Fast and scalable non-parametric Bayesian inference for Poisson point processes

Downloaded from: https://research.chalmers.se, 2026-01-15 08:00 UTC

Citation for the original published paper (version of record):
Gugushvili, S., van der Meulen, F., Schauer, M. et al (2019). Fast and scalable non-parametric Bayesian inference for Poisson point processes. RESEARCHERS.ONE

N.B. When citing this work, cite the original published paper.

research.chalmers.se offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all kind of research output: articles, dissertations, conference papers, reports etc. since 2004. research.chalmers.se is administrated and maintained by Chalmers Library

(article starts on next page)

Fast and scalable non-parametric Bayesian inference for Poisson point processes

Shota Gugushvili*, Frank van der Meulen†, Moritz Schauer‡ and Peter Spreij§

February 14, 2020

Abstract

We study the problem of non-parametric Bayesian estimation of the intensity function of a Poisson point process. The observations are n independent realisations of a Poisson point process on the interval $[0, T]$. We propose two related approaches. In both approaches we model the intensity function as piecewise constant on N bins forming a partition of the interval $[0, T]$. In the first approach the coefficients of the intensity function are assigned independent gamma priors, leading to a closed form posterior distribution. On the theoretical side, we prove that as $n \to \infty$, the posterior asymptotically concentrates around the "true", data-generating intensity function at an optimal rate for h-Hölder regular intensity functions ($0 < h \leq 1$).

In the second approach we employ a gamma Markov chain prior on the coefficients of the intensity function. The posterior distribution is no longer available in closed form, but inference can be performed using a straightforward version of the Gibbs sampler. Both approaches scale well with sample size, but the second is much less sensitive to the choice of N.

Practical performance of our methods is first demonstrated via synthetic data examples. We compare our second method with other existing approaches on the UK coal mining disasters data. Furthermore, we apply it to the US mass shootings data and Donald Trump's Twitter data.

Keywords and phrases: Empirical Bayes, Intensity function, gamma Markov chain prior, Gibbs sampler, Independent gamma prior, Markov Chain Monte Carlo, Metropolis-within-Gibbs, Non-homogeneous Poisson process, Non-parametric Bayesian estimation, Poisson point process, Posterior contraction rate.

*Biometris, Wageningen University & Research, Postbus 16, 6700 AA Wageningen, The Netherlands, shota.gugushvili@wur.nl

†Delft Institute of Applied Mathematics, Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands, f.h.vandermeulen@tudelft.nl

‡Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE-412 96 Göteborg, Sweden, smoritz@chalmers.se

§Korteweg-de Vries Institute for Mathematics, University of Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands, and Institute for Mathematics, Astrophysics and Particle Physics, Radboud University, Nijmegen, spreij@uva.nl

1 Introduction

Poisson point processes (see Kingman (1993)) are among the basic modelling tools in areas as different as astronomy, biology, geology, image analysis, medicine, operations research, physics, reliability theory, social media and others; see, e.g., Brémaud (1981), Møller & Waagepetersen (2004) and Streit (2010). Their probabilistic properties are completely determined by their intensity measure Λ, or its density, the intensity function λ. In this paper, we study a non-parametric Bayesian approach to estimation of the intensity function λ for univariate Poisson point processes. We propose two related approaches that lead to simple implementations, and also enjoy favourable performance in practice.

1.1 Setting

We introduce our statistical setting in greater detail. We restrict our attention to a Poisson point process X on an interval $[0, T]$ of the real line $\mathbb{R}$, which we equip with the Borel σ-field $\mathcal{B}([0, T])$. Such Poisson point processes are also often called non-homogeneous Poisson processes. The process X is a random integer-valued measure on $[0, T]$ (we assume the underlying (complete) probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ in the background), such that

(i) for any disjoint subsets $B_1, B_2, \ldots, B_m \in \mathcal{B}([0, T])$, the random variables $X(B_1)$, $X(B_2), \ldots, X(B_m)$ are independent, and

(ii) for any $B \in \mathcal{B}([0, T])$, the random variable $X(B)$ is Poisson distributed with parameter $\Lambda(B)$. Here Λ is a finite measure on $([0, T], \mathcal{B}([0, T]))$, called the intensity measure of the process X. Moreover, it is assumed that Λ admits a density λ with respect to the Lebesgue measure on $\mathcal{B}([0, T])$.

Intuitively, the process X can be thought of as random scattering of points in $[0, T]$, where the way the scattering occurs is determined by properties (i)–(ii).

A popular assumption in the statistical literature (see, e.g., Karr (1986) and Kutoyants (1998)) is that one has independent copies $X_1, \ldots, X_n$ of the process X at her disposal. Based on these observations, an estimator of the intensity function λ has to be constructed. Intensity functions that are periodic with a known period also lead to this statistical setting.

1.2 Literature overview

Theoretical results for a kernel-type estimator of the intensity function are given in Kutoyants (1998), where further references can be found. See also Diggle (1985), where a practical implementation is discussed in a closely related problem of estimation of the intensity of a stationary Cox process. Some recent references treating various types of the kernel method are Baddeley et al. (2016), Diggle (2014), and Cronie & van Lieshout (2018).

Works dealing with non-parametric[1] Bayesian intensity function estimation include, among others, Adams et al. (2009), Heikkinen & Arjas (1998), Hensman et al. (2015), John & Hensman (2018), Kom Samo & Roberts (2015), Møller et al. (1998) and Teh &

[1] As put characteristically in Efron & Hastie (2016), page 172: "Nonparametric" is another name for "very highly parametrized".

Rao (2011), and concentrate primarily on computational aspects. On the other hand, Kirichenko & van Zanten (2015) is a theoretical contribution analysing the approach in Adams et al. (2009) (cf. also Gugushvili & Spreij (2013)). Belitser et al. (2015) deals both with practical and theoretical aspects of the problem. Another recent paper dealing with derivation of posterior contraction rates in the Poisson point process setting using Gaussian priors is Grant & Leslie (2019).

Advantages of a non-parametric Bayesian approach over the kernel method are succinctly summarised in the discussion given in Adams et al. (2009). One obvious drawback of a "naive" kernel estimator is that it is inconsistent at the boundary of the set $[0, T]$ on which the process is defined. In practice this estimator will also behave poorly close to boundary points. This has to do with the well-known boundary bias problem of the kernel estimator. Simulation examples in Adams et al. (2009) demonstrate that even after the correction for edge effects following the method in Diggle (1985) and with an optimal choice of the bandwidth parameter, the kernel method is still outperformed by the Bayesian approach of Adams et al. (2009). Secondly, if a kernel of order higher than two is used, a non-negative estimate of the intensity function is not guaranteed by the kernel method. On the other hand, the Bayesian approaches studied in the above cited works are often computationally intense, in that their practical implementation is based on advanced Markov Chain Monte Carlo (MCMC) or optimisation methods, and are also not trivial to implement.

1.3 Our contribution

In this paper we propose two simple non-parametric Bayesian approaches to estimation of the intensity function. We model the intensity function as piecewise constant on the domain of its definition $[0, T]$. In our first approach we equip the coefficients of the intensity function with independent gamma priors, and obtain a posterior distribution that is known in closed form and is also of gamma type. Posterior inference with this approach is computationally elementary, with no need to recourse to simulation methods such as MCMC, or optimisation techniques such as variational Bayes. Hence the method scales extremely well with sample size. This method can be thought of as a simple to use Bayesian tool for preliminary, exploratory data analysis, that is much akin to a histogram or a regressogram; see Wasserman (2006). The approach requires the choice of the number of coefficients of the intensity function. A simple practical method to select this hyperparameter is to use the empirical Bayes method.

On the theoretical side, we derive the contraction rate of the posterior distribution around the "true" intensity function λ_0 under $\mathbb{P}_{\lambda_0}^{(n)}$, where $\mathbb{P}_{\lambda_0}^{(n)}$ denotes the law of the observation $X^{(n)} = (X_1, \ldots, X_n)$ under the true parameter λ_0. This concerns taking a sequence of shrinking neighbourhoods of λ_0 and determining the fastest rate, at which these neighbourhoods can shrink, while still capturing most of the posterior mass (the precise definition will be given below). This rate can be thought of as an analogue of the convergence rate of a frequentist estimator. As we will demonstrate, our approach attains the optimal posterior contraction rate for estimating an h-Hölder regular intensity function, $0 < h \leq 1$. We stress the fact that the data-generating λ_0 in our approach is not necessarily assumed to be piecewise constant with known break points.

Inspired by ideas in the audio signal processing literature, see, e.g., Cemgil & Dikmen (2007), Peeling et al. (2008) and Dikmen & Cemgil (2010), as well as inference in diffusion models, see Gugushvili et al. (2018) and Gugushvili et al. (2019b), we next propose a

second non-parametric Bayesian method. This method extends the first and relies on the gamma Markov chain (GMC) prior. Specifically, as in our first approach, we model a priori the intensity function as piecewise constant, but now its coefficients under a prior form a gamma Markov chain. Unlike our first method, the posterior is not available in closed form. However, this second method can be implemented in a straightforward way using the Gibbs sampler. We initialise our Gibbs sampler using the information obtained from our first method, to ensure quick convergence of the resulting Markov chain. We provide a comparison of both methods via simulations, and show that the first one is more sensitive to the appropriate choice of the number of bins and is outperformed by the second one in practice. Finally, like the first method, also the second one scales well with sample size: once the number of Poisson events falling within each bins has been determined, the computational complexity per MCMC step of the second method is linear in the number of bins.

The methodology developed in this paper is applied on three real data examples: the UK coal mining disasters data, the US mass shootings data and Donald Trump's Twitter data. The first dataset is a classical benchmark in point process inference and is used to compare our method with results from Adams et al. (2009) and Lloyd et al. (2015).

1.4 Structure of the paper

The rest of the paper is organised as follows: in Section 2 we introduce in detail our first Bayesian procedure, study its frequentist asymptotics, and discuss one method for practical selection of a hyperparameter (the number of coefficients), which governs properties of the prior. In Section 3 we introduce our second Bayesian estimation technique based on the GMC prior. In Section 4 we present simulations examples, while in Section 5 we apply our approach on real datasets. In Section 6 we summarise contributions of our paper. Appendix A contains the proofs of technical results. Finally, in Appendices B and C we include some results obtained based on Gaussian processes and equipping a prior on the model index respectively.

1.5 Notation

For two sequences $\{a_n\}$ and $\{b_n\}$ of positive real numbers, the notation $a_n \lesssim b_n$ (or $b_n \gtrsim a_n$) means that there exists a constant $C > 0$ that is independent of n and such that $a_n \leq Cb_n$. We write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. $\|\cdot\|_2$ denotes the L_2-norm with respect to the Lebesgue measure on the Borel sets of $[0, T]$. We denote a prior (depending on the hyperparameter N) by Π_N. Given data $X^{(n)}$, the corresponding posterior measure is denoted by $\Pi_N(\cdot \mid X^{(n)})$, and $\mathbb{E}_{\Pi_N}[\cdot \mid X^{(n)}]$ stands for expectation with respect to the posterior measure, while $\mathbb{V}\mathrm{ar}_{\Pi_N}(\cdot \mid X^{(n)})$ is the corresponding variance. The gamma distribution with shape parameter α and rate parameter β ($\alpha, \beta > 0$) is denoted by $\mathrm{G}(\alpha, \beta)$. Its density is given by $x \mapsto \frac{\beta^\alpha}{\gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, $x > 0$, where Γ is the gamma function. The inverse gamma distribution with shape parameter α and scale parameter β is denoted by $\mathrm{IG}(\alpha, \beta)$. Its density is $x \mapsto \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$, $x > 0$. A normal (Gaussian) density with mean μ and standard deviation σ is denoted by $\phi(\cdot; \mu, \sigma)$. We use the notation $\mathrm{Uniform}(a, b)$ for a uniform distribution on $[a, b]$, whereas $\mathrm{Exp}(\beta)$ stands for an exponential distribution with mean $1/\beta$. In conformance with standard Bayesian notation, we will often use lowercase letters for random variables and write a density of a random variable x as $p(x)$. A conditional density of x given y will be written as $p(x \mid y)$,

while conditioning of x on y will be denoted by $x \mid y$. Finally, $\lceil a \rfloor$ will stand for the integer nearest to the real number a.

2 Independent gamma prior

In this section we present in detail our first method of estimation of the intensity function.

2.1 Likelihood

In a Bayesian approach to estimation of λ one starts with putting a prior Π on λ. This might be thought of as reflecting one's prior knowledge or beliefs on λ. In formal terms, the prior is a measure Π defined on the parameter set Θ (a set of functions $\lambda\colon [0,T] \mapsto \mathbb{R}_+$) equipped with a σ-field $\sigma(\Theta)$. By Proposition 6.11 in Karr (1986) or Theorem 1.3 in Kutoyants (1998), the law of X under the parameter value λ admits a density $p(\cdot\,;\lambda)$ with respect to the measure induced by a standard Poisson point process of rate 1. This density is given by

$$p(\xi;\lambda) = \exp\left(\int_{[0,T]} \log\lambda(x)\mathrm{d}\xi(x) - \int_{[0,T]} (\lambda(x)-1)\mathrm{d}x\right),$$

where $\xi = \sum_{i=1}^m \delta_{x_i}$ is a realisation of X (here δ_{x_i} denotes the Dirac measure at x_i). We assume independent observations $X_1, \ldots, X_n$ with the same distribution as X and write $X_j = \sum_{i=1}^{m_j} \delta_{X_{ij}}$ for $j = 1, \ldots, n$. If we define $X^{(n)} = (X_1, X_2, \ldots, X_n)$, then it follows that the likelihood $L(X^{(n)};\lambda)$ can be written as

$$L(X^{(n)};\lambda) = \prod_{j=1}^n \exp\left(\int_{[0,T]} \log\lambda(x)\mathrm{d}X_j(x) - \int_{[0,T]} (\lambda(x)-1)\,\mathrm{d}x\right). \tag{1}$$

2.2 Prior

Fix a positive integer N. Let $0 = b_0 < b_1 < \cdots < b_{N-1} < b_N = T$ be a grid of points on the interval $\mathcal{X} = [b_0, b_N]$, for instance a uniform grid. Using this grid, define the bins $B_k = [b_{k-1}, b_k)$, $k = 1, \ldots, N-1$, with the last bin $B_N = [b_{N-1}, b_N]$. In order to define a prior on λ, introduce a collection of piecewise constant functions λ on bins B_k,

$$\lambda(x) = \sum_{k=1}^N \psi_k \mathbf{1}_{B_k}(x), \quad x \in [0,T]. \tag{2}$$

Assumption 1. *Assume that the coefficients* $\psi_k \overset{i.i.d.}{\sim} \mathrm{G}(\alpha,\beta)$. *Define the prior* Π_N *on* λ *to be the law of the random function* (2).

We will refer to the prior Π_N as the independent gamma prior. If the grid on $\mathcal{X} = [0,T]$ is taken to be uniform, the prior has three hyperparameters: α, β, and N. Depending on the amount of available data and the degree of prior knowledge on λ, the hyperparameters α, β can for instance be chosen to render the $\mathrm{G}(\alpha,\beta)$ prior diffuse (non-informative). This corresponds to the case when one has little information on the magnitude and shape of λ. The hyperparameter N can be viewed as a smoothing parameter. We discuss one practical method for its choice in Subsection 2.5.

Using a prior with piecewise constant realisations as in (2) is not unnatural, as follows by a comparison to a histogram-like frequentist procedure studied in detail from the theoretical point of view in Henderson (2003) and Leemis (2004). In fact, in practical applications it is often the case that Poisson point processes are only partially observed through aggregate counts of points over successive intervals (cf. Henderson (2003)). The intensity function cannot be learned beyond the resolution of these intervals, which lends support to using priors with piecewise constant realisations to model the intensity function.

2.3 Posterior

Since the function λ in our case is parametrised by the coefficients $\psi_1, \dots, \psi_N$, the posterior distribution of the intensity function λ given the data $X^{(n)}$ can be equivalently described through the posterior distribution $\psi_1, \dots, \psi_N \mid X^{(n)}$. Transition from the prior to the posterior can be thought of as updating our prior opinion on λ upon seeing the data $X^{(n)}$.

Lemma 1. *Define*

$$H_k = \sum_{j=1}^{n} \sum_{i=1}^{m_j} \mathbf{1}_{\{X_{ij} \in B_k\}} \tag{3}$$

and $\Delta_k = b_k - b_{k-1}$ *for* $k = 1, \dots, N$. *Then* $\psi_1, \dots, \psi_N$ *are a posteriori independent and*

$$\psi_k \mid X^{(n)} \sim \mathrm{G}(\alpha + H_k, \beta + n\Delta_k), \quad k = 1, \dots, N. \tag{4}$$

Thus, the posterior for λ is known in closed form. The posterior mean is given by $\hat{\lambda}(x) = \sum_{k=1}^{N} \hat{\psi}_k \mathbf{1}_{B_k}(x)$, $x \in [0, T]$, with

$$\hat{\psi}_k = \frac{H_k + \alpha}{n\Delta_k + \beta}, \quad k = 1, \dots, N. \tag{5}$$

Marginal $1-\gamma$ credible bands for λ can be obtained by producing $1-\gamma$ credible intervals for ψ_k, using the lower and upper $\gamma/2$-quantiles of the gamma distribution. This gives Bayesian uncertainty quantification for estimation of λ.

If we were to ignore the hyperparameters α, β (or alternatively, if we were to consider the limit $n \to \infty$ in such a way that $n \min_k \Delta_k \to \infty$), then the posterior mean $\hat{\lambda}$ would have coincided with a frequentist estimator of the intensity function from Henderson (2003); cf. also Leemis (2004). Thus our approach sheds additional light on the latter method, in that it yields its Bayesian interpretation.

Returning to an interpretation of the prior distribution on λ, our view aligns with that in Martin & Walker (2019): "[I]n high-dimensional problems... the role [of prior distributions] is simply to facilitate efficient posterior inference, and, therefore, only priors whose corresponding posterior has good properties are used". That the latter is indeed the case, is the subject of the next subsection. Computational efficiency is clear from the above discussion and will be further demonstrated below.

2.4 Bayesian asymptotics

In this subsection we perform the asymptotic frequentist analysis of the Bayesian procedure we introduced above. This concerns the study of the asymptotic properties of the posterior measure as the sample size $n \to \infty$.

We first formalise our assumption on the bins B_k.

Assumption 2. *Assume that the grid $\{b_k\}$ defining the bins B_k is uniform: $b_k = Tk/N$, $k = 0, \ldots, N$, so that the bins are of equal width $\Delta_k = \Delta = T/N$.*

The assumption that the grid $\{b_k\}$ is uniform is made for simplicity in the proofs, and can in fact be relaxed. It is not necessary for our method to work in synthetic and real data examples.

The next condition deals with the "true", data-generating intensity function λ_0. It places a modest smoothness assumption on λ_0, which will often be satisfied in practice.

Assumption 3. *The intensity function $\lambda_0 \colon [0, T] \to (0, \infty)$ is Hölder continuous: there exist constants $L > 0$ and $0 < h \leq 1$, such that*

$$|\lambda_0(x) - \lambda_0(y)| \leq L|x - y|^h, \quad \forall x, y \in [0, T].$$

Our first result shows that the posterior mean $\hat{\lambda}$ is a consistent estimator of λ_0, and establishes its convergence rate. The expectation $\mathbb{E}[\,\cdot\,]$ here and in the sequel is always under the law of the observations with the "true" parameter value λ_0.

Theorem 1. *Let Assumption 3 hold, and assume $N \asymp n^{1/(2h+1)}$. Then*

$$\mathbb{E}[\|\hat{\lambda} - \lambda_0\|_2^2] \lesssim n^{-2h/(2h+1)}.$$

The right-hand side is the optimal rate for estimating an h-Hölder-regular intensity function, see Kutoyants (1998).

The next result gives a posterior contraction rate in the L_2-metric. Whereas Theorem 1 dealt with the 'centre' of the posterior distribution, the theorem below deals with the entire posterior distribution.

Theorem 2. *Let the assumptions of Theorem 1 hold, and let $\varepsilon_n \asymp n^{-h/(2h+1)}$. Then, for any sequence $M_n \to \infty$,*

$$\mathbb{E}[\Pi_N(\|\lambda - \lambda_0\|_2 \geq M_n \varepsilon_n \mid X^{(n)})] \to 0$$

as $n \to \infty$.

The first conclusion that follows from Theorem 2 is that the proposed Bayesian procedure is consistent: as the sample size $n \to \infty$, the posterior puts most of its mass on (shrinking) L_2-neighbourhoods around the true parameter λ_0. Furthermore, the rate $\varepsilon_n \asymp n^{-h/(2h+1)}$ in Theorem 2 is the optimal posterior contraction rate for h-Hölder-regular intensity functions.

Note that the posterior contraction rate in Theorem 2 does not involve a $\log n$ factor, that often appears when studying the frequentist asymptotics of non-parametric Bayesian procedures. The reason for this is that our proof does not appeal to the powerful and general-purpose machinery from Ghosal et al. (2000), which might yield a superfluous log factor in the contraction rate. Instead, it relies on elementary and direct calculations and estimates.

Although not concerned with direct inference in Poisson point process models, a work related to ours is Grant et al. (2019). There the authors study the problem of adaptively placing sensors along an interval to detect stochastically-generated events. A histogram-type Bayesian prior on the intensity function plays a significant role in the developments in that paper. The authors derive an $O(M^{2/3})$ bound on Bayesian regret in M rounds, which in their work is comparable to the statements of Theorems 1 and 2 that we gave above.

2.5 Empirical Bayes for bin number selection

According to the asymptotic results in Subsection 2.4, the proposed Bayesian approach is guaranteed to be consistent and asymptotically optimal for estimating an h-Hölder intensity function (with $0 < h \leq 1$) if the number of bins satisfies $N \asymp n^{1/(2h+1)}$. However, this choice of the hyperparameter N depends on a proportionality constant. In practice the resulting performance of our Bayesian procedure may turn out to be suboptimal for a given sample size and given dataset, if the constant is not chosen appropriately. Essentially N plays the role of a smoothing parameter. In general, choice of a smoothing parameter constitutes the biggest challenge in non-parametric estimation (see, e.g., Loader (1999)). Chaudhuri & Marron (2000) goes as far as to propose a method 'agnostic' of such a choice. Frequentist papers Henderson (2003) and Leemis (2004), that are related to our work, concentrate on asymptotics of the kernel estimator of the intensity function and do not provide specific practical guidance for selecting the number of bins.

The main idea behind our approach in this section is that the marginal likelihood can be viewed as model evidence in Bayesian statistics. Maximising it over N, roughly speaking, selects a model that is most compatible with the data at hand. This can be seen as an instance of the well-known empirical Bayes method (see, e.g., Efron (2010)).

The marginal likelihood is given by

$$\begin{aligned} \mathrm{ML}_N(X^{(n)}) &= \int_{[0,\infty)^N} L(X^{(n)}; \psi_1, \ldots, \psi_N) \prod_{k=1}^{N} \pi(\psi_k)\, \mathrm{d}\psi_1 \cdots \mathrm{d}\psi_N \\ &= e^{Tn} \frac{\beta^{\alpha N}}{\Gamma(\alpha)^N} \prod_{k=1}^{N} \frac{\Gamma(\alpha + H_k)}{(n\Delta_k + \beta)^{\alpha + H_k}}. \end{aligned} \tag{6}$$

Viewed as a function $N \mapsto \mathrm{ML}_N(X^{(n)})$ (with α and β fixed), the marginal likelihood can thus be easily evaluated and maximised graphically. For numerical stability, it is advisable to work with $\mathrm{LML}_N(X^{(n)}) = \log \mathrm{ML}_N(X^{(n)})$. We study the behaviour of this procedure for selecting N in simulation examples in Section 4.

Alternatively, at first sight, the marginal likelihood can also be used to optimise the hyperparameters α, β of the prior (keeping N fixed). Cf. Clayton & Kaldor (1987) for a somewhat similar idea in a different context than ours. However, we advise against this approach in our setting, as the resulting procedure is plagued by numerical problems. Instead, a numerically stable empirical Bayes procedure can be obtained by maximisation of the marginal likelihood over β for a fixed α (and N). The (unique) maximiser can be found upon setting the derivative of the criterion function with respect to β to zero, which leads to the relation

$$\frac{\alpha}{\beta} = \frac{1}{N} \sum_{k=1}^{N} \frac{H_k + \alpha}{n\Delta_k + \beta}. \tag{7}$$

This has an intuitive interpretation. Namely, the left-hand side is the prior mean of ψ_k, whereas the right-hand side is the average of the posterior means $\hat{\psi}_k$'s, see equation (5). Thus β chosen according to the above rule implies a stability property, whereby the prior mean matches the (average) posterior mean.

As small values of hyperparameters correspond to diffuse priors, one may want to fix α and β at small positive values (from (7) it follows that for small α also β should be small).

3 Gamma Markov chain prior

In this section we propose an alternative Bayesian approach to intensity function estimation. Ideas we use to that end have appeared in various works in the audio signal modelling literature (see, e.g., Cemgil & Dikmen (2007), Dikmen & Cemgil (2010) and Peeling et al. (2008)). They were applied in the volatility estimation setting for diffusion processes in Gugushvili et al. (2019b), where a prior resembling the one below was employed as a conjugate prior for a Gaussian likelihood.

Our starting point is the same as in Section 2. Namely, as in equation (2), we model the intensity function as piecewise constant on bins B_k forming a partition of the interval $[0, T]$. However, instead of assuming that the coefficients ψ_k of the intensity function λ are a priori independent and gamma distributed, we postulate that under a prior they form a gamma Markov chain (GMC). This chain is defined as follows: introduce auxiliary variables $\zeta_k, k = 2, \ldots, N$, use the time ordering $\psi_1, \zeta_2, \psi_2, \ldots, \zeta_N, \psi_N$, and set

$$\psi_1 \sim \mathrm{G}(\alpha_1, \beta_1), \quad \zeta_k \mid \psi_{k-1} \sim \mathrm{IG}(\alpha_\zeta, \alpha_\zeta \psi_{k-1}), \quad \psi_k \mid \zeta_k \sim \mathrm{G}\left(\alpha_\psi, \frac{\alpha_\psi}{\zeta_k}\right), \tag{8}$$

where $k \in \{2, \ldots, N\}$. The name of the chain reflects the fact that its transition distributions are (inverse) gamma. The parameters α_1, β_1, α_ζ and α_ψ are the hyperparameters of the GMC prior. The hyperparameters α_1, β_1 allow one to 'release' the chain at the origin. This is important to avoid possible edge effects in non-parametric estimation due to a strong specification of the prior at the time origin $t = 0$. Next, a principal aim in using latent variables ζ_k's in (8) is to attain positive correlation between ψ_k's. In the intensity function modelling context this induces smoothing across different bins. Depending on whether the ratio α_ζ/α_ψ is less than one, equal to one, or greater than one, the subsequence $\{\psi_k\}$ extracted from the gamma Markov chain exhibits in the limit $N \to \infty$ a decreasing trend, no trend, or an increasing trend, respectively; cf. Cemgil & Dikmen (2007). This feature is attractive in case one possesses prior information on the monotonicity properties of the "true" intensity function λ. Furthermore, large values of the hyperparameters $\alpha_\zeta, \alpha_\psi$ correspond to a strong correlation between ψ_k's and a slow decay of the autocorrelation function, see Figure 1 for an illustration. The GMC prior with large values of $\alpha_\zeta, \alpha_\psi$ has thus a stronger smoothing effect.

3.1 Sampling from the posterior

The posterior distribution with the GMC prior is not available in closed form. This necessitates the use of some approximate posterior inference technique. In this subsection we give the details of the Gibbs sampler to draw (dependent) samples from the posterior.

3.1.1 Full conditional distributions

By the Markov property in (8), the joint distribution of $\{\psi_k\}$ and $\{\zeta_k\}$ factorises as

$$p(\psi_1) \prod_{k=2}^{N} p(\zeta_k \mid \psi_{k-1}) p(\psi_k \mid \zeta_k). \tag{9}$$

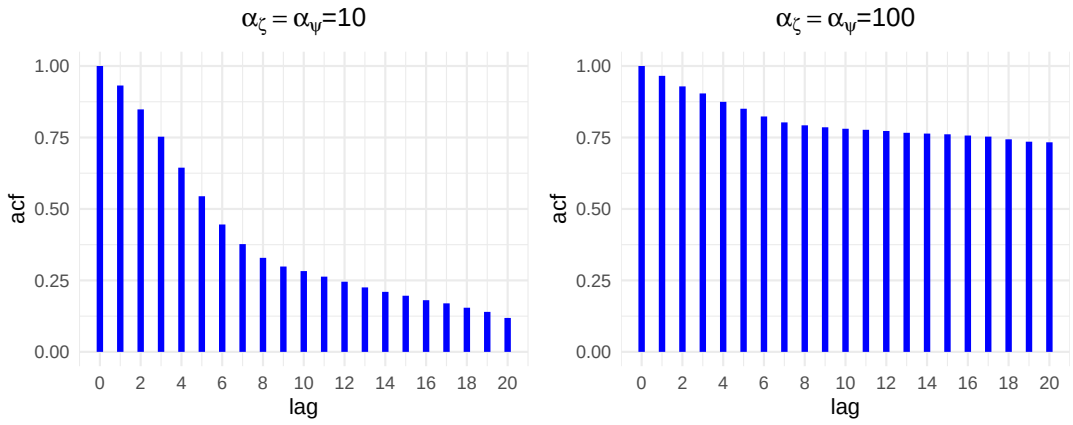

Figure 1: Sample autocorrelation function of $\{\psi_k\}$ based on two realisations of the gamma Markov chain with $N = 1000$. The left plot corresponds to a realisation with $\alpha_\zeta = \alpha_\psi = 10$, the right one to $\alpha_\zeta = \alpha_\psi = 100$. In both cases, $\alpha_1 = 4$, $\beta_1 = 1$.

Using this formula and (8), it can be seen that the full conditional distributions are determined by

$$\begin{aligned}
\zeta_k \mid \psi_k, \psi_{k-1} &\sim \mathrm{IG}\left(\alpha_\zeta + \alpha_\psi, \alpha_\zeta \psi_{k-1} + \alpha_\psi \psi_k\right), \quad k = 2, \ldots, N, \\
\psi_k \mid \zeta_{k+1}, \zeta_k &\sim \mathrm{G}\left(\alpha_\psi + \alpha_\zeta, \frac{\alpha_\psi}{\zeta_k} + \frac{\alpha_\zeta}{\zeta_{k+1}}\right), \quad k = 2, \ldots, N-1, \\
\psi_1 \mid \zeta_2 &\sim \mathrm{G}\left(\alpha_1 + \alpha_\zeta, \beta_1 + \frac{\alpha_\zeta}{\zeta_2}\right), \\
\psi_N \mid \zeta_N &\sim \mathrm{G}\left(\alpha_\psi, \frac{\alpha_\psi}{\zeta_N}\right).
\end{aligned}$$

3.1.2 Gibbs sampler

Combining the preceding results with formula (15), we deduce that given the data $X^{(n)}$, the full conditional distributions of ψ_k, ζ_k are

$$\zeta_k \mid \psi_k, \psi_{k-1} \sim \mathrm{IG}\left(\alpha_\zeta + \alpha_\psi, \alpha_\zeta \psi_{k-1} + \alpha_\psi \psi_k\right), \quad k = 2, \ldots, N, \tag{10}$$

$$\psi_k \mid \zeta_{k+1}, \zeta_k, X^{(n)} \sim \mathrm{G}\left(\alpha_\psi + \alpha_\zeta + H_k, \frac{\alpha_\psi}{\zeta_k} + \frac{\alpha_\zeta}{\zeta_{k+1}} + n\Delta_k\right), \quad k = 2, \ldots, N-1, \tag{11}$$

$$\psi_1 \mid \zeta_2, X^{(n)} \sim \mathrm{G}\left(\alpha_1 + \alpha_\zeta + H_1, \beta_1 + \frac{\alpha_\zeta}{\zeta_2} + n\Delta_1\right), \tag{12}$$

$$\psi_N \mid \zeta_N, X^{(n)} \sim \mathrm{G}\left(\alpha_\psi + H_N, \frac{\alpha_\psi}{\zeta_N} + n\Delta_N\right). \tag{13}$$

The Gibbs sampler (see Geman & Geman (1984) and Gelfand & Smith (1990)) cycles through formulae (10)–(13) to generate (approximate) samples from the posterior distribution of $\{\psi_k\}$, $\{\zeta_k\}$ given the data $X^{(n)}$. One can initialise the sampler e.g. by providing values for $\psi_1, \ldots, \psi_N$.

Remark 1. Comparison of formula (11) to Lemma 1 shows that our two methods for estimating the intensity function match each other when diffuse limits $\alpha = \beta \to 0$ and

$\alpha_\psi = \alpha_\zeta \to 0$ are taken, so that there is a type of continuous transition between the two methods.

3.1.3 Sampler initialisation

Quick convergence of the Gibbs sampler from Subsection 3.1.2 can be facilitated by a good initialisation of the Markov chain. In our simulations we obtained starting values for $\psi_1, \ldots, \psi_N$ by drawing from their posterior distribution based on the independent gamma prior. However, in the simulation examples we considered, even in those cases when we used rather poor initial values for the sampler, it stabilised quite quickly to the 'correct' part of the parameter space (results not shown in this paper).

3.2 Hyperparameters

Assume the number of bins N is fixed. The GMC prior in (8) depends on hyperparameters α_1, β_1, α_ζ and α_ψ. In practice we recommend to use a diffuse prior on ψ_1. Hyperparameters α_1, β_1 can be either both fixed manually, or obtained in a data-driven way from formula (7). There remain two other hyperparameters α_ζ, α_ψ, which we suggest to equip with a prior, and incorporate updates for these hyperparameters in the Gibbs sampler derived in Subsection 3.1.2; cf. Dikmen & Cemgil (2009). Taking for concreteness $\alpha_\zeta = \alpha_\psi$ (we will write α for simplicity, though this leads to a clash with our notation in Section 2 in the case of the independent gamma prior) and denoting this prior by π, we find that the joint density of α, $\{\psi_k\}$, $\{\zeta_k\}$ is $\pi(\alpha)p(\psi_1)\prod_{k=2}^N p(\zeta_k \mid \psi_{k-1})p(\psi_k \mid \zeta_k)$. Using this formula, the unnormalised full conditional density of α given the remaining parameters $\{\psi_k\}$, $\{\zeta_k\}$ is seen to be

$$q(\alpha) \;=\; \pi(\alpha) \times \left(\frac{\alpha^\alpha}{\Gamma(\alpha)}\right)^{2(N-1)} \times \prod_{k=2}^{N} \left(\psi_{k-1}\psi_k\zeta_k^{-2}\right)^\alpha \times \exp\left(-\alpha \sum_{k=2}^{N} \frac{1}{\zeta_k}(\psi_{k-1} + \psi_k)\right).$$

The corresponding normalised density is nonstandard. Hence the full conditional of α is not easily accessible. Therefore we will use a Metropolis-within-Gibbs step (see Tierney (1994)) to update the parameter α when running the Gibbs sampler from Subsection 3.1.2.

As α is nonnegative, we reparametrise α as $\widetilde{\alpha} = \log(\alpha)$ and note that the unnormalised full conditional density of $\widetilde{\alpha}$ is $\widetilde{\alpha} \mapsto e^{\widetilde{\alpha}} q(e^{\widetilde{\alpha}}) = \widetilde{q}(\widetilde{\alpha})$. Once one designs an update algorithm for $\widetilde{\alpha}$, samples for α can be obtained by exponentiation.

We propose to use a Gaussian random walk proposal on $\widetilde{\alpha}$. As a prior on the parameter α, we recommend to use a distribution with not too light right tail, next to having sufficient mass near zero. The former because large values of α may be necessary to adequately regularise the non-parametric estimation problem if the number of bins is large. On the other hand, if the prior puts mass close to zero, then the prior allows the method to choose a model similar to the independent gamma prior.

3.3 Bin number selection

There remains a choice to be made for the hyperparameter N, i.e. the number of bins. As with our first approach using independent gamma priors on the coefficients ψ_k's, here too the bin number can in principle be optimised via the marginal likelihood. However, unlike our first approach, the marginal likelihood is not available in closed form. Although for

any fixed N it can be estimated from the posterior simulation output (see, e.g., Chib (1995), Chib & Jeliazkov (2001) and references therein), this is far from trivial. However, as we demonstrate in Section 4, the inferences are quite stable with respect to the choice of N. In fact, for a fixed value of N, the GMC method will balance the amount of smoothing by tuning the hyperparameter α from the data. Such a shift from the discrete model selection problem to tuning a single continuous hyperparameter is often advantageous from the computational point of view.

In the examples which deal with small or moderate size datasets (of several hundred Poisson points), we use the rule-of-thumb

$$N = \min\left(50, \left\lceil \frac{\mathcal{H}}{4} \right\rceil\right). \tag{14}$$

Here

$$\mathcal{H} = \sum_{j=1}^{n} \sum_{i=1}^{m_j} \mathbf{1}_{\{X_{ij} \in [0,T]\}}$$

is the total number of Poisson points. It is advisable not to make the bins B_k too small. While increasing α gives to realisations from the prior more regularity, the exact trade-off between this and the number of bins is difficult to quantify. The upper truncation with 50 in (14) is somewhat arbitrary. It was sufficient to obtain convincing results in our simulation examples. This upper bound can be easily modified to a larger number and does not involve a significant computational overhead. Cf. our results in Subsection 4.1.1.

4 Simulation examples

In this section we study the performance of our non-parametric Bayesian procedures on simulated data examples. We implemented our procedures in **Julia**, see Bezanson et al. (2017). The computer code and datasets for replication of our examples forms part of the PointProcessInference package, see Gugushvili et al. (2019)[2]. For plotting we used functionalities of the **ggplot2** package (see Wickham (2016)) in **R** (see R Core Team (2018)). The computations were performed on a MacBook Pro, with a 2.7GHz Intel Core i5 with 8 GB RAM.

Full specifications used for synthetic data generation and posterior inference are given in each example below. For the method based on the GMC prior, we ran the Gibbs sampler for 30000 iterations. The first half of the generated posterior samples was then discarded as a burn-in, and the posterior inference was based on the second half of the samples. A Gaussian random walk proposal was used to update the hyperparameter α, with variance scaled in such a way so as to ensure the acceptance rate between $25\% - 50\%$ in the Metropolis-within-Gibbs step.

In general, in the ensuing plots the "true" intensity function is represented by a solid red line. The posterior mean is given by a solid black line, while 95% marginal credible bands are shaded in light blue.

4.1 Oscillating exponential function

In our first example, we consider

$$\lambda_0(x) = 2e^{-x/5}(5 + 4\cos(x)), \quad x \in [0, 10].$$

[2]For additional details see `https://github.com/mschauer/PointProcessInference.jl`

A principal challenge in inferring this function consists in the fact that it takes small values in the middle part of its domain and has a slope of changing sign. In Figure 2 we plot our estimation results with sample size $n = 1$ (the total number of Poisson points was $\mathcal{H} = 46$). In this figure, as well as in subsequent ones, the rug plot on the top displays the event times. For the independent gamma prior we took $\alpha = \beta = 0.1$. For the GMC prior we took $\alpha_1 = \beta_1 = 0.1$ and $\alpha \sim \text{Exp}(0.1)$. For the independent gamma prior, the optimal choice for N obtained by maximising the log marginal likelihood was $N = 4$. See the top panels in Figure 2. In Figure 3 a plot of this log marginal likelihood (ignoring the irrelevant term e^{Tn}) versus N is shown. Quick convergence of the Gibbs sampler can be seen from the trace and autocorrelation plots for several parameters in Figure 4.

Estimation results for the sample size $n = 5$ are reported in Figure 5 (the total number of Poisson points was $\mathcal{H} = 215$). Here in the top panel N for the independent gamma prior was determined using the empirical Bayes method.

The following conclusions can be gleaned from the simulation results:

- Performance of the empirical Bayes method for selecting N is not particularly encouraging. For moderate sample sizes that we considered, it oversmooths by selecting a visually too small N.
- For larger N, posterior means obtained with the independent gamma prior tend to show more fluctuation than those obtained with the GMC prior.
- For larger N, marginal posterior bands with the GMC method tend to be narrower than those obtained with the independent gamma prior.
- Inferential conclusions with the GMC prior, as reflected in marginal posterior bands, appear to be robust with respect to the choice of N, provided this is not chosen exceedingly large or small. In particular, the rule-of-thumb (14) appears to work in practice. On the other hand, this robustness property is not shared by the method based on the independent gamma prior.

If the independent gamma prior for ψ_k is chosen in a more informative way, determining N via the empirical Bayes method seems to perform better compared to the case when a diffuse gamma prior is used on ψ_k. In Figure 6 we chose $\alpha = 2$ and $\beta = 1$, which led to the optimal value for N being equal to 9 (which is close to the value 12, chosen by the rule-of-thumb (14) for the GMC prior). Interestingly enough, in this case the posterior band obtained with the independent gamma prior is somewhat narrower than the one with the GMC prior, although visually the latter reflects uncertainty in estimation results better than the former. A difficulty with the empirical Bayes method is that a sensible informative prior on ψ_k might not be available in practice.

4.1.1 Scalability

Our next goal was to illustrate scalability of our methods with big data. In Figures 7 and 8 we plot estimation results with a very large sample size $n = 4000$, that resulted in $\mathcal{H} = 177781$ Poisson points (we omit posterior means from the plots, as they obfuscated the resulting (narrow) marginal credible bands). As noted e.g. in Adams et al. (2009) and Teh & Rao (2011), samples of this size are far beyond the computational reach of their methods. On the other hand, both our methods perform excellently in terms of estimation quality. That the method based on the independent gamma prior is very fast

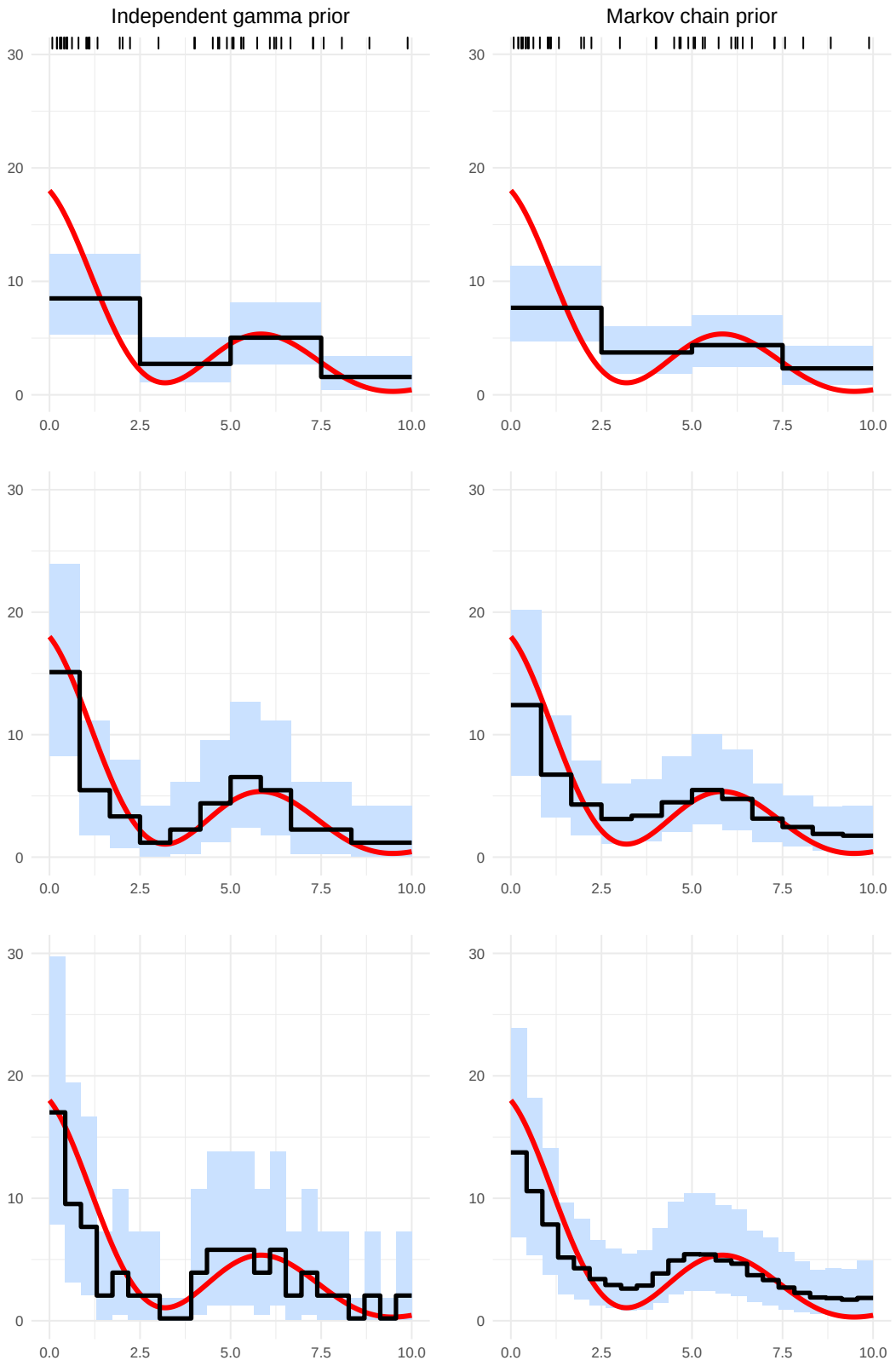

Figure 2: Estimation results for the oscillating exponential function λ_0 from Subsection 4.1 with $n = 1$. In the top row, $N = 4$ is selected via maximising the marginal likelihood as in Subsection 2.5, in the middle row $N = 12$ via (14), and in the bottom row $N = \mathcal{H}/2 = 23$. The prior settings were: $\alpha = 0.1$ and β determined from (7) for the independent gamma prior, and $\alpha_1 = \beta_1 = 0.1$ and $\alpha \sim \text{Exp}(0.1)$ for the GMC prior.

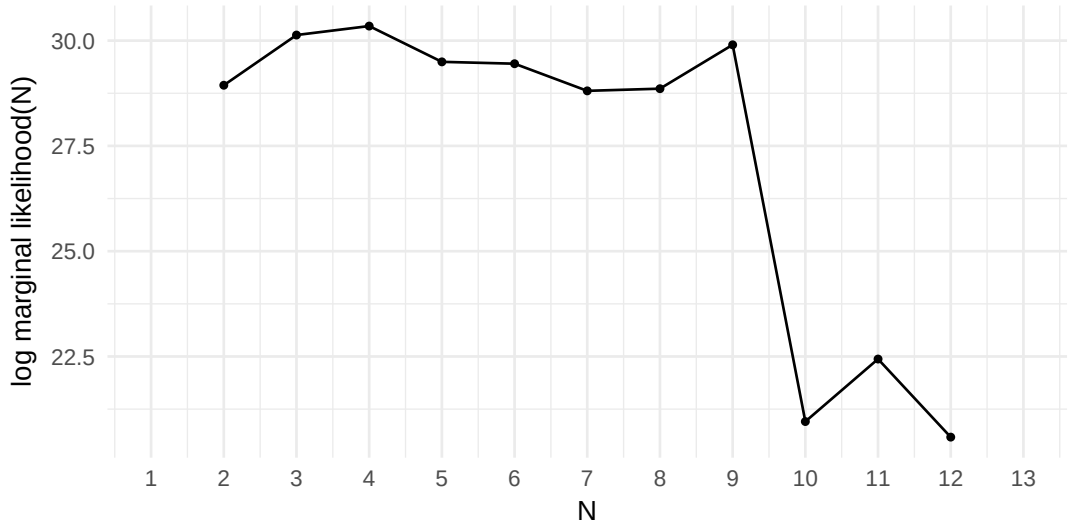

Figure 3: Simulation example of the oscillating exponential function λ_0 from Subsection 4.1 with $n = 1$: logarithm of the marginal likelihood as a function of N (up to an additive constant independent of N), with $\alpha = \beta = 0.1$. The maximum is attained at $N = 4$.

(for Figure 7 we used $N = 48$) comes as no surprise. We also timed the method based on the GMC prior, specifically its part for running the Gibbs sampler, which task was completed in ca. 4 seconds for $N = 200$ (used for Figure 8).

In this example, given the huge amount of data, it makes sense to experiment with a large number of bins. We took $N = 1000$ and compared two choices for a prior on α in Figure 9. In the right panel we took the Lévy distribution (which is just the IG$(1/2, 1/2)$ distribution), while in the left panel we used the Exp(10) distribution as prior on α. These distributions are very different, but only after zooming in and carefully studying the credible bands can one see a small difference: the band produced with the Lévy distribution appears to be smoother. These results were expected: a larger value for α induces a stronger positive correlation in a realisation from the gamma Markov chain prior. With 1000 bins, the computing time was approximately 9 seconds. We conclude that our methods scale well with sample size.

4.1.2 Sensitivity

In Figure 10 we illustrate sensitivity of the GMC procedure with respect to the choice of the prior on α by comparing estimation results with three different choices of the prior (we took $N = 12$, which follows from (14)). All three priors lead to qualitatively similar estimation results. In in all cases the GMC-based procedure calibrates the parameter α from the data, as evidenced by comparison of the prior density of α to the corresponding posterior density. Cf. also Figure 9.

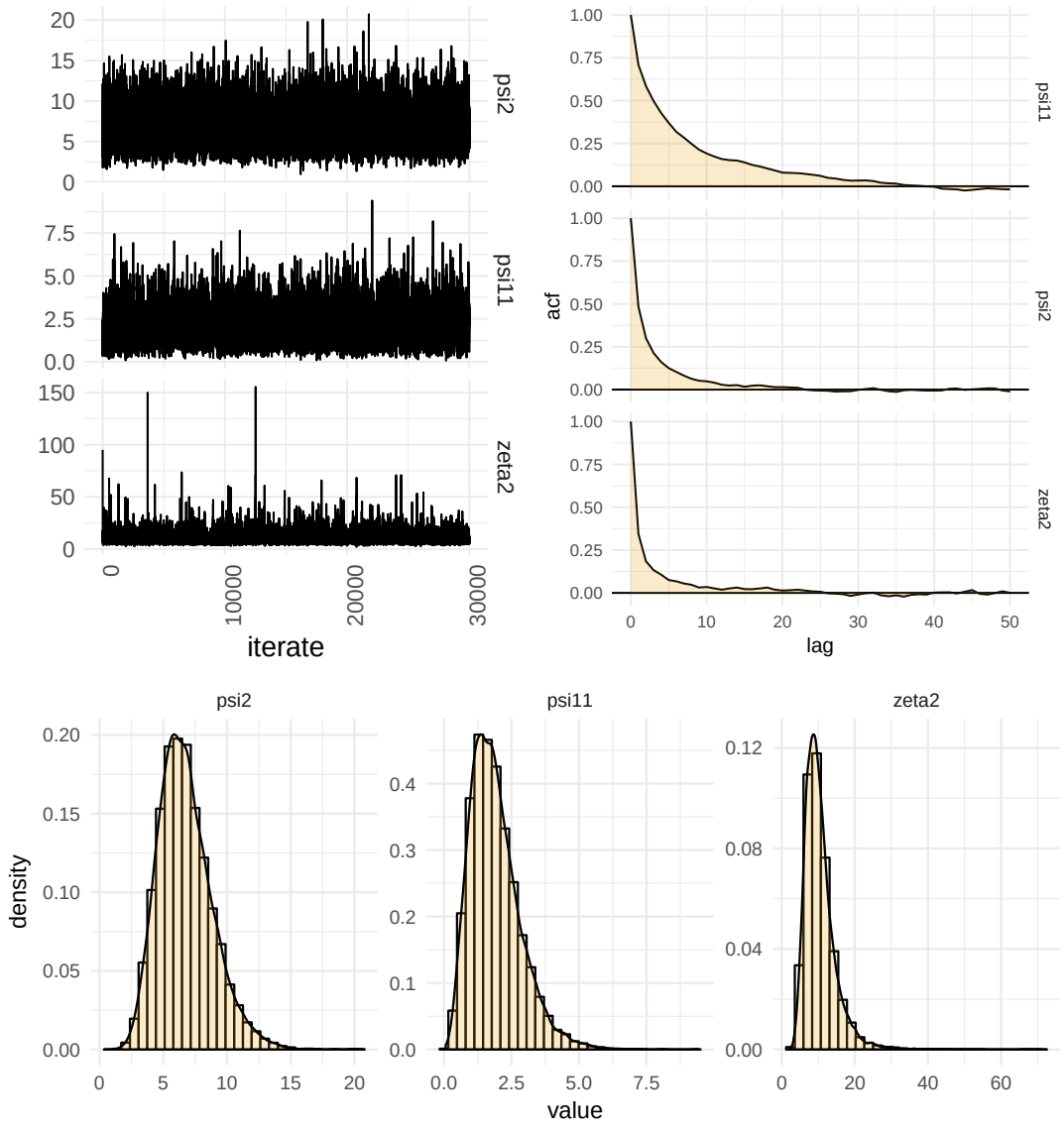

Figure 4: Posterior simulation output for estimating the oscillating exponential function λ_0 from Subsection 4.1, with $n = 1$. Trace plots, autocorrelation plots and histograms for ψ and ζ. For the histograms and autocorrelation plots the first half of the samples has been discarded as burn-in.

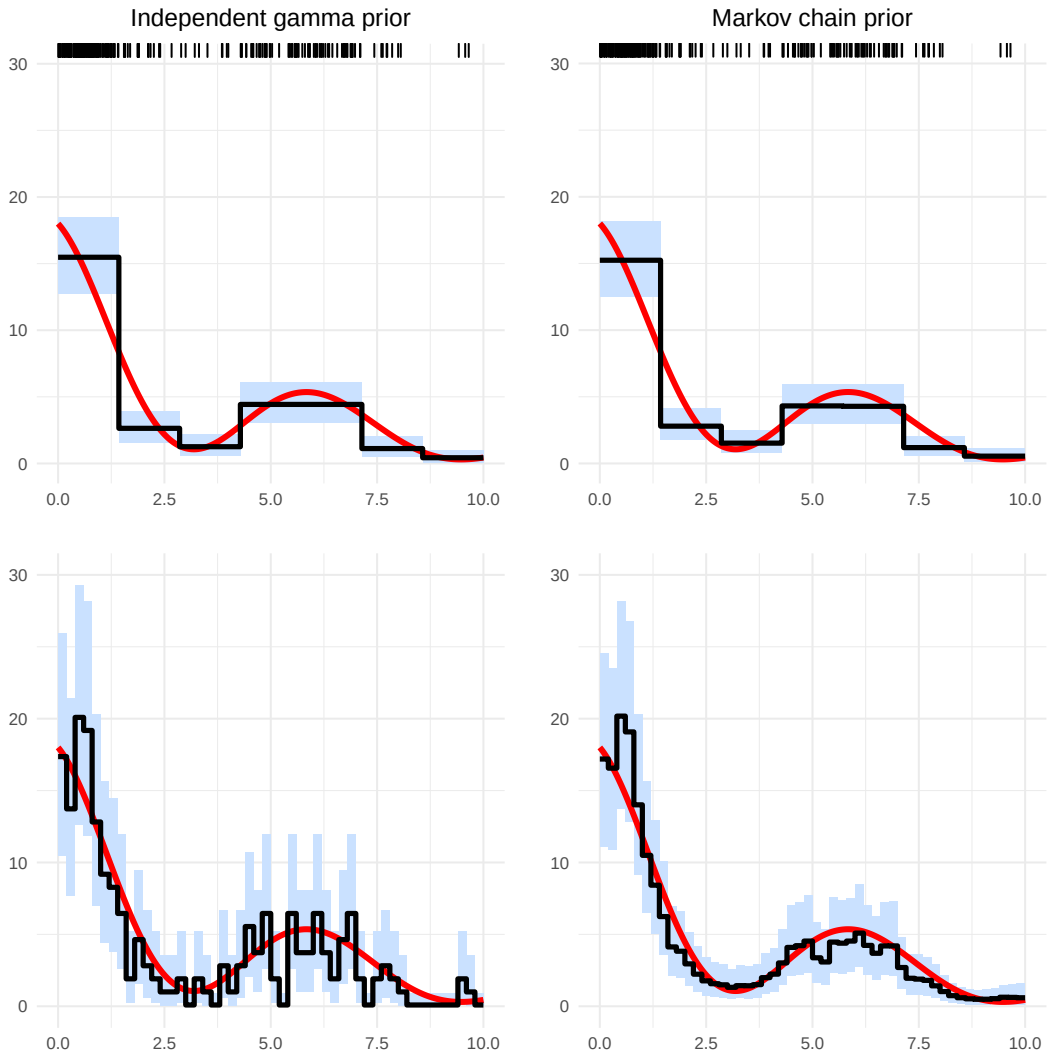

Figure 5: Estimation results for the oscillating exponential function λ_0 from Subsection 4.1 with $n = 5$. The top plot corresponds to the method based on the independent gamma prior, with $N = 7$ chosen via the empirical Bayes method. The bottom plot corresponds to the method based on the GMC prior, with $N = 50$ chosen via (14). Other settings are as in the case $n = 1$ as in Figure 2.

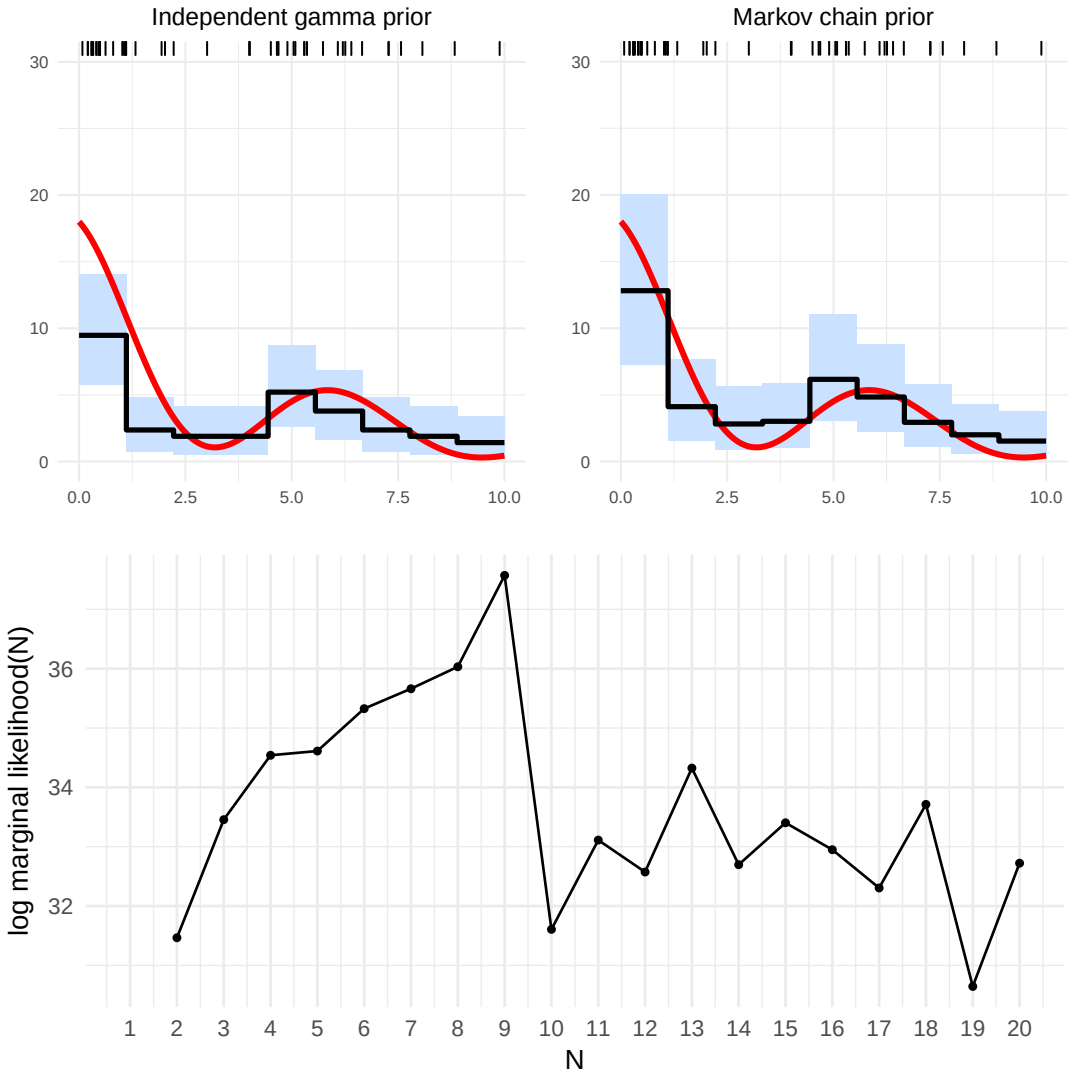

Figure 6: Estimation results for the oscillating exponential function λ_0 from Subsection 4.1 with $n = 1$. The prior distribution on each ψ_k is taken to be $\mathrm{G}(2, 1)$. The log marginal likelihood is plotted as a function of N in the bottom panel (up to an additive constant independent of N). The optimal number of bins chosen via the empirical Bayes method is $N = 9$.

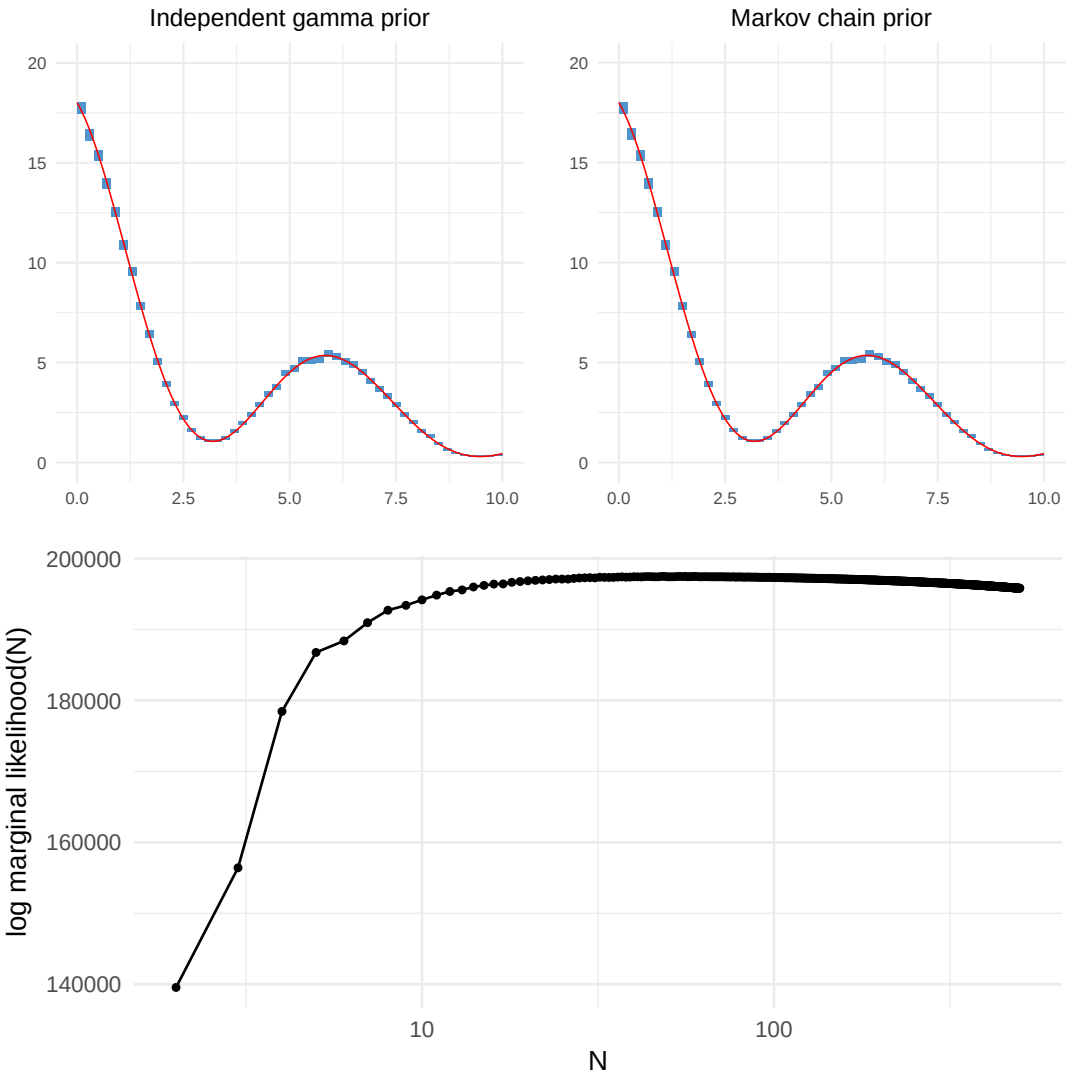

Figure 7: Estimation results for the oscillating exponential function λ_0 from Subsection 4.1 with $n = 4000$. The bin number $N = 48$ chosen via the empirical Bayes method, other hyperparameters determined as in the case $n = 1$. The top panel gives the plots of λ_0 and the 95% marginal credible bands. The bottom panel gives the plot of the log marginal likelihood as a function of N (up to an additive constant independent of N).

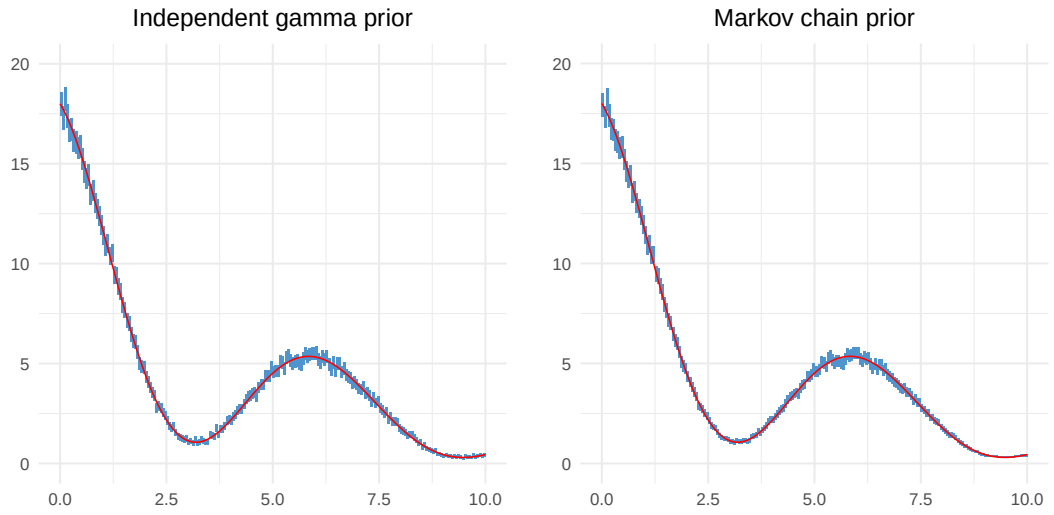

Figure 8: Estimation results for the oscillating exponential function λ_0 from Subsection 4.1 with $n = 4000$. $N = 200$, other hyperparameters determined as in the case $n = 1$.

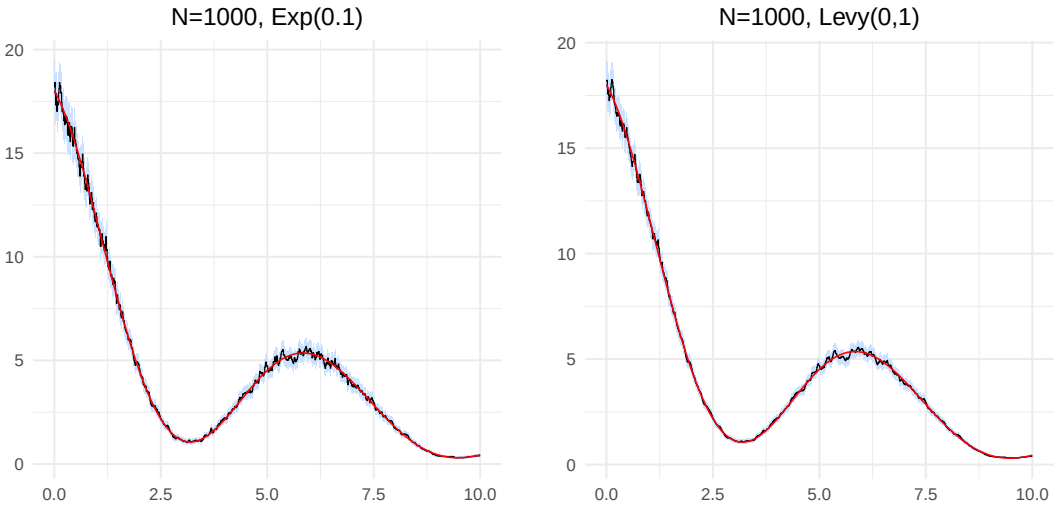

Figure 9: Estimation results for the oscillating exponential function λ_0 from Subsection 4.1 with $n = 4000$ using the gamma Markov chain prior. The bin number was fixed at $N = 1000$. Both plots show λ_0 and 95% marginal credible bands, where the true intensity is the red curve. In the left panel we took the Exp(10) distribution as prior on α, whereas in the right panel we used the standard Lévy distribution as prior on α.

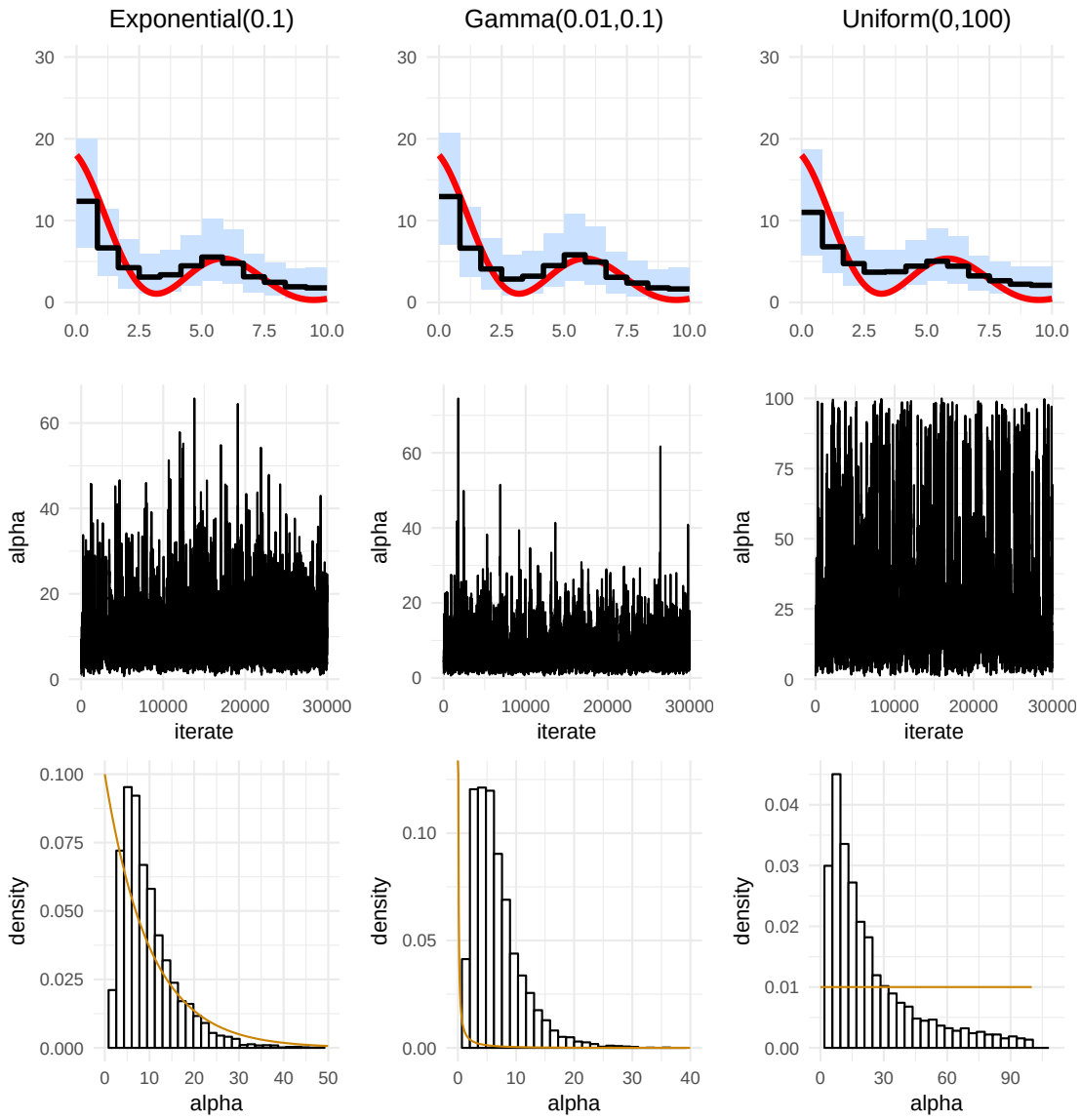

Figure 10: Estimation results for the oscillating exponential function λ_0 from Subsection 4.1 with $n = 1$ using the GMC prior. The hyperparameters were $N = 12$ and $\alpha_1 = \beta_1 = 0.1$. The plots show the effect the prior on α has on inference results. In the left column, $\alpha \sim \text{Exp}(0.1)$, in the middle $\alpha \sim \text{G}(0.01, 0.1)$, and in the right $\alpha \sim \text{Uniform}(0, 100)$. Top row: true intensity function (red solid), posterior mean (black solid) and 95% marginal credible bands (light blue). Middle row: trace plots of α. Bottom row: histogram of posterior samples of α (first half of the samples was discarded as a burn-in), with a prior density of α is added in orange.

4.2 Bart Simpson function

As our second example, we consider the Bart Simpson intensity function defined as

$$\lambda_0(x) = \frac{1}{2}\phi(x; 3, 1) + \frac{1}{10}\sum_{j=0}^{4} \phi\left(x; \frac{j}{2} + 2, \frac{1}{10}\right), \quad x \in [0, 6].$$

This example has been adapted from the non-parametric density estimation context in Chapter 6 in Wasserman (2006). The Bart Simpson function is a mixture of Gaussian densities, interpreted as a half times the density $\phi(x; 3, 1)$ with a number of superimposed peaks. The latter can be thought of as corresponding to bursts of increased activity of a Poisson point process.

We give estimation results with sample sizes $n = 200$ and $n = 500$ in Figure 11 (the number of Poisson points was $\mathcal{H} = 200$ and $\mathcal{H} = 491$, respectively) for both the independent gamma prior and the GMC prior. Additionally, in Figure 12 we report the results when the number of bins with the independent gamma prior is selected via the empirical Bayes method. We note that the same optimal number of bins $N = 7$ was obtained when we used a more informative prior $\psi_k \overset{\text{i.i.d.}}{\sim} \mathrm{G}(2, 1)$ too. We do not provide the corresponding plots, as they did not indicate a message different from Figure 12.

The following conclusions can be deduced from the figures:

- As in the case of the oscillating exponential function from Subsection 4.1, the empirical Bayes method to select N performs visually unsatisfactorily.
- Posterior means of both of our estimation methods pick up nicely the peaks and valleys of the Bart Simpson function (provided for the independent gamma prior we use a sufficiently large number of bins N; both for the independent gamma prior and the GMC prior N is chosen using our rule-of-thumb (14)). For the sample size $n = 200$ and $n = 500$, the posterior mean corresponding to the independent gamma prior does this somewhat better than the one corresponding to the GMC prior. On the other hand, the marginal credible bands for the GMC prior are tighter, and they also show far less spurious variability.

4.3 Practical recommendation

The synthetic data examples considered in previous subsections allow us to formulate a practical recommendation. Specifically, we saw that both our methods are fast, scale with the sample size and perform well in non-trivial inference problems. However, this good performance is conditional on an appropriate choice of the bin number N. In that respect, results we obtained for the independent gamma prior, when N was chosen via the empirical Bayes method, were mixed. On the other hand, the method based on the GMC prior showed substantial stability with respect to a choice of N. In particular, a simple rule-of-thumb (14) led to good practical results. In light of these observations, we recommend the method based on the GMC prior as a default estimation strategy.

5 Real data examples

In this section, we apply our approach with the GMC prior on three real datasets. In light of Subsection 4.3, we do not consider the method based on the independent gamma prior.

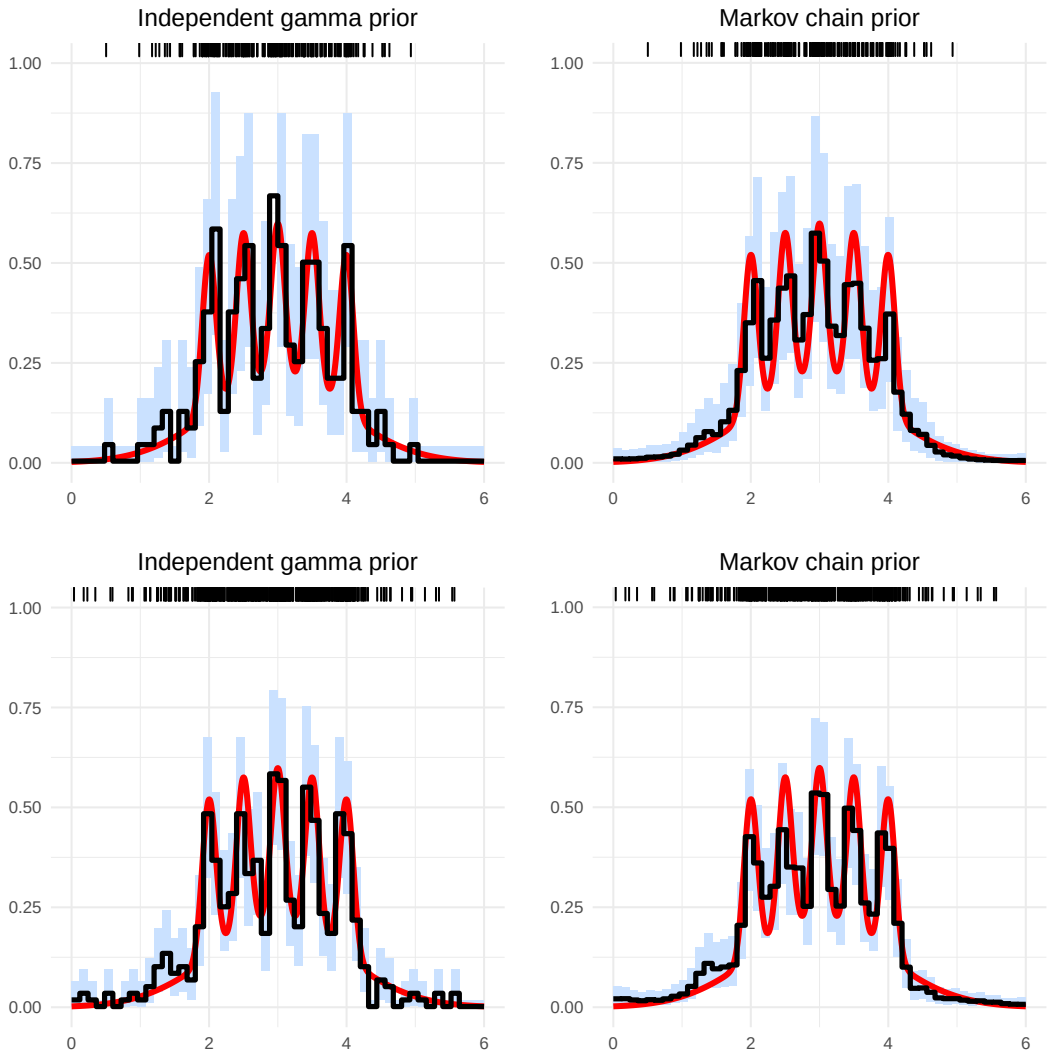

Figure 11: Estimation results for the Bart Simpson function λ_0 from Subsection 4.2. The hyperparameters are $N = 50$, $\alpha = \beta = 0.1$ for the independent gamma prior, and $N = 50$ chosen via (14), $\alpha_1 = \beta_1 = 0.1$ and $\alpha \sim \mathrm{G}(1, 0.1)$ for the GMC prior. Top row: $n = 200$, bottom row: $n = 500$.

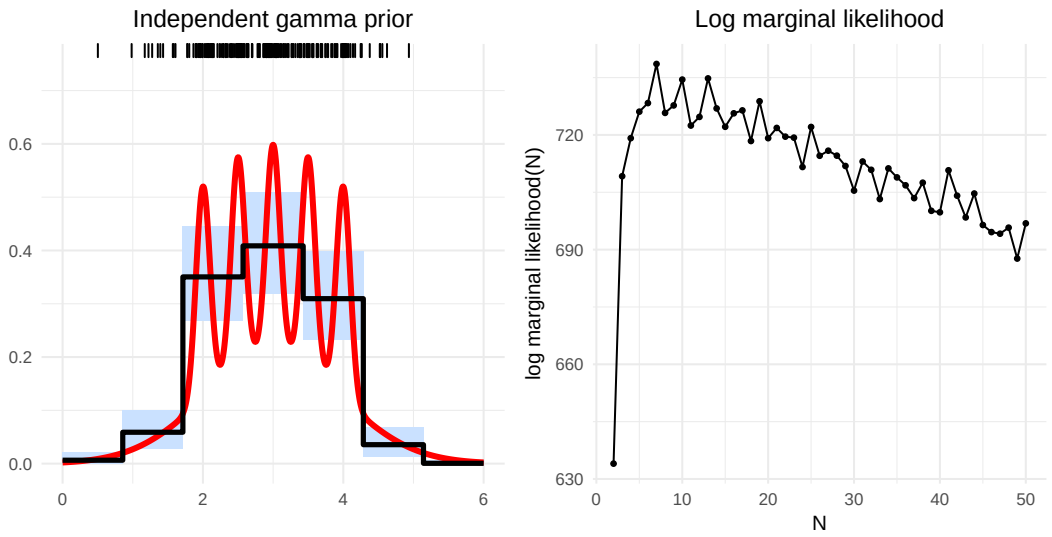

Figure 12: Estimation results for the Bart Simpson function λ_0 from Subsection 4.2, using the independent gamma prior. The sample size was $n = 200$, the bin number $N = 7$ was determined via the empirical Bayes method from Subsection 2.5, and the remaining hyperparameters were $\alpha = \beta = 0.1$. The right panel displays the logarithm of the marginal likelihood as a function of N (up to an additive constant independent of N).

Unless explicitly stated otherwise, we took $\alpha_1 = \beta_1 = 0.1$, N as in (14) and $\alpha \sim \text{Exp}(0.1)$. In each case, we ran the Gibbs sampler for 30000 iterations, and based the posterior inference on the second half of the generated posterior samples, with the first half dropped as a burn-in.

5.1 UK coal mining disasters data

The dates of coal mining disasters (defined as accidents with 10 or more fatal casualties, 191 events in total) in Britain between 15 March 1851 and 22 March 1962 provided in Jarrett (1979) serve as a modern benchmark for point process inference. The dataset is accessible in **R** as *coal* in the **boot** package; see Canty & Ripley (2017).

The data have been analysed in Green (1995) for change-point detection (it should be noted that change-point estimation is a different inferential task from the one studied in this paper). A historical perspective on change-point analysis for this problem is given in Raftery & Akman (1986), where it is suggested that an observed decrease in the accident rate over 1851–1962 is mainly due to an abrupt decrease around the years 1887–1895, possibly associated with changes in the coal industry around that time, such as decline in labour productivity (e.g., due to overtime) starting at the end of 1880s, and the emergence of the Miners' Federation in 1889. As a further possible reason, Lloyd et al. (2015) indicate passing of the Coal Mines Regulation Acts of 1872 and 1887 by the UK parliament with the aim of improving safety for mine workers. A more precise treatment would require adjustment for the coal mining industry size[3]. This is not attempted neither in the above,

[3]Some relevant historical data are available at `https://www.gov.uk/government/statistical-data-sets/historical-coal-data-coal-production-availability-and-consumption` (Accessed on 22 June 2019), although the figures start to become annual only from 1913 on.

nor in the following works, that give a machine learning perspective on inference for the coal mining disasters data: Adams et al. (2009), Hensman et al. (2015), Kom Samo & Roberts (2015) and Teh & Rao (2011).

The rule-of-thumb (14) led to $N = 48$ bins. The resulting posterior mean and the 75% and 95% marginal posterior credible bands are shown in Figure 13. The computing time for the Gibbs sampler was under a second. Our estimation results are, broadly speaking, similar to those already reported in the literature. Hence we only present a brief comparison with two state-of-the-art Bayesian methods from Adams et al. (2009) and Lloyd et al. (2015) (the corresponding posterior means were read off Figure 3 in Lloyd et al. (2015) via WebPlotDigitizer 4.1, see Rohatgi (2017)). Our conclusions are as follows:

- We believe that the method from Adams et al. (2009) does not pick up well enough the behaviour of the intensity function in the neighbourhood of the year 1851, which follows upon a comparison of the posterior mean to the rug plot of data points. Likewise, the posterior mean seems to drop off too sharply starting from mid-1950ies, at the right boundary point of the observation window, this being corroborated also by the method from Lloyd et al. (2015). Both these features of the posterior mean from Adams et al. (2009) are possibly instances of edge effects (see, e.g., Fan & Gijbels (1996) for a discussion of edge or boundary effects in non-parametric estimation). This boundary behaviour, but only at the left endpoint, is demonstrated by the method from Lloyd et al. (2015) as well.

- The method from Adams et al. (2009) appears to oversmooth in the years 1900–1925, which also follows by a comparison to the method from Lloyd et al. (2015). As a consequence, the method from Adams et al. (2009) does not differentiate the World War I period in any particular way from the immediately preceding or successive years. This is unlike the World War II period, where it is in global agreement with the other two methods.

- The posterior means from Adams et al. (2009) and Lloyd et al. (2015) do pass through the 95% marginal credible band constructed via our GMC method.

- The 75% marginal posterior bands for both our method and the method from Lloyd et al. (2015) look similar, except a disagreement on the first two bins, and a somewhat higher accident rate during the World War II period. As explained above, we believe the former to be due to an edge effect.

5.2 US mass shootings data

In this subsection we will apply our GMC Bayesian method to analyse the US mass shootings data collected by the US nonprofit organisation Mother Jones, see Follman et al. (2018a) and Follman et al. (2018b). The data cover the years 1982–2018 and provide extensive information on mass shootings in the US over that period. A mass shooting is defined as a single attack in public place in which four or more victims were killed (see Follman et al. (2018a)). This definition is essentially based on the FBI crime classification report from 2005. The definition excludes armed robbery, gang violence, or domestic violence, and focusses on shooting incidents where the motive appears to be indiscriminate mass murder (see Follman (2012)). We note that for the data collected over

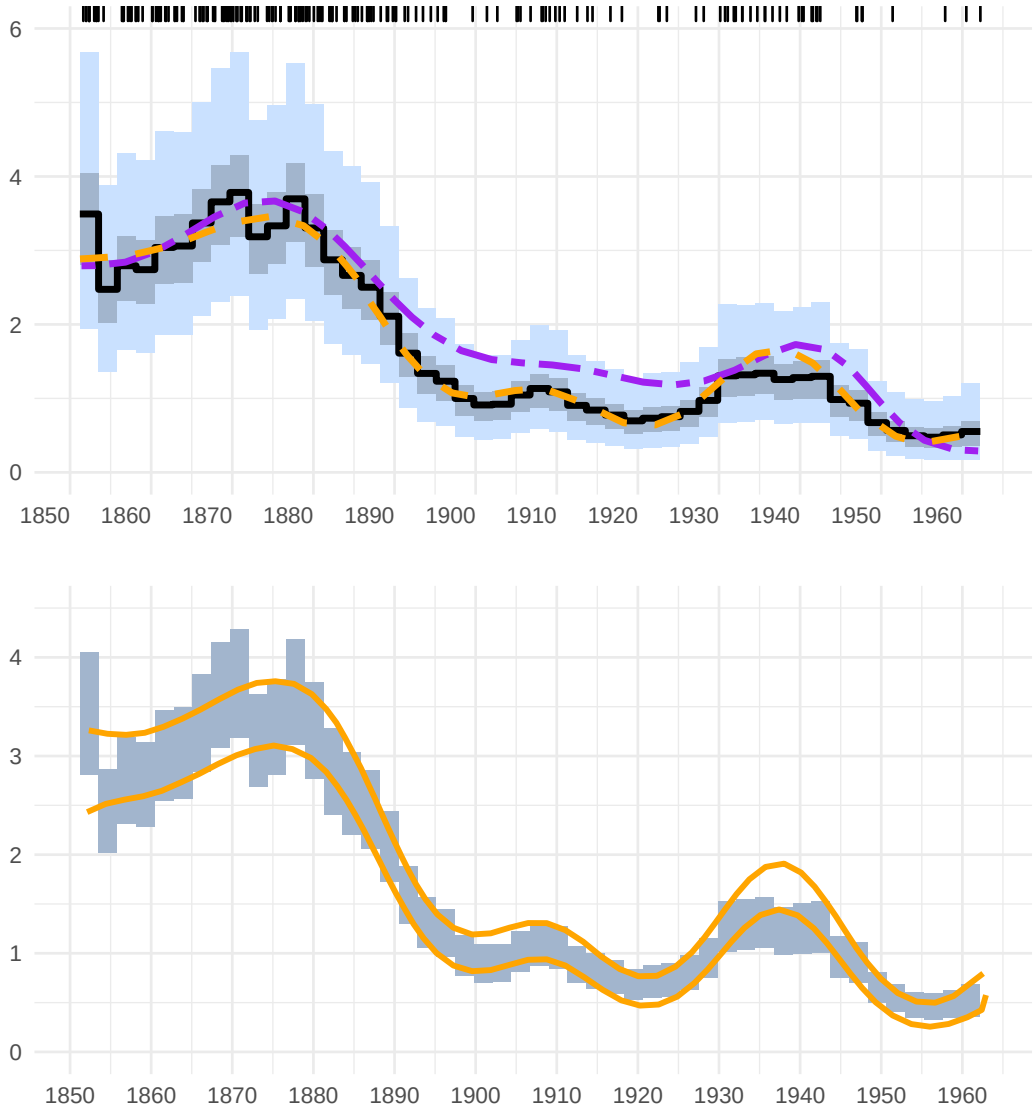

Figure 13: Coal mining disasters data. In the top plot displayed are the posterior mean (black curve) with 75% (grey) and 95% (light blue) credible bands. The dashed purple curve is the estimate from Adams et al. (2009), the dashed yellow line is an estimate from Lloyd et al. (2015). In the bottom plot the 75% marginal credible band is compared to that from Lloyd et al. (2015) (visualised via two yellow lines).

2013–2018, Mother Jones lowered the baseline of four fatal victims to three, reflecting a revised federal law.[4] In our analysis, for uniformity in methods for data collection, we maintained the older definition. However, we acknowledge the fact that a precise definition of a mass shooting is somewhat elusive. This will affect inclusion or exclusion of specific shooting cases in any quantitative analysis.

The dataset used in our analysis consists of 85 shooting incidents over the period of 20 August 1982 – 14 February 2018, that lead to 4 or more lethal casualties each. We model occurrences of mass shootings as realisations from a Poisson point process. This approach is in line with modelling the coal mining disasters data via a Poisson point process, see Subsection 5.1, as well as using the Poisson distribution in Spiegelhalter (2019), pages 248–251, to model the UK homicide rates. We set 1 January 1982 and 14 March 2018 as the start and end times of the observation period.

We note that the US mass shootings data have already been analysed in Cohen et al. (2014) using a different approach (Shewhart's chart), and data up to 2014. A conclusion reached in Cohen et al. (2014) is that the rate of mass shooting incidents increased since September 2011. Such a conclusion is further corroborated by the FBI study of active shooter incidents in the 2000–2013 period, see Blair & Schweit (2014).

We present our estimation results in Figure 14 both for $N = 9$ bins and $N = 21$ bins. The first setting corresponds to considering bins of 4 years, whereas the second relies on the rule-of-thumb (14). The Gibbs sampler runs were completed under a second. Both of the presented plots suggest, roughly speaking, an increasing intensity function, which can be taken as a substantial evidence of an increasing trend in occurrences of mass shootings. In fact, posterior means indicate a four-fold increase in the Poisson process intensity over 1982–2018. Part of this increase can possibly be correlated with an increase in the US population between 1982 to 2018 (from 232 to 327 million), but nevertheless the relative increase in population is considerably smaller than that in the intensity of mass shooting incidents. Contrast this to Figure 15, where we display the number of homicides by firearms in the US over the years 1999–2017 (data for years prior to 1999 was not accessible to us)[5]; this has stayed relatively stable. Especially after around 2005 the shooting incidents appear to occur at an ever accelerating rate. A fuller treatment of the question would require inclusion of the population size as a covariate in the model, perhaps in combination with other covariates too. See, e.g., the right panel of Figure 15, where we display the yearly average numbers of homicides per 100,000 inhabitants in the US[6]. This, however, lies outside the scope of the present work, that deals exclusively with the Poisson point process inference.

[4]Investigative Assistance for Violent Crimes Act of 2012. Pub. L. 112–265. 126 Stat. 2435–2436. 14 January 2013. https://www.gpo.gov/fdsys/pkg/PLAW-112publ265/content-detail.html (Accessed on 24 March 2018)

[5]Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying Cause of Death 1999–2017 on CDC WONDER Online Database, released December, 2018. Data are from the Multiple Cause of Death Files, 1999–2017, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov/ucd-icd10.html on June 22, 2019.

[6]We downloaded the population data from: World Bank, Population, Total for United States [POPTOTUSA647NWDB], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/POPTOTUSA647NWDB,September14,2019.

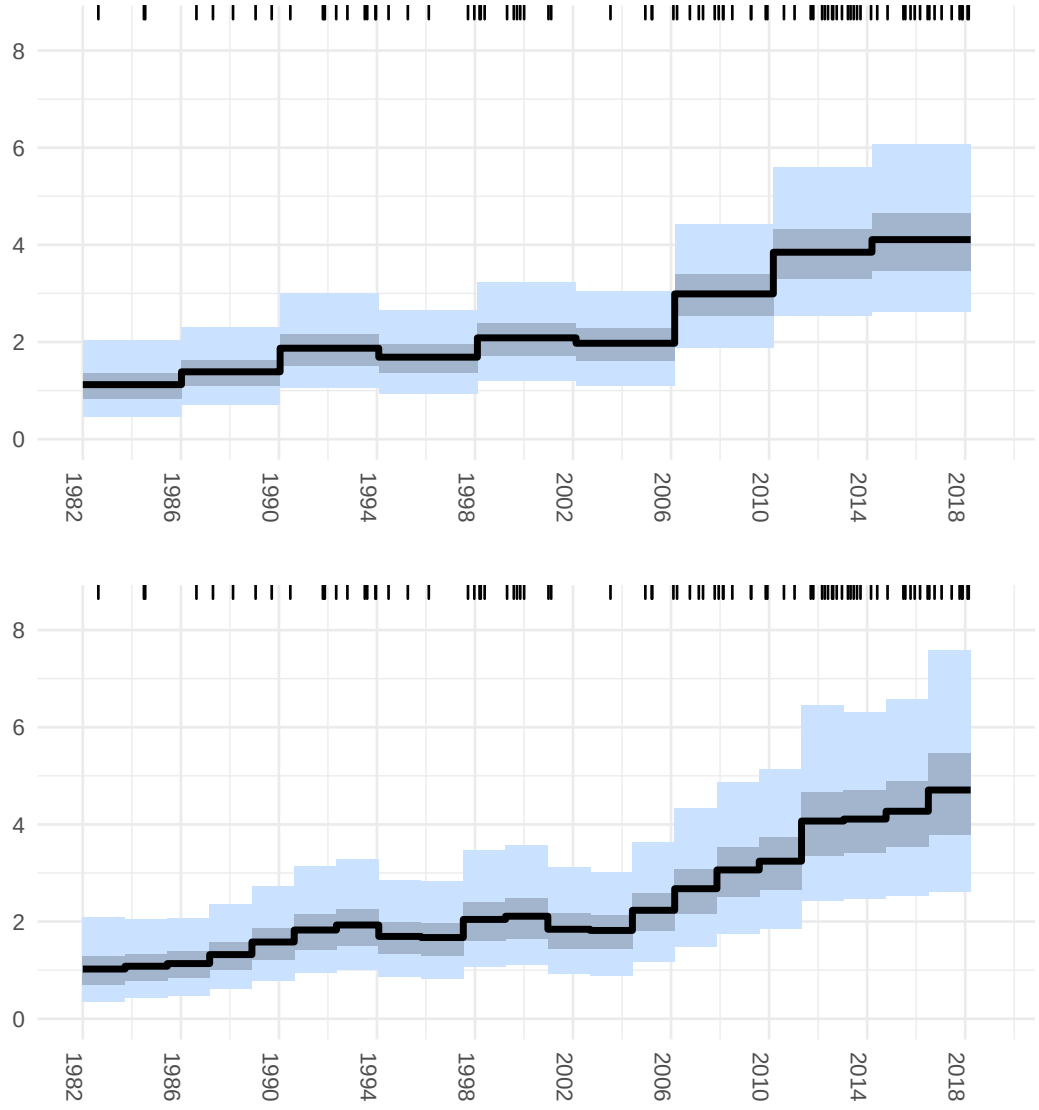

Figure 14: US mass shootings data. The rug plot on the top displays the event times. Displayed are the posterior mean (black curve) with 75% (grey) and 95% (light blue) credible bands. Top: $N = 9$ bins (one bin for each 4 years). Bottom: $N = 21$ bins (according to the rule for determining N given in equation (14)).

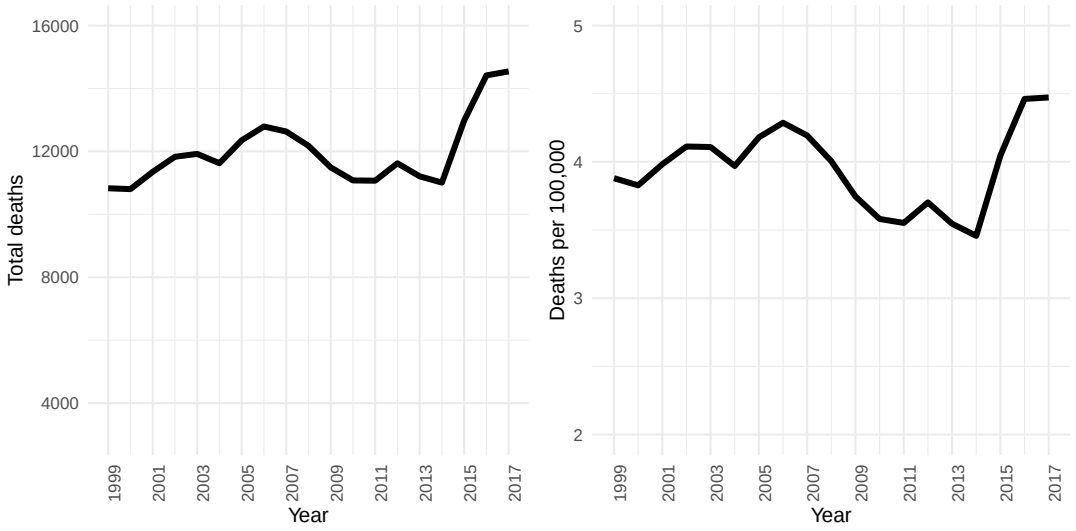

Figure 15: Left panel: yearly number of homicides by firearms in the US over the years 1999–2017. Right panel: average number of homicides per 100,000 inhabitants.

5.3 Trump's Twitter data

Social media constitutes a natural field for application of point processes (cf. Lloyd et al. (2015) and Kom Samo & Roberts (2015)). In this subsection, we will analyse the tweet data from Donald J. Trump's personal Twitter account `@realDonaldTrump`. President Trump is an active Twitter user[7]. His tweet data has already been a subject of quantitative studies in the past. Thus, using text mining tools, David Robinson (see Robinson (2016) and Robinson (2017)) analysed the question whether tweets posted on `@realDonaldTrump` from Android and iPhone devices belong to different persons: Android tweets were posted or dictated by Donald Trump personally,[8] whereas the iPhone tweets were written by his staff members. This question can be approached from the point process angle as well. Specifically, for each device type (Android and iPhone) we will model tweet arrival times as realisations from a Poisson point process. We will then infer the intensity functions and compare them. Significant differences in shapes of the intensity functions can be taken as an indicator of differing tweeting habits of the Android and iPhone users.

We model tweet times as realisations from a periodic Poisson process with a period of one day. The data we used in our analysis were all the tweets made from the Android and iPhone devices in the period between 16 June 2015 (the official launch of Donald Trump's presidential campaign) and 9 November 2016 (the date he won the presidential elections). For interpretability, we took $N = 48$ bins, which corresponds to bins of half an hour (note that the rule (14) would lead to $N = 50$, a minor difference). The number of analysed tweets using Android and iPhone equals 563 and 952 respectively. Completion of each of the Gibbs samplers took less than a second. Our estimation results are given in Figure 16. The most prominent differences between the two intensity functions we

[7] A complete archive of Donald Trump's tweets, updated on an hourly basis, is available at `https://github.com/bpb27/trump_tweet_data_archive`

[8] It is known that Donald Trump's personal mobile device was an Android, possibly a Samsung Galaxy S3. See, e.g., `https://www.androidcentral.com/which-android-phone-does-donald-trump-use` (Accessed on 26 March 2018)

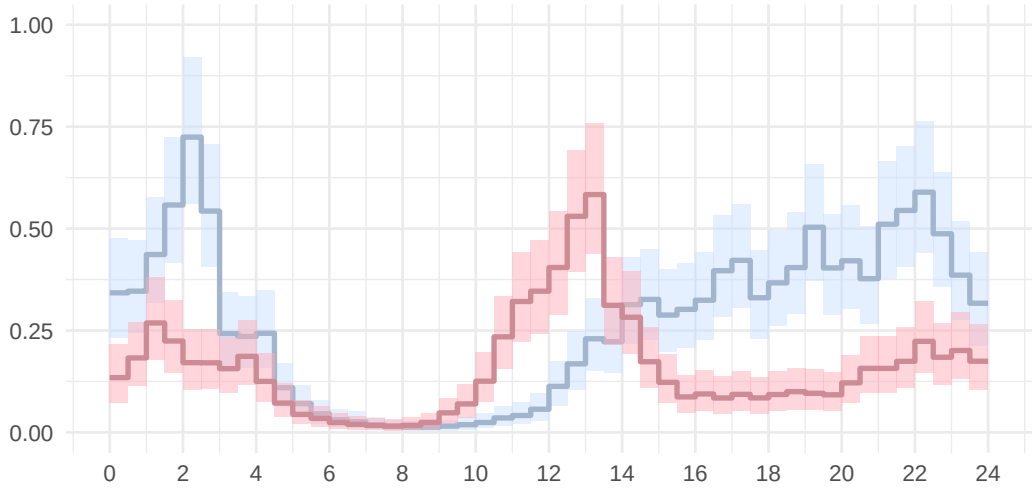

Figure 16: Estimation results for the Twitter data from Subsection 5.3. Displayed are the posterior means with 95% credible bands. Blue: iPhone tweet data, pink: Android tweet data. The observation period in both cases is 16 June 2015 – 9 November 2016.

inferred are that the iPhone user shows a high tweeting activity during the night hours and in the second half of the day. On the other hand, the Android user's main activity falls in the morning and during early afternoon hours.

In Figure 17 we compare the intensity function inferred from iPhone tweets made after 25 March 2017 (the date of the last Android tweet. As confirmed a few days later by Dan Scavino Jr., the White House Director of Social Media and Assistant to the President, Donald Trump stopped using an Android device and switched to an iPhone[9]) until 1 January 2018 to the intensity function inferred from the combined Android and iPhone data over the period 16 June 2015 – 9 November 2016, as well as only the Android data. A conclusion that emerges from the graphs is that although the inferred intensity functions do not match each other exactly, general tweeting habits of the Android user during 16 June 2015 – 9 November 2016 and the iPhone user after 25 March 2017 are qualitatively similar. Specifically, tweeting activity over the night hours is largely the same for both these datasets, which also show a surge in tweeting around 11:00–13:00, and a subsequent decrease until a few hours prior to the midnight. On the other hand, observed differences between the two intensity functions can be plausibly explained by the fact that past the 25 March 2017 date, the iPhone is possibly still used by President Trump's aides, primarily in the morning and during early afternoon hours.

6 Discussion

In this paper we studied two related non-parametric Bayesian approaches to estimation of the intensity function of a Poisson point process. Our methods are easy to implement, in that the posterior with the first method is available in closed form, whereas with the

[9]See, e.g., `https://www.businessinsider.nl/donald-trump-switches-apple-iphone-from-unsecure-android-2017-3/` (Accessed on 26 March 2018)

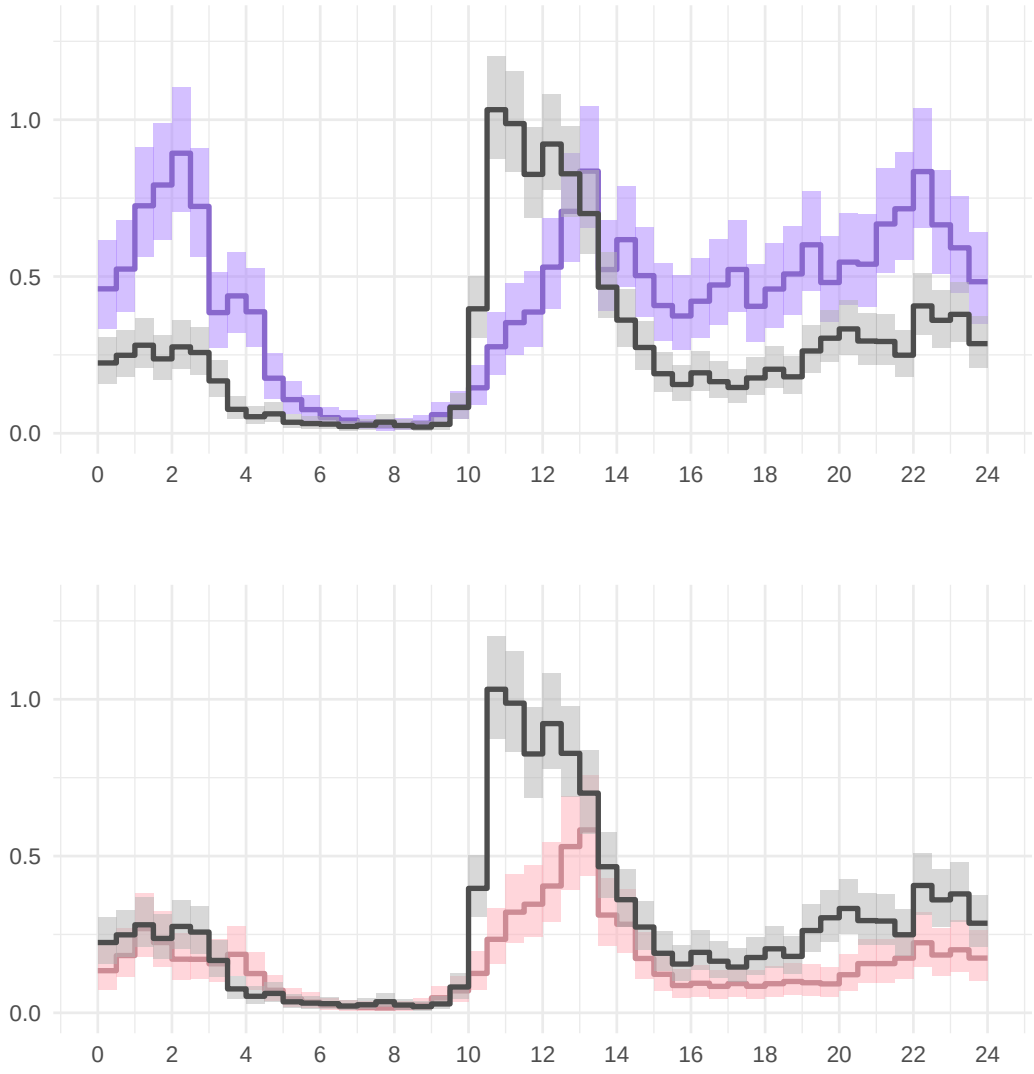

Figure 17: Estimation results for the Twitter data from Subsection 5.3. Top plot: displayed are in black the posterior mean with a 95% credible band for all iPhone tweets over the period 3 March 2017 (date of last Android tweet) – 1 January 2018, and in purple the posterior mean with a 95% credible band for the combined iPhone and Android tweets over the period 16 June 2015 – 9 November 2016. Bottom plot: displayed are in black the posterior mean with a 95% credible band for all iPhone tweets over the period 3 March 2017 – 1 January 2018, and in pink the posterior mean with a 95% credible band for the iPhone tweets over the period 16 June 2015 – 9 November 2016.

second method the posterior inference can be performed using a straightforward implementation of the Gibbs sampler. We believe that our methods are conceptually simpler than those previously developed in the statistical literature. There also exists a solid body of machine learning research dedicated to inference in Poisson point processes. The methods used in this strand of the literature are largely based on optimisation techniques and the variational Bayes approach (see Lloyd et al. (2015)), or combinations of these techniques with MCMC (see Hensman et al. (2015)). However, they may underreport uncertainties in parameter estimates, which is an issue with the variational inference (see Blei et al. (2017)). In that sense, the MCMC-based machine learning approaches to inference in point process models, such as Adams et al. (2009) and Teh & Rao (2011), hold an advantage over those employing optimisation techniques. On the downside, they typically do not scale well with large amounts of data. In contrast, implementation of our methods in **Julia** is fast, with hundreds of thousands of data points and a large number of bins N posing no significant computational challenges. Finally, we demonstrated good performance of our approach on simulated data examples and three real datasets: coal mining disasters data, the US mass shootings data and Donald Trump's Twitter data.

Two promising future research directions that we envision are an extension of our methods to a multi-dimensional setting and incorporation of covariates in the model. On the theoretical side, working out frequentist asymptotics for the method based on the GMC prior is interesting.

Acknowledgments

The research leading to the results in this paper has received funding from the European Research Council under ERC Grant Agreement 320637. We would like to thank Adeline Leclercq Samson (Université Grenoble Alpes, Grenoble, France), Ryan Martin (North Carolina State University, Raleigh, North Carolina, USA) and Paulo Serra (Eindhoven University of Technology, Eindhoven, The Netherlands) for their constructive comments, that led to various improvements in the draft version of the paper. We have benefited from discussions on the GMC prior with Ali Taylan Cemgil (Boğaziçi University, Istanbul, Turkey).

A Proofs of the technical results

Proof of Lemma 1. With $\lambda(x) = \sum_{k=1}^{N} \psi_k \mathbf{1}_{B_k}(x)$, we get for the likelihood (as displayed in (1))

$$\begin{aligned} L(X^{(n)};\lambda) &= \left(\prod_{j=1}^{n}\prod_{i=1}^{m_j}\lambda(X_{ij})\right) \times \exp\left(nT - n\int_0^T \lambda(x)\,\mathrm{d}x\right) \\ &\propto \left(\prod_{j=1}^{n}\prod_{i=1}^{m_j}\left[\sum_{k=1}^{N}\psi_k \mathbf{1}_{B_k}(X_{ij})\right]\right) \times \prod_{k=1}^{N} e^{-n\Delta_k\psi_k}. \end{aligned}$$

As the factor with the double product can be rewritten as $\prod_{k=1}^{N}\psi_k^{H_k}$, we get

$$L(X^{(n)};\lambda) \propto \prod_{k=1}^{N}\psi_k^{H_k} e^{-n\Delta_k\psi_k}. \tag{15}$$

The result now follows in a straightforward way by conjugacy of the prior, see Assumption 1. □

Proof of Theorem 1. Recall from Subsection 2.2 that $B_k = [b_{k-1}, b_k)$. We will use the fact that

$$H_k \sim \text{Poisson}\left(n \int_{B_k} \lambda_0(x)\,\mathrm{d}x\right), \quad k = 1, \ldots, N,$$

which follows by independence of $X_1, \ldots, X_n$. Hence,

$$\mathbb{E}[H_k] = \mathbb{V}\text{ar}\,[H_k] = n \int_{B_k} \lambda_0(x)\,\mathrm{d}x, \quad k = 1, \ldots, N. \tag{16}$$

For notational simplicity we assume that $T = 1$. This entails no loss of generality. Recall from (5) that the posterior mean is piecewise constant with value equal to $\hat{\psi}_k = (n\Delta + \beta)^{-1}(H_k + \alpha)$ on the k-th bin B_k.

Introduce

$$\bar{\lambda}_0(x) = \sum_{k=1}^{N} \lambda_0(b_{k-1}) \mathbf{1}_{B_k}(x).$$

First we note

$$\mathbb{E}[\|\hat{\lambda} - \lambda_0\|_2^2] \leq 2\left(\|\lambda_0 - \bar{\lambda}_0\|_2^2 + \mathbb{E}[\|\hat{\lambda} - \bar{\lambda}_0\|_2^2]\right). \tag{17}$$

Using Hölder-continuity of λ_0 (see Assumption 3), we get for the first term on the right-hand side that

$$\|\lambda_0 - \bar{\lambda}_0\|_2^2 = \sum_{k=1}^{N} \int_{B_k} (\lambda_0(x) - \lambda(b_{k-1}))^2\,\mathrm{d}x \leq L^2 \Delta^{2h}.$$

For the second term on the right-hand side of (17) we have

$$\begin{aligned}
\mathbb{E}[\|\hat{\lambda} - \bar{\lambda}_0\|_2^2] &= \Delta \sum_{k=1}^{N} \mathbb{E}[(\hat{\psi}_k - \lambda_0(b_{k-1}))^2] \\
&= \Delta \sum_{k=1}^{N} \{\mathbb{E}[\hat{\psi}_k] - \lambda_0(b_{k-1})\}^2 + \Delta \sum_{k=1}^{N} \mathbb{V}\text{ar}\,[\hat{\psi}_k].
\end{aligned}$$

By Hölder-continuity of λ_0, for $x \in B_k$, $\lambda_0(x) = \lambda_0(b_{k-1}) + O\left(\Delta^h\right)$. Combining this with (16), we obtain

$$\mathbb{E}[H_k] = n\Delta\lambda_0(b_{k-1}) + nO(\Delta^{1+h}), \quad k = 1, \ldots, N, \tag{18}$$

so that

$$\mathbb{E}[\hat{\psi}_k] - \lambda_0(b_{k-1}) = O\left(\frac{1}{n\Delta} + \Delta^h\right) = O\left(\Delta^h\right), \quad k = 1, \ldots, N.$$

Here the last equality follows from the choice of Δ. Furthermore, again by (16) and Assumption 3,

$$\mathbb{V}\text{ar}\,[\hat{\psi}_k] = \frac{1}{(n\Delta + \beta)^2} \mathbb{V}\text{ar}\,[H_k] \lesssim \frac{n\Delta}{(n\Delta + \beta)^2} \lesssim \frac{1}{n\Delta}.$$

Thus,

$$\mathbb{E}[\|\hat{\lambda} - \bar{\lambda}_0\|_2^2] \lesssim \Delta^{2h} + \frac{1}{n\Delta}.$$

The statement of the theorem follows from the fact that $\Delta \asymp n^{-1/(2h+1)}$. □

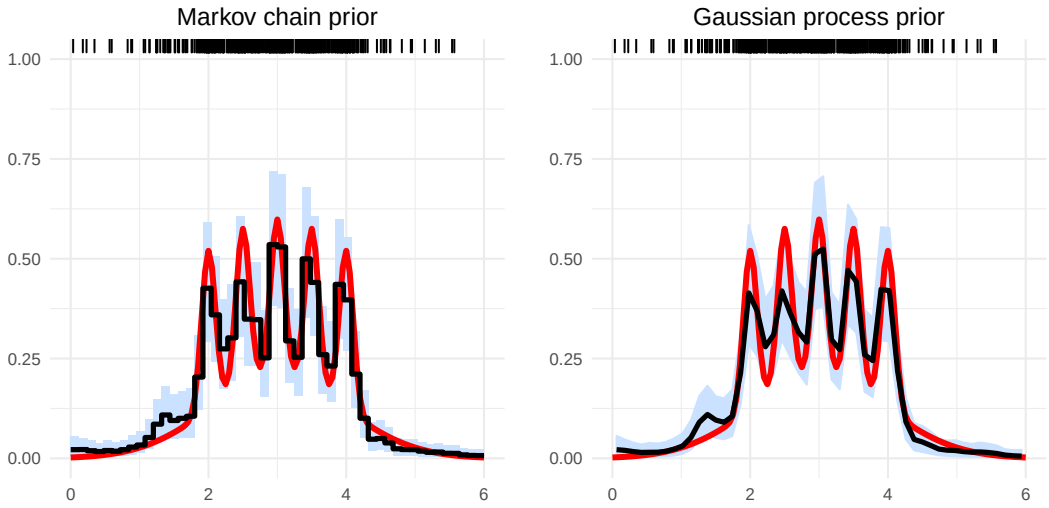

Figure 18: Estimation results for the Bart Simpson function λ_0 from Subsection 4.2, with the same $n = 500$ replicated observations as used in that section. The settings for the GMC prior are described in the caption of Figure 11. The Gaussian process prior is taken to have zero mean and covariance given by a Matérn 3/2 kernel, with its hyperparameters (on the log scale) assigned the independent $N(-2, 4)$ prior distributions. $N = 50$ bins are used to bin the count data and the GP posterior summaries are assigned to the midpoints of these bins.

Proof of Theorem 2. By Chebyshev's inequality,

$$\mathbb{E}[\Pi_N(\|\lambda - \lambda_0\|_2 \geq M_n\varepsilon_n \mid X^{(n)})] \leq \frac{1}{M_n^2\varepsilon_n^2}\mathbb{E}\left[\mathbb{E}_{\Pi_N}\left(\|\lambda - \lambda_0\|_2^2 \mid X^{(n)}\right)\right]. \tag{19}$$

Then, by the posterior bias-variance decomposition,

$$\mathbb{E}\left[\mathbb{E}_{\Pi_N}\left(\|\lambda - \lambda_0\|_2^2 \mid X^{(n)}\right)\right] = \mathbb{E}\left[\|\hat{\lambda} - \lambda_0\|_2^2\right] + \Delta\sum_{k=1}^{N}\mathbb{E}[\mathbb{V}\mathrm{ar}_{\Pi_N}(\psi_k \mid X^{(n)})]. \tag{20}$$

We have for the second term on the right-hand side, using (4), a formula for the variance of a gamma distribution and (18) that

$$\begin{aligned}\Delta\sum_{k=1}^{N}\mathbb{E}\left[\mathbb{V}\mathrm{ar}_{\Pi_N}\left(\psi_k \mid X^{(n)}\right)\right] &= \Delta\sum_{k=1}^{N}\mathbb{E}\left[\frac{H_k + \alpha}{(n\Delta + \beta)^2}\right] \\ &= \frac{\Delta}{(n\Delta+\beta)^2}\sum_{k=1}^{N}\mathbb{E}[H_k] + O\left(\frac{1}{(n\Delta)^2}\right) = O\left(\frac{1}{n\Delta}\right) = O(n^{-2h/(2h+1)}).\end{aligned}$$

As far as the first term on the right-hand side of (20) is concerned, by Theorem 1 it is also of the order $n^{-2h/(2h+1)}$. Hence so is (20). Using this fact, the statement of the theorem now follows from equation (19). □

B Gaussian processes

Although the emphasis of the present paper is not on a comparison with other methods for point process inference per se, in this appendix we briefly report on the results we obtained with intensity estimation based on a Gaussian process prior (see the literature overview in Subsection 1.2). As a test model, we chose the Bart Simpson function and used the same simulated dataset as in Subsection 4.2, corresponding to $n = 500$ observations. For numerical evaluation, we relied on the **GaussianProcesses.jl** package of Fairbrother et al. (2018). As detailed in Section 4.3 of that technical report, the approach requires binning of the point process data (similar to our methods), and the model for the intensity of the process assumes the form $\Lambda_k = \exp(f_k)$, for $k = 1, \ldots, N$, where Λ_k is the (integrated) intensity of the process on the kth bin, whereas f_k's are assigned the Gaussian process (GP) prior. Posterior simulation relies on the Hamiltonian Monte Carlo (HMC) algorithm. From the assumed model it follows that in our notation, $\psi_k = e^{f_k}/(n\Delta)$ for $k = 1, \ldots, N$.

To make results comparable with ours in Subsection 4.2, we used $N = 50$ bins. The GP had mean zero and a Matérn 3/2 kernel, with its hyperparameters (on the log scale) assigned the independent $N(-2, 4)$ prior distributions. These choices correspond to the ones made in Section 4.3 in Fairbrother et al. (2018), and are not claimed to be optimal. After a trial-and-error procedure, we set the integration step size of HMC to 0.025. This yielded a ca. 50% acceptance rate of HMC. We kept other tuning quantities at their default values in **GaussianProcesses.jl**. We ran the Markov chain for 30000 iterations and dropped the first third of the posterior samples as burn-in; no thinning was used (cf. our settings in Section 4). The posterior summaries were assigned to the midpoints of the bins.

The results are displayed in Figure 18. Overall, the performance of the GMC prior-based method is quite similar to that of the GP-based one in this example and for this dataset. Our method appears to retain the heights of the sharp peaks of the Bart Simpson function somewhat better than the GP-based one, though admittedly the difference is marginal to be relevant in practice. More importantly, our method turned out to be much faster, with its computation completed in a fraction of a second, whereas the timing was in minutes for the GP-based approach.

C Imposing a prior on the number of bins

Rather than fixing the number of bins N, it can be treated as an unknown parameter. From a Bayesian point of view this is a natural approach and the computational problem is commonly tackled using the reversible jump algorithm. Investigation of the performance of this approach, combined with the independent gamma prior, was suggested to us by Ryan Martin.

As different values of N correspond to different binning of the data we write $H_k^{(N)}$ and $\Delta_k^{(N)}$ to highlight dependence on N. The marginal likelihood of the model indexed by N is given by (see Equation (6))

$$\mathrm{ML}_N(X^{(n)}) = e^{Tn} \frac{\beta^{\alpha N}}{\Gamma(\alpha)^N} \prod_{k=1}^{N} \frac{\Gamma(\alpha + H_k^{(N)})}{(n\Delta_k^{(N)} + \beta)^{\alpha + H_k^{(N)}}},$$

Let π_N be the prior probability of the model indexed by N. Define

$$R(N) = \log \mathrm{ML}_N(X^{(n)}) + \log \pi_N.$$

Suppose $q(N^\circ \mid N)$ is a Markov kernel on the model indices. Let the current iterate be N with value $R(N)$. The reversible jump algorithm iterates over the following steps:

(i) Select a new model N° from $q(\cdot \mid N)$.

(ii) Compute
$$A = R(N^\circ) - R(N) + \log q(N \mid N^\circ) - \log q(N^\circ \mid N).$$

(iii) Draw $U \sim \mathrm{Uniform}(0,1)$. If $\log U < A$, set N to N°.

This iterative scheme yields a sequence $\{N_i\}$ from the posterior of the model index (here i is the Monte-Carlo iteration number). Within a model indexed by N, coefficients can be sampled from their posterior

$$\psi_k^{(N)} \sim \mathrm{G}\left(\alpha + H_k^{(N)}, \beta + n\Delta_k^{(N)}\right), \qquad k \in \{1, \ldots, N\}.$$

Note that existence of a closed form expression for $\mathrm{ML}_N(X^{(n)})$ greatly simplifies derivation of the reversible jump algorithm.

For each proposed model N the binning vector $(H_1^{(N)}, \ldots, H_N^{(N)})$ needs to be computed. However, for a particular value of N this only needs to be done once over all MCMC iterations.

To illustrate performance of the method, we revisit the example of Subsection 4.1 and compare the results to those that were displayed in Figure 2.

The proposal q on the model index was defined as follows: Fix $\eta \in (0, 1/2)$. If the present model index is ≥ 2, we propose to go one index up or one index down with probability η. With probability $1 - 2\eta$ we propose to stay within the present model. When the present model index is 1, with equal chances we stay in this model or move up to model 2.

We ran the sampler for 30000 iterations, discarding the first half of the iterates as burn-in. We chose $\eta = 0.45$. Within each model we assumed the $\mathrm{G}(0.1, 0.1)$ prior distribution on ψ_k's. We considered two different choices of the prior on N:

(i) $N \sim \mathrm{DiscreteUniform}(\{1, \ldots, 50\})$;

(ii) $N \sim \mathrm{ShiftPoisson}(23)$, by which we mean the Poisson(23) distribution restricted to $\{1, 2, \ldots\}$.

The choice 23 was motivated by the lower panel in Figure 2, where the number of bins equals 23. Plots of the posterior along with marginal credible bands are in Figure 19. The results do not appear to be satisfactory. Especially for the uniform prior on N the posterior mean is visually oversmoothing. Under both priors, realisations from the posterior do not have any visual smoothness property, in the sense that conditional on the model index, the heights over the bins are both a priori and a posteriori independent. Hence our preference for the GMC prior.

To illustrate difficulties with using a prior on the number of bins, we report a numerical experiment where we took $\lambda(x) = 2 + 0.2\sin(30x) + \mathbf{1}_{[0.7,1]}(x)$ and $n = 10000$. We employed the gamma Markov chain prior with $\mathrm{Exp}(0.1)$ on the smoothing parameter

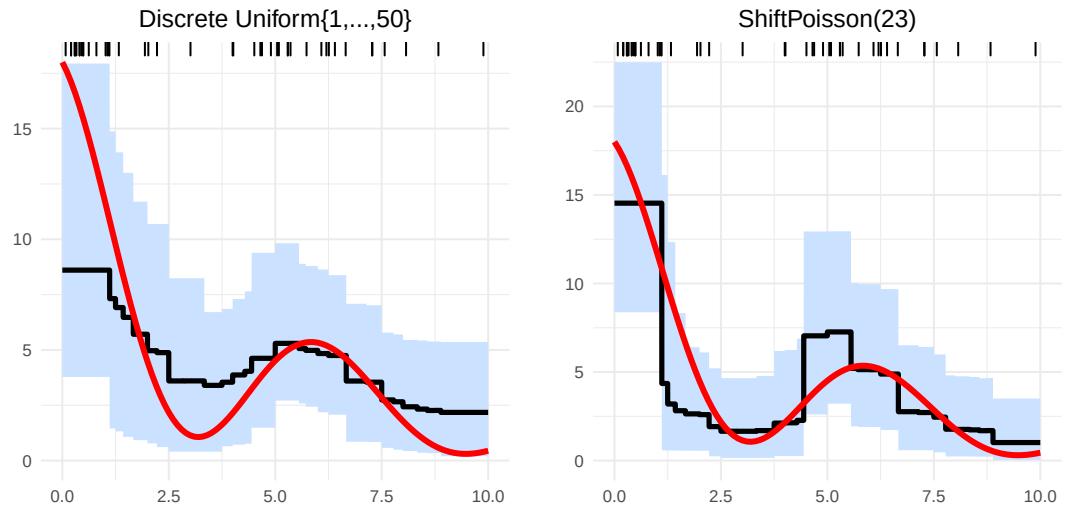

Figure 19: Estimation results for the oscillating exponential function λ_0 from Subsection 4.1 with $n = 1$ using a prior on the model index N. The prior on ψ_k's is $\mathrm{G}(0.1, 0.1)$. Left: $N \sim \mathrm{DiscreteUniform}(\{1, \ldots, 50\})$. Right: $N \sim \mathrm{ShiftPoisson}(23)$.

α and fixed number of bins $N = 200$, as well as the independent gamma prior with $N \sim \mathrm{DiscreteUniform}(\{1, \ldots, 50\})$. The results are shown in Figure 22. As can be seen from the bottom right panel there, the log-likelihood has many local maxima. The resulting difference in plots in the top row can be attributed to a defect of a sampler that only proposes visits to neighbouring sites. Having multiple local maxima in the marginal likelihood is not surprising, as any "good" approximation to λ must be able to capture the discontinuity at 0.7. One could instead use more advanced samplers, though this will come at a further computational cost. As a side remark, we note that estimation of discontinuous functions might be a delicate matter with the GMC prior. See Gugushvili et al. (2019a) for further remarks.

Rather than employing a prior on N for equal-sized bins, one could instead consider bins obtained from a non-equidistant grid of points. We conjecture that this will give visually more appealing results. However, such an approach requires a reversible jump algorithm over the locations of the bin endpoints. This is disadvantageous: the computing times of such algorithms do not scale well with the number of observations.

References

Adams, R. P., Murray, I., & MacKay, D. J. C. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In A. Danyluk, L. Bottou, & M. Littman (Eds.) *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, (pp. 9–16). New York, NY, USA: ACM.

Baddeley, A., Rubak, E., & Turner, R. (2016). *Spatial point patterns: methodology and applications with R*. Chapman Hall/CRC Interdiscip. Stat. Ser. Boca Raton, FL: CRC Press.

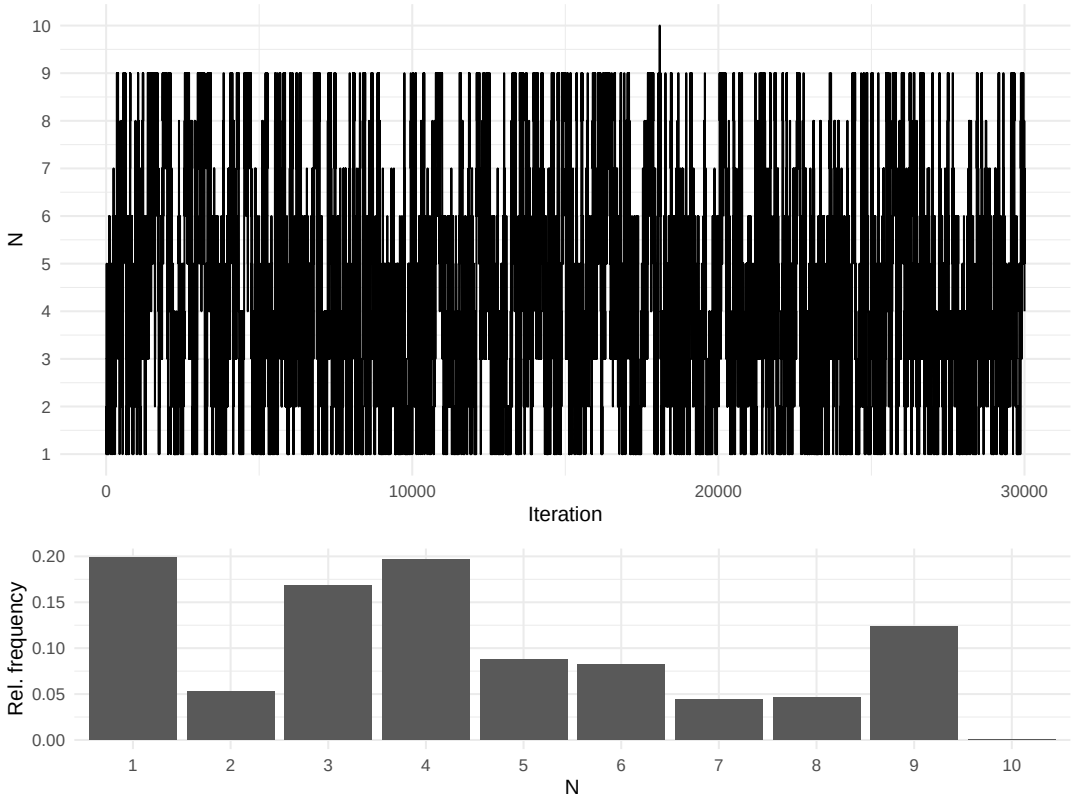

Figure 20: Results with $N \sim \text{DiscreteUniform}(\{1, \ldots, 50\})$. Top: traceplot on the model index N. Bottom: relative frequencies of models.

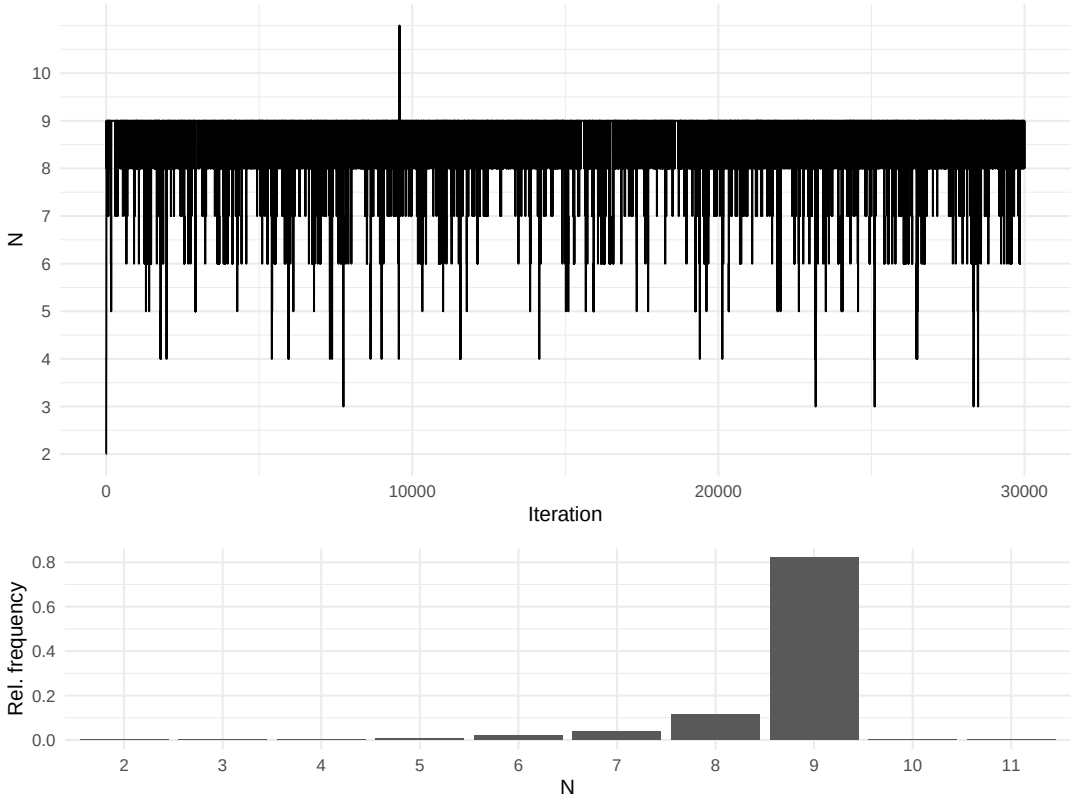

Figure 21: Results with $N \sim \text{ShiftPoisson}(23)$ prior. Top: traceplot on the model index N. Bottom: relative frequencies of models.

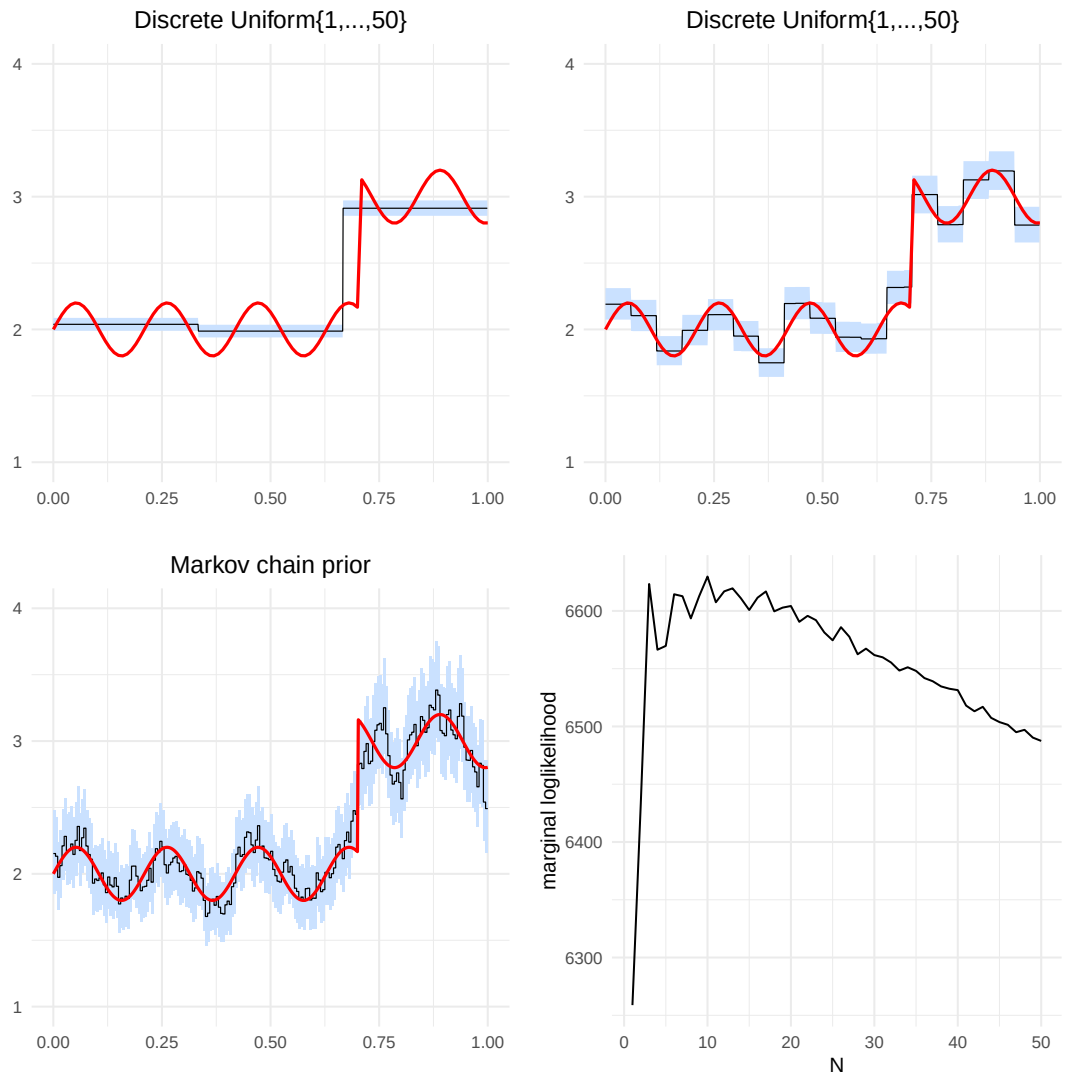

Figure 22: Estimation results for true intensity $\lambda(x) = 2 + 0.2\sin(30x) + \mathbf{1}_{[0.7,1]}(x)$ with $n = 10000$. Top left: independent gamma prior with reversible jump algorithm initialised at model index 2. Top right: independent gamma prior with reversible jump algorithm initialised at model index 20. Bottom left: GMC prior with $N = 200$ bins and Exp(0.1) prior on the smoothing parameter α. Bottom right: marginal log-likelihood as function of the model index.

Belitser, E., Serra, P., & van Zanten, H. (2015). Rate-optimal Bayesian intensity smoothing for inhomogeneous Poisson processes. *J. Statist. Plann. Inference*, *166*, 24–35.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: a fresh approach to numerical computing. *SIAM Rev.*, *59*(1), 65–98.

Blair, J. P., & Schweit, K. W. (2014). *A study of active shooter incidents in the United States between 2000 and 2013*. Texas State University and Federal Bureau of Investigation, U.S. Department of Justice, Washington D.C.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Amer. Statist. Assoc.*, *112*(518), 859–877.

Brémaud, P. (1981). *Point processes and queues: Martingale dynamics*. Springer-Verlag, New York-Berlin. Springer Series in Statistics.

Canty, A., & Ripley, B. D. (2017). *boot: Bootstrap R (S-Plus) functions*. R package version 1.3-20.

Cemgil, A. T., & Dikmen, O. (2007). Conjugate Gamma Markov random fields for modelling nonstationary sources. In M. E. Davies, C. J. James, S. A. Abdallah, & M. D. Plumbley (Eds.) *Independent Component Analysis and Signal Separation*, (pp. 697–705). Berlin, Heidelberg: Springer.

Chaudhuri, P., & Marron, J. S. (2000). Scale space view of curve estimation. *Ann. Statist.*, *28*(2), 408–428.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.*, *90*(432), 1313–1321.

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *J. Amer. Statist. Assoc.*, *96*(453), 270–281.

Clayton, D., & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, *43*, 671–81.

Cohen, A. P., Azrae, D., & Miller, M. (2014). Rate of mass shootings has tripled since 2011, Harvard research shows. Mother Jones.
URL `https://www.motherjones.com/politics/2014/10/mass-shootings-increasing-harvard-research/`

Cronie, O., & van Lieshout, M.-C. N. M. (2018). A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. *Biometrika*, *105*(2), 455–462.

Diggle, P. (1985). A kernel method for smoothing point process data. *J. R. Stat. Soc., Ser. C*, *34*, 138–147.

Diggle, P. J. (2014). *Statistical analysis of spatial and spatio-temporal point patterns*, vol. 128 of *Monogr. Stat. Appl. Probab.*. Boca Raton, FL: CRC Press, 3rd ed.

Dikmen, O., & Cemgil, A. T. (2009). Unsupervised single-channel source separation using Bayesian NMF. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (pp. 93–96).

Dikmen, O., & Cemgil, A. T. (2010). Gamma Markov random fields for audio source modeling. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(3), 589–601.

Efron, B. (2010). *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction*, vol. 1 of *Institute of Mathematical Statistics (IMS) Monographs*. Cambridge University Press, Cambridge.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science*. Institute of Mathematical Statistics Monographs. Cambridge, UK: Cambridge University Press.

Fairbrother, J., Nemeth, C., Rischard, M., Brea, J., & Pinder, T. (2018). GaussianProcesses.jl: A nonparametric Bayes package for the Julia language. *arXiv e-prints*.
URL https://arxiv.org/abs/1812.09064

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*, vol. 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Follman, M. (2012). What exactly is a mass shooting? Mother Jones.
URL https://www.motherjones.com/crime-justice/2012/08/what-is-a-mass-shooting/

Follman, M., Aronsen, G., & Pan, D. (2018a). A guide to mass shootings in America. Mother Jones.
URL https://www.motherjones.com/politics/2012/07/mass-shootings-map/

Follman, M., Aronsen, G., & Pan, D. (2018b). US mass shootings, 1982-2018: Data from Mother Jones' investigation. Mother Jones.
URL https://www.motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data/

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, *85*(410), 398–409.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, *6*(6), 721–741.

Ghosal, S., Ghosh, J. K., & van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, *28*(2), 500–531.

Grant, J., Boukouvalas, A., Griffiths, R.-R., Leslie, D., Vakili, S., & De Cote, E. M. (2019). Adaptive sensor placement for continuous spaces. In K. Chaudhuri, & R. Salakhutdinov (Eds.) *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, (pp. 2385–2393). Long Beach, California, USA: PMLR.

Grant, J. A., & Leslie, D. S. (2019). Posterior contraction rates for Gaussian Cox processes with non-identically distributed data. *arXiv e-prints*.
URL https://arxiv.org/abs/1906.08799

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Gugushvili, S., & Spreij, P. (2013). A note on non-parametric Bayesian estimation for Poisson point processes. *arXiv e-prints*.
URL https://arxiv.org/abs/1304.7353

Gugushvili, S., van der Meulen, F., Schauer, M., & Spreij, P. (2018). Nonparametric Bayesian volatility learning under microstructure noise. *ArXiv e-prints*.
URL https://arxiv.org/abs/1805.05606

Gugushvili, S., van der Meulen, F., Schauer, M., & Spreij, P. (2019a). Bayesian wavelet de-noising with the caravan prior. *ESAIM Probab. Stat.*, *23*, 947–978.

Gugushvili, S., van der Meulen, F., Schauer, M., & Spreij, P. (2019b). Nonparametric Bayesian volatility estimation. In J. de Gier, C. E. Praeger, & T. Tao (Eds.) *2017 MATRIX Annals*, (pp. 279–302). Cham: Springer International Publishing.

Gugushvili, S., van der Meulen, F., Schauer, M., & Spreij, P. (2019). PointProcessInference 0.1.0 – Code and Julia package accompanying the article "Gugushvili, van der Meulen, Schauer, Spreij (2018): Fast and scalable non-parametric Bayesian inference for Poisson point processes" (http://arxiv.org/abs/1804.03616).
URL https://doi.org/10.5281/zenodo.2591395

Heikkinen, J., & Arjas, E. (1998). Non-parametric Bayesian estimation of a spatial Poisson intensity. *Scandinavian Journal of Statistics*, *25*, 435–450.

Henderson, S. G. (2003). Estimation for nonhomogeneous Poisson processes from aggregated data. *Oper. Res. Lett.*, *31*(5), 375–382.

Hensman, J., de G. Matthews, A. G., Filippone, M., & Ghahramani, Z. (2015). MCMC for variationally sparse Gaussian processes. In *Proceedings of the 28th International Conference on Neural Information Processing Systems – Volume 1*, NIPS'15, (pp. 1648–1656). Cambridge, MA, USA: MIT Press.

Jarrett, R. G. (1979). A note on the intervals between coal-mining disasters. *Biometrika*, *66*(1), 191–193.

John, S. T., & Hensman, J. (2018). Large-scale Cox process inference using variational Fourier features. In J. Dy, & A. Krause (Eds.) *Proceedings of the 35th International Conference on Machine Learning*, vol. 80 of *Proceedings of Machine Learning Research*, (pp. 2362–2370). Stockholmsmässan, Stockholm, Sweden: PMLR.

Karr, A. F. (1986). *Point processes and their statistical inference*. Probability: Pure and Applied, 2. New York-Basel: Marcel Dekker, Inc.

Kingman, J. F. C. (1993). *Poisson processes*. Oxford: Clarendon Press.

Kirichenko, A., & van Zanten, H. (2015). Optimality of Poisson processes intensity learning with Gaussian processes. *Journal of Machine Learning Research*, *16*, 2909–2919.

Kom Samo, Y.-L., & Roberts, S. (2015). Scalable nonparametric Bayesian inference on point processes with Gaussian processes. In F. Bach, & D. Blei (Eds.) *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37 of *Proceedings of Machine Learning Research*, (pp. 2227–2236). Lille, France: PMLR.

Kutoyants, Y. A. (1998). *Statistical inference for spatial Poisson processes*, vol. 134 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Leemis, L. M. (2004). Technical note: Nonparametric estimation and variate generation for a nonhomogeneous Poisson process from event count data. *IIE Transactions*, *36*, 1155–1160.

Lloyd, C., Gunter, T., Osborne, M. A., & Roberts, S. J. (2015). Variational inference for Gaussian process modulated Poisson processes. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning – Volume 37*, ICML'15, (pp. 1814–1822). JMLR.org.

Loader, C. R. (1999). Bandwidth selection: classical or plug-in? *Ann. Statist.*, *27*(2), 415–438.

Martin, R., & Walker, S. G. (2019). Data-driven priors and their posterior concentration rates. *Electron. J. Statist.*, *13*(2), 3049–3081.

Møller, J., Syversveen, A. R., & Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scand. J. Statist.*, *25*(3), 451–482.

Møller, J., & Waagepetersen, R. P. (2004). *Statistical inference and simulation for spatial point processes*, vol. 100 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL.

Peeling, P., Cemgil, A. T., & Godsill, S. (2008). Bayesian hierarchical models and inference for musical audio processing. In *2008 3rd International Symposium on Wireless Pervasive Computing*, (pp. 278–282).

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL `https://www.R-project.org/`

Raftery, A. E., & Akman, V. E. (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika*, *73*(1), 85–89.

Robinson, D. (2016). Text analysis of Trump's tweets confirms he writes only the (angrier) Android half. Variance Explained.
URL `http://varianceexplained.org/r/trump-tweets/`

Robinson, D. (2017). Trump's Android and iPhone tweets, one year later. Variance Explained.
URL `http://varianceexplained.org/r/trump-followup/`

Rohatgi, A. (2017). WebPlotDigitizer, Version 4.1.
URL `https://automeris.io/WebPlotDigitizer/`

Spiegelhalter, D. (2019). *The Art of Statistics: Learning from Data*. Pelican books. London, UK: Pelican.

Streit, R. (2010). *Poisson point processes: Imaging, tracking and sensing*. Springer US.

Teh, Y. W., & Rao, V. (2011). Gaussian process modulated renewal processes. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.) *Advances in Neural Information Processing Systems 24*, (pp. 2474–2482). Curran Associates, Inc.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.*, *22*(4), 1701–1762. With discussion and a rejoinder by the author.

Wasserman, L. (2006). *All of Nonparametric Statistics (Springer Texts in Statistics)*. Berlin, Heidelberg: Springer-Verlag.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag, 2nd ed.