficient inference with high robustness to tuning parameter misspecification.

Another extension would be the use of Bayesian nonparametric modelling in the form of a Cox process model. This, however, would not be trivial as even in the simplest of scenarios, exact Cox process inference is a very difficult problem due to the intractability of the likelihood term. The random-weight particle filter along with the pseudo-marginal MH algorithm, as outlined in Chapter 4, could be applied in this context, provided sufficient attention is paid to the problem-specific modifications. In Chapter 4, we focus on a setup involving data composed of a single time sequence. Extending methodology to be applicable in clinical trial recruitment modelling would itself be challenging. The first challenge would be the interval censoring of arrivals; in the data used for the analyses in Chapter 3, the censoring was on a daily scale, but it is not uncommon to for recruitment drives to only report weekly or even monthly counts. Provided the intensity would not change much over that period of time one could either apply small jitter to arrivals as a preprocessing step or estimate the inter-censored likelihood using additional auxiliary variables (see Section 5.3 of Gonçalves

et al., 2020). The second challenge would be the incorporation of random effects. It is possible that a discrete distribution, rather than the continuous gamma used in Chapter 3, for the random effects could allow for a more straightforward inference.

The pseudo-marginal scheme used in Chapter 4 relied on random walk proposals. One could apply the AAPS methodology (Chapter 5) to the pseudo-marginal inference carried out in Chapter 4. This would involve a modification of the existing pseudo-marginal Hamiltonian Monte Carlo (Alenlöv et al., 2021). Utilising the notion of apogees when defining the integration time could make for a more robust algorithm without a substantial loss to the peak performance.

Appendix A

Appendix for Chapter 3

A.1 Non-parametric bootstrapped test

4.1 Non-parametric test for detection of change in recruitment rates: _

1. **Input:** Series of counts $\{N_c(t)\}_{t=1}^{\tau_c}$, $c = 1, \ldots, C$; number of bootstrapped samples B
2. **Output:** Probability of observed difference in means, $\hat{p}$, under H_0
3. Calculate observed difference $\Delta = \sum_{c=1}^{C} \left(\sum_{t=1}^{\tau_c/2} N_c(t) - \sum_{t=\tau_c/2+1}^{\tau_c} N_c(t) \right)$
4. **for** $b = 1, \ldots, B$ **do**
   (a) **for** $C = 1, \ldots, C$ **do**
      i. Resample $\{N_c^{(b)}(t)\}_{t=1}^{\tau_c}$ with replacement;
      ii. Calculate difference $\Delta^{(b)} = \sum_{c=1}^{C} \left(\sum_{t=1}^{\tau_c/2} N_c^{(b)}(t) - \sum_{t=\tau_c/2+1}^{\tau_c} N_c^{(b)}(t) \right)$
   (b) Calculate approximate p-value: $\hat{p} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}_{\{\Delta \geq \Delta^{(b)}\}}$;
5. **return** $\hat{p}$;

A.2 Curve-shape

The integrated, normalised parametric intensities are:

$$\begin{aligned}
G_0(t) &= t, \\
G_1(t;\theta) &= \frac{\mathsf{log}(1+\theta t)}{\mathsf{log}(1+\theta\tau)}\tau, \\
G_\kappa(t;\theta) &= \frac{(1+\theta t/\kappa)^{1-\kappa}-1}{(1+\theta\tau/\kappa)^{1-\kappa}-1}\tau, \quad \kappa \notin \{0,1,\infty\}, \\
G_\infty(t;\theta) &= \frac{1-\mathsf{exp}(-\theta t)}{1-\mathsf{exp}(-\theta\tau)}\tau.
\end{aligned}$$

In two instances, the flexible-tail form can give rise to identifiability problems:

$\underline{t,\tau >> \kappa/\theta \text{ and } \kappa < 1}$

$$G_\kappa(t;\theta) = \frac{(1+\theta t/\kappa)^{1-\kappa}-1}{(1+\theta\tau/\kappa)^{1-\kappa}-1}\tau \approx \frac{(\theta t/\kappa)^{1-\kappa}-1}{(\theta\tau/\kappa)^{1-\kappa}-1}\tau \approx \left(\frac{t}{\tau}\right)^{1-\kappa}\tau,$$

which does not depend on θ.

$\underline{t,\tau >> \kappa/\theta \text{ and } \kappa >> 1}$

$$\begin{aligned}
G_\kappa(t;\theta) &= \frac{(1+\theta t/\kappa)^{1-\kappa}-1}{(1+\theta\tau/\kappa)^{1-\kappa}-1}\tau \\
&\approx \frac{(1+\theta t/\kappa)\mathsf{exp}(-\theta t)-1}{(1+\theta\tau/\kappa)\mathsf{exp}(-\theta\tau)-1}\tau \\
&\approx \frac{\mathsf{exp}(-\theta t)-1}{\mathsf{exp}(-\theta\tau)-1}\tau,
\end{aligned}$$

which does not depend on κ.

A.3 Maximum likelihood inference

In the frequentist setting, we aim to find estimators which maximise the likelihood surface (4.2) in the main paper. This is equivalent to maximising the log-likelihood surface (up to a constant)

$$\ell(\alpha, \phi, \theta; \mathbf{n}, \boldsymbol{\tau}) = C\left(\alpha \log \frac{\alpha}{\phi} - \log \Gamma(\alpha)\right) - \sum_{c=1}^{C}\left\{\left(\alpha + n_c^{(\cdot)}\right) \log\left(G(\tau_c;\theta) + \frac{\alpha}{\phi}\right) - \log \Gamma\left(\alpha + n_c^{(\cdot)}\right) - \sum_{t=1}^{\tau_c} n_c^{(t)} \log(G(t;\theta) - G(t-1;\theta))\right\}.$$

The log-likelihood function can be optimised using a range of methods, for example, the Nelder-Mead (Nelder and Mead, 1965) method used in `R`. The inverse of the negative Hessian at the mode can then be used as the covariance matrix for the asymptotic normal distribution of the MLEs.

The α and ϕ parameters are asymptotically orthogonal for a homogeneous Poisson-gamma model (Huzurbazar, 1950). A time contraction argument can be used to extend the result to the inhomogeneous case. As discussed in Section 4 of the main paper and visible from (4.3), in the special case where $\tau_c \equiv \tau \ \forall c$, θ is orthogonal to both α and ϕ. When carrying out maximum likelihood inference, different model selection criteria such as AIC (Akaike, 1973) and BIC (Schwarz, 1978) can be used. Alternatively, one could employ frequentist model averaging methods (see Hjort and Claeskens (2003), for instance).

The only pair of parameters which are not asymptotically orthogonal when centres have not been open for the same length of time are ϕ and θ.

Score and observed and expected information

Here we provide the score function and the observed and expected information, for frequentist inference.

The score function is the gradient of the log-likelihood of the model,

$$\nabla \ell(\alpha, \phi, \theta; \mathbf{n}, \boldsymbol{\tau}) =$$

$$= \begin{bmatrix} C\left(1 + \log \frac{\alpha}{\phi} - \psi(\alpha)\right) - \sum_{c=1}^{C}\left(\frac{\alpha + n_c^{(\cdot)}}{\alpha + \phi G(\tau_c;\theta)} + \log\left(G(\tau_c;\theta) + \frac{\alpha}{\phi}\right) - \psi\left(\alpha + n_c^{(\cdot)}\right)\right)\tau \\ -C\alpha/\phi + \sum_{c=1}^{C} \frac{\alpha\left(\alpha + n_c^{(\cdot)}\right)}{\phi(\alpha + \phi G(\tau_c;\theta))}\tau \\ -\sum_{c=1}^{C}\left[\partial_\theta G(\tau_c;\theta)\left(\frac{\alpha + n_c^{(\cdot)}}{G(\tau_c;\theta) + \frac{\alpha}{\phi}}\right) - \sum_{t=1}^{\tau_c} n_c^{(t)}\left(\frac{\partial_\theta G(t;\theta) - \partial_\theta G(t-1;\theta)}{G(t;\theta) - G(t-1;\theta)}\right)\right] \end{bmatrix}.$$

The observed information matrix is made up of the negative Hessian elements

$$\begin{aligned}
-\partial^2_{\alpha\alpha}\ell(\alpha, \phi, \theta; \mathbf{n}, \boldsymbol{\tau}) =& C\left(\psi'(\alpha) - \frac{1}{\alpha}\right) + \sum_{n=1}^{C}\left\{\frac{\phi G(\tau_c;\theta) - n_c^{(\cdot)} + 1}{\alpha + \phi G(\tau_c;\theta)} - \psi'\left(\alpha + n_c^{(\cdot)}\right)\right\}, \\
-\partial^2_{\phi\phi}\ell(\alpha, \phi, \theta; \mathbf{n}, \boldsymbol{\tau}) =& -C\alpha/\phi^2 + \sum_{c=1}^{C}\frac{\alpha\left(\alpha + 2\phi G(\tau_c;\theta)\right)\left(\alpha + n_c^{(\cdot)}\right)}{\phi^2(\alpha + \phi G(\tau_c;\theta))^2}, \\
-\partial^2_{\theta\theta}\ell(\alpha, \phi, \theta; \mathbf{n}, \boldsymbol{\tau}) =& \sum_{c=1}^{C}\left[\frac{\left(\alpha + n_c^{(\cdot)}\right)\left\{\partial^2_{\theta\theta}G(\tau_c;\theta)(G(\tau_c;\theta) + \alpha/\phi) - (\partial_\theta G(\tau_c;\theta))^2\right\}}{(G(\tau_c;\theta) + \alpha/\phi)^2}\right. \\
& \left. - \sum_{t=1}^{\tau_c} n_c^{(t)}\frac{H_t\partial^2_{\theta\theta}H_t - (\partial_\theta H_t)^2}{(H_t)^2}\right], \\
-\partial^2_{\alpha\phi}\ell(\alpha, \phi, \theta; \mathbf{n}, \boldsymbol{\tau}) =& \frac{1}{\phi}\left\{C - \sum_{c=1}^{C}\frac{\alpha^2 + 2\alpha\phi G(\tau_c;\theta) + \phi G(\tau_c;\theta)n_c^{(\cdot)}}{\left(\alpha + \phi G(\tau_c;\theta)\right)^2}\right\}, \\
-\partial^2_{\alpha\theta}\ell(\alpha, \phi, \theta; \mathbf{n}, \boldsymbol{\tau}) =& \sum_{c=1}^{C}\partial_\theta G(\tau_c;\theta)\frac{G(\tau_c;\theta) - n_c^{(\cdot)}/\phi}{\{G(\tau_c;\theta) - \alpha/\phi\}^2}, \\
-\partial^2_{\phi\theta}\ell(\alpha, \phi, \theta; \mathbf{n}, \boldsymbol{\tau}) =& -\alpha\sum_{c=1}^{C}\partial_\theta G(\tau_c;\theta)\frac{\alpha + n_c^{(\cdot)}}{\{\alpha + \phi G(\tau_c;\theta)\}^2},
\end{aligned}$$

where $\psi(x) = \Gamma'(x)/\Gamma(x)$ and $H_t = G(t;\theta) - G(t-1;\theta)$ to simplify the notation. Noting that $E\left[N_c^{(\cdot)}\right] = \phi G(\tau_c;\theta)$, we obtain the entries of the Fisher information

matrix,

$$\begin{aligned}
\mathbb{E}[-\partial^2_{\alpha\alpha}\ell(\alpha,\phi,\theta;\mathbf{N},\boldsymbol{\tau})] =& C\left(\psi'(\alpha)-\frac{1}{\alpha}\right)+\sum_{n=1}^{C}\left[\frac{1}{\alpha+\phi G(\tau_c;\theta)}-E\left\{\psi'\left(\alpha+n_c^{(\cdot)}\right)\right\}\right],\\
\mathbb{E}[-\partial^2_{\phi\phi}\ell(\alpha,\phi,\theta;\mathbf{N},\boldsymbol{\tau})] =& \frac{\alpha}{\phi}\sum_{c=1}^{C}\frac{G(\tau_c;\theta)}{\alpha+\phi G(\tau_c;\theta)},\\
\mathbb{E}[-\partial^2_{\theta\theta}\ell(\alpha,\phi,\theta;\mathbf{N},\boldsymbol{\tau})] =& \sum_{c=1}^{C}\left[\frac{\phi\left\{\partial^2_{\theta\theta}G(\tau_c;\theta)(G(\tau_c;\theta)+\alpha/\phi)-(\partial_\theta G(\tau_c;\theta))^2\right\}}{\phi G(\tau_c;\theta)+\alpha}\right.\\
&\left.-\sum_{t=1}^{\tau_c}n_c^{(t)}\partial^2_{\theta\theta}H_t-\frac{\left(\partial_\theta H_t\right)^2}{H_t}\right],\\
\mathbb{E}[-\partial^2_{\alpha\phi}\ell(\alpha,\phi,\theta;\mathbf{N},\boldsymbol{\tau})] =& 0,\\
\mathbb{E}[-\partial^2_{\alpha\theta}\ell(\alpha,\phi,\theta;\mathbf{N},\boldsymbol{\tau})] =& 0,\\
\mathbb{E}[-\partial^2_{\phi\theta}\ell(\alpha,\phi,\theta;\mathbf{N},\boldsymbol{\tau})] =& -\alpha\sum_{c=1}^{C}\frac{\partial_\theta G(\tau_c;\theta)}{\alpha+\phi G(\tau_c;\theta)}.
\end{aligned}$$

A.4 Curve-shape prior

The flexible form (4.4) in Section 4 of the main paper, leads to the following prior density for $\tilde{\theta}$,

$$\pi_0\left(\tilde{\theta}\,|\kappa,a,b\right)=\begin{cases}t_0\mathsf{exp}\left(\tilde{\theta}-t_0\mathsf{e}^{\tilde{\theta}}\right)f_{\mathsf{B}}\left(\mathsf{exp}\left(-t_0\mathsf{e}^{\tilde{\theta}}\right);a,b\right), & \kappa=\infty\\ t_0\mathsf{exp}(\tilde{\theta})\left(1+t_0\mathsf{e}^{\tilde{\theta}}/\kappa\right)^{-\kappa-1}f_{\mathsf{B}}\left(\left(1+t_0\mathsf{e}^{\tilde{\theta}}/\kappa\right)^{-\kappa};a,b\right), & \kappa\in(0,\infty)\end{cases},$$

where $f_{\mathsf{B}}(\,\cdot\,;a,b)$ is a density of a beta variate with shape parameters a and b.

A.5 Sampling time to completion via model averaging

In Lan et al. (2019), the time to recruit the required number of patients is sampled by repeatedly simulating the whole system until the condition is satisfied, which is inefficient because each iteration involves a (random) large number of expensive simulations. Additionally, it only provides an approximate distribution due to the discretisation in the time domain; the discretisation of recruitment to monthly increments might also affect the precision of any predictions. To sample the time to completion exactly, we use the integrated intensity function of the whole trial $\Lambda(t)$. If T is the time to the mth arrival of an inhomogeneous Poisson process with integrated intensity $\Lambda(t)$ then (Devroye, 1986)

$$\Lambda(T) \sim \mathsf{Gamma}(m, 1).$$

Given (α, β, θ), we can sample rates the λ_c^o for all the centres and construct one realisation of the integrated intensity Λ for the whole trial. Then, to obtain a single realisation of T, we sample a $\mathsf{Gamma}(m, 1)$ variate and use an inverse-transform of Λ on it. Unless all the centres had been open for the same length of time, the inversion procedure will involve some root-finding algorithm, such as Nelder-Mead (Nelder and Mead, 1965). As $\Lambda(t)$ in our framework is a monotonically increasing function, the non-linear equation will have a unique solution. Parameter uncertainty can be incorporated into this predictive by using a different sample from the posterior at each iteration.

Given C^+ centres with the first C already opened before the census time and the remaining $C^+ - C$ to be open, as well as known centre opening times $t_0^{(c)}$, $c = 1, \ldots, C^+$, we construct the integrated intensity for modelling the recruitment since the census time τ,

$$\Lambda(t) = \sum_{c=1}^{C} \lambda_c^o \left\{ G\left(t - t_0^{(c)}; \theta\right) - G(\tau_c; \theta) \right\} + \sum_{c=C+1}^{C^+} \lambda_c^o G\left(t - t_0^{(c)}; \theta\right) \mathbb{I}_{\left\{t > t_0^{(c)}\right\}}, \quad t \geq \tau,$$

where $\chi_{\{\cdot\}}$ is the indicator function and

$$\lambda_c^o | \alpha, \phi, \theta, \mathbf{n} \sim \begin{cases} \mathsf{Gamma}\left(\alpha + n_c^{(\cdot)}, \alpha/\phi + G(\tau_c; \theta)\right), & c = 1, \ldots, C \\ \mathsf{Gamma}\left(\alpha, \alpha/\phi\right), & c = C+1, \ldots, C^+ \end{cases} . \quad \text{(A.5.1)}$$

The algorithm below outlines the sampling procedure to obtain the distribution of the time needed to recruit the target number of patients m.

4.2 Model-averaged time-to-completion sampling:

1. **Input:** Models $\mathcal{M}_1, \ldots, \mathcal{M}_k$ with posterior probabilities $\pi(\mathcal{M}_1|\boldsymbol{n}), \ldots, \pi(\mathcal{M}_K|\boldsymbol{n})$ and posterior samples from each model, number of samples from the predictive B, target number of recruitments m
2. **Output:** Distribution of the time to completion $\left\{T^{(b)}\right\}_{b=1}^{B}$
3. **for** $b = 1, \ldots, B$ **do**
   (a) Sample $\mathcal{M}^{(b)} \sim \pi(\mathcal{M}_k|\boldsymbol{n})$;
   (b) Sample $(\alpha, \phi, \theta)^{(b)} \sim \pi(\cdot\,|\mathcal{M}^{(b)}, \boldsymbol{n})$;
   (c) Sample rates $\lambda_c^o|(\alpha, \phi, \theta)^{(b)}$ from distributions ((A.5.1)) and construct $\Lambda^{(b)}(t)$;
   (d) Sample $\tilde{T} \sim \mathsf{Gamma}(m, 1)$ and solve $\Lambda^{(b)}(T) = \tilde{T}$;
   (e) Set $T^{(b)} = T$;
4. **return** $\left\{T^{(b)}\right\}_{b=1}^{B}$;

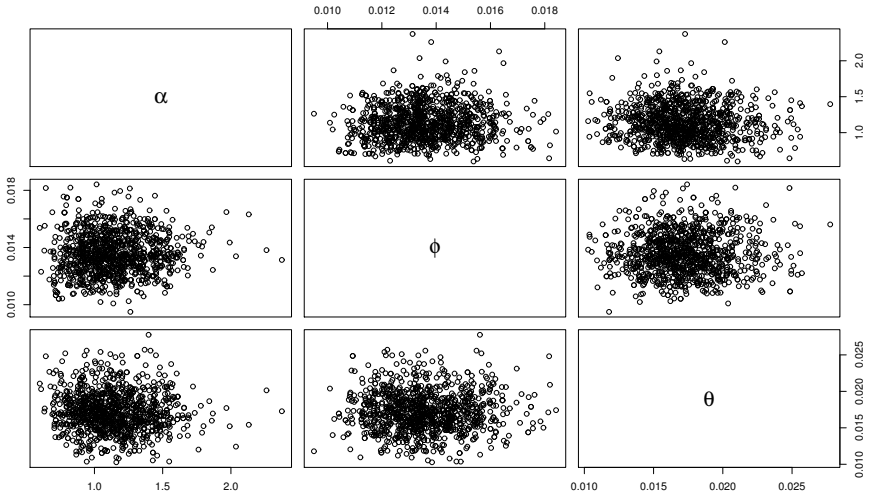


Figure A.6.1: Matrix scatterplot of the parameter posterior of the model with highest posterior probability.

A.6 Additional details from the simulation study and data analysis

A.6.1 Simulation study

Figure A.6.1 shows the plots of posterior samples of the model. The three parameters are close to orthogonal as discussed in Sections 4 and 5 of the paper, and this approximate independence was also observed in the posteriors of other models.

Figure A.6.2 shows a QQ-plot of the hierarchical gamma distribution compared to the posterior means of the random effects. The approximately straight line indicates that generating rates for newly opened centres from the gamma distribution will be consistent with what has been observed thus far. Figure A.6.3 shows a QQ-plot of the theoretical, negative binomial distribution of recruitments in the first 2 months compared the observed distribution ($t_* = 60$). The theoretical distribution used the

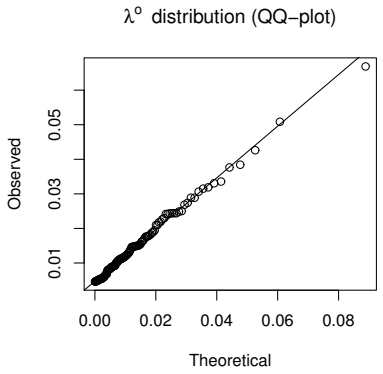


Figure A.6.2: Re-estimated λ_c^o expectations compared to $\mathsf{Gamma}\left(\hat{\alpha}, \hat{\alpha}/\hat{\phi}\right)$ distribution.

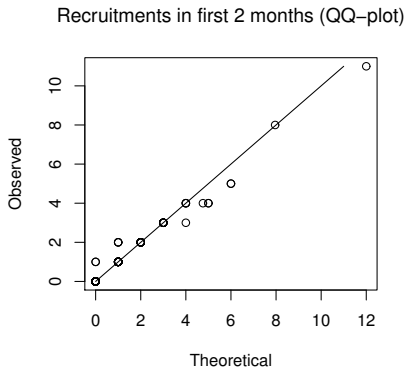


Figure A.6.3: Observed recruitments compared to the theoretical negative binomial distribution.

posterior means of the parameters, and the prior random effect distribution was used. The straight line shows that the model can predict the recruitment in the first two months of a centre sufficiently well. In practice, the two diagnostics would indicate that the mixing gamma distribution is sufficient and that the model is capable of accurately predicting recruitments in the early days of a new centre. Figures A.6.4, A.6.5, A.6.6 and A.6.7 show the diagnostic plots for models fit to simulated datasets at the census $t = 360$ with the true random-effect distribution being a mixture. For $\mathbb{E}[\lambda_c^o] = 0.01$, the relationship is close to linear and is reflected in the reasonably accurate predictions shown in the article. The QQ-plots for $\mathbb{E}[\lambda_c^o] = 0.03$ show stronger non-linearity and informing us of the potential misspecification, thus showing that the diagnostics can be used to validate the model.

Figures A.6.8a and A.6.8b show examples of recruitment predictions when the random effects have a mixture distribution and the centre opening times are "clumped" together. The clumping accentuates the effect of the misspecification; the fitted model

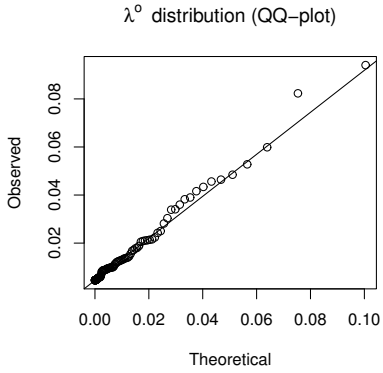


Figure A.6.4: Re-estimated λ_c^o expectations compared to $\mathsf{Gamma}\left(\hat{\alpha}, \hat{\alpha}/\hat{\phi}\right)$ distribution; true random-effect distribution is a mixture with $\mathbb{E}[\lambda_c^o] = 0.01$.

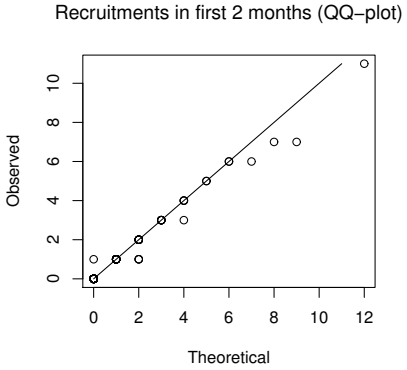


Figure A.6.5: Observed recruitments compared to the theoretical negative binomial distribution; true random-effect distribution is a mixture with $\mathbb{E}[\lambda_c^o] = 0.01$

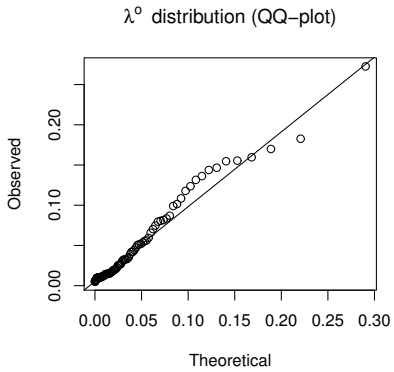


Figure A.6.6: Re-estimated λ_c^o expectations compared to $\mathsf{Gamma}\left(\hat{\alpha}, \hat{\alpha}/\hat{\phi}\right)$ distribution; true random-effect distribution is a mixture with $\mathbb{E}[\lambda_c^o] = 0.03$.

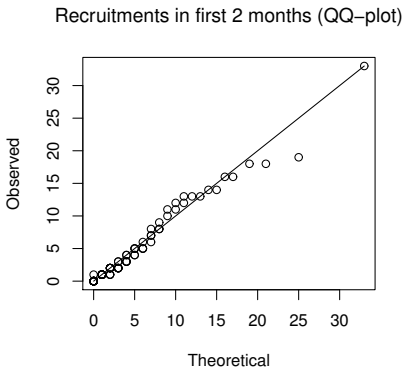
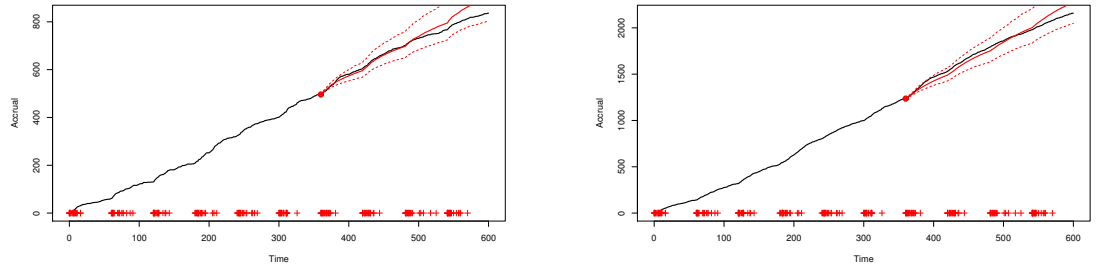


Figure A.6.7: Observed recruitments compared to the theoretical negative binomial distribution; true random-effect distribution is a mixture with $\mathbb{E}[\lambda_c^o] = 0.03$



(a) Clumped openings, $\mathbb{E}[\lambda_c^o] = 0.01$; p-value $= 0.666$

(b) Clumped openings, $\mathbb{E}[\lambda_c^o] = 0.03$; p-value $= 0.312$

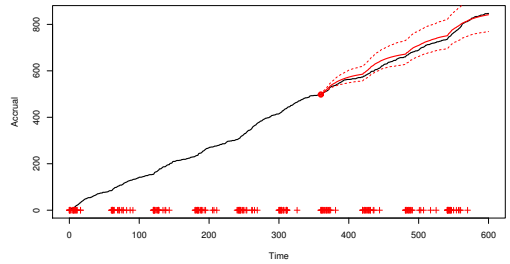
Figure A.6.8: Accruals (black, solid) with predictive means (red, solid) and 95% prediciton bands (red, dashed) when the true random-effect distribution is a mixture, in various scenarios. Prediction bands are based on the 2.5% and 97.5% quantiles. The "+" symbols on the x-axis indicate centre opening times.

relies on the "incorrect" prior gamma distribution when simulating rates for unopened centres.

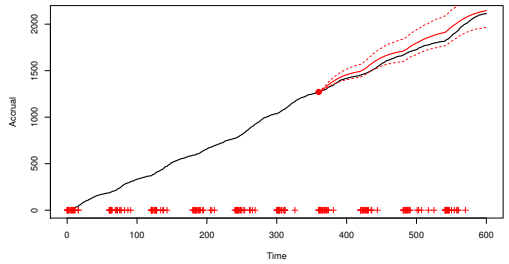
Figures A.6.9a and A.6.9b show the predictions made when using data simulated from a Weibull-shape intensity with centre opening times clumped together. With repeated simulations, we found a consistent correspondence between linear QQ-plots and accurate predictions.

A.6.2 Data analysis

In the dataset examined in Section 7 of the main paper, we encountered an unexpected surge in recruitments at a global scale. Figure A.6.10 shows the accrual along with 2 sets of forecasts, focusing on the surge at a time of around 0.7. Once this has been observed, and forward predictions are needed, one possibility is modelling this as a global surge in recruitment; that is, during the period between 0.6 and 0.75 all recruitment rates are multiplied by $\mathsf{exp}(\beta)$ for some unknown β, which would be an



(a) Clumped openings, $\mathbb{E}[\lambda_c^o] = 0.01$; p-value $= 0.294$



(b) Clumped openings, $\mathbb{E}[\lambda_c^o] = 0.03$; p-value $= 0.064$

Figure A.6.9: Accruals (black, solid) with predictive means (red, solid) and 95% prediciton bands (red, dashed) when the true intensity shape is Weibull and opening times are clumped, with two different values for $\mathbb{E}[\lambda_c^o]$. Prediction bands are based on the 2.5% and 97.5% quantiles. The "+" symbols on the x-axis indicate centre opening times.

extra parameter to be estimated via importance sampling.

A.7 Stochastic centre-initiation times

The framework, as presented in the main paper, is conditioned on the set of initiation times both for clarity of presentation and because it is the methodological contribution from the paper. In practice, the exact future initiation times would be unknown; instead, the practitioners would have proposed initiation schedules, contingency plans and recruitment data up to the census time. Here we present a simulation study similar to that in Section 6 of the main paper which illustrates how a stochastic centre-initiation model can be seamlessly incorporated. The centres are not initiated exactly on schedule but, instead, there is a Weibull-distributed initiation delay for each centre. Following information provided to us from a large meta-analysis, we set the Weibull parameters such that the 5th and 95th percentiles are 10 and 322 days respectively; the

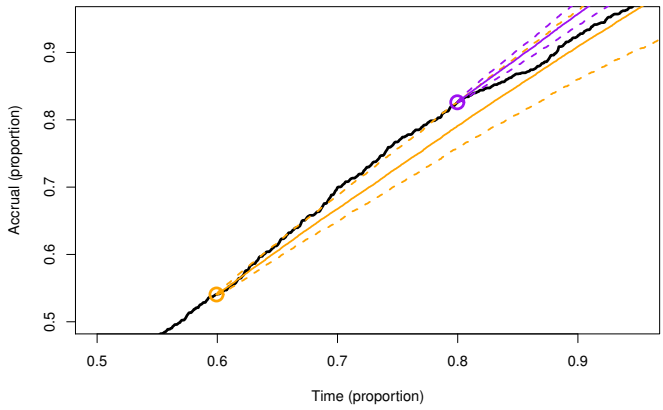


Figure A.6.10: Accrual predictions, zoomed-in to focus on the unexpected surge in recruitment at around the time of 0.7. Only interim forecasts from times 0.6 and 0.8 are shown.

median delay is 90 days. At the census, the observed day-censored delays are used for maximum-likelihood estimation of the Weibull parameters; additionally centres which were planned to initiate before the census but did not do so contribute with a censored likelihood. The estimates are then used in the Monte Carlo simulations. Figure A.7.1 compare the predictions under three different approaches; (i) the correct Weibull distribution for delays (with parameters estimated from the data), (ii) a constant, avergae delay taken to be the sample mean of the observed delays, and (iii) an assumption of no delays. It is clear that assuming no future delays given historical evidence of the contrary leads to poor forecasts. However, even very simple delay predictions based on the empirical average can achieve desirable forecasts. Of course, fitting the true model results in predictions which capture the truth extremely well. This illustrates that our site-level prediction method can be easily combined with site-initiation models.

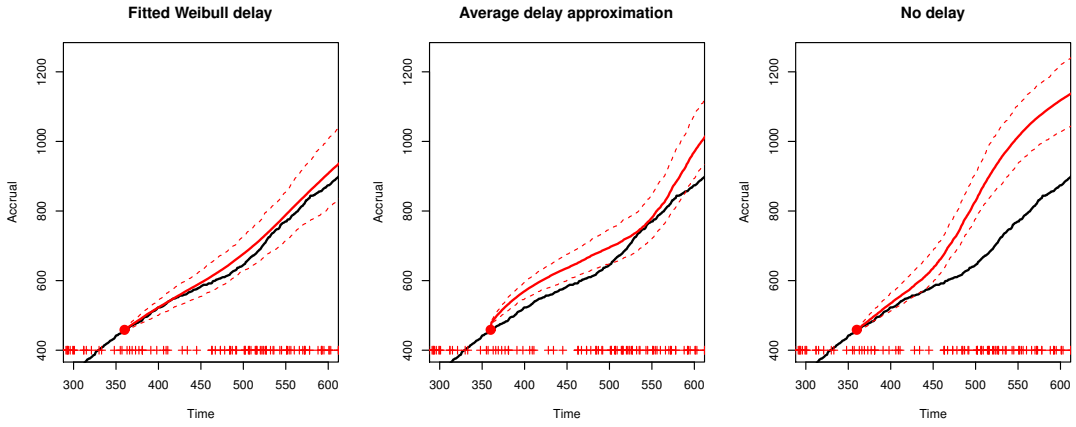


Figure A.7.1: Comparison of predictions for recruitment data with stochastically-delayed centre-initiation times. Three modelling approaches are considered: correct Weibull-distributed delay fitted (left); constant, historical average delay added to each initiation time (centre); and no delay considered in predictions (right).

Appendix B

Appendix for Chapter 4

B.1 Expectation and variance of a truncated Poisson random variable

Let Y be a $(k-1)$-truncated $\mathsf{Pois}(\mu)$ random variable, $Y \sim \mathsf{TruncPois}(\mu; k)$, with a mass function

$$\mathbb{P}(Y = y) = \frac{\mu^y \mathsf{e}^{-\mu}}{S_{\mathsf{P}}(k-1;\mu) y!}, \quad y = k, k+1, \ldots, \tag{B.1.1}$$

where $S_{\mathsf{P}}(\,\cdot\,;\mu)$ is the survival function of a standard Poisson random variable with expectation μ. The expectation is

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{y=k}^{\infty} \frac{y\mu^y \mathsf{e}^{-\mu}}{S_{\mathsf{P}}(k-1;\mu) y!} \\
&= \frac{\mu}{S_{\mathsf{P}}(k-1;\mu)} \sum_{y=k}^{\infty} \frac{\mu^{y-1} \mathsf{e}^{-\mu}}{(y-1)!} \\
&= \mu \frac{S_{\mathsf{P}}(k-2;\mu)}{S_{\mathsf{P}}(k-1;\mu)}
\end{aligned}$$

Since $S_{\mathsf{P}}(k;\mu)$ is non-increasing in k for any μ, $\mathbb{E}[Y] \geq \mu$. To obtain the variance,

we use the fact that $\mathbb{V}[Y] = \mathbb{E}[Y(Y-1)] - \mathbb{E}[Y]^2 + \mathbb{E}[Y]$. The second factorial moment is

$$\begin{aligned}
\mathbb{E}\left[Y(Y-1)\right] &= \sum_{y=0}^{\infty} y(y-1)\mathbb{P}\left(Y=y\right) \\
&= \sum_{y=2}^{\infty} y(y-1)\mathbb{P}\left(Y=y\right) \\
&= \sum_{y=\mathsf{max}\{k,2\}}^{\infty} \frac{y(y-1)\mu^y \mathsf{e}^{-\mu}}{S_{\mathsf{P}}(k-1;\mu)y!} \\
&= \frac{\mu^2}{S_{\mathsf{P}}(k-1;\mu)} \sum_{y=\mathsf{max}\{k,2\}}^{\infty} \frac{\mu^{y-2}\mathsf{e}^{-\mu}}{(y-2)!} \\
&= \mu^2 \frac{S_{\mathsf{P}}(k-3;\mu)}{S_{\mathsf{P}}(k-1;\mu)},
\end{aligned}$$

which gives the variance

$$\mathbb{V}[Y] = \mu^2 \frac{S_{\mathsf{P}}(k-3;\mu)}{S_{\mathsf{P}}(k-1;\mu)} - \mu \frac{S_{\mathsf{P}}(k-2;\mu)}{S_{\mathsf{P}}(k-1;\mu)} \left(\mu \frac{S_{\mathsf{P}}(k-2;\mu)}{S_{\mathsf{P}}(k-1;\mu)} - 1 \right).$$

Due to the presence of the survival functions, the variance is not easy to manipulate. Figure B.1.1 shows the variance for different choices of k and μ. The plot suggests that for a given μ the variance seems to decrease as we increase the truncation $k-1$.

The estimator introduced in Section 4.4 involves the quantity $\binom{Y}{k}$. The first two

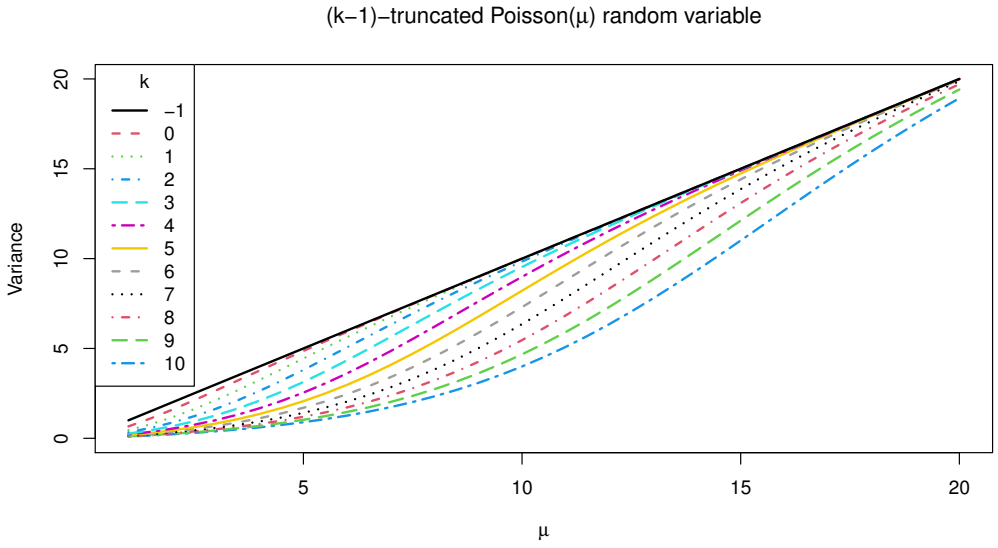


Figure B.1.1: Variance of a truncated Poisson random variable; $k = -1$ corresponds to the standard untruncated variate.

moments are

$$\begin{aligned}
\mathbb{E}\left[\binom{Y}{k}\right] &= \mathbb{E}\left[\frac{Y!}{(Y-k)!k!}\right] \\
&= \frac{1}{S_{\mathsf{P}}(k-1;\mu)} \sum_{y=k}^{\infty} \frac{y!}{(y-k)!k!} \cdot \frac{\mathsf{e}^{-\mu}(\mu)^y}{y!} \\
&= \frac{\mathsf{e}^{-\mu}(\mu)^k}{S_{\mathsf{P}}(k-1;\mu)k!} \sum_{y=k}^{\infty} \frac{(\mu)^{y-k}}{(r-k)!} \\
&= \frac{(\mu)^k}{S_{\mathsf{P}}(k-1;\mu)k!},
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[\binom{Y}{k}^2\right] &= \mathbb{E}\left[\left(\frac{Y!}{(Y-k)!k!}\right)^2\right] \\
&= \frac{1}{S_{\mathsf{P}}(k-1;\lambda_0\mathcal{T})}\sum_{y=k}^{\infty}\frac{(y!)^2}{((y-k)!)^2(y!)^2}\cdot\frac{\mathsf{e}^{-\mu}\mu^y}{y!} \\
&= \frac{\mathsf{e}^{-\mu}}{S_{\mathsf{P}}(k-1;\mu)(n!)^2}\sum_{y=k}^{\infty}\frac{\mu^y y!}{((y-k)!)^2} \\
&= \frac{\mathsf{e}^{-\mu}}{S_{\mathsf{P}}(k-1;\mu)(k!)^2}\cdot k!\mu^k{}_1\mathcal{F}_1(k+1;1;\mu) \\
&= \mathbb{E}\left[\binom{Y}{k}\right]\mathsf{e}^{-\mu}{}_1\mathcal{F}_1(k+1;1;\mu) < \infty,
\end{aligned}$$

where ${}_1\mathcal{F}_1$ is the Kummer confluent hypergeometric function (Kummer, 1837). As a result, we have the variance

$$\mathbb{V}\left[\binom{Y}{k}\right] = \mathbb{E}\left[\binom{Y}{k}\right]\left(\mathsf{e}^{-\mu}{}_1\mathcal{F}_1(k+1;1;\mu) - \mathbb{E}\left[\binom{Y}{k}\right]\right).$$

B.2 Proof of Proposition 4.2.1

Proof. To solve the SDE (4.2.3), we first let G_t be the fundamental matrix of the deterministic ordinary differential equation $\frac{\mathsf{d}\boldsymbol{x}}{\mathsf{d}t} = -A\boldsymbol{x}$, that is $\frac{\mathsf{d}G_t}{\mathsf{d}t} = -AG_t$ and $G_0 = I$. We note that $\frac{\mathsf{d}G_t^{-1}}{\mathsf{d}t} = G_t^{-1}A$. Here, the fundamental matrix is $G_t = \mathsf{exp}(-At)$ with its inverse $G_t^{-1} = \mathsf{exp}(At)$. We use a substitution $\boldsymbol{Y}_t = G_t^{-1}\boldsymbol{X}_t$ and note that

$\boldsymbol{Y}_0 = \boldsymbol{X}_0 \sim \mathsf{N}(\boldsymbol{\mu}_0, \Sigma_0)$. After this linear transformation the SDE becomes

$$\begin{aligned}
\mathsf{d}Y_t &= \mathsf{d}(G_t^{-1}\boldsymbol{X}_t) \\
&= (\mathsf{d}G_t^{-1})\boldsymbol{X}_t + G_t^{-1}\,\mathsf{d}\boldsymbol{X}_t \\
&= G_t^{-1}A\boldsymbol{X}_t\,\mathsf{d}t - G_t^{-1}A\boldsymbol{X}_t\,\mathsf{d}t + G_t^{-1}h\,\mathsf{d}W_t \\
&= G_t^{-1}h\,\mathsf{d}W_t.
\end{aligned}$$

Hence, $\boldsymbol{Y}_t - \boldsymbol{Y}_0 = \int_0^t G_s^{-1}h\,\mathsf{d}\boldsymbol{W}_s$. By linearity, $\boldsymbol{Y}_t - \boldsymbol{Y}_0$ has a Gaussian distribution. The expectation at time t is $\mathbf{0}$, and the variance can be obtained via Itô isometry. We find that

$$\boldsymbol{Y}_t - \boldsymbol{Y}_0 \sim \mathsf{N}(\mathbf{0}, \Psi_t), \text{ where } \Psi_t = \int_0^t G_s^{-1}hh^\top (G_s^{-1})^\top\,\mathsf{d}s,$$

and thus $\boldsymbol{Y}_t$ has a Gaussian distribution. As $\boldsymbol{X}_t = G_t\boldsymbol{Y}_t$ is a linear transformation,

$$\boldsymbol{X}_t \sim \mathsf{N}\left(G_t\boldsymbol{\mu}_0, G_t\left(\Psi_t + \Sigma_0\right)G_t^\top\right).$$

□

We now assume the setting studied in the main body,

$$A = \begin{bmatrix} \theta_1 & -1 \\ 0 & \theta_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{h} = \begin{bmatrix} 0 \\ \sigma \end{bmatrix}.$$

Using spectral decomposition it can be shown that

$$A^k = \begin{bmatrix} \theta_1^k & \frac{\theta_2^k - \theta_1^k}{\theta_2 - \theta_1} \\ 0 & \theta_2^k \end{bmatrix},$$

and hence we obtain an analytical expression for the exponential,

$$\begin{aligned}
\mathsf{exp}(At) &= \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k \\
&= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \sum_{k=1}^{\infty} \frac{t^k}{k!} \begin{bmatrix} \theta_1^k & \frac{\theta_2^k - \theta_1^k}{\theta_2 - \theta_1} \\ 0 & \theta_2^k \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} \sum_{k=1}^{\infty} \frac{t^k}{k!} \theta_1^k & \frac{1}{(\theta_2 - \theta_1)} \sum_{k=1}^{\infty} \frac{t^k (\theta_2^k - \theta_1^k)}{k!} \\ 0 & \sum_{k=1}^{\infty} \frac{t^k}{k!} \theta_2^k \end{bmatrix} \\
&= \begin{bmatrix} \mathsf{e}^{\theta_1 t} & \frac{\mathsf{e}^{\theta_2 t} - \mathsf{e}^{\theta_1 t}}{\theta_2 - \theta_1} \\ 0 & \mathsf{e}^{\theta_2 t} \end{bmatrix}.
\end{aligned}$$

This gives

$$\begin{aligned}
G_t^{-1} \boldsymbol{h} &= \mathsf{e}^{At} \boldsymbol{h} \\
&= \begin{bmatrix} \mathsf{e}^{\theta_1 t} & \frac{\mathsf{e}^{\theta_2 t} - \mathsf{e}^{\theta_1 t}}{\theta_2 - \theta_1} \\ 0 & \mathsf{e}^{\theta_2 t} \end{bmatrix} \begin{bmatrix} 0 \\ \sigma \end{bmatrix} \\
&= \begin{bmatrix} \sigma \left(\frac{\mathsf{e}^{\theta_2 t} - \mathsf{e}^{\theta_1 t}}{\theta_2 - \theta_1} \right) \\ \sigma \mathsf{e}^{\theta_2 t} \end{bmatrix},
\end{aligned}$$

and

$$\begin{aligned}
G_t^{-1} \boldsymbol{h} \boldsymbol{h}^\top (G_t^{-1})^\top &= \mathsf{e}^{At} h h^\top \mathsf{e}^{A^\top t} \\
&= \begin{bmatrix} \sigma \left(\frac{\mathsf{e}^{\theta_2 t} - \mathsf{e}^{-\theta_1 t}}{\theta_2 - \theta_1} \right) \\ \sigma \mathsf{e}^{\theta_2 t} \end{bmatrix} \begin{bmatrix} \sigma \left(\frac{\mathsf{e}^{\theta_2 t} - \mathsf{e}^{\theta_1 t}}{\theta_2 - \theta_1} \right) & \sigma \mathsf{e}^{\theta_2 t} \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2 \left(\frac{\mathsf{e}^{\theta_2 t} - \mathsf{e}^{\theta_1 t}}{\theta_2 - \theta_1} \right)^2 & \sigma^2 \mathsf{e}^{\theta_2 t} \left(\frac{\mathsf{e}^{\theta_2 t} - \mathsf{e}^{\theta_1 t}}{\theta_2 - \theta_1} \right) \\ \sigma^2 \mathsf{e}^{\theta_2 t} \left(\frac{\mathsf{e}^{\theta_2 t} - \mathsf{e}^{\theta_1 t}}{\theta_2 - \theta_1} \right) & \sigma^2 \mathsf{e}^{2\theta_2 t} \end{bmatrix}.
\end{aligned}$$

Integrating the expression we arrive at the required covariance matrix,

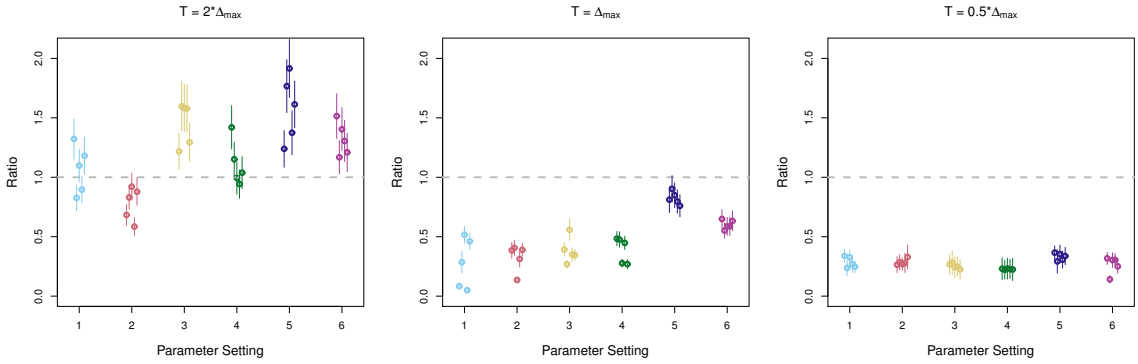
$$\Psi_t = \begin{bmatrix} \left(\frac{\sigma}{\theta_2-\theta_1}\right)^2 \left(\frac{\mathrm{e}^{2\theta_2 t}}{2\theta_2} + \frac{\mathrm{e}^{2\theta_1 t}}{2\theta_1} - \frac{2\mathrm{e}^{(\theta_1+\theta_2)t}}{\theta_1+\theta_2} - \frac{\theta_1^2-2\theta_1\theta_2+\theta_2^2}{2\theta_1\theta_2(\theta_2+\theta_1)}\right) & \frac{\sigma^2}{\theta_2-\theta_1}\left(\frac{1}{\theta_1+\theta_2} - \frac{1}{2\theta_2} - \frac{\mathrm{e}^{(\theta_1+\theta_2)t}}{(\theta_1+\theta_2)} + \frac{\mathrm{e}^{2\theta_2 t}}{2\theta_2}\right) \\ \frac{\sigma^2}{\theta_2-\theta_1}\left(\frac{1}{\theta_1+\theta_2} - \frac{1}{2\theta_2} - \frac{\mathrm{e}^{(\theta_1+\theta_2)t}}{(\theta_1+\theta_2)} + \frac{\mathrm{e}^{2\theta_2 t}}{2\theta_2}\right) & \frac{\sigma^2}{2\theta_2}\left(\mathrm{e}^{2\theta_2 t} - 1\right) \end{bmatrix}.$$

B.2.1 Additional results for effect of merging adjacent subintervals

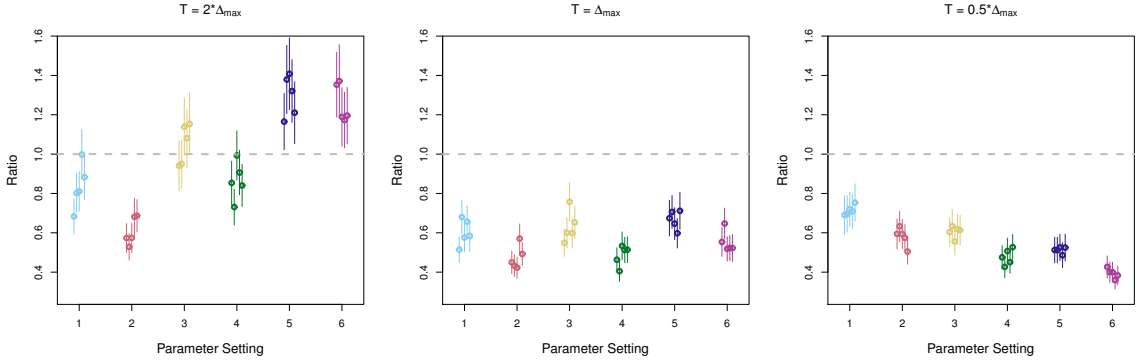
Figures B.2.1 - B.2.5 provide additional results for the simulation study carried out in Section 4.4.

B.2.2 MCMC diagnostics

Figures B.2.6 and B.2.7 provide the traceplots and autocorrelation function estimates of the Markov chains in the real-world data examples in Section 4.5; coal mining disasters and USD-EUR exchange rates. The burn-in was appropriately discarded beforehand.

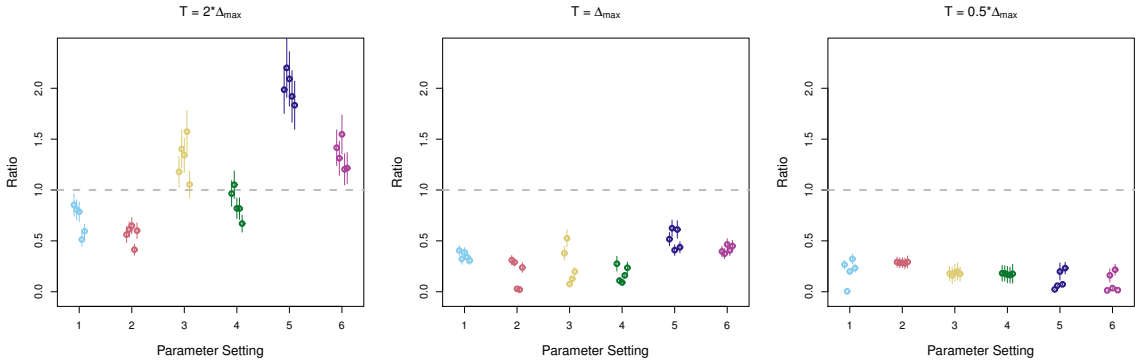


(a) $n_1 = 0,\ n_2 = 2$

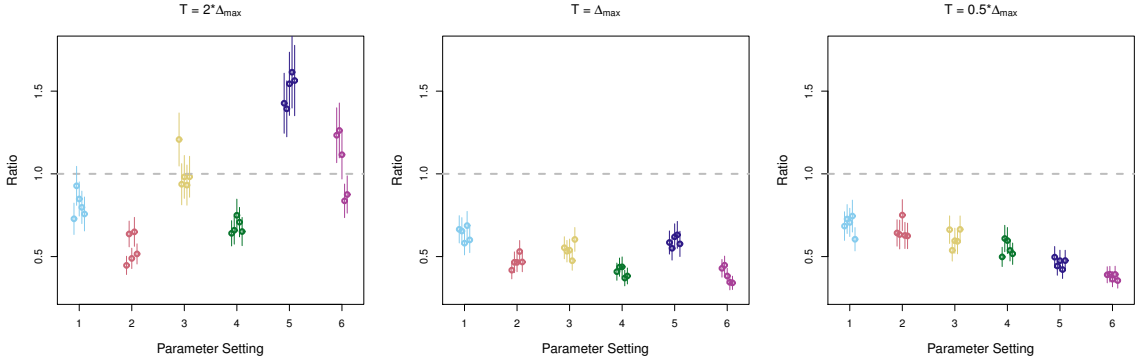


(b) $n_1 = 2,\ n_2 = 0$

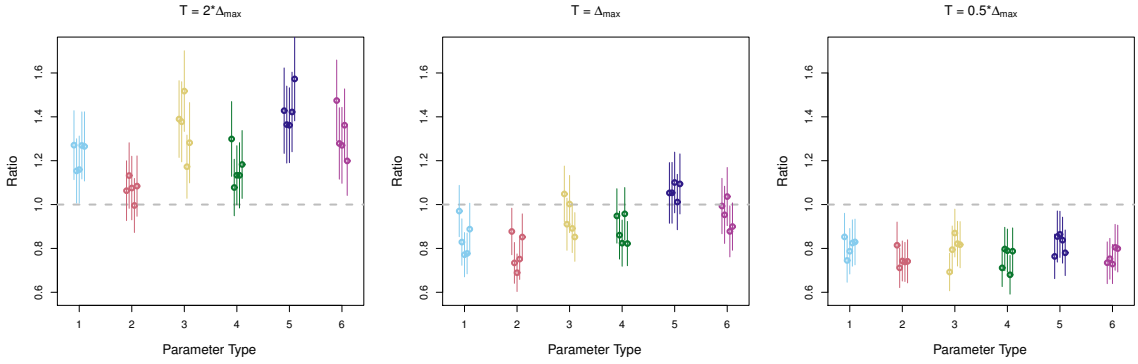
Figure B.2.1: Effect of merging adjacent equal-width subintervals on the variance contribution to the pseudo-marginal log-likelihood estimate $\mathsf{log}\,\widehat{L}_{\mathsf{PM}}$. The ratio represents the relative variance of estimates produced by a random-weight particle filter run on merged subintervals compared to a filter on split subintervals. Respective Monte Carlo means are denoted by "∘", and "ǀ" is used to represent intervals within 2 standard errors from the mean. Subinterval widths are denoted by Δ with $\Delta_{\mathsf{max}} = \mathsf{inf}\{s : \mathsf{Corr}\,[X_{1,t}, X_{1,t+s}] < 0.5\}$ and parameter settings are given in Table 4.4.1.



(a) $n_1 = 0,\ n_2 = 3$

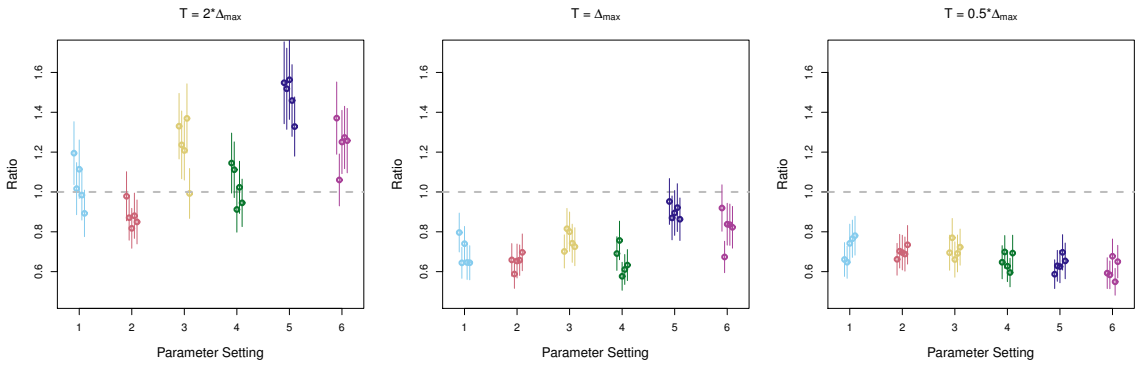


(b) $n_1 = 3,\ n_2 = 0$

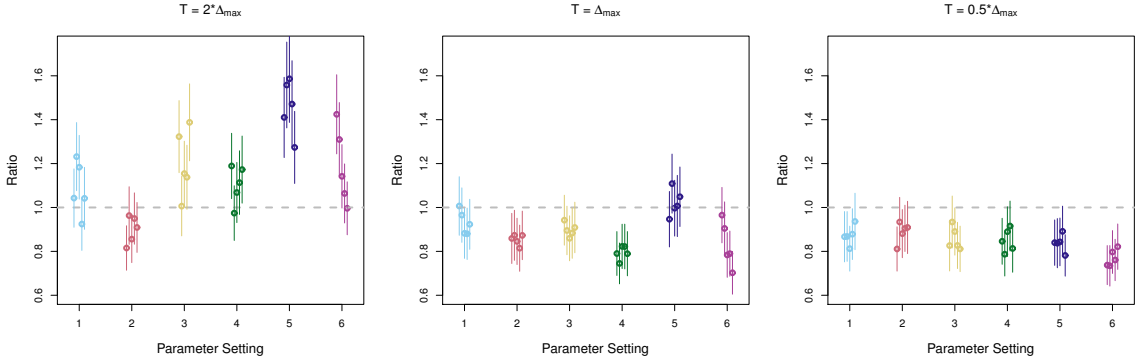


(c) $n_1 = 1,\ n_2 = 1$

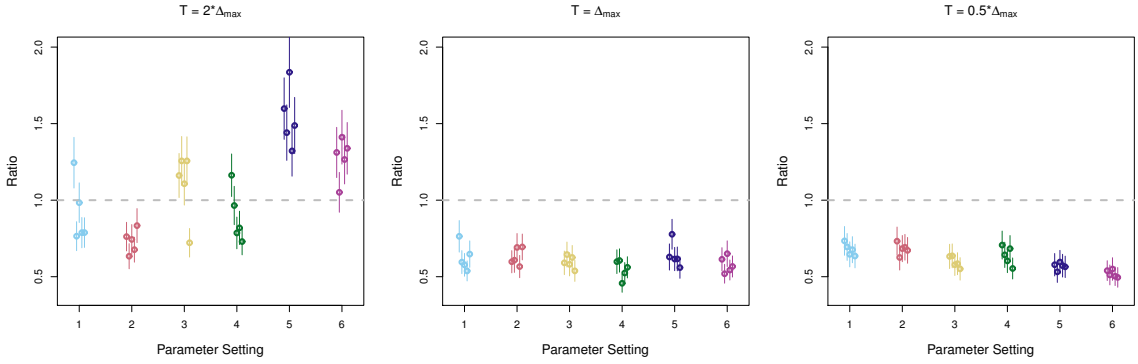
Figure B.2.2: Effect of merging adjacent equal-width subintervals on the variance contribution to the pseudo marginal log-likelihood estimate $\mathsf{log}\,\widehat{L}_{\mathsf{PM}}$. The ratio represents the relative variance of estimates produced by a random-weight particle filter run on merged subintervals compared to a filter on split subintervals. Respective Monte Carlo means are denoted by "○", and "|" is used to represent intervals within 2 standard errors from the mean. Subinterval widths are denoted by Δ with $\Delta_{\mathsf{max}} = \mathsf{inf}\{s : \mathsf{Corr}\,[X_{1,t}, X_{1,t+s}] < 0.5\}$ and parameter settings are given in Table 4.4.1.



(a) $n_1 = 1,\ n_2 = 2$

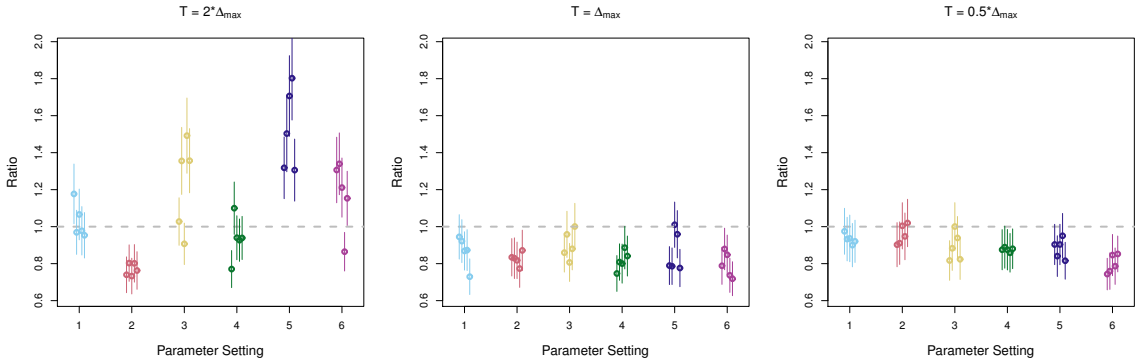


(b) $n_1 = 2,\ n_2 = 1$

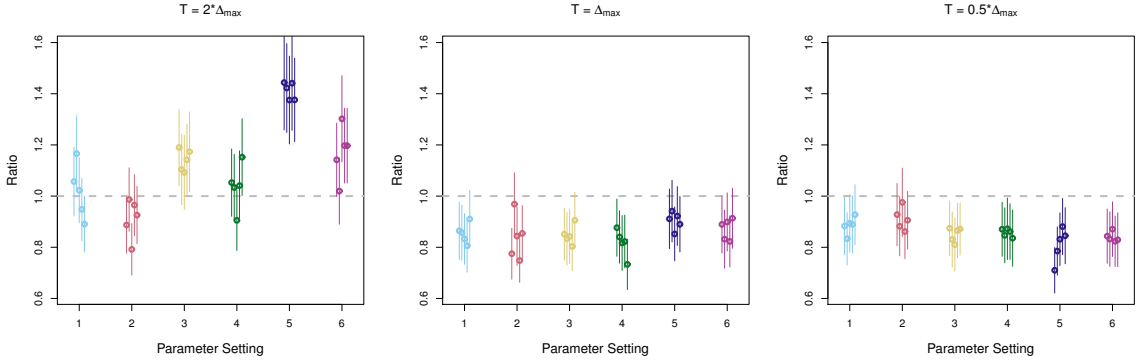


(c) $n_1 = 1,\ n_2 = 3$

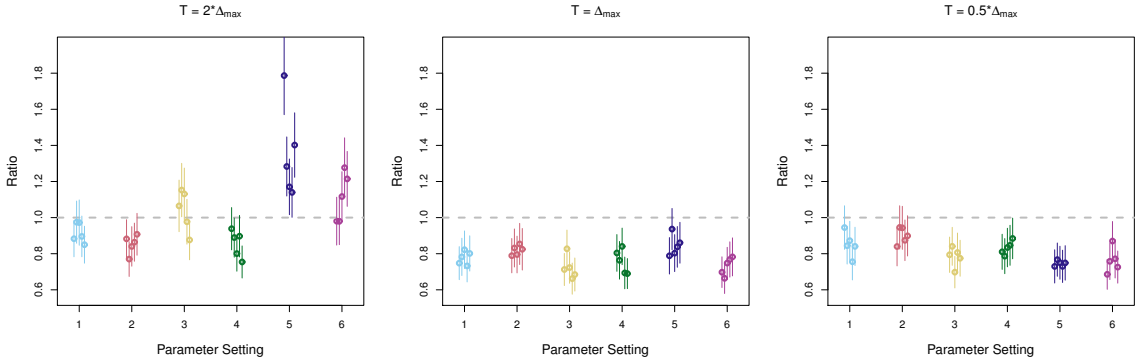
Figure B.2.3: Effect of merging adjacent equal-width subintervals on the variance contribution to the pseudo-marginal log-likelihood estimate $\log \widehat{L}_{\mathsf{PM}}$. The ratio represents the relative variance of estimates produced by a random-weight particle filter run on merged subintervalsc compared to a filter on split subintervals. Respective Monte Carlo means are denoted by "○", and "|" is used to represent intervals within 2 standard errors from the mean. Subinterval widths are denoted by Δ with $\Delta_{\mathsf{max}} = \inf\{s : \mathsf{Corr}\,[X_{1,t}, X_{1,t+s}] < 0.5\}$ and parameter settings are given in Table 4.4.1.



(a) $n_1 = 3,\ n_2 = 1$

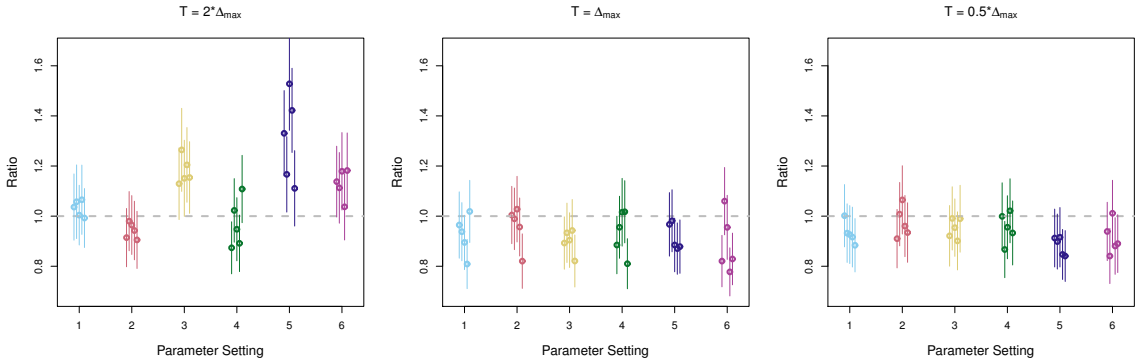


(b) $n_1 = 2,\ n_2 = 2$

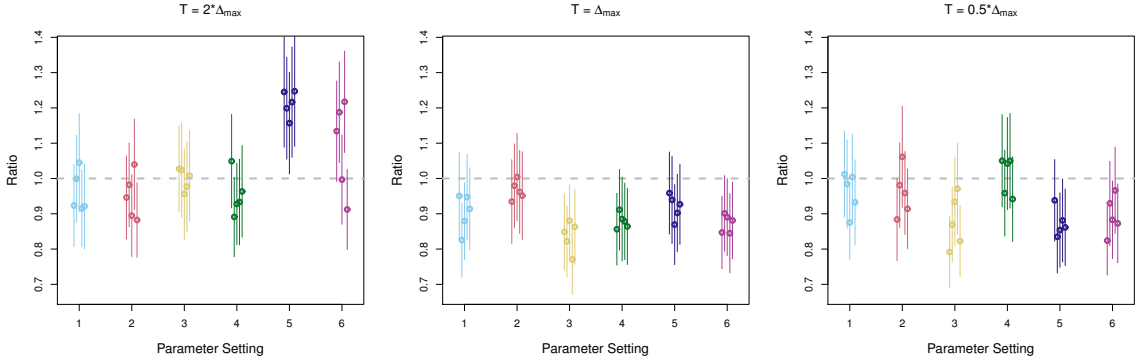


(c) $n_1 = 2,\ n_2 = 3$

Figure B.2.4: Effect of merging adjacent equal-width subintervals on the variance contribution to the pseudo-marginal log-likelihood estimate $\log \widehat{L}_{\mathsf{PM}}$. The ratio represents the relative variance of estimates produced by a random-weight particle filter run on merged subintervals compared to a filter on split subintervals. Respective Monte Carlo means are denoted by "○", and "|" is used to represent intervals within 2 standard errors from the mean. Subinterval widths are denoted by Δ with $\Delta_{\mathsf{max}} = \inf\{s : \mathsf{Corr}\left[X_{1,t}, X_{1,t+s}\right] < 0.5\}$ and parameter settings are given in Table 4.4.1.



(a) $n_1 = 3,\ n_2 = 2$



(b) $n_1 = 3,\ n_2 = 3$

Figure B.2.5: Effect of merging adjacent equal-width subintervals on the variance contribution to the pseudo-marginal log-likelihood estimate $\log \widehat{L}_{\mathsf{PM}}$. The ratio represents the relative variance of estimates produced by a random-weight particle filter run on merged subintervals compared to a filter on split subintervals. Respective Monte Carlo means are denoted by "∘", and "|" is used to represent intervals within 2 standard errors from the mean. Subinterval widths are denoted by Δ with $\Delta_{\mathsf{max}} = \mathsf{inf}\{s : \mathsf{Corr}\left[X_{1,t}, X_{1,t+s}\right] < 0.5\}$ and parameter settings are given in Table 4.4.1.

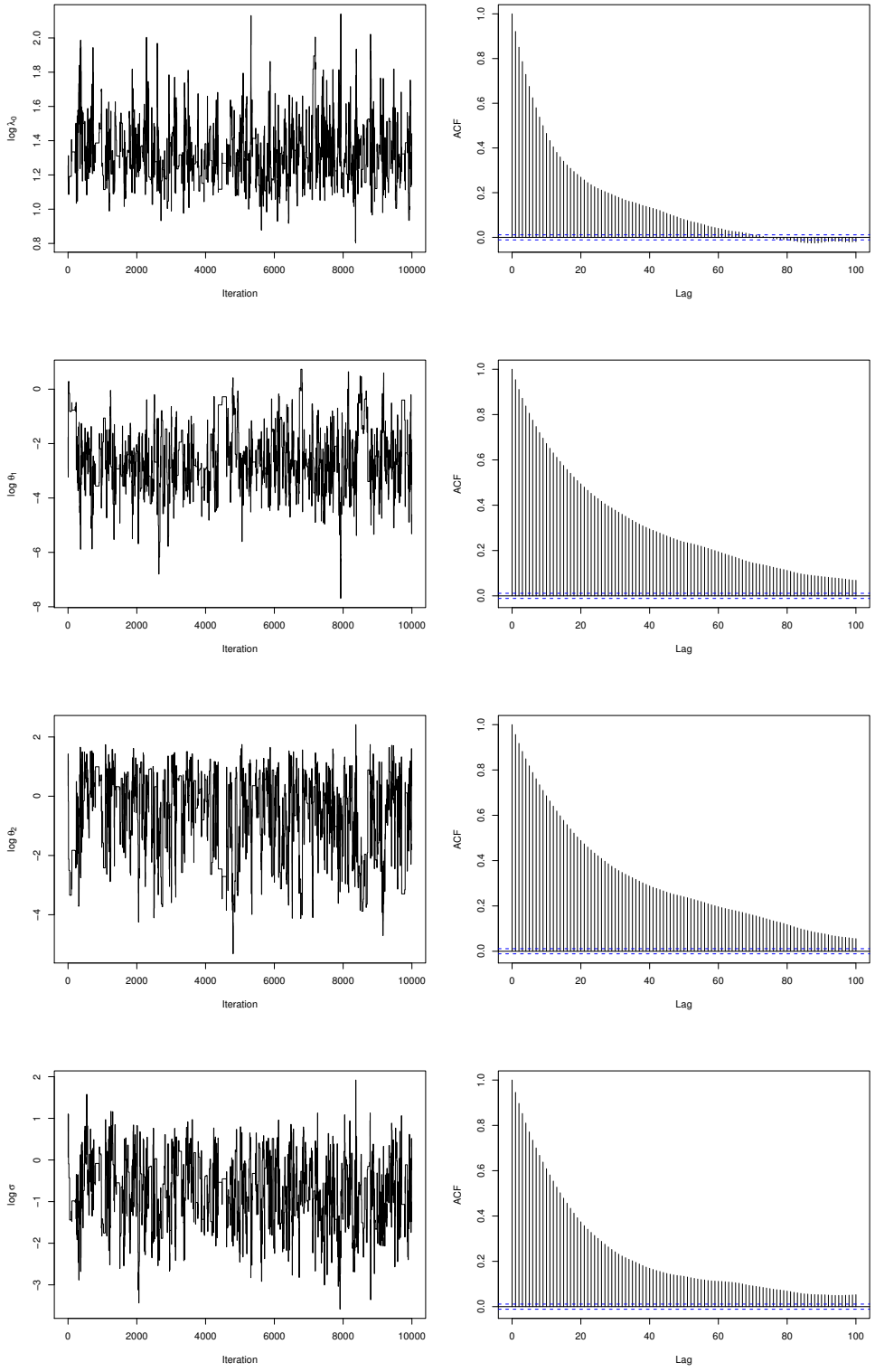


Figure B.2.6: Traceplots and corresponding autocorrelation function estimates for the pseudo-marginal Metropolis-Hastings algorithm on model hyperparameters in the coal-mining disasters dataset example (Section 4.5.2).

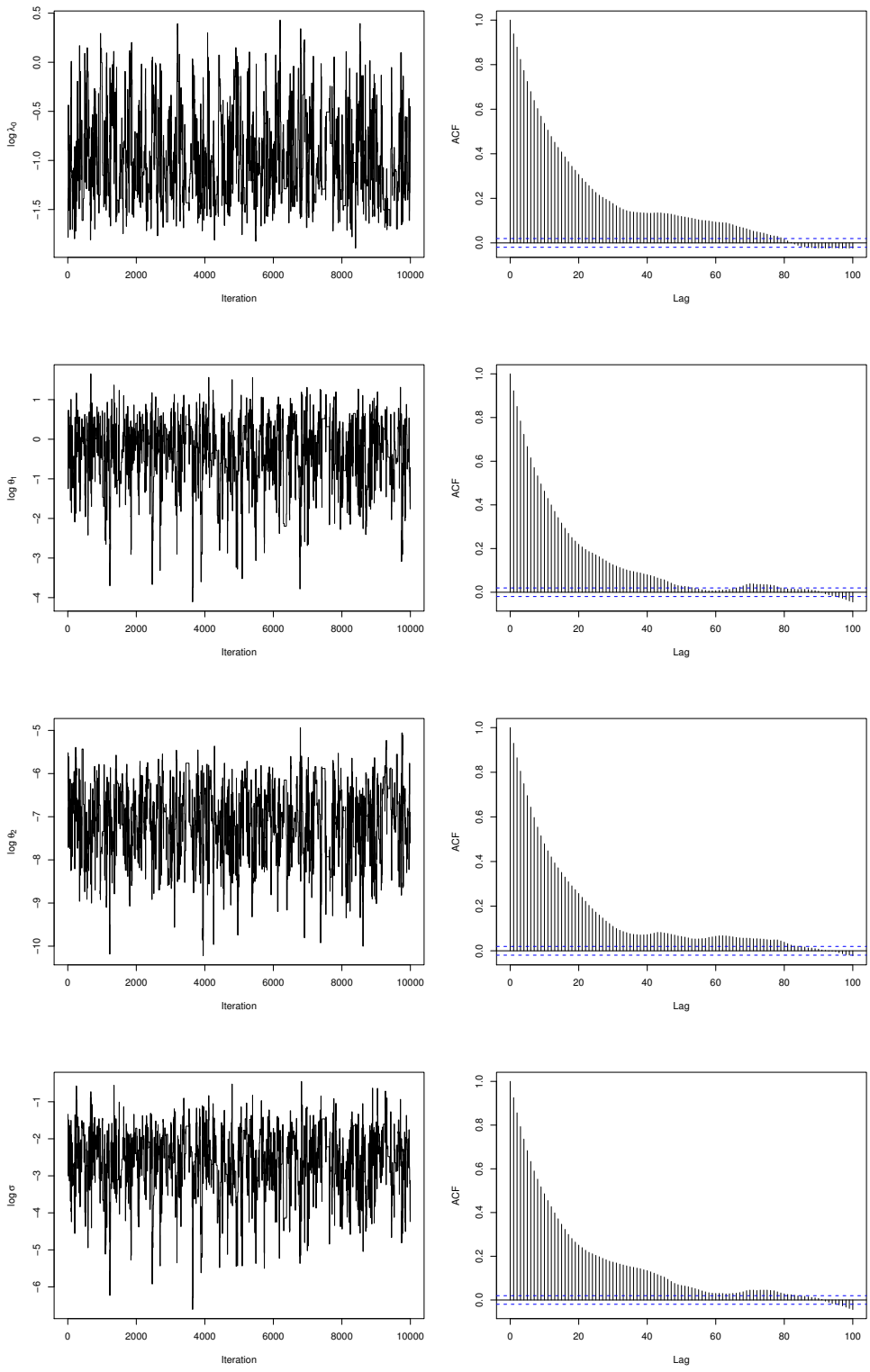


Figure B.2.7: Traceplots and corresponding autocorrelation function estimates for the pseudo-marginal Metropolis-Hastings algorithm on model hyperparameters in the USD-EUR dataset example (Section 4.5.3).

Appendix C

Appendix for Chapter 5

C.1 Additional tuning comparisons

Figure C.1.1 provides additional comparisons between tuning HMC and AAPS_K, this time for Gaussian, logistic and skewed-Gaussian targets. The plots are analogous to those in Figure 5.3.1. For each target, over all components, the smallest scale parameter is $\sigma_1 = 1$. For a Gaussian target this implies that the leapfrog integrator is only stable provided $\varepsilon < 2\sigma_1 = 2$ (e.g. Neal, 2011). Even at $\varepsilon = 1.9$ we encounter undesirable behaviour as seen in Figure C.1.1b. For the skewed-Gaussian distribution with σ_1 and $\alpha = 3$, the density of the narrowest component in the left tail is

$$2\phi(x)\Phi(3x) \approx \frac{2}{3x}\phi(x)\phi(3x) \propto \frac{1}{x}\phi(\sqrt{10}x),$$

where ϕ and Φ are the density and distribution functions of a standard normal random variable, respectively. Since in the left tail, the density resembles that of a Gaussian with $\sigma = 1/\sqrt{10}$, we expect the critical value for ε to be $2/\sqrt{10} \approx 0.632$. As demon-

strated by the plots, for any sensible choice of ε, the performance of AAPS_K is much more robust to the choice of K than HMC (standard or blurred) is to the choice of L, or equivalently, $\mathcal{T}$.

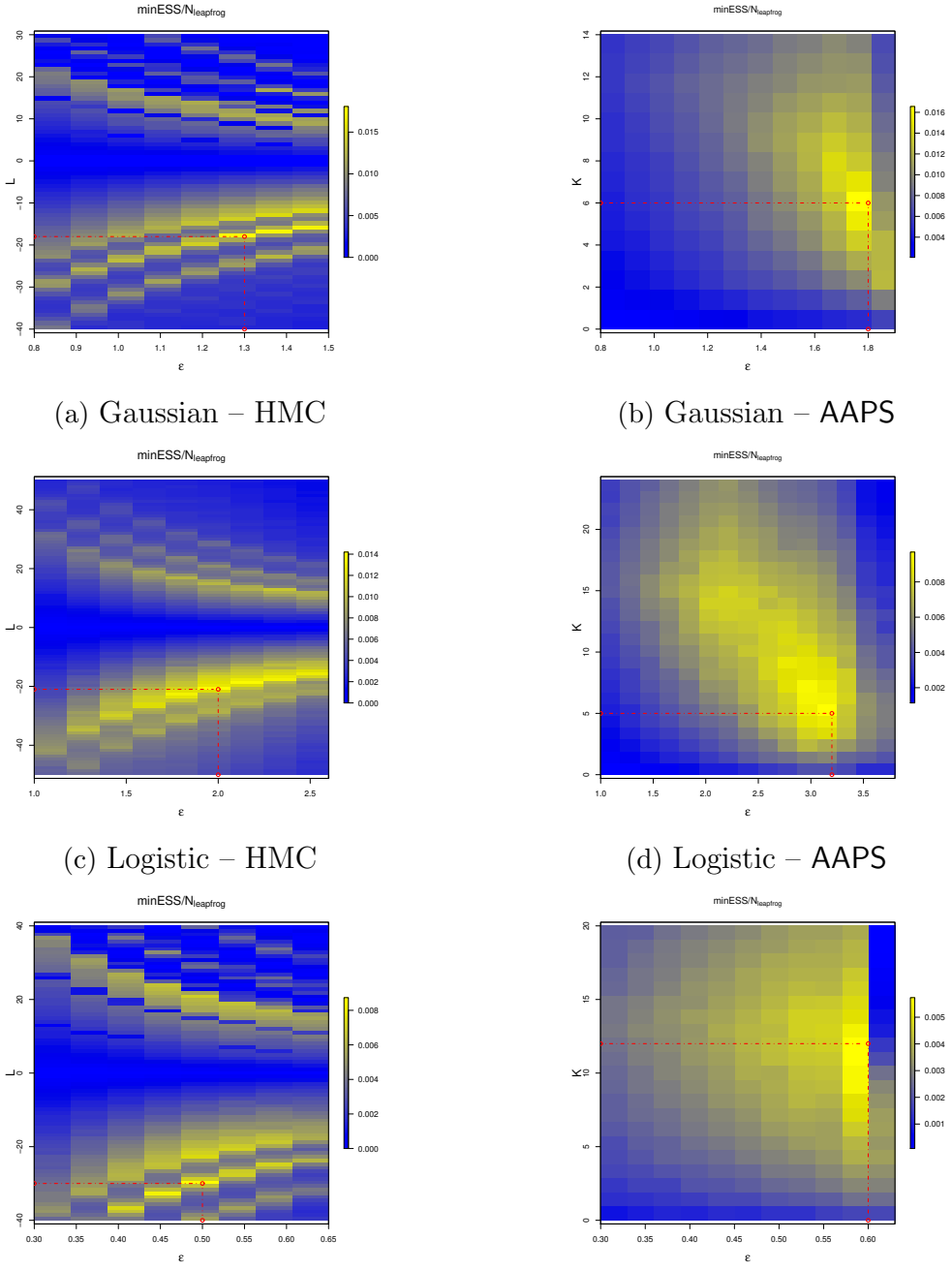
C.2 The diagnostic for K: Heuristic explanation

The following section is from Sherlock et al. (2021) and the work was carried out by my supervisor, Prof Chris Sherlock. It is included for its relevance to the results presented in this chapter.

For a particular segment, j segments from the current point, let $s_*(j; x, p)$ represent the average (over the segment) squared distance in position between the current point, (x, p), and points in the segment. To explain, heuristically, how the diagnostic works we make three simplifying assumptions:

1. $s_*(j; x, p) = c(x, p)s(j)$, for some functions c and s with $s(0) = 0$.
2. The number of points in segment j does not depend on j; *i.e.*, $|\mathcal{S}_j(X, P)| = N(X, P)$ for some integer-valued function N.
3. Acceptance probabilities are generally large enough that variation in these is a secondary effect.

The first assumption seems reasonable as, at least for low j, $s_*(j; x, p) \propto j^2$, approximately, although, strictly $s_*(0; x, p) > 0$ unless there is a single point in the initial segment. The second assumption is strictly incorrect, but in the limit as $\varepsilon \downarrow 0$, $\mathbb{E}[|\mathcal{S}_j(X, P)|]$ does not depend on j since, at stationarity, all points in the Hamiltonian



(a) Gaussian – HMC

(b) Gaussian – AAPS

(c) Logistic – HMC

(d) Logistic – AAPS

(e) Skewed-Gaussian – HMC

(f) Skewed-Gaussian – AAPS

Figure C.1.1: Efficiency w.r.t. tuning parameters for HMC and AAPS on three targets: π_G^{VAR}, π_L^{VAR} and π_{SG}^{VAR}, with $d = 40$ and $\xi = 20$.

path have the same density as the initial point. Thus, the second assumption is reasonable provided values do not vary too much from $N(X, P)$. The third assumption is certainly correct in the limit as the step size, $\varepsilon \downarrow 0$, but is reasonable empirically more generally.

Since proposals are made in proportion to squared jumping distance,

$$m_{K_*}(j) \approx \mathbb{E}\tilde{\pi} s_*(j; X, P) N(X, P) = c_1 s(j), \quad \text{where } c_1 := \mathbb{E}_{\tilde{\pi}}[c(X, P) N(X, P)].$$

We now assume that the tuning parameter has been set to K. The absolute segment number k is proposed with a probability proportional to $p_K(k) N(X, P) c(X, P) s(k)$. The mean squared jumping distance resulting from a tuning parameter K is, therefore

$$MSJD_K := \frac{\sum_{j=0}^{K} p_K(j) N(X, P) c(X, P)^2 s(j)^2}{\sum_{j=0}^{K} p_K(j) N(X, P) c(X, P) s(j)} = c(X, P) \frac{\sum_{j=0}^{K} p_K(j) s(j)^2}{\sum_{j=0}^{K} p_K(j) s(j)}.$$

Denoting $\mathbb{E}[c(X, P)]$ by c_2, the expectation over all initial values is

$$\mathsf{ESJD}_K := c_2 \frac{\sum_{j=0}^{K} p_K(j) s(j)^2}{\sum_{j=0}^{K} p_K(j) s(j)} = c_2 \frac{\sum_{j=1}^{K} p_K(j) s(j)^2}{\sum_{j=1}^{K} p_K(j) s(j)} = c_2 \frac{\sum_{j=1}^{K} (K + 1 - j) s(j)^2}{\sum_{j=1}^{K} (K + 1 - j) s(j)}$$

since $s(0) = 0$. We can simplify this to

$$\mathsf{ESJD}_K = \mathbb{E} s(J), \quad \text{where } \mathbb{P}(J = j) \propto r_j := (K + 1 - j) s(j),\ j = 1, \ldots, K. \quad \text{(C.2.1)}$$

ESJD does not take into account the computational effort, which is proportional to the number of segments, $K + 1$. Hence, we define the efficiency as

$$\mathsf{Eff}_K := \frac{1}{K + 1} \mathbb{E}[S(J)].$$

Intuitively, for small j, $s(j) \propto j^2$, approximately, which motivates the assumptions

in the following.

Proposition C.2.1. *If $s(0) = s'(0) = 0$ and K is small enough that $s(j)$ is convex on $\{0, \dots, K\}$, then*

$$\mathsf{Eff}_K \geq \mathsf{Eff}_K^{\mathrm{lb}} := \frac{1}{K+1} s\left(\frac{K+1}{2}\right),$$

and $\mathsf{Eff}_j^{\mathrm{lb}}$ is non-decreasing on $j \in \{0, \dots, K\}$.

Proof. Jensen's inequality gives $\mathsf{ESJD}_K \geq s(\mathbb{E}[J])$. Further, as s is convex, for any $b \geq a$, the slope of the chord from 0 to b is at least as large as that of the chord from 0 to a:

$$\frac{s(b) - s(0)}{b - 0} \geq \frac{s(a) - s(0)}{a - 0} \implies \frac{s(b)}{b} \geq \frac{s(a)}{a}, \tag{C.2.2}$$

since $s(0) = 0$. Thus, for $j \leq (K+1)/2$, setting $b = K + 1 - j$ and $a = j$, we have

$$r_{K+1-j} \geq r_j,$$

and so $\mathbb{E}[J] \geq (K+1)/2$. Since $s'(j) \geq 0$ on $\{0, \dots, K\}$, we have

$$\mathsf{ESJD}_K \geq s\left(\frac{K+1}{2}\right),$$

proving the first part. Reusing (C.2.2) with $b = (K+1)/2$ and $a = K/2$ gives, as required,

$$\mathsf{Eff}_K^{\mathrm{lb}} \geq \mathsf{Eff}_{K-1}^{\mathrm{lb}}.$$

□

Indeed, straightforward algebra shows that if $s(j) = \lambda j$ then $\mathsf{Eff}_K = \lambda/2$; the inequality is tight for convex functions. However, in practice, $s(j)$ is *strictly* convex initially, which suggests the maximum efficiency may occur after the first point when s

is no longer convex. Plots from various starting points of the mean squared Euclidean distance from the start to points in segment j show behaviour approximately similar to $s(j) \propto 1-\cos(\pi j/b)$, for some b, although there is, typically, less oscillation between peaks and troughs once the first peak has been passed. For small j, this formulation gives, approximately, $s(j) \propto j^2$ which fits with intuition.

Figure C.2.1 plots $s(j)$ against j when $b = 15$, and the resulting Eff_K against K. The efficiency is maximised at a value close to $\mathsf{inf}_{j\geq 0}\mathsf{argmax}\ s(j)$ and the relative difference between the efficiencies at the two points is small (here less that 0.5%). The $1/(K+1)$ penalty term means that damping of the oscillations of $s(j)$ after the first peak will make no difference to the point of maximum efficiency, only to the tail of the efficiency curve.

C.3 Stochastic-volatility model details

To ensure numerical stability we transform the parameters

$$\alpha = 2\mathsf{tanh}^{-1}(\phi), \quad \beta = \mathsf{log}(\kappa), \quad \text{and } \gamma = \mathsf{log}(\sigma^2),$$

with resulting Jacobians

$$\left|\frac{\mathsf{d}\phi}{\mathsf{d}\alpha}\right| = \frac{1}{2}\mathsf{sech}^2(\alpha/2), \quad \left|\frac{\mathsf{d}\beta}{\mathsf{d}\kappa}\right| = \mathsf{e}^{\beta}, \quad \text{and } \left|\frac{\mathsf{d}\sigma^2}{\mathsf{d}\gamma}\right| = \mathsf{e}^{\gamma}.$$

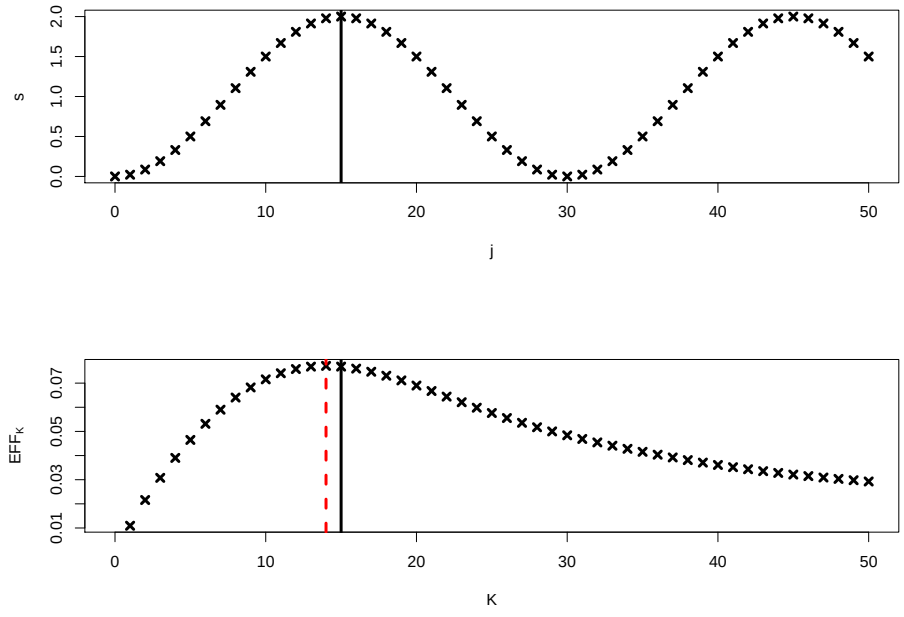


Figure C.2.1: Top panel: plot of $s(j) = 1 - \cos(\pi j/b)$ against j. Bottom panel: plot of the resulting Eff_K against K. The vertical black line in each plot shows the value of b (here, 15), whilst the dashed vertical red line shows the K value at which Eff_K is maximised.

The prior densities then become

$$\begin{aligned}
\pi_0(\beta) &\propto \frac{1}{\mathsf{e}^{\beta}} \cdot \mathsf{e}^{\beta} \\
&= 1, \\
\pi_0(\gamma) &\propto \mathsf{exp}\left(-\frac{10(0.05)}{2\mathsf{e}^{\gamma}}\right) (\mathsf{e}^{\gamma})^{-(1-10/2)} \cdot \mathsf{e}^{\gamma} \\
&= \mathsf{exp}\left(-\frac{1}{4}\mathsf{e}^{-\gamma} - 5\gamma\right), \\
\pi_0(\alpha) &\propto (1+\mathsf{tanh}(\alpha/2))^{19} (1+\mathsf{tanh}(\alpha/2))^{\frac{1}{2}} \cdot \mathsf{sech}^2(\alpha/2) \\
&\propto \mathsf{exp}\left(20\alpha - \frac{43}{2}\mathsf{log}(\mathsf{e}^{\alpha}+1)\right).
\end{aligned}$$

For the last density, recall that

$$\mathsf{tanh}(z) = \frac{\mathsf{e}^{2z}-1}{\mathsf{e}^{2z}+1} \quad \text{and} \quad \mathsf{sech}(z) = \frac{2\mathsf{e}^{z}}{\mathsf{e}^{2z}+1}.$$

The transition densities for the likelihood are

$$\begin{aligned}
p(x_1|\alpha,\gamma) &\propto \mathsf{exp}\left(-\frac{1}{2}x_1^2\mathsf{sech}^2(\alpha/2)\mathsf{e}^{-\gamma} - \frac{1}{2}\gamma\right)\frac{\mathsf{e}^{\alpha/2}}{\mathsf{e}^{\alpha}-1}, \\
p(x_t|x_{t-1},\alpha,\gamma) &\propto \mathsf{exp}\left(-\frac{1}{2}(x_t - \mathsf{tanh}(\alpha/2)x_{t-1})^2\mathsf{e}^{-\gamma} - \frac{1}{2}\gamma\right), \qquad t = 2,\ldots,T \\
p(y_t|x_t,\beta) &\propto \mathsf{exp}\left(-\frac{1}{2}y_t\mathsf{e}^{-2\beta}\mathsf{e}^{-x_t} - \beta - x_t/2\right), \qquad t = 1,\ldots,T.
\end{aligned}$$

Thus, the full log-posterior of the model is, up to an additive constant,

$$\begin{aligned}
\mathsf{log}\,\pi(\alpha,\beta,\gamma,x_{1:T}|y_{1:T}) =& \frac{41}{2}\alpha - \frac{45}{2}\mathsf{log}(\mathsf{e}^{\alpha}+1) - T\beta - \frac{1}{4}\mathsf{e}^{-\gamma} - \frac{2x_1^2\mathsf{e}^{-\gamma}\mathsf{e}^{\alpha}}{(\mathsf{e}^{\alpha}+1)^2} - \left(\frac{T}{2}+5\right)\gamma \\
&- \frac{1}{2}\sum_{t=1}^{T} x_t - \frac{\mathsf{e}^{-2\beta}}{2}\sum_{t=1}^{T} y_t^2\mathsf{e}^{-x_t} - \frac{\mathsf{e}^{-\gamma}}{2}\sum_{t=2}^{T}\left(x_t - \mathsf{tanh}(\alpha/2)x_{t-1}\right)^2,
\end{aligned}$$

with partial derivatives

$$\begin{aligned}
\partial_\alpha \log \pi(\alpha,\beta,\gamma,x_{1:T}|y_{1:T}) =& \frac{41}{2} - \frac{45\mathsf{e}^{\alpha}}{2(\mathsf{e}^{\alpha}+1)} + \frac{2x_1^2\mathsf{e}^{-\gamma}\mathsf{e}^{\alpha}(\mathsf{e}^{\alpha}-1)}{(\mathsf{e}^{\alpha}+1)^3} \\
&+ \frac{\mathsf{e}^{-\gamma}\mathsf{sech}^2(\alpha/2)}{2}\sum_{t=2}^{T}\left(x_t - \mathsf{tanh}(\alpha/2)x_{t-1}\right)x_{t-1}, \\
\partial_\beta \log \pi(\alpha,\beta,\gamma,x_{1:T}|y_{1:T}) =& -T + \mathsf{e}^{-2\beta}\sum_{t=1}^{T} y_t^2\mathsf{e}^{-x_t}, \\
\partial_\gamma \log \pi(\alpha,\beta,\gamma,x_{1:T}|y_{1:T}) =& \frac{e^{-\gamma}}{2}\left(\frac{1}{2} + \sum_{t=2}^{T}\left(x_t - \mathsf{tanh}(\alpha/2)x_{t-1}\right)^2\right) + \frac{2x_1^2\mathsf{e}^{-\gamma}\mathsf{e}^{\alpha}}{(\mathsf{e}^{\alpha}+1)^2} - \left(\frac{T}{2}+5\right), \\
\partial_{x_1} \log \pi(\alpha,\beta,\gamma,x_{1:T}|y_{1:T}) =& -\frac{1}{2} - \frac{4x_1\mathsf{e}^{-\gamma}\mathsf{e}^{\alpha}}{(\mathsf{e}^{\alpha}+1)^2} + \frac{1}{2}y_1^2\mathsf{e}^{-x_1}\mathsf{e}^{-2\beta} + \mathsf{e}^{-\gamma}\mathsf{tanh}(\alpha/2)(x_2 - \mathsf{tanh}(\alpha/2)x_1), \\
\partial_{x_t} \log \pi(\alpha,\beta,\gamma,x_{1:T}|y_{1:T}) =& \frac{1}{2}(y_t^2\mathsf{e}^{-2\beta}\mathsf{e}^{-x_t} - 1) \\
&- \mathsf{e}^{-\gamma}\left[x_t(\mathsf{tanh}^2(\alpha/2)+1)\right. \\
&\qquad \left. -\mathsf{tanh}(\alpha/2)(x_{t+1}+x_{t-1})\right], \quad t = 2,\ldots,T-1 \\
\partial_{x_T} \log \pi(\alpha,\beta,\gamma,x_{1:T}|y_{1:T}) =& \frac{1}{2}(y_T^2\mathsf{e}^{-2\beta}\mathsf{e}^{-x_T} - 1) - \mathsf{e}^{-\gamma}\left(x_T - \mathsf{tanh}(\alpha/2)x_{T-1}\right).
\end{aligned}$$

Bibliography

Christopher P. Adams and Van V. Brantner. Estimating the cost of new drug development: is it really $802 million? *Health affairs*, 25(2):420–428, 2006.

Ryan P. Adams, Iain Murray, and David J. C. MacKay. Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 9–16, 2009.

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. B. N.Petrov, and F. Csaki, editors. Second International Symposium on Information Theory, pages 267–281. Budapest (Hungary): Akademiai Kiado, 1973.

Johan Alenlöv, Arnaud Doucet, and Fredrik Lindsten. Pseudo-Marginal Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 22, 2021.

Renaud Alie, David A. Stephens, and Alexandra M. Schmidt. On data augmentation in point process models based on thinning. *arXiv preprint arXiv:2203.06743*, 2022.

Christophe Andrieu and Gareth O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

Christophe Andrieu and Johannes Thoms. A tutorial on adaptive MCMC. *Statistics and computing*, 18(4):343–373, 2008.

Christophe Andrieu and Matti Vihola. Establishing some order amongst exact approximations of MCMCs. *The Annals of Applied Probability*, 26(5):2661–2696, 2016.

Christophe Andrieu, Arnaud Doucet, Sumeetpal S. Singh, and Vladislav B. Tadic. Particle methods for change detection, system identification, and control. *Proceedings of the IEEE*, 92(3):423–438, 2004.

Vladimir V. Anisimov. Predictive modelling of recruitment and drug supply in multicenter clinical trials. In *Proc. of joint statistical meeting*, pages 1248–1259, 2009.

Vladimir V. Anisimov and Valerii V. Fedorov. Modelling, prediction and adaptive adjustment of recruitment in multicentre trials. *Statistics in Medicine*, 26(27): 4958–4975, 2007.

M. Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

Edward R. Beadle and Petar M. Djuric. A fast-weighted Bayesian bootstrap filter for nonlinear model state estimation. *IEEE Transactions on Aerospace and Electronic Systems*, 33(1):338–343, 1997.

Mark A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160, 2003.

Alexandros Beskos, Omiros Papaspiliopoulos, Gareth O. Roberts, and Paul Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):333–382, 2006.

Alexandros Beskos, Natesh Pillai, Gareth O. Roberts, Jesus-Maria Sanz-Serna, and Andrew Stuart. Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013. ISSN 13507265. URL http://www.jstor.org/stable/42919328.

Michael Betancourt. Identifying the optimal integration time in Hamiltonian Monte Carlo, 2016. URL https://arxiv.org/abs/1601.00225.

Mogens Bladt, Samuel Finch, and Michael Sørensen. Simulation of multivariate diffusion bridges. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(2):343–369, 2016.

Salomon Bochner. Monotone Funktionen, Stieltjessche Integrale und harmonische Analyse. *Mathematische Annalen*, 108(1):378–410, 1933.

Cancer Research UK. Phases of clinical trials, Oct 2022. URL https://www.cancerresearchuk.org/about-cancer/find-a-clinical-trial/what-clinical-trials-are/phases-of-clinical-trials.

Jessica Cariboni and Wim Schoutens. Jumps in intensity models: investigating the performance of Ornstein-Uhlenbeck processes in credit risk modeling. *Metrika*, 69 (2):173–198, 2009.

Jean-Paul Chiles and Pierre Delfiner. *Geostatistics: modeling spatial uncertainty*, volume 497. John Wiley & Sons, 2009.

Nicolas Chopin. Central limit theorem for sequential monte carlo methods and its application to bayesian inference. *The Annals of Statistics*, 32(6):2385–2411, 2004.

Erhan Çinlar. *Introduction to stochastic processes*. Prentice-Hall,Englewood Cliffs, New Jersey, 1975.

David R. Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):129–157, 1955.

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

Carles M. Cuadras. On the covariance between functions. *Journal of Multivariate Analysis*, 81(1):19–27, 2002.

Pierre Del Moral. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems With Applications*, volume 100. 05 2004. ISBN 0387202684. doi: 10.1007/978-1-4684-9393-1.

Pierre Del Moral and Spiridon Penev. *Stochastic Processes: From Applications to Theory*. Chapman and Hall/CRC, 2017.

Luc Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, 1986. URL http://cg.scs.carleton.ca/~luc/rnbookindex.html.

Peter Diggle. A kernel method for smoothing point process data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 34(2):138–147, 1985.

Joseph L. Doob. *Stochastic processes*, volume 7. Wiley New York, 1953.

Randal Douc and Olivier Cappé. Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, pages 64–69. IEEE, 2005.

Arnaud Doucet, Nando De Freitas, and Neil James Gordon. *Sequential Monte Carlo methods in practice*, volume 1. Springer, 2001.

Arnaud Doucet, Nando de Freitas, and Neil Gordon. *Sequential Monte Carlo methods in practice*. New York: Springer Science, 2013.

Arnaud Doucet, Michael K. Pitt, George Deligiannidis, and Robert Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313, 2015.

Simon Duane, A.D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(87)91197-X. URL https://www.sciencedirect.com/science/article/pii/037026938791197X.

Paul Fearnhead, Omiros Papaspiliopoulos, and Gareth O. Roberts. Particle filters for partially observed diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):755–777, 2008.

Paul Fearnhead, Omiros Papaspiliopoulos, Gareth O. Roberts, and Andrew Stuart. Random-weight particle filtering of continuous time processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):497–512, 2010.

William Feller. *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons Inc., New York, 1971.

Lawrence M. Friedman, Curt Furberg, David L. DeMets, David Reboussin, and Christopher B. Granger. *Fundamentals of clinical trials*, volume 3. Springer, 1998.

Byron J. Gajewski, Stephen D. Simon, and Susan E. Carlson. Predicting accrual in clinical trials with Bayesian posterior predictive distributions. *Statistics in Medicine*, 27(13):2328–2340, 2008.

Kenneth Getz and Mary Jo Lamberti. 89% of trials meet enrolment, but timelines slip, half of sites under-enrol. *Tufts CSDD Impact Report*, 15:1–4, 2013.

Charles J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4): 473–483, 1992. ISSN 08834237. URL http://www.jstor.org/stable/2246094.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011. doi: https://doi.org/10.1111/j.1467-9868.2010.00765.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2010.00765.x.

Flávio B. Gonçalves and Dani Gamerman. Exact Bayesian inference in spatiotemporal

Cox processes driven by multivariate Gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):157–175, 2018.

Flavio B. Gonçalves, Krzysztof G. Łatuszyński, and Gareth O. Roberts. Exact Bayesian inference for diffusion driven Cox processes. *arXiv preprint arXiv:2007.05812*, 2020.

Neil J. Gordon, David J. Salmond, and Adrian F.M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE proceedings F (radar and signal processing)*, volume 140, pages 107–113. IET, 1993.

Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Kangxia Gu, Hon Keung Tony Ng, Man Lai Tang, and William R. Schucany. Testing the ratio of two Poisson rates. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(2):283–298, 2008.

Tom Gunter, Chris Lloyd, Michael A. Osborne, and Stephen J. Roberts. Efficient Bayesian nonparametric modelling of structured point processes. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 310–319, 2014.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. Geometric numerical integration illustrated by the Störmer–Verlet method. *Acta Numerica*, 12:399—450, 2003. doi: 10.1017/S0962492902000144.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration: Structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media, 2006.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970. ISSN 0006-3444.

Philip Heidelberger and Peter D. Welch. A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, 24 (4):233–245, 1981.

Daniel F. Heitjan, Zhiyun Ge, and Gui Shuang Ying. Real-time prediction of clinical trial enrollment and event counts: a review. *Contemporary Clinical Trials*, 45: 26–33, 2015.

Nils Lid Hjort and Gerda Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899, 2003.

Matthew D. Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

Matthew D. Hoffman, Alexey Radul, and Pavel Sountsov. An adaptive-MCMC scheme for setting trajectory lengths in Hamiltonian Monte Carlo. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning*

Research, pages 3907–3915. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/hoffman21a.html.

David W. Hogg and Daniel Foreman-Mackey. Data analysis recipes: Using Markov chain Monte Carlo. *The Astrophysical Journal Supplement Series*, 236(1):11, 2018.

Alan M. Horowitz. A generalized guided Monte Carlo algorithm. *Physics Letters B*, 268(2):247–252, 1991.

Vasant Shankar Huzurbazar. Probability distributions and orthogonal parameters. volume 46 of *Mathematical Proceedings of the Cambridge Philosophical Society*, pages 281–284. Cambridge University Press, 1950.

Richard G. Jarrett. A note on the intervals between coal-mining disasters. *Biometrika*, 66(1):191–193, 1979.

Yu Jiang, Steve Simon, Matthew S. Mayo, and Byron J. Gajewski. Modeling and validating Bayesian accrual models on clinical data and simulations using adaptive priors. *Statistics in Medicine*, 34(4):613–629, 2015.

Ruiyang Jin, Sumeetpal S. Singh, and Nicolas Chopin. De-biasing particle filtering for a continuous time hidden Markov model with a Cox process observation model. *arXiv preprint arXiv:2206.10478*, 2022.

Rudolf E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960. ISSN 0021-9223. doi: 10.1115/1.3662552. URL https://doi.org/10.1115/1.3662552.

Benjamin Kasenda, Erik Von Elm, John You, Anette Blümle, Yuki Tomonaga, Ramon Saccilotto, Alain Amstutz, Theresa Bengough, Joerg J. Meerpohl, and Mihaela Stegert. Prevalence, characteristics, and publication of discontinued randomized trials. *JAMA*, 311(10):1045–1052, 2014.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

Peter E. Kloeden and Eckhard Platen. Stochastic differential equations. In *Numerical solution of stochastic differential equations*, pages 103–160. Springer, 1992.

Augustine Kong. A note on importance sampling using standardized weights. *University of Chicago, Dept. of Statistics, Tech. Rep*, 348, 1992.

Augustine Kong, Jun S. Liu, and Wing Hung Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, 1994.

Kalimuthu Krishnamoorthy and Jessica Thomson. A more powerful test for comparing two Poisson means. *Journal of Statistical Planning and Inference*, 119(1):23–35, 2004.

EE Kummer. De integralibus quibusdam definitis et seriebus infinitis. *Journal für die reine und angewandte Mathematik*, 16:228–242, 1837.

Yu Lan, Gong Tang, and Daniel F. Heitjan. Statistical modeling and prediction of clinical trial recruitment. *Statistics in Medicine*, 38:945–955, 2019.

Louis Lasagna. Problems in publication of clinical trial methodology. *Clinical Pharmacology & Therapeutics*, 25(5part2):751–753, 1979.

Radka Lechnerová, Kateřina Helisová, and Vit Beneš. Cox point processes driven by Ornstein–Uhlenbeck type processes. *Methodology and Computing in Applied Probability*, 10(3):315–335, 2008.

Young Jack Lee. Interim recruitment goals in clinical trials. *Journal of Chronic Diseases*, 36(5):379–389, 1983.

Benedict Leimkuhler and Sebastian Reich. Simulating Hamiltonian dynamics, Feb 2005. URL http://dx.doi.org/10.1017/CBO9780511614118.

P. A. W. Lewis and Gerald S. Shedler. Simulation of nonhomogeneous Poisson processes by thinning. *Naval research logistics quarterly*, 26(3):403–413, 1979.

Chenhao Li and Simon Godsill. Sequential inference methods for non-homogeneous Poisson processes with state-space prior. *IEEE Transactions on Signal Processing*, 69:1154–1168, 2021.

Chenhao Li and Simon J. Godsill. Sequential inference methods for non-homogeneous Poisson processes with state-space prior. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2856–2860. IEEE, 2018.

D. V. Lindley. A statistical paradox. *Biometrika*, 44(1-2):187–192, 06 1957. ISSN 0006-3444. doi: 10.1093/biomet/44.1-2.187. URL https://doi.org/10.1093/biomet/44.1-2.187.

Jun S. Liu and Rong Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044, 1998.

Samuel Livingstone, Michael Betancourt, Simon Byrne, and Mark Girolami. On the geometric ergodicity of Hamiltonian Monte Carlo. *Bernoulli*, 25(4A):3109 – 3138, 2019. doi: 10.3150/18-BEJ1083. URL https://doi.org/10.3150/18-BEJ1083.

Chris Lloyd, Tom Gunter, Michael Osborne, and Stephen Roberts. Variational inference for Gaussian process modulated Poisson processes. In *International Conference on Machine Learning*, pages 1814–1822. PMLR, 2015.

Lynn Harold Loomis. An introduction to abstract harmonic analysis. 1953.

Paul B. Mackenze. An improved hybrid Monte Carlo method. *Physics Letters B*, 226 (3-4):369–371, 8 1989. doi: 10.1016/0370-2693(89)91212-4.

James Martin, Lucas C. Wilcox, Carsten Burstedde, and Omar Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log Gaussian Cox processes. *Scandinavian journal of statistics*, 25(3):451–482, 1998.

Wilson Tsakane Mongwe, Rendani Mbuvha, and Tshilidzi Marwala. Antithetic magnetic and shadow Hamiltonian Monte Carlo. *IEEE Access*, 9:49857–49867, 2021. doi: 10.1109/ACCESS.2021.3069196.

Lucy E. Morgan, Barry L. Nelson, Andrew C. Titman, and David J. Worthington. A spline-based method for modelling and generating a nonhomogeneous Poisson process. In *2019 Winter Simulation Conference (WSC)*, pages 356–367. IEEE, 2019.

Gerhard Nahler. *Lasagna's law. In: Dictionary of Pharmaceutical Medicine*. Springer, Vienna, 2009.

Radford M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

Radford M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

John A. Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

Christopher Nemeth, Chris Sherlock, and Paul Fearnhead. Particle Metropolis-adjusted Langevin algorithms. *Biometrika*, 103(3):701–717, 2016.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

Tin Lok J. Ng and Thomas B. Murphy. Estimation of the intensity function of an in-

homogeneous Poisson process with a change-point. *Canadian Journal of Statistics*, 47(4):604–618, 2019.

Bernt Øksendal. Stochastic differential equations. In *Stochastic differential equations*, pages 65–84. Springer, 2003.

Filippo Pagani, Martin Wiegand, and Saralees Nadarajah. An n-dimensional Rosenbrock distribution for Markov chain Monte Carlo testing. *Scandinavian Journal of Statistics*, 2021. doi: https://doi.org/10.1111/sjos.12532. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12532.

Steven Piantadosi and Blossom Patterson. A method for predicting accrual, cost, and paper flow in clinical trials. *Controlled Clinical Trials*, 8(3):202–215, 1987.

Michael K. Pitt and Neil Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599, 1999.

Michael K. Pitt, Ralph dos Santos Silva, Paolo Giordani, and Robert Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151, 2012.

Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. CODA: convergence diagnosis and output analysis for MCMC. *R news*, 6(1):7–11, 2006.

Stuart J. Pocock. *Clinical trials: a practical approach*. John Wiley & Sons, 1983.

Emilia Pompe, Chris Holmes, and Krzysztof Łatuszyński. A framework for adaptive

MCMC targeting multimodal distributions. *The Annals of Statistics*, 48(5):2930–2952, 2020.

Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Vinayak Rao, Ryan P. Adams, and David D. Dunson. Bayesian inference for Matérn repulsive processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):877–897, 2017.

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387212396.

Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16:351–367, 2001.

Tim Robertson, F. T. Wright, and R. L. Dykstra. *Order restricted statistical inference*. Wiley: New York, 1988.

Sheldon M. Ross, John J. Kelly, Roger J. Sullivan, William James Perry, Donald Mercer, Ruth M. Davis, Thomas Dell Washburn, Earl V. Sager, Joseph B. Boyce, and Vincent L. Bristow. *Stochastic processes*, volume 2. Wiley New York, 1996.

Tushar Sakpal. Sample size estimation in clinical trial. *Perspectives in clinical research*, 1(2):67–67, 2010.

John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6 (2):461–464, 1978.

Chris Sherlock, Alexandre H. Thiery, Gareth O. Roberts, and Jeffrey S. Rosenthal. On the efficiency of pseudo-marginal random walk Metropolis algorithms. *The Annals of Statistics*, 43(1):238–275, 2015.

Chris Sherlock, Szymon Urbas, and Matthew Ludkin. The apogee to apogee path sampler. *arXiv preprint arXiv:2112.08187*, 2021.

Seymour Shlien and M. Nafi Toksöz. A clustering model for earthquake occurrences. *Bulletin of the Seismological Society of America*, 60(6):1765–1787, 1970.

Stan Development Team. Stan modeling language users guide and reference manual, 2020. URL http://mc-stan.org/. Version 2.28.

Daniel W. Stroock. *Markov processes from K. Itô's perspective*. Annals of mathematics studies ; Number 155. Princeton University Press, Princeton, New Jersey ; Oxfordshire, England, 2003. ISBN 0-691-11542-7.

Gong Tang, Yuan Kong, Chung-Chou Ho Chang, Lan Kong, and Joseph P. Costantino. Prediction of accrual closure date in multi-center clinical trials with discrete-time Poisson process models. *Pharmaceutical Statistics*, 11(5):351–356, 2012.

Yee Teh and Vinayak Rao. Gaussian process modulated renewal processes. *Advances in Neural Information Processing Systems*, 24, 2011.

Szymon Urbas. *Predicting Recruitment in Clinical Trials.* MRes dissertation, Lancaster University, 2018.

Wolfgang Wagner. Unbiased Monte Carlo estimators for functionals of weak solutions of stochastic differential equations. *Stochastics: An International Journal of Probability and Stochastic Processes*, 28(1):1–20, 1989.

Ziyu Wang, Shakir Mohamed, and Nando Freitas. Adaptive Hamiltonian and Riemann manifold Monte Carlo. In *International conference on machine learning*, pages 1462–1470. PMLR, 2013.

David I. Warton and Leah C. Shepherd. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, pages 1383–1402, 2010.

Lee Jen Wei. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15):1871–1879, 1992.

Jonathan Weinberg, Lawrence D. Brown, and Jonathan R. Stroud. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of the American Statistical Association*, 102(480):1185–1198, 2007.

William O. Williford, Stephen F. Bingham, David G. Weiss, Joseph F. Collins, Keith T. Rains, and William F. Krol. The "constant intake rate" assumption

in interim recruitment goal methodology for multicenter clinical trials. *Journal of Chronic Diseases*, 40(4):297–307, 1987.

Changye Wu, Julien Stoehr, and Christian P. Robert. Hamiltonian Monte Carlo by learning leapfrog scale, 2019.

N. Donald Ylvisaker. The expected number of zeros of a stationary Gaussian process. *Annals of Mathematical Statistics*, 36:1043–1046, 1965.

Xiaoxi Zhang and Qi Long. Stochastic modeling and prediction for accrual in clinical trials. *Statistics in Medicine*, 29(6):649–658, 2010.