Bayesian Density Regression

David B. Dunson[1,2] **and Natesh Pillai**[2]

[1] *Biostatistics Branch*

MD A3-03, National Institute of Environmental Health Sciences

P.O. Box 12233, Research Triangle Park, NC 27709

E-mail: dunson1@niehs.nih.gov

[2] *Institute of Statistics and Decision Sciences*

Duke University

Summary. This article considers Bayesian methods for density regression, allowing a random probability distribution to change flexibly with multiple predictors. The conditional response distribution is expressed as a nonparametric mixture of parametric densities, with the mixture distribution changing according to location in the predictor space. A new class of priors for dependent random measures is proposed for the collection of random mixing measures at each location. The conditional prior for the random measure at a given location is expressed as a mixture of a Dirichlet process (DP) distributed innovation measure and neighboring random measures. This specification results in a coherent prior for the joint measure, with the marginal random measure at each location being a finite mixture of DP basis measures. Integrating out the infinite-dimensional collection of mixing measures, we obtain a simple expression for the conditional distribution of the subject-specific random variables, which generalizes the Pólya urn scheme. Properties are considered and a simple Gibbs sampling algorithm is developed for posterior computation. The methods are illustrated using simulated data examples and epidemiologic studies.

Keywords: Conditional density function; Dirichlet process; Mixture model; Nonparametric Bayes; Pólya urn Gibbs sampler; Random probability measure field; Semiparametric; Smoothing.

1. Introduction

1.1. *Problem Formulation & Background*

This article addresses the problem of density regression, investigating changes in the distribution of a random variable $Y \in \mathcal{Y}$ according to predictors $\mathbf{x} = (x_1, \ldots, x_p)' \in \mathcal{X}$ using a Bayesian semiparametric approach. The sample space $\mathcal{Y}$ can be either discrete or continuous and bounded or unbounded. Although we focus on the case in which $\mathcal{X} = \mathcal{X}_C \subset \Re^p$ is a continuous sample space, the proposed framework can also be applied when there are both continuous and discrete predictors so that $\mathcal{X} = \mathcal{X}_C \times \mathcal{X}_D$, where $\mathcal{X}_D$ is a discrete space. A challenge is that the distribution function of Y given $\mathbf{x}$ is unknown, and there can be unanticipated changes in the shape of the distribution according to the location of $\mathbf{x} \in \mathcal{X}$. Thus, it is not appropriate to assume that the residual distribution in a mean or quantile regression model is constant over $\mathcal{X}$.

A recent article by De Iorio et al. (2004) proposed a Bayesian nonparametric approach for modeling of dependence across random distributions $G_{\mathbf{x}}$ indexed by a vector $\mathbf{x} \in \mathcal{X}_D$ of categorical covariates. In particular, they defined a prior for the array of random measures $\mathcal{G}_{\mathbf{X}} = \{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}_D\}$, which maintains a marginal Dirichlet process (DP) (Ferguson, 1973; 1974) structure for the distribution at each value of $\mathbf{x}$. This is accomplished using the dependent Dirichlet process (DDP) approach of MacEachern (1999; 2000; 2001), which relies on expressing the DP in stick-breaking form (Sethuraman, 1994):

$$G_{\mathbf{x}} = \sum_{h=1}^{\infty} p_{\mathbf{x},h} \delta_{\theta_{\mathbf{x},h}}, \quad \text{with } p_{\mathbf{x},h} / \textstyle\prod_{l=1}^{h-1}(1 - p_{\mathbf{x},l}) \overset{iid}{\sim} \text{beta}(1, \alpha),$$

where $\{p_{\mathbf{x},h}\}$ are random weights, δ_θ is a degenerate distribution with all its mass at 0, and $\{\theta_{\mathbf{x},h}\}$ are atoms generated from the base measure $G_{0,\mathbf{x}}$. Sethuraman (1994) showed that this characterization is equivalent to assuming $G_{\mathbf{x}} \sim DP(\alpha G_{0,\mathbf{x}})$, where $DP(\alpha G_0)$ denotes the Dirichlet process centered on base measure G_0 with precision α.

Assuming a common set of weights, $p_{\mathbf{x},h} = p_h$ for all $\mathbf{x} \in \mathcal{X}_D$, MacEachern (1999; 2001) allows for dependency by defining a stochastic process for the atoms $\{\theta_{\mathbf{x},h}\}$. De Iorio et al. (2004) used the DDP to produce an ANOVA-type dependency structure, while Gelfand et al. (2005) applied the DDP to spatial modeling applications by using a Gaussian process for the atoms. Recently, Griffin

and Steel (2005) proposed an order-based DDP, which allows the weights to vary with covariates. An alternative dynamic form of the DP was proposed by Dunson (2004) to model changes in a random distribution across levels of an ordered categorical predictor. Cifarelli and Regazzini (1978) instead introduce dependence across related measures by using DP priors linked through a common regression component in the base measure. Related approaches have been considered by Muliere and Petrone (1993), Mira and Petrone (1996), Giudici, Mezzetti, and Muliere (2003), and Griffin and Steel (2004).

Müller, Erkanli and West (1996) instead used a DP mixture of normals for the joint distribution of y and $\mathbf{x}$, and then focused on the implied conditional density of y given $\mathbf{x}$ in estimating the mean regression function. Another strategy for allowing dependence in random measures is to allow the measures to depend on a shared set of latent factors, which are assigned Dirichlet process priors. Gelfand and Kottas (2001) proposed such an approach to address the problem of modeling of stochastically ordered distributions, expressing random variables as products of DP distributed factors. Pennell and Dunson (2004) used a conceptually related idea to model dependence in time-dependent frailty distributions within a multiple event time model. Such approaches are not straightforward to extend to general regression settings. For a recent overview of Bayesian nonparametric inference, refer to Müller and Quintana (2004).

1.2. *Mixture Modeling Structure*

This article proposes a different type of approach. For subject i $(i = 1, \ldots, n)$, express the conditional density of the response y_i given $\mathbf{x}_i$ as a nonparametric mixture of parametric densities as follows:

$$f(y_i \,|\, \mathbf{x}_i) = \int f(y_i \,|\, \mathbf{x}_i, \phi_i)\, dG_{\mathbf{x}_i}(\phi_i), \tag{1}$$

where $f(y \,|\, \mathbf{x}, \phi)$ is a known kernel that depends on the finite-dimensional parameter ϕ, and $G_{\mathbf{x}_i}$ is a random mixing measure that can vary according to the location of $\mathbf{x}_i \in \mathcal{X}$. It is well known that, given sufficient numbers of components, mixtures of Gaussian or exponential family distributions are extremely flexible.

In the special case in which $G_{\mathbf{x}} \equiv \delta_\theta$, expression (1) reduces to the parametric model $f(y_i \,|\, \mathbf{x}_i) = f(y_i \,|\, \mathbf{x}_i, \phi_i = \theta)$. Now suppose $G_{\mathbf{x}} \equiv G = \sum_{h=1}^{\infty} p_h \delta_{\theta_h}$, with $\theta_h \overset{iid}{\sim} G_0$ and $p_h / \prod_{l=1}^{h-1}(1-p_l) \overset{iid}{\sim}$ beta$(1, \alpha)$, $h = 1, \ldots, \infty$. Then, a single, unknown mixing measure, $G \sim DP(\alpha G_0)$, holds for all $\mathbf{x} \in \mathcal{X}$, and expression (1) is a standard Dirichlet process mixture model (DPM). Unfortunately, the assumption of a single mixing measure may be overly restrictive in certain applications. For example, if ϕ consists of coefficients in a regression model, a few distinct values may be sufficient within a particular local region of $\mathcal{X}$. However, the distribution of the coefficients may need to change across subregions of $\mathcal{X}$ to accommodate evolving deviations from the parametric model.

In regression problems involving a continuous response, with $\mathcal{Y} \equiv \Re$, a simple default choice for $f(y_i \,|\, \mathbf{x}_i, \phi_i)$ would be $\mathrm{N}(y_i; \mathbf{x}_i' \boldsymbol{\beta}_i, \sigma^2)$, with $\phi_i = (\boldsymbol{\beta}_i', \sigma^2)'$, possibly with σ^2 also varying with i. In this case, expression (1) is a mixture of normal linear regression models. By allowing the mixture distribution for the slopes, $\boldsymbol{\beta}_i$, to vary according to the predictor values, $\mathbf{x}_i$, one can allow an unknown nonlinear mean regression function and an unknown non-stationary residual distribution. For different choices of $\mathcal{Y}$, one can appropriately modify the default choice of $f(y_i \,|\, \mathbf{x}_i, \phi_i)$. For example, when the outcome is a count, we can choose a Poisson log-linear model, and when the outcome is Bernoulli, we can choose a logistic regression model.

Section 2 provides details on the proposed prior specification. Section 3 outlines a Gibbs sampling algorithm for posterior computation. Section 4 illustrates the approach through simulated data examples. Section 5 contains an application to data from an epidemiologic study, and Section 6 discusses the results.

2. Priors for Density Regression

2.1 *Proposed Formulation and Properties*

There is an infinite collection of random probability measures $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}}, \forall \mathbf{x} \in \mathcal{X}\}$. Focusing initially on the observed predictor values, we have $\mathcal{G}_{\mathbf{X}} = \{G_{\mathbf{x}_i}, i = 1, \ldots, n\}$. We assume that $\phi_i \overset{ind}{\sim} G_{\mathbf{x}_i}$ conditionally on $\mathbf{X}$ and $\mathcal{G}_{\mathbf{X}}$ and that each $G_{\mathbf{x}_i}$ can be expressed as:

$$G_{\mathbf{x}_i} = \sum_{h=1}^{\infty} p_{hi} \delta_{\theta_{hi}}, \quad \text{for } i = 1, \ldots, n, \tag{2}$$

with atoms $\mathbf{\Theta}_i = \{\theta_{hi}, h = 1, 2, \ldots, \infty\}$, random weights $\mathbf{p}_i = \{p_{hi}, h = 1, 2, \ldots, \infty\}$, and $\sum_{h=1}^{\infty} p_{hi} = 1$ a.s. Dependence in the random measures within the collection $\mathcal{G}_{\mathbf{X}}$ arises through common atoms and dependent weights. Overall, the set of atoms is denoted $\mathbf{\Theta} = \bigcup_{i=1}^{n} \mathbf{\Theta}_i$, with the elements of $\mathbf{\Theta}$ assumed to be generated independently from non-atomic base measure G_0. Re-expressing (2) in terms of the atoms in $\mathbf{\Theta}$, we have

$$G_{\mathbf{x}_i} = \sum_{h=1}^{\infty} P_{hi} \delta_{\Theta_h}, \quad \text{for } i = 1, \ldots, n, \tag{3}$$

where $\mathbf{P}_i = \{P_{hi}, h = 1, 2, \ldots, \infty\}$, $\sum_{h=1}^{\infty} P_{hi} = 1$ a.s., and the random weights $\mathbf{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_n\} \sim \mathcal{P}$ are generated independently from the atoms $\mathbf{\Theta}$.

Let $\mathcal{Y}$ be a complete separable metric space with $\mathcal{B}$ the corresponding Borel $\sigma-$algebra on $\mathcal{Y}$. Then, let $\mathcal{M}(\mathcal{Y})$ denote the space of probability measures on $(\mathcal{Y}, \mathcal{B})$. For each collection of disjoint Borel sets $(B_1, \ldots, B_L)$, the prior for $G_{\mathbf{x}_i}$ in expression (3) prescribes that

$$\begin{aligned} \Big(G_{\mathbf{x}_i}(B_1), G_{\mathbf{x}_i}(B_2), \ldots, G_{\mathbf{x}_i}(B_L)\Big) = \\ \Big(\sum_{h=1}^{\infty} P_{hi} 1(\Theta_h \in B_1), \sum_{h=1}^{\infty} P_{hi} 1(\Theta_h \in B_2), \ldots, \sum_{h=1}^{\infty} P_{hi} 1(\Theta_h \in B_L)\Big), \end{aligned} \tag{4}$$

which places a probability measure $\Pi^{\mathbf{x}_i}_{B_1,\ldots,B_L}$ on the L-dimensional probability simplex S_L. The expectation of $\Pi^{\mathbf{x}_i}_{B_1,\ldots,B_L}$ integrating out the unknown atoms and weights is

$$\begin{aligned} \mathrm{E}(\Pi^{\mathbf{x}_i}_{B_1,\ldots,B_L}) &= \int \Big(\sum_{h=1}^{\infty} P_{hi} 1(\Theta_h \in B_1), \ldots, \sum_{h=1}^{\infty} P_{hi} 1(\Theta_h \in B_L)\Big)\Big\{\prod_{h=1}^{\infty} dG_0(\Theta_h)\Big\} d\mathcal{P}(\mathbf{P}) \\ &= \Big(\sum_{h=1}^{\infty} P_{hi} G_0(B_1), \ldots, \sum_{h=1}^{\infty} P_{hi} G_0(B_L)\Big) = \Big(G_0(B_1), \ldots, G_0(B_L)\Big), \end{aligned}$$

which directly implies that $\mathrm{E}\{G_{\mathbf{x}_i}\} = G_0$, so that the prior for $G_{\mathbf{x}_i}$ is centered on G_0.

As a general approach for characterizing dependency in the random measures in the collection $\mathcal{G}_{\mathbf{X}}$, we propose the following conditional mixture structure:

$$G_{\mathbf{x}_i} = a_{ii} G^*_{\mathbf{x}_i} + \sum_{j \sim i} a_{ij} G_{\mathbf{x}_j}, \tag{5}$$

which expresses $G_{\mathbf{x}_i}$ as a mixture of an innovation random measure $G^*_{\mathbf{x}_i} \sim DP(\alpha G_0)$, which is assigned a Dirichlet process prior, and neighboring random measures $\{G_{\mathbf{x}_j}, j \sim i\}$. Here, $\mathbf{a}_i = (a_{i1}, \ldots, a_{in})'$ are mixture probabilities, with $0 \leq a_{ij} \leq 1$, $a_{ii} + \sum_{j \sim i} a_{ij} = 1$, and $\{a_{ij} = 0 \; \forall \; j \not\sim i\}$.

In addition, $j \sim i$ indexes subjects $j \in \mathcal{N}_i \subset \{1, \ldots, n\}/i$ with $\mathcal{N}_i = \{j : d(\mathbf{x}_i, \mathbf{x}_j) < \epsilon, j \neq i\}$, where $d(\mathbf{x}_i, \mathbf{x}_j)$ is a known measure of distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ and ϵ is a positive constant. The innovation measures within $\mathcal{G}^*_{\mathbf{X}} = \{G^*_{\mathbf{x}_1}, \ldots, G^*_{\mathbf{x}_n}\}$ are assumed to be drawn independently.

The conditional prior distribution on $G_{\mathbf{x}_i}$ given $\{G_{\mathbf{x}_j}, j \neq i\}$ can be characterized by deriving the conditional prior for $\Pi^{\mathbf{x}_i}_{B_1,\ldots,B_L}$ for each collection of disjoint Borel sets $(B_1, \ldots, B_L)$:

$$\Big(\Pi^{\mathbf{x}_i}_{B_1,\ldots,B_L} \,|\, G_{\mathbf{x}_j}, j \neq i\Big) \sim a_{ii}\text{Dirichlet}\Big(\alpha G_0(B_1), \ldots, \alpha G_0(B_L)\Big) + \sum_{j\sim i} a_{ij}\delta_{\left(G_{\mathbf{x}_j}(B_1),\ldots,G_{\mathbf{x}_j}(B_L)\right)},$$

so that the conditional expectation and variance of $G_{\mathbf{x}_i}(B)$, for any Borel set $\mathcal{B}$, are as follows:

$$\begin{aligned}
\mathrm{E}\{G_{\mathbf{x}_i}(B) \,|\, G_{\mathbf{x}_j}, j \sim i\} &= a_{ii}G_0(B) + \sum_{j\sim i} a_{ij}G_{\mathbf{x}_j}(B) \\
\mathrm{Var}\{G_{\mathbf{x}_i}(B) \,|\, G_{\mathbf{x}_j}, j \sim i\} &= a_{ii}G_0(B)\Big(\frac{1+\alpha G_0(B)}{1+\alpha}\Big) + \sum_{j\sim i} a_{ij}G_{\mathbf{x}_j}(B)^2 \\
&\quad -\Big(a_{ii}G_0(B) + \sum_{j\sim i} a_{ij}G_{\mathbf{x}_j}(B)\Big)^2.
\end{aligned}$$

Note that under expression (5), the support of $G_{\mathbf{x}_i}$ with respect to the weak topology is at least as large as the set of all distributions whose support is contained in the support of G_0. This property results directly from properties of the Dirichlet process prior for $G^*_{\mathbf{x}_i}$.

Because the random probability measures for $G_{\mathbf{x}_i}$, $i = 1, \ldots, n$, in (5) are expressed conditionally on $\{G_{\mathbf{x}_j} : j \neq i\}$, it is necessary to prove the existence of an implied joint probability measure for $\mathcal{G}_{\mathbf{X}}$ in order for the specification to be coherent. Theorem 1 establishes existence by demonstrating that the joint measure can be expressed as a finite mixture of independent DPs introduced at each location.

Theorem 1. Let $\mathbf{A}$ denote the $n \times n$ matrix with elements $\{a_{ij}\}_{i,j=1}^n$ satisfying $0 \leq a_{ij} < 1$ and $\mathbf{a}_i'\mathbf{1}_n = 1$ for all row vectors $\mathbf{a}_i'$. Suppose we have

$$\begin{aligned}
G_1 &= a_{11}G^*_1 + a_{12}G_2 + \ldots + a_{1n}G_n \\
G_2 &= a_{21}G_1 + a_{22}G^*_2 + \ldots + a_{2n}G_n \\
\vdots &= \vdots \qquad\qquad \ddots \qquad\qquad \vdots \\
G_n &= a_{n1}G_1 + a_{n2}G_2 + \ldots + a_{nn}G^*_n
\end{aligned}$$

Then, for each $\mathbf{A}$ there exists a corresponding $n \times n$ matrix $\mathbf{B}$, with elements $\{b_{ij}\}_{i,j=1}^{n}$ satisfying $0 \leq b_{ij} \leq 1$ and $\mathbf{b}_i'\mathbf{1}_n = 1$ for all row vectors $\mathbf{b}_i'$, such that

$$G_i = \sum_{j=1}^{n} b_{ij} G_j^*.$$

The proof is in Appendix A. It follows from Theorem 1 and expression (5) that

$$(\phi_i \,|\, \mathbf{x}_i) \sim G_{\mathbf{x}_i} = \sum_{j=1}^{n} b_{ij} G^*_{\mathbf{x}_j}, \quad G^*_{\mathbf{x}_j} \overset{ind}{\sim} DP(\alpha G_0), \quad \text{for } j = 1, \ldots, n, \tag{6}$$

where $\mathbf{b}_i$ is an $n \times 1$ vector of probabilities summing to 1, and the matrix $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_n)' = h(\mathbf{A}) = (\mathbf{I}_n + \mathbf{C})^{-1}\mathbf{D}$, with $\mathbf{A} = \mathbf{C} + \mathbf{D}$, $c_{ij} = a_{ij}$ for $i \neq j$, $c_{ii} = 0$ for $i = 1, \ldots, n$, and $\mathbf{D} = \text{diag}(a_{11}, \ldots, a_{nn})$. Hence, for any model specified as in (5), there is a corresponding model specified as in (6), with the probability weights $\mathbf{B}$ calculated from $\mathbf{A}$ using the simple deterministic function $h(\cdot)$.

Theorem 1 and expression (6) describe the prior for $G_{\mathbf{X}}$. However, for purposes of estimation and prediction at values of $\mathbf{x}$ not represented in the study sample, $\mathbf{X}$, it is of interest to define a prior for $G_{\mathcal{X}}$ consistent with expressions (5) and (6). We establish existence of such a prior in Theorem 2, and then explicitly derive its form.

Theorem 2. There exists an equivalence class of priors for $G_{\mathcal{X}} = \{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$ corresponding to each prior for $G_{\mathbf{X}} = \{G_{\mathbf{x}_i} : i = 1, \ldots, n\}$ specified from expression (5).

To prove Theorem 2, first let $G_{\mathbf{x}} = \sum_{j=1}^{n} b_j(\mathbf{x}) G^*_{\mathbf{x}_j}$ for all $\mathbf{x} \in \mathcal{X}$, where $\mathbf{b}(\mathbf{x}) = [b_1(\mathbf{x}), \ldots, b_n(\mathbf{x})]'$ is chosen subject to $b_j(\mathbf{x}) \geq 0$, $j = 1, \ldots, n$, $\mathbf{b}(\mathbf{x})'\mathbf{1}_n = 1$, for all $\mathbf{x} \in \mathcal{X}$, and $\mathbf{b}(\mathbf{x}_i) = \mathbf{b}_i = (b_{i1}, \ldots, b_{in})'$. These restrictions imply that $G_{\mathbf{x}}$ is a well defined random probability measure satisfying expression (6) and hence expression (5). There are clearly infinitely many different choices of $\mathbf{b}(\mathbf{x})$ satisfying these restrictions, and Theorem 2 follows directly. Thus, we have explicitly described a class of priors for the infinite-dimensional collection of random probability measures $\mathcal{G}_{\mathcal{X}}$, which are consistent with expressions (5) and (6).

Dependency in the random measures contained in $\mathcal{G}_{\mathcal{X}}$ arises through shared dependency on the DP-distributed basis distributions $\mathcal{G}^*_{\mathbf{X}}$. To derive explicit properties of this formulation, it is useful

to re-express (6) in the following hierarchical form:

$$\begin{aligned}(\phi_i \,|\, Z_i = j, \mathbf{x}_i = \mathbf{x}) &\sim G^*_{\mathbf{x}_j}, \\ (Z_i \,|\, \mathbf{x}_i = \mathbf{x}) &\sim \text{Multinomial}(\{1,\ldots,n\}; \mathbf{b}(\mathbf{x})), \\ G^*_{\mathbf{x}_j} &\sim DP(\alpha G_0), j = 1,\ldots,n, \end{aligned} \tag{7}$$

where $Z_i = j$ denotes that ϕ_i was drawn from the jth *basis* distribution, $G^*_{\mathbf{x}_j}$, for $j = 1,\ldots,n$. Hence, the marginal distribution of ϕ_i is represented as a finite mixture of DPs, with $Z_i \in \{1,\ldots,n\}$ indexing the mixture component. Note that this expression holds not only for subjects $i \in \{1,\ldots,n\}$ in the sample, but also for future subjects $i = n+1$ having $\mathbf{x}_{n+1} \notin \mathbf{X}$.

This formulation is useful in deriving properties. It is immediately apparent that

$$\mathrm{E}\{G_{\mathbf{x}_i}(B)\} = \sum_{j=1}^{n} \Pr(Z_i = j)\mathrm{E}\{G^*_{\mathbf{x}_j}(B)\} = \sum_{j=1}^{n} b_j(\mathbf{x}_i)G_0(B) = G_0(B), \tag{8}$$

$$\mathrm{V}\{G_{\mathbf{x}_i}(B)\} = \sum_{j=1}^{n} \left(\frac{b_j(\mathbf{x}_i)^2}{1+\alpha}\right) G_0(B)\{1-G_0(B)\} = \frac{\mathbf{b}(\mathbf{x}_i)'\mathbf{b}(\mathbf{x}_i)G_0(B)\{1-G_0(B)\}}{1+\alpha}, \tag{9}$$

In addition, the dependency between $G_{\mathbf{x}_i}$ and $G_{\mathbf{x}_{i'}}$ can be characterized using Theorem 3.

> **Theorem 3**. For any $\mathbf{x}_i, \mathbf{x}_{i'} \in \mathcal{X}$, including values not represented in $\mathbf{X}$, $G_{\mathbf{x}_i}$ and $G_{\mathbf{x}_{i'}}$ are dependent random probability measures, with $\mathrm{Cor}\{G_{\mathbf{x}_i}(B), G_{\mathbf{x}_{i'}}(B)\} = \rho_{i,i'} = \rho(\mathbf{x}_i, \mathbf{x}_{i'}) = \mathbf{b}(\mathbf{x}_i)'\mathbf{b}(\mathbf{x}_{i'})$ for any Borel set $B \subset \Re$.

The proof of Theorem 3 is in Appendix B. Due to the lack of dependency on B, this expression is particularly useful. In the limiting case as $b_i(\mathbf{x}_i) \to 1$, for $i \in \{1,\ldots,n\}$, $Z_i = i$, $\phi_i \sim G^*_{\mathbf{x}_i}$, $\mathbf{A} = \mathbf{B} = \mathbf{I}_n$, and $\rho_{ij} = \rho(\mathbf{x}_i, \mathbf{x}_j) = 0$ for all $i, j \in \{1,\ldots,n\}$, with $i \neq j$. This special case corresponds to introducing independent DPs at each location.

Note that $\rho(\mathbf{x}_i, \mathbf{x}_{i'})$ is a bounded kernel function, and it may prove useful to consider specifications that rely on choosing a particular kernel instead of explicitly specifying the functions $\mathbf{b}(\mathbf{x})$. Such an approach is conceptually related to the use of reproducing kernel Hilbert spaces (RKHS), which are function spaces defined through choice of a kernel, which implicitly implies a particular set of orthogonal basis functions (Cristianini and Shawer-Taylor, 2000). This avoids the need to specify the basis functions directly, simplifying implementation.

It is useful to derive the prior distribution for ϕ_i given $\boldsymbol{\phi}^{(i)} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_n)'$ and $\mathbf{X}$ marginalizing across the prior for $\mathcal{G}_{\mathcal{X}}$. In the special case in which $\phi_i \sim G_{\mathbf{x}_i} \equiv G \sim DP(\alpha G_0)$, the Pólya urn scheme of Blackwell and MacQueen (1973) prescribes that

$$(\phi_i \,|\, \boldsymbol{\phi}^{(i)}, \mathbf{X}, \alpha) \sim \left(\frac{\alpha}{\alpha + n - 1}\right) G_0 + \left(\frac{1}{\alpha + n - 1}\right) \sum_{j \neq i} \delta_{\phi_j}, \tag{10}$$

which generates new values from $\phi_i \sim G_0$ with probability $\alpha/(\alpha + n - 1)$ and otherwise sets ϕ_i equal to an element of $\boldsymbol{\phi}^{(i)}$ chosen by sampling from a discrete uniform. Derivation of (10) relies on exchangeability of the elements of $\boldsymbol{\phi}$, which no longer holds in the general case.

Relying on the formulation in expression (7), let $\mathcal{I}_j = \{i : Z_i = j\} \subset \{1, \ldots, n\}$ denote an index set for the subjects drawn from the jth mixture component, for $j = 1, \ldots, n$. Then, we have $\phi_i \overset{iid}{\sim} G^*_{\mathbf{x}_j}$ for $i \in \mathcal{I}_j$. Conditioning on the allocation of subjects to mixture components $\mathbf{Z} = (Z_1, \ldots, Z_n)'$, we can use the Pólya urn result to obtain the following conditional prior:

$$\begin{aligned}(\phi_i \,|\, \mathbf{Z}, \boldsymbol{\phi}^{(i)}, \mathbf{X}, \alpha) \sim & \left(\frac{\alpha}{\alpha + \sum_{j \neq i} 1(Z_j = Z_i)}\right) G_0 \\ & + \left(\frac{1}{\alpha + \sum_{j \neq i} 1(Z_j = Z_i)}\right) \sum_{j \neq i} 1(Z_j = Z_i) \delta_{\phi_j}. \end{aligned} \tag{11}$$

Hence, only the subvector of elements of $\boldsymbol{\phi}^{(i)}$ belonging to $\mathcal{I}_{Z_i}$ are informative. Let $M_{ij} = 1(Z_i = Z_j)$ be a 0/1 indicator that subjects i and j belong to the same mixture component. Then, the probability of $\mathbf{M}_i = \{M_{ij}, j \neq i\} = \mathbf{m}_i = \{m_{ij}, j \neq i\}$, for $\mathbf{m}_i \in \{0, 1\}^{n-1}$, is

$$\begin{aligned}\Pr(\mathbf{M}_i = \mathbf{m}_i) &= \sum_{j=1}^{n} \Pr(Z_i = j) \prod_{h \neq i} \Pr(Z_h = j)^{m_{ih}} \{1 - \Pr(Z_h = j)\}^{1 - m_{ih}} \\ &= \sum_{j=1}^{n} b_j(\mathbf{x}_i) \prod_{h \neq i} b_j(\mathbf{x}_h)^{m_{ih}} \{1 - b_j(\mathbf{x}_h)\}^{1 - m_{ih}} = \sum_{j=1}^{n} b_{ij} \prod_{h \neq i} b_{hj}^{m_{ih}} (1 - b_{hj})^{1 - m_{ih}} \end{aligned} \tag{12}$$

Marginalizing across the distribution for $\mathbf{M}_i$, we obtain the following generalization of the Blackwell and MacQueen (1973) Pólya urn scheme of expression (10):

$$\begin{aligned}(\phi_i \,|\, \boldsymbol{\phi}^{(i)}, \mathbf{X}, \alpha, \mathbf{B}) \sim & \sum_{h \neq i} \sum_{m_{ih}=0}^{1} \left\{ \sum_{j=1}^{n} b_{ij} \prod_{l \neq i} b_{lj}^{m_{il}} (1 - b_{lj})^{1 - m_{il}} \right\} \\ & \times \left\{ \left(\frac{\alpha}{\alpha + \sum_{l \neq i} m_{il}}\right) G_0 + \left(\frac{1}{\alpha + \sum_{l \neq i} m_{il}}\right) \sum_{l \neq i} m_{il} \delta_{\phi_l} \right\}. \end{aligned} \tag{13}$$

To illustrate this expression, consider the special case in which $n = 4$ and interest is in the conditional distribution of ϕ_i given $\boldsymbol{\phi}^{(i)}$. In this case, we have

m_{i1}	m_{i2}	m_{i3}	$\Pr\{\mathbf{M}_i = (m_{i1}, m_{i2}, m_{i3})\}$	$(\phi_4 \vert \boldsymbol{\phi}^{(i)}, \mathbf{m}_i)$
0	0	0	$\sum_j b_{ij}(1-b_{1j})(1-b_{2j})(1-b_{3j})$	G_0
1	0	0	$\sum_j b_{ij}b_{1j}(1-b_{2j})(1-b_{3j})$	$\left(\frac{\alpha}{\alpha+1}\right)G_0 + \left(\frac{1}{\alpha+1}\right)\delta_{\phi_1}$
0	1	0	$\sum_j b_{ij}(1-b_{1j})b_{2j}(1-b_{3j})$	$\left(\frac{\alpha}{\alpha+1}\right)G_0 + \left(\frac{1}{\alpha+1}\right)\delta_{\phi_2}$
0	0	1	$\sum_j b_{ij}(1-b_{1j})(1-b_{2j})b_{3j}$	$\left(\frac{\alpha}{\alpha+1}\right)G_0 + \left(\frac{1}{\alpha+1}\right)\delta_{\phi_3}$
1	1	0	$\sum_j b_{ij}b_{1j}b_{2j}(1-b_{3j})$	$\left(\frac{\alpha}{\alpha+2}\right)G_0 + \left(\frac{1}{\alpha+2}\right)(\delta_{\phi_1} + \delta_{\phi_2})$
1	0	1	$\sum_j b_{ij}b_{1j}(1-b_{2j})b_{3j}$	$\left(\frac{\alpha}{\alpha+2}\right)G_0 + \left(\frac{1}{\alpha+2}\right)(\delta_{\phi_1} + \delta_{\phi_3})$
0	1	1	$\sum_j b_{ij}(1-b_{1j})b_{2j}b_{3j}$	$\left(\frac{\alpha}{\alpha+2}\right)G_0 + \left(\frac{1}{\alpha+2}\right)(\delta_{\phi_2} + \delta_{\phi_3})$
1	1	1	$\sum_j b_{ij}b_{1j}b_{2j}b_{3j}$	$\left(\frac{\alpha}{\alpha+3}\right)G_0 + \left(\frac{1}{\alpha+3}\right)(\delta_{\phi_1} + \delta_{\phi_2} + \delta_{\phi_3})$

The expression for $(\phi_i \,|\, \boldsymbol{\phi}^{(i)}, \mathbf{X}, \alpha, \mathbf{B})$ is obtained by summing over the distributions in the last column using the probability weights in the fourth column. Let

$$\boldsymbol{\Gamma}_0 = \left(1, \frac{\alpha}{\alpha+1}, \frac{\alpha}{\alpha+2}, \ldots, \frac{\alpha}{\alpha+n-1}\right)', \; \boldsymbol{\Gamma}_1 = \left(\frac{1}{\alpha+1}, \frac{1}{\alpha+2}, \ldots, \frac{1}{\alpha+n-1}\right)',$$

let $\mathbf{p}_{i0} = \mathbf{p}_0(\mathbf{x}_i)$ denote the $n \times 1$ vector of probabilities corresponding to $\Pr(M_{i+} = m \,|\, \mathbf{x}_i)$, for $m = 0, \ldots, n-1$ with $M_{i+} = \sum_{j \neq i} M_{ij}$, and let $\mathbf{p}_{ij} = \mathbf{p}_j(\mathbf{x}_i)$ denote the $(n-1) \times 1$ vector of probabilities corresponding to $\Pr(M_{ij} = 1, M_{i+} = m \,|\, \mathbf{x}_i)$, for $m = 1, \ldots, n-1$. For example, in the special case considered in the above table, letting $p_{000}, p_{100}, p_{010}, p_{001}, p_{110}, p_{101}, p_{011}, p_{111}$ denote the probabilities in column 4, we have $\mathbf{p}_{i0} = (p_{000}, p_{100} + p_{010} + p_{001}, p_{110} + p_{011} + p_{101}, p_{111})'$, $\mathbf{p}_{i1} = (p_{100}, p_{110} + p_{101}, p_{111})'$, $\mathbf{p}_{i2} = (p_{010}, p_{110} + p_{011}, p_{111})'$, and $\mathbf{p}_{i3} = (p_{001}, p_{011} + p_{101}, p_{111})'$. In general, using this notation, we can express (13) as

$$(\phi_i \,|\, \boldsymbol{\phi}^{(i)}, \mathbf{X}, \alpha, \mathbf{B}) = \mathbf{p}_{i0}'\boldsymbol{\Gamma}_0 G_0 + \sum_{j \neq i} \mathbf{p}_{ij}'\boldsymbol{\Gamma}_1 \delta_{\phi_j} = \mathbf{p}_0(\mathbf{x}_i)'\boldsymbol{\Gamma}_0 G_0 + \sum_{j \neq i} \mathbf{p}_j(\mathbf{x}_i)'\boldsymbol{\Gamma}_1 \delta_{\phi_j}, \tag{14}$$

where $\mathbf{p}_{i0}'\mathbf{1}_n = 1$ and $\mathbf{p}_{ij}'\mathbf{1}_{n-1} \leq 1$. This expression is in the form of a weighted average of Blackwell and MacQueen (1973) Pólya urn distributions. To further simplify this expression, we rely on Theorem 4 (proof in Appendix C).

Theorem 4. For every $n \times n$ matrix $\mathbf{B}$, with elements $\{b_{ij}\}_{i,j=1}^n$ satisfying $0 \leq b_{ij} \leq 1$ and $\mathbf{b}_i'\mathbf{1}_n = 1$, there exists a unique $n \times (n-1)$ matrix $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)'$ having row vectors $\mathbf{w}_i' = (w_{i,1}, \ldots, w_{i,i-1}, w_{i,i+1}, \ldots, w_{i,n})$, with $0 \leq w_{ij} \leq 1 \; \forall \; i, j$, satisfying the following system of equations:

$$\mathbf{p}_{i0}'\boldsymbol{\Gamma}_0 = \left(\frac{\alpha}{\alpha + \mathbf{w}_i'\mathbf{1}_{n-1}}\right) \quad \text{and} \quad \mathbf{p}_{ij}'\boldsymbol{\Gamma}_1 = \left(\frac{w_{ij}}{\alpha + \mathbf{w}_i'\mathbf{1}_{n-1}}\right), \; \forall \; j \neq i,$$

for $i = 1, \ldots, n$, where $\mathbf{p}_{i0}, \mathbf{p}_{ij}, j \neq i$, are calculated from $\mathbf{B}$ as described above. In particular, we have $w_{ij} = \alpha \mathbf{p}'_{ij}\boldsymbol{\Gamma}_1 / \mathbf{p}'_{i0}\boldsymbol{\Gamma}_0$, for all i, j.

Hence, from Theorem 4, expression (14) is equivalent to

$$(\phi_i \,|\, \boldsymbol{\phi}^{(i)}, \mathbf{X}, \alpha, \mathbf{B}) = \left(\frac{\alpha}{\alpha + w_{i+}}\right) G_0 + \sum_{j \neq i} \left(\frac{w_{ij}}{\alpha + w_{i+}}\right) \delta_{\phi_j}, \tag{15}$$

where $0 \leq w_{ij} \leq 1$ and $w_{i+} = \sum_{j \neq i} w_{ij} \leq n$. This simple form is both intuitively appealing and computationally-convenient. The Pólya urn conditional distribution in expression (10) is obtained as a special case by letting $w_{ij} = 1$ for all i, j. In general, viewing the w_{ij}'s as weights and examining expression (35), the weight for the jth subject ($j \neq i$) in the conditional distribution for ϕ_i will depend on the relative values of $\mathbf{p}_{ij}$ and $\mathbf{p}_{i0}$. In the limit as $p_{ijm} = \Pr(M_{i+} = m, M_{ij} = 1) \rightarrow p_{i0,m+1} = \Pr(M_{i+} = m)$, for $m = 1, \ldots, n-1$, which implies $\Pr(M_{ij} = 1) \rightarrow 1$, we have $w_{ij} \rightarrow 1$, while in the limit as $p_{ijm} \rightarrow 0$, for $m = 1, \ldots, n-1$, $w_{ij} \rightarrow 0$. Subjects that have a high probability of being assigned to the same mixture component as subject i will be given high weight. In this manner, expression (15) relaxes the exchangeability assumption implicit in expression (10) to incorporate information on the distance between subjects.

In order to derive the conditional distribution of ϕ_i for a new subject $i = n+1$ having a predictor value $\mathbf{x}_{n+1}$ that may or may not be represented in the sample $\mathbf{X}$, we rely on the Theorem:

> **Theorem 5**. For every set of functions $\mathbf{b}(\mathbf{x}) = [b_1(\mathbf{x}), \ldots, b_n(\mathbf{x})]$ satisfying $b_j(\mathbf{x}) \geq 0$ and $\mathbf{b}(\mathbf{x})'\mathbf{1}_n = 1$, for all $\mathbf{x} \in \mathcal{X}$, there exists a bounded kernel function $w(\mathbf{x}, \mathbf{x}_j) \leq 1$, for all $\mathbf{x} \in \mathcal{X}$ and $\mathbf{x}_j \in \mathbf{X}$, such that $w(\mathbf{x}, \mathbf{x}_j) = \alpha \mathbf{p}_j(\mathbf{x})'\boldsymbol{\Gamma}_1 / \mathbf{p}_0(\mathbf{x})'\boldsymbol{\Gamma}_0$.

The proof is straightforward following the same approach as in Theorem 4. Following the approach used in deriving expression (15), we obtain the prior predictive distribution

$$\begin{aligned}
(\phi_{n+1} \,|\, \mathbf{x}_{n+1}, \boldsymbol{\phi}, \mathbf{X}, \alpha) &= \left(\frac{\alpha}{\alpha + \sum_{j=1}^n w(\mathbf{x}_{n+1}, \mathbf{x}_j)}\right) G_0 + \sum_{j=1}^n \left(\frac{w(\mathbf{x}_{n+1}, \mathbf{x}_j)}{\alpha + \sum_{j=1}^n w(\mathbf{x}_{n+1}, \mathbf{x}_j)}\right) \delta_{\phi_j}, \\
&= \left(\frac{\alpha}{\alpha + w_{n+1}}\right) G_0 + \sum_{j=1}^n \left(\frac{w_{n+1,j}}{\alpha + w_{n+1,j}}\right) \delta_{\phi_j},
\end{aligned} \tag{16}$$

where $w_{ij} = w(\mathbf{x}_i, \mathbf{x}_j)$ and $w_i = \sum_{j \neq i} w_{ij}$, for $i = 1, \ldots, n+1$, is used as shorthand.

As discussed above, because good choices of $\mathbf{b}(\mathbf{x})$ are not immediately apparent, it may be more convenient to choose the kernel function $\rho(\mathbf{x}_i, \mathbf{x}_{i'})$. An alternative, which from the above results implies a prior for $G_{\mathcal{X}}$, is to specify the bounded kernel function $w(\mathbf{x}, \mathbf{x}_j)$. Because expressions (15) and (16) form the starting points for posterior computation and predictions, this approach greatly simplifies implementation, resulting in a procedure which is essentially no more difficult to implement than a standard DP mixture model. Note that $w_{ij} = w(\mathbf{x}_i, \mathbf{x}_j)$ $(0 \leq w_{ij} \leq 1)$ provides a standardized measure of similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$, with $w_{ij} = w_{ji}$, $\lim_{\mathbf{x}_j \to \mathbf{x}_i} w_{ij} = 1$, and $w_{ij} \to 0$ monotonically as $d(\mathbf{x}_i, \mathbf{x}_j)$ increases, with $d(\mathbf{x}_i, \mathbf{x}_j)$ a distance measure specific to $w(\cdot)$. Some possibilities for $w(\cdot)$ are described in subsection 2.3.

2.2 *Clustering Properties and Prediction*

Due to the Pólya urn property, the Ferguson Dirichlet process has been widely used not only for density and function estimation but also for clustering. Hence, it is very interesting to consider clustering properties of the class of priors proposed in subsection 2.1, which we refer to as density regression priors (DRPs). First, note that the conditional distribution (15) implies

1. The prior probability of setting $\phi_i = \phi_j$ decreases in proportion to the distance between the predictor values $\mathbf{x}_i$ and $\mathbf{x}_j$.

2. The prior probability of $\phi_i \in \boldsymbol{\phi}^{(i)}$ increases as more neighbors are added that have predictor values $\mathbf{x}_j$ close to $\mathbf{x}_i$.

3. The expected prior probability of $\phi_i \notin \boldsymbol{\phi}^{(i)}$ (i.e., the ith subject belongs to its own cluster) increases in proportion to the hyperparameter α.

Hence, allocation of the n subjects to $k < n$ clusters is controlled both by the hyperparameter α and by how close the subject's predictor values are to one another.

The n subjects will be allocated into $k \leq n$ clusters, with $k << n$ when α is small and $w(\mathbf{x}_i, \mathbf{x}_j) \to 0$ slowly as $d(\mathbf{x}_i, \mathbf{x}_j)$ increases. Because the DP Pólya urn scheme places more weight on $\boldsymbol{\phi}^{(i)}$, the prior for k under (15) will be stochastically larger (given the same α, n) than the prior described by Antoniak (1974) for the DP. More formally, as $w(\mathbf{x}_i, \mathbf{x}_j) \to 0$ more rapidly as $d(\mathbf{x}_i, \mathbf{x}_j)$

increases, the prior for k will increase stochastically between the Antoniak (1974) prior, in the case where $w_{ij} \equiv 1\ \forall i, j$, to $k \sim \delta_n$, in the case where $w_{ij} = 0\ \forall i, j$. Note that subjects are more likely to be assigned to the same cluster if they are located *close* to each other in terms of the distance measure $d(\cdot)$. Individuals at isolated regions of $\mathcal{X}$ with $N_i = \emptyset$ will be assigned $(\phi_i \,|\, \boldsymbol{\phi}^{(i)}, \mathbf{X}, \alpha) \sim G_0$, so that the base parametric mixture model will be used in extrapolating across sparse data regions. This structure addresses the curse of dimensionality problem by allowing larger model deviations in regions with more data.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ denote the $k \leq n$ unique values of $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)'$, and let $\mathbf{S} = (\mathcal{S}_1, \ldots, \mathcal{S}_n)'$ be a vector of indicators denoting the global configuration of subjects to unique values $\boldsymbol{\theta}$, with $\mathcal{S}_i = h$ if $\phi_i = \theta_h$ indexing the location of the ith subject within the $\boldsymbol{\theta}$ vector. Excluding the ith subject, $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta} \setminus \phi_i$ denotes the $k^{(i)}$ unique values of $\boldsymbol{\phi}^{(i)}$ and $\mathbf{S}^{(i)}$ denotes the configuration of subjects $\{1, \ldots, n\} \setminus i$ to these values. Grouping subjects with common values of $\phi_j, j \neq i$, into clusters, expression (15) is equivalent to

$$(\phi_i \,|\, \boldsymbol{\phi}^{(i)}, \mathbf{X}, \alpha) \sim \left(\frac{\alpha}{\alpha + w_i}\right) G_0 + \left(\frac{1}{\alpha + w_i}\right) \sum_{h=1}^{k^{(i)}} w^*_{ih} \delta_{\theta_h^{(i)}}, \tag{17}$$

where $w^*_{ih} = \sum_{j \neq i} 1(\mathcal{S}_j^{(i)} = h)\, w_{ij}$. Similarly, the prior predictive distribution of ϕ_{n+1} is

$$(\phi_{n+1} \,|\, \mathbf{x}_{n+1}, \boldsymbol{\phi}, \mathbf{X}, \alpha) \sim \left(\frac{\alpha}{\alpha + w_{n+1}}\right) G_0 + \left(\frac{1}{\alpha + w_{n+1}}\right) \sum_{h=1}^{k} w^*_{n+1,h} \delta_{\theta_h}, \tag{18}$$

where $w^*_{n+1,h} = \sum_{j=1}^{n} 1(\mathcal{S}_j = h)\, w_{n+1,j}$ The predictive density of y_{n+1} given $\mathbf{x}_{n+1}$, $\boldsymbol{\phi}$ and $\mathbf{X}$ is

$$\begin{aligned}(y_{n+1} \,|\, \mathbf{x}_{n+1}, \boldsymbol{\phi}, \mathbf{X}, \alpha) \sim{}& \left(\frac{\alpha}{\alpha + w_{n+1}}\right) \int f(y_{n+1} \,|\, \mathbf{x}_{n+1}, \phi)\, dG_0(\phi) \\ &+ \left(\frac{1}{\alpha + w_{n+1}}\right) \sum_{h=1}^{k} w^*_{n+1,h} f(y_{n+1} \,|\, \mathbf{x}_{n+1}, \phi = \theta_h).\end{aligned} \tag{19}$$

This density is a mixture of the base parametric model, obtained by integrating $f(y \,|\, \mathbf{x}, \phi)$ across the mixing measure G_0, and a finite mixture of k densities having distinct values for ϕ. Because the probability weights assigned to the different components of the mixture depend on the location of $\mathbf{x}_{n+1} \in \mathcal{X}$, it is clear that the prior allows model deviations to vary systematically with predictors. In addition, as long as there are no discontinuities in $w(\cdot)$, the following continuity property holds:

$$\lim_{\mathbf{x}_{n+1} \to \mathbf{x}_0} (y_{n+1} \,|\, \mathbf{x}_{n+1}, \boldsymbol{\phi}, \mathbf{X}) \stackrel{d}{=} (y \,|\, \mathbf{x}_0, \boldsymbol{\phi}, \mathbf{X}),$$

implying that the predictive densities converge as the predictor values move closer together.

2.3 *Choice of Kernel Function*

Note that the approach relies on the choice of a bounded kernel function, $w(\cdot)$, which impacts the degree of borrowing of information from the neighbors in estimating the density at any particular predictor value, $\mathbf{x}$. The issues involved in choosing $w(\cdot)$ are related to those arising in choosing a kernel function in density estimation (Härdle, 1991), though the mechanics of how the function impacts the density estimator are very different.

In the special case in which $p = 1$, $w(\cdot)$ can be taken to be the standardized kernel of an arbitrary density. For example, a natural choice, which is easily generalized to multiple dimensions ($p > 1$), is the Gaussian kernel, which would result in

$$w_{ij} = w(x_i, x_j) = \exp\Big\{ -\frac{\psi}{2}(x_i - x_j)^2 \Big\}, \tag{20}$$

where $\psi^{-1/2}$ is a pre-specified bandwidth parameter. In this case, all subjects would technically be in the same neighborhood, though for small to moderate $\psi^{-1/2}$, $w_{ij} \to 0$ rapidly as $|x_i - x_j|$ increases. Another possibility would be the kernel of a triangular distribution

$$w_{ij} = w(x_i, x_j) \quad = \quad \begin{cases} 1 - \psi^{1/2}|x_i - x_j| & \text{for } |x_i - x_j| \leq \psi^{-1/2} \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

where any two subjects, i and j, having predictor values within $\pm\psi^{-1/2}$ are in the same neighborhood, and within the neighborhood w_{ij} decreases linearly with $|x_i - x_j|$.

For categorical predictors, $x_i \in \mathcal{X} = \{1, \ldots, C\}$, it may be more natural to let

$$w_{ij} = w(x_i, x_j) = \begin{cases} 1 & \text{if } x_i = x_j \\ \psi & \text{if } |x_i - x_j| = 1 \end{cases} \tag{22}$$

which assigns a weight of one to subjects with the same predictor value and a weight of $0 \leq \psi \leq 1$ to subjects with a predictor value differing by one unit (can be extended to assign a lower weight to subjects differing by two units, etc). It is interesting to compare the resulting procedure to the DDP of MacEachern (1999; 2001). The DDP allows for dependency in the distributions at $x = c$ and $x = c + 1$ by assuming the distributions have the same number of atoms and allowing dependency in these atoms. In contrast, under our proposed approach, the distributions share a

subset of their atoms, but the number of atoms can change to accommodate evolving deviations from the base measure.

Although an arbitrary choice of $w(\cdot)$ can be used without complications in implementation, we focus on a multivariate Gaussian kernel for ease in generalization to $p > 1$ cases:

$$w_{ij} = w(\mathbf{x}_i, \mathbf{x}_j) = \exp\Big\{ -\frac{\psi}{2}(\mathbf{x}_1 - \mathbf{x}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_1 - \mathbf{x}_2)\Big\}, \tag{23}$$

where $\boldsymbol{\Sigma}$ is taken to be the empirical covariance matrix of $\mathbf{X}$. By plugging in the empirical covariance matrix, prior elicitation is simplified and only the value of the smoothing parameter ψ needs to be specified. In a variety of simulated and real data examples, we have found $\psi = 1/4$ to provide a reasonable default choice, though sensitivity analyses should be conducted in practice.

3. Posterior Computation

3.1 *Gibbs Sampling Algorithm for Mixture Parameters*

For posterior computation, we recommend an adaptation of the Polyá urn Gibbs sampler (MacEachern, 1994; West et al., 1994; Escobar and West, 1998, among others) developed for Dirichlet process mixture models. The Pólya urn scheme has been widely used in implementing Gibbs sampling for DP mixture models. For related approaches for species sampling models, refer to Ishwaran and James (2001; 2003). To improve mixing of the Markov chain, our algorithm updates the cluster-specific parameters, $\boldsymbol{\theta}$, separately from the cluster membership indicators, $\mathbf{S}$, and number of clusters, k. We initially assume that the base parametric mixing measure, G_0, is known, though generalizations will be considered.

Note that the full conditional posterior distribution of ϕ_i can be derived as follows:

$$(\phi_i \,|\, \boldsymbol{\phi}^{(i)}, \mathbf{y}, \mathbf{X}) \quad \propto \quad q_{i,0} G_{i,0} + \sum_{h=1}^{k^{(i)}} q_{i,h} \delta_{\theta_h^{(i)}}, \tag{24}$$

where the posterior obtained by updating the prior $G_0(\phi)$ with the likelihood $f(y_i \,|\, \mathbf{x}_i, \phi)$ is

$$G_{i,0}(\phi) = \frac{G_0(\phi)\, f(y_i \,|\, \mathbf{x}_i, \phi)}{\int f(y_i \,|\, \mathbf{x}_i, \phi)\, dG_0(\phi)} = \frac{G_0(\phi)\, f(y_i \,|\, \mathbf{x}_i, \phi)}{h_i(y_i \,|\, \mathbf{x}_i)},$$

$q_{i,0} = c\,\alpha\, h_i(y_i \,|\, \mathbf{x}_i)$, $q_{i,h} = c\, w^*_{ih} f(y_i \,|\, \mathbf{x}_i, \theta_h)$, and c is a normalizing constant.

Instead of sampling directly from (24) in implementing Gibbs sampling, we alternate between (1) updating the configuration of subjects to clusters, $\mathbf{S}$, and the number of clusters, k; and (2) updating the cluster-specific parameters $\boldsymbol{\theta}$:

1. To update $\mathbf{S}$ and k, we sequentially sample from the full conditional posterior distributions of $\mathcal{S}_i$, for $i = 1, \ldots, n$. From expression (24), it follows that

$$\Pr(\mathcal{S}_i = h \,|\, \boldsymbol{\phi}^{(i)}, \mathbf{y}, \mathbf{X}) = q_{ih}, \quad \text{for } h = 0, 1, 2, \ldots, k, \tag{25}$$

   Whenever $\mathcal{S}_i = 0$, ϕ_i has a value different from those in the existing clusters and we generate this value from $G_{i,0}$.

2. To update the parameters $\boldsymbol{\theta}$ conditional on $\mathbf{S}$ and k, we sample θ_h, for $h = 1, \ldots, k$, from the full conditional posterior distribution:

$$(\theta_h \,|\, \boldsymbol{\theta}^{(h)}, \mathbf{S}, k, \mathbf{y}, \mathbf{X}) \propto \Big\{ \prod_{i:\mathcal{S}_i = h} f(y_i \,|\, \mathbf{x}_i, \theta_h) \Big\} G_0(\theta_h). \tag{26}$$

These steps can be incorporated within an MCMC algorithm that also has steps for updating additional model unknowns and latent variables within a larger hierarchical model.

Because $f(y_i \,|\, \mathbf{x}_i) = \int f(y_i \,|\, \mathbf{x}_i, \phi_i)\, dG_0(\phi_i)$ is a parametric mixture model for the conditional distribution of y_i given $\mathbf{x}_i$, there will typically be unknown parameters characterizing G_0. Extending our formulation to explicitly allow for dependency of G_0 on parameters $\boldsymbol{\gamma}$, we let $\pi(\boldsymbol{\gamma})$ denote the hyperprior distribution for $\boldsymbol{\gamma}$. Then, from the above formulation, it follows that the full conditional posterior distribution of $\boldsymbol{\gamma}$ can be expressed as:

$$(\boldsymbol{\gamma} \,|\, \boldsymbol{\phi}, \mathbf{y}, \mathbf{X}) \propto \pi(\boldsymbol{\gamma}) \Big\{ \prod_{h=1}^{k} G_0(\theta_h; \boldsymbol{\gamma}) \Big\}. \tag{27}$$

For mixtures of normal linear regression models, expressions (25) - (27) have simple closed forms, as described in Section 3.2.

3.2 *Mixtures of Normal Linear Models*

It is interesting to consider the simple special case in which

$$f(y_i \,|\, \mathbf{x}_i, \phi_i) \;\; = \;\; (2\pi\sigma^2)^{-1/2} \exp\Big\{ -\frac{1}{2\sigma^2} (y_i - \mathbf{x}_i' \boldsymbol{\beta}_i)^2 \Big\},$$

so that $f(y_i \,|\, \mathbf{x}_i)$ is characterized by a nonparametric mixture of normal linear regression models. In this case, we fix the normal residual variance, but allow the regression coefficients to vary by letting $\phi_i = (\boldsymbol{\beta}_i', \sigma^{-2})'$ and $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})'$. It is straightforward to modify the approach to allow σ^{-2} to also vary with i, but we focus on the simpler case for ease in presentation. Also, because we are allowing the mixture distribution to change with $\mathbf{x} \in \mathcal{X}$, the approach is very flexible even assuming fixed variance.

To complete a Bayesian specification of the model, the error precision $\tau = \sigma^{-2}$ is assigned a gamma prior, $\pi(\tau) = \mathcal{G}(\tau; a_\tau, b_\tau)$, and we choose a multivariate normal for the base parametric mixture distribution, $G_0(\boldsymbol{\beta}_i; \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta) = \mathrm{N}_p(\boldsymbol{\beta}_i; \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta)$. For additional flexibility, we choose hyperprior distributions for $\boldsymbol{\gamma} = \{\boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta\}$, the parameters characterizing G_0. In particular, let $\pi(\boldsymbol{\gamma}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{\Sigma}_\beta)$, $\pi(\boldsymbol{\beta}) = \mathrm{N}_p(\boldsymbol{\beta}; \boldsymbol{\beta}_0, V_{\beta_0})$ and $\pi(\boldsymbol{\Sigma}_\beta^{-1}) = \mathcal{W}(\boldsymbol{\Sigma}_\beta^{-1}; (\nu_0 \boldsymbol{\Sigma}_0)^{-1}, \nu_0)$, the Wishart density with degrees of freedom ν_0 and expectation $\boldsymbol{\Sigma}_0^{-1}$.

The conditional probabilities in expression (25) can be calculated plugging in:

$$h_i(y_i \,|\, \mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta, \tau) = \frac{\mathrm{N}_p(\mathbf{0}; \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta)\, N(0; y_i, \tau^{-1})}{\mathrm{N}_p(\mathbf{0}; \widehat{\boldsymbol{\beta}}_i, \widehat{V}_{\beta_i})},$$

for $h_i(y_i \,|\, \mathbf{x}_i)$, where $\widehat{V}_{\beta_i} = (\boldsymbol{\Sigma}_\beta^{-1} + \tau \mathbf{x}_i \mathbf{x}_i')^{-1}$ and $\widehat{\boldsymbol{\beta}}_i = \widehat{V}_{\beta_i}(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} + \tau y_i \mathbf{x}_i)$. In addition, letting $\theta_h = \boldsymbol{\beta}_h$, the value of $\boldsymbol{\beta}_i$ for subjects in the hth cluster, expression (26) simplifies to

$$(\boldsymbol{\beta}_h \,|\, \boldsymbol{\beta}^{(h)}, \mathbf{S}, k, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta, \tau, \mathbf{y}, \mathbf{X}) \sim \mathrm{N}_p(\boldsymbol{\beta}_h; \widehat{\boldsymbol{\beta}}_h, \widehat{V}_{\beta_h}), \tag{28}$$

where $\widehat{V}_{\beta_h} = (\boldsymbol{\Sigma}_\beta^{-1} + \tau \sum_{i:\mathcal{S}_h=1} \mathbf{x}_i \mathbf{x}_i')^{-1}$ and $\widehat{\boldsymbol{\beta}}_h = \widehat{V}_{\beta_h}(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} + \tau \sum_{i:\mathcal{S}_h=1} \mathbf{x}_i y_i)$. The full conditional posterior distributions of the remaining unknowns can be expressed as follows:

$$(\tau \,|\, \mathbf{S}, k, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta, \mathbf{y}, \mathbf{X}) \;\sim\; \mathcal{G}\Big(a_\tau + \frac{n}{2}, b_\tau + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta}_i)^2\Big), \tag{29}$$

$$(\boldsymbol{\beta} \,|\, \mathbf{S}, k, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k, \tau, \boldsymbol{\Sigma}_\beta, \mathbf{y}, \mathbf{X}) \;\sim\; \mathrm{N}_p(\widehat{\boldsymbol{\beta}}, \widehat{V}_\beta), \tag{30}$$

$$(\boldsymbol{\Sigma}_\beta^{-1} | \mathbf{S}, k, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k, \tau, \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) \;\sim\; \mathcal{W}\Big(\Big\{\sum_{h=1}^{k}(\boldsymbol{\beta}_h - \boldsymbol{\beta})(\boldsymbol{\beta}_h - \boldsymbol{\beta})' + \nu_0 \boldsymbol{\Sigma}_0\Big\}^{-1}, k + \nu_0\Big) \tag{31}$$

where $\widehat{V}_\beta = (V_{\beta_0}^{-1} + k\boldsymbol{\Sigma}_\beta^{-1})^{-1}$ and $\widehat{\boldsymbol{\beta}} = \widehat{V}_\beta(V_{\beta_0}^{-1}\boldsymbol{\beta}_0 + \sum_{h=1}^{k} \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta}_h)$. Gibbs sampling can proceed by sequentially sampling from (25), which follows a multinomial closed form in this case, and (28)-(31). This algorithm is applied in Sections 4 and 5 to simulated and real data examples.

In the mixture of normal linear regression models case, the conditional predictive density of a future observation y_{n+1} from a subject with predictors $\mathbf{x}_{n+1}$ can be expressed as follows:

$$\begin{aligned}(y_{n+1} \,|\, \mathbf{x}_{n+1}, \boldsymbol{\phi}, \mathbf{y}, \mathbf{X}) \;&\sim\; \left(\frac{\alpha}{\alpha + w_{n+1}}\right) \int f(y_{n+1} \,|\, \mathbf{x}_{n+1}, \boldsymbol{\beta}_{n+1}, \tau)\, dG_0(\boldsymbol{\beta}_{n+1}; \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta) \\ &\quad + \left(\frac{1}{\alpha + w_{n+1}}\right) \sum_{h=1}^{k} w^*_{n+1,h} f(y_{n+1} \,|\, \mathbf{x}_{n+1}, \boldsymbol{\beta}_h, \tau) \\ &\stackrel{d}{=}\; \left(\frac{\alpha}{\alpha + w_{n+1}}\right) N(y_{n+1}; \mathbf{x}'_{n+1}\boldsymbol{\beta}, \tau^{-1} + \mathbf{x}'_{n+1}\boldsymbol{\Sigma}_\beta\, \mathbf{x}_{n+1}) \\ &\quad + \left(\frac{1}{\alpha + w_{n+1}}\right) \sum_{h=1}^{k} w^*_{n+1,h} \mathrm{N}(y_{n+1}; \mathbf{x}'_{n+1}\boldsymbol{\beta}_h, \tau^{-1}) \qquad (32)\end{aligned}$$

which is a finite mixture of normals. Note that large values of α will lead to a high degree of shrinkage towards the first normal component. The first component will also receive high probability weight when there are few subjects in the data set close to $\mathbf{x}_{n+1}$, because at such locations $w_{n+1} \approx 0$. For the remaining k normal components, each of which has a distinct set of regression coefficients, the weights will depend on the number of subjects in the data set that have predictor values close to $\mathbf{x}_{n+1}$ and are allocated to that component. In this manner, the weights on the different components are spatially adaptive according to the location of $\mathbf{x}_{n+1} \in \mathcal{X}$. In addition, the number of components k is treated as unknown and will change across the MCMC samples.

Interest commonly focuses on estimating the predictive density of y_{n+1} for a variety of $\mathbf{x}_{n+1}$ values, possibly to investigate changes in the density across $\mathcal{X}$. To remove the conditioning on the unknowns, $\mathbf{S}, k, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k, \tau, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta$, one can calculate the expected predictive density averaging over the posterior distribution by using a large number of iterates collected after apparent convergence of the Gibbs sampling algorithm. In particular, letting $t = 1, \ldots, T$ index the iteration number, one can use the estimator:

$$\begin{aligned}\widehat{f}(y_{n+1} \,|\, \mathbf{x}_{n+1}) = \frac{1}{T}\Bigg[&\sum_{t=1}^{T} \left(\frac{\alpha}{\alpha + w_{n+1}}\right) N(y_{n+1}; \mathbf{x}'_{n+1}\boldsymbol{\beta}^{(t)}, {\tau^{(t)}}^{-1} + \mathbf{x}'_{n+1}\boldsymbol{\Sigma}_\beta^{(t)}\, \mathbf{x}_{n+1}) \\ &+ \left(\frac{1}{\alpha + w_{n+1}}\right) \sum_{h=1}^{k(t)} w^*_{n+1,h} \mathrm{N}(y_{n+1}; \mathbf{x}'_{n+1}\boldsymbol{\beta}_h^{(t)}, {\tau^{(t)}}^{-1}) \Bigg]. \qquad (33)\end{aligned}$$

Note that this estimator is defined for a particular y_{n+1} value. In practice, one can estimate the density at a dense grid of possible y_{n+1} values.

4. Simulation Examples

In order to assess the computational performance of the Gibbs sampling algorithm and whether the approach seems to give reasonable results, we analyzed data from a series of simulated examples. Although we also considered sample sizes of $n = 200$ and $n = 1000$, we focus on $n = 500$ for sake of brevity (other cases yielded similar results). We consider $p = \{2, 3\}$, with $\mathbf{x}_i = (1, x_{i2}, x_{i3})'$ and x_{i2}, x_{i3} simulated from independent uniform distributions. For the hyperparameters, we let $\alpha = \{0.5, 1\}$, $\boldsymbol{\beta}_0 = \mathbf{0}$, $V_{\beta_0} = (\mathbf{X}'\mathbf{X})^{-1}/n$, $\nu_0 = p$, $\boldsymbol{\Sigma}_0^{-1} = \mathrm{I}_{p\times p}$, and $a_\tau = b_\tau = 0.1$. We considered a range of values for ψ, $\psi \in \{0.125, 0.25, 0.5\}$, using weight function (23). In the primary analysis, $\alpha = 0.5$ and $\psi = 0.25$.

As a null case, we first simulated data under the normal linear regression model,

$$f(y_i \,|\, \mathbf{x}_i) = \mathrm{N}(y_i; -1 + 2x_{i2}, 0.01)$$

[trying also a variety of alternative values for the true regression function and error variance]. We then analyzed the simulated data using the proposed Gibbs sampling algorithm run for 10,000 iterations with a 1,000 iteration burn-in. Based on examination of trace plots for the different unknowns, convergence was rapid and mixing was efficient. In each case, the predictive mean regression function closely approximated the true linear regression function, which was entirely enclosed in pointwise 99% credible intervals. In addition, estimates of the predictive density of y_{n+1} at the 10th, 25th, 50th, 75th, and 90th percentiles of the empirical distribution of x_{i2} were essentially indistinguishable from the true normal density. These results held regardless of the value of ψ, though estimated predictive mean curves and densities were slightly bumpier for $\psi = 0.125$.

As a more interesting case, we simulated data from a mixture of two normal linear regression models, with the mixture weights depending on the predictor, with the error variance differing, and with a non-linear mean function for the second component:

$$f(y_i \,|\, \mathbf{x}_i) = e^{-2x_{i2}} N(y_i; x_{i2}, 0.01) + (1 - e^{-2x_{i2}}) N(y_i; x_{i2}^4, 0.04).$$

Figure 1 shows the true density (dotted line), estimated predictive density (solid line), and pointwise 99% credible intervals (dashed lines) for a range of values of x_{i2}. The estimates correspond

approximately to the true densities in each case. The bottom right panel contains an $x - y$ plot of the data along with the estimated predictive mean curve (solid line), which is indistinguishable from the true mean curve (dotted line). Essentially identical results were obtained for analyses with $\psi = 0.125$, $\psi = 0.5$, and $\alpha = 1$. Repeating the analysis as described above, but with $\phi_i \overset{iid}{\sim} G$ and $G \sim DP(\alpha G_0)$, we obtained poor results (density estimates diverged substantially from true densities, posterior mean curve failed to capture true non-linear function), suggesting that a DP mixture model is inadequate.

5. Application: Epidemiologic Study

5.1 *Data Structure and Scientific Problem*

The methods are applied to a study of reproductive hormones and obesity. Study participants were premenopausal 35-50 year old women randomly selected from the membership list of a Washington, DC health plan. Luteinizing hormone (LH) was measured in urine collected by the women on the first or last five days of the menstrual cycle to avoid mid-cycle variability due to the rapid rise in LH at the time of ovulation. Appropriately-timed urine samples assayed for LH and a current body mass index (BMI) were available for 522 women.

An association between LH and BMI would be interesting for several reasons. First, there is growing evidence that LH has a proliferative effect on uterine smooth muscle cells, possibly leading to fibroid growth. An abnormally elevated LH level among obese women may indicate a greater risk of developing fibroids, a common reproductive tract tumor which leads to substantial morbidity in the U.S. On the other hand, LH also has a critical role in ovulation and menstrual cycling, and abnormally low LH levels may indicate reproductive dysfunction. Hence, it is interesting to assess how the distribution of BMI changes as LH changes. Of course, it is important to adjust for the potentially confounding effect of age.

We do not expect the distribution of BMI among women of a given age with a particular value of LH to be normally distributed, and there is likely to be some degree of positive skewness. In addition, given the above biological considerations, it seems plausible that the shape of the BMI density may change as LH changes, with a possible differential effect for the more obese women in

the right tail of the distribution. Hence, the density regression approach proposed in this article seems ideal for these data.

5.2 *Univariate Analysis*

For woman i ($i = 1, \ldots, 522$), let y_i, x_{i2} and x_{i3} denote BMI, LH and age, respectively. Variables were normalized prior to analysis, but transformed back to the original scale in presenting the results. Prior specification and posterior computation proceeded as in Section 4. We initially considered LH as the only predictor. As previously, samples appeared to converge rapidly to a stationary distribution and mixing was good. Figure 2 presents the estimated predictive density of BMI for LH values corresponding to the 1, 10, 25, 50, 75 and 90th percentiles of the empirical distribution. As expected, the BMI densities tend to be right-skewed. Interestingly, the distributions are more highly skewed, with a greater proportion of morbidly obese women (BMI > 40), when LH values are low. In fact, there is even evidence of a second mode at high BMI values when LH is low.

These results suggest that obese women, particularly morbidly obese women, tend to have low LH levels. This goes against the hypothesis that obese women may be at greater risk of uterine smooth muscle cell proliferation and fibroid development due to increased LH. However, it is consistent with our secondary hypothesis that obese women may have diminished reproductive functioning, which is manifest by low LH levels. The raw LH and BMI data are plotted in Figure 3, along with the posterior predictive mean curve and pointwise 99% credible intervals. The second mode at low LH levels, which was picked up by our density regression estimator, is also apparent in the raw data. Overall, there is a decreasing trend in mean BMI with increasing LH, with the nonlinear curve flattening out at higher LH levels. Conclusions were robust to the degree of smoothing. For $\psi = 0.125$, the density estimates were slightly bumpier and for $\psi = 0.5$ the estimates were slightly smoother, but the differences were barely noticeable. The second mode at high BMI for low LH was present in all analyses.

5.3 *Bivariate Analysis*

The above results did not consider age, which is an important predictor of BMI. To adjust for possible confounding, we repeated the analysis including age as an additional predictor. For multivariate predictors, we are faced with the problem of how to summarize changes that occur in the response distribution across the predictor space. We chose to estimate the predictive density of BMI given age and LH for a range of LH values, with age fixed at its sample mean value. We then repeated this exercise to estimate the predictive density for a range of age values with LH fixed at its sample mean. To assess interactions between age and LH, we estimated the predictive densities for the same range of LH values, but with age fixed at a low or high value.

The age main effect is illustrated in Figure 4. The BMI densities have a slight increase in positive skewness between the mid and late 40s as the proportion of morbidly obese women increases. Interestingly, there are minimal changes with age in the mode of the BMI density, and changes in the mean are primarily attributable to an increasing subset of obese and morbidly obese women, an observation not apparent from mean regression curves. The LH main effect results are essentially identical to those from the univariate analysis, so adjustment for age did not impact our conclusions about the relationship between LH and BMI. These results were robust to moderate changes in the choice of ψ and α.

6. Discussion

This article has proposed a Bayesian approach to the density regression problem, relying on a mixture model with a novel prior specification for the mixing measure. In particular, the mixing measure is allowed to vary depending on the location of a multivariate (potentially continuous) predictor $\mathbf{x} \in \mathcal{X}$. The proposed structure has a number of appealing properties, including the availability of a simple weighted Pólya urn-type scheme, which facilitates posterior computation via a simple Gibbs sampling algorithm. Results of simulated and real data applications are promising.

In future research, it will be interesting to consider additional properties of the proposed prior specification and to generalize the approach. One conceptually straightforward extension would be to allow the smoothing parameter, ψ, and precision parameter, α, to vary depending on the location of $\mathbf{x} \in \mathcal{X}$. In addition, because these hyperparameters play such an important role, it would be

interesting to consider methods that allow the data to inform about their values. Other areas in need of additional consideration, include efficient computation in non-conjugate cases (refer to Neal, 2000 for related work in DP mixture models) and approaches allowing full nonparametric Bayesian inference on the collection of unknown distributions (see Gelfand and Kottas, 2002; Ishwaran and James, 2002 for related approaches based on approximating the DP).

Appendix A: Proof of theorem 1

Let $\mathbf{G} = (G_1, G_2, ..., G_n)'$ and $\mathbf{G}^* = (G_1^*, G_2^*, ..., G_n^*)'$. Notice that we have

$$\mathbf{G} = \mathbf{C}\mathbf{G} + \mathbf{D}\mathbf{G}^*$$

where $\mathbf{A} = \mathbf{C} + \mathbf{D}$, with $c_{ij} = a_{ij}$ for $i \neq j$, $c_{ij} = 0$ for $i = j$, and $\mathbf{D}$ a diagonal matrix with $d_{ii} = a_{ii}$. Hence we have

$$(\mathbf{I}_n - \mathbf{C})\mathbf{G} = \mathbf{D}\mathbf{G}^* \Rightarrow \mathbf{G} = (\mathbf{I}_n - \mathbf{C})^{-1}\mathbf{D}\mathbf{G}^*$$

Letting $\mathbf{B} = (\mathbf{I}_n - \mathbf{C})^{-1}\mathbf{D}$, it suffices to prove the following Lemmas:

1. The matrix $(\mathbf{I}_n - \mathbf{C})$ is invertible
2. $\mathbf{B}$ is row stochastic, so that $\mathbf{b}_i'\mathbf{1}_n = 1$ (rows sum to 1)
3. $\mathbf{B}$ has non-negative entries

Lemma 1: $(\mathbf{I}_n - \mathbf{C})$ is invertible

Proof: Let $\tilde{\mathbf{C}} = \mathbf{I}_n - \mathbf{C}$. Then we have

$$\sum_{j=1, j\neq i}^{n} |\tilde{c}_{ij}| = \sum_{j=1, j\neq i}^{n} |c_{ij}| = 1 - c_{ii} < 1 \qquad \forall i \in \{1, 2, ..., n\}$$

Hence, the matrix $\tilde{\mathbf{C}} = \mathbf{I}_n - \mathbf{C}$ is strictly diagonally dominant. Note that a square matrix $\mathbf{S}$ is strictly diagonally dominant if $|s_{ii}| > \sum_{j\neq i} |s_{ij}|, 1 \leq i \leq n$. From Serre (2002, p73), strictly diagonally dominant matrices are invertible, so lemma 1 holds.

Lemma 2: The matrix $\mathbf{B}$ is row stochastic.

Proof: Notice that $\mathbf{B} = (\mathbf{I}_n - \mathbf{C})^{-1}\mathbf{D} = [\mathbf{D}^{-1}(\mathbf{I}_n - \mathbf{C})]^{-1}$, and hence $\mathbf{B} = \tilde{\mathbf{B}}^{-1}$, where $\tilde{\mathbf{B}} = \mathbf{D}^{-1}(\mathbf{I}_n - \mathbf{C})$. Hence we have

$$\tilde{b}_{ij} = \frac{h_{ij}}{c_{ii}}, \quad h_{ij} = 1 \text{ for } \quad i = j, \text{ and } h_{ij} = -c_{ij} \text{ for } \quad i \neq j$$

Hence we have for $i \in \{1, 2, ..., n\}$,

$$\sum_{j=1}^{n} \tilde{b}_{ij} = \frac{1 - \sum_{j=1, j\neq i}^{n} c_{ij}}{c_{ii}} = \frac{c_{ii}}{c_{ii}} = 1.$$

Thus, $\tilde{\mathbf{B}}$ has 1 as an eigenvalue and $\mathbf{1}_n$ as an eigenvector. Because the eigenvectors are preserved during the inverse operation and 1 is an eigenvalue of $\widetilde{\mathbf{B}}^{-1} = \mathbf{B}$, $\mathbf{B}$ is row stochastic.

Lemma 3: $\mathbf{B} = (\mathbf{D}^{-1}(\mathbf{I}_n - \mathbf{C}))^{-1}$ has non negative entries.

Proof: Let $\mathbf{b} = (b_1, b_2, ..., b_n)'$, such that $b_i > 0$ for $i \in \{1, 2, ..., n\}$. Let $\mathbf{x}$ be the solution of the equation, $\mathbf{Bx} = \mathbf{b}$ and $i = \text{argmin } x_i$. Then we have

$$\frac{1}{c_{ii}} x_i = b_i - \sum_{j=1, j\neq i}^{n} \frac{c_{ij}}{c_{ii}} x_j > x_i \left(\sum_{j=1, j\neq i}^{n} \frac{c_{ij}}{c_{ii}} \right)$$

$$\Rightarrow \left(\frac{1}{c_{ii}} - \sum_{j=1, j\neq i}^{n} \frac{c_{ij}}{c_{ii}} \right) x_i > 0 \Rightarrow x_i > 0$$

and since x_i was the minimum, we have $x > 0$. From Serre (2002, page 80), a matrix $\mathbf{S}$ is nonnegative if and only if $x \leq 0$ implies $\mathbf{Sx} \leq 0$. Hence, Lemma 3 follows directly.

Appendix B: Proof of theorem 3

The correlation between $G_{\mathbf{x}_i}(B)$ and $G_{\mathbf{x}_{i'}}(B)$ has the following form:

$$\text{Cor}\{G_{\mathbf{x}_i}(B), G_{\mathbf{x}_{i'}}(B)\} \quad = \quad \frac{\text{E}\{G_{\mathbf{x}_i}(B) G_{\mathbf{x}_{i'}}(B)\} - \text{E}\{G_{\mathbf{x}_i}(B)\}\text{E}\{G_{\mathbf{x}_{i'}}(B)\}}{\Big[\text{V}\{G_{\mathbf{x}_i}(B)\}\text{V}\{G_{\mathbf{x}_{i'}}\}\Big]^{1/2}}.$$

The numerator can be expressed as follows:

$$\begin{aligned}
&\text{E}\Big[\{b_{i1}G^*_{\mathbf{x}_1} + b_{i2}G^*_{\mathbf{x}_2} + \ldots + b_{in}G^*_{\mathbf{x}_n}\}\{b_{i'1}G^*_{\mathbf{x}_1} + b_{i'2}G^*_{\mathbf{x}_2} + \ldots + b_{i'n}G^*_{\mathbf{x}_n}\} - G_0(B)^2 \\
&= \quad \text{E}\left(\Big\{ \sum_{h=1}^{n} b_{ih}b_{i'h}G^*_{\mathbf{x}_h}(B)^2 \Big\} + \Big[\sum_{h=1}^{n} b_{ih}G^*_{\mathbf{x}_h}(B) \Big\{ \sum_{l\neq h} b_{i'l}G^*_{\mathbf{x}_l}(B) \Big\} \Big] \right) - G_0(B)^2
\end{aligned}$$

$$
\begin{aligned}
&= \Big\{ \sum_{h=1}^{n} b_{ih} b_{i'h} \mathrm{E}\{G^*_{\mathbf{x}_h}(B)^2\} \Big\} + \Big\{ \sum_{h=1}^{n} b_{ih}(1 - b_{i'h}) G_0(B)^2 \Big\} - G_0(B)^2 \\
&= \Big(\sum_{h=1}^{n} b_{ih} b_{i'h} \Big[\frac{G_0(B)\{1 - G_0(B)\}}{1+\alpha} + G_0(B)^2 \Big] \Big) + G_0(B)^2 \sum_{h=1}^{n} b_{ih} - G_0(B)^2 \sum_{h=1}^{n} b_{ih} b_{i'h} - G_0(B)^2 \\
&= \Big(\sum_{h=1}^{n} b_{ih} b_{i'h} \Big) \Big[\frac{G_0(B)\{1 - G_0(B)\}}{1+\alpha} \Big]. \qquad (34)
\end{aligned}
$$

Because the term in $[\cdot]$ equals $\mathrm{V}\{G_{\mathbf{x}_h}(B)\}$ for $h = i, i'$, it follows directly that $\mathrm{Cor}\{G_{\mathbf{x}_i}(B), G_{\mathbf{x}_{i'}}(B)\} = \sum_{h=1}^{n} b_{ih} b_{i'h}$.

Appendix C: Proof of theorem 4

Letting $w_{i+} = \mathbf{w}_i'\mathbf{1}_{n-1}$, for $i = 1, \ldots, n$, we first show that there exists a unique vector $\mathbf{w}_+ = (w_{1+}, \ldots, w_{n+})'$ corresponding to $\mathbf{B}$. In particular, the elements of this vector correspond to the solution to the following system of equations:

$$\frac{\alpha}{\alpha + w_{i+}} = p_{i01} + p_{i02}\left(\frac{\alpha}{\alpha+1}\right) + \ldots + p_{i0n}\left(\frac{\alpha}{\alpha+n-1}\right), \quad \text{for } i = 1, \ldots, n.$$

It is straightforward to obtain the simple closed form solution $w_{i+} = \alpha(1 - \mathbf{p}_{i0}'\mathbf{\Gamma}_0)/(\mathbf{p}_{i0}'\mathbf{\Gamma}_0)$, for $i = 1, \ldots, n$, where $\Pr(M_{i+} = \sum_{j \neq i} M_{ij} = m) = p_{i0m}$, for $m = 0, \ldots, n-1$, is the probability mass function for M_{i+}, which can be calculated from $\mathbf{B}$ using expression (12).

Following a similar route to solve for w_{ij}, for all i, j, holding w_{i+} as fixed:

$$\frac{w_{ij}}{\alpha + w_{i+}} = p_{ij1}\left(\frac{1}{\alpha+1}\right) + \ldots + p_{ij,n-1}\left(\frac{1}{\alpha+n-1}\right) = \mathbf{p}_{ij}'\mathbf{\Gamma}_1$$

we obtain $w_{ij} = (\alpha + w_{i+})\mathbf{p}_{ij}'\mathbf{\Gamma}_1 = \alpha\mathbf{p}_{ij}'\mathbf{\Gamma}_1/\mathbf{p}_{i0}'\mathbf{\Gamma}_0$. It remains to show $0 \leq \alpha\mathbf{p}_{ij}'\mathbf{\Gamma}_1/\mathbf{p}_{i0}'\mathbf{\Gamma}_0 \leq 1$. Letting $R_{ij} = \mathbf{p}_{ij}'\mathbf{\Gamma}_1/\mathbf{p}_{i0}'\mathbf{\Gamma}_0$, we have

$$\begin{aligned} R_{ij} &= \frac{p_{ij1}\frac{1}{\alpha+1} + \ldots + p_{ij,n-1}\frac{1}{\alpha+n-1}}{p_{i01} + p_{i02}\frac{\alpha}{\alpha+1} + \ldots + p_{i0n}\frac{\alpha}{\alpha+n-1}} \\ &= \frac{0 \times \frac{1}{\alpha} + p_{ij1}\frac{1}{\alpha+1} + \ldots + p_{ij,n-1}\frac{1}{\alpha+n-1}}{\alpha\left[p_{i01}\frac{1}{\alpha} + p_{i02}\frac{1}{\alpha+1} + \ldots + p_{i0n}\frac{1}{\alpha+n-1}\right]} \end{aligned}$$

Thus, letting $\tilde{\mathbf{p}}_{ij} = (0, \mathbf{p}_{ij}')'$, $w_{ij} = \alpha R_{ij}$ can be expressed as

$$w_{ij} = \frac{\sum_{m=1}^{n} \tilde{p}_{ijm}\left(\frac{1}{\alpha+m-1}\right)}{\sum_{m=1}^{n} p_{i0m}\left(\frac{1}{\alpha+m-1}\right)}. \tag{35}$$

Recalling that $p_{i0m} = \Pr(M_{i+} = m-1)$ while $p_{ijm} = \Pr(M_{i+} = m, M_{ij} = 1)$, we have $p_{i0,m+1} \geq p_{ijm}$, for $m = 1, \ldots, n-1$, which implies that $p_{i0m} \geq \tilde{p}_{ijm}$, for $m = 1, \ldots, n$. It follows that $\sum_{m=1}^{n} \tilde{p}_{ijm}/(\alpha+m-1) \leq \sum_{m=1}^{n} p_{i0m}/(\alpha+m-1)$. Hence, $0 \leq w_{ij} \leq 1$.

References

Antoniak, C.E. (1974) Mixtures of Dirichlet processes with application to nonparametric problems. *The Annals of Statistics*, **2**, 1152-1174.

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

Cifarelli, D., and Regazzini, E. (1978) Problemi statistici non parametrici in condizioni di scambialbilita parziale e impiego di medie associative. Technical Report, Quaderni Istituto Matematica Finanziaria, Torino.

Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK.

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.

Dunson, D.B. (2004) Semiparametric Bayesian latent response models. *ISDS Working Paper Series*, **04-10**, Duke University.

Escobar, M.D. (1994) Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, **89**, 268-277.

Escobar, M.D. and West, M. (1998) Computing nonparametric hierarchical models," In *Practical Nonparametric and Semiparametric Bayesian Statistics* (eds. D. Dey, P. Müller and D. Sinha), pp. 1-22. New York: Springer-Verlag.

Ferguson, T.S. (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**, 615-629.

Gelfand, A.E. and Kottas, A. (2001) Nonparametric Bayesian modeling for stochastic order. *Annals of the Institute of 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., Kottas, A., and MacEachern, S.N. (2004) Bayesian nonparametric spatial modeling using dependent Dirichlet processes. *Journal of the American Statistical Association*, to appear.

Giudici, P., Mezzetti, M., and Muliere, P. (2003), "Mixtures of Dirichlet Process Priors for Variable Selection in Survival Analysis," *Journal of Statistical Planning and Inference*, 111, 101-115.

Griffin, J.E. and Steel, M.F.J. (2004), "Semiparametric Bayesian inference for stochastic frontier models," *Journal of Econometrics*, **123**, 121-152.

Griffin, J.E. and Steel, M.F.J. (2005), "Order-based dependent Dirichlet processes," *Journal of the American Statistical Association*, to appear.

Härdle, W. (1991) *Smoothing Techniques: With Implementation in S*. Springer Verlag.

Ishwaran, H. and James, L.F. (2002) Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information. *Journal of Computational and Graphical Statistics*, **11**, 508-532.

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

Ishwaran, H. and James, L.F. (2003) Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, **13**, 1211-1235.

MacEachern, S.N. (1994) Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation*, **23**, 727-741.

MacEachern, S.N. (1999) Dependent Nonparametric Processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association.

MacEachern, S.N. (2000) Dependent Dirichlet processes. Unpublished manuscript, Department of Statistics, The Ohio State University.

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: ISBA, pp. 551-560.

Mira, A. and Petrone, S. (1996) Bayesian hierarchical nonparametric inference for change-point problems. In *Bayesian Statistics* 5 (eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith), Oxford: Oxford University Press.

Muliere, P. and Petrone, S. (1993) A Bayesian predictive approach to sequential search for an optimal dose: parametric and nonparametric models. *Journal of the Italian Statistical Society*, **2**, 349-364.

Müller, P., Erkanli, A. and West, M. (1996), "Bayesian Curve Fitting using Multivariate Normal Mixtures," *Biometrika*, 83, 67-79.

Müller, P. and Quintana, F.A. (2004) Nonparametric Bayesian Data Analysis. *Statistical Science*, **19**, 95-110.

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

Pennell, M. and Dunson, D.B. (2004) Bayesian semiparametric dynamic frailty models for multiple event time data. *ISDS Working Paper Series*, **04-27**, Duke University.

Pitman, J. (1995) Exchangeable and partially exchangeable random partitions. *Probability Theory Related Fields*, **102**, 145-158.

Pitman, J. (1996) Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory* (eds T.S. Ferguson, L.S. Shapley and J.B. MacQueen), pp. 245-267. IMS Lecture Notes-Monograph Series, Vol. **30**.

Serre, D. (2002). Matrices: Theory and Application. New York: Springer-Verlag.

Sethuraman, J. (1994) A constructive definition of the Dirichlet process prior. *Statistica Sinica*, **2**, 639-650.

West, M., Müller, P. and Escobar, M.D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. In *A Tribute to D. V. Lindley* (A.F.M. Smith and P.R. Freeman). John Wiley and Sons.

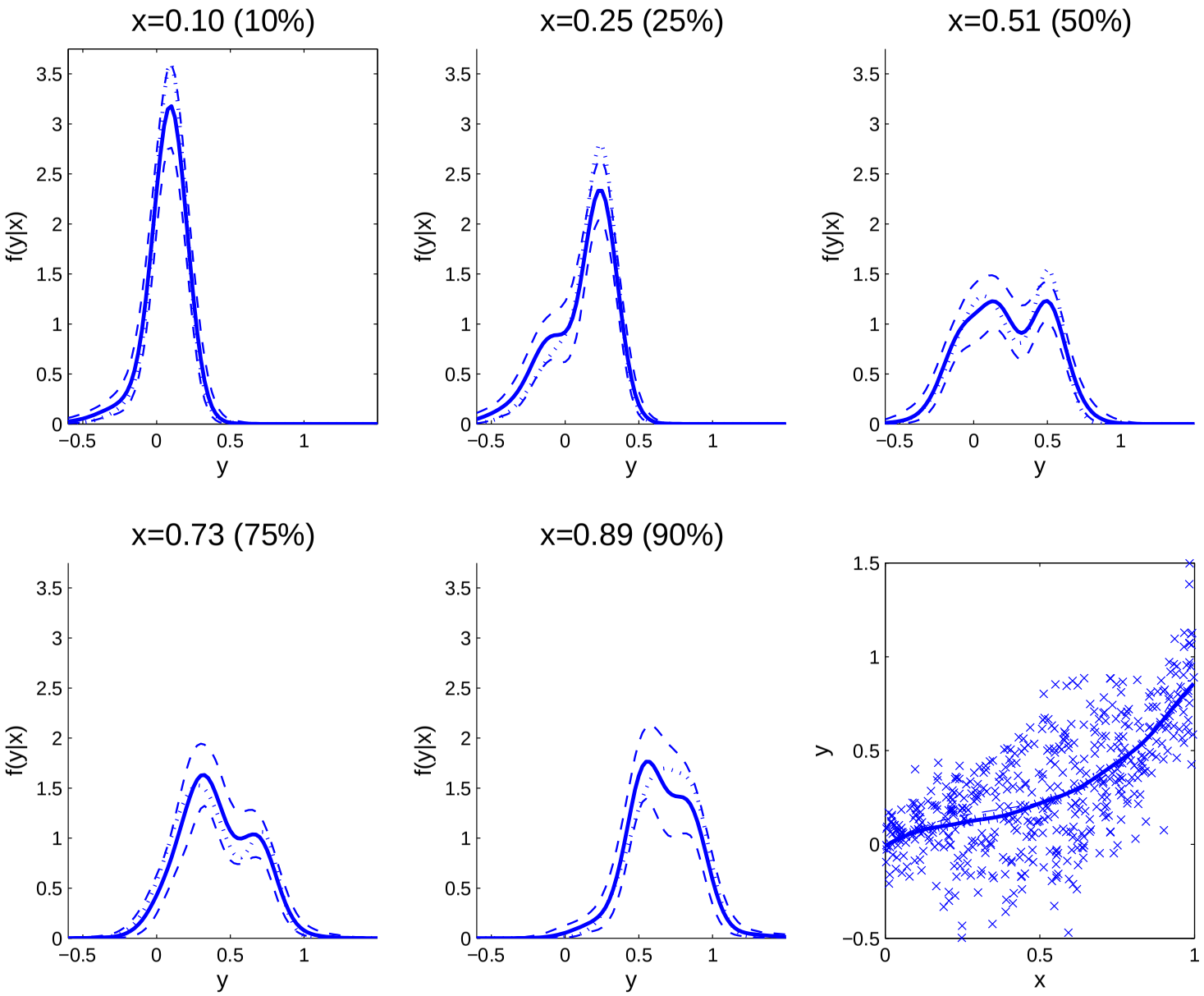
Figure Captions

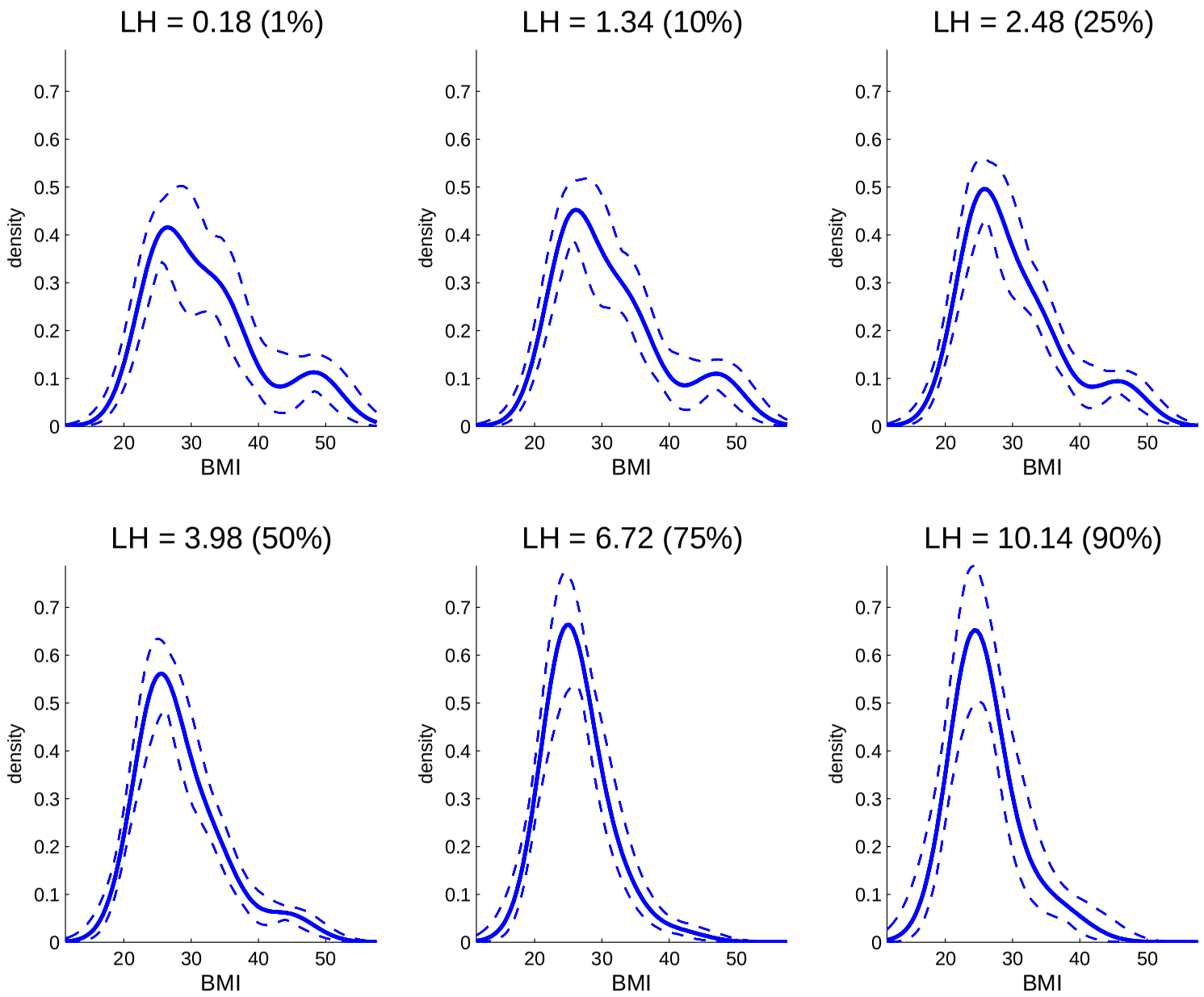
1. True conditional densities of $y|x$ (dotted lines) and estimated predictive densities $\hat{f}(y \mid x)$ (solid lines) for a range of x values in the simulation example. Dashed lines correspond to 99% pointwise credible intervals.

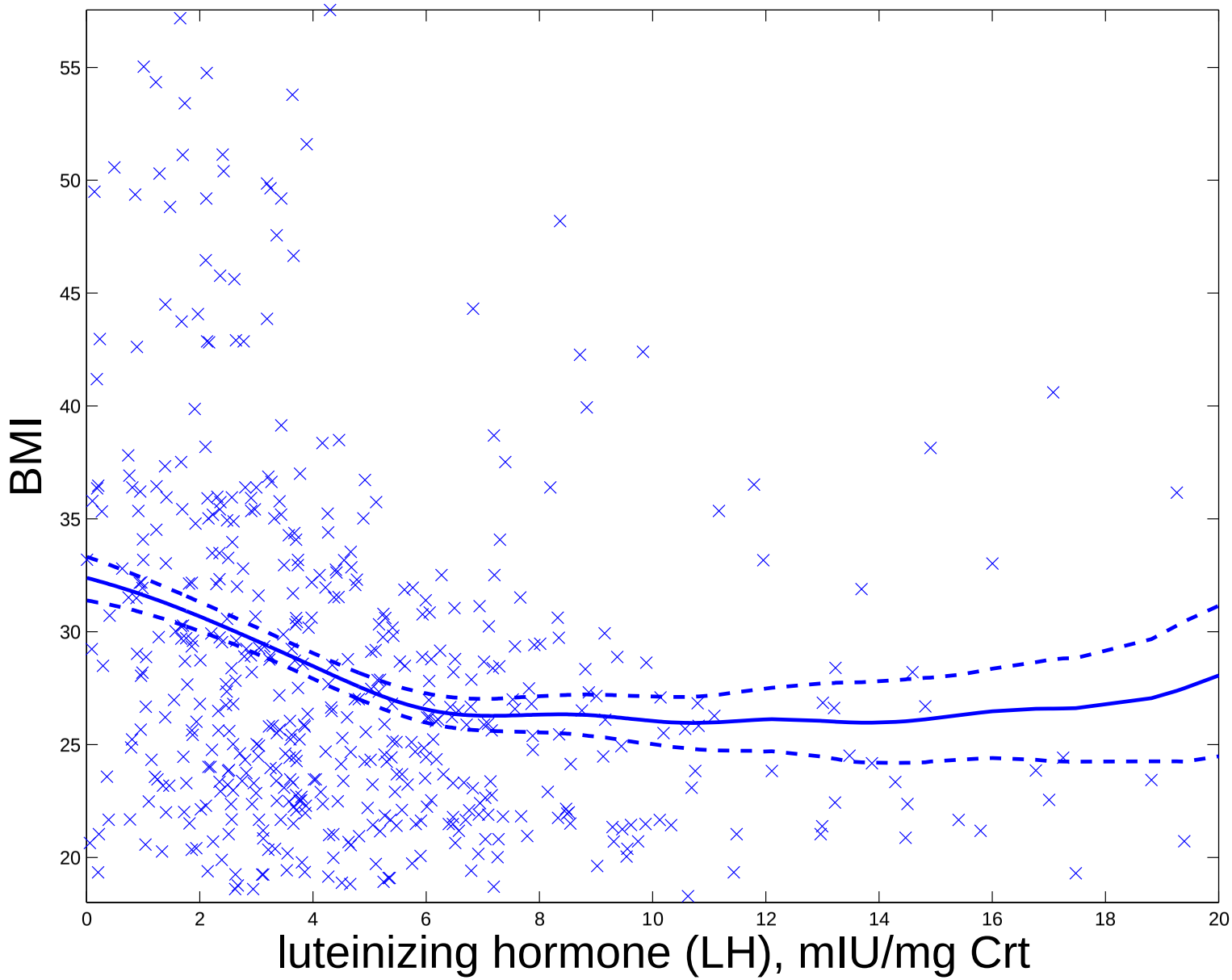
2. Predictive densities for body mass index (BMI) conditional on a range of values for luteinizing hormone (LH). Posterior predictive means (solid lines) and 99% credible intervals (dashed lines) are shown.

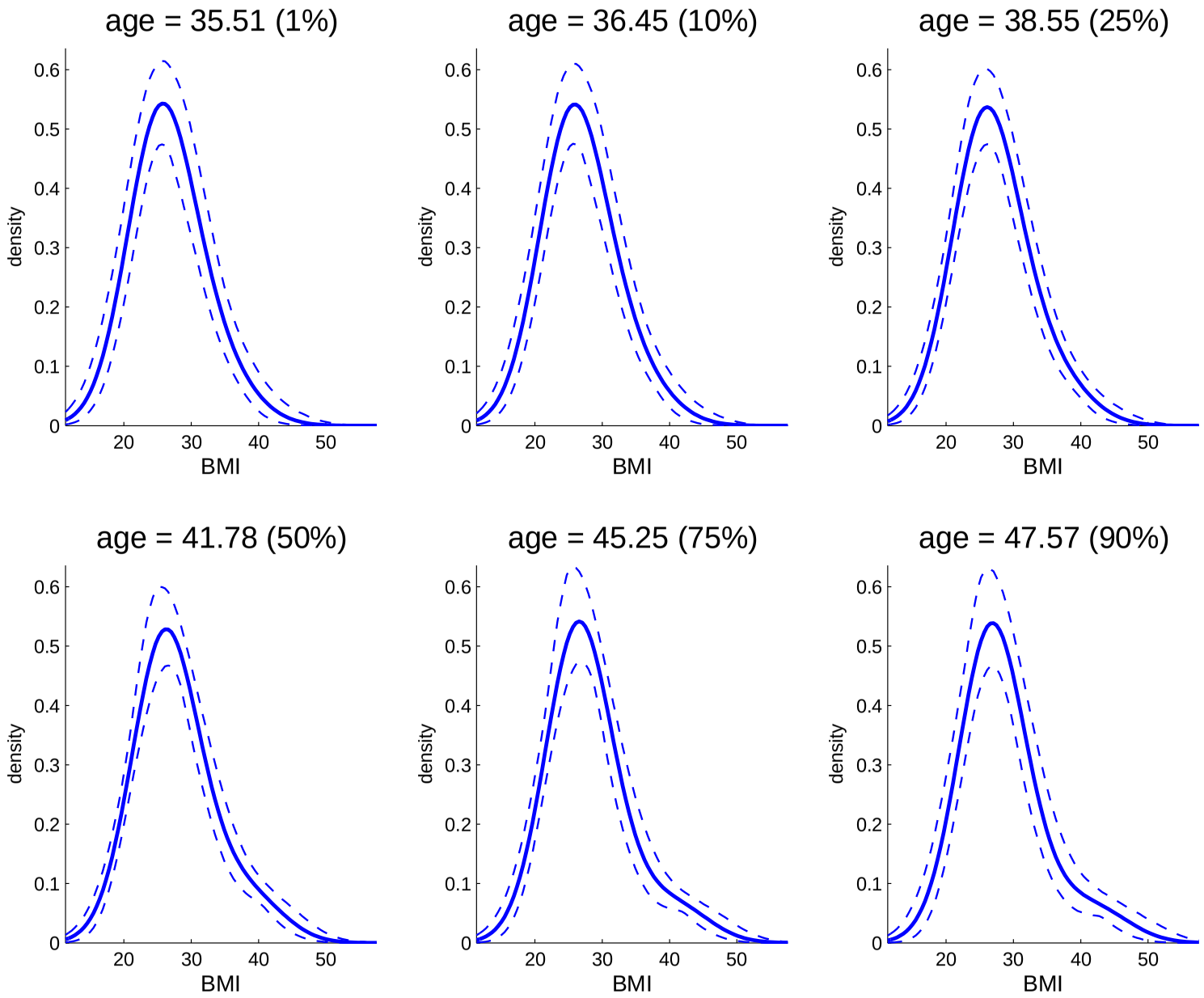
3. Plot of data values for luteinizing hormone (LH) against body mass index. Predictive mean curve (solid line) and 99% pointwise credible intervals (dotted lines) are shown.

4. Predictive densities for body mass index (BMI) conditional on a range of values for age adjusting for LH.


x=0.10 (10%)
x=0.25 (25%)
x=0.51 (50%)
x=0.73 (75%)
x=0.89 (90%)
f(y|x)
y
x


LH = 0.18 (1%)
LH = 1.34 (10%)
LH = 2.48 (25%)
LH = 3.98 (50%)
LH = 6.72 (75%)
LH = 10.14 (90%)
density
BMI
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
20
30
40
50


BMI
55
50
45
40
35
30
25
20
0
2
4
6
8
10
12
14
16
18
20
luteinizing hormone (LH), mIU/mg Crt


age = 35.51 (1%)
age = 36.45 (10%)
age = 38.55 (25%)
age = 41.78 (50%)
age = 45.25 (75%)
age = 47.57 (90%)
density
BMI
0
0.1
0.2
0.3
0.4
0.5
0.6
20
30
40
50