Tractable Nonparametric Bayesian Inference in Poisson Processes with Gaussian Process Intensities

Ryan Prescott Adams RPA23@CAM.AC.UK
Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, UK

Iain Murray MURRAY@CS.TORONTO.EDU
Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3G4

David J.C. MacKay MACKAY@MRAO.CAM.AC.UK
Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, UK

Abstract

The inhomogeneous Poisson process is a point process that has varying intensity across its domain (usually time or space). For nonparametric Bayesian modeling, the Gaussian process is a useful way to place a prior distribution on this intensity. The combination of a Poisson process and GP is known as a Gaussian Cox process, or doubly-stochastic Poisson process. Likelihood-based inference in these models requires an intractable integral over an infinite-dimensional random function. In this paper we present the first approach to Gaussian Cox processes in which it is possible to perform inference without introducing approximations or finite-dimensional proxy distributions. We call our method the Sigmoidal Gaussian Cox Process, which uses a generative model for Poisson data to enable tractable inference via Markov chain Monte Carlo. We compare our methods to competing methods on synthetic data and apply it to several real-world data sets.

1. Introduction

The Poisson process is a widely-used model for point data in temporal and spatial settings. The inhomogeneous variant of the Poisson process allows the rate of arrivals to vary in time (or space), but typically we do not have a preconceived idea of the appropriate functional form for this variation. In this setting, it is often desirable to use another stochastic process to describe nonparametrically the variation in the Poisson intensity function. This construction is called a *doubly-stochastic* Poisson process, or a Cox process (Cox, 1955), and has been applied in a variety of settings, e.g. neuroscience (Cunningham et al., 2008b), astronomy (Gregory & Loredo, 1992), and forestry (Heikkinen & Arjas, 1999).

One variant of the Cox process is the Gaussian Cox process, where the intensity function is a transformation of a random realization from a Gaussian process (GP). From a modeling perspective, this is a particularly convenient way to specify general prior beliefs about the intensity function via a kernel, without having to choose a particular functional form. Unfortunately, however, likelihood-based inference in this model is generally intractable, due to the need to integrate an infinite-dimensional random function. Various approximations have been introduced to deal with this intractability. The classic approach of Diggle (1985) uses Parzen-type kernel densities to construct a nonparametric estimator, with the bandwidth chosen via the empirical Ripley's function (Ripley, 1977). Nonparametric Bayesian approaches to the Gaussian Cox process have been studied in works by Rathbun and Cressie (1994) and Møller et al. (1998), which both introduced tractable finite-dimensional proxy distributions via discretization. There have also been nonparametric Bayesian approaches to inhomogeneous Poisson Process inference that do not use underlying Gaussian processes, e.g. Dirichlet process mixtures of Beta distributions (Kottas & Sansó, 2007).

In this paper we present the first approach to a Gaussian Cox process model that enables fully-nonparametric Bayesian inference via Markov chain Monte Carlo (MCMC), without requiring either numeric approximation or a finite-dimensional proxy dis-

Appearing in *Proceedings of the 26^{th} International Conference on Machine Learning*, Montreal, Canada, 2009. Copyright 2009 by the author(s)/owner(s).

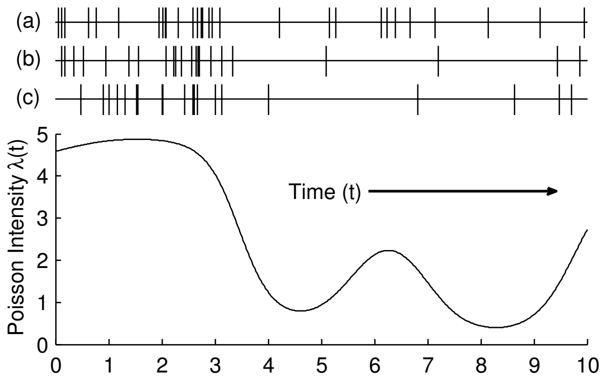


Figure 1. Three realizations of events from an inhomogeneous Poisson process in time, along with the associated intensity function.

tribution. We achieve this tractability by extending recently-developed MCMC methods for probability density functions (Adams et al., 2009).

2. The Model

In this section we review the Poisson process and specify our model, the Sigmoidal Gaussian Cox Process (SGCP), which transforms a Gaussian process into a nonparametric prior distribution on intensity functions. We then show that the SGCP allows exact simulation of Poisson data from a random infinite-dimensional intensity function, without performing intractable integrals. This approach is similar to that taken for general density modeling by Adams et al. (2009).

2.1. The Poisson Process

We consider the inhomogeneous Poisson process on a domain $\mathcal{S}$ which we will take to be $\mathbb{R}^D$. The Poisson process is parameterized by an intensity (or rate) function $\lambda(s) : \mathcal{S} \to \mathbb{R}^+$ where $\mathbb{R}^+$ indicates the nonnegative real numbers. The random number of events $N(\mathcal{T})$ within a subregion $\mathcal{T} \subset \mathcal{S}$ is Poisson-distributed with parameter $\lambda_{\mathcal{T}} = \int_{\mathcal{T}} \lambda(s) \mathrm{d}s$. The number of events in disjoint subsets are independent. Figure 1 shows an example of a one-dimensional Poisson intensity function along with three independently-drawn event sequences.

2.2. A GP Prior on Poisson Intensities

We introduce a random scalar function $g(s) : \mathcal{S} \to \mathbb{R}$. This function has a Gaussian process prior, which means that the prior distribution over any discrete set of function values $\{g(s_n)\}_{n=1}^N$ is a multivariate Gaussian distribution. These distributions can be consistently defined with a positive definite covariance function $C(\cdot,\cdot) : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ and a mean function $m(\cdot) : \mathcal{S} \to \mathbb{R}$. The mean and covariance functions are parameterized by hyperparameters θ. For a more detailed review of Gaussian processes see, e.g., Rasmussen and Williams (2006).

In the Sigmoidal Gaussian Cox Process, we transform the random $g(s)$ into a random intensity function $\lambda(s)$ via

$$\lambda(s) = \lambda^\star \, \sigma(g(s)) \tag{1}$$

where $\lambda^\star$ is an upper bound on $\lambda(s)$ and $\sigma(\cdot)$ is the logistic function, $\sigma(z) = (1 + e^{-z})^{-1}$. Equation 1 squashes $g(s)$ through a sigmoid function so that each $g(s)$ corresponds to a random function with outputs between zero and $\lambda^\star$. In this paper we use a logistic function as the sigmoid. The choice of $\lambda^\star$ is part of the prior and can be included in inference, as described in Section 3.5.

2.3. Generating Poisson Data from Random Gaussian Process Intensities

We use the transformation of Equation 1 because it allows us to simulate exact Poisson data from a random intensity function drawn from the prior provided by the Gaussian process. By *exact* we mean that the data are not biased by, for example, the starting state of a finite Markov chain. We generate these exact data via *thinning*, which is a point-process variant of rejection sampling introduced by Lewis and Shedler (1979). We extend the thinning procedure to simultaneously sample the function $g(s)$ from the Gaussian process.

We wish to generate a set of events $\{s_k\}_{k=1}^K$ on some subregion $\mathcal{T}$ of $\mathcal{S}$ which are drawn according to a Poisson process whose intensity function $\lambda(s)$ is the result of applying Equation 1 to a random function $g(s)$ drawn from the GP. We do this by first simulating a set of events $\{\hat{s}_j\}_{j=1}^J$ from a homogeneous Poisson process with intensity $\lambda^\star$. If $\mu(\cdot)$ is a measure on $\mathcal{S}$, then we first sample J, the *number* of events in $\mathcal{T}$, by drawing it from a Poisson distribution with parameter $\lambda^\star \mu(\mathcal{T})$. Next, the J events $\{\hat{s}_j\}_{j=1}^J$ are distributed uniformly within $\mathcal{T}$.

The $\{\hat{s}_j\}_{j=1}^J$ are now events randomly drawn from a homogeneous Poisson process with intensity $\lambda^\star$ on $\mathcal{T}$. Next, we treat these $\{\hat{s}_j\}_{j=1}^J$ as input points for a Gaussian process and sample the function $g(s)$ at these locations, to generate a corresponding set of function values, denoted $\{g(\hat{s}_j)\}_{j=1}^J$. We now use the thinning procedure to choose which $K \leq J$ points of $\{\hat{s}_j\}_{j=1}^J$ we will keep so that the kept points, denoted $\{s_k\}_{k=1}^K$, are drawn from an inhomogeneous Poisson process with an

Algorithm 1 Simulate data from a Poisson process on region $\mathcal{T}$ with random $\lambda(s)$ drawn as in Equation 1

```
Inputs: Region $\mathcal{T}$, Upper-bound $\lambda^\star$, GP functions $m(s)$ and $C(s,s')$
Outputs: Exact Poisson events $\mathcal{E} = (s_1, s_2, \ldots)$
$V \leftarrow \mu(\mathcal{T})$                                  ▷ Compute the measure of $\mathcal{T}$.
$J \sim \mathsf{Poisson}(V\lambda^\star)$                         ▷ Draw the number of events.
$\{\hat{s}_j\}_{j=1}^{J} \sim \mathsf{Uniform}(\mathcal{T})$      ▷ Distribute the events uniformly in $\mathcal{T}$.
$\{g(\hat{s}_j)\}_{j=1}^{J} \sim \mathcal{GP}\left(C(\cdot,\cdot), m(\cdot), \theta, \{\hat{s}_j\}_{j=1}^{J}\right)$   ▷ Sample the function at the events from the GP.
$\mathcal{E} \leftarrow \emptyset$                                ▷ Initialize the set of accepted events.
for $j \leftarrow 1 \ldots J$ do
    $r_j \sim \mathsf{Uniform}(0,1)$                              ▷ Draw a uniform random variate on the unit interval.
    if $r_j < \sigma(g(\hat{s}_j))$ then                          ▷ Apply acceptance rule.
        $\mathcal{E} \leftarrow \mathcal{E} \cup \hat{s}_j$       ▷ Add $\hat{s}_j$ to accepted events.
    end if
end for
return $\mathcal{E}$
```

intensity function $\lambda(s)$ consistent with the $\{g(\hat{s}_j)\}_{j=1}^{J}$ we have just simulated from the Gaussian process. We do this by generating J uniform random variates on $(0,1)$, denoted $\{r_j\}_{j=1}^{J}$. We only accept the events for which $r_j < \sigma(g(\hat{s}_j))$. These accepted events form the set $\{s_k\}_{k=1}^{K}$. This procedure is shown in Algorithm 1 and graphically in Figure 2.

3. Inference

We have so far defined a model for generating data from an inhomogeneous Poisson process using a GP-based prior for the intensity function. We now address the problem of inference: given a set of K events, denoted $\{s_k\}_{k=1}^{K}$, within a region $\mathcal{T}$, and using the SGCP model of Section 2 as the prior, what is the posterior distribution over $\lambda(s)$? The Poisson process likelihood function is

$$p(\{s_k\}_{k=1}^{K} \,|\, \lambda(s)) = \exp\left\{ -\int_{\mathcal{T}} \mathrm{d}s\ \lambda(s) \right\} \prod_{k=1}^{K} \lambda(s_k). \quad (2)$$

For random infinite-dimensional $\lambda(s)$, such as the Log Gaussian Cox Process or the SGCP model presented in Section 2, the integral inside the exponential cannot be evaluated. We write Bayes' theorem for our model, using $\boldsymbol{g}$ to indicate the infinite-dimensional object corresponding to $g(s)$:

$$p(\boldsymbol{g} \,|\, \{s_k\}_{k=1}^{K}) = \quad (3)$$
$$\frac{\mathcal{GP}(\boldsymbol{g}) \exp\left\{-\int_{\mathcal{T}} \lambda^\star \sigma(g(s))\,\mathrm{d}s\right\} \prod_k \lambda^\star \sigma(g(s_k))}{\int \mathrm{d}\boldsymbol{g}\ \mathcal{GP}(\boldsymbol{g}) \exp\left\{-\int_{\mathcal{T}} \lambda^\star \sigma(g(s))\,\mathrm{d}s\right\} \prod_k \lambda^\star \sigma(g(s_k))}.$$

This posterior distribution is *doubly-intractable* (Murray et al., 2006), due to the presence of an intractable integral over $\mathcal{T}$ in the numerator and an intractable integral over $\boldsymbol{g}$ in the denominator. Standard Markov chain Monte Carlo methods are unable to deal with intractability in the likelihood as in Equations 2 and 3. We also have the basic intractability that we cannot naïvely represent the posterior distribution over the infinite-dimensional $\boldsymbol{g}$, even if we could perform the integral calculations.

3.1. Tractability Via Latent Variables

Rather than performing MCMC inference directly via the posterior in Equation 3, we augment the posterior distribution to make the Markov chain tractable. Specifically, we consider the Poisson data to have been generated as in Section 2, and the additional latent variables are 1) the total number of "thinned" events M; 2) the locations of the thinned events, $\{\tilde{s}_m\}_{m=1}^{M}$; 3) the values of the function $g(s)$ at the thinned events, denoted $\boldsymbol{g}_M = \{g(\tilde{s}_m)\}_{m=1}^{M}$; 4) the values of the function $g(s)$ at the observed events, denoted $\boldsymbol{g}_K = \{g(s_k)\}_{k=1}^{K}$. The generative procedure did not require integrating an infinite-dimensional random function, nor did it require knowledge of $g(s)$ or $\lambda(s)$ at more than a finite number of locations. By considering the procedure as a latent variable model, we inherit these convenient properties for inference. The joint distribution over the fixed data $\{s_k\}_{k=1}^{K}$, the number of thinned events M, the location of the thinned events $\{\tilde{s}_m\}_{m=1}^{M}$, and the function value vectors $\boldsymbol{g}_M$ and $\boldsymbol{g}_K$, is

$$p(\{s_k\}_{k=1}^{K}, M, \{\tilde{s}_m\}_{m=1}^{M}, \boldsymbol{g}_{M+K} \,|\, \lambda^\star, \mathcal{T}, \theta) =$$
$$(\lambda^\star)^{K+M} \exp\{-\lambda^\star \mu(\mathcal{T})\} \prod_{k=1}^{K} \sigma(g(s_k)) \prod_{m=1}^{M} \sigma(-g(\tilde{s}_m))$$
$$\times\, \mathcal{GP}\left(\boldsymbol{g}_{M+K} \,|\, \{s_k\}_{k=1}^{K}, \{\tilde{s}_m\}_{m=1}^{M}, \theta\right) \quad (4)$$

where $\boldsymbol{g}_{M+K}$ denotes a vector concatenating $\boldsymbol{g}_M$ and $\boldsymbol{g}_K$. Note that $\sigma(-z) = 1 - \sigma(z)$, and that we

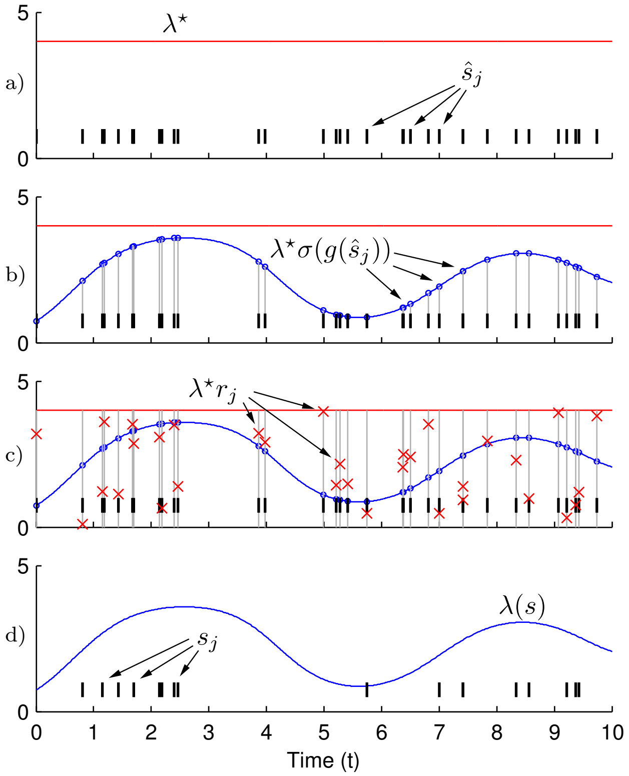


Figure 2. This sequence of figures shows the generative procedure for the SGCP. a) The upper-bounding intensity $\lambda^\star$ is shown, along with a Poisson time series generated from it. b) At each point of the time series, a sample is drawn from the Gaussian process. This function is squashed through the logistic function so that it is everywhere positive and upper-bounded by $\lambda^\star$. c) Variates are drawn uniformly on $(0, \lambda^\star)$ in the vertical coordinate. d) If the uniform variates are greater than the random function, then the corresponding events are discarded. The kept events are drawn from the inhomogeneous Poisson process corresponding to the random intensity $\lambda(s)$.

have integrated out the vertical coordinates r that determined acceptance and rejection. We use three kinds of Markov transitions to sample from this joint distribution: 1) changing M, the number of latent thinned events, 2) changing the locations $\{\tilde{s}_m\}_{m=1}^M$ of the thinned events, and 3) changing the latent function vector $\boldsymbol{g}_{M+K}$. We also address hyperparameter inference in Section 3.5.

3.2. Sampling the Number of Thinned Events

We use Metropolis–Hastings to sample from the number of thinned events M. We define a function $b(K, M) : \mathbb{N} \times \mathbb{N} \rightarrow (0, 1)$ that gives the Bernoulli probability of proposing an insertion or a deletion. An insertion move consists of proposing a new $\tilde{s}'$ drawn uniformly from $\mathcal{T}$, followed by a draw of the corresponding function value $g(\tilde{s}')$ from the Gaussian process, conditioned on the current state $\boldsymbol{g}_{M+K}$. This proposal is

$$q_{\mathsf{ins}}(M{+}1 \leftarrow M) = \frac{b(K, M)}{\mu(\mathcal{T})}\, \mathcal{GP}\left(g(\tilde{s}') \,|\, \tilde{s}', \boldsymbol{g}_{M+K}\right). \tag{5}$$

A deletion move for a latent thinned event consists of selecting the event m to remove randomly and uniformly from the M events in the current state. This proposal has density

$$q_{\mathsf{del}}(M{-}1 \leftarrow M) = \frac{1 - b(K, M)}{M}. \tag{6}$$

We incorporate the joint distribution of Equation 4 to find the Metropolis–Hastings acceptance ratios of each type of proposal, integrating out the vertical coordinate r:

$$a_{\mathsf{ins}} = \frac{(1{-}b(K, M{+}1))\ \mu(\mathcal{T})\ \lambda^\star}{(M{+}1)\ b(K, M)\ (1{+}\exp\{g(\tilde{s}')\})} \tag{7}$$

$$a_{\mathsf{del}} = \frac{M\ b(K, M{-}1)\ (1{+}\exp\{g(\tilde{s}_m)\})}{(1{-}b(K, M))\ \mu(\mathcal{T})\ \lambda^\star}. \tag{8}$$

The proposal probability $b(K, M)$ can safely be set to $\frac{1}{2}$. Tuning and domain knowledge can almost certainly yield better choices, however. We have not extensively explored this topic. In practice we have found it useful to make several (≈ 10) of these transitions for each of the transitions made in Sections 3.3 and 3.4.

3.3. Sampling the Locations of Thinned Events

Given the number of thinned events, M, we also wish to sample from the posterior distribution on the locations of the events, $\{\tilde{s}_m\}_{m=1}^M$. We use Metropolis–Hastings to perform this sampling. We iterate over each of the M thinned events and propose a new location $\tilde{s}'_m$ via the proposal density $q(\tilde{s}'_m \leftarrow \tilde{s}_m)$. We then draw a function value $g(\tilde{s}'_m)$ from the Gaussian process, conditioned on the current state $\boldsymbol{g}_{M+K}$. The Metropolis–Hastings acceptance ratio for this proposal, integrating out the vertical coordinate r, is

$$a_{\mathsf{loc}} = \frac{q(\tilde{s}_m \leftarrow \tilde{s}'_m)\ (1 + \exp\{g(\tilde{s}_m)\})}{q(\tilde{s}'_m \leftarrow \tilde{s}_m)\ (1 + \exp\{g(\tilde{s}'_m)\})}. \tag{9}$$

Typically, perturbative proposals on the order of the Gaussian process length scale are appropriate for these Metropolis–Hastings steps. If the move is accepted, the old values $\tilde{s}_m$ and $g(\tilde{s}_m)$ can safely be forgotten.

3.4. Sampling the Function

We use Hamiltonian Monte Carlo (Duane et al., 1987) for inference of the function values $\boldsymbol{g}_{M+K}$, to take advantage of gradient information and make efficient proposals. We perform gradient calculations in the "whitened" space resulting from linearly transforming $\boldsymbol{g}_{M+K}$ with the inverse Cholesky decomposition of the covariance matrix $\boldsymbol{\Sigma}$, as this results in a better-conditioned space for calculations. The log conditional posterior distribution is

$$\begin{aligned}
&\ln p(\boldsymbol{g}_{M+K} \mid M, \{s_k\}_{k=1}^{K}, \{s_m\}_{m=1}^{M}, \theta) = \\
&-\frac{1}{2}\boldsymbol{g}_{M+K}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{g}_{M+K} - \sum_{k=1}^{K}\ln\left(1+\exp\{-g(s_k)\}\right) \\
&\qquad - \sum_{m=1}^{M}\ln\left(1+\exp\{g(\tilde{s}_k)\}\right) + \text{const.} \qquad (10)
\end{aligned}$$

3.5. Hyperparameter Inference

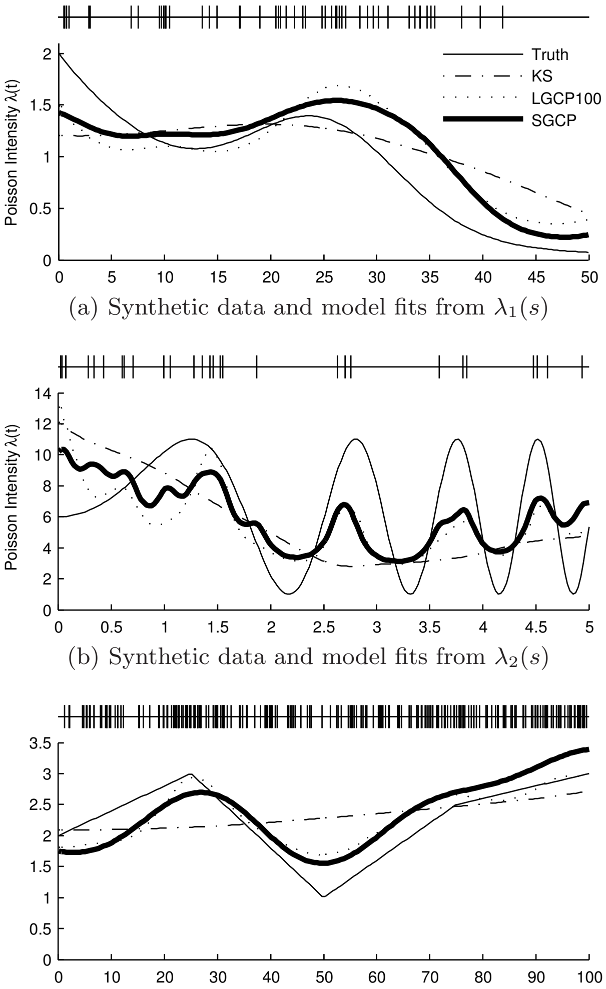
Given the data, the thinned events $\{\tilde{s}_m\}_{m=1}^{M}$ and the function values $\boldsymbol{g}_{M+K}$, we might also like to take advantage of hierarchical Bayesian inference to sample from the posterior distribution on any hyperparameters θ in the covariance and mean functions. This can be performed straightforwardly using Hamiltonian Monte Carlo as described by Neal (1998).

The upper-bound parameter $\lambda^\star$ can also be inferred as part of the MCMC procedure. Conditioned on M, K, and the thinned event locations, the union of $\{s_k\}_{k=1}^{K}$ and $\{s_m\}_{m=1}^{M}$ are drawn from a homogeneous Poisson process on $\mathcal{T}$ with rate $\lambda^\star$. The Gamma distribution with shape parameter α and inverse-scale parameter β provides a conditionally-conjugate prior for $\lambda^\star$. We can sample from the conditional posterior distribution, which is Gamma with parameters

$$\alpha_{\mathsf{post}} = \alpha + K + M \qquad \beta_{\mathsf{post}} = \beta + \mu(\mathcal{T}).$$

3.6. Predictive Samples

The predictive distribution is often of interest in Bayesian modeling. This distribution is the one arising on sets of events, integrating out the posterior over intensity functions and hyperparameters. For the SGCP model, this corresponds to generating an entirely new time series from the model, integrating out $g(s)$. It is straightforward to generate data from this distribution as a part of the MCMC inference: after any given step in the Markov chain, run the generative procedure of Algorithm 1, but condition on the current $\boldsymbol{g}_{M+K}$ when drawing from the Gaussian process. That is, in Line 4 of Algorithm 1, condition on $\{s_k, g(s_k)\}_{k=1}^{K}$ and $\{s_m, g(s_m)\}_{m=1}^{M}$. This provides new time series that are drawn from the predictive distribution.



(a) Synthetic data and model fits from $\lambda_1(s)$

(b) Synthetic data and model fits from $\lambda_2(s)$

(c) Synthetic data and model fits from $\lambda_3(s)$

Figure 3. These are synthetic time series and mean functions from analyzed estimation methods. The thin solid lines show the true intensities, the thick solid lines show the Sigmoidal Gaussian Cox Process (SGCP) posterior mean, the dot-dashed lines show the kernel estimator (KS), and the dotted lines show the posterior mean of the Log Gaussian Cox Process with 100 bins (LGCP100).

4. Empirical Results

We performed two types of empirical analysis of our approach. We used synthetic data with known $\lambda(s)$ to compare the SGCP to other comparable nonparametric approaches. We also applied our method to two real-world data sets, one temporal and one spatial.

4.1. Synthetic Data

We created three one-dimensional data sets using the following intensity functions:

1. A sum of an exponential and a Gaussian bump:

Table 1. The Sigmoidal Gaussian Cox Process (SGCP) is compared with the main frequentist kernel smoothing (KS) method (Diggle, 1985) and with the Log Gaussian Cox Process (Møller et al., 1998). The LGCP requires binning and the table shows the results with ten (LGCP10), 25 (LGCP25) and 100 (LGCP100) bins. The comparisons were done against time series from known intensity functions and compared on ℓ_s norm to the true function and the mean predictive log probability (lp) of ten unseen time series from the same intensity function.

		SGCP	KS	LGCP10	LGCP25	LGCP100
$\lambda_1(s)$	ℓ_2	**4.20**	6.65	5.96	6.12	5.44
	lp	**-45.11**	-46.41	-46.00	-46.80	-45.24
$\lambda_2(s)$	ℓ_2	**38.38**	73.71	70.34	53.27	43.51
	lp	24.45	**28.19**	23.36	22.89	25.29
$\lambda_3(s)$	ℓ_2	11.41	30.56	90.76	22.14	**10.79**
	lp	**-43.39**	-46.47	-53.67	-52.31	-47.16

$\lambda_1(s) = 2\exp\{-s/15\} + \exp\{-((s-25)/10)^2\}$ on the interval $[0, 50]$. 53 events.

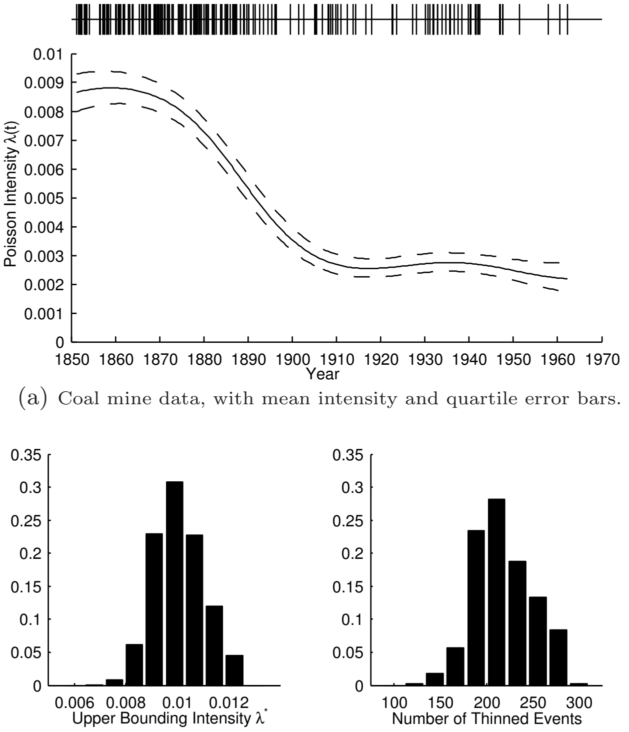
2. A sinusoid with increasing frequency: $\lambda_2(s) = 5\sin(s^2){+}6$ on $[0, 5]$. 29 events.

3. $\lambda_3(s)$ is the piecewise linear function shown in Figure 3(c), on the interval $[0, 100]$. 235 events.

We compared the SGCP to the classical kernel smoothing (KS) approach of Diggle (1985). We performed edge-corrected kernel smoothing using a quartic kernel and the recommended mean-square minimization technique for bandwidth selection. We also compared to the most closely-related nonparametric Bayesian technique, the Log Gaussian Cox Process of Rathbun and Cressie (1994) and Møller et al. (1998). To implement this method, we used discretization to make a finite-dimensional approximation and applied Markov chain Monte Carlo. We ran the Log Gaussian Cox Process method with 10 bins (LGCP10), 25 bins (LGCP25) and 100 bins (LGCP100).

We used the squared-exponential kernel for both the SGCP and the LGCP, and sampled from the hyperparameters for both models.

We report the numerically-estimated ℓ_2 distance between the mean $\lambda(s)$ provided by each method and the known true function. We also report the mean log predictive probability of ten additional held-out time series generated from the same (known) $\lambda(s)$. These predictive probabilities were calculated by numerical integrations of the functions. These results are provided in Table 1, and the resulting estimates (excluding LGCP10 and LGCP25) are shown in Figure 3.



(a) Coal mine data, with mean intensity and quartile error bars.

(b) Normalized histogram of sampled $\lambda^\star$ values.

(c) Normalized histogram of the sampled number of thinned events.

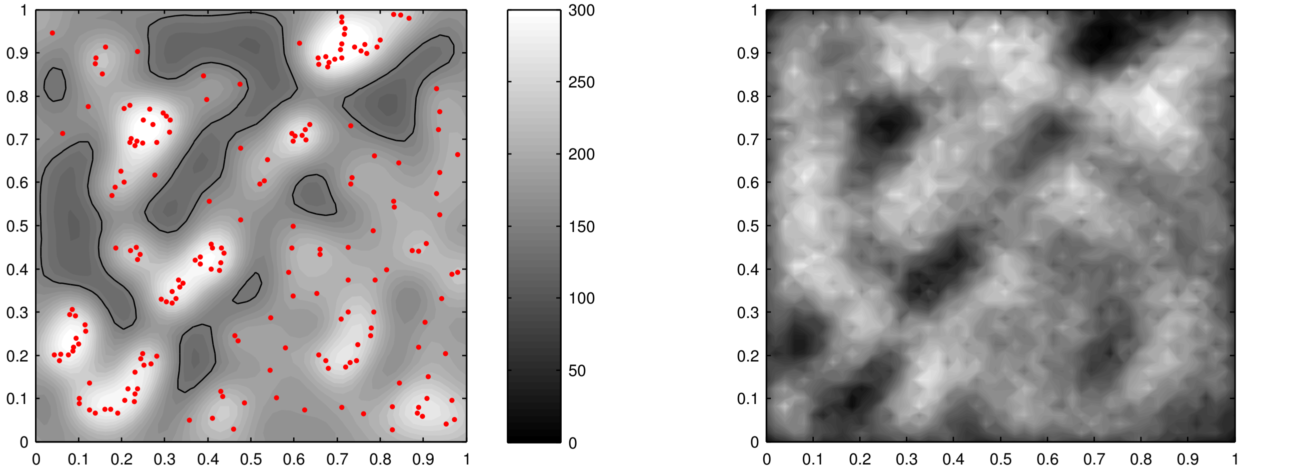
Figure 5. These three figures show the result of MCMC inference applied to coal mine disaster data. There were 191 events between 15 March 1875 and 22 March 1962. The top figure shows the inferred function with quartile error bars. The bottom two figures are normalized histograms of the sampled $\lambda^\star$ and the number of thinned events M.

4.2. Coal Mining Disaster Data

We ran the Markov chain Monte Carlo inference procedure on the classic coal mine disaster data of Jarrett (1979). These data are the dates of 191 coal mine explosions that killed ten or more men in Britain between 15 March 1875 and 22 March 1962. Figure 5(a) shows the events along the top, and the inferred mean intensity function. Also shown are approximate quartile error bars. In Figure 5(b) is a normalized histogram of the inferred upper bounding intensity, $\lambda^\star$. Figure 5(c) is a normalized histogram of the number of latent thinned events, M.

4.3. Redwoods Data

We used a standard data set from spatial statistics to demonstrate the Sigmoidal Gaussian Cox Process in two dimensions. These data are the locations of redwood trees studied by Ripley (1977) and others. There are 195 points and, as in previous studies, they have been scaled to the unit square. Figure 4(a) shows the data along with the inferred mean intensity function.



(a) Redwood locations and mean intensity.

(b) Normalized histogram of thinned event locations.

Figure 4. The left figure shows the redwood locations that have been scaled to a unit square as in Ripley (1977). The mean intensity estimate from the SGCP is also shown. The black line shows the contour at $\lambda(s) = 150$. On the right is a normalized histogram of the locations of thinned events. This illustrates the tendency of these events to migrate to areas where $\lambda(s)$ is small and "peg down" the GP.

These data are useful for examining the placement of latent thinned events. Figure 4(b) shows a normalized histogram of where the these events tended to be located during the MCMC run. As expected, it is approximately a "negative" of the mean intensity; the thinned events are moving to places where it is necessary to "peg down" the intensity function.

5. Discussion

5.1. Computational Concerns

Gaussian processes have significant computational demands: they have $O(n^3)$ time complexity for n input points, and $O(n^2)$ space complexity. When performing inference in the SGCP model, this means that each MCMC step costs $O((K+M)^3)$ as the thinned events must be included in the GP. Thus the approach we present is infeasible for data sets that have more than several thousand events. Generally, mixing of the Markov chain is a potential computational concern, but by using Hamiltonian Monte Carlo we have had good results. Autocorrelation plots from the $\lambda_1(s)$ inference are shown in Figure 6.

5.2. Contrasting With Other Methods

The motivation for the Gaussian Cox process is primarily the ease with which one can specify prior beliefs about the variation of the intensity function of a Poisson process, without specifying a particular functional form. There have been several methods proposed for approximation to this model, most notably the discretization method of Rathbun and Cressie (1994) and Møller et al. (1998), to which we have compared the SGCP. Cunningham et al. (2008a) also use a discretization-based approximation for the Cox process and report performance improvements on their domain of interest. This method, however, suffers from three deficiencies for general application: it is a finite-dimensional proxy method, it uses the renewal-process formalism that cannot be easily generalized beyond the time domain, and its model is inconsistent in that it assigns prior mass to negative intensity functions.

The Dirichlet process mixture of Beta distributions (Kottas & Sansó, 2007) has the appealing characteristic that it allows tractable nonparametric Bayesian inference via relatively standard MCMC methods. However, it is unclear how to choose the hyperparameters of an infinite mixture of fixed densities to represent beliefs about a Poisson process intensity. The underlying infinite-dimensional stochastic process is the Dirichlet, and it is fixed throughout space and/or time. Variations in the intensity only arise out of the *parametric* variations in the distributions being mixed. Also, it is straightforward to understand the marginalization properties of the Gaussian Cox Process if the region of interest changes, but a mixture of Betas appears to have discontinuities when expanding the studied region.

The Sigmoidal Gaussian Cox Process is superior to the frequentist kernel density approach (Diggle, 1985) in several ways. First, we obtain samples from the posterior rather than a point estimate of the unknown intensity function. Second, we are able to perform bandwidth selection in a principled way by sampling from the hyperparameters of the Gaussian process. Third, we are able to incorporate arbitrary nonstationary Gaussian processes into the framework without modification. Finally, our method does not suffer from detrimental edge effects. These improvements do, however, come at some computational cost.

5.3. Variations on the Model

There are several ways in which the Sigmoidal Gaussian Cox Process we have presented could be modified for different modeling situations. For example, to arrive at bounded random intensities we used a constant dominating function $\lambda^\star$, but other tractable parametric forms would be suitable and potentially more efficient. Also, we use the logistic function in Equation 1 to map from arbitrary functions to the bounded domain, but other functions could be used to achieve different properties. For example, if $\sigma(\cdot)$ was the normal cumulative distribution function and the Gaussian process prior was zero-mean and stationary with unit amplitude, then the random intensities $\lambda(s)$ from the model would be marginally uniform beneath $\lambda^\star$.

5.4. Summary of Contributions

We have introduced a novel method of inference for the Gaussian Cox process that avoids the intractability typical to such models. Our model, the Sigmoidal Gaussian Cox Process, uses a generative prior that allows exact Poisson data to be generated from a random intensity function drawn from a transformed Gaussian process. With the ability to generate exact data, we can simulate a Markov chain on the posterior distribution of infinite-dimensional intensity functions without approximation and with finite computational resources. Our approach stands in contrast to other nonparametric Bayesian approaches to the inhomogeneous Poisson process in that it requires neither a crippling of the model, nor a finite-dimensional approximation.

Acknowledgements

The authors wish to thank Maneesh Sahani and Zoubin Ghahramani for valuable comments. Ryan Adams thanks the Gates Cambridge Trust and the Canadian Institute for Advanced Research.

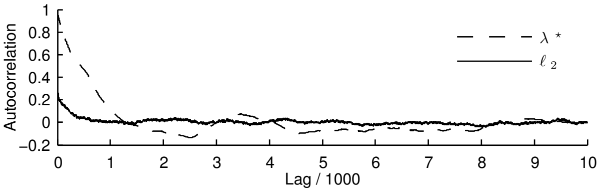


Figure 6. Two autocorrelation functions from the MCMC run from function $\lambda_1(s)$. The dashed line shows the autocorrelation of $\lambda^\star$ and the solid line shows the autocorrelation of the numerically-estimated ℓ_2 distance between the estimated function and the true function. We use the ℓ_2 distance as a global measure of function mixing.

References

Adams, R. P., Murray, I., & MacKay, D. J. C. (2009). The Gaussian process density sampler. *NIPS 21*.

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society, Series B*, *17*, 129–164.

Cunningham, J. P., Shenoy, K. V., & Sahani, M. (2008a). Fast Gaussian process methods for point process intensity estimation. *ICML 2008*.

Cunningham, J. P., Yu, B. M., Shenoy, K. V., & Sahani, M. (2008b). Inferring neural firing rates from spike trains using Gaussian processes. *NIPS 20*.

Diggle, P. (1985). A kernel method for smoothing point process data. *Applied Statistics*, *34*, 138–147.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, *195*, 216–222.

Gregory, P. C., & Loredo, T. J. (1992). A new method for the detection of a periodic signal of unknown shape and period. *The Astrophysical Journal*, *398*, 146–168.

Heikkinen, J., & Arjas, E. (1999). Modeling a Poisson forest in variable elevations: a nonparametric Bayesian approach. *Biometrics*, *55*, 738–745.

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

Kottas, A., & Sansó, B. (2007). Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. *Journal of Statistical Planning and Inference*, *137*, 3151–3163.

Lewis, P. A. W., & Shedler, G. S. (1979). Simulation of a nonhomogeneous Poisson process by thinning. *Naval Research Logistics Quarterly*, *26*, 403–413.

Møller, J., Syversveen, A. R., & Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, *25*, 451–482.

Murray, I., Ghahramani, Z., & MacKay, D. J. (2006). MCMC for doubly-intractable distributions. *UAI 22*.

Neal, R. M. (1998). Regression and classification using Gaussian process priors. *Bayesian Statistics 6* (pp. 475–501).

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.

Rathbun, S. L., & Cressie, N. (1994). Asymptotic properties of estimators for the parameters of spatial inhomogeneous Poisson point processes. *Advances in Applied Probability*, *26*, 122–154.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society, Series B*, *39*, 172–212.