

The Local Dirichlet Process

Yeonseung Chung^{1,2} and David B. Dunson¹

¹*Biostatistics Branch*

National Institute of Environmental Health Sciences

U.S. National Institute of Health

P.O. Box 12233, RTP, NC 27709, U.S.A

²*Department of Biostatistics*

University of North Carolina at Chapel Hill

E-mail: chungy@email.unc.edu

Summary. As a generalization of the Dirichlet process to allow predictor dependence, we propose a local Dirichlet process (IDP). The IDP provides a prior distribution for a collection of random probability measures indexed by predictors. This is accomplished by assigning stick-breaking weights and atoms to random locations in a predictor space. The probability measure at a given predictor value is then formulated using the weights and atoms located in a neighborhood about that predictor value. This construction results in a marginal Dirichlet process prior for the random measure at any specific predictor value. Dependence is induced through local sharing of random components. Theoretical properties are considered and a blocked Gibbs sampler is proposed for posterior computation in IDP mixture models. The methods are illustrated using simulated examples and an epidemiologic application.

Keywords: Dependent Dirichlet process; Blocked Gibbs sampler; Mixture model; Nonparametric Bayes; Stick-breaking representation.

1. Introduction

In recent years, there has been a dramatic increase in applications of nonparametric Bayes methods, motivated largely by the availability of simple and efficient methods for posterior computation in Dirichlet process mixture (DPM) models (Lo, 1984; Escobar, 1994; Escobar and West, 1995). The DPM models incorporate Dirichlet process (DP) priors (Ferguson, 1973, 1974) for components in Bayesian hierarchical models, resulting in an extremely flexible class of models. Due to the flexibility and ease in implementation, DPM models are now routinely implemented in a wide variety of applications, ranging from machine learning (Beal et al., 2002 and Blei et al., 2004) to genomics (Xing et al., 2004 and Kim et al., 2006).

In many settings, it is natural to consider generalizations of the DP and DPM-based models to accommodate dependence. For example, one may be interested in studying changes in a density with predictors. Following Lo (1984), one can use a DPM for Bayes inference on a single density as follows:

$$f(y) = \int_{\Omega} k(y, u) G(du), \quad (1)$$

where $k(y, u)$ is a non-negative valued kernel defined on $(\mathcal{D} \times \Omega, \mathcal{F} \times \mathcal{B})$ such that for each $u \in \Omega$, $\int_{\mathcal{D}} k(y, u) dy = 1$ and for each $y \in \mathcal{D}$, $\int_{\Omega} k(y, u) G(du) < \infty$ with \mathcal{D}, Ω Borel subsets of Euclidean spaces and \mathcal{F}, \mathcal{B} the corresponding σ -fields, and G is a finite random probability measure on (Ω, \mathcal{B}) following a DP. A natural extension for modeling of a conditional density $f(y|\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$, with \mathcal{X} a Lebesgue measurable subset of \mathbb{R}^p , is as follows:

$$f(y|\mathbf{x}) = \int_{\Omega} k(y, u) G_{\mathbf{x}}(du), \quad (2)$$

where the mixing measure $G_{\mathbf{x}}$ is now indexed by the predictor value. We are then faced with modeling a collection of random mixing measures denoted as $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$.

Recent work on defining priors for collections of random probability measures has primarily relied on extending the stick-breaking representation of the DP (Sethuraman, 1994). This literature was stimulated by the dependent Dirichlet process (DDP) framework proposed by MacEachern (1999, 2000, 2001), which replaces the atoms in the Sethuraman (1994) representation with stochastic processes. The DDP framework has been adopted to develop ANOVA-type models for random

probability measures (De Iorio et al., 2004), for flexible spatial modeling (Gelfand et al., 2004), in time series applications (Caron et al., 2006), and for inferences on stochastic ordering (Dunson and Peddada, 2007). The specification of the DDP used in applications incorporates dependence only through the atoms while assuming fixed weights. In other recent work, Griffin and Steel (2006) proposed an order-based DDP (π DDP) which allows varying weights, while Duan et al. (2005) developed a multivariate stick-breaking process for spatial data.

Alternatively, convex combinations of independent DPs can be used for modeling collections of dependent random measures. Müller et al.(2004) proposed this idea to allow dependence across experiments and discrete dynamic settings were considered by Pennell and Dunson (2006) and Dunson(2006). Recently, the idea has been extended to continuous covariate cases by Dunson et al. (2006) and Dunson and Park (2006).

Some desirable properties of a prior for a collection, $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, of predictor-dependent probability measures are: (1) increasing dependence in $G_{\mathbf{x}}$ and $G_{\mathbf{x}'}$ with decreasing distance between \mathbf{x} and \mathbf{x}' ; (2) simple and interpretable expressions for the expectation and variance of each $G_{\mathbf{x}}$ as well as the correlation between $G_{\mathbf{x}}$ and $G_{\mathbf{x}'}$; (3) $G_{\mathbf{x}}$ has a marginal DP prior for all $\mathbf{x} \in \mathcal{X}$; (4) posterior computation can proceed efficiently through a straightforward MCMC algorithm in a broad variety of applications. Although the DDP, π DDP and the prior proposed by Duan et al. (2005) achieve (1), π DDP and Duan et al. (2005) approaches are not straightforward to implement in general applications. The fixed stick-breaking weights version of the DDP tends to be easy to implement, but has the disadvantage of not allowing locally adaptive mixture weights. The kernel mixture approaches of Dunson et al.(2006) and Dunson and Park (2006) lack the marginal DP property (3). Property (3) is appealing in that there is rich theoretical literature on DPs, showing posterior consistency (Ghosal et al., 1999 and Lijoi et al., 2005) and rates of convergence (Ghosal and Van der Vaart, 2006).

This article proposes a simple extension of the DP, which provides an alternative to the fixed weights DDP in order to allow local adaptivity, while also achieving properties (1)-(4). The prior is constructed by first assigning stick-breaking weights and atoms to random locations in a predictor

space. Each predictor-dependent random probability measure is formulated using the random weights and atoms located in a neighborhood about that predictor value. Dependence is induced by local sharing of random components. We call this prior the local Dirichlet process (IDP).

Section 2 describes stick-breaking priors for collections of predictor-dependent random probability measures. Section 3 introduces the IDP and discusses properties. Computation is described in section 4. Sections 5 and 6 include simulation studies and an epidemiologic application. Section 7 concludes with a discussion. Proofs are included in appendices.

2. Predictor-Dependent Stick-Breaking Priors

2.1 Stick-Breaking Priors

Ishwaran and James (2001) proposed a general class of stick-breaking priors for random probability measures. This class provides a useful starting point in considering extensions to allow predictor dependence.

Definition 1. A random probability measure, G , has a *stick-breaking prior* if

$$G = \sum_{h=1}^N p_h \delta_{\theta_h}, \quad 0 \leq p_h \leq 1, \quad \sum_{h=1}^N p_h = 1 \quad a.s., \quad (3)$$

where δ_{θ} is a discrete measure concentrated at θ , $p_h = V_h \prod_{l < h} (1 - V_l)$ are random weights with $V_h \stackrel{iid}{\sim} \text{Beta}(a_h, b_h)$ independently from $\theta_h \stackrel{iid}{\sim} G_0$ with G_0 a non-atomic base probability measure. For $N = \infty$, the condition $\sum_{h=1}^N p_h = 1$ a.s. is satisfied by Lemma 1 in Ishwaran and James (2001). For finite N , the condition is satisfied by letting $V_N = 1$.

There are many processes that fall into this class. The DP corresponds to the special case in which $N = \infty$, $a_h = 1$ and $b_h = \alpha$ as established in Sethuraman (1994). The two-parameter Poisson-Dirichlet process corresponds to the case where $N = \infty$, $a_h = 1 - a$, and $b_h = b + ha$ with $0 \leq a < 1$ and $b > -1$ (Pitman 1995, 1996). Additional special cases are listed in Ishwaran and James (2001).

2.2 Predictor-Dependent Stick-Breaking Priors

Consider an uncountable collection of predictor-dependent random probability measures, $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$. The predictor space \mathcal{X} is a Lebesgue measurable subset of Euclidian space and the

random measures $G_{\mathbf{x}}$ are defined on (Ω, \mathcal{B}) where Ω is a sample space and \mathcal{B} is a corresponding Borel σ -algebra. Let \mathcal{P} be a probability measure on $(\mathcal{M}, \mathcal{N})$ where \mathcal{M} is the space of uncountable collections of random probability measures $G_{\mathbf{x}}$ and \mathcal{N} is the corresponding σ -algebra. Then, $\mathcal{G}_{\mathcal{X}} \sim \mathcal{P}$ denotes that \mathcal{P} is a prior on the random collection $\mathcal{G}_{\mathcal{X}}$.

We call \mathcal{P} a *predictor-dependent stick-breaking prior* ($\text{SBP}_{\mathcal{X}}$) if it can be represented as:

$$G_{\mathbf{x}} = \sum_{h=1}^{N(\mathbf{x})} p_h(\mathbf{x}) \delta_{\theta_h(\mathbf{x})}$$

with $0 \leq p_h(\mathbf{x}) \leq 1$ and $\sum_{h=1}^{N(\mathbf{x})} p_h(\mathbf{x}) = 1 \quad a.s., \quad \forall \mathbf{x} \in \mathcal{X},$ (4)

where the random weights $p_h(\mathbf{x})$ have a stick-breaking representation, $p_h(\mathbf{x})$ and $\theta_h(\mathbf{x})$ are predictor-dependent, and $N(\mathbf{x})$ is also indexed by the predictor value \mathbf{x} . Depending on how we form $p_h(\mathbf{x})$, $\theta_h(\mathbf{x})$ and $N(\mathbf{x})$, different dependencies among $G_{\mathbf{x}}$ are induced. Several interesting priors, such as the DDP, π DDP and the prior proposed by Duan et al. (2005) fall into the $\text{SBP}_{\mathcal{X}}$ class. In the next section, we propose a new choice of $\text{SBP}_{\mathcal{X}}$ deemed the local Dirichlet process (lDP).

3. Local Dirichlet Process

3.1 Formulation

The lDP is constructed from the following three sequences of mutually independent global random components:

$$\mathbf{\Gamma} = \{\Gamma_h, h = 1, \dots, \infty\}, \quad \mathbf{V} = \{V_h, h = 1, \dots, \infty\}, \quad \mathbf{\Theta} = \{\theta_h, h = 1, \dots, \infty\}, \quad (5)$$

where $\Gamma_h \stackrel{iid}{\sim} H$ are locations, $V_h \stackrel{iid}{\sim} \text{Beta}(1, \alpha)$ are probability weights, and $\theta_h \stackrel{iid}{\sim} G_0$ are atoms. G_0 is a probability measure on (Ω, \mathcal{B}) on which $G_{\mathbf{x}}$ will be defined and H is a probability measure on $(\mathcal{X}', \mathcal{A})$ where \mathcal{A} is a Borel σ -algebra of subsets of \mathcal{X}' and \mathcal{X}' is a Lebesgue measurable subset of Euclidian space that may or may not correspond to the predictor space \mathcal{X} . In addition, H satisfies the following regularity condition:

Condition 1. Letting $d : \mathcal{X} \times \mathcal{X}' \rightarrow \mathbb{R}^+$ be a distance measure, $\eta_{\mathbf{x}}^{\psi} = \{\mathbf{x}' : d(\mathbf{x}, \mathbf{x}') < \psi, \mathbf{x}' \in \mathcal{X}'\}$ defines a ψ -neighborhood around a point $\mathbf{x} \in \mathcal{X}$. Then, $H(\eta_{\mathbf{x}}^{\psi}) > 0, \forall \mathbf{x} \in \mathcal{X}$ and $\psi > 0$.

Next, let us define sets of local random components, for any $\mathbf{x} \in \mathcal{X}$:

$$\mathbf{\Gamma}(\mathbf{x}) = \{\Gamma_h, h \in \mathcal{L}_{\mathbf{x}}\}, \mathbf{V}(\mathbf{x}) = \{V_h, h \in \mathcal{L}_{\mathbf{x}}\}, \mathbf{\Theta}(\mathbf{x}) = \{\theta_h, h \in \mathcal{L}_{\mathbf{x}}\}, \quad (6)$$

where $\mathcal{L}_{\mathbf{x}} = \{h : d(\mathbf{x}, \Gamma_h) < \psi, h = 1, \dots, \infty\}$ is a predictor-dependent set indexing the locations belonging to the neighborhood of \mathbf{x} . Hence, the sets $\mathbf{V}(\mathbf{x})$ and $\mathbf{\Theta}(\mathbf{x})$ contain the random weights and atoms that are assigned to the locations $\mathbf{\Gamma}(\mathbf{x})$ in a neighborhood of \mathbf{x} . Here, ψ controls the neighborhood size. For simplicity, we treat ψ as fixed throughout the paper, though one can obtain a more flexible class of priors by assuming a hyper prior for ψ .

Using the local random components in (6), let us define the following form for $G_{\mathbf{x}}$:

$$G_{\mathbf{x}} = \sum_{l=1}^{N(\mathbf{x})} p_l(\mathbf{x}) \delta_{\theta_{\pi_l(\mathbf{x})}} \quad \text{with} \quad p_l(\mathbf{x}) = V_{\pi_l(\mathbf{x})} \prod_{j<l} (1 - V_{\pi_j(\mathbf{x})}), \quad (7)$$

where $N(\mathbf{x})$ is the cardinality of $\mathcal{L}_{\mathbf{x}}$ and $\pi_l(\mathbf{x})$ is the l th ordered index in $\mathcal{L}_{\mathbf{x}}$. Then, Condition 1 ensures that the following lemma holds.

Lemma 1. For all $\mathbf{x} \in \mathcal{X}$, $N(\mathbf{x}) = \infty$ and $\sum_{l=1}^{N(\mathbf{x})} p_l(\mathbf{x}) = 1$ almost surely.

Therefore, $G_{\mathbf{x}}$ is a well-defined stick-breaking probability measure for all $\mathbf{x} \in \mathcal{X}$ and the formulation (5)-(7) provides a new choice of $\text{SBP}_{\mathcal{X}}$, deemed the local Dirichlet process (IDP). We use the shorthand notation $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\} \sim \text{IDP}(\alpha, G_0, H, \psi)$ to denote that the collection of probability measures $\mathcal{G}_{\mathcal{X}}$ is assigned an IDP with hyperparameters α, G_0, H, ψ .

Figure 1 illustrates the IDP formulation graphically. Suppose $\mathcal{G}_{\mathcal{X}} \sim \text{IDP}(\alpha, G_0, H, \psi)$, with $\mathcal{X} = [0, 1]^2$, $H = \text{Uniform}([0, 1]^2)$, and $\psi = 0.2$. Random locations in $[0, 1]^2$ are generated from a uniform distribution, with the first 100 locations plotted as ' \star ' in Figure 1. The random pair (V_h, θ_h) is placed at location Γ_h . For the predictor value $\mathbf{x} = (0.5, 0.3)'$, the red dashed circle indicates the neighborhood of \mathbf{x} , $\eta_{\mathbf{x}}^{\psi}$, with $d(\cdot)$ the Euclidian distance. Then, $G_{\mathbf{x}}$ at $\mathbf{x} = (0.5, 0.3)'$ is constructed using the weights and atoms within the dashed circle in the order of the index to formulate the stick-breaking representation. From Figure 1, it is apparent that the dependence between $G_{\mathbf{x}}$ and $G_{\mathbf{x}'}$ increases as the distance between \mathbf{x} and \mathbf{x}' decreases. For closer \mathbf{x} and \mathbf{x}' , their neighborhoods overlap more so that similar components are used for constructing $G_{\mathbf{x}}$ and $G_{\mathbf{x}'}$, while if \mathbf{x} and \mathbf{x}' are far apart, there will be at most a small area of intersection so that few to none of the random

components are shared. In the non-overlapping case, $G_{\mathbf{x}}$ and $G'_{\mathbf{x}}$ are assigned independent DP priors, as is clear from Theorem 1 and the subsequent development.

Theorem 1. If $\mathcal{G}_{\mathcal{X}} \sim \text{IDP}(\alpha, G_0, H, \psi)$, for any $\mathbf{x} \in \mathcal{X}$, $G_{\mathbf{x}} \sim \text{DP}(\alpha G_0)$.

The marginal DP property shown in Theorem 1 is appealing in allowing one to rely directly on the rich literature on properties of the DP to obtain insight into the prior for the random probability measure at any particular predictor value. However, unlike the DP, the IDP allows the probability measure to vary with predictors, while borrowing information across local regions of the predictor space. This is accomplished through incorporating shared random components. Due to the sharing and to the almost sure discreteness property of each $G_{\mathbf{x}}$, the IDP will induce local clustering of subjects according to their predictor values. Theorem 2 illustrates this local clustering property more clearly.

Theorem 2. Suppose $\mathcal{G}_{\mathcal{X}} \sim \text{IDP}(\alpha, G_0, H, \psi)$ and $\phi_i \sim G_{\mathbf{x}_i}$, for $i = 1, \dots, n$, with \mathbf{x}_i denoting the predictor value for subject i . Then,

$$\kappa_{\mathbf{x}_i, \mathbf{x}_j} = \Pr(\phi_i = \phi_j \mid \mathbf{x}_i, \mathbf{x}_j, \alpha, \psi) = \frac{2P_{\mathbf{x}_i, \mathbf{x}_j}}{(1 + P_{\mathbf{x}_i, \mathbf{x}_j})\alpha + 2}, \text{ for any } \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$$

where $P_{\mathbf{x}_i, \mathbf{x}_j} = \frac{H(\eta_{\mathbf{x}_i}^\psi \cap \eta_{\mathbf{x}_j}^\psi)}{H(\eta_{\mathbf{x}_i}^\psi \cup \eta_{\mathbf{x}_j}^\psi)}$ is the conditioned probability of Γ_h falling within the intersection region $\eta_{\mathbf{x}_i}^\psi \cap \eta_{\mathbf{x}_j}^\psi$ given $\Gamma_h \in \eta_{\mathbf{x}_i}^\psi \cup \eta_{\mathbf{x}_j}^\psi$.

The clustering probability $\kappa_{\mathbf{x}_i, \mathbf{x}_j}$ increases from 0 when $\eta_{\mathbf{x}_i}^\psi \cap \eta_{\mathbf{x}_j}^\psi = \emptyset$ to $1/(\alpha + 1)$ when $\mathbf{x}_i = \mathbf{x}_j$ which is the case of $P_{\mathbf{x}_i, \mathbf{x}_j} = 1$. This implies that, for fixed α , the clustering probability under $\mathcal{G}_{\mathcal{X}} \sim \text{IDP}(\alpha, G_0, H, \psi)$ is bounded above by the clustering probability under the global DP, which takes $G_{\mathbf{x}} \equiv G \sim \text{DP}(\alpha G_0)$, leading to $\Pr(\phi_i = \phi_j \mid \alpha) = 1/(1 + \alpha)$. Also, note that small values of the precision parameter α will induce V_h values that are close to one. This in turn causes a small number of atoms in each neighborhood to dominate, inducing few local clusters. However, when ψ is small and hence neighborhood sizes are small, there will still be many clusters across \mathcal{X} .

It is interesting to consider relationships between the IDP and other priors proposed in the literature in limiting special cases. First, note that the IDP converges to the DP as $\psi \rightarrow \infty$, so that the neighborhoods around each of the predictor values encompass the entire predictor space. Also, the $\text{IDP}(\alpha, G_0, H, \psi)$ corresponds to a limiting case of the kernel stick-breaking process (KSBP)

(Dunson and Park, 2006), in which the kernel is defined as $K(\mathbf{x}, \Gamma) = 1(d(\mathbf{x}, \Gamma) < \psi)$ and the DP placed at each location have precision parameters $\rightarrow 0$. In addition, if we condition on the locations Γ as fixed, the IDP corresponds to a special case of the π DDP, in which the region of locations relevant for determining the ordering at \mathbf{x} is fixed and the stochastic random ordering $\{\pi(\mathbf{x})\}_{\mathbf{x} \in \mathcal{X}}$ is replaced by fixed orderings for all $\mathbf{x} \in \mathcal{X}$.

3.2 Moments and Correlation

From Theorem 1 and properties of the DP, $\mathcal{G}_{\mathcal{X}} \sim lDP(\alpha, G_0, H, \psi)$ implies, for any $\mathbf{x} \in \mathcal{X}$,

$$E\{G_{\mathbf{x}}(B)\} = G_0(B) \quad \text{and} \quad Var\{G_{\mathbf{x}}(B)\} = \frac{G_0(B)(1 - G_0(B))}{1 + \alpha}, \quad \forall B \in \mathcal{B} \quad (8)$$

Next, let us consider the correlation between $G_{\mathbf{x}_1}$ and $G_{\mathbf{x}_2}$, for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$. First, we show the correlation conditionally on the locations Γ but marginalizing out the weights \mathbf{V} and atoms Θ . As discussed in section 3.1, if Γ is given, the IDP can be regarded as a special case of the π DDP. Hence, following Theorem 1 in Griffin and Steel (2006), for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$,

$$\begin{aligned} \rho_{\mathbf{x}_1, \mathbf{x}_2}(\Gamma) &= Corr\{G_{\mathbf{x}_1}(B), G_{\mathbf{x}_2}(B) | \Gamma\} \\ &= \frac{2}{\alpha + 2} \sum_{h \in \mathcal{L}_{\mathbf{x}_1} \cap \mathcal{L}_{\mathbf{x}_2}} \left(\frac{\alpha}{\alpha + 2}\right)^{\#\mathcal{S}_h} \left(\frac{\alpha}{\alpha + 1}\right)^{\#\mathcal{S}'_h}, \quad \forall B \in \mathcal{B}, \end{aligned} \quad (9)$$

where $\#\mathcal{S}$ is the cardinality of the set \mathcal{S} , $\mathcal{S}_h = \mathcal{A}_{1h} \cap \mathcal{A}_{2h}$, $\mathcal{S}'_h = \mathcal{A}_{1h} \cup \mathcal{A}_{2h} - \mathcal{S}_h$, and $\mathcal{A}_{kh} = \{\pi_j(\mathbf{x}_k) : j < l, \pi_l(\mathbf{x}_k) = h\}$ for $h \in \mathcal{L}_{\mathbf{x}_1} \cap \mathcal{L}_{\mathbf{x}_2}$. In other words, $\#\mathcal{S}_h$ is the number of indices on the locations Γ that are below h and are shared in the neighborhoods of \mathbf{x}_1 and \mathbf{x}_2 , while $\#\mathcal{S}'_h$ is the number of indices that are below h and belong to the neighborhoods of either \mathbf{x}_1 or \mathbf{x}_2 but not both. For a given h , reducing $\#\mathcal{S}_h$ by one induces adding two elements to \mathcal{S}'_h , thus reducing the correlation, as expected. From expression (9), it is clear that the neighborhoods around \mathbf{x}_1 and \mathbf{x}_2 are increasingly overlapping and the correlation between $G_{\mathbf{x}_1}$ and $G_{\mathbf{x}_2}$ increases as $\mathbf{x}_1 \rightarrow \mathbf{x}_2$. Expression (9) is particularly useful in being free of dependence on B .

Marginalizing the correlation in (9) over the prior for the random locations Γ is equivalent to marginalizing out the $\#\mathcal{S}_h$ and $\#\mathcal{S}'_h$ for $h \in \mathcal{L}_{\mathbf{x}_1} \cap \mathcal{L}_{\mathbf{x}_2}$. In considering the correlation between $G_{\mathbf{x}_1}$ and $G_{\mathbf{x}_2}$, we can ignore the Γ_h for $h \in \{1, \dots, \infty\} \setminus \mathcal{L}_{\mathbf{x}_1} \cup \mathcal{L}_{\mathbf{x}_2}$ and focus on the Γ_h only for

$h \in \mathcal{L}_{\mathbf{x}_1} \cup \mathcal{L}_{\mathbf{x}_2}$. Let γ_j be the j th ordered component of $\mathcal{L}_{\mathbf{x}_1} \cup \mathcal{L}_{\mathbf{x}_2}$. For example, if $\mathcal{L}_{\mathbf{x}_1} \cup \mathcal{L}_{\mathbf{x}_2} = \{1, 3, 5, 6, 7, 8, \dots\}$, $\gamma_1 = 1, \gamma_2 = 3, \gamma_3 = 5, \gamma_4 = 6, \dots$. Let $Z_{\gamma_j} = 1(\gamma_j \in \mathcal{L}_{\mathbf{x}_1} \cap \mathcal{L}_{\mathbf{x}_2})$ be an indicator for whether Γ_{γ_j} are shared by the neighborhoods of \mathbf{x}_1 and \mathbf{x}_2 or not. Then, the formula in (9) can be reexpressed with respect to Z_{γ_j} as follows:

$$\begin{aligned} \rho_{\mathbf{x}_1, \mathbf{x}_2}(\mathbf{\Gamma}) &= \text{Corr}\{G_{\mathbf{x}_1}(B), G_{\mathbf{x}_2}(B) | \mathbf{\Gamma}\} \\ &= \frac{2}{\alpha + 2} \sum_{j=1}^{\infty} Z_{\gamma_j} \left(\frac{\alpha}{\alpha + 2}\right)^{\sum_{k=1}^{j-1} Z_{\gamma_k}} \left(\frac{\alpha}{\alpha + 1}\right)^{j-1 - \sum_{k=1}^{j-1} Z_{\gamma_k}} \end{aligned} \quad (10)$$

Note that it is straightforward to show that $Z_{\gamma_j} \stackrel{iid}{\sim} \text{Bernoulli}(P_{\mathbf{x}_1, \mathbf{x}_2})$, for $j = 1, \dots, \infty$, with $P_{\mathbf{x}_1, \mathbf{x}_2} = \frac{H(\eta_{\mathbf{x}_1}^{\psi} \cap \eta_{\mathbf{x}_2}^{\psi})}{H(\eta_{\mathbf{x}_1}^{\psi} \cup \eta_{\mathbf{x}_2}^{\psi})}$ the conditioned probability of Γ_h falling within the intersection region $\eta_{\mathbf{x}_1}^{\psi} \cap \eta_{\mathbf{x}_2}^{\psi}$ given $\Gamma_h \in \eta_{\mathbf{x}_1}^{\psi} \cup \eta_{\mathbf{x}_2}^{\psi}$. Finally, marginalizing out $\{Z_{\gamma_j}\}_{j=1}^{\infty}$ results in the following Theorem.

Theorem 3. If $\mathcal{G}_{\mathcal{X}} \sim \text{LDP}(\alpha, G_0, H, \psi)$, for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$,

$$\rho_{\mathbf{x}_1, \mathbf{x}_2} = \text{Corr}\{G_{\mathbf{x}_1}(B), G_{\mathbf{x}_2}(B)\} = \frac{2P_{\mathbf{x}_1, \mathbf{x}_2}(\alpha + 1)}{(1 + P_{\mathbf{x}_1, \mathbf{x}_2})\alpha + 2}, \quad \forall B \in \mathcal{B}$$

The correlation is expressed only in terms of $P_{\mathbf{x}_1, \mathbf{x}_2}$ and α . Regardless of α , the correlation is 1 if $P_{\mathbf{x}_1, \mathbf{x}_2} = 1$, which implies that the neighborhoods around $\mathbf{x}_1, \mathbf{x}_2$ are identical, which only occurs if $\mathbf{x}_1 = \mathbf{x}_2$. Also, the correlation is 0 when $P_{\mathbf{x}_1, \mathbf{x}_2} = 0$, meaning that the neighborhoods are non-overlapping. In addition, $P_{\mathbf{x}_1, \mathbf{x}_2} \leq \rho_{\mathbf{x}_1, \mathbf{x}_2} \leq 1$ and $\rho_{\mathbf{x}_1, \mathbf{x}_2}$ increases as α increases for fixed $P_{\mathbf{x}_1, \mathbf{x}_2}$. When $\alpha \rightarrow 0$, the correlation converges to $P_{\mathbf{x}_1, \mathbf{x}_2}$. Meanwhile, when $\alpha \rightarrow \infty$, the correlation converges to $\frac{2P_{\mathbf{x}_1, \mathbf{x}_2}}{1 + P_{\mathbf{x}_1, \mathbf{x}_2}}$.

Note that $P_{\mathbf{x}_1, \mathbf{x}_2}$ depends on H and the locations \mathbf{x}_1 and \mathbf{x}_2 . When \mathcal{X}' is chosen to satisfy Condition 2, some appealing properties result.

Condition 2. For all $\mathbf{x} \in \mathcal{X}$, the points $\mathbf{x}' \in \eta_{\mathbf{x}}^{\psi}$ belonging to the ψ -neighborhood of \mathbf{x} all fall within \mathcal{X}' .

Condition 2 implies that $\mathcal{X} \subset \mathcal{X}'$, with \mathcal{X}' consisting of the union of the local neighborhoods around every point in \mathcal{X} . Under condition 2, with H chosen to be the uniform probability measure on \mathcal{X}' , $P_{\mathbf{x}_1, \mathbf{x}_2}$ depends only on $d(\mathbf{x}_1, \mathbf{x}_2)$ and not on the location of \mathbf{x}_1 or \mathbf{x}_2 within \mathcal{X} . Hence, upon examination of Theorem 3, it is apparent that condition 2 implies an isotropic correlation structure, which is an appealing default in the absence of prior knowledge of changes in the correlation

structure according to location in \mathcal{X} . Figure 2 shows how the correlation $\rho_{\mathbf{x}_1, \mathbf{x}_2}$ changes as a function of the distance between \mathbf{x}_1 and \mathbf{x}_2 , $d(\mathbf{x}_1, \mathbf{x}_2)$, in the case where \mathbf{x} is univariate, with $\mathbf{x} \in \mathcal{X} = [0, 1]$, $H \text{ Uniform}([- \psi, 1 + \psi])$, and $d(\cdot)$ corresponding to the Euclidian distance. The $\rho_{\mathbf{x}_1, \mathbf{x}_2}$ decays from 1 to 0 as $d(\mathbf{x}_1, \mathbf{x}_2)$ increases and the decay is faster for smaller ψ . As $\psi \rightarrow \infty$, the decay line gets closer to a horizontal line at $\rho_{\mathbf{x}_1, \mathbf{x}_2} = 1$, which is the case of IDP=DP. Also, for a given ψ and $d(\mathbf{x}_1, \mathbf{x}_2)$, the $\rho_{\mathbf{x}_1, \mathbf{x}_2}$ is higher as $\alpha \rightarrow \infty$. Although the choice of $d(\cdot)$ being Euclidian makes the curves in Figure 2 close to linear, the curvature can easily be changed by choosing a different distance measure $d(\cdot)$.

3.3 Truncation Approximation

Finite approximations to infinite stick-breaking priors form the basis for commonly-used computational algorithms (Ishwaran and James, 2001). In this subsection, we discuss a finite dimensional approximation to the IDP.

Since the IDP has the marginal DP property, let us recall the finite dimensional DP. Ishwaran and James (2001) defines an N-truncation of the DP (DP^N) by discarding the $N + 1, N + 2, \dots, \infty$ terms and replacing p_N with $1 - \sum_{h=1}^{N-1} p_h$ in the DP stick-breaking form in (3). They show that the DP^N approximates the DP well in terms of the total variation (tv) norm of the marginal densities of the data obtained from the corresponding DPM models. According to their Theorem 2,

$$\|\mu_N - \mu_\infty\| \leq 4 \left(1 - E \left[\left(\sum_{h=1}^{N-1} p_h \right)^n \right] \right) \approx 4n \exp(-(N-1)/\alpha), \quad (11)$$

where $\|\cdot\|$ is tv norm, μ_N and μ_∞ are the marginal probability measures for the data from the DPM^N and DPM models, and n is the sample size. Note that the sample size has a modest effect on the bound for a reasonably large value of N and the bound decreases exponentially with N increasing, implying that even for a fairly large sample size, the DPM^N approximates the DP well with moderate N .

Following a similar route, let us define an N-truncation of the IDP (IDP^N) as follows:

Definition 3. For a finite N , let $\mathbf{\Gamma}^N = \{\Gamma_h, h = 1, \dots, N\}$, $\mathbf{V}^N = \{V_h, h = 1, \dots, N\}$, and $\mathbf{\Theta}^N = \{\theta_h, h = 1, \dots, N\}$ be the sets of random locations, weights, and atoms, respectively.

Distributional assumptions for Γ_h , V_h , and θ_h are the same as in (5) and the corresponding local sets are defined as in (6). Then, $\mathcal{G}_{\mathcal{X}} \sim \text{IDP}^N(\alpha, G_0, H, \psi)$ if

$$G_{\mathbf{x}} = \sum_{l=1}^{N(\mathbf{x})-1} p_l(\mathbf{x}) \delta_{\theta_{\pi_l(\mathbf{x})}} + \left(1 - \sum_{l=1}^{N(\mathbf{x})-1} p_l(\mathbf{x})\right) \delta_{\theta_{\pi_{N(\mathbf{x})}(\mathbf{x})}}$$

with $p_l(\mathbf{x}) = V_{\pi_l(\mathbf{x})} \prod_{j<l} (1 - V_{\pi_j(\mathbf{x})})$ for $l = 1, \dots, N(\mathbf{x}) - 1$ (12)

The $G_{\mathbf{x}}$ in (12) has a similar form to $G = \sum_{h=1}^N p_h \delta_{\theta_h}$ obtained from the DP^N except that N in G is replaced by $N(\mathbf{x})$ in $G_{\mathbf{x}}$. The difference is that N in DP^N is fixed but $N(\mathbf{x})$ in IDP^N is random. Focusing on a particular predictor value \mathbf{x} , it is easy to show that $N(\mathbf{x}) \sim \text{Binomial}(N, P_{\mathbf{x}})$, where N is the total number of global components and $P_{\mathbf{x}} = H(\eta_{\mathbf{x}}^{\psi})$ is the probability that a component belongs to the neighborhood around \mathbf{x} , $\eta_{\mathbf{x}}^{\psi}$. Then, marginalizing out $N(\mathbf{x})$ in the bound on the tv distance between the marginal densities of an observation obtained at a particular predictor value \mathbf{x} from the IDPM and IDPM^N models results in Theorem 4.

Theorem 4. Define a model (2) with $\mathcal{G}_{\mathcal{X}} \sim \text{IDP}(\alpha, G_0, H, \psi)$ as local Dirichlet process mixture (IDPM) model. IDPM^N corresponds to (2) with $\mathcal{G}_{\mathcal{X}} \sim \text{IDP}^N$. Suppose an observation is obtained from IDPM^N and IDPM models. Then,

$$\|\mu_N(\mathbf{x}) - \mu_{\infty}(\mathbf{x})\| \leq 4 \left(\frac{\alpha + 1}{\alpha} \right) \left[1 - \left(\frac{1}{\alpha + 1} \right) P_{\mathbf{x}} \right]^N,$$

where $\mu_N(\mathbf{x})$ and $\mu_{\infty}(\mathbf{x})$ are the marginal probability measures for the observation. Notice that the bound decreases exponentially with N increasing, suggesting that we can obtain a good approximation to the IDP using a moderate N , as long as α is small and the neighborhood size is not too small. In particular, we require a large N for a given level of accuracy as $\psi \rightarrow 0$, since $P_{\mathbf{x}}$ decreases as the size of $\eta_{\mathbf{x}}^{\psi}$ decreases.

4. Posterior Computation

We develop an MCMC algorithm based on the blocked Gibbs sampler (Ishwaran and James, 2001) for an IDPM^N model. For simplicity in exposition, we describe a Gibbs sampling algorithm for a particular hierarchical model, though the approach can be easily adapted for computation in a

broad variety of other settings. In particular, we let

$$\begin{aligned} f(y_i | \mathbf{x}_i, \tau) &= \int f(y_i | \mathbf{x}_i, \boldsymbol{\beta}_i, \tau) dG_{\mathbf{x}_i}(\boldsymbol{\beta}_i) \quad \text{for } i = 1, \dots, n \\ \mathcal{G}_{\mathcal{X}} &\sim \text{LDPN}(\alpha, G_0, H, \psi), \end{aligned} \quad (13)$$

where $f(y_i | \mathbf{x}_i, \boldsymbol{\beta}_i, \tau) = N(y_i; \mathbf{x}_i' \boldsymbol{\beta}_i, \tau^{-1})$ with $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{ip})'$. For simplicity, we consider a univariate predictor case where $p = 2$ and $\mathbf{x}_i' = (1, x_i)$, so that we take a univariate probability measure for H . G_0 is assumed to be $N_p(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$ and additional conjugate priors are assigned for τ , α , $\boldsymbol{\mu}_\beta$ and $\boldsymbol{\Sigma}_\beta$.

Let K_i be an indicator variable denoting that $K_i = h$ implies i th subject is assigned to the h th mixture component. Then, the hierarchical structure of the model (13) with respect to the random variables is recast as follows.

$$\begin{aligned} (y_i | \mathbf{x}_i, \boldsymbol{\beta}^*, \tau, \mathbf{K}) &\sim N(\mathbf{x}_i' \boldsymbol{\beta}_{K_i}^*, \tau^{-1}), \quad i = 1, \dots, n \\ (K_i | \mathbf{V}, \boldsymbol{\Gamma}) &\sim \sum_{l=1}^{N(\mathbf{x}_i)} p_l(\mathbf{x}_i) \delta_{\pi_l(\mathbf{x}_i)}(\cdot), \quad i = 1, \dots, n \\ (V_h | \alpha) &\sim \text{Beta}(1, \alpha), \quad h = 1, \dots, N \\ (\Gamma_h) &\sim \text{Uniform}(a_\Gamma, b_\Gamma), \quad h = 1, \dots, N \\ (\boldsymbol{\beta}_h^* | \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) &\sim N_p(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta), \quad h = 1, \dots, N \\ \boldsymbol{\mu}_\beta &\sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_\mu) \\ \boldsymbol{\Sigma}_\beta^{-1} &\sim \text{Wishart}(\{\nu_0 \boldsymbol{\Sigma}_0\}^{-1}, \nu_0) \\ \tau &\sim \text{Gamma}(\nu_1, \nu_2) \\ \alpha &\sim \text{Gamma}(\eta_1, \eta_2), \end{aligned} \quad (14)$$

where $\boldsymbol{\beta}^* = \{\boldsymbol{\beta}_h^*, h = 1, \dots, N\}$, $\mathbf{K} = \{K_i, i = 1 \dots, n\}$, $\mathbf{V} = \{V_h, h = 1, \dots, N\}$, and $\boldsymbol{\Gamma} = \{\Gamma_h, h = 1, \dots, N\}$. The full conditionals for each of the random components are based on the following joint distribution.

$$\begin{aligned} &(\mathbf{y}, \mathbf{K}, \mathbf{V}, \boldsymbol{\Gamma}, \boldsymbol{\beta}^*, \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta, \tau, \alpha) \\ &\propto (\mathbf{y} | \boldsymbol{\beta}^*, \tau, \mathbf{K})(\mathbf{K} | \mathbf{V}, \boldsymbol{\Gamma})(\mathbf{V} | \alpha)(\boldsymbol{\Gamma})(\boldsymbol{\beta}^* | \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)(\boldsymbol{\mu}_\beta)(\boldsymbol{\Sigma}_\beta)(\tau)(\alpha) \end{aligned} \quad (15)$$

Then, the Gibbs sampler proceeds by sampling from the following conditional posterior distributions:

(a) Conditional for $K_i, i = 1, \dots, n$

$$\begin{aligned}
(K_i | \mathbf{y}, \mathbf{V}, \mathbf{\Gamma}, \boldsymbol{\beta}^*, \tau) &\sim \sum_{l=1}^{N(\mathbf{x}_i)} p'_l(\mathbf{x}_i) \delta_{\pi_l(\mathbf{x}_i)}(K_i) \\
p'_l(\mathbf{x}_i) &= \frac{N(y_i; \mathbf{x}'_i \boldsymbol{\beta}^*_{\pi_l(\mathbf{x}_i)}, \tau^{-1}) p_l(\mathbf{x}_i)}{\sum_{l=1}^{N(\mathbf{x}_i)} N(y_i; \mathbf{x}'_i \boldsymbol{\beta}^*_{\pi_l(\mathbf{x}_i)}, \tau^{-1}) p_l(\mathbf{x}_i)} \\
p_l(\mathbf{x}_i) &= V_{\pi_l(\mathbf{x}_i)} \prod_{j < l} (1 - V_{\pi_j(\mathbf{x}_i)}) \quad \text{for } l < N(\mathbf{x}_i) \\
p_l(\mathbf{x}_i) &= \prod_{j < l} (1 - V_{\pi_j(\mathbf{x}_i)}) \quad \text{for } l = N(\mathbf{x}_i)
\end{aligned}$$

(b) Conditional for $V_h, h = 1, \dots, N$

$$(V_h | \mathbf{K}, \mathbf{\Gamma}, \alpha) \sim \text{Beta}(1 + \sum_{i=1}^n I_{(K_i=h \& K_i \neq \pi_{N(\mathbf{x}_i)}(\mathbf{x}_i))}, \alpha + \sum_{i=1}^n I_{(K_i > h)})$$

(c) Conditional for $\Gamma_h, h = 1, \dots, N$

$$(\Gamma_h | \mathbf{K}, \mathbf{V}) \sim \text{Uniform}(\max[\max_{i; K_i=h} (x_i - \psi), a_\Gamma], \min[\min_{i; K_i=h} (x_i + \psi), a_\Gamma])$$

(d) Conditional for $\boldsymbol{\beta}_h^*, h = 1, \dots, N$

$$\begin{aligned}
(\boldsymbol{\beta}_h^* | \mathbf{y}, \mathbf{K}, \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta, \tau) &\sim N_p(\hat{\boldsymbol{\mu}}_{\beta h}, \hat{\boldsymbol{\Sigma}}_{\beta h}) \\
\hat{\boldsymbol{\mu}}_{\beta h} &= \hat{\boldsymbol{\Sigma}}_{\beta h} [\boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\mu}_\beta + \tau \mathbf{X}_{ih} \mathbf{y}_{ih}] \\
\hat{\boldsymbol{\Sigma}}_{\beta h} &= [\boldsymbol{\Sigma}_\beta^{-1} + \tau \mathbf{X}_{ih} \mathbf{X}'_{ih}]^{-1},
\end{aligned}$$

where \mathbf{y}_{ih} is $n_h \times 1$ response vector and \mathbf{X}'_{ih} is $n_h \times p$ design matrix for the subjects with $K_i = h$ and n_h is the number of those subjects.

(e) Conditional for $\boldsymbol{\mu}_\beta$

$$\begin{aligned}
(\boldsymbol{\mu}_\beta | \boldsymbol{\beta}^*, \boldsymbol{\Sigma}_\beta) &\sim N_p(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_\mu) \\
\hat{\boldsymbol{\mu}}_0 &= \hat{\boldsymbol{\Sigma}}_\mu [\boldsymbol{\Sigma}_\mu^{-1} \boldsymbol{\mu}_0 + \boldsymbol{\Sigma}_\beta^{-1} \sum_{h=1}^N \boldsymbol{\beta}_h^*] \\
\hat{\boldsymbol{\Sigma}}_\mu &= [\boldsymbol{\Sigma}_\mu^{-1} + N \boldsymbol{\Sigma}_\beta^{-1}]^{-1}
\end{aligned}$$

(f) Conditional for $\boldsymbol{\Sigma}_\beta^{-1}$

$$(\boldsymbol{\Sigma}_\beta^{-1} | \boldsymbol{\beta}^*, \boldsymbol{\mu}_\beta) \sim \text{Wishart}([\sum_{h=1}^N (\boldsymbol{\beta}_h^* - \boldsymbol{\mu}_\beta)(\boldsymbol{\beta}_h^* - \boldsymbol{\mu}_\beta)' + \nu_0 \boldsymbol{\Sigma}_0]^{-1}, N + \nu_0)$$

(g) Conditional for τ

$$(\tau | \mathbf{y}, \boldsymbol{\beta}^*, \mathbf{K}) \sim \text{Gamma}(\nu_1 + \frac{n}{2}, \nu_2 + \frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}_i' \boldsymbol{\beta}_{K_i}^*)^2)$$

(h) Conditional for α

$$(\alpha | \mathbf{V}) \sim \text{Gamma}(\eta_1 + N, \eta_2 - \sum_{h=1}^N \log(1 - V_h))$$

Note that this Gibbs sampling algorithm consists only of simple steps for sampling from standard distributions and is no more complex than blocked Gibbs samplers for DPMs. In addition, we have observed good computational performance, in terms of mixing and convergence rates, in simulated and real data applications.

5. Simulation Examples

We obtained data from two simulated examples, where $n = 500$ and a univariate predictor x_{i1} was simulated from Uniform(0,1). Case 1 was a null case where y_i was generated from a normal regression model $N(y_i; -1 + 2x_{i1}, 0.01)$. Case 2 was 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:

$$f(y_i | \mathbf{x}_i) = e^{-2x_{i1}} N(y_i; x_{i1}, 0.01) + (1 - e^{-2x_{i1}}) N(y_i; x_{i1}^4, 0.04) \quad (16)$$

We applied the IDPM^N model in (13) to the simulated data with N=50. Based on the results, N=50 seems to be chosen to be large enough since the higher clusters having higher indices are not used in any of the subjects or are used in only a small proportion of them. Also, repeating the analysis for twice N and for half of the N, we obtained very similar results, suggesting that the results are robust to the choice of N, as long as N is not chosen to be small.

For the hyperparameters, we let $\nu_1 = \nu_2 = 0.01$, $\eta_1 = \eta_2 = 2$, $\nu_0 = p$, $\boldsymbol{\Sigma}_0 = I_p$, $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_\mu = (\mathbf{X}'\mathbf{X})^{-1}/n$, $a_\Gamma = -0.05$, and $b_\Gamma = 1.05$. The neighborhood size $\psi = 0.05$ was chosen such that the average number of subjects belonging to the neighborhoods around each predictor value in the sample is $\approx n/10$. We analyzed the simulated data using the proposed Gibbs sampling algorithm run for 10,000 iterations with a 5,000 iteration burn-in. The convergence and mixing

of the MCMC algorithm were good (Trace plots not shown). Also, results tended to be robust to repeating analysis with reasonable alternative hyperparameter values.

For case 1, as shown in Figure 3, the predictive mean regression curve (blue dashed, right bottom panel) closely approximated the true linear regression function (red solid), which was entirely enclosed in pointwise 95% credible intervals (green dashed). Figure 3 also shows the predictive densities (blue dashed) of y_{n+1} at the 10th, 25th, 50th, 75th, and 90th percentiles of the empirical distribution of x_{i1} , with these densities almost indistinguishable from the true densities (red solid).

For case 2, Figure 4 shows an x - y plot (right bottom panel) of the data along with the estimated predictive mean curve (blue dashed), which closely follows the true mean curve (red solid). Figure 4 also shows the true densities (red solid), estimated predictive density (blue dashed), and pointwise 95% credible intervals (green dashed) for the empirical percentiles of x_{i1} . The estimated densities correspond approximately to the true densities in each case.

Repeating the analysis for case 2, but with $\beta_i \stackrel{iid}{\sim} G$ 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.

6. Epidemiological Application

6.1 Background and Motivation

In diabetic studies, interest has been on the relationship of 2-hour (post-75 g oral glucose) serum insulin levels with plasma glucose level for many different ethnic groups (Dowse et al., 1992). Although most studies focus on the mean change of insulin response versus glucose level, it would be more interesting to assess the whole distributional change of insulin level across the glucose level.

We obtained data from a study which followed a sample of Pima Indians from a population near Phoenix, Arizona since 1965. This study was conducted by the National Institute of Diabetes and Digestive and Kidney Disease, with the Pima Indians chosen as a group at high risk of diabetes. These data have been widely used to test algorithms for prediction, such as the ADAP learning algorithm (Smith et al., 1988). Here, our focus is quite different in that we are interested in conducting inferences on changes in the insulin response density with changes in glucose level

without making restrictive assumptions, such as normality or a constant residual density. Certainly, it is biologically plausible that the shape of the insulin density should change as glucose level changes, with this density being non-Gaussian having a skewed form.

6.2 Analysis and Results

For woman i ($i = 1, \dots, 393$), let y_i correspond to the 2-hour serum insulin level measured in $\mu\text{U}/\text{ml}$ (micro Units per milliliter) and let x_i denote the 2-hour plasma glucose level measured in mg/dl (milligrams per deciliter). We applied the IDPM^N model described in (15), after standardizing y and x by dividing by 100. Hyperparameters were set to be the same as in the simulation study except that $\psi = 0.08$ such that $n/10$ subjects belong to each neighborhood on average and $a_\Gamma = \min(x_i) - \psi$, and $b_\Gamma = \max(x_i) + \psi$. We analyzed the simulated data using the proposed Gibbs sampling algorithm run for 10,000 iterations with a 5,000 iteration burn-in. The convergence and mixing of the MCMC algorithm were good (Trace plots not shown) and results were robust with reasonable alternative hyperparameter values.

Figure 5 shows the predictive densities for the insulin level at various percentiles of the empirical distribution of glucose level. As the glucose level increases, there is a nonlinear change in the mean insulin level (refer to Figure 4) and a dramatic increase in the heaviness of the right tail of the insulin distribution. This shift in the shape of the insulin distribution is exactly what one would expect biologically. In particular, the different individuals have different sensitivities to the glucose stimuli, with women with severe diabetes expected to be extremely sensitive and hence to fall out further in the right tail. Such heterogeneity in an exposure or treatment response is pervasive in epidemiologic studies, but is not at all well characterized by standard regression models that do not allow the whole density to shift with predictors.

7. Discussion

This article proposed a new stick-breaking prior for the collection of predictor-dependent random probability measures. The prior, called the IDP, is a useful alternative to recently developed prior models that induce predictor-dependence among distributions. In addition to maintaining a marginal DP structure, the IDP is appealing in its simplicity. In particular, the construction

is intuitive and leads to simple expressions for the dependence in random measures at different locations, while also leading to straightforward posterior computation. The marginal DP structure and the simplicity of the construction should also be useful in considering theoretical properties, such as posterior consistency and rates of convergence.

Although we have focused on a conditional density estimation application, there are many interesting applications of the IDP to be considered in future work. First, the DP is widely used to induce a prior on a random partition or clustering structure (Quintana, 2006; Kim et al., 2006). In such settings, the DP has the potential disadvantage of requiring an exchangeability assumption, which may be violated when predictors are available that can inform about the clustering. The IDP provides a straightforward mechanism for local, predictor-dependent clustering, which can be used as an alternative to product partition models (Quintana and Iglesias, 2003) and model-based clustering approaches (Fraley and Raftery, 2002). It is of interest to explore the theoretical properties of the induced prior on the random partition. In this respect, it is likely that the hyperparameter ψ plays a key role. Hence, as a more robust data-driven approach one may consider fully Bayes or empirical Bayes methods for allowing uncertainty in ψ .

Appendix.

Proof of Lemma 1

An infinite number of locations $\Gamma = \{\Gamma_h, h = 1, \dots, \infty\}$ are generated from H on \mathcal{X}' . Any ψ -neighborhood of \mathbf{x} defined as $\eta_{\mathbf{x}}^{\psi} = \{\mathbf{x}' : d(\mathbf{x}, \mathbf{x}') < \psi, \mathbf{x}' \in \mathcal{X}'\}$ with $\psi > 0$ is a subset of \mathcal{X}' . The regularity condition 1 for H ensures that there is a positive probability for a location Γ_h to be generated in any $\eta_{\mathbf{x}}^{\psi}$. Therefore, there are also an infinite number of locations in $\eta_{\mathbf{x}}^{\psi}$ for all $\mathbf{x} \in \mathcal{X}$ and $\psi > 0$, which implies $N(\mathbf{x}) = \infty$. Then, $\sum_{l=1}^{N(\mathbf{x})} p_l(\mathbf{x})$ almost surely by lemma 1 in Ishwaran and James (2001).

Proof of Theorem 1

Assume that $\mathcal{G}_{\mathcal{X}} \sim lDP(\alpha, G_0, H, \psi)$. Then, from the definition of the IDP in (5)-(7), we can

reexpress (7) as follows:

$$G_{\mathbf{x}} = \sum_{l=1}^{N(\mathbf{x})} V_l^{(\mathbf{x})} \prod_{j<l} (1 - V_j^{(\mathbf{x})}) \delta_{\theta_l^{(\mathbf{x})}},$$

where $V_l^{(\mathbf{x})}$ is the l th element of $\mathbf{V}(\mathbf{x})$ and $\theta_l^{(\mathbf{x})}$ is the l th element of $\Theta(\mathbf{x})$. Note that it follows from the proof of Lemma 1 that $N(\mathbf{x}) = \infty$. Since the random weights and atoms are generated by iid sampling from $\text{Beta}(1, \alpha)$ and G_0 , respectively, independently from the location, we have $V_l^{(\mathbf{x})} \stackrel{iid}{\sim} \text{Beta}(1, \alpha)$ independently from $\Theta_l^{(\mathbf{x})} \stackrel{iid}{\sim} G_0$, for $l = 1, \dots, \infty$. Hence, it follows directly from Sethuraman's (1994) representation of the DP, that $G_{\mathbf{x}} \sim DP(\alpha G_0)$, $\forall \mathbf{x} \in \mathcal{X}$.

Proof of Theorem 2

Given Γ and \mathbf{V} ,

$$\begin{aligned} Pr(\phi_i = \phi_j | \mathbf{x}_i, \mathbf{x}_j, \Gamma, \mathbf{V}, \psi) &= \sum_{\{(k,l): \pi_k(\mathbf{x}_i) = \pi_l(\mathbf{x}_j)\}} p_k(\mathbf{x}_i) p_l(\mathbf{x}_j) \\ &= \sum_{h \in \mathcal{L}_{\mathbf{x}_1} \cap \mathcal{L}_{\mathbf{x}_2}} V_h^2 \prod_{m \in \mathcal{S}_h} (1 - V_m)^2 \prod_{m \in \mathcal{S}'_h} (1 - V_m) \end{aligned}$$

Marginalizing out \mathbf{V} over the Beta distribution,

$$Pr(\phi_i = \phi_j | \mathbf{x}_i, \mathbf{x}_j, \Gamma, \alpha, \psi) = \frac{2}{(\alpha + 1)(\alpha + 2)} \sum_{h \in \mathcal{L}_{\mathbf{x}_i} \cap \mathcal{L}_{\mathbf{x}_j}} \left(\frac{\alpha}{\alpha + 2} \right)^{\#\mathcal{S}_h} \left(\frac{\alpha}{\alpha + 1} \right)^{\#\mathcal{S}'_h}$$

For the definition of \mathcal{S}_h and \mathcal{S}'_h , refer to the equation (9) in section 3.2. Marginalizing out \mathcal{S}_h and \mathcal{S}'_h by introducing $Z_{\gamma_j} \stackrel{iid}{\sim} \text{Bernoulli}(P_{\mathbf{x}_i, \mathbf{x}_j})$ (Refer to the formulations from (9) through (10) in section 3.2),

$$\begin{aligned} Pr(\phi_i = \phi_j | \mathbf{x}_i, \mathbf{x}_j, \alpha, \psi) &= \left(\frac{2}{(\alpha + 1)(\alpha + 2)} \right) \left(\frac{P_{\mathbf{x}_i, \mathbf{x}_j}(\alpha + 2)(\alpha + 1)}{\alpha(1 + P_{\mathbf{x}_i, \mathbf{x}_j}) + 2} \right) \\ &= \frac{2P_{\mathbf{x}_i, \mathbf{x}_j}}{(1 + P_{\mathbf{x}_i, \mathbf{x}_j})\alpha + 2} \end{aligned}$$

Proof of Theorem 3

From (10),

$$\begin{aligned} Corr\{G_{\mathbf{x}_1}(B), G_{\mathbf{x}_2}(B) | \Gamma\} &= \frac{2}{\alpha + 2} \sum_{j=1}^{\infty} Z_{\gamma_j} \left(\frac{\alpha}{\alpha + 2} \right)^{\sum_{k=1}^{j-1} Z_{\gamma_k}} \left(\frac{\alpha}{\alpha + 1} \right)^{j-1 - \sum_{k=1}^{j-1} Z_{\gamma_k}} \\ &= \frac{2}{\alpha + 2} \sum_{j=1}^{\infty} Z_{\gamma_j} \left(\frac{\alpha + 1}{\alpha + 2} \right)^{\sum_{k=1}^{j-1} Z_{\gamma_k}} \left(\frac{\alpha}{\alpha + 1} \right)^{j-1} \end{aligned}$$

where Z_{γ_j} are iid draws from Bernoulli($P_{\mathbf{x}_1, \mathbf{x}_2}$). Taking expectation of $\{Z_{\gamma_j}\}_{j=1}^{\infty}$ with respect to Bernoulli($P_{\mathbf{x}_1, \mathbf{x}_2}$),

$$E[Corr\{G_{\mathbf{x}_1}(B), G_{\mathbf{x}_2}(B)\}] = \frac{2}{\alpha+2} P_{\mathbf{x}_1, \mathbf{x}_2} \sum_{j=1}^{\infty} \left(\frac{\alpha}{\alpha+1}\right)^{j-1} E\left[\left(\frac{\alpha+1}{\alpha+2}\right)^{Y_j}\right],$$

where $Y_j \sim \text{Binomial}(j-1, P_{\mathbf{x}_1, \mathbf{x}_2})$. Using the Binomial Theorem, the expectation on the right is marginalized out with respect to Binomial($j-1, P_{\mathbf{x}_1, \mathbf{x}_2}$), which results in

$$\begin{aligned} Corr\{G_{\mathbf{x}_1}(B), G_{\mathbf{x}_2}(B)\} &= \frac{2}{\alpha+2} P_{\mathbf{x}_1, \mathbf{x}_2} \sum_{j=1}^{\infty} \left(\frac{\alpha}{\alpha+1}\right)^{j-1} \left[\left(\frac{\alpha+1}{\alpha+2}\right) P_{\mathbf{x}_1, \mathbf{x}_2} + (1 - P_{\mathbf{x}_1, \mathbf{x}_2})\right]^{j-1} \\ &= \frac{2}{\alpha+2} P_{\mathbf{x}_1, \mathbf{x}_2} \sum_{j=1}^{\infty} \left[\frac{-\alpha P_{\mathbf{x}_1, \mathbf{x}_2}}{(\alpha+2)(\alpha+1)} + \frac{\alpha}{\alpha+1}\right]^{j-1} \end{aligned}$$

Since $\left|\frac{-\alpha P_{\mathbf{x}_1, \mathbf{x}_2}}{(\alpha+2)(\alpha+1)} + \frac{\alpha}{\alpha+1}\right| \leq 0$, the infinite sum on the right converges. Then,

$$\begin{aligned} Corr\{G_{\mathbf{x}_1}(B), G_{\mathbf{x}_2}(B)\} &= \left(\frac{2P_{\mathbf{x}_1, \mathbf{x}_2}}{\alpha+2}\right) \left(\frac{(\alpha+2)(\alpha+1)}{\alpha(1+P_{\mathbf{x}_1, \mathbf{x}_2})+2}\right) \\ &= \frac{2P_{\mathbf{x}_1, \mathbf{x}_2}(\alpha+1)}{(1+P_{\mathbf{x}_1, \mathbf{x}_2})\alpha+2} \end{aligned}$$

Proof of Theorem 4

Due to the marginal DP property and using the inequality on the left in (11) with $n=1$, we get

$$\|\mu_N(\mathbf{x}) - \mu_{\infty}(\mathbf{x})\| \leq 4 \left(1 - E\left[\left(\sum_{h=1}^{N(\mathbf{x})-1} p_h\right)\right]\right),$$

where μ_N , μ_{∞} , and N in (11) are replaced by $\mu_N(\mathbf{x})$, $\mu_{\infty}(\mathbf{x})$, and $N(\mathbf{x})$ and n is substituted by 1.

Here, $N(\mathbf{x})$ is random differently from N in (11). Conditioned on $N(\mathbf{x})$ but marginalizing out p_h ,

we get

$$\begin{aligned} \|\mu_N(\mathbf{x}) - \mu_{\infty}(\mathbf{x})\| &\leq 4 \left(1 - E\left[1 - \left(\frac{\alpha}{1+\alpha}\right)^{N(\mathbf{x})-1}\right]\right) \\ &= 4E\left[\left(\frac{\alpha}{1+\alpha}\right)^{N(\mathbf{x})-1}\right] \end{aligned}$$

Note that $N(\mathbf{x}) \sim \text{Binomial}(N, P_{\mathbf{x}})$ as discussed in section 3.3. Using Binomial Theorem, we obtain

$$\|\mu_N(\mathbf{x}) - \mu_{\infty}(\mathbf{x})\| \leq 4 \left(\frac{\alpha+1}{\alpha}\right) \left[1 - \left(\frac{1}{\alpha+1}\right) P_{\mathbf{x}}\right]^N$$

References

- Beal, M., Ghahramani, Z., and Rasmussen, C. (2002) The infinite hidden Markov model. *Neural Information Processing Systems*, **14**, Cambridge, MA: MIT press.
- Blei, D., Griffiths, T., Jordan, M., and Tenenbaum, J. (2004) Hierarchical topic models and the nested Chinese restaurant process. *Neural Information Processing Systems*, **16**, Cambridge, MA: MIT press.
- Caron, F., Davy, M., Doucet, A., Duflos, E., and Vanheeghe, P. (2006) Bayesian inference for dynamic models with Dirichlet process mixtures. *International Conference on Information Fusion*, Italia, July 10-13.
- 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.
- Duan, J.A., Guidani, M. and Gelfand, A.E. (2005). Generalized spatial Dirichlet process models. *ISDS Discussion Paper*, **05-23**, Duke University, Durham, NC.
- Dunson, D.B. (2005). Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association*, **100**, 618-627.
- Dunson, D.B. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics*, **7**, 551-568.
- Dunson, D.B. and Herring, A.H. (2006). Semiparametric Bayes latent trajectory models. *ISDS Discussion Paper*, **06-16**, Duke University, Durham, NC, USA.
- Dunson, D.B. and Park, J-H. (2006). Kernel Stick-Breaking Process. *ISDS Discussion Paper*, **06-22**, Duke University, Durham, NC, USA.
- Dunson, D.B. and Peddada, S.D. (2007). Bayesian nonparametric inference on stochastic ordering. *ISDS Discussion Paper*, **07-02**, Duke University, Durham, NC, USA.

- Dunson, D.B., Pillai, N., and Park, J-H. (2006). Bayesian density regression. *Journal of the Royal Statistical Society, Series B*, in press.
- 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. (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577-588.
- 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.
- Fraley, C. and Raftery, A.E. (2002) Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, **97**, 611-631.
- Gelfand, A.E., Kottas, A., and MacEachern, S.N. (2004) Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, **100**, 1021-1035.
- Ghosal, S., Ghosh, J.K., and Ramamoorthi, R.V. (1999) Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, **27**, 143-158.
- Ghosal, S. and Van der Vaart, A.W. (2006) Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, in press.
- Griffin, J.E. and Steel, M.F.J. (2006) Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, **101**, 179-194.
- Ishwaran, H. and James, L.F. (2001) Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161-173.

- Kim, S., Tadesse, M. G., and Vannucci, M. (2006) Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, **94**, 877-893.
- Lijoi, A., Prünster, I. and Walker, S.G. (2005) On consistency of non-parametric normal mixtures for Bayesian density estimation. *Journal of the American Statistical Association*, **100**, 1292-1296.
- Lo, A. Y. (1984) On a class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics*, **12**, 351-357.
- 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.
- Müller, P., Quintana, F. and Rosner, G. (2004) A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society B*, **66**, 735-749.
- Pennell, M.L. and Dunson, D.B. (2006) Bayesian semiparametric dynamic frailty models for multiple event time data. *Biometrics*, , **62**, 1044-1052.
- Pitman, J. (1995) Exchangeable and partially exchangeable random partitions. *Probability Theory and 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, IMS Lecture Notes-Monograph Series (Vol. 30), Hayward, CA; Institute of Mathematical Statistics, pp. 245-267.

- Quintana, F.A. (2006) A Predictive View of Bayesian Clustering. *Journal of Statistical Planning and Inference*, **136**, 2407-2429.
- Quintana, F.A. and Iglesias, P.L. (2003) Bayesian Clustering and Product Partition Models. *Journal of the Royal Statistical Society B*, **65**, 557-574.
- Sethuraman, J. (1994) A constructive definition of the Dirichlet process prior. *Statistica Sinica*, **2**, 639-650.
- Xing, E. P., Sharan, R., and Jordan, M. (2004) Bayesian haplotype inference via the Dirichlet process. In *Proceedings of the International Conference on Machine Learning (ICML)*

Figure 1. Graphical Illustration for IDP formulation

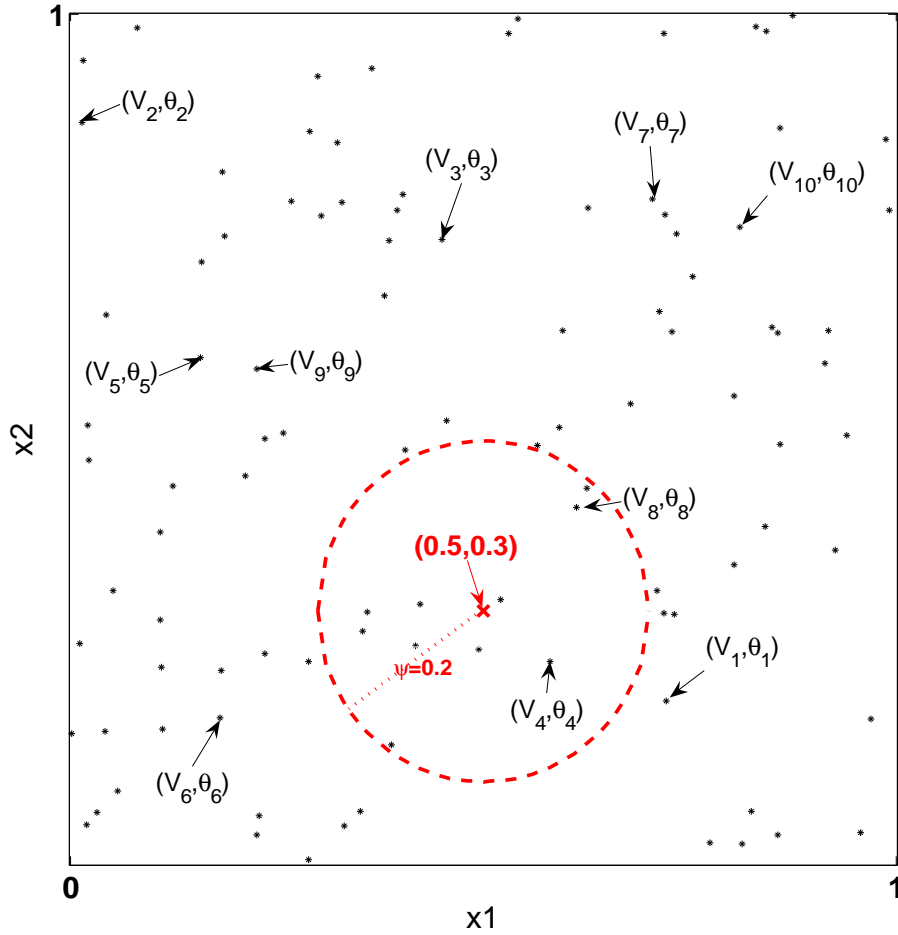


Figure 1: Graphical illustration for the IDP formulation: Black stars are the first 100 random locations generated on $\mathcal{X}' = [0, 1]^2$ from $H = \text{Uniform}([0, 1]^2)$. Red dashed circle indicates the neighborhood of the red crossed predictor point $\mathbf{x} = (0.5, 0.3)'$ determined by Euclidian distance $d(\cdot)$ and $\psi = 0.2$.

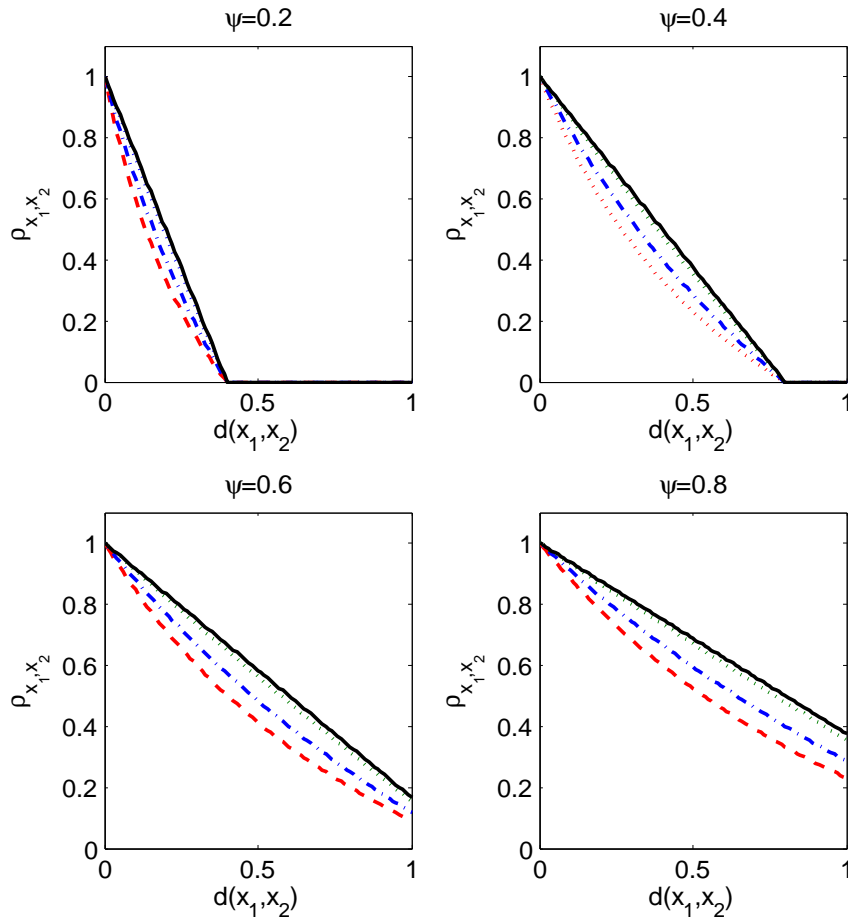


Figure 2: Change in correlation $\rho_{\mathbf{x}_1, \mathbf{x}_2}$ over the change in distance $d(\mathbf{x}_1, \mathbf{x}_2)$ for different α and ψ : $\alpha = 0.0001$ (red dashed), $\alpha = 1$ (blue dot-dashed), $\alpha = 10$ (green dotted), $\alpha = 10000$ (black solid).

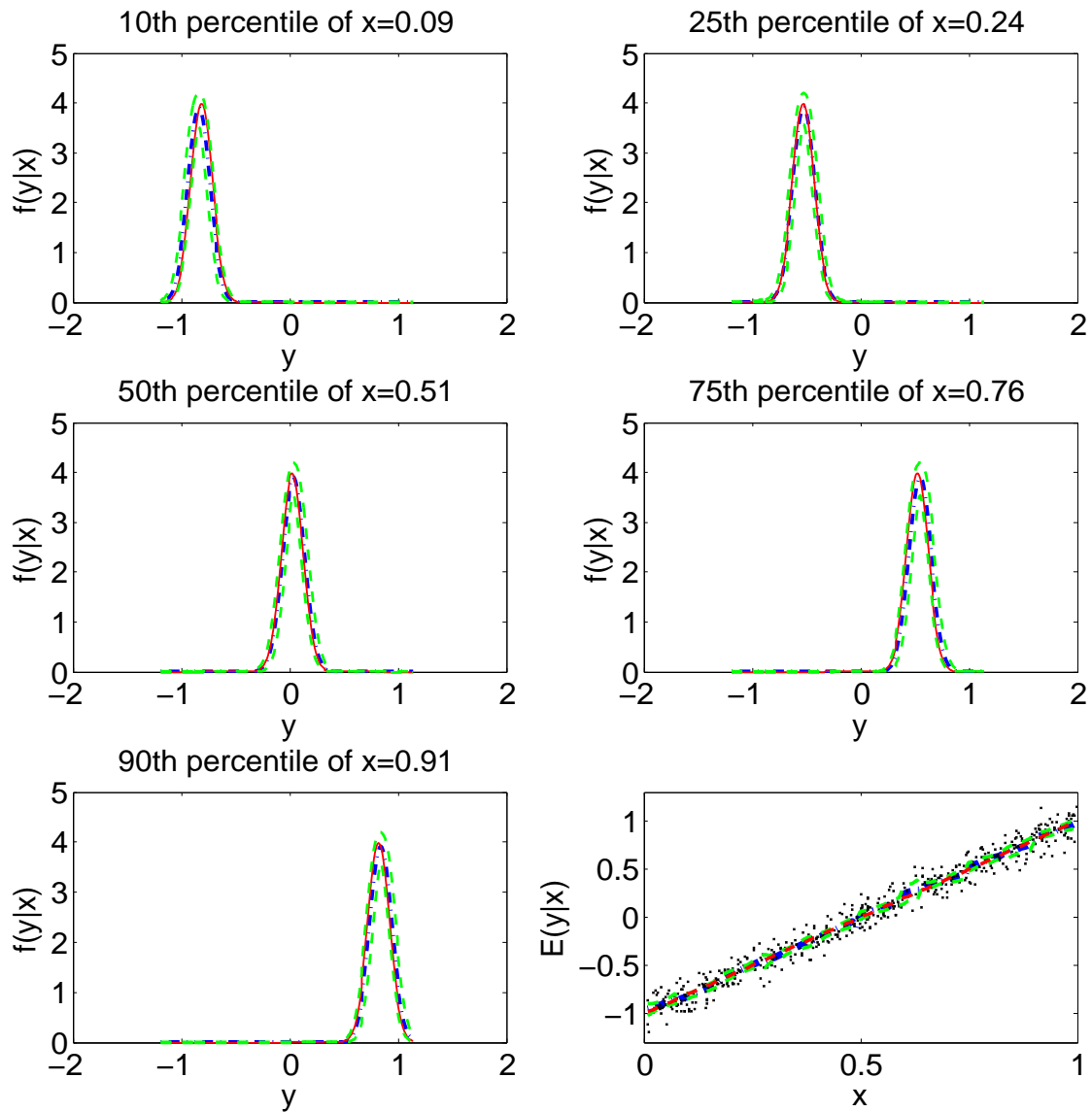


Figure 3: Results for simulation case 1: True conditional densities of $y|x$ (red solid), predictive conditional densities (blue dashed), and 95% pointwise credible intervals (green dashed). The lower right panel shows the data (black dots), along with true (red solid) and estimated mean (blue dashed) regression curves superimposed with 95% credible line (green dashed).

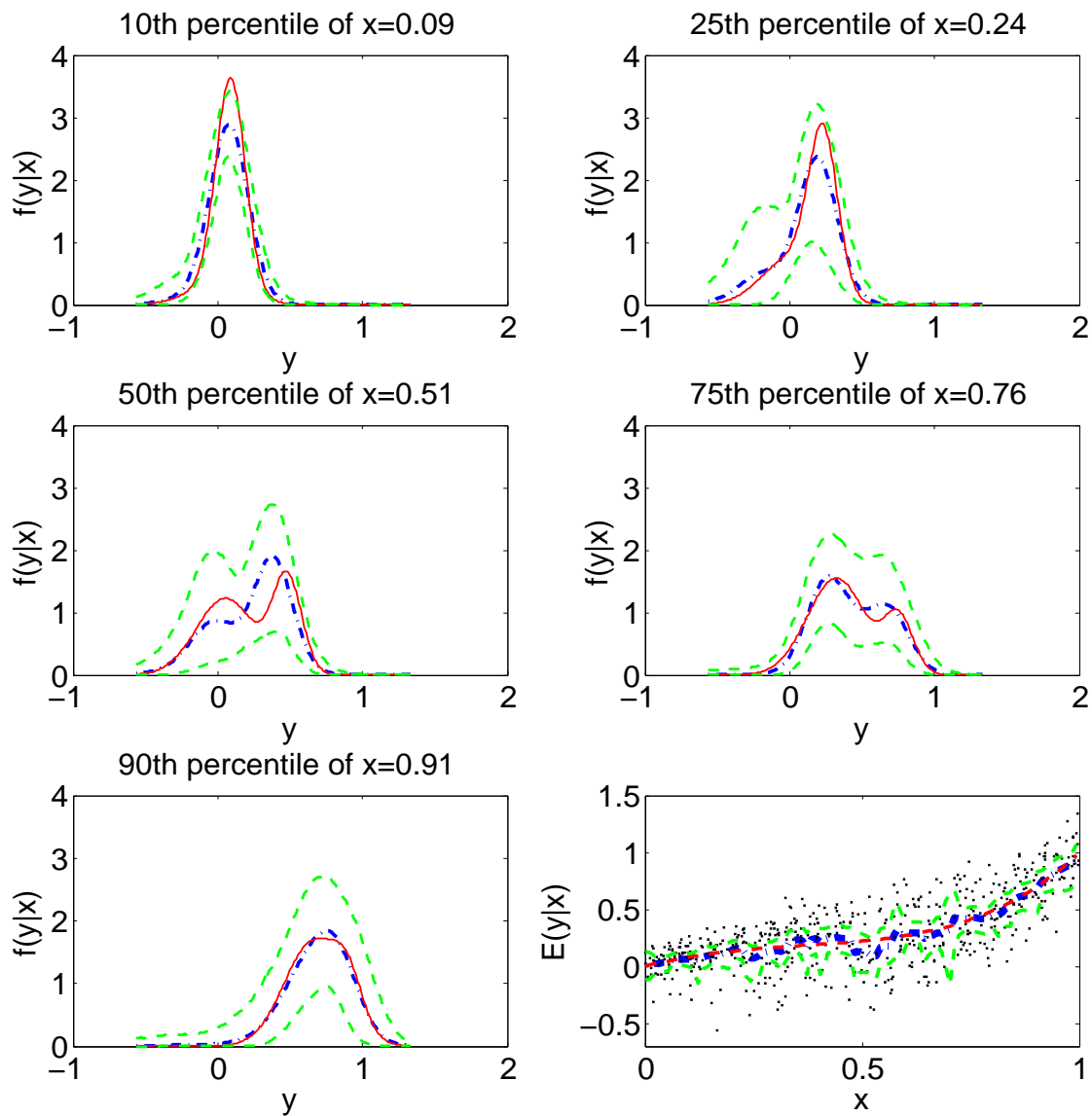


Figure 4: Results for simulation case 2: True conditional densities of $y|x$ (red solid), predictive conditional densities (blue dashed), and 95% pointwise credible intervals (green dashed). The lower right panel shows the data (black dots), along with true (red solid) and estimated mean (blue dashed) regression curves superimposed with 95% credible line (green dashed).

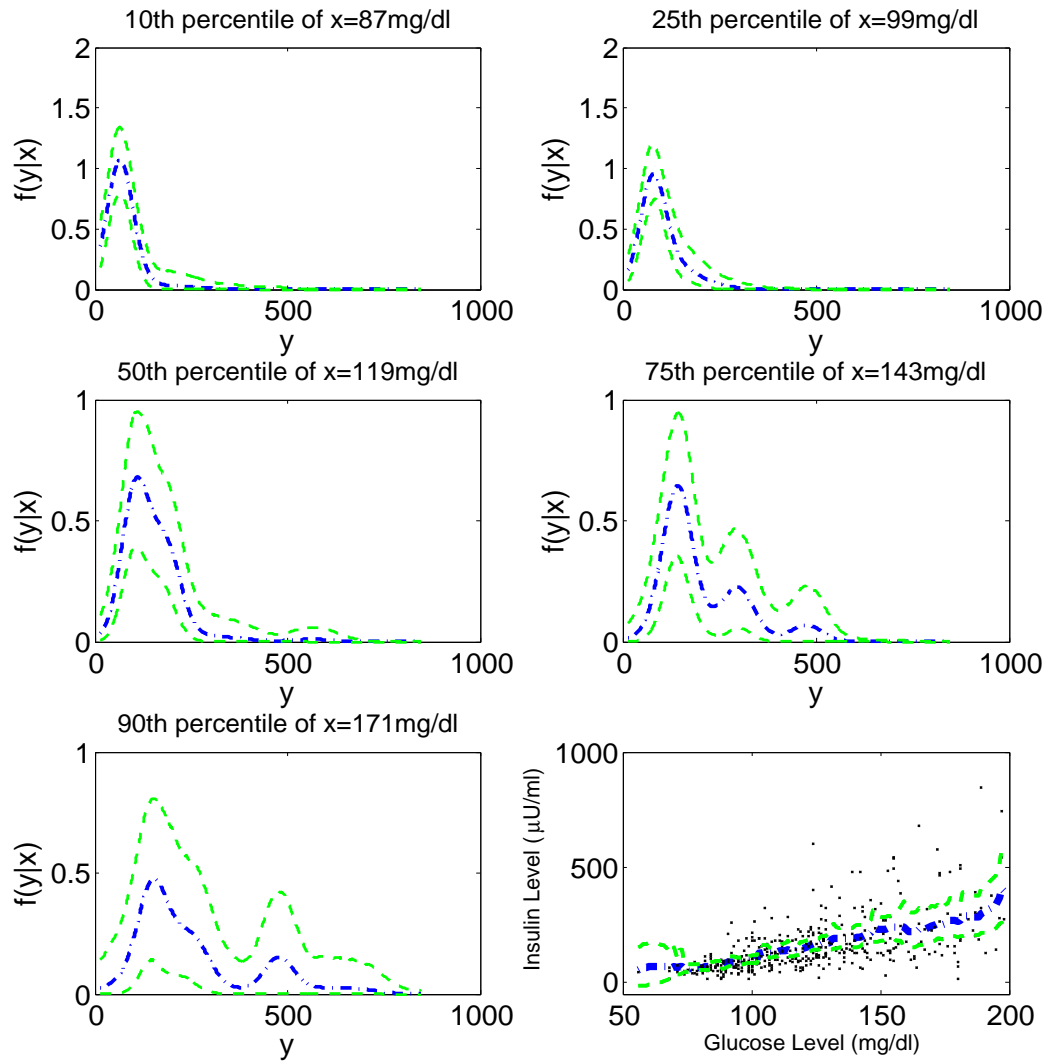


Figure 5: Results for Pima Indian Example: Predictive conditional densities (blue dashed), and 95% pointwise credible intervals (green dashed). The lower right panel shows the data (black dots), along with estimated mean (blue dashed) regression curves superimposed with 95% credible line (green dashed).