

Autoregressive Mixture Models for Dynamic Spatial Poisson Processes: Application to Tracking Intensity of Violent Crime

M. A. Taddy

The University of Chicago Booth School of Business

This article develops a set of tools for smoothing and prediction with dependent point event patterns. The methodology is motivated by the problem of tracking weekly maps of violent crime events, but is designed to be straightforward to adapt to a wide variety of alternative settings. In particular, a Bayesian semiparametric framework is introduced for modeling correlated time series of marked spatial Poisson processes. The likelihood is factored into two independent components: the set of total integrated intensities and a series of process densities. For the former it is assumed that Poisson intensities are realizations from a dynamic linear model. In the latter case, a novel class of dependent stick-breaking mixture models are proposed to allow nonparametric density estimates to evolve in discrete time. This, a simple and flexible new model for dependent random distributions, is based on autoregressive time series of marginally beta random variables applied as correlated stick-breaking proportions. The approach allows for marginal Dirichlet process priors at each time and adds only a single new correlation term to the static model specification. Sequential Monte Carlo algorithms are described for on-line inference with each model component, and marginal likelihood calculations form the basis for inference about parameters governing temporal dynamics. Simulated examples are provided to illustrate the methodology, and we close with results for the motivating application of tracking violent crime in Cincinnati.

M. A. Taddy is Assistant Professor of Econometrics and Statistics and Robert L. Graves Faculty Fellow at the University of Chicago Booth School of Business, 5807 S Woodlawn Ave, Chicago, IL 60637 (email: taddy@chicagobooth.edu) and is partially supported by the IBM Corporation Faculty Research Fund at Chicago Booth. This article contains research from the author's Ph.D. dissertation at University of California, Santa Cruz, and the author thanks his thesis adviser, Athanasios Kottas, for much guidance and discussion related to the methodology herein. The material is also related to work under Lawrence Livermore National Laboratory subcontract B565602; Herbie Lee, Bruno Sansó, and Luis Acevedo-Arreguin all collaborated on the original data analysis project, and are to thank for helpful insight and assistance with the data. Finally, two anonymous referees provided advice which greatly improved the presentation.

1 Introduction and Data

This article describes flexible Bayesian inference for a discrete series of marked spatial point patterns. The motivating data set consists of a year-long weekly record of urban violent crime events, mapped by location and classified by degree of violence. Analysis of this data will be designed to, in as simple a fashion as possible, provide Bayesian nonparametric inference for the weekly spatial intensity of extremely violent crime, with estimated intensity surfaces that are correlated with event patterns from previous weeks and for less violent crime.

We begin by highlighting the application: Section 1.1 details the crime data and tracking problem, including specifics on data collection, data processing, and event categorization, and outlines our general motivation and modeling goals. Section 2 then details modeling methodology. 2.1 reviews the approach of Taddy and Kottas (2009) and specifies a Dirichlet process mixture model for static marked crime event data (i.e., the marginal model for each weekly point pattern) which factors Poisson intensity into the product of integrated intensity and normalized process density. Section 2.2 provides an outline for extending these models to a dynamic setting, including description of a dynamic linear model for Poisson random variables and discussion of some general approaches to nonparametric inference for series of correlated distributions. Finally, 2.3 details our proposed model for dynamic point process densities – a novel version of the dependent Dirichlet process that is especially well suited to the application of interest. This final model element, which is termed the BAR stick-breaking process and involves a simple scheme for autoregressive beta random variables in correlated stick-breaking, is completely new and could have wider applicability. The model is illustrated through a short simulation study.

Posterior inference is based on a sequential Monte Carlo approach described in Section 3.1 for the Poisson intensity series and in 3.2 for our nonparametric dynamic density models. Additional implementation specific details are in the Appendix, and Section 3 closes in 3.3 with a simple procedure for sequential marginal likelihood estimation. Application and inference for the motivating dataset are presented in Section 4: 4.1 applies to crime counts a first-order version of the dynamic linear model

from Section 2.2, 4.2 contains results for an analysis of crime event densities based on the dynamic density model of Section 2.3, and 4.3 describes combined results for both density and integrated intensity analysis. Finally, general contributions of the framework introduced in this article will be discussed and summarized in Section 5.

1.1 Mapping Crime in Cincinnati

Tracking violent crime is a conceptually simple data analysis problem of considerable public interest that is, however, commonly associated with controversy and re-analysis. The notorious difficulty of crime tracking is due partially to the politicized nature of the data, but comes also from inherent problems in specifying appropriate levels of spatial and temporal aggregation. At the same time, the data-intensive nature of modern law enforcement and a growing availability of on-line crime maps (e.g., www.crimereports.com) have created an abundance of spatio-temporal crime data. As such, the field should benefit from development of relatively automatic tools for dimension reduction and inference.

The source for our motivating application is an on-line repository of crime reports from Cincinnati, OH, at <http://www.cincinnati-oh.gov/police/pages/-4258-/>. The complete database reports date, time, and location of all arrests in Hamilton county (which includes Cincinnati), as well as other data that might be useful to characterize a reported event. We consider violent crime arrests that occurred during 2006 within a rectangular observation window that includes most of greater Cincinnati.

Each arrest is assigned one of the more than 170 FBI Uniform Crime Reporting codes that describe type of crime, with labels ranging from telephone harassment to murder. This variable was used to reclassify and group the arrests into two levels of crimes against people (excluding other types of crime): that with extreme violence (e.g., murder, rape, or aggravated assault), and that without extreme violence (e.g., non-aggravated assault, mugging). Although this categorization is somewhat arbitrary (there could easily be more refined levels of violence), it served our need to have a relatively more prevalent type of event inform inference about the less common events (less than a third were classified as extreme) and there is support among criminologists

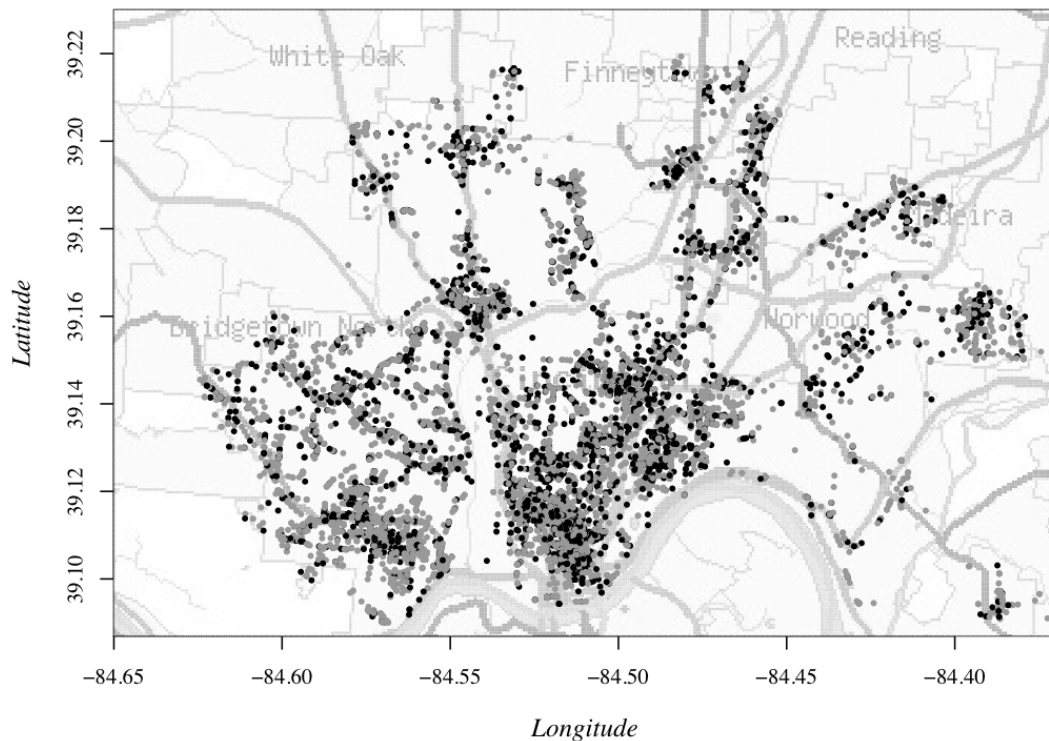


Figure 1: Cincinnati crime data and observation window: each violent crime arrest in 2006 is plotted, with the darker circles indicating extreme violence.

for dependence between different levels of crime (e.g., Kelling and Coles, 1998).

Figure 1 shows the 12,431 violent crime arrests, 3,857 of which were extremely violent. The observation window – the area plotted in Figure 1 – is contained within $[-84.65, -84.37]$ degrees longitude and $[39.09, 39.23]$ degrees latitude. Although this rectangle includes a portion of Kentucky (south of the Ohio river) not covered by the Cincinnati crime database, it will be seen that the proposed model is able to assign this region posterior intensity of very nearly zero. Finally, crime event times were used to map each crime to sequential Sunday-Monday weeks – the highest fidelity that can be modeled without the introduction of structural model elements (e.g., night or weekend effects) that are beyond the intended scope of application.

A note on the motivating application: although the paper is focused on crime event data, crime tracking is actually a proxy for more general data problems. This work grows out of a project with Lawrence Livermore National Laboratory, and the crime data was presented as an application similar to other problems in which LLNL was interested. Since we would only ever be working with the crime data, the task was to

devise a state-of-the-art analysis that would be straightforward for others to generalize in alternate settings. Therefore, emphasis is placed on maintaining ease of application while providing a general model that can take advantage of dependencies in space, time, and the mark domain and is sufficiently flexible to capture a wide variety of point pattern behavior. The resulting inference framework can be applied in on-line tracking of generic dynamic marked Poisson process intensity surfaces.

2 Modeling Methodology

The following three subsections describe each aspect of the general modelling framework: a static Poisson process model in 2.1, a general model for dynamic intensities in 2.2, and the autoregressive mixture model for correlated densities in 2.3.

2.1 A Mixture Model for Marked Spatial Point Patterns

The static point patterns – each individual week’s crime event map – can be characterized as the realization of a spatial Poisson process with categorical marks. That is, for any given week, the set of n observed crime locations, $\{\mathbf{x}_i\}_{i=1}^n$, and associated crime types, $\{y_i\}_{i=1}^n$, are assumed to arise from a marked Poisson process defined on the observation window \mathcal{R} (shown in Figure 1) with intensity function $\lambda(\mathbf{x})$ that is non-negative and locally integrable for all bounded $\mathcal{B} \subseteq \mathcal{R}$, and with conditional mark probability function $f(y|\mathbf{x})$ for y in mark-space \mathcal{M} . Following from standard results concerning Poisson processes (see, e.g., Møller and Waagepetersen, 2004), this specification leads to the following assumed pattern characteristics:

- i.* For any $\mathcal{B} \subseteq \mathcal{R}$, the number of points in \mathcal{B} , denoted $n(\mathcal{B})$, is a Poisson random variable with mean $\int_{\mathcal{B}} \lambda(\mathbf{x}) d\mathbf{x}$. In particular, $n \sim \text{Po}(\Lambda)$ where $\text{Po}(\cdot)$ denotes the Poisson distribution and total integrated intensity is defined as $\Lambda = \int_{\mathcal{R}} \lambda(\mathbf{x}) d\mathbf{x}$.
- ii.* Conditional on $n(\mathcal{B})$, the point locations within any $\mathcal{B} \subseteq \mathcal{R}$ are independent and identically distributed with density $\lambda(\mathbf{x}) / \int_{\mathcal{B}} \lambda(\mathbf{x}) d\mathbf{x}$. Hence, for a given n , observed point locations $\{\mathbf{x}_i\}_{i=1}^n$ are treated as independent draws from \mathcal{R} with density function $f(\mathbf{x}) = \lambda(\mathbf{x}) / \Lambda$.

iii. Given the point locations, event marks $\{y_i\}_{i=1}^n$ are independently distributed with conditional density functions $f(y_i|\mathbf{x}_i)$, $i = 1, \dots, n$.

Assumptions *i* and *ii* yield a marginal data likelihood that factorizes as $p(\{\mathbf{x}_i\}_{i=1}^n | \Lambda, f) \propto \Lambda^n \exp(-\Lambda) \prod_{i=1}^n f(\mathbf{x}_i)$, and this serves to render a generally tractable data analysis problem. Indeed, it is for this practical reason that Poisson point processes form the basis of the majority of inference about spatial point patterns (see, e.g., Diggle, 2003). Similarly, the conditional mark independence of *iii* enables use of generic regression techniques in modeling $f(y|\mathbf{x})$ and, as will be discussed in more detail below, allows for re-expression of the model as a joint process over $\mathcal{R} \times \mathcal{M}$. It is easy to imagine scenarios where the crime events of Section 1.1 (or similar point patterns) do not conform completely to restrictions *i-iii*. However, the marked Poisson process provides a powerful first approximation to many more complicated pattern generating mechanisms, and it is often applied in difficult settings and as the foundation for more complex analysis schemes (see Best, Ickstadt, and Wolpert, 2000; Liang, Banerjee, and Carlin, 2009, for some recent Bayesian examples). Section 4.3 revisits the implications of these simplifying Poisson assumptions in the context of our crime tracking application.

Since *i* forces a $\text{Po}(n; \Lambda)$ total count likelihood, it remains only to specify the model for process density, $f(\mathbf{x})$, and conditional mark probability, $f(y|\mathbf{x})$. Our approach follows directly from the framework in Taddy and Kottas (2009), and relies on two main insights. First, the marked process can also be viewed as a joint Poisson process over $\mathcal{R} \times \mathcal{M}$ with intensity $\phi(\mathbf{x}, y) = \lambda(\mathbf{x})f(y | \mathbf{x}) = \Lambda f(\mathbf{x})f(y | \mathbf{x})$, such that we can define the marked process through model specification for $f(\mathbf{x}, y)$, the joint process density. Second, as described in Kottas and Sansó (2007), the likelihood factorization of *ii* enables general density estimation techniques for determining process intensity; in particular, $f(\mathbf{x}, y)$ can be assigned a Dirichlet Process (DP) mixture prior (Antoniak, 1974) and inferred under a Bayesian nonparametric paradigm. That is, one can specify $f(\mathbf{x}, y; G) = \int k(\mathbf{x}, y; \theta) dG(\theta)$, where G is assigned a DP prior and k is some parametric density kernel. This model is most easily understood through the constructive definition of the DP (e.g., Perman, Pitman, and Yor, 1992), according to

which $G \sim \text{DP}(\alpha, G_0(\psi))$ is almost surely of the form

$$dG(\theta) = \sum_{l=1}^{\infty} p_l \delta_{\vartheta_l}(\theta), \text{ where } \vartheta_l \sim G_0(\psi), \quad v_l \sim \text{be}(1, \alpha), \text{ and } p_l = v_l \left(1 - \sum_{i < l} p_i\right) \quad \forall l, \quad (1)$$

for joint process density $f(\mathbf{x}, y; G) = \sum_{l=1}^{\infty} p_l k(\mathbf{x}, y; \vartheta_l)$ and intensity function $\phi(\mathbf{x}, y; G) = \Lambda \sum_{l=1}^{\infty} p_l k(\mathbf{x}, y; \vartheta_l)$. A sampled prior intensity is thus based on probability weights realized through repeated $\text{be}(1, \alpha)$ portions of a remaining “probability stick”. Finally, the conditional mark density is obtained after dividing $f(\mathbf{x}, y; G)$ by marginal $f(\mathbf{x}; G)$ which, due to discreteness of G , should be easily available.

The weekly crime event patterns form a binary type process, such that $\mathcal{M} = \{0, 1\}$ where $y = 1$ indicates that violence was deemed extreme. Thus $f(\mathbf{x}, y)$ is a mixed continuous/discrete density, and the implied conditional $f(y|\mathbf{x})$ is a discrete probability function. In addition, it is assumed throughout that spatial coordinates are scaled so that $\mathcal{R} = (0, 1) \times (0, 1)$ and a logit transformation (i.e., $\text{logit}(x) = \log(x) - \log(1 - x)$) of each spatial dimension will map $\mathbf{x} \rightarrow \mathbb{R}^2$. Accordingly, the joint density will be modeled as a DP mixture with kernel that is the product of a Bernoulli density for y and a bivariate normal kernel function on logit transformations of the scaled \mathbf{x} locations. A primary feature of this specification is that it allows for conditional conjugacy – here, a base measure that is conjugate for the kernel parametrization – and Rao-Blackwellization of posterior inference by marginalizing over kernel parameters. Although this product kernel structure implies that dependence between \mathbf{x} and y is induced only through reweighting via G , the approach proves to be very flexible in practice; see, e.g., Taddy and Kottas (2009, 2010).

To summarize, the complete model for marginal weekly crime intensity holds that event pattern $\{\mathbf{x}_i, y_i\}_{i=1}^n$ is the realization of a Poisson process with intensity $\phi(\mathbf{x}, y) = \Lambda f(\mathbf{x}, y)$ on $(0, 1)^2 \times \{0, 1\}$, and the process density model is

$$f(\mathbf{x}, y; G) = \int \text{N}(\text{logit}(\mathbf{x}); \boldsymbol{\mu}, \boldsymbol{\Sigma}) \frac{q^y (1 - q)^{1-y}}{\mathbf{x}'(1 - \mathbf{x})} dG(\boldsymbol{\mu}, \boldsymbol{\Sigma}, q), \quad (2)$$

$$G \sim \text{DP}(\alpha, G_0), \quad dG_0 = \text{N}(\boldsymbol{\mu}; \boldsymbol{\gamma}, \boldsymbol{\Sigma}/\kappa) \text{W}(\boldsymbol{\Sigma}^{-1}; \nu, \boldsymbol{\Omega}) \text{be}(q; a_q, b_q),$$

where Wishart $W(\cdot)$ is such that $\mathbb{E}[\Sigma^{-1}] = \nu \mathbf{\Omega}^{-1}$ and $\text{be}(a, b)$ denotes a beta distribution with mean $a/(a+b)$. For G_0 , it will generally suffice to place a Wishart $W(\nu_{\Omega}, \Psi_{\Omega}^{-1})$ hyperprior on $\mathbf{\Omega}$, such that $\mathbb{E}[\mathbf{\Omega}] = \nu_{\Omega} \Psi_{\Omega}$, and fix the remaining parameters.

Taddy and Kottas (2009) describe much more detailed model properties, provide a full framework for inference about intensities and conditional mark densities, and outline related approaches from the literature, including the relationship between our mixture intensity model and the more common log-Gaussian Cox process model (Møller et al., 1998). Briefly, practical advantages to the framework used herein include an immediately available joint model for marks and spatially changing intensity surface variability (which requires non-stationary Gaussian process modeling under the standard approach). Taddy and Kottas also provide straightforward guidelines for analogous specification of static marked process density models in a wide variety of settings. For example, alternative DP base measures may better capture variability in Σ (e.g., Manolopoulou et al., 2009) and it is possible to incorporate any number of additional (continuous or categorical) process marks. Finally, the dynamic innovations of Sections 2.2 and 2.3 operate independent of the marginal process model, such that adaptations of the specification in this section will not require any additional methodology.

2.2 *Extension to Dynamic Point Processes*

Following from the intensity factorization of Section 2.1, a dynamic Poisson process model will be characterized by a series of correlated intensity functions, $\{\phi_t\}_{t=1}^T$, consisting of the integrated intensity series, $\Lambda^T = \{\Lambda_t\}_{t=1}^T$, multiplied by a series of process densities, $f^T = \{f_t\}_{t=1}^T$. This section will address general inference for both, with a specific autoregressive mixture density model for f^T following in Section 2.3. Particle learning algorithms for posterior simulation are detailed in Section 3.

In modeling Λ^T , the data of interest are the set of event counts – in this context, the weekly crime pattern sizes, $n^t = \{n_t\}_{t=1}^{52}$. These weekly counts are each distributed $n_t \sim \text{Po}(\Lambda_t)$ such that, in the language of state-space modeling, $\{\Lambda_t\}_{t=1}^{52}$ is the series of dynamic states corresponding to a Poisson observation equation. The model proposed here follows directly out of the conditionally Gaussian dynamic linear model (DLM)

framework, as proposed by Cargnoni, Müller, and West (1997). For Poisson distributed count data, this yields the Poisson DLM

$$\text{Observation Equation : } n_t | \Lambda_t \sim \text{Po}(n_t; \Lambda_t) \quad (3)$$

$$\text{Structural Equation : } \log(\Lambda_t) = \mathbf{F}'\eta_t + v_t, \quad v_t \sim \text{N}(0, \mathbf{V})$$

$$\text{State Equation : } \eta_t = \mathbf{G}\eta_{t-1} + \omega_t, \quad \omega_t \sim \text{N}(0, \mathbf{W}_t),$$

with initial information $\eta_0 \sim \text{N}(m_0, C_0)$, where η_t is the $d \times 1$ state vector, \mathbf{F} is the $d \times 1$ design vector, and \mathbf{G} is a $d \times d$ evolution matrix. A gamma prior is conditionally conjugate for \mathbf{V}^{-1} , and it is often convenient to reflect the (usually block diagonal) structure of \mathbf{G} by specifying \mathbf{W}_t through one or more discount factors, $\delta \in (0, 1]$, such that $\text{var}[\eta_t | \Lambda^{t-1}] = \text{var}[\mathbf{G}\eta_{t-1} | \Lambda^{t-1}]/\delta$. More generally, there is huge flexibility in model specification through changes to $\{\mathbf{F}, \mathbf{G}, \mathbf{V}, \mathbf{W}_t\}$, and any of these parameters may be made time-dependent to yield non-constant DLMs (refer to West and Harrison, 1997, for a complete treatment of model specification and design).

It remains to build a model for the time series of process densities, $f^T = \{f_t\}_{t=1}^T$. Recalling the specification in equation (2), each marginal process density can be written as $f(\mathbf{x}, y; G) = \int k(\mathbf{x}, y; \theta) dG(\theta)$, where k is the logit-normal and Bernoulli product kernel and $\theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, q\}$, while G is assigned a DP prior. Assuming a constant kernel function, f^T is fully specified by $G^T = \{G_t\}_{t=1}^T$, a series of random mixing distributions such that $f_t(\mathbf{x}, y; G_t) = \int k(\mathbf{x}, y; \theta) dG_t(\theta)$. Our task is thus to develop an extension of the DP prior for each G_t that allows correlation across a discrete time index.

The formulation in (1) places the DP within a general class of *stick-breaking* processes that recursively specify each point-mass as some random portion (the v_l) of remaining probability, and there is a large recent literature concerned with dependent extensions for these models. Much of the earlier work is derived from a report by MacEachern (2000), which introduced and outlined the Dependent Dirichlet Process (DDP). This general model formulation holds that $\mathbf{G} = \{G_{s_1}, \dots, G_{s_R}\}$, the set of random distributions corresponding to points $\{s_r\}_{r=1}^R \in \mathcal{S}$ and collectively distributed as a DDP ($p(V; \boldsymbol{\alpha}), G_0(\boldsymbol{\Theta}; \psi)$), are realized in the form $dG_s(\theta) = \sum_{l=1}^{\infty} p_{ls} \delta_{\vartheta_{ls}}(\theta)$, where $p_{ls} = v_{ls} (1 - \sum_{i < l} p_{is})$. In this, the proportions $V_l = [v_{ls_1}, \dots, v_{ls_R}] \in [0, 1]^R$ and loca-

tions $\Theta_l = [\vartheta_{ls_1}, \dots, \vartheta_{ls_R}]$ are, for $l = 1, \dots, \infty$, *iid* draws from $p(V; \boldsymbol{\alpha})$ and $G_0(\Theta; \psi)$ respectively, with each the finite dimensional distribution induced by a measurable stochastic process over \mathcal{S} . Dependence is only specified at the level of random distributions, rather than between observations, such that $p(\{\theta_{s_r}\}_{r=1}^R | \mathbf{G}) = \prod_{r=1}^R dG_{s_r}(\theta_{s_r})$ in the absence of further modeling constraints.

The BAR stick-breaking process that will be presented in Section 2.3 is a member of the class of single- θ DDP models for which realized locations are constant in \mathcal{S} and $dG_s(\theta) = \sum_{l=1}^{\infty} p_{ls} \delta_{\vartheta_l}(\theta)$. With a single- θ model, it is only ever necessary to specify dependencies across the one-dimensional p_{ls} parameters, offering a more easily generalizable scheme than those which depend upon stochastic process specification for θ_{ls} , the full set of kernel parameters. This provides a clear practical advantage even for the fairly simple kernel in (2). There is also an intuitive justification for the use of such processes in the temporal modeling of Poisson process densities: it is plausible that regions of higher intensity are caused by practically static factors (e.g., geography and history), and that dynamics are dominated by change in the relative weighting of intensity in these regions (due, e.g., to a reallocation of resources or intermigration). Furthermore, it is straightforward to adapt the results in MacEachern (2000), through construction of near independent random measures, to show that prior full support over the space of random distributions \mathbf{G} is not restricted by the constraint of θ_l that are held constant in time. As a final point, we note that there is nothing that inherently limits our modeling framework to single- θ dynamics. Indeed, it may be desirable to allow for dynamics in the mark distribution when extending this model to more complicated processes. The key point is that dependent stick-breaking induces basic temporal dynamics for the entire set of kernel parameters, allowing any additional required variation to be focused only on a subset of parameters of interest.

Before moving to development of the BAR process prior in 2.3, we note that there already exists a bulk of literature concerned with building dependent nonparametric priors through correlated stick-breaking processes. Two prominent lines of research are models related to the probit stick-breaking of Rodriguez and Dunson (2009), and the order based dependent DP in Griffin and Steel (2009). In the former case, a real-

valued stochastic process over \mathcal{S} is transformed through a probit link function to form distributions on a series of correlated probabilities, and in the latter case ordering of the stick-breaking procedure is regressed onto locations within \mathcal{S} . Although both probit and order-based stick-breaking models have much to recommend them and can be set in a more general space than the BAR model, the probit scheme does not allow for DP marginal distributions, and is thus disadvantaged by not affording the intuition associated with a very well understood model, and the order-based procedure requires a more complex machinery than is desired here.

A third modeling option, the generalized Pólya urn framework of Caron, Davy, and Doucet (2007), is perhaps closest in spirit to the approach proposed herein. These authors develop a model for time-varying DP mixtures that is based on an algorithmic description of the dynamics between posterior predictive distributions at each time step. The procedure focuses on latent allocation variables that assign each observation to a particular mixture component, and the prior predictive distribution at each successive time step is built by removing a random subset of allocated observations from the posterior predictive for the previous time point. This defines a discounting of individual elements in successive DP centering measures, and is thus an extension of work that induces correlation through linear combinations of DPs (e.g., Müller, Quintana, and Rosner, 2004). Despite the practical and conceptual simplicity of this approach, we found it preferable to adopt a model-based framework that assigns an explicit prior correlation structure to model parameters. (Caron *et al* also require that the entire data-set be stored during filtering, as opposed to the algorithm in 3.2 which tracks only conditional sufficient statistics).

2.3 *BAR Stick-Breaking Processes*

As discussed above, we seek to define a model for correlated Poisson process densities $f_t(\mathbf{x}, y; G_t) = \int k(\mathbf{x}, y; \theta) dG_t(\theta)$ for $t = 1, \dots, T$, and will do so by extending the DP to model $\{G_t\}_{t=1}^T$. The autoregressive Beta (BAR) stick-breaking process defines a prior for correlated random distributions indexed by discrete time, whereby each G_t is marginally distributed as a DP but the underlying stick-breaking proportions form a

dependent surface of beta random variables – that is, the G_t are realized as in (1), but with each successive proportion v_{lt} the member of a correlated multivariate random vector, $V_l = [v_{l1}, \dots, v_{lT}]$. In particular, each independent time series of stick-breaking proportions is modeled as a BAR process, introduced by McKenzie (1985) as a model for positively correlated beta random variables, which can be defined through the evolution equation for a new v_t given v_{t-1} ,

$$v_t = 1 - u_t(1 - w_tv_{t-1}) \text{ where } u_t \sim \text{be}(b, a - \rho), \ w_t \sim \text{be}(\rho, a - \rho), \text{ and } u_t \perp\!\!\!\perp w_t, \quad (4)$$

for positive a and b and $0 < \rho < a$. Since $v_t = (w_t u_t)v_{t-1} + (1 - u_t)$, the autocorrelation for this $\text{BAR}(a, b, \rho)$ process is $\text{corr}(v_t, v_{t-k}) = r(k) = (\mathbb{E}[w]\mathbb{E}[u])^k = [(b\rho/a)/(a + b - \rho)]^k$, such that, as a function of ρ , r is strictly increasing and has a range of the entire unit interval. Note that the BAR process is stationary: if $v_{t-1} \sim \text{be}(a, b)$, then $1 - w_tv_{t-1} \sim \text{be}(a + b - \rho, \rho) \Rightarrow u_t(1 - w_tv_{t-1}) \sim \text{be}(b, a)$, such that $v_t \sim \text{be}(a, b)$.

The BAR stick-breaking model for $G^T = \{G_1, \dots, G_T\}$ is specified through $dG_t(\theta)$

$$= \sum_{l=1}^{\infty} \left(v_{lt} \prod_{i < l} (1 - v_{it}) \right) \delta_{\vartheta_l}(\theta) \text{ with } \vartheta_l \sim G_0(\vartheta; \psi) \text{ and } V_l \sim \text{BAR}(V; a, b, \rho) \ \forall \ l, \quad (5)$$

where $\text{BAR}(V; a, b, \rho)$ denotes the T -dimensional distribution induced by the corresponding BAR process. In general, the marginal distribution for each G_t is a two-parameter beta process, as described by Ishwaran and Zarepour (2000). However, of particular interest is the special case with $V_l \sim \text{BAR}(V; 1, \alpha, \rho) \ \forall \ l$, which implies the v_{lt} are marginally distributed $\text{be}(1, \alpha)$ and the prior for each $f_t = \int k(z; \theta) dG_t(\theta)$ is a $\text{DP}(\alpha, G_0(\psi))$ mixture. The $\text{BAR}(V; 1, \alpha, \rho)$ stick-breaking process is thus able to represent time series of correlated densities while retaining the DP mixture as a marginal model, and hence allows one to make use of extensive existing work related to understanding and prior specification for the DP. Furthermore, this extension of the DP mixture introduces only a single additional prior parameter, $0 < \rho < 1$, and leads to straightforward dependence between stick-breaking weights: $r(k) = [\rho\alpha/(1 + \alpha - \rho)]^k$.

Two short simulations are presented to illustrate this dynamic density estimation framework (both are available as demos in the `Bmix` package for R). In each case, a BAR stick-breaking mixture of normal kernels was fit, through the algorithm of 3.2, to data drawn from a sequence of changing densities. The normal mixtures were specified as for the relevant components of (2), with G_0 scaled appropriately and marginal DP precision $\alpha = 4$. Inference below refers to posterior density expectations, which were obtained at each time interval as $\mathbb{E}[f(z; G)|z^r] \approx (1/N) \sum_{i=1}^N p\left(z | (\mathbf{s}, \mathbf{m})_r^{(i)}, \mathbf{v}_t^{(i)}\right)$, with r the most recent observation corresponding to time t . Examples were fit with $N = 1000$ particles.

For the first example, shown in Figure 2, the BAR mixture model with $\rho = 0.8$ was fit to data from a mixture of two bivariate normal distributions. In this case, the normal densities were static but the mixture probabilities were changing in time: with 20 observations for each of 10 time intervals, the mixture moved from having 0.7 probability assigned to a peaked density centered at (0,0), to having 0.9 probability assigned to a more diffuse density at (4,0). The plotted images show filtered density estimates at time $t = 10$. Comparison against an independent DP fit to the same marginal data shows the advantages of BAR stick-breaking; in particular, the independent fit is unable to discern separation between distinct mixture components.

In the second example, shown in Figure 3, 1-dimensional data was realized from an even-probability mixture of two normal distributions with both mean and variance changing in time. There are 10 observations for each of 100 time points, and the BAR correlation was set at $\rho = 0.5$. Filtered results show that the BAR mixture model was able to adapt to these dynamic densities through a reweighting of static mixture kernels; this was achieved despite having fairly rapidly moving densities with only 10 observations per time point. The figure also provides an image of the corresponding estimates obtained through independent DP mixture fits for each time point; these densities are highly variable and generally inaccurate, thus emphasizing the clear benefit of shrinkage afforded by a correlated mixture approach.

We will also use the data from this latter 1D example to briefly explore related roles of α and ρ . For each marginal density, the standard DP interpretation of α

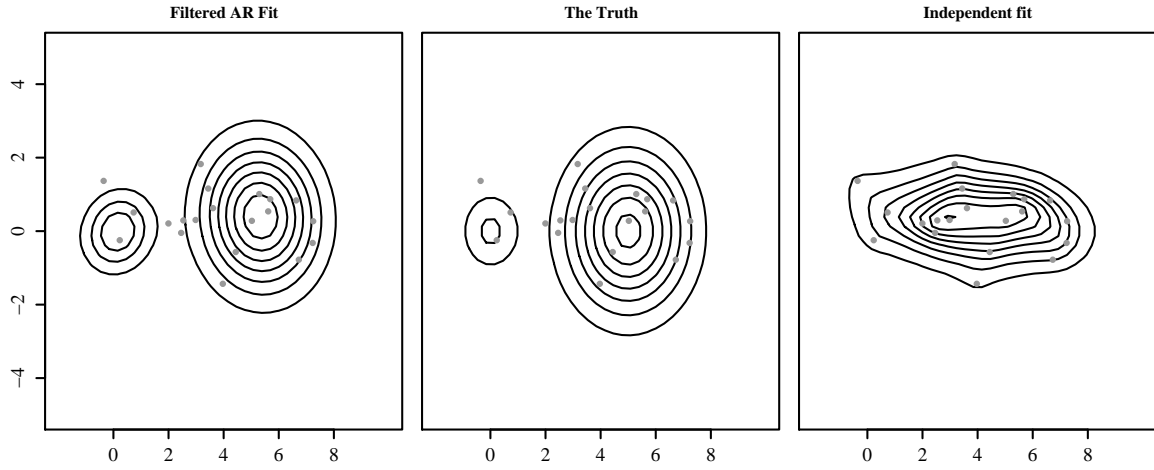


Figure 2: Filtered 2D densities. Left and right plots show posterior density expectations at the final time point (data in grey). The left panel corresponds to BAR with $\rho = 0.8$, center is the true density, and the right is from an independent DP.

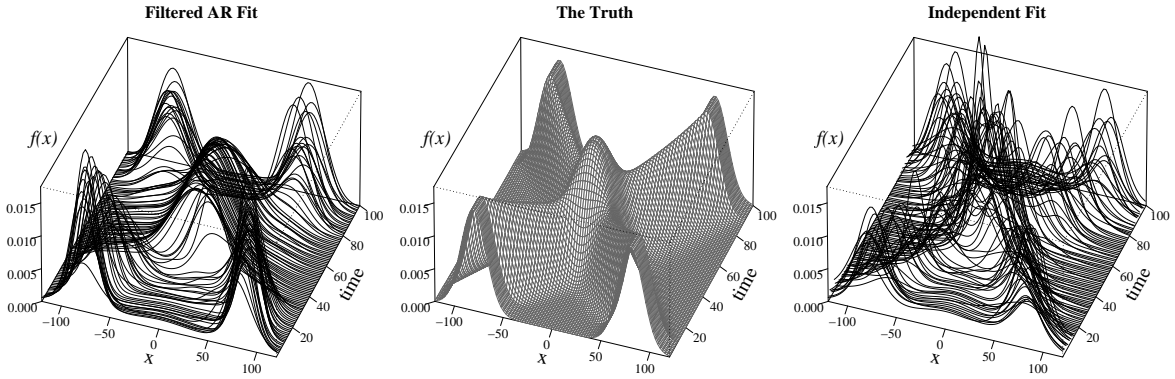


Figure 3: Filtered 1D densities. Left and right plots show posterior density expectations at each time point, for BAR with $\rho = 0.5$ and independent DPs respectively. Data generating densities are plotted in the center.

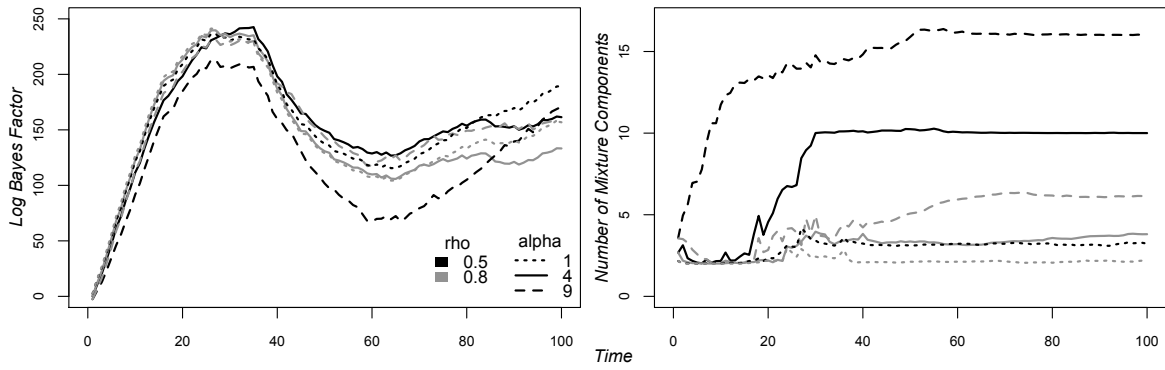


Figure 4: Log Bayes factor and mean number of mixture components after filtering for each time step of 1D data, for models with $\rho \in \{0.5, 0.8\}$ and $\alpha \in \{1, 4, 9\}$. Null model for the Bayes factors has $\rho = 0$ and $\alpha = 4$ (i.e., the right panel of Figure 3).

still holds. However, the autocorrelation function for stick-breaking proportions is an increasing function of both α and ρ . In particular, $r(k)$ approaches ρ from below as α increases. In terms of the number of observed mixture components, the marginal prior expected number of components to which observations *from each time period* are assigned increases with α , but for higher ρ these are more likely to be the same components that were observed at previous times. This behavior is illustrated on the right panel of Figure 4, which shows the mean number of allocated components after filtering for 10 observations from each time period: increasing α from 1 to 9 and increasing ρ from 0.5 to 0.8 both lead to higher numbers. The change is more dramatic over our chosen values of ρ than α , and in both cases the number of components is fairly constant after $t = 50$ (as it should be, since all densities for $t > 50$ have already been sampled). The left panel of Figure 4 shows Bayes factors for each $[\alpha, \rho]$ parametrization, calculated as in Section 3.3 for a null model of independent DP mixtures with $\alpha = 4$, and illustrates that one-step ahead predictive performance is not nearly as sensitive to parameter changes as is the number of mixture components. This should caution use of the DP as a clustering mechanism: dramatically different cluster structures can correspond to relatively similar predictive density surfaces.

3 Posterior Simulation

As the data considered here are ordered by time and inference will naturally occur in an on-line setting, we provide sequential Monte Carlo algorithms for posterior inference. In particular, we develop two new versions of the particle learning approach introduced by Carvalho, Johannes, Lopes, and Polson (2010); refer to that paper for discussion on general efficiency of particle learning and closely related particle filtering algorithms. If an MCMC alternative is preferred, see Cargnoni et al. (1997) for Gibbs sampling that is straightforward to adapt to the Poisson DLM and Taddy (2008) for a Metropolis-Hastings approach to sampling from the BAR stick-breaking posterior.

3.1 Particle Learning for a Poisson DLM

Consider data $y_t \sim N(y_t; F'\eta_t, V)$, with parameters ψ and accompanied by sufficient statistics S_t for η_t given y^t and D_t for ψ given $\{y^t, \eta^t\}$. The particle learning updates of Carvalho et al. (2010) are

$$\begin{aligned} p(S_t, D_t, \eta_t | y^t) &= \int p(S_t, D_t, \eta_t | y_t, \psi, S_{t-1}, D_{t-1}) dP(\psi, S_{t-1}, D_{t-1} | y^t) \\ &\propto \int p(S_t, D_t, \eta_t | y_t, \psi, S_{t-1}, D_{t-1}) p(y_t | S_{t-1}, \psi) dP(\psi, S_{t-1}, D_{t-1} | y^{t-1}), \end{aligned} \quad (6)$$

where further factorization yields $p(\psi, S_{t-1}, D_{t-1} | y^{t-1}) = p(\psi | D_{t-1}) p(S_{t-1}, D_{t-1} | y^{t-1})$ and $p(S_t, D_t, \eta_t | y_t, \psi, S_{t-1}, D_{t-1}) = p(D_t | \eta_t, y_t, D_{t-1}) p(\eta_t | S_t, \psi) p(S_t | y_t, S_{t-1}, \psi)$. Given $\{(\psi, S_{t-1}, D_{t-1})^{(i)}\}_{i=1}^N$ as a particle approximation to $p(\psi, S_{t-1}, D_{t-1} | y^{t-1})$, uncertainty update steps for a new y_t observation are to *resample* with-replacement N particles with weights proportional to the predictive $p(y_t | \{S_{t-1}, \psi\}^{(i)})$, and *propagate* each re-sampled particle (j) through a draw from $p(S_t, D_t, \eta_t | y_t, \{\psi, S_{t-1}, D_{t-1}\}^{(j)})$. Model parameters are sampled off-line from $p(\psi | D_t)$. Note that the term “conditional sufficient statistic” is used loosely, and that S_t and D_t may be augmented to include any auxiliary variables for sampling from $p(S_t, D_t, \eta_t | y_t, \psi, S_{t-1}, D_{t-1})$.

To adapt the basic algorithm for the Poisson DLM in (3), a simple approximation to the Poisson likelihood is used to identify a lower variance slice variable that effectively propagates each Λ_t before the resampling step. In the notation from (6), conditional sufficient statistics for η_t given Λ^t are the Kalman filter moments $S_t = \{m_t, C_t\}$, such that $p(\eta_t | \Lambda^t) = N(\eta_t; m_t, C_t)$. Through standard DLM recursions, these can be combined with model parameters (e.g., V and δ) to obtain forecast moments f_t and Q_t , such that $p(\log(\Lambda_t) | m_{t-1}, C_{t-1}, \psi) = N(\log(\Lambda_t); f_t, Q_t)$. Hence, the uncertainty update of (6) can be expanded to include Λ_t so $p(m_t, C_t, D_t, \eta_t | n^t)$ is proportional to

$$\iint p(m_t, C_t, D_t, \eta_t | (\Lambda, m, C, D)_{t-1}, \psi) p(\Lambda_t, n_t | (m, C)_{t-1}, \psi) d\Lambda_t dP(\psi, (m, C, D)_{t-1} | n^{t-1}), \quad (7)$$

where $p(\Lambda_t, n_t | m_{t-1}, C_{t-1}, \psi) = \text{Po}(n_t; \Lambda_t) N(\log(\Lambda_t); f_t, Q_t)$ is the joint predictive probability function for both crime intensity and count at time t .

The integral with respect to Λ_t in (7) is intractable and must be evaluated stochastically through inclusion of predicted intensity as an auxiliary variable. A naive approach would blindly propagate from $N(\log(\Lambda_t^{(i)}); (f_t, Q_t)^{(i)})$ and use resampling weights proportional to $\text{Po}(n_t; \Lambda_t^{(i)})$. However, $N(\log(\Lambda_t); f_t, Q_t)$ will usually be much more diffuse than the desired conditional slice of $p(\Lambda_t, n_t | m_{t-1}, C_{t-1}, \psi)$, which is inefficient. Instead, note that the normal approximation to the Poisson likelihood for $\log(\Lambda_t)$ is $\text{Po}(n_t; \Lambda_t) \approx N(\log(\Lambda_t); \log(n_t), 1/n_t)/n_t$, such that $\Lambda_t \approx n_t + \varepsilon_t$ and $\varepsilon_t \sim N(0, 1/n_t)$ provides a convenient independent variable for obtaining a draw of $\Lambda_t \propto p(\Lambda_t, n_t | m_{t-1}, C_{t-1}, \psi)$. Hence, $\{(\psi, m_{t-1}, C_{t-1}, D_{t-1})^{(i)}\}_{i=1}^N$ is updated through

Resample: Augment with $\varepsilon_t^{(i)} \sim N(0, 1/n_t)$, set $\Lambda_t^{(i)} = n_t + \varepsilon_t^{(i)}$, and resample indices $\{\zeta(j)\}_{j=1}^N$ with $p(\zeta(j) = i) = \frac{\text{Po}(n_t; \Lambda_t^{(i)})}{N(\log(\Lambda_t^{(i)}); \log(n_t), 1/n_t)} N(\log(\Lambda_t^{(i)}); (f_t, Q_t)^{(i)})$.

Propagate: Standard DLM recursions for each $p((m_t, C_t)^{(j)} | (\Lambda_t, m_{t-1}, C_{t-1}, \psi)^{(j)})$ lead to deterministic updates for state sufficient statistics, and a new state is then propagated from $p(\eta_t | (m_t, C_t)^{(j)})$. Each set of parameter sufficient statistics is updated $p(D_t | (\eta_t, \Lambda_t, D_{t-1})^{(j)})$, and new $\psi^{(j)}$ is drawn from $p(\psi | D_t^{(j)})$.

This basic algorithm provides a general solution to posterior simulation for dynamic models of Poisson intensity, and the ε_t slice trick allows for direct application of standard techniques for updating various DLM specifications. Note that the normal approximation improves as n_t increases, and that in Section 4 our reweighting to account for this approximation yields weights very close to one (> 0.95).

3.2 Particle Learning for the BAR DDP

Carvalho, Lopes, Polson, and Taddy (2009) introduce a particle learning framework for general nonparametric mixture models. This section will develop a version that applies when correlated random mixing distributions are specified through the BAR stick-breaking process of equation (5). The algorithm is completely generic, in the sense that it does not depend on choice of mixture kernel; a more concrete illustration of the methodology is contained in Section 4 and in the Appendix.

The model holds that $z_r \sim f_{\tau(r)}(z; G_{\tau(r)}) = \sum_{l=1}^{\infty} (v_{l\tau(r)} \prod_{i < l} (1 - v_{i\tau(r)})) k(z_{\tau(r)}; \vartheta_l)$, for $r = 1, \dots, R$ and $\tau(r) \in \{1, \dots, T\}$, where the v_{lt} evolve in time according to a $\text{BAR}(a, b, \rho)$ process and each ϑ_l is *iid* from $G_0(\psi)$. In this, r denotes the ordered index of consecutive observations, each of which occurs during the time interval denoted by $\tau(r)$, such that $\sum_{r=1}^R \mathbb{1}_{[\tau(r)=t]} = n_t$ and $\sum_{t=1}^T n_t = R$. It is convenient to break the mixture with latent variables, $\theta^R = \{\theta_1, \dots, \theta_R\}$, such that $z_r \sim k(z_r; \theta_r)$ and $\theta_r \sim G_{\tau(r)}$, for $r = 1, \dots, R$. These latent parameters can be re-written as $\boldsymbol{\vartheta} = \{\vartheta_1^*, \dots, \vartheta_{L_r}^*\}$, the L_r members of $\{\vartheta_l\}_{l=1}^{\infty}$ that are allocated to at least one observation, and the indicator $k^r = \{k_1, \dots, k_r\}$ such that $\theta_i = \vartheta_{k_i}^*$ and $z_i \sim f(z; \vartheta_{k_i}^*)$. Since location parameters are constant in time, it is possible to summarize information available at time r about the allocated $\boldsymbol{\vartheta}$ through the number of observations allocated to each unique component, $\mathbf{m}_r = (m_{r1}, \dots, m_{rL_r})$ such that $m_{rj} = \sum_{i=1}^r \mathbb{1}_{[k_i=j]}$ and $\sum_{j=1}^{L_r} m_{rj} = r$, and conditional sufficient statistics $\mathbf{s}_r = (s_{r1}, \dots, s_{rL_r})$, where $p(\vartheta_j^* | \{z_i, k_i\}_{i=1}^r) = p(\vartheta_j^* | s_{rj}, n_{rj})$.

A basic particle learning approach to mixtures, as in Carvalho, Lopes, Polson, and Taddy (2009), provides filtering for \mathbf{m}_r and \mathbf{s}_r through integration over a distribution G that is constant in time. For BAR stick-breaking mixtures, however, it is necessary to track the dynamic stick-breaking proportions that control dependence between densities (hence, the standard Pólya urn approach to inference is impossible). Although the full set of $v_{l\tau(r)}$ is infinite dimensional, it is possible to work with only the L_r stick-breaking proportions associated with $\boldsymbol{\vartheta}$, while integrating over the rest. The BAR stick-breaking mixture model holds that each new z , within time period t , is allocated to either the parametrization $\vartheta_l \in \boldsymbol{\vartheta}$ with probability $v_{lt} \prod_{i < l} (1 - v_{it})$, or assumed drawn from $k(z; \vartheta_{L_r+1})$ with probability $\prod_{l \leq L_r} (1 - v_{lt})$, where unallocated ϑ_{L_r+1} is drawn from $G_0(\psi)$. This leads to the *conditional* posterior predictive probability function, given sampled L_r weights,

$$p(z | \{v_{l\tau(r)}, \vartheta_{l\tau(r)}\}_{l=1}^{L_r}) = \sum_{l=1}^{L_r} \left[v_{l\tau(r)} \prod_{i < l} (1 - v_{i\tau(r)}) k(z; \vartheta_l) \right] + \prod_{l \leq L_r} (1 - v_{l\tau(r)}) \int k(z; \theta) dG_0(\theta; \psi). \quad (8)$$

Thus, $\boldsymbol{\vartheta} = \{\vartheta_l\}_{l=1}^{L_r}$ and $\mathbf{v}_{\tau(r)} = \{v_{l\tau(r)}\}_{l=1}^{L_r}$ are representative of the posterior after

observation r , and functionals of interest are marginalized over remaining parameters. This approach is of the same spirit as the slice sampling in Walker (2007), but sequential particle inference allows for a more simple formulation.

Since it is not practical to fully integrate over stick-breaking proportions, the augmented filtering algorithm will track $\mathcal{V}_r = \{\mathbf{v}_{\tau(r)-1}, \mathbf{c}_r\}$, where $\mathbf{v}_{\tau(r)-1}$ is the set of stick-breaking proportions from the *previous* time period (corresponding to L^* allocated ϑ_l at the end of period $\tau(r) - 1$) and $\mathbf{c}_r = \{c_{r1}, \dots, c_{rL_r}\}$ is the count of observations during time period $\tau(r)$ allocated to each of $L_r \geq L^*$ different ϑ_l . Given prior information from $\mathbf{v}_{\tau(r)-1}$, \mathbf{c}_r is sufficient for $\mathbf{v}_{\tau(r)}$. In particular, $p(\mathbf{v}_{\tau(r)}|\mathcal{V}_{r-1})$ is proportional to $p(\mathbf{c}_r|\mathbf{v}_{\tau(r)})p(\mathbf{v}_{\tau(r)}|\mathbf{v}_{\tau(r)-1})$

$$= \prod_{l=1}^{L_{r-1}} \text{bin}\left(c_{rl}; \sum_{i=l}^{L_{r-1}} c_{ri}, v_{l\tau(r)}\right) \prod_{l=1}^{L^*} \text{BAR}(v_{l\tau(r)}|v_{l\tau(r)-1}; a, b, \rho) \prod_{l=L^*+1}^{L_{r-1}} \text{be}(v_{l\tau(r)}; a, b), \quad (9)$$

where $\text{bin}(\cdot; n, p)$ is a binomial with probability p and n trials, $\text{BAR}(v_{l\tau(r)}|v_{l\tau(r)-1}; a, b, \rho)$ is the evolution in (4), and the beta priors for $l > L^*$ are implied by BAR process marginals. Sequential updates are thus

$$\begin{aligned} p(\mathbf{s}_r, \mathbf{m}_r, \mathcal{V}_r|z^r) &= \int p(\mathbf{s}_r, \mathbf{m}_r, \mathcal{V}_r|\mathbf{s}_{r-1}, \mathbf{m}_{r-1}, \mathcal{V}_{r-1}, z_r) dP(\mathbf{s}_{r-1}, \mathbf{m}_{r-1}, \mathcal{V}_{r-1}|z^r) \\ &\propto \int p(k_r, \mathcal{V}_r|(\mathbf{s}, \mathbf{m}, \mathcal{V})_{r-1}, z_r) p(z_r|(\mathbf{s}, \mathbf{m}, \mathcal{V})_{r-1}) dP((\mathbf{s}, \mathbf{m}, \mathcal{V})_{r-1}|z^{r-1}) \\ &\equiv \iint p(k_r|(\mathbf{s}, \mathbf{m})_{r-1}, \mathbf{v}_{\tau(r)}, z_r) p(z_r, \mathbf{v}_{\tau(r)}|(\mathbf{s}, \mathbf{m}, \mathcal{V})_{r-1}) d\mathbf{v}_{\tau(r)} dP((\mathbf{s}, \mathbf{m}, \mathcal{V})_{r-1}|z^{r-1}), \end{aligned} \quad (10)$$

where $p(z_r, \mathbf{v}_{\tau(r)}|(\mathbf{s}, \mathbf{m}, \mathcal{V})_{r-1}) = p(z_r|(\mathbf{s}, \mathbf{m})_{r-1}, \mathbf{v}_{\tau(r)}) p(\mathbf{v}_{\tau(r)}|\mathcal{V}_{r-1})$, and the predictive is as in (8), but integrated over the posterior for $\boldsymbol{\vartheta}$ given \mathbf{m}_{r-1} and \mathbf{s}_{r-1} .

The extra level of complexity in (10), when compared to a basic particle learning update as in (6), is due to the need to propagate stick-breaking proportions in evaluation of the predictive probability function. However, the mechanics of this approach are fairly simple in practice, and are most easily understood in the context of a particle update iteration. Consider an existing particle representation of the posterior, $p(\mathbf{s}_{r-1}, \mathbf{m}_{r-1}, \mathcal{V}_{r-1}, \psi|z^{r-1}) \propto \sum_{i=1}^N \mathbb{1}_{\{(\mathbf{s}, \mathbf{m}, \mathcal{V})_{r-1}^{(i)}, \psi^{(i)}\}}$. Upon arrival of observation z_r ,

Augment: Incorporate current auxiliary stick-breaking proportions through a sampling and re-weighting procedure, with the targeted conditional posterior $p(\mathbf{v}_{\tau(r)}|\mathcal{V}_{r-1})$ from equation (9). In detail, with each particle (i)

- For $l \leq L^*$, each $v_{l\tau(r)}$ is propagated from $\text{BAR}(v_{l\tau(r)} | v_{l\tau(r)-1}; a, b, \rho)$.
- For $L^* < l \leq L_{r-1}$, draw $v_{l\tau(r)}$ from $\text{be}\left(v; a + c_{rl}, b + \sum_{i=l+1}^{L_{r-1}} c_{ri}\right)$.
- New weights are $p_c^{(i)} \propto p(\{c_r\}_{r=1}^{L^*} | \mathbf{v}_{\tau(r)}) = \prod_{l=1}^{L^*} \text{bin}\left(c_{rl}; \sum_{i=l}^{L_{r-1}} c_{ri}, v_{l\tau(r)}\right)$.

Resample: The particle set is replaced with N resampled particles corresponding to indices $\{\zeta(j)\}_{j=1}^N$, where $p(\zeta(j) = i) \propto p_c^{(i)} p(z_r | (\mathbf{s}_{r-1}, \mathbf{m}_{r-1}, \mathbf{v}_{\tau(r)})^{(i)})$, with $p_c^{(i)}$ calculated in the above augment step and $p(z_r | (\mathbf{s}_{r-1}, \mathbf{m}_{r-1}, \mathbf{v}_{\tau(r)})^{(i)}) = \int p(z_r | \mathbf{v}_{\tau(r)}^{(i)}, \boldsymbol{\vartheta}) dP(\boldsymbol{\vartheta} | (\mathbf{s}, \mathbf{m})_{r-1}^{(i)})$, the posterior predictive from equation (8) integrated over conditional uncertainty in $\boldsymbol{\vartheta}$.

Propagate: Draw for each $k_r^{(i)}$ according to probability function,

$$\text{for } l = 1, \dots, L_{r-1}, \quad p(k_r = l) = v_{l\tau(r)} \prod_{i < l} (1 - v_{i\tau(r)}) \int k(z_r; \vartheta_l) dP(\vartheta_l | s_{rl}, n_{rl}),$$

$$\text{and } p(k_r = L_{r-1} + 1) = \prod_{l \leq L_{r-1}} (1 - v_{l\tau(r)}) \int k(z; \theta) dG_0(\theta; \psi).$$

This draw for k_r is followed by a series of deterministic sufficient statistic updates. Set $m_{rk_r}^{(i)} = m_{r-1k_r}^{(i)} + 1$ and copy the rest, while the update for $\mathbf{s}_{r-1} \rightarrow \mathbf{s}_r$ is model specific. If $\tau(r) \neq \tau(r-1)$ (new time period) update each $\mathcal{V}_r^{(i)}$ such that $\mathbf{v}_{\tau(r)-1}^{(i)} = \mathbf{v}_{\tau(r-1)}^{(i)}$ and $\mathbf{c}_r^{(i)}$ is empty except for $c_{rk_r}^{(i)} = 1$; otherwise, $\mathcal{V}_r^{(i)} = \mathcal{V}_{r-1}^{(i)}$ but with new $c_{rk_r}^{(i)} = c_{r-1k_r}^{(i)} + 1$. Finally, resample each $\psi^{(i)}$ from $p(\psi | (\mathbf{s}, \mathbf{m})_r^{(i)})$.

3.3 Sequential Marginal Likelihood Estimation

Before moving to the motivating application and results, it is useful to note the sequential marginal data likelihood calculation (and the associated notion of Bayes factors) that is relied upon in both 4.1 and 4.2 for comparing different levels of “smoothness”

specification. For general data $z^T = \{z_t\}_{t=1}^T$, with sufficient statistics S_t and parameters ψ , marginal data likelihood is $p(z^T) = \prod_{t=1}^T p(z_t|z^{t-1})$ and the factors can be approximated within PL as

$$p(z_t|z^{t-1}) = \mathbb{E} [p(z_t|\psi, S_{t-1})p(\psi, S_{t-1}|z^{t-1})] \approx \frac{1}{N} \sum_{i=1}^N p(z_t|\psi^{(i)}, S_{t-1}^{(i)}). \quad (11)$$

This simple calculation will be applied in 4.1, where $z_t = n_t$, and in 4.2, where $z_t = (\mathbf{x}_t, y_t)$, to compare discount factors (δ) and correlation parameters (ρ) respectively. Since total intensity and process density are independent in the posterior, inference can proceed independently for each.

It is possible that the parameters of interest could have been sampled *during* particle learning through use of, say, rejection sampling (there are no simple conjugate prior schemes for these parameters). However, as with any sequential Monte Carlo sampler, it is best to limit the number of sampled hyperparameters so as to avoid particle degeneracy. Furthermore, it is conceptually helpful to have these key model elements summarized in the clear manner provided by Bayes factors and (11) shows the explicit connection between marginal likelihood and predictive performance. Alternatively, one could obtain posterior summaries through use of discrete hyperpriors and an augmented draw for parameters of interest (in the case of ρ , however, this would require potentially prohibitive calculations involving all sampled BAR weights).

4 Application and Inference

This section details application of the methodology developed in this article to the crime event data introduced in Section 1.1. Computational aspects follow the algorithms outlined in Section 3, with specific details relegated to the Appendix. Results are initially presented separately for each of the integrated intensity in 4.1 and the dynamic process density in 4.2, with combined inference in 4.3.

All results are based on algorithms applied with 1000 particles. This was judged to be a large enough set to allow for sufficient exploration of high-probability areas of the state-space, but small enough for fast computation with the large dataset and

multiple models under consideration. Experience with the data would support this judgement. In particular, a subset of the Bayes factor calculations were found to be practically unchanged for repeated filtering runs with a larger number of particles.

4.1 *A First-Order DLM Applied to Crime Counts*

A simple first-order version of the Poisson DLM was applied to the time series of 52 weekly crime counts. In detail, the model holds that

$$\log(\Lambda_t) = \eta_t + v_t, \quad v_t \sim N(0, V) \quad \text{and} \quad \eta_t = \eta_{t-1} + w_t \quad \text{for } t = 1, \dots, 52, \quad (12)$$

where $V \sim \text{ga}(V^{-1}; \nu_0/2, D_0/2)$, $\text{var}(w_t)$ is specified through the discount factor δ , and initial information is summarized by $\eta_0 \sim N(m_0, C_0)$. This “random walk plus noise” was chosen in lieu of more complicated models due to both the limited data size (only 52 weeks) and the fact that change in η_t provides an easily interpretable and robust first-order approximation to underlying fundamental change. Furthermore, although a seasonal component should probably be incorporated for data extensions that span multiple years, experimentation with other models failed to indicate usefulness of more structured dependencies; for example, the DLM specification corresponding to a traditional AR(1) model leads to a posterior for the relevant coefficient that is peaked at zero. The applied parametrization has $\nu_0 = 6$ and $D_0 = 1/5$, for prior expected V of about 3/4 the sample variance of $\log(n_t)$, and $m_0 = \log(210)$ and $C_0 = 1/10$, for a diffuse initial state. Analysis was robust to reasonable changes in this parametrization.

The discount factor plays a central role in governing behavior of the random walk (e.g., $W_t/V \xrightarrow{t} (1 - \delta)^2/\delta$ quickly for most models), which is itself representative of persistent change in crime intensity. Inference for δ was addressed through the use of Bayes factors, based on the sequential marginal likelihood calculation. The model was fit for 6 values of δ from 0.4 to 0.9 as well as for a pure random walk (with $V = 0$ and W assigned the $\text{ga}(W^{-1}; \nu_0/2, D_0/2)$ prior) and a constant mean (i.e., $\delta = 1$).

Results are shown in Figures 5 and 6. The first plots illustrate filtered posterior distributions for η_t corresponding to each discount factor, and it is easy to see the

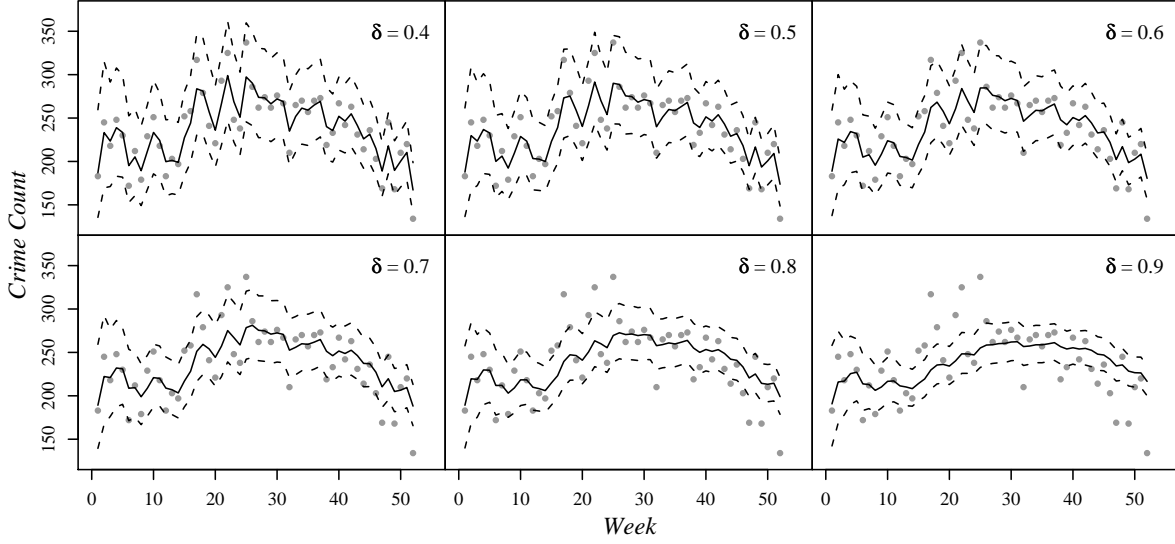


Figure 5: Integrated Poisson intensity DLM: filtered moving average state $\exp(\eta_t)$ corresponding to different discount factors. Solid line is posterior mean, and the dashed lines are posterior 5th and 95th percentiles.

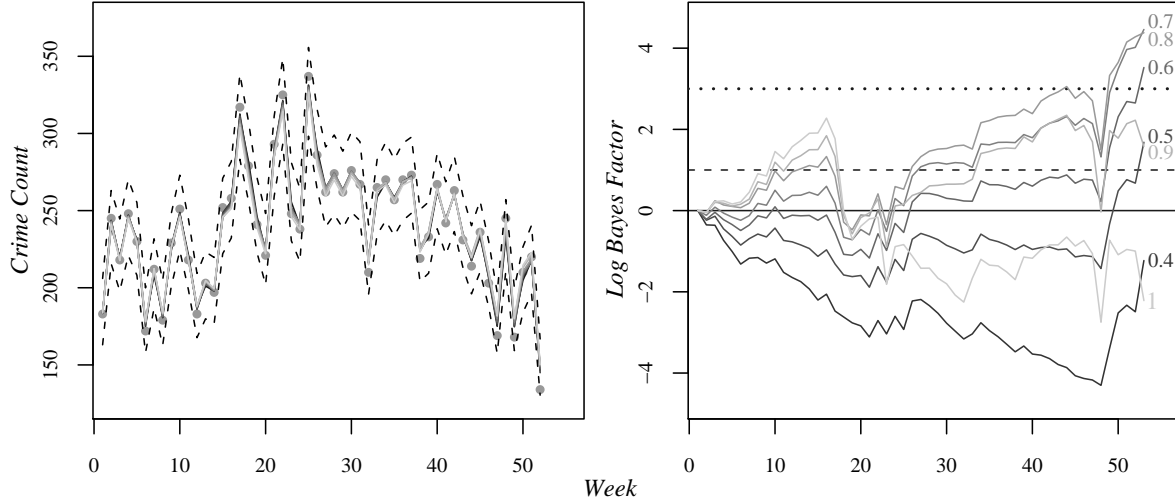


Figure 6: Filtered predictions of Poisson intensity Λ_t and log Bayes factors for the model corresponding to each discount factor. On the left plot, black and dashed lines show mean and 90% posterior interval for the pure random walk model, with $\log(\Lambda)_t = \eta_t$ (i.e., the same quantities plotted in figure (5), but with $\delta = 0$). Filtered posterior expected Λ_t for each discount are then plotted on top, shaded according to the legend in the right panel. Bayes factors are calculated relative to the null hypothesis of a pure random walk, shown as the solid line at zero. The dashed line at one and dotted line at three represent thresholds for positive and strong evidence, respectively, as accorded by Kass and Raftery (1995).

effect of δ in determining smoothness for expected crime intensity. In each case, it appears that the series in η_t is offering more or less smoothed versions of a weather dependent (i.e., annual) crime trend: high crime in warm months and lower crime in winter. At the same time, the left panel of Figure 6 shows that the different models lead to virtually identical estimates for Λ_t , providing evidence that various values of δ do not restrict ability to fit the data. Finally, filtered log Bayes factors, evaluated for each δ against the pure random walk model, are shown in the right panel of 6. There is convincing evidence of a discount between 0.5 and 0.9, with δ of 0.7 or 0.8 most probable. As there is a higher risk for predictive performance associated with a too small W (i.e., a too large δ ; refer to West and Harrison, 1997, 2.4.1), the value $\delta = 0.7$ was adopted for analysis in Section 4.3.

4.2 *A BAR Stick-Breaking Model for Dynamic Crime Event Densities*

The series of weekly process densities was modeled under the kernel and base measure specification of (2), with prior for time series of stick-breaking weights induced through a $\text{BAR}(1, \alpha, \rho)$ process (implying DP marginal weekly priors). Recalling that the multivariate normal kernel applies on transformed logit locations, the base distribution is parametrized with $\gamma = 0$, $\kappa = 1/1000$, $\nu = 3$, and hyperparameters $\nu_\Omega = 3$, and $\Psi_\Omega = \mathbf{I}/10$. Specification of G_0 is completed with a $\text{be}(0.5, 0.5)$ centering distribution for q . It was possible to make major changes to this parametrization (e.g., scaling up κ by an order of magnitude or adopting $\text{be}(q; 2, 3)$ as the base distribution for mark probability) without perceptible change to marginal weekly density estimates.

The applied version of the BAR stick-breaking process is parametrized by only two parameters, a precision α and correlation term ρ . The former, which is also the precision parameter for associated marginal DP mixture priors, is fixed at $\alpha = 4$ for $1/(\alpha + 1) = 0.2$ marginal prior probability that any single pair of observations are allocated to the same mixture component. Again, for a dataset of this size, it was observed that inference about the predictive density surface is very robust to changes in α ; moreover, the straightforward marginal interpretability of α makes it a natural candidate for user specification. The correlation parameter plays a much larger role

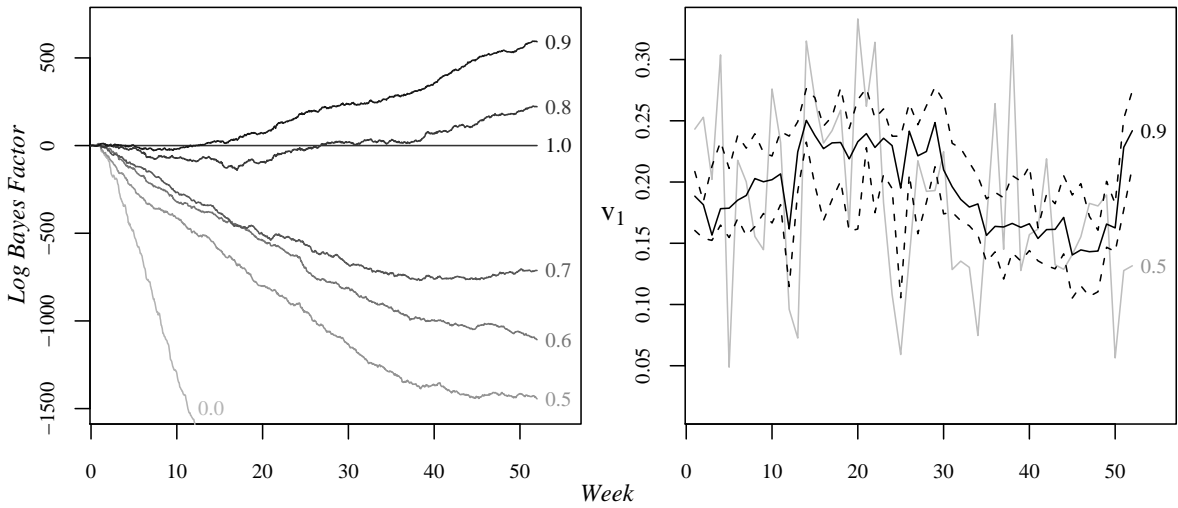


Figure 7: Density dynamics. The left plot shows log Bayes factors for models corresponding to a set of possible BAR correlation parameters. The Bayes factors are calculated relative to the null hypothesis of a static density ($\rho = 1$), shown as the solid line at zero. The second plot shows evolution of the first stick-breaking weight, with mean and 90% posterior interval plotted in black for the $\rho = .9$ model. For comparison, the filtered mean for v_1 corresponding to $\rho = .5$ is shown in grey.

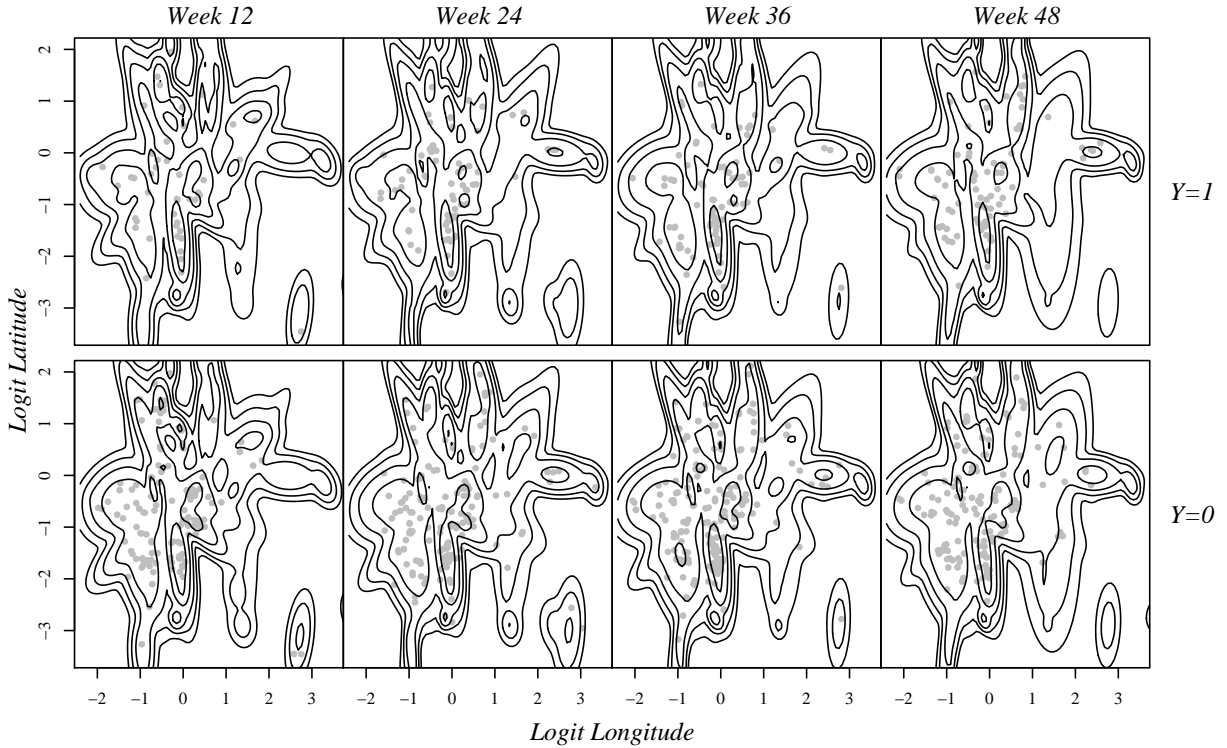


Figure 8: Dynamic log densities for transformed event locations. Filtered posterior log mean density contours for weeks 12, 24, 36, and 48, for point processes corresponding to both extremely violent crime ($Y=1$) and non-extreme crime ($Y=0$). Contour lines move from -10 to 0 in steps of 2. Events for each week are plotted in in grey.

in posterior inference, as ρ determines how much information is shared across weeks. Hence, a Bayes factor approach was applied to choose among 5 possible values for ρ between 0.5 and 0.9, as well as independent weekly densities ($\rho = 0$) or a single unchanging density ($\rho = 1$).

Figure 7 shows sequential log Bayes factors corresponding to each model, with the constant model as baseline. The more than 12,000 observations allow for clear rejection of $\rho < 0.8$, indicating a relatively high persistence in crime densities. The most probable choice is $\rho = 0.9$, for weight autocorrelation of 0.88, and final analysis was based on this model. As an illustration of BAR dynamics, and of the contrast between high and low correlation processes, the right plot of Figure 7 shows filtered posterior distributions for the first stick-breaking weight under $\rho = 0.9$, as well as the mean for this weight when $\rho = 0.5$ (since the first observation is allocated to this component, v_1 is weakly identified in practice).

Finally, although this application is focused on extremely violent crime, the space-time dependent correlation between non-extreme and extreme violence is essential to our plan of having the former inform inference about the latter. Figure 8 compares posterior mean density surfaces corresponding to each type of crime, for weeks 12, 24, 36, and 48. Density values for individual particles are calculated, for each of $y = 0$ or 1 and a set of grid locations, as the Rao-Blackwellized conditional posterior predictive density, marginalized over components $l > L_r$, as detailed in equation (13) of the appendix; refer to Taddy and Kottas (2009) for further guidance on Poisson process inference under this type of model. As would be expected, major features of the densities appear constant in time and across crime types. However, it is possible to discern subtle distinction in area details, indicating that the relationship between crime types is non-constant in both time and space.

4.3 Inference about Crime Intensities

Filtered results from Sections 4.1 and 4.2 are combined to obtain full posterior inference about weekly intensity functions for extremely violent crime, detailed in Figure 9. Such sequential intensity maps are the most obvious product of our framework; they

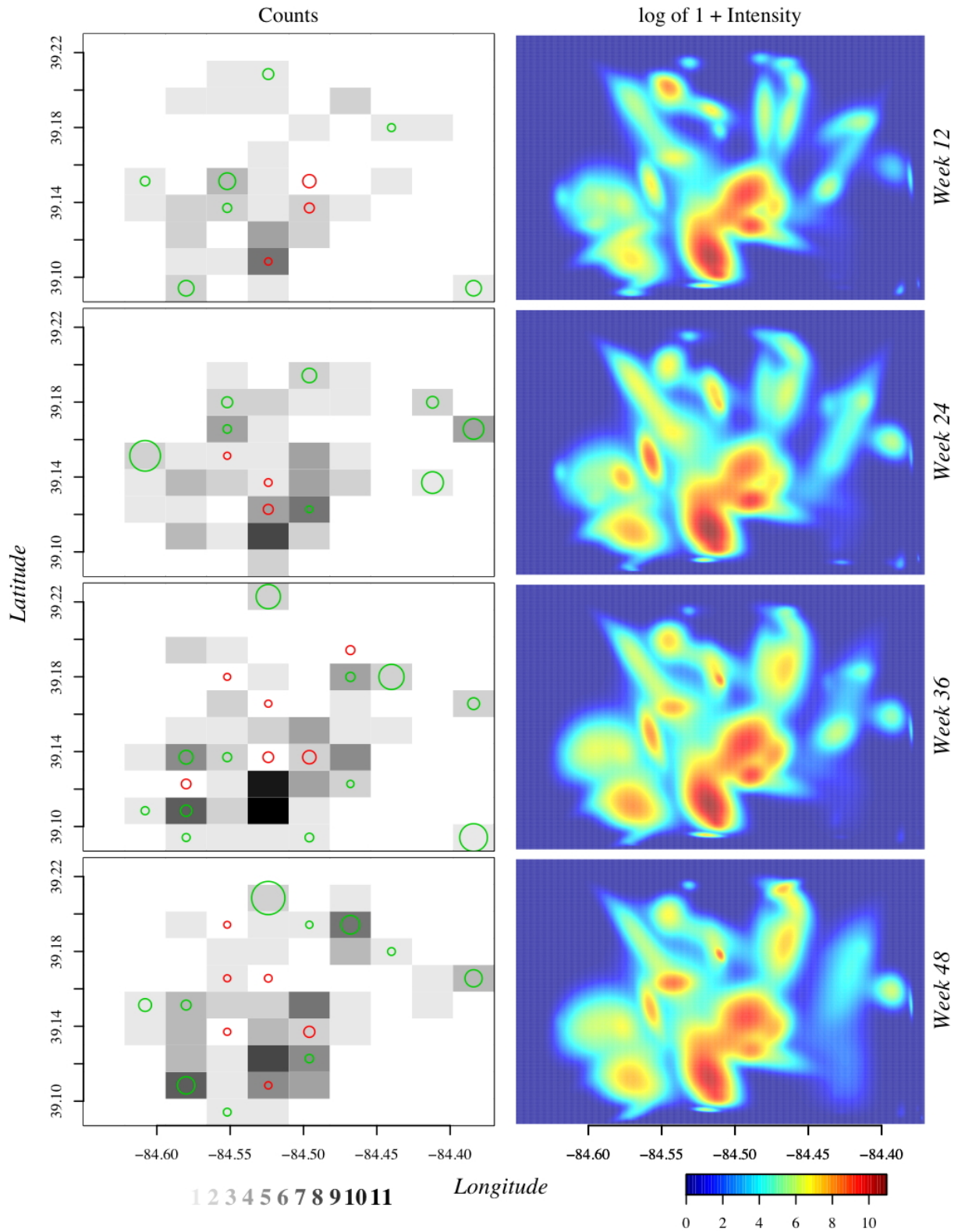


Figure 9: Filtered posterior log mean intensity for point processes corresponding to extremely violent crime events. Crime counts for 100 bins are shown left of each intensity map. In addition, standardized Pearson residuals are plotted with point diameter proportional to the absolute size of the residual (only residuals with $|r_{ti}| \geq 1$ are shown). Red indicates negative residuals and green are positive residuals.

provide a more informed visualization of extremely violent crime activity which can then be used in, for example, police resource allocation and dissemination of statistics to the wider public. Bayesian mean-inference estimation allows for information to be aggregated across weeks and over model uncertainty, and dependent nonparametric smoothing leads to images that can be obtained in a relatively automatic manner without detailed guidance from specific ground-level factors. When required, the intensity surfaces can also form the basis for more focused location specific inference such as numerically integrated intensity or conditional mark density plots.

The maps in Figure 9 vary only subtly in time, as dictated by the persistence in process density inferred during Section 4.2. However, there is noticeable change to some areas of relatively high intensity. We do not have here the scope to consider any conclusions about dynamics of crime in Cincinnati, but anecdotal sources of change include natural seasonality due to weather as well as the effect of added police presence in the Over-the-Rhine neighborhood (just north of downtown).

The left panel of Figure 9 shows counts of extremely violent crime events in 100 evenly spaced bins – a raw “model free” analogy to the corresponding Bayesian mean intensity maps. For comparison, note that transformed intensity $\log(\lambda(\mathbf{x}) + 1) = 5.5$ implies an expected count of about 0.1 extremely violent crimes per bin area, $\log(\lambda(\mathbf{x}) + 1) = 7.8 \Rightarrow \mathbb{E}[\text{count/bin}] = 1$, and $\log(\lambda(\mathbf{x}) + 1) = 10.1 \Rightarrow \mathbb{E}[\text{count/bin}] = 10$. There are a variety of different count residual measures for nonhomogeneous Poisson processes, and Baddeley, Turner, Møller, and Hazelton (2005) provide a survey. Their Pearson residuals are appealing for having a parameter free theoretical variance, equal to the area of the count bin. Through simple numerical integration of the posterior mean intensity, scaled Pearson residuals can be calculated here, for each bin \mathcal{B}_i at time t , as $r_{ti} = \left(\sum_{j: \mathbf{x}_j \in \mathcal{B}_i, y_j=1} 1/\sqrt{\Lambda_t f_t(\mathbf{x}_j, 1)} - \int_{\mathcal{B}_i} \sqrt{\Lambda_t f_t(\mathbf{s}, 1)} d\mathbf{s} \right) / |\mathcal{B}_i|$.

The resulting residuals are plotted on top of each corresponding bin on the left panel of Figure 9. In support of a decent model fit, the residual mean is near zero (0.005) and 95% are of absolute value less than 1.65. However, the residual distribution is skewed to the right by large positive residuals (i.e. higher than expected counts) appearing in areas of relatively low intensity. It seems probable that this behavior

is caused by correlation between crime events: even though extreme violence occurs with low intensity in these areas, when it does happen the events come in bunches. If so, we are encountering the general limitations of a Poisson process assumption, which restricts the model to conditions *i-iii* from Section 2.1. As was discussed in that section, the nonhomogeneous Poisson process can only provide a first-order approximation to point patterns with dependent events. However, any improvements in this respect will require modeling for individual event dependencies, leading to a considerably more elaborate modeling enterprise. From the outset, our intention was only to provide a simple Bayesian framework that provides accurate results despite avoiding any detailed modeling for pairwise dependencies. Hence, we are pleased with the ability of the Bayesian auto-regressive mixture model to provide such inference for correlated sequences of complicated marked point process.

5 Discussion

This article presents a flexible Bayesian framework for smoothing and short term prediction of time series of spatial nonhomogeneous point patterns. The goal throughout has been to develop easily generalizable tools for efficient sharing of information across time, space, and the mark domain. To this end, we describe a novel single- θ version of the DDP that provides a dynamic extension to stick-breaking mixture models. When combined with a DLM for correlated Poisson random variables, this results in a powerful setup for tracking arbitrary marked spatial Poisson processes. Since the BAR dynamic extension does not depend upon kernel or base measure specification, and since the DLM framework has been well studied and is hugely flexible, this work should apply in a wide variety of settings.

Basic sequential Monte Carlo algorithms have been provided to allow on-line uncertainty updating for model components, and these straightforward algorithms should appeal to many practitioners. Although the notation can get complicated, posterior simulation involves familiar posterior conditional calculations and the algorithms will apply in fairly general settings. As with any particle based filtering, there is always a risk of particle degeneracy and Monte Carlo error. This is mitigated through carefully

chosen particle states (e.g., not storing each k_r allocation or current stick-breaking weights), with propagation steps that are designed specifically for the data and model at hand (e.g., via the normal approximation to Poisson likelihood for the DLM), and by marginal likelihood inference for sensitive prior parameters (i.e., δ and ρ).

It is helpful to note some common themes underlying the approach in this article. The likelihood factorization enables a directed relaxation of standard model constraints: Poisson conditions are maintained and Λ_t is modelled parametrically, but a nonparametric scheme is adopted for flexible learning of the process density. This Poisson assumption is in the same spirit as the decision to use a random walk DLM: each model choice leads to a robust first-order approximation that avoids added complication of more detailed dependencies. Furthermore, each major modeling component is designed to allow for hierarchical smoothing: the raw data, n_t or $(\mathbf{x}, y)_r$, are associated with a top-level random variable, Λ_t or k_r , that links the observations to a smooth underlying mechanism – either the random walk in η or the density $\sum p_l k(\mathbf{x}, y; \vartheta_l)$. Finally, models are parametrized to isolate a single parameter, δ or ρ , that governs degree of smoothness, and this parameter is chosen based upon a sequential Bayes factor analysis. The end result is a straightforward framework for automatic information sharing across data dimensions, and we are encouraged by success of this approach to flexible Bayesian semiparametric modeling.

Appendix: Details of the Applied Particle Learning Algorithm

Section 3.1 details the particle learning algorithm for a Poisson intensity DLM. In the context of Section 4, with $\log(\Lambda_t) = \eta_t + v_t$ and W_t specified through the discount δ , the recursions are straightforward. First, $p(\eta_t | n^{t-1}) = N(m_{t-1}, R_t)$, with $R_t = \frac{1}{\delta} C_{t-1}$, such that forecast moments are $f_t = m_{t-1}$ and $Q_t = R_t + V$. Parameter updates are then $m_t = m_{t-1} + (R_t/Q_t)(\log(\Lambda_t) - f_t)$ and $C_t = R_t V/Q_t$. Finally, the off-line draw for V is from $\text{ga}(V^{-1}; \nu_t/2, D_t/2)$, where $\nu_t = \nu_{t-1} + 1$ and $D_t = D_{t-1} + (\log(\Lambda_t) - \eta_t)^2$.

Section 3.2 describes the general particle learning algorithm for BAR stick-breaking process mixtures. It only remains to outline details related to the kernel and base measure specification of (2), as applied in Section 4. It is convenient to work on the

logit transformed scale, with $\mathbf{z} = \text{logit}(\mathbf{x})$. Conditional sufficient statistics, given r observations, for each unique mixture component (i.e., each element $s_{rl} \in \mathbf{s}_r$) are the location means $\bar{\mathbf{z}}_{rl} = \sum_{j:k_j=l} \mathbf{z}_j / m_{r,j}$, sums of squares $S_{rl} = \sum_{j:k_j=l} (\mathbf{z}_j - \bar{\mathbf{z}}_{rl})(\mathbf{z}_j - \bar{\mathbf{z}}_{rl})' = \sum_{j:k_j=l} \mathbf{z}_j \mathbf{z}_j' - n_{rl} \bar{\mathbf{z}}_{rl} \bar{\mathbf{z}}_{rl}'$, and mark counts $R_{rl} = \sum_{j:k_j=l} y_j$ (recall that $y = 1$ for extreme violence, and is otherwise 0).

The resample step relies upon the Rao-Blackwellized predictive probability function, for \mathbf{z} and y realized at time t and conditional on r observations,

$$p(\mathbf{z}, y | \mathbf{s}_r, \mathbf{m}_r, \mathbf{v}_t) = \sum_{j=1}^{L_r} \left(v_{jt} \prod_{i < j} (1 - v_{it}) \right) \text{St}(\mathbf{z}; a_{rl}, B_{rl}, c_{rl}) \hat{q}_{rl}^y (1 - \hat{q}_{rl})^{1-y} \quad (13) \\ + \left(\prod_{i \leq L_r} (1 - v_{it}) \right) \text{St}(\mathbf{z}; a_0, B_0, c_0) \hat{q}_0^y (1 - \hat{q}_0)^{1-y},$$

where the Student's t distributions are parametrized by $a_0 = \gamma$, $B_0 = 2\Omega(\kappa + 1)/(\kappa c_0)$, $c_0 = 2\nu - 1$, $a_{rl} = (\kappa\gamma + m_{rl}\bar{\mathbf{z}}_{rl})/(\kappa + m_{rl})$, $B_{rl} = (2\Omega + D_{rl})(\kappa + m_{rl} + 1)/[(\kappa + m_{rl})c_{rl}]$, $c_{rl} = 2\nu + m_{rl} - 1$, and $D_{rl} = S_{rl} + \frac{\kappa m_{rl}}{(\kappa + m_{rl})}(\gamma - \bar{\mathbf{z}}_{rl})(\gamma - \bar{\mathbf{z}}_{rl})'$; the predictive mark probabilities are $\hat{q}_0 = a_q/(a_q + b_q)$ and $\hat{q}_{rl} = (a_q + R_{rl})/(a_q + b_q + \sum_{j=1}^{L_r} R_{rj})$. The propagation distribution for allocation k associated this new (\mathbf{z}, y) data follows easily as $p(k = l | \mathbf{s}_r, \mathbf{m}_r, \mathbf{v}_t, \mathbf{z}, y) = v_{lt} \prod_{j < l} (1 - v_{jt}) \text{St}(\mathbf{z}; a_{rl}, B_{rl}, c_{rl}) \hat{q}_{rl}^y (1 - \hat{q}_{rl})^{1-y}$ for $l \leq L_r$ and $p(k = L_r + 1 | \mathbf{s}_r, \mathbf{m}_r, \mathbf{v}_t) = \prod_{j < L_r} (1 - v_{jt}) \text{St}(\mathbf{z}; a_0, B_0, c_0) \hat{q}_0^y (1 - \hat{q}_0)^{1-y}$.

This leads to straightforward sufficient statistic updates with each new $(\mathbf{z}_{r+1}, y_{r+1})$ observation and propagated allocation k_{r+1} : If $k_{r+1} = L_r + 1$, the new sufficient statistics are defined by $L_{r+1} = L_r + 1$, $\bar{\mathbf{z}}_{r+1L_{r+1}} = \mathbf{z}_{r+1}$, $S_{r+1L_{r+1}} = 0$, and $R_{r+1L_{r+1}} = y_{r+1}$. If $k_{r+1} = l \leq L_r$, then $m_{r+1l} = m_{rl} + 1$ and s_{r+1l} is updated such that $\bar{\mathbf{z}}_{r+1l} = (m_{rl}\bar{\mathbf{z}}_{rl} + \mathbf{z}_{r+1})/m_{r+1l}$ and $S_{r+1l} = S_{rl} + \mathbf{z}_{r+1}\mathbf{z}_{r+1}' + m_{rl}\bar{\mathbf{z}}_{rl}\bar{\mathbf{z}}_{rl}' - m_{r+1l}\bar{\mathbf{z}}_{r+1l}\bar{\mathbf{z}}_{r+1l}'$. Remaining sufficient statistics are the same as at iteration r . Finally, the conditional posterior for Ω is $\Omega \sim W\left(\gamma\Omega + L_r\nu, \sum_{l=1}^{L_r} \Sigma_l^{\star-1} + \Psi_\Omega^{-1}\right)$, where each $\Sigma_l^{\star-1}$ (a member of the ϑ_l^* which are marginalized out elsewhere) is simulated from $W(\Sigma_l^{\star-1}; \nu + m_{rl}, \Omega + D_{rl})$.

References

Antoniak, C. (1974), "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *Annals of Statistics*, 2, 1152 – 1174.

- Baddeley, A., Turner, R., Møller, J., and Hazelton, M. (2005), “Residual analysis for spatial point processes,” *Journal of the Royal Statistical Society, Series B*, 67, 617–666.
- Best, N. G., Ickstadt, K., and Wolpert, R. L. (2000), “Spatial Poisson regression for health and exposure data measured at disparate resolutions,” *Journal of the American Statistical Association*, 95, 1076–1088.
- Cargnoni, C., Müller, P., and West, M. (1997), “Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models,” *Journal of the American Statistical Association*, 92, 640–647.
- Caron, F., Davy, M., and Doucet, A. (2007), “Generalized Pólya urn for time-varying Dirichlet process mixtures,” in *23rd Conference on Uncertainty in Artificial Intelligence*, Vancouver, Canada.
- Carvalho, C. M., Lopes, H. F., Polson, N., and Taddy, M. A. (2009), “Particle learning for general mixtures,” Tech. Rep. 09-02, Duke Department of Statistical Science.
- Carvalho, C. M., Johannes, M., Lopes, H. F., and Polson, N. (2010), “Particle learning and smoothing,” *Statistical Science*, to appear.
- Diggle, P. J. (2003), *Statistical Analysis of Spatial Point Patterns*, Arnold, 2nd edn.
- Griffin, J. E. and Steel, M. F. J. (2009), “Time-dependent stick-breaking processes,” Tech. Rep. 09-05, CRiSM.
- Ishwaran, H. and Zarepour, M. (2000), “Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models,” *Biometrika*, 87, 371–390.
- Kass, R. E. and Raftery, A. E. (1995), “Bayes factors,” *Journal of the American Statistical Association*, 90, 773–795.
- Kelling, G. L. and Coles, C. M. (1998), *Fixing Broken Windows: Restoring Order And Reducing Crime In Our Communities*, Touchstone, New York.
- 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.

- Liang, S., Banerjee, S., and Carlin, B. P. (2009), “Bayesian wombling for spatial point processes,” *Biometrics*, 65, 1243–1253.
- MacEachern, S. N. (2000), “Dependent Dirichlet processes,” Tech. rep., Ohio State University Department of Statistics.
- Manolopoulou, I., Wang, X., Ji, C., Lynch, H. E., Stewart, S., Sempowski, G. D., Alam, S. M., West, M., and Kepler, T. B. (2009), “Statistical analysis of cellular aggregates in immunofluorescence histology,” Tech. Rep. 09-19, Duke Department of Statistical Science.
- McKenzie, E. (1985), “An autoregressive process for beta random variables,” *Management Science*, 31, 988–997.
- Møller, J. and Waagepetersen, R. P. (2004), *Statistical Inference and Simulation for Spatial Point Processes*, Chapman & Hall/CRC.
- Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998), “Log Gaussian Cox processes,” *Scandinavian Journal of Statistics*, 25, 451–452.
- Müller, P., Quintana, F. A., and Rosner, G. (2004), “A method for combining inference across related nonparametric Bayesian models,” *Journal of the Royal Statistical Society, Series B*, 66, 735–749.
- Perman, M., Pitman, J., and Yor, M. (1992), “Size-biased sampling of Poisson point processes and excursions,” *Probability Theory and Related Fields*, 92, 21–39.
- Rodriguez, A. and Dunson, D. B. (2009), “Nonparametric Bayesian models through probit stick-breaking processes,” Tech. Rep. 09-12, UCSC Baskin School of Engineering.
- Taddy, M. and Kottas, A. (2010), “A Bayesian nonparametric approach to inference for quantile regression,” *Journal of Business and Economic Statistics*, to appear.
- Taddy, M. A. (2008), “Bayesian nonparametric analysis of conditional distributions and inference for Poisson point processes,” Ph.D. thesis, UC Santa Cruz.
- Taddy, M. A. and Kottas, A. (2009), “Dirichlet process mixture modelling for marked Poisson processes,” Tech. Rep. 09-31, UCSC Baskin School of Engineering.

- Walker, S. G. (2007), “Sampling the Dirichlet mixture model with slices,” *Communications in Statistics – Simulation and Computation*, 36, 45–54.
- West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, Springer, 2nd edn.