Journal of Machine Learning Research 15 (2014) 1041-1071 Submitted 2/13; Revised 11/13; Published 3/14

Improving Prediction from Dirichlet Process Mixtures via Enrichment*†

Sara Wade SARA.WADE@ENG.CAM.AC.UK
Department of Engineering
University of Cambridge
Cambridge, CB2 1PZ, UK

David B. Dunson DUNSON@STAT.DUKE.EDU
Department of Statistical Science
Duke University
Durham, NC 27708-0251, USA

Sonia Petrone SONIA.PETRONE@UNIBOCCONI.IT
Department of Decision Sciences
Bocconi University
Milan, 20136, Italy

Lorenzo Trippa LTRIPPA@JIMMY.HARVARD.EDU
Department of Biostatistics
Harvard University
Boston, MA 02115, USA

Editor: David Blei

Abstract

Flexible covariate-dependent density estimation can be achieved by modelling the joint density of the response and covariates as a Dirichlet process mixture. An appealing aspect of this approach is that computations are relatively easy. In this paper, we examine the predictive performance of these models with an increasing number of covariates. Even for a moderate number of covariates, we find that the likelihood for x tends to dominate the posterior of the latent random partition, degrading the predictive performance of the model. To overcome this, we suggest using a different nonparametric prior, namely an enriched Dirichlet process. Our proposal maintains a simple allocation rule, so that computations remain relatively simple. Advantages are shown through both predictive equations and examples, including an application to diagnosis Alzheimer's disease.

Keywords: Bayesian nonparametrics, density regression, predictive distribution, random partition, urn scheme

∗. For the Alzheimer's Disease Neuroimaging Initiative.

†. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.ucla.edu`). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`.

1. Introduction

Dirichlet process (DP) mixture models have become popular tools for Bayesian nonparametric regression. In this paper, we examine their behavior in prediction and aim to highlight the difficulties that emerge with increasing dimension of the covariate space. To overcome these difficulties, we suggest a simple extension based on a nonparametric prior developed in Wade et al. (2011) that maintains desirable conjugacy properties of the Dirichlet process and leads to improved prediction. The motivating application is to Alzheimer's disease studies, where the focus is prediction of the disease status based on biomarkers obtained from neuroimages. In this problem, a flexible nonparametric approach is needed to account for the possible nonlinear behavior of the response and complex interaction terms, resulting in improved diagnostic accuracy.

DP mixtures are widely used for Bayesian density estimation, see, for example, Ghosal (2010) and references therein. A common way to extend these methods to nonparametric regression and conditional density estimation is by modelling the joint distribution of the response and the covariates (X, Y) as a mixture of multivariate Gaussians (or more general kernels). The regression function and conditional density estimates are indirectly obtained from inference on the joint density, an idea which is similarly employed in classical kernel regression (Scott, 1992, Chapter 8).

This approach, which we call the joint approach, was first introduced by Müller et al. (1996), and subsequently studied by many others including Kang and Ghosal (2009); Shahbaba and Neal (2009); Hannah et al. (2011); Park and Dunson (2010); and Müller and Quintana (2010). The DP model uses simple local linear regression models as building blocks and partitions the observed subjects into clusters, where within clusters, the linear regression model provides a good fit. Even though within clusters the model is parametric, globally, a wide range of complex distributions can describe the joint distribution, leading to a flexible model for both the regression function and the conditional density.

Another related class of models is based on what we term the conditional approach. In such models, the conditional density of Y given x, $f(y|x)$, is modelled directly, as a convolution of a parametric family $f(y|x, \theta)$ with an unknown mixing distribution $\mathbf{P}_x$ for θ. A prior is then given on the family of distributions $\{\mathbf{P}_x, x \in \mathcal{X}\}$ such that the $\mathbf{P}_x$'s are dependent. Examples for the law of $\{\mathbf{P}_x, x \in \mathcal{X}\}$ start from the dependent DPs of MacEachern (1999); Gelfand et al. (2005) and include Griffin and Steel (2006); Dunson and Park (2008); Ren et al. (2011); Chung and Dunson (2009); and Rodriguez and Dunson (2011), just to name a few. Such conditional models can approximate a wide range of response distributions that may change flexibly with the covariates. However, computations are often quite burdensome. One of the reasons the model examined here is so powerful is its simplicity. Together, the joint approach and the clustering of the DP provide a built-in technique to allow for changes in the response distribution across the covariate space, yet it is simple and generally less computationally intensive than the nonparametric conditional models based on dependent DPs.

The random allocation of subjects into groups in joint DP mixture models is driven by the need to obtain a good approximation of the joint distribution of X and Y. This means that subjects with similar covariates and similar relationship between the response and covariates will tend to cluster together. However, difficulties emerge as p, the dimension of

the covariate space, increases. As we will detail, even for moderately large p the likelihood of x tends to dominate the posterior of the random partition, so that clusters are based mainly on similarity in the covariate space. This behaviour is quite unappealing if the marginal density of X is complex, as is typical in high dimensions, because it causes the posterior to concentrate on partitions with many small clusters, as many kernels are needed to describe $f(x)$. This occurs even if the conditional density of Y given x is more well behaved, meaning, a few kernels suffice for its approximation. Typical results include poor estimates of the regression function and conditional density with unnecessarily wide credible intervals due to small clusters and, consequently, poor prediction.

This inefficient performance may not disappear with increasing samples. On one hand, appealing to recent theoretical results (Wu and Ghosal, 2008, 2010; Tokdar, 2011), one could expect that as the sample size increases, the posterior on the unknown density $f(x, y)$ induced by the DP joint mixture model is consistent at the true density. In turn, posterior consistency of the joint is likely to have positive implications for the behavior of the random conditional density and regression function; see Rodriguez et al. (2009); Hannah et al. (2011); and Norets and Pelenis (2012) for some developments in this direction. However, the unappealing behaviour of the random partition that we described above could be reflected in worse convergence rates. Indeed, recent results by Efromovich (2007) suggest that if the conditional density is smoother than the joint, it can be estimated at a faster rate. Thus, improving inference on the random partition to take into account the different degree of smoothness of $f(x)$ and $f(y|x)$ appears to be a crucial issue.

Our goal in this paper is to show that a simple modification of the nonparametric prior on the mixing distribution, that better models the random partition, can more efficiently convey the information present in the sample, leading to more efficient conditional density estimates in term of smaller errors and less variability, for finite samples. To achieve this aim, we consider a prior that allows *local* clustering, that is, the clustering structure for the marginal of X and the regression of Y on x may be different. We achieve this by replacing the DP with the enriched Dirichlet process (EDP) developed in Wade et al. (2011). Like the DP, the EDP is a conjugate nonparametric prior, but it allows a nested clustering structure that can overcome the above issues and lead to improved predictions. An alternative proposal is outlined in Petrone and Trippa (2009) and in unpublished work by Dunson et al. (2011). However, the EDP offers a richer parametrization. In a Bayesian nonparametric framework, several extensions of the DP have been proposed to allow local clustering (see, e.g., Dunson et al. 2008; Dunson 2009; Petrone et al. 2009). However, the greater flexibility is often achieved at the price of more complex computations. Instead, our proposal maintains an analytically computable allocation rule, and therefore, computations are a straightforward extension of those used for the joint DP mixture model. Thus, our main contributions are to highlight the problematic behavior in prediction of the joint DP mixture model for increasing p and also offer a simple solution based on the EDP that maintains computational ease. In addition, we give results on random nested partitions that are implied by the proposed prior.

This paper is organized as follows. In Section 2, we review the joint DP mixture model, discuss the behavior of the random partition, and examine the predictive performance. In Section 3, we propose a joint EDP mixture model, derive its random partition model, and emphasize the predictive improvements of the model. Section 4 covers computational

procedures. We provide a simulated example in Section 5 to demonstrate how the EDP model can lead to more efficient estimators by making better use of information contained in the sample. Finally, in Section 6, we apply the model to predict Alzheimer's disease status based on measurements of various brain structures.

2. Joint DP Mixture Model

A joint DP mixture model for multivariate density estimation and nonparametric regression assumes that

$$(X_i, Y_i)|P \overset{iid}{\sim} f(x, y|P) = \int K(x, y|\xi) dP(\xi),$$

where X_i is p-dimensional, Y is usually univariate, and the mixing distribution P is given a DP prior with scale parameter $\alpha > 0$ and base measure P_0, denoted by $\mathbf{P} \sim \mathrm{DP}(\alpha P_0)$. Due to the a.s. discrete nature of the DP, the model reduces to a countable mixture

$$f(x, y|P) = \sum_{j=1}^{\infty} w_j K(x, y|\tilde{\xi}_j),$$

where the mixing weights w_j have a stick breaking prior with parameter α and $\tilde{\xi}_j \overset{iid}{\sim} P_0$, independently of the w_j. This model was first developed for Bayesian nonparametric regression by Müller et al. (1996), who assume multivariate Gaussian kernels $\mathrm{N}_{p+1}(\mu, \Sigma)$ with $\xi = (\mu, \Sigma)$ and use a conjugate normal inverse Wishart prior as the base measure of the DP. However, for even moderately large p, this approach is practically unfeasible. Indeed, the computational cost of dealing with the full $p+1$ by $p+1$ covariance matrix greatly increases with large p. Furthermore, the conjugate inverse Wishart prior is known to be too poorly parametrized; in particular, there is a single parameter ν to control variability, regardless of p (see Consonni and Veronese, 2001).

A more effective formulation of this model has been recently proposed by Shahbaba and Neal (2009), based on two simple modifications. First, the joint kernel is decomposed as the product of the marginal of X and the conditional of $Y|x$, and the parameter space consequently expressed in terms of the parameters ψ of the marginal and the parameters θ of the conditional. This is a classic reparametrization which, in the Gaussian case, is the basis of generalizations of the inverse Wishart conjugate prior; see Brown et al. (1994). Secondly, they suggest using simple kernels, assuming local independence among the covariates, that is, the covariance matrix of the kernel for X is diagonal. These two simple modifications allow several important improvements. Computationally, reducing the covariance matrix to p variances can greatly ease calculations. Regarding flexibility of the base measure, the conjugate prior now includes a separate parameter to control variability for each of the p variances. Furthermore, the model still allows for local correlations between Y and X through the conditional kernel of $Y|x$ and the parameter θ. In addition, the factorization in terms of the marginal and conditional and the assumption of local independence of the covariates allow for easy inclusion of discrete or other types of response or covariates. Note that even though, within each component, we assume independence of the covariates, globally, there may be dependence.

This extended model, in full generality, can be described through latent parameter vectors as follows:

$$Y_i|x_i,\theta_i \overset{ind}{\sim} F_y(\cdot|x_i,\theta_i), \quad X_i|\psi_i \overset{ind}{\sim} F_x(\cdot|\psi_i), \tag{1}$$
$$(\theta_i,\psi_i)|P \overset{iid}{\sim} P, \quad \mathbf{P} \sim \mathrm{DP}(\alpha P_{0\theta} \times P_{0\psi}).$$

Here the base measure $P_{0\theta} \times P_{0\psi}$ of the DP assumes independence between the θ and the ψ parameters, as this is the structurally conjugate prior for (θ,ψ), and thus, results in simplified computations, but the model could be extended to more general choices. We further assume that $P_{0\theta}$ and $P_{0\psi}$ are absolutely continuous, with densities $p_{0\theta}$ and $p_{0\psi}$. Since $\mathbf{P}$ is discrete a.s., integrating out the subject-specific parameters (θ_i,ψ_i), the model for the joint density is

$$f(y_i,x_i|P) = \sum_{j=1}^{\infty} w_j K(y_i|x_i,\tilde{\theta}_j)K(x_i|\tilde{\psi}_j), \tag{2}$$

where the kernels $K(y|x,\theta)$ and $K(x|\psi)$ are the densities associated to $F_y(\cdot|x,\theta)$ and $F_x(\cdot|\psi)$. In the Gaussian case, $K(x|\psi)$ is the product of $\mathrm{N}(\mu_{x,h},\sigma^2_{x,h})$ Gaussians, $h=1,\ldots,p$, and $K(y|x,\theta)$ is $\mathrm{N}(\underline{x}\beta,\sigma^2_{y|x})$, where $\underline{x}=(1,x')$. Shahbaba and Neal (2009) focus on the case when Y is categorical and the local model for $Y|x$ is a multinomial logit. Hannah et al. (2011) extend the model to the case when, locally, the conditional distribution of $Y|x$ belongs to the class of generalized linear models (GLM), that is, the distribution of the response belongs to the exponential family and the mean of the response can be expressed a function of a linear combination of the covariates.

Model (2) allows for flexible conditional densities

$$f(y|x,P) \;=\; \frac{\sum_{j=1}^{\infty} w_j K(x|\tilde{\psi}_j)K(y|x,\tilde{\theta}_j)}{\sum_{j'=1}^{\infty} w_{j'} K(x|\tilde{\psi}_{j'})} \equiv \sum_j w_j(x)K(y|x,\tilde{\theta}_j),$$

and nonlinear regression

$$\mathrm{E}[Y|x,P] = \sum_{j=1}^{\infty} w_j(x)\,\mathrm{E}[Y|x,\tilde{\theta}_j],$$

with $\mathrm{E}[Y|x,\tilde{\theta}_j] = \underline{x}\beta^*_j$ for Gaussian kernels. Thus, the model provides flexible kernel based density and regression estimation, and MCMC computations are standard. However, the DP only allows joint clusters of the parameters (θ_i,ψ_i), $i=1,\ldots,n$. We underline drawbacks of such joint clustering in the following subsections.

2.1 Random Partition and Inference

One of the crucial features of DP mixture models is the dimension reduction and clustering obtained due to the almost sure discreteness of $\mathbf{P}$. In fact, this implies that a sample (θ_i,ψ_i), $i=1,\ldots,n$ from a DP presents ties with positive probability and can be conveniently described in terms of the random partition and the distinct values. Using a standard notation, we denote the random partition by a vector of cluster allocation labels $\rho_n =$

$(s_1, \ldots, s_n)$, with $s_i = j$ if (θ_i, ψ_i) is equal to the j^{th} unique value observed, (θ_j^*, ψ_j^*), for $j = 1, \ldots, k$, where $k = k(\rho_n)$ is the number of groups in the partition ρ_n. Additionally, we will denote by $S_j = \{i : s_i = j\}$ the set of subject indices in the j^{th} cluster and use the notation $y_j^* = \{y_i : i \in S_j\}$ and $x_j^* = \{x_i : i \in S_j\}$. We also make use of the short notation $x_{1:n} = (x_1, \ldots, x_n)$.

Often, the random probability measure $\mathbf{P}$ is integrated out and inference is based on the posterior of the random partition ρ_n and the cluster-specific parameters $(\theta^*, \psi^*) \equiv (\theta_j^*, \psi_j^*)_{j=1}^k$;

$$p(\rho_n, \theta^*, \psi^* | x_{1:n}, y_{1:n}) \propto p(\rho_n) \prod_{j=1}^{k} p_{0\theta}(\theta_j^*) p_{0\psi}(\psi_j^*) \prod_{j=1}^{k} \prod_{i \in S_j} K(x_i|\psi_j^*) K(y_i|x_i, \theta_j^*).$$

As is well known (Antoniak, 1974), the prior induced by the DP on the random partition is $p(\rho_n) \propto \alpha^k \prod_{j=1}^k \Gamma(n_j)$, where n_j is the size of S_j. Marginalizing out the θ^*, ψ^*, the posterior of the random partition is

$$p(\rho_n | x_{1:n}, y_{1:n}) \propto \alpha^k \prod_{j=1}^{k} \Gamma(n_j)\, h_x(x_j^*)\, h_y(y_j^* | x_j^*). \tag{3}$$

Notice that the marginal likelihood component in (3), say $h(x_{1:n}, y_{1:n}|\rho_n)$, is factorized as the product of the cluster-specific marginal likelihoods $h_x(x_j^*) h_y(y_j^*|x_j^*)$ where $h_x(x_j^*) = \int_\Psi \prod_{i \in S_j} K(x_i|\psi) dP_{0\psi}(\psi)$ and $h_y(y_j^*|x_j^*) = \int_\Theta \prod_{i \in S_j} K(y_i|x_i, \theta) dP_{0\theta}(\theta)$.

Given the partition and the data, the conditional distribution of the distinct values is simple, as they are independent across clusters, with posterior densities

$$\begin{aligned} p(\theta_j^* | \rho_n, x_{1:n}, y_{1:n}) &\propto p_{0\theta}(\theta_j^*) \prod_{i \in S_j} K(y_i|x_i, \theta_j^*), \\ p(\psi_j^* | \rho_n, x_{1:n}, y_{1:n}) &\propto p_{0\psi}(\psi_j^*) \prod_{i \in S_j} K(x_i|\psi_j^*). \end{aligned} \tag{4}$$

Thus, given ρ_n, the cluster-specific parameters (θ_j^*, ψ_j^*) are updated based only on the observations in cluster S_j. Computations are simple if $p_{0\theta}$ and $p_{0\psi}$ are conjugate priors.

The above expressions show the crucial role of the random partition. From Equation (3), we have that given the data, subjects are clustered in groups with similar behaviour in the covariate space and similar relationship with the response. However, even for moderate p the likelihood for x tends to dominate the posterior of the random partition, so that clusters are determined only by similarity in the covariate space. This is particularly evident when the covariates are assumed to be independent locally, that is, $K(x_i|\psi_j^*) = \prod_{h=1}^p K(x_{i,h}|\psi_{j,h}^*)$. Clearly, for large p, the scale and magnitude of changes in $\prod_{h=1}^p K(x_{i,h}|\psi_{j,h}^*)$ will wash out any information given in the univariate likelihood $K(y_i|\theta_j^*, x_i)$. Indeed, this is just the behavior we have observed in practice in running simulations for large p (results not shown).

For a simple example demonstrating how the number of components needed to approximate the marginal of X can blow up with p, imagine X is uniformly distributed on a cuboid of side length $r > 1$. Consider approximating

$$f_0(x) = \frac{1}{r^p} \mathbf{1}(x \in [0, r]^p) \quad \text{by} \quad f_k(x) = \sum_{j=1}^{k} w_j \mathrm{N}_p(x|\mu_j, \sigma_j^2 I_p).$$

Since the true distribution of x is uniform on the cube $[0,r]^p$, to obtain a good approximation, the weighted components must place most of their mass on values of x contained in the cuboid. Let $B_\sigma(\mu)$ denote a ball of radius σ centered at μ. If a random vector V is normally distributed with mean μ and variance $\sigma^2 I_p$, then for $0<\epsilon<1$,

$$P(V \in B_{\sigma z(\epsilon)}(\mu)) = 1-\epsilon, \quad \text{where} \quad z(\epsilon)^2 = (\chi^2_p)^{-1}(1-\epsilon),$$

that is, the square of $z(\epsilon)$ is the $(1-\epsilon)$ quantile of the chi-squared distribution with p degrees of freedom. For small ϵ, this means that the density of V places most of its mass on values contained in a ball of radius $\sigma z(\epsilon)$ centered at μ. For $\epsilon>0$, define

$$\tilde{f}_k(x) = \sum_{j=1}^{k} w_j \mathrm{N}(x;\mu_j,\sigma_j^2 I_p) * \mathbf{1}(x \in B_{\sigma_j z(\epsilon_j)}(\mu_j)),$$

where $\epsilon_j = \epsilon/(kw_j)$. Then, $\tilde{f}_k$ is close to f_k (in the L_1 sense):

$$\int_{\mathbb{R}^p} |f_k(x) - \tilde{f}_k(x)|dx = \int_{\mathbb{R}^p} \sum_{j=1}^{k} w_j \mathrm{N}(x;\mu_j,\sigma_j^2 I_p) * \mathbf{1}(x \in B^c_{\sigma_j z(\epsilon_j)}(\mu_j))dx = \epsilon.$$

For $\tilde{f}_k$ to be close to f_0, the parameters μ_j, σ_j, w_j need to be chosen so that the balls $B_{\sigma_j z(\epsilon/(kw_j))}(\mu_j)$ are contained in the cuboid. That means that centers of the balls are contained in the cuboid,

$$\mu_j \in [0,r]^p, \tag{5}$$

with further constraints on σ_j^2 and w_j, so that the radius is small enough. In particular,

$$\sigma_j z\left(\frac{\epsilon}{kw_j}\right) \leq \min(\mu_1, r-\mu_1, \ldots, \mu_p, r-\mu_p) \leq \frac{r}{2}. \tag{6}$$

However, as p increases the volume of the cuboid goes to infinity, but the volume of any ball $B_{\sigma_j z(\epsilon/(kw_j))}(\mu_j)$ defined by (5) and (6) goes to 0 (see Clarke et al., 2009, Section 1.1). Thus, just to reasonably cover the cuboid with the balls of interest, the number of components will increase dramatically, and more so, when we consider the approximation error of the density estimate. Now, as an extreme example, imagine that $f_0(y|x)$ is a linear regression model. Even though one component is sufficient for $f_0(y|x)$, a large number of components will be required to approximate $f_0(x)$, especially as p increases.

It appears evident from (4) this behavior of the random partition also negatively affects inference on the cluster-specific parameters. In particular, when many kernels are required to approximate the density of X with few observations within each cluster, the posterior for θ_j^* may be based on a sample of unnecessarily small size, leading to a flat posterior with an unreliable posterior mean and large influence of the prior.

2.2 Covariate-dependent Urn Scheme and Prediction

Difficulties associated to the behavior of the random partition also deteriorate the predictive performance of the model. Prediction in DP joint mixture models is based on a covariate-dependent urn scheme (Park and Dunson, 2010; Müller and Quintana, 2010), such that

conditionally on the partition ρ_n and $x_{1:n+1}$, the cluster allocation s_{n+1} of a new subject with covariate value x_{n+1} is determined as

$$s_{n+1}|\rho_n, x_{1:n+1} \sim \frac{\omega_{k+1}(x_{n+1})}{c_0}\,\delta_{k+1} + \sum_{j=1}^{k} \frac{\omega_j(x_{n+1})}{c_0}\delta_j, \tag{7}$$

where

$$\omega_{k+1}(x_{n+1}) = \frac{\alpha\, h_{0,x}(x_{n+1})}{\alpha+n},$$
$$\omega_j(x_{n+1}) = \frac{n_j \int K(x_{n+1}|\psi)p(\psi|x_j^*)d\psi}{\alpha+n} \equiv \frac{n_j\, h_{j,x}(x_{n+1})}{\alpha+n},$$

and $c_0 = p(x_{n+1}|\rho_n, x_{1:n})$ is the normalizing constant. This urn scheme is a generalization of the classic Pólya urn scheme that allows the cluster allocation probability to depend on the covariates; the probability of allocation to cluster j depends on the similarity of x_{n+1} to the x_i in cluster j as measured by the predictive density $h_{j,x}$. See Park and Dunson (2010) for more details.

From the urn scheme (7) one obtains the structure of the prediction. The predictive density at y for a new subject with a covariate of x_{n+1} is computed as

$$\begin{aligned}
&f(y|y_{1:n}, x_{1:n+1}) \\
&= \sum_{\rho_n}\sum_{s_{n+1}} f(y|y_{1:n}, x_{1:n+1}, \rho_n, s_{n+1})p(s_{n+1}|y_{1:n}, x_{1:n+1}, \rho_n)\, p(\rho_n|y_{1:n}, x_{1:n+1}) \\
&= \sum_{\rho_n}\left(\frac{\omega_{k+1}(x_{n+1})}{c_0}\, h_{0,y}(y|x_{n+1}) + \sum_{j=1}^{k}\frac{\omega_j(x_{n+1})}{c_0}\, h_{j,y}(y|x_{n+1})\right)\frac{c_0 p(\rho_n|x_{1:n}, y_{1:n})}{p(x_{n+1}|x_{1:n}, y_{1:n})} \\
&= \sum_{\rho_n}\left(\frac{\omega_{k+1}(x_{n+1})}{c}\, h_{0,y}(y|x_{n+1}) + \sum_{j=1}^{k}\frac{\omega_j(x_{n+1})}{c}\, h_{j,y}(y|x_{n+1})\right) p(\rho_n|x_{1:n}, y_{1:n}),
\end{aligned} \tag{8}$$

where $c = p(x_{n+1}|x_{1:n}, y_{1:n})$. Thus, given the partition, the conditional predictive density is a weighted average of the prior guess $h_{0,y}(y|x) \equiv \int K(y|x,\theta)dP_{0\theta}(\theta)$ and the cluster-specific predictive densities of y at x_{n+1},

$$h_{j,y}(y|x_{n+1}) = \int K(y|x_{n+1},\theta)p(\theta|x_j^*, y_j^*)d\theta,$$

with covariate-dependent weights. The predictive density is obtained by averaging with respect to the posterior of ρ_n. However, for moderate to large p, the posterior of the random partition suffers the drawbacks discussed in the previous subsection. In particular, too many small x-clusters lead to unreliable within cluster predictions based on small sample sizes. Furthermore, the measure which determines similarity of x_{n+1} and the j^{th} cluster will be too rigid. Consequently, the resulting overall prediction may be quite poor.

These drawbacks will also affect the point prediction, which, under quadratic loss, is

$$\mathrm{E}[Y_{n+1}|y_{1:n}, x_{1:n+1}]$$
$$= \sum_{\rho_n} \left(\frac{\omega_{k+1}(x_{n+1})}{c} \mathrm{E}_0[Y_{n+1}|x_{n+1}] + \sum_{j=1}^{k} \frac{\omega_j(x_{n+1})}{c} \mathrm{E}_j[Y_{n+1}|x_{n+1}] \right) p(\rho_n|x_{1:n}, y_{1:n}), \quad (9)$$

where $\mathrm{E}_0[Y_{n+1}|x_{n+1}]$ is the expectation of Y_{n+1} with respect to $h_{0,y}$ and $\mathrm{E}_j[Y_{n+1}|x_{n+1}] \equiv \mathrm{E}[\mathrm{E}[Y_{n+1}|x_{n+1}, \theta_j^*]|x_j^*, y_j^*]$ is the expectation of Y_{n+1} with respect to $h_{j,y}$.
Example. When $K(y|x, \theta) = \mathrm{N}(y; \underline{x}\beta, \sigma^2)$ and the prior for (β, σ^2) is the multivariate normal inverse gamma with parameters $(\beta_0, C^{-1}, a_y, b_y)$, (9) is

$$\sum_{\rho_n} \left(\frac{\omega_{k+1}(x_{n+1})}{c} \underline{x}_{n+1}\beta_0 + \sum_{j=1}^{k} \frac{\omega_j(x_{n+1})}{c} \underline{x}_{n+1}\hat{\beta}_j \right) p(\rho_n|x_{1:n}, y_{1:n}),$$

where $\hat{\beta}_j = \hat{C}_j^{-1}(C\beta_0 + \underline{X}_j' y_j^*)$, $\hat{C}_j = C + \underline{X}_j'\underline{X}_j$, $\underline{X}_j$ is a n_j by $p+1$ matrix with rows $\underline{x}_i$ for $i \in S_j$, and (8) is

$$\frac{\omega_{k+1}(x_{n+1})}{c} \mathcal{T}\left(y|\underline{x}_{n+1}\beta_0, \frac{b_y}{a_y} W^{-1}, 2a_y\right) + \sum_{j=1}^{k} \frac{\omega_j(x_{n+1})}{c} \mathcal{T}\left(y|\underline{x}_{n+1}\hat{\beta}_j, \frac{\hat{b}_{y,j}}{\hat{a}_{y,j}} W_j^{-1}, 2\hat{a}_{y,j}\right),$$

where $\mathcal{T}(\cdot; \mu, \sigma^2, \nu)$ denotes the density of a random variable V such that $(V-\mu)/\sigma$ has a t-distribution with ν degrees of freedom; $\hat{a}_{y,j} = a_y + n_j/2$;

$$\hat{b}_{y,j} = b_y + \frac{1}{2}(y_j^* - \underline{X}_j\beta_0)'(I_{n_j} - \underline{X}_j\hat{C}_j^{-1}\underline{X}_j')(y_j^* - \underline{X}_j\beta_0);$$

$$W = 1 - \underline{x}_{n+1}(C + \underline{x}_{n+1}'\underline{x}_{n+1})^{-1}\underline{x}_{n+1}';$$

and W_j is defined as W with C replaced by $\hat{C}_j$.

3. Joint EDP Mixture Model

As seen, the global clustering of the DP prior on $\mathcal{P}(\Theta \times \Psi)$, the space of probability measures on $\Theta \times \Psi$, does not allow one to efficiently model the types of data discussed in the previous section. Instead, it is desirable to use a nonparametric prior that allows many ψ-clusters, to fit the complex marginal of X, and fewer θ-clusters. At the same time, we want to preserve the desirable conjugacy properties of the DP, in order to maintain fairly simple computations. To these aims, our proposal is to replace the DP with the more richly parametrized *enriched Dirichlet process* (Wade et al., 2011). The EDP is conjugate and has an analytically computable urn scheme, but it gives a nested partition structure that can model the desired clustering behavior.

Recall that the model (1) was obtained by decomposing the joint kernel as the product of the marginal and conditional kernels. The EDP is a natural alternative for the mixing distribution of this model, as it is similarly based on the idea of expressing the unknown

random joint probability measure $\mathbf{P}$ of (θ, ψ) in terms of the random marginal and conditionals. This requires the choice of an ordering of θ and ψ, and this choice is problem specific. In the situation described here, it is natural to consider the random marginal distribution $\mathbf{P}_\theta$ and the random conditional $\mathbf{P}_{\psi|\theta}$, to obtain the desired clustering structure. Then, the EDP prior is defined by

$$\begin{aligned} \mathbf{P}_\theta &\sim \mathrm{DP}(\alpha_\theta P_{0\theta}), \\ \mathbf{P}_{\psi|\theta}(\cdot|\theta) &\sim \mathrm{DP}(\alpha_\psi(\theta) P_{0\psi|\theta}(\cdot|\theta)), \quad \forall \theta \in \Theta, \end{aligned}$$

and $\mathbf{P}_{\psi|\theta}(\cdot|\theta)$ for $\theta \in \Theta$ are independent among themselves and from $\mathbf{P}_\theta$. Together these assumptions induce a prior for the random joint $\mathbf{P}$ through the joint law of the marginal and conditionals and the mapping $(P_\theta, P_{\psi|\theta}) \to \int P_{\psi|\theta}(\cdot|\theta) dP_\theta(\theta)$. The prior is parametrized by the base measure P_0, expressed as

$$P_0(A \times B) = \int_A P_{0\psi|\theta}(B|\theta) dP_{0\theta}(\theta)$$

for all Borel sets A and B, and by a precision parameter α_θ associated to θ and a collection of precision parameters $\alpha_\psi(\theta)$ for every $\theta \in \Theta$ associated to $\psi|\theta$. Note the contrast with the DP, which only allows one precision parameter to regulate the uncertainty around P_0.

The proposed EDP mixture model for regression is as in (1), but with

$$\mathbf{P} \sim \mathrm{EDP}(\alpha_\theta, \alpha_\psi(\theta), P_0)$$

in place of $\mathbf{P} \sim \mathrm{DP}(\alpha P_{0\theta} \times P_{0\psi})$. In general, P_0 is such that θ and ψ are dependent, but here we assume the same structurally conjugate base measure as for the DP model (1), so $P_0 = P_{0\theta} \times P_{0\psi}$. Using the square breaking representation of the EDP (Wade et al., 2011, Proposition 4) and integrating out the (θ_i, ψ_i) parameters, the model for the joint density is

$$f(x, y|P) = \sum_{j=1}^{\infty} \sum_{l=1}^{\infty} w_j w_{l|j} K(x|\tilde{\psi}_{l|j}) K(y|x, \tilde{\theta}_j).$$

This gives a mixture model for the conditional densities with more flexible weights

$$f(y|x, P) = \sum_{j=1}^{\infty} \frac{\sum_{l=1}^{\infty} w_{l|j} K(x|\tilde{\psi}_{l|j})}{\sum_{j'=1}^{\infty} w_{j'} \sum_{l'=1}^{\infty} w_{l'|j'} K(x|\tilde{\psi}_{l'|j'})} K(y|x, \tilde{\theta}_{j'}) \equiv \sum_{j=1}^{\infty} \tilde{w}_j(x) K(y|x, \tilde{\theta}_j).$$

3.1 Random Partition and Inference

The advantage of the EDP is the implied nested clustering. The EDP partitions subjects in θ-clusters and ψ-subclusters within each θ-cluster, allowing the use of more kernels to describe the marginal of X for each kernel used for the conditional of $Y|x$. The random partition model induced from the EDP can be described as a nested Chinese Restaurant Process (nCRP).

First, customers choose restaurants according to the CRP induced by $\mathbf{P}_\theta \sim \mathrm{DP}(\alpha_\theta P_{0\theta})$, that is, with probability proportional to the number n_j of customers eating at restaurant j,

the $(n+1)^{\text{th}}$ customer eats at restaurant j, and with probability proportional to α_θ, she eats at a new restaurant. Restaurant are then colored with colors $\theta_j^* \overset{iid}{\sim} P_{0\theta}$. Within restaurant j, customers sit at tables as in the CRP induced by the $\mathbf{P}_{\psi|\theta_j^*} \sim \text{DP}(\alpha_\psi(\theta_j^*)P_{0\psi|\theta}(\cdot|\theta_j^*))$. Tables in restaurant j are then colored with colors $\psi_{l|j}^* \overset{iid}{\sim} P_{0\psi|\theta}(\cdot|\theta_j^*)$.

This differs from the nCRP proposed by Blei et al. (2010), which had the alternative aim of learning topic hierarchies by clustering parameters (topics) hierarchically along a tree of infinite CRPs. In particular, each subject follows a path down the tree according to a sequence of nested CRPs and the parameters of subject i are associated with the cluster visited at a latent subject-specific level l of this path. Although related, the EDP is not a special case with the tree depth fixed to 2; the EDP defines a prior on a multivariate random probability measure on $\Theta \times \Psi$ and induces a nested partition of the multivariate parameter (θ, ψ), where the first level of the tree corresponds to the clustering of the θ parameters and the second corresponds to the clustering of the ψ parameters. A generalization of the EDP to a depth of $D \leq \infty$ is related to Blei et al.'s nCRP with depth $D \leq \infty$, but only if one regards the parameters of each subject as the vector of parameters $(\xi_1, \ldots, \xi_D)$ associated to each level of the tree. Furthermore, this generalization of the EDP would allow a more flexible specification of the mass parameters and possible correlation among the nested parameters.

The nested partition of the EDP is described by $\rho_n = (\rho_{n,y}, \rho_{n,x})$, where $\rho_{n,y} = (s_{y,1}, ..., s_{y,n})$ and $\rho_{n,x} = (s_{x,1}, ..., s_{x,n})$ with $s_{y,i} = j$ if $\theta_i = \theta_j^*$, the j^{th} distinct θ-value in order of appearance, and $s_{x,i} = l$ if $\psi_i = \psi_{l|j}^*$, the l^{th} color that appeared inside the j^{th} θ-cluster. Additionally, we use the notation $S_{j+} = \{i : s_{y,i} = j\}$, with size n_j, $j = 1, \ldots, k$, and $S_{l|j} = \{i : s_{y,i} = j, s_{x,i} = l\}$, with size $n_{l|j}$, $l = 1, \ldots, k_j$. The unique parameters will be denoted by $\theta^* = (\theta_j^*)_{j=1}^k$ and $\psi^* = (\psi_{1|j}^*, \ldots, \psi_{k_j|j}^*)_{j=1}^k$. Furthermore, we use the notation $\rho_{n_j,x} = (s_{x,i} : i \in S_{j+})$ and $y_j^* = \{y_i : i \in S_{j+}\}$, $x_j^* = \{x_i : i \in S_{j+}\}$, $x_{l|j}^* = \{x_i : i \in S_{l|j}\}$.

Proposition 1 *The probability law of the nested random partition defined from the EDP is*

$$p(\rho_n) = \frac{\Gamma(\alpha_\theta)}{\Gamma(\alpha_\theta + n)} \alpha_\theta^k \prod_{j=1}^k \int_\Theta \alpha_\psi(\theta)^{k_j} \frac{\Gamma(\alpha_\psi(\theta))\Gamma(n_j)}{\Gamma(\alpha_\psi(\theta) + n_j)} dP_{0\theta}(\theta) \prod_{l=1}^{k_j} \Gamma(n_{l|j}).$$

Proof From independence of random conditional distributions among $\theta \in \Theta$,

$$p(\rho_n, \theta^*) = p(\rho_{n,y}) \prod_{j=1}^k p_{0\theta}(\theta_j^*) p(\rho_{n,x}|\rho_{n,y}, \theta^*) = p(\rho_{n,y}) \prod_{j=1}^k p_{0\theta}(\theta_j^*) p(\rho_{n_j,x}|\theta_j^*).$$

Next, using the results of the random partition model of the DP (Antoniak, 1974), we have

$$p(\rho_n, \theta^*) = \frac{\Gamma(\alpha_\theta)}{\Gamma(\alpha_\theta + n)} \alpha_\theta^k \prod_{j=1}^k p_{0\theta}(\theta_j^*) \alpha_\psi(\theta_j^*)^{k_j} \frac{\Gamma(\alpha_\psi(\theta_j^*))\Gamma(n_j)}{\Gamma(\alpha_\psi(\theta_j^*) + n_j)} \prod_{l=1}^{k_j} \Gamma(n_{l|j}).$$

Integrating out θ^* leads to the result. ■

From Proposition 1, we gain an understanding of the types of partitions preferred by the

EDP and the effect of the parameters. If for all θ, $\alpha_\psi(\theta) = \alpha_\theta P_{0\theta}(\{\theta\})$, that is $\alpha_\psi(\theta) = 0$ if $P_{0\theta}$ is non-atomic, we are back to the DP random partition model, see Proposition 2 of Wade et al. (2011). In the case when $P_{0\theta}$ is non-atomic, this means that the conditional $\mathbf{P}_{\psi|\theta}$ is degenerate at some random location with probability one (for each restaurant—one table).

In general, $\alpha_\psi(\theta)$ may be a flexible function of θ, reflecting the fact that within some θ-clusters more kernels may be required for good approximation of the marginal of X. In practice, a common situation that we observe is a high value of $\alpha_\psi(\theta)$ for average values of θ and lower values of $\alpha_\psi(\theta)$ for more extreme θ values, capturing homogeneous outlying groups. In this case, a small value of α_θ will encourage few θ-clusters, and, given θ^*, a large $\alpha_\psi(\theta_j^*)$ will encourage more ψ-clusters within the j^{th} θ-cluster. The term $\prod_{j=1}^k \prod_{l=1}^{k_j} \Gamma(n_{l|j})$ will encourage asymmetrical (θ, ψ)-clusters, preferring one large cluster and several small clusters, while, given θ^*, the term involving the product of beta functions contains parts that both encourage and discourage asymmetrical θ-clusters. In the special case when $\alpha_\psi(\theta) = \alpha_\psi$ for all $\theta \in \Theta$, the random partition model simplifies to

$$p(\rho_n) = \frac{\Gamma(\alpha_\theta)}{\Gamma(\alpha_\theta + n)} \alpha_\theta^k \prod_{j=1}^{k} \alpha_\psi^{k_j} \frac{\Gamma(\alpha_\psi)\Gamma(n_j)}{\Gamma(\alpha_\psi + n_j)} \prod_{l=1}^{k_j} \Gamma(n_{l|j}).$$

In this case, the tendency of the term involving the product of beta functions is to slightly prefer asymmetrical θ-clusters with large values of α_ψ boosting this preference.

As discussed in the previous section, the random partition plays a crucial role, as its posterior distribution affects both inference on the cluster-specific parameters and prediction. For the EDP, it is given by the following proposition.

Proposition 2 *The posterior of the random partition of the EDP model is*

$$p(\rho_n|\ x_{1:n}, y_{1:n}) \propto \alpha_\theta^k \prod_{j=1}^{k} \int_\Theta \frac{\Gamma(\alpha_\psi(\theta))\Gamma(n_j)}{\Gamma(\alpha_\psi(\theta) + n_j)} \alpha_\psi(\theta)^{k_j} dP_{0\theta}(\theta) h_y(y_j^*|x_j^*) \prod_{l=1}^{k_j} \Gamma(n_{l|j}) h_x(x_{l|j}^*).$$

The proof relies on a simple application of Bayes theorem. In the case of constant $\alpha_\psi(\theta)$, the expression for the posterior of ρ_n simplifies to

$$p(\rho_n|\ x_{1:n}, y_{1:n}) \propto \alpha_\theta^k \prod_{j=1}^{k} \frac{\Gamma(\alpha_\psi)\Gamma(n_j)}{\Gamma(\alpha_\psi + n_j)} \alpha_\psi^{k_j} h_y(y_j^*|x_j^*) \prod_{l=1}^{k_j} \Gamma(n_{l|j}) h_x(x_{l|j}^*).$$

Again, as in (3), the marginal likelihood component in the posterior distribution of ρ_n is the product of the cluster-specific marginal likelihoods, but now the nested clustering structure of the EDP separates the factors relative to x and $y|x$, being $h(x_{1:n}, y_{1:n}|\rho_n) = \prod_{j=1}^k h_y(y_j^*|x_j^*) \prod_{l=1}^{k_j} h_x(x_{l|j}^*)$. Even if the x-likelihood favors many ψ-clusters, now these can be obtained by subpartitioning a coarser θ-partition, and the number k of θ-clusters can be expected to be much smaller than in (3).

Further insights into the behavior of the random partition are given by the induced covariate-dependent random partition of the θ_i parameters given the covariates, which is detailed in the following propositions. We will use the notation $\mathcal{P}_n$ to denote the set of all possible partitions of the first n integers.

Proposition 3 *The covariate-dependent random partition model induced by the EDP prior is*

$$p(\rho_{n,y}|x_{1:n}) \propto \alpha_\theta^k \prod_{j=1}^{k} \sum_{\rho_{n_j,x} \in \mathcal{P}_{n_j}} \int_\Theta \frac{\Gamma(\alpha_\psi(\theta))\Gamma(n_j)}{\Gamma(\alpha_\psi(\theta)+n_j)} \alpha_\psi(\theta)^{k_j} dP_{0\theta}(\theta) \prod_{l=1}^{k_j} \Gamma(n_{l|j}) h_x(x^*_{l|j}).$$

Proof An application of Bayes theorem implies that

$$p(\rho_n|x_{1:n}) \propto \alpha_\theta^k \prod_{j=1}^{k} \int_\Theta \frac{\Gamma(\alpha_\psi(\theta))\Gamma(n_j)}{\Gamma(\alpha_\psi(\theta)+n_j)} \alpha_\psi(\theta)^{k_j} dP_{0\theta}(\theta) \prod_{l=1}^{k_j} \Gamma(n_{l|j}) h_x(x^*_{l|j}). \tag{10}$$

Integrating over $\rho_{n,x}$, or equivalently summing over all $\rho_{n_j,x}$ in $\mathcal{P}_{n_j,x}$ for $j = 1, \ldots, k$ leads to,

$$p(\rho_{n,y}|x_{1:n}) \propto \sum_{\rho_{n_{1+},x}} \cdots \sum_{\rho_{n_{k+},x}} \alpha_\theta^k \prod_{j=1}^{k} \int_\Theta \frac{\Gamma(\alpha_\psi(\theta))\Gamma(n_j)}{\Gamma(\alpha_\psi(\theta)+n_j)} \alpha_\psi(\theta)^{k_j} dP_{0\theta}(\theta) \prod_{l=1}^{k_j} \Gamma(n_{l|j}) h_x(x^*_{l|j}),$$

and, finally, since (10) is the product over the j terms, we can pull the sum over $\rho_{n_j,x}$ within the product. ■

This covariate-dependent random partition model will favor θ-partitions of the subjects which can be further partitioned into groups with similar covariates, where a partition with many desirable subpartitions will have higher mass.

Proposition 4 *The posterior of the random covariate-dependent partition induced from the EDP model is*

$$\begin{aligned} p(\rho_{n,y}|x_{1:n}, y_{1:n}) \propto & \alpha_\theta^k \prod_{j=1}^{k} h_y(y^*_j|x^*_j) \\ & \times \sum_{\rho_{n_j,x} \in \mathcal{P}_{n_j}} \int_\Theta \frac{\Gamma(\alpha_\psi(\theta))\Gamma(n_j)}{\Gamma(\alpha_\psi(\theta)+n_j)} \alpha_\psi(\theta)^{k_j} dP_{0\theta}(\theta) \prod_{h=1}^{k_j} \Gamma(n_{l|j}) h_x(x^*_{l|j}). \end{aligned}$$

The proof is similar in spirit to that of Proposition 3. Notice the preferred θ-partitions will consist of clusters with a similar relationship between y and x, as measured by marginal local model h_y for $y|x$ and similar x behavior, which is measured much more flexibly as a mixture of the previous marginal local models.

The behavior of the random partition, detailed above, has important implications for the posterior of the unique parameters. Conditionally on the partition, the cluster-specific parameters (θ^*, ψ^*) are still independent, their posterior density being

$$p(\theta^*, \psi^*|y_{1:n}, x_{1:n}, \rho_n) = \prod_{j=1}^{k} p(\theta^*_j|y^*_j, x^*_j) \prod_{l=1}^{k_j} p(\psi^*_{l|j}|x^*_{l|j}),$$

where

$$p(\theta_j^*|y_j^*, x_j^*) \propto p_{0\theta}(\theta_j^*) \prod_{i \in S_{j+}} K(y_i|\theta_j^*, x_i), \quad p(\psi_{l|j}^*|x_{l|j}^*) \propto p_{0\psi}(\psi_{l|j}^*) \prod_{i \in S_{j,l}} K(x_i|\psi_{l|j}^*).$$

The important point is that the posterior of θ_j^* can now be updated with much larger sample sizes if the data determines that a coarser θ-partition is present. This will result in a more reliable posterior mean, a smaller posterior variance, and larger influence of the data compared with the prior.

3.2 Covariate-dependent Urn Scheme and Prediction

Similar to the DP model, computation of the predictive estimates relies on a covariate-dependent urn scheme. For the EDP, we have

$$s_{y,n+1}|\rho_n, x_{1:n+1}, y_{1:n}) \sim \frac{\omega_{k+1}(x_{n+1})}{c_0}\, \delta_{k+1} + \sum_{j=1}^{k} \frac{\omega_j(x_{n+1})}{c_0} \delta_j, \tag{11}$$

where $c_0 = p(x_{n+1}|\rho_n, x_{1:n})$, but now the expression of the $\omega_j(x_{n+1})$ takes into account the possible allocation of x_{n+1} in subgroups, being

$$\omega_j(x_{n+1}) = \sum_{s_{x,n+1}} p(x_{n+1}|\rho_n, x_{1:n}, y_{1:n}, s_{y,n+1} = j, s_{x,n+1}) p(s_{x,n+1}|\rho_n, x_{1:n}, y_{1:n}, s_{y,n+1} = j).$$

From this, it can be easily found that

$$\omega_{k+1}(x_{n+1}) = \frac{\alpha_\theta}{\alpha_\theta + n}\, h_{0,x}(x_{n+1}),$$
$$\omega_j(x_{n+1}) = \frac{n_j}{\alpha_\theta + n} \left(\pi_{k_j+1|j} h_{0,x}(x_{n+1}) + \sum_{l=1}^{k_j} \pi_{l|j} h_{l|j,x}(x_{n+1}) \right),$$

where

$$\pi_{l|j} = \int \frac{n_{l|j}}{\alpha_\psi(\theta) + n_j} dP_{0\theta}(\theta|x_j^*, y_j^*), \quad \pi_{k_j+1|j} = \int \frac{\alpha_\psi(\theta)}{\alpha_\psi(\theta) + n_j} dP_{0\theta}(\theta).$$

In the case of constant $\alpha_\psi(\theta) = \alpha_\psi$, these expressions simplify to

$$\pi_{l|j} = \frac{n_{l|j}}{\alpha_\psi + n_j}, \quad \pi_{k_j+1|j} = \frac{\alpha_\psi}{\alpha_\psi + n_j}.$$

Notice that (11) is similar to the covariate-dependent urn scheme of the DP model. The important difference is that the weights, which measure the similarity between x_{n+1} and the j^{th} cluster, are much more flexible.

It follows, from similar computations as in Section 2.2, that the predictive density at y for a new subject with a covariate value of x_{n+1} is

$$f(y|y_{1:n}, x_{1:n+1})$$
$$= \sum_{\rho_n} \left(\frac{\omega_{k+1}(x_{n+1})}{c} h_{0,y}(y|x_{n+1}) + \sum_{j=1}^{k} \frac{\omega_j(x_{n+1})}{c} h_{j,y}(y|x_{n+1}) \right) p(\rho_n|x_{1:n}, y_{1:n}), \tag{12}$$

where $c = p(x_{n+1}|x_{1:n}, y_{1:n})$.

Under the squared error loss function, the point prediction of y_{n+1} is

$$\begin{aligned}&\mathrm{E}[Y_{n+1}|y_{1:n}, x_{1:n+1}]\\&= \sum_{\rho_n} \left(\frac{\omega_{k+1}(x_{n+1})}{c} \mathrm{E}_0[Y_{n+1}|x_{n+1}] + \sum_{j=1}^{k} \frac{\omega_j(x_{n+1})}{c} \mathrm{E}_j[Y_{n+1}|x_{n+1}] \right) p(\rho_n|x_{1:n}, y_{1:n}). \quad (13)\end{aligned}$$

The expressions for the prediction density (12) and point prediction (13) are quite similar to those of the DP, (8) and (9), respectively; in both cases, the cluster-specific predictive estimates are averaged with covariate-dependent weights. However, there are two important differences for the EDP model. The first is that the weights in (11) are defined with a more flexible kernel; in fact, it is a mixture of the original kernels used in the DP model. This means that we have a more flexible measure of similarity in the covariate space. The second difference is that k will be smaller and n_j will be larger with a high posterior probability, leading to a more reliable posterior distribution of θ_j^* due to larger sample sizes and better cluster-specific predictive estimates. We will demonstrate the advantage of these two key differences in simulated and applied examples of Sections 5 and 6.

4. Computations

Inference for the EDP model cannot be obtained analytically and must therefore be approximated. To obtain approximate inference, we rely on Markov Chain Monte Carlo (MCMC) methods and consider an extension of Algorithm 2 of Neal (2000) for the DP mixture model. In this approach, the random probability measure, $\mathbf{P}$, is integrated out, and the model is viewed in terms of $(\rho_n, \theta^*, \psi^*)$. This algorithm requires the use of conjugate base measures $P_{0\theta}$ and $P_{0\psi}$. To deal with non-conjugate base measures, the approach used in Algorithm 8 of Neal (2000) can be directly adapted to the EDP mixture model.

Algorithm 2 is a Gibbs sampler which first samples the cluster label of each subject conditional on the partition of all other subjects, the data, and (θ^*, ψ^*), and then samples (θ^*, ψ^*) given the partition and the data. The first step can be easily performed thanks to the Pólya urn which marginalizes the DP.

Extending Algorithm 2 for the EDP model is straightforward, since the EDP maintains a simple, analytically computable urn scheme. In particular, the conditional probabilities $p(s_i|\rho_{n-1}^{-i}, \theta^*, \psi^*, x_{1:n}, y_{1:n})$ (provided in the Appendix) have a simple closed form, which allows conditional sampling of the individual cluster membership indicators s_i, where ρ_{n-1}^{-i} denotes the partition of the $n-1$ subjects with the i^{th} subject removed. To improve mixing, we include an additional Metropolis-Hastings step; at each iteration, after performing the n Gibbs updates for each s_i, we propose a shuffle of the nested partition structure obtained by moving a ψ-cluster to be nested within a different or new θ-cluster. This move greatly improves mixing. A detailed description of the sampler, including the Metropolis-Hastings step, can be found in the Appendix.

MCMC produces approximate samples, $\{\rho_n^s, \psi^{*s}, \theta^{*s}\}_{s=1}^S$, from the posterior. The prediction given in Equation (13) can be approximated by

$$\frac{1}{S}\sum_{s=1}^{S}\frac{\omega_{k+1}^s(x_{n+1})}{\hat{c}}\mathrm{E}_{h_y}[Y_{n+1}|x_{n+1}] + \sum_{j=1}^{k^s}\frac{\omega_j^s(x_{n+1})}{\hat{c}}\mathrm{E}_{F_y}[Y_{n+1}|x_{n+1},\theta_j^{*s}],$$

where $\omega_j^s(x_{n+1})$ for $j = 1,\ldots,k^s+1$, are as previously defined in (11) with $(\rho_n, \psi^*, \theta^*)$ replaced by $(\rho_n^s, \psi^{*s}, \theta^{*s})$ and

$$\hat{c} = \frac{1}{S}\sum_{s=1}^{S}\omega_{k+1}^s(x_{n+1}) + \sum_{j=1}^{k^s}\omega_j^s(x_{n+1}).$$

For the predictive density estimate at x_{n+1}, we define a grid of new y values and for each y in the grid, we compute

$$\frac{1}{S}\sum_{s=1}^{S}\frac{\omega_{k+1}^s(x_{n+1})}{\hat{c}}h_y(y|x_{n+1}) + \sum_{j=1}^{k^s}\frac{\omega_j^s(x_{n+1})}{\hat{c}}K(y|x_{n+1},\theta_j^{*s}). \tag{14}$$

Note that hyperpriors may be included for the precision parameters, α_θ and $\alpha_\psi(\cdot)$, and the parameters of the base measures. For the simulated examples and application, a Gamma hyperprior is assigned to α_θ, and $\alpha_\psi(\theta)$ for $\theta \in \Theta$ are assumed to be i.i.d. from a Gamma hyperprior. At each iteration, α_θ^s and $\alpha_\psi^s(\theta)$ at θ_j^{*s} for $j = 1,\ldots,k^s$ are approximate samples from the posterior using the method described in Escobar and West (1995).

5. Simulated Example

We consider a toy example that demonstrates two key advantages of the EDP model; first, it can recover the true coarser θ-partition; second, improved prediction and smaller credible intervals result. The example shows that these advantages are evident even for a moderate value of p, with more drastic differences as p increases. A data set of $n = 200$ points were generated where only the first covariate is a predictor for Y. The true model for Y is a nonlinear regression model obtained as a mixture of two normals with linear regression functions and weights depending only on the first covariate;

$$Y_i|x_i \overset{ind}{\sim} p(x_{i,1})\mathrm{N}(y_i|\beta_{1,0}+\beta_{1,1}x_{i,1},\sigma_1^2) + (1-p(x_{i,1}))\mathrm{N}(y_i|\beta_{2,0}+\beta_{2,1}x_{i,1},\sigma_2^2),$$

where

$$p(x_{i,1}) = \frac{\tau_1\exp\left(-\frac{\tau_1^2}{2}(x_{1,i}-\mu_1)^2\right)}{\tau_1\exp\left(-\frac{\tau_1^2}{2}(x_{1,i}-\mu_1)^2\right)+\tau_2\exp\left(-\frac{\tau_2^2}{2}(x_{1,i}-\mu_2)^2\right)},$$

with $\beta_1 = (0,1)'$, $\sigma_1^2 = 1/16$, $\beta_2 = (4.5, 0.1)'$, $\sigma_2^2 = 1/8$ and $\mu_1 = 4$, $\mu_2 = 6$, $\tau_1 = \tau_2 = 2$. The covariates are sampled from a multivariate normal,

$$X_i = (X_{i,1},\ldots,X_{i,p})' \overset{iid}{\sim} \mathrm{N}(\mu,\Sigma), \tag{15}$$

centered at $\mu = (4, \ldots, 4)'$ with a standard deviation of 2 along each dimension, that is, $\Sigma_{h,h} = 4$. The covariance matrix Σ models two groups of covariates: those in the first group are positively correlated among each other and the first covariate, but independent of the second group of covariates, which are positively correlated among each other but independent of the first covariate. In particular, we take $\Sigma_{h,l} = 3.5$ for $h \neq l$ in $\{1, 2, 4, \ldots, 2\lfloor p/2 \rfloor\}$ or $h \neq l$ in $\{3, 5, \ldots, 2\lfloor (p-1)/2 \rfloor + 1\}$ and $\Sigma_{h,l} = 0$ for all other cases of $h \neq l$.

We examine both the DP and EDP mixture models;

$$Y_i | x_i, \beta_i, \sigma^2_{y,i} \overset{ind}{\sim} \mathrm{N}(\underline{x}_i \beta_i, \sigma^2_{y,i}), \quad X_i | \mu_i, \sigma^2_{x,i} \overset{ind}{\sim} \prod_{h=1}^{p} \mathrm{N}(\mu_{i,h}, \sigma^2_{x,h,i}),$$
$$(\beta_i, \sigma^2_{y,i}, \mu_i, \sigma^2_{x,i}) | P \overset{iid}{\sim} P$$

with $P \sim$ DP or $P \sim$ EDP. The conjugate base measure is selected; $P_{0\theta}$ is a multivariate normal inverse gamma prior and $P_{0\psi}$ is the product of p normal inverse gamma priors, that is

$$p_{0\theta}(\beta, \sigma^2_y) = \mathrm{N}(\beta; \beta_0, \sigma^2_y C^{-1}) \mathrm{IG}(\sigma^2_y; a_y, b_y),$$

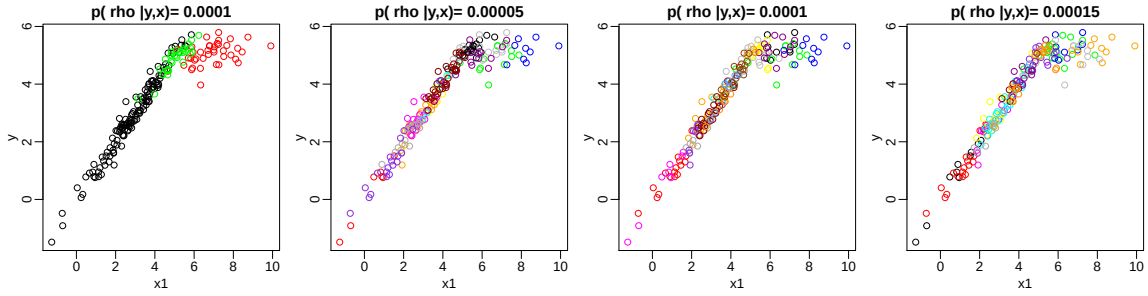
and

$$p_{0\psi}(\mu, \sigma^2_x) = \prod_{h=1}^{p} \mathrm{N}(\mu_h; \mu_{0,h}, \sigma^2_{x,h} c_h^{-1}) \mathrm{IG}(\sigma^2_{x,h}; a_{x,h}, b_{x,h}).$$

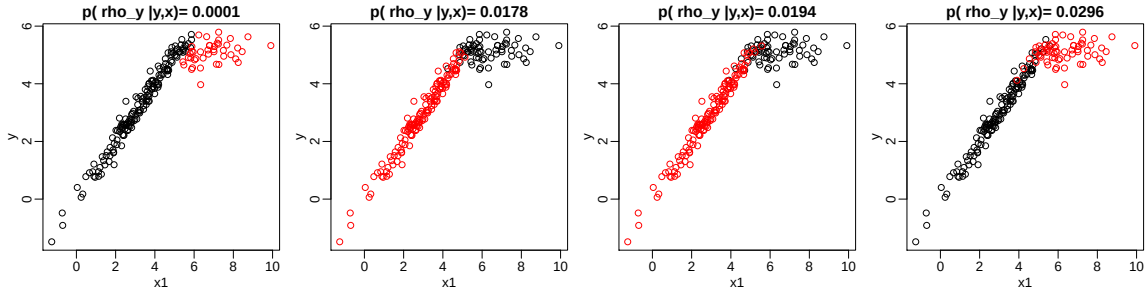
For both models, we use the same subjective choice of the parameters of the base measure. In particular, we center the base measure on an average of the true parameters values with enough variability to recover the true model. A list of the prior parameters can be found in the Appendix. We assign hyperpriors to the mass parameters, where for the DP model, $\alpha \sim \mathrm{Gamma}(1, 1)$, and for the EDP model, $\alpha_\theta \sim \mathrm{Gamma}(1, 1)$, $\alpha_\psi(\beta, \sigma^2_y) \overset{iid}{\sim} \mathrm{Gamma}(1, 1)$ for all $\beta, \sigma^2_y \in \mathbb{R}^{p+1} \times \mathbb{R}_+$.

The computational procedures described in Section 4 were used to obtain posterior inference with 20,000 iterations and burn in period of 5,000. An examination of the trace and autocorrelation plots for the subject-specific parameters $(\beta_i, \sigma^2_{y,i}, \mu_i, \sigma^2_{x,i})$ provided evidence of convergence. Additional criteria for assessing the convergence of chain, in particular, the Geweke diagnostic, also suggested convergence, and the results are given in Table 6 of the Appendix (see the **R** package *coda* for implementation and further details of the diagnostic). It should be noted that running times for both models are quite similar, although slightly faster for the DP.

The first main point to emphasize is the improved behavior of the posterior of the random partition for the EDP. We note that for both models, the posterior of the partition is spread out. This is because the space of partitions is very large and many partitions are very similar, differing only in a few subjects; thus, many partitions fit the data well. We depict representative partitions of both models with increasing p in Figure 1. Observations are plotted in the x_1–y space and colored according to the partition for the DP and the θ-partition for the EDP. As expected, for $p = 1$ the DP does well at recovering the true partition, but as clearly seen from Figure 1, for large values of p, the DP partition is comprised of many clusters, which are needed to approximate the multivariate density of X. In fact, the density of $Y|x$ can be recovered with only two kernels regardless of p, and



(a) DP, p=1 (b) DP, p=5 (c) DP, p=10 (d) DP, p=15



(e) EDP, p=1 (f) EDP, p=5 (g) EDP, p=10 (h) EDP, p=15

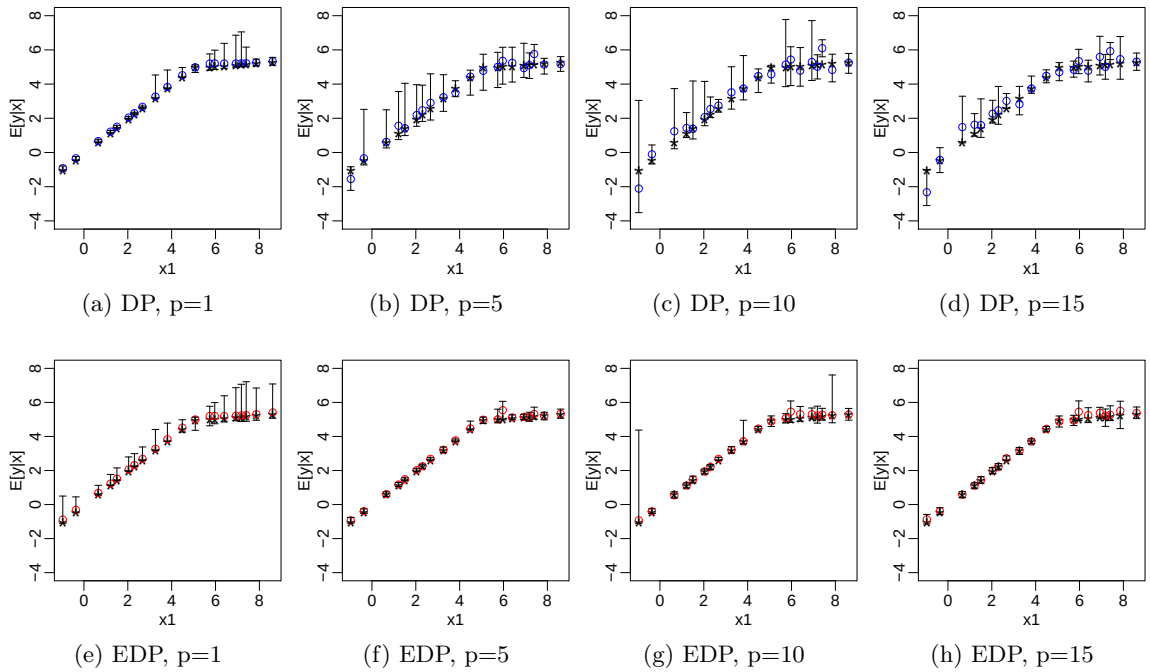
Figure 1: The y partition with the highest estimated posterior probability. Data points are plotted in the x_1 vs. y space and colored by cluster membership with estimated posterior probability included in the plot title.

	$p=1$		$p=5$		$p=10$		$p=15$	
	$\hat{k}$	$\hat{\alpha}$	$\hat{k}$	$\hat{\alpha}$	$\hat{k}$	$\hat{\alpha}$	$\hat{k}$	$\hat{\alpha}$
DP	3	0.51	14	2.75	16	3.25	15	3.09
EDP	3	0.56	2	0.35	2	0.39	2	0.36

Table 1: The posterior mode number of y clusters, denoted $\hat{k}$ for both models, and the posterior mean of α, α_θ, denoted $\hat{\alpha}$ for both models, as p increases.

the θ-partitions of the EDP depicted in Figure 1, with only two θ-clusters, are very similar to the true configuration, even for increasing p. On the other hand, the (θ, ψ)-partition of the EDP (not shown) consists of many clusters and resembles the partition of the DP model.

This behavior is representative of the posterior distribution of the random partition that, for the DP, has a large posterior mode of k and large posterior mean of α for larger values of p, while most of the EDP θ-partitions are composed of only 2 clusters with only a handful of subjects placed in the incorrect cluster and the posterior mean of α_θ is much smaller for large p. Table 1 summarizes the posterior of k for both models and the posterior of the precision parameters α, α_θ. It is interesting to note that posterior samples of $\alpha_\psi(\theta)$



(a) DP, p=1 (b) DP, p=5 (c) DP, p=10 (d) DP, p=15

(e) EDP, p=1 (f) EDP, p=5 (g) EDP, p=10 (h) EDP, p=15

Figure 2: The point predictions for 20 test samples of the covariates are plotted against x_1 and represented with circles (DP in blue and EDP in red) with true prediction as black stars. The bars about the prediction depict the 95% credible intervals.

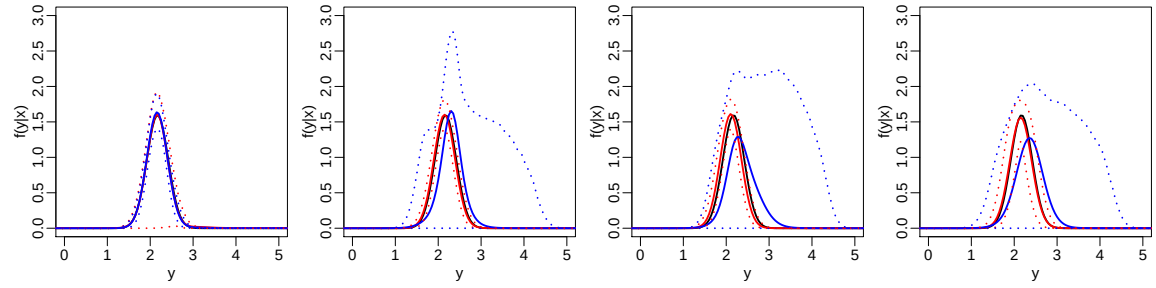
	$p=1$		$p=5$		$p=10$		$p=15$	
	$\hat{l}_1$	$\hat{l}_2$	$\hat{l}_1$	$\hat{l}_2$	$\hat{l}_1$	$\hat{l}_2$	$\hat{l}_1$	$\hat{l}_2$
DP	**0.03**	**0.05**	0.16	0.2	0.25	0.34	0.26	0.34
EDP	0.04	**0.05**	**0.06**	**0.1**	**0.09**	**0.16**	**0.12**	**0.21**

Table 2: Prediction error for both models as p increases.

for θ characteristic of the first cluster tend to be higher than posterior samples of $\alpha_\psi(\theta)$ for the second cluster; that is, more clusters are needed to approximate the density of X within the first cluster. A non-constant $\alpha_\psi(\theta)$ allows us to capture this behavior.

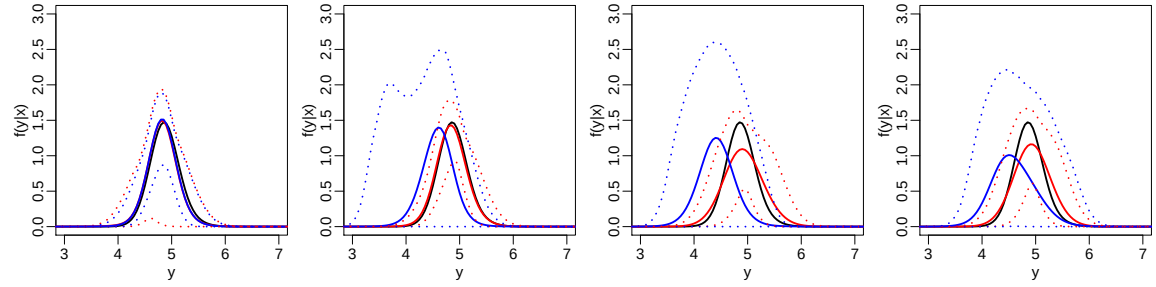
As discussed in Section 3, a main point of our proposal is the increased finite sample efficiency of the EDP model. To illustrate this, we create a test set with $m = 200$ new covariates values simulated from (15) with maximum dimension $p = 15$ and compute the true regression function $\mathrm{E}[Y_{n+j}|x_{n+j}]$ and conditional density $f(y|x_{n+j})$ for each new subject. To quantify the gain in efficiency of the EDP model, we calculate the point prediction and predictive density estimates from both models and compare them with the truth.

Judging from both the empirical l_1 and l_2 prediction errors, the EDP model outperforms the DP model, with greater improvement for larger p; see Table 2. Figure 2 displays the prediction against x_1 for 20 new subjects. Recall that each new x_1 is associated with different values of $(x_2, \ldots, x_p)$, which accounts for the somewhat erratic behavior of the



(a) p=1 (b) p=5 (c) p=10 (d) p=15

Figure 3: Predictive density estimate (DP in blue and EDP in red) for the first new subject with true conditional density in black. The pointwise 95% credible intervals are depicted with dashed lines (DP in blue and EDP in red).



(a) p=1 (b) p=5 (c) p=10 (d) p=15

Figure 4: Predictive density estimate (DP in blue and EDP in red) for the fifth new subject with true conditional density in black. The pointwise 95% credible intervals are depicted with dashed lines (DP in blue and EDP in red).

	$p=1$	$p=5$	$p=10$	$p=15$
DP	0.13	0.5	0.66	0.68
EDP	**0.12**	**0.15**	**0.24**	**0.29**

Table 3: l_1 density regression error for both models as p increases.

prediction as a function of x_1 for increasing p. The comparison of the credible intervals is quite interesting. For $p > 1$, the unnecessarily wide credible intervals for the DP regression model stand out in the first row of Figure 2. This is due to small cluster samples sizes for the DP model with $p > 1$.

The density regression estimates for all new subjects were computed by evaluating (14) at a grid of y-values. As a measure of the performance of the models, the empirical l_1 distance between the true and estimated conditional densities for each of the new covariate values is shown in Table 3. Again the EDP model outperforms the DP model. Figures

3 and 4 display the true conditional density (in black) for two new covariate values with the estimated conditional densities in blue for the DP and red for the EDP. It is evident that for $p > 1$ the density regression estimates are improved and that the pointwise 95% credible intervals are almost uniformly wider both in y and x for the DP model, sometimes drastically so. It is important to note that while the credible intervals of the EDP model are considerably tighter, they still contain the true density.

6. Alzheimer's Disease Study

Our primary goal in this section is to show that the EDP model leads to improved inference in a real data study with the important goal of diagnosing Alzheimer's disease (AD). In particular, EDP leads to improved predictive accuracy, tighter credible intervals around the predictive probability of having the disease, and a more interpretable clustering structure.

Alzheimer's disease is a prevalent form of dementia that slowly destroys memory and thinking skills, and eventually even the ability to carry out the simplest tasks. Unfortunately, a definitive diagnosis cannot be made until autopsy. However, the brain may show severe evidence of neurobiological damage even at early stages of the disease before the onset of memory disturbances. As this damage may be difficult to detect visually, improved methods for automatically diagnosing disease based on MRI neuroimaging is needed.

Data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, which is publicly accessible at UCLA's Laboratory of Neuroimaging.[1] The data set consists of summaries of fifteen brain structures computed from structural MRI obtained at the first visit for 377 patients, of which 159 have been diagnosed with AD and 218 are cognitively normal (CN). The covariates include whole brain volume (BV), intracranial volume (ICV), volume of the ventricles (VV), left and right hippocampal volume (LHV, RHV), volume of the left and right inferior lateral ventricle (LILV, RILV), thickness of the left and right middle temporal cortex (LMT, RMT), thickness of the left and right inferior temporal cortex (LIT, RIT), thickness of the left and right fusiform cortex (LF, RF), and thickness of the left and right entorhinal cortex (LE, RE). Volume is measured in cm^3 and cortical thickness is measured in mm.

The response is a binary variable with 1 indicating a cognitively normal subject and 0 indicating a subject who has been diagnosed with AD. The covariate is the 15-dimensional

1. The ADNI was launched in 2003 by the National Institute on Ageing (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $ 60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California-San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see `www.adni-info.org`.

vector of measurements of various brain structures. Our model builds on local probit models and can be stated as follows:

$$Y_i|x_i,\beta_i \overset{ind}{\sim} \text{Bern}(\Phi(\underline{x}_i\beta_i)), \quad X_i|\mu_i,\sigma_i^2 \overset{ind}{\sim} \prod_{h=1}^{p} \text{N}(\mu_{i,h},\sigma_{i,h}^2),$$
$$(\beta_i,\mu_i,\sigma_i^2)|P \overset{iid}{\sim} P, \quad \mathbf{P}\sim Q.$$

The analysis is first carried using a DP prior for $\mathbf{P}$ with mass parameter α and base measure $P_{0\beta}\times P_{0\psi}$, with $P_{0\beta}=\text{N}(0_p,C^{-1})$ and $P_{0\psi}$ defined as the product of p normal inverse gamma measures. A list of the prior parameters can be found in the Appendix. The mass parameter is given a hyperprior of $\alpha\sim\text{Gamma}(1,1)$.

The prior parameters were carefully selected based on prior knowledge of the brain structures and their relationship with the disease and empirical evidence. The base measure for β was chosen to be centered on zero because even though we have prior belief about how each structure is related to AD individually, the joint relationship may be more complex. For simplicity, the covariance matrix is diagonal. The variances were chosen to reflect belief in the maximum range of the coefficient for each brain structure. We also explored the idea of defining C through a g-prior, where $C^{-1}=g(\underline{X}'\underline{X})^{-1}$ with g fixed or given a hyperprior. However, this proposal was unsatisfactory because prior information about the maximum range of the coefficient for each brain structure is condensed in a single parameter g. For example, there was no way to incorporate the belief that while the variability of hippocampal volume and inferior lateral ventricular volume are similar, the correlation between hippocampal volume and disease status is stronger. The parameters of the base measure for X were chosen based on prior knowledge and exploratory analysis of the average volume and cortical thickness of the brain structures (μ_0) and variability (b_x). The parameter a_x was chosen to equal 2, so that mean of the inverse gamma prior is properly defined and the variance is relatively large. The parameter c_x is equal to 1/2 to increase variability of μ given σ_x.

In this example, correlation between the measurements of the brain structures is expected. Furthermore, univariate histograms of the covariates show non-normal behavior. These facts suggest that many Gaussian kernels with local independence of the covariates will be needed to approximate the density of X. The conditional density of the response, on the other hand, may not be so complicated. This motivates the choice of an EDP prior. We emphasize that the same conjugate base measure is used with the identical subjective choice of parameters. The hyperpriors for the mass parameter of

$$\alpha_\beta\sim\text{Gamma}(1,1), \quad \alpha_\psi(\beta)\overset{iid}{\sim}\text{Gamma}(1,1) \quad \forall\beta\in\mathbb{R}^{p+1}.$$

If $\alpha_\psi(\beta)\approx 0$ for all $\beta\in\mathbb{R}^{p+1}$ the model converges to a DP mixture model, suggesting that the extra flexibility of the EDP is unnecessary. On the other hand, $\alpha_\beta\approx 0$ suggests that a linear model is sufficient for modelling the conditional response distribution.

The data were randomly split into a training sample of size 185 and a test sample of size 192. Inference is based on the algorithm explained in Section 4 with the added step of sampling a latent normal variable to deal with the binary response. For both results the number of iterations is 50,000 with burn in period of 10,000. An examination of the trace

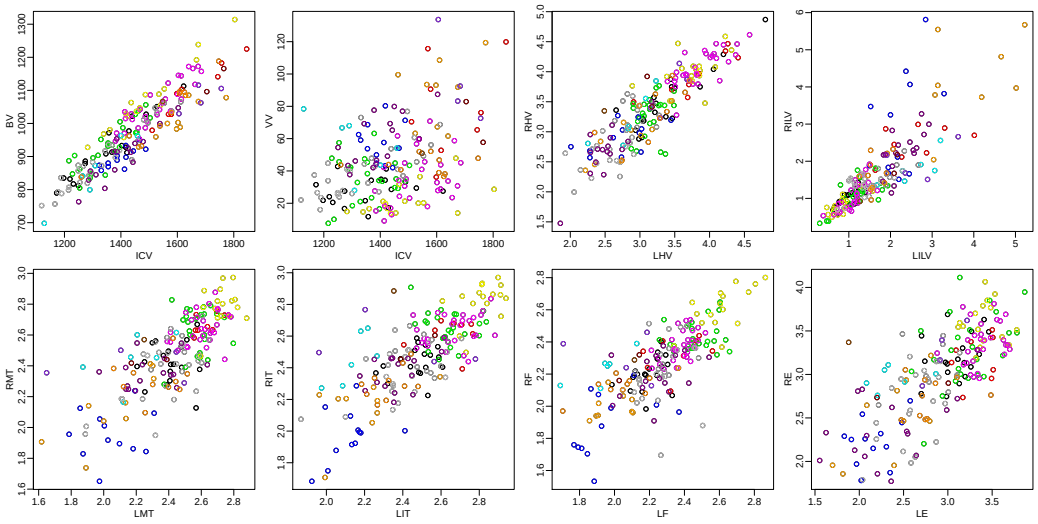


Figure 5: Data points are plotted in the covariate space and colored by the partition with the highest posterior probability for the DP model.

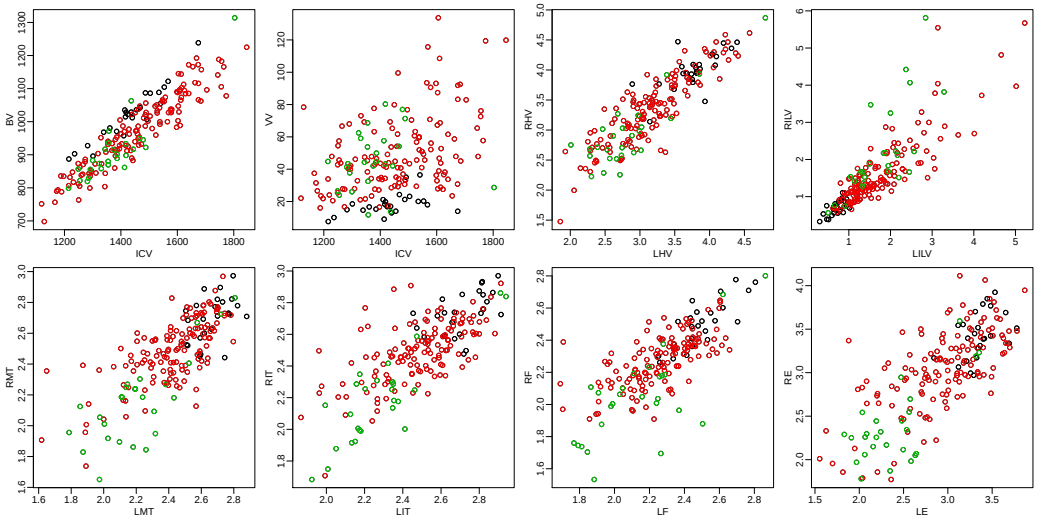


Figure 6: Data points are plotted in the covariate space and colored by the β-partition with the highest posterior probability for the EDP model.

and autocorrelation plots for the subject-specific parameters $(\beta_i, \mu_i, \sigma_i^2)$ provided evidence of convergence, which was further checked via the Geweke and Raftery and Lewis methods. Computation times are quite similar for both models, although slightly faster for the DP.

The DP based model requires many kernels to approximate the joint distribution, while the EDP prefers a coarser β-partition for the conditional density of $Y|x$. The posterior of k and the precision parameters α and α_β are summarized and compared for the two models in

	$\hat{k}$	$\tilde{k}$	$[k_l, k_u]$	$\hat{\alpha}$	$[\alpha_l, \alpha_u]$
DP	16	16	[14,19]	3.41	[1.79, 5.59]
EDP	3	4	[2,7]	0.71	[0.11, 1.76]

Table 4: The posterior of k for both models is summarized through the posterior mode, denoted $\hat{k}$; the posterior median, denoted $\tilde{k}$; and the 95% credible intervals, denoted $[k_l, k_u]$. The posterior of the precision parameter α for the DP and α_β for the EDP is summarized through the posterior mode, denoted $\hat{\alpha}$, and 95% credible intervals, denoted $[\alpha_l, \alpha_u]$.

	Accuracy	AUC	MXE
DP	82.81%	0.88	0.42
EDP	86.46%	0.90	0.38

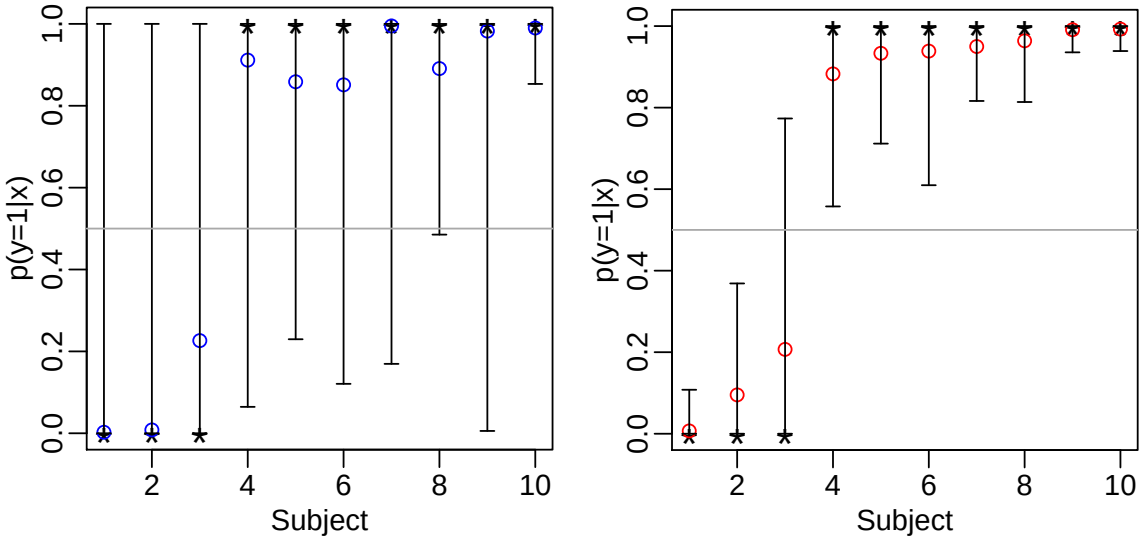
Table 5: Predictive accuracy, area under the ROC curve (AUC), and mean cross entropy (MXE) for the test set.

Table 4. Posterior samples of $\alpha_\psi(\beta)$ tend to be higher for average values of β, meaning that more kernels are needed for the density of X within such β-clusters. The added flexibility of a β-dependent precision parameter for ψ allows us to capture this behavior.

The posterior of the partition is quite spread out across many similar partitions for the DP and EDP models. A representative partition, the partition with the highest estimated posterior probability, for the DP mixture models is depicted in Figure 5, where the data points are plotted in the covariate space and colored by the partition. Notice the high number of kernels with small sample sizes within each cluster. Figure 6 depicts a representative β-partition for the EDP mixture model, where the data points are colored by the β-partition. Not only are there fewer clusters with larger sample sizes, but the clusters are much more interpretable as well. In particular, the posterior concentrates on partitions similar to the one depicted in Figure 6 with a general cluster and two extreme clusters of 100% AD and 100% non-AD patients (the green and black clusters, respectively). The black cluster of non-AD patients display high brain volume compared to intracranial volume, low ventricular volume, high hippocampal volume, and high cortical thickness; this behavior could be expected as AD is associated with shrinking brain tissue and increased cerebrospinal fluid, while intracranial volume remained fixed. The green cluster of AD patients display lower hippocampal volume and interestingly, low intracranial volume and low cortical thickness with a "right-less-than-left" pattern.

To quantify the gain in efficiency with the EDP model, we computed the predictive accuracy, area under the ROC curve (AUC), and mean cross entropy (MXE) for the test set, using the *ROCR* package in **R**. The larger sample sizes of the EDP model improve all predictive criteria when compared to the DP model (Table 5).

Finally, we note that by allowing for a coarser β-partition when appropriate, the increased cluster sample sizes not only result in improved accuracy for the EDP model but also allow for much tighter credible intervals. This is shown in Figure 7 which depicts the



(a) DP (b) EDP

Figure 7: Predicted probability of being healthy against subject index for 10 new subjects represented with circles (DP in blue and EDP in red) with the true outcome as black stars. The bars about the prediction depict the 95% credible intervals.

estimated probability of being healthy for 10 subjects along with lower and upper bounds for 95% credible intervals, as a function of subject index. Notice the tighter credible intervals for the EDP model with some dramatic examples given by subjects 1 and 8.

We also compared with Gaussian process (GP), support vector machine (SVM), and random forest (RF) models, which are implemented in the *kernlab* and *randomForest* packages in **R**. The results of the EDP are comparable with these other standard nonparametric classification methods, in particular the predictive accuracy of the GP, SVM, and RF models are 85.42%, 86.46%, and 86.46%, respectively. The results remain quite comparable with different random splits into training and test sets, which also confirmed the conclusions suggested by Figures 5, 6, and 7 and Tables 4 and 5.

7. Discussion

In this paper, we have highlighted a drawback of DP mixture models when the aim is estimation of the regression function and conditional distribution. We have proposed a simple, but efficient, solution based on the EDP, which overcomes the problems of the DP mixture model by introducing a nested partition structure. An important feature of the proposed EDP mixture model is that computations remain relatively simple; unlike other modifications of the DP for conditional distribution modeling we maintain the ability to marginalize out the random measure and induce a simple urn scheme. In scaling up to larger numbers of predictors p, this simplified structure should be advantageous.

Acknowledgments

We thank the referees and associate editor for their helpful comments. Data collection and sharing for the application in Section 6 of this work was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). We acknowledge the funding contributions of ADNI supporters (`adni-info.org/Scientists/ADNISponsors.aspx`).

Appendix A. Computations

This appendix contains further details of the MCMC algorithm described in Section 4.

The conditional distribution of $s_i = (s_{i,y}, s_{i,x})$, which denotes the vector containing the y-cluster and x-cluster membership for subject i, is

$$s_i|\rho_{n-1}^{-i}, \theta^*, \psi^*, x_{1:n}, y_{1:n} \sim \frac{\omega_{k^{-i}+1,1}(y_i, x_i)}{c_1}\delta_{(k^{-i}+1,1)} + \sum_{j=1}^{k^{-i}}\left(\frac{\omega_{j,k_j^{-i}+1}(y_i, x_i)}{c_1}\delta_{(j,k_j^{-i}+1)} + \sum_{l=1}^{k_j^{-i}}\frac{\omega_{j,l}(y_i, x_i)}{c_1}\delta_{(j,l)}\right), \tag{16}$$

where for $j = 1, \ldots, k^{-i}$ and $l = 1, \ldots, k_j^{-i}$,

$$\omega_{j,l}(y_i, x_i) = \frac{n_j^{-i} n_{l|j}^{-i}}{\alpha_\psi(\theta_j^{*-i}) + n_j^{-i}} K(y_i|x_i, \theta_j^{*-i}) K(x_i|\psi_{l|j}^{*-i}),$$

for $j = 1, \ldots, k^{-i}$,

$$\omega_{j,k_j^{-i}+1}(y_i, x_i) = \frac{n_j^{-i}\alpha_\psi(\theta_j^{*-i})}{\alpha_\psi(\theta_j^{*-i}) + n_j^{-i}} K(y_i|x_i, \theta_j^{*-i}) h_x(x_i),$$

$$\omega_{k^{-i}+1,1}(y_i, x_i) = \alpha_\theta h_y(y_i|x_i) h_x(x_i),$$

and

$$c_1 = \omega_{k^{-i}+1,1}(y_i, x_i) + \sum_{j=1}^{k^{-i}}\left(\omega_{j,k_j^{-i}+1}(y_i, x_i) + \sum_{l=1}^{k_j^{-i}}\omega_{j,l}(y_i, x_i)\right).$$

Here, ρ_{n-1}^{-i} represents the partition of the $n-1$ subjects with the i^{th} subject removed where $k^{-i}, k_j^{-i}, n_j^{-i}, n_{l|j}^{-i}$ are defined from ρ_{n-1}^{-i}. Similarly, θ_j^{*-i} and $\psi_{l|j}^{*-i}$ are the unique cluster parameters associated to the clusters of ρ_{n-1}^{-i}.

Next, we describe the Metropolis-Hastings step, which proposes to move a ψ-cluster to be nested with a different or new θ-cluster. This step is separated into three possible moves: 1) a ψ-cluster, among those within θ-clusters with more than one ψ-cluster, is moved to a different θ-cluster; 2) a ψ-cluster, among those within θ-clusters with more than one ψ-cluster, is moved to a new θ-cluster; 3) a ψ-cluster, among those within θ-clusters with only one ψ-cluster, is moved to a different θ-cluster.

Let $k_{x,2+}$ be the number of ψ-clusters within a θ-cluster with more than one ψ-cluster and $k_{x,1}$ be the number of ψ-clusters within a θ-cluster with only one ψ-cluster. The proposal distributions for the three moves are as follows. For the first move, the ψ-cluster is uniformly selected with probability $k_{x,2+}^{-1}$ and moved within a different θ-cluster selected uniformly with probability $(k-1)^{-1}$. For the second, the ψ-cluster is again uniformly selected with probability $k_{x,2+}^{-1}$ and moved to a new cluster. Lastly, for the third, the ψ-cluster is uniformly selected with probability $k_{x,1}^{-1}$ and moved within a different θ-cluster selected uniformly with probability $(k-1)^{-1}$.

Let ρ_n^* be the proposed partition defined by moving ψ-cluster l in θ-cluster j to θ-cluster h. For the first move, $h \in \{1, \ldots, j-1, j+1, \ldots, k\}$ and the acceptance probability is $\min(1, p)$, where

$$p = \frac{\Gamma(n_j - n_{l|j})\Gamma(n_h + n_{l|j})}{\Gamma(n_j)\Gamma(n_h)} \frac{\Gamma(\alpha_\psi(\theta_j^*) + n_j)\Gamma(\alpha_\psi(\theta_h^*) + n_h)}{\Gamma(\alpha_\psi(\theta_j^*) + n_j - n_{l|j})\Gamma(\alpha_\psi(\theta_h^*) + n_h + n_{l|j})} \frac{\alpha_\psi(\theta_h^*)}{\alpha_\psi(\theta_j^*)} \frac{\prod_{i \in S_{l|j}} K(y_i|x_i, \theta_h^*)}{\prod_{i \in S_{l|j}} K(y_i|x_i, \theta_j^*)} \frac{k_{x,2+}}{k_{x,2+}^*},$$

and $k_{x,2+}^*$ is the number of ψ-clusters within a θ-cluster with more than one ψ-cluster under the proposed partition. For the second move, $h = k+1$ and let $\theta_{k+1}^* \sim P(\theta|y_{l|j}^*, x_{l|j}^*)$ be the proposed parameter of the $k+1$ θ-cluster. The acceptance probability is $\min(1, p)$, where

$$p = \frac{\Gamma(n_j - n_{l|j})\Gamma(n_{l|j})}{\Gamma(n_j)} \frac{\Gamma(\alpha_\psi(\theta_j^*) + n_j)\Gamma(\alpha_\psi(\theta_{k+1}^*))}{\Gamma(\alpha_\psi(\theta_j^*) + n_j - n_{l|j})\Gamma(\alpha_\psi(\theta_{k+1}^*) + n_{l|j})} \alpha_\theta \frac{\alpha_\psi(\theta_{k+1}^*)}{\alpha_\psi(\theta_j^*)} \frac{h_y(y_{l|j}^*|x_{l|j}^*)}{\prod_{i \in S_{l|j}} K(y_i|x_i, \theta_j^*)} \frac{k_{x,2+}}{k_{x,1}^* k},$$

and $k_{x,1}^*$ is the number of ψ-clusters within a θ-cluster with only one ψ-cluster under the proposed partition. Finally, for the third move, $h \in \{1, \ldots, j-1, j+1, \ldots, k\}$ and the acceptance probability is $\min(1, p)$, where

$$p = \frac{\Gamma(n_h + n_{l|j})}{\Gamma(n_{l|j})\Gamma(n_h)} \frac{\Gamma(\alpha_\psi(\theta_j^*) + n_{l|j})\Gamma(\alpha_\psi(\theta_h^*) + n_h)}{\Gamma(\alpha_\psi(\theta_j^*))\Gamma(\alpha_\psi(\theta_h^*) + n_h + n_{l|j})} \frac{1}{\alpha_\theta} \frac{\alpha_\psi(\theta_h^*)}{\alpha_\psi(\theta_j^*)} \frac{\prod_{i \in S_{l|j}} K(y_i|x_i, \theta_h^*)}{h_y(y_{l|j}^*|x_{l|j}^*)} \frac{k_{x,1}k - 1}{k_{x,2+}^*}.$$

Each iteration of the MCMC algorithm is summarized as follows:

- For $i = 1, \ldots, n$,
  - if $s_{i,y} = j$ and $n_j^{-i} = 0$,
    - then remove θ_j^* and $\psi_{l|j}^*$ from (θ^*, ψ^*).
  - Otherwise, if $s_{i,y} = j$, $s_{i,x} = l$ and $n_{l|j}^{-i} = 0$,
    - then remove $\psi_{l|j}^*$ from ψ^*.
  - Next, sample s_i given $\rho_{n-1}^{-i}, \theta^*, \psi^*, x_{1:n}, y_{1:n}$ as defined by Equation (16).
  - If $s_{i,y} = k^{-i} + 1$,

		$p=1$	$p=5$	$p=10$	$p=15$
$\beta_{0,i}$	DP	-0.91	-0.13	-1.22	-3.15
	EDP	-1.86	0.59	1.24	-0.67
$\beta_{1,i}$	DP	0.61	-0.22	0.79	3.25
	EDP	1.54	-0.64	2.08	-0.42
$\sigma^2_{y,i}$	DP	-1.51	1.35	-0.13	1.06
	EDP	0.49	1.1	-2.09	-3.51

Table 6: The Z-score from Geweke diagnostic for the subject-specific θ-parameters of one subject.

    - $*$ sample $\theta^*_{k^{-i}+1}$ given y_i, x_i and $\psi^*_{1|k^{-i}+1}$ given x_i and concatenate them to (θ^*, ψ^*).
  - Otherwise, if $s_{i,y} = j$ and $s_{i,x} = k_j^{-i} + 1$,
    - $*$ sample $\psi^*_{k_j^{-i}+1|j}$ given x_i and concatenate it to ψ^*.

- Carry out the first move described in the Metropolis-Hastings step.
- Sample $u \sim \mathrm{U}(0,1)$. If $u < 0.5$, perform move 2, otherwise perform move 3.
- For $j = 1, \ldots, k$,
  - sample θ^*_j given (y^*_j, x^*_j), that is, from the posterior based on $p_{0\theta}(\theta^*_j)$ and $\prod_{i \in S_{j+}} K(y_i|x_i, \theta^*_j)$,
  - and for $l = 1, \ldots, k_j$,
    - $*$ sample $\psi^*_{l|j}$ given $x^*_{l|j}$, that is, from the posterior based on $p_{0\psi}(\psi^*_{l|j})$ and $\prod_{i \in S_{j,l}} K(x_i|\psi^*_{l|j})$.

Appendix B. Simulation Study

The prior parameters used for the simulation study in Section 5 are $\beta_0 = (2.25, 0.55, 0, \ldots, 0)'$, $C = \mathrm{diag}\,(0.05, 1, \ldots, 1)$; $a_y = 2$, $b_y = 0.1$, $\mu_0 = (4, \ldots, 4)'$, $c = (0.25, \ldots, 0.25)'$; $a_x = (2, \ldots, 2)'$, $b_x = (1, \ldots, 1)'$.

Table 6 lists the Z-scores for the θ parameters of the one subject from the Geweke diagnostic for assessing convergence of the MCMC chain, which are slightly high for $p = 10$ and $p = 15$ but a thinning of 2 improves the scores.

Appendix C. Alzheimer's Disease Study

For the prior parameters for the AD study in Section 6, C^{-1} is a diagonal matrix with diagonal elements $(400, .0001, .0001, 0.0004, 4, 4, .25, .25, 4, 4, 4, 4, 1, 1, 1, 1)$, and $\mu_0 = (1000, 1450, 45, 3.25, 3.25, 2, 2, 2.4, 2.4, 2.5, 2.5, 2.3, 2.3, 2.75, 2.75)'$; $c_{x,h} = 1/2, a_{x,h} = 2\ \forall h$; and $b_x = (10000, 10000, 150, .25, .25, .25, .25, .04, .04, .04, .04, .04, .04, .1, .1)'$.

References

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

D.M. Blei, T.L. Griffiths, and M.I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57:1–30, 2010.

P.J. Brown, N.D. Le, and J.V. Zidek. Inference for a covariance matrix. In P.R. Freeman and A.F.M. Smith, editors, *Aspects of Uncertainty. A Tribute to D.V. Lindley*, pages 77–92. Wiley, Chichester, 1994.

Y. Chung and D.B. Dunson. Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104:1646–1660, 2009.

B.S. Clarke, E. Fokoué, and H.H. Zhang. *Principles and Theory for Data Mining and Machine Learning*. Springer Series in Statistics, New York, 2009.

G. Consonni and P. Veronese. Conditionally reducible natural exponential families and enriched conjugate priors. *Scandinavian Journal of Statistics*, 28:377–406, 2001.

D.B. Dunson. Nonparametric Bayes local partition models for random effects. *Biometrika*, 96:249–262, 2009.

D.B. Dunson and J.H. Park. Kernel stick-breaking processes. *Biometrika*, 95:307–323, 2008.

D.B. Dunson, J. Xue, and L. Carin. The matrix stick breaking process: Flexible Bayes meta analysis. *Journal of the American Statistical Association*, 103:317–327, 2008.

D.B. Dunson, S. Petrone, and L. Trippa. Partially hierarchical Dirichlet mixtures for flexible clustering and regression. 2011. Unpublished manuscript.

S. Efromovich. Conditional density estimation in a regression setting. *Annals of Statistics*, 35:2504–2535, 2007.

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

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

S. Ghosal. Dirichlet process, related priors, and posterior asymptotics. In N.L. Hjort, C. Holmes, P. Müller, and S.G. Walker, editors, *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, 2010.

J.E. Griffin and M.F.J. Steel. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 10:179–194, 2006.

L.A. Hannah, D.M. Blei, and W.B. Powell. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12:1923–1953, 2011.

C. Kang and S. Ghosal. Clusterwise regression using Dirichlet process mixtures. In A. Sengupta, editor, *Advances in Multivariate Statistical Methods*, pages 305–325. 2009.

S.N. MacEachern. Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pages 50–55, Alexandria, VA, 1999. American Statistical Association.

P. Müller and F.A. Quintana. Random partition models with regression on covariates. *Journal of Statistical Planning and Inference*, 140:2801–2808, 2010.

P. Müller, A. Erkanli, and M. West. Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 88:67–79, 1996.

R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistcs*, 9:249–265, 2000.

A. Norets and J. Pelenis. Bayesian modeling of joint and conditional distributions. *Journal of Econometrics*, 168:332–346, 2012.

J.H. Park and D.B. Dunson. Bayesian generalized product partition model. *Statistica Sinica*, 20:1203–1226, 2010.

S. Petrone and L. Trippa. Bayesian modeling via nested random partitions. In *Proceedings of the International Conference on Complex Data Modelling and Computationally Intensive Statistical Methods*, Milan, Italy, 2009. Politecnico di Milano.

S. Petrone, M. Guindani, and A.E. Gelfand. Hybrid Dirichlet mixture models for functional data. *Journal of the Royal Statistical Society, Series B*, 71:755–782, 2009.

L. Ren, L. Du, D.B. Dunson, and L. Carin. The logistic stick-breaking process. *Journal of Machine Learning and Research*, 12:203–239, 2011.

A. Rodriguez and D.B. Dunson. Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6:145–178, 2011.

A. Rodriguez, D.B. Dunson, and A.E. Gelfand. Bayesian nonparametric functional data analysis through density estimation. *Biometrika*, 96:149–162, 2009.

D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, Inc., Hoboken, NJ, 1992.

B. Shahbaba and R.M. Neal. Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*, 10:1829–1850, 2009.

S.T. Tokdar. Adaptive convergence rates of a Dirichlet process mixture of multivariate normals. 2011. `arXiv:1111.4148 [math.ST]`.

S.K. Wade, S. Mongelluzzo, and S. Petrone. An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Analysis*, 6:359–386, 2011.

Y. Wu and S. Ghosal. Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics*, 2:298–331, 2008.

Y. Wu and S. Ghosal. The L_1-consistency of Dirichlet mixtures in multivariate density estimation. *Journal of Multivariate Analysis*, 101:2411–2419, 2010.