Bayesian Analysis (2012) **7**, Number 4, pp. 813–840

Nonparametric Bayesian Segmentation of a Multivariate Inhomogeneous Space-Time Poisson Process

Mingtao Ding *, Lihan He †, David Dunson ‡ and Lawrence Carin §

Abstract. A nonparametric Bayesian model is proposed for segmenting time-evolving multivariate spatial point process data. An inhomogeneous Poisson process is assumed, with a logistic stick-breaking process (LSBP) used to encourage piecewise-constant spatial Poisson intensities. The LSBP explicitly favors spatially contiguous segments, and infers the number of segments based on the observed data. The temporal dynamics of the segmentation and of the Poisson intensities are modeled with exponential correlation in time, implemented in the form of a first-order autoregressive model for uniformly sampled discrete data, and via a Gaussian process with an exponential kernel for general temporal sampling. We consider and compare two different inference techniques: a Markov chain Monte Carlo sampler, which has relatively high computational complexity; and an approximate and efficient variational Bayesian analysis. The model is demonstrated with a simulated example and a real example of space-time crime events in Cincinnati, Ohio, USA.

Keywords: Bayesian hierarchical model, spatial segmentation, temporal dynamics, Gaussian process, logistic stick breaking process, inhomogeneous Poisson process

1 Introduction

1.1 Motivating application

Assume access to the locations of various types of crimes occurring in a given city, as a function of time. As a motivating example, in Figure 1(a) data are shown for 3090 crimes (of 17 crime types) in Cincinnati in Jan 2008. Our focus is on obtaining a spatial segmentation, such as that shown in Figure 1(b). In addition to the spatial dependence of point process data, we wish to simultaneously explore time dynamics. For example, in the crime data analysis, the crime intensity in summer may be different statistically from that in winter, and this intensity may change smoothly over seasons; consequently, the spatial segmentation of the city may also vary smoothly over time.

The analysis of time dynamics helps to discover the temporal pattern of the events

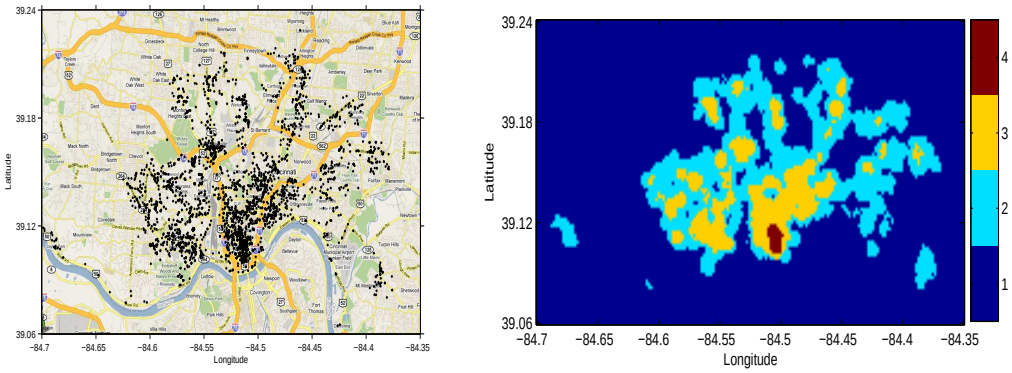
*Department of Electrical & Computer Engineering, Duke University, Durham, NC mingtao.ding@ece.duke.edu

†Department of Electrical & Computer Engineering, Duke University, Durham, NC lihan@ece.duke.edu

‡Statistical Sciences Department, Duke University, Durham, NC dunson@stats.duke.edu

§Department of Electrical & Computer Engineering, Duke University, Durham, NC lcarin@ece.duke.edu

© 2012 International Society for Bayesian Analysis DOI:10.1214/12-BA727



(a) Crime events in Cincinnati during Jan., 2008

(b) Segmentation of Cincinnati

Figure 1: Crime events and the segmentation of the city. In (a) 3090 crime events are shown as black dots; in (b) each color indexes a segment with associated crime intensities in 17 crime types (see Section 4 for details).

and to predict the spatial segmentation at an unobserved time instance or in the future. We desire that the analysis provide a simple summary that is useful to police forces and city planners in targeting resources, as well as to researchers in studying crime trends. We would like to obtain this space-time segmentation quickly, utilizing data from different types of events, while allowing temporal interpolation and forecasting.

1.2 Summary of proposed model

Consider the data $\mathcal{D} = \{\boldsymbol{s}_i, \boldsymbol{v}_{it}\}_{i=1,\ldots,M,\ t=1,\ldots,T}$, where $\boldsymbol{v}_{it}$ is a d-dimensional vector of the counts of d types of events, occurring in a (small) spatial region $\Delta(\boldsymbol{s}_i)$, with the center of the region being $\boldsymbol{s}_i \in \mathbb{R}^2$. In the context of Figure 1, we are interested in d types of crime. The contiguous grid of spatial regions $\Delta(\cdot)$ is fixed in advance, and the size of $\Delta(\cdot)$ is very small relative to the size of the entire spatial domain, providing justification for an approximation in which we index regions by the center point and assume homogeneity within regions (using the model developed below, in the limit $\Delta \rightarrow 0$ we have a Poisson process). There are T time points at which data are observed, not necessarily uniformly spaced in time. Although not done here, one may envision aligning the grid $\Delta(\cdot)$ with the geometry of the terrain (*e.g.*, roads).

The proposed space-time model may be summarized as

$$\boldsymbol{v}_{it} \sim \prod_{j=1}^{d} \text{Poisson}(\lambda_{ijt}), \qquad \boldsymbol{\lambda}_{it} \sim \sum_{k=1}^{K} w_k(\boldsymbol{s}_i; \boldsymbol{\theta}_{kt}) \delta_{\boldsymbol{\lambda}^*_{kt}} \tag{1}$$

where $w_k(\boldsymbol{s}_i; \boldsymbol{\theta}_{kt}) \geq 0$, $\sum_{k=1}^{K} w_k(\boldsymbol{s}_i; \boldsymbol{\theta}_{kt}) = 1$ for all $\boldsymbol{s}_i$, $\delta_{\boldsymbol{\lambda}^*_{kt}}$ is a unit measure concentrated at $\boldsymbol{\lambda}^*_{kt}$, and λ_{ijt} is the jth component of $\boldsymbol{\lambda}_{it}$. This corresponds to a mixture

model, with space-time varying mixture weights $w_k(\boldsymbol{s}_i;\boldsymbol{\theta}_{kt})$ and time-varying atoms $\boldsymbol{\lambda}^*_{kt}$.

Expression $w_k(\boldsymbol{s};\boldsymbol{\theta}_{kt})$ represents a general parametric function capable of modeling the probability of cluster k at spatial location $\boldsymbol{s}$. In the details of the proposed model, one of the $\{w_k(\boldsymbol{s};\boldsymbol{\theta}_{kt})\}_{k=1,K}$ is likely to be dominant (large probability) over a contiguous region, yielding a segmentation. Since the parameters $\boldsymbol{\theta}_{kt}$ change in general with time t, a probabilistic space-time segmentation is manifested. Within the proposed model, the prior encourages that $\{\boldsymbol{\theta}_{kt}\}$ and $\boldsymbol{\lambda}^*_{kt}$ vary smoothly as a function of time, and hence the model imposes smooth space-time variation in the shape/form of the segments, and smooth temporal variation of the Poisson rates associated with a given segment.

Two methods are considered for imposing temporal smoothness, representing two perspectives on imposing the same temporal structure. For discrete-time data with uniform temporal spacing, it is natural to consider the first-order autoregressive model, *i.e.*, AR(1), as $\theta_{kpt} \sim \mathcal{N}(\zeta\theta_{kp(t-1)}, \alpha_0^{-1})$, with θ_{kpt} the pth component of $\boldsymbol{\theta}_{kt}$, ζ the AR(1) coefficient (with $|\zeta| < 1$), and α_0 a precision to be inferred (ζ and α_0 could also be extended to depend on k and p). The log of each component of $\boldsymbol{\lambda}^*_{kt}$ may be similarly modeled.

We also consider a Gaussian process (GP) model (Rasmussen and Willams 2006) in time for each component θ_{kpt}, and for the log of each component of $\boldsymbol{\lambda}^*_{kt}$, thus allowing non-uniform temporal sampling. To make the AR(1) and GP models consistent, we assume an exponential model for the GP covariance between times t_i and t_l, $c_0 c_1^{|t_i - t_l|}$, with c_1 playing a role analogous to ζ in the AR(1) model, and the variance c_0 corresponding to $[(1-\zeta^2)\alpha_0]^{-1}$ from the AR(1) model. The AR(1) and chosen GP representations are therefore essentially different means of imposing the same temporal prior, with the former restricted to uniform temporal sampling.

In addition to developing a new model for multivariate inhomogeneous space-time Poisson process data, a contribution of this paper concerns computations, in the form of a detailed comparison of Markov chain Monte Carlo (MCMC) and variational Bayesian (VB) inference for this class of models. The former is widely used, but it can be computationally prohibitive for the motivating large-scale problems considered here. Computations based on VB are attractive for large-scale modeling studies, but many simplifying assumptions must be made.

1.3 Related research

A natural model for exploiting spatial information, and to model point process data, is the inhomogeneous Poisson process (Diggle 2003),(Møller and Waagepetersen 2004). Researchers have recently studied nonparametric Bayesian approaches for such applications. One of these approaches models the Poisson intensity function by a variation of a Gaussian process (GP) (Adams et al. 2009; Rathbun and Cressie 1994; Møller et al. 1998). The log-Gaussian Cox process (Møller et al. 1998), corresponding to an intensity function modeled as an exponentiated GP, has proven highly successful in point process

(Hossain and Lawson 2009) and geostatistical modeling (Diggle et al. 2010; Pati et al. 2010). Mixture models provide another approach to representing the Poisson intensity function (Wolpert and Ickstadt 1998). Kottas and Sansó (2007) proposed a Dirichlet process (DP) mixture model of bivariate beta densities to model heterogeneity in intensity functions. Dirichlet process mixture models of multivariate normal densities can be also found in (Ji et al. 2009; Chakraborty and Gelfand 2010).

In Taddy (2008, 2010); Taddy and Kottas (2012) a dynamic model was proposed for Poisson point processes, based on a novel version of the dependent Dirichlet process. Models of this type have been applied to the data considered in Figure 1, although the problem of segmentation was not considered. In Achcar et al. (2011) a time inhomogeneous Poisson model was proposed, with change-points to estimate the number of times that a given environmental standard is violated in a time interval of interest.

Rather than modeling the Poisson intensity via a GP or a DP mixture model, the model in (1) constitutes a mixture model with space-time mixture weights, and the spatial locations $\{\boldsymbol{s}_i\}$ of the grid are modeled as covariates. The details of how $w_k(\boldsymbol{s};\boldsymbol{\theta}_{kt})$ is modeled encourages contiguous regions in space and time for which a single component (cluster) dominates, encouraging a piecewise-constant Poisson intensity function. In Heikkinen and Arjas (1998) the authors similarly build a piecewise constant prior model for spatial Poisson intensities, using Voronoi tessellations. We model $w_k(\boldsymbol{s};\boldsymbol{\theta}_{kt})$ via an extension of the logistic stick-breaking process (LSBP) (Ren et al. 2011). The region of interest is partitioned into a set of contiguous small square cells, with related ideas considered in Hossain and Lawson (2009). Within the context of the aforementioned GP construction for the temporal dependence of $\boldsymbol{\theta}_{kt}$, related ideas were presented in the context of factor analysis (Luttinen and Ilin 2009), where GPs were used to describe the smoothness of both spatial locations and time. An AR model for temporal dynamics was considered in Taddy (2008, 2010).

2 Model Details

2.1 Basic construction

The proposed space-time model for data $\mathcal{D} = \{\boldsymbol{s}_i, \boldsymbol{v}_{it}\}_{i=1,\ldots,M,t=1,\ldots,T}$ is summarized as

$$\boldsymbol{v}_{it} \sim \prod_{j=1}^{d} \text{Poisson}(\lambda_{ijt}), \qquad \boldsymbol{\lambda}_{it} \sim \sum_{k=1}^{K} w_k(\boldsymbol{s}_{it})\delta_{\boldsymbol{\lambda}^*_{kt}} \tag{2}$$

$$w_k(\boldsymbol{s}_{it}) = p_k(\boldsymbol{s}_{it}) \prod_{h=1}^{k-1} [1 - p_h(\boldsymbol{s}_{it})] \tag{3}$$

$$p_k(\boldsymbol{s}_{it}) = \sigma(g_k(\boldsymbol{s}_{it})), \text{ for } k = 1, ..., K-1, \qquad p_K(\boldsymbol{s}_{it}) = 1 \tag{4}$$

$$g_k(\boldsymbol{s}_{it}) = \sum_{j=1}^{J} \beta_{kjt} \mathcal{K}(\boldsymbol{s}_{it}, \tilde{\boldsymbol{s}}_j; \psi_k) + \beta_{k0t} \tag{5}$$

where (2) is repeated here from (1), for convenience. Below we explain and motivate each term in this construction. Parameters $\boldsymbol{\theta}_{kt}$ from the Introduction correspond here to $\{\beta_{kjt}\}_{j=0,J}$ and ψ_k. In what follows, the notation $\boldsymbol{s}_{it}$ is meant to assign statistics to spatial location $\boldsymbol{s}_i$ at time t; for example, $w_k(\boldsymbol{s}_{it})$ is the kth mixture weight as observed at $\boldsymbol{s}_i$ and time t. The spatial grid defining the regions $\{\Delta(\boldsymbol{s}_i)\}_{i=1,M}$ is not changing with time.

The expression in (3), with $p_k(\boldsymbol{s}_{it}) \in [0,1]$ for all $\boldsymbol{s}_{it}$, is suggestive of the stick-breaking representation of the Dirichlet process (Sethuraman 1994). The function $\sigma(x) = \exp(x)/(1+\exp(x))$ is associated with a logistic model, and $p_K(\boldsymbol{s}_{it}) = 1$ such that $\sum_{k=1}^{K} w_k(\boldsymbol{s}_{it}) = 1$ for all $\boldsymbol{s}_{it}$. By the construction of $g_k(\boldsymbol{s}_{it})$ in (5), the probabilities $p_k(\boldsymbol{s}_{it})$ have space-time variation, with such variation transferred to the mixture weights $w_k(\boldsymbol{s}_{it})$ via (3). Therefore, via mixture weights $w_k(\boldsymbol{s}_{it})$ in (2) we constitute a multivariate Poisson mixture model, with weights that vary as a function of $\boldsymbol{s}_{it}$.

Function $\mathcal{K}(\boldsymbol{s}, \tilde{\boldsymbol{s}}_j; \psi_k)$ denotes a kernel with parameter ψ_k. Here we employ the radial basis function $\mathcal{K}(\boldsymbol{s}, \tilde{\boldsymbol{s}}_j; \psi_k) = \exp(-\|\boldsymbol{s} - \tilde{\boldsymbol{s}}_j\|_2^2/\psi_k)$, with J predefined kernel centers $\{\tilde{\boldsymbol{s}}_j\}_{j=1,J}$; for convenience these J centers are here aligned with the centers of the spatial grid defined by $\Delta(\tilde{\boldsymbol{s}}_j)$ (recall discussion in the Introduction). The appropriate kernel parameters $\{\psi_k\}$ will be inferred. To ease computations, we assume a discrete set of parameters $\{\psi_1^*, \ldots, \psi_L^*\}$ over which a uniform prior is placed; each kernel parameter ψ_k is assumed drawn from this finite library of parameters.

The space-time dependence of the model is manifested in how $\{\beta_{kjt}\}_{j=0,J}$ and $\{\boldsymbol{\lambda}_{kt}^*\}$ are modeled.

2.2 Temporal modeling

When the data are sampled uniformly in time, an autoregressive (AR) temporal model is natural. Specifically, we consider

$$\beta_{kjt} \sim \mathcal{N}(\zeta \beta_{kj(t-1)}, \alpha_\beta^{-1}) \ , \quad j = 0, \ldots, J \tag{6}$$

$$\log \lambda_{kjt}^* \sim \mathcal{N}(\xi \log \lambda_{kj(t-1)}^*, \alpha_\lambda^{-1}) \ , \quad j = 1, \ldots, J \tag{7}$$

with $\beta_{kj0} = \log \lambda_{kj0}^* = 0$. Gamma priors are placed on α_β and α_λ. Further, ζ and ξ are drawn from a truncated normal $\mathcal{N}_{(0,1)}(0,1)$ with $0 < \zeta, \xi < 1$.

The collection of data may be expensive, and there may be situations for which nonuniform temporal sampling is desired (*e.g.*, to provide fine-scale sampling in particular regions – seasons – of time that may be interesting). This suggests using a Gaussian process (GP) model (Rasmussen and Willams 2006) for the temporal variation of β_{kjt} and $\log \lambda_{kjt}^*$.

For the kth mixture component, we let

$$\boldsymbol{B}_k \sim \mathcal{N}(\boldsymbol{B}_k | \mathbf{0}, \boldsymbol{\Omega}_k) = \prod_{j=0}^{J} \mathcal{N}(\boldsymbol{\beta}_{kj:} | \mathbf{0}, \boldsymbol{\Sigma}_{kj}), \qquad [\boldsymbol{\Sigma}_{kj}]_{il} = c_0 {c_1}^{|t_i - t_l|} \tag{8}$$

where $\boldsymbol{\beta}_{kj:} = [\beta_{kj1}, ..., \beta_{kjT}]^T$, and $\boldsymbol{B}_k \in \mathbb{R}^{T(J+1)}$ denotes a vector formed by concatenating $\boldsymbol{\beta}_{kj:}$ for $j = 0, ..., J$. The covariance $\boldsymbol{\Omega}_k$ is a block-diagonal matrix of size $T(J+1) \times T(J+1)$, and each block $\boldsymbol{\Sigma}_{kj}$ is a $T \times T$ covariance matrix; the entry at row i and column l, denoted as $[\boldsymbol{\Sigma}_{kj}]_{il}$, is evaluated using the GP covariance function with the hyperparameters $\{c_0, c_1\}$. A gamma prior is placed on c_0. Since c_1 plays the same role as ζ, we also draw c_1 from the truncated normal $\mathcal{N}_{(0,1)}(0, 1)$ with $0 < c_1 < 1$.

The Gaussian process priors are also placed on $\log \lambda^*_{kjt}$. For mixture component k

$$\log(\boldsymbol{\lambda}^*_{kj:}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_{kj}), \qquad [\boldsymbol{\Gamma}_{kj}]_{il} = d_0 d_1{}^{|t_i - t_l|} \tag{9}$$

where $\log(\boldsymbol{\lambda}^*_{kj:}) = [\log(\lambda^*_{kj1}), ..., \log(\lambda^*_{kjT})]^T$, and the covariance matrix $\boldsymbol{\Gamma}_{kj} \in \mathbb{R}^{T \times T}$, with the entries defined by the GP covariance function with the hyperparameters $\{d_0, d_1\}$. A gamma prior and truncated normal prior are placed on d_0 and d_1. As discussed in the Introduction, the considered AR(1) and GP priors are consistent, and provide different modeling strategies for the same imposed temporal dynamics.

2.3 Model interpretation

Equations (3)-(5) are of the form of the logistic stick-breaking process (LSBP) introduced in Ren et al. (2011); however, that paper did not consider Poisson data, and space-time processes were not addressed. Recall that $\sigma(x) \approx 1$ for $x > 4$; we refer to this as the "clipping" property of the logistic, as all x larger than about 4 contribute effectively in the same manner to $\sigma(x)$; one may alternatively use a probit model, to achieve the same end. If $\beta_{kjt} > 4$, then $p_k(\boldsymbol{s}) \approx 1$ for $\|\boldsymbol{s} - \tilde{\boldsymbol{s}}_j\|_2^2 < \psi_k$. This implies via (3) that within region $\|\boldsymbol{s} - \tilde{\boldsymbol{s}}_j\|_2^2 < \psi_k$, if $\beta_{kjt} > 4$ mixture component k is highly probable (assuming that other clusters $k' \neq k$ do not have large $p_{k'}(\boldsymbol{s})$ in the vicinity of $\tilde{\boldsymbol{s}}_j$). The "clipping" nature of the logistic function, and large values of $\beta_{kjt} > 4$, encourage contiguous regions for which a given cluster k has high space-time probability of being manifested (all locations $\boldsymbol{s}$ at which $g_k(\boldsymbol{s}) > 4$ have similarly high probability of being associated with cluster k, regardless of the exact value of $g_k(\boldsymbol{s})$). The weights $\{\beta_{kjt}\}$ play the role of assigning which regions in space-time are most likely to be associated with a given cluster k, and ψ_k defines the size scale of the cluster. Note that while we truncate the model to K mixture components, this does not mean that all components need actually be used to represent the data. For example, if a given β_{k0t} is large and negative, then the kth mixture component is unlikely to be utilized at all spatial locations at time t; K is simply an upper bound on the number of mixture components (segment types).

3 Posterior inference

The posterior distribution of the model parameters is inferred via an MCMC sampler and via variational Bayesian (VB) inference (Beal 2003). The VB inference typically converges quickly and is computationally efficient; by contrast, MCMC convergence may be difficult to diagnose, and a large number of iterations is required to collect samples

representing the joint posterior distribution. The detailed MCMC and VB update equations are provided in the Appendix (we provide equations for the GP model, with minor changes manifested for the AR case). Since VB analysis is not as widely used in the statistics literature, for completeness we provide details on its modeling assumptions.

Let $\boldsymbol{\Theta}$ represent a vector of all model parameters; the goal is to infer the posterior $p(\boldsymbol{\Theta}|\mathcal{D})$. The likelihood of the data is represented by $p(\mathcal{D}|\boldsymbol{\Theta})$ and the prior on the model parameters is denoted by $p(\boldsymbol{\Theta})$. Let $q(\boldsymbol{\Theta};\boldsymbol{\Gamma})$ be a parametric distribution with hyperparameters $\boldsymbol{\Gamma}$, and consider the variational expression

$$\mathcal{F}(\boldsymbol{\Gamma}) = \int d\boldsymbol{\Theta} q(\boldsymbol{\Theta};\boldsymbol{\Gamma}) \ln \frac{q(\boldsymbol{\Theta};\boldsymbol{\Gamma})}{p(\mathcal{D}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})} = D_{KL}[q(\boldsymbol{\Theta};\boldsymbol{\Gamma})\|p(\boldsymbol{\Theta}|\mathcal{D})] - \ln p(\mathcal{D}). \tag{10}$$

In VB analysis the goal is to optimize the hyperparameters $\boldsymbol{\Gamma}$ to minimize the Kullback-Leibler divergence between $q(\boldsymbol{\Theta};\boldsymbol{\Gamma})$ and the true posterior $p(\boldsymbol{\Theta}|\mathcal{D})$; this corresponds to adjusting $\boldsymbol{\Gamma}$ in $q(\boldsymbol{\Theta};\boldsymbol{\Gamma})$ such that $\mathcal{F}(\boldsymbol{\Gamma})$ is minimized. Note that $\int d\boldsymbol{\Theta} q(\boldsymbol{\Theta};\boldsymbol{\Gamma}) \ln \frac{q(\boldsymbol{\Theta};\boldsymbol{\Gamma})}{p(\mathcal{D}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})}$ is only a function of the likelihood $p(\mathcal{D}|\boldsymbol{\Theta})$ and the prior $p(\boldsymbol{\Theta})$, and *not* the unknown posterior; with careful selection of $q(\boldsymbol{\Theta};\boldsymbol{\Gamma})$, numerical techniques akin to expectation-maximization (EM) (Beal 2003) can be employed to minimize $\mathcal{F}(\boldsymbol{\Gamma})$, with assurance of convergence to a local-optimal solution.

Focusing on the GP temporal model (the AR case is very similar), the model parameters are

$$\boldsymbol{\Theta} = \{\{\boldsymbol{\lambda}^*_{kj:}\}_{\substack{j=1,\ldots,d,\\ k=1,\ldots,K}}, \{\boldsymbol{B}_k\}_{k=1,\ldots,K}, \{Z_k(\boldsymbol{s}_{it})\}_{\substack{t=1,\ldots,T,\\ i=1,\ldots M,\\ k=1,\ldots,K}}, c_0, c_1, d_0, d_1\} \tag{11}$$

where $Z_k(\boldsymbol{s}_{it}) \sim \text{Bernoulli}(p_k(\boldsymbol{s}_{it}))$, with $p_k(\boldsymbol{s}_{it})$ defined in (4). Completing the generative process, $\boldsymbol{v}_{it} \sim \prod_{j=1}^{d} \text{Poisson}(\lambda^*_{\hat{k}jt})$ if $Z_k(\boldsymbol{s}_{it}) = 0$ for $k < \hat{k}$ and $Z_{\hat{k}}(\boldsymbol{s}_{it}) = 1$; $\lambda^*_{\hat{k}jt}$ is the jth component of vector $\boldsymbol{\lambda}^*_{\hat{k}t}$.

In VB one typically assumes a factorized form for $q(\boldsymbol{\Theta};\boldsymbol{\Gamma})$, *i.e.*, $q(\boldsymbol{\Theta};\boldsymbol{\Gamma}) = \prod_l q_l(\boldsymbol{\Theta}_l;\boldsymbol{\Gamma}_l)$, where $\boldsymbol{\Theta}_l$ represents the lth set of model parameters and $q_l(\boldsymbol{\Theta}_l;\boldsymbol{\Gamma}_l)$ is a parametric density function with hyperparameters $\boldsymbol{\Gamma}_l$; the union of all $\boldsymbol{\Theta}_l$ corresponds to $\boldsymbol{\Theta}$. Through careful selection of $q_l(\boldsymbol{\Theta}_l;\boldsymbol{\Gamma}_l)$ one may iteratively optimize the variational expression $\mathcal{F}(\boldsymbol{\Theta})$.

For the proposed model, $q(\boldsymbol{B}_k)$ is a multivariate normal distribution, $q(Z_k(\boldsymbol{s}_{it}))$ is Bernoulli (with Bernoulli probability defined by a logistic function), $q(\psi_k)$ is multinomial based upon a finite library of possible parameters $\{\psi^*_l\}_{l=1,L}$, and $q(c_0)$ and $q(d_0)$ are gamma distributions. It is not possible to define a $q(\boldsymbol{\lambda}^*_{kj:})$ that yields closed-form updates. Therefore, the parameters $\boldsymbol{\lambda}^*_{kj:}$ within the VB analysis are also approximated at each iteration via a point estimate that maximizes the functional $\mathcal{F}(\boldsymbol{\Gamma})$. Similarly, $q(c_1)$ and $q(d_1)$ cannot be obtained in closed form. The parameters c_1 and d_1 are updated on each VB iteration by defining parameters that maximize the functional $\mathcal{F}(\boldsymbol{\Gamma})$.

4 Example Results

While the proposed model may appear relatively complicated, the number of hyperparameters that need be set is actually modest. We compare the AR-LSBP and GP-LSBP models for imposing a prior on the temporal dependence with a simpler model in which the priors for each time point t are independent. In the context of this independent LSBP (ind-LSBP), we impose

$$\beta_{kjt} \sim \mathcal{N}(0, \alpha_{kjt}^{-1}) \ , \quad \alpha_{kjt} \sim \text{Gamma}(a_0, b_0) \tag{12}$$

and we set $a_0 = b_0 = 10^{-6}$ as in the relevance vector machine (RVM) (Tipping 2001). The same gamma priors are placed on α_β and α_λ for the AR-LSBP model, and on c_0 and c_1 for the GP-LSBP model. In all examples the truncation level on the LSBP was set at $K = 20$, and the results are insensitive to this parameter, as long as it is large relative to the actual number of clusters/segments inferred by the model. Finally, we must specify the library for kernel parameters $\{\psi_k\}_{k=1,K}$; the manner in which these are specified is discussed when presenting the specific examples.

For uniform temporal sampling, the AR(1) and GP imposition of temporal dynamics are theoretically identical, for the imposed GP covariance. Nevertheless, even for uniform temporal sampling we show results for both of these implementations, because the details of the numerics dictates that the two models are slightly different in practice. Specifically, within the GP model a point estimate is employed for the kernel hyperparameters, with this obviously unnecessary for the direct AR(1) model. The comparison allows examination of the accuracy of this approximation within the GP inference, relative to the direct AR(1) implementation; this sheds light on the quality of the computations for non-uniform temporal sampling, where the GP implementation is required.

4.1 Simulation Example

We assume the data are constructed by a total of 9 equally spaced time instances, $t = 1, 2, ..., 9$. At each time we randomly draw 50 spatial locations in one-dimensional space from a uniform distribution with support $[0, 20]$, denoted as $s_{it} \sim \text{Uniform}[0, 20]$, $i = 1, ..., 50, t = 1, ...9$. For each location, we draw an event count v_{it} from a Poisson distribution with the intensity parameter λ_{it}. To represent the time dynamics, we let $\lambda_{it} = 20$ when $5 + \frac{5}{8}(t-1) \leq s_{it} \leq 10 + \frac{5}{8}(t-1)$, and $\lambda_{it} = 1$ otherwise. By this setting the high-intensity window moves gradually from [5,10] to [10, 15] when time t increases. Note that here $s_{it} \in \mathbb{R}^1$ and $v_{it} \in \mathbb{R}^1$. The kernel centers are defined as $\tilde{s}_j = 0.5(j-1)$ for $j = 1, ..., J$. The data are depicted in Figure 2. Within the analysis, the library of kernel parameters is the union of the following two sets: $\{0.05, 0.1, 0.05, \ldots, 0.5\}$ and $\{0.5, 1, 1.5, \ldots, 5\}$.

The mean results from VB are shown in Figure 3, in which the inferred Poisson rate is constituted; for these and all VB results the computations were stopped when the change in the variational bound changed by 10^{-4}. Further, all VB results are initialized at random. The VB results presented below represent a local-optimal solution, which

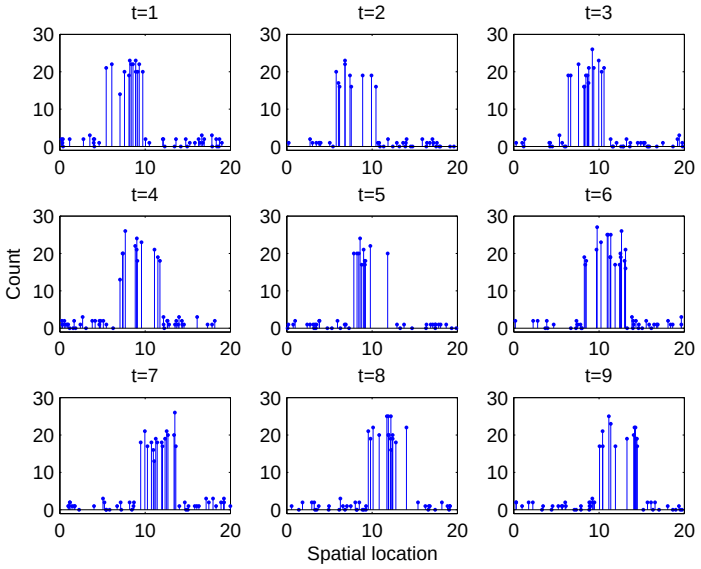
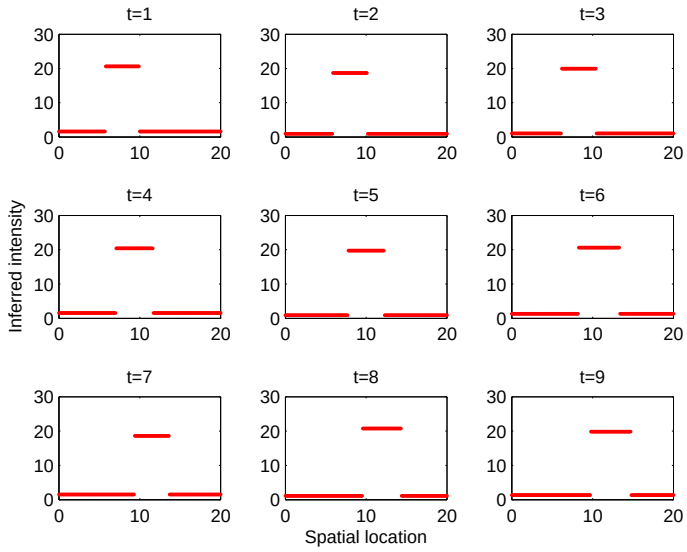


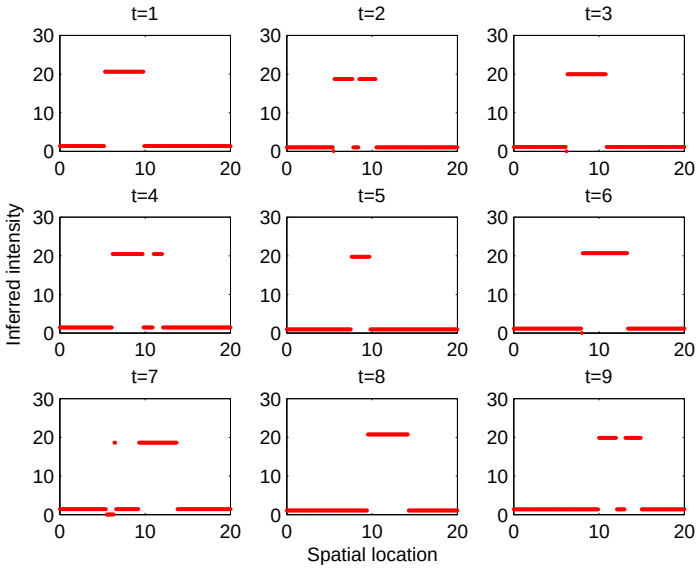
Figure 2: Simulation example. The high-intensity window moves gradually from [5,10] to [10, 15] when time increases.

forms one source of error, and this is compounded by the factorized approximation to the posterior. Nevertheless, the VB implementation of the GP-LSBP and AR-LSBP model yields results comparable to that of the MCMC implementation. When implementing MCMC, a total of 10,000 iterations are run, with the first 1000 discarded as burn-in. On the same PC (and both codes written in Matlab), the VB GP-LSBP and AR-LSBP results required approximately 158 seconds of CPU time, while the VB ind-LSBP results required approximately 96 seconds. In contrast, the GP-LSBP and AR-LSBP results based on the MCMC sampler required 6517 seconds, and ind-LSBP required 2913 seconds (109 and 48 minutes, respectively). The software was *not* optimized, and these numbers therefore represent a *relative* view of computational expense of the VB and MCMC solutions.

From Figure 3 it is observed that, for the VB solution, incorporation of temporal smoothness in the GP-LSBP model yields significant improvements in the inferred Poisson rate, as compared to the VB ind-LSBP solution (with temporal dependence not accounted for in the prior); the AR-LSBP model performed similarly to GP-LSBP. It appears that the prior constraint imposed by GP/AR within the VB solution plays an important role in mitigating the underlying VB approximations. By contrast, for the MCMC results improvements are manifested via GP-LSBP and AR-LSBP relative to ind-LSBP, but in this case the differences are less dramatic (plots of MCMC results are not shown, for brevity).



(a) GP-LSBP inferred based on VB



(b) ind-LSBP inferred based on VB

Figure 3: Segmentation and latent intensity inferred by VB: Comparison between GP-LSBP and ind-LSBP, considering the simulated-data example. The AR-LSBP results are similar to the GP-LSBP results, and are omitted for brevity.

We next examine the generative performance of the proposed model. After the model has been learned, either via VB or MCMC, we randomly generate 100 new test data, following the same procedure that generated the training data. We then compute the average log-likelihood and the accuracy rate of segmentation from the learned GP-LSBP, AR-LSBP and ind-LSBP models. The accuracy rate of segmentation is defined as the number of test data points segmented correctly as a fraction of the total number of test data points. The results are summarized in Table 1. We find that the GP-LSBP and AR-LSBP models achieve a higher likelihood and accuracy of segmentation compared to the ind-LSBP. Note that the differences between GP-LSBP, AR-LSBP and ind-LSBP are relatively modest for the MCMC solution, while there are again marked advantages in the GP-LSBP and AR-LSBP solutions relative to ind-LSBP when employing VB inference.

Table 1: Comparison of generative performance between AR-LSBP, GP-LSBP and ind-LSBP, on simulated data.

Method	Average log-likelihood		Accuracy rate of segmentation	
	VB	MCMC	VB	MCMC
AR-LSBP	-3.702	-1.749	0.9796	0.9801
GP-LSBP	-3.882	-2.082	0.9765	0.9757
ind-LSBP	-15.544	-2.274	0.9478	0.9741

Table 2: Comparison of prediction performance between AR-LSBP, GP-LSBP and ind-LSBP.

N_{miss}	Average log-likelihood					
	AR-LSBP		GP-LSBP		ind-LSBP	
	VB	MCMC	VB	MCMC	VB	MCMC
1	-3.948	-1.975	-4.102	-2.123	-21.194	-2.641
2	-4.211	-2.241	-4.526	-2.473	-27.195	-3.077
3	-4.468	-2.573	-4.718	-2.652	-27.776	-3.507
4	-4.882	-2.740	-5.133	-3.108	-26.682	-3.963
5	-5.801	-3.014	-5.987	-3.521	-31.217	-4.316

N_{miss}	Accuracy rate of segmentation					
	AR-LSBP		GP-LSBP		ind-LSBP	
	VB	MCMC	VB	MCMC	VB	MCMC
1	0.9792	0.9794	0.9767	0.9758	0.7165	0.9545
2	0.9787	0.9786	0.9761	0.9754	0.6669	0.9581
3	0.9787	0.9785	0.9763	0.9752	0.6458	0.9379
4	0.9780	0.9783	0.9752	0.9740	0.6647	0.9274
5	0.9763	0.9770	0.9741	0.9633	0.6131	0.9066

Finally we test the prediction performance of the model. We first generate data $\mathcal{D} = \{s_i, v_{it}\}_{i=1,...,50,\ t=1,...,9}$ as discussed above, and then randomly select N_{miss} time instances $\hat{t}_1, ..., \hat{t}_{N_{miss}}$ from $t = 1, ..., 9$, and this constructs our test data $\mathcal{D}_{tst}$; the train-

ing data $\mathcal{D}_{trn}$ is composed of the data in $\mathcal{D}$ but not in $\mathcal{D}_{tst}$. We learn the model based on VB or MCMC analysis with $\mathcal{D}_{trn}$, and predict the kernel weights $\hat{\beta}_{kj\hat{t}}$ and Poisson intensities $\hat{\lambda}^*_{k\hat{t}}$ at time $\hat{t}$. The average log-likelihood and accuracy of segmentation are evaluated based on the prediction results of $\mathcal{D}_{tst}$, given only the spatial locations $\hat{s}_{i\hat{t}}$. We perform 100 trials, and at each trial N_{miss} time instances are selected randomly to construct $\mathcal{D}_{tst}$. The average results are shown in Table 2.

Only the GP-LSBP results are fully principled in this analysis, where we use the learned parameters of the GP covariance matrix to interpolate to new time points (Rasmussen and Willams 2006). The AR model implicitly assumes that the data are sampled uniformly in time, while the ind-LSBP has no principled means of interpolating to missing time points. Nevertheless, as a comparison, for the AR-LSBP computations in this test the AR component was simply applied to consecutive observed time points, essentially assuming that the temporal variation was smooth, even if not sampled uniformly. To interpolate to new points using the learned AR-LSBP and ind-LSBP results, to obtain model parameters at any new point $\hat{t}$, we average the learned model parameters from the two closest observed points, before and after $\hat{t}$. From Table 2 it is observed that again for the VB solution, there is a marked advantage manifested via the GP-LSBP and AR-LSBP priors, as compared to ind-LSBP. For the MCMC solution, there is also a noticeable advantage manifested via the GP-LSBP and AR-LSBP solutions, particularly for segmentation accuracy for relatively large N_{miss}. Based upon the average log-likelihood, we note a small but consistent advantage of the AR-LSBP model over the GP-LSBP counterpart, for both VB and MCMC computations. This observation on simulated data will carry over to the analysis of real data.

4.2 Crime Data

We investigate crime events in Cincinnati, Ohio, USA; the data are available online at `http://www.cincinnati-oh.gov`. The data include the date, time, location and other information of all reported crimes in Cincinnati since 2006. This data set was first studied in Taddy (2008, 2010), where a mixture of beta distributions was employed to model the event density $\nu(\boldsymbol{s})$, and to discover the evolution of the density with time. In our problem we seek to segment the city into contiguous regions, with crime events at each region characterized by a common constant Poisson intensity vector.

We consider 117,314 crime events within the city, reported from January 2006 to December 2008. Each crime is assigned a uniform crime reporting (UCR) code. In total more than 170 different UCR codes describe a variety of crimes. These crime events can be categorized into 17 different crime types, based on the prefix of their UCR codes. They are: 1) murder, 2) rape, 3) robbery, 4) assault with weapon, 5) burglary, 6) nonvehicle theft, 7) vehicle theft, 8) general assault, 9) arson, 10) forgery, 11) fraud, 12) receiving stolen property 13) vandalism, 14) weapons related but no physical harm, 15) sexual crime, 16) children related, 17) general harassment. As an example, the locations (latitude and longitude coordinates) of the 3090 crime events in January 2008 are shown in Figure 1(a). Based on the locations of all the 117,314 crime events, the

observation window is considered within a rectangular region of [39.06°, 39.24°] latitude and [-84.70°, -84.35°] longitude.

We construct the data $\mathcal{D} = \{\boldsymbol{s}_i, \boldsymbol{v}_{it}\}_{i=1,...,M,\ t=1,...,T}$ as follows. The total crime events within one month are considered as one time instance, and therefore there are in total 36 time points. At each time, the observation window is divided into 15,750 small square grids (90 rows by 175 columns) of size $0.002° \times 0.002°$, and the event location $\boldsymbol{s}_{it}$ is defined as the center of each small square area, with this denoted as $\Delta(\boldsymbol{s}_i)$. The count v_{ijt} is then the number of Type j crimes within $\Delta(\boldsymbol{s}_i)$ over the corresponding month indexed by t. This produces a 17-dimensional count vector $\boldsymbol{v}_{it}$ at $\boldsymbol{s}_i$ for $i = 1, ..., 15750$ and $t = 1, ..., 36$. Related research in Taddy (2008, 2010) applied marked Poisson processes to address the crime types, regarding each crime type at $\boldsymbol{s}_{it}$ as a random mark. Here we attempt to segment the city by considering all the crime types within a local region $\Delta(\boldsymbol{s}_{it})$ as a correlated variable (a vector), instead of treating each event as a random type.

The proposed GP-LSBP, AR-LSBP and ind-LSBP models are inferred via VB and MCMC, with truncation level $K = 20$. The kernel centers are uniformly spaced every 0.04° (latitude and longitude) in the observation window, with a total of 60 kernel centers defined. The library of kernel parameters $\{\psi_l^*\}_{l=1,L}$ is the union of the following sets: $\{0.006°, 0.012°, 0.018°, \ldots, 0.06°\}$ and $\{0.06°, 0.12°, 0.18°, \ldots, 0.6°\}$. On the same PC, the VB GP-LSBP and AR-LSBP results require approximately 2.8 hours of CPU time, while the VB ind-LSBP results required approximately 1.3 hours. By contrast, due to the large size of the data, 3000 MCMC samples are employed, with 1000 discarded as burn-in. With the same PC, the MCMC GP-LSBP and AR-LSBP results required approximately 47.5 hours. We also considered 10,000 MCMC samples, with 1000 discarded as burn-in (at very significant computational cost), with little change in the results relative to those presented below.

Figure 4(a) shows the VB-based segmentation of the entire spatial observation window at 36 time instances, using GP-LSBP (similar results were found using AR-LSBP, omitted for brevity). The city is segmented into 4 regions (inferred by the model), and the segmentation changes smoothly with time. For comparison, Figure 4(b) shows the segmentation results obtained by applying an independent LSBP (VB computations) at each time instance. It is observed that with GP priors the proposed model presents a spatial segmentation more consistently over time and spatially more contiguously than ind-LSBP.

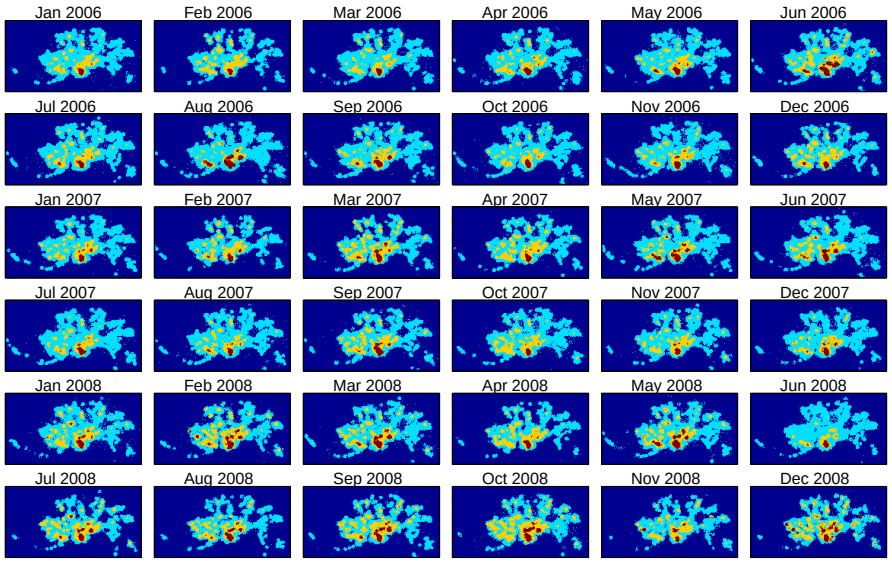
We are also interested in examining the clustering manifested by the MCMC computations, with this complicated by label switching between samples. We compute an MCMC clustering that may be compared to the VB results as follows. We consider one spatial location from Segment 1 in Figure 4, denoted $\boldsymbol{s}_1^*$. Based upon the MCMC collection samples, for each other spatial location in the scene $\boldsymbol{s} \neq \boldsymbol{s}_1^*$, we compute the probability that position $\boldsymbol{s}$ and $\boldsymbol{s}_1^*$ are in the same cluster. All positions $\boldsymbol{s}$ with high probability of such clustering should (ideally) constitute a spatial region similar to Segment 1 inferred via VB. In Figure 5(a) we show MCMC results for Segment 1, and the high-probability regions (red) do indeed align well with the VB results in Fig-

ure 4. In Figure 5(b) we compute similar MCMC results for Segment 2, and in this case the high-probability spatial locations are aligned well with the VB results for Segment 2 in Figure 4. We found in general good agreement between the VB and MCMC segmentation results for GP-LSBP and AR-LSBP for these data.

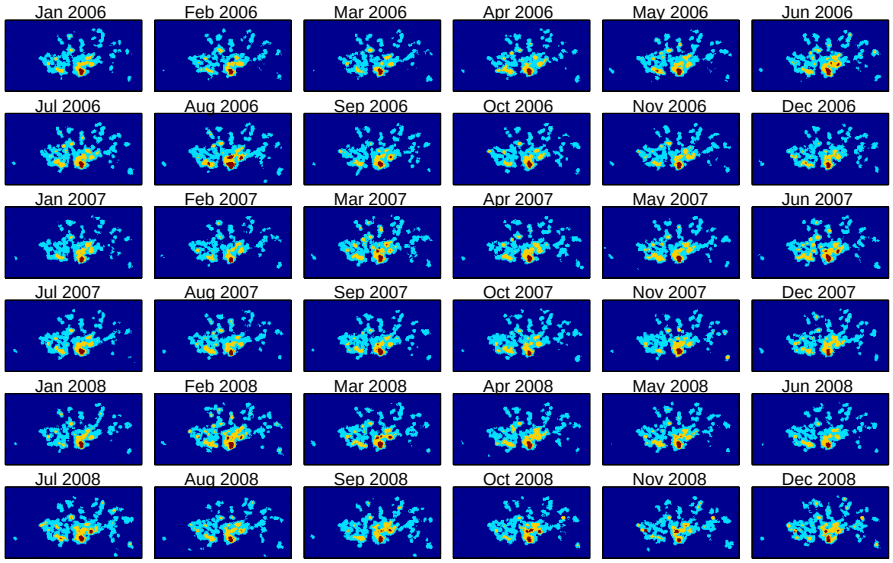
Figures 6(a)-(d) show the dynamic change of the VB-inferred Poisson intensities for each segment. To make the figure easier to read, we only plot components 3, 5 and 6 from the 17-dimensional vector $\boldsymbol{\lambda}_{kt}^*$; these components correspond to crime types "robbery", "burglary", and "nonvehicle theft", respectively. From these figures we observed that in all segments the crime intensities fluctuated periodically over season. Generally in summer there were more crime events of all types than than in winter. The overall crime intensities varied with regions. Segment 4 was in the downtown region, and had much more crime events compared to other regions. In all four regions Type 6 crime (nonvehicle theft) was dominant. In addition, the crime patterns were different in different regions. For example, Segment 4 had relatively less Type 5 crime (burglary), while in other 3 segments, the intensity of Type 5 crime was almost half of Type 6 crime. In Segment 4, Type 3 crime (robbery) was prevalent, while Segment 1 had relatively less Type 3 crime. For a comparison, we also present the MCMC-inferred Poisson intensities of Segment 3, as a representative (typical) example. It is observed that the MCMC and VB results are in generally good agreement for the GP-LSBP and AR-LSBP models.

These results may be used by police to assign resources (personnel) to segmented regions in a consistent manner, to address varying levels of crimes. The segments typically change with season, and the spatial distribution of resources may be temporally adjusted as well. By relating the demographics of regions to the spatial segments (we didn't have access to such demographics), one may deduce relationships between types of crimes and the types of people living and working in given regions, of interest to criminologists and city planners.

Following the same procedure as in the simulated example, we now examine the prediction performance of our model for the crime data. We randomly select N_{miss} time instances to construct a test set, and let the remaining data be the training set. Ten random trials are performed and the comparison of average log-likelihood between GP-LSBP, AR-LSBP and ind-LSBP inferred by VB is shown in Table 3. Since in this real application there is no ground truth, we cannot evaluate the accuracy rate of segmentation as done in the simulated example. From Table 3 GP-LSBP and AR-LSBP consistently achieve higher likelihood than the independent LSBP for various N_{miss} values. Note also that for these real data there is less of a difference between the AR/GP-LSBP and ind-LSBP results for the VB solution, as compared to the synthetic data considered above. We do not perform this experiment for MCMC inference, as the computational requirements needed to perform this many experiments are prohibitive with this large data set (however, in isolated tests, the results were slightly better than the VB-based GP-LSBP and AR-LSBP models, consistent with the simulated example above).

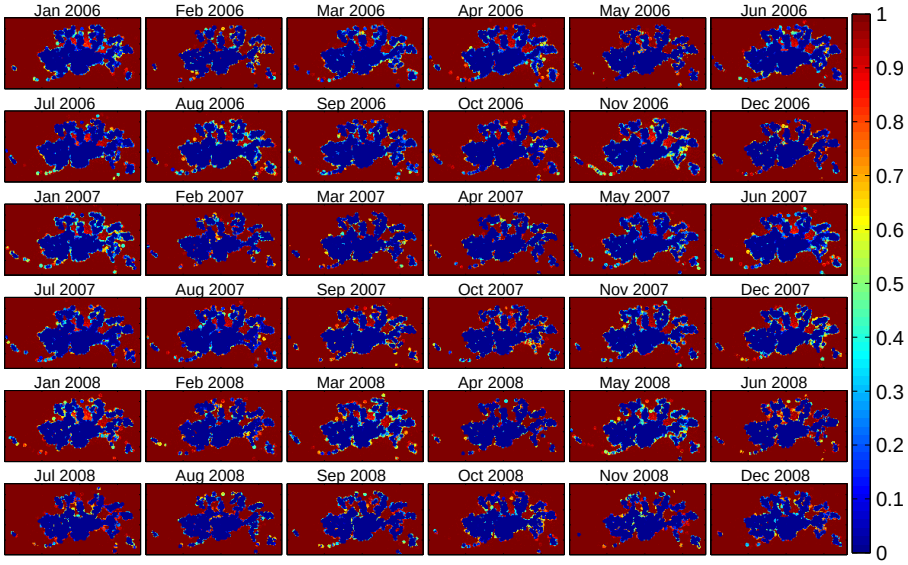


(a)

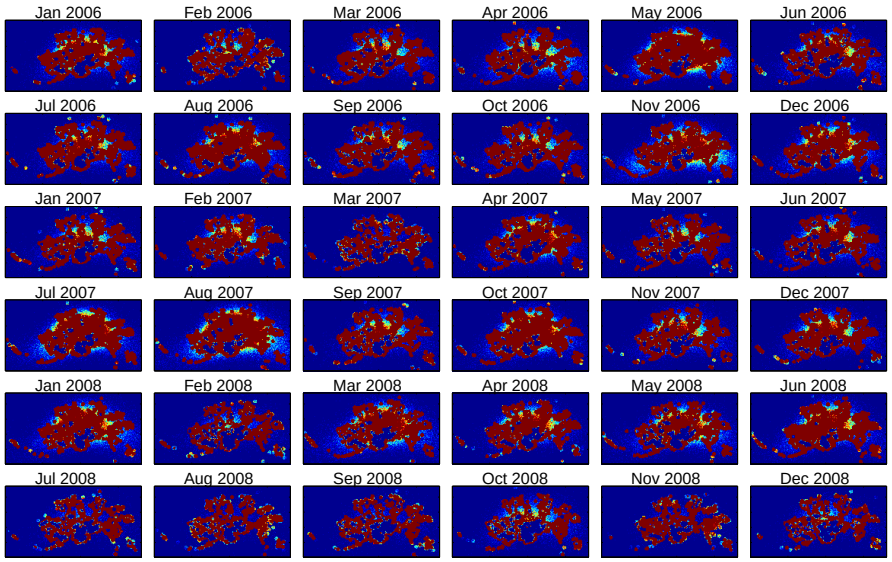


(b)

Figure 4: Comparison of spatial segmentation for crime data in Cincinnati, Ohio from January 2006 to December 2008 (VB results). Each color represents a segment with an associated intensity vector $\boldsymbol{\lambda}_{kt}^*$, and there are a total of four segments inferred: 1 - dark blue, 2 - light blue, 3 - yellow, and 4 - dark red. (a) GP-LSBP, (b) ind-LSBP.

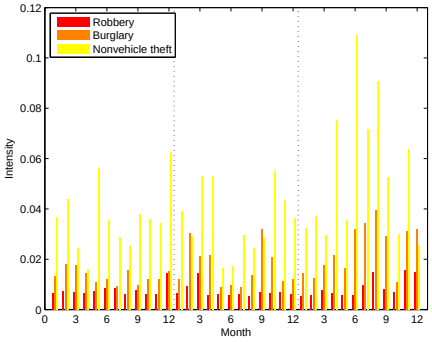


(a)

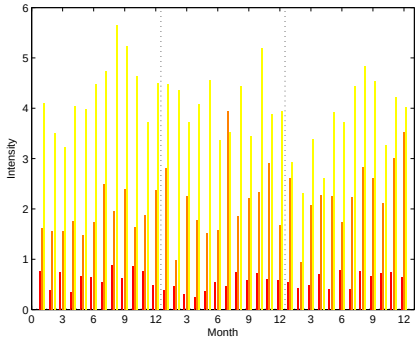


(b)

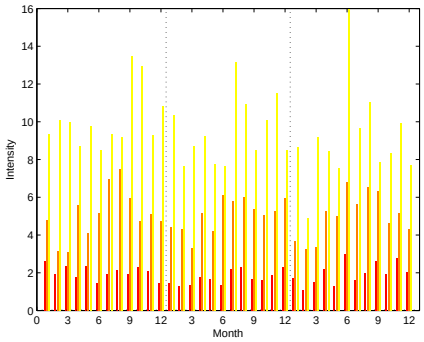
Figure 5: Comparison of spatial segmentation for crime data in Cincinnati, Ohio from January 2006 to December 2008 (MCMC results). (a) Segment 1, (b) Segment 2, where these segments are related to the results in Figure 4(a). The color scale is the same in (a) and (b).



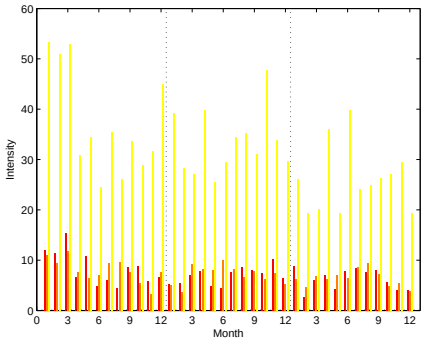
(a) Segment 1: Dark blue region in Fig. 4(a)



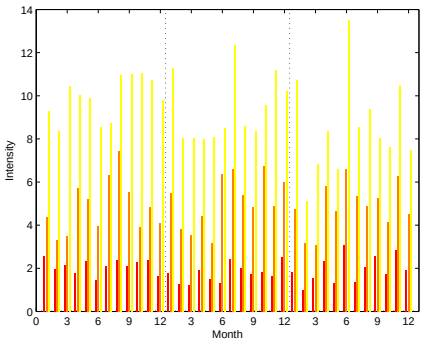
(b) Segment 2: Light blue region



(c) Segment 3: Yellow region



(d) Segment 4: Dark red region



(e) Segment 3 inferred by MCMC

Figure 6: Inferred intensity vector $\boldsymbol{\lambda}_{kt}^{*}$ associated with the segments shown in Figure 4(a). Only 3 crime types are shown here to make the figure easy to read.

Table 3: Comparison of average log-likelihood in the prediction for the crime data (VB inference).

N_{miss}	1	2	3	4	5	6
AR-LSBP	-6.131	-6.352	-7.204	-7.631	-7.957	-8.338
GP-LSBP	-6.570	-6.762	-7.713	-7.965	-8.426	-8.721
ind-LSBP	-8.666	-9.247	-9.595	-8.840	-9.848	-8.762

4.3 Pearson residuals

Following Taddy (2010), we check model quality via computation of Pearson residuals (see Turner et al. (2005) for a detailed discussion of residuals for spatial point processes). For the modeling framework considered here, the Pearson residual reduces to

$$R(\Delta(\boldsymbol{s}_{it}), \hat{\lambda}_{it}) = \frac{n_{it}}{\sqrt{\hat{\lambda}_{it}}} - \sqrt{\hat{\lambda}_{it}} \tag{13}$$

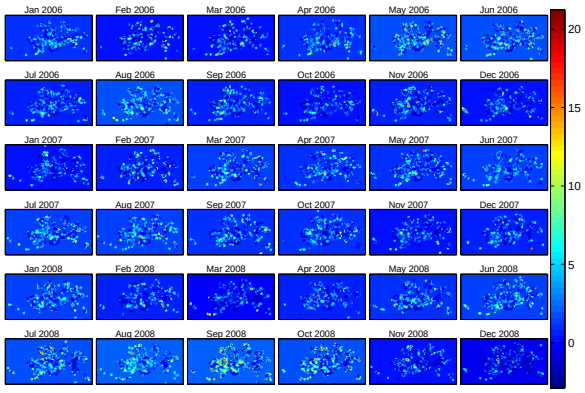
where n_{it} is the number of events in region $\Delta(\boldsymbol{s}_{it})$ and $\hat{\lambda}_{it}$ is the inferred Poisson rate parameter in small region $\Delta(\boldsymbol{s}_{it})$. Ideally the residual should be close to zero, if the underlying Poisson assumption is valid. Note that within the proposed model we have a *vector* of counts $\boldsymbol{v}_{it}$, and therefore we may compute the residual for each of the different types of crimes.

From Figure 7, which is based upon VB inference, we observe that the Pearson residuals tend to decrease substantially based upon a model that explicitly imposes temporal smoothness (note that the residuals are significantly lower for GP-LSBP and AR-LSBP, relative to ind-LSBP). Further, the AR-LSBP residuals are smaller than those of the GP-LSBP. Although we omit the MCMC results for brevity, similar phenomena were observed in that case. The residuals tend to be small, in the range [-2,2], with the larger values manifested on the edges of segments, as might be expected (segment interfaces are characterized typically by abrupt changes in statistical properties).

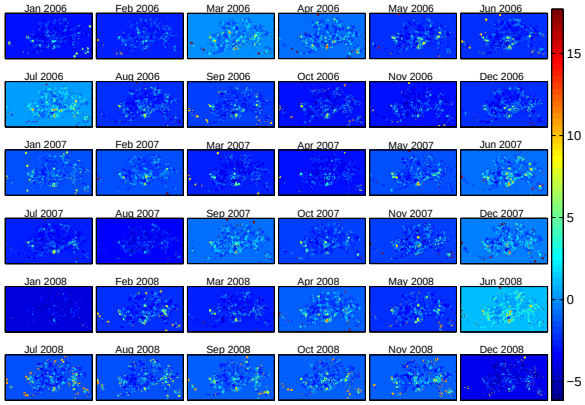
5 Conclusions

A Bayesian hierarchical model has been presented for segmenting time-evolving point process data, when the events are in vector form. The spatial-dependent point process is modeled using a generalization of a Poisson process, with piecewise constant Poisson intensities defined within the observation window. The logistic stick-breaking process is employed to favor spatially contiguous segments, and GP and AR models are considered for imposition of temporal smoothness of the segmentation and the Poisson intensity.

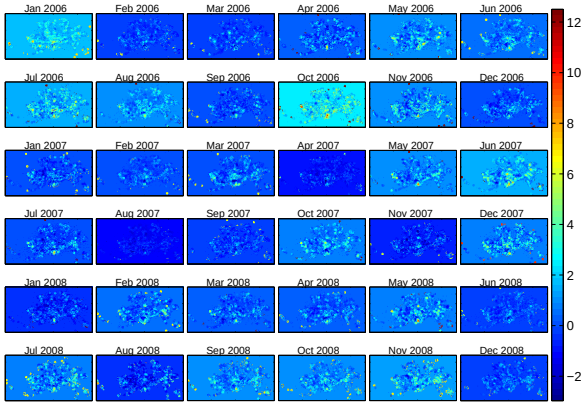
In addition to developing the model, a contribution of this paper concerns a detailed comparison between MCMC sampling and a VB approximation. For both the synthetic and real data, it was found that the GP-LSBP and AR-LSBP results computed via VB and MCMC were in close agreement, and the imposition of temporal smoothness manifested via GP/AR (compared to treating the different temporal samples indepen-



(a)



(b)



(c)

Figure 7: Pearson residuals for "nonvehicle theft," using VB inference; best viewed electriconically, zoomed in. (a) ind-LSBP, (b) GP-LSBP, (c) AR-LSBP.

dently) yielded significant improvements in the VB results. While the VB results are approximate, and are subject to local-optimal solutions (although the GP/AR models seemed to mitigate this to some extent), the VB approach provides significant advantages with regard to computations. For the large crime data set considered, while the MCMC results are in principle convergent, if run for enough samples, this attractiveness is tempered by the very significant computation time required to realize a number of collection samples to assure that we are indeed sampling from the posterior. Given that computational requirements will in practice mitigate the ability to collect as many MCMC samples as desired (and therefore MCMC is also an approximation), the VB solution appears to be an attractive option. However, the results presented here indicate that imposition of as much information as possible (here smoothness via GP/AR) is desirable. In future research it is of interest to consider *online* VB analysis (Hoffman et al. 2010), which provides further acceleration for large datasets, and it is appropriate for time-dependent data observed in an online/sequential manner, like the time-evolving crime data considered here.

Appendix: MCMC and VB Update Equations

A.1 MCMC Inference

The MCMC computations are performed using Gibbs sampling where the conditional density functions are analytic, and samples are drawn from the conditional density functions via Metropolis-Hastings when not analytic. The update equations are summarized as follows.

- Sample $\boldsymbol{\lambda}^*_{kj:}$ from their respective posteriors conditional on $\{Z_k(\boldsymbol{s}_{it})\}$ and $\{\nu_{ijt}\}$:

$$p\left(\boldsymbol{\lambda}^*_{kj:}\middle|-\right) \propto \prod_{t=1}^{T}\prod_{i=1}^{M} \text{Poisson}\left(\nu_{ijt}\middle|\,\lambda^*_{kjt}\right)^{\mathbf{I}(c_i=k)} \ln\mathcal{N}\left(\boldsymbol{\lambda}^*_{kj:}\middle|\,0,\boldsymbol{\Gamma}_{kj}\right). \tag{14}$$

  It is not possible to sample $\boldsymbol{\lambda}^*_{kj:}$ from the full conditionals. We update each $\boldsymbol{\lambda}^*_{kj:}$ by the Metropolis-Hastings algorithm. When updating $\boldsymbol{\lambda}^*_{kj:}$, the proposed $\boldsymbol{\lambda}^{*(\tau+1)}_{kj:}$ is generated from the following distribution

$$q\left(\ln\boldsymbol{\lambda}^{*(\tau+1)}_{kj:}\middle|\ln\boldsymbol{\lambda}^{*(\tau)}_{kj:}\right) = \mathcal{N}\left(\ln\boldsymbol{\lambda}^{*(\tau)}_{kj:}, (d_0+d_2)\mathbf{I}_{\mathbf{T}}\right) \tag{15}$$

  where $\mathbf{I}_{\mathbf{T}}$ is the $T\times T$ identity matrix, and T denotes the number of time points. The acceptance probability for the proposed $\boldsymbol{\lambda}^{*(\tau+1)}_{kj:}$ is $\min\left(1,\alpha\left(\boldsymbol{\lambda}^{*(\tau+1)}_{kj:},\boldsymbol{\lambda}^{*(\tau)}_{kj:}\right)\right)$, where

$$\alpha\left(\boldsymbol{\lambda}_{kj:}^{*(\tau+1)},\boldsymbol{\lambda}_{kj:}^{*(\tau)}\right)=\exp\left(-\frac{1}{2}(\boldsymbol{\lambda}_{kj:}^{*(\tau+1)})^{T}\boldsymbol{\Gamma}_{kj}^{-1}\boldsymbol{\lambda}_{kj:}^{*(\tau+1)}+\frac{1}{2}\boldsymbol{\lambda}_{kj:}^{*(\tau)T}\boldsymbol{\Gamma}_{kj}^{-1}\boldsymbol{\lambda}_{kj:}^{*(\tau)}\right)\cdot$$
$$\prod_{t=1}^{T}\left[\left(\frac{\lambda_{kjt}^{*(t+1)}}{\lambda_{kjt}^{*(t)}}\right)^{\sum_{i=1}^{M}w_k(\boldsymbol{s}_{it})v_{ij1}-1}\exp\left[\sum_{i=1}^{M}w_k(\boldsymbol{s}_{it})\left(\lambda_{kjt}^{*(\tau+1)}-\lambda_{kjt}^{*(\tau)}\right)\right]\right]. \quad (16)$$

- Sample $\boldsymbol{\beta}_{k:i}$ from their respective posteriors conditional on $\{Z_k(\boldsymbol{s}_{it})\}$:

$$p(\boldsymbol{B}_k|-)\propto\prod_{t=1}^{T}\prod_{i=1}^{M}p(Z_k(\boldsymbol{s}_{it})|B_k)\prod_{j=1}^{J}\mathcal{N}(\beta_{kj:}|\mathbf{0},\Sigma_{kj}). \quad (17)$$

Reorder the entries of $\boldsymbol{B}_k$ (and the associated Ω_k) in (8) such that $\boldsymbol{B}_k=[\boldsymbol{\beta}_{k:1},\cdots,\boldsymbol{\beta}_{k:T}]^T$, then we obtain

$$\begin{aligned}p(\boldsymbol{B}_k|-) \quad\propto\quad & \exp\left\{-\sum_{t=1}^{T}\sum_{i=1}^{M}f(\eta_{kit})\boldsymbol{\beta}_{k:t}^{T}\varphi_{kit}\varphi_{kit}^{T}\boldsymbol{\beta}_{k:t}\right\}\\ & \times\exp\left\{-\frac{1}{2}\boldsymbol{B}_k^{T}\Omega_K^{-1}\boldsymbol{B}_k+\sum_{t=1}^{T}\sum_{i=1}^{M}(2Z_k(\boldsymbol{s}_{it})-1)\varphi_{kit}^{T}\boldsymbol{\beta}_{k:t}\right\}. \quad (18)\end{aligned}$$

So, $\boldsymbol{B}_k$ can be drawn from a normal distribution as

$$p(\boldsymbol{B}_k|-)=\mathcal{N}\left(\boldsymbol{B}_k;\left(\boldsymbol{\Omega}_k^{-1}+\boldsymbol{U}_k\right)^{-1}\boldsymbol{Y}_k,\left(\boldsymbol{\Omega}_k^{-1}+\boldsymbol{U}_k\right)^{-1}\right), \quad (19)$$

where $\boldsymbol{U}_k$ is a $(J+1)T\times(J+1)T$ block-diagonal matrix with the t-th $(J+1)\times(J+1)$ block expressed as $\boldsymbol{u}_{kt}=2\sum_{i=1}^{M}f(\eta_{kit})\boldsymbol{\phi}_{kit}\boldsymbol{\phi}_{kit}^{T}$ and $\boldsymbol{Y}_k$ is a $(J+1)T\times1$ vector formed by concatenating the T vectors $\boldsymbol{y}_{kt}=\sum_{i=1}^{M}\left(Z_k(\boldsymbol{s}_{it})-\frac{1}{2}\right)\boldsymbol{\phi}_{kit}$, $t=1,\cdots,T$. In these expressions $\boldsymbol{\phi}_{kit}=[1,\mathcal{K}(\boldsymbol{s}_{it},\tilde{\boldsymbol{s}}_1;\psi_k),\cdots,\mathcal{K}(\boldsymbol{s}_{it},\tilde{\boldsymbol{s}}_J;\psi_k)]^T$. The parameter $f(\eta_{kit})=\varphi_{kit}^{T}\boldsymbol{\beta}_{k:t}$.

- Sample $Z_k(\boldsymbol{s}_{it})$ from their respective posteriors conditional on $\boldsymbol{B}_k$ and $\{\nu_{ijt}\}$. According to the definition of LSBP,

$$\begin{aligned}&p(Z_k(\boldsymbol{s}_{it})=1|-)\\ =&\begin{cases}\frac{\sigma(g_k(\boldsymbol{s}_{it}))p(\nu_{it}|\boldsymbol{\lambda}_{kt}^{*})}{\sigma(g_k(\boldsymbol{s}_{it}))p\left(v_{it}|\boldsymbol{\lambda}_{kt}^{*}\right)+\sigma(-g_k(\boldsymbol{s}_{it}))p\left(\nu_{it}|\boldsymbol{\lambda}_{k't}^{*}\right)}, & \text{if } Z_l(\boldsymbol{s}_{it})=0 \text{ for } l<k\\ \sigma(g_k(\boldsymbol{s}_{it})), & \text{if } \exists\ l<k, \text{ such that } Z_l(\boldsymbol{s}_{it})=1\end{cases}\end{aligned} \quad (20)$$

where k' is the first integer larger than k, associated with non-zero indicator. The equation can be expressed as

$$p\left(Z_k\left(\boldsymbol{s}_{it}\right)=1|-\right)=\frac{1}{1+\exp\left(-\rho_{kit}\right)}, \tag{21}$$

with

$$\begin{aligned}\rho_{kit} &= \prod_{l<k}\left(1-Z_l\left(\boldsymbol{s}_{it}\right)\right)\log p\left(\nu_{it}|\,\boldsymbol{\lambda}^*_{kt}\right)- \\ &\quad \sum_{k'>k} Z_l\left(\boldsymbol{s}_{it}\right)\prod_{\substack{l<k'\\ l\neq k}}\left(1-Z_l\left(\boldsymbol{s}_{it}\right)\right)\log p\left(\left(\nu_{it}|\,\boldsymbol{\lambda}^*_{k't}\right)\right)+\varphi^T_{kit}\boldsymbol{\beta}_{k:t}. \end{aligned}\tag{22}$$

- With a uniform prior assumed on the kernel parameter library (a predefined finite set), the posterior distribution for each ψ_k can be represented as

$$p(\psi_k=\psi^*_l)\propto\prod_{t=1}^{T}\prod_{i=1}^{M}\sigma(g^l_k(\boldsymbol{s}_{it}))^{w_k(\boldsymbol{s}_{it})}\prod_{t=1}^{T}\prod_{i=1}^{M}\prod_{k'>k}\left(1-\sigma(g^l_k(\boldsymbol{s}_{it}))\right)^{w_{k'}(\boldsymbol{s}_{it})}. \tag{23}$$

For each specific k from $k=1,...,K$, we have the following update equation

$$\psi_k=\psi^*_{r_k}, r_k\sim\text{Mult}\left(p_{k1},...,p_{kL}\right), p_{kj}=\frac{p(\psi_k=\psi^*_j)}{\sum_{l=1}^{L}p(\psi_k=\psi^*_l)}. \tag{24}$$

We sample the kernel parameters based on the multinomial distributions from a given discrete set in each MCMC iteration.

- Sample c_0 from its posteriors conditional on $\{\boldsymbol{B}_k\}$ and $\{a_0,b_0\}$:

$$p(c_0)\propto\text{Gamma}\left(c_0;a_0,b_0\right)\prod_{k=1}^{K}\mathcal{N}\left(\boldsymbol{B}_k;\boldsymbol{0},\boldsymbol{\Omega}_k\right). \tag{25}$$

Therefore, c_0 can be drawn from a Gamma distribution

$$p(c_0)=\text{Gamma}\left(c_0;\tilde{a}_0,\tilde{b}_0\right), \tag{26}$$

where $\tilde{a}_0=a_0+0.5KT(J+1)$ and $\tilde{b}_0=b_0+0.5\sum_{k=1}^{K}\sum_{j=0}^{J}\boldsymbol{\beta}^T_{kj:}\tilde{\Sigma}^{-1}_{kj}\boldsymbol{\beta}_{kj:}$ with $[\tilde{\boldsymbol{\Sigma}}_{kj}]_{il}=c_1{}^{|t_i-t_l|}$.

- Sample c_1 from its posterior conditional on $\{\boldsymbol{B}_k\}$:

$$p(c_1)\propto\mathcal{N}_{(0,1)}(c_1;0,1)\prod_{k=1}^{K}\mathcal{N}\left(\boldsymbol{B}_k;\boldsymbol{0},\boldsymbol{\Omega}_k\right). \tag{27}$$

When updating c_1, the proposed $c_1^{(\tau+1)}$ is generated from the following distribution:

$$q\left(c_1^{(\tau+1)}|c_1^{\tau}\right) = \mathcal{N}_{(0,1)}\left(c_1^{(\tau+1)}; c_1^{\tau}, 1\right). \tag{28}$$

The acceptance probability for the proposed $c_1^{(\tau+1)}$ is $\min\left(1, \alpha(c_1^{(\tau+1)}, c_1^{\tau})\right)$, where

$$\alpha(c_1^{(\tau+1)}, c_1^{\tau}) = \frac{|\boldsymbol{\Sigma}_{kj}^{-1}(c_1^{\tau})|^{\frac{K(J+1)}{2}}}{|\boldsymbol{\Sigma}_{kj}^{-1}(c_1^{(\tau+1)})|^{\frac{K(J+1)}{2}}} \exp\left\{\frac{1}{2}\left(c_1^{(\tau+1)^2} - c_1^{(\tau+1)^2}\right)\right\}$$
$$\times \quad \exp\left\{\frac{1}{2}\left(\sum_{k=1}^{K}\sum_{j=0}^{J}\boldsymbol{\beta}_{kj:}^{T}\boldsymbol{\Sigma}_{kj}^{-1}(c_1^{\tau})\boldsymbol{\beta}_{kj:} - \sum_{k=1}^{K}\sum_{j=0}^{J}\boldsymbol{\beta}_{kj:}^{T}\boldsymbol{\Sigma}_{kj}^{-1}(c_1^{(\tau+1)})\boldsymbol{\beta}_{kj:}\right)\right\} \tag{29}$$

- Similarly, d_0 can be drawn from a Gamma distribution

$$p(d_0) = \text{Gamma}\left(d_0; \tilde{a}_0, \tilde{b}_0\right), \tag{30}$$

where $\tilde{a}_0 = a_0 + 0.5dKT$ and $\tilde{b}_0 = b_0 + 0.5\sum_{k=1}^{K}\sum_{j=1}^{d}\ln\boldsymbol{\lambda}_{kj:}^{*T}\tilde{\Gamma}_{kj}^{-1}\ln\boldsymbol{\lambda}_{kj:}^{*}$ with $[\tilde{\boldsymbol{\Gamma}}_{kj}]_{il} = d_1^{|t_i - t_l|}$.

- Similar to c_1, we update d_1 by the Metropolis-Hastings algorithm. The proposed $d_1^{(\tau+1)}$ is generated from the following distribution:

$$q\left(d_1^{(\tau+1)}|d_1^{\tau}\right) = \mathcal{N}_{(0,1)}\left(d_1^{(\tau+1)}; d_1^{\tau}, 1\right). \tag{31}$$

The acceptance probability for the proposed $d_1^{(\tau+1)}$ is $\min\left(1, \alpha(d_1^{(\tau+1)}, d_1^{\tau})\right)$, where

$$\alpha(d_1^{(\tau+1)}, d_1^{\tau}) = \frac{|\boldsymbol{\Gamma}_{kj}^{-1}(c_1^{\tau})|^{\frac{dK}{2}}}{|\boldsymbol{\Gamma}_{kj}^{-1}(c_1^{(\tau+1)})|^{\frac{dK}{2}}} \exp\left\{\frac{1}{2}\left(d_1^{(\tau+1)^2} - d_1^{(\tau+1)^2}\right)\right\}$$
$$\times\exp\left\{\frac{1}{2}\left(\sum_{k=1}^{K}\sum_{j=1}^{d}\ln\boldsymbol{\lambda}_{kj:}^{*T}\boldsymbol{\Gamma}_{kj}^{-1}(d_1^{\tau})\ln\boldsymbol{\lambda}_{kj:} - \sum_{k=1}^{K}\sum_{j=1}^{d}\ln\boldsymbol{\lambda}_{kj:}^{*T}\boldsymbol{\Sigma}_{kj}^{-1}(d_1^{(\tau+1)})\ln\boldsymbol{\lambda}_{kj:}\right)\right\}. \tag{32}$$

A2. VB inference

The log-normal priors placed on the Poisson intensities introduce non-conjugacy, which results in difficulty for VB inference. Therefore, we employ a point estimate for the Poisson intensities, by maximizing the lower bound $\mathcal{F}$. For the GP hyperparameters c_1 and d_1, the truncated normal prior also introduces non-conjugacy. Their posteriors are also inferred from point estimation by maximizing the VB lower bound. The update equations of the posterior inference of $\boldsymbol{\Theta}$ are summarized below. In our model,

$$\boldsymbol{\Theta} = \{\{\boldsymbol{\lambda}_{kj:}^{*}\}_{\substack{j=1,\ldots,d,\\ k=1,\ldots,K}}, \{\boldsymbol{B}_k\}_{k=1,\ldots,K}, \{Z_k(\boldsymbol{s}_{i,t})\}_{\substack{t=1,\ldots,T,\\ i=1,\ldots M,\\ k=1,\ldots,K}}, c_0, c_1, d_0, d_1\}.$$

- The lower bound for the Poisson intensity $\boldsymbol{\lambda}^*_{kj:}$ may be derived as

$$\mathcal{F}(\boldsymbol{\lambda}^*_{kj:}) \propto -\frac{1}{2}\boldsymbol{\Lambda}^T_{k,j}\boldsymbol{\Gamma}^{-1}_{kj}\boldsymbol{\Lambda}_{kj} - \boldsymbol{Q}^T_{kj}e^{\boldsymbol{\Lambda}_{kj}} + \boldsymbol{R}^T_{kj}\boldsymbol{\Lambda}_{kj} + \text{constant} \tag{33}$$

where $\boldsymbol{\Lambda}_{kj} = \log(\boldsymbol{\lambda}^*_{kj:})$, $\boldsymbol{R}_{kj} = [\sum_{i=1}^{M_1}\langle w_k(\boldsymbol{s}_{i1})\rangle\nu_{ij1} - 1, \cdots, \sum_{i=1}^{M}\langle w_k(\boldsymbol{s}_{iT})\rangle\nu_{ijT} - 1]^T$, and $\boldsymbol{Q}_{kj} = [\sum_{i=1}^{M_1}\langle w_k(\boldsymbol{s}_{i1})\rangle, \cdots, \sum_{i=1}^{M}\langle w_k(\boldsymbol{s}_{iT})\rangle]^T$, with $\langle\cdot\rangle$ denoting the expectation such that $\langle w_k(\boldsymbol{s}_{it})\rangle = q(w_k(\boldsymbol{s}_{it}) = 1)$ (see Section 2 for detail of $w_k(\boldsymbol{s}_{it})$). The point estimate for $\boldsymbol{\lambda}^*_{kj:}$ can be updated at each VB iteration by maximizing the lower bound $\mathcal{F}(\boldsymbol{\lambda}^*_{kj:})$. One may easily examine that $\mathcal{F}(\boldsymbol{\lambda}^*_{kj:})$ is a concave function, and therefore a global maximum can be obtained by any appropriate convex optimization method. Note that if $\boldsymbol{\Gamma}^{-1}_{kj} \to 0$ (setting large variance for the prior distribution), by taking the derivative of (33) and setting it to zero, we have $\boldsymbol{\lambda}^*_{kj:} = e^{\boldsymbol{\Lambda}_{kj}} \to \boldsymbol{R}_{kj}/\boldsymbol{Q}_{kj}$, which is consistent with the update equation if independent gamma priors are placed on λ^*_{kjt} for $t = 1, ..., T$. Therefore, the GP priors represented in $\boldsymbol{\Gamma}_{kj}$ introduce the correlation among the components of $\boldsymbol{\lambda}^*_{kj:}$.

- To update the variational distribution for the kernel weights β_{kjt}, note that the logistic link function $\sigma(\cdot)$ is not within the exponential family and therefore introduces the nonconjugacy. We here follow Jaakkola and Jordan (1998) by introducing a variational bound using the inequality

$$\sigma(y)^z[1-\sigma(y)]^{1-z} = \sigma(x) \geq \sigma(\eta)\exp(\frac{x-\eta}{2} - f(\eta)(x^2-\eta^2))$$

where $x = (2z-1)y$, $f(\eta) = \frac{\tanh(\eta/2)}{4\eta}$, and η is a variational parameter. An exact bound is achieved as $\eta = \pm x$.

If we reorder the entries of $\boldsymbol{B}_k$ (and the associated $\boldsymbol{\Omega}_k$) in (8) such that $\boldsymbol{B}_k = [\boldsymbol{\beta}_{k:1}, ..., \boldsymbol{\beta}_{k:T}]^T$, the update equation for $\boldsymbol{B}_k$ can be expressed as

$$q(\boldsymbol{B}_k) = \mathcal{N}\left((\boldsymbol{\Omega}^{-1}_k + \boldsymbol{U}_k)^{-1}\boldsymbol{Y}_k, (\boldsymbol{\Omega}^{-1}_k + \boldsymbol{U}_k)^{-1}\right) \tag{34}$$

where $\boldsymbol{U}_k$ is a $(J+1)T \times (J+1)T$ block-diagonal matrix with the tth $(J+1)\times(J+1)$ block expressed as

$$\boldsymbol{u}_{kt} = 2\sum_{i=1}^{M} f(\eta_{kit})\boldsymbol{\phi}_{kit}\boldsymbol{\phi}^T_{kit}$$

and $\boldsymbol{Y}_k$ is a $(J+1)T \times 1$ vector formed by concatenating the T vectors

$$\boldsymbol{y}_{kt} = \sum_{i=1}^{M}\left(\langle Z_k(\boldsymbol{s}_{it})\rangle - \frac{1}{2}\right)\boldsymbol{\phi}_{kit}, \qquad t = 1, ..., T.$$

In the above expressions $\boldsymbol{\phi}_{kit} = [1, \mathcal{K}(\boldsymbol{s}_{it}, \tilde{\boldsymbol{s}}_1; \psi_k), ..., \mathcal{K}(\boldsymbol{s}_{it}, \tilde{\boldsymbol{s}}_J; \psi_k)]^T$.

The variational parameters η_{kit} are then updated as

$$\eta^2_{kit} = \boldsymbol{\phi}^T_{kit}\langle\boldsymbol{\beta}^T_{k:t}\boldsymbol{\beta}_{k:t}\rangle\boldsymbol{\phi}_{kit} \tag{35}$$

where $\langle\boldsymbol{\beta}^T_{k:t}\boldsymbol{\beta}_{k:t}\rangle = COV(\boldsymbol{\beta}_{k:t}, \boldsymbol{\beta}_{k:t}) + \langle\boldsymbol{\beta}_{k:t}\rangle\langle\boldsymbol{\beta}_{k:t}\rangle^T$ and it may be evaluated from $q(\boldsymbol{B}_k)$ with the mean and variance associated with time t.

- The variational distribution for the binary indicator $Z_k(\boldsymbol{s}_{it})$ may be updated as

$$q\left(Z_k(\boldsymbol{s}_{it})=1\right)=\frac{1}{1+\exp(-\rho_{kit})} \tag{36}$$

with

$$\begin{aligned}\rho_{kit} &= \prod_{l<k}(1-\langle Z_l(\boldsymbol{s}_{it})\rangle)\log p(\nu_{it}|\boldsymbol{\lambda}^*_{kt})\\ &- \sum_{k'>k}\langle Z_{k'}(\boldsymbol{s}_{it})\rangle\prod_{\substack{l<k'\\ l\neq k}}(1-\langle Z_l(\boldsymbol{s}_{it})\rangle)\log p(\nu_{it}|\boldsymbol{\lambda}^*_{k't})\\ &+ \sum_{j=1}^{J}\langle\beta_{kjt}\rangle\mathcal{K}(\boldsymbol{s}_{it},\tilde{\boldsymbol{s}}_j;\psi_k)+\langle\beta_{k0t}\rangle\end{aligned}$$

where $\log p(\nu_{it}|\boldsymbol{\lambda}^*_{kt})$ is the data log-likelihood from the Poisson distribution such that $\log p(\boldsymbol{v}_{it}|\boldsymbol{\lambda}^*_{kt})=\log\left(\prod_{j=1}^{d}\text{Poisson}(\nu_{ijt}|\lambda^*_{kjt})\right)$, and the expectation $\langle\beta_{kjt}\rangle$ can be obtained from $q(\boldsymbol{B}_k)$.

- Due to the non-conjugacy of the sigmoid function, we cannot acquire a variational distribution for ψ_k. However, we can sample it from its posterior distribution by establishing a discrete set of potential kernel widths $\{\psi^*_l\}_{l=1,\cdots,L}$. The posterior distribution for each ψ_k is represented as

$$\begin{aligned}p(\psi_k=\psi^*_l) \propto\ & \exp\{\sum_{t=1}^{T}\sum_{i=1}^{M}\langle w_k(\boldsymbol{s}_{it})\rangle\langle\log\sigma\left(g^l_k(\boldsymbol{s}_{it})\right)\rangle\}\\ &\times\exp\{\sum_{t=1}^{T}\sum_{i=1}^{M}\sum_{k'>k}\langle w_{k'}(\boldsymbol{s}_{it})\rangle\langle\log\left(1-\sigma\left(g^l_k(\boldsymbol{s}_{it})\right)\right)\rangle\},\end{aligned} \tag{37}$$

where $g^l_k(\boldsymbol{s}_{it})=\sum_{j=1}^{J}\beta_{kjt}\mathcal{K}(\boldsymbol{s}_{it},\tilde{\boldsymbol{s}}_j;\psi^*_l)+\beta_{k0t}$. The detailed calculations of $\langle\log\sigma\left(g^l_k(\boldsymbol{s}_{it})\right)\rangle$ and $\langle\log\left(1-\sigma\left(g^l_k(\boldsymbol{s}_{it})\right)\right)\rangle$ can be found in Ren et al. (2011).

- The variational distribution for c_0 may be updated as

$$q(c_0)=\text{Gamma}\left(c_0;\tilde{a}_0,\tilde{b}_0\right), \tag{38}$$

with $\tilde{a}_0=a_0+0.5KT(J+1)$ and $\tilde{b}_0=b_0+0.5\sum_{k=1}^{K}\sum_{j=0}^{J}\sum_{i=1}^{T}\sum_{l=1}^{T}[\tilde{\Sigma}^{-1}_{kj}]_{il}\langle\boldsymbol{\beta}_{kji}\boldsymbol{\beta}_{kjl}\rangle$ with $[\tilde{\boldsymbol{\Sigma}}_{kj}]_{il}=c_1{}^{|t_i-t_l|}$.

- The VB lower bound for c_1 may be derived as

$$\mathcal{F}(c_1)=\log\mathcal{N}_{(0,1)}(c_1;0,1)+\sum_{k=1}^{K}\log\mathcal{N}\left(\boldsymbol{B}_k;\boldsymbol{0},\boldsymbol{\Omega}_k\right)+\text{constant}. \tag{39}$$

The point estimate for c_1 can be updated at each VB iteration by maximizing the lower bound $\mathcal{F}(c_1)$.

- Since a point estimate of $\boldsymbol{\lambda}_{kj:}$ is employed at each VB iteration, the variational distribution for d_0 may be the same as (30)

$$q(d_0) = \text{Gamma}\left(d_0; \tilde{a}_0, \tilde{b}_0\right), \tag{40}$$

  where $\tilde{a}_0 = a_0 + 0.5dKT$ and $\tilde{b}_0 = b_0 + 0.5 \sum_{k=1}^{K} \sum_{j=1}^{d} \ln \boldsymbol{\lambda}_{kj:}^{*T} \tilde{\boldsymbol{\Gamma}}_{kj}^{-1} \ln \boldsymbol{\lambda}_{kj:}^{*}$.

- Similarly, the lower bound for d_1 is

$$\mathcal{F}(d_1) = \log \mathcal{N}_{(0,1)}(d_1; 0, 1) + \sum_{k=1}^{K} \sum_{j=1}^{d} \log \mathcal{N}\left(\boldsymbol{\Lambda}_{kj}; \mathbf{0}, \boldsymbol{\Gamma}_{kj}\right) + \text{constant}. \tag{41}$$

  and the point estimation for d_1 is obtained by maximizing $\mathcal{F}(d_1)$.

By following (33)-(41), the model parameters and GP hyperparameters can be updated iteratively until convergence. In our experiment, we observed fast convergence; typically the relative change of the lower bound reduces to 10^{-4} within 100 iterations.

References

Achcar, J. A., Rodrigues, E. R., and Tzintzun, G. (2011). "Using non-homogeneous Poisson models with multiple change-points to estimate the number of ozone exceedances in Mexico City." *Environmetrics*, 22: 1–12. 816

Adams, R. P., Murray, I., and MacKay, D. (2009). "Tractable nonparametric Bayesian Inference in Poisson processes with Gaussian process intensities." In *International Conference on Machine Learning*, 9–16. Montreal, Quebec. 815

Beal, M. J. (2003). "Variational algorithms for approximate Bayesian inference." Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London. 818, 819

Chakraborty, A. and Gelfand, A. E. (2010). "Analyzing spatial point patterns subject to measurement error." *Bayesian Analysis*, 5: 97–122. 816

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. Arnold, 2 edition. 815

Diggle, P. J., Menezes, R., and Su, T. (2010). "Geostatistical inference under preferential sampling (with discussion)." *Journal of the Royal Statistical Society - Series C*, 59: 191–232. 816

Heikkinen, J. and Arjas, E. (1998). "Non-parametric Bayesian estimation of a spatial Poisson Intensity." *Scandinavian Journal of Statistics*, 25: 435–450. 816

Hoffman, M., Blei, D., and Bach, F. (2010). "Online learning for latent Dirichlet allocation." In *Advances in Neural Information Processing Systems*, 993–1022. Vancouver, Canada. 832

Hossain, M. M. and Lawson, A. B. (2009). "Approximate methods in Bayesian point process spatial models." *Computational Statistics and Data Analysis*, 53: 2831–2842. 816

Jaakkola, T. and Jordan, M. I. (1998). "Bayesian parameter estimation through variational methods." *Statistics and Computing*, 10: 25–37. 836

Ji, C., Merl, D., and Kepler, T. B. (2009). "Spatial mixture modeling for unobserved point processes: Examples in immunofluorescence histology." *Bayesian Analysis*, 4: 297–315. 816

Kottas, A. and 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. 816

Luttinen, J. and Ilin, A. (2009). "Variational Gaussian-process factor analysis for modeling spatio-temporal data." In *Advances in Neural Information Processing Systems*, 1177–1185. Vancouver, Canada. 816

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). "Log Gaussian Cox process." *Scandinavian Journal of Statistics*, 25: 451–482. 815

Møller, J. and Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC. 815

Pati, D., Reich, B. J., and Dunson, D. B. (2010). "Bayesian geostatistical modeling with informative sampling locations." *Biometrika*, 98: 35–48. 816

Rasmussen, C. E. and Willams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press. 815, 817, 824

Rathbun, S. L. and Cressie, N. (1994). "Asymptotic propertes of estimators for the parameters of spatial inhomogeous Poisson point processes." *Advances in Applied Probability*, 26: 122–154. 815

Ren, L., Du, L., Carin, L., and Dunson, D. B. (2011). "Logistic stick-breaking process." *Journal of Machine Learning Research*, 12: 203–239. 816, 818, 837

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4: 639–650. 817

Taddy, M. (2010). "Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime." *Journal of the American Statistical Association*, 105: 1403–1417. 816, 824, 825, 830

Taddy, M. and Kottas, A. (2012). "Mixture modeling for marked Poisson processes." *Bayesian Analysis*, 7: 335–362. 816

Taddy, M. A. (2008). "Bayesian nonparametric analysis of conditional distributions and inference for Poisson point processes." Ph.D. thesis, Statistics and Stochastic Modeling, University of California, Santa Cruz. 816, 824, 825

Tipping, M. E. (2001). "Sparse Bayesian learning and the relevance vector machine." *Journal of Machine Learning Research*, 1: 211–244. 820

Turner, R., Møller, J., and Hazelton, M. (2005). "Residual analysis for spatial point processes (with discussion)." *Journal of the Royal Statistical Society - Series B*, 67: 617–666. 830

Wolpert, R. and Ickstadt, K. (1998). "Poisson/Gamma random field models for spatial statistics." *Biometrika*, 85: 251–267. 816

Acknowledgments

The authors wish to thank the reviewers and editors for their comments, which have substantially improved the paper. The research reported here was supported by the Army Research Office (Dr. Liyi Dai) and the Office of Naval Research (Dr. Wen Masters).