A Bayesian Nonparametric Regression Model With Normalized Weights - A Study of Hippocampal Atrophy in Alzheimer’s Disease

Citation for published version:
Antoniano-Villalobos, I, Wade, S & Walker, SG 2014, 'A Bayesian Nonparametric Regression Model With Normalized Weights - A Study of Hippocampal Atrophy in Alzheimer’s Disease' Journal of the American Statistical Association, vol. 109, no. 506, pp. 477-490. DOI: 10.1080/01621459.2013.879061

Digital Object Identifier (DOI):
10.1080/01621459.2013.879061

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of the American Statistical Association

A Bayesian Nonparametric Regression Model with Normalized Weights: A Study of Hippocampal Atrophy in Alzheimer's Disease

Isadora Antoniano-Villalobos [*] Sara Wade [†]

Stephen G. Walker [‡]

For the Alzheimer's Disease Neuroimaging Initiative. [§]

Abstract

Hippocampal volume is one of the best established biomarkers for Alzheimer's disease. However, for appropriate use in clinical trials research, the evolution of hippocampal volume needs to be well understood. Recent theoretical models propose a sigmoidal pattern for its evolution. To support this theory, the use of Bayesian nonparametric regression mixture models seems particularly suitable due to the flexibility that models

[*] Bocconi University, Milan, Italy. PhD research funded by CONACyT

[†] University of Cambridge, Cambridge, UK.

[‡] University of Texas at Austin, USA.

[§] 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 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`.

of this type can achieve and the unsatisfactory predictive properties of semiparametric methods. In this paper, our aim is to develop an interpretable Bayesian nonparametric regression model which allows inference with combinations of both continuous and discrete covariates, as required for a full analysis of the data set. Simple arguments regarding the interpretation of Bayesian nonparametric regression mixtures lead naturally to regression weights based on normalized sums. Difficulty in working with the intractable normalizing constant is overcome thanks to recent advances in MCMC methods and the development of a novel auxiliary variable scheme. We apply the new model and MCMC method to study the dynamics of hippocampal volume, and our results provide statistical evidence in support of the theoretical hypothesis.

Keywords: Mixture model; Dependent Dirichlet process; Latent model.

1 Introduction

Alzheimer's disease (AD) is an irreversible, progressive brain disease that slowly destroys memory and thinking skills, and eventually even the ability to carry out the simplest tasks (ADEAR, 2011). Due to its damaging effects and increasing prevalence, it has become a major public health concern. Thus, the development of disease-modifying drugs or therapies is of great importance. In a clinical trial setting, with the purpose of assessing the effectiveness of any proposed drugs or therapies, accurate tools for monitoring disease progression are needed. Unfortunately, a definite measure of disease progression is unavailable, as even a definitive diagnosis requires histopathologic examination of brain tissue, an invasive procedure typically only performed at autopsy. Non-invasive methods can be used to produce neuroimages and biospecimens which provide evidence of the changes in the brain associated with AD. Moreover, biomarkers based on neuroimaging or biological data may present a higher sensitivity to changes due to drugs or therapies over shorter periods of time than clinical measures, making them better suited tools for monitoring disease progression in clinical trials. However, before biomarkers based on neuroimaging or biological data can be useful in clinical trials, their evolution over time needs to be well understood. The biomarkers which change earliest and fastest should be used as inclusion criteria for the trials and those which change the most in the disease stage of interest should be used for disease monitoring.

In this work, we focus on hippocampal volume, one of the best established neuroimaging biomarkers for AD. Jack et al. (2010), in a recent paper, propose a theoretical model for the evolution of hippocampal volume, which is further discussed in Frisoni et al. (2010). They hypothesize that hippocampal volume evolves sigmoidally with changes beginning early and continuing into late stages of the disease. This theoretical model needs to be validated, before the use of hippocampal volume as a measure for disease severity in clinical trials can be appropriately considered. Thus, in the present paper, we focus on the validation of Jack et al.'s proposed model. Caroli and Frisoni (2010) and Sabuncu et al. (2011) assess the fit of parametric sigmoidal curves, and Jack et al. (2012) consider a more flexible model based on cubic splines with three chosen knot points. This last approach is the most flexible among the three, but they all impose significant restrictions which favor a sigmoidal shape. To provide strong statistical support for the sigmoidal shape hypothesis, a flexible nonparametric regression model is needed that would remove all restrictions on the regression curve allowing the data to choose the shape that provides the best fit and predictive properties for unobserved values.

There are many methods for nonparametric regression, and most standard approaches, such as splines, wavelets, or regression trees (Denison et al., 2002; Dimatteo et al., 2001), achieve flexibility by representing the regression function as a linear combination of basis functions. Another increasingly popular practice is to place a Gaussian process prior on the unknown regression function (Rasmussen and Williams, 2006). While these models are able to capture a wide range of regression functions, the assumptions on the distribution of the errors about the mean is quite restrictrive; typically, independent and identically distributed additive Gaussian errors are assumed, and thus, these models are often referred to as semiparametric. In the hippocampal volume study, we not only expect a non linear behaviour for the evolution of the AD biomarker with age, but also suspect the presence of multimodality, heavy tails, and evolving variance in the error distribution due to variability in the onset of the disease and unobserved factors, such as enhanced cognitive reserve or neuroprotective genes. Indeed, in a semiparametric analysis of the data, we observe a non-normal behavior in the errors that depends on the covariates, which raises suspicions about the estimated regression curve. To correctly model the data, a nonparametric approach for modelling the conditional density in its entirety is needed. In this way, no specific structure is imposed on the regression function or error distribution, so a fit confirming the hypothesized sigmiodal shape would provide strong statistical support for the theoretical model.

In this paper, we investigate the dynamics of hippocampal volume as a function of age, disease status, and gender. To do so, we construct a flexible and interpretable nonparametric mixture model for the conditional density of hippocampal volume which incorporates both continuous and discrete covariates. Simple arguments regarding the interpretation of Bayesian nonparametric regression mixtures lead naturally to regression weights based on normalized sums. To overcome the difficulties in working with the intractable normalizing constant, a novel auxiliary variable Markov chain Monte Carlo (MCMC) scheme is developed. The novel model and MCMC algorithm are applied to study the behavior of hippocampal volume, and the results provide strong support for the theoretical model.

The layout of the paper is as follows. In Section 2, we describe the model and provide its unique provision of interpretability. In Section 3, we introduce the latent variables necessary for estimating the model via MCMC methods. Section 4 describes the MCMC algorithm for posterior inference with further details in the Appendix, and in Section 5, we present a comprehensive simulation study outlining precisely how the model works and what it is capable of achieving, particularly, in comparison to simpler semiparametric models. In Section 6, we present the study of the Alzheimer's disease data. In addition to the detailed calculations required for the MCMC algorithm, the Appendix also includes a discussion of parameter choices and a sensitivity analysis.

2 The regression model

For independent and identically distributed observations, the standard mixture model for density estimation is given by

$$f_P(y) = \int K(y|\theta)\mathrm{d}P(\theta), \tag{1}$$

where $K(\cdot|\theta)$ is a parametric family of density functions defined on $\mathbb{Y}$ and P is a probability measure on the parameter space Θ. In a Bayesian setting, this model is completed with a prior distribution on the mixing measure P. A common prior choice, the stick-breaking prior, assumes P is a discrete random measure and can be represented as

$$P = \sum_{j=1}^{\infty} w_j \delta_{\theta_j},$$

for atoms $\theta_j \in \Theta$, taken i.i.d. from some probability measure P_0, known as the base measure; and weights $w_j \geq 0$, such that $\sum_j w_j = 1$ (a.s.), constructed from a sequence $v_j \overset{ind}{\sim} \mathrm{Beta}(\zeta_{1,j}, \zeta_{2,j})$ with $w_j = v_j \prod_{j'<j}(1 - v_{j'})$. The mixture model (Lo, 1984) can then be expressed as a countable convex combination of kernels

$$f_P(y) = \sum_{j=1}^{\infty} w_j K(y|\theta_j).$$

For the covariate dependent density estimation problem in which we are interested, the mixture model (1) can be adapted by allowing the mixing distribution P_x to depend on the covariate x and replacing the density model $K(y|\theta)$ with a regression model $K(y|x,\theta)$, such as a linear model. Hence, for every $x \in \mathbb{X}$,

$$f_{P_x}(y|x) = \int K(y|x,\theta)\mathrm{d}P_x(\theta).$$

Once again, the Bayesian model is completed by assigning a prior distribution on the family $(P_x)_{x\in\mathbb{X}}$ of covariate dependent mixing probability measures. If the prior gives probability one to the set of discrete probability measures, then

$$P_x = \sum_{j=1}^{\infty} w_j(x)\delta_{\theta_j(x)}, \quad \text{and} \quad f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j(x) K(y|x, \theta_j(x)), \tag{2}$$

where $\theta_j(x) \in \Theta$, and the $w_j(x) \geq 0$ are such that $\sum_j w_j(x) = 1$ (a.s.) for all $x \in \mathbb{X}$. This general model was introduced by MacEachern (1999; 2000), who focused on the case when the weights are constant functions of x, $w_j(x) = w_j$, defined in accordance with a Dirichlet process (DP). This simplified version of the model is popular, as inference can be carried out using any of the well established algorithms for DP mixture models (see e.g. Neal, 2000; Papaspiliopoulos and Roberts, 2008; Kalli et al., 2011).

Recent developments explore the use of covariate dependent weights. To simplify computations and ease interpretation, atoms are usually assumed not to depend on the covariates. The main constraint for prior specification, in this case, is the condition, $\sum_j w_j(x) = 1$ for all $x \in \mathbb{X}$, which is non trivial for an infinite number of positive weights. The only technique currently in use for directly defining the covariate dependent weights is through the stick-breaking representation, given by

$$w_1(x) = v_1(x) \text{ and for } j > 1 \quad w_j(x) = v_j(x) \prod_{j' < j} (1 - v_{j'}(x)), \tag{3}$$

where the $(v_j(\cdot))$ are independent processes on $\mathbb{X}$ and independent of the atoms, (θ_j). There are various proposals for the construction of the $v_j(x)$, see e.g. Griffin and Steel (2006); Dunson and Park (2008); Rodriguez and Dunson (2011); Chung and Dunson (2009); Ren et al. (2011); or Dunson (2010) and Müller and Quintana (2010) for reviews of nonparametric regression mixture models.

The stick-breaking definition poses challenges in terms of the various choices that need to be made for functional shapes and hyperparameters when defining the $(v_j(x))$. The difficulties are amplified by the lack of interpretation of the quantities involved. Moreover, combining continuous and discrete covariates in a useful way is not straightforward. We, therefore, propose a different construction of the covariate dependent weights, which follows from an alternative perspective on mixture models. The idea is to realize that, in the i.i.d. setting, each weight contains information about the applicability of each parametric component, within the sample space $\mathbb{Y}$. In a regression setting, covariate dependent weights are necessary because it is not reasonable to assume that such importance is equal throughout the entire covariate space $\mathbb{X}$; rather, it depends on the value x. Since the nature of such dependence is unknown, the uncertainty about it should be incorporated through prior specification.

In the nonparametric regression mixture model

$$f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j(x) K(y|x, \theta_j),$$

each covariate dependent weight $w_j(x)$ represents the probability that an observation with a covariate value of x comes from the j^{th} parametric regression model $K(y|x, \theta_j)$. Thus, letting d be the random variable indicating the component from which an observation is generated, we have that $w_j(x) = p(d = j|x)$. A simple application of Bayes theorem implies

$$p(d = j|x) \propto p(d = j)p(x|d = j),$$

where $p(d = j)$ represents the probability that an observation, regardless of the value of the covariate, comes from parametric regression model j; and $p(x|d = j)$ describes how likely it is that an observation generated from regression model j has a covariate value of x. Therefore, $p(x|d = j)$ can be defined to reflect prior beliefs as to where in the covariate space the regression model j will have the largest relative applicability. A natural and simple way to achieve this is to define it through a parametric kernel function $K(x|\psi_j)$ and with some prior on the ψ_j. Uncertainty about the $p(d = j) :=$ w_j is expressed through a prior on the infinite dimensional simplex. Putting things together, and incorporating the normalizing constant, we have that

$$w_j(x) = \frac{w_j K(x|\psi_j)}{\sum_{j'=1}^{\infty} w_{j'} K(x|\psi_{j'})}, \tag{4}$$

where $0 \leq w_j \leq 1$ for all j and $\sum_{j=1}^{\infty} w_j = 1$.

Note that the conditional densities $p(x|j)$ are not related to whether the covariates are picked by an expert or sampled from some distribution, which itself could be known or unknown. They only indicate the prior belief about where, in $\mathbb{X}$, regression model j best applies. Moreover, the density $p(x) = \sum_{j=1}^{\infty} P(j)\, p(x|j)$ does not correspond to the distribution from which the covariates are sampled, if indeed they are sampled; it simply represents the likelihood that an observation has a covariate value of x. The key element that must be defined is the kernel $K(x|\psi_j)$. If x is a continuous covariate, a natural choice is the normal density function. In this case, the interpretation would be that there is some central location $\mu_j \in \mathbb{X}$ where regression model j applies best, and a parameter τ_j describing the rate at which the applicability of the model decays around μ_j. On the other hand, if x is discrete, then a standard distribution on discrete spaces can be used, such as the Bernoulli or its generalization, the categorical distribution. Even if x is a combination of both discrete and continuous covariates, it is still possible to specify a joint density by combining both discrete and continuous distributions. This will be explained and demonstrated later on in the paper.

It is to be noted that the infinite sum in the denominator of (4) introduces an intractable normalizing constant for which no posterior simulation methods are currently available. Only finite versions of this type of model have been introduced in the literature (see e.g. Pettitt et al., 2003; Møller et al., 2006; Murray et al., 2006; Adams et al., 2008), since simulation methods are available only for the finite case. In the next section, we introduce a suitable set of latent variables, that solves the infinite dimensional intractable normalizing constant problem.

3 The latent model

The aim of this section is to re-express the model in terms of latent variables, which are essential for Bayesian inference. For a sample $\big((y_1, x_1), \ldots, (y_n, x_n)\big)$, the likelihood for the proposed model is given by

$$f_P(y_{1:n} \,|\, x_{1:n}) = \prod_{i=1}^{n} \left(\sum_{j=1}^{\infty} w_j(x_i)\, K(y_i|x_i, \theta_j) \right), \tag{5}$$

with covariate dependent weights given by expression (4). The infinite sum in the denominator constitutes an intractable normalizing constant, which makes inference infeasible. However, through a simple trick, which relies on the series expansion,

$$\sum_{k=0}^{\infty} (1-r)^k = r^{-1}, \text{ for } 0 < r < 1, \tag{6}$$

we can move the infinite sum from the denominator to the numerator, thus making inference possible, following the introduction of auxiliary variables.

In order to illustrate the ideas with a simplified notation, we start by considering the likelihood of a single data point. We assume that the first q elements of x represent discrete covariates, each x_h taking values in $\{0, \ldots, G_h\}$, for $h = 1 \ldots, q$; the last p elements of x represent continuous covariates. In this case, we let

$$K(y|x, \theta_j) = \mathrm{N}(y|X\beta_j, \sigma_j^2),$$

$$K(x|\psi_j) = \prod_{h=1}^{q} \mathrm{Cat}(x_h|\rho_{j,h}) \prod_{h=1}^{p} \mathrm{N}(x_{h+q}|\mu_{j,h}, \tau_h^{-1}),$$

where $\theta_j = (\beta_j, \sigma_j)$, $\psi_j = (\rho_j, \mu_j, \tau)$, $X = (1, x')$; and $\mathrm{Cat}(\cdot|\rho_h)$ represents the categorical distribution,

$$\mathrm{Cat}(x_h|\rho_h) = \prod_{g=0}^{G_h} \rho_{h,g}^{\mathbf{1}\,(x_h = g)}.$$

For simplicity, in the above expression we have $\tau_j \equiv \tau$ for all j, but this restriction may be removed with some realistic assumptions on τ_j.

The likelihood of the single data point (y, x) may be written as

$$f_P(y \,|\, x) = \frac{1}{r(x)} \sum_{j=1}^{\infty} w_j K(x|\psi_j)\, K(y \,|\, x, \theta_j),$$

where

$$r(x) = \sum_{j=1}^{\infty} w_j\, K(x|\psi_j); \quad K(x|\psi_j) = \prod_{h=1}^{q+p} K(x_h|\psi_{j,h});$$

and

$$K(x_h|\psi_{j,h}) = \begin{cases} \prod_{g=0}^{G_h} \rho_{h,g}^{\mathbf{1}\,(x_h=g)} & h = 1, \ldots, q \\ \\ \exp\{-\frac{1}{2}\tau_{h-q}(x_h - \mu_{j,h-q})^2\} & h = q+1, \ldots, q+p. \end{cases}$$

Notice that we have redefined the kernel function $K(x|\psi_j)$ by cancelling the precision term τ from the normal density, which appears both in the numerator and the denominator of the normalized weights expression. In this way, we guarantee that $0 < r(x) < 1$ for all $x \in \mathbb{X}$, so we can apply the series expansion (6) to write

$$\frac{1}{r(x)} = \sum_{k=0}^{\infty} \left[1 - \sum_{j=1}^{\infty} w_j\, K(x|\psi_j)\right]^k = \sum_{k=0}^{\infty} \left[\sum_{j=1}^{\infty} w_j (1 - K(x|\psi_j))\right]^k,$$

where the last equality relies on the fact that $\sum_{j=1}^{\infty} w_j = 1$ almost surely. This trick allows us to move the infinite sum from the denominator to the numerator and equivalently express the likelihood as

$$f_P(y \,|\, x) = \sum_{j=1}^{\infty} w_j K(x|\psi_j)\, K(y \,|\, x, \theta_j) \sum_{k=0}^{\infty} \left[\sum_{j=1}^{\infty} w_j (1 - K(x|\psi_j))\right]^k. \qquad (7)$$

We now introduce a latent variable k taking values in $\{0, \ldots, \infty\}$, where the joint density of (y, k) given x and the model parameters is

$$f_P(y, k \,|\, x) = \sum_{j=1}^{\infty} w_j K(x|\psi_j)\, K(y \,|\, x, \theta_j) \left[\sum_{j=1}^{\infty} w_j (1 - K(x|\psi_j))\right]^k.$$

This allows us to deal with the mixture in the usual way, by introducing a latent variable d to indicate the mixture component to which a given observation is associated. Thus, we obtain

$$f_P(y,k,d\,|\,x) = w_d K(x|\psi_d)\, K(y\,|\,x,\theta_d) \left[\sum_{j=1}^{\infty} w_j (1 - K(x|\psi_j)) \right]^k .$$

For the remaining sum, we have the exponent k to consider. We first re-write this term as the product of k copies of the infinite sum,

$$f_P(y,k,d\,|\,x) = w_d K(x|\psi_d)\, K(y\,|\,x,\theta_d) \prod_{l=1}^{k} \sum_{j_l=1}^{\infty} w_{j_l} (1 - K(x|\psi_{j_l})),$$

and then, introduce k latent variables, $D_1, \ldots, D_k$, arriving at the full latent model

$$f_P(y,k,d,D\,|\,x) = w_d K(x|\psi_d)\, K(y\,|\,x,\theta_d) \prod_{l=1}^{k} w_{D_l} (1 - K(x|\psi_{D_l})).$$

It is easy to check that the original likelihood (7) is recovered by marginalizing over the d, k and $D = (D_1, \ldots, D_k)$.

For a sample of size $n \geq 1$ we simply need n copies of the latent variables. Therefore, the full latent model is given by

$$\begin{aligned} f_P(y_{1:n}, k_{1:n}, d_{1:n}, D_{1:n}\,|\,x_{1:n}) = &\prod_{i=1}^{n} w_{d_i} K(x_i|\psi_{d_i}) K(y_i \mid x_i, \theta_{d_i}) \\ &\prod_{l=1}^{k_i} w_{D_{l,i}} \left(1 - K(x_i|\psi_{D_{l,i}})\right). \end{aligned} \tag{8}$$

Once again, we note that the original likelihood (5) can be easily recovered by marginalizing over the $d_{1:n}, k_{1:n}$, and $D_{1:n}$. However, the introduction of these latent variables makes Bayesian inference possible, via posterior simulation of the (w_j), the (θ_j) and the (ψ_j), as we show in the next section.

4 Posterior inference via MCMC

A prior for P, defined by a prior specification for the weights (w_j) and the parameters, (θ_j) and (ψ_j), completes the Bayesian model. Our focus for the prior on the weights

(w_j) is on stick-breaking priors (Ishwaran and James (2001)). Therefore, for some positive sequence $(\zeta_{1,j}, \zeta_{2,j})_{j=1}^{\infty}$ and independent $v_j \sim \mathrm{Beta}(\zeta_{1,j}, \zeta_{2,j})$ variables, we have

$$w_1 = v_1, \quad \text{and for } j > 1, \quad w_j = v_j \prod_{j'<j} (1 - v_{j'}).$$

Some important examples of this type of prior are the Dirichlet process, when $\zeta_{1,j} = 1$ and $\zeta_{2,j} = \zeta$ for all j; the Poisson-Dirichlet process, when $\zeta_{1,j} = 1-\zeta_1$ and $\zeta_{2,j} = \zeta_2 + j\zeta_1$ for $0 \leq \zeta_1 < 1$ and $\zeta_2 > -\zeta_1$; and the two parameter stick-breaking process where $\zeta_{1,j} = \zeta_1$ and $\zeta_{2,j} = \zeta_2$ for all j.

To complete the prior specification, the (θ_j, ψ_j) are i.i.d. from some fixed distribution F_0 and independent from the (v_j). We define F_0 through its associated density f_0, which in this case is defined by the product of the following components,

$$f_0(\beta_j, \sigma_j^2) = \mathrm{N}(\beta_j \,|\, \beta_0, \sigma_j^2 C^{-1}) \mathrm{Ga}(1/\sigma_j^2 \,|\, \alpha_1, \alpha_2);$$

$$f_0(\mu_j, \tau) = \prod_{h=1}^{p} \mathrm{N}(\mu_{j,h} \,|\, \mu_{0,h}, (\tau_h c_h)^{-1}) \mathrm{Ga}(\tau_h \,|\, a_h, b_h); \quad \text{and} \quad f_0(\rho_j) = \prod_{h=1}^{q} \mathrm{Dir}(\rho_{j,h} \,|\, \gamma_h).$$

Together with the joint latent model, this provides a joint density for all the variables which need to be sampled for posterior estimation, i.e. the $(w_j, \theta_j, \psi_j, k_i, d_i, D_{l,i})$.

However, there is still an issue due to the infinite choice of the $(d_i, D_{l,i})$, which we overcome through the slice sampling technique of Kalli et al. (2011). Accordingly, in order to reduce the choices represented by $(d_i, D_{l,i})$ to a finite set, we introduce new latent variables, $(\nu_i, \nu_{l,i})$, which interact with the model through the indicating functions $\mathbf{1}\big(\nu_i < \exp(-\xi d_i)\big)$ and $\mathbf{1}\big(\nu_{l,i} < \exp(-\xi D_{l,i})\big)$, for some $\xi > 0$. Hence, the full conditional distributions for the index variables are given by

$$\mathbb{P}(d_i = j | \cdots) \propto w_j \exp(\xi j)\, K(x_i | \psi_j) K(y_i \,|\, x_i, \theta_j)\, \mathbf{1}(1 \leq j \leq J_i),$$
$$\mathbb{P}(D_{l,i} = j | \cdots) \propto w_j \exp(\xi j)\, (1 - K(x_i | \psi_j))\, \mathbf{1}(1 \leq D_{l,i} \leq J_{l,i}),$$

where $J_i = \lfloor -\xi^{-1} \log \nu_i \rfloor; \quad J_{l,i} = \lfloor -\xi^{-1} \log \nu_{l,i} \rfloor$. Note that, at any given iteration, the full conditional densities for the variables involved in the MCMC algorithm do not depend on values beyond $J = \max_{l,i}\{J_i, J_{l,i}\}$, so we only need to sample a finite number of the (ψ_j, θ_j, w_j).

The $(w_j)_{j=1}^J$ can be updated at each iteration of the MCMC algorithm in the usual way, that is, by making $w_1 = v_1$ and, for $j > 1$, $w_j = v_j \prod_{j'<j}(1 - v_{j'})$, where the (v_j) are sampled independently from Beta distributions with updated parameters (specified in the Appendix). The variables involved in the linear regression kernel, that is, the (β_j, σ_j^2), are also updated in the standard way. Since the normal-inverse gamma base measure is conjugate, we simply need to sample from a normal-inverse gamma distribution with updated parameters, detailed in the Appendix.

The full conditional distribution for the $(\psi_j)_{j=1}^J$ seems somewhat more complicated, due to the additional product term in the latent model (equation (8)), involving the latent variables (k_i) and $(D_{l,i})$. However, such a product can be easily transformed into a truncation term, by the introduction of additional auxiliary variables. Thus, posterior simulation for the $(\psi_j)_{j=1}^J$ is achieved by sampling from standard truncated distributions with updated parameters, which can be easily calculated due to the choice of conjugate base measure. The details of this procedure, as well as the resulting updated parameters and truncations are presented in the Appendix. At this point, we only mention that the introduction of the additional variables does not pose a problem, since they are all conditionally independent given the $(\psi_j)_{j=1}^J$, and hence can be sampled in parallel, using the "parfor" routine in Matlab.

Finally, for the update of each k_i, we use ideas involving a version of reversible jump MCMC (see Green, 1995) introduced by Godsill (2001), to deal with the change of dimension in the sampling space. We start by proposing a move from k_i to $k_i + 1$ with probability 1/2, and accepting it with probability

$$\min\left\{1, \sum_{j=1}^{J} w_j \left(1 - K(x_i|\psi_j)\right)\right\}.$$

In this case, we need to sample the additional index D_{i,k_i+1}, and we choose $D_{i,k_i+1} = j$ with probability proportional to $w_j\,(1 - K(x_i|\psi_j))$, for $j = 1, \ldots, J$. Similarly, if $k_i > 0$, a move from k_i to $k_i - 1$ is proposed with probability $1/2$, and accepted with probability

$$\min\left\{1, \left[\sum_{j=1}^{J} w_j\,(1 - K(x_i|\psi_j))\right]^{-1}\right\}.$$

It is therefore possible to perform posterior inference for the nonparametric regression model proposed, via an MCMC scheme applied to the latent model. We have successfully implemented the method in Matlab (R2012a), and present some results in the next section. In the following examples, the aim is prediction and predictive density estimation, which under the quadratic loss are, respectively, given by

$$\mathrm{E}[Y_{n+1}|y_{1:n}, x_{1:n+1}] = \mathrm{E}\left[\sum_{j=1}^{\infty} w_j(x_{n+1}) X_{n+1}\beta_j \Big| y_{1:n}, x_{1:n}\right], \tag{9}$$

$$f(y_{n+1}|y_{1:n}, x_{1:n+1}) = \mathrm{E}\left[\sum_{j=1}^{\infty} w_j(x_{n+1}) \mathrm{N}(y|X_{n+1}\beta_j, \sigma_j^2) \Big| y_{1:n}, x_{1:n}\right], \tag{10}$$

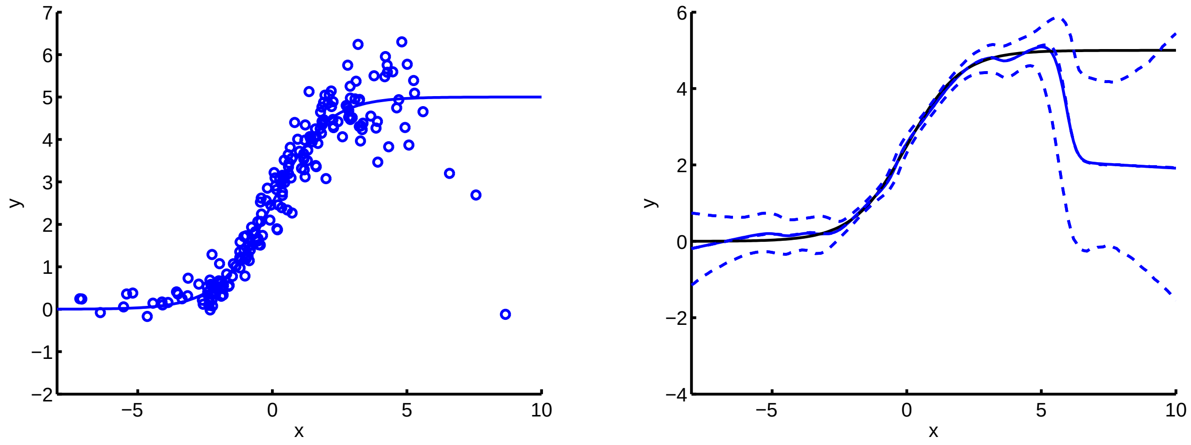
where and $X_{n+1} = (1, x'_{n+1})$; and the expectation is taken with respect to the posterior distribution of (w_j, θ_j, ψ_j). MCMC estimates for these quantities are used, as specified in the Appendix.

5 Simulation Study

To demonstrate the ability of the model to recover a complex regression function with covariate dependent errors, we simulate $n = 200$ data points (depicted in Figure 1a) through the following formula,

$$x_i \overset{iid}{\sim} \mathrm{N}(\cdot|0, 2.5^2), \quad y_i|x_i \overset{ind}{\sim} \mathrm{N}\left(\cdot \Big| \frac{5}{1 + \exp(-x_i)}, \left[\frac{1}{4} + \exp\left(\frac{x_i - 6}{3}\right)\right]^2\right).$$

Our model is given by



(a) Simulated data

(b) Predicted regression function

Figure 1: The left panel depicts the data and the true regression mean. The right panel depicts the predicted regression function (in blue) for a grid of new covariate values, along with 95% pointwise credible intervals; the black line represents the true mean function.

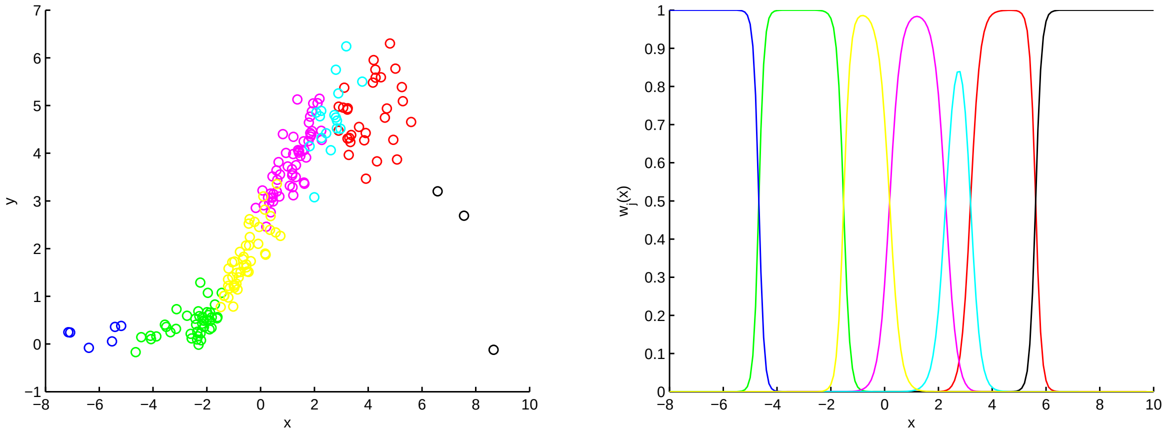
$$f_P(y|x) = \sum_{j=1}^{\infty} w_j(x) N(y|X\beta_j, \sigma_j^2), \quad \text{with} \quad w_j(x) = \frac{w_j \exp(-\tau/2(x-\mu_j)^2)}{\sum_{j'=1}^{\infty} w_{j'} \exp(-\tau/2(x-\mu_{j'})^2)}.$$

The prior for (w_j) and (θ_j, ψ_j) is described in Section 4. The prior choice for the (w_j) is a Dirichlet process with unit mass, i.e. $\zeta_{1,j} = \zeta_{2,j} = 1$, and for the prior of (θ_j, ψ_j), we set

$$\beta_0 = (5/2, 5/8)'; \quad C^{-1} = \text{diag}(4, 1/4); \quad \alpha_1 = 1; \quad \alpha_2 = 1;$$
$$\mu_0 = 0; \quad c = 1/8; \quad a = 1; \quad b = 1.$$

An explanation for the choice of these quantities can be found in the Appendix, along with a sensitivity analysis. Inference is carried out via the algorithm discussed in Section 4 with 5,000 iterations after a burn-in period of 5,000.

Figure 1b depicts, in blue, the estimated regression function for a grid of unobserved x values, along with 95% pointwise credible intervals. The true regression function is



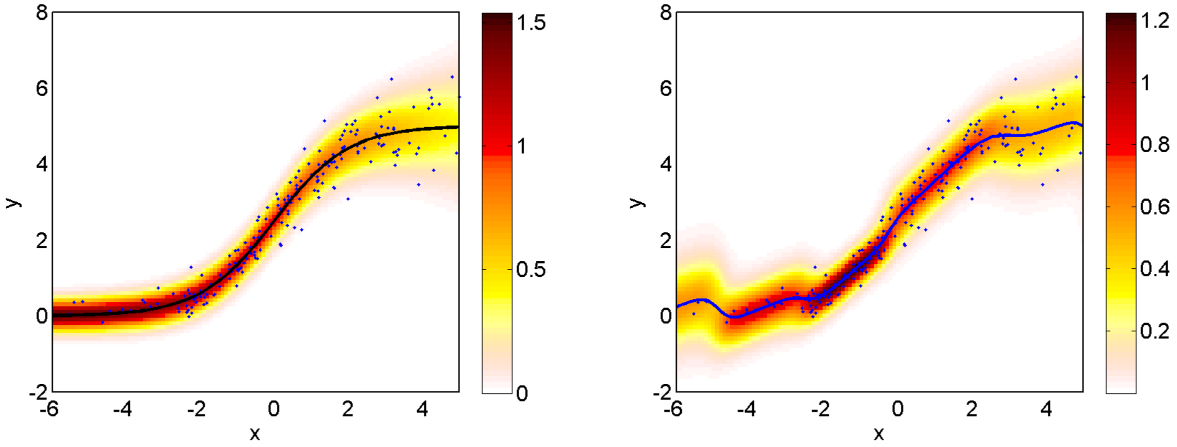
(a) Partition with highest probability

(b) Covariate-dependent weights

Figure 2: The left panel depicts the partition with the highest posterior probability, where the data are colored by component membership. The right panel depicts a sample of the covariate-dependent weights associated to this partition.

shown in black. For large values of x we can observe a deterioration of the curve estimate, which is pulled down by some extreme observations. This is to be expected due to a lack of data for large x-values. Indeed, with an increased sample size, this behavior is corrected (analysis not shown).

The flexibility in estimating the regression function relies heavily on the posterior distribution of the covariate dependent weights. The left panel of Figure 2 depicts the partition with highest estimated posterior probability, with data points coloured by component membership. The right panel of Figure 2 shows a posterior sample of the covariate-dependent weights as a function of x, given this partition. It is important to observe that *aposteriori* the weights are able to peak close to one in areas of high applicability of their associated linear regression models and decay smoothly or sharply, as needed, when the covariates move away from this area. For values of x



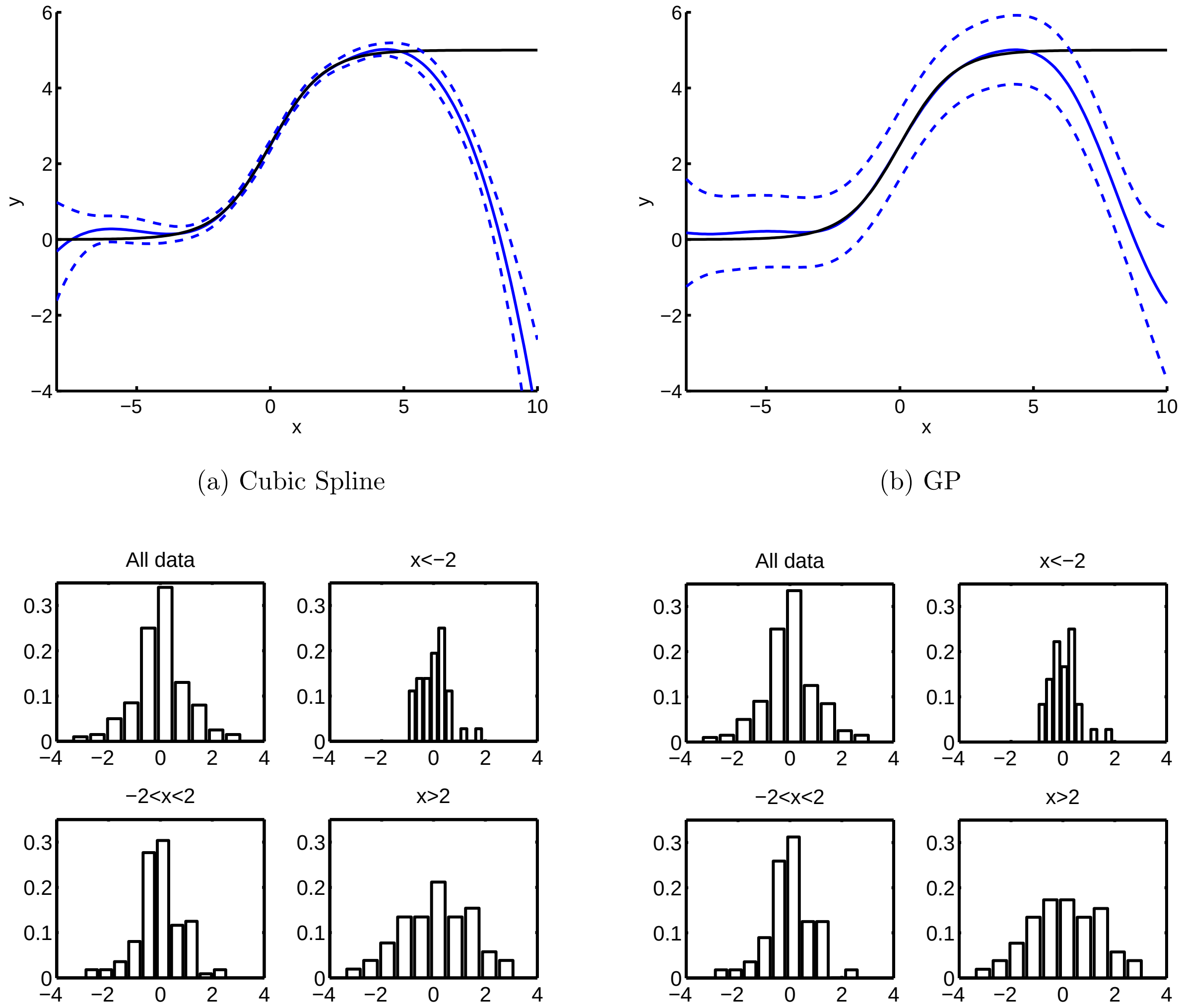
(a) True conditional densities

(b) Estimated conditional densities

Figure 3: The left panel depicts a heat plot of the true conditional densities $f(y|x)$ for a grid of covariate values; the right panel corresponds to the estimated conditional densities. In both cases, the corresponding mean curve is shown, along with the data.

around -3 (green cluster), a single linear regression model dominates; for values around 3 (cyan cluster), the dominance is less clear; while, for values around 0 a combined effect of two linear models is indicated by the dependent weights. We emphasize that Figure 2 clearly shows that the kernels in the covariate space are not modelling the density of x, which is a simple Gaussian, but reflect the regions in the covariate space where each linear regression model applies.

We are also able to produce estimates of the predictive densities, that is, the entire conditional density $f(y|x)$ at any value of x in the covariate space. Results are shown in Figure 3b. The estimated densities are represented through heat maps, where a darker color indicates higher density values. The estimated densities can be compared with the true conditional densities, shown in Figure 3a. As is expected, the estimated variance is higher than the true for small values of x where less data is observed.



(a) Cubic Spline

(b) GP

(c) Cubic Spline

(d) GP

Figure 4: The predicted regression function (in blue) and 95% pointwise confidence (spline) and credible (GP) intervals for a grid of new covariate values along with true mean function in black for the cubic spline and GP models, respectively, and histograms of the standardized residuals for restricted ranges of covariate values.

However, it is clear from the picture that the change, with x, of the variance of $y|x$ is recovered by the model.

Finally, we consider a comparison with semiparametric models. Figure 4 plots the predicted regression function for a grid of new covariate values for two competing models, namely, the cubic spline (CS) and Gaussian process (GP) models, implemented in the *crs* package in **R** and the *GPML* toolbox in **Matlab**, respectively. The *crs* package includes an automatic tool which selects the "best" spline model over a range of degrees, number of knot points, and the choice of equally-spaced knots or knots placed at the quantiles. When restricted to cubic splines (Figure 4a), the selected spline model contained 6 knot points placed at the quantiles. Without this restriction, the selected spline model had a degree of 7 with 8 knot points placed at the quantiles. Results in this case are not shown since the model was outperformed by the cubic spline restricted version under the performance metrics that we consider. The Gaussian process model (Figure 4b) assumed a squared exponential covariance function.

For these semiparametric models, the poor mean function predictions for large values of x with overly narrow confidence/credible intervals are clearly observed in Figures 4a and 4b. It is important to note that these simpler models assume i.i.d. standard normal errors, an assumption that is clearly violated in this case, as can be observed in the histograms of the corresponding standardized residuals obtained after fitting these models (Figures 4c and 4d). This raises questions about the use of these models for this dataset, particularly for prediction. Table 1 provides a numerical comparison of the models with respect to some commonly used distance measures between the true and estimated regression curves, as well as between the true and estimated conditional densities. In terms of fit (error calculated on observed covariate values), the models are quite comparable, although measures which are more sensitive to outliers (L_2 and $\max(L_1)$) are improved for the proposed model. However, in terms of prediction (estimation for unobserved x values), our proposed nonparametric model with normalized weights is superior, as would be expected given the non-normality of the errors.

Estimated Item		**Error measure**	**CS**	**GP**	**NW**
Regression mean	**Fit**	$\hat{L}_1$	**0.08**	**0.08**	0.10
		$\hat{L}_2$	0.44	0.41	**0.38**
	Predictive	$\hat{L}_1$	0.98	0.88	**0.79**
		$\hat{L}_2$	2.33	1.95	**1.46**
Conditional densities	**Fit**	$\text{avg}(\hat{L}_1)$	**0.15**	0.30	0.22
		$\max(\hat{L}_1)$	5.12	1.79	**1.25**
	Predictive	$\text{avg}(\hat{L}_1)$	1.07	0.70	**0.25**
		$\max(\hat{L}_1)$	9.81	1.79	**1.25**

Table 1: Model Comparison: NW (normalized weights) stands for our model

To summarize the simulation study; it is important to model $f(y|x)$ in its entirety as a density rather than just a mean, for example. It is also important to model the weights as a function of x. While we do not believe other Bayesian nonparametric models could improve on things, but could do as well as our model, the model proposed here does have full interpretation for the parameters.

6 Alzheimer's disease study

Hippocampal volume is one of the best established and most studied biomarkers because of its known association with memory skills and relatively easy identification in sMRI. In two recent papers, Jack et al. (2010) and Frisoni et al. (2010) discussed a hypothetical model for the dynamics of hippocampal volume as a function of age and disease severity. If confirmed, this model would have important implications for the use of hippocampal volume to measure the efficacy of treatments in clinical trials.

The clinical stages of the AD are divided into three phases (Jack et al. (2010)); the pre-symptomatic phase, the prodromal phase, and the dementia phase. During the pre-symptomatic phase, some AD pathological changes are present, but patients do not exhibit clinical symptoms. This phase may begin possibly 20 years before the onset of clinical symptoms. The pre-prodromal stage of AD is known as mild cognitive impairment (MCI); patients diagnosed with MCI exhibit early symptoms of cognitive impairment, but do not meet the dementia criteria. The final stage of AD is dementia, when patients are officially diagnosed AD. Jack et al. (2010) and Frisoni et al. (2010) hypothesized that hippocampal volume evolves sigmoidally over time, with changes starting slightly before the MCI stage and occurring until late in dementia phase. The steepest changes are supposed to occur shortly after the dementia threshold has been crossed.

To provide validation for this model, we study the evolution of hippocampal volume as a function of age, gender, and disease status. Data was obtained from the Alzheimer's Disease Neuroimaging Initiative database which is publicly accessible at UCLA's Laboratory of Neuroimaging[1]. The ADNI database contains neuroimaging,

[1] The ADNI was launched in 2003 by the National Institute on Aging (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

biological, and clinical data, along with summaries of neuroimages, including the volume of various brain structures. The dataset analysed here consists of the volume hippocampus obtained from the sMRI performed at the first visit for 736 patients. Of the 736 patients in our study, 159 have been diagnosed with AD, 357 have MCI, and 218 are cognitively normal (CN). Figure 5 displays the data.

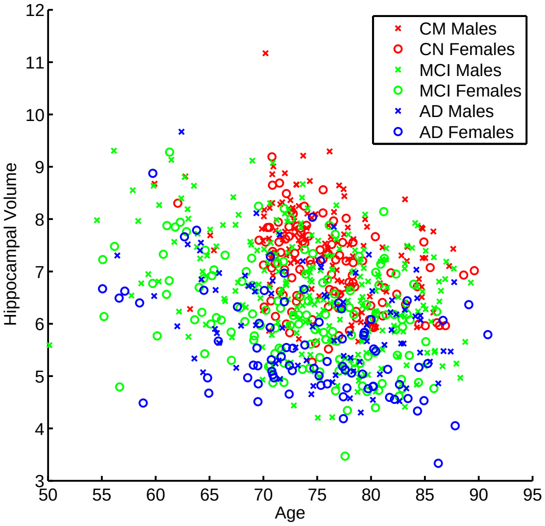
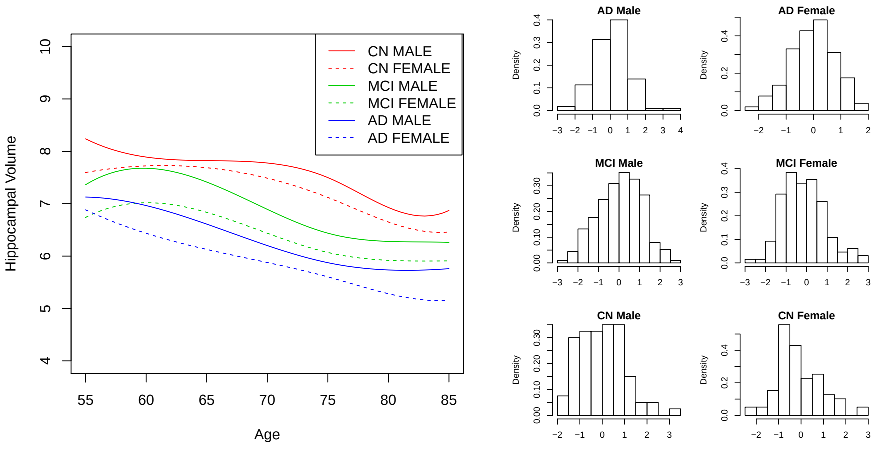


Figure 5: Hippocampal volume plotted against age. The data are colored by disease status with circles representing females and crosses representing males.

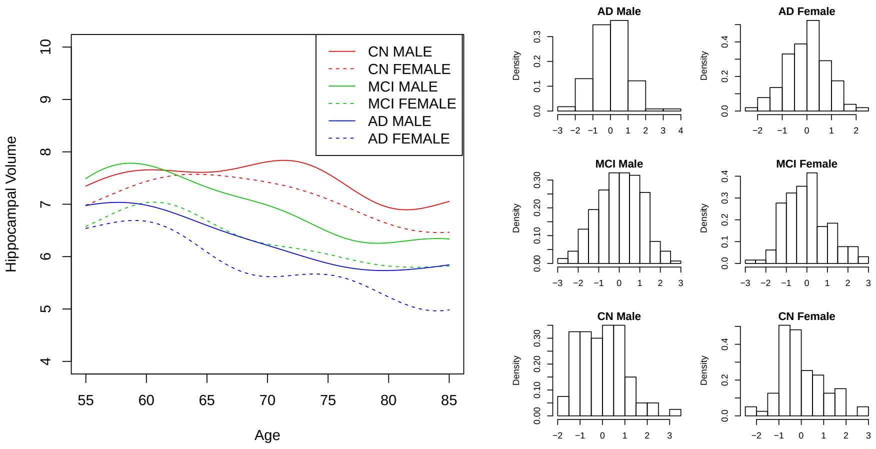
As discussed in Jack et al. (2010), we not only expect non-linearity in the regression function, but also suspect the possibility of non-normal and covariate dependent errors,

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`.



(a) Regression function (b) Standardized errors

Figure 6: Cubic spline model: (6a) estimated regression function and (6b) histogram of the standardized errors as a function of sex and disease status.



(a) Regression function (b) Standardized errors

Figure 7: Gaussian process model: (7a) estimated regression function and (7b) histogram of the standardized errors as a function of sex and disease status.

for example due to the presence of unobserved neuroprotective genes. Indeed, in a preliminary semiparametric analysis where the errors are assumed to be i.i.d. normal, we find some peculiarities in the model fit. Figures 6 and 7 display the estimated regression function and histogram of the standardized errors within each combination of sex and disease status for the semi-parametric cubic spline and Gaussian process models, respectively, which are implemented in the *crs* and *kernlab* packages in **R**. Notice that both of these models tend to overfit the data to overcome the rigid assumption on the errors. Furthermore, we find some abnormal behaviour in the errors that depends on sex and disease status. As we learned in the simulation study, this odd behavior in the fitted errors for the semiparametric models raises doubts about their use for this dataset and can be a signal for poor prediction.

In order to fully capture the dynamics of the data, a nonparametric approach which flexibly models both the regression function and the error distribution is needed. To this aim, we consider the model developed in this paper, specifically, the infinite Gaussian kernel mixture model with covariate dependent weights given by

$$w_j(x) = \frac{w_j \prod_{h=1}^{2} \prod_{g=0}^{G_h} \rho_{j,h,g}^{1_{x_h=g}} \exp(-\tau/2(x_3-\mu_j)^2)}{\sum_{j'=1}^{\infty} w_{j'} \prod_{h=1}^{2} \prod_{g=0}^{G_h} \rho_{j',h,g}^{1_{x_h=g}} \exp(-\tau/2(x_3-\mu_{j'})^2)},$$

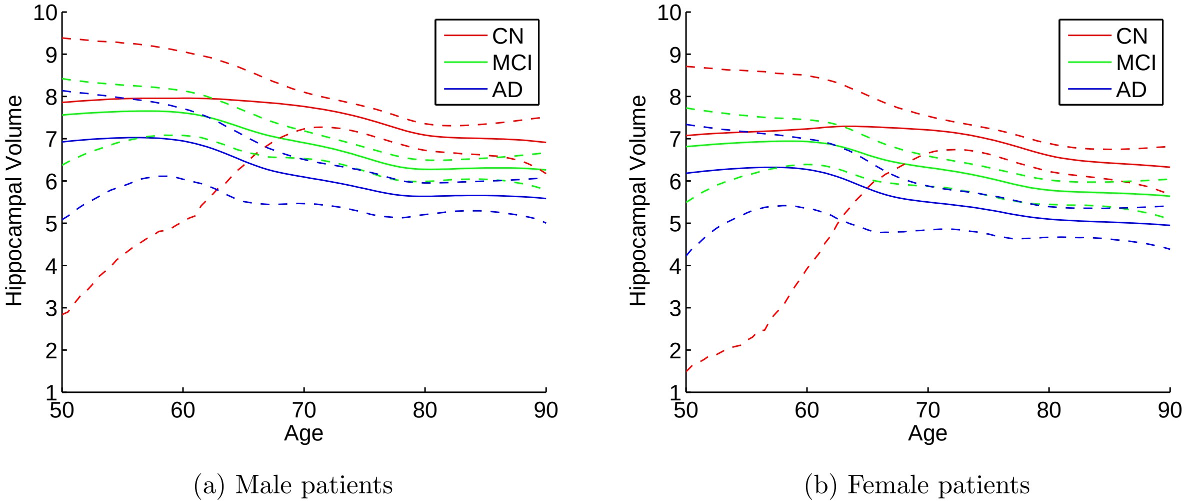
where $G_1 = 1$ (x_1 represents gender) and $G_2 = 2$ (x_2 represents disease status). Note that here age (x_3) is a real number measuring time from birth to exam date and thus, is treated as a continuous covariate.

The prior distribution for w_j and (θ_j, ψ_j) is described in Section 4. The prior parameters for w_j are $\zeta_{1,j} = 1$ and $\zeta_{2,j} = 1$, corresponding to a Dirichlet process prior with a precision parameter of 1. For the prior of (θ_j, ψ_j), we set

$$\beta_0 = (8, -1, -1, -1/4)'; \quad C^{-1} = \text{diag}(4, 1/4, 1/4, 1/50); \quad \alpha_1 = 1; \quad \alpha_2 = 1;$$

$$\gamma_1 = (1, 1)'; \quad \gamma_2 = (1, 1, 1)'; \quad \mu_0 = 72.5; \quad c = 1/4; \quad a_1 = 1; \quad b_h = 1.$$

See the Appendix for an explanation of these parameter choices. Inference is carried out via the algorithm discussed in Section 4 with 23,000 iterations after a burn-in period of 7,000.

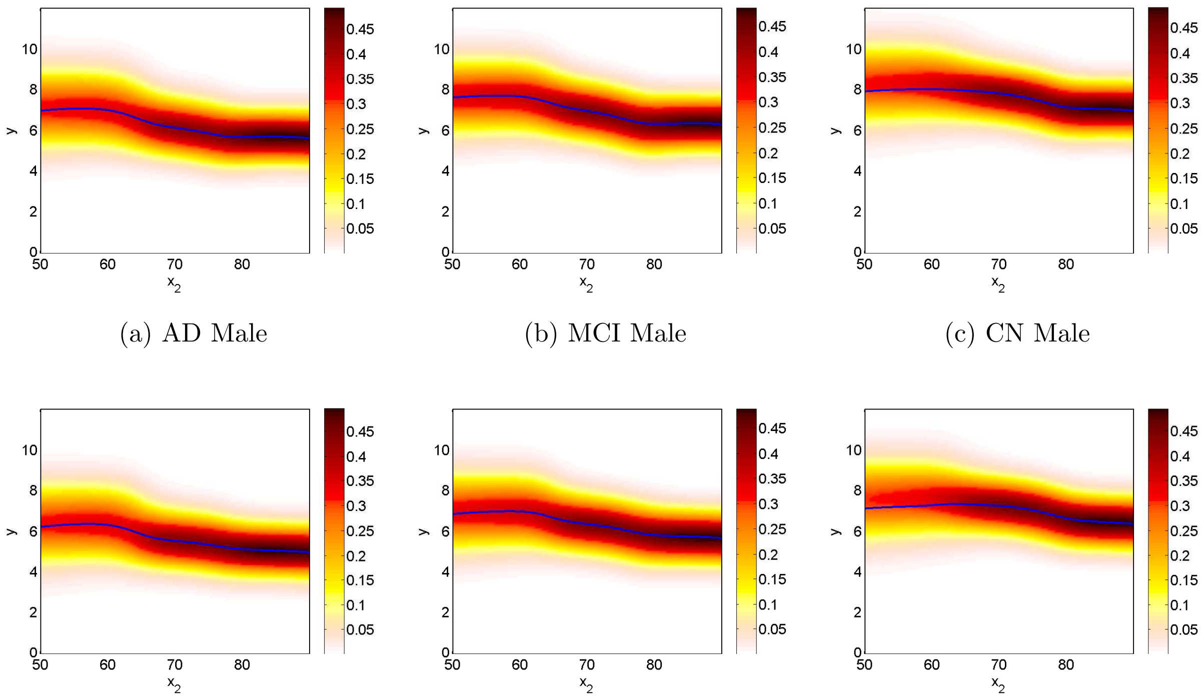


(a) Male patients

(b) Female patients

Figure 8: Estimated mean hippocampal volume as a function of age, disease, and sex. The curves are colored by disease status with dashed lines representing 95% pointwise credible intervals around the estimated regression function.

Figure 8 displays the estimated mean regression function for a grid of ages with all possible combinations of disease status and sex. Interestingly, we observe a confirmation of the hypothesized sigmoidal evolution of hippocampal volume with increasing age. The estimated mean function coincides with the point predictor under the quadratic loss function. In this sense, cognitively normal subjects are predicted to have highest values of hippocampal volume at all ages, and MCI patients are predicted to have higher values of hippocampal volume at all ages when compared with AD patients. This indicates that hippocampal volume may be useful in disease staging during both the MCI and AD phases. With careful examination of Figure 8, we observe that CN patients are predicted to show the most gradual decline with increasing age, while AD patients display the greatest. Notice that, as expected, females are predicted to have lower values of hippocampal volume. We should comment that there is little data for the subgroup of CN subjects under 60, which reflects on the greater uncertainty in the estimation.



(a) AD Male

(b) MCI Male

(c) CN Male

(d) AD Female

(e) MCI Female

(f) CN Female

Figure 9: Heat map of conditional density estimates, i.e. predictve density, for new covariates with a grid of ages between 50 and 90 and all combinations of disease status and sex.

Figure 9 displays the heat map of conditional density estimates, i.e. the predictive densities, for a grid of new ages between 50 and 90 and all combinations of disease status and sex. In a clinical trial setting, the preference is for reliable outcome measures, i.e. biomarkers with small variability. In general, we observe that variance decreases with increasing age, indicating that hippocampal volume is more reliable for elderly patients. The difference is slightly more pronounced for females as opposed to males. In particular, hippocampal volume is predicted to have a large variability for young females across all disease stages, with the largest for young CN females (the subgroup with no data). Instead, for older females, the variance is much smaller for all disease stages. When comparing males across disease status, we notice that young CN patients are predicted to show a large variability compared with young MCI and AD patients, while old MCI patients are predicted to show the largest variability when compared with their CN and AD counterparts.

This figure clearly illustrates a feature which provides a strong motivation for our model, rather than a simpler one which assumes, for example, constant variance and skewness. The data suggest that it is important to model mean, variance, skewness and possibly also kurtosis as being dependent on the covariate values. Hence, a standard model such as $y = m(x) + \sigma\varepsilon$, $\varepsilon \overset{iid}{\sim} \mathrm{N}(0,1)$ will fail to reproduce the results we have obtained for the more general $f(y|x)$ model. Even though the model is necessarily more complicated, all the elements in it are interpretable.

7 Discussion

In this paper, we have described and implemented a fully Bayesian nonparametric approach to examine the evolution of hippocampal volume as a function of age, gender, and disease status. We find that with increasing age, hippocampal volume is predicted to display a sigmoidal decline for cognitively normal, MCI, and AD patients. We also observe the most gradual decline for CN patients, while AD patients are predicted to show the steepest decline. As the approach was nonparametric, no structure was assumed for the regression function, yet our results confirm the hypothetical dynamics of hippocampal volume proposed by Jack et al. (2010). This provides strong statistical support for their model of hippocampal atrophy. A comparison with two commonly used semiparametric models suggest the superiority of the proposed model for prediction, i.e. estimation of the regression curve and conditional densities $f(y|x)$ for unobserved covariate values. Future work in this application will involve examining the dynamics of various biomarkers jointly, which could be accomplished by replacing the normal linear regression component for y with a multivariate linear regression component. Another important future study will consist of combining the cross-sectional data with the longitudinal data for each patient.

In our analysis of the dynamics of hippocampal volume, we have developed a novel Bayesian nonparametric regression model based on normalized covariate dependent weights. The important contributions of this approach are a natural and interpretable structure for the weights, a novel algorithm for exact posterior inference, and the inclusion of both continuous and discrete covariates. We have focused on a univariate and continuous response, but the model and algorithm can be easily extended to accommodate other types of responses by, for example, replacing the normal linear regression component for y with a generalized linear model. Future work will consist of examining theoretical properties of this model.

Acknowledgments. The authors would like to thank the Editor, Associate Editor and two referees for their valuable comments and suggestions.

Appendix

Section 3 details

We specify the full conditional distributions for the MCMC posterior sampling scheme used for inference on the latent model constructed in Section 3.

The sampling of the weights is obtained via the Stick Breaking definition, where the (v_j) must be independently sampled from the corresponding full conditionals,

$$f(v_j | \cdots) = \mathrm{Be}(\zeta_{1,j} + n_j + N_j, \zeta_{2,j} + n_j^+ + N_j^+),$$

where

$$n_j = \sum_i \mathbf{1}(d_i = j); \quad N_j = \sum_{l,i} \mathbf{1}(D_{l,i} = j);$$
$$n_j^+ = \sum_i \mathbf{1}(d_i > j); \quad N_j^+ = \sum_{l,i} \mathbf{1}(D_{l,i} > j).$$

Each of the (β_j, σ_j^2) can be sampled independently across j, from the full conditional density

$$f(\beta_j, \sigma_j^2 | \cdots) \;\; = \;\; \mathrm{N}(\beta_j \,|\, \hat{\beta}_j, \sigma_j^2 \hat{C}_j^{-1}) \mathrm{Ga}(1/\sigma_j^2 \,|\, \hat{\alpha}_{1j}, \hat{\alpha}_{2j}),$$

where

$$\hat{\alpha}_{1j} = \alpha_1 + n_j/2; \qquad \hat{\alpha}_{2j} = \alpha_2 + \frac{1}{2}(\underline{y}_j - \underline{X}_j \beta_0)' W_j (\underline{y}_j - \underline{X}_j \beta_0);$$
$$\hat{\beta}_j = \hat{C}_j^{-1}(C\beta_0 + \underline{X}_j' \underline{y}_j); \qquad \hat{C}_j = C + \underline{X}_j' \underline{X}_j; \qquad W_j = I_j - \underline{X}_j \hat{C}_j^{-1} \underline{X}_j'.$$

Here, $\underline{X}_j$ denotes the matrix with rows given by $X_i = (1, x_i')$ for $d_i = j$; $\underline{y}_j$ is defined analogously; and I_j denotes the identity matrix of size n_j.

We now show how the introduction of an additional set of latent variables enables the update of the $(\psi_j)_{j=1}^J$, as explained in Section 4, and specify the resulting posterior densities and truncation regions. Observe that, for any integer H and vector $(c_1, \ldots, c_H) \in (0,1)^H$, the following identity holds

$$1 - \prod_{h=1}^H c_h = \sum_{u \in \mathbb{U}} \int_{(0,1)^H} \prod_{h=1}^H \left[u_h \mathbf{1}\left(U_h < c_h\right) + (1 - u_h) \mathbf{1}\left(U_h > c_h\right) \right] \mathrm{d}U,$$

where $U = (U_1, \ldots, U_H)$, $u = (u_1, \ldots, u_H)$ and $\mathbb{U}$ is the set of H-dimensional $\{0, 1\}$ vectors of which at least one entry is 0. We can, therefore, introduce latent variables $(u_{i,l,h}, U_{i,l,h})$, for $i = 1, \ldots, n$, $l = 1, \ldots, k_i$ and $h = 1, \ldots, q + p$, to deal with the terms $(1 - \prod_h K(x_{i,h}|\psi_{j,h}))$ in the latent likelihood (equation (8)). The full conditional density for $(\psi_j)_{j=1}^J$ is thus extended to the latent model

$$f(\psi_{1:J}, \{u_{i,l,h}\}, \{U_{i,l,h}\}| \cdots) \propto \prod_{j=1}^{J} f_0(\psi_j) \prod_{i=1}^{n} \prod_{h=1}^{q+p} K(x_{i,h}|\psi_{d_i,h})$$
$$\prod_{l=1}^{k_i} \left[u_{i,l,h}\mathbf{1}\left(U_{i,l,h} < K_{i,l,h}\right) + (1 - u_{i,l,h})\mathbf{1}\left(U_{i,l,h} > K_{i,l,h}\right)\right],$$

where $K_{i,l,h} = K(x_{i,h}|\psi_{D_{i,l},h})$, from which the original conditional density can be recovered by marginalizing over the $(u_{i,l,h}, U_{i,l,h})$.

The latent variables $(u_{i,l,h}, U_{i,l,h})$ can be sampled from their full conditional density by first observing that they are independent across $i = l, \ldots, n$ and $l = 1, \ldots, k_i$. For each i, l, the variable $u_{i,l}$ is a $(q + p)$-dimensional vector of zeros and ones with at least one zero entry. There are $2^{p+q} - 1$ such vectors, and for any u in this set, the update must be done according to the following distribution

$$\mathbb{P}(u_{i,l} = u| \cdots) \propto \prod_{h=1}^{q+p} \left[u_h K(x_{i,h}|\psi_{D_{i,l},h}) + (1 - u_h)(1 - K(x_{i,h}|\psi_{D_{i,l},h}))\right].$$

Conditional on $u_{i,l}$, the latent variables $U_{i,l,h}$ for $h = 1, \ldots, p + q$ are independent and uniformly distributed in the region

$$\left[K(x_{i,h}|\psi_{D_{i,l},h})(1 - u_{i,l,h}), K(x_{i,h}|\psi_{D_{i,l},h})^{u_{i,l,h}}\right].$$

Therefore, the additional variables do not pose a problem for posterior simulation. Furthermore, the introduction of these new variables transforms the latent term, introduced to deal with the intractable normalizing constant, into a product of truncation terms which is multiplied by the usual posterior density for the nonparametric mixture.

We first consider the update of the $(\rho_j)_{j=1}^J$, which is achieved by sampling each $\rho_{j,h}$ independently from a truncated Dirichlet distribution,

$$f(\rho_{j,h} \,|\, \cdots) \propto \mathrm{Dir}(\rho_{j,h} \,|\, \hat{\gamma}_{j,h}) \, \mathbf{1} \, (\rho_{j,h} \in R_{j,h}) \, , \quad \text{where} \quad \hat{\gamma}_{j,h,g} = \gamma_{j,h,g} + \sum_{d_i = j} \mathbf{1} \, (x_{i,h} = g) \, .$$

The truncation region for each of the $(\rho_j)_{j=1}^J$ is given by

$$R_{j,h} = \left\{ \rho \in (0,1)^{G_h} : r_{j,h,g}^- < \rho_g < r_{j,h,g}^+, \; g = 1, \ldots, G_h \right\}$$

and for $g = 0 \ldots, G_h$,

$$\begin{aligned} r_{j,h,g}^- &= \max \left\{ U_{i,l,h} \mathbf{1} \, (x_{i,h} = g) : D_{i,l} = j, u_{i,l,h} = 1 \right\} , \\ r_{j,h,g}^+ &= \min \left\{ U_{i,l,h}^{\mathbf{1} \, (x_{i,h} = g)} : D_{i,l} = j, u_{i,l,h} = 0 \right\} . \end{aligned}$$

We then consider the $(\mu_j, \tau_j)_{j=1}^J$. Recall that $\tau_j = \tau$ for every j, so we update this variable by sampling each τ_h independently from a truncated gamma density,

$$f(\tau_h \,|\, \cdots) \propto \mathrm{Ga}(\tau_h \,|\, \hat{a}_h, \hat{b}_h) \mathbf{1} \, (\tau_h \in T_h),$$

where

$$\hat{a}_h = a_h + J/2 \quad \text{and} \quad \hat{b}_h = b_h + \frac{1}{2} \sum_{i=1}^n (x_{i,h+q} - \mu_{d_i,h})^2 + \frac{1}{2} c_h \sum_{j=1}^J (\mu_{j,h} - \mu_{0,h})^2.$$

The truncation region for each τ_h is an interval $T_h = (\tau_h^-, \tau_h^+)$, where

$$\begin{aligned} \tau_h^- &= \max \left\{ \frac{-2 \log U_{i,l,h+q}}{(x_{i,h+q} - \mu_{D_{i,l},h})^2} : u_{i,l,h+q} = 0 \right\} , \\ \tau_h^+ &= \min \left\{ \frac{-2 \log U_{i,l,h+q}}{(x_{i,h+q} - \mu_{D_{i,l},h})^2} : u_{i,l,h+q} = 1 \right\} . \end{aligned}$$

We then sample each $\mu_{j,h}$ independently from a truncated normal

$$f(\mu_{j,h} \,|\, \cdots) \propto \mathrm{N}(\mu_{j,h} \,|\, \hat{\mu}_{j,h}, (\tau_h \hat{c}_{j,h})^{-1}) \, \mathbf{1} \, (\mu_{j,h} \in A_{j,h}) \, ,$$

where

$$\hat{\mu}_{j,h} = \frac{1}{\hat{c}_{j,h}} \left(c_h \mu_{0,h} + \sum_{d_i = j} x_{i,h+q} \right) ; \quad \hat{c}_{j,h} = c_h + n_j.$$

The truncation region for each of the $\mu_{j,h}$ is an intersection of sets,

$$A_{j,h} = \bigcap_{D_{i,l}=j} A_{i,l,h},$$

where each $A_{i,l,h}$ is defined in terms of the intervals,

$$I_{i,l,h} = \left(x_{i,h+q} - \sqrt{\frac{-2\log U_{i,l,h+q}}{\tau_h}},\, x_{i,h+q} + \sqrt{\frac{-2\log U_{i,l,h+q}}{\tau_h}} \right),$$

as $A_{i,l,h} = I_{i,l,h}$ when $u_{i,l,h+p} = 1$, and $A_{i,l,h} = I^c_{i,j,h}$ when $u_{i,l,h+p} = 0$.

Finally, in order to improve the mixing of the algorithm we applied the label switching moves introduced by Papaspiliopoulos and Roberts (2008). The Markov Chain scheme detailed here and explained in Section 4, produces posterior samples $(w^s_j, \theta^s_j, \psi^s_j)$ for $s = 1, \ldots, S$, which can be used to estimate the regression mean (9) and predictive density (10) via

$$\mathrm{E}[Y_{n+1}|y_{1:n}, x_{1:n+1}] \approx \sum_{s=1}^{S} \sum_{j=1}^{J^s} w^s_j(x_{n+1}) \underline{X}_{n+1} \beta^s_j,$$

$$f(y_{n+1}|y_{1:n}, x_{1:n+1}) \approx \sum_{s=1}^{S} \sum_{j=1}^{J^s} w^s_j(x_{n+1}) \mathrm{N}(y|\underline{X}_{n+1}\beta^s_j, \sigma^{2s}_j),$$

where

$$w^s_j(x_{n+1}) = \frac{w^s_j K(x_{n+1}|\psi^s_j)}{\sum_{j'=1}^{J^s} w^s_{j'} K(x_{n+1}|\psi^s_{j'})}.$$

Prior Specification and Sensitivity Analysis

We discuss the specification of the prior parameters in Sections 5 and 6 and provide a sensitivity analysis with respect to the prior parameters of the simulation study, where a comparison of the results to the true data-generating model is possible. For both examples, based on a visual analysis of the data set and prior knowledge, we were able to determine of maximum range for the parameters, which was then used to select the prior parameters.

We first consider the simulated dataset analysed in Section 5. In order to fit the scatter plot of the data, the local linear components must be allowed to have a slope between $[0, 5/2]$ and an intercept between $[0, 5]$. Thus, we chose to center the prior for the regression kernel parameters on $\beta_0 = (5/2, 5/4)'$ with a variability 4 and $1/4$ for the intercept and slope, respectively, thus allowing them to cover the specified range. The variance σ around the local regression lines should range between $[1/4, 4]$, and the choice of a inverse gamma prior with parameters $(1, 1)$ is sufficiently diffuse to cover that range. Since most of the observed covariates are concentrated in the interval $[-5, 5]$, we chose to center the covariate-related location parameters on $\mu_0 = 0$ with a variability increased by a factor of 8 with respect to the component variability, thus making $c = 1/8$; the precision τ linked to the range of applicability of each regression kernel in the covariate space is given a gamma prior with parameters $(1, 1)$. These choices reflect the fact that true model can be approximated by dividing the x-space into regions (with little overlap) of moderate range with a normal linear regression component within each region.

For the ADNI dataset, many studies have shown that hippocampal volume shrinks with age with greater decreases for diseased patients. A sensible range for the slope, as observed in the scatter plot of the data, is $[-1/2, 0]$, i.e. between a minimum of no shrinkage and a maximum decrease of .5 cm^3 in one year. We center therefore the prior for the slope on $-1/4$ with a variability of $1/50$ to cover this range. We chose to center the intercept on 8 cm^3, as it reflects the average hippocampal volume of cognitively normal males, with a variability of 4 to cover the range of intercepts that we could anticipate. Women tend to have lower brain volume than men, and a range of $[-2, 0]$ reflects the belief that hippocampal volume could be equal or up to 2 cm^3 less for women and men of the same age and disease status; the slope of the gender indicator is centered on -1 with a variability of $1/4$. We assume that when compared to cognitively normal subjects of the same age and gender, MCI patients may have a minimum of no decrease in hippocampal volume to a maximum decrease of 2 cm^3, while AD patients compared to cognitively normal subjects of the same age and gender have a minimum of no decrease and a maximum decrease of 4 cm^3; the slope of the AD indicator is centered on -1 with a variability of 1/4. Finally, we selected vague uniform prior to describe prior information about the regions where the components best apply in the discrete x-space. In the continous x-space, the conjugate normal gamma prior was centered on the average age of 72.5 with parameters $(1, 1)$ for the precision and the variability of locations relative to the range of best applicability was increased by a factor of 4; this was chosen to encourage fairly well separated x-regions of moderate range.

Rather than using a hyper-prior for the precision parameter of the Dirichlet process, we fix it to be $\zeta_{2,j} = 1$ in both examples. Due to the unidentifiability of the weights, such a practice corresponds to the standard solution of fixing the location of one of the variables for models with identifiability issues. The unidentifiability of the weights arises from the fact that they are given by $w_j(x) \propto w_j K(x|\psi_j)$. We resolve this in the usual way by fixing the locations of the (w_j) rather than assigning a hyper-prior to the precision parameter. Note that the model is fundamentally different from the usual DP mixture model, where the weights corresponding to each component in the mixture are simply the (w_j), without any multiplicative factors. Hence in the DP model the use of a hyper-prior for the precision parameter is known to be important, while in the present model that need is overcome by the effect of the kernels $K(x|\psi_j)$.

Additionally, we performed a sensitivity analysis with regards to the prior specification for the simulation study. Table 2 lists performance metrics for the proposed model under various modifications of the prior hyperparameters. Specifically, we explored decreasing and increasing the mass parameter of the Dirichlet process prior; decreasing

Estimated item		**Error**	**NW**	**NW**	**NW**	**NW**	**NW**	**NW**	**NW**	**NW**	**NW**
		measure		$\downarrow m$	$\uparrow m$	$\downarrow C^{-1}$	$\uparrow C^{-1}$	$\uparrow \alpha$	$\downarrow c^{-1}$	$\uparrow c^{-1}$	$\uparrow a_1$
Regression mean	**Fit**	$\hat{L_1}$	0.10	0.10	0.10	0.11	**0.09**	0.11	0.11	0.10	0.11
		$\hat{L_2}$	0.38	0.38	0.38	0.37	0.38	0.38	**0.36**	0.37	0.39
	Predictive	$\hat{L_1}$	0.79	0.79	0.79	0.76	0.81	0.80	**0.75**	0.76	0.80
		$\hat{L_2}$	1.46	1.46	1.45	**1.38**	1.57	1.46	**1.38**	1.43	1.48
Conditional densities	**Fit**	avg($\hat{L_1}$)	0.22	0.22	0.25	0.24	**0.20**	0.26	0.24	0.22	0.23
		max($\hat{L_1}$)	1.25	1.25	1.25	1.23	1.23	1.30	**1.12**	1.18	1.29
	Predictive	avg($\hat{L_1}$)	0.25	0.25	0.28	0.27	**0.22**	0.29	0.27	0.25	0.26
		max($\hat{L_1}$)	1.25	1.25	1.25	1.24	1.24	1.30	**1.14**	1.20	1.29

Table 2: Sensitivity Analysis: comparison of fitted and predictive errors in the regression mean and conditional density for varying choices of the prior parameters. The first two prior parameter modifications explore decreasing and increasing the mass parameter, m, of the Dirichlet process to 0.5 and 5. The following prior parameter changes explore decreasing and increasing the variability of the local regression coefficients (with $\downarrow C^{-1}$ corresponding to $C^{-1} = \text{diag}(1.5, 1/10)$ and $\uparrow C^{-1}$ corresponding to $C^{-1} = \text{diag}(10, 1)$); decreasing the variability of the precision around the local regression lines (with $\uparrow \alpha$ corresponding to $\alpha_1 = 2, \alpha_2 = 2$); decreasing and increasing the variability of the locations of the components in the x-space (with $\downarrow c^{-1}$ corresponding to $c^{-1} = 4$ and $\uparrow c^{-1}$ corresponding to $c^{-1} = 16$); and increasing the location and variability of the precision associated to the components in the x-space ($a_1 = 2$).

and increasing the variability of the local regression coefficients; decreasing the variability of the precision of the local linear regression models; decreasing and increasing the variability of the locations of the components in the x-space; and increasing the location and variability of precision of the components in the x-space. We find that the results are quite robust to these choices.

References

R.P. Adams, I. Murray, and D.J.C. MacKay. Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, 95:169–186, 2008.

Alzheimer's Disease Education & Referral Center ADEAR. Alzheimer's disease fact sheet. *NIH Publication*, 11-6423, 2011.

A. Caroli and G.B. Frisoni. The dynamics of Alzheimer's disease biomarkers in the Alzheimer's Disease Neuroimaging Initiative cohort. *Neurobiology of Aging*, 31:1263–1274, 2010.

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

D.G.T. Denison, C.C. Holmes, B.K. Mallick, and A.F.M Smith. *Bayesian Methods for Nonlinear Classification and Regression*. John Wiley & Sons, 2002.

I. Dimatteo, D.R. Genovese, and R.E. Kass. Bayesian curve fitting with free-knot splines. *Biometrika*, 88:1055–1071, 2001.

D. B. Dunson. *Nonparametric Bayes applications to biostatistics*, chapter 7, pages 223–273. Cambridge University Press, Cambridge, 2010.

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

G.B. Frisoni, N.C. Fox, C.R. Jr Jack, P. Scheltens, and P.M. Thompson. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6:67–77, 2010.

S.J. Godsill. On the relationship between Markov chain Monte Carlo Methods for model uncertainty. *Journal of Computational and Graphical Statistics*, 10(2):230–248, 2001.

P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

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

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

C.R. Jr Jack, D.S. Knopman, W.J. Jagust, L.M. Shaw, Aisen P.S., M.W. Weiner, R.C. Petersen, and J.Q. Trojanowski. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurology*, 9:119–128, 2010.

C.R. Jr Jack, P. Vemuri, H.J. Wiste, S.D. Weigand, T.G. Lesnick, V. Lowe, K. Kantarci, M.A. Bernstein, M.L. Senjem, J.L. Gunter, B.F. Boeve, J.Q. Trojanowski, L.M. Shaw, P.S. Aisen, M.W. Weiner, R.C. Petersen, and D.S. Knopman. Shapes of the trajectories of 5 major biomarkers of Alzheimer disease. *Archives of Neurology*, 69: 856–867, 2012.

M. Kalli, J.E. Griffin, and S.G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21:93–105, 2011.

A.Y. Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1):351–357, 1984.

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.

S.N. MacEachern. Dependent dirichlet processes. *Technical Report, Department of Statistics, Ohio State University*, 2000.

J. Møller, A.N. Pettitt, R. Reeves, and K.K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.

P. Müller and F. Quintana. *More nonparametric Bayesian models for biostatistics*, chapter 8, pages 274–291. Cambridge University Press, Cambridge, 2010.

I. Murray, Z. Ghahramani, and D.J.C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 359–366. AUAI Press, 2006.

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

O. Papaspiliopoulos and G.O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

A.N. Pettitt, N. Friel, and R. Reeves. Efficient calculation of the normalizing constant of the autologistic and related models on the cylinder and lattice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):235–246, 2003.

C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning.* the MIT Press, 2006.

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.

M.R. Sabuncu, R.S. Desikan, J. Sepulcre, B.T.T. Yeo, H. Liu, N.J. Schmansky, M. Reuter, M.W. Weiner, R.L. Buckner, R.A. Sperling, and B. Fischl. The dynamics of cortical and hippocampal atrophy in Alzheimer disease. *Archives of Neurology*, 68:1040–1048, 2011.