15 Revisiting Bayesian curve fitting using multivariate normal mixtures

STEPHEN G. WALKER
AND GEORGE KARABATSOS

15.1 Introduction

There has been a significant amount of recent research on Bayesian nonparametric regression models. This has primarily focused on developing models of the form

$$f(y|\mathbf{x}) = \sum_{j=1}^{\infty} w_j(\mathbf{x})\, K(y|\mathbf{x}, \boldsymbol{\theta}_j(\mathbf{x}))$$

where $K(y|\mathbf{x}, \boldsymbol{\theta})$ is a chosen parametric density function, the $w_j(\mathbf{x})$ are mixture weights that sum to 1 at every value of the covariate vector $\mathbf{x} \in \mathcal{X}$, and with a prior distribution on weights $\{w_j(\mathbf{x})\}_{j=1,2,\ldots}$ and atoms $\{\boldsymbol{\theta}_j(\mathbf{x})\}_{j=1,2,\ldots}$, that are an infinite collection of processes indexed by $\mathcal{X}$. The literature on these models has exploded even in the last few years and so we cite a number of the key papers: MacEachern (1999 [13]; 2000 [14]; 2001 [15]); De Iorio *et al.* (2005) [3]; Gelfand *et al.* (2005) [6]; Griffin & Steel, (2006[7], 2010)[8]; Dunson & Park, (2008[4]); Chung & Dunson, (2009[2]); Rodríguez & Dunson, (2011[20]); Fuentes-García, *et al.* (2010[5]); Karabatsos & Walker, (2011[11]).

Our thesis is that it is a very difficult task to specify these components; i.e. the $w_j(\mathbf{x})$ and $K(y|\mathbf{x}, \boldsymbol{\theta})$. It is almost limitless in possibilities and over-fitting and un-identifiability are serious issues. There needs to be some guide as to how to choose the components of the regression model.

An intuitive approach to Bayesian nonparametric regression has been proposed by Müller *et al.* (1996[17]). The idea is to specify a Dirichlet process mixture model for the joint density $f(y, \mathbf{x})$, with mixture weights and atoms independent of $\mathbf{x}$ (Lo, 1984[12]). This would lead to a standard infinite mixture model treating the $(y_i, \mathbf{x}_i)$ as independent and identically distributed observations. This would not be a controversial choice of model.

In this case one would employ the likelihood function

$$\prod_i f(y_i, \mathbf{x}_i)$$

as used by Müller *et al.* (1996[17]). However, when the aim is regression the appropriate likelihood function is given by

$$\prod_i f(y_i, \mathbf{x}_i)/f(\mathbf{x}_i) = \prod_i f\,(y_i|\mathbf{x}_i)$$

To see this we simply note that there are two likelihood functions here and they are not the same. It is also clear that for regression purposes we are interested in the conditional density $f(y|\mathbf{x})$ and it is this that should form the basis of the likelihood function.

It is then possible to note, and we will see this later, that the weights $w_j(\mathbf{x})$ take a particular form which has not appeared in the literature:

$$w_j(\mathbf{x}) = \frac{w_j\, K(\mathbf{x}|\boldsymbol{\theta}_j)}{\sum_j\, w_j\, K(\mathbf{x}|\boldsymbol{\theta}_j)}$$

The reason why such a simple, motivated and useful regression model has not appeared in the literature is due to the posterior distribution having an intractable normalizing constant. Inference is complicated by the uncomputable integrals in the normalizing constant. Müller *et al.* (2004[18], Section 3.3) avoid this complication by proposing a modified prior distribution such that, when combined with the correct likelihood, it yields a posterior distribution that is identical to the posterior of the original model.

The aim in this paper is to show how it is possible to use the correct likelihood for regression by describing how to deal with the problem of the normalizing constant. This uses ideas recently introduced in Walker (2011[22]), who shows how latent variables can be used to construct a latent model which can be studied using Markov chain Monte Carlo (MCMC) methods. The aim is not to compare the model with other models; we take it for granted, based on the work of Müller *et al.* (1996[17]), that the model is going to be useful.

The layout of the paper is as follows: Section 15.2 fully describes our regression model, and the methods for sampling the posterior distribution of the model. To obtain full posterior inference of the model, a reversible-jump sampling algorithm (Walker, 2011[22]) is used to deal with the uncomputable normalizing constant. In Section 15.3 we illustrate our model through data analysis.

15.2 The regression model and modelling methods

15.2.1 The model

First we describe the model for $(y, \mathbf{x})$ which is a standard Bayesian nonparametric mixture model based on the mixture of Dirichlet process model. If we take the normal density as the kernel density; i.e.

$$\mathrm{n}(y|\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$$

so

$$f(y, \mathbf{x}) = \sum_j w_j\, \mathrm{n}\,(y, \mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$$

where the (w_j) are stickbreaking weights, that is, $w_j = \lambda_j \prod_{l=1}^{j-1}(1-\lambda_l)$, $\lambda_j \in [0,1]$, for $j \geq 1$ (Sethuraman, 1994[21]). As has been mentioned, such a modelling approach was taken by Müller *et al.* (1996[17]). However, for regression modelling, the 'correct' likelihood function is

$$\prod_{i=1}^{n} f(y_i|\mathbf{x}_i)$$

Hence, given data $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, we arrive at

$$f(y|\mathbf{x}) = \frac{f(y,\mathbf{x})}{p(\mathbf{x})} = \frac{\sum_j \mathrm{n}(y,\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma})\, w_j}{\sum_j \mathrm{n}(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{x}j}, \Sigma_{\mathbf{x}})\, w_j} \tag{15.1}$$

To elaborate, the $f(y,\mathbf{x})$ has a numerator as though

$$\begin{pmatrix} y \\ \mathbf{x} \end{pmatrix} = \mathrm{n}\left(\begin{pmatrix} \mu_{yj} \\ \boldsymbol{\mu}_{\mathbf{x}j} \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \rho_{y\mathbf{x}} \\ \rho_{y\mathbf{x}} & \Sigma_{\mathbf{x}} \end{pmatrix}\right)$$

and $\boldsymbol{\mu}_j = (\mu_{yj}, \boldsymbol{\mu}_{\mathbf{x},j})$ and

$$\Sigma = \begin{pmatrix} \sigma_y^2 & \rho_{y\mathbf{x}} \\ \rho_{y\mathbf{x}} & \Sigma_{\mathbf{x}} \end{pmatrix}$$

But the likelihood does not assume that the $\mathbf{x}$s are randomly generated. It is obviously a valid likelihood since if we integrate out the (y_i) we get 1. But it is strictly a regression model which has motivation when it is acknowledged that constructing $f(y|\mathbf{x})$ is problematic and guidelines are required. Of course, if the $\mathbf{x}$s are randomly generated then the model is fully motivated.

The model is completed by the specification of prior densities $\lambda_j \sim_{ind} \mathrm{beta}(a_j, b_j)$ and $\boldsymbol{\mu}_j \sim_{ind} \pi_j(\boldsymbol{\mu}_j)$, $j = 1, 2, \ldots$, and $\Sigma \sim \pi(\Sigma)$. The choice of (a_j, b_j) will be discussed later but we point out now that the choice of $a_j = 1$ and $b_j = b > 0$ leads to the Dirichlet process. Also, a default choice of priors for the mean and covariance matrix $(\boldsymbol{\mu}_j, \Sigma)$ is given by a multivariate normal and inverted-Wishart prior densities, i.e. $\boldsymbol{\mu}_j \sim_{iid} \mathrm{n}_q(\boldsymbol{\mu}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}})$ and $\Sigma \sim \mathrm{iw}_q(\nu_\Sigma, \mathbf{T})$, with $q = \dim(y,\mathbf{x})$, with $(1/\{\nu_\Sigma - q - 1\})\mathbf{T}$ as the mean of the inverted-Wishart density.

We need the denominator of the likelihood, as it contains parameters. It would appear that we would not be able to do inference due to the intractable nature of the denominator, but it turns out that inference is possible by defining a suitable latent model.

If we write

$$f(\mathbf{x}) = |\Sigma_{\mathbf{x}}|^{-1/2} \sum_{j=1}^{\infty} w_j\, m(\mathbf{x}; \boldsymbol{\mu}_{\mathbf{x}j}, \Sigma_{\mathbf{x}})$$

where

$$m(\mathbf{x}; \boldsymbol{\mu}_{\mathbf{x}j}, \Sigma_{\mathbf{x}}) = \exp\left\{-0.5(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}j})^\top \Sigma_{\mathbf{x}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}j})\right\}$$

then we can easily note that

$$m(\mathbf{x}) = \sum_{j=1}^{\infty} w_j\, m(\mathbf{x}; \boldsymbol{\mu}_{\mathbf{x}j}, \Sigma_{\mathbf{x}}) < 1$$

And this will be the key to appropriately dealing with the denominator. The idea is that for any $0 < \zeta < 1$ it is that

$$\sum_{k=0}^{\infty}(1-\zeta)^k = \zeta^{-1}$$

Hence, we can represent the denominator at a generic $\mathbf{x}$ value as

$$|\Sigma_{\mathbf{x}}|^{1/2}\sum_{k=0}^{\infty}(1-m(\mathbf{x}))^k$$

and hence we can represent the $\prod_i f(\mathbf{x}_i)$ as

$$|\Sigma_{\mathbf{x}}|^{n/2}\prod_{i=1}^{n}\prod_{l=1}^{k_i} w_{d_{il}}(1-m(\mathbf{x}_i;\boldsymbol{\mu}_{\mathbf{x}_i\, d_{il}},\Sigma_{\mathbf{x}}))$$

For now if we sum over the d_{il}, i.e. for each i and $l \in \{1,\ldots,k_i\}$, we have $d_{il} \in \{1,2,\ldots\}$, then we recover

$$|\Sigma_{\mathbf{x}}|^{n/2}\prod_{i=1}^{n}(1-m(\mathbf{x}_i))^{k_i}$$

and if we now sum over each $k_i \in \{0,1,2,\ldots\}$ then we recover

$$|\Sigma_{\mathbf{x}}|^{n/2}\prod_{i=1}^{n} m^{-1}(\mathbf{x}_i)$$

which is precisely

$$\prod_{i=1}^{n} f^{-1}(\mathbf{x}_i)$$

Specifically, after introducing latent variables $(u_i, u_{il}, v_{il}, d_i, d_{il}, k_i)$, the likelihood becomes:

$$|\Sigma_{\mathbf{x}}|^{n/2}\prod_{i=1}^{n}\Big[\mathbf{1}\left(0<u_i<\xi_{d_i}\right)w_{d_i}\xi_{d_i}^{-1}\mathrm{n}(y_i,\mathbf{x}_i;\boldsymbol{\mu}_{d_i},\Sigma) \tag{15.2}$$

$$\times\left\{\prod_{l=1}^{k_i}\mathbf{1}(0<u_{il}<1-\exp[-.5(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x}d_{il}})^{\top}\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x}d_{il}})]),\right.$$

$$\left.\times\,\mathbf{1}(0<v_{il}<\xi_{d_{il}})w_{d_{il}}\,\xi_{d_{il}}^{-1}\right\}\Big] \tag{15.3}$$

where $\mathbf{1}(\cdot)$ is the indicator function, and ξ_j is a fixed decreasing function, such as $\xi_j = \exp(-j)$. This slicing strategy is to force the (d_i), and also the (d_{il}) to be bounded and hence can be sampled in a MCMC algorithm. See Kalli *et al.* (2010 [10]) for a discussion on the choice of (ξ_j). It is easy to show that marginalizing over the latent variables yields the correct likelihood $f(y|\mathbf{x})$ of the regression

model, as in (15.1). Importantly, the combined use of ξ with the latent variables facilitates MCMC sampling of the posterior distribution of the infinite-dimensional regression model.

15.2.2 MCMC sampling

For the regression model, an MCMC sampling algorithm is used to iteratively and repeatedly sample from its full conditional posterior distributions. Let $\mathbf{1}_N(j)$ be the function indicating whether $j \in \{1, \ldots, N\}$. Given the likelihood (15.2), the full conditional posterior distributions are given below, for $i = 1, \ldots, n$, $j = 1, \ldots, N$, and $l = 1, \ldots, k_i$.

$$\pi(u_i|\cdots) \propto \mathbf{1}(0 < u_i < \xi_{d_i});$$

$$\pi(u_{il}|\cdots) \propto \mathbf{1}(0 < u_{il} < 1 - \exp\{-.5(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}d_{il}})^\top \Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}d_{il}})\});$$

$$\pi(v_{il}|\cdots) \propto \mathbf{1}(0 < v_{il} < \xi_{d_{il}});$$

$$\pi(\lambda_j|\cdots) = \text{beta}\begin{pmatrix} a_j + \#(d_i = j) + \#(d_{il} = j), \\ b_j + \#(d_i > j) + \#(d_{il} > j) \end{pmatrix} \mathbf{1}_N(j)$$

$$\Pr\left(d_i = j|\cdots\right) \propto w_j \xi_j^{-1} \text{n}(y_i, \mathbf{x}_i; \boldsymbol{\mu}_j, \Sigma)\mathbf{1}_{N_i}(j);$$

$$\Pr(d_{il} = j|\cdots) \propto w_j \xi_j^{-1}\{1 - \exp[-.5(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}j})^\top \Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}j})]\}\mathbf{1}_{N_{il}}(j);$$

$$\pi(\boldsymbol{\mu}_j|\ldots) \propto \pi_j(\boldsymbol{\mu}_j)\mathbf{1}_N(j) \prod_{d_i=j} \text{n}(y_i, \mathbf{x}_i|\boldsymbol{\mu}_{d_i}, \Sigma)$$
$$\times \prod_{d_{il}=j} \mathbf{1}\left(u_{il} < 1 - \exp[-.5(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}d_{il}})^\top \Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}d_{il}})]\right);$$

$$\pi(\Sigma|\cdots) \propto \pi(\Sigma)|\Sigma_{\mathbf{x}}|^{n/2}\left[\prod_{i=1}^{n} \text{n}(y_i, \mathbf{x}_i|\boldsymbol{\mu}_{d_i}, \Sigma)\right]$$
$$\times \prod_{i=1}^{n}\prod_{l=1}^{k_i} \mathbf{1}\left(u_{il} < 1 - \exp[-.5(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}d_{il}})^\top \Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}d_{il}})]\right)$$

Importantly, since ξ_j is a decreasing function, we can define finite values of $N = \max[\max_i N_i, \max_{i,l} N_{il}]$, $N_i = \max_j j\mathbf{1}(0 < u_i < \xi_j)$, $N_{il} = \max_j j\mathbf{1}(0 < v_{il} < \xi_j)$. Therefore, conditional on latent variables $(d_i, d_{il}, u_i, u_{il}, v_{il}, k_i)$, posterior sampling proceeds as if the infinite mixture model were a finite-dimensional model, as in Kalli *et al.* (2010[10]). All full conditionals can be easily sampled. The nonstandard full conditional densities $p(\boldsymbol{\mu}_j|\ldots)$ are each sampled using the random-walk Metropolis–Hastings algorithm, with normal proposal density $\text{n}_q(\boldsymbol{\mu}_j, \text{diag}(v_1, \ldots, v_q))$ having variances $(v_1, \ldots, v_q)$ automatically adapted to achieve the desired acceptance rate of 0.44 for each respective component of $\boldsymbol{\mu}$ over MCMC iterations, using a Robbins–Monro algorithm (see Atchadé & Rosenthal, 2005[1]). Also, when the model is assigned prior $\Sigma \sim \text{iw}_q(\nu_\Sigma, \mathbf{T})$, the nonstandard full conditional posterior density $\pi(\Sigma|\cdots)$ can be sampled using an independent Metropolis–Hastings algorithm, with $\text{iw}_q(\nu_\Sigma, c\mathbf{T})$ as the proposal density for some chosen constant $c > 0$.

The sampling of the full conditional distribution $\Pr(k_i|\cdots)$ requires reversible-jump MCMC methods (Walker, 2011[22]), because the dimensionality of the model parameters changes with k_i. In particular, using the full set $(u_{il}, v_{il}, d_{il})_{l=1}^{\infty}$ we construct the following joint density

$$p(k_i,(u_{il},v_{il},d_{il})_{l=1}^{\infty}|\cdots)\propto\prod_{l=1}^{k_i}\mathbf{1}\left(0<u_{id_{il}}<1-\exp\left[\begin{array}{c}-.5(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x}d_{il}})^{\top}\\ \times\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x}d_{il}})\end{array}\right]\right)$$
$$\times\prod_{l=1}^{k_i}\mathbf{1}_{N_{il}}(d_{il})\mathbf{1}(0<v_{il}<\xi_{d_{il}})\,w_{d_{il}}\xi_{d_{il}}^{-1}$$
$$\times\prod_{l=k_i}^{\infty}p(u_{i,l+1},v_{i,l+1},d_{i,l+1}|u_{il},v_{il},d_{il}),$$

where the last term is a product of densities, each serving as a proposal density, which could be chosen as an independent (proposal) density:

$$p(u_{i,l+1},v_{i,l+1},d_{i,l+1}|u_{il},v_{il},d_{il})=p(u_{i,l+1},v_{i,l+1},d_{i,l+1})=\mathbf{1}_{N_{il}}(d_{i,l+1})\mathbf{1}(0<v_{i,l+1}<\xi_{d_{i,l+1}})$$
$$\times\frac{\mathbf{1}(0<u_{id_{i,l+1}}<1-\exp[-.5(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x}d_{i,l+1}})^{\top}\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x}d_{i,l+1}})])}{1-\exp[-.5(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x}d_{i,l+1}})^{\top}\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x}d_{i,l+1}})]}.$$

Then given that the MCMC chain is at state k_i, a proposal is made to move to state k_i+1 with probability $q(k_i+1|k_i)$, and given a sample $(u_{i,k_i+1},v_{i,k_i+1},d_{i,k_i+1})$ from $p(u_{i,k_i+1},v_{i,k_i+1},d_{i,k_i+1})$, this proposal is accepted with probability

$$\min\left[1,\frac{w_{d_{i,k_i+1}}\xi_{d_{i,k_i+1}}^{-1}\{1-\exp[-.5(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x}d_{i,k_i+1}})^{\top}\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x}d_{i,k_i+1}})]\}}{q(k_i+1|k_i)/q(k_i|k_i+1)}\right]$$

Otherwise, with probability $q(k_i-1|k_i)$, a proposal is made to move to state k_i-1, and this proposal is accepted with probability

$$\min\left[1,\frac{q(k_i|k_i-1)/q(k_i-1|k_i)}{w_{d_{ik_i}}\xi_{d_{ik_i}}^{-1}\{1-\exp[-.5(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x}d_{ik_i}})^{\top}\Sigma_{\mathbf{x}}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{x}d_{ik_i}})]\}}\right].$$

We can define the proposal distribution by $q(1|0)=1, q(0|1)=0$, and $q(k'|k)=.5$ for $k>0$ and for all $|k'-k|=1$.

Finally, the full conditional posterior (predictive) density of Y_i $(i=1,\ldots,n)$ is:

$$\mathrm{n}(\mu_{yd_i}+\textstyle\sum_{s=1}^{p}(x_{si}-\mu_{x_sd_i})/\Delta_y,1/\Delta_y),$$

with $\Delta=\Sigma^{-1}$, following known results involving a univariate conditional distribution of a multivariate normal density. Also, the predictive accuracy of the model can proceed via the evaluation of standardized residuals $(y_i-\mu_{yd_i})\Delta_y^{1/2}$, $i=1,\ldots,n$.

To conduct MCMC sampling of the model, we wrote code in MATLAB (2011, The MathWorks, Natick, MA).

15.3 Illustrations

In this section we illustrate our model through the analysis of simulated and real data. In each illustration, we have rescaled y and each covariate x_s $(s = 1, \ldots, p)$ to have mean 0 and variance 1 prior to model fitting, assigned prior densities $\lambda_j \sim_{ind} \text{beta}(1/2, 1/2 + j/2)$, $\boldsymbol{\mu}_j \sim_{iid} \text{n}_q(\mathbf{0}, \mathbf{I})$ $(j = 1, 2, \ldots)$, and $\Sigma \sim \text{iw}_q(q + 2, \mathbf{I})$ to the parameters of the regression model, and chosen an inverted-Wishart $\text{iw}_q(\nu_\Sigma, .5\mathbf{T})$ proposal density for the independent Metropolis sampling of Σ in the MCMC algorithm. Hence, a Poisson–Dirichlet (Pitman–Yor) process prior was assigned to the mixture weights λ_j (e.g., Ishwaran & James, 2001[9]). Also, all results are reported on the original scale of y and $\mathbf{x}$. Finally, for each data illustration, we have estimated the regression model based on 20 000 samples of the MCMC algorithm, long after the predictions of the model seemed to stabilize over the MCMC iterations. We estimate the predictions of the model from every fifth MCMC sample of its predictive distribution.

15.3.1 Simulation

We first illustrate the model through the analysis of a simulated dataset with 61 observations, coming from $Y_i \sim \text{n}(0.2x_i^3, 0.25)$, with $x_1 = -3, x_2 = -2.9, \ldots, x_{n-1} = 2.9, x_n = 3$. Figure 15.1 presents the simulated data, and for the model, presents the estimate of the posterior predictive mean and interquartile range of Y conditional on each of the observed values of x. We see that the range captures the small number of simulated data points.

15.3.2 Crime Data

Here we present an analysis of a dataset described in McNeil (1977[16]), consisting of information on urban population percentage, the number of murder arrests, assault arrests, and rape arrests per

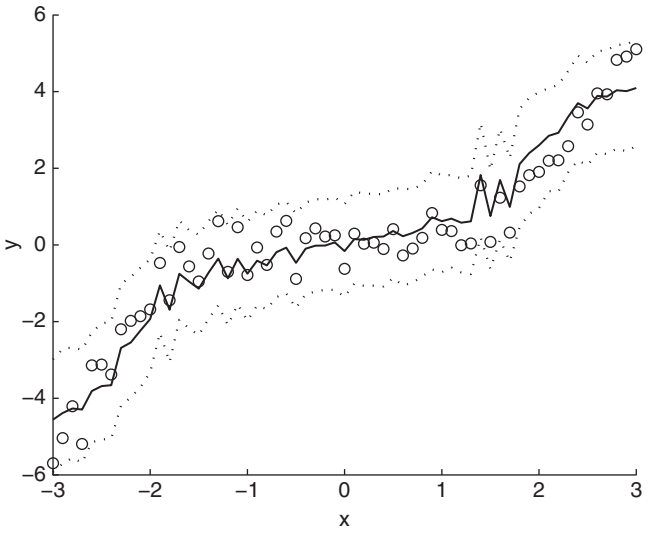

Figure 15.1 Simulated example. From the posterior predictive distribution of the model, the mean (solid line) and interquartile range (dashed lines) of Y, conditional on values of x.

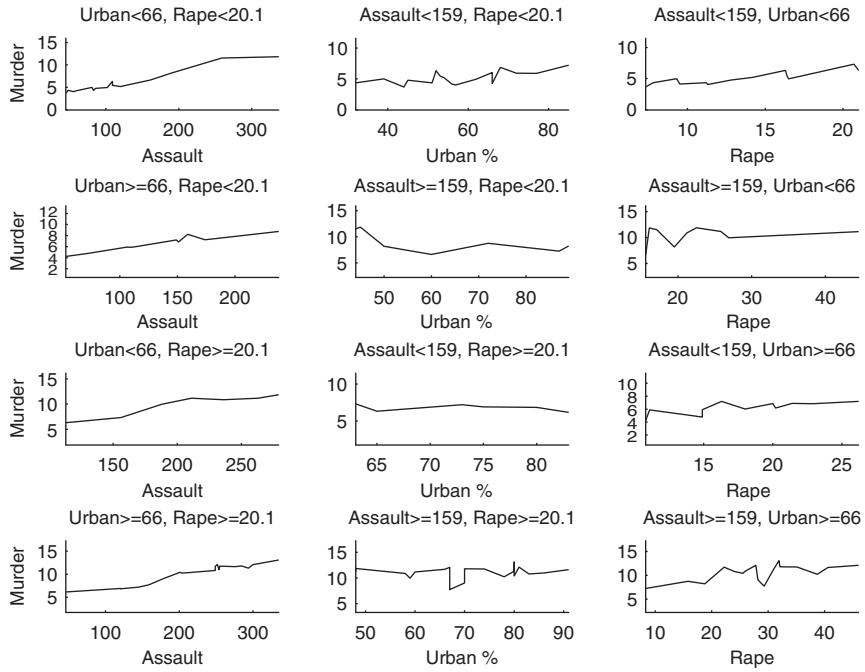

Figure 15.2 Crime data. Estimates of the posterior predictive mean of Y, conditional on values of the three covariates.

100 000 people, for each of the $n = 50$ U.S. states during the year 1973. This 'USArrests' dataset was obtained from the datasets package of the R software (R Development Core Team[19]). We treat murder rate as the dependent variable, and the other three variables as covariates.

From the regression model, Figure 15.2 presents estimates of the posterior predictive mean of Y, conditional on values of the three covariates. The figure shows that there is generally a positive correlation between assault arrests and rape arrests with murder arrests, and that there are nonlinear and interactive relationships between each of the three covariates and murder arrests. Also, an inspection of the posterior predictive distribution of the standardized residuals revealed that there were no outliers in the model.

15.4 Discussion

In this paper, we have developed a Bayesian nonparametric regression model which relies on a standard Bayesian nonparametric form for the joint distribution of both the dependent and independent variables. The regression model then is available as a conditional density which can only be written as a ratio of two infinite-dimensional mixture models.

While this model has been acknowledged as desirable, its drawback has been the intractability of the likelihood, specifically the denominator of the likelihood. However, recent advances in both latent models for dealing with normalizing constants and simulation techniques linking slice sampling and infinite mixture models have now rendered this model tractable.

References

[1] Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, **11**, 815–828.

[2] Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modelling with variable selection. *Journal of the American Statistical Association*, **104**, 1646–1660.

[3] DeIorio, 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.

[4] Dunson, D. and Park, J.-H. (2008). Kernel stick breaking processes. *Biometrika*, **95**, 307–323.

[5] Fuentes-García, R., Mena, R. H. and Walker, S. G. (2010). A new Bayesian nonparametric mixture model. *Communications In Statistics*, **39**, 669–682.

[6] Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005). Bayesian nonparametric spatial modelling with Dirichlet processes mixing. *Journal of the American Statistical Association*, **100**, 1021–1035.

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

[8] Griffin, J. E. and Steel, M. F. J. (2010). Bayesian nonparametric modelling with the Dirichlet process regression smoother. *Statistica Sinica*, **20**, 1507–1527.

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

[10] Kalli, M., Griffin, J. and Walker, S. G. (2010). Slice sampling mixture models. *Statistics and Computing*, **21**, 93–105.

[11] Karabatsos, G. and Walker, S. G. (2011). Bayesian unimodal density regression. Technical report, University of Illinois, Chicago.

[12] Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. *Annals of Statistics*, **12**, 351–357.

[13] MacEachern, S. N. (1999). Dependent nonparametric processes. *Proceedings of the Bayesian Statistical Sciences Section of the American Statistical Association*, 50–55.

[14] MacEachern, S. N. (2000). Dependent Dirichlet Processes. Technical report, Department of Statistics, The Ohio State University.

[15] MacEachern, S. N. (2001). Decision theoretic aspects of dependent nonparametric processes. In *Bayesian Methods with Applications to Science, Policy and Official Statistics* (ed. E. George), Creta, pp. 551–560. International Society for Bayesian Analysis.

[16] McNeil, D. R. (1977). *Interactive Data Analysis*. John Wiley, New York.

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

[18] Müller, P., Quintana, F. A. and Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society, Series B*, **66**, 735–749.

[19] R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[20] Rodriguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, **6**, 1–34.

[21] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.

[22] Walker, S. G. (2011). Posterior sampling when the normalizing constant is unknown. *Communications in Statistics: Simulation and Computation*, **40**, 784–792.