UNIVERSITY OF CALIFORNIA
SANTA CRUZ

BAYESIAN NONPARAMETRIC ANALYSIS OF CONDITIONAL DISTRIBUTIONS AND INFERENCE FOR POISSON POINT PROCESSES

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS AND STOCHASTIC MODELING

by

Matthew Alan Taddy

June 2008

The Dissertation of Matthew Alan Taddy is approved:

Professor Athanasios Kottas, Chair and Primary Advisor

Professor Herbert K. H. Lee, Secondary Advisor

Professor Bruno Sansó

Professor Abel Rodríguez

Lisa C. Sloan
Vice Provost and Dean of Graduate Studies

Table of Contents

Abstract

Bayesian Nonparametric Analysis of Conditional Distributions

and Inference for Poisson Point Processes

by

Matthew Alan Taddy

This thesis provides a suite of flexible and practical nonparametric Bayesian analysis frameworks, together related under a particular approach to Dirichlet process (DP) mixture modeling based on joint density estimation with well chosen kernels and inference through finite stick-breaking approximation to the random mixing measure. Development of a novel nonparametric mean regression estimator serves as an introduction to a general modeling approach for nonparametric analysis of conditional distributions through initial inference about joint probability distributions. Three novel regression modeling frameworks are proposed: quantile regression, hidden Markov switching regression, and regression for survival data. A related approach is adopted in modeling for marked spatial Poisson processes. This class of models is then expanded to a full nonparametric framework for inference about marked or unmarked dynamic spatial Poisson processes which occur at discrete time intervals. This involves the development of a version of the dependent DP as a prior on the space of correlated sets of probability distributions. Posterior simulation methodology is contained throughout and numerous data examples have been provided in illustration.

For Tim, Jane, and Kirsty.

Acknowledgments

I am fortunate to have been guided and assisted by a wonderful group of people throughout my education as a statistician and in production of this dissertation.

When I decided to join the young department of Applied Math and Statistics at UCSC, it was with the hope that I would be able to work closely with many different people. I am happy to confirm that my hopes were realized and I am incredibly lucky to have experienced the collaboration and friendship which I found in AMS. From the moment I first arrived, Bobby Gramacy welcomed me as a partner in research and software development, and from this I have gained a skillset without which the results in this thesis would be much impoverished. Marc Mangel is to thank for equipping me with powerful new ways to think about science and the role of mathematics in science. I have profited greatly from my interaction with all of the statisticians in AMS, but Bruno Sansó in particular has repeatedly offered insight and inspiration that allows me to grow as a Bayesian and I would consider myself fortunate if I am able to emulate his approach to statistics. I am indebted to both Bruno and Abel Rodriguez for serving as readers on my thesis committee, and each of them has contributed valuable advice that much improved this document.

Within the AMS faculty, I was lucky enough to find two advisors who have taught me huge amounts while managing to provide both mentorship and friendship. Herbert Lee is a large part of both my motivation to come to Santa Cruz as well as my ability to find success here. We have worked together from the moment that I

arrived, and he has been a both a role model and a great friend during my immersion into Bayesian statistics. Herbie has worked very hard to involve me in his research and I have gained an invaluable breadth of experience working alongside someone with incredible talent for data analysis. Athanasios Kottas, who has been my advisor for most of the work contained in this thesis, completed my education by enabling the leap to Bayesian nonparametrics. My fundamental approach to model development and inference is deeply influenced by his work, and this finished product would have been impossible without his tireless editing and guidance. Thanasis has led me through the final stages in development of an approach to statistics that will serve as the foundation for an entire career, and I cannot thank him enough.

My achievements at Santa Cruz are rooted in the fundamental knowledge and experience that I gained during my studies at McGill. In particular, David Wolfson, Alain Vandal, and Russ Steel managed to introduce me to the world of statistics while enforcing a level of rigor in my work that has served me well throughout my education. I thank Luis Acevedo Arreguin for cleaning the Cincinnati crime data. My sister, Heather Taddy, edited my discussion of the AIDS clinical trial data to remove complete nonsense. Tracie Tucker is to thank for keeping me organized and out of trouble. And my closest colleagues at Sandia national laboratories, Genetha Gray and Josh Griffin, allowed me to experience the very best of collaborative research and real teamwork.

Finally, during this process, I have often asked for sacrifice by the people closest to me, my family and friends. Everything I have managed to achieve has only been possible through the support of the people I love.

Chapter 1

Introduction

The essential inference procedure of a Bayesian is to, conditional on data observations, update prior belief about the stochastic mechanisms under which the data arise and predict future observables. In parametric Bayesian analysis, the prior beliefs are formally defined as probability distributions for parameters of an assumed model. Nonparametric Bayesian analysis instead dispenses with the parametric model assumption, and prior beliefs are expressed as probability measures assigned directly to the set of probability distributions which could possibly have generated the data.

Contemporary work on Bayesian nonparametrics may be classified into four main categories: theory, modeling, computation, and application. Theoretical work is concentrated on the construction of classes of measures on spaces of random functions and on study of the asymptotic properties of resulting posterior point estimates. Modeling research seeks to develop general frameworks for inference built around the use of a nonparametric prior measure. The computational research is required to develop

efficient posterior simulation algorithms for these models, and the end goal of this whole effort is that the modeling frameworks will be put to use in careful application to challenging data analysis problems. This thesis falls firmly within the modeling category, and should be read with this motivation in mind. Hence, the development of nonparametric prior models and of posterior simulation algorithms was always undertaken as a step towards a particular flexible and practical framework for nonparametric Bayesian inference. And although the presented methodology will be immediately applicable in practical data analysis, the examples contained herein were not intended to provide complete case studies.

There is, of course, a spectrum of Bayesian analysis methodologies which spans the divide between standard parametric analysis (such as mean estimation for a normal distribution) and full nonparametric analysis (with priors induced by stochastic processes defined over the space of all possible distribution functions). For instance, a regression model with a Gaussian process prior for the regression function may be referred to as a nonparametric regression model inasmuch as there is no parametric form assumed for the mean regression line. However, the Gaussian process covariance is specified through use of a parametric covariance family with process stationarity typically assumed. Similarly, finite mixture models supplant the assumption of a single parametric family for the data density with the assumption of a finite mixture of parametric density kernels. The model flexibility increases with the number of mixture components or with modeling for an unknown number of components. And with the specification of a nonparametric prior for the random mixing distribution (thus assigning probabilities

to the set of possible mixing measures for parameters of the density kernel) the analysis is formally nonparametric.

This latter approach, termed Bayesian nonparametric mixture analysis, is the one utilized throughout this thesis. In particular, the focus is placed on density models consisting of mixtures of parametric kernels with a Dirichlet Process (DP) prior assumed for the random mixing distribution. The historical context of this modeling approach is reviewed, with reference to the proposed methodology, throughout this document.

DP mixture priors are huge. They have become the flexible prior *du jour*, and are appearing as the default nonparametric modeling extension in fields as diverse as biometry, econometrics, and machine learning. Much of the rise to prominence has been caused by the same factors which account for increased usage of Bayesian nonparametrics and Bayesian statistics in general; namely, the declining cost of computing power coupled with accessible Markov Chain Monte Carlo (MCMC) simulation algorithms and a proven record for predictive ability. But the appeal of DP priors, in particular, owes much to the elegance and intuitive simplicity of the Pólya urn formulation for the posterior predictive distribution as a mixture of point masses on the observed data and the underlying prior base measure αG_0. Here, $\alpha > 0$ corresponds to prior precision and G_0 is a probability distribution with the same support as the modeled random distribution. Thus for the DP mixture model, in notation analogous to that introduced and used throughout the thesis, with basic prior model

$$f(z) \equiv f(z; G) = \int \mathrm{k}(z; \theta) dG(\theta), \quad G \sim \mathrm{DP}(\alpha, G_0),$$

inference about the posterior expectation for the density function f is possible without posterior realizations of the random mixing distribution G. In application, it does sometimes appear that the Pólya urn (now the posterior predictive distribution for the kernel parameters θ) is regarded as if it *is* the DP. Indeed, a popular view of the DP as a clustering mechanism is rooted in the Pólya urn structure. And in certain cases, such as for the now-standard robust modeling extension of having additive error specified through a DP mixture, posterior predictive inference typically suffices. However, it is inevitable that simply marginalizing over the random mixing measure, a parameter which is essential to the model formulation, will limit possibilities for inference.

The earliest origin of this thesis lies in an investigation of what is lost by not including G in the inferential scheme for a generic DP mixture model for multivariate distributions. The discussion of Section 2.1 outlines why, in general, a reliance solely on DP posterior predictive inference about a joint density for some set of random variables precludes correct estimation of related conditional densities. In response to this, Chapter 2 defines a novel mean regression framework that relies upon truncated realizations of the infinite dimensional G. The simple underlying idea is that full inference (i.e., dealing with G) about multivariate random distributions assigned DP mixture priors allows for any desired inference about related conditional distributions.

This simple insight serves as an introduction to the general modeling approach of nonparametric analysis for conditional distributions through initial inference about joint probability distributions. Chapter 3 contains three novel modeling frameworks that fall within the general regression paradigm: quantile regression in Section 3.1, hidden

Markov model switching regression in Section 3.2, and regression for survival data in Section 3.3. In each case, flexible nonparametric modeling for joint distributions informs the entire inference procedure, and regression modeling for related conditional densities is born of this.

In Chapter 4, we turn to modeling for spatial Poisson point processes. The modeling framework continues in the same spirit as that for regression, and indeed the inference for marked processes in Section 4.1.2 is another example of analysis of conditional distributions based on nonparametric modeling for a joint distribution. But the conditional modeling for marks is only a part of the framework for Bayesian nonparametric inference about Poisson point processes developed in this chapter. We describe and specify, with illustration through data examples, modeling for both marked and unmarked spatial Poisson processes and modeling for multiple related realizations of such processes that are observed over discrete time intervals. This latter dynamic modeling framework involves specification of a novel version of the Dependent Dirichlet Process prior measure.

A characteristic of much of the work contained in this thesis is that inference is required for the random mixing distributions (e.g., G in the simple model above). Thus, all of the modeling frameworks are accompanied by a finite truncation for this random distribution, so as to facilitate posterior sampling. The reason for sampling G in the analysis of conditional distributions (either for regression response or point process marks) has already been discussed. Furthermore, correctly estimated posterior uncertainty intervals, whether around individual density function expectations or about

a spatial region's integrated Poisson intensity, are only available on the basis of posterior sampling for the associated random mixing distributions. In addition, all inference for models based on the temporally related random mixing measures developed in Chapter 4 requires posterior realizations of this set of measures. Sampling for truncated approximations to G also provides the basis for alternative posterior simulation algorithms. Forward-backward recursive sampling for the Markov switching regression model in Section 3.2 and the particle filtering approach to sampling dynamic random measures in Section 4.1.3 each rely upon truncated realizations for the random mixing distribution. Also, the increased efficiency of such algorithms allows for nonparametric analysis of relatively large data sets, such as the 35,000 crime events considered in Section 4.3.2.

Thus, although the model development is always fully nonparametric, almost all of the practical implementation and inference relies upon finite truncations of the infinite dimensional prior model. In one sense, we are presenting very flexible high dimensional prior families which contain fully nonparametric models as limiting cases. However, the work presented herein is more accurately characterized as the development of nonparametric modeling frameworks accompanied by the truncation methodology and guidance required for practical implementation. As mentioned above, this document should be viewed as a thesis on modeling. The primary goal throughout has been to develop modeling frameworks that may serve as the basis for a wide variety of inference, and my hope is that such inference will benefit distinctly from the proposed methodology.

Chapter 2

Bayesian Nonparametric Regression Using Dirichlet Process Mixtures

This chapter introduces our general framework for the analysis of conditional distributions through the example of a novel mean regression estimator. In the Bayesian regression literature, two dominant trends have been to attempt to find increasingly flexible regression function models and to accompany these models with more comprehensive uncertainty quantification. Typically, Bayesian nonparametric modeling focuses on either the regression function or the error distribution. Müller and Quintana (2004) provide an overview of the respective methodologies. The starting point for our approach is a novel Bayesian nonparametric extension of implied conditional regression, wherein Dirichlet process mixtures (Ferguson, 1973; Antoniak, 1974) are used to model the joint distribution of response and covariates, from which full inference is obtained for the desired conditional distribution for response given covariates. Both the response

distribution and, implicitly, the regression function are modeled nonparametrically, thus providing a flexible framework for the general regression problem.

2.1 Multivariate Normal Mixture Model

We introduce in this section the DP mixture curve fitting approach to nonparametric regression, beginning with a canonical Normal mixture model built around the assumption of real-valued continuous response and covariates. The data, denoted by $\mathcal{D}$, consist of realizations of covariates $\mathbf{X} = (X_1, ..., X_{d_\mathbf{x}})$ and response Y, occurring according to some underlying random joint distribution, and interest lies in the conditional distribution for Y given $\mathbf{X}$. Although we refer throughout this chapter to the case of a univariate response, there is no such restriction in the methodology. Indeed, Section 3.3 includes an example of multivariate regression for survival data.

One possible approach for nonparametric regression is to estimate the joint and marginal densities, $f(\mathbf{x},\, y)$ and $f(\mathbf{x})$, and then obtain inference from the conditional density $f(y \mid \mathbf{x}) = f(\mathbf{x}, y)/f(\mathbf{x})$. This *implied conditional regression* approach dates back at least to Nadaraya (1964) and Watson (1964), where estimation was based on kernel smoothing methods. In the machine learning literature, members of the general class of such techniques are referred to as generative estimation methods (see Bishop and Lasserre, 2007, for discussion in the context of classification). An attractive implementation arises from the use of Bayesian nonparametric density estimation. We use DP mixtures of multivariate normal distributions to model the joint distribution of

the response and covariates (as in, e.g., Müller et al., 1996). The multivariate normal distribution is a natural choice for the mixture kernel in the presence of (real-valued) continuous variables, due to both its flexibility and the relative ease of application. Finite mixtures of multivariate normals have proven successful in many problems where the number of mixture components is assumed to be known, as well as in situations where it is estimated from the data (see, e.g., Lopes et al., 2003; Dellaportas and Papageorgiou, 2006). These models can be extended by placing a DP prior on the random mixing measure for the multivariate normal kernels, an approach that yields both methodological and practical advantages, due to a more general modeling framework and posterior simulation techniques that are typically less complex than for finite mixture models with an unknown number of components.

The DP was developed by Ferguson (1973) as a prior probability model for random distributions (equivalently, distribution functions) G. DP models have enjoyed considerable popularity due to the ready availability of posterior simulation techniques, the analytic tractability of almost surely discreet realized probability functions, as well as the theoretical elegance of the model formulations. A $\mathrm{DP}(\alpha, G_0)$ prior for G is defined in terms of two parameters, a parametric base distribution G_0 (the mean of the process) and a positive scalar parameter α, which can be interpreted as a precision parameter; larger values of α result in realizations G that are closer to G_0. We will write $G \sim \mathrm{DP}(\alpha, G_0)$ to indicate that a DP prior is used for the random distribution G. In fact, DP-based modeling typically utilizes mixtures of DPs (Antoniak, 1974), i.e., a more general version of the DP prior that involves hyperpriors for α and/or the

parameters of G_0. The most commonly used DP definition is its constructive definition (Sethuraman, 1994), which characterizes DP realizations as countable mixtures of point masses (and thus as random discrete distributions). Specifically, a random distribution G generated from $\mathrm{DP}(\alpha, G_0)$ is (almost surely) of the form

$$G(\cdot) = \sum_{\ell=1}^{\infty} p_\ell \, \delta_{\vartheta_\ell}(\cdot) \tag{2.1}$$

where $\delta_\vartheta(\cdot)$ denotes a point mass at ϑ. The locations of the point masses, ϑ_ℓ, are i.i.d. realizations from G_0; the corresponding weights, p_ℓ, arise from a *stick-breaking* mechanism based on i.i.d. draws $\{v_k : k = 1, 2, ...\}$ from a $\mathrm{Be}(1, \alpha)$ distribution (Here, $\beta(a, b)$ denotes the Beta distribution with mean $a/(a + b)$). In particular, $p_1 = v_1$, and, for each $\ell = 2, 3, ..., \, p_\ell = v_\ell \prod_{k=1}^{\ell-1}(1 - v_k)$. Moreover, the sequences $\{\vartheta_\ell, \ell = 1, 2, \ldots\}$ and $\{v_k : k = 1, 2, ...\}$ are independent.

In the $d = d_{\mathbf{x}} + 1$ dimensional setting, with data $\mathcal{D} = \{\mathbf{z}_i = (x_i^1, \ldots, x_i^{d_{\mathbf{x}}}, y_i)$: $i = 1, \ldots, n\}$, the location-scale normal DP mixture model can be described as follows:

$$\mathbf{z}_i \mid G \overset{ind}{\sim} f(\mathbf{z}_i; G) = \int \mathrm{N}(\mathbf{z}_i; \mu, \Sigma) dG(\mu, \Sigma), \quad G \mid \alpha, \psi \sim \mathcal{D}P\left(\alpha, G_0(\psi)\right), \tag{2.2}$$

with the DP centering distribution given by $G_0(\mu, \Sigma; \psi) = \mathrm{N}(\mu; m, V)\mathrm{W}_\nu(\Sigma^{-1}; S^{-1})$, where $\psi = (m, V, S)$, and $W_v(\cdot; M)$ denotes the Wishart distribution with v degrees of freedom and expectation vM. We place hyperpriors on ψ and the DP precision parameter α. In particular, we take $\pi(\psi) = \mathrm{N}\left(m; a_m, B_m\right) \, \mathrm{W}_{a_V}(V^{-1}; B_V^{-1}) \, \mathrm{W}_{a_S}(S; B_S)$ and $\pi(\alpha) = \mathrm{Ga}(\alpha; a_\alpha, b_\alpha)$, where $\mathrm{Ga}(a, b)$ denotes the gamma distribution with expectation a/b. The values for hyperparameters of ψ are usually chosen only to scale the mixture to the data, and prior specification is discussed in Section 2.2 below.

This mixture specification provides also the prior model for the marginal density for $\mathbf{X}$, $f(\mathbf{x}; G) = \int \mathrm{N}(\mathbf{x}; \mu^{\mathbf{x}}, \Sigma^{\mathbf{xx}}) dG(\mu, \Sigma)$, after the mean vector and covariance matrix of the normal kernel have been partitioned. In particular, μ comprises $(d_{\mathbf{x}} \times 1)$ vector $\mu^{\mathbf{x}}$ and scalar μ^{y}, and Σ is a square block matrix with diagonal elements given by $(d_{\mathbf{x}} \times d_{\mathbf{x}})$ covariance matrix $\Sigma^{\mathbf{xx}}$ and scalar variance Σ^{yy}, and above and below diagonal vectors $\Sigma^{\mathbf{x}y}$, and $\Sigma^{y\mathbf{x}}$, respectively.

A set of latent parameters $\boldsymbol{\theta} = \{\theta_i = (\mu_i, \Sigma_i) : i = 1, \ldots, n\}$ are introduced to break the mixture in model (2.2) such that

$$\begin{aligned} \mathbf{z}_i | \mu_i, \Sigma_i &\overset{ind}{\sim} \mathrm{N}_{L+1}(\mathbf{z}_i; \mu_i, \Sigma_i), \;\; i = 1, ..., n \\ (\mu_i, \Sigma_i) | G &\overset{iid}{\sim} G, \;\; i = 1, ..., n \\ G \mid \alpha, \psi &\sim \mathrm{DP}(\alpha, G_0(\psi)). \end{aligned} \tag{2.3}$$

Most typically, DP mixture models are fit by marginalizing the random mixing distribution G over its DP prior and using the resulting Pólya urn representation for the latent mixing parameters $\boldsymbol{\theta}$ (Blackwell and MacQueen, 1973) in sampling from the posterior. Without posterior realizations of G, inference for the densities $f(\mathbf{x}, y; G)$ and $f(\mathbf{x}; G)$ is available only in the form of point estimates through the posterior predictive density for $\mathbf{x}$ and y, $P_0(\mathbf{x}, y \mid \mathcal{D})$, which can be estimated using only the posterior distribution $\Pr(\boldsymbol{\theta}, \alpha, \psi \mid \mathcal{D})$ for $\{\boldsymbol{\theta}, \alpha, \psi\}$. Specifically,

$$P_0(\mathbf{x}, y \mid \mathcal{D}) = \int P_0(\mathbf{x}, y \mid \boldsymbol{\theta}, \alpha, \psi) d \Pr(\boldsymbol{\theta}, \alpha, \psi \mid \mathcal{D}), \tag{2.4}$$

where the predictive density conditional on model parameters has a convenient Pólya

urn structure,

$$P_0(\mathbf{x}, y \mid \boldsymbol{\theta}, \alpha, \psi) = \frac{\alpha}{\alpha + n} \int N(\mathbf{x}, y; \theta_0) dG_0(\theta_0; \psi) + \sum_{j=1}^{n^\star} \frac{H_j^\star}{\alpha + n} N(\mathbf{x}, y; \theta_j^\star). \qquad (2.5)$$

Here, $\theta_0 = (\mu_0, \Sigma_0)$, $\boldsymbol{\theta}^\star = \{\theta_j^\star = (\mu_j^\star, \Sigma_j^\star) : j = 1, \ldots, n^\star\}$ is the set of $n^\star$ distinct parameter values in $\boldsymbol{\theta}$, and $H_j^\star$ is the number of data observations allocated to unique component $\theta_j^\star$.

Müller et al. (1996) developed a DP mixture implied conditional regression method based on multivariate normal mixtures as in model (2.2). However, their curve fitting approach relies on a point estimate for the conditional density,

$$\int \frac{P_0(\mathbf{x}, y \mid \boldsymbol{\theta}, \alpha, \psi)}{P_0(\mathbf{x} \mid \boldsymbol{\theta}, \alpha, \psi)} d\Pr(\boldsymbol{\theta}, \alpha, \psi \mid \mathcal{D}), \qquad (2.6)$$

which is not $\mathbb{E}[f(y \mid \mathbf{x}; G) \mid \mathcal{D}]$. That is, (2.6) is not the posterior expectation for the random conditional density $f(y \mid \mathbf{x}; G) = f(\mathbf{x}, y; G)/f(\mathbf{x}; G)$, which would be the natural point estimate for the regression function at any specified combination of values $(\mathbf{x}, y)$. Note that $P_0(\mathbf{x}, y \mid \boldsymbol{\theta}, \alpha, \psi)$ in (2.5) arises from

$$\begin{aligned} \int \mathrm{N}(\mathbf{x}, y; \theta_0) \left[\int dG(\theta_0) d\mathcal{F}(G \mid \alpha^\star, G_0^\star) \right] &= \int \left[\int \mathrm{N}(\mathbf{x}, y; \theta_0) dG(\theta_0) \right] d\mathcal{F}(G \mid \alpha^\star, G_0^\star) \\ &= \mathbb{E}[f(\mathbf{x}, y; G) \mid \boldsymbol{\theta}, \alpha, \psi], \qquad (2.7) \end{aligned}$$

where, in general, $\mathcal{F}(G \mid \alpha, G_0)$ denotes the distribution over random distribution functions G implied by a $\mathrm{DP}(\alpha, G_0)$. Here, the precision parameter $\tilde{\alpha} = \alpha + n$, and the centering distribution $\tilde{G}_0(\cdot) \equiv \tilde{G}_0(\cdot \mid \boldsymbol{\theta}, \alpha, \psi) = (\alpha + n)^{-1} [\alpha dG_0(\cdot; \psi) + \sum_{i=1}^n \delta_{\theta_i}(\cdot)]$, where δ_u denotes a point mass at u (Antoniak, 1974). Therefore, we also have $P_0(\mathbf{x} \mid \mathcal{D}) = \mathbb{E}[f(\mathbf{x}; G) \mid \mathcal{D}]$, and thus joint and marginal posterior predictive densities $P_0(\mathbf{x}, y \mid \mathcal{D})$

and $P_0(\mathbf{x} \mid \mathcal{D})$ provide point estimates (posterior expectations) for $f(\mathbf{x}, y; G)$ and $f(\mathbf{x}; G)$, respectively. Hence, the regression estimator proposed by Müller et al. (1996), based on (2.6), as well as that proposed in the more recent work of Rodriguez et al. (2008), based upon $P_0(\mathbf{x}, y \mid \mathcal{D})$ / $P_0(\mathbf{x} \mid \mathcal{D})$, are approximating the expectation of a ratio with a ratio of expectations.

Indeed, through calculations similar to that of (2.7), it will be true in general that the marginal posterior predictive density estimate will lead to incorrect estimation if used as a basis for conditional inference. Any inference for functionals of a conditional density estimate based on (2.6) (or even based on the true posterior predictive conditional density) cannot be formally related to the posterior expectation of the random conditional distribution. Evidently, there is much to be gained from taking the extra step to obtain the posterior distribution of G. Primarily, this allows for direct draws from the posterior of the conditional density $f(y \mid \mathbf{x}; G)$, which will provide the exact point estimate $\mathbb{E}\left[f(y \mid \mathbf{x}; G) \mid \mathcal{D}\right]$ and, most importantly, quantification of posterior uncertainty about the implied conditional density.

Posterior sampling for the infinite dimensional parameter G is possible through the constructive definition in (2.1), wherein a realization G from the $\mathrm{DP}(\alpha, G_0(\psi))$ is almost surely a discrete distribution with a countable number of possible values drawn i.i.d. from $G_0(\psi)$, and corresponding weights that are built from i.i.d. $\mathrm{Be}(1, \alpha)$ variables through stick-breaking. Hence, a truncation approximation to G can be defined as

follows,

$$G^L(\cdot) \;=\; \sum_{l=1}^{L} p_l \delta_{\vartheta_l}(\cdot) \quad \text{with} \quad \mathbf{p}, \boldsymbol{\vartheta} \sim \mathcal{P}_L(\mathbf{p} \mid 1, \alpha) \prod_{l=1}^{L} dG_0(\vartheta_l; \psi), \tag{2.8}$$

where $\mathbf{p} = (p_1, \ldots, p_L)$, $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_L)$, and the finite stick-breaking prior $\mathcal{P}_L(\mathbf{p} \mid a, b)$ is defined constructively by

$$\begin{aligned} v_1, \ldots, v_{L-1} &\overset{iid}{\sim} \text{Be}(a, b), \; v_L = 1; \\ p_1 = v_1, \text{ and for } l = 2, \ldots, L: \; p_l &= v_l \prod_{s=1}^{l-1} (1 - v_s), \end{aligned} \tag{2.9}$$

(see, e.g., Ishwaran and James, 2001).

The truncated distribution G^L may be used only for the conditional posterior of G given $\{\boldsymbol{\theta}, \alpha, \psi\}$ after the Pólya urn marginalization, in which case, G^L is a truncation approximation to a DP with parameters $\tilde{\alpha}$ and $\tilde{G}_0$ given above. Alternatively, one can utilize the truncated DP, defined by (2.8) and (2.10), throughout and draw parameters conditional on this truncation (as in the Blocked Gibbs posterior simulation method of Section 3.2.3). In either case, given a posterior draw of the truncated DP parameters, $G^L = \{(p_l, (\mu_l, \Sigma_l)) : l = 1, ..., L\}$, posterior realizations for the joint and marginal densities are readily available through $f(\mathbf{x}, y; G^L) = \sum_{l=1}^{L} p_l \text{N}(\mathbf{x}, y; \mu_l, \Sigma_l)$ and $f(\mathbf{x}; G^L) = \sum_{l=1}^{L} p_l \text{N}(\mathbf{x}; \mu_l^{\mathbf{x}}, \Sigma_l^{\mathbf{xx}})$. Then, the posterior realization for the conditional response density at any value $(\mathbf{x}, y)$ is given by

$$f(y \mid \mathbf{x}; G^L) = \frac{f(\mathbf{x}, y; G^L)}{f(\mathbf{x}; G^L)} = \frac{\sum_{l=1}^{L} p_l \text{N}(\mathbf{x}, y; \mu_l, \Sigma_l)}{\sum_{l=1}^{L} p_l \text{N}(\mathbf{x}; \mu_l^{\mathbf{x}}, \Sigma_l^{\mathbf{xx}})}. \tag{2.10}$$

It is thus possible to obtain the posterior of the conditional density for response given covariates over a grid in $\mathbf{x}$ and y, which yields full inference for the implied conditional

regression relationship, for example, through posterior point and interval estimates.

Under the modeling framework defined by (2.2) and (2.3), the discreteness of G, induced by its DP prior, is a key feature as it enables flexible shapes for the joint distribution of the response and covariates through data-driven *clustering* of the mixing parameters (μ_i, Σ_i). Note, however, that we employ the DP mixture setting to model random distributions (as it was originally intended) and not as a clustering mechanism (as used, to some extent, in the more recent literature). In this regard, although it may be of methodological interest to study some of the recent extensions of the DP (e.g. Ishwaran and James, 2001; Lijoi et al., 2005) as alternative priors for G, these prior models would, arguably, not lead to practical advantages over the DP with regard to the resulting inference.

We note that the structure of conditional moments for the normal mixture kernel enables posterior sampling of the conditional mean regression function without having to compute the conditional density. Specifically,

$$\begin{aligned} \mathbb{E}\left[Y \mid \mathbf{x}; G^L\right] &= \frac{1}{f(\mathbf{x}; G^L)} \sum_{l=1}^{L} p_l \int y \mathrm{N}\left(y, \mathbf{x}; \mu_l, \Sigma_l\right) dy \qquad (2.11) \\ &= \frac{1}{f(\mathbf{x}; G^L)} \sum_{l=1}^{L} p_l \mathrm{N}(\mathbf{x}; \mu_l^{\mathbf{x}}, \Sigma_l^{\mathbf{xx}}) \left[\mu_l^{y} + \Sigma_l^{y\mathbf{x}} (\Sigma_l^{\mathbf{xx}})^{-1} (\mathbf{x} - \mu_l^{\mathbf{x}})\right], \end{aligned}$$

which, evaluated over a grid in $\mathbf{x}$, yields posterior realizations of the conditional mean function. Thus, we have an unbiased version of the regression estimator first proposed by Müller et al. (1996). Since it is the inverse covariance Σ_l^{-1} that is drawn during MCMC, a rapid implementation of the above mean regression is facilitated by noting that Σ_l^{-1} has diagonal elements $A = \Sigma^{\mathbf{xx}-1} + \Sigma^{\mathbf{xx}-1}\Sigma^{\mathbf{x}y} E \Sigma^{\mathbf{x}y}\Sigma^{\mathbf{xx}-1}$ and $E = (\Sigma^{yy} - \Sigma^{y\mathbf{x}}\Sigma^{\mathbf{xx}-1}\Sigma^{\mathbf{x}y})^{-1}$,

with above diagonal $B = -\Sigma^{\mathbf{xx}-1}\Sigma^{\mathbf{x}y}E$ and B' below, such that $\Sigma^{\mathbf{x}y}\Sigma^{\mathbf{xx}-1} = -E^{-1}B'$ and $\Sigma^{\mathbf{xx}-1} = A - BE^{-1}B'$.

Note that Rodriguez et al. (2008) (details contained in Rodriguez, 2007) proved posterior consistency for the conditional density estimator $P_0(\mathbf{x}, y \mid \mathcal{D}) \,/\, P_0(\mathbf{x} \mid \mathcal{D})$, with $P_0(\,\cdot\,|\mathcal{D})$ the posterior predictive density of (2.4), which is the true posterior predictive conditional density. As a corollary, they show consistency for the associated mean regression estimators. This would indicate that, as the sample size tends to infinity, point estimators based on either the true posterior predictive conditional density or the Müller et al. (1996) approach will converge to the expectation of conditional density and mean regression functions sampled from their full posterior as in (2.10) and (2.11) respectively. This, however, does not change the fact that neither posterior predictive conditional density estimator represents the posterior expectation of the conditional density for finite sample sizes and that it is impossible to accurately account for posterior uncertainty about these point estimates without sampling G^L.

A simple illustration of this mean regression estimator is shown in Figure 2.1, where the DP mixture of normals model with the default prior specification of Section 2.2 is fit to a simple nonlinear function with additive normal noise. Further use of the estimator is illustrated in Section 3.2.5 in the context of switching regression. However, through posterior sampling of the truncated G^L, a wide range of inference about the conditional distribution is possible, and modeling throughout the remainder of this thesis is proposed in the spirit of taking full advantage of this complete inference. In Section 3.1, we describe a quantile regression framework which arises naturally from the

DP implied regression modeling approach, and in Section 3.2 the complete G^L measures are used to estimate hidden Markov model states. Before this, however, the following sections describe prior specification and a first approach to posterior simulation.

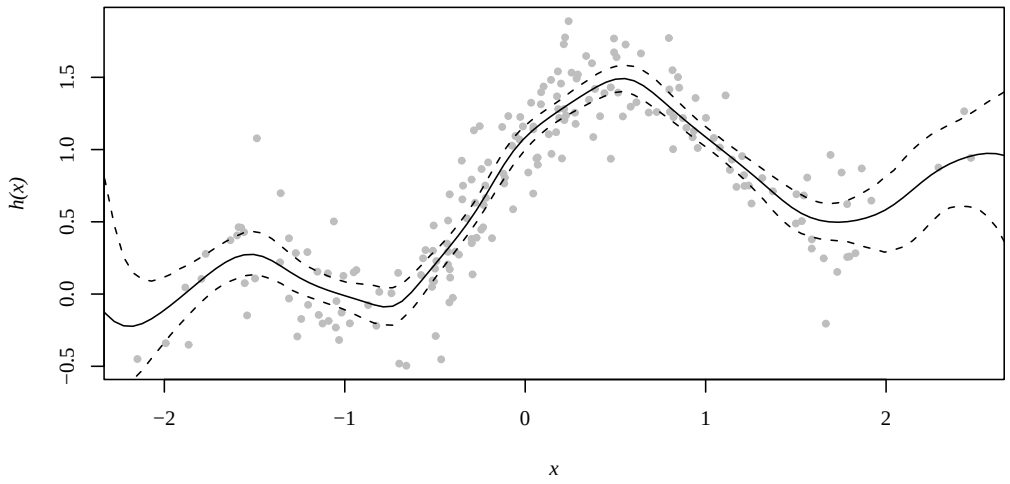


Figure 2.1: Posterior mean (solid line) and 90% interval (dashed lines) for the mean regression estimate $\mathbb{E}[h|x;G]$, where $h(x) = 0.4x + 0.5\sin(2.7x) + 1.1(1+x^2)^{-1} +$ $\mathrm{N}(0,.25)$ (from Neal, 1997). The 200 covariate locations were drawn from a standard normal.

2.2 Prior Specification

Here, we discuss the choice of hyperpriors for the DP mixture model of Section 2.1. The general approach to hyperprior specification for DP mixture models is influenced by a guiding principle holding that the base measure for kernel parameters should be appropriate for a simplified model with a single kernel serving as the density for all observations. This approach ensures that the hyperprior is appropriately diffuse

and is scaled to the data, while requiring only a small amount of prior information. In particular, only rough prior guesses at the center of the response and covariate variables, say, e_y and e_{x_j}, $j = 1, ..., d_x$, as well as at their corresponding ranges, say, r_y and r_{x_j}, $j = 1, ..., d_x$. Let $\mathbf{e} = (e_y, e_{x_1}, ..., e_{x_{d_x}})$ and denote by R the $(d_x + 1) \times (d_x + 1)$ diagonal matrix with diagonal elements $(r_y/4)^2$ and $(r_{x_j}/4)^2$, $j = 1, ..., d_x$, which are prior estimates for the variability of the response and covariates. For a default specification we consider a single component in the mixture, $\mathrm{N}_{d_x+1}(\cdot\ ;\ \mu, \Sigma)$, i.e., the limiting case of model (2.3) with $\alpha \to 0^+$. Therefore, we effectively seek to roughly center and scale the mixture model, using prior information that identifies the subset of $\mathbb{R}^{d_x+1}$ where the data are expected to be supported. Next, based on the form of G_0 and the hyperpriors for its parameters ψ, we can obtain marginal prior moments for μ, i.e., $\mathrm{E}(\mu) = a_{\mathbf{m}}$, and $\mathrm{Cov}(\mu) = (a_V - d_x - 2)^{-1} B_V + B_{\mathbf{m}}$, which are matched with $\mathbf{e}$ and R. Specifically, we take $a_{\mathbf{m}} = \mathbf{e}$, and, using a variance inflation factor of 2, set $B_{\mathbf{m}} = R$ and $(a_V - d_x - 2)^{-1} B_V = R$. We use R to specify also the prior for S through $R = \mathrm{E}(\Sigma) = (\nu - d_x - 2)^{-1} a_S B_S$. Finally, ν, a_V, and a_S are chosen to scale appropriately the hyperpriors, e.g., note that smaller values of $(\nu - d_x - 2)^{-1} a_S$ yield more dispersed priors for S, and that $a_V = d_x + 3$ is the (integer) value that yields the largest possible dispersion while ensuring finite prior expectation for V. For the data analysis presented in Section 3.1.2, we used $\nu = a_V = a_S = 2(d_x + 2)$; we have also empirically observed this choice to work well for other data sets that we have studied with model (2.3).

Regarding the prior choice for the DP precision α, guidelines are available based on the role this parameter plays with regard to the number of distinct components

in the DP mixture model. Note that, marginalizing G over its DP prior, the second and third stages of model (2.3) collapse into a joint prior distribution for the mixing parameters $\boldsymbol{\theta} = \{(\mu_i, \Sigma_i) : i = 1, ..., n\}$, which arises according to a particular Pólya urn scheme. Specifically, as shown by Blackwell and MacQueen (1973), conditional on the DP hyperparameters,

$$p(\boldsymbol{\theta} \mid \alpha, \psi) = g_0(\mu_1, \Sigma_1; \psi) \prod_{i=2}^{n} \left\{ \frac{\alpha}{\alpha + i - 1} g_0(\mu_i, \Sigma_i; \psi) + \frac{1}{\alpha + i - 1} \sum_{\ell=1}^{i-1} \delta_{(\mu_\ell, \Sigma_\ell)}(\mu_i, \Sigma_i) \right\} \tag{2.12}$$

where g_0 is the density of G_0. This expression indicates the DP-induced clustering of the mixing parameters. In particular, $\boldsymbol{\theta}$ is partitioned into $n^*(\leq n)$ distinct components, where the prior distribution for n^* is controlled by α (see Antoniak, 1974; Escobar and West, 1995, for example). In practice, larger values of α yield higher prior probabilities for larger n^*. For instance, under a $\text{Ga}(a_\alpha, b_\alpha)$ prior for α (with mean a_α/b_α), a useful approximation, for moderately large n, to the prior expectation for n^* is given by $(a_\alpha/b_\alpha)\log\{1 + (nb_\alpha/a_\alpha)\}$.

2.3 Incorporating Categorical Covariates

Here, we briefly discuss possible extensions of the modeling framework of Section 2.1 to incorporate both continuous covariates, $\mathbf{x}_c$, and categorical covariates, $\mathbf{x}_d$, where $\mathbf{x} = (\mathbf{x}_c, \mathbf{x}_d)$. Section 3.1.4 introduces in detail one such model in the context of quantile regression, accompanied by discussion of posterior simulation methodology. In addition, the data example of Section 3.3.2 involves regression for survival data with

bivariate response in the presence of a single binary covariate.

A natural extension of the DP mixture model in (2.2) and (2.3) involves replacing the multivariate normal distribution with a mixed continuous/discrete specification for the mixture kernel $k(y, \mathbf{x}_c, \mathbf{x}_d; \theta)$. One possible specification emerges from independent components for $(y, \mathbf{x}_c)$ and $\mathbf{x}_d$. The former can be a multivariate normal distribution, as in Section 2.1, and the latter would be assigned an appropriate multivariate discrete distribution. In its simplest form, this discrete distribution would comprise independent components for the individual elements of $\mathbf{x}_d$. More generally, $k(y, \mathbf{x}_c, \mathbf{x}_d; \theta)$ can be built from a conditional distribution for either the categorical or continuous part given the other variables. Dropping the kernel parameters from the notation, in the former case, $k(y, \mathbf{x}_c, \mathbf{x}_d) = \Pr(\mathbf{x}_d \mid y, \mathbf{x}_c) k(y, \mathbf{x}_c)$, where, for example, with a single binary covariate x_d, a (linear) logistic form could be used for $\Pr(x_d = 1 \mid y, \mathbf{x}_c)$. The latter setting will perhaps be more appropriate given the direction of conditioning involving the response variable. In this case, we could have $k(y, \mathbf{x}_c, \mathbf{x}_d) = k(y, \mathbf{x}_c \mid \mathbf{x}_d) \Pr(\mathbf{x}_d)$, and use a multivariate normal density for $k(y, \mathbf{x}_c \mid \mathbf{x}_d)$ with parameters that are functions of $\mathbf{x}_d$. A simpler formulation would be $k(y, \mathbf{x}_c, \mathbf{x}_d) = k(y \mid \mathbf{x}_c, \mathbf{x}_d) k(\mathbf{x}_c) \Pr(\mathbf{x}_d)$, using, say, a normal density for $k(y \mid \mathbf{x}_c, \mathbf{x}_d)$ with mean that is a function of $\mathbf{x}_c$ and $\mathbf{x}_d$.

2.4 Posterior Simulation

We describe here an approach to MCMC sampling from the posterior of model (2.3) with G marginalized over its DP prior. As discussed in Section 2.1, this marginal-

ization yields a model with a finite-dimensional parameter vector consisting of the mixing parameters $\boldsymbol{\theta} = \{(\mu_i, \Sigma_i) : i = 1, ..., n\}$ and the DP hyperparameters α and ψ. An alternative "blocked Gibbs" method for posterior simulation involves direct approximation of G in model (2.3), using the constructive definition of its DP(α, G_0) prior, and then application of an MCMC technique for the induced discrete mixture model (see, e.g. Ishwaran and James, 2001). Although this alternative approach is preferable in many situations, we focus now on the more common Pólya urn methodology and save the blocked Gibbs for Section 3.2.3, where it will be introduced in the context of switching regression.

We update each (μ_i, Σ_i) using algorithm 5 from Neal (2000), which is based on Metropolis-Hastings steps with proposal distribution given by the prior full conditional of (μ_i, Σ_i) implied by (2.12). Updating all the (μ_i, Σ_i), $i = 1, ..., n$, generates a posterior realization for the partition of $\boldsymbol{\theta}$ comprising $n^\star$ distinct components $(\mu_j^\star, \Sigma_j^\star)$, $j = 1, ..., n^\star$. The $(\mu_j^\star, \Sigma_j^\star)$, along with configuration indicators $\mathbf{k} = (k_1, ..., k_n)$ defined such that $k_i = j$ if and only if $(\mu_i, \Sigma_i) = (\mu_j^\star, \Sigma_j^\star)$, determine $\boldsymbol{\theta}$. Hence, an equivalent representation for $\boldsymbol{\theta}$ is given by $(n^\star, \{(\mu_j^\star, \Sigma_j^\star) : j = 1, ..., n^\star\}, \mathbf{k})$. The Metropolis-Hastings approach to update the (μ_i, Σ_i) can potentially lead to poor mixing. However, it is straightforward to implement and, combined with the technique from Bush and MacEachern (1996) to resample the $(\mu_j^\star, \Sigma_j^\star)$, yields an efficient MCMC method. For each $j = 1, ..., n^\star$, the posterior full conditional for $(\mu_j^\star, \Sigma_j^\star)$ is proportional to $g_0(\mu_j^\star, \Sigma_j^\star; \psi) \prod_{\{i:k_i=j\}} \mathrm{N}_{d_x+1}(\mathbf{z}_i; \mu_j^\star, \Sigma_j^\star)$, and is sampled by drawing from the full conditionals for $\mu_j^\star$ and $\Sigma_j^\star$. The former is $(d_x + 1)$-variate normal with mean vector

$(V^{-1}+n_j^\star\Sigma_j^{\star-1})^{-1}(V^{-1}\mathbf{m}+\Sigma_j^{\star-1}\sum_{\{i:k_i=j\}}\mathbf{z}_i)$ and covariance matrix $(V^{-1}+n_j^\star\Sigma_j^{\star-1})^{-1}$, where $n_j^\star=|\{i:k_i=j\}|$. The latter is inverse Wishart with scalar parameter $\nu+n_j^\star$ and matrix parameter $S+\sum_{\{i:k_i=j\}}(\mathbf{z}_i-\mu_j^\star)(\mathbf{z}_i-\mu_j^\star)'$.

Regarding the DP hyperparameters, we update α using the auxiliary variable method from Escobar and West (1995). The posterior full conditional for $\mathbf{m}$ is (d_x+1)-variate normal with mean vector $(B_\mathbf{m}^{-1}+n^\star V^{-1})^{-1}(B_\mathbf{m}^{-1}a_\mathbf{m}+V^{-1}\sum_{j=1}^{n^\star}\mu_j^\star)$ and covariance matrix $(B_\mathbf{m}^{-1}+n^\star V^{-1})^{-1}$. The full conditional for V is inverse Wishart with scalar parameter $a_V+n^\star$ and matrix parameter $B_V+\sum_{j=1}^{n^\star}(\mu_j^\star-\mathbf{m})(\mu_j^\star-\mathbf{m})'$. Finally, the full conditional for S is given by a Wishart distribution with scalar parameter $a_S+\nu n^\star$ and matrix parameter $(B_S^{-1}+\sum_{j=1}^{n^\star}\Sigma_j^{\star-1})^{-1}$.

Next, note that, based on Antoniak (1974), the full posterior of model (2.3) is given by

$$p(G,\boldsymbol{\theta},\alpha,\psi|\text{data})=p(G|\boldsymbol{\theta},\alpha,\psi)p(\boldsymbol{\theta},\alpha,\psi|\text{data}). \tag{2.13}$$

Here, the distribution for $G|\boldsymbol{\theta},\alpha,\psi$ corresponds to a DP with precision parameter $\alpha+n$ and mean $\tilde{G}_0(\cdot\,;\,\boldsymbol{\theta},\alpha,\psi)$, which is a mixed distribution with point masses $n_j^\star(\alpha+n)^{-1}$ at $(\mu_j^\star,\Sigma_j^\star)$, $j=1,...,n^\star$, and continuous mass $\alpha(\alpha+n)^{-1}$ on $G_0(\psi)$.

Hence, we can draw from the full posterior in (2.13) by augmenting each posterior sample from $p(\boldsymbol{\theta},\alpha,\psi|\text{data})$ with a draw from $p(G|\boldsymbol{\theta},\alpha,\psi)$. The latter requires simulation from the DP with parameters given above, which we implement using the DP constructive definition (discussed in Section 2.1) with a truncation approximation (Gelfand and Kottas, 2002; Kottas, 2006). Therefore, this approach yields realizations of $\{G^L,\boldsymbol{\theta},\alpha,\psi\}$ from the full posterior (2.13). Each posterior realization G^L is

a discrete distribution with point masses at $\vartheta_l = (\mu_l, \Sigma_l)$, $l = 1, ..., L$, drawn i.i.d. from $\tilde{G}_0(\cdot\ ;\ \boldsymbol{\theta}, \alpha, \psi)$, and corresponding weights p_l, $l = 1, ..., L$, generated using the stick-breaking construction based on i.i.d. Be$(1, \alpha)$ draws, and normalized so that $\sum_{l=1}^{L} p_l = 1$. Here, L is the number of terms used in the truncation series approximation to the countable series representation for the DP. In general, L may depend on the particular posterior realization, and the approximation can be specified up to any desired accuracy (see Kottas, 2006, for a specific rule to choose L).

For any specific combination of response and covariate values, say, $(y_0,\ \mathbf{x}_0)$,

$$\begin{aligned} f(y_0, \mathbf{x}_0; G^L) &= \int \mathrm{N}_{d_x+1}(y_0, \mathbf{x}_0; \mu, \Sigma) \mathrm{d}G^L(\mu, \Sigma) \\ &= \sum_{l=1}^{L} p_l \mathrm{N}_{d_x+1}(y_0, \mathbf{x}_0; \mu_l, \Sigma_l) \end{aligned}$$

is a realization from the posterior of the random mixture density $f(y, \mathbf{x}; G^L)$ in (2.2) at point $(y, \mathbf{x}) = (y_0,\ \mathbf{x}_0)$. Analogously, we can obtain the draw from the posterior of the marginal density $f(\mathbf{x}; G^L)$ at point $\mathbf{x} = \mathbf{x}_0$ by computing $f(\mathbf{x}_0; G^L) = \int \mathrm{N}_{d_x}(\ \mathbf{x}_0;\ \mu^{\mathbf{x}},\ \Sigma^{\mathbf{xx}})\ \mathrm{d}G^L(\mu, \Sigma)$. Thus, we obtain $f(y_0 \mid \mathbf{x}_0; G^L) = f(y_0, \mathbf{x}_0; G^L)\ /\ f(\mathbf{x}_0; G^L)$, which is a realization from the posterior of the conditional density $f(y \mid \mathbf{x}; G^L)$, at point $(y, \mathbf{x}) = (y_0, \mathbf{x}_0)$.

Chapter 3

Novel Nonparametric Inference for Regression

This chapter contains three new nonparametric regression frameworks, each taking advantage of full inference about a joint density function, through posterior sampling of truncated random distributions, as a basis for inference about related conditional probability density functions and functionals thereof. Section 3.1 expands on the models proposed in Chapter 2 and develops a comprehensive approach to quantile regression. The methodology is introduced with reference to the canonical model of DP mixtures of multivariate normal distributions, before turning to an example which incorporates categorical covariates and censored response. Data examples are provided in Sections 3.1.2 and 3.1.4. Section 3.2 introduces a nonparametric approach to switching regression wherein a DP mixture model for the joint distribution of covariates and response provides the basis of inference about both the state regression functions and

the hidden Markov model underlying state membership. As part of this framework, we develop in Section 3.2.4 a semiparametric model that can be used to link, through the hidden Markov model, nonparametric analysis with parametric inference about external covariates. The motivating data example is provided in Section 3.2.5. Finally, Section 3.3 applies our nonparametric analysis of conditional distributions to survival data and introduces, in this context, the modeling extension to multivariate regression.

3.1 A Model-Based Approach to Quantile Regression

Quantile regression can be used to quantify the relationship between quantiles of the response distribution and available covariates. It offers a practically important alternative to traditional mean regression, since, in general, a set of quantiles provides a more complete description of the response distribution than the mean. In many regression examples (e.g., in econometrics, educational studies, and environmental applications), we might expect a different structural relationship for the higher (or lower) responses than the *average* responses. In such applications, mean, or median, regression approaches would likely overlook important features that could be uncovered by a more general quantile regression analysis.

There is a fairly extensive literature on classical estimation for the standard p-th quantile regression model, $y_i = \mathbf{x}_i^T\beta + \epsilon_i$, where y_i denotes the response observations, $\mathbf{x}_i$ the corresponding covariate vectors, and ϵ_i the errors, which are typically assumed independent from a distribution (with density, say, $f_p(\cdot)$) that has p-th quantile equal to

0 (see, e.g., Koenker, 2005). This literature is dominated by *semiparametric* techniques where the error density $f_p(\cdot)$ is left unspecified (apart from the restriction $\int_{-\infty}^{0} f_p(\epsilon)\mathrm{d}\epsilon = p$). Hence, since there is no probability model for the response distribution, point estimation for β proceeds by optimization of some *loss* function. For instance, under the standard setting with independent and uncensored responses, the point estimates for β minimize $\sum \rho_p(y_i - \mathbf{x}_i^T\beta)$, where $\rho_p(u) = up - u1_{(-\infty,0)}(u)$; this form yields the least absolute deviations criterion for $p = 0.5$, i.e., for the special case of median regression. Any inference beyond point estimation is based on asymptotic arguments or resampling methods. The classical literature includes also work that relaxes the parametric (linear) regression form for the quantile regression function (see, e.g. He et al., 1998; Horowitz and Lee, 2005).

By comparison with the existing volume of classical work, the Bayesian literature on quantile regression is relatively limited. The special case of median regression has been considered in Walker and Mallick (1999), Kottas and Gelfand (2001), and Hanson and Johnson (2002). This work is based on a parametric form for the median regression function and nonparametric modeling for the error distribution, using either Pólya tree or Dirichlet process (DP) priors. (see, e.g. Müller and Quintana, 2004; Hanson et al., 2005, for reviews of these nonparametric prior models.) Regarding quantile regression, Yu and Moyeed (2001) and Tsionas (2003) discuss parametric inference based on linear regression functions and the asymmetric Laplace distribution for the errors; Kottas and Krnjajić (2008) develop Bayesian semiparametric models using DP mixtures for the error distribution; and Hjort and Petrone (2005) study nonparametric

inference for the quantile function based on DP priors, including brief discussion of the semiparametric extension to quantile regression. Moreover, Chamberlain and Imbens (2003) and Dunson and Taylor (2005) propose *semi-Bayesian* inference methods for linear quantile regression, which, in contrast to the work discussed above, do not involve probabilistic modeling for the response distribution.

A practical limitation of the Bayesian semiparametric modeling approaches developed in Walker and Mallick (1999), Kottas and Gelfand (2001), Hanson and Johnson (2002), and Kottas and Krnjajić (2008) is that, although they provide flexible shapes for the error distribution, they are based on parametric (in fact, linear) quantile regression functions. Regarding inference for non-linear quantile regression functions, Scaccia and Green (2003) model the conditional distribution of the response given a single continuous covariate with a discrete normal mixture with covariate-dependent weights. Moreover, Yu (2002) discusses a semi-Bayesian estimation method based on a piecewise polynomial representation for the quantile regression function corresponding, again, to a single continuous covariate, but without a probability model for the error distribution. We note that both of these approaches involve relatively complex Markov chain Monte Carlo (MCMC) methods for inference (specifically, certain forms of reversible jump MCMC techniques); moreover, their extension to handle problems with more than one covariate appears to be non-trivial.

The quantile regression approach proposed here is founded on Bayesian probabilistic modeling for the underlying unknown (random) distributions. The joint distribution of the response and the covariates is modeled with a flexible nonparametric

mixture, as outlined in Section 2.1, and we develop inference for different quantile curves based on the induced conditional distribution of the response given the covariates. To our knowledge, this presents the first attempt to develop a model-based, fully inferential framework for Bayesian nonparametric quantile regression. The methodology enables exact and full inference for any quantile regression function that may be of interest.

3.1.1 Posterior Inference Framework

We describe here the approach to estimate quantile curves based on the posterior for the conditional response density $f(y|\mathbf{x}; G)$ implied by DP mixture model (2.2).

Through use of the procedures of Section 2.4, it is possible to obtain an approximate realization of $f(y_0 \mid \mathbf{x}_0; G) = f(y_0, \mathbf{x}_0; G)/f(\mathbf{x}_0; G)$ at any point $(y, \mathbf{x}) = (y_0, \mathbf{x}_0)$. Repeating over a grid in y, that covers the range of response values of interest, we obtain a posterior realization from the random conditional density function $f(\cdot \mid \mathbf{x}_0; G)$ for the specific covariate values $\mathbf{x}_0$. Note that this is a posterior realization for the entire function, obtained, of course, up to the grid approximation. Now, for any $0 < p < 1$, the conditional quantile $q_p(\mathbf{x}_0) \equiv q_p(\mathbf{x}_0; G)$ satisfies $\int^{q_p(\mathbf{x}_0)} f(y \mid \mathbf{x}_0; G)\, \mathrm{d}y = p$. Hence, using numerical integration (with interpolation) of the posterior realizations from the conditional density $f(\cdot \mid \mathbf{x}_0; G)$, yields draws from the posterior of $q_p(\mathbf{x}_0)$ for any set of percentiles that might be of interest.

Therefore, for any $\mathbf{x}_0$, and for any $0 < p < 1$, we obtain samples from $p(q_p(\mathbf{x}_0) \mid \mathcal{D})$ that can be used to summarize the information from these conditional quantiles in any desired form. In particular, for any set of p values, working with a grid

over the covariate space, we can compute point and interval estimates for the corresponding quantile curves $q_p(\cdot; G)$. Evidently, graphical depiction of these estimates for the entire curve is not feasible for problems with more than two covariates. As shown in Section 3.1.2, for such applications, one can focus on illustrations involving the quantile regression function given subsets of the covariate vector including specific choices of one or two covariates.

Because of the need to obtain the posterior of $f(\cdot \mid \mathbf{x}_0; G)$ over a sufficiently dense grid of $\mathbf{x}_0$ values, implementation of inference becomes computationally intensive for high-dimensional covariate spaces. However, if interest focuses on the posterior of conditional response densities $f(y \mid \mathbf{x}_0; G)$ (e.g., Figure 3.3), or corresponding conditional quantiles, for a small number of specified $\mathbf{x}_0$ values, the approach is feasible in higher dimensions. Moreover, as discussed above, for inference on conditional quantile regression functions for a small subset of the covariates (e.g., Figures 3.1 and 3.2), the input grid is over a lower dimensional space and the computational expense is reduced. Regardless, the proposed approach to inference for quantile regression is well-suited for problems with small to moderate number of covariates, and there is indeed a wide variety of such regression problems that are of interest in, for example, economics and public health research. For such settings, the methodology is very flexible as it allows both non-linear quantile curves as well as non-standard shapes for the conditional distribution of the response given the covariates. Moreover, the model does not rely on the additive nonparametric regression formulation and therefore can uncover interactions between covariates that might influence certain quantile regression curves. Finally, a key

feature of the approach is that it enables simultaneous inference for any set of quantile curves of interest in a particular application.

3.1.2 Moral Hazard Data Example

To illustrate this quantile regression methodology, we consider data used by Yafeh and Yoshua (2003) to investigate the relationship between shareholder concentration and several indices for managerial moral hazard in the form of expenditure with scope for private benefit. The data set includes a variety of variables describing 185 Japanese industrial chemical firms listed on the Tokyo stock exchange. (The data set is available online through the *Economic Journal* at *http://www.res.org.uk*.) A subset of these data was also considered by Horowitz and Lee (2005) in application of their classical nonparametric estimation technique for an additive quantile regression model. As was done there, we consider a single model proposed by Yafeh and Yoshua (2003) in which index *MH5*, consisting of general sales and administrative expenses deflated by sales, is the response y related to a four-dimensional covariate vector $\mathbf{x}$, which includes *Leverage* (ratio of debt to total assets), $\log(Assets)$, the *Age* of the firm, and $TOPTEN$, the percent of ownership held by the ten largest shareholders. The response and all four covariates are continuous and, although *Leverage* and $TOPTEN$ occur over subsets of the real line, the data lies far enough from support boundaries to render the multivariate normal distribution a suitable choice for the kernel of the DP mixture model in (2.2).

The model is implemented using the prior specification approach outlined in Section 2.2. In the absence of genuine prior information in our illustrative analysis,

we take values from the data for the *prior* guesses of the center and range for the response and four covariates. Results were insensitive to reasonable changes in the prior specification, e.g., doubling the observed data range for the response and covariates did not affect the posterior estimates in Figures 3.1 – 3.3. A $\text{Ga}(1, 0.2)$ prior is placed on the DP precision parameter α, implying $\text{E}(n^\star) \approx 18$. Experimentation with alternative gamma priors, yielding smaller prior estimates for the number of distinct mixture components, has resulted in essentially identical posterior inference. Results are based on an MCMC sample of 150,000 parameter draws recorded on every tenth iteration following a (conservative) burn-in of 50,000 iterations.

Although it is not possible to show the response quantile functions over all four variables, as discussed in Section 3.1.1, it is straightforward to obtain quantile curves for the response given any one-dimensional or two-dimensional subset of the covariates. In Figure 3.1, we plot posterior point and 90% interval estimates for the response median and 90-th percentile as a function of each individual covariate. In addition, Figure 3.2 provides inference for the response median and 90-th percentile surfaces over the two-dimensional covariate space defined by *Leverage* and *TOPTEN*. (Note that Yafeh and Yoshua, 2003, found these two covariates to be the most significant.) In particular, shown are point estimates, through the posterior mean, and a measure of the related uncertainty, through the posterior interquartile range.

These two figures indicate the capacity of the model to capture non-linear shapes in the estimated quantile curves as well as to quantify the associated uncertainty. Figure 3.1 shows that the marginal relationship between each covariate and

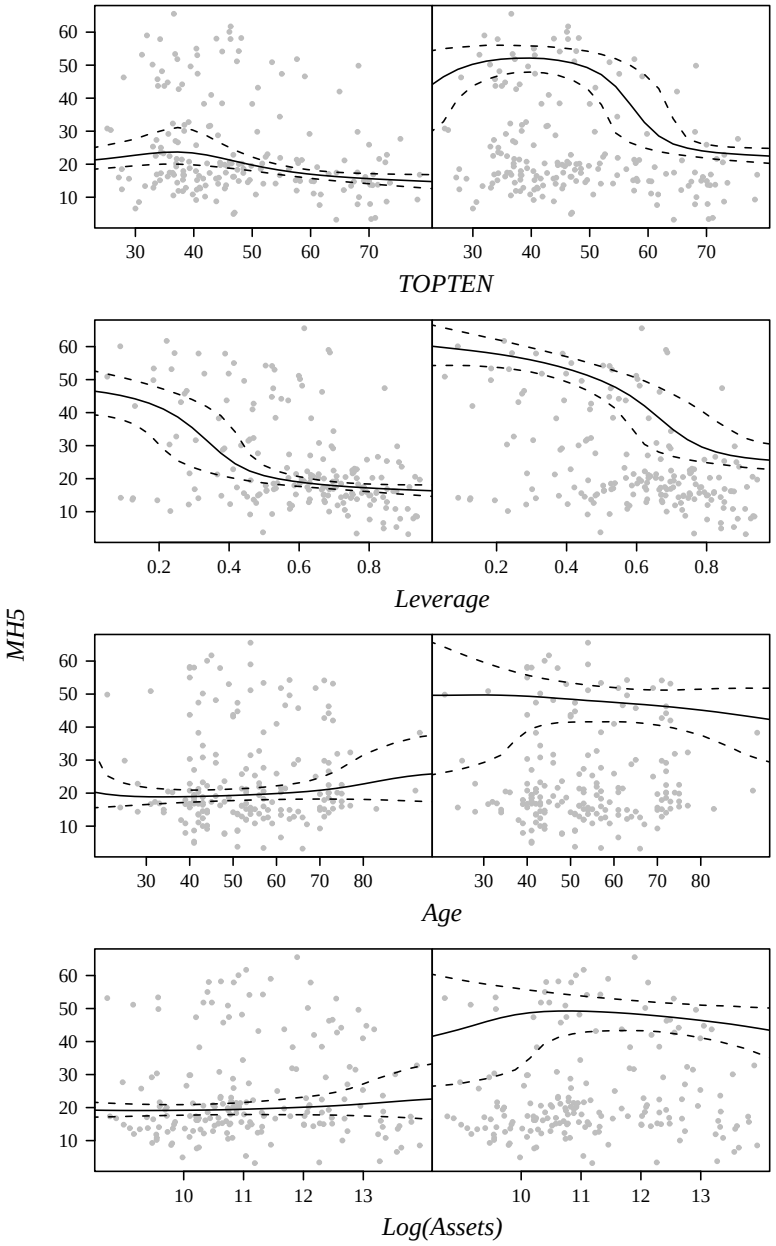


Figure 3.1: Moral hazard data. Posterior estimates for median regression (left column) and 90-th percentile regression (right column) for *MH5* conditional on each individual covariate. The solid lines are posterior mean estimates and dashed lines contain a 90% posterior interval. Data scatterplots are shown in grey.

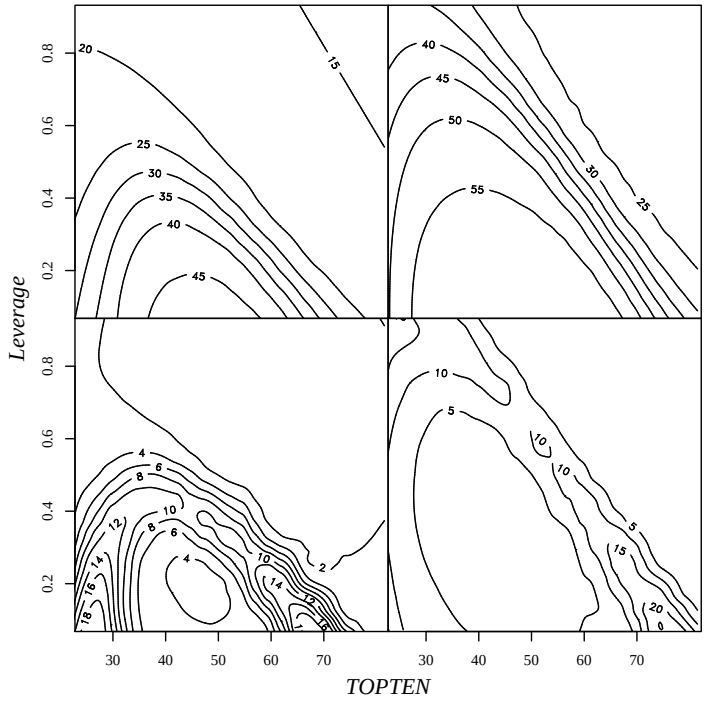


Figure 3.2: Moral hazard data. Posterior estimates of median surfaces (left column) and 90-th percentile surfaces (right column) for *MH5* conditional on *Leverage* and *TOPTEN*. The posterior mean is shown on the top row and the posterior interquartile range on the bottom.

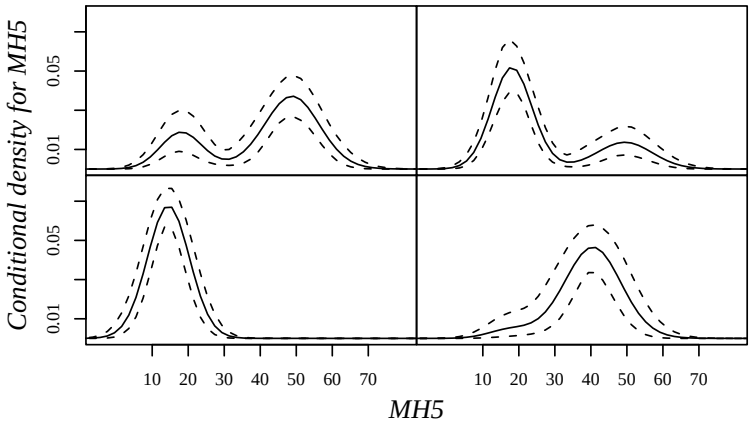


Figure 3.3: Moral hazard data. Posterior mean estimates (solid lines) and 90% interval estimates (dashed lines) for four conditional densities $f(y \mid \mathbf{x}_0; G)$ (see Section 4 for the values of $\mathbf{x}_0$).

MH5 may differ significantly depending upon the quantile of interest; this is particularly clear in the contrast between median and 90th percentile curves for *MH5* conditional on $TOPTEN$. The inference results displayed in Figure 3.2 show an interaction between the effects of *Leverage* and $TOPTEN$ in both the median and 90th percentile functions, suggesting that it is useful to relax the assumption of additivity over the covariate space (which forms the basis of the method in Horowitz and Lee, 2005). The same picture shows that posterior uncertainty about the quantile functions is highly variable throughout the covariate space; for each quantile, regions of steep change in the quantile function correspond to significantly higher uncertainty around the function estimate. In addition, it is interesting to note that all of the figures show a monotonically decreasing *MH5* 90th percentile and median with respect to *Leverage*, but that these quantiles do not appear to be strictly decreasing with respect to increases in $TOPTEN$. In particular, for $TOPTEN$ less than 50, our findings do not support the hypothesis of Yafeh and Yoshua (i.e., that increased shareholder concentration leads to lower managerial moral hazard). However, the *MH5* quantiles do seem to be monotonically decreasing with $TOPTEN$ greater than 50, indicating perhaps that the relationship hypothesized by Yafeh and Yoshua only manifests itself after a small group of shareholders has amassed a significant stake in the firm.

Figure 3.3 illustrates inference for the conditional response density $f(y \mid \mathbf{x}_0; G)$. Included are results for four values, $\mathbf{x}_0$, of the covariate vector $\mathbf{x} = (TOPTEN$, *Leverage*, *Age*, $\log(Assets))$. Specifically, clockwise from top left, the plots correspond to $\mathbf{x}_0 = (40, 0.3, 55, 11)$, $(35, 0.6, 55, 11)$, $(40, 0.3, 70, 13)$, and $(70, 0.8, 55, 11)$. This type of infer-

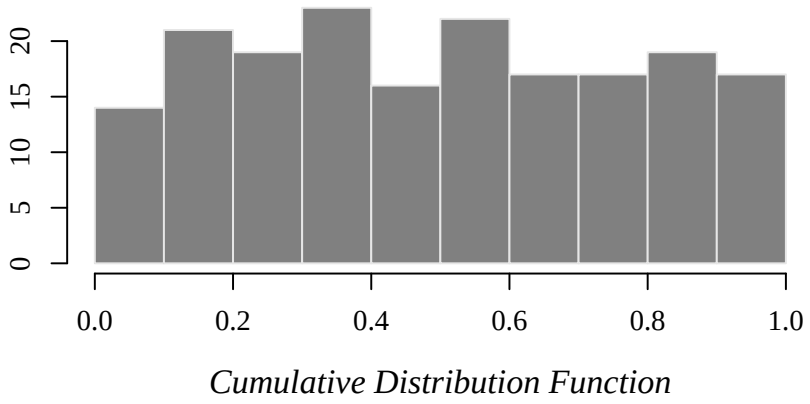


Figure 3.4: Moral hazard data. Histogram of the posterior mean for the cumulative distribution function evaluated at the data: $\{\mathbb{E}\left[F(y_1|\mathbf{x}_1; G)|\mathcal{D}\right], \ldots, \mathbb{E}\left[F(y_n|\mathbf{x}_n; G)|\mathcal{D}\right]\}$. The number of observations allocated to each bin is shown on the vertical axis.

ence highlights the ability of the model to capture non-standard distributional features such as heavy tails, skewness, and multimodality. The posterior estimates in Figure 3.3 clearly indicate that the response distribution changes significantly throughout the covariate space in ways that can not be modeled with standard parametric forms. Inspection of the data scatterplots in Figure 3.1 makes it clear that the non-standard features captured in the posterior estimates from the DP mixture model are driven by the data and are not simply an artifact of the flexible nonparametric prior mixture model.

It is also possible to perform a graphical goodness-of-fit test based upon posterior sampling results and the probability integral transform theorem. Full inference is available for the conditional cumulative distribution function $F(y|\mathbf{x}; G)$ at any de-

sired points $(\mathbf{x}, y)$ and, if the model is performing as desired, the values $\{F(y_1|\mathbf{x}_1; G)$, $\ldots$, $F(y_n|\mathbf{x}_n; G)\}$ should be uniformly distributed. Indeed, Figure 3.4 shows the posterior mean sample for these values (i.e., the histogram of $\{\mathbb{E}\left[F(y_1|\mathbf{x}_1; G)|\mathcal{D}\right], \ldots, \mathbb{E}\left[F(y_n|\mathbf{x}_n; G)|\mathcal{D}\right]\}$) and it is clear that the distribution is roughly uniform.

3.1.3 Tobit Quantile Regression

There are several regression applications that involve constrained observations for the response variable, and possibly also for the covariates. For instance, different types of censoring or truncation are commonly present in survival analysis data. In econometrics applications, a standard scenario involves certain forms of partially observed responses leading to what is typically referred to as Tobit regression models, after the work by Tobin (1958) (see, e.g. Amemiya, 1984, for a thorough review of various types of Tobit models).

The standard Tobit model is formulated through latent random variables $\tilde{y}_i$, which are assumed independent and normally distributed with mean $\mathbf{x}_i^T\beta$ and variance σ^2. Tobit quantile regression arises by modeling a specific quantile of the latent response distribution as a function of the covariates. The covariate vectors $\mathbf{x}_i$ are observed for all subjects in the data. However, the observed responses, y_i, are constrained according to $y_i = \max\{y_i^0, \tilde{y}_i\}$, where the y_i^0 are fixed threshold points. In applications, the threshold value is typically the same for all data subjects, and we can thus set without loss of generality $y_i^0 = 0$ (as in our data example of Section 3.1.4). Formally, this data structure corresponds to (fixed) left censoring. However, there is a subtle difference with more

traditional survival analysis applications, since in economics settings, the latent variable $\tilde{y}$ may exist only conceptually, e.g., as a particular *utility* functional formulated based on empirical and/or theoretical studies.

The classical semiparametric literature includes several estimation techniques for both the mean regression and quantile regression Tobit models (see, e.g. Buchinsky and Hahn, 1998, and further references therein). Again, these approaches do not include probabilistic modeling for the latent response distribution and are thus limited in terms of the range of inferences that they can provide. Bayesian approaches to Tobit regression for econometrics applications appear to have focused on parametric modeling with linear regression functions. For instance, the early work of Chib (1992) developed Bayesian inference for linear Tobit regression with normal errors whereas, more recently, Yu and Stander (2007) studied linear Tobit quantile regression with asymmetric Laplace errors.

The modeling framework developed in Chapter 2 and Section 3.1.1 can be utilized to provide a flexible nonparametric approach to inference for Tobit quantile regression. Again, we start with a DP mixture model, $f(\tilde{y}, \mathbf{x}; G) = \int k(\tilde{y}, \mathbf{x}\ ; \theta)\ \mathrm{d}G(\theta)$, $G \mid \alpha,\ \psi \sim \mathrm{DP}(\alpha, G_0(\psi))$, for the joint distribution of the latent response variable $\tilde{y}$ and the vector of covariates $\mathbf{x}$. The mixture kernel can be defined by a multivariate normal with continuous covariates (as in Section 2.1) or involve discrete components when categorical covariates are available (as discussed in Section 2.3). The first stage of the hierarchical model for the data, $(y_i, \mathbf{x}_i)$, $i = 1, ..., n$, is built again from conditional independence given the mixing parameters θ_i, $i = 1, ..., n$, but is modified with respect to (2.3) to replace the (conditional) response kernel density with its corresponding dis-

tribution function for all i with $y_i = 0$. The analogous modifications to the MCMC posterior simulation method of Section 2.4 yield the full posterior for G, α, ψ and the θ_i, $i = 1, ..., n$. We provide more details in Section 3.1.4 with the concrete DP mixture model used for our data illustration.

As in Section 3.1.1, full and exact inference for any set of quantile regression curves emerges from the posterior realizations for the conditional response density $f(\cdot \mid \mathbf{x}_0; G)$ over grid values $\mathbf{x}_0$ in the covariate space. Note that here, for any specified point $y_0 > 0$ associated with fully observed responses, $f(y_0 \mid \mathbf{x}_0; G)$ in the notation of Section 3 is given through $f(y_0 \mid \tilde{y} = y_0 > 0, \mathbf{x}_0; G)$. Hence, inference for Tobit quantile regression is based on the conditional response density, given $\mathbf{x}$, arising from the underlying DP mixture $f(\tilde{y}, \mathbf{x}; G)$, conditionally also on $\tilde{y} > 0$. Moreover, using the posterior realizations for $f(\tilde{y} \mid \mathbf{x}; G)$, we can obtain the posterior for $\Pr(\tilde{y} \leq 0 \mid \mathbf{x}_0; G)$. A collection of these posteriors for a set of specified $\mathbf{x}_0$ provides information on the relationship between the covariates and the censoring mechanism for the response. Because of the flexibility of the mixture model for the joint distribution of $\tilde{y}$ and $\mathbf{x}$, the proposed modeling approach enables potentially different structure for the relationship between the response and the covariates across different quantile regression curves as well as for the relationship between the covariates and the underlying mechanism that constrains the response. This is a practically important advantage over parametric formulations (as in, e.g., Yu and Stander, 2007) that postulate a linear regression form for all the relationships above.

3.1.4 Female Labor Supply Data Example

To illustrate the extensions developed in Section 3.1.3, we consider a subset of the data on female labor supply corresponding to the University of Michigan Panel Study of Income Dynamics for year 1975. Using this data set, Mroz (1987) presents a systematic analysis of theoretical and statistical assumptions used in empirical models of female labor supply. The sample considered by Mroz (1987) consists of 753 married white women between the ages of 30 and 60, with 428 of them working at some time during year 1975. The 428 fully observed responses, y_i, are given by the wife's *work* (in 100 hours) during year 1975. For the remaining 325 women, the observed *work* of $y_i = 0$ corresponds to negative values for the latent *labor supply* response, $\tilde{y}_i$. The data set includes covariate information on family income, wife's wage, education, age, number of children of different age groups, and mother's and father's educational attainment, as well as on husband's age, education, wage, and hours of work. For our purely illustrative analysis, we consider *number of children* as the single covariate, x. This covariate combines observations from two variables in the data set, "number of children less than 6 years old in household" and "number of children between ages 6 and 18 in household".

Although the response variable can be treated as continuous (non-zero observed responses range from 12 to 4950 hours), the covariate is a categorical variable (with values that range from 0 to 8 children). As discussed in Section 2.3, there are several possible choices for the DP mixture kernel. Here, we consider the simple setting with $k(\tilde{y}, x; \theta)$ comprising independent normal and Poisson components, a version that is

sufficient for our illustrative purposes. (In other applications, a similar model based on negative binomial, rather than Poisson, components for the mixture kernel could be considered as a robust alternative.) Specifically, we work with the following DP mixture model,

$$\begin{aligned} f(\tilde{y}, x; G) &= \int \mathrm{N}(\tilde{y}; \mu, \sigma^2)\mathrm{Po}(x; \lambda)\, \mathrm{d}G(\mu, \sigma^2, \lambda), \\ G \mid \alpha, \psi &\sim \mathrm{DP}(\alpha, G_0(\psi)), \end{aligned} \tag{3.1}$$

for the latent labor supply response and number of children covariate. Here, $\mathrm{N}(\cdot\,;\, \mu,\, \sigma^2)$ denotes the density of the normal distribution with mean μ and variance σ^2, and $\mathrm{Po}(\cdot;\, \lambda)$ the probability mass function of the Poisson distribution with mean λ. Moreover, G_0 is built from independent components, specifically, $\mathrm{N}(\psi_1,\, \psi_2)$ for μ, $\mathrm{Ga}(c, \psi_3)$ for σ^{-2}, and $\mathrm{Ga}(d, \psi_4)$ for λ, with hyperpriors placed on $\psi = (\psi_1,\, \psi_2,\, \psi_3,\, \psi_4)$.

The results reported below are based on a $\mathrm{Ga}(1, 0.2)$ prior for α, and $\mathrm{N}(10, 40)$, $\mathrm{Ga}(2, 40)$, $\mathrm{Ga}(2, 0.2)$, and $\mathrm{Ga}(3, 3)$ priors for ψ_1, ψ_2^{-1}, ψ_3, and ψ_4, respectively. The remaining parameters of G_0 are set to $c = 2$ and $d = 1$. We have experimented increasing and decreasing the variability around α and ψ_1 and the prior expectations for ψ_2 and ψ_3, as well as with alternative specifications for ψ_4, and have not found this to affect the analysis. Results are based on an MCMC sample of 100,000 parameter draws recorded on every fifth iteration following a (conservative) burn-in period of 50,000 iterations.

Let $p(\theta_1, ..., \theta_n \mid \alpha, \psi)$ be the, analogous to (2.12), Pólya urn prior for the mixing parameters $\theta_i = (\mu_i, \sigma_i^2, \lambda_i)$, that results after integrating G over its DP prior, and set $I_0 = \{i : y_i = 0\}$ and $I_1 = \{i : y_i > 0\}$. Then, the posterior for α, ψ and the θ_i

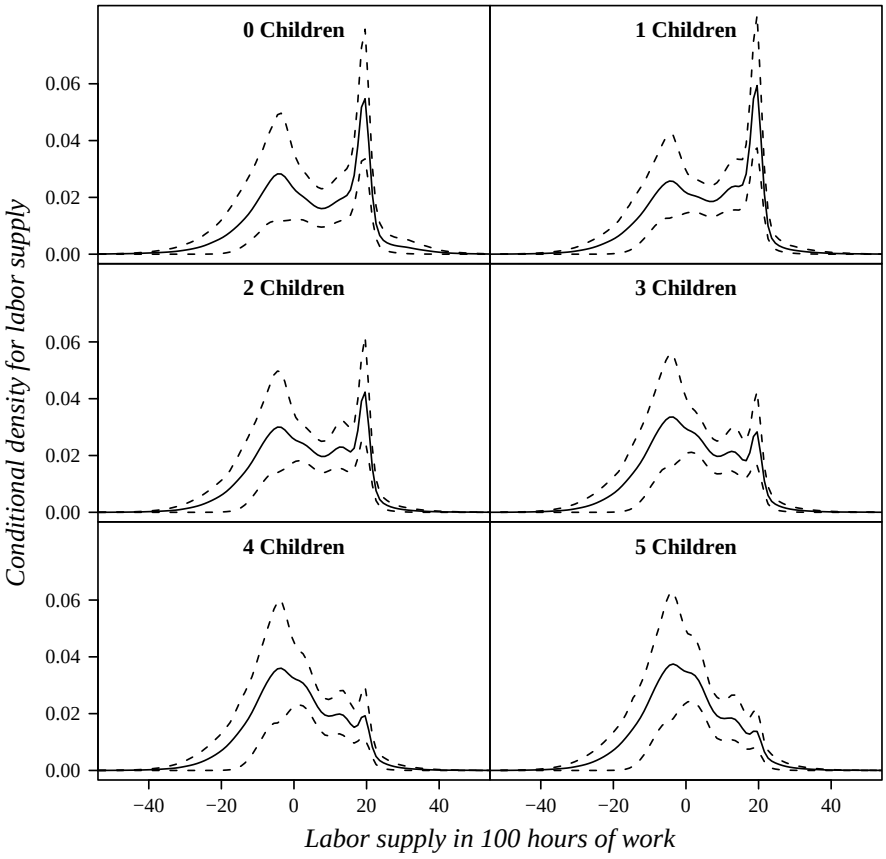


Figure 3.5: Female labor supply data. Posterior estimates for $f(\tilde{y} \mid x; G)$ given $x = 0, ..., 5$ children. Solid and dashed lines correspond to posterior mean and 90% posterior interval estimates, respectively.

is proportional to

$$p(\alpha)p(\psi)p(\theta_1, ..., \theta_n \mid \alpha, \psi) \prod_{i \in I_0} \Phi(-\mu_i/\sigma_i) \prod_{i \in I_1} \mathrm{N}(y_i; \mu_i, \sigma_i^2) \prod_{i=1}^{n} \mathrm{Po}(x_i; \lambda_i)$$

where $\Phi(\cdot)$ is the standard normal distribution function. We sample from the full posterior, that includes also G, using an MCMC method similar to the one described in Section 2.4. The structure of the Metropolis-Hastings steps for the θ_i remains the same. However, when resampling, for $j = 1, ..., n^\star$, from $g_0(\mu_j^\star, \sigma_j^{\star 2}, \lambda_j^\star; \psi) \prod_{\{i:w_i=j\}} \mathrm{Po}(x_i; \lambda_j^\star)$ $\prod_{i \in I_0 \cap \{i:w_i=j\}} \Phi(-\mu_j^\star/\sigma_j^\star) \prod_{i \in I_1 \cap \{i:w_i=j\}} \mathrm{N}(y_i; \mu_j^\star, \sigma_j^{\star 2})$, the posterior full conditionals for

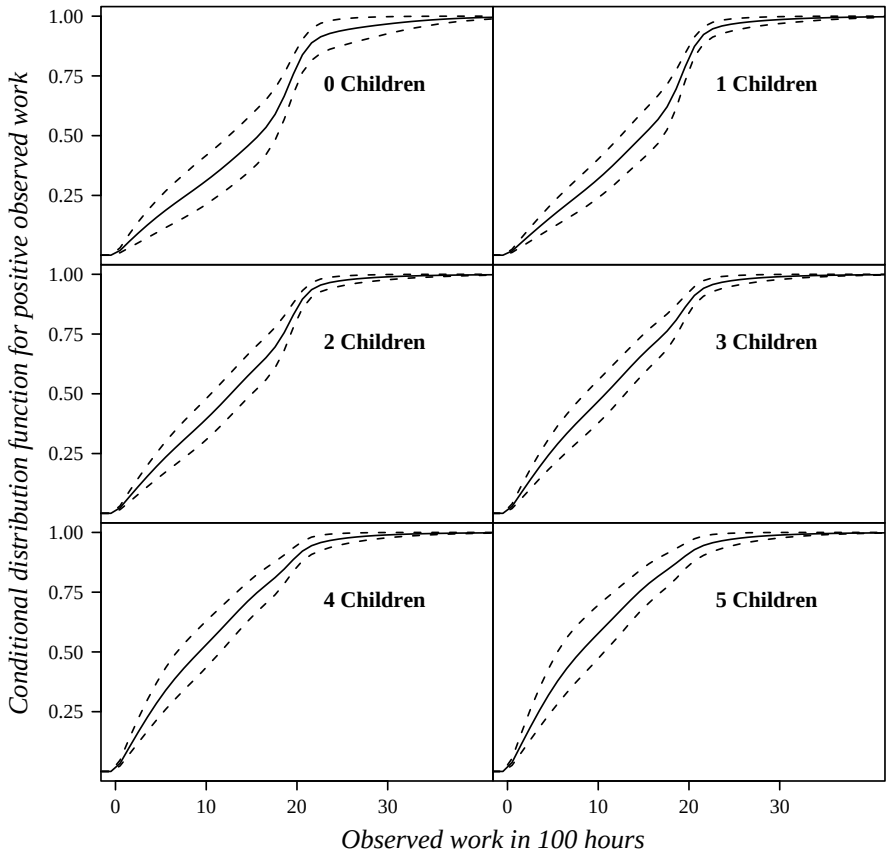


Figure 3.6: Female labor supply data. Posterior estimates for $\Pr(\tilde{y} < u \mid \tilde{y} > 0, x; G)$ for $x = 0, \ldots, 5$ children. The solid lines are posterior mean estimates and dashed lines indicate 90% posterior interval estimates.

$\mu_j^\star$ and $\sigma_j^{\star 2}$ are no longer available in a form from which it is easy to draw. Sampling proceeds through Metropolis-Hastings steps with normal proposals for $\mu_j^\star$ and gamma proposals for $\sigma_j^{\star 2}$. The posterior full conditional for $\lambda_j^\star$ is a gamma distribution with shape parameter $d + \sum_{\{i:w_i=j\}} x_i$ and rate parameter $\psi_4 + n_j$. The DP precision parameter is, again, updated using the method from Escobar and West (1995). Finally, the posterior full conditionals for all four hyperparameters in ψ have standard forms.

As in Section 2.4, the posterior samples for G can be used to obtain the poste-

rior of the conditional distribution for the latent labor supply response given a specific value for the number of children covariate. Posterior estimates for the conditional densities $f(\tilde{y} \mid x; G)$, corresponding to $x = 0, ..., 5$ children, are shown in Figure 3.5. The estimated latent response densities have non-standard shapes that change with the covariate value in a fashion that is difficult to describe with a parametric regression relationship. The peak around 2000 hours of work, which is seen in conditional response densities for lower numbers of children, corresponds to a traditional full-time job (50 weeks of 40 hours). The nonparametric DP mixture model is exposing density structure that would have been missed under standard parametric assumptions for the latent response distribution, e.g., the models developed by Chib (1992) and Yu and Stander (2007) based on normal and asymmetric Laplace distributions, respectively.

Non-standard features are also seen in response distributions for positive observed work. This is illustrated in Figure 3.6, which shows posterior estimates for $\Pr(\tilde{y} < u \mid \tilde{y} > 0, x; G) = \Pr(0 < \tilde{y} < u, x; G) / \Pr(\tilde{y} > 0, x; G)$, i.e., the conditional distribution function at $u > 0$, given positive observed work and given x; results are plotted for $x = 0, ..., 5$ children. For any value of x, working with a grid of u values, posterior realizations for $\Pr(\tilde{y} < u \mid \tilde{y} > 0, x; G)$ are given by

$$\Pr(\tilde{y} < u \mid \tilde{y} > 0, x; G^L) = \frac{\sum_{l=1}^{L} p_l \mathrm{Po}(x; \lambda_l) \left[\Phi((u - \mu_l)/\sigma_l) - \Phi(-\mu_l/\sigma_l)\right]}{\sum_{l=1}^{L} p_l \mathrm{Po}(x; \lambda_l) \left[1 - \Phi(-\mu_l/\sigma_l)\right]}, \qquad (3.2)$$

where, following the notation of Section 2.4, $G = \{p_l, (\mu_l, \sigma_l^2, \lambda_l) : l = 1, ..., L\}$ is the truncated posterior realization for G.

Next, inference about conditional quantiles $q_p(x)$ for positive observed work

proceeds based on these posterior realizations. In particular, for any specified p and any value x for the number of children, the posterior samples $\{q_{pb}(x) : b = 1, ..., B\}$ for $q_p(x)$ are obtained (with interpolation) from $p = \Pr(\tilde{y} < q_{pb}(x) \mid \tilde{y} > 0, x; G)$. As an illustration, Figure 3.7 plots boxplots of the posterior samples for $q_{0.5}(x)$ and $q_{0.9}(x)$. (Boxplots are constructed such that the boxes contain the interquartile sample range and the whiskers extend to the most extreme sample point that is no more than 1.5 times the interquartile range outside the central box.) Noteworthy is the different rate of decrease in the median and 90-th percentile regression relationships between positive observed work and number of children. Note also that the posteriors for $q_{0.9}(x)$ at $x = 1, 2, 3, 4$ children are more concentrated than the posterior for $q_{0.9}(0)$, whereas such a difference is substantially less pronounced in the posteriors for $q_{0.5}(x)$. This difference in the posterior uncertainty around the right tail of the conditional distribution functions at $x = 0$ and at $x = 1, 2, 3, 4$ children is also reflected in the corresponding posterior estimates in Figure 3.6.

Finally, as discussed in Section 3.1.3, of interest might be inference for $\Pr(\tilde{y} \leq 0 \mid x; G)$, i.e., the probability of zero hours of observed work given the number of children. For any value of $x = 0, ..., 8$, posterior samples for this probability arise from $\Pr(\tilde{y} \leq 0 \mid x; G^L) = \left[\sum_{l=1}^{L} p_l \text{Po}(x; \lambda_l)\Phi(-\mu_l/\sigma_l)\right] / \sum_{l=1}^{L} p_l \text{Po}(x; \lambda_l)$. Boxplots of these posterior samples are shown in Figure 3.8, indicating fairly similar relationship between the covariate and the censoring mechanism for the response when $x = 0, 1$ children; a noticeable increase in the probability of zero hours of observed work with $x = 2, 3, 4$ children; and similar probabilities, albeit with increased posterior uncertainty,

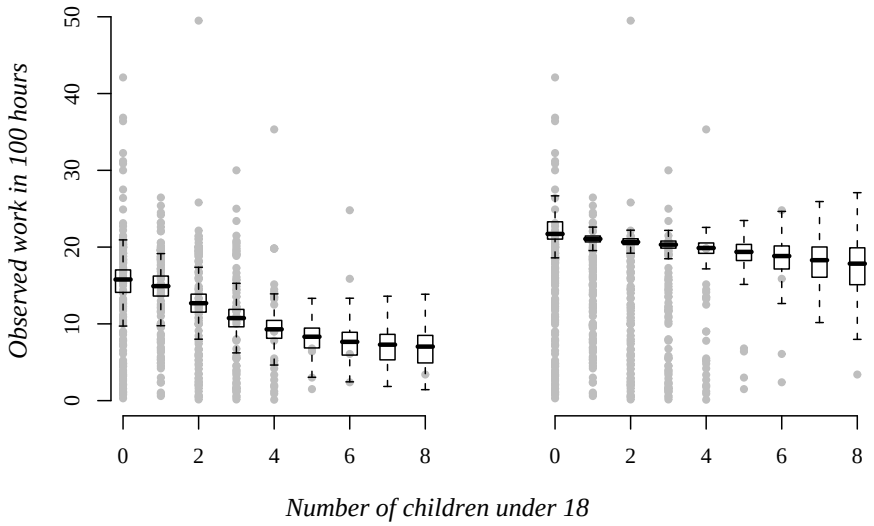


Figure 3.7: Female labor supply data. Posterior samples of positive observed work median (left panel) and 90-th percentile (right panel) given the realized values of the covariate. The positive data observations are shown in grey.

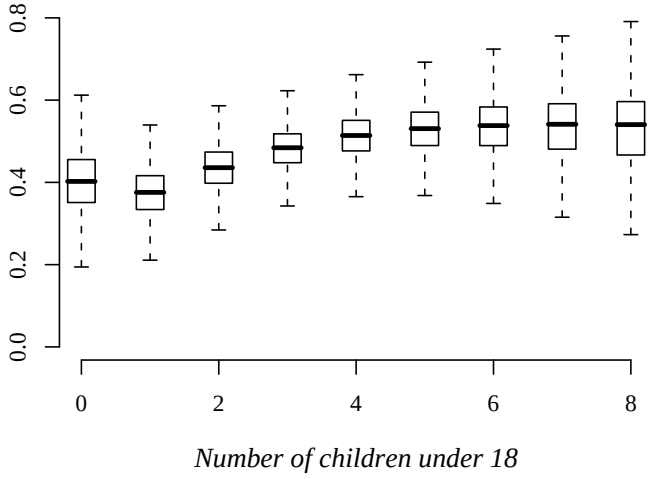


Figure 3.8: Female labor supply data. Posterior samples for $\Pr(\tilde{y} \leq 0 \mid x; G)$.

for $x = 5, 6, 7, 8$ children.

To summarize, posterior estimates for the conditional response distribution of $\tilde{y}$ and related functionals, including inference about the distribution for y conditional on $\tilde{y} > 0$, exhibit behavior that would be very difficult to capture with existing Bayesian estimation methods. From an economic perspective, the figures suggests that the main effect of an increase in offspring on labor supply is to reduce the proportion of women working full-time. This is especially clear in Figure 3.5, where the density peak corresponding to full-time labor decreases in magnitude as the number of children increases and the probability mass is redistributed in the region $\tilde{y} < 2000$ hours of work.

3.2 Markov Switching Regression

The focus of this section is to develop a flexible approach to nonparametric switching regression combining DP mixture nonparametric regression with a hidden Markov model. A modeling framework for data that has been drawn from a number of unobserved *states*, where each state defines a different relationship between response and covariates, switching regression was originally developed in the context of econometrics (Goldfeld and Quandt, 1973; Quandt and Ramsey, 1978) and has primarily been approached through likelihood-based estimation. When response and covariates occur in time, temporal dependence for state membership can be used to guide the modeling. A hidden Markov mixture model in this context holds that the state vector constitutes a Markov chain, and thus introduces an underlying dependence into the data. Robert

et al. (1993) and Chib (1996) discuss hidden Markov models in the estimation of mixtures of parametric densities. Our methodological framework involves a known small number of states where prior information is available on the properties of the underlying state Markov chain, but there is a need for nonparametric modeling within each subpopulation. Hence, the motivation for our approach is distinct from that of hidden Markov modeling with an unknown number of states as in, e.g., the work by Robert et al. (2000) and Beal et al. (2002). Our assumption that the number of mixture states is known fits within the general premise of an informative state estimation coupled with flexible nonparametric modeling for density and regression estimation.

Mixtures of regressions are used to study multiple populations, each of which involves a different conditional relationship between response and covariates. The basic switching regression model defines distinct regression functions for data that have been drawn from populations corresponding to a number of unobserved *states*. Following the early work of Goldfeld and Quandt (1973) and Quandt and Ramsey (1978), the more recent literature includes, for instance, approaches for switching dynamic linear models (Shumway and Stoffer, 1991) and switching ARMA models (Billio et al., 1999). Moreover, Hurn et al. (2003) describe a Bayesian decision theoretic approach to estimation for mixtures of linear regressions, whereas the approach of Shi et al. (2005) offers a departure from the linear regression assumption through a mixture of Gaussian process regressions.

The generic mixtures of regressions setting holds that the response y given covariates $\mathbf{x}$ has been drawn from a member of a heterogeneous set of R conditional

distributions defined by the densities $f_1(y \mid \mathbf{x}), \ldots, f_R(y \mid \mathbf{x})$, and hence that $\Pr(y \mid \mathbf{x}) = w_1 f_1(y \mid \mathbf{x}) + \ldots + w_R f_R(y \mid \mathbf{x})$, where $\sum_{r=1}^{R} w_r = 1$. We propose a departure from this standard form, wherein the response and covariates are jointly distributed according to one of the densities $f_1(\mathbf{x}, y), \ldots, f_R(\mathbf{x}, y)$ – i.e., now $\Pr(\mathbf{x}, y) = w_1 f_1(\mathbf{x}, y) + \ldots + w_R f_R(\mathbf{x}, y)$ – and therefore $\Pr(y \mid \mathbf{x}) = w_1^\star f_1(\mathbf{x}, y) + \ldots + w_R^\star f_R(\mathbf{x}, y)$, where $w_r^\star = w_r / \sum_{r=1}^{R} w_r f_r(\mathbf{x})$. Thus, the approach is appropriate whenever mixture component probabilities for a given $\mathbf{x}$ and y should be dependent upon the joint distribution for response and covariates, even though primary interest is in the regression relationship for response given covariates.

Section 3.2.1 presents the hidden Markov DP mixture model, and Sections 3.2.2 and 3.2.3 develop an efficient Markov chain Monte Carlo (MCMC) algorithm for posterior simulation. Effective sampling of the hidden chain states is essential to success of the MCMC algorithm, and we thus propose a method based on forward-backward sampling (see, e.g., Scott, 2002). An extension of the hidden Markov DP mixture model to include external variables which are correlated with the underlying Markov chain, but conditionally independent of the joint covariate-response distribution, is described in Section 3.2.4. In Section 3.2.5, the methods are illustrated with an application from fisheries research involving analysis of stock-recruitment data under shifts in the ecosystem state.

3.2.1 Model Specification

Here, we develop the extension of DP mixture implied conditional regression to the context of time dependent switching regression. The data consist of a set of covariate vectors $\mathbf{x}_t$ and corresponding responses y_t observed at times $t = 1, \ldots, T$. The data from each time point are associated with a hidden state variable, $h_t \in \{1, \ldots, R\}$, such that, given h_t, the response-covariate joint distribution is defined by a state-specific density $f_{h_t}(\mathbf{x}_t, y_t)$. As in Section 2.1, we begin by describing density estimation in the $d = d_{\mathbf{x}} + 1$ dimensional setting, with data $\mathcal{D} = \{\mathbf{z}_t = (x_t^1, \ldots, x_t^{d_{\mathbf{x}}}, y_t) : \ t = 1, \ldots, T\}$. Now, however, the successive observations $\mathbf{z}_t$ are correlated through dependence in state membership $\mathbf{h} = (h_1, \ldots, h_T)$, which constitutes a stationary Markov chain defined by an $R \times R$ transition matrix $\mathbf{Q}$. Although we consider only first-order dependence in the Markov chain, the model and posterior simulation methods can be extended to handle higher order Markov chains.

The first-order hidden Markov location-scale normal DP mixture model can then be expressed as follows,

$$\begin{aligned}
\mathbf{z}_t \mid h_t, G_{h_t} &\overset{ind}{\sim} f_{h_t}(\mathbf{z}_t; G_{h_t}) = \int \mathrm{N}(\mathbf{z}_t; \mu, \sigma) dG_{h_t}(\mu, \sigma), \quad t = 1, \ldots, T \\
G_r \mid \alpha_r, \psi_r &\overset{ind}{\sim} \mathrm{DP}\left(\alpha_r, G_0(\psi_r)\right), \quad r = 1, \ldots, R \\
\mathbf{h} \mid \mathbf{Q} &\sim \Pr(\mathbf{h} \mid \mathbf{Q}) = \prod_{t=2}^{T} Q_{h_{t-1}, h_t},
\end{aligned} \tag{3.3}$$

where we denote the r-th row of $\mathbf{Q}$ by $Q_r = (Q_{r,1}, \ldots, Q_{r,R})$, with $Q_{r,s} = \Pr(h_t = s \mid h_{t-1} = r)$, for $r, s = 1, ..., R$. We assume that, in the prior, each state is equally likely for h_1. Moreover, the DP centering distributions, $G_0(\mu, \sigma; \psi_r)$, are equal to $\mathrm{N}(\mu \ ; m_r,$

$V_r)$ $\mathrm{W}_{\nu_r}(\ \sigma^{-1};\ S_r^{-1})$, with $\psi_r = (m_r, V_r, S_r)$. For $r = 1, \ldots, R$, we place hyperpriors on ψ_r and α_r such that $\pi(\psi_r) = \mathrm{N}\left(m_r; a_{m_r}, B_{m_r}\right) \mathrm{W}_{a_{V_r}}(V_r^{-1}; B_{V_r}^{-1}) \mathrm{W}_{a_{S_r}}(S_r; B_{S_r})$, and $\pi(\alpha_r) = \mathrm{Ga}(\alpha_r; a_{\alpha_r}, b_{\alpha_r})$. The prior for $\mathbf{Q}$ is built from independent Dirichlet distributions, $\pi(Q_r) = \mathrm{Dir}(Q_r; \lambda_r)$, where $\mathrm{Dir}(Q_r; \lambda_r)$, with $\lambda_r = (\lambda_{r,1}, \ldots, \lambda_{r,R})$, denotes the Dirichlet distribution such that $\mathbb{E}[Q_{r,s}] = \lambda_{r,s}/(\sum_{i=1}^{R} \lambda_{r,i})$.

Applying the regression approach of Section 2.1, the joint response-covariate density specification in (3.3) yields our proposed hidden Markov switching regression model. In particular, conditional on the (truncated) state-specific random mixing distribution, G_r^L, whose posterior distribution is obtained using the MCMC method developed in Section 3.2.2, the conditional density for y given $\mathbf{x}$ drawn from the population corresponding to state r is

$$f_r(y \mid \mathbf{x}; G_r^L) = \frac{f_r(\mathbf{x}, y; G_r^L)}{f_r(\mathbf{x}; G_r^L)} = \frac{\sum_{l=1}^{L} p_{r,l} \mathrm{N}\left(\mathbf{x}, y; \mu_{r,l}, \Sigma_{r,l}\right)}{\sum_{l=1}^{L} p_{r,l} \mathrm{N}\left(\mathbf{x}; \mu_{r,l}^{\mathbf{x}}, \Sigma_{r,l}^{\mathbf{xx}}\right)}. \tag{3.4}$$

The conditional mean regression for state r, $\mathbb{E}_r\left[y \mid \mathbf{x}; G_r^L\right]$, can be estimated through a state-specific application of equation (2.11).

In practice, the hyperparameters for the α_r, ψ_r and for $\mathbf{Q}$ need to be carefully chosen. We are motivated by a setting where prior information is available on the state vector $\mathbf{h}$, and the λ_r parameters of $\pi(Q_r)$ are chosen based on prior expectation for the probabilities of moving from state r to each state in a single time step. However, this prior information pertains only to the transition probabilities between states and does not fully identify the state components. Thus, we need to provide enough information to facilitate identification of the mixture components and ensure that the

transition probabilities defined by $\mathbf{Q}$ refer to the intended states. On the other hand, the nonparametric regression is motivated by a desire to be noninformative about each regression component and we thus seek a more automatic prior specification for each ψ_r.

Within the framework of our DP mixture implied conditional regression, it is possible to have each state-specific centering distribution, $G_0(\psi_r)$, associate the densities $\int \mathrm{N}(\mathbf{z}; \mu, \sigma) dG_r(\mu, \sigma)$ with specific regions of the joint response-covariate space, without putting prior information on the shape of the conditional response density or regression curve within each region. Since the prior parameters m_r and V_r control the location of the normal kernels, the hyperparameters a_{m_r}, B_{m_r}, a_{V_r}, and B_{V_r} can be used to express prior belief about the state-specific joint response-covariate distributions. Specifically, assume a prior guess for the mean and covariance matrix corresponding to the population for state r, where prior information for the covariance may only be available in the form of a diagonal matrix. Then, we can set a_{m_r} equal to the prior mean, B_{m_r} to the prior covariance, and choose a_{V_r} and B_{V_r} such that $\mathbb{E}[V_r]$ is equal to the prior covariance (alternatively, $\mathbb{E}[V_r^{-1}]$ can be set equal to the inverse of the prior covariance matrix and we have observed the method to be robust to either specification). In the absence of such prior information, one can use a data-dependent prior specification technique. Given a prior allocation of observations expressed as the state vector $\mathbf{h}^\pi = (h_1^\pi, ..., h_T^\pi)$, each set $\{a_{m_r}, B_{m_r}, B_{V_r}\}$ can be specified through the mean and covariance of the data subset $\{\mathbf{z}_t : h_t^\pi = r\}$. In particular, a_{m_r} is set to the state-specific data mean and both B_{m_r} and $\mathbb{E}[V_r] = (a_{V_r} - d - 1)^{-1} B_{V_r}$ are set to the state-specific data covariance. With care

taken to ensure that it does not overly restrict the component locations, this approach provides an automatic prior specification that combines strong state allocation beliefs with weak information about the state-specific regression functions.

For the S_r we seek only to scale the mixture components to the data, and thus we set all the $\mathbb{E}(S_r) = a_{S_r} B_{S_r}$ equal to a diagonal matrix with each diagonal entry a quarter of the full data range for the respective dimension. The precision parameters a_{V_r}, a_{S_r}, and ν_r, for $r = 1, \ldots, R$, are set to values slightly larger than $d+2$; in practice, we have found $2(d+1)$ to work well. Working with various data sets, including the one in Section 3.2.5, we have observed results to be insensitive to reasonable changes in this specification. In particular, experimentation with a variety of choices for the matrices B_{S_r}, indicating prior expectation of either more or less diffuse normal kernel components, resulted in robust posterior inference.

Specification of the hyperpriors on DP precision parameters is facilitated by the role that each α_r plays in the prior distribution for the number of unique components in the set of n_r latent mixing parameters $\theta_t = (\mu_t, \sigma_t)$ corresponding to state r. For a given n_r (i.e., conditional on $\mathbf{h}$), we can use results from Antoniak (1974) to explore properties of this prior for different α_r values. For instance, the prior expected number of unique components in the set $\{\theta_t : h_t = r\}$ (i.e., the set of latent mixing parameters, as in Section 2.4, belonging to state r) is approximately $\alpha_r \log[(n_r + \alpha_r)/\alpha_r]$, and this expression may be used to guide prior intuition about the α_r.

3.2.2 Posterior Simulation

Here, we present an MCMC method for posterior inference under the model developed in Section 3.2.1. Detailed expressions for the algorithm follow in Section 3.2.3. We first discuss an adaptation to switching regression of the more standard MCMC approach, involving marginalization of the random mixing distributions G_r in (3.3) over their DP priors, as outlined in Section 2.4. Because it necessitates individual conditional updates for each element of the state vector $\mathbf{h}$, this approach is inefficient. Hence, we propose an alternative MCMC algorithm, which is based on a truncation approximation of each G_r to facilitate a more efficient forward-backward recursive sampling of state vector $\mathbf{h}$.

To obtain the full probability model corresponding to (3.3), we introduce latent parameters $\boldsymbol{\theta} = \{\theta_t = (\mu_t, \sigma_t) : t = 1, ..., T\}$ such that the first stage in (3.3) is replaced with $\mathbf{z}_t \mid \theta_t \overset{ind}{\sim} \mathrm{N}(\mathbf{z}_t; \theta_t)$ and $\theta_t \mid h_t, G_{h_t} \overset{ind}{\sim} G_{h_t}$, for $t = 1, ..., T$. Then, the full posterior, comprising $\boldsymbol{\alpha} = \{\alpha_r : r = 1, ..., R\}$, $\boldsymbol{\psi} = \{\psi_r : r = 1, ..., R\}$, $\mathbf{Q}$, $\mathbf{h}$, $\boldsymbol{\theta}$, and $\{G_r : r = 1, ..., R\}$, is proportional to $\left(\prod_{r=1}^{R} \pi(\alpha_r)\pi(\psi_r)\pi(Q_r)\right) \Pr(\mathbf{h} \mid \mathbf{Q}) \left(\prod_{r=1}^{R} d\mathcal{F}(G_r \mid \alpha_r,\ G_0(\psi_r))\ \prod_{\{t:h_t=r\}} dG_r(\theta_t)\right)\ \left(\prod_{t=1}^{T} \mathrm{N}(\mathbf{z}_t; \theta_t)\right)$, using the fact that, given $\mathbf{h}$, the specification $\theta_t \mid h_t, G_{h_t} \overset{ind}{\sim} G_{h_t}$, $t = 1, ..., T$, can be expressed as $\prod_{r=1}^{R} \prod_{\{t:h_t=r\}} dG_r(\theta_t)$. That is, conditionally on $\mathbf{h}$, the vector of latent mixing parameters breaks down into state-specific subvectors $\{\theta_t : h_t = r\}$, $r = 1, ..., R$, such that the distribution of each is built from independent G_r distributions for the θ_t corresponding to state r. Therefore, for each $\Pr(\{\theta_t : h_t = r\}, G_r \mid \mathbf{h}, \alpha_r, \psi_r) = d\mathcal{F}(G_r \mid \alpha_r, G_0(\psi_r)) \prod_{\{t:h_t=r\}} dG_r(\theta_t)$, we

can apply results from Blackwell and MacQueen (1973) and Antoniak (1974) to write

$$\Pr(\{\theta_t : h_t = r\}, G_r \mid \mathbf{h}, \alpha_r, \psi_r) \quad = \quad d\mathcal{F}(G_r \mid \alpha_r^\star, G_{r0}^\star)\Pr(\{\theta_t : h_t = r\} \mid \mathbf{h}, \alpha_r, \psi_r).$$

Here, $\Pr(\{\theta_t : h_t = r\} \mid \mathbf{h}, \alpha_r, \psi_r)$ is the Pólya urn marginal prior for $\{\theta_t : h_t = r\}$ (induced by marginalizing G_r in $\Pr(\{\theta_t : h_t = r\}, G_r \mid \mathbf{h}, \alpha_r, \psi_r)$ over its DP prior); $\tilde{\alpha}_r = \alpha_r + n_r$ (where $n_r = |\{t : h_t = r\}|$); and $\tilde{G_{r0}}(\cdot) \equiv \tilde{G_{r0}}(\cdot \mid \mathbf{h}, \{\theta_t : h_t = r\}, \alpha_r, \psi_r) = (\alpha_r + n_r)^{-1}\left[\alpha_r dG_0(\cdot\,; \psi_r) + \sum_{\{t:h_t=r\}} \delta_{\theta_t}(\cdot)\right]$.

Hence, the full posterior assumes the form $\Pr(\boldsymbol{\alpha}, \boldsymbol{\psi}, \mathbf{Q}, \mathbf{h}, \boldsymbol{\theta} \mid \mathcal{D}) \prod_{r=1}^R d\mathcal{F}(G_r \mid \tilde{\alpha}_r, \tilde{G}_{r0})$, where $\Pr(\alpha, \psi, \mathbf{Q}, \mathbf{h}, \boldsymbol{\theta} \mid \mathcal{D})$ is equal to

$$\Pr(\mathbf{h} \mid \mathbf{Q}) \left(\prod_{r=1}^R \pi(\alpha_r)\pi(\psi_r)\pi(Q_r)\Pr(\{\theta_t : h_t = r\} \mid \mathbf{h}, \alpha_r, \psi_r)\right) \left(\prod_{t=1}^T \mathrm{N}(\mathbf{z}_t; \theta_t)\right),$$

the marginal posterior corresponding to the finite-dimensional portion of the full parameter vector. This posterior can be sampled by extending standard MCMC techniques for DP mixtures (see, e.g., Neal, 2000). Most of the algorithm involves a straightforward adaptation of the methodology presented in Section (2.4), with the only major change being a joint full conditional draw for each (h_t, θ_t) pair given the incomplete parameter vectors $\mathbf{h}_{-t}$ and $\boldsymbol{\theta}_{-t}$. In particular, given a present value of (h_t, θ_t), a Metropolis-Hastings step is to first propose a new h_t such that

$$\Pr(h_t = r) = \frac{Q_{h_{t-1},r} Q_{r,h_{t+1}}}{\sum_{s=1}^R Q_{h_{t-1},s} Q_{s,h_{t+1}}} \tag{3.5}$$

followed by a proposal for a new $\theta_t = \theta'$, given r, with probability density $(\alpha_r + n_r^-)^{-1}\,[\alpha_r\, dG_0(\theta'\,; \psi_r) + \sum_{\{i \neq t : h_i = r\}} \delta_{\theta_i}(\theta')]$. The move from (h_t, θ_t) to (r, θ') is then accepted with probability $\min\{\mathrm{N}(\mathbf{z}_t; \theta')/\mathrm{N}(\mathbf{z}_t; \theta_t), 1\}$. Using an extension of the approach in Section

(2.4), posterior samples for the full parameter vector can be obtained by augmenting each posterior sample from $\Pr(\boldsymbol{\alpha}, \boldsymbol{\psi}, \mathbf{Q}, \mathbf{h}, \boldsymbol{\theta} \mid \mathcal{D})$ with posterior realizations for the G_r drawn from a truncation approximation to $\mathrm{DP}(\tilde{\alpha}_r, \tilde{G}_{r0})$, for $r = 1, ..., R$.

In general, although sampling for $\boldsymbol{\theta}$ conditional on $\mathbf{h}$ is not difficult, there will be no possible marginal update for $\mathbf{h}$ conditional on $\boldsymbol{\theta}$ (i.e. without conditioning on the G_r). In the posterior simulation approach described above, this forces sampling each individual (h_t, θ_t) from its posterior full conditional even though forward-backward sampling is a substantially more efficient method for exploring the state space (see, e.g., Scott, 2002). Forward-backward sampling in this setting would first require forward calculating for each t the joint probability mass function for states h_t and h_{t+1} conditional on the incomplete dataset $\{\mathbf{z}_1, \ldots, \mathbf{z}_t\}$, the prior transition matrix $\mathbf{Q}$, and the random mixing distributions $\{G_1, \ldots, G_R\}$. This is followed by backward sampling h_t conditional on h_{t+1} for $t = T-1, \ldots, 1$, after first sampling for h_T from its marginal posterior full conditional. It is thus necessary to evaluate state probabilities with respect to the entire G_r distributions, and the necessary calculations must be made using the finite stick-breaking G_r^L, such that the posterior sample for $\mathbf{h}$ is dependent upon the level of truncation. The blocked Gibbs sampling method for DP mixture models (Ishwaran and James, 2001) provides a natural approach wherein the entire MCMC algorithm is based on a finite stick-breaking approximation of the DP. As well as being the consistent choice if the truncated distributions are used in state vector draws, blocked Gibbs can lead to very efficient sampling for the complete posterior (for further discussion of the base algorithm and its properties, look to Ishwaran and James, 2001; Ishwaran and

Zarepour, 2000, 2002).

Using the DP stick-breaking representation, we replace each G_r in model (3.3) with a truncation approximation of the form in (2.8). Specifically, for specified (finite) L, we work with $G_r^L(\cdot) = \sum_{l=1}^{L} p_{r,l}\delta_{\vartheta_{r,l}}(\cdot)$, where the $\vartheta_{r,l} = (\mu_{r,l}, \Sigma_{r,l})$, $l = 1, ..., L$, are i.i.d. $G_0(\psi_r)$, and $\mathbf{p}_r = (p_{r,1}, ..., p_{r,L})$ has distribution $\mathcal{P}_L(\mathbf{p}_r \mid 1, \alpha_r)$ defined in (2.10). Hence, each G_r^L is defined by the set of L location-scale parameters $\boldsymbol{\vartheta}_r = (\vartheta_{r,1}, ..., \vartheta_{r,L})$ and weights $\mathbf{p}_r$. Guidelines to choose the truncation level L for the DP approximation, up to any desired accuracy, can be obtained, e.g., from Ishwaran and Zarepour (2000). For instance, conditional on α_r, the quantity $\sum_{l=B}^{\infty} p_{r,l}$ corresponding to the full DP mixture model has expectation $(\alpha_r/(1+\alpha_r))^{B-1}$ and variance $(\alpha_r/(2+\alpha_r))^{B-1} - (\alpha_r/(1+\alpha_r))^{2(B-1)}$. Thus, based upon these moments and prior guesses for α_r, the truncation L may be chosen to ensure that the residual probability $\sum_{l=L}^{\infty} p_{r,l}$ for each state is acceptably small. Although the modeling and inference framework does not restrict us to a common truncation level L for all states, we consider this setting that results in simpler notation. In practice, it is convenient to choose a single L that is greater than all the desired state-specific truncation levels.

Now the first stage of model (3.3) is replaced with $\mathbf{z}_t \mid h_t, (\mathbf{p}_{h_t}, \boldsymbol{\vartheta}_{h_t}) \overset{ind}{\sim} \sum_{l=1}^{L} p_{h_t,l}\mathrm{N}(\mathbf{z}_t; \vartheta_{h_t,l})$, $t = 1, ..., T$. The limiting case of this finite mixture model (as $L \to \infty$) is the countable DP mixture model $f_{h_t}(\mathbf{z}_t; G_{h_t}) = \int \mathrm{N}(\mathbf{z}_t; \theta) dG_{h_t}(\theta)$ in (3.3). We introduce configuration variables $\mathbf{k} = (k_1, ..., k_T)$, where each k_t takes values in $\{1, ..., L\}$, such that, conditionally on h_t, $\mathbf{z}_t$ given k_t is distributed $\mathrm{N}(\mathbf{z}_t; \vartheta_{h_t,k_t})$. Hence, model (3.3) with the finite stick-breaking approximation can be expressed in the follow-

ing hierarchical form

$$\begin{aligned}
\mathbf{z}_t \mid \boldsymbol{\vartheta}_{h_t}, k_t &\overset{ind}{\sim} \mathrm{N}(\mathbf{z}_t; \vartheta_{h_t,k_t}), \;\; t = 1, ..., T \\
k_t \mid h_t, \mathbf{p}_{h_t} &\overset{ind}{\sim} \sum_{l=1}^{L} p_{h_t,l}\delta_l(k_t), \;\; t = 1, ..., T \\
\mathbf{p}_r, \boldsymbol{\vartheta}_r \mid \alpha_r, \psi_r &\overset{ind}{\sim} \mathcal{P}_L(\mathbf{p}_r \mid 1, \alpha_r) \prod_{l=1}^{L} dG_0(\vartheta_{r,l}; \psi_r), \;\; r = 1, ..., R
\end{aligned} \tag{3.6}$$

with $\mathbf{h} \mid \mathbf{Q} \sim \Pr(\mathbf{h} \mid \mathbf{Q}) = \prod_{t=2}^{T} Q_{h_{t-1},h_t}$, and the hyperpriors for $\{(\alpha_r, \psi_r) : r = 1, ..., R\}$ and $\mathbf{Q}$ given in Section 3.2.1. The full posterior corresponding to model (3.6) is now proportional to

$$\begin{aligned}
\Pr(\mathbf{h} \mid \mathbf{Q}) \prod_{r=1}^{R} \Bigg\{ \pi(\alpha_r)\pi(\psi_r)\pi(Q_r)\mathcal{P}_L(\mathbf{p}_r \mid 1, \alpha_r) \left(\prod_{l=1}^{L} dG_0(\vartheta_{r,l}; \psi_r) \right) \\
\cdot \left(\prod_{\{t:h_t=r\}} \mathrm{N}(\mathbf{z}_t; \vartheta_{r,k_t}) \sum_{l=1}^{L} p_{r,l}\delta_l(k_t) \right) \Bigg\}.
\end{aligned} \tag{3.7}$$

Here, again, the key observation is that, conditionally on $\mathbf{h}$, the first two stages of model (3.6), $\prod_{t=1}^{T} \Pr(\mathbf{z}_t, k_t \mid h_t, (\mathbf{p}_{h_t}, \boldsymbol{\vartheta}_{h_t})) = \prod_{t=1}^{T} \mathrm{N}(\mathbf{z}_t; \vartheta_{h_t,k_t})(\sum_{l=1}^{L} p_{h_t,l}\delta_l(k_t))$, can be expressed in the state-specific form, $\prod_{r=1}^{R} \left\{ \prod_{\{t:h_t=r\}} \mathrm{N}(\mathbf{z}_t; \vartheta_{r,k_t})(\sum_{l=1}^{L} p_{r,l}\delta_l(k_t)) \right\}$. To explore the full posterior, we develop an MCMC approach that combines Gibbs sampling steps for parameters k_t, for $t = 1, ..., T$, and $(\alpha_r, \psi_r, Q_r, \mathbf{p}_r, \boldsymbol{\vartheta}_r)$, for $r = 1, ..., R$, with forward-backward sampling for the state vector $\mathbf{h}$. We discuss the latter next, deferring to Section 3.2.3 the details of a Gibbs sampler for all other parameters.

Regarding the state vector $\mathbf{h}$, by virtue of our sampling the truncated random mixing distributions for each state, we are able to use forward-backward recursive sampling for the full conditional distribution, $\Pr(\mathbf{h}|\mathbf{Q}, \{\boldsymbol{\vartheta}_r, \mathbf{p}_r : r = 1, \ldots, R\}, \mathcal{D})$. Define the $R \times R$ filtering matrix $P^t = P(h_t, h_{t+1}|\ \mathbf{z}_1,\ \ldots,\ \mathbf{z}_t,\ \{\boldsymbol{\vartheta}_r,\ \mathbf{p}_r : r = 1, \ldots, R\},\ \mathbf{Q})$

representing a joint probability mass function for states h_t and h_{t+1} conditional on the incomplete dataset $\{\mathbf{z}_1, \ldots, \mathbf{z}_t\}$, such that row r of P^t defines the vector of probabilities assigned to possible next states for present state $h_t = r$. The forward step is to recursively calculate the matrices $P^3, P^4, \ldots, P^T$ such that

$$\begin{aligned} P^t_{r,s} &= \Pr(h_{t-1} = r, h_t = s | \mathbf{z}_1, \ldots, \mathbf{z}_t, [\boldsymbol{\vartheta}_r, \mathbf{p}_r], [\boldsymbol{\vartheta}_s, \mathbf{p}_s]) \\ &= C_t^{-1} \Pr(\mathbf{z}_t | h_t = s, \boldsymbol{\vartheta}_s, \mathbf{p}_s) \Pr(h_t = s | h_{t-1} = r) \Pr(h_{t-1} = r | \mathbf{z}_1, \ldots, \mathbf{z}_{t-1}, \boldsymbol{\vartheta}_r, \mathbf{p}_r) \\ &= C_t^{-1} \sum_{l=1}^{L} p_{s,l} \mathrm{N}(\mathbf{z}_t; \vartheta_{s,l}) Q_{r,s} \sum_{i=1}^{R} P^{t-1}_{i,r}, \end{aligned} \tag{3.8}$$

with P^2 such that each element $P^2_{r,s}$ is equal to $\Pr(h_1 = r,\ h_2 = s \mid \mathbf{z}_1,\ \mathbf{z}_2,\ [\boldsymbol{\vartheta}_r,\ \mathbf{p}_r],\ [\boldsymbol{\vartheta}_s,\ \mathbf{p}_s]) = C_2^{-1} \sum_{l=1}^{L} p_{s,l}\ \mathrm{N}(\mathbf{z}_2;\ \vartheta_{s,l})\ Q_{r,s}\ \sum_{l=1}^{L} p_{r,l}\ \mathrm{N}(\mathbf{z}_1;\ \vartheta_{r,l})$, and where each C_t is a constant such that the sum of the elements of P^t is one. The backward sampling begins by drawing h_T from $\Pr(h_T = r | \{\boldsymbol{\vartheta}_r, \mathbf{p}_r : r = 1..R\}, \mathbf{Q}, \mathcal{D}) = \sum_{s=1}^{R} P^T_{s,r}$. Then, for $t = T-1, \ldots, 1$, we have that $\Pr(h_t = r | h_{t+1}, \{\boldsymbol{\vartheta}_r, \mathbf{p}_r : r = 1, \ldots, R\}, \mathbf{Q}, \mathcal{D}) \propto P^{t+1}_{r,h_{t+1}}$. In this way, $\mathbf{h}$ is sampled at once conditional on the entire data set and conditional on the truncated random mixing distributions.

3.2.3 Blocked Gibbs MCMC Algorithm Details

Here, we detail the approach to MCMC posterior simulation discussed in Section 3.2.2. Recall that the key to the finite stick-breaking algorithm is that we are able to use forward-backward recursive sampling of the posterior conditional distribution for $\mathbf{h}$ as described in Section 3.2.2. Gibbs sampling details for all other parameters of model (3.6) are provided below, in an extension of the general blocked Gibbs MCMC scheme

of Ishwaran and James (2001).

First, for each $t = 1, ..., T$, k_t has a discrete posterior full conditional distribution with values in $\{1, ..., L\}$ and corresponding probabilities $p_{h_t,l}$ $\mathrm{N}(\mathbf{z}_t;\ \vartheta_{h_t,l})$ / $\{\sum_{b=1}^{L} p_{h_t,b}\ \mathrm{N}(\mathbf{z}_t;\ \vartheta_{h_t,b})\}$, for $l = 1, ..., L$.

For each $r = 1, ..., R$, the posterior full conditional distribution for $\mathbf{p}_r$ is proportional to $\mathcal{P}_L(\mathbf{p}_r \mid 1, \alpha_r) \prod_{\{t:h_t=r\}} \left(\sum_{l=1}^{L} p_{r,l}\delta_l(k_t)\right) = \mathcal{P}_L(\mathbf{p}_r \mid 1, \alpha_r) \prod_{l=1}^{L} p_{r,l}^{H_{r,l}}$, where $H_{r,l} = \sum_{t=1}^{T} \delta_{[h_t=r,k_t=l]}$. Note that the $\mathcal{P}_L(\mathbf{p}_r \mid 1, \alpha_r)$ prior for $\mathbf{p}_r$, defined constructively in (2.10), is given by

$$\mathcal{P}_L(\mathbf{p}_r \mid 1, \alpha_r) = \alpha_r^{L-1} p_{r,L}^{\alpha_r - 1}(1-p_{r,1})^{-1}(1-(p_{r,1}+p_{r,2}))^{-1} \ldots \left(1 - \sum_{l=1}^{L-2} p_{r,l}\right)^{-1}. \quad (3.9)$$

Recall the generalized Dirichlet distribution $\mathrm{GD}(\mathbf{p}; \mathbf{a}, \mathbf{b})$ (Connor and Mosimann, 1969) for random vector $\mathbf{p} = (p_1, ..., p_L)$, supported on the L dimensional simplex, with density proportional to

$$p_1^{a_1-1} \ldots p_{L-1}^{a_{L-1}-1} p_L^{b_{L-1}-1} (1-p_1)^{b_1-(a_2+b_2)} \ldots (1-(p_1+\ldots+p_{L-2}))^{b_{L-2}-(a_{L-1}+b_{L-1})},$$

where the parameters are $\mathbf{a} = (a_1, ..., a_{L-1})$ and $\mathbf{b} = (b_1, ..., b_{L-1})$. Then, $\mathcal{P}_L(\mathbf{p}_r \mid 1, \alpha_r) \equiv \mathrm{GD}(\mathbf{p}_r; \mathbf{a}, \mathbf{b})$ with $\mathbf{a} = (1, ..., 1)$ and $\mathbf{b} = (\alpha_r, ..., \alpha_r)$. Moreover, the $\prod_{l=1}^{L} p_{r,l}^{H_{r,l}}$ form is also proportional to a $\mathrm{GD}(\mathbf{p}_r; \mathbf{a}, \mathbf{b})$ distribution with $\mathbf{a} = (H_{r,1}+1, ..., H_{r,L-1}+1)$ and $\mathbf{b} = ((L-1) + \sum_{l=2}^{L} H_{r,l}, ..., 2 + H_{r,L-1} + H_{r,L}, 1 + H_{r,L})$. Hence, the posterior full conditional for $\mathbf{p}_r$ can be completed to a generalized Dirichlet distribution with parameters $\mathbf{a} = (H_{r,1}+1, ..., H_{r,L-1}+1)$ and $\mathbf{b} = (\alpha_r + \sum_{l=2}^{L} H_{r,l}, \alpha_r + \sum_{l=3}^{L} H_{r,l}, ..., \alpha_r + H_{r,L})$. This distribution can be sampled constructively by first drawing independent

$v_{r,l} \sim \beta(1 + H_{r,l}, \alpha_r + \sum_{b=l+1}^{L} H_{r,b})$, for $l = 1, ..., L-1$, and then setting $p_{r,1} = v_{r,1}$; $p_{r,l} = v_{r,l} \prod_{b=1}^{l-1}(1 - v_{r,b})$, $l = 2, ..., L-1$; and $p_{r,L} = 1 - \sum_{l=1}^{L-1} p_{r,l}$.

Next, for each $r = 1, ..., R$, the posterior full conditional distribution for $\boldsymbol{\vartheta}_r$ is

$$\Pr\left(\boldsymbol{\vartheta}_r | \mathbf{k}, \alpha_r, \psi_r, \{\mathbf{z}_t : h_t = r\}\right)$$

$$\propto \prod_{j=1}^{n_r^\star} \left[dG_0(\vartheta_{r,j}^\star; \psi_r) \prod_{t:h_t=r, k_t^\star=j} \mathrm{N}(\mathbf{z}_t^r; \vartheta_{r,j}^\star) \right] \prod_{j=1}^{n_r^u} dG_0(\vartheta_{r,j}^u | \psi_r). \tag{3.10}$$

Here, each $\boldsymbol{\vartheta}_r$ has been partitioned into subsets $\boldsymbol{\vartheta}_r^\star = \{\vartheta_{r,l} : H_{r,l} > 0\}$ and $\boldsymbol{\vartheta}_r^u = \{\vartheta_{r,l} : H_{r,l} = 0\}$, and we introduce the index vector $\mathbf{k}^\star = [k_1^\star, \ldots, k_T^\star]$ defined such that observation $\mathbf{z}_t$ is allocated to component $\vartheta_{h_t,k_t^\star}^\star$. For $r = 1, ..., R$, $n_r^\star$ is the number of elements in $\boldsymbol{\vartheta}_r^\star$ (i.e., the number of distinct values of k_t that correspond to the r-th state), and n_r^u denotes the number of elements in $\boldsymbol{\vartheta}_r^u$ (i.e., the number of unallocated parameters corresponding to the r-th state). We can thus sample independently each $\vartheta_{r,j}^u \sim G_0(\psi_r)$ for $j = 1, \ldots, n_r^u$. With respect to the remaining (allocated) parameters, the posterior full conditional for each $\vartheta_{r,j}^\star \equiv (\mu_{r,j}^\star, \Sigma_{r,j}^\star)$ is proportional to

$$\mathrm{N}(\mu_{r,j}^\star; m_r, V_r)\mathrm{W}_{\nu_r}(\Sigma_{r,j}^{\star -1}; S_r^{-1}) \prod_{\{t:h_t=r, k_t^\star=j\}} \mathrm{N}(\mathbf{z}_t; \mu_{r,j}^\star, \Sigma_{r,j}^\star),$$

and can be sampled through draws from the full conditional for $\mu_{r,j}^\star$ and for $\Sigma_{r,j}^{\star -1}$. The former is normal with covariance matrix $T_j = (V_r^{-1} + H_{r,j}^\star \Sigma_{r,j}^{\star -1})^{-1}$, where $H_{r,j}^\star = \sum_{t=1}^{T} \delta_{[h_t=r, k_t^\star=j]}$, and mean vector $T_j(V_r^{-1} m_r + \Sigma_{r,j}^{\star -1} \sum_{\{t:h_t=r, k_t^\star=j\}} \mathbf{z}_t)$. The latter is $\mathrm{W}_{\nu_r + H_{r,j}^\star}(\cdot\ ;\ (S_r + \sum_{\{t:h_t=r, k_t^\star=j\}} (\mathbf{z}_t - \mu_{r,j}^\star)(\mathbf{z}_t - \mu_{r,j}^\star)^T)^{-1})$.

The draw for $\psi_r = (m_r, V_r, S_r)$ for each r is facilitated by noticing that, as seen in (3.10), the unallocated $\vartheta_{r,l}$ have just been drawn from $G_0(\psi_r)$ and we can marginalize the joint conditional posterior for $\{\boldsymbol{\vartheta}_r, \psi_r\}$ over $\boldsymbol{\vartheta}_r^u$ to get the posterior

full conditional $\Pr(\psi_r|\boldsymbol{\vartheta}_r^\star,\ n_r^\star) \propto \pi(\psi_r)\ \prod_{j=1}^{n_r^\star}\ dG_0(\vartheta_{r,j}^\star;\psi_r)$ Hence, ψ_r can be updated by separate draws from the posterior full conditionals for m_r, V_r, and S_r. The full conditional for m_r is normal with covariance matrix $\tilde{B}_{m_r} = (B_{m_r}^{-1} + n_r^\star V_r^{-1})^{-1}$ and mean vector $\tilde{B}_{m_r}(B_{m_r}^{-1}a_{m_r} + V_r^{-1}\sum_{j=1}^{n_r^\star}\mu_{r,j}^\star)$. The full conditional for V_r^{-1} is $\mathrm{W}_{n_r^\star+a_{V_r}}(\cdot\,;\ (B_{V_r} + \sum_{j=1}^{n_r^\star}(\mu_{r,j}^\star - m_r)(\mu_{r,j}^\star - m_r)^T)^{-1})$, and the full conditional for S_r is $\mathrm{W}_{\nu_r n_r^\star + a_{S_r}}(\cdot\,;\ [B_{S_r}^{-1} + \sum_{j=1}^{n_r^\star}\Sigma_{r,j}^{\star -1}]^{-1})$.

Regarding the DP precision parameters, combining the $\mathrm{Ga}(a_{\alpha_r}, b_{\alpha_r})$ prior for α_r with the relevant terms from (3.9), we obtain that, for each $r = 1, ..., R$, the posterior full conditional for α_r is a $\mathrm{G}(a_{\alpha_r} + L - 1, b_{\alpha_r} - \log(p_{r,L}))$ distribution.

Finally, with the $\mathrm{Dir}(Q_r; \lambda_r)$ prior on each row Q_r of the transition matrix $\mathbf{Q}$, the posterior full conditional for Q_r is $\mathrm{Dir}(Q_r; \lambda_r + J_r)$, where $J_r = (J_{r,1}, ..., J_{r,R})$ with $J_{r,s}$ denoting the number of transitions from state r to state s, which are defined by the currently imputed state vector $\mathbf{h}$.

3.2.4 Extension to Semiparametric Modeling with External Covariates

In the spirit of allowing the switching probabilities to drive the nonparametric regression, we extend here the methodology to include additional information about the state vector. This leads to a nonhomogeneous hidden Markov mixture where, interpreted in the context of our model, the hidden state provides a link between the joint covariate-response random variable $\mathbf{z}$ and an external covariate vector $\mathbf{u} = \{u_1, \ldots, u_T\}$. Although we develop the methodology with respect to a single covariate, the work is

easily extendible to the setting of multiple external covariates. The standard nonhomogeneous hidden Markov model holds that the switching probabilities are dependent upon the external covariates, such that $\Pr(h_t \mid h_1, \ldots, h_{t-1}, \mathbf{u}) = \Pr(h_t \mid h_{t-1}, u_t)$. Berliner and Lu (1999) present a Bayesian approach to nonhomogeneous hidden Markov models in which $\Pr(h_t|h_{t-1}, u_t)$ is estimated through *probit* regression. Our work is more closely related to the likelihood analysis of Hughes and Guttorp (1994), wherein a heuristic argument using Bayes theorem is proposed to justify the model $\Pr(h_t|h_{t-1}, u_t) \propto \Pr(h_t|h_{t-1})\mathcal{L}(h_t; u_t)$, where the likelihood $\mathcal{L}(h_t; u_t)$ in their example is normal with state dependent mean (they also consider a more complicated form based on $\mathcal{L}(h_t, h_{t-1}; u_t)$). Treating each u_t as the realization of a random variable is quite natural in the context of our approach. Hence, we obtain a similar model by adding a further stage, $u_t|h_t \overset{ind}{\sim} \mathrm{k}_u(u_t; \varphi_{h_t})$ where k_u is a density function and $\boldsymbol{\varphi} = \{\varphi_1, \ldots, \varphi_R\}$ are state-specific parameterizations, to the prior specification for model (3.3) and assuming that $\mathbf{u}$ is conditionally independent of $\mathbf{z}_1, \ldots, \mathbf{z}_T$ given $\mathbf{h}$. Thus, for $t = 1, \ldots, T$, the top level specification of the model becomes

$$\mathbf{z}_t, u_t \mid h_t, G_{h_t}, \boldsymbol{\varphi} \overset{ind}{\sim} f_{h_t}(\mathbf{z}_t; G_{h_t})\mathrm{k}_u(u_t; \varphi_{h_t}). \tag{3.11}$$

Clearly this implies that the hidden Markov chain is nonhomogeneous conditional on $\mathbf{u}$. But unconditionally in the prior, it is more accurate to say that both $\mathbf{z}_1, \ldots, \mathbf{z}_T$ and $\mathbf{u}$ are dependent upon a shared homogeneous Markov chain and that they are conditionally independent given $\mathbf{h}$.

This extension is easily implemented within the MCMC algorithm of Section

3.2.2. In general, the $\boldsymbol{\varphi}$ parameter set will be sampled conditional on only $\mathbf{u}$ and the state allocation $\mathbf{h}$. And given $\boldsymbol{\varphi}$, only the MCMC draws that involve $\mathbf{h}$ need to be altered. In detail, the forward probability matrix calculation of (3.8) becomes

$$\begin{aligned} P_{r,s}^{t} &= \Pr(\mathbf{z}_t|h_t=s,\boldsymbol{\vartheta}_s,\mathbf{p}_s)\Pr(h_t=s|h_{t-1}=r,u_t)\Pr(h_{t-1}=r|\boldsymbol{\vartheta}_r,\mathbf{p}_r,\{\mathbf{z}_i,u_i\}_{i=1}^{t-1}) \\ &= \frac{1}{\mathrm{C_t}}\sum_{l=1}^{L} p_{s,l}\mathrm{N}(\mathbf{z}_t|\vartheta_{s,l})\mathrm{k}_u(u_t|\varphi_s)Q_{r,s}\sum_{i=1}^{R} P_{ir}^{t-1}, \end{aligned} \tag{3.12}$$

with P^2 defined such that each $P_{r,s}^2$ is equal to $C_2^{-1}\sum_{l=1}^{L} p_{s,l}\ \mathrm{N}(\mathbf{z}_2;\ \vartheta_{s,l})\ \mathrm{k}_u(u_2;\varphi_s)\ Q_{r,s}$ $\sum_{l=1}^{L} p_{r,l}\ \mathrm{N}(\mathbf{z}_1;\ \vartheta_{r,l})\ \mathrm{k}_u(u_1;\varphi_r)$ and the C_t updated as normalizing constants.

3.2.5 Analysis of Stock-Recruitment Relationships Subject to Environmental Regime Shifts

The relationship between the number of mature individuals of a species (stock) and the production of offspring (recruitment) is fundamental to the behavior of any ecological system. This has special relevance in fisheries research, where the stock-recruitment relationship applies directly to decision problems of fishery management with serious policy implications (Quinn and Derisio, 1999; Bravington et al., 2000). A standard ecological modeling assumption holds that as stock abundance increases, successful recruitment per individual (reproductive success) decreases. However, a wide variety of parameters will influence this reproductive relationship and there are many competing models for the influence of biological and physical mechanisms. Munch et al. (2005) present an overview of the literature on parametric modeling for stock-recruitment functions, arguing for the utility of standard semiparametric Gaussian pro-

cess regression modeling. In the same spirit, albeit under the more general DP mixture modeling framework developed in Sections 2.1, 3.2.1, and 3.2.4, our focus is to allow flexible regression to capture the nature of recruitment dependence upon stock without making parametric assumptions for either the stock-recruitment function or the errors around it.

An added complexity in studying stock-recruitment relationships is introduced by ecosystem *regime switching*. It has been observed that rapid shifts in the ecosystem state can occur, during which the parameters governing relationships such as that between stock and recruitment will undergo major change. This has been observed in the North Pacific in particular (McGowan et al., 1998; Hare and Mantua, 2000). Although empirical evidence of regime shifts is well documented and there have been attempts to establish mechanisms for the effect of this switching on stock-recruitment (e.g. Jacobson et al., 2005), the relationship between the physical effects of regime shifts and their biological manifestation is still unclear. This presents an ideal setting for Markov-dependent switching regression models due to their ability to link observed processes that occur on different scales (in this case, biological and physical) and are correlated in an undetermined manner.

Annual stock and recruitment for Japanese sardine from 1951-1991 will be used to illustrate the DP hidden Markov switching regression model. Wada and Jacobson (1998) use modeling of catch abundance and egg count samples to estimate R, the successful recruits of age less than one (in multiples of 10^6 fish). With estimated annual egg production E (in multiples of 10^{12} eggs) used as a proxy for stock abundance, they

investigate the relationship between $\log(E)$ and $\log(R/E)$, the log of the proportion of eggs that survive. Japanese sardine have been been observed to switch between *favorable* and *unfavorable* feeding regime states related to the North Pacific environmental regime switching discussed above. Based upon a predetermined regime allocation, Wada and Jacobson fit a linear regression relationship for $\log(E)$ vs $\log(R/E)$ within each regime.

We consider an analysis of the Japanese sardine data using the modeling framework developed in Section 3.2.1, which relaxes the parametric linear regression assumption and allows for simultaneous estimation of regime state allocation and regime-specific stock-recruitment relationships. The joint distribution of $\mathbf{z} = (\log(E), \log(R/E))$ is assigned the hierarchical prior specification of model (3.3), where the underlying state of either favorable or unfavorable feeding regimes constitutes a first order Markov chain. The DP precision hyperparameters are $a_\alpha = 2$ and $b_\alpha = 0.2$. The prior for ψ_r is specified as outlined in Section 3.2.2 such that, conditional on the prior regime allocation taken from Wada and Jacobson, a_{m_1} and a_{m_2} are set to data means $(5, 3)$ and $(5, 5)$ for the unfavorable and favorable regime observations respectively while B_{m_1} and $(a_{V_1} - 3)^{-1}B_{V_1}$, with diagonal $(5.3, 2.6)$ and off-diagonal -3.1; and B_{m_2} and $a_{V_2}^{-1}B_{V_2}$, with diagonal $(4.5, 1.4)$ and off-diagonal -2.0; are the observed covariance matrix for each regime. The B_{S_r} are each a diagonal matrix with diagonal $(7.8, 7.7)$, one quarter of the data range, and $\nu, a_v^1, a_s^1, a_v^2, a_s^2 = 2(d + 1) = 6$. The prior for $\mathbf{Q}$ is implied by assuming a $\text{Be}(3, 1.5)$ prior for the probability of staying in the same state, which reflects the relative rarity of regime shifts. The data and prior allocation are shown in Figure 3.9 along with draws from a normal distribution for each regime state with mean

$\mathbb{E}_{G_0}[\mu] = a_{m_r}$ and variance $\text{var}_{G_0}(\mu \mid V_r = \mathbb{E}_\pi[V_r]) = B_{m_r} + (a_{V_r} - 3)^{-1} B_{V_r}$. These draws illustrate prior expectation of the random mixing distribution for the μ_r and, noting that this does not include prior uncertainty in the μ_r due to the DP mixture or variability in the prior for V_r, clearly show that the prior specification has not overly restricted mixture components.

As described above, the sardine feeding regime is part of a larger ecosystem state for this region of the North Pacific. The physical variables that are linked to the ecosystem state switching can be used as external covariates for the hidden Markov chain, as outlined in Section 3.2.4. For the purpose of illustration, we choose a single physical variable, the winter average Pacific decadal oscillation (PDO) index, which is highly correlated with biological regime switching (Hare and Mantua, 2000), to act as an external covariate. The PDO index provides the first principle component of an aggregate of North Pacific sea surface temperatures. Although it is not directly responsible, sea surface temperature is believed to be a proxy for mechanisms such as current flow (MacCall, 2002) that control the regime switching. The sardine data were analyzed for the model proposed in (3.2.4) including winter average PDO from 1951 to 1991 as the vector $\mathbf{u}$, a single external covariate assumed to be conditionally independent of $\log(E)$ and $\log(R/E)$ given the regime allocation $\mathbf{h}$. The above formulation for model (3.3) was thus augmented by the statement, $u_t \mid h_t \overset{ind}{\sim} \text{N}(u_t; \gamma_{h_t}, \tau^{-2})$, with conjugate normal priors for $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2\}$ and gamma prior for τ^{-2}, such that $\gamma_1 \sim \text{N}(-0.44, 0.26)$, $\gamma_2 \sim \text{N}(0.73, 0.26)$, and $\tau^2 \sim \text{Ga}(0.5, 0.125)$. The γ_r mean values are average winter PDO for two ten year periods that are generally accepted to fall within each ecosystem

regime (Hare and Mantua, 2000), γ_r variance is the pooled variance for these mean estimates, and the expectation for τ^{-2} is chosen to give some overlap between prior PDO densities for each regime. During posterior Gibbs sampling, we are able to draw directly from normal and gamma posterior distributions for $\boldsymbol{\gamma}$ and τ^2 conditional on $\mathbf{u}$ and $\mathbf{h}$. Posterior samples of these parameters are shown in Figure 3.11.

Our analyses of the data, both with and without PDO as an external variable, are based on MCMC runs of 30,000 iterations, with prediction after a burn-in of 5000 iterations. A truncation of $L = 100$ was used in the stick-breaking prior. The results are presented in Figures 3.10 through 3.13. The posterior mean implied conditional densities for each regime, evaluated over a 50×50 grid, are shown in Figure 3.10. These act as a point estimate of the conditional relationship between stock and recruitment for each regime. Figure 3.12 shows both the posterior mean for the state vector $\mathbf{h}$, which can be interpreted as the mean probability for inclusion of each year in the favorable regime, and the posterior sample of mean regression functions for each regime, which were sampled through equation (2.11). These Figures illustrate the ability of the models to fit nonlinear curves and capture nonstandard features of the response distribution, such as heavy tails and bimodality. In each figure, we see a clear separation between the conditional densities corresponding to the two recruitment regimes. The impact of inclusion of PDO as an external variable is also clearly evident. In the absence of such information, the observations for years 1988-1991 are likely to be allocated in the favorable regime due to the rarity of regime shifting (i.e., due to posterior realizations of $\mathbf{Q}$ which put a high probability on staying in the same state). However, given the inclusion

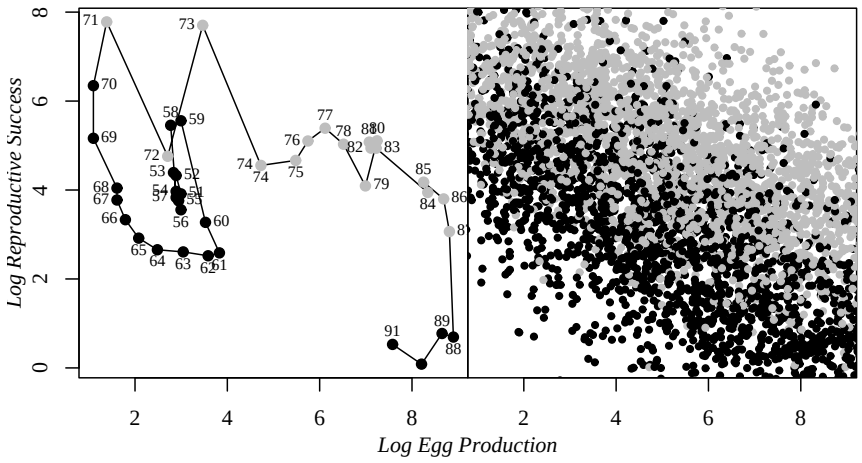


Figure 3.9: Japanese sardine data (left) and draws from $\mathrm{N}(a_m^r, B_m^r + (a_V^r - 3)^{-1} B_v^r)$, with favorable regime in grey.

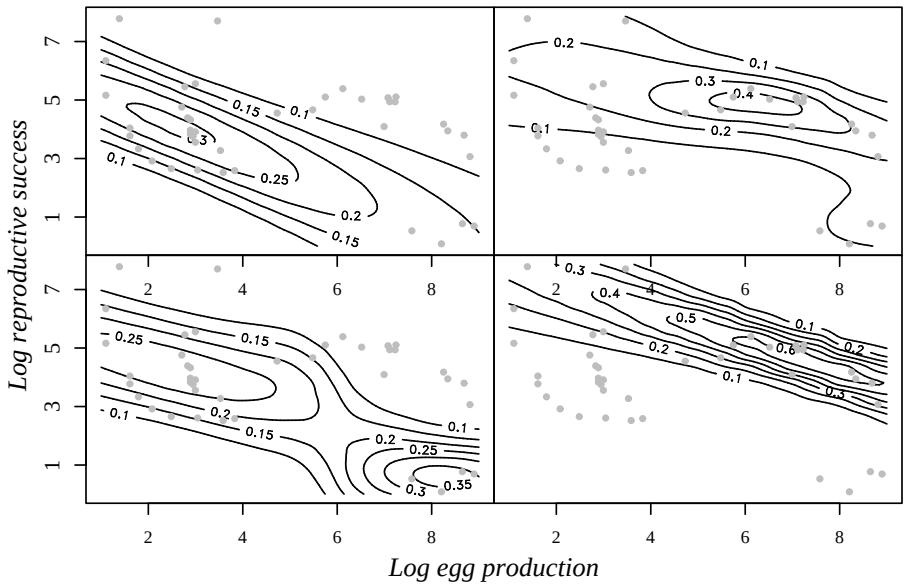


Figure 3.10: Japanese sardine data. Mean posterior conditional density surface, $\mathbb{E}[\, f(\log(R/E) \mid \log(E); G_r) \mid \mathcal{D}]$, for each regime. The unfavorable regime is plotted on the left and favorable on the right, while the top row corresponds to analysis from the base model and the bottom row to the extended model including PDO as an external covariate.

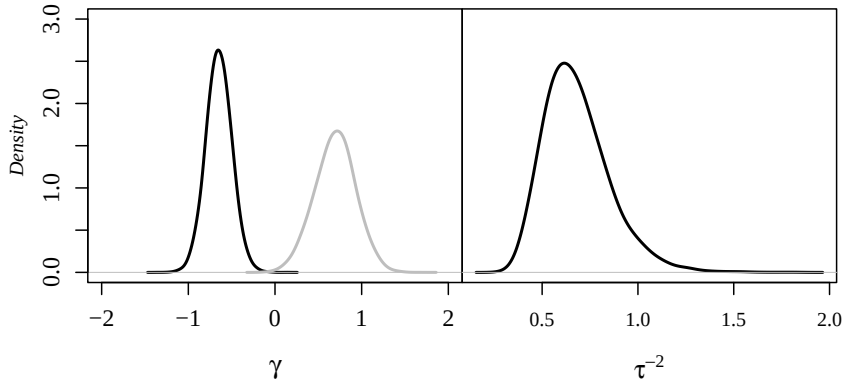


Figure 3.11: Japanese sardine data. Two component normal mixture fit to the winter PDO averages, with posterior distributions for mean parameters on the left (favorable regime in grey), and for variance on the right.

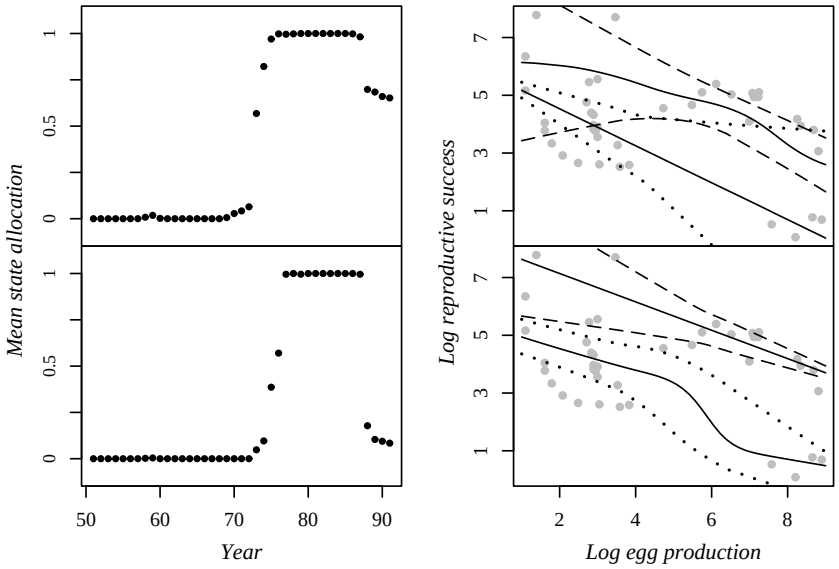


Figure 3.12: Japanese sardine data. The posterior mean regime membership by year is plotted on the left (zero = unfavorable), and the posterior mean regression samples from $\Pr\left(\mathbb{E}[\ \log(R/E) \mid \log(E);\ G_r]\,|\mathcal{D}\right)$ for each regime are on the right (90% posterior intervals are included within dashed lines for the favorable regime, and within dotted lines for the unfavorable). The top row corresponds to analysis from the base model and the bottom row to the extended model including PDO as an external covariate.

of PDO in our model, these years are more probably associated with the unfavorable regime. Also, posterior intervals for the mean regression curves corresponding to each regime do not exclude the possibility of a linear mean relationship between log egg production and log reproductive success. These last two points would indicate that we do not have sufficient evidence to reject the original assumptions of Wada and Jacobson.

Finally, Figure 3.13 shows prediction for 1992, the year following the end of our dataset, both with and without the inclusion of winter PDO as an external covariate. Prediction for both the regime state $h_{T+\Delta}$ and the conditional density $f(\log(R/E) \mid \log(E)\ ;\ G^L_{h_{T+\Delta}})$ in future years is possible conditional on each draw of the transition matrix $\mathbf{Q}$, the state h_T, and the finite stick-breaking mixtures G^L_1 and G^L_2. Posterior mean estimates of $f(\log(R/E) \mid \log(E)\ ;\ G^L_{h_{1992}})$ are shown on the left hand side of figure 3.13 for modeling both with and without inclusion of PDO as an external covariate ($u_{1992} = 0.26$). In addition, the data include egg production estimates for the years 1992 to 1995 with $\log(E) = \log(675) = 6.515$ in 1992. Wada and Jacobson found a recruitment estimate based on partial stock assessment for 1992 of $R = 20591$ ($\log(20591/675) \approx 3.4$), but they were unable to quantify uncertainty about this value. We are able to show full posterior samples for the density $f(y|x = \log(675); G_{h_{1992}})$. A comparison of plots on the right hand side of Figure 3.13 shows the considerable change in predicted response distribution (conditional on $\log(E) = 6.515$ in 1992) from the inclusion of winter PDO in the model. The mean posterior conditional uncertainty is actually increased with the inclusion of PDO, as it is now more likely that the regime state will be unfavorable in 1992 and there is little information from the data to inform

the unfavorable regime regression curve around $\log(E) = 6.515$. In each case, the right-hand plots for prediction conditional on $E = 675$ illustrate the posterior variability of the implied conditional density, and the quantification of this variability is an important aspect of the fully nonparametric modeling.

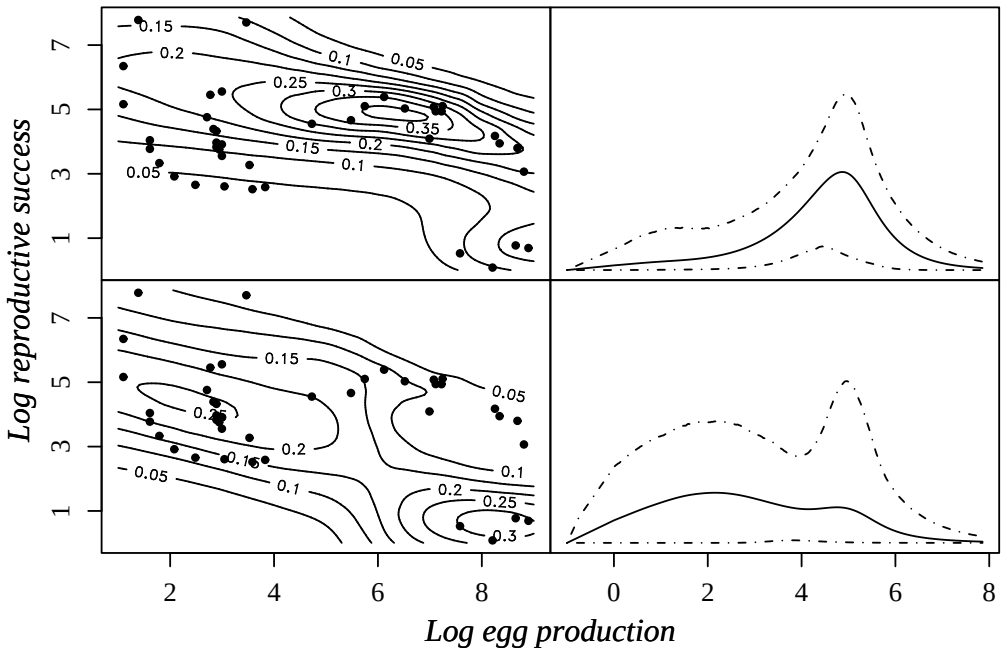


Figure 3.13: Japanese sardine data; prediction for 1992. The mean conditional density $\mathbb{E}\left[f(\log(R/E) \mid \log(\ E\ )\ ;\ G_{h_{1992}}) \mid \mathcal{D}\right]$ is shown on the left and a posterior sample, conditional on an egg count of 675, from $\Pr\left(f(\log(R/E) \mid \log(E) = 6.515;\ G_{h_{1992}}) \,|\mathcal{D}\right)$ is plotted on the right (the point estimate is shown as a solid line and dashed lines contain a 90% posterior interval). The bottom plots correspond to analysis with the inclusion of PDO as an external covariate, and prediction is conditional on a winter average PDO of 0.26 in 1992.

3.3 Regression for Survival Data

Survival data is characterized through realizations of random variables with positive support. The data tends to be partially observed due to censoring, which arises in a variety of different manners. Right censoring presents the most common issue, and this occurs when the study of interest terminates while some of the observed processes continue to survive. Left and interval censoring are also prevalent, especially in the setting of clinical trials or when continuous lifetime processes are not observed continuously.

Bayesian nonparametric modeling for survival data has an extensive history. Indeed, much of the early inferential work in Bayesian nonparametrics was focused on modeling with random priors over the space of survival functions. Susarla and Van Ryzin (1976) presented one of the first DP-based approaches, and the work of Ferguson and Phadia (1979) is an early example of inference through the use of more general neutral-to-the-right process priors. More recent work has involved the use of Pólya tree priors (Muliere and Walker, 1997) and beta-Stacy processes (Walker and Damien, 1998) to model the survival function, or DP mixture priors for the density function of the survival distribution (Kottas, 2006). Other authors have investigated nonparametric modeling through use of priors on the space of cumulative hazard functions (e.g., Hjort, 1990; Damien et al., 1996) or hazard functions (e.g., Nieto-Barajas and Walker, 2002).

We seek to develop a unified fully nonparametric framework (unified in that all functionals of interest are available through a single inferential process for response

densities) for multivariate survival data in the presence of a regression component. Previous Bayesian nonparametric work related to regression for survival data has generally been limited to specification of nonparametric prior components within a parametric regression model. In particular, many authors have considered semiparametric modeling for accelerated failure time models, typically assuming a parametric form for the regression function conditional on a nonparametrically modeled additive error term. DP priors form the basis for early work along this line by Johnson and Christensen (1989), whereas DP mixture priors are utilized in the more recent work of Kuo and Mallick (1997), Kottas and Gelfand (2001), Merrick et al. (2003), and Hanson (2006). Another semiparametric approach arises through nonparametric hazard function estimation combined with a parametric proportional hazards model for the effect of covariates. Gelfand and Mallick (1995), Laud et al. (1998), Mallick and Walker (2003), and De Blasi and Hjort (2007) describe model development in this spirit, while Hanson and Yang (2007) present a related approach wherein the proportional hazards assumption is replaced by a proportional odds restriction. Finally, Mallick et al. (1999) present a nonlinear hazard regression model based on multivariate adaptive regression splines.

In contrast to these approaches, we obtain inference for survival and hazard functions based upon fully nonparametric modeling for the conditional distribution for survival responses given covariates. This section outlines a Bayesian nonparametric framework, again based upon DP mixture joint covariate-response density estimation. Distinguishing features of the proposed approach for survival regression include model adjustments to handle censoring in the response as well as illustration of the general

approach with multivariate responses. The general modeling framework will be presented in Section 3.3.1 and illustration of the methodology is presented through a data example in Section 3.3.2. The example of this latter section involves interval, right, and left censored bivariate response for a single binary covariate. Posterior simulation details will be provided in the context of this example.

3.3.1 Model Development

As in previous sections, we propose an approach to the general regression problem based on DP mixture modeling for the joint distribution of responses and covariates. In this development, we will consider data corresponding to a multivariate response $\mathbf{Y} = \{Y_1, \ldots, Y_{d_y}\}$ with covariates $\mathbf{X} = \{X_1, \ldots, X_{d_x}\}$. Each response variable is assumed to have support on the positive real line and may be either fully observed, interval censored, right censored, or left censored. Thus the data sets that can be handled under the proposed modeling framework are of the form $\mathcal{D} = \{\mathbf{x}_i = (x_{i1}, \ldots, x_{id_x}), \mathbf{z}_i = (z_{i1}, \ldots, z_{id_y})\colon i = 1, \ldots, n\}$ with response j for observation $i \in \mathcal{I}_O^j$ fully observed, $i \in \mathcal{I}_{LC}^j$ left censored, $i \in \mathcal{I}_{RC}^j$ right censored, and $i \in \mathcal{I}_{IC}^j$ interval censored. Accordingly, $z_{ij} = y_{ij}$ for $i \in \mathcal{I}_O^j$ and for censored observations, $z_{ij} = (a_{ij}, b_{ij})$ such that $y_{ij} \in (a_{ij}, b_{ij})$ where $a_{ij} = 0$ for $i \in \mathcal{I}_{LC}^j$ and $b_{ij} = \infty$ for $i \in \mathcal{I}_{RC}^j$.

The data generating probability density is modeled as a random mixture of kernels, where the joint kernel for $\mathbf{y}$ and $\mathbf{x}$ is built through d_y independent probability densities for each response variable and a d_x dimensional probability density (possibly also a product of independent components) for $\mathbf{x}$. The d_y response kernel components

may assume the form of any probability density with support over the positive real line. Ibrahim et al. (2001) discuss multiple parametric models for univariate survival data, including gamma, log-normal, and Weibull probability densities, and any of these may be used to build the individual response kernel components. The kernel component for $\mathbf{x}$ can be constructed according to the framework of Chapter 2, which contains guidance for the modeling of both continuous and discrete covariates.

The Weibull density, $\mathrm{k}_w(y;\gamma,\lambda) = \lambda^{-1}\gamma y^{\gamma-1}\exp\left[-y^{\gamma}/\lambda\right]$, is a very convenient choice for the modeling of censored data due to the closed form of its cumulative distribution function ($\mathrm{K}_w(y;\gamma,\lambda) = 1 - \exp\left[-y^{\gamma}/\lambda\right]$). Indeed, the Weibull is a standard choice for parametric modeling of survival data. As such, we will use d_y Weibull densities in construction of our joint covariate-response mixture kernel. Hence, the mixture model is expressed as

$$\begin{aligned} \mathbf{x},\mathbf{y}|G &\sim \int \left(\prod_{j=1}^{d_y} \mathrm{k}_w(y_j;\gamma_j,\lambda_j)\right) \mathrm{k}_{\mathbf{x}}(\mathbf{x};\theta)\, dG\left(\Gamma,\Lambda,\theta\right) \\ G &\sim \mathrm{DP}\left(\alpha, G_0^{\mathbf{y}}(\Gamma,\Lambda;\boldsymbol{\phi},\boldsymbol{\psi}) G_0^{\mathbf{x}}(\theta;\rho)\right), \end{aligned} \tag{3.13}$$

where $\Gamma = \{\gamma_1,\ldots,\gamma_{d_y}\}$ and $\Lambda = \{\lambda_1,\ldots,\lambda_{d_y}\}$. The centering distribution component corresponding to the response vector is $G_0^{\mathbf{y}}(\Gamma,\Lambda;\,\boldsymbol{\phi},\,\boldsymbol{\psi}) = \prod_{j=1}^{d_y}\mathrm{Ga}(\gamma_j;\,c_j,\,\phi_j)\;\mathrm{Ga}(\lambda_j^{-1};\,d_j,\,\psi_j)$, with independent hyperpriors $\pi(\phi_j) = \mathrm{Ga}(\phi_j;\,q_{j1},\,q_{j2})$ and $\pi(\psi_j) = \mathrm{Ga}(\psi_j;\,s_{j1},\,s_{j2})$ for $j = 1,\ldots,d_y$. The centering distribution $G_0^{\mathbf{x}}(\theta;\rho)$ may be specified, depending upon the type of covariates, following the approach of Chapter 2. Finally, we assume an independent $\mathrm{Ga}(a_\alpha, b_\alpha)$ prior for α.

Inference for the model of (3.13) will, of course, depend upon the partially

observed $\mathbf{z}_i$ rather than the fully observed $\mathbf{y}_i$. Due to the availability of closed form Weibull distribution functions, this requires only a straightforward adaptation of the simulation algorithms described elsewhere in this thesis (either the Pólya urn based scheme of Chapter 2 or the stick-breaking blocked Gibbs of Section 3.2). We detail in the following section the inference procedure for an example consisting of bivariate response and a single binary covariate; the extension of this methodology to other survival regression settings should be clear. Posterior simulation will be based upon a finite stick-breaking approximation to the DP, although a version of the Pólya urn scheme could also be developed. Before turning to this example, we introduce some general aspects of the inference framework in the common context of a univariate response. In this particular setting, the finite stick-breaking truncation of the DP prior (with truncation level L) leads to the hierarchical model, where the relationship between the underlying response y_i and observed z_i is as described above, for $i = 1, \ldots, n$,

$$\begin{aligned} \mathbf{x}_i, y_i | \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{k} &\overset{ind}{\sim} \mathrm{k}_w(y_i; \gamma_{k_i}, \lambda_{k_i}) \mathrm{k}_{\mathbf{x}}(\mathbf{x}_i; \theta_{k_i}) \\ k_i | \mathbf{p} &\sim \sum_{l=1}^{L} p_l \delta_{[l]}(k_i) \\ \mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\theta} | \alpha, \phi, \psi, \rho &\sim \mathcal{P}_L\left(\mathbf{p}; \mathrm{Be}(v; 1, \alpha)\right) \prod_{l=1}^{L} dG_0^{\mathbf{y}}(\gamma_l, \lambda_l; \phi, \psi) dG_0^{\mathbf{x}}(\theta_l; \rho) \end{aligned} \tag{3.14}$$

where $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_L\}$, $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_L\}$, $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_L\}$, the stick-breaking prior $\mathcal{P}_L(\mathbf{p}; \mathrm{Be}(v; 1, \alpha))$ is defined constructively as in equation (2.10), and the base measure and hyperprior specification are described following (3.13).

Conditional on realizations of the truncated random mixing measure, $G^L = \{\mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\theta}\}$, it is possible to sample any desired functional of the conditional density

for response, which is itself available as

$$f(y|\mathbf{x}; G^L) = \frac{\sum_{l=1}^{L} p_l \mathrm{k}_w(y; \gamma_l, \lambda_l)\mathrm{k}_{\mathbf{x}}(\mathbf{x}; \theta_l)}{\sum_{l=1}^{L} p_l \mathrm{k}_{\mathbf{x}}(\mathbf{x}; \theta_l)}. \tag{3.15}$$

In particular, conditional survival and hazard functions will be available. The survival function, $S(y|\mathbf{x}; G^L) = \Pr(Y > y|\mathbf{x}; G^L)$ may be calculated at any $(\mathbf{x}, y)$ location as

$$S(y|\mathbf{x}; G^L) = \frac{\sum_{l=1}^{L} p_l \exp\left[\frac{-y^{\gamma_l}}{\lambda_l}\right] \mathrm{k}_{\mathbf{x}}(\mathbf{x}; \theta_l)}{\sum_{l=1}^{L} p_l \mathrm{k}_{\mathbf{x}}(\mathbf{x}; \theta_l)} \tag{3.16}$$

and the hazard function $h(y|\mathbf{x}; G^L) = f(y|\mathbf{x}; G^L) \ / \ S(y|\mathbf{x}; G^L)$ is

$$h(y|\mathbf{x}; G^L) = \frac{\sum_{l=1}^{L} p_l \mathrm{k}_w(y; \gamma_l, \lambda_l)\mathrm{k}_{\mathbf{x}}(\mathbf{x}; \theta_l)}{\sum_{l=1}^{L} p_l \exp\left[\frac{-y^{\gamma_l}}{\lambda_l}\right] \mathrm{k}_{\mathbf{x}}(\mathbf{x}; \theta_l)}. \tag{3.17}$$

For examples with a multivariate response, the marginal survival and hazard functions remain available in the same form as equations (3.16) and (3.17) respectively, with the only change being that in inference for response j, γ_{lj} and λ_{lj} will replace γ_l and λ_l. In the next section, we consider posterior simulation and inference for a data example based on a version of the model consisting of bivariate response and a single binary covariate.

3.3.2 Model Illustration with an AIDS Clinical Trial Data Example

This section focuses on data from an AIDS observational study conducted by the AIDS clinical trial group (the ACTG), a National Institutes of Health-sponsored multi-center initiative. The purpose of the ACTG 181 study, corresponding to the data considered here, was to determine the natural history of the opportunistic infection cytomeglovirus (CMV) in an HIV-infected individual. In ACTG 181, patients supplied

urine and blood samples periodically over the duration of the study, and these were tested for the presence (or *shedding*) of CMV. Although the sampling was scheduled to occur at regular intervals (every 12 weeks for blood and every 4 weeks for urine), many patients missed or moved their scheduled visits, such that the observed data is censored to irregular intervals. Not all participants exhibited CMV shedding in blood or urine at the study conclusion, resulting in right censoring. In addition, some patients exhibited shedding in blood or urine at the time of their first visit, resulting in left censoring.

Goggins and Finkelstein (2000) consider the problem of regression for the blood and urine shedding of patients in ACTG 181 conditional on an indicator variable related to their CD4 cell counts. The stage of HIV infection is commonly classified by CD4 (cluster of differentiation 4, a protein found on a variety of different cells) count falling below certain threshold, as this is an indication of immune system deterioration. In the example considered by Goggins and Finkelstein (2000) and herein (as well as in Sun (2006) in the context of classical nonparametric regression), a single binary covariate x is set to one for patients with baseline CD4 count less than 75 cells per 10^{-6}L, and zero otherwise.

Thus the responses y_B and y_U, indicating time to CMV shedding in blood and urine respectively, are observed as the intervals $z_B = (a_B, b_B)$ and $z_U = (a_U, b_U)$, where $a_j = 0$ for left censored response j and $b_j = \infty$ for right censored response j, with $j = U$ or B. Each observation is accompanied by the covariate x, indicating whether or not CD4 count has fallen below the specified threshold. There are a total 204 patients in the study. For blood shedding, 7 observations were left censored and 174 were right

censored. For urine shedding, 50 observations were left censored and 87 were right censored.

We model the joint density for responses and covariate as the DP mixture specified by (3.13), where the covariate kernel component is a Bernoulli parameterized by probability θ and the corresponding centering distribution is U(θ; 0, 1). Posterior simulation and inference will be based upon a finite stick-breaking truncation of the DP prior, such that the full hierarchical model is, with the relationship between y_{ij} and z_{ij} described above, for $i = 1, \ldots, n = 204$,

$$\begin{aligned} y_{iB}, y_{iU}, x_i | \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \boldsymbol{\theta}, \mathbf{k} &\overset{ind}{\sim} \mathrm{k}_w(y_{iB}; \gamma_{k_i B}, \lambda_{k_i B}) \mathrm{k}_w(y_{iU}; \gamma_{k_i U}, \lambda_{k_i U}) \theta_{k_i}^{x_i} (1 - \theta_{k_i})^{1 - x_i} \\ k_i | \mathbf{p} &\sim \sum_{l=1}^{L} p_l \delta_{[l]}(k_i) \\ \mathbf{p}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \boldsymbol{\theta} | \alpha, \boldsymbol{\phi}, \boldsymbol{\psi} &\sim \mathcal{P}_L\left(\mathbf{p}; \mathrm{Be}(v; 1, \alpha)\right) \prod_{l=1}^{L} dG_0^{\mathbf{y}}(\Gamma_l, \Lambda_l; \boldsymbol{\phi}, \boldsymbol{\psi}) \delta_{[\theta_l \in (0,1)]}. \end{aligned} \tag{3.18}$$

Here, $\Gamma_l = \{\gamma_{lB}, \gamma_{lU}\}$ and $\Lambda_l = \{\lambda_{lB}, \lambda_{lU}\}$. The response base measure component is specified as for the model defined in (3.13) and thereafter, such that $G_0^{\mathbf{y}} = \mathrm{Ga}(\gamma_B;\ c_B,\ \phi_B)\ \mathrm{Ga}(\lambda_B^{-1};\ d_B,\ \psi_B)\ \mathrm{Ga}(\gamma_U;\ c_U,\ \phi_U)\ \mathrm{Ga}(\lambda_U^{-1};\ d_U,\ \psi_U)$. A Ga(2, 0.2) prior is assumed for the precision parameter α. Prior and hyperprior specification is completed, as elsewhere in this thesis, by considering the limiting case of the model as $\alpha \to 0^+$ wherein the distribution corresponding to a single density kernel has generated all of the observations. Hence, $c_B = c_U = 2$, $d_B = d_U = 2$, $\mathbf{q}_B = \mathbf{q}_U = [2, 1]$, $\mathbf{s}_B = [2, 2/20]$, and $\mathbf{s}_U = [2, 2/10]$.

Conditional on $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\theta}$, and $\mathbf{k}$, the data likelihood is $\mathcal{L}(\ \boldsymbol{\Gamma},\ \boldsymbol{\Lambda},\ \boldsymbol{\theta},\ \mathbf{k}\ |\ \mathcal{D}) =$

$\prod_{i=1}^{n} \Pr(\mathbf{z}_i \mid \Gamma_{k_i}, \Lambda_{k_i}) \Pr(x_i \mid \theta_{k_i})$, where $\Pr(x_i \mid \theta_{k_i}) = \theta_{k_i}^{x_i} (1-\theta_{k_i})^{1-x_i}$ and

$$\Pr(\mathbf{z}_i|\Gamma_{k_i}, \Lambda_{k_i}) = \left(\exp\left[-\frac{a_{iB}^{\gamma_{k_i B}}}{\lambda_{k_i B}} \right] - \exp\left[-\frac{b_{iB}^{\gamma_{k_i B}}}{\lambda_{k_i B}} \right] \right) \left(\exp\left[-\frac{a_{iU}^{\gamma_{k_i U}}}{\lambda_{k_i U}} \right] - \exp\left[-\frac{b_{iU}^{\gamma_{k_i U}}}{\lambda_{k_i U}} \right] \right).$$

Thus, posterior sampling for this model is possible through a version of the blocked Gibbs algorithm developed in Sections 3.2.3 and 4.2, with changes made only to adapt for the censoring. Sampling for α and $\mathbf{p}$ is unchanged by the presence of partially observed data. In particular, we can sample directly from the full conditional posterior for $\mathbf{p}$ given $\mathbf{k}$ and α by drawing $v_l \sim \text{Be}(1 + H_l, \alpha + \sum_{j=l+1}^{L} H_j)$ for $l = 1, \ldots, L-1$, where $H_l = \sum_{i=1}^{n} \delta_{[k_i=l]}$, before setting $v_L = 1$, $p_1 = v_1$, and $p_l = \prod_{j=1}^{l-1}(1-v_j)v_l$ for $l = 2, \ldots, L$. The posterior full conditional for α is $\text{Ga}(a_\alpha + L - 1, b_\alpha - \log(p_L))$.

The independent posterior full conditional for each k_i is proportional to $\sum_{l=1}^{L} p_l \Pr(\mathbf{z}_i \mid \Gamma_l, \Lambda_l) \Pr(x_i \mid \theta_l) \delta_{[l]}(k_i)$. The joint full conditional for the location parameters is

$$\Pr(\mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{\theta}|\mathbf{k}, \mathcal{D}) \propto \prod_{l=1}^{L} \left[dG_0^{\mathbf{y}}(\Gamma_l, \Lambda_l; \boldsymbol{\phi}, \boldsymbol{\psi}) \prod_{i:k_i=l} \Pr(\mathbf{z}_i|\Gamma_l, \Lambda_l) \theta_l^{x_i} (1-\theta_l)^{1-x_i} \right].$$

For l such that $H_l = 0$, the unallocated Γ_l, Λ_l, and θ_l are just sampled from the centering distribution. The independent posterior full conditional for each allocated θ_l is $\text{Be}(\sum_{i:k_i=l} x_i + 1, n - \sum_{i:k_i=l} x_i + 1)$. Sampling for the $n^\star$ allocated Γ_l and Λ_l proceeds through conditional Metropolis-Hastings draws for each of γ_{lj} given λ_{lj} and λ_{lj} given γ_{lj}, for $j = B, U$. In the case of the full conditional draw for each λ_{lj} given γ_{lj}, this is facilitated by the use of independent proposal distributions built through an approximation to $\Pr(z_{ij}|\gamma_{k_i j}, \lambda_{k_i j})$ for left and interval censored observations. In particular, replacing the true likelihood for such z_{ij} with $\text{k}_w(m_{ij}|\gamma_{k_i j}, \lambda_{k_i j})$, where $m_{ij} \in (a_{ij}, b_{ij})$, leads to a

$\text{Ga}(n^j_{lIC} + n^j_{lLC} + d_j,\ \psi_j + \sum_{\{i \in \mathcal{I}^j_{LC} \cup \mathcal{I}^j_{IC} : k_i = l\}} m_{ij}^{\gamma_{lj}} + \sum_{\{i \in \mathcal{I}^j_{RC} : k_i = l\}} a_{ij}^{\gamma_{lj}})$ approximate posterior full conditional for λ_{lj} given γ_{lj} and $\{z_{ij} : k_i = l\}$, where n^j_{lIC} and n^j_{lLC} are respectively the number of interval censored and left censored j responses allocated to component l. Finally, the posterior full conditional for each ϕ_j is $\text{Ga}(q_{j1} + n^\star c_j,\ q_{j2} + \sum_{l:H_l>0} \gamma_{lj})$, and the posterior full conditional for each ψ_j is $\text{Ga}(s_{j1} + n^\star d_j,\ s_{j2} + \sum_{l:H_l>0} \lambda_{lj}^{-1})$.

The following results are based on an MCMC sample of 20,000 parameter draws recorded on every fourth iteration following a burn-in period of 10,000 iterations. Posterior mean and interval estimates for marginal survival and hazard functions, calculated as in equations (3.16) and (3.17), are shown in Figures 3.15 and 3.16 respectively. The marginal hazard functions do not appear to be proportional, as was assumed in Goggins and Finkelstein (2000), especially for CMV shedding in urine. This is more pronounced in Figure 3.17, which shows posterior inference for the ratio of marginal hazard functions (i.e. the marginal hazard for time y_j conditional on CD4 count less than 75 cells/10^{-6}L, divided by the marginal hazard at the same time conditional on CD4 count equal to or above that threshold). Due to heavy right censoring of the data, we observe wide posterior uncertainty intervals around these marginal hazard ratios. However, it does seem clear from Figure 3.15 that having a CD4 count of less than 75 cells/10^{-6}L leads to significantly shorter time to CMV shedding, as there is little or no overlap between the posterior interval for survival functions conditional on $x = 0$ (top row) and conditional on $x = 1$ (bottom row). Finally, posterior mean bivariate density functions for shedding in blood and urine, conditional on the binary CD4 indicator, are

shown in Figure 3.18 and illustrate both the flexibility of our model and the potential for multivariate inference. This figure indicates that CMV shedding times in blood and urine are clearly correlated, as was found by Sun (2006) through the use of a bivariate Copula model. We note, however, that an extension of this Copula model for regression analysis requires an assumption of proportional hazards, which is unlikely to be true.

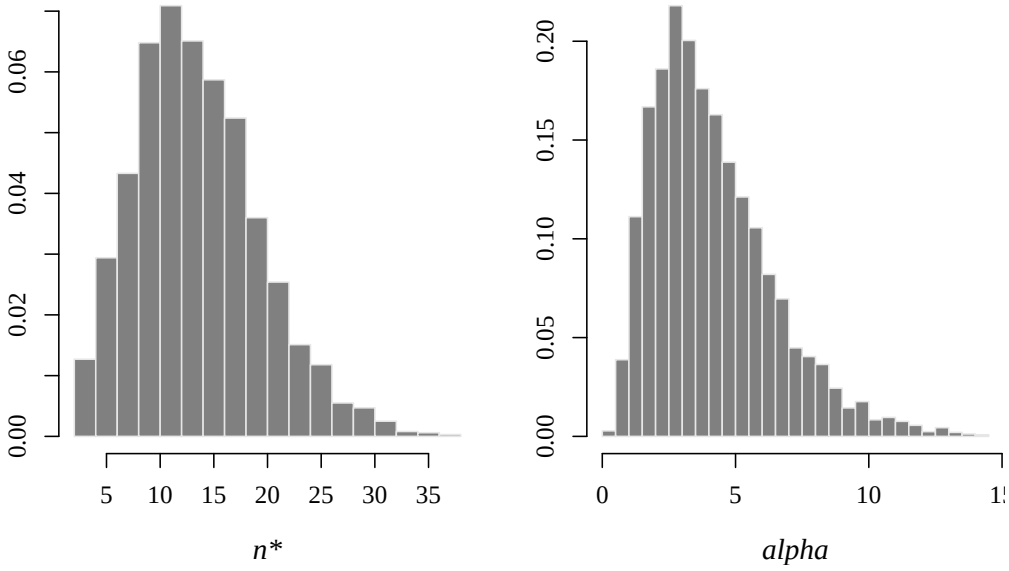


Figure 3.14: AIDS clinical trial group data. Posterior samples for the number of distinct clusters and the DP prior precision.

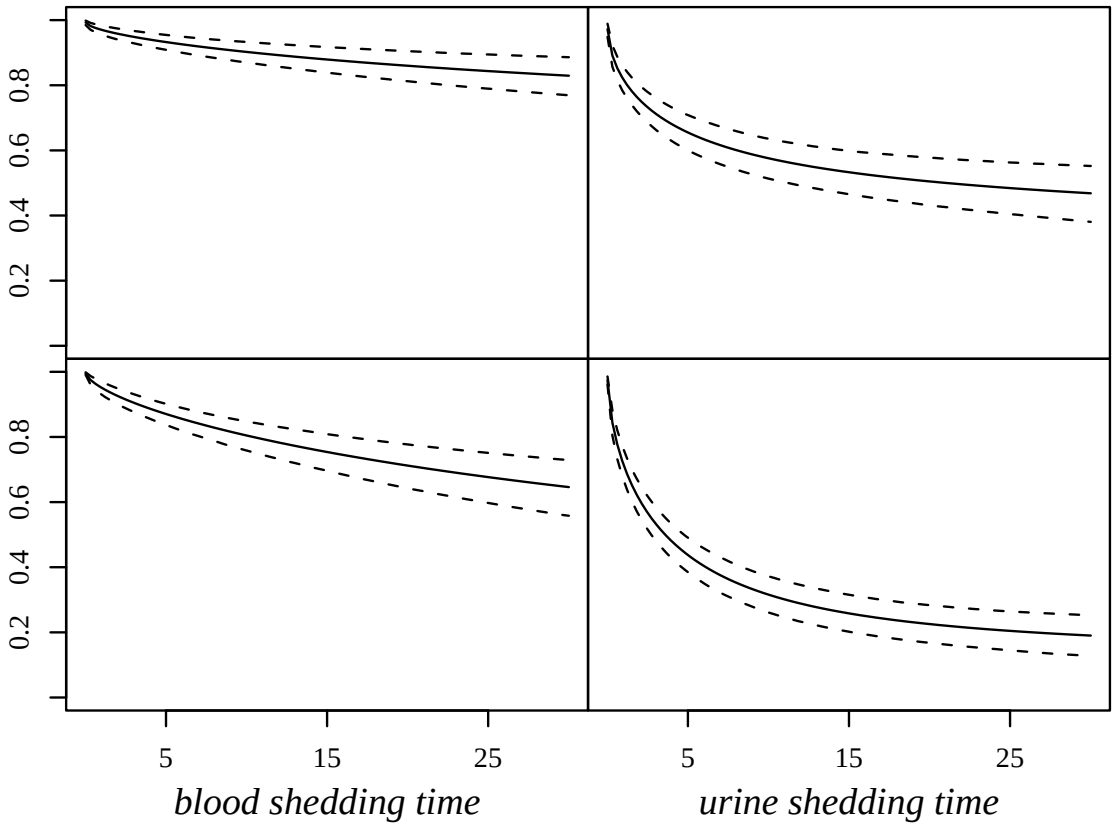


Figure 3.15: AIDS clinical trial group data. Posterior mean (solid lines) and 90% interval (dashed lines) for marginal survival functions $S(y_j|x;G) = \Pr[Y_j > y_j|x;G]$ for CMV shedding time (in weeks) for blood (left column) and urine (right column), conditional on CD4 cells/10^{-6}L ≥ 75 (top row) and CD4 < 75 cells/10^{-6}L (bottom row).

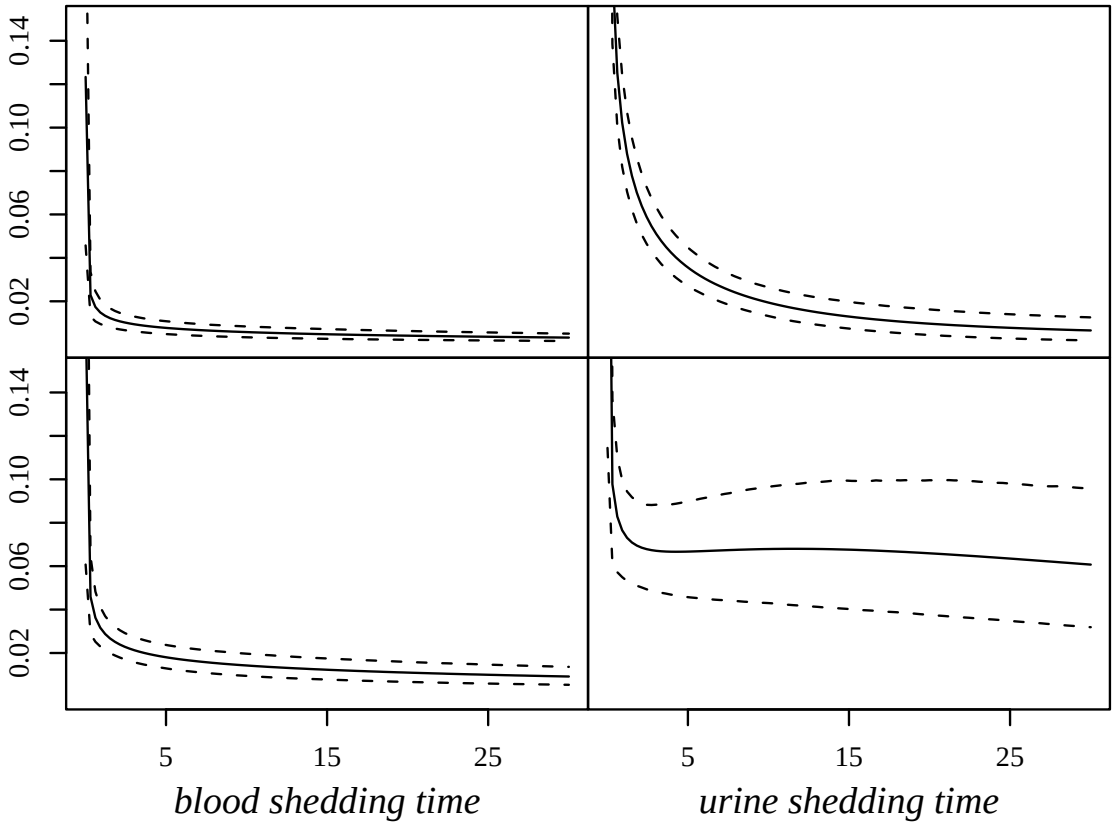


Figure 3.16: AIDS clinical trial group data. Posterior mean (solid lines) and 90% interval (dashed lines) for marginal hazard functions $h(y_j|x;G) = f(y_j|x;G) \,/\, S(y_j|x;G)$ for CMV shedding time (in weeks) for blood (left column) and urine (right column), conditional on CD4 cells/10^{-6}L $\geq$ 75 (top row) and CD4 $<$ 75 cells/10^{-6}L (bottom row).

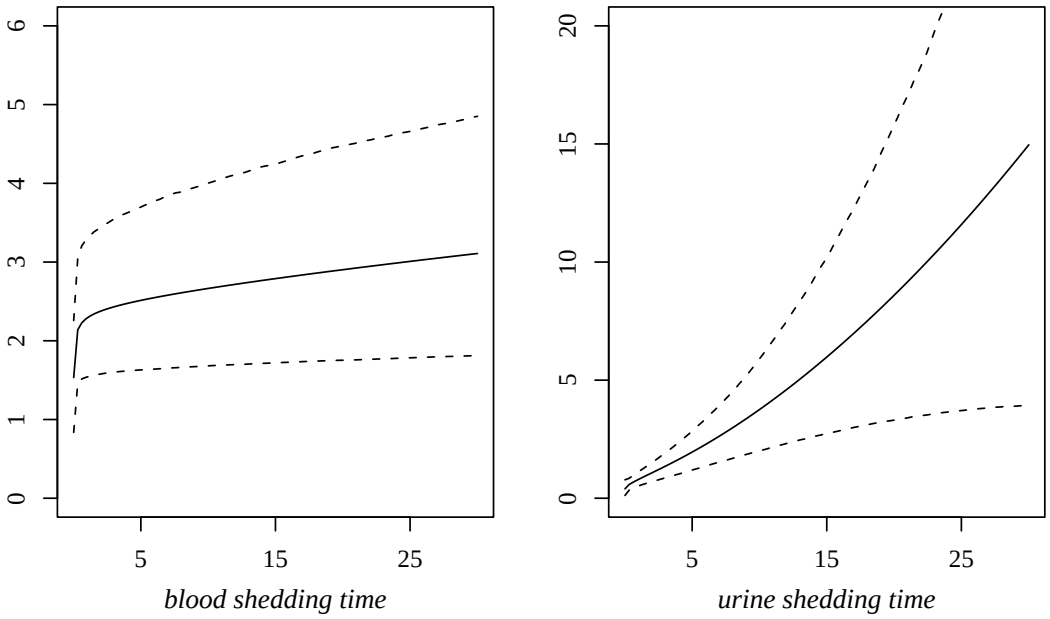


Figure 3.17: AIDS clinical trial group data. Posterior mean (solid lines) and 90% interval (dashed lines) for the ratio of marginal hazard functions, $h(y_j|x=1;G)$ / $h(y_j|x=0;G)$ for CMV shedding time (in weeks) for blood (left) and urine (right).

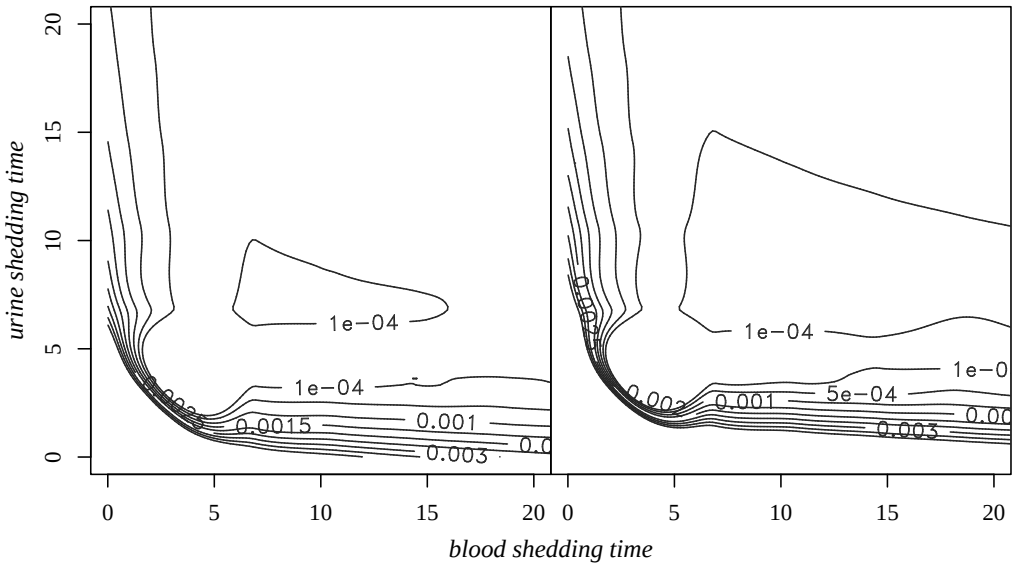


Figure 3.18: AIDS clinical trial group data. Posterior mean for the bivariate conditional response density for weeks to CMV shedding in blood and urine, $f(y_B, y_U \mid x\ ;\ G)$, conditional on CD4 cells/10^{-6}L ≥ 75 (left) and CD4 < 75 cells/10^{-6}L (right).

Chapter 4

Modeling Framework for Dynamic Spatial Marked Poisson Processes

This chapter proposes a general framework for modeling Poisson point processes, through separation of the process intensity into a total intensity and a normalized process density. The process density will be modeled nonparametrically via DP mixture models with bounded kernels, and development of the model for marked processes parallels our modeling for conditional distributions from previous chapters. The framework builds on a basic model for nonhomogeneous Poisson processes described by Kottas and Sansó (2007), generalizing the work to alternative kernel choices and extending the framework to include temporal dynamics and random marks.

While the classical literature on point processes is extensive, fully Bayesian work is much more limited. The book by Diggle (2003) contains a review of parametric likelihood and classical nonparametric inference approaches for spatial Poisson

processes, and Møller and Waagepetersen (2004) serves as an excellent reference on basic theory for Poisson point processes and, in addition, contains a review of more recent work on simulation-based inference for spatial point processes. For more detailed theoretical background on spatial Poisson processes, see, for instance, Cressie (1993), Kingman (1993), and Daley and Vere-Jones (2003). The intensity histogram approach outlined in Diggle (1985) provides an example of classical nonparametric process intensity estimation.

Early examples of Bayesian nonparametric inference for spatial Poisson processes can be found in Heikkinen and Arjas (1998, 1999), where piecewise constant functions, driven by Voronoi tesselations and Markov random field priors, were used to model the intensity function. A more common approach is to rely upon log-Gaussian Cox process models (note that a Cox process is just a Poisson process with a random intensity function, and thus the distinction is arguably irrelevant from a Bayesian perspective), wherein the random intensity function is modeled on logarithmic scale as a Gaussian process. Møller et al. (1998) study properties of log-Gaussian Cox processes and discuss empirical Bayesian inference for the intensity surface. Extensions to spatio-temporal settings are considered in Brix and Diggle (2001) and Brix and Møller (2001).

The Bayesian nonparametric approaches developed by Wolpert and Ickstadt (1998) and Ishwaran and James (2004) are closest in spirit to the work presented here. Both of these approaches utilize a mixture representation for the intensity function based upon a convolution of nonnegative kernels with a gamma process (or in the case of Wolpert and Ickstadt (1998), any Levy random field). Applications to regression

settings are discussed by Ickstadt and Wolpert (1999) and Best et al. (2000), and the former reference provides a connection to modeling for marked processes through an additive intensity formulation. Brix (1999) developed the related shot-noise G-Cox processes model – a Cox processes with intensity built of kernel smoothed generalized gamma measures – which includes the Poisson-gamma random field model as a special case. Even more general related probability models for Cox processes are studied by Møller (2003) and Møller and Torrisi (2005), though this work deals primarily with probabilistic aspects of the spatial processes. Finally, there is a connection between the gamma-process models and DP mixture models, due to the connection between the DP and the gamma process (see Ferguson, 1973, 1974). However, the approach of Kottas and Sansó (2007) and that presented herein differs in that the mixture representation is used directly for the process density instead of process intensity.

4.1 Model Development

This section outlines the various models for Poisson processes underlying our general framework. Posterior sampling methodology for a selection of fully specified hierarchical models follows in Section 4.2 and illustrative data examples are presented in Section 4.3.

4.1.1 Dirichlet Process Mixture Models for Poisson Processes

A non-homogeneous Poisson process $\text{PoP}(\mathcal{R}, \lambda)$ on a bounded observation window $\mathcal{R} \subset \mathbb{R}^2$, with intensity $\lambda(\mathbf{y})$ for $\mathbf{y} \in \mathcal{R}$ that is locally integrable for all bounded

$\mathcal{B} \subseteq \mathcal{R}$, is defined such that

i. For any such $\mathcal{B}$, the number of points in $\mathcal{B}$, $N(\mathcal{B}) \sim \text{Po}\left(\Lambda(\mathcal{B}) = \int_{\mathcal{B}} \lambda(\mathbf{y}) d\mathbf{y}\right)$.

ii. Given $N(\mathcal{B})$, the points are *iid* with density $\lambda(\mathbf{y})/\Lambda(\mathcal{B})$.

Although Poisson processes may be defined over an unbounded space, the observation window is almost always bounded and we refer throughout only to the spatial processes as defined on the bounded observation window. When considering marked processes, below, the observation window may be unbounded. Regardless, through a reparameterization in terms of the density $f(\mathbf{y}) = \lambda(\mathbf{y})/\Lambda(\mathcal{R})$, inference about process properties can be made through the use of density estimation methodology. For data $\mathcal{D} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \sim \text{PoP}(\mathcal{R}, \lambda)$, if we set $\gamma = \Lambda(\mathcal{R})$, the likelihood is

$$\mathcal{L}(f, \gamma | \mathcal{D}) \quad \propto \quad \exp(-\gamma)\, \gamma^n \prod_{i=1}^{n} f(\mathbf{y}_i). \tag{4.1}$$

The conjugate prior for γ is a gamma distribution and the improper reference prior is $\pi(\gamma) \propto \gamma^{-1}$. Since the count, $N(\mathcal{R})$, is sufficient for the integrated intensity, γ, inference about this parameter is independent of estimation for f. Thus, the methodology presented below will apply in the case of either reference or conjugate prior modeling for γ.

DP mixture models are an attractive option for nonparametric estimation of f. The model proposed by Kottas and Sansó (2007) holds that the normalized intensity

arises as a DP mixture of bivariate beta density kernels with Sarmanov dependence,

$$\begin{aligned} f(\mathbf{y}) &= \int \mathrm{B}_2(\tilde{\mathbf{y}}; \boldsymbol{\mu}, \boldsymbol{\tau}, \varphi) dG(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi), \\ \mathrm{B}_2(\tilde{\mathbf{y}}; \boldsymbol{\mu}, \boldsymbol{\tau}, \varphi) &= \mathrm{Be}(\tilde{y}_1; \mu_1\tau_1, (1-\mu_1)\tau_1)\mathrm{Be}(\tilde{y}_2; \mu_2\tau_2, (1-\mu_2)\tau_2) r(\tilde{\mathbf{y}}, \boldsymbol{\mu}; \varphi) \\ G &\sim DP(\alpha, G_0(\boldsymbol{\beta})), \end{aligned} \tag{4.2}$$

where $\tilde{\mathbf{y}}$ is the result of a mapping for $\mathbf{y}$ from a rectangular observation window $\mathcal{R}$ to the unit square $\mathcal{I} = [0,1] \times [0,1]$, and $r(\tilde{\mathbf{y}}; \boldsymbol{\mu}, \varphi) = 1 + \varphi(\tilde{y}_1 - \mu_1)\ (\tilde{y}_2 - \mu_2)$. The centering distribution is

$$G_0(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi; \boldsymbol{\beta}) = \left(\prod_{i=1,2} \mathrm{U}\left(\mu_i; 0, 1\right) \mathrm{Ga}\left(\tau_i^{-1}; c_\beta, \beta_i\right) \right) \mathrm{U}(\varphi; C_{\boldsymbol{\mu}}, C^{\boldsymbol{\mu}}) \tag{4.3}$$

where $\mathrm{U}(\cdot\ ;\ a, b)$ is the uniform distribution over (a, b), and $C_{\boldsymbol{\mu}} = -[\max\{\mu_1\mu_2, (1-\mu_1)(1-\mu_2)\}]^{-1}$ and $C^{\boldsymbol{\mu}} = -[\min\{\mu_1(\mu_2-1), \mu_2(\mu_1-1)\}]^{-1}$ are conditional bounds such that $r(\tilde{\mathbf{y}}, \boldsymbol{\mu}; \varphi) \geq 0$ for all $\tilde{\boldsymbol{y}} \in I$. Hyperpriors are such that $\alpha \sim \mathrm{Ga}(a_\alpha, b_\alpha)$ and $\pi(\boldsymbol{\beta}) = \mathrm{Ga}(\beta_1;\ a_\beta,\ b_\beta)\ \mathrm{Ga}(\beta_2;\ a_\beta,\ b_\beta)$.

This suggests a more general family of models built through DP mixtures of arbitrary kernels,

$$f(\mathbf{y}) = \int \mathrm{k}(\mathbf{y}; \theta) dG(\theta), \quad G \sim DP(\alpha, G_0(\boldsymbol{\psi})) \tag{4.4}$$

where support for k is bounded to the observation window $\mathcal{R}$. The Sarmanov bivariate beta mixture model is obviously of this sort, as is the related case with independent beta kernel components such that $\mathrm{k}(\mathbf{y}; \theta) = \mathrm{Be}(\tilde{y}_1; a_1, b_1)\mathrm{Be}(\tilde{y}_2; a_2, b_2)$. The extra flexibility of the bivariate B_2 kernel is especially desirable when the process intensity is strong near borders of the observation window.

A further generalization of this approach is to allow the support for k to be restricted to subsets of the observation window, where the random mixing distribution for θ places positive prior probability on the support including the entire window. This will allow for discontinuities in the intensity surface. For example, if the spatial surface is expected to include aerial units with relatively uniform properties, the kernel could be built as the product of uniform densities such that,

$$f(\mathbf{y}) = \int \mathrm{U}(y_1; l_1, u_1)\mathrm{U}(y_2; l_2, u_2)dG(\mathbf{l}, \mathbf{u}) \tag{4.5}$$

with $G \sim \mathrm{DP}(\alpha, \mathrm{U}_2(l_1, u_1; a_1, b_1)\mathrm{U}_2(l_2, u_2; a_2, b_2))$, where a_1 is the western boundary for $\mathcal{R}$, b_1 is the eastern boundary, a_2 is the southern boundary, b_2 is the northern boundary, and $\mathrm{U}_2(u, l;\ a, b)$ denotes a uniform distribution over the triangle $a < l < u < b$. Posterior simulation for this model will not be considered in detail, however it is straightforward to devise an MCMC algorithm built around algorithm 5 from Neal (2000) with sampling for truncated G, as in Section 2.4 for the multivariate normal DP mixture model (realizations from U_2 are easily drawn through rejection sampling). This model is loosely connected to the univariate "random histograms" proposed by Gasparini (1996) and outlined in Ghosh and Ramamoorthi (2003), but in that semi-parametric model the uniform width of the histogram bins is modeled parametrically and the DP prior is assumed only for bin probabilities. In addition, Brunner and Lo (1989) proposed DP mixtures of uniform densities in estimation of symmetric unimodal densities, however this only involved mixing on upper (or lower) bounds of intervals between zero and a parameter θ (or $-\theta$) for $\theta \in (0, \infty)$. A perhaps closer connection

exists between the model of (4.5) and mixtures of Pólya tree priors (see, e.g., Lavine, 1992; Walker et al., 1999), as in each case a finite version of the model (i.e. for truncated G or for partially specified Pólya trees) leads to realized measures on the set of interest consisting of random probabilities assigned to axisymmetric intervals of random width.

Before moving to models for marked and dynamic Poisson processes, we revisit the use of kernels restricted to rectangular support. Although the issue is seldom addressed in the modeling literature, in practice it is not uncommon to encounter observation windows which are not rectangular. Due to the flexibility of the DP mixture approach to modeling f, a large sample size will lead to inference that is practically unaffected by conveniently increasing the modeled observation window to a rectangle containing the true window. The process density fit accounts for the lack of points outside of the true window by assigning to those areas a density value very near to zero. However, when this simplification is unacceptable, it is possible to adapt our modeling framework to account for a nonrectangular observation window.

One approach is to define a one-to-one transformation of variables which maps from the irregular observation window to a rectangular region, allowing for process density to be modeled over the transformed space through use of a kernel with rectangular support. A standard approach to such problems is to search for a conformal mapping (i.e. a transformation which preserves angles between vectors). Trefethen (1980) describes numerical computation of the Schwarz-Christoffel transformation between an arbitrary polynomial and the unit disc (implemented in the Fortran package `SCPACK` and available in `MATLAB`), and Trefethen (1984) applies the technique in a composition of

transformations for conformal mapping between arbitrary polynomials and a rectangle. The transformations themselves are evaluated analytically, as it is only the coefficients of the Schwarz-Christoffel transformation which are estimated numerically, such that the Jacobian is available and the posterior f density realizations can be transformed back to the original observation window. However, this approach should be undertaken with care to ensure that the spatial dependence structure in the transformed space does not lead to unreasonable densities in the original observation window. And, of course, any prior intuition about the process density must be translated to the transformed coordinate system.

An obvious alternative is to simply use kernels which are bounded to the nonrectangular observation window. If the kernel k is defined with support over a larger rectangle $\mathcal{W}$ that includes all of the nonrectangular $\mathcal{R}$, the process density may be specified through the truncated kernel model $f(\mathbf{y}) = \int \mathrm{k}_{\mathcal{R}}(\mathbf{y};\theta)\, dG(\theta)$ where $\mathrm{k}_{\mathcal{R}}(\mathbf{y};\theta)$ $= \mathrm{k}(\mathbf{y};\theta)\, \delta_{[\mathbf{y}\in\mathcal{R}]} \,/\, (1 - \int_{\mathcal{W}/\mathcal{R}} \mathrm{k}(\mathbf{s};\theta)\, d\mathbf{s})$. Model fitting will be able to follow many of the same steps as outlined below for kernels with rectangular support, except that numerical integration of the normalizing constant is required for kernel evaluation at every new θ value. This computational complexity is avoided if one is able to construct a kernel which is analytically bounded to the irregular support. In particular, the DP mixture of uniform densities in (4.5) lends itself to such constructions. The process density model

may be re-stated for arbitrary $\mathcal{R}$ as

$$\begin{aligned} f(\mathbf{y}) &= \int \frac{\delta_{[\mathbf{y}\in(\mathbf{l},\mathbf{u})\cap\mathcal{R}]}}{\mathrm{A}\left[(\mathbf{l},\mathbf{u})\cap\mathcal{R}\right]} dG(\mathbf{l},\mathbf{u}), \\ G &\sim DP\left(\alpha, \mathrm{U}_2(l_1,u_1;a_1,b_1)\mathrm{U}_2(l_2,u_2;a_2,b_2)\right), \end{aligned} \tag{4.6}$$

where a_1 is the western-most point on the boundary of $\mathcal{R}$, b_1 is the eastern-most point, a_2 is the southern-most point, b_2 is the northern-most point, $(\mathbf{l},\mathbf{u})$ is the interior of the rectangle defined by $\mathbf{l}$ and $\mathbf{u}$ and $\mathrm{A}\left[(\mathbf{l},\mathbf{u})\cap\mathcal{R}\right]$ is the area of intersection between $(\mathbf{l},\mathbf{u})$ and the observation window.

4.1.2 Marked Spatial Poisson Processes

A marked Poisson process consists of points from a spatial Poisson point process, $\mathbf{y} \sim \mathrm{PoP}(\mathcal{R},\phi)$, and a random mark m at each location drawn from the conditional distribution $\Pr(m|\mathbf{y})$. If a $\mathrm{PoP}(\mathcal{R}\times\mathcal{M},\lambda)$ process defined over the joint location-mark observation window is such that $\int_{\mathcal{M}}\lambda(\mathbf{y},m)dm = \phi(\mathbf{y})$ is locally integrable, then the joint process is such a marked process. We make use of this property of Poisson processes to model marked processes and, as a result, model the conditional distribution for marks. Define a process over the joint location-mark space with intensity function

$$\lambda(\mathbf{y},m) = \gamma \int \mathrm{k}_{\mathbf{y}}(\mathbf{y};\theta_{\mathbf{y}})\mathrm{k}_m(m;\theta_m)dG(\theta_{\mathbf{y}},\theta_m) = \gamma f(\mathbf{y},m;G), \tag{4.7}$$

where $\gamma = \int_{\mathcal{R}} \int_{\mathcal{M}} \lambda(\mathbf{y}, m) dm d\mathbf{y}$ and the mark kernel $\mathrm{k}_m(m; \theta_m)$ has support on $\mathcal{M}$. Then, due to the almost sure discreteness of G,

$$\begin{aligned} \int_{\mathcal{M}} \lambda(\mathbf{y}, m) dm &= \gamma \int_{\mathcal{M}} \int_{\theta_{\mathbf{y}}} \mathrm{k}_{\mathbf{y}}(\mathbf{y}; \theta_{\mathbf{y}}) \int_{\theta_m} \mathrm{k}_m(m; \theta_m) dG(\theta_{\mathbf{y}}, \theta_m) dm \\ &= \gamma \int_{\theta_{\mathbf{y}}} \mathrm{k}_{\mathbf{y}}(\mathbf{y}; \theta_{\mathbf{y}}) \int_{\theta_m} \left[\int_{\mathcal{M}} \mathrm{k}_m(m; \theta_m) dm \right] dG(\theta_{\mathbf{y}}, \theta_m) \\ &= \gamma \int_{\theta_{\mathbf{y}}} \mathrm{k}_{\mathbf{y}}(\mathbf{y}; \theta_{\mathbf{y}}) dG(\theta_{\mathbf{y}}) = \gamma f(\mathbf{y}; G) = \phi(\mathbf{y}). \end{aligned} \tag{4.8}$$

Since $\gamma = \int_{\mathcal{R}} \int_{\mathcal{M}} \lambda(\mathbf{y}, m) dm d\mathbf{y} = \int_{\mathcal{R}} \phi(\mathbf{y}) d\mathbf{y}$, we have recovered the original DP mixture model of Section 4.1.1 for the marginal location Poisson point process $\mathrm{PoP}(\mathcal{R}, \phi)$. The conditional distribution for marks is thus

$$\Pr(m|\mathbf{y}; G) = \frac{f(\mathbf{y}, m; G)}{f(\mathbf{y}; G)} = \frac{\int \mathrm{k}_{\mathbf{y}}(\mathbf{y}; \theta_{\mathbf{y}}) \mathrm{k}_m(m; \theta_m) dG(\boldsymbol{\theta})}{\int \mathrm{k}_{\mathbf{y}}(\mathbf{y}; \theta_{\mathbf{y}}) dG(\boldsymbol{\theta})}. \tag{4.9}$$

Note that, through an argument analogous to that of (4.8), the marginal mark intensity defined as $\int_{\mathcal{R}} \lambda(\mathbf{y}, m) d\mathbf{y}$ is locally integrable. This, combined with the fact that conditional on G, $\lambda(\mathbf{y}, m) = \phi(\mathbf{y}) \Pr(m|\mathbf{y})$, implies that our DP mixture joint location-mark process of (4.7) satisfies the requirements of proposition 3.9 in Møller and Waagepetersen (2004) and hence the marks alone are marginally distributed as a Poisson process $\mathrm{PoP}\big(\mathcal{M}, \int_{\mathcal{R}} \lambda(\mathbf{y}, m) d\mathbf{y}\big)$.

Assuming a Sarmanov bivariate beta kernel for the location process, all that remains is to specify the independent kernel for marks. In modeling for categorical marks, the multinomial kernel is a straightforward option which, due to the flexibility of the DP mixture model, will be appropriate for a wide variety of applications. In this case, for mark space $\mathcal{M} = \{1, 2, \ldots, M\}$, the $\mathrm{PoP}(\mathcal{R} \times \mathcal{M}, \lambda)$ process is defined through

a hierarchical model for the intensity function,

$$\begin{aligned} \lambda(\mathbf{y}, m) &= \gamma \int \mathrm{B}_2(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\tau}, \varphi) q_m dG(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi, Q) \\ G &\sim \mathrm{DP}\left(\alpha, G_0^{\mathbf{y}}(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi; \boldsymbol{\beta}) \mathrm{Dir}(Q; \mathbf{a})\right), \end{aligned} \tag{4.10}$$

where $G_0^{\mathbf{y}}$ and $\pi(\boldsymbol{\beta})$ are defined as in the location process specification of (4.3) and thereafter, $Q = [q_1, \ldots, q_M]$ and $\mathrm{Dir}(Q; \mathbf{a})$ is the Dirichlet distribution such that $\mathbb{E}(q_m|\mathbf{a})$ $= a_m / \sum_{i=1}^{M} a_i$. It is possible to place a hyperprior on the parameter vector $\mathbf{a}$ (e.g. Leonard, 1977), but in many cases this can simply be fixed to the prior expectation of unconditional mark proportions.

Similarly, continuous marks can be modeled through an appropriate choice for the independent mark kernel. In the case of real-valued continuous marks (i.e. $\mathcal{M} = \mathbb{R}$), the choice of a normal density kernel leads to the intensity function $\lambda(\mathbf{y}, m) =$ $\gamma \int \mathrm{B}_2(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\tau}, \varphi) N(m; q_1, q_2) dG(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi, \mathbf{q})$. In the more common case of positive continuous marks (i.e. $\mathcal{M} = \mathbb{R}^+$), a lognormal density mark kernel leads to the full model

$$\begin{aligned} \lambda(\mathbf{y}, m) &= \gamma \int \mathrm{B}_2(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\tau}, \varphi) \mathrm{N}(\log(m); q_1, q_2) dG(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi, q_1, q_2) \\ G &\sim \mathrm{DP}\left(\alpha, G_0^{\mathbf{y}}(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi; \boldsymbol{\beta}) \mathrm{N}(q_1; s_1, s_2) \mathrm{Ga}(q_2^{-1}; s_4, s_3)\right). \end{aligned} \tag{4.11}$$

The gamma density provides an alternative kernel component for positive continuous marks (for example, in the case when m is a wait time or distance to travel).

4.1.3 Dynamic Poisson Processes

In many situations, realizations from spatial point processes will be observed repeatedly at discrete time intervals. In such cases, the temporal dependence between

realized point patterns needs to be accounted for in the modeling framework. One possible approach would be to model the combined set of points as a Poisson process over both space and time (and possibly the mark support). However, when spatial patterns are observed after discrete time intervals, it will usually be more natural to make explicit the temporal dependence of the spatial process through modeling of a time-dependent intensity function λ_t.

In our modeling framework, the intensity function is built as the product of integrated intensity, $\gamma = \int_{\mathcal{R}} \lambda(\mathbf{y}) d\mathbf{y}$, and a density function, $f(\mathbf{y}) = \int \mathrm{k}(\mathbf{y}; \theta) dG(\theta)$, which arises as a random mixture of bounded kernels with support in the observation window $\mathcal{R}$. In order to account for temporal dependence, both the integrated intensity and the process density will become functions over the discrete timeframe $\mathcal{T} = \{1, \ldots, T\}$. The dynamic integrated intensity, $\{\gamma_t : t = 1, \ldots, T\}$, forms a univariate time series and can be approached with parametric time series methodology; we will return to this in Section 4.1.4. Temporal dependence for f_t will be introduced via dynamic modeling for the random mixing distribution, such that $f_t(\mathbf{y}) = \int \mathrm{k}(\mathbf{y}; \theta) dG_t(\theta)$. This is achieved through the use of dependent Dirichlet process (DDP) priors, as introduced in the technical report by MacEachern (2000).

DDP priors are prevalent in the Bayesian nonparametrics literature, and have been used to model random measures which are related in space (Gelfand et al., 2005) and over categorical covariates (de Iorio et al., 2004), among other applications. The original specification of MacEachern's DDP is based upon the stick-breaking characterization of the DP. In the context of time dependent random measures, the general

model formulation holds that the set of random measures corresponding to time points in $\mathcal{T}$, $\boldsymbol{G} = \{G_1, \ldots, G_T\}$, collectively distributed as a DDP$\big(S_{\mathcal{T}}(\mathbf{v};\boldsymbol{\alpha}), G_0^{\mathcal{T}}(\boldsymbol{\theta};\boldsymbol{\psi})\big)$, are realized in the form

$$G_t \quad = \quad \sum_{l=1}^{\infty} p_{l,t}\delta_{\theta_{l,t}} \quad \text{for } t \in \mathcal{T}, \tag{4.12}$$

$$\text{where} \quad p_{1,t} = v_{1,t} \quad \text{and} \quad p_{l,t} = v_{l,t}\prod_{i=1}^{l-1}(1-v_{i,t}), \quad \text{for } l = 2, \ldots, \infty.$$

The stick-breaking proportions $\mathbf{v}_l = \{v_{l,1}, \ldots, v_{l,T}\}$ and the locations $\boldsymbol{\theta}_l = \{\theta_{l,1}, \ldots, \theta_{l,T}\}$ are respectively *iid* realizations over the indices in $\mathcal{T}$ from the finite dimensional distributions $S_{\mathcal{T}}(\mathbf{v};\boldsymbol{\alpha})$ parameterized by $\boldsymbol{\alpha}$ and $G_0^{\mathcal{T}}(\boldsymbol{\theta};\boldsymbol{\psi})$ parameterized by $\boldsymbol{\psi}$. The finite dimensional distributions for both $\mathbf{v}$ and $\boldsymbol{\theta}$ are induced by measurable stochastic processes indexed by time. Note that the temporal dependence is only specified at the level of the random measures rather than at the observation level, and thus in the absence of further modeling constraints we have conditional independence such that $\Pr(\theta_1, \ldots, \theta_T) = \prod_{t=1}^{T}\sum_{l=1}^{\infty} p_{l,t}\delta_{\theta_{l,t}}(\theta_t)$.

Alternative approaches for the modeling of dependent random measures include the order-based dependent DP of Griffin and Steel (2006), which allows for the ordering of the stick-breaking proportions to be correlated across related populations, and the kernel stick-breaking process of Dunson and Park (2008), which builds a dependence upon covariates into the Pólya urn posterior predictive distribution that would result from a standard DP. The generalized spatial Dirichlet process of Duan et al. (2007) presents a multivariate stick-breaking process that can be viewed as an extension of the standard full DDP. In the context of discrete time de-

pendence, the generalized spatial DP allows for observation vectors to be correlated across time conditional on a realized set of random distributions $G_1, \ldots, G_T$, such that $\Pr(\theta_1, \ldots, \theta_T) = \sum_{l_1=1}^{\infty} \ldots \sum_{l_T=1}^{\infty} p_{l_1 \ldots l_T} \delta_{\theta_{l_1,1}}(\theta_1) \ldots \delta_{\theta_{l_T,T}}(\theta_T)$. Although the generalized prior is attractively flexible and has many very nice properties, inference about this model is relatively computationally intensive and, in many cases, the extra flexibility of the prior will be unnecessary.

Existing applications of the DDP have focused on the simplified single-p version wherein the stick-breaking proportions are assumed constant over the index set, such that $\mathbf{v}_l = v_l$. Hence, members of the set of random measures, $\boldsymbol{G}$, collectively distributed as a single-p DDP$\big(\alpha, G_0^{\mathcal{T}}(\boldsymbol{\theta}; \boldsymbol{\psi})\big)$, are realized in the form $G_t = \sum_{l=1}^{\infty} p_l \delta_{\theta_{l,t}}$ where $p_1 = v_1$ and $p_l = v_l \prod_{i=1}^{l-1} v_i$ for $l = 2, \ldots, \infty$, with locations $\boldsymbol{\theta}_l = \{\theta_{l,1}, \ldots, \theta_{l,T}\}$ constructed as above for (4.12) and the v_l drawn *iid* from a Be$(1, \alpha)$ distribution. The simplification of constant stick-breaking proportions allows for substantially more tractable posterior simulation, and in many applications the flexibility of the single-p version is more than adequate. The spatial DP of Gelfand et al. (2005) is an example of the single-p DDP where $G_0^{\mathcal{T}}(\boldsymbol{\theta}; \boldsymbol{\psi})$ is induced by a Gaussian process. Note that, in this case, the simplified model provides a clear way to impose dependence over the index set at the observation level, such that $\Pr(\theta^1, \ldots, \theta^T) = \sum_{l=1}^{\infty} p_l \delta_{\theta_l^1}(\theta^1) \ldots \delta_{\theta_l^T}(\theta^T) = \sum_{l=1}^{\infty} p_l \delta_{\boldsymbol{\theta}_l}(\boldsymbol{\theta})$.

The single-p DDP in the context of discrete-time dependent Poisson processes would have $\boldsymbol{G} \sim \text{DDP}\left(\alpha, G_0^{\mathcal{T}}(\boldsymbol{\theta}; \boldsymbol{\psi})\right)$ for the resultant dynamic process density, $f_t(\mathbf{y}) = \int \mathrm{k}(\mathbf{y}; \theta) dG_t(\theta)$. For the Sarmanov bivariate beta mixture model, where $\mathrm{k}(\mathbf{y}; \theta) = \mathrm{B}_2(\mathbf{y};$

$\boldsymbol{\mu}$, $\boldsymbol{\tau}$, φ), a possible prior formulation for $\boldsymbol{G} = \{G_1, \ldots, G_T\}$ would have

$$\begin{aligned} \boldsymbol{G} &\sim \text{DDP}\left(\alpha, G_0(\boldsymbol{\mu}, \varphi, \mathbf{T} = \{\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_T\}; \boldsymbol{\beta}, \psi)\right) \qquad (4.13) \\ G_0(\boldsymbol{\mu}, \varphi, \mathbf{T}; \boldsymbol{\beta}, \psi) &= \delta_{[0<\mu_1,\mu_2<1]}\text{U}(\varphi; C_{\boldsymbol{\mu}}, C^{\boldsymbol{\mu}})\text{S}(\mathbf{T}; \psi), \end{aligned}$$

with $C_{\boldsymbol{\mu}}$, $C^{\boldsymbol{\mu}}$, and hyperpriors for $\boldsymbol{\beta}$ defined as for (4.3) above. The finite dimensional distribution $\text{S}(\mathbf{T}; \psi)$ is induced by a bivariate temporal process for $\boldsymbol{\tau}$ with realizations on $\mathbb{R}^+$; for example, the logarithm of two standard dynamic linear models. Thus, the centering distribution is the product of a T dimensional distribution for $\mathbf{T} = \{\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_T\}$ and a distribution that is constant in time for $\boldsymbol{\mu}$ and φ. Following in the spirit of the spatial DP, prior realizations of the joint distribution $\boldsymbol{G}$ would be of the form $\boldsymbol{G} = \sum_{l=1}^{\infty} p_l \delta_{\boldsymbol{\theta}_l}$, where the $\boldsymbol{\theta}_l = \{\boldsymbol{\mu}_l, \varphi_l, \mathbf{T}_l\}$ are *iid* realizations from the distribution $G_0(\boldsymbol{\mu}, \varphi, \mathbf{T}; \boldsymbol{\beta}, \psi)$. Alternatively, one may define a single-p DDP such that the mixing measure for $\boldsymbol{\mu}$ varies in time while $\boldsymbol{\tau}$ and φ remain constant, or even have all parameters modeled dynamically. A related time-dependent single-p DDP model, for the estimation of densities which evolve in discrete time, has been recently introduced by Rodriguez and ter Horst (2008). In this formulation, the random mixing measure for kernel location parameters is distributed as a DP with centering distribution provided by a dynamic linear model with normal error components.

Instead of following the single-p DDP route for our modeling of dynamic Poisson processes, we look to the alternative single-θ simplification of the DDP, wherein the realized locations $\theta_{l,t}$ are assumed to be constant in time, such that $\boldsymbol{\theta}_l = \theta_l$. Hence, members of the set of random measures, $\boldsymbol{G}$, collectively distributed as a single-

θ DDP$(S_{\mathcal{T}}(\mathbf{v};\boldsymbol{\alpha}), G_0(\boldsymbol{\psi}))$, are realized in the form $G_t = \sum_{l=1}^{\infty} p_{l,t}\delta_{\theta_l}$ where the θ_l are *iid* draws from $G_0(\theta;\boldsymbol{\psi})$ and the $p_{l,t}$ are constructively defined as in (4.12) through *iid* sample path realizations of the stick-breaking proportions $\mathbf{v}_l = \{v_{l,1}, \ldots, v_{l,T}\}$ from the finite dimensional distribution $S_{\mathcal{T}}(\mathbf{v};\boldsymbol{\alpha})$ induced by a measurable stochastic processes on $(0,1)$ indexed by time.

In analogy to MacEachern's development of the single-p model, it is necessary to argue that the restriction to constant θ locations has not reduced the support of the prior distribution. This is the minimum requirement of any nonparametric prior.

Proposition *Assume that* $\boldsymbol{G} \sim \text{DDP}\left(S_{\mathcal{T}}(\mathbf{v};\boldsymbol{\alpha}), G_0(\boldsymbol{\psi})\right)$*, with* $\mathcal{T} = t_1, \ldots, t_T$ *a set of distinct points in either a countable set or an open set in* $\mathbb{R}^+$ *(i.e. either discrete or continuous time domain) and* $S_{\mathcal{T}}(\mathbf{v};\boldsymbol{\alpha})$ *the corresponding finite dimensional distribution induced by a measurable stochastic process* S *over this domain. If* S *has support* $(0,1)^T$ *and the distributions* $G_1, \ldots, G_T$ *are each absolutely continuous with respect to* $G_0(\psi)$*, then for any* $\varepsilon > 0$ *and for any* $t \in \{t_1, \ldots, t_T\}$*, with* θ^{ε} *denoting the* ε*-neighborhood of* θ*, the set of distributions* $\{G : G(\theta) \leq G_t(\theta^{\varepsilon}) + \varepsilon \ \forall\ \theta \in \Theta\}$ *(i.e. within the Lévy distance* ε *of* G_t*), has positive probability under the single-*θ *DDP prior for* $\boldsymbol{G}$*.*

Proof. We consider the case where Θ, the σ-field over which G_0 is defined, consists of the Borel sets generated by the intervals of (w.l.o.g.) $\mathbb{R}^p$. The corresponding proof for categorical θ is a more straightforward version of this argument. First, for each $t = t_1, \ldots, t_T$, we can construct a discrete approximation to the probability measure defined by G_t: say $\mathcal{C}_t = [-c_t, c_t]^p$ such that $\Pr_{G_t}(\mathcal{C}_t) > 1 - \varepsilon/4$ and c_t is a multiple of $\varepsilon/4$, and define

hypercubes (open to the southwest) $\{Q_{t,1}, \ldots, Q_{t,N}\}$ with side-length $\varepsilon/4$ which cover $\mathcal{C}_t$ (such that $N = (8c_t/\varepsilon)^p$). Then the measure H_t defined by assigning mass $\Pr_{G_t}(Q_{t,i})$ to the midpoint, $q_{t,i}$, of each hypercube is such that $H_t(\theta) \leq G_t(\theta^{\frac{\varepsilon}{2}}) + \varepsilon/2 \ \forall \ \theta \in \Theta$. Second, we show that members of a set of measures assigned positive prior probability will each be suitably close to the discrete approximation H_t. For each t, the distributions $F_t \in \mathcal{F}_t$ are built by assigning appropriate mass to a finite sequence of locations, $\boldsymbol{\theta}^\star = \{\theta_1^\star, \ldots, \theta_n^\star\}$, which can be re-numbered as $\{\theta_{0,1}, \ldots, \theta_{0,n_0}, \ldots, \theta_{N,1}, \ldots, \theta_{N,n_N}\}$ such that $\{\theta_{j,1}, \ldots, \theta_{j,n_j}\} \subset Q_{t,j}$ for $j = 0, \ldots, N$ and $\sum_{j=0}^{N} n_j = n$, with $Q_{t,0} = \Theta \setminus \mathcal{C}_t$. Since $G_t << G_0$, any such sequence has positive probability as an *iid* draw from G_0 (i.e. as a finite subset of an infinite *iid* sequence). Indeed, assume that $\boldsymbol{\theta}^\star$ arises during the location-sequence part of the stick-breaking single-θ DDP construction and was preceded by an arbitrarily large $n_{<t}$ locations (i.e. enough locations to allow for a similar construction of distributions arbitrarily close to $G_1, \ldots, G_{t-1}$). Now, since $S(\mathbf{v}; \boldsymbol{\alpha})$ has full support over $(0,1)^T$, the event $0 < v_{l,t} < \delta_0 < 1$ has positive prior probability for each $l \leq n_{<t}$, where δ_0 has been chosen such that $v_{l,t} < \delta_0$ for $l = 1, \ldots, n_{<t}$ guarantees $\sum_{l=1}^{n_{<t}} p_{l,t} = \sum_{l=1}^{n_{<t}} \left[v_{l,t} \prod_{i=1}^{l-1} v_{i,t} \right] < \varepsilon/8$. Similarly, the event $0 < \delta_L < v_{l,t} < \delta_U < 1$ for $l = n_{<t}, \ldots, n_{<t} + n$ has positive prior probability where δ_L, δ_U, and each n_j have been chosen together to ensure that $|\sum_{i=1}^{n_j} p_{i,j} - \Pr_{G_t}(Q_{t,j})| < \varepsilon/4N$ for $j = 0, \ldots, N$. The distribution $F_t \in \mathcal{F}_t$ built by assigning mass $p_{i,j}$ to each location $\theta_{i,j}$ is then such that $F_t(\theta) \leq H_t(\theta^{\frac{\varepsilon}{2}}) + \varepsilon/2 \ \forall \ \theta \in \Theta$, and the triangle inequality implies our final result. $\square$

Note that this proof is a straightforward adaptation of the argument used by

MacEachern (2000) to show full support under the stated weak topology for the full DDP. Also note that the restriction of support $(0,1)^T$ for S can be weakened to require only $\exists\delta \in (0,1]$ such that every interval in the region less than δ has positive prior mass at each time t. Thus the only extra condition (over the full DDP) for full prior support of the single-θ DDP is that each G_t is absolutely continuous with respect to the single distribution G_0. Finally, the argument uses the construction of practically independent random measures to show full support for each individual G_t. The prior will obviously be much more efficient in representation of correlated random measures, but it is interesting to note that this limiting case of near independence between the G_t is not possible under the single-p DDP (think about the positive probability of a large point mass at any single location; this structure must be represented in each correlated random measure).

The motivation for our use of single-θ processes is both practical and conceptual. As for the single-p DDP, simplification of the full DDP leads to considerably more straightforward posterior sampling without unduly restricting the flexibility of the model. The single-θ model also leads to a practical modeling advantage over single-p version: since the stochastic process underlying the temporal dynamics of the random measure is restricted to only the stick-breaking proportions, it is only ever necessary to specify a univariate time series on $(0,1)$. This is especially attractive in the context of more complex mixture kernels. The model of (4.13) requires specification of only a two dimensional time series, however this is achieved by restricting some kernel components to be held constant in time. And in the model proposed by (4.5), the uniform

base measure does not lend itself easily to a dynamic extension. For non-rectangular observation windows and for the modeling of marked processes, the kernel complexity may need to increase considerably and it becomes much more difficult to specify the stochastic process underlying $G_0^{\mathcal{T}}(\boldsymbol{\theta};\boldsymbol{\psi})$. In contrast, a carefully designed process for stick-breaking proportions may be used in a variety of different applications concerning both marked and unmarked processes. In situations where the restriction of the single-θ process is not sufficiently flexible to efficiently model the process dynamics (e.g., if the conditional mark distribution is governed by different dynamics than the underlying location point process), it may be possible to adapt the model incrementally by introducing time dependence in the random mixing distribution for individual kernel parameters whose behavior is not adequately modeled. The resultant full DDP would have a centering distribution that is constant over time for some parameters and dynamic for other parameters, analogous to the structure of the centering distribution in (4.13); however, the dynamic stick-breaking proportions will allow the realized measures for all parameters to evolve in time.

In Bayesian nonparametric modeling, a wide variety of models will be able to adapt to capture similar behavior. The distinction between models is thus based largely on the practical ease of implementation and the efficiency with which a model is able to capture the data behavior. In addition to the practical benefits of the single-θ process, there are potentially inherent aspects of point processes for which it is uniquely well suited and thus lead to its increased efficiency over alternative priors. In modeling for spatial Poisson processes, it is often the case that the shape of the major spatial factors

affecting process intensity, such as geographical features or socio-economic variables, remains constant over the time period covered by $\mathcal{T}$. Hence, the temporal intensity variability is largely limited to relative changes in intensity across roughly defined aerial units. As an illustration, in our motivating example of crime event data, the neighborhood structure of Cincinnati remains constant throughout the year. Hence, intensity dynamics are caused by crime increasing or decreasing in entire neighborhoods due to external factors such as increased police presence or sporting events. The single-θ model provides an efficient representation for such behavior through a dynamic re-weighting of static kernels.

The finite dimensional distribution $S_{\mathcal{T}}(\mathbf{v}; \boldsymbol{\alpha})$ is determined through specification of a stochastic process on the open unit interval. There are numerous different possible approaches for the modeling of such time series of proportions. The most common frameworks are based upon a transformation of a real-valued time series. For example, any normal dynamic linear model can, through either a *logit* or *probit* transformation, be expressed as a time series of proportions (see, e.g., Cargnoni et al., 1997). Our proposed Poisson DLM model for monthly integrated intensity, presented in Section 4.1.4, is an example of methodology in this spirit (albeit with respect to a log transform). There are, however, reasons why these models are inappropriate in the context of stick-breaking proportions. It is not usually possible to specify a transformation which will lead to beta distribution marginals for each $v_{l,t}$. Although this is not a theoretical constraint, beta marginals for the proportions at each time point lead to marginal beta process priors for the G_t and allow the modeler to make use of existing work related to these models (e.g.,

Ishwaran and Zarepour, 2000) in hyperprior specification and interpretation. More importantly, the standard transformations lead to autocorrelation functions for successive proportions that vary dramatically throughout the unit-interval. In particular, for both *logit* and *probit* transformations of normal AR processes, proportions near zero or one are much more highly correlated with the successive proportions than for proportions closer to 0.5. This behavior is especially undesirable for $S_{\mathcal{T}}(\mathbf{v};\boldsymbol{\alpha})$, as the possibility for θ_l components that are heavily weighted at one time point to become insignificant later is a key aspect of efficient prior full support.

One possible option would be to apply the power discount steady state model introduced by Smith (1979). A scheme for time series with exponential family observation distributions, this results in filtering and forecasting equations analogous to those of Kalman filtering. Indeed, Grunwald et al. (1993) specify a time series model on the simplex with Dirichlet marginal distributions through use of this approach. In our context of univariate proportions, a simple version of the model could be specified with observation equation $\Pr(v_t|\omega_t,\kappa) = \text{Be}(v_t;\, \omega_t\kappa,\, (1-\omega_t)\kappa)$ and evolution such that $\Pr(\omega_{t+1} = \omega \mid \mathbf{v}_1^t) = \left[\Pr(\omega_t = \omega|\mathbf{v}_1^t)\right]^{\delta} = [\text{BeC}(\omega;\alpha_t,\kappa,\sigma_t)]^{\delta}$, where the beta conjugate distribution $\text{BeC}(\omega\ ;\ \alpha,\kappa,\sigma)$ is proportional to $\exp\{\sigma\,(\kappa\omega\alpha - \log[\beta(\omega\kappa,(1-\omega)\kappa)])\}$ and here β denotes the beta function. Although filtering and forecasting are straightforward for this model ($\sigma_{t+1} = 1+\delta\sigma_t$ and $\alpha_{t+1} = [\delta\alpha_t + \text{logit}(v_t)]\,/[1+\delta]$), it is difficult to combine with the binomial likelihood for the $v_{l,t}$ in posterior sampling and the necessary process smoothing is computationally expensive. The framework of Grunwald et al. (1993) does, however, allow for the inclusion of external covariates and seasonal factors and

this model could provide a basis in the future for more complicated single-θ DDP prior models.

Our chosen approach is to model the time series of stick-breaking proportions as stationary positively correlated beta autoregressive processes (PBAR; introduced by McKenzie, 1985). The PBAR$(V_t|V_{t-1}; a, b, \rho)$ process evolution is defined

$$V_t = 1 - U_t(1 - W_t V_{t-1}) \text{ where } U_t \sim \text{Be}(b, a-\rho), W_t \sim \text{Be}(\rho, a-\rho), \tag{4.14}$$

with each U_t and W_t independent, for a and b positive and $0 < \rho < a$. The autocorrelation for a PBAR(a, b, ρ) process is

$$r(k) = (\mathbb{E}[W]\mathbb{E}[U])^k = \left(\frac{b\rho}{a(a+b-\rho)}\right)^k, \tag{4.15}$$

such that, as a function of ρ, $r(1)$ is strictly increasing and has a range of the entire unit interval. Note that the PBAR process is stationary: if $V_{t-1} \sim \text{Be}(a, b)$, then

$$\begin{aligned} W_t V_{t-1} \sim \text{Be}(\rho, a+b-\rho) &\Rightarrow 1 - W_t V_{t-1} \sim \text{Be}(a+b-\rho, \rho) \\ \Rightarrow U_t(1 - W_t V_{t-1}) \sim \text{Be}(b, a) &\Rightarrow 1 - U_t(1 - W_t V_{t-1}) \sim \text{Be}(a, b). \end{aligned} \tag{4.16}$$

Thus if we specify $S_{\mathcal{T}}(\mathbf{v}; \boldsymbol{\alpha}) = \text{PBAR}(\mathbf{v}; 1, \alpha, \rho)$ as the T dimensional distribution induced by PBAR evolution as in (4.14) and the assumption that $v_{l,1} \overset{iid}{\sim} \text{Be}(1, \alpha)$ for $l = 1, \ldots, \infty$, then each $v_{l,t}$ is marginally *iid* distributed $\text{Be}(1, \alpha)$ and the marginal prior for each G_t is $\text{DP}(\alpha, G_0(\psi))$. For this process parameterization, with $a = 1$ and $b = \alpha$ and $0 < \rho < 1$, the autocorrelation of (4.15) simplifies to $r(k) = [\rho\alpha/(1 + \alpha - \rho)]^k$.

The single-θ DDP mixture model based upon this process is

$$\begin{aligned} f_t(\mathbf{y}) &= \int \mathrm{k}(\mathbf{y};\theta)dG_t(\theta), \text{ for } t = 1,\ldots,T \\ \boldsymbol{G} &\sim \mathrm{DDP}(\mathrm{PBAR}(\mathbf{v};1,\alpha,\rho), G_0(\psi)). \end{aligned} \tag{4.17}$$

A PBAR prior specification for stick-breaking proportions offers flexible sample paths from a simple stationary process, easily accessible beta marginal distributions, and a clear correlation structure. For these reasons, we have adopted PBAR as the default form for S in all that follows.

4.1.4 Time Series Modeling for Integrated Poisson Intensity

We now turn to address parametric modeling for the integrated Poisson intensity $\boldsymbol{\gamma} = \boldsymbol{\gamma}_1^T = [\gamma_1,\ldots,\gamma_T]$, where $\gamma_t = \int_{\mathcal{Z}} \lambda_t(\mathbf{z})d\mathbf{z}$ for either $\mathcal{Z} = \mathcal{R}$ or $\mathcal{Z} = \mathcal{R}\times\mathcal{M}$ corresponding to unmarked or marked processes respectively. Recall that integrated Poisson intensity is independent of the process density in the likelihood (refer to (4.1)), such that the event counts $\mathbf{n} = \mathbf{n}_1^T = \{n_1,\ldots,n_T\}$ are the sufficient statistics for $\boldsymbol{\gamma}$. Thus, n_t given γ_t is distributed $\mathrm{Po}(n_t;\gamma_t)$ for $t = 1,\ldots,T$ and the count vector $\mathbf{n}$ is viewed as a time series of Poisson random variables correlated through the mean vector.

In modeling for non-normal time series, a basic approach is to apply a variant of the steady state power discount scheme proposed by Smith (1979) (refer to Section 4.1.3 for discussion of this model in the context of correlated beta random variables). Assuming that each n_t given γ_t is conditionally independent of the other counts and distributed $\mathrm{Po}(n_t;\gamma_t)$, the power discount time series for marginally Poisson distributed count data

is specified through the state equation $\Pr(\gamma_t = \gamma \mid \mathbf{n}_1^{t-1}) \propto \left[\Pr(\gamma_{t-1} = \gamma|\mathbf{n}_1^{t-1})\right]^{\delta}$ and the initial information $\pi(\gamma_0) = \text{Ga}(\gamma_0 \; ; \; \kappa_0, \; \eta_0)$, where $\delta \in (0, 1)$ is the power discount factor. If $\Pr(\gamma_{t-1} = \gamma \mid \mathbf{n}_1^{t-1}) = \text{Ga}(\gamma; \; \kappa_{t-1}, \; \eta_{t-1})$, then $\Pr(\gamma_t = \gamma \mid \mathbf{n}_1^{t-1}) = \text{Ga}(\gamma; \; \delta\kappa_{t-1} - \delta + 1, \; \delta\eta_{t-1})$ and the recursive filtering equation for $\Pr(\gamma_t|\mathbf{n}_1^t)$ holds that $\text{Po}(n_t; \; \gamma_t) \; \Pr(\gamma_t \mid \mathbf{n}_1^{t-1}) \propto \gamma_t^{n_t + \delta\kappa_{t-1} - \delta} \exp\left[-(1 + \delta\eta_{t-1})\gamma_t\right]$, such that $\kappa_t = n_t + \delta\kappa_{t-1} - \delta + 1$ and $\eta_t = 1 + \delta\eta_{t-1}$. The forecast distribution for n_{t+1} given $\mathbf{n}_1^t$ is straightforward to obtain as a negative binomial.

Our focus is largely on smoothing the time series; that is, obtaining the probability distribution for $\boldsymbol{\gamma}$ given $\mathbf{n}$. Recursive backwards sampling equations for this model are found through the assumption of conditional independence between γ_t and $\mathbf{n}_{t+1}^T$ given γ_{t+1}, and between γ_{t+1} and $\mathbf{n}_1^t$ given γ_t, such that $\Pr(\gamma_t|\mathbf{n}_1^T, \gamma_{t+1})$ is equal to $\Pr(\gamma_t \mid \gamma_{t+1}, \; \mathbf{n}_1^t) \propto \Pr(\gamma_t \mid \mathbf{n}_1^t) \; \Pr(\gamma_{t+1} \mid \gamma_t)$. The density $\Pr(\gamma_t|\mathbf{n}_1^t)$ is gamma, with parameters specified as above. Following the approach taken in Wheeler (2001), if we assume that the evolution $\Pr(\gamma_{t+1} \mid \gamma_t)$ depends only on the discount δ and the distance $\gamma_{t+1} - \gamma_t$, then $\Pr(\gamma_{t+1} \mid \gamma_t) = h_\delta(\gamma_{t+1} - \gamma_t)$ where h_δ is defined such that $\text{Ga}(\gamma_{t+1} \; ; \; \kappa_t, \; \eta_t) = \int_0^\infty h_\delta(\gamma_{t+1} - \gamma_t) \; \text{Ga}(\gamma_t \; ; \; \kappa_t, \; \eta_t) d\gamma_t$. Thus, backwards sampling is possible through repeated solution of a simple convolution integral equation.

The power discount scheme is appealingly simple, but it is not amenable to the introduction of more structured stochastic elements such as polynomial or periodic trends. Furthermore, due to the need for numerical evaluation of a convolution equation, smoothing for the model is relatively expensive and this negates much of the computational efficiency gained through accessible forward filtering equations. We

thus concentrate on an alternative modeling framework provided through conditionally Gaussian dynamic linear models (DLM), as proposed by Cargnoni et al. (1997) in the context of multinomial time series. The conditionally Gaussian DLM can itself be seen as an extension of the generalized DLM proposed by West et al. (1985), and both approaches seek to extend the powerful DLM framework beyond marginally Gaussian time series data.

In our context of a time series for Poisson distributed count data, we propose the conditionally Gaussian constant DLM, denoted PoDLM($\mathbf{n}, \boldsymbol{\gamma}; \{\mathrm{F}, \mathrm{G}, \kappa, W\}$),

$$\begin{aligned}
\text{Observation Equation :} \quad & n_t|\gamma_t \sim \mathrm{Po}(n_t; \gamma_t) \\
\text{Structural Equation :} \quad & \log(\gamma_t) = \mathrm{F}'\eta_t + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{N}(0, \kappa) \\
\text{State Equation :} \quad & \eta_t = \mathrm{G}\eta_{t-1} + \omega_t, \quad \omega_t \sim \mathrm{N}(0, W) \\
\text{Initial Information :} \quad & \eta_0 \sim \mathrm{N}(\mathrm{m}_0, \mathrm{C}_0),
\end{aligned}$$

where η is the $r \times 1$ state vector, F is the $r \times 1$ design vector, and G is a $r \times r$ evolution matrix. There is huge flexibility in specification of the model through changes to $\{\mathrm{F}, \mathrm{G}, \kappa, W\}$, and in non-constant models any element may be allowed to vary in time (refer to West and Harrison, 1997, for a complete treatment of DLM model specification and design). We discuss a simple 2nd order polynomial/seasonal model in Section 4.2.4 below.

Conditional on $\log(\boldsymbol{\gamma})$, κ, and W, the structural and state equations above specify a standard dynamic linear model with known variance components. Thus, conditional sampling for $\boldsymbol{\eta} = \{\eta_1, \ldots, \eta_T\}$ is a straightforward application of the forward-

filtering, backward-sampling algorithm proposed by Frühwirth-Schnatter (1994) and Carter and Kohn (1994). Full conditional posterior distributions for κ and W are available directly with the use of conditionally conjugate priors (or through use of a discount factor specification for W). Finally, sampling for the posterior $\boldsymbol{\gamma}$ given $\mathbf{n}$, $\boldsymbol{\eta}$, and κ, requires a simple Metropolis-Hastings step. Full details for this posterior simulation procedure are provided in Section 4.2.4.

4.2 Model Specification and Posterior Simulation

In this section we present detailed model specification under the framework of Section 4.1 for various types of Poisson processes, accompanied by MCMC algorithms for posterior simulation based on a finite stick-breaking truncation of the prior. For simplicity of notation, we assume throughout that $\mathcal{R} = (0,1) \times (0,1)$ (i.e. $\mathbf{y} = \tilde{\mathbf{y}}$ in the notation of Section 4.1.1). Illustration of these models follows in the data examples of Section 4.3.

In specification of prior and hyperprior parameters for each model, the general guidelines are unchanged from the discussion in Section 2.2. Recall that, under this approach, DP prior specification arises from consideration of a simplified model with a single kernel serving as the process density (i.e., the limiting case of a generic DP mixture model as $\alpha \to 0^+$). The data analysis of Section 4.3 includes particular examples of model parameterization which follow from this approach. In addition, prior parameterization issues unique to the modeling of dynamic processes are discussed in

the example of Section 4.3.3.

4.2.1 Nonhomogeneous Spatial Poisson Processes

We assume that the data $\mathcal{D} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ form a pattern of point events that are randomly distributed throughout $\mathcal{R}$. This spatial point pattern is modeled as a realization from $\text{PoP}(\mathcal{R}, \lambda)$, such that the intensity function is $\lambda(\mathbf{y}) = \gamma f(\mathbf{y})$, where $\pi(\gamma) \propto \gamma^{-1}$ and the prior for $f(\mathbf{y})$ is specified as in (4.2) and thereafter.

Model 1

$$\begin{aligned}
\{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \mid \gamma, f &\sim \text{PoP}\left(\mathcal{R}, \gamma f(\mathbf{y})\right) \\
\pi(\gamma) \propto \gamma^{-1} \quad \text{and} \quad f(\mathbf{y}; G) &= \int \text{B}_2(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\tau}, \varphi) dG(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi) \\
G|\alpha, \boldsymbol{\beta} &\sim DP(\alpha, G_0^{\mathbf{y}}(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi; \boldsymbol{\beta})) \\
G_0^{\mathbf{y}}(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi; \boldsymbol{\beta}) &= \prod_{i=1,2} \text{U}\left(\mu_i; 0, 1\right) \text{Ga}\left(\tau_i^{-1}; c_\beta, \beta_i\right) \text{U}(\varphi; C_{\boldsymbol{\mu}}, C^{\boldsymbol{\mu}})
\end{aligned}$$

where $C_{\boldsymbol{\mu}} = -[\max\{\mu_1\mu_2, (1-\mu_1)(1-\mu_2)\}]^{-1}$, $C^{\boldsymbol{\mu}} = -[\min\{\mu_1(\mu_2-1), \mu_2(\mu_1-1)\}]^{-1}$, $\pi(\alpha) = \text{Ga}(\alpha; a_\alpha, b_\alpha)$, and $\pi(\boldsymbol{\beta}) = \text{Ga}(\beta_1; a_\beta, b_\beta)\text{Ga}(\beta_2; a_\beta, b_\beta)$.

4.2.1.1 MCMC Posterior Simulation

First, the integrated intensity γ has posterior density $\text{Ga}(n, 1)$ independent of f. This holds true for all of the models specified in this chapter, save for the dynamic Poisson process model of Section 4.2.4 where we apply the conditionally Gaussian DLM methodology of Section 4.1.4.

Turning to posterior sampling for f, all of the MCMC simulation algorithms presented in this chapter are based upon a finite stick-breaking approximation to the

infinite dimensional DP distributed random mixing distribution, G. Sampling then proceeds through an extended version of the blocked Gibbs algorithm described in Section (3.2.3). As a result of this truncation, we will not in practice have an infinite dimensional, fully nonparametric, prior for the random mixing measure. However, the modeling is entirely motivated by the nonparametric prior framework, and may be used in conjunction with a level of finite truncation which ensures that the resultant inference is practically indistinguishable from that which would result from the infinite dimensional prior model. Obviously, the reduction of an infinite dimensional parameter space to a finite one will lead to more efficient posterior sampling. But more importantly, any inference about Poisson intensity integrated over a subregion of $\mathcal{R}$ relies upon a truncated approximation to G, regardless of the method used to simulate process parameters, and modeling based upon such finite truncation throughout provides a more consistent approach to inference. Thus, the framework depends upon a very high-dimensional prior model which is motivated by an infinite dimensional process.

In the context of **Model 1**, the hierarchical specification for $f(\mathbf{y}; G)$ and for G is replaced by

$$\begin{aligned}
\mathbf{y}_i|\boldsymbol{\vartheta}, \mathbf{k} &\sim \mathrm{B}_2\left(\mathbf{y}_i; \vartheta_{k_i} = \left[\boldsymbol{\mu}_{k_i}, \boldsymbol{\tau}_{k_i}, \varphi_{k_i}\right]\right) \\
k_i|\mathbf{p} &\sim \sum_{l=1}^{L} p_l \delta_{[l]}(k_i), \text{ for } i = 1, \ldots, n \\
\mathbf{p}, \boldsymbol{\vartheta}|\alpha, \boldsymbol{\beta} &\sim \mathcal{P}_L\left(\mathbf{p}; \mathrm{Be}(v; 1, \alpha)\right) \prod_{l=1}^{L} dG_0^{\mathbf{y}}(\vartheta_l; \boldsymbol{\beta})
\end{aligned}$$

where the stick-breaking prior $\mathcal{P}_L\left(\mathbf{p}; \mathrm{Be}(v; 1, \alpha)\right)$ is defined constructively as in equation (2.10).

Introducing the variable $H_l = \sum_{i=1}^{n} \delta_{[k_i=l]}$, representing the number of data observations allocated to each mixture component, the set of kernel parameters, $\boldsymbol{\vartheta} = \{\vartheta_1, \ldots, \vartheta_L\}$, is then partitioned into two subsets. The first, $\boldsymbol{\vartheta}^\star = \{\vartheta_l : H_l > 0\}$, includes the $n^\star$ parameters that are allocated to data observations and the complement of this subset, $\boldsymbol{\vartheta}^u = \{\vartheta_l : H_l = 0\}$, includes the n^u unallocated parameters such that $n^\star + n^u = n$ and $\boldsymbol{\vartheta}^\star \cap \boldsymbol{\vartheta}^u = \emptyset$. Introduce the index vector $\mathbf{k}^\star$ such that observation $\mathbf{y}_i$ is allocated to component $\vartheta^\star_{k^\star_i}$. Then sampling from the posterior full conditional for $\boldsymbol{\vartheta}$,

$$\Pr\left(\boldsymbol{\vartheta}|\mathbf{k}, \boldsymbol{\beta}, \mathcal{D}\right) \propto \prod_{j=1}^{n^\star} \left[dG_0(\vartheta^\star_j; \boldsymbol{\beta}) \prod_{i:k^\star_i=j} \mathrm{B}_2(\mathbf{y}_i; \vartheta^\star_j) \right] \prod_{j=1}^{n^u} dG_0(\vartheta^u_j; \boldsymbol{\beta}), \tag{4.18}$$

is possible through independent draws of each $\vartheta^u_j \sim G_0(\vartheta^u_j; \boldsymbol{\beta})$, for $j = 1, \ldots, n^u$, and from $\Pr(\vartheta^\star_j|\mathbf{k}, \boldsymbol{\beta}, \mathcal{D}) \propto dG_0(\vartheta^\star_j; \boldsymbol{\beta}) \prod_{i:k^\star_i=j} \mathrm{B}_2(\mathbf{y}_i; \vartheta^\star_j)$, for $j = 1, \ldots, n^\star$.

Independent draws for each $\vartheta^\star_j$ from this latter distribution will be broken into two full conditional draws. First, draw each $(\boldsymbol{\mu}^\star_j, \varphi^\star_j)$ from

$$\Pr(\boldsymbol{\mu}^\star_j, \varphi^\star_j|\boldsymbol{\tau}^\star_j, \mathbf{k}, \mathcal{D}) \propto \prod_{i:k^\star_i=j} \mathrm{B}_2(\mathbf{y}_i; \boldsymbol{\mu}^\star_j, \boldsymbol{\tau}^\star_j, \varphi^\star_j) \frac{1}{C^{\boldsymbol{\mu}^\star_j} - C_{\boldsymbol{\mu}^\star_j}} \tag{4.19}$$

by proposing $\boldsymbol{\mu}' \sim g_{\boldsymbol{\mu}}(\boldsymbol{\mu}'; \boldsymbol{\mu}^\star_j)$, where g is symmetric such that $g(\boldsymbol{\mu}'; \boldsymbol{\mu}^\star_j) = g_{\boldsymbol{\mu}}(\boldsymbol{\mu}^\star_j; \boldsymbol{\mu}')$, followed by a proposal $\rho' \sim \mathrm{U}(C^{\boldsymbol{\mu}'}, C_{\boldsymbol{\mu}'})$. The MCMC move $(\boldsymbol{\mu}^\star_j, \varphi^\star_j) \to (\boldsymbol{\mu}', \varphi')$ is then accepted with probability $\max\left\{1, \prod_{i:k^\star_i=j} \mathrm{B}_2(\mathbf{y}_i; \boldsymbol{\mu}', \boldsymbol{\tau}^\star_j, \varphi') / \mathrm{B}_2(\mathbf{y}_i; \boldsymbol{\mu}^\star_j, \boldsymbol{\tau}^\star_j, \varphi^\star_j)\right\}$. A suitable option for $g_{\boldsymbol{\mu}}$ is to use independent $\mathrm{PBAR}(\mu'_i|\mu^\star_{j,i}; 1, 1, \rho_\mu)$ for each of $i = 1, 2$ as a symmetric proposal scheme over the unit square (i.e. draw $U \sim \mathrm{Be}(1, 1-\rho_\mu)$ and $W \sim \mathrm{Be}(\rho_\mu, 1-\rho_\mu)$ and set $\mu'_i = 1 - U(1 - W\mu^\star_{j,i})$); recall that the PBAR process is

time-reversible. The next full conditional draw, for each $\boldsymbol{\tau}_j^\star$ from

$$\Pr(\boldsymbol{\tau}_j^\star|\boldsymbol{\mu}_j^\star,\varphi_j^\star,\mathbf{k},\boldsymbol{\beta},\mathcal{D}) \propto \prod_{i:k_i^\star=j} \mathrm{B}_2(\mathbf{y}_i;\boldsymbol{\mu}_j^\star,\boldsymbol{\tau}_j^\star,\varphi_j^\star)\mathrm{Ga}(\tau_{j1}^{\star-1};c_\beta,\beta_1)\mathrm{Ga}(\tau_{j2}^{\star-1};c_\beta,\beta_2), \quad (4.20)$$

proceeds with a proposed move to $\boldsymbol{\tau}' \sim g_{\boldsymbol{\tau}}(\boldsymbol{\tau}';\boldsymbol{\tau}_j^\star)$, where $g_{\boldsymbol{\tau}}$ is symmetric such that $g_{\boldsymbol{\tau}}(\boldsymbol{\tau}';\boldsymbol{\tau}_j^\star) = g_{\boldsymbol{\tau}}(\boldsymbol{\tau}_j^\star;\boldsymbol{\tau}')$, which is then accepted with probability

$$\max\left\{1, \prod_{i:k_i^\star=j}\left[\frac{\mathrm{B}_2(\mathbf{y}_i;\boldsymbol{\mu}_j^\star,\boldsymbol{\tau}',\varphi_j^\star)}{\mathrm{B}_2(\mathbf{y}_i;\boldsymbol{\mu}_j^\star,\boldsymbol{\tau}_j^\star,\varphi_j^\star)}\right]\frac{\mathrm{Ga}(\tau_1'^{-1};c,\beta_1)\mathrm{Ga}(\tau_2'^{-1};c,\beta_2)}{\mathrm{Ga}(\tau_{j1}^{\star-1};c,\beta_1)\mathrm{Ga}(\tau_{j2}^{\star-1};c,\beta_2)}\right\}.$$

We use $g_{\boldsymbol{\tau}}(\boldsymbol{\tau}';\boldsymbol{\tau}_j^\star)$ defined such that $\log(\tau_i'^{-1}) \sim \mathrm{N}(\log(\tau_{ji}^{\star-1}),\sigma_\tau^2)$ for each of $i = 1, 2$.

The draw for $\mathbf{k}$ conditional on ϑ, $\mathbf{p}$, and $\mathcal{D}$ is straightforward, since for $i = 1,\ldots,n$, each k_i is independently distributed $\Pr(k_i = j|\boldsymbol{\vartheta},\mathbf{p},\mathbf{y}_i) \propto \sum_{l=1}^{L} p_l\ \mathrm{B}_2(\mathbf{y}_i;\vartheta_l)\ \delta_l(j)$. Finally, by conjugacy of the stick-breaking prior to multinomial sampling and the fact that $\mathbf{p}$ is conditionally independent of the data given $\mathbf{k}$ (see Ishwaran and James (2001) and refer to the relevant discussion in Section 3.2.3), we can sample directly from the conditional posterior for $\mathbf{p}$ given $\mathbf{k}$ and α by drawing $v_l \sim \mathrm{Be}(1+H_l, \alpha+\sum_{j=l+1}^{L} H_j)$ for $l = 1,\ldots,L-1$, where H_l is (as above) the number of observations allocated to kernel parameter component ϑ_l, before setting $v_L = 1$, $p_1 = v_1$, and $p_l = \prod_{j=1}^{l-1}(1-v_j)v_l$ for $l = 2,\ldots,L$.

Direct sampling from full conditional posterior distributions is possible for each of the hyperparameters α and $\boldsymbol{\beta}$. As outlined in Ishwaran and Zarepour (2000) and in Section 3.2.3, α is independent from $\mathbf{k}$ and $\boldsymbol{\vartheta}$ given $\mathbf{p}$ and, following from the generalized Dirichlet density representation in Connor and Mosimann (1969) for random variables drawn from a finite stick-breaking prior, $\Pr(\alpha \mid \mathbf{p}) \propto \alpha^{L-1}\ p_L^{\alpha-1}\ \pi(\alpha) \propto \alpha^{a_\alpha+L-2}$ $\exp\left[\alpha\log(p_L) - \alpha b_\alpha\right]$, such that α given $\mathbf{p}$ is sampled from $\mathrm{Ga}(a_\alpha + L - 1, b_\alpha - \log(p_L))$.

The draw for $\boldsymbol{\beta}$ is facilitated by noticing that since, as seen in (4.18), the unallocated $\boldsymbol{\vartheta}^u$ have just been drawn from $G_0(\boldsymbol{\beta})$, the conditional joint posterior for $\{\boldsymbol{\beta}, \boldsymbol{\vartheta}\}$ can be marginalized over $\boldsymbol{\vartheta}^u$ to obtain the full conditional posterior for $\boldsymbol{\beta}$,

$$\Pr(\boldsymbol{\beta}|\boldsymbol{\tau}^\star, n^\star) \propto \pi(\boldsymbol{\beta}) \prod_{j=1}^{n^\star} \mathrm{Ga}(\tau_{j1}^{\star -1}; c_\beta, \beta_1)\mathrm{Ga}(\tau_{j2}^{\star -1}; c_\beta, \beta_2), \tag{4.21}$$

such that each β_i is drawn independently from a $\mathrm{Ga}(n^\star c_\beta + a_\beta, \sum_{j=1}^{n^\star} \tau_{ji}^{\star -1} + b_\beta)$ distribution.

4.2.2 Spatial Poisson Processes with Categorical Marks

We assume that the data $\mathcal{D} = \{(\mathbf{y}_1, m_1), \ldots, (\mathbf{y}_n, m_n)\}$ form a pattern of point events distributed throughout $\mathcal{R}$ and accompanied by random marks with support $\mathcal{M} = \{1, \ldots, M\}$. This spatial point pattern is modeled as a realization from the joint location-mark Poisson process $\mathrm{PoP}(\mathcal{R} \times \mathcal{M}, \lambda)$, with joint intensity such that $\lambda(\mathbf{y}, m) = \gamma f(\mathbf{y}, m)$, where the marginal intensity for location is modeled as above in Section 4.2.1 and the mark kernel is a multinomial density. Again, $\pi(\gamma) \propto \gamma^{-1}$ and the posterior for γ is $\mathrm{Ga}(n, 1)$.

Model 2

$$\begin{aligned}
\{(\mathbf{y}_1, m_1), \ldots, (\mathbf{y}_n, m_n)\} \mid \gamma, f &\sim \mathrm{PoP}\left(\mathcal{R} \times \mathcal{M}, \gamma f(\mathbf{y}, m)\right), \quad \pi(\gamma) \propto \gamma^{-1} \\
f(\mathbf{y}, m; G) &= \int \mathrm{B}_2(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\tau}, \varphi) q_m dG(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi, Q) \\
G(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi, Q) &\sim \mathrm{DP}\left(\alpha, G_0^{\mathbf{y}}(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi; \boldsymbol{\beta})\mathrm{Dir}(Q; \mathbf{a}^q)\right)
\end{aligned}$$

where $G_0^{\mathbf{y}}$ and $\pi(\alpha, \boldsymbol{\beta})$ are defined as for **Model 1**, $Q = [q_1, \ldots, q_M]$ and $\mathbf{a}^q = \{ a_1^q, \ldots, a_M^q\}$ is fixed.

4.2.2.1 MCMC Posterior Simulation

The finite stick-breaking truncation for this model is obtained by replacing the hierarchical specification for $f(\mathbf{y}, m; G)$ and for G with

$$\begin{aligned}
\mathbf{y}_i, m_i|\boldsymbol{\vartheta}, \mathbf{Q}, \mathbf{k} &\sim \mathrm{B}_2\left(\mathbf{y}_i; \vartheta_{k_i}\right) q_{k_i m_i} \\
k_i|\mathbf{p} &\sim \sum_{l=1}^{L} p_l \delta_{[l]}(k_i), \text{ for } i = 1, \ldots, n \\
\mathbf{p}, \boldsymbol{\vartheta}, \mathbf{Q}|\alpha, \boldsymbol{\beta} &\sim \mathcal{P}_L\left(\mathbf{p}; \mathrm{Be}(v; 1, \alpha)\right) \prod_{l=1}^{L} dG_0^{\mathbf{y}}(\vartheta_l; \boldsymbol{\beta}) \mathrm{Dir}(Q_l; \mathbf{a}^q),
\end{aligned}$$

where $\boldsymbol{\vartheta}$ is as in Section 4.2.1 and $\mathbf{Q} = \{Q_1, \ldots, Q_L\}$ with $Q_l = [q_{l1}, \ldots, q_{lM}]$.

The MCMC posterior simulation algorithm is very similar to that outlined in 4.2.1.1. Due to independence of the kernel components, $\mathbf{Q}$ is conditionally independent of $\boldsymbol{\vartheta}$ given $\mathbf{k}$. Defining $\mathbf{Q}^\star = \{Q_l : H_l > 0\}$ and $\mathbf{Q}^u = \{Q_l : H_l = 0\}$ in analogue to $\boldsymbol{\vartheta}^\star$ and $\boldsymbol{\vartheta}^u$ from Section 4.2.1, the full conditional joint posterior for $\boldsymbol{\vartheta}$ and $\mathbf{Q}$ given $\mathbf{k}$ and $\mathcal{D}$ is obtained through multiplication of (4.18) by

$$\prod_{j=1}^{n^\star} \left[\mathrm{Dir}(Q_j^\star; \mathbf{a}^q) \prod_{i:k_i^\star = j} q_{jm_i}^\star \right] \prod_{j=1}^{n^u} \mathrm{Dir}(Q_j^u; \mathbf{a}^q), \tag{4.22}$$

Thus, full conditional posterior sampling for $\boldsymbol{\vartheta}$ and $\mathbf{Q}$ resolves into independent draws from the full conditional posterior for $\boldsymbol{\vartheta}$, exactly as described in Section 4.2.1.1, and a draw for $\mathbf{Q}$ proportional to (4.22). In this latter draw, the conditional independence structure of (4.18) remains intact such that we can sample each Q_j^u and $Q_j^\star$ independently. From (4.22), we see that $Q_j^u \overset{iid}{\sim} \mathrm{Dir}(Q_j^u; \mathbf{a}^q)$ for $j = 1, \ldots, n^u$ and that the full conditional posterior for each $Q_j^\star$, $j = 1, \ldots, n^\star$, is $\mathrm{Dir}(Q_j^\star; \mathbf{a}^q + \mathbf{s}_j)$, where $s_{j,r} = \sum_{i:k_i^\star = j} \delta_{[m_i = r]}$, for $r = 1, \ldots, M$, is the number of components allocated to $Q_j^\star$

with $m_i = r$.

Sampling of $\mathbf{k}$ is structurally unchanged from the procedure for unmarked processes, and for $i = 1, \ldots, n$, each k_i is independently distributed $\Pr = (k_i = j |\ \boldsymbol{\vartheta},\ \mathbf{Q},\ \mathbf{p},\ \mathbf{y}_i,\ m_i) \propto \sum_{l=1}^{L} p_l\ \mathrm{B}_2(\mathbf{y}_i; \vartheta_l) q_{lm_i}\ \delta_l(j)$. The posterior full conditional distributions of the remaining parameter vector, $\mathbf{p}$, and the hyperparameters, α and $\boldsymbol{\beta}$, are unchanged by the introduction of marks and may be sampled exactly as in Section 4.2.1.1.

4.2.3 Spatial Poisson Processes with Positive Continuous Marks

We assume that the data $\mathcal{D} = \{(\mathbf{y}_1, m_1), \ldots, (\mathbf{y}_n, m_n)\}$ form a pattern of point events distributed throughout $\mathcal{R}$ and accompanied by random marks with support $\mathcal{M} = \mathbb{R}^+$. Model specification follows closely that of Section 4.2.2. Again, the spatial point pattern is modeled as a realization from the joint location-mark Poisson process $\mathrm{PoP}(\mathcal{R} \times \mathcal{M}, \lambda)$, with joint intensity such that $\lambda(\mathbf{y}, m) = \gamma f(\mathbf{y}, m)$. The marginal intensity for location is modeled as above in Section 4.2.1, but now the mark kernel is a log-normal density. As above, $\pi(\gamma) \propto \gamma^{-1}$ and the posterior for γ is $\mathrm{Ga}(n, 1)$.

Model 3

$$
\begin{aligned}
\{(\mathbf{y}_1, m_1), \ldots, (\mathbf{y}_n, m_n)\} \mid \gamma, f &\sim \mathrm{PoP}\left(\mathcal{R} \times \mathcal{M}, \gamma f(\mathbf{y}, m)\right) \quad \pi(\gamma) \propto \gamma^{-1} \\
f(\mathbf{y}, m; G) &= \int \mathrm{B}_2(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\tau}, \varphi) \mathrm{N}(\log(m); q_1, q_2) dG\left(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi, Q\right) \\
G(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi, Q) &\sim \mathrm{DP}\left(\alpha, G_0^{\mathbf{y}}(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi; \boldsymbol{\beta}) \mathrm{N}(q_1; s_1, s_2) \mathrm{Ga}(q_2^{-1}; s_4, s_3)\right)
\end{aligned}
$$

where $G_0^{\mathbf{y}}$ and $\pi(\alpha, \boldsymbol{\beta})$ are defined as for **Model 1**, $Q = [q_1, q_2]$, $\pi(s_1) = \mathrm{N}(s_1; a_1, b_1)$, $\pi(s_2^{-1}) = \mathrm{Ga}(s_2; a_2, b_2)$, $\pi(s_3) = \mathrm{Ga}(s_3; a_3, b_3)$, and s_4 is fixed.

4.2.3.1 MCMC Posterior Simulation

The finite stick-breaking truncation for this model is obtained by replacing the hierarchical specification for $f(\mathbf{y}, m; G)$ and for G with

$$\begin{aligned}
\mathbf{y}_i, m_i|\boldsymbol{\vartheta}, \mathbf{Q}, \mathbf{k} &\sim \mathrm{B}_2\left(\mathbf{y}_i; \vartheta_{k_i}\right) \mathrm{N}(\log(m_i); q_{k_i 1}, q_{k_i 2}) \\
k_i|\mathbf{p} &\sim \sum_{l=1}^{L} p_l \delta_{[l]}(k_i), \text{ for } i = 1, \ldots, n \\
\mathbf{p}, \boldsymbol{\vartheta}, \mathbf{Q}|\alpha, \boldsymbol{\beta}, \mathbf{s} &\sim \mathcal{P}_L\left(\mathbf{p}; \mathrm{Be}(v; 1, \alpha)\right) \prod_{l=1}^{L} dG_0^{\mathbf{y}}(\vartheta_l; \boldsymbol{\beta}) \mathrm{N}(q_{l1}; s_1, s_2) \mathrm{Ga}(q_{l2}^{-1}; s_4, s_3),
\end{aligned}$$

where $\boldsymbol{\vartheta}$ is as in Section 4.2.1, and $\mathbf{Q} = \{Q_1, \ldots, Q_L\}$ with $Q_l = [q_{l1}, q_{l2}]$.

MCMC posterior simulation is developed following the same approach taken in Section 4.2.2.1 to extend the algorithm from Section 4.2.1.1. Now, the full conditional joint posterior for $\boldsymbol{\vartheta}$ and $\mathbf{Q}$ given $\mathbf{k}$, $\boldsymbol{\beta}$, $\mathbf{s}$, and $\mathcal{D}$ is obtained through multiplication of (4.18) by

$$\begin{aligned}
\prod_{j=1}^{n^\star} \left[\mathrm{N}(q_{j1}^\star; s_1, s_2) \mathrm{Ga}(q_{j2}^{\star -1}; s_4, s_3) \prod_{i:k_i^\star = j} \mathrm{N}(\log(m_i); q_{j1}^\star, q_{j2}^\star) \right] \\
\cdot \prod_{j=1}^{n^u} \mathrm{N}(q_{j1}^u; s_1, s_2) \mathrm{Ga}(q_{j2}^{u-1}; s_4, s_3)
\end{aligned} \tag{4.23}$$

after partitioning $\mathbf{Q}$ into allocated $\mathbf{Q}^\star$ and unallocated $\mathbf{Q}^u$. The location kernel parameters, $\boldsymbol{\vartheta}$, can be sampled as in Section 4.2.1.1, and the unallocated $\mathbf{Q}^u$ are sampled through *iid* draws of $[q_{j1}^u, q_{j2}^u]$ from $\mathrm{N}(q_{j1}^u; s_1, s_2)\mathrm{Ga}(q_{j2}^{u-1}; s_4, s_3)$ for $j = 1, \ldots, n^u$. For $j = 1, \ldots, n^\star$, the independent posterior full conditional for each $[q_{j1}^\star, q_{j2}^\star]$ is proportional to $\mathrm{N}(q_{j1}^\star; s_1, s_2)$ $\mathrm{Ga}(q_{j2}^{\star -1}; s_4, s_3)$ $\prod_{i:k_i^\star = j} \mathrm{N}(\log(m_i); q_{j1}^\star, q_{j2}^\star)$. This leads to two

conditional draws

$$\begin{aligned} q^{\star}_{j1}|q^{\star}_{j2}, \{m_i : k^{\star}_i = j\} &\sim \mathrm{N}\left(q^{\star}_{j1};\ D_j\left[\frac{\psi_1}{\psi_2} + \frac{E_j}{q^{\star}_{j2}}\right],\ D_j\right) \\ q^{\star -1}_{j2}|q^{\star}_{j1}, \{m_i : k^{\star}_i = j\} &\sim \mathrm{Ga}\left(q^{\star -1}_{j2};\ s_4 + \frac{H^{\star}_j}{2}, s_3 + \frac{1}{2}\sum_{i:k^{\star}_i=j}(\log(m_i) - q^{\star}_{j1})^2\right) \end{aligned} \tag{4.24}$$

where $D_j = \left[s_2^{-1} + H^{\star}_j q^{\star -1}_{j2}\right]^{-1}$, $E_j = \sum_{i:k^{\star}_i=j}\log(m_i)$, and $H^{\star}_j$ is the number of observations allocated to $Q^{\star}_j$.

For $i = 1, \ldots, n$, each k_i is independently distributed $\Pr\left(k_i = j|\boldsymbol{\vartheta}, \mathbf{Q}, \mathbf{p}, \mathbf{y}_i, m_i\right)$ $\propto \sum_{l=1}^{L} p_l\ \mathrm{B}_2(\mathbf{y}_i; \vartheta_l)\mathrm{N}(\log(m_i); q_{l1}, q_{l2})\ \delta_l(j)$. As in Section 4.2.2, the posterior full conditional distributions of the remaining parameter vector, $\mathbf{p}$, and the hyperparameters, α and $\boldsymbol{\beta}$, are unchanged by the introduction of marks and may be sampled exactly as in Section 4.2.1.1. Finally, since the unallocated $\mathbf{Q}^u$ have just been drawn from $\mathrm{N}(q^u_1; s_1, s_2)\mathrm{Ga}(q^{u-1}_2; s_4, s_3)$, the the conditional joint posterior for $\{\mathbf{s}, \mathbf{Q}\}$ may be marginalized over $\mathbf{Q}^u$ to obtain the full conditional posterior for $\mathbf{s}$,

$$\Pr(\mathbf{s}|\mathbf{Q}^{\star}, n^{\star}) \propto \pi(\mathbf{s})\prod_{j=1}^{n^{\star}}\mathrm{N}(q^{\star}_{j1}; s_1, s_2)\mathrm{Ga}(q^{\star -1}_{j2}; s_4, s_3). \tag{4.25}$$

Thus $\mathbf{s}$ is sampled through three conditional draws

$$\begin{aligned} s_1|s_2, \mathbf{q_1}^{\star}, n^{\star} &\sim \mathrm{N}\left(s_1;\ D_s\left[\frac{a_1}{b_1} + \frac{E_s}{s_2}\right], D_s\right) \\ s_2^{-1}|s_1, \mathbf{q_2}^{\star}, n^{\star} &\sim \mathrm{Ga}\left(s_2^{-1};\ a_2 + \frac{n^{\star}}{2}, b_2 + \frac{1}{2}\sum_{j=1}^{n^{\star}}(q^{\star}_{j1} - s_1)\right) \\ s_3|\mathbf{q_2}^{\star}, n^{\star} &\sim \mathrm{Ga}\left(s_3;\ a_3 + n^{\star}s_4, b_3 + \sum_{j=1}^{n^{\star}} q^{\star -1}_{j2}\right) \end{aligned} \tag{4.26}$$

where $D_s = \left[b_1^{-1} + n^{\star}s_2^{-1}\right]^{-1}$ and $E_s = \sum_{j=1}^{n^{\star}} q^{\star}_{j1}$.

4.2.4 Dynamic Discrete Time Spatial Poisson Processes

Our basic model for dynamic Poisson processes (both marked and unmarked) invokes a single-θ DDP extension of the static models above. Thus, the canonical model holds that the data $\mathcal{D} = \{(\mathbf{z}_1^1, \ldots, \mathbf{z}_{n_1}^1), \ldots, (\mathbf{z}_1^T, \ldots, \mathbf{z}_{n_T}^T)\}$ consisting of T point patterns realized on the observation window (possibly including mark support) $\mathcal{Z}$ and observed over time indices $t \in \mathcal{T} = \{1, \ldots, T\}$, are the realization of a dynamic Poisson process $\mathrm{PoP}(\mathcal{Z}, \boldsymbol{\gamma}\mathbf{f}(\mathbf{z})) = \{\ \mathrm{PoP}\left(\mathcal{Z}, \gamma_1 f_1(\mathbf{z})\right), \ldots, \mathrm{PoP}\left(\mathcal{Z}, \gamma_T f_T(\mathbf{z})\right)\ \}$, where $\boldsymbol{\gamma}$ is modeled as in Section 4.1.4 as a conditionally Gaussian DLM and the prior for process densities, f_t, is a single-θ DDP extension of the appropriate DP mixture model for $\mathbf{z}$ (one of the models in Sections 4.2.1, 4.2.2, and 4.2.3 above). We specify here the most basic model, for a dynamic unmarked nonhomogeneous Poisson process observed in discrete time.

Model 4

$$\begin{aligned} \{\mathbf{y}_1^t, \ldots, \mathbf{y}_{n_t}^t\} \mid \gamma_t, f_t \quad &\overset{ind}{\sim} \quad \mathrm{PoP}\left(\mathcal{R}, \gamma_t f_t(\mathbf{y})\right) \text{ for } t = 1, \ldots, T \\ \boldsymbol{\gamma} \sim \mathrm{DLM}(\log(\boldsymbol{\gamma}); \{\mathrm{F}, \mathrm{G}, \kappa, W\}) \quad &\text{and} \quad f(\mathbf{y}; G_t) = \int \mathrm{B}_2(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\tau}, \varphi) dG_t(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi) \\ \boldsymbol{G} \quad &\sim \quad \mathrm{DDP}\left(\mathrm{PBAR}(\mathbf{v}; 1, \alpha, \rho), G_0^{\mathbf{y}}(\boldsymbol{\mu}, \boldsymbol{\tau}, \varphi; \boldsymbol{\beta})\right) \end{aligned}$$

where $\boldsymbol{G} = \{G_1, \ldots, G_t\}$, $G_0^{\mathbf{y}}$ and $\pi(\boldsymbol{\beta})$ are defined as for **Model 1**, $\pi(\alpha) = \mathrm{Ga}(\alpha; a_\alpha, b_\alpha)$, and $\pi(\rho) = \mathrm{U}(\rho; 0, 1)$. Recall that PBAR($\mathbf{v}$; 1, α, ρ) is the finite dimensional distribution for $\mathbf{v} = [v_1, \ldots, v_T]$ induced by a PBAR(1, α, ρ) process over the indices in $\mathcal{T}$. DLM($\cdot$; $\{F, G, \kappa, W\}$) refers to the standard normal DLM, such that the implied model for $\mathbf{n}$ is the conditionally Gaussian Poisson DLM specified in Section 4.1.4. Thus, $\log(\gamma_t)$ $\sim \mathrm{N}(\mathrm{F}'\eta_t, \kappa)$ where $\eta_t \sim \mathrm{N}(\mathrm{G}\eta_{t-1}, W)$. The first variance component is assigned prior

$\pi(\kappa) = \text{Ga}(\kappa^{-1}; a_\kappa, b_\kappa)$. The prior for W must reflect the (usually block diagonal) structure of G, and the conditionally conjugate form is inverse Wishart. Alternatively, and this is the approach utilized below, the system variance may be implied through the use of a discount factor, $\delta \in [.9, .99]$, such that $\text{var}[\eta_t|\boldsymbol{\gamma}_1^{t-1}] = \text{var}[G\eta_{t-1}|\boldsymbol{\gamma}_1^{t-1}]/\delta$ (refer to West and Harrison, 1997, for a complete account of the discount factor approach to DLM modeling). Specification is completed with a prior for the initial state vector, $\pi(\eta_0) = \text{N}(\mathbf{m}_0, \mathbf{C}_0)$.

4.2.4.1 MCMC Posterior Simulation

We first describe posterior simulation for parameters related to the process density, f, with methodology corresponding to the Poisson DLM prior for $\mathbf{n}$ and $\boldsymbol{\gamma}$ to follow. The finite stick-breaking truncation of **Model 4** is obtained by replacing the hierarchical specification for $\mathbf{f}(\mathbf{y}, m; \boldsymbol{G})$ and for $\boldsymbol{G}$ with

$$\begin{aligned}
\mathbf{y}_i^t|\boldsymbol{\vartheta}, \mathbf{k}_t &\overset{ind}{\sim} \text{B}_2(\mathbf{y}_i^t; \boldsymbol{\mu}_{k_{t,i}}, \boldsymbol{\tau}_{k_{t,i}}, \varphi_{k_{t,i}}) \text{ for } i = 1, \ldots, n_t,\ t = 1, \ldots, T \\
k_{t,i}|\mathbf{P} &\overset{ind}{\sim} \sum_{l=1}^{L} p_{l,t}\delta_{[l]}(k_{t,i}),\ \text{for } i = 1, \ldots, n_t,\ t = 1, \ldots, T \\
\vartheta_l|\boldsymbol{\beta} &\overset{iid}{\sim} G_0^{\mathbf{y}}(\vartheta_l; \boldsymbol{\beta}) \text{ for } l = 1, \ldots, L \\
\mathbf{P}|\alpha, \rho &\sim \mathcal{P}_L\left(\mathbf{P}; \text{PBAR}(\mathbf{v}; 1, \alpha, \rho)\right)
\end{aligned}$$

where the multivariate stick-breaking prior $\mathcal{P}_L\left(\mathbf{P}; \text{PBAR}(\mathbf{v}; 1, \alpha, \rho)\right)$ for $\mathbf{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_L\}$ is defined constructively such that $\mathbf{v}_1, \ldots, \mathbf{v}_{L-1} \overset{iid}{\sim} \text{PBAR}(\mathbf{v}; 1, \alpha, \rho)$, $\mathbf{v}_L = \mathbf{1}$, $\mathbf{p}_1 = \mathbf{v}_1$, and for $l = 2, \ldots, L$; $t \in \mathcal{T}$: $p_{l,t} = \prod_{j=1}^{l-1}(1 - v_{j,t})v_{l,t}$.

After relabeling $\{(\mathbf{y}_1^1, \ldots, \mathbf{y}_{n_1}^1), \ldots, (\mathbf{y}_1^T, \ldots, \mathbf{y}_{n_T}^T)\}$ as $\{\mathbf{y}_1^\star, \ldots, \mathbf{y}_{n_\mathcal{T}}^\star\}$ where $n_1 +$

$\ldots + n_T = n_{\mathcal{T}}$, partitioning $\boldsymbol{\vartheta}$ into allocated $\boldsymbol{\vartheta}^\star$ and unallocated $\boldsymbol{\vartheta}^u$ as in Section 4.2.1.1, and defining $\mathbf{k}^\star$ such that $\mathbf{y}_i^\star$ given $\{\boldsymbol{\vartheta}^\star, \mathbf{k}^\star\}$ is distributed $\mathrm{B}_2(\mathbf{y}_i^\star; \vartheta_{k_i^\star}^\star)$, sampling from the posterior full conditional for $\boldsymbol{\vartheta}$ follows exactly the same procedure as in Section 4.2.1.1. The posterior full conditional draw for $\boldsymbol{\beta}$ given $\boldsymbol{\vartheta}^\star$ is also unchanged. The draw for $\{\mathbf{k}_t : t \in \mathcal{T}\}$ is straightforward since for $i = 1, \ldots, n_t;\ t = 1, \ldots, T$, each $k_{t,i}$ is independently distributed $\Pr\left(k_{t,i} = j | \boldsymbol{\vartheta}, \mathbf{P}, \mathbf{y}_i^t\right) \propto \sum_{l=1}^L p_{l,t}\ \mathrm{B}_2(\mathbf{y}_i^t; \vartheta_l)\ \delta_l(j)$. All that remains for parameters related to the f_t is to sample from the posterior full conditional for $\mathbf{P}$, $\Pr(\mathbf{P}|\ \{\mathbf{k}_1, \ldots, \mathbf{k}_T\}, \alpha, \rho)$, and from $\Pr(\alpha, \rho\ |\mathbf{P})$.

With respect to the posterior full conditional for $\mathbf{P}$, the sufficient statistic is the L by T matrix $\mathbf{h}$, where $h_{l,t}$ is the number of data observations at time t allocated to mixture component ϑ_l (hence, $\sum_{l=1}^L h_{l,t} = n_t$). The draw for $\mathbf{P}$ is then

$$\begin{aligned}
\Pr(\mathbf{P}|\mathbf{h}, \alpha, \rho) &\propto \prod_{t=1}^T \left[\prod_{l=1}^L (v_{l,t})^{h_{l,t}} (1 - v_{l,t})^{\sum_{i=l+1}^L h_{i,t}}\right] \prod_{l=1}^L \mathrm{PBAR}\left(\mathbf{v}_l; 1, \alpha, \rho\right) \\
&\propto \prod_{l=1}^L \left[\mathrm{PBAR}\left(\mathbf{v}_l; 1, \alpha, \rho\right) \prod_{t=1}^T \mathrm{Bin}\left(h_{l,t}; \sum_{i=l}^L h_{i,t}, v_{l,t}\right)\right] \\
&\propto \prod_{l=1}^L \Pr(\mathbf{v}_l | \{h_{l,t}, \ldots, h_{L,t} :\ t \in \mathcal{T}\}, \alpha, \rho),
\end{aligned} \tag{4.27}$$

where $\mathrm{Bin}(\cdot\ ; n, p)$ denotes a binomial distribution with mean np, such that the prior independence between $\mathbf{v}_i$ and $\mathbf{v}_j$ for $i \neq j$ is maintained in the conditional posterior.

We define the variable $L^\star = L - 1$ if $\sum_{t=1}^T h_{L,t} \neq 0$, otherwise set $L^\star = \inf\left\{l \in 1, \ldots, L-1 : \sum_{i=l+1}^L \sum_{t=1}^T h_{i,t} = 0\right\}$. Then the vectors $\mathbf{v}_l$ for $L^\star < l < L$ are simply drawn from the PBAR($\mathbf{v}_l$; 1, α, ρ) prior ($\mathbf{v}_L$ is a vector of ones). For $l = 1, \ldots, L^\star$, the draw from each $\Pr(\mathbf{v}_l | \{h_{l,t}, \ldots, h_{L,t} :\ t \in \mathcal{T}\}, \alpha, \rho)$ proceeds through

forward-filtering and backwards-sampling based on sequential Monte Carlo techniques. Forward filtering of the likelihood, to obtain $\Pr(v_{l,T}|\ \{h_{l,t},\ldots,h_{L,t}:\ t\in\mathcal{T}\},\ \alpha,\rho)$, is possible with particle filtering. A wide ranging discussion of such methodology is available in Doucet et al. (2001). We present a basic approach here, making use of the bootstrap filter variation of sequential importance sampling. The smoothing step, recursive sampling for $\Pr(v_{l,t}|v_{l,t+1},\ \{h_{l,t},\ldots,h_{L,t}:\ t\in\mathcal{T}\},\ \alpha,\rho)$ for $t=T-1,\ldots,1$, is then possible through application of the particle smoothing algorithm for nonlinear time series described in Godsill et al. (2004). Thus, for $l=1,\ldots,L^{\star}$, proceed as follows.

Filtering

- Sample $\tilde{v}_{1,1},\ \ldots,\ \tilde{v}_{1,C}$ *iid* from $\Pr(v_{l,1}|\{h_{l,1},\ \ldots,\ h_{L,1}\},\ \alpha,\rho)=\mathrm{Be}(1+h_{l,1},\ \alpha+\sum_{i>l}h_{i,1})$ (recall the conjugacy of the generalized Dirichlet discussed in Section 3.2.3), and set $\omega_{1,1},\ldots,\omega_{1,C}$ equal to $1/C$.

- For $t=2,\ldots,T$:

  - Resample $\tilde{v}_{t-1,i_1},\ldots,\tilde{v}_{t-1,i_C}$ *iid* with replacement from $\tilde{v}_{t-1,1},\ \ldots,\ \tilde{v}_{t-1,C}$ with sampling probabilities $\Pr(\tilde{v}_{t-1,i})=\omega_{t-1,i}$.

  - For $j=1,\ldots,C$, draw $\tilde{v}_{t,j}\sim\mathrm{PBAR}(\tilde{v}_{t,j}|\tilde{v}_{t-1,i_j};1,\alpha,\rho)$ (i.e., draw $U\sim\mathrm{Be}(\alpha,1-\rho)$ and $W\sim\mathrm{Be}(\rho,1-\rho)$ and set $\tilde{v}_{t,j}=1-U(1-W\tilde{v}_{t-1,i_j})$) and set weights $\omega_{t,j}\propto\mathrm{Bin}(h_{l,t};\sum_{i\geq l}h_{i,t},\tilde{v}_{t,j})$.

Smoothing

- Sample $v_{l,T}$ from $\tilde{v}_{T,1},\ldots,\tilde{v}_{T,C}$ with probabilities $\omega_{T,1},\ldots,\omega_{T,C}$.

- For $t=T-1,\ldots,1$, draw $v_{l,t}$ from $\tilde{v}_{t,1},\ldots,\tilde{v}_{t,C}$ such that $\Pr(v_{l,t}=\tilde{v}_{t,i})=$

$\omega_{t|t+1,i}$, where $\omega_{t|t+1,i} \propto \omega_{t,i}\text{PBAR}(v_{l,t+1}|\tilde{v}_{t,i}; 1, \alpha, \rho)$, and the evolution density is

$\text{PBAR}(v_{t+1}|v_t; 1, \alpha, \rho)$

$$= \int_0^{\min\left\{1, \frac{v_{t+1}}{v_t}\right\}} \frac{1}{1 - wv_t}\text{Be}\left(\frac{1 - v_{t+1}}{1 - wv_t}; \alpha, 1 - \rho\right)\text{Be}\left(w; \rho, 1 - \rho\right) dw \quad (4.28)$$

$$= \int_{1-v_{t+1}}^{\min\left\{1, \frac{1-v_{t+1}}{1-v_t}\right\}} \frac{1}{uv_t}\text{Be}\left(\frac{u + v_{t+1} - 1}{uv_t}; \rho, 1 - \rho\right)\text{Be}\left(u; \alpha, 1 - \rho\right) du. \quad (4.29)$$

Evaluation of the conditional density thus requires numerical integration. Since the integrands in both (4.28) and (4.29) tend to infinity at the bounds of integration, adaptive quadrature methods were found to be unstable. However, straightforward Monte Carlo has proven to be quite successful in practice. A sequence in either w or u is randomly sampled from $\text{Be}(w;\ \rho,\ 1 - \rho)$ or $\text{Be}(u;\ \alpha,\ 1 - \rho)$ respectively, and the average of the remainder of the respective integrand evaluated over this sequence is used as an estimate of the integral. A preliminary sample of both w and u can be used to decide which of (4.28) or (4.29) produces a less variable estimator.

Sampling for the PBAR hyperparameters α and ρ conditional on $\mathbf{P}$ (or the untransformed $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_L\}$) proceeds through a Metropolis Hastings step. First, as described above, $\mathbf{v}_{L^\star+1}, \ldots, \mathbf{v}_L$ have just been sampled from the $\text{PBAR}(1, \alpha, \rho)$ prior. Hence, the joint posterior for $\{\mathbf{V}, \alpha, \rho\}$ may be marginalized over these values to obtain the posterior full conditional, $\Pr(\alpha,\ \rho \mid \mathbf{v}_1, \ldots, \mathbf{v}_{L^\star})$

$$\propto \text{U}(\rho;\ 0, 1)\text{Ga}(\alpha;\ a_\alpha, b_\alpha)\prod_{l=1}^{L^\star}\left[\text{Be}(v_{l,1}|1, \alpha)\prod_{t=2}^{T}\text{PBAR}(v_{l,t}|v_{l,t-1}; 1, \alpha, \rho)\right]$$

$$\propto \Pr(\alpha|v_{l,1}, \ldots, v_{L^\star,1})\text{U}(\rho;\ 0, 1)\prod_{t=2}^{T}\prod_{l=1}^{L^\star}\text{PBAR}(v_{l,t}|v_{l,t-1}; 1, \alpha, \rho), \quad (4.30)$$

where $\Pr(\alpha|v_{l,1},\ldots,v_{L^\star,1}) = \mathrm{Ga}\Big(L^\star + a_\alpha, b_\alpha + \log\left(\prod_{l=1}^{L^\star}(1-v_{l,1})\right)\Big) = \mathrm{Ga}(a_\alpha + L^\star, b_\alpha + \log(p_{L^\star+1,1}/v_{L^\star+1,1}))$. Also, note that since the PBAR is a time reversible process, $\mathrm{PBAR}(\rho'|\rho;a,b,r) = \mathrm{PBAR}(\rho|\rho';a,b,r)$ presents a symmetric proposal distribution for ρ on the unit interval. Thus a Metropolis-Hastings draw for (4.30), given the current state of the MCMC at (α,ρ), proceeds by proposing α' from $\mathrm{Ga}(\alpha';\ a_\alpha + L^\star,\ b_\alpha + \log(p_{L^\star+1,1}\ /\ v_{L^\star+1,1}))$, ρ' from $\mathrm{PBAR}(\rho'|\rho;1,1,r)$, and accepting a move from (α,ρ) to (α',ρ') with probability set to

$$\min\left\{1, \frac{\prod_{t=2}^{T}\prod_{l=1}^{L^\star}\mathrm{PBAR}(v_{l,t}|v_{l,t-1};1,\alpha',\rho')}{\prod_{t=2}^{T}\prod_{l=1}^{L^\star}\mathrm{PBAR}(v_{l,t}|v_{l,t-1};1,\alpha,\rho)}\right\}. \tag{4.31}$$

Note that the denominator of the acceptance probability has already been calculated during the draw for $\mathbf{P}$.

Finally, we detail the MCMC algorithm for sampling $\boldsymbol{\gamma}$ and the related prior and hyperprior parameters. In this, we largely follow the procedure outlined by Cargnoni et al. (1997), adapted for a Poisson observation equation. Conditional on the Poisson means $\boldsymbol{\gamma}$, the model specifies a standard normal DLM with $\log(\boldsymbol{\gamma}_1^T) = \{\log(\gamma_1), \ldots, \log(\gamma_T)\}$ as observations. First, conditional on κ and $\boldsymbol{\gamma}$ and with fixed δ (the discount factor through which the system variance is implicitly defined), we sample the state vector $\boldsymbol{\eta}$ through use of the forward-filtering, backward sampling algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994).

- For $t = 1,\ldots,T$, compute $\Pr(\eta_t|\log(\boldsymbol{\gamma}_1^t),\kappa,\delta) = \mathrm{N}(\eta_t;\mathrm{m}_t,\mathrm{C}_t)$ through direct application of the sequential updating equations for a normal DLM (West and Harrison, 1997, chap. 4). In detail, with $a_t = \mathrm{G}\mathrm{m}_{t-1}$; $R_t = \mathrm{G}\mathrm{C}_{t-1}\mathrm{G}'/\delta$; $Q_t = \mathrm{F}'R_t\mathrm{F} + \kappa$;

and $A_t = R_t \mathrm{F} Q_t^{-1}$; $\mathrm{m}_t = a_t + A_t\,[\log(\gamma_t) - \mathrm{F}' a_t]$ and $\mathrm{C}_t = R_t - A_t Q_t A_t'$.

- Sample $\eta_T \sim \mathrm{N}(\mathrm{m}_T, \mathrm{C}_T)$.

- For $t = T-1, \ldots, 0$, sample η_t from $\Pr(\eta_t|\ \eta_{t+1},\ \log(\boldsymbol{\gamma}_1^T),\ \kappa,\ \delta) \propto \mathrm{N}(\eta_t\ ;\ \mathrm{m}_t,\ \mathrm{C}_t)\ \Pr(\eta_{t+1}|\kappa,\ \delta)$, which is also a normal distribution, with moments as specified in West (1995). In detail, $\Pr(\eta_t|\ \eta_{t+1},\ \log(\boldsymbol{\gamma}_1^T),\ \kappa,\ \delta) = \mathrm{N}(\eta_t; \mathrm{m}_t + B_t(\eta_{t+1} - a_{t+1}), \mathrm{C}_t - B_t R_{t+1} R_t')$, where $B_t = \mathrm{C}_t \mathrm{G} R_{t+1}^{-1}$.

The posterior full conditional for $\boldsymbol{\gamma}$ is

$$\Pr(\boldsymbol{\gamma}|\mathbf{n}_1^T, \boldsymbol{\eta}, \kappa) \quad \propto \quad \prod_{t=1}^{T} \mathrm{N}\left(\log(\gamma_t);\ \mathrm{F}'\eta_t,\ \kappa\right) \mathrm{Po}(n_t;\ \gamma_t), \tag{4.32}$$

and the entire vector will be sampled in a single Metropolis-Hastings step. The Laplace approximation to the Poisson likelihood for $\log(\gamma_t)$ is $\mathrm{Po}(n_t; \gamma_t) \approx \mathrm{N}(\log(\gamma_t);\ \log(n_t),\ 1/n_t)$, and we make use of this to build an independent proposal distribution. Given a present state of the MCMC at $\boldsymbol{\gamma}$, propose each $\tilde{\gamma}_t$ from $\mathrm{N}\left(\log(\tilde{\gamma}_t); DE, D\right)$, where $D = (n + 1/\kappa)^{-1}$ and $E = n\log(n) + \mathrm{F}'\eta_t/\kappa$. Each move $\gamma_t \to \tilde{\gamma}_t$ is then accepted with probability

$$\min\left\{1, \frac{\mathrm{Po}(n_t;\ \tilde{\gamma}_t)\mathrm{N}\left(\gamma_t;\ DE, D\right)}{\mathrm{Po}(n_t;\ \gamma_t)\mathrm{N}\left(\tilde{\gamma}_t;\ DE, D\right)}\right\}. \tag{4.33}$$

Finally, the posterior full conditional for κ is sampled directly from $\mathrm{Ga}\big(\kappa^{-1};\ a_\kappa + T/2,\ b_\kappa + \sum_{t=1}^{T}(\log(\gamma_t) - \mathrm{F}'\eta_t)^2\big)$.

4.3 Data Examples

We turn to two examples of spatial point pattern data to illustrate the DP mixture modeling framework. The first data set, consisting of information about a longleaf pine forest in southern Georgia, USA, is used to illustrate the modeling for a Poisson process with positive continuous marks. The second data set consists of crime event information for the city of Cincinnati, OH, during 2006. The crime events have been classified by type, and as such offer an example of a Poisson process with categorical marks. As well, we are able to view the monthly crime event patterns as an example of a discrete-time dynamic spatial point process. The data illustrations make reference to the model specification and posterior simulation methodology of Section 4.2, in each case assuming a transformation of observed locations to coordinates within the unit square through normalization of the observation window.

4.3.1 Longleaf Pine Forest with Tree Diameter Marks

The data record the locations and diameters of 584 Longleaf pine (*Pinus Palustris*) trees in a 200×200 meter patch of forest in Thomas County in the state of Georgia. The trees were surveyed in 1979 and the measured mark is diameter at breast height (1.5 m), or dbh. The data is available as part of the `spatstat` package for `R`, and was introduced in a study by Platt et al. (1988). A detailed description of the data may be found in this article. In addition, the data were analyzed by Rathburn and Cressie (1994) as part of a space-time survival point process. Poisson processes are generally

viewed as an inadequate model for forest patterns, due to the dependent birth process by which trees occur. However, at a single time point, the nonhomogeneous Poisson process should be flexible enough to account for the variability in tree counts throughout the observation window. The combination of a high density for juveniles (trees < 10 *cm dbh*) with a more even dispersal of larger trees leads to multimodal conditional mark densities and nonhomogeneous variability about this density.

The model of Section 4.2.3 was applied to the data and posterior simulation results were obtained following the MCMC algorithm of Section 4.2.3.1. Prior specification with respect to parameters of the location base distribution $G_0^{\mathbf{y}}$ was $a_\beta = 2$, $b_\beta = 1$, and $c_\beta = 2$ after scaling the observation window to a unit square, and $\pi(\alpha) = \text{Ga}(2, 0.2)$. Following again in the spirit of Section 2.2, the prior for the base distribution of kernel parameters related to the *dbh* marks (measured in centimeters) was such that $\pi(s_1) = \text{N}(2.9, 0.37)$, $\pi(s_2^{-1}) = \text{Ga}(2, 0.37)$, $\pi(s_3) = \text{Ga}(2, 2/0.37)$, and s_4 is fixed at 2, where 2.9 is the mean of the log *dbh* and $\sqrt{.37}$ is one sixth of the log *dbh* data range. Results are based on an MCMC sample of 10,000 parameter draws recorded on every second iteration following a burn-in period of 5000 iterations. $L = 100$ for the finite stick-breaking approximation.

Posterior samples of the DP precision α and of the number of unique allocated $\boldsymbol{\vartheta}$ components, $n^\star$, are shown in figure 4.1. Posterior realizations of the marginal location process density at any point $\mathbf{y}_0$ is available, conditional on a realization the finite stick-breaking G^L, as $f(\mathbf{y}_0; G^L) = \sum_{l=1}^{L} p_l \, \text{B}_2(\, \mathbf{y}_0;\, \boldsymbol{\mu}_l,\, \boldsymbol{\tau}_l,\, \varphi_l)$. As seen in figure 4.2, this marginal process density is dominated by a few peaks which arise due to high

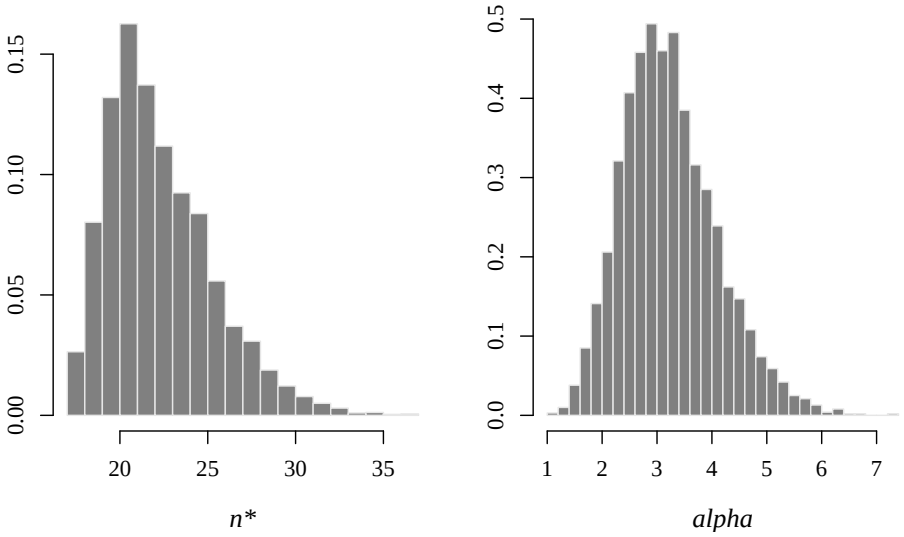


Figure 4.1: Longleaf Pine Data. Posterior samples for the number of distinct clusters and the DP prior precision.

concentrations of small juvenile trees. Despite this highly variable process intensity, the analysis is able to capture the conditional mean *dbh* surface shown on the right side of Figure 4.2, where we see that the peaks in process density correspond to valleys for expected *dbh*. This is to be expected, as smaller trees are able to survive in much higher density. Note that realisations of this mean are available analytically (up to the truncation approximation), since

$$\begin{aligned} \mathbb{E}[m|\mathbf{y}; G^L] &= \int_0^\infty m \Pr(m|\mathbf{y}; G^L) dm \\ &= \frac{\sum_{l=1}^L p_l \mathrm{B}_2(\mathbf{y}; \vartheta_l) \exp\left[q_{l1} + \frac{1}{2} q_{l2}\right]}{\sum_{l=1}^L p_l \mathrm{B}_2(\mathbf{y}; \vartheta_l)}. \end{aligned} \tag{4.34}$$

Finally, posterior sampling of the conditional mark density at any point m_0 given $\mathbf{y}_0$ is available, for a realization of G^L, as

$$f(m_0 \mid \mathbf{y}_0; G^L) = \frac{\sum_{l=1}^L p_l \mathrm{B}_2(\mathbf{y}_0; \boldsymbol{\mu}_l, \boldsymbol{\tau}_l, \varphi_l) \mathrm{N}(\log(m_0); q_{l1}, q_{l2})}{f(\mathbf{y}_0; G^L)}. \tag{4.35}$$

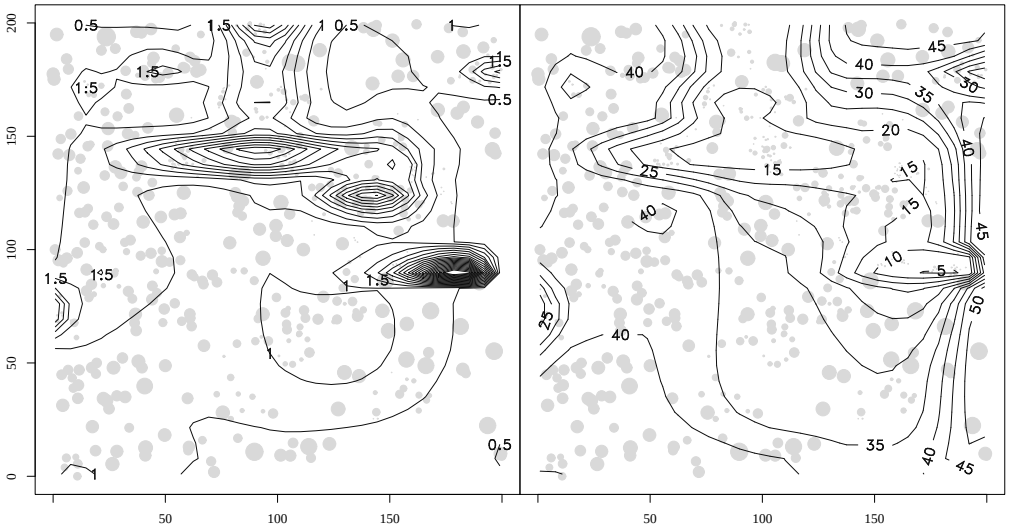


Figure 4.2: Longleaf Pine Data. Mean posterior Poisson process marginal location density $\mathbb{E}[\ f(\mathbf{y}; G) \mid \mathcal{D}]$ (left) and the mean posterior conditional mean for tree *dbh*, $\mathbb{E}\left[\mathbb{E}[m|\mathbf{y};\ G]\mathcal{D}\right]$ (right).

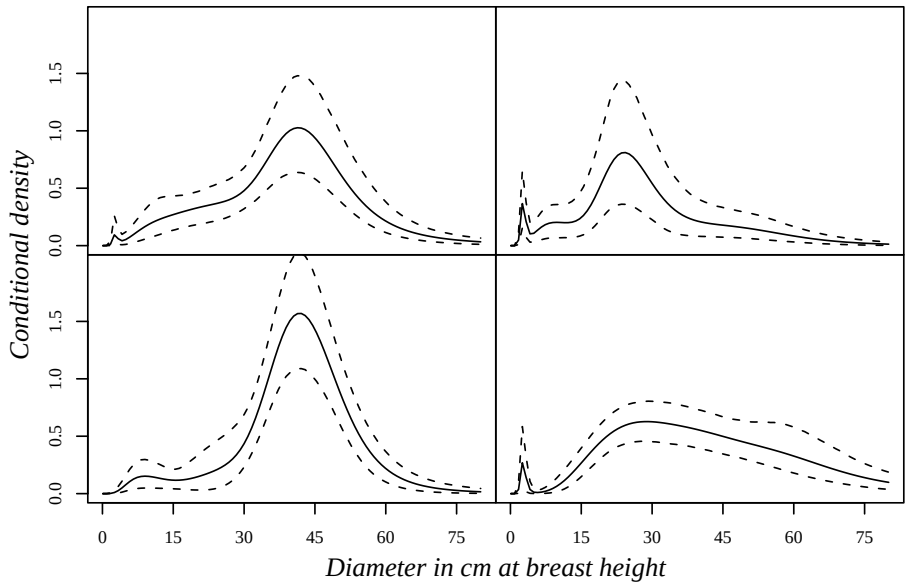


Figure 4.3: Longleaf Pine Data. Posterior mean estimates (solid lines) and 90% interval estimates (dashed lines) for four conditional mark densities $f(m \mid \mathbf{y}_0; G)$ at (clockwise from top-left) $\mathbf{y}_0 = [100, 100]$, $[150, 150]$, $[150, 50]$, and $[25, 150]$.

In Figure 4.3, the expected multimodal behavior is clearly exhibited in posterior samples of the conditional density for *dbh* at four different locations. Although the conditional densities vary in shape over the different locations, each appears to show the mixture of a relatively smooth density component for mature trees combined with a sharp peak at low *dbh* values corresponding to collections of juvenile trees. It is notable that we are able to infer this structure nonparametrically, in contrast to existing approaches where the effect of a tree-age threshold is assumed *a priori* (as in Rathburn and Cressie, 1994).

4.3.2 Crime Event Data with Categorical Classification

The city of Cincinnati maintains an online database of detailed crime statistics. For illustration of the DP mixture model methodology for a Poisson point process with categorical marks, we consider 34,651 crime events within the city during 2006. The database of arrests in Hamilton County (which contains Cincinnati) reports date, time, and location of crimes, as well as other data that might be useful to characterize the magnitude of the reported event. Note that crimes south of the Ohio River (i.e., in Kentucky) are not contained in the database. Crimes are reported by addresses, and the geocoding to convert these locations to longitude and latitude coordinates was conducted using the website `www.gpsvisualizer.com`.

Crimes have been assigned one of more than 170 different Uniform Crime Reporting codes describing a variety of events such as telephone harassment, vehicle theft, murder, and the like. This variable was used to reclassify and group the data into 3 main categories (excluding other types of crime):

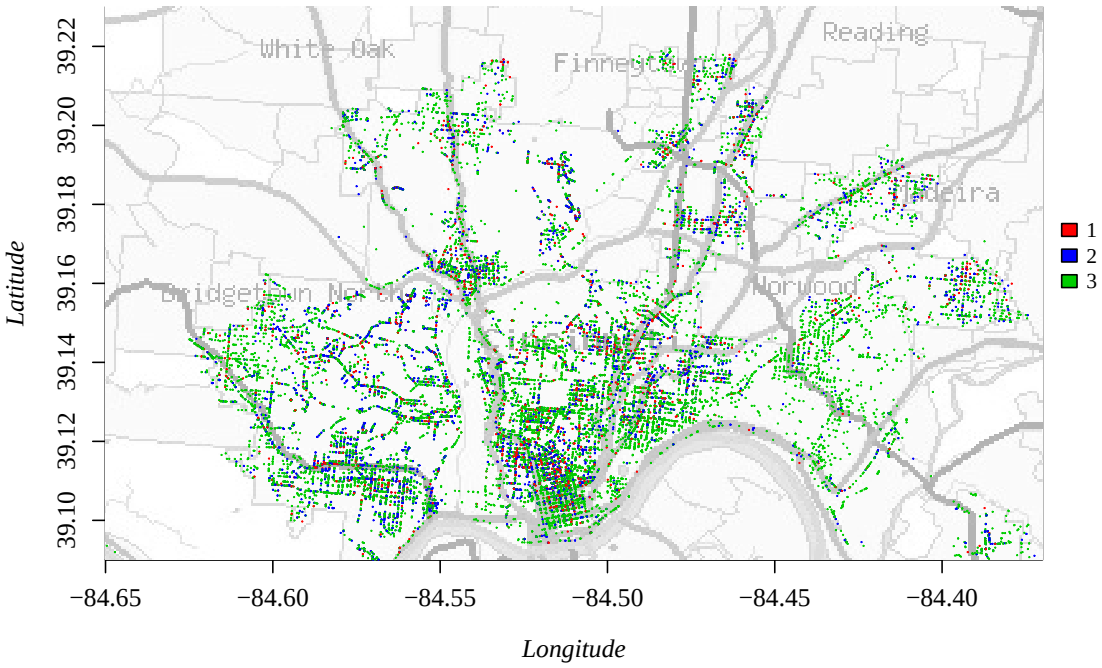


Figure 4.4: Crime Event Data with Categorical Marks. Observation window and data locations, color coded by type of crime.

1. Crimes against people with extreme violence (e.g. murder, rape)

2. Crimes against people with minor violence (e.g. assault, mugging)

3. Crimes against property (e.g. burglary, arson)

Thus the mark m associated with each crime is membership in one of these three categories.

The 34,651 data locations and the observation window are shown in Figure 4.4. For purposes of our analysis, the observation window has been defined to be the area contained within $[-84.65, -84.37]$ degrees longitude and $[39.09, 39.23]$ degrees latitude. While this includes a region south of the Ohio river, and thus not technically part of the observation window, we have assumed for illustrative purposes that the

rectangular window holds. In practice, due to the large sample size, the DP mixture model assigns posterior intensity of very nearly zero to areas south of the river, and with the conditional mark density defined as simply zero wherever the joint location-mark density is zero, the results are unaffected by this simplification. A more complete analysis of this data, however, may benefit from use of a bounded kernel. In particular, due to other discontinuities inherent in an urban landscape, crime event pattern analysis could be an ideal application for the bivariate uniform kernel of (4.6).

The model of Section 4.2.2 was applied to the data and posterior simulation results were obtained following the procedure of Section 4.2.2.1. Prior specification for parameters of the location base distribution $G_0^{\mathbf{y}}$ was $a_\beta = 2$, $b_\beta = 0.2$, and $c_\beta = 2$ after scaling the observation window to a unit square, and $\pi(\alpha) = \text{Ga}(2, 0.2)$. The Dirichlet base distribution for kernel mark probabilities Q was parameterized by $\mathbf{a}^q = [1, 2, 4]$, representing expected relative frequency for the three crime classes. Results are based on an MCMC sample of 8000 parameter draws recorded on every second iteration following a burn-in period of 2000 iterations. $L = 300$ for the finite stick-breaking approximation, and inference occurred over a 30×30 grid of locations.

Posterior samples of the DP precision α and of the number of unique allocated $\boldsymbol{\vartheta}$ components, $n^\star$, are shown in Figure 4.5. It is notable that only 60 to 70 of the mixture components were ever allocated to observations, despite the huge number of data points, indicating a considerable efficiency in modeling. Posterior realizations of the marginal location process density at any point $\mathbf{y}_0$ is again available, calculated as in Section 4.3.1, and the posterior mean is shown in Figure 4.6. This marginal intensity estimate is able

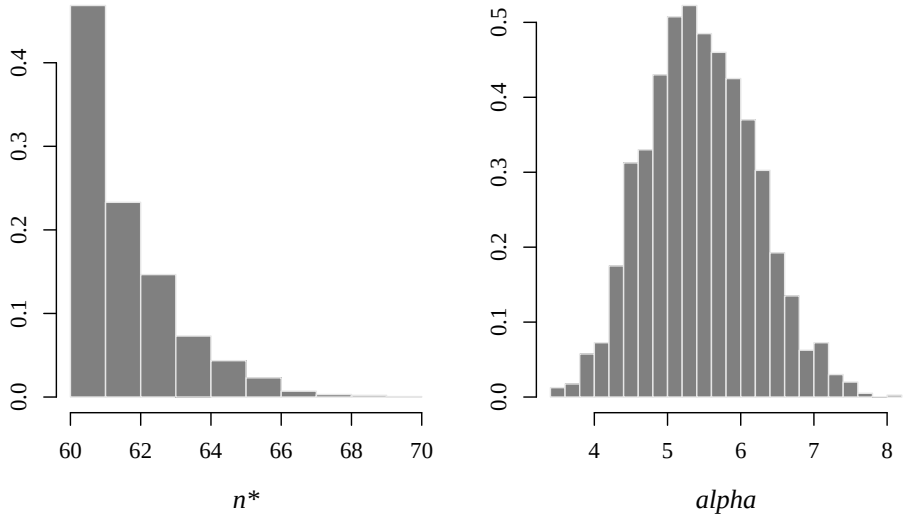


Figure 4.5: Crime Event Data with Categorical Marks. Posterior samples for the number of distinct clusters and the DP prior precision.

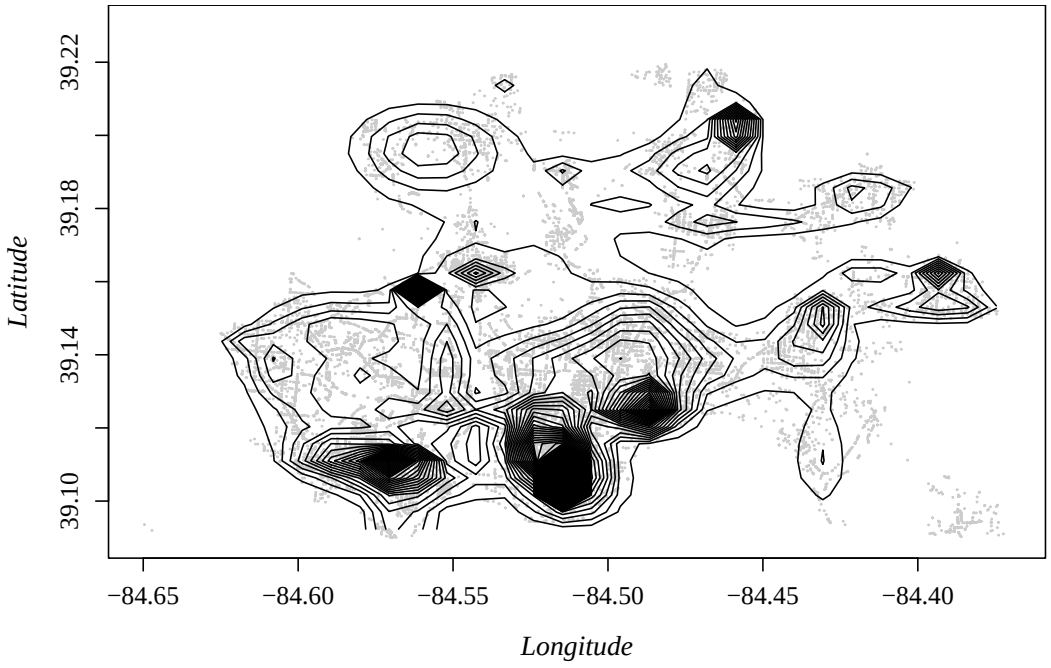


Figure 4.6: Crime Event Data with Categorical Marks. Mean posterior Poisson process marginal location density $\mathbb{E}[\, f(\mathbf{y}; G) \mid \mathcal{D}]$. The data locations are plotted in grey.

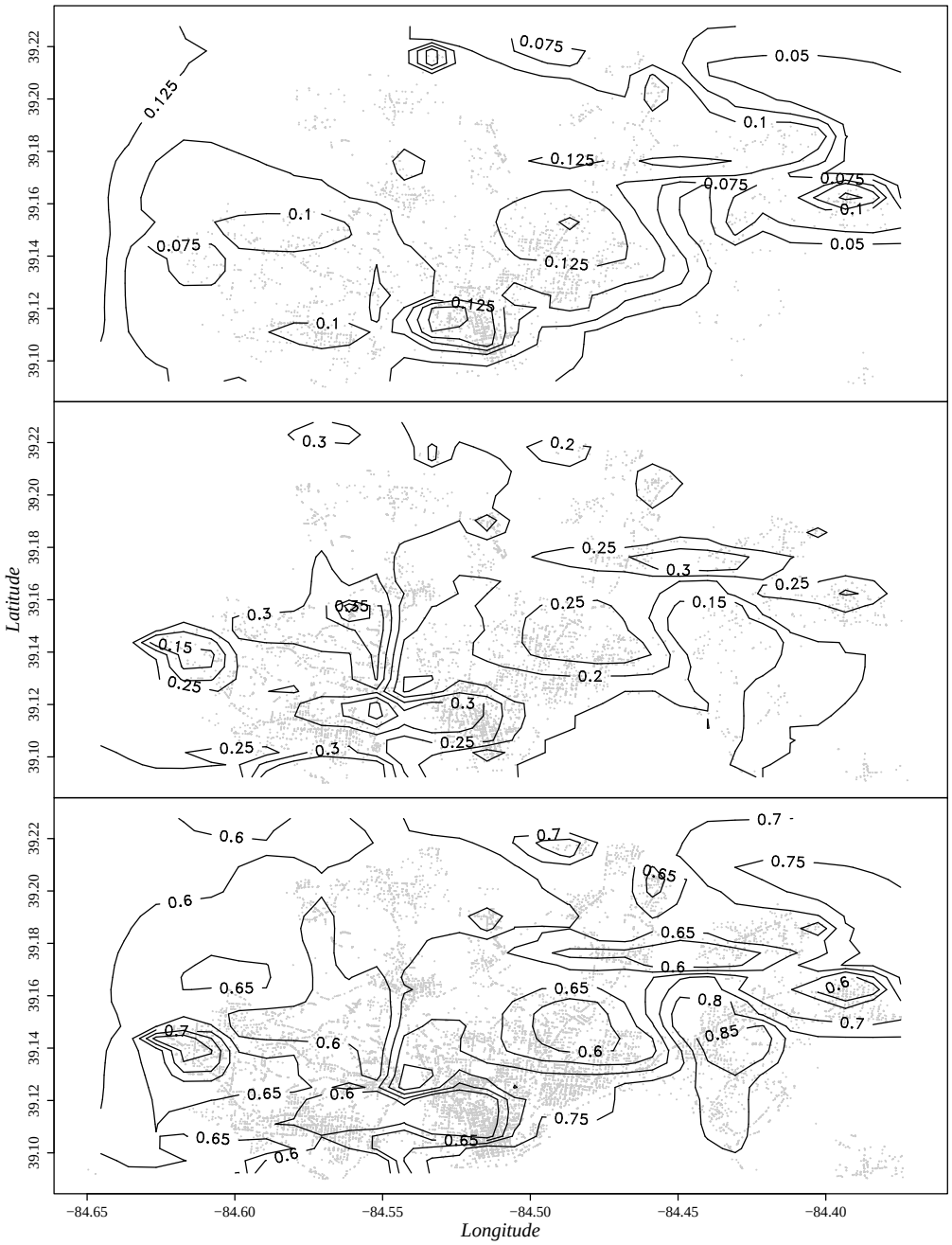


Figure 4.7: Crime Event Data with Categorical Marks. Mean posterior conditional probability for each crime category (for $m = 1$ to 3, from top to bottom), $\mathbb{E}\left[\Pr[m|\mathbf{y}; G]\mathcal{D}\right]$. Observations corresponding to each category are plotted in grey.

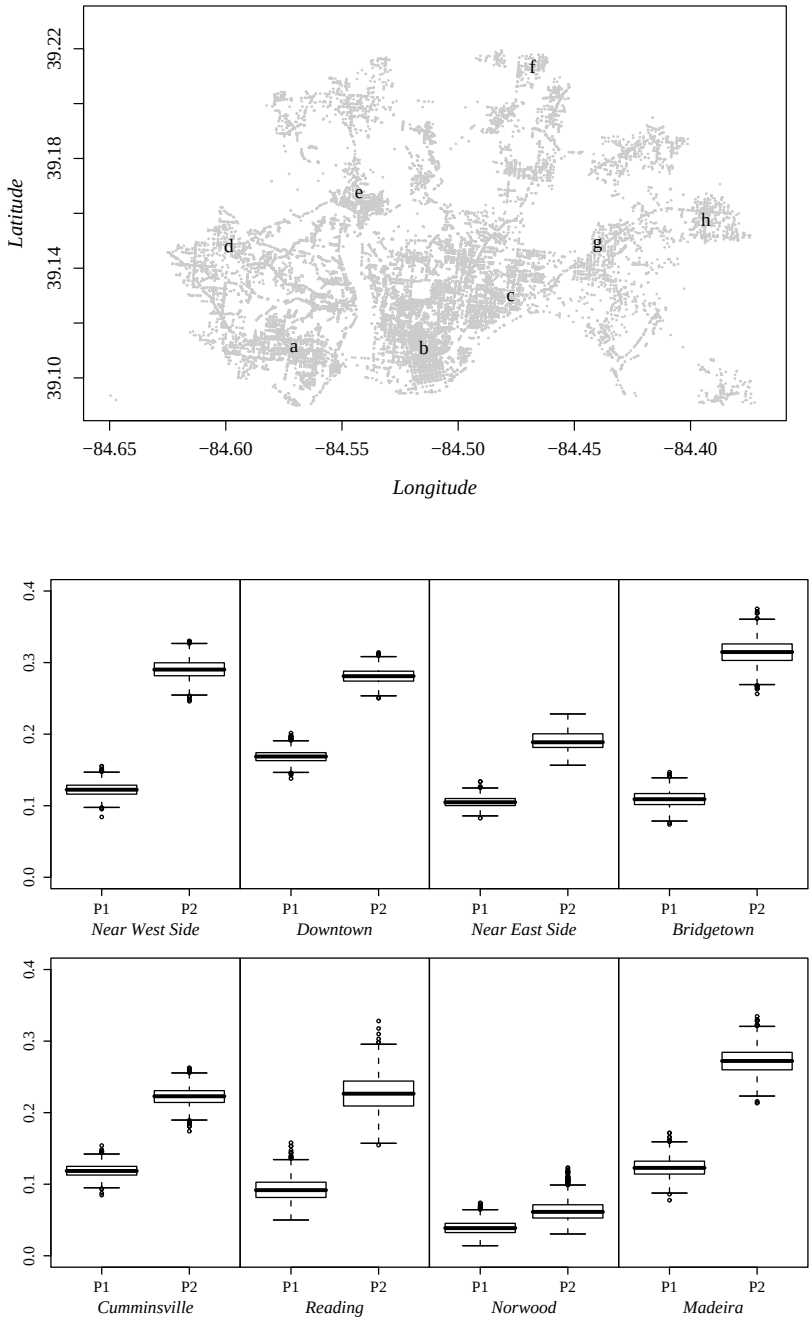


Figure 4.8: Crime Event Data with Categorical Marks. Posterior samples of the conditional probability for crime against persons with or without extreme violence at eight locations within Hamilton County. The top panel shows the exact locations corresponding to each conditional probability function, labeled (a) to (h) reading from left to right, top to bottom.

to capture the broad neighborhood structure and account for geographical boundaries within the observation window, despite a lack of any prior information related to these features. Posterior sampling for the location-dependent probability for each crime class m_0 is calculated, given G^L, as

$$f(m_0 \mid \mathbf{y}_0; G^L) \quad = \quad \frac{\sum_{l=1}^{L} p_l \mathrm{B}_2(\mathbf{y}_0; \boldsymbol{\mu}_l, \boldsymbol{\tau}_l, \varphi_l) q_{m_0}}{f(\mathbf{y}_0; G^L)}. \tag{4.36}$$

Posterior mean category probabilities corresponding to $m_0 = 1$, 2, and 3 are mapped in Figure 4.7, and posterior samples for $\Pr(m = 1|\mathbf{y}_0)$ and $\Pr(m = 2|\mathbf{y}_0)$ at specific $\mathbf{y}_0$ locations are in Figure 4.8. The mean conditional probability surface for $m = 1$ is considerably more uniform than those for $m = 2$ and $m = 3$, as would be expected due to the general rarity of extremely violent crime. It also appears that the probability of a crime event involving violence at all (i.e., $m = 1$ or $m = 2$) is higher in the more central neighborhoods. In detail, we see in the boxplots of Figure 4.8 that crimes in both the near West side and downtown are more likely to involve any level of violence than those committed elsewhere in the city, while crimes committed in Madeira are most likely property crimes. Finally, the probability of a crime involving extreme violence is significantly higher in the downtown area than anywhere else monitored. In fact, the location corresponding to downtown in this figure is located just north of the central business district, in the Over-the-Rhine neighborhood, which was the site of violent rioting in 2001. The riots began in reaction to specific incidents of police brutality, and the aftermath of this event led to the public dissemination of the crime statistics analyzed here.

4.3.3 Monthly Violent Crime Event Data

To illustrate the single-p DDP model for discrete time Poisson processes, we again consider the Cincinnati crime event data, but now restrict ourselves to 3857 events of extremely violent crime against persons (crime category 1). The monthly data is plotted in Figure 4.9.

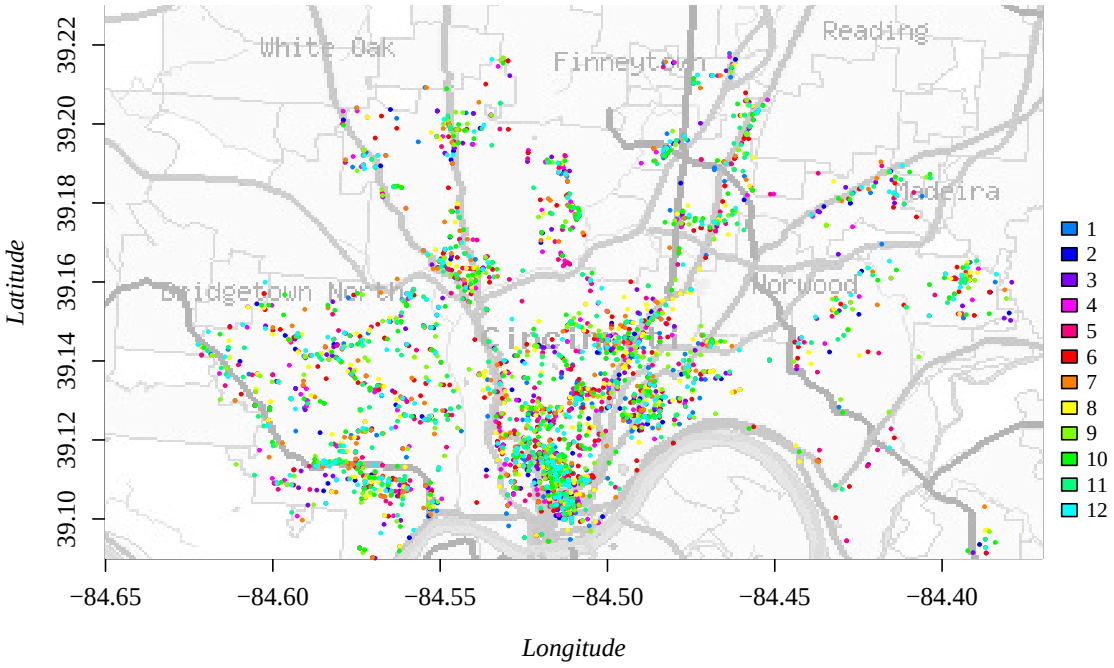


Figure 4.9: Monthly Crime Event Data. Observation window and data locations, color coded by month.

The model of Section 4.2.4 was applied to the data for $\mathcal{T} = \{1, \ldots, 12\}$, and posterior simulation results were obtained following the procedure of Section 4.2.4.1. Prior specification for parameters of the location base distribution $G_0^{\mathbf{y}}$ was the same as for Section 4.3.2, and the hyperprior for parameters underlying the PBAR prior was $\pi(\alpha) = \text{Ga}(2, 0.2)$ and $\pi(\rho) = \text{U}(0, 1)$. The DLM model for log monthly intensity is

specified with

$$\mathrm{F} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad \mathrm{G} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cos(\frac{2\pi}{4}) & \sin(\frac{2\pi}{4}) & 0 & 0 \\ 0 & 0 & -\sin(\frac{2\pi}{4}) & \cos(\frac{2\pi}{4}) & 0 & 0 \\ 0 & 0 & 0 & 0 & \cos(\frac{2\pi}{12}) & \sin(\frac{2\pi}{12}) \\ 0 & 0 & 0 & 0 & -\sin(\frac{2\pi}{12}) & \cos(\frac{2\pi}{12}) \end{bmatrix}. \quad (4.37)$$

In the block diagonal $\mathrm{G} = \mathrm{Diag}[\mathbf{J}_2(1), \mathbf{J}_2(1, 2\pi/4), \mathbf{J}_2(1, 2\pi/12)]$, the first block corresponds to a linear trend and the other blocks to seasonal components consisting of persistent harmonic oscillations of period 4 and 12 respectively (triannual and annual trends). The variance components are specified through a $\mathrm{Ga}(1, 1)$ prior for κ and use of a discount factor $\delta = 0.99$. The prior state parameters are $\mathrm{m}_0 = [5.8,0,0,0,0,0]'$ and $\mathrm{C}_0 = \mathrm{Diag}[1,2,0.5,0.5,0.5,0.5]$.

Results are again based on an MCMC sample of 8000 parameter draws recorded on every second iteration following a burn-in period of 2000 iterations. $L = 300$ for the finite stick-breaking approximation, and inference occurred over a 30×30 grid of locations for each of the 12 months in 2006. The monthly posterior mean process densities are shown in Figure 4.11, and we see that the crime density has become more diffuse throughout the year. Anecdotally, this may be a result of warming throughout the spring combined with a campaign of extra police officers dispatched to the worst neighborhoods which began in May. As the neighborhood structure remains static over the entire year, but the relative magnitude of crime intensity is clearly dynamic, this

presents an efficient application of the single-θ DDP model. We note that these mean process density surfaces all appear to be more difffuse that the marginal process density (for all crime classes) of Figure 4.6; this is consistent with the general rarity of extremely violent crime. The DLM fit for log integrated intensity is shown in Figure 4.10, showing a negative linear trend along with triannual and annual seasonal effects (however, since there are only 12 observations, the posterior is quite diffuse). This figure also contains an intensity forecast for 2007.

The PBAR process underlying the single-p DDP monthly correlation is exhibited in the posterior sample of the first three weights of the related stick-breaking measures, $\mathbf{p}_1$, $\mathbf{p}_2$, and $\mathbf{p}_3$. Figure 4.13 shows posterior quartiles for these three vectors as well as a thinned sample of realized monthly paths for each weight. Figure 4.12 contains histograms for the relevant parameters. Note that the posterior mean for α is 2.7 and the posterior mean ρ is 0.76, for an expected temporal autocorrelation between stick-breaking proportions of about 0.7. This indicates a temporal dependence between monthly crime event densities, as would be expected. However, it is also clear that the densities evolve qualitatively over time, and as such it would be insufficient to model this data with a dynamic integrated intensity but a static process density.

Within four subregions of the observation window, shown in the top panel of Figure 4.14, the intensity surface was numerically integrated at each MCMC iteration to obtain an integrated monthly intensity (i.e., for each t the product of γ_t and $\int_{\mathcal{B}} f(\mathbf{y}; G_t^L)$, with $\mathcal{B}$ the region of interest). The posterior sample of these intensities is shown in the bottom panel of Figure 4.14. In addition, the posterior mean intensity path for

each subregion was compared to the monthly crime counts for the respective area, and the Anscombe residuals (see McCullagh and Nelder, 1989, for a discussion with respect to Poisson GLM) were calculated as $3\left(n_t^{2/3} - \Lambda_t(\mathcal{B})^{2/3}\right) \,/\, \left(2\Lambda_t(\mathcal{B})^{1/6}\right)$. We see in the table of Figure 4.14 that most of the residuals fall within $[-2, 2]$, indicating a reasonable model fit for these subregions of the observation window. This is a considerable accomplishment, as these estimates are derived from a complex nonparametric estimation procedure over the larger space without any explicit modeling for the individual subregions.

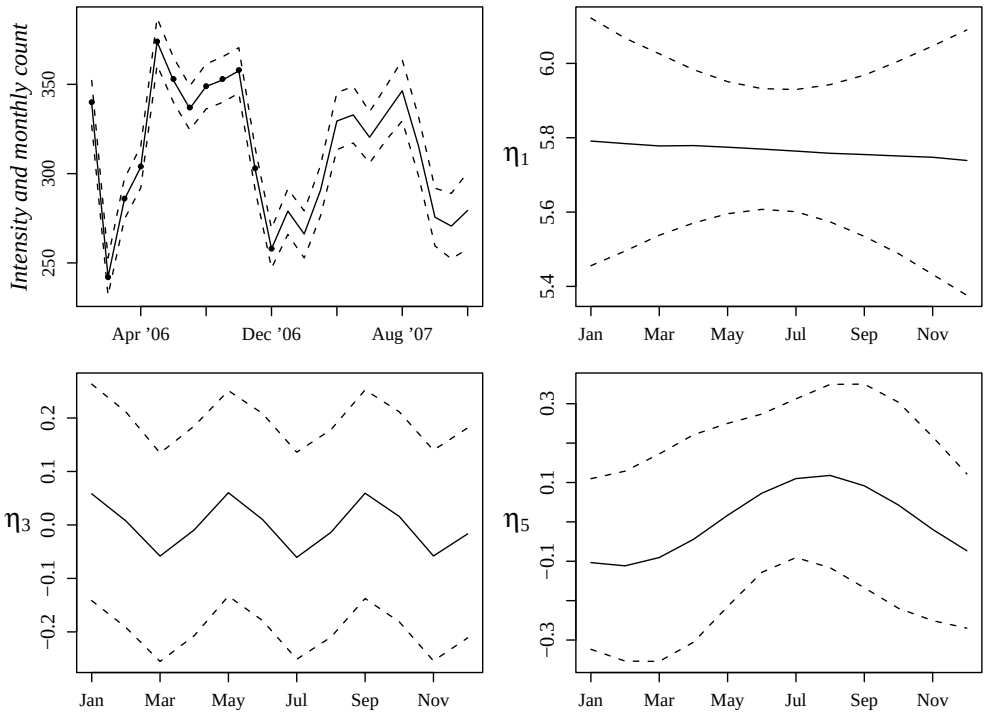


Figure 4.10: Monthly Crime Event Data. The top left plot shows posterior quartiles for monthly total intensity $\gamma_1, \ldots, \gamma_{12}$ and forecast mean intensity for 2007. Actual crime event counts for 2006 are plotted in the background. The remaining plots show posterior quartiles for the DLM state components. Clockwise from top right, we show the linear trend, the annual trend, and the triannual trend.

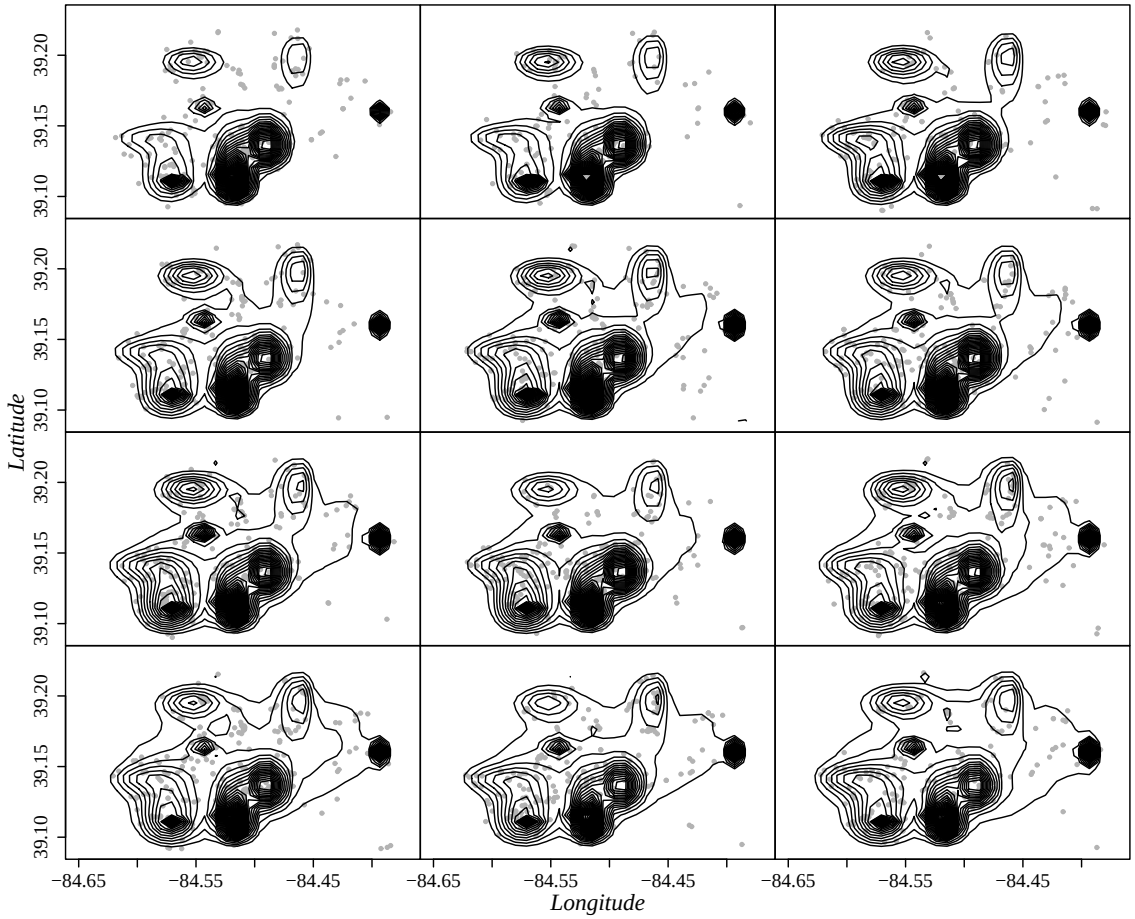


Figure 4.11: Monthly Crime Event Data. Mean posterior Poisson process marginal location density $\mathbb{E}[\ f(\mathbf{y}; G_t) \mid \mathrm{D}]$ for $t = 1,\ldots,12$, reading from top to bottom and from left to right.

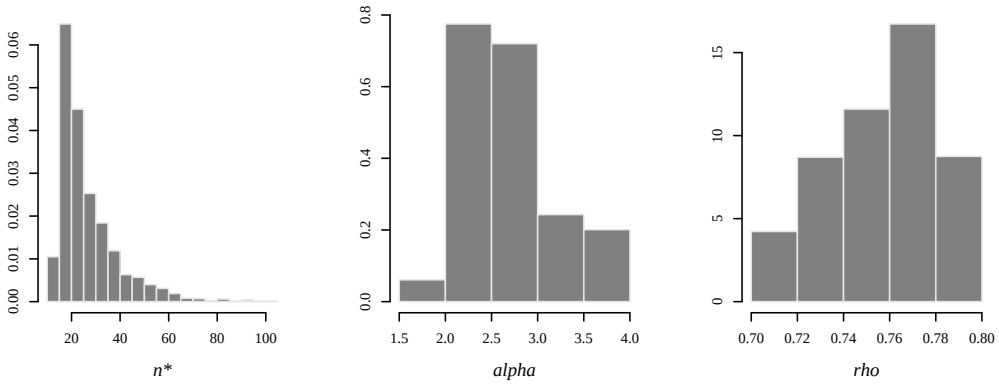


Figure 4.12: Monthly Crime Event Data. Posterior samples for the number of distinct clusters, the DP prior precision, and the PBAR correlation parameter ρ.

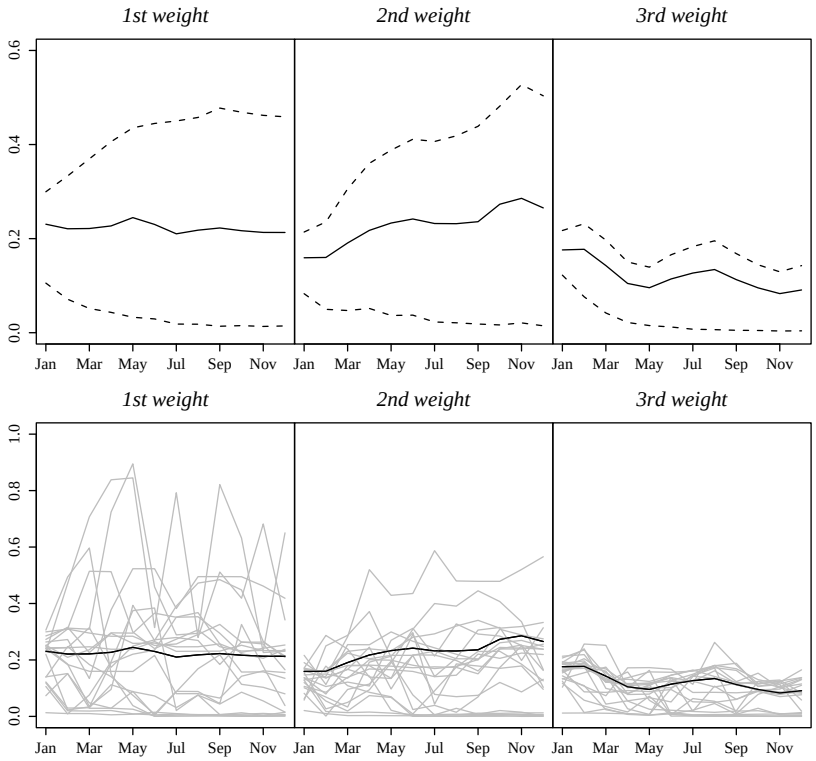
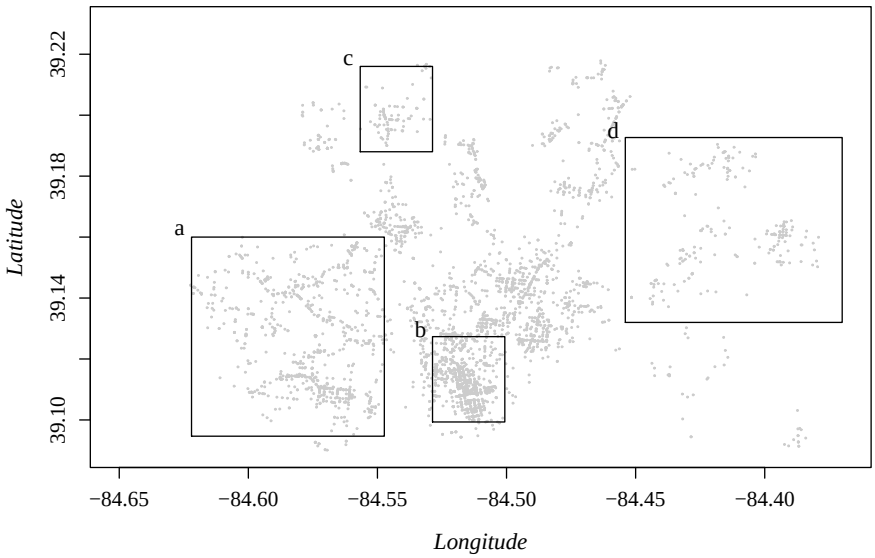
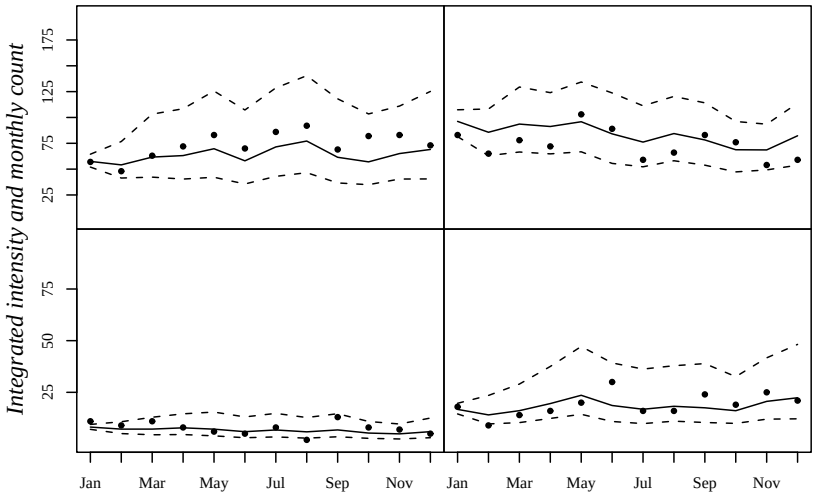


Figure 4.13: Monthly Crime Event Data. The top panel contains posterior median (solid line) and quartile (dashed lines) monthly values for the first three stick-breaking weights, and the bottom panel shows a thinned sample of these same weights along with the posterior mean in black.





	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
a	-0.06	-0.85	0.17	1.095	1.54	1.52	1.66	1.64	0.94	3.10	2.13	0.47
b	-1.38	-2.33	-1.66	-2.09	0.72	0.53	-2.05	-2.09	0.53	0.86	-1.83	-2.70
c	0.93	0.65	1.31	0.06	-0.45	-0.40	0.45	-1.86	2.11	1.10	0.89	-0.41
d	0.30	-1.45	-0.542	-0.83	-0.76	2.42	-0.20	-0.54	1.47	0.70	0.93	-0.29

Figure 4.14: Monthly Crime Event Data, integrated intensity. The top panel shows the four subregions of study, the middle panel contains Posterior median (solid line) and quartile (dashed lines) monthly integrated intensity plotted against actual crime event counts for each subregion, and the bottom panelshows monthly Anscombe residuals with respect to posterior mean integrated intensity.

Chapter 5

Conclusion

We conclude with a discussion of some implications and possible extensions of the presented work. First, each of the general regression frameworks of Chapters 2 and 3 may be extended through the use of dependent DP priors in modeling for correlated random measures. As in the discussion of Section 4.1.3, use of DDP priors for the random mixing measure will not require fundamental changes to either kernel specification or the inference framework. Indeed, the Appendix contains a general outline of posterior sampling methodology for single-θ DDP mixtures, and this prior model may be appropriate for new settings other than modeling for Poisson point processes. For example, we have considered an extension of the regression scheme of Chapter 2 to the setting of geographically related, but distinct, populations of covariates and response. In this case, a conditional autoregressive dependent DP prior may be used to induce correlation between nonparametric regression estimators for adjacent populations.

In the particular setting of regression, it may be useful to have only a subset of

the covariates modeled as part of a joint density kernel, while making the random mixing measure dependent upon the remaining covariates. A regression model for a combination of both continuous and discrete covariates arises by maintaining the multivariate normal kernel structure of Chapter 2, but assigning a dependent DP prior on the collection of mixing distributions corresponding to the different levels of the categorical covariates. For instance, with a single binary covariate, the data vector can be decomposed into two groups, $\{\mathbf{z}_{ij} = (y_{ij}, \mathbf{x}_{ij}) : i = 1, ..., n_j\}$, $j = 1, 2$, associated with the two levels of the categorical covariate. Then, for $j = 1, 2$, the $\mathbf{z}_{ij}$ are assumed to arise from the DP mixture in (2.2) given group-specific mixing distributions G_j. The model is completed with a dependent DP prior for (G_1, G_2), say, in the spirit of Tomlinson and Escobar (1999) and de Iorio et al. (2004), or Gelfand and Kottas (2001) if stochastic order restrictions for the categorical covariate levels are plausible (as may be the case in treatment-control settings for the survival regression models of Section 3.2). This approach is also an option when the marginal distribution for a subset of the covariate vector is known *a priori* (e.g., in a designed experiment).

Also within the context of regression, it will be possible to develop inference frameworks for conditional functionals other than quantiles or means. In particular, sensitivity analysis is performed to resolve the sources of response variability by apportioning elements of this variation to different sets of covariates (refer to Saltelli et al., 2000, for an introduction and references). Such global sensitivity analysis is based on inference for sensitivity indices which measure variability of the response with respect to uncertainty in different subsets of the covariates (see Taddy et al., 2007, for a Bayesian

example based on Gaussian process regression). This presents a natural application for our approach to implied conditional regression, as we are already modeling for covariate uncertainty in sampling of the joint covariate-response density function.

The general approach in Section 3.2 of having informative parametric modeling linked with nonparametric models through an underlying hidden stochastic process is both theoretically appealing and practically powerful. We believe that there is great potential for such models, since they provide an efficient way to bridge the difference in scale between two observed processes, and the MCMC algorithms presented here can be the basis for extended techniques in other settings. Also with reference to Section 3.2, the methodology presented therein is applicable in more general settings involving hidden Markov model structure. In particular, since the switching occurs at the level of the joint distribution for response and covariates, the algorithms are directly applicable to nonparametric density estimation through DP mixtures of multivariate normal distributions for heterogeneous populations where switching between subpopulations occurs as a Markov chain.

Unlike the Pólya urn marginalization, posterior simulation algorithms built around a finite truncation of the DP do not rely upon a $\text{Be}(1, \alpha)$ prior for stick-breaking proportions. Thus, much of the model development and posterior simulation methodology may be extended to alternative stick breaking priors. In particular, corresponding frameworks based on the general beta two-parameter process of Ishwaran and Zarepour (2000) present an obvious continuation of our work. In a similar spirit, the sequential Monte Carlo methodology for single-θ DDP priors presented in Section 4.2.4 will apply

to multivariate stick-breaking proportion priors other than the PBAR induced density, and many more elaborate constructions for time series of random measures are possible.

This thesis provides a suite of flexible and practical nonparametric Bayesian analysis frameworks, together related under a particular approach to DP mixture modeling based on joint density estimation with carefully chosen kernels and inference through finite stick-breaking approximation. As it is applicable for DDP mixtures of generic kernels and requires specification of only a single univariate stochastic process, the single-θ DDP prior and posterior simulation development is very much in keeping with this approach. It was a stated goal in the introduction that the methodology contained herein would find usage in a wide variety of data analysis applications, and I believe that the preference towards relative simplicity in model specification and generality in posterior simulation methodology (combined with forthcoming software in publicly available packages for `R`) will do much to achieve this goal.

Appendix

Posterior Simulation for Generic Finite Stick-Breaking Models

While Pólya urn based posterior simulation techniques for DP mixture models have been detailed extensively elsewhere (e.g., Neal, 2000), sampling algorithms for the finite stick-breaking models presented throughout this thesis may be less familiar. Thus, we now present a quick outline of a general approach to simulation for this class of models.

Consider data $\mathcal{D} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, where $\mathbf{z}_i = [z_{i,1}, \ldots, z_{i,d}]$, accompanied by an index vector $\mathbf{t} = \{t_1, \ldots, t_n\}$, where $t_i \in \{1, \ldots, T\}$ indicates the state (e.g., time) corresponding to observation $\mathbf{z}_i$. A generic dynamic finite stick-breaking mixture model

holds that, for $i = 1, \ldots, n$, $\mathbf{z}_i \sim \mathrm{k}(\mathbf{z}_i; \Theta_{k_i})$ and

$$\begin{aligned} k_i|\mathbf{P} &\sim \sum_{l=1}^{L} p_{l,t_i}\delta_{[l]}(k_i) \\ \mathbf{P}, \boldsymbol{\Theta}|\boldsymbol{\alpha}, \boldsymbol{\psi} &\sim \mathcal{P}_L\left(\mathbf{P}; S(\mathbf{v}; \boldsymbol{\alpha})\right) \prod_{l=1}^{L} dG_0(\Theta_l; \boldsymbol{\psi}), \end{aligned}$$

where $\boldsymbol{\Theta} = \{\Theta_1, \ldots, \Theta_L\}$, S is a T-dimensional distribution induced by a stochastic process with realizations in $(0, 1)$, and the multivariate stick-breaking prior $\mathcal{P}_L\left(\mathbf{P}; S(\mathbf{v}; \boldsymbol{\alpha})\right)$ for $\mathbf{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_L\}$, is defined constructively such that $\mathbf{v}_1, \ldots, \mathbf{v}_{L-1} \overset{iid}{\sim} S(\mathbf{v}; \boldsymbol{\alpha})$, $\mathbf{v}_L = \mathbf{1}$, $\mathbf{p}_1 = \mathbf{v}_1$, and for $l = 2, \ldots, L$; $t = 1, \ldots, T$: $p_{l,t} = \prod_{j=1}^{l-1}(1 - v_{j,t})v_{l,t}$.

This model represents a finite stick-breaking truncation of the single-θ DDP as defined in Section 4.1.3. A generic static stick-breaking model is defined in the special case of $T = 1$. In full generality, assume that $\Theta_l = \{\theta_{l,1}, \ldots, \theta_{l,B}\}$ for $B \leq d$ such that $\mathrm{k}(\mathbf{z}; \Theta) = \prod_{b=1}^{B} \mathrm{k}_b(\mathbf{z}^b; \theta_b)$ and $G_0(\Theta; \boldsymbol{\psi}) = \prod_{b=1}^{B} G_0^b(\theta_b; \psi_b)$, where $\mathbf{z} = [\mathbf{z}^1, \ldots, \mathbf{z}^B]$. Thus the B kernel components (corresponding to subsets of $\mathbf{z}$; e.g., continuous and discrete variables) are conditionally independent given $\mathbf{k}$, and posterior sampling methodology will exploit this fact whenever possible. The hyperprior is $\pi(\boldsymbol{\alpha}, \boldsymbol{\psi}) = \pi(\boldsymbol{\alpha}) \prod_{b=1}^{B} \pi(\psi_b)$.

Introduce the $L \times T$ indicator matrix $\mathbf{h}$, where $h_{l,t} = \sum_{i=1}^{n} \delta_{[k_i=l, t_i=t]}$, and the related vector $\mathbf{H} = \{H_1, \ldots, H_L\}$, where $H_l = \sum_{i=1}^{n} \delta_{[k_i=l]} = \sum_{t=1}^{T} h_{l,t}$. Hence, the location parameters may be partitioned into allocated $\boldsymbol{\Theta}^\star = \{\Theta_1^\star, \ldots, \Theta_{n^\star}^\star\} = \{\Theta_l : H_l > 0\}$ and unallocated $\boldsymbol{\Theta}^u = \{\Theta_1^u, \ldots, \Theta_{n^u}^u\} = \{\Theta_l : H_l = 0\}$. Additionally, this allows us to define $L^\star \in \{1, \ldots, L-1\}$ as the lowest index value such that $\sum_{l=L^\star+1}^{L} H_l = 0$, with $L^\star = L-1$ if this is impossible. Finally, it is convenient to specify $\mathbf{k}^\star = \{k_1^\star, \ldots, k_n^\star\}$ such that $\mathbf{z}_i \sim \mathrm{k}(\mathbf{z}_i; \Theta_{k_i^\star}^\star)$.

Gibbs sampling from posterior full conditional distributions then proceeds as follows:

- For $j = 1, \ldots, n^u$, simulate independently each $\Theta_j^u \sim G_0(\boldsymbol{\psi})$.
- For $j = 1, \ldots, n^\star$ and for $b = 1, \ldots, B$, simulate independently each $\theta_{j,b}^\star$ from the distribution with density proportional to $dG_0^b(\theta_{j,b}^\star; \psi_b) \prod_{\{i:k_i^\star = j\}} \mathrm{k}_B(\mathbf{z}_i^b; \theta_{j,b}^\star)$.
- For $b = 1, \ldots, B$, sample ψ_b with density proportional to $\pi(\psi_b) \prod_{j=1}^{n^\star} dG_0(\theta_{j,b}^\star; \psi_b)$.
- For $i = 1, \ldots, n$, each k_i is independently sampled with probability function proportional to $\sum_{1=1}^{L} p_{l,t_i} \mathrm{k}(\mathbf{z}_i; \Theta_l) \delta_{[l]}(k_i)$.
- For $l = 1, \ldots, L^\star$, each $\mathbf{v}_l$ is sampled independently from the density proportional to $S(\mathbf{v}_l; \boldsymbol{\alpha}) \prod_{t=1}^{T} \mathrm{Bin}\left(h_{l,t}; \sum_{j=l}^{L^\star} h_{j,t}, v_{l,t}\right)$. If $T = 1$ and $S(\boldsymbol{\alpha}) = \mathrm{Be}(1, \alpha)$ (as for a truncated DP prior), this simplifies to a $\mathrm{Be}(1 + H_l, \alpha + \sum_{j=l+1}^{L} H_j)$ distribution.
- For l such that $L^\star < l < L$, draw independently each $\mathbf{v}_l$ from the prior $S(\mathbf{v}_l; \boldsymbol{\alpha})$. $\mathbf{v}_L$ is a vector of ones.
- Sample $\boldsymbol{\alpha}$ with density proportional to $\pi(\boldsymbol{\alpha}) \prod_{l=1}^{L^\star} S(\mathbf{v}_l; \boldsymbol{\alpha})$. In the static case with $S(\boldsymbol{\alpha}) = \mathrm{Be}(1, \alpha)$ and $\pi(\alpha) = \mathrm{Ga}(a, b)$, the posterior full conditional for α is $\mathrm{Ga}\Big(a + L^\star, b + \prod_{l=1}^{L^\star}(1 - v_l)\Big)$.

Detailed descriptions of sampling algorithms based upon this general structure are contained throughout this thesis, immediately following the relevant model specification.

Bibliography

Amemiya, T. (1984), "Tobit models: A survey," *Journal of Econometrics*, 24, 3–61.

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

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002), "The infinite hidden Markov model," in *Advances in Neural Information Processing Systems*, eds. T. Dietterich, S. Becker, and Z. Ghahramani, vol. 14, MIT Press.

Berliner, L. M. and Lu, Z.-Q. (1999), "Markov switching time series models with application to a daily runoff series," *Water Resources Research*, 35, 523–534.

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.

Billio, M., Monfort, A., and Robert, C. P. (1999), "Bayesian estimation of switching ARMA models," *Journal of Econometrics*, 93, 229–255.

Bishop, C. M. and Lasserre, J. (2007), "Generative or discriminative? Getting the best

of both worlds," in *Bayesian Statistics 8*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, pp. 3–24, Oxford University Press.

Blackwell, D. and MacQueen, J. (1973), "Ferguson distributions via Pólya urn schemes," *The Annals of Statistics*, 1, 353–355.

Bravington, M. V., Stokes, T. K., and O'Brien, C. M. (2000), "Sustainable fishing: the bottom line," *Marine and Freshwater Research*, 51, 465–475.

Brix, A. (1999), "Generalized gamma measures and shot-noise cox processes," *Advances in Applied Probability*, 31, 929–953.

Brix, A. and Diggle, P. J. (2001), "Spatiotemporal prediction for log-gaussian cox processes," *Journal of the Royal Statistical Society, Series B*, 63, 823–841.

Brix, A. and Møller, J. (2001), "Space-time multi type log Gaussian Cox processes with a view to modeling weeds," *Scandinavian Journal of Statistics*, 28, 471–488.

Brunner, L. J. and Lo, A. Y. (1989), "Bayes methods for a symmetric uni-modal density and its mode," *The Annals of Statistics*, 17, 1550–1566.

Buchinsky, M. and Hahn, J. (1998), "An alternative estimator for the censored quantile regression model," *Econometrica*, 66, 653–671.

Bush, C. and MacEachern, S. (1996), "A semiparametric Bayesian model for randomized block designs," *Biometrika*, 83, 275–285.

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.

Carter, C. K. and Kohn, R. (1994), "On Gibbs sampling for state space models," *Biometrika*, 81, 541–543.

Chamberlain, G. and Imbens, G. W. (2003), "Nonparametric applications of Bayesian inference," *Journal of Business and Economic Statistics*, 21, 12–18.

Chib, S. (1992), "Bayes ineference in the Tobit censored regression model," *Journal of Econometrics*, 51, 79–99.

Chib, S. (1996), "Calculating posterior distributions and modal estimates in Markov mixture models," *Journal of Econometrics*, 75, 79–97.

Connor, R. and Mosimann, J. (1969), "Concepts of independence for proportions with a generalization of the Dirichlet distribution," *Journal of the American Statistical Association*, 64, 194–206.

Cressie, N. A. C. (1993), *Statistics for Spatial Data*, Wiley, revised edn.

Daley, D. J. and Vere-Jones, D. (2003), *An Introduction to the Theory of Point Processes*, Springer-Verlag, 2nd edn.

Damien, P., Laud, P. W., and Smith, A. M. (1996), "Implementation of Bayesian non-parametric inference based on beta processes," *Scandinavian Journal of Statistics*, 23, 27–36.

De Blasi, P. and Hjort, N. L. (2007), "Bayesian survival analysis in proportional hazard models with logistic relative risk," *Scandinavian Journal of Statistics*, 34, 229–257.

de Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004), "An ANOVA model for dependent random measures," *Journal of the American Statistical Association*, 99, 205–215.

Dellaportas, P. and Papageorgiou, I. (2006), "Multivariate mixtures of normals with unknown number of components," *Statistics and Computing*, 16, 57–68.

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

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

Doucet, A., de Freitas, N., and Gordon, N. (eds.) (2001), *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science, Springer-Verlag.

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007), "Generalized spatial Dirichlet process models," *Biometrika*, 94, 809–825.

Dunson, D. B. and Park, J.-H. (2008), "Kernel stick-breaking processes," *Biometrika*, In press.

Dunson, D. B. and Taylor, J. A. (2005), "Approximate Bayesian inference for quantiles," *Journal of Nonparametric Statistics*, 17, 385–400.

Escobar, M. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Ferguson, T. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1, 209–230.

Ferguson, T. S. (1974), "Prior distributions on spaces of probability measures," *The Annals of Statistics*, 2, 209–230.

Ferguson, T. S. and Phadia, E. G. (1979), "Bayesian nonparametric estimation based on censored data," *The Annals of Statistics*, 7, 163–186.

Frühwirth-Schnatter, S. (1994), "Data augmentation and dynamic linear models," *Journal of Time Series Analysis*, 15, 183–202.

Gasparini, M. (1996), "Bayesian density estimation via Dirichlet density processes," *Nonparametric Statistics*, 6, 355–366.

Gelfand, A. E. and Kottas, A. (2001), "Nonparametric Bayesian modeling for stochastic order," *Annals of the Institute for Statistical Mathematics*, 53, 865–876.

Gelfand, A. E. and Kottas, A. (2002), "A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, 11, 289–305.

Gelfand, A. E. and Mallick, B. K. (1995), "Bayesian analysis of proportional hazards models built from monotone functions," *Biometrics*, 51, 843–852.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005), "Bayesian nonparametric spatial modeling with Dirichlet process mixing," *Journal of the American Statistical Association*, 100, 1021–1035.

Ghosh, J. K. and Ramamoorthi, R. V. (2003), *Bayesian Nonparametrics*, Springer-Verlag.

Godsill, S. J., Doucet, A., and West, M. (2004), "Monte Carlo smoothing for nonlinear time series," *Journal of the American Statistical Association*, 99, 156–168.

Goggins, W. B. and Finkelstein, D. M. (2000), "A proportional hazards model for multivariate interval-censored failure time data," *Biometrics*, 56, 940–943.

Goldfeld, S. M. and Quandt, R. E. (1973), "A Markov model for switching regression," *Journal of Econometrics*, 1, 3–16.

Griffin, J. E. and Steel, M. F. J. (2006), "Order-based dependent Dirichlet processes," *Journal of the American Statistical Association*, 101, 179–194.

Grunwald, G. K., Raftery, A. E., and Guttorp, P. (1993), "Time series of continuous proportions," *Journal of the Royal Statistical Society, Series B*, 55, 103–116.

Hanson, T. and Johnson, W. O. (2002), "Modeling regression error with a mixture of Pólya trees," *Journal of the American Statistical Association*, 97, 1020–1033.

Hanson, T. and Yang, M. (2007), "Bayesian semiparametric proportional odds models," *Biometrics*, 63, 88–95.

Hanson, T., , Branscum, A., and Johnson, W. O. (2005), "Bayesian nonparametric modeling and data analysis: An introduction," in *Handbook of Statistics*, vol. 25, pp. 245–278, Elsvier.

Hanson, T. E. (2006), "Modeling censored lifetime data using a mixture of gammas baseline," *Bayesian Analysis*, 1, 575–594.

Hare, S. R. and Mantua, N. J. (2000), "Emperical evidence for North Pacific regime shifts in 1977 and 1989," *Progress in Oceanography*, 47, 103–145.

He, X., Ng, P., and Portnoy, S. (1998), "Bivariate quantile smoothing spline," *Journal of the Royal Statistical Society, Series B*, 60, 537–550.

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

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

Hjort, N. L. (1990), "Nonparametric Bayes estimators based on beta processes in models for life history data," *The Annals of Statistics*, 3, 1259–1294.

Hjort, N. L. and Petrone, S. (2005), "Nonparametric quantile inference using Dirichlet processes," in *Festschrift for Kjell Doksum*, IMS Lecture Notes, Institute of Mathematical Statistics.

Horowitz, J. L. and Lee, S. (2005), "Nonparametric quantile inference using Dirichlet processes," *Journal of the American Statistical Association*, 100, 1238–1249.

Hughes, J. P. and Guttorp, P. (1994), "A class of stochastic model for relating synoptic atmospheric patters to regional hydrologic phenomena," *Water Resources Research*, 30, 1535–1546.

Hurn, M., Justel, A., and Robert, C. P. (2003), "Estimating mixtures of regressions," *Journal of Computational and Graphical Statistics*, 12, 55–79.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001), *Bayesian Survival Analysis*, Springer-Verlag.

Ickstadt, K. and Wolpert, R. L. (1999), "Spatial regression for marked point processes," in *Bayesian Statistics*, eds. J. M. Bernardo, J. O. Berger, P. Dawid, and A. F. M. Smith, pp. 232–341, Oxford University Press.

Ishwaran, H. and James, L. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96, 161–173.

Ishwaran, H. and James, L. F. (2004), "Computational methods for multiplicative intensity models using weighted gamma processes," *Journal of the American Statistical Association*, 99, 175–190.

Ishwaran, H. and Zarepour, M. (2000), "Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models," *Biometrika*, 87, 371–390.

Ishwaran, H. and Zarepour, M. (2002), "Exact and approximate sum representations for the Dirichlet process," *Canadian Journal of Statistics*, 30, 269–283.

Jacobson, L. D., Bograd, S. J., Parrish, R. H., Mendelssohn, R., and Schwing, F. B. (2005), "An ecosystem-based hypothesis for climatic effects on surplus production in california sardine and environmentally dependent surplus production models," *Canadian Journal of Fisheries and Aquatic Sciences*, 62, 1782–1796.

Johnson, W. O. and Christensen, R. (1989), "Bayesian nonparametric survival analysis for the accelerated failure time model," *Statistics & Probability Letters*, 8, 179–184.

Kingman, J. F. C. (1993), *Poisson Processes*, Clarendon Press.

Koenker, R. (2005), *Quantile Regression*, Cambridge University Press.

Kottas, A. (2006), "Nonparametric Bayesian survival analysis using mixtures of Weibull distributions," *Journal of Statistical Planning and Inference*, 136, 578–596.

Kottas, A. and Gelfand, A. E. (2001), "Bayesian semiparametric median regression modeling," *Journal of the American Statistical Association*, 96, 1458–1468.

Kottas, A. and Krnjajić, M. (2008), "Bayesian semiparametric modeling in quantile regression," *to appear in the Scandinavian Journal of Statistics*.

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, (Special Issue on Bayesian Inference in Stochastic Processes)*, 137, 3151–3163.

Kuo, L. and Mallick, B. (1997), "Bayesian semiparametric inference for the accelerated failure-time model," *Canadian Journal of Statistics*, 25, 457–472.

Laud, P. W., Damien, P., and Smith, A. F. M. (1998), "Bayesian nonparametric and covariate analysis of failure time data," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, Springer-Verlag.

Lavine, M. (1992), "Some aspects of Pólya tree distributions for statistical modeling," *The Annals of Statistics*, 20, 1222–1235.

Leonard, T. (1977), "A Bayesian approach to some multinomial estimation and pretesting problems," *Journal of the American Statistical Association*, 72, 869–874.

Lijoi, A., Mena, R. H., and Prunster, I. (2005), "Hierarchical mixture modeling with normalized inverse-Gaussian priors," *Journal of the American Statistical Association*, 100, 1278–1291.

Lopes, H. F., Müller, P., and Rosner, G. L. (2003), "Bayesian meta-analysis for longitudinal data models using multivariate mixture priors," *Biometrics*, 59, 66–75.

MacCall, A. D. (2002), "An hypothesis explaining biological regimes in sardine-producing pacific boundary current systems," in *Climate and fisheries: interacting paradigms, scales, and policy approaches: the IRI-IPRC Pacific Climate-Fisheries Workshop*, pp. 39–42, International Research Institute for Climate Prediction, Columbia University.

MacEachern, S. N. (2000), "Dependent Dirichlet processes," Tech. rep., Ohio State University Department of Statistics.

Mallick, B. K. and Walker, S. (2003), "A Bayesian semiparametric transformation model incorporating frailties," *Journal of Statistical Planning and Inference*, 112, 159–174.

Mallick, B. K., Denison, D. G. T., and Smith, A. F. M. (1999), "Bayesian survival analysis using a MARS model," *Biometrics*, 55, 1071.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, Chapman & Hall/CRC, 2nd edn.

McGowan, J. A., Cayan, D. R., and Dorman, L. M. (1998), "Climate-ocean variability and ecosystem response in the Northeast Pacific," *Science*, 281, 210–217.

McKenzie, E. (1985), "An autoregressive process for beta random variables," *Management Science*, 31, 988–997.

Merrick, J. R. W., Soyer, R., and Mazzuchi, T. A. (2003), "A Bayesian semiparametric analysis of the reliability and maintenance of machine tools," *Technometrics*, 45, 58–69.

Møller, J. (2003), "Shot noise Cox processes," *Advances in Applied Probability*, 35, 614–640.

Møller, J. and Torrisi, G. L. (2005), "Generalized shot noise Cox processes," *Advances in Applied Probability*, 37, 48–74.

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.

Mroz, T. A. (1987), "The sensitivity of an emperical model of married women's hours of work to economic and statistical assumptions," *Econometrica*, 55, 765–799.

Muliere, P. and Walker, S. (1997), "A Bayesian non-parametric approach to survival analysis using Pólya trees," *Scandinavian Journal of Statistics*, 24, 331–340.

Müller, P. and Quintana, F. A. (2004), "Nonparametric Bayesian data analysis," *Statistical Science*, 19, 95–110.

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian curve fitting using multivariate normal mixtures," *Biometrika*, 83, 67–79.

Munch, S. B., Kottas, A., and Mangel, M. (2005), "Bayesian nonparametric analysis of stock-recruitment relationships," *Canadian Journal of Fisheries and Aquatic Sciences*, 62, 1808–1821.

Nadaraya, E. (1964), "On estimating regression," *Theory of Probability and its Applications*, 1, 141–142.

Neal, R. (1997), "Monte Carlo implementation of Gaussian process models for Bayesian regression and classification," Tech. Rep. CRG-TR-97-2, University of Toronto, Dept. of Computer Science.

Neal, R. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

Nieto-Barajas, L. E. and Walker, S. (2002), "Markov beta and gamma processes for modelling hazard rates," *Scandinavian Journal of Statistics*, 29, 413–424.

Platt, W. J., Evans, G. W., and Rathburn, S. L. (1988), "The population dynamics of a long-lived conifer (Pinus Palustris)," *The American Naturalist*, 131, 491–525.

Quandt, R. E. and Ramsey, J. B. (1978), "Estimating mixtures of normal distributions and switching regressions (with discussion)," *Journal of the American Statistical Association*, 73, 730–752.

Quinn, T. J. I. and Derisio, R. B. (1999), *Quantitative Fish Dynamics*, Oxford University Press.

Rathburn, S. L. and Cressie, N. (1994), "A space-time survival point process for a longleaf pine forest in Southern Georgia," *Journal of the American Statistical Association*, 89, 1164–1174.

Robert, C., Celeux, G., and Diebolt, J. (1993), "Bayesian estimation of hidden Markov chains: a stochastic implementation," *Statistics & Probability Letters*, 16, 77–83.

Robert, C. P., Rydén, T., and Titterington, D. M. (2000), "Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method," *Journal of the Royal Statistical Society, Series B*, 62, 57–75.

Rodriguez, A. (2007), "Some advances in Bayesian nonparametric modeling," Ph.D. thesis, Duke University, Department of Statistical Science.

Rodriguez, A. and ter Horst, E. (2008), "Bayesian dynamic density estimation," *Bayesian Analysis*, 3, 339–366.

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008), "Nonparametric functional data analysis through Bayesian density estimation," *to appear in Biometrica*.

Saltelli, A., Chan, K., and Scott, E. (eds.) (2000), *Sensitivity Analysis*, John Wiley and Sons.

Scaccia, L. and Green, P. J. (2003), "Bayesian growth curves using normal mixtures with nonparametric weights," *Journal of Computational and Graphical Statistics*, 12, 308–331.

Scott, S. (2002), "Bayesian methods for hidden Markov models: recursive computing for the 21st century," *Journal of the American Statistical Association*, 97, 337–351.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–6650.

Shi, J. Q., Murray-Smith, R., and Titterington, D. M. (2005), "Hierarchical Gaussian process mixtures for regression," *Statistics and Computing*, 15, 31–41.

Shumway, R. H. and Stoffer, D. S. (1991), "Dynamic linear models with switching," *Journal of the American Statistical Association*, 86, 763–769.

Smith, J. Q. (1979), "A generalization of the Bayesian steady forecasting model," *Journal of the Royal Statistical Society, Series B*, 41, 375–387.

Sun, J. (2006), *The Statistical Analysis of Interval-Censored Data*, Springer-Verlag.

Susarla, V. and Van Ryzin, J. (1976), "Nonparametric Bayesian estimation of survival curves from incomplete observaions," *Journal of the American Statistical Association*, 71, 897–902.

Taddy, M. A., Lee, H. K. H., Gray, G. A., and Griffin, J. D. (2007), "Bayesian guided pattern search for robust local optimization," Tech. rep., Dept. of Applied Math & Statistics, University of California, Santa Cruz.

Tobin, J. (1958), "Estimation of relationships for limited dependent variables," *Econometrica*, 26, 24–36.

Tomlinson, G. and Escobar, M. (1999), "Analysis of densities," Tech. rep., University of Toronto Department of Public Health Sciences.

Trefethen, L. N. (1980), "Numerical computation of the Schwarz–Christoffel transformation," *SIAM Journal on Scientific and Statistical Computing*, 1, 82–102.

Trefethen, L. N. (1984), "Analysis and design of polygonal resistors by conformal mapping," *Zeitschrift fr Angewandte Mathematik und Physik (ZAMP)*, 35, 692–704.

Tsionas, E. G. (2003), "Bayesian quantile inference," *Journal of Statistical Computation and Simulation*, 73, 659–674.

Wada, T. and Jacobson, L. D. (1998), "Regimes and stock-recruitment relationships in Japanese sardine, 1951-1995," *Canadian Journal of Fisheries and Aquatic Sciences*, 55, 2455–2463.

Walker, S. and Damien, P. (1998), "A full Bayesian non-parametric analysis involving a neutral to the right process," *Scandinavian Journal of Statistics*, 25.

Walker, S. G. and Mallick, B. K. (1999), "A Bayesian semiparametric accelerated failure time model," *Biometrics*, 55, 477–483.

Walker, S. G., Damien, P., Laud, P. W., and Smith, A. F. M. (1999), "Bayesian non-parametric inference for random distributions and related functions," *Journal of the Royal Statistical Society, Series B*, 61, 485–527.

Watson, G. (1964), "Smooth regression analysis," *Sankhya; Series A*, 26, 359–372.

West, M. (1995), "Bayesian inference in cyclical component dynamic linear models," *Journal of the American Statistical Association*, 90, 1301–1312.

West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, Springer Series in Statistics, Springer-Verlag, 2nd edn.

West, M., Harrison, P. J., and Migon, H. S. (1985), "Dynamic generalized linear models and Bayesian forecasting," *Journal of the American Statistical Association*, 80, 73–83.

Wheeler, F. P. (2001), "A Bayesian approach to service level performance monitoring in supplier provider relationships," *The Journal of the Operational Research Society*, 52, 383–390.

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

Yafeh, Y. and Yoshua, O. (2003), "Large shareholders and banks: Who monitors and how?" *The Economic Journal*, 113, 128–146.

Yu, K. (2002), "Quantile regression using the RJMCMC algorithm," *Computational Statistics and Data Analysis*, 40, 303–315.

Yu, K. and Moyeed, R. A. (2001), "Bayesian quantile regression," *Statistics & Probability Letters*, 54, 437–447.

Yu, K. and Stander, J. (2007), "Bayesian analysis of a Tobit quantile regression model," *Journal of Econometrics*, 137, 260–276.