

Bayesian Density Regression

David B. Dunson^{1,2}, Natesh Pillai², and Ju-Hyun Park³

¹*Biostatistics Branch*

MD A3-03, National Institute of Environmental Health Sciences

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

E-mail: dunson1@niehs.nih.gov

²*Institute of Statistics and Decision Sciences*

Duke University

³*Department of Biostatistics*

University of North Carolina at Chapel Hill

Summary. This article considers Bayesian methods for density regression, allowing a random probability distribution to change flexibly with multiple predictors. The conditional response distribution is expressed as a nonparametric mixture of regression models, with the mixture distribution changing with predictors. A class of weighted mixture of Dirichlet process (WMDP) priors is proposed for the uncountable collection of mixture distributions. It is shown that this specification results in a generalized Pólya urn scheme, which incorporates weights dependent on the distance between subjects' predictor values. To allow local dependency in the mixture distributions, we propose a kernel-based weighting scheme. A Gibbs sampling algorithm is developed for posterior computation. The methods are illustrated using simulated data examples and an epidemiologic application.

Keywords: Conditional density function; Dirichlet process; Local smoothing; Mixture model; Non-parametric Bayes; Generalized Pólya urn.

1. Introduction

This article addresses the problem of density regression, investigating changes in the distribution of a random variable $Y \in \mathcal{Y}$ according to predictors $\mathbf{x} = (x_1, \dots, x_p)' \in \mathcal{X}$ using a Bayesian semiparametric approach. The challenge is that a parametric form for the density of Y is unknown, and there can be unanticipated changes in the shape of the density according to the predictor values \mathbf{x} . Thus, it is not appropriate to assume that the residual distribution in a mean or quantile regression model is constant over \mathcal{X} .

Motivated by the well known result that mixtures of a sufficiently-large number of normal densities can be used to accurately approximate any smooth density, we focus on the following mixture of regression models:

$$f(y_i | \mathbf{x}_i) = \int f(y_i | \mathbf{x}_i, \phi_i) G_{\mathbf{x}_i}(d\phi_i), \quad (1)$$

where $f(y_i | \mathbf{x}_i, \phi_i) = N(y_i; \mathbf{x}_i' \boldsymbol{\beta}_i, \sigma_i^2)$, with $\phi_i = (\boldsymbol{\beta}_i', \sigma_i^2)'$, and $G_{\mathbf{x}_i}$ is an unknown mixture distribution that can vary according to the location of $\mathbf{x}_i \in \mathcal{X}$. This model encompasses a wide variety of regression and mixture models as special cases, including normal linear regression, linear regression with the residual density modeled as a finite or infinite mixture of Gaussians, and the finite mixture of linear regressions:

$$f(y_i | \mathbf{x}_i) = \sum_{h=1}^k \pi_h(\mathbf{x}_i) N(y_i; \mathbf{x}_i' \boldsymbol{\beta}_h, \sigma^2). \quad (2)$$

Expression (2) is often referred to as a latent class regression model, in which case $\pi_h(\mathbf{x}_i)$ is characterized using a parametric regression model, commonly polytomous logistic regression. Hierarchical mixtures-of-experts models (Jordan and Jacobs, 1994; Viele and Tong, 2002) widely used in the machine learning literature instead use a flexible form for $\pi_h(\mathbf{x}_i)$.

Our focus is instead on treating $G_{\mathbf{x}_i}$ nonparametrically by using an infinite-dimensional Bayesian model. More formally, $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$ will be treated as an uncountable collection of random probability measures. A recent article by De Iorio et al. (2004) proposed a Bayesian nonparametric approach for modeling of dependence across random distributions $G_{\mathbf{x}}$ indexed by a vector $\mathbf{x} \in \mathcal{X}_D$ of categorical covariates. In particular, they defined a prior for the array of random measures

$\mathcal{G}_{\mathbf{x}} = \{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}_D\}$, which maintains a marginal Dirichlet process (DP) (Ferguson, 1973; 1974) structure for the distribution at each value of \mathbf{x} . This is accomplished using the dependent Dirichlet process (DDP) approach of MacEachern (1999; 2000; 2001), which relies on expressing the DP in stick-breaking form (Sethuraman, 1994):

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

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

Assuming a common set of weights, $\pi_{\mathbf{x},h} = \pi_h$ for all $\mathbf{x} \in \mathcal{X}_D$, MacEachern (1999; 2001) allows for dependency by defining a stochastic process for the atoms $\{\theta_{\mathbf{x},h}\}$. De Iorio et al. (2004) used the DDP to produce an ANOVA-type dependency structure, while Gelfand et al. (2005) applied the DDP to spatial modeling applications by using a Gaussian process for the atoms. Recently, Griffin and Steel (2006) proposed an order-based DDP, which allows the weights to vary with covariates. Müller, Quintana and Rosner (2004) instead induced a hierarchical dependency structure through mixtures of independent DP basis measures. Dunson (2006) extended this idea to a times series setting, defining a dynamic mixture of DPs. A simpler, but less flexible approach, is to induce dependency through a regression in the DP base measures (Cifarelli and Regazzini, 1978). Related approaches have been considered by Muliere and Petrone (1993), Mira and Petrone (1996), Giudici, Mezzetti, and Muliere (2003), and Griffin and Steel (2004). Müller, Erkanli and West (1996) instead used a DP mixture of normals for the joint distribution of y and \mathbf{x} , and then focused on the implied conditional density of y given \mathbf{x} in estimating the mean regression function. For a recent overview of Bayesian nonparametric inference, refer to Müller and Quintana (2004).

Section 2 discusses DP mixtures of linear regression models. Section 3 proposes a class of weighted mixture of DP (WMDP) priors, and considers properties. Section 4 proposes an efficient Markov chain Monte Carlo (MCMC) algorithm for posterior computation. Section 5 illustrates the methods through simulated data examples, Section 6 applies the approach to an epidemiologic

application, and Section 7 discusses the results. Proofs are included in Appendices.

2. Dirichlet Process Mixtures of Regression Models

Under the assumption that $G_{\mathbf{x}} \equiv G$, with $G \sim DP(\alpha G_0)$, expression (1) is in the form of a Dirichlet process mixture of normal linear regression models. Under the Sethuraman (1994) stick-breaking representation, this implies that

$$f(y_i | \mathbf{x}_i) = \sum_{h=1}^{\infty} \pi_h N(y_i; \mathbf{x}_i' \boldsymbol{\beta}_h, \sigma_h^2), \quad (3)$$

with $\boldsymbol{\pi} = (\pi_h, h = 1, \dots, \infty)$ an infinite sequence of stick-breaking weights and $\boldsymbol{\theta} = (\boldsymbol{\theta}_h, h = 1, \dots, \infty)$, with $\boldsymbol{\theta}_h = (\boldsymbol{\beta}_h', \sigma_h^2)$, atoms sampled independently from G_0 . For example, G_0 could be chosen as the normal-inverse gamma probability measure, or one could let $\sigma_h^2 = \sigma^2$ and then let G_0 correspond to a multivariate normal.

The DPM described in (3) seems very flexible in incorporating an infinite number of normal linear regression components. However, by assuming the weights $\boldsymbol{\pi}$ are constant, one greatly restricts the conditional density. For example, note that the mean regression structure is linear:

$$E(y_i | \mathbf{x}_i) = \sum_{h=1}^{\infty} \pi_h \mathbf{x}_i' \boldsymbol{\beta}_h = \sum_{h=1}^{\infty} \pi_h \sum_{j=1}^p x_{ij} \beta_{hj} = \sum_{j=1}^p x_{ij} \sum_{h=1}^{\infty} \pi_h \beta_{hj} = \mathbf{x}_i' \bar{\boldsymbol{\beta}},$$

where $\bar{\boldsymbol{\beta}} = \sum_h \pi_h \boldsymbol{\beta}_h$. Potentially, we could extend model (3) to allow greater flexibility through a predictor-dependent stick-breaking process, as described by Griffin and Steel (2006) and Duan et al. (2005). However, such approaches are quite demanding computationally and it would be appealing to have an approach that can be implemented as easily as the DPM model.

One of the primary reasons for the great success and rapidly expanding use of DPMS is the development of simple and efficient Markov chain Monte Carlo (MCMC) algorithms for posterior computation (MacEachern, 1994; Escobar and West, 1998; MacEachern and Müller, 1998, Ishwaran and James, 2001; among others). The most popular algorithms rely on the Blackwell and MacQueen (1973) Pólya urn scheme, which integrates out the infinite-dimensional G to obtain the conditional prior of ϕ_i given $\boldsymbol{\phi}^{(i)} = (\phi_1, \dots, \phi_{i-1}, \phi_{i+1}, \dots, \phi_n)'$:

$$(\phi_i | \boldsymbol{\phi}^{(i)}, \alpha) \sim \left(\frac{\alpha}{\alpha + n - 1} \right) G_0 + \left(\frac{1}{\alpha + n - 1} \right) \sum_{j \neq i} \delta_{\phi_j}, \quad (4)$$

which generates new values from $\phi_i \sim G_0$ with probability $\alpha/(\alpha + n - 1)$ and otherwise sets ϕ_i equal to an element of $\boldsymbol{\phi}^{(i)}$ chosen by sampling from a discrete uniform.

The tendency of the DP to cluster subjects into groups having identical coefficients is quite apparent from this structure. In particular, the n subjects will be allocated to $k \leq n$ unique values, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$, with the induced prior on k stochastically increasing with α and n (Antoniak, 1974). Hence, although expression (3) is an infinite mixture, we obtain a finite number k of clusters upon integrating out the random weights $\boldsymbol{\pi}$ and random atoms $\boldsymbol{\theta}$. For this reason, DP methods are commonly used to allow uncertainty in the number of mixture components and for clustering. However, the implicit assumption is that the subjects are exchangeable, which does not hold if a subject's predictor values are informative about the clustering.

3. Weighted Mixtures of Dirichlet Process Priors

3.1 Proposed formulation

Let $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ denote the uncountable collection of random probability measures $G_{\mathbf{x}}$ on $(\mathcal{Y}, \mathcal{B})$, with \mathcal{B} the Borel σ -algebra of subsets of \mathcal{Y} . Note that $G_{\mathbf{x}}$ is indexed by values \mathbf{x} in the covariate space \mathcal{X} , with the focus being on continuous predictors. Our goal is to specify a prior probability model for $\mathcal{G}_{\mathcal{X}}$. Although it is conceptually appealing to specify a prior that does not depend on the data, many authors have argued in regression settings to define priors dependent on the predictor values in the sample. One notable example is Zellner's (1986) g-prior, which has had wide use in linear regression and model selection applications.

To define a general class of priors for $\mathcal{G}_{\mathcal{X}}$, we propose the following formulation:

$$G_{\mathbf{x}} = \sum_{j=1}^n b_j(\mathbf{x}) G_{\mathbf{x}_j}^*, \quad G_{\mathbf{x}_j}^* \stackrel{ind}{\sim} DP(\alpha G_0), \quad \text{for } j = 1, \dots, n, \forall \mathbf{x} \in \mathcal{X}, \quad (5)$$

where $\mathbf{b}(\mathbf{x}) = [b_1(\mathbf{x}), \dots, b_n(\mathbf{x})]'$ is a weight function mapping from $\mathcal{X} \rightarrow \mathcal{P}_n$, with \mathcal{P}_n the n -dimensional probability simplex, so that $b_j(\mathbf{x}) \geq 0$, $j = 1, \dots, n$, and $\mathbf{b}(\mathbf{x})' \mathbf{1}_n = 1$, for all $\mathbf{x} \in \mathcal{X}$. The collection $\mathcal{G}_{\mathbf{X}}^* = \{G_{\mathbf{x}_i}^*, i = 1, \dots, n\}$ consists of independent samples from the Dirichlet process with common base measure αG_0 . Hence, the prior for $\mathcal{G}_{\mathcal{X}}$ is defined by placing a DP-distributed random basis measure at each of the sample predictor values, and then mixing across these basis

measures to construct an uncountable collection of random probability measures for all possible predictor values $\mathbf{x} \in \mathcal{X}$. We refer to this prior as a weighted mixture of DPs (WMDP).

The weight function, $\mathbf{b}(\mathbf{x})$, can potentially be defined explicitly as follows:

$$b_j(\mathbf{x}) = \frac{\gamma_j K(\mathbf{x}, \mathbf{x}_j)}{\sum_{l=1}^n \gamma_l K(\mathbf{x}, \mathbf{x}_l)}, \quad j = 1, \dots, n, \forall \mathbf{x} \in \mathcal{X}, \quad (6)$$

where $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_n)'$ represent weights on the different basis locations, and $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathfrak{R}^+$ is a kernel function, such as $K(\mathbf{x}, \mathbf{x}') = \exp(-\psi \|\mathbf{x} - \mathbf{x}'\|^2)$. For this choice, basis distributions located close to \mathbf{x} are automatically assigned relatively high weight in the prior for $G_{\mathbf{x}}$, particularly if these locations have high γ values. Because the random basis distributions are located at the sample predictor values, one automatically allows more flexibility in data-rich regions of \mathcal{X} .

The weights, $\boldsymbol{\gamma}$, are an important component of specification (6), allowing additional flexibility. In particular, we can favor a sparser representation by allowing the weights for some or most of the locations to be close to zero. An extreme case of this approach corresponds to $\gamma_j / \sum_l \gamma_l \rightarrow 1$, which results in $b_j(\mathbf{x}) = 1$ and $b_{j'}(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}$ and $j' \neq j$, assuming $K(\mathbf{x}, \mathbf{x}') \neq 0$, for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. This case is equivalent to letting $G_{\mathbf{x}} = G \sim DP(\alpha G_0)$, so that we obtain the DP mixture model described in Section 2. In practice, $\boldsymbol{\gamma}$ will be considered unknown by choosing a prior of the form, $\gamma_j \stackrel{iid}{\sim} g$, with the choice of g discussed in Section 3.4.

Dependency in the random measures contained in $\mathcal{G}_{\mathcal{X}}$ arises through shared dependency on the DP-distributed random basis measures $G_{\mathbf{x}_j}^*$. Note that $\phi \sim G_{\mathbf{x}}$ can be equivalently expressed in hierarchical form as:

$$\begin{aligned} (\phi | Z = j, \mathbf{x}) &\sim G_{\mathbf{x}_j}^*, \\ (Z | \mathbf{x}) &\sim \text{Multinomial}(\{1, \dots, n\}; \mathbf{b}(\mathbf{x})), \\ G_{\mathbf{x}_j}^* &\sim DP(\alpha G_0), j = 1, \dots, n, \end{aligned} \quad (7)$$

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

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

$$\mathbb{E}\{G_{\mathbf{x}}(B)\} = \sum_{j=1}^n \Pr(Z = j) \mathbb{E}\{G_{\mathbf{x}_j}^*(B)\} = \sum_{j=1}^n b_j(\mathbf{x}) G_0(B) = G_0(B), \quad (8)$$

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

with $B \in \mathcal{B}$. The dependency between $G_{\mathbf{x}}$ and $G_{\mathbf{x}'}$ can be characterized using Theorem 1.

Theorem 1. For any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, including values not represented in \mathbf{X} , $G_{\mathbf{x}}$ and $G_{\mathbf{x}'}$ are dependent random probability measures, with

$$\text{Cor}\{G_{\mathbf{x}}(B), G_{\mathbf{x}'}(B)\} = \rho(\mathbf{x}, \mathbf{x}') = \frac{\langle \mathbf{b}(\mathbf{x}), \mathbf{b}(\mathbf{x}') \rangle}{\|\mathbf{b}(\mathbf{x})\| \cdot \|\mathbf{b}(\mathbf{x}')\|} \quad \text{for any Borel set } B \in \mathcal{B},$$

with \mathcal{B} the Borel σ -algebra of subsets of \mathcal{Y} .

The proof is in Appendix A. Due to the lack of dependency on B , this expression is particularly useful. Note that $\rho : \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$ is a bounded kernel expressed as the normed inner product of the weight functions, $\mathbf{b}(\mathbf{x})$ and $\mathbf{b}(\mathbf{x}')$.

3.2 Conditional formulation

It is interesting to consider properties of the induced prior for $\mathcal{G}_{\mathbf{X}} = \{G_{\mathbf{x}_i}, i = 1, \dots, n\}$, the finite collection of random measures within $\mathcal{G}_{\mathcal{X}}$ at the sample predictor values. Starting with (5), we obtain

$$G_{\mathbf{x}_i} = \sum_{j=1}^n b_{ij} G_{\mathbf{x}_j}^*, \quad G_{\mathbf{x}_j}^* \stackrel{\text{ind}}{\sim} DP(\alpha G_0), \quad \text{for } j = 1, \dots, n, \quad (10)$$

where $\mathbf{b}_i = (b_{i1}, \dots, b_{in})' = \mathbf{b}(\mathbf{x}_i)$, with $b_{ij} \geq 0$ and $\sum_{j=1}^n b_{ij} = 1$.

Interestingly, Theorem 2 demonstrates that we can also obtain (10) by starting with a very different conditional mixture structure:

$$(G_{\mathbf{x}_i}(B) | G_{\mathbf{x}_j}, j \sim i) \sim a_{ii} G_{\mathbf{x}_i}^*(B) + \sum_{j \sim i} a_{ij} G_{\mathbf{x}_j}(B), \quad G_{\mathbf{x}_i}^* \sim DP(\alpha G_0), \quad \text{for all } B \in \mathcal{B}, \quad (11)$$

where $j \sim i$ indexes subjects $j \in \mathcal{N}_i \subset \{1, \dots, n\}/i$ with $\mathcal{N}_i = \{j : d(\mathbf{x}_i, \mathbf{x}_j) < \epsilon, j \neq i\}$, where $d(\mathbf{x}_i, \mathbf{x}_j)$ is a known measure of distance between \mathbf{x}_i and \mathbf{x}_j and ϵ is a positive constant. This expression considers the random measure, $G_{\mathbf{x}_i}$, to be equal to a weighted mixture of neighboring

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

Theorem 2. Let \mathbf{A} denote the $n \times n$ matrix with elements $\{a_{ij}\}_{i,j=1}^n$ satisfying $0 \leq a_{ij} < 1$ and $\mathbf{a}'_i \mathbf{1}_n = 1$ for all row vectors \mathbf{a}'_i . Suppose we have

$$\begin{aligned} G_1 &= a_{11}G_1^* + a_{12}G_2 + \dots + a_{1n}G_n \\ G_2 &= a_{21}G_1 + a_{22}G_2^* + \dots + a_{2n}G_n \\ &\vdots = \vdots \quad \quad \quad \ddots \quad \quad \quad \vdots \\ G_n &= a_{n1}G_1 + a_{n2}G_2 + \dots + a_{nn}G_n^* \end{aligned}$$

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

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

The proof is in Appendix B. Note that the matrix $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_n)' = h(\mathbf{A}) = (\mathbf{I}_n - \mathbf{C})^{-1}\mathbf{D}$, with $\mathbf{A} = \mathbf{C} + \mathbf{D}$, $c_{ij} = a_{ij}$ for $i \neq j$, $c_{ii} = 0$ for $i = 1, \dots, n$, and $\mathbf{D} = \text{diag}(a_{11}, \dots, a_{nn})$. Hence, for any model specified as in (11), there is a corresponding model specified as in (10), with the probability weights \mathbf{B} calculated from \mathbf{A} using the simple deterministic function $h(\cdot)$.

The fact that (11) implies (10) is interesting, because (11) has a intuitive conditional dependency structure, which is in a similar form to that commonly used to define spatial dependency in parametric models. Hence, (11) may form the basis for extensions to spatially-dependent priors and other interesting cases. In addition, formulation (11) provides some insight into the mechanism of borrowing of information through the conditional moments:

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

For example, as $a_{ii} \rightarrow 1$, we have $E\{G_{\mathbf{x}_i}(B) | G_{\mathbf{x}_j}, j \sim i\} = G_0(B)$, so the distributions at adjacent locations are not taken into account, and one relies on the base parametric model.

Note that, although (10) does not imply (11) without further restrictions, we can show that:

Theorem 3. An uncountably infinite class of priors for $\mathcal{G}_{\mathcal{X}}$ of the form shown in (5) can be obtained, for any prior defined as in (11).

As proof, it is necessary to show that there are infinitely many different choices of $\mathbf{b}(\mathbf{x})$ subject to $b_j(\mathbf{x}) \geq 0$, $j = 1, \dots, n$, $\mathbf{b}(\mathbf{x})' \mathbf{1}_n = 1$, for all $\mathbf{x} \in \mathcal{X}$, and $\mathbf{b}(\mathbf{x}_i) = \mathbf{b}_i = (b_{i1}, \dots, b_{in})'$, with \mathbf{b}_i constants derived from \mathbf{a}_i . Consider the case in which $p = 1$. Let $c_l(x) = \sum_{h=1}^l b_h(x)$, for $l = 1, \dots, n$. Then, $c_l(x_i) = c_{il} = \sum_{h=1}^l b_{ih}$, for $i = 1, \dots, n$, are fixed points along the function c_l , with $c_n(x) = 1$ for all $x \in \mathcal{X}$. Any collection of continuous functions $\{c_l(x), l = 1, \dots, n-1\}$ extrapolating through these fixed points subject to the restriction that $c_l(x) \leq c_{l'}(x)$ for all $l < l'$ will result in functions $\mathbf{b}(x)$ satisfying the conditions, with linear extrapolation providing one example among uncountably infinitely many possibilities. This argument is trivially generalized to $p \geq 2$.

3.3 Generalized Pólya urn scheme

As discussed in Section 2, one of the most useful properties of the DP is the Pólya urn scheme shown in expression (4). In this subsection, we show that marginalizing across the WMDP prior shown in expression (5) results in a generalization of the Pólya urn scheme, which incorporates weights that depend on the distance between subjects' predictor values. In this manner, we relax the exchangeability assumption, facilitating local learning and predictor-dependent clustering.

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

$$\begin{aligned} (\phi_i | \mathbf{Z}, \phi^{(i)}, \mathbf{X}, \alpha) &\sim \left(\frac{\alpha}{\alpha + \sum_{j \neq i} \mathbf{1}(Z_j = Z_i)} \right) G_0 \\ &+ \left(\frac{1}{\alpha + \sum_{j \neq i} \mathbf{1}(Z_j = Z_i)} \right) \sum_{j \neq i} \mathbf{1}(Z_j = Z_i) \delta_{\phi_j}. \end{aligned} \quad (12)$$

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

$$\begin{aligned} \Pr(\mathbf{M}_i = \mathbf{m}_i) &= \sum_{j=1}^n \Pr(Z_i = j) \prod_{h \neq i} \Pr(Z_h = j)^{m_{ih}} \{1 - \Pr(Z_h = j)\}^{1-m_{ih}} \\ &= \sum_{j=1}^n b_j(\mathbf{x}_i) \prod_{h \neq i} b_j(\mathbf{x}_h)^{m_{ih}} \{1 - b_j(\mathbf{x}_h)\}^{1-m_{ih}} = \sum_{j=1}^n b_{ij} \prod_{h \neq i} b_{hj}^{m_{ih}} (1 - b_{hj})^{1-m_{ih}} \end{aligned} \quad (13)$$

Marginalizing across the distribution for \mathbf{M}_i , we obtain the following generalization of the Blackwell and MacQueen (1973) Pólya urn scheme of expression (4):

$$\begin{aligned} (\phi_i | \phi^{(i)}, \mathbf{X}, \alpha, \mathbf{B}) &\sim \sum_{h \neq i} \sum_{m_{ih}=0}^1 \left\{ \sum_{j=1}^n b_{ij} \prod_{l \neq i} b_{lj}^{m_{il}} (1 - b_{lj})^{1-m_{il}} \right\} \\ &\times \left\{ \left(\frac{\alpha}{\alpha + \sum_{l \neq i} m_{il}} \right) G_0 + \left(\frac{1}{\alpha + \sum_{l \neq i} m_{il}} \right) \sum_{l \neq i} m_{il} \delta_{\phi_l} \right\}. \end{aligned} \quad (14)$$

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

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

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

$$\mathbf{\Gamma}_0 = \left(1, \frac{\alpha}{\alpha+1}, \frac{\alpha}{\alpha+2}, \dots, \frac{\alpha}{\alpha+n-1} \right)', \quad \mathbf{\Gamma}_1 = \left(\frac{1}{\alpha+1}, \frac{1}{\alpha+2}, \dots, \frac{1}{\alpha+n-1} \right)',$$

let $\mathbf{p}_{i0} = \mathbf{p}_0(\mathbf{x}_i)$ denote the $n \times 1$ vector of probabilities corresponding to $\Pr(M_{i+} = m | \mathbf{x}_i)$, for $m = 0, \dots, n-1$ with $M_{i+} = \sum_{j \neq i} M_{ij}$, and let $\mathbf{p}_{ij} = \mathbf{p}_j(\mathbf{x}_i)$ denote the $(n-1) \times 1$ vector of probabilities corresponding to $\Pr(M_{ij} = 1, M_{i+} = m | \mathbf{x}_i)$, for $m = 1, \dots, n-1$. For example, in the

special case considered in the above table, letting $p_{000}, p_{100}, p_{010}, p_{001}, p_{110}, p_{101}, p_{011}, p_{111}$ denote the probabilities in column 4, we have $\mathbf{p}_{i0} = (p_{000}, p_{100} + p_{010} + p_{001}, p_{110} + p_{011} + p_{101}, p_{111})'$, $\mathbf{p}_{i1} = (p_{100}, p_{110} + p_{101}, p_{111})'$, $\mathbf{p}_{i2} = (p_{010}, p_{110} + p_{011}, p_{111})'$, and $\mathbf{p}_{i3} = (p_{001}, p_{011} + p_{101}, p_{111})'$. In general, using this notation, we can express (14) as

$$(\phi_i | \phi^{(i)}, \mathbf{X}, \alpha, \mathbf{B}) = \mathbf{p}'_{i0} \Gamma_0 G_0 + \sum_{j \neq i} \mathbf{p}'_{ij} \Gamma_1 \delta_{\phi_j} = \mathbf{p}_0(\mathbf{x}_i)' \Gamma_0 G_0 + \sum_{j \neq i} \mathbf{p}_j(\mathbf{x}_i)' \Gamma_1 \delta_{\phi_j}, \quad (15)$$

where $\mathbf{p}'_{i0} \mathbf{1}_n = 1$ and $\mathbf{p}'_{ij} \mathbf{1}_{n-1} \leq 1$. This expression is in the form of a weighted average of Blackwell and MacQueen (1973) Pólya urn distributions.

To further simplify this expression, we rely on Theorem 4 (proof in Appendix C).

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

$$\mathbf{p}'_{i0} \Gamma_0 = \left(\frac{\alpha}{\alpha + \mathbf{w}'_i \mathbf{1}_{n-1}} \right) \quad \text{and} \quad \mathbf{p}'_{ij} \Gamma_1 = \left(\frac{w_{ij}}{\alpha + \mathbf{w}'_i \mathbf{1}_{n-1}} \right), \quad \forall j \neq i,$$

for $i = 1, \dots, n$, where $\mathbf{p}_{i0}, \mathbf{p}_{ij}, j \neq i$, are calculated from \mathbf{B} as described above. In particular, we have $w_{ij} = \alpha \mathbf{p}'_{ij} \Gamma_1 / \mathbf{p}'_{i0} \Gamma_0$, for all i, j .

Hence, from Theorem 4, expression (15) is equivalent to

$$(\phi_i | \phi^{(i)}, \mathbf{X}, \alpha, \mathbf{B}) = \left(\frac{\alpha}{\alpha + w_i} \right) G_0 + \sum_{j \neq i} \left(\frac{w_{ij}}{\alpha + w_i} \right) \delta_{\phi_j}, \quad (16)$$

where $\mathbf{W} = \{w_{ij}, i = 1, \dots, n, j \neq i\}$ is a set of weights between 0 and 1 that depend on \mathbf{B} and α deterministically, and $w_i = \sum_{j \neq i} w_{ij} \leq n$.

This simple form is intuitively-appealing. Instead of assuming the subjects are exchangeable, weights $\{w_{ij}\}$ are incorporated which depend on the subjects' relative predictor values. The Pólya urn conditional distribution in expression (4) is obtained as a special case by letting $\mathbf{b}(\mathbf{x}) = (1, 0, \dots, 0)'$, for all $\mathbf{x} \in \mathcal{X}$, which implies $w_{ij} = 1$ for all i, j . In general, the weight for the j th subject ($j \neq i$) in the conditional distribution for ϕ_i will depend on the relative values of \mathbf{p}_{ij} and \mathbf{p}_{i0} . In the limit as $p_{ijm} = \Pr(M_{i+} = m, M_{ij} = 1) \rightarrow p_{i0,m+1} = \Pr(M_{i+} = m)$, for

$m = 1, \dots, n - 1$, which implies $\Pr(M_{ij} = 1) \rightarrow 1$, we have $w_{ij} \rightarrow 1$, while in the limit as $p_{ijm} \rightarrow 0$, for $m = 1, \dots, n - 1$, $w_{ij} \rightarrow 0$. Subjects that have a high probability of being assigned to the same mixture component as subject i will be given high weight. For previous work on exchangeable generalizations of the DP-based Pólya urn scheme, refer to Ishwaran and James (2003).

3.4 Hyperprior for Weights on Locations

The properties of the prior for $\mathcal{G}_{\mathcal{X}}$, defined in expressions (5) and (6), depend on the kernel function $K(\cdot)$ and the choice of prior for the elements of γ . Focusing first on the kernel function, we let $K(\mathbf{x}, \mathbf{x}') = \exp(-\psi\|\mathbf{x} - \mathbf{x}'\|)$, with ψ an unknown smoothing parameter, which is assigned a log-normal hyperprior, $\psi \sim \log\text{-N}(\mu_\psi, \sigma_\psi^2)$. In the limit as $\psi \rightarrow 0$, $K(\mathbf{x}, \mathbf{x}') = 1$ for all \mathbf{x}, \mathbf{x}' , while as ψ increases the kernel decreases increasingly rapidly as the distance between \mathbf{x} and \mathbf{x}' increases. By choosing a hyperprior for ψ , we allow the data to inform about its value.

The weights, γ , have a more subtle role. Assume that ψ is not large for sake of discussion. If a single location, \mathbf{x}_j , is assigned very high weight, γ_j , relative to other locations, then we say that the location dominates. If subjects are all assigned to the same location, \mathbf{x}_j , then $\phi_i \sim G_{\mathbf{x}_j}^* \sim DP(\alpha G_0)$, and one is simply fitting a DP mixture model. It is only through the allocation of subjects to multiple basis locations that one allows the mixture distribution to change with predictors. However, if a single subject is assigned to each location, then $\phi_i \sim G_0$, for $i = 1, \dots, n$, and one is effectively fitting the base parametric mixture model, which is even more restrictive than the single-location DP mixture model. Hence, assigning uniform weights, say $\gamma_j = 1/n$, for $j = 1, \dots, n$, is also a poor choice, as the number of subjects at any location will tend to be small.

Intuitively, a good choice of prior for γ_j would favor a few dominate locations, with this number increasing slowly with sample size. With this goal in mind, we propose the following prior:

$$\gamma_j \sim \text{gamma}(\kappa, n \times \kappa), \quad \kappa \sim \log\text{-N}(\mu_\kappa, \sigma_\kappa^2), \quad (17)$$

with μ_κ and σ_κ^2 chosen to place high probability on small, but not minute values. In particular, we recommend letting $\mu_\kappa = -2.5$ and $\sigma_\kappa^2 = 1$, motivated by results in simulating from the prior. By choosing a hyperprior for κ , we allow the data to inform about the number of dominate locations.

Because choosing κ too small or too large results in a poor fit, the data are highly informative about κ and the choice of μ_κ and σ_κ^2 will tend to have a minimal impact on the posterior, particularly when σ_κ^{-2} is small relative to the sample size n . The updating and tendency of the posterior to automatically favor a few dominate locations is illustrated in Sections 5 and 6.

4. Posterior Computation

4.1 MCMC algorithm

We propose a data augmentation MCMC algorithm for posterior computation, relying on a generalization of the accelerated Pólya urn Gibbs sampler (MacEachern 1994; West et al. 1994; Escobar and West, 1998). Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$ denote the $k \leq n$ unique values of $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)'$, and let $\mathbf{S} = (\mathcal{S}_1, \dots, \mathcal{S}_n)'$ be a vector of indicators denoting the global configuration of subjects to unique values $\boldsymbol{\theta}$, with $\mathcal{S}_i = h$ if $\phi_i = \theta_h$ indexing the location of the i th subject within the $\boldsymbol{\theta}$ vector. In addition, let $\mathbf{C} = (\mathcal{C}_1, \dots, \mathcal{C}_k)'$, with $\mathcal{C}_h = j$ denoting that θ_h is an atom from the j th basis distribution, $G_{\mathbf{x}_j}^*$. Hence, $\mathcal{C}_{\mathcal{S}_i} = Z_i = j$ denotes that subject i was drawn from j th basis distribution.

Excluding the i th subject, $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta} \setminus \phi_i$ denotes the $k^{(i)}$ unique values of $\boldsymbol{\phi}^{(i)}$, $\mathbf{S}^{(i)}$ denotes the configuration of subjects $\{1, \dots, n\} \setminus i$ to these values, and $\mathbf{C}^{(i)} = (\mathcal{C}_1^{(i)}, \dots, \mathcal{C}_{k^{(i)}}^{(i)})'$ indexes the DP component numbers for the elements of $\boldsymbol{\theta}^{(i)}$. Conditioning on $\mathbf{Z}^{(i)}$ but marginalizing over Z_i , we obtain the following conditional prior for ϕ_i :

$$(\phi_i | \mathbf{Z}^{(i)}, \boldsymbol{\phi}^{(i)}, \mathbf{X}, \alpha) \sim \sum_{j=1}^n \left(\frac{\alpha b_{ij}}{\alpha + \sum_{l \neq i} 1(Z_l = j)} \right) G_0 + \sum_{m \neq i} \left(\frac{b_{ij} 1(Z_m = j)}{\alpha + \sum_{l \neq i} 1(Z_l = j)} \right) \delta_{\phi_m}.$$

Grouping together the subjects in the same cluster, we obtain the expression:

$$(\phi_i | \mathbf{S}^{(i)}, \mathbf{C}^{(i)}, \boldsymbol{\theta}^{(i)}, \mathbf{X}, \alpha) \sim w_{i0} G_0 + \sum_{h=1}^{k^{(i)}} w_{ih} \delta_{\theta_h^{(i)}}, \quad (18)$$

where the probability weights on the different components are defined as:

$$w_{i0} = \left\{ \sum_{j=1}^n \frac{\alpha b_{ij}}{\alpha + \sum_{l \neq i} 1(\mathcal{C}_{S_l^{(i)}} = j)} \right\}, \quad w_{ih} = \frac{b_{i, \mathcal{C}_h^{(i)}} \sum_{m \neq i} 1(\mathcal{S}_m^{(i)} = h)}{\alpha + \sum_{l \neq i} 1(\mathcal{C}_{S_l^{(i)}} = \mathcal{C}_h)}, \quad h = 1, \dots, k^{(i)}.$$

Updating prior with the likelihood for the data \mathbf{y} , we obtain the conditional posterior:

$$(\phi_i | \mathbf{S}^{(i)}, \mathbf{C}^{(i)}, \boldsymbol{\theta}^{(i)}, \mathbf{X}, \alpha) \sim q_{i0} G_{i,0} + \sum_{h=1}^{k^{(i)}} q_{ih} \delta_{\theta_h^{(i)}}, \quad (19)$$

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

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

$q_{i0} = c w_{i0} h_i(y_i|\mathbf{x}_i)$, $q_{ih} = c w_{ih} f(y_i|\mathbf{x}_i, \theta_h)$, and c is a normalizing constant.

Our MCMC algorithm then alternates between the following steps:

1. Update \mathcal{S}_i , for $i = 1, \dots, n$, by sampling from the multinomial conditional posterior having $\Pr(\mathcal{S}_i = h) = q_{ih}$, for $h = 1, \dots, k^{(i)}$. When $\mathcal{S}_i = 0$, sample $\phi_i \sim G_{i,0}$ and $\mathcal{C}_{\mathcal{S}_i} \sim \text{multinomial}(\{1, \dots, n\}, \mathbf{b}_i)$.
2. Update the θ_h , for $h = 1, \dots, k$, by sampling from the conditional posterior distribution

$$(\theta_h | \mathbf{S}, \mathbf{C}, \boldsymbol{\theta}^{(h)}, k, \mathbf{y}, \mathbf{X}) \propto \left\{ \prod_{i:\mathcal{S}_i=h} f(y_i|\mathbf{x}_i, \theta_h) \right\} G_0(\theta_h), \quad (20)$$

which follows a simple form when G_0 is chosen to be conjugate.

3. Update \mathcal{C}_h , for $h = 1, \dots, k$, by sampling from the multinomial conditional with

$$\Pr(\mathcal{C}_h = j | \mathbf{C}^{(h)}, \mathbf{S}, k, \boldsymbol{\theta}, \mathbf{y}, \mathbf{X}) = \frac{\prod_{i:\mathcal{S}_i=h} b_{ij}}{\sum_{l=1}^n \prod_{i:\mathcal{S}_i=h} b_{il}}, \quad j = 1, \dots, n.$$

4. Update γ_j , for $j = 1, \dots, n$, using a data augmentation approach motivated by Dunson and Stanford (2005) and Holmes and Held (2006). Letting $K_{ij} = \exp(-\psi \|\mathbf{x}_i - \mathbf{x}_j\|^2)$ and $K_{ij}^* = K_{ij}/(\sum_{l \neq j} \gamma_l K_{il})$, the conditional likelihood for γ_j is

$$L(\gamma_j) = \prod_{i=1}^n \left(\frac{\gamma_j K_{ij}^*}{1 + \gamma_j K_{ij}^*} \right)^{1(\mathcal{C}_{\mathcal{S}_i}=j)} \left(\frac{1}{1 + \gamma_j K_{ij}^*} \right)^{1(\mathcal{C}_{\mathcal{S}_i} \neq j)}.$$

This likelihood can be obtained using $1(\mathcal{C}_{\mathcal{S}_i} = j) = 1(Z_{ij}^* > 0)$, with $Z_{ij}^* \sim \text{Poisson}(\gamma_j \xi_{ij} K_{ij}^*)$ and $\xi_{ij} \sim \exp(1)$. Updating $\{Z_{ij}^*, \xi_{ij}\}$ and $\{\gamma_j\}$ in Gibbs steps:

- (a) Let $Z_{ij}^* = 0$ if $\mathcal{C}_{\mathcal{S}_i} \neq j$ and otherwise $Z_{ij}^* \sim \text{Poisson}(\gamma_j \xi_{ij} K_{ij}^*) 1(Z_{ij}^* > 0)$, for all i, j .
- (b) $\xi_{ij} \sim \text{gamma}(1 + Z_{ij}^*, 1 + \gamma_j K_{ij}^*)$, for all i, j .
- (c) Letting $\text{gamma}(a_\gamma, b_\gamma)$ denote the prior for γ_j ,

$$\gamma_j \sim \text{gamma}\left(a_\gamma + \sum_{i=1}^n Z_{ij}^*, b_\gamma + \sum_{i=1}^n \xi_{ij} K_{ij}^*\right).$$

The algorithm is related to Pólya urn Gibbs samplers commonly used for posterior computation in DP mixture models, with the exceptions that (i) n DP components are concatenated so that one simultaneously updates the assignment to DP components and the cluster allocation within a component in step (1); (ii) although the assignment to components \mathbf{C} is updated in step (1) we include an acceleration step (3) to reshuffle the assignment and improve mixing; and (iii) to update the n -dimensional weight parameter $\boldsymbol{\gamma}$, we propose an efficient data augmentation scheme.

These steps can be incorporated within an MCMC algorithm that also has steps for updating additional unknowns within a larger hierarchical model. In our implementation used in Sections 5 and 6, we also updated ψ and κ using Metropolis random walk steps, and allowed for unknown parameters $\boldsymbol{\nu}$ within G_0 using standard Gibbs steps derived from the full conditional:

$$(\boldsymbol{\nu} \mid \boldsymbol{\phi}, \mathbf{y}, \mathbf{X}) \propto \pi(\boldsymbol{\nu}) \left\{ \prod_{h=1}^k G_0(\theta_h; \boldsymbol{\nu}) \right\}. \quad (21)$$

Extension to allow unknown α by generalizing the approach of West (1992) is also straightforward.

4.2 Mixtures of normal linear models

It is interesting to consider the simple special case in which

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

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

To complete a Bayesian specification of the model, the error precision $\tau = \sigma^{-2}$ is assigned a gamma prior, $\pi(\tau) = \mathcal{G}(\tau; a_\tau, b_\tau)$, and we choose a multivariate normal for the base parametric mixture distribution, $G_0(\boldsymbol{\beta}_i; \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta) = N_p(\boldsymbol{\beta}_i; \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta)$. For additional flexibility, we choose hyperprior distributions for $\boldsymbol{\nu} = \{\boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta\}$, the parameters characterizing G_0 . In particular, let

$\pi(\boldsymbol{\nu}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{\Sigma}_\beta)$, $\pi(\boldsymbol{\beta}) = N_p(\boldsymbol{\beta}; \boldsymbol{\beta}_0, V_{\beta_0})$ and $\pi(\boldsymbol{\Sigma}_\beta^{-1}) = \mathcal{W}(\boldsymbol{\Sigma}_\beta^{-1}; (\nu_0 \boldsymbol{\Sigma}_0)^{-1}, \nu_0)$, the Wishart density with degrees of freedom ν_0 and expectation $\boldsymbol{\Sigma}_0^{-1}$.

The conditional probabilities in expression (19) can be calculated plugging in:

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

for $h_i(y_i | \mathbf{x}_i)$, where $\hat{V}_{\beta_i} = (\boldsymbol{\Sigma}_\beta^{-1} + \tau \mathbf{x}_i \mathbf{x}_i')^{-1}$ and $\hat{\boldsymbol{\beta}}_i = \hat{V}_{\beta_i} (\boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\beta} + \tau y_i \mathbf{x}_i)$. In addition, letting $\theta_h = \boldsymbol{\beta}_h$, the value of $\boldsymbol{\beta}_i$ for subjects in the h th cluster, expression (20) simplifies to

$$(\boldsymbol{\beta}_h | \boldsymbol{\beta}^{(h)}, \mathbf{S}, k, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta, \tau, \mathbf{y}, \mathbf{X}) \sim N_p(\boldsymbol{\beta}_h; \hat{\boldsymbol{\beta}}_h, \hat{V}_{\beta_h}), \quad (22)$$

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

$$(\tau | \mathbf{S}, k, \mathbf{C}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_k, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta, \mathbf{y}, \mathbf{X}) \sim \mathcal{G}\left(a_\tau + \frac{n}{2}, b_\tau + \frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}_i' \boldsymbol{\beta}_i)^2\right), \quad (23)$$

$$(\boldsymbol{\beta} | \mathbf{S}, k, \mathbf{C}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_k, \tau, \boldsymbol{\Sigma}_\beta, \mathbf{y}, \mathbf{X}) \sim N_p(\hat{\boldsymbol{\beta}}, \hat{V}_\beta), \quad (24)$$

$$(\boldsymbol{\Sigma}_\beta^{-1} | \mathbf{S}, k, \mathbf{C}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_k, \tau, \boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) \sim \mathcal{W}\left(\left\{ \sum_{h=1}^k (\boldsymbol{\beta}_h - \boldsymbol{\beta})(\boldsymbol{\beta}_h - \boldsymbol{\beta})' + \nu_0 \boldsymbol{\Sigma}_0 \right\}^{-1}, k + \nu_0\right) \quad (25)$$

where $\hat{V}_\beta = (V_{\beta_0}^{-1} + k \boldsymbol{\Sigma}_\beta^{-1})^{-1}$ and $\hat{\boldsymbol{\beta}} = \hat{V}_\beta (V_{\beta_0}^{-1} \boldsymbol{\beta}_0 + \sum_{h=1}^k \boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\beta}_h)$.

4.3 Predictive density

One of the primary goals of the methodology is to estimate the predictive density of a future observation y_{n+1} from a new subject with predictors $\mathbf{x}_{n+1} = \mathbf{x}$. Conditional on parameters, which can be marginalized out in the MCMC algorithm, and focusing on the normal linear regression case of subsection 4.2, we obtain

$$\begin{aligned} & (y_{n+1} | \mathbf{x}_{n+1} = \mathbf{x}, \mathbf{S}, k, \mathbf{C}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta, \tau, \mathbf{y}, \mathbf{X}) \\ & \sim \left(\sum_{j=1}^n \frac{\alpha b_j(\mathbf{x})}{\alpha + \sum_{i=1}^n 1(\mathcal{C}_{\mathcal{S}_i} = j)} \right) \int f(y_{n+1} | \mathbf{x}, \boldsymbol{\beta}_{n+1}, \tau) dG_0(\boldsymbol{\beta}_{n+1}; \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta) \\ & \quad + \sum_{h=1}^k \left(\frac{b_{\mathcal{C}_h}(\mathbf{x}) \sum_{i=1}^n 1(\mathcal{S}_i = h)}{\alpha + \sum_{i=1}^n 1(\mathcal{C}_{\mathcal{S}_i} = \mathcal{C}_h)} \right) N(y_{n+1}; \mathbf{x}' \boldsymbol{\beta}_h, \tau^{-1}) \\ & \stackrel{d}{=} \omega_{n,0}(\mathbf{x}) N(y_{n+1}; \mathbf{x}' \boldsymbol{\beta}, \tau^{-1} + \mathbf{x}' \boldsymbol{\Sigma}_\beta \mathbf{x}) + \sum_{h=1}^k \omega_{n,h}(\mathbf{x}) N(y_{n+1}; \mathbf{x}' \boldsymbol{\beta}_h, \tau^{-1}), \quad (26) \end{aligned}$$

which is a finite mixture of normal linear regression models, with $\boldsymbol{\omega}_n(\mathbf{x}) = [\omega_{n,h}(\mathbf{x}), h = 0, 1, \dots, k]'$ predictor-dependent mixture weights. Note that this expression is closely-related to the hierarchical mixtures of experts model (Jacobs and Jordan, 1994), though it is obtained in a fundamentally different manner.

Note that large values of α will lead to a high degree of shrinkage towards the first normal component. The first component will also receive high probability weight when there are few subjects in the data set close to \mathbf{x}_{n+1} , because then $\{b_{C_h}(\mathbf{x}_{n+1}), h = 1, \dots, k\}$ tend to be close to zero. For the remaining k normal components, each of which has a distinct set of regression coefficients, the weights will depend adaptively on location of $\mathbf{x}_{n+1} \in \mathcal{X}$. The number of components k is treated as unknown and will change across the MCMC samples. To remove the conditioning on the unknowns in expression (26), one can calculate the expected predictive density averaging over the posterior distribution by using a large number of iterates collected after apparent convergence of the Gibbs sampling algorithm.

5. Simulation Examples

In order to assess the computational performance of the Gibbs sampling algorithm and whether the approach seems to give reasonable results, we analyzed data from two simulated examples. We let $n = 500$ and $p = 2$, with $\mathbf{x}_i = (1, x_{i2})$ and x_{i2} simulated from a uniform(0,1) density. For the hyperparameters, we let $\alpha = 0.1$ to favor the introduction of few clusters per location, $\mu_\psi = \log(30)$, $\sigma_\psi^2 = 0.5$, $\mu_\kappa = -2.5$, $\sigma_\kappa^2 = 1$, $\boldsymbol{\beta}_0 = \mathbf{0}$, $V_{\beta_0} = (\mathbf{X}'\mathbf{X})^{-1}/n$, $\nu_0 = p$, $\boldsymbol{\Sigma}_0^{-1} = \mathbf{I}_{p \times p}$, and $a_\tau = b_\tau = 0.1$. Rather than rerunning for many different hyperparameters, we assessed sensitivity by examining how well conditional density estimates approximated the truth in a variety of cases. We also investigated Bayesian learning about the key hyperparameters, ψ and κ .

As a null case, we first simulated data under the normal linear regression model, $f(y_i | \mathbf{x}_i) = N(y_i; -1 + 2x_{i2}, 0.01)$. We then analyzed the simulated data using the proposed Gibbs sampling algorithm run for 30,000 iterations with a 10,000 iteration burn-in. Trace plots of ψ , κ , the number of occupied locations, and the number of clusters across locations are shown in Figure 1. It is apparent that each of these unknowns rapidly converged to a stationary distribution, suggesting good

performance of the proposed MCMC algorithm, though there was some evidence of slow-mixing in the κ chain. Interestingly, all subjects were assigned to one or two locations with high probability, and the posterior for κ was tightly distributed about 0.0053 (95% interval = [0.0041,0.0066]), with the posterior std of 0.0028 much lower than the prior variance of 0.705. Figure 2 shows the predictive density of y_{n+1} at the 10th, 25th, 50th, 75th, and 90th percentiles of the empirical distribution of x_{i2} . The predictive mean regression function closely approximated the true linear regression function, which was entirely enclosed in pointwise 99% credible intervals. In addition, the predictive densities were essentially indistinguishable from the true densities.

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

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

Figure 3 shows the true density (dotted line), estimated predictive density (solid line), and pointwise 99% credible intervals (dashed lines) for a range of values of x_{i2} . The estimates correspond approximately to the true densities in each case. The bottom right panel contains an $x - y$ plot of the data along with the estimated predictive mean curve (solid line), which is indistinguishable from the true mean curve (dotted line). The estimated value of ψ was 70.2, with a 95% interval of [26.9,94.3], while the estimated value of κ was 0.0049, with a 95% interval of [0.0038,0.0063]. The value of κ was similar to that obtained in case 1, while ψ was considerably higher, as expected. The number of occupied locations was higher than in case 1, with a 96.7% posterior probability of one occupied location in case 1, and a 95.7% probability of two occupied locations in case 2.

Repeating the analysis as described above, but with $\phi_i \stackrel{iid}{\sim} G$ and $G \sim DP(\alpha G_0)$, we obtained poor results (density estimates diverged substantially from true densities, posterior mean curve failed to capture true non-linear function), suggesting that a DP mixture model is inadequate.

6. Application: Epidemiologic Study

6.1 Data structure and scientific problem

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

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

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

6.2 Analysis and Results

For woman i ($i = 1, \dots, 522$), let y_i , x_{i2} and x_{i3} denote BMI, LH and age, respectively. Variables were normalized prior to analysis, but transformed back to the original scale in presenting the results. Prior specification and posterior computation proceeded as in Section 4. As demonstrated in Figure 4, samples appeared to converge rapidly to a stationary distribution and mixing was acceptable. As in simulation case 2, the estimated value of κ was close to 0.005 and all subjects were allocated to either 2 or 3 locations with high probability. However, unlike in the simulation

case, ψ was small, suggesting it is appropriate to borrow information widely across the predictor space in estimating the mixture distributions.

Figure 5 presents the estimated predictive density of BMI for LH values corresponding to the 1, 10, 25, 50, 75 and 90th percentiles of the empirical distribution, with age fixed at the sample mean value to simplify presentation of the results. To assess interactions with age, we also obtained plots for age fixed at a low or high value, but the estimates were essentially identical to Figure 5. As expected, the BMI densities tend to be right skewed. Interestingly, the distributions are more highly skewed, with a greater proportion of morbidly obese women ($\text{BMI} > 40$), when LH values are low. In fact, there is even evidence of three modes at low LH values. As an empirical check, we obtained kernel density estimates in R using the *density* function, for different ranges of LH, without adjustment for age. Although the estimates were sensitive to subjective choice of the bandwidth and to the range of LH values chosen, we obtained results in agreement with Figure 5.

These results suggest that obese women, particularly morbidly obese women, tend to have low LH levels. This goes against the hypothesis that obese women may be at greater risk of uterine smooth muscle cell proliferation and fibroid development due to increased LH. However, it is consistent with our secondary hypothesis that obese women may have diminished reproductive functioning, which is manifest by low LH levels. The raw LH and BMI data are plotted in Figure 6, along with the age-adjusted posterior predictive mean regression curve and pointwise 99% credible intervals. The second mode at low LH levels, which was picked up by our density regression estimator, is also apparent in the raw data. Overall, there is a decreasing trend in mean BMI with increasing LH, with the nonlinear curve flattening out at higher LH levels. The corresponding plot for age vs BMI is shown in Figure 7.

7. Discussion

This article has proposed a Bayesian approach to the density regression problem, relying on a nonparametric mixture of parametric regression models. The statistically novel aspect was the proposed WMDP prior for the uncountable collection of unknown mixture distributions indexed by the predictors. This specification has a number of appealing theoretic properties, including a simple

form for the dependency in random measures at different predictor values and a generalization of the Pólya urn scheme to incorporate predictor-dependent weights. In addition, the specification leads to straightforward computation using a generalization of MCMC algorithms commonly used for DP mixture models.

In future research, it will be interesting to consider additional properties of the prior specification and generalizations. The current specification relies on placing random basis measures at the sample predictor values, so a natural question is how to avoid sample-dependency. This can potentially be accomplished by placing a random probability measure on the distribution of basis locations, allowing them to be assigned to any location in the continuous predictor space.

Another area of interest is to develop methods that rely directly on the generalized Pólya urn scheme by using a kernel to borrow information about the weights $\{w_{ij}\}$ without explicitly specifying $\mathbf{b}(\mathbf{x})$. For example, one could let $w_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, avoiding the need to update the γ weights and keep track of the assignment to DP components in implementing posterior computation.

Appendix A: Proof of theorem 1

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

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

The numerator can be expressed as follows:

$$\begin{aligned} & \text{E}\left[\{b_{i1}G_{\mathbf{x}_1}^* + b_{i2}G_{\mathbf{x}_2}^* + \dots + b_{in}G_{\mathbf{x}_n}^*\}\{b_{i'1}G_{\mathbf{x}_1}^* + b_{i'2}G_{\mathbf{x}_2}^* + \dots + b_{i'n}G_{\mathbf{x}_n}^*\} - G_0(B)^2\right] \\ &= \text{E}\left(\left\{\sum_{h=1}^n b_{ih}b_{i'h}G_{\mathbf{x}_h}^*(B)^2\right\} + \left[\sum_{h=1}^n b_{ih}G_{\mathbf{x}_h}^*(B)\left\{\sum_{l \neq h} b_{i'l}G_{\mathbf{x}_l}^*(B)\right\}\right]\right) - G_0(B)^2 \\ &= \left\{\sum_{h=1}^n b_{ih}b_{i'h}\text{E}\{G_{\mathbf{x}_h}^*(B)^2\}\right\} + \left\{\sum_{h=1}^n b_{ih}(1 - b_{i'h})G_0(B)^2\right\} - G_0(B)^2 \\ &= \left(\sum_{h=1}^n b_{ih}b_{i'h}\left[\frac{G_0(B)\{1 - G_0(B)\}}{1 + \alpha} + G_0(B)^2\right]\right) + G_0(B)^2 \sum_{h=1}^n b_{ih} - G_0(B)^2 \sum_{h=1}^n b_{ih}b_{i'h} - G_0(B)^2 \\ &= \left(\sum_{h=1}^n b_{ih}b_{i'h}\right)\left[\frac{G_0(B)\{1 - G_0(B)\}}{1 + \alpha}\right]. \end{aligned} \tag{27}$$

Note that the term in $[\cdot]$ equals $\text{V}\{G_{\mathbf{x}_h}^*(B)\}$ while $\text{V}\{G_{\mathbf{x}_h}(B)\} = \|\mathbf{b}(\mathbf{x}_i)\|^2 \text{V}\{G_{\mathbf{x}_h}^*(B)\}$, for $h = i, i'$.

Hence, $\text{corr}\{G_{\mathbf{x}_i}(B), G_{\mathbf{x}_{i'}}(B)\} = \langle \mathbf{b}(\mathbf{x}_i), \mathbf{b}(\mathbf{x}_{i'}) \rangle / \{\|\mathbf{b}(\mathbf{x}_i)\| \cdot \|\mathbf{b}(\mathbf{x}_{i'})\|\}$.

Appendix B: Proof of theorem 2

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proof: Our main argument for this proof is the following lemma from Serre (2002, page 80). A matrix S is non negative if and only if $x \geq 0$ implies $Sx \geq 0$.

Note that, by lemma 1, the matrix \mathbf{B} is invertible. Now to show that \mathbf{B} is non negative, by the above lemma, it is enough to show that for any $\mathbf{b} \geq 0$, $\mathbf{B}\mathbf{b} = x \geq 0$. However, $\mathbf{b} = \mathbf{B}^{-1}x$, and hence it is enough to show that for any $\mathbf{b} \geq 0$ such that $\mathbf{B}^{-1}x = \mathbf{b}$, then we have $x \geq 0$.

Indeed, let $\mathbf{b} = (b_1, b_2, \dots, b_n)'$, such that $b_i \geq 0$ for $i \in \{1, 2, \dots, n\}$. Let \mathbf{x} be the solution of the equation, $\mathbf{B}^{-1}\mathbf{x} = \mathbf{b}$ and $i = \operatorname{argmin} x_i$. Then we have

$$\begin{aligned} b_i &= \frac{1}{a_{ii}}x_i - \sum_{j=1, j \neq i}^n \frac{a_{ij}}{a_{ii}}x_j \\ \Rightarrow \frac{1}{a_{ii}}x_i &= b_i + \sum_{j=1, j \neq i}^n \frac{a_{ij}}{a_{ii}}x_j \geq \sum_{j=1, j \neq i}^n \frac{a_{ij}}{a_{ii}}x_j \geq \sum_{j=1, j \neq i}^n \frac{a_{ij}}{a_{ii}}x_i \\ &\Rightarrow \left(\frac{1}{a_{ii}} - \sum_{j=1, j \neq i}^n \frac{a_{ij}}{a_{ii}} \right) x_i \geq 0 \Rightarrow x_i \geq 0 \end{aligned}$$

and since x_i was the minimum, we have $\mathbf{x} \geq 0$. Hence the matrix \mathbf{B} has non negative entries.

Appendix C: Proof of theorem 4

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

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

It is straightforward to obtain the simple closed form solution $w_{i+} = \alpha(1 - \mathbf{p}'_{i0}\mathbf{\Gamma}_0)/(\mathbf{p}'_{i0}\mathbf{\Gamma}_0)$, for $i = 1, \dots, n$, where $\Pr(M_{i+} = \sum_{j \neq i} M_{ij} = m) = p_{i0m}$, for $m = 0, \dots, n - 1$, is the probability mass function for M_{i+} , which can be calculated from \mathbf{B} using expression (13).

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

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

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

Letting $R_{ij} = \mathbf{p}'_{ij} \mathbf{\Gamma}_1 / \mathbf{p}'_{i0} \mathbf{\Gamma}_0$, we have

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

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

$$w_{ij} = \frac{\sum_{m=1}^n \tilde{p}_{ijm} \left(\frac{1}{\alpha+m-1} \right)}{\sum_{m=1}^n p_{i0m} \left(\frac{1}{\alpha+m-1} \right)}. \quad (28)$$

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

References

- Antoniak, C.E. (1974) Mixtures of Dirichlet processes with application to nonparametric problems. *The Annals of Statistics*, **2**, 1152-1174.
- Blackwell, D. and MacQueen, J.B. (1973) Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, **1**, 353-355.
- Cifarelli, D., and Regazzini, E. (1978) Problemi statistici non parametrici in condizioni di scambiabilità parziale e impiego di medie associative. Technical Report, Quaderni Istituto Matematica Finanziaria, Torino.
- De Iorio, M., Müller, P., Rosner, G.L. and MacEachern, S.N. (2004) An Anova model for dependent random measures. *Journal of the American Statistical Association*, **99**, 205-215.
- Duan, J.A., Guidani, M. and Gelfand, A.E. (2005). Generalized spatial Dirichlet process models. *ISDS Discussion Paper 2005-23*, Duke University, Durham, NC.

- Dunson, D.B. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics*, in press.
- Dunson, D.B. and Stanford, J.B. (2005). Bayesian inferences on predictors of conception probabilities. *Biometrics*, 61, 126-133.
- Escobar, M.D. and West, M. (1998) Computing nonparametric hierarchical models,” In *Practical Nonparametric and Semiparametric Bayesian Statistics* (eds. D. Dey, P. Müller and D. Sinha), pp. 1-22. New York: Springer-Verlag.
- Ferguson, T.S. (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1, 209-230.
- Ferguson, T.S. (1974) Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2, 615-629.
- Gelfand, A.E., Kottas, A., and MacEachern, S.N. (2004) Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100, 1021-1035.
- Giudici, P., Mezzetti, M., and Muliere, P. (2003) Mixtures of Dirichlet process priors for variable selection in survival analysis. *Journal of Statistical Planning and Inference*, 111, 101-115.
- Griffin, J.E. and Steel, M.F.J. (2004) Semiparametric Bayesian inference for stochastic frontier models. *Journal of Econometrics*, 123, 121-152.
- Griffin, J.E. and Steel, M.F.J. (2006) Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101, 179-194.
- Holmes, C.C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* 1, 145-168.
- Jordan, M.I. and Jacob, R.A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6, 181-214.

- Ishwaran, H. and James, L.F. (2001) Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161-173.
- Ishwaran, H. and James, L.F. (2003) Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, **13**, 1211-1235.
- MacEachern, S.N. (1994) Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation*, **23**, 727-741.
- MacEachern, S.N. (1999) Dependent Nonparametric Processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association.
- MacEachern, S.N. (2000) Dependent Dirichlet processes. Unpublished manuscript, Department of Statistics, The Ohio State University.
- MacEachern, S.N. (2001) Decision theoretic aspects of dependent nonparametric processes. In *Bayesian Methods with Applications to Science, Policy and Official Statistics*, ed. E. George. Creta: ISBA, pp. 551-560.
- MacEachern, S.N. and Müller, P. (1998) Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223-238.
- Mira, A. and Petrone, S. (1996) Bayesian hierarchical nonparametric inference for change-point problems. In *Bayesian Statistics 5* (eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith), Oxford: Oxford University Press.
- Muliere, P. and Petrone, S. (1993) A Bayesian predictive approach to sequential search for an optimal dose: parametric and nonparametric models. *Journal of the Italian Statistical Society*, **2**, 349-364.
- Müller, P., Erkanli, A. and West, M. (1996), "Bayesian Curve Fitting using Multivariate Normal Mixtures," *Biometrika*, 83, 67-79.

- Müller, P. and Quintana, F.A. (2004) Nonparametric Bayesian Data Analysis. *Statistical Science*, **19**, 95-110.
- Müller, P., Quintana, F. and Rosner, G. (2004) A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society B*, **66**, 735-749.
- Serre, D. (2002). *Matrices: Theory and Application*. New York: Springer-Verlag.
- Sethuraman, J. (1994) A constructive definition of the Dirichlet process prior. *Statistica Sinica*, **2**, 639-650.
- West, M. (1992). Hyperparameter estimation in Dirichlet process mixture model. *ISDS Discussion Paper #92-03*, Duke University.
- West, M., Müller, P. and Escobar, M.D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. In *A Tribute to D. V. Lindley* (A.F.M. Smith and P.R. Freeman). John Wiley and Sons.
- Viele, K and Tong, B. (2002). Modeling with mixtures of linear regressions. *Statistics and Computing* **12**, 315-330.
- Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pp. 233-243. North-Holland/Elsevier.

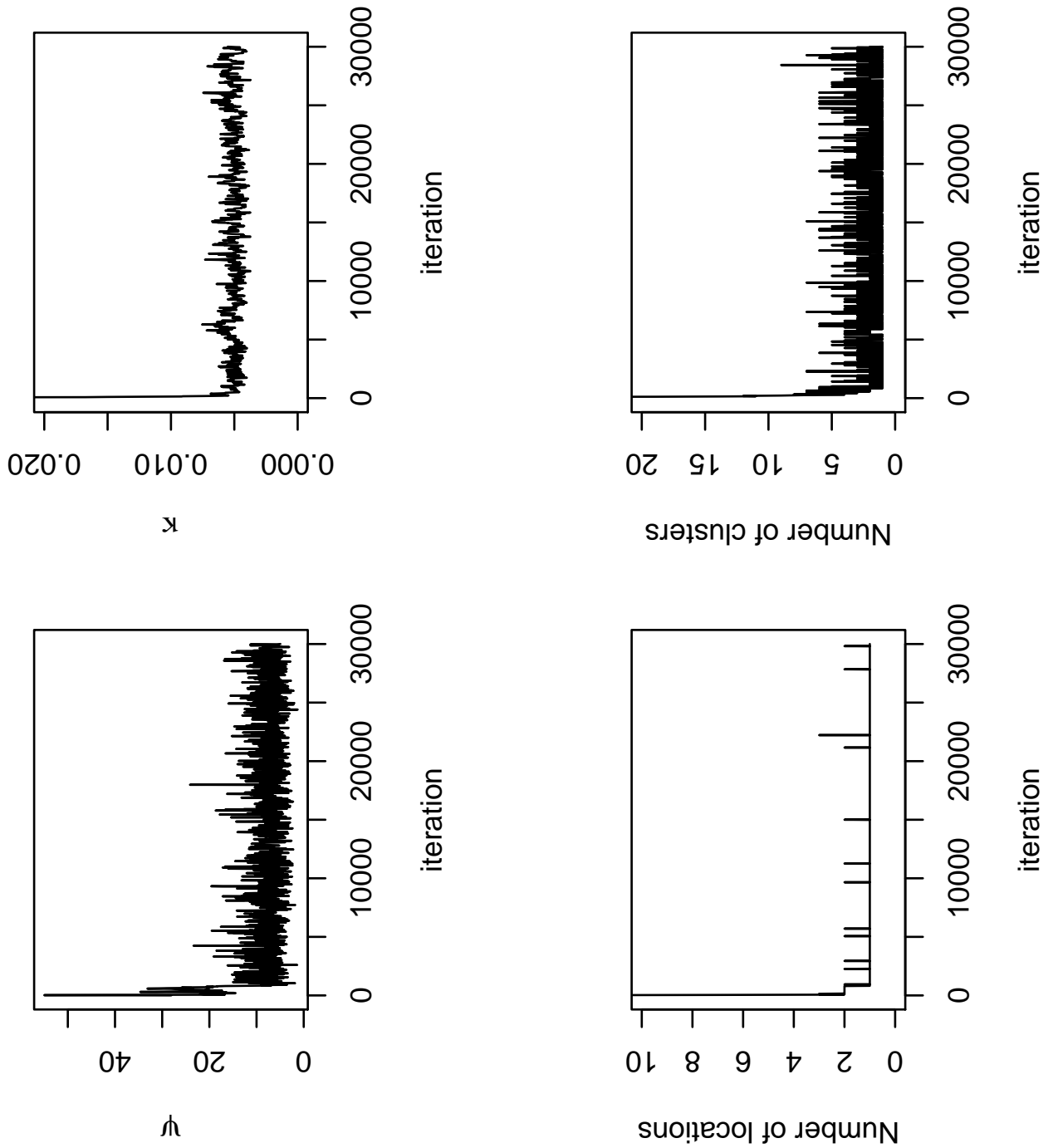


Figure 1: Trace plots for ψ , κ , the number of basis locations occupied, and the total number of clusters obtained for 30,000 MCMC iterations in simulation case 1.

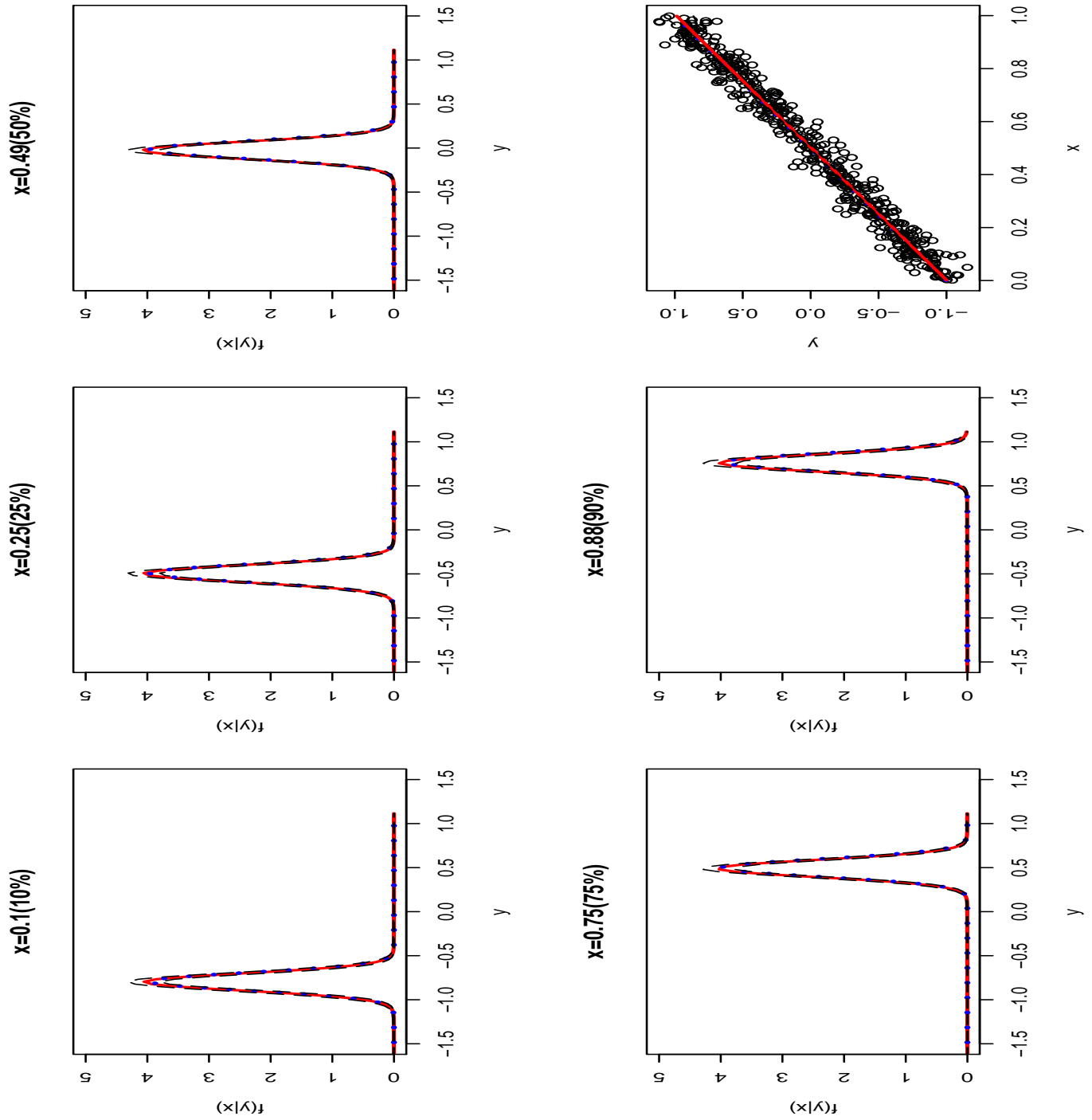


Figure 2: True conditional densities of $y|x$ (dotted lines), posterior mean estimates (solid lines), and 99% pointwise credible intervals (dashed lines) in simulation case 1. The lower right panel shows the data, along with true and estimated mean regression curves.

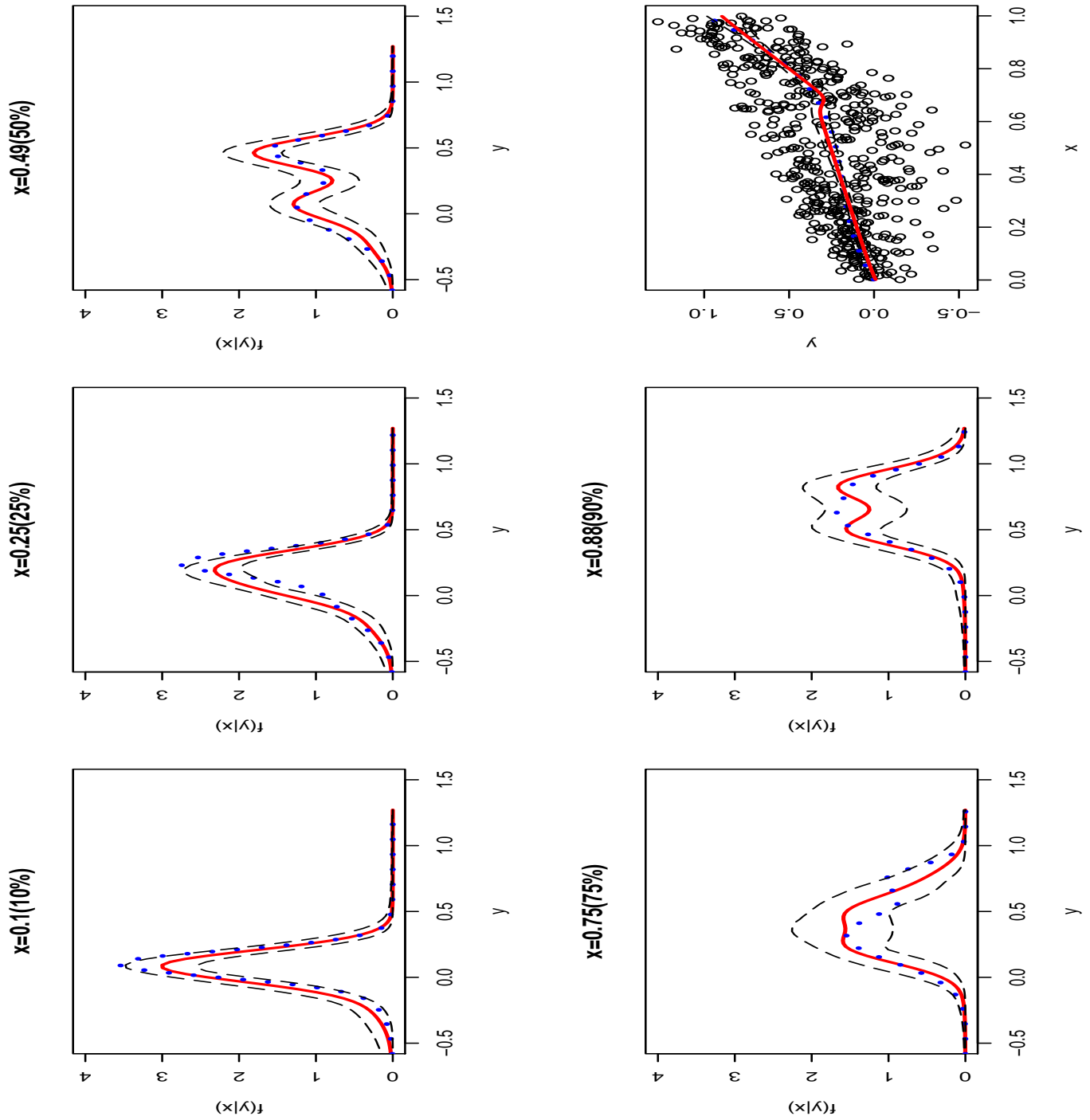


Figure 3: True conditional densities of $y|x$ (dotted lines), posterior mean estimates (solid lines), and 99% pointwise credible intervals (dashed lines) in simulation case 2. The lower right panel shows the data, along with true and estimated mean regression curves.

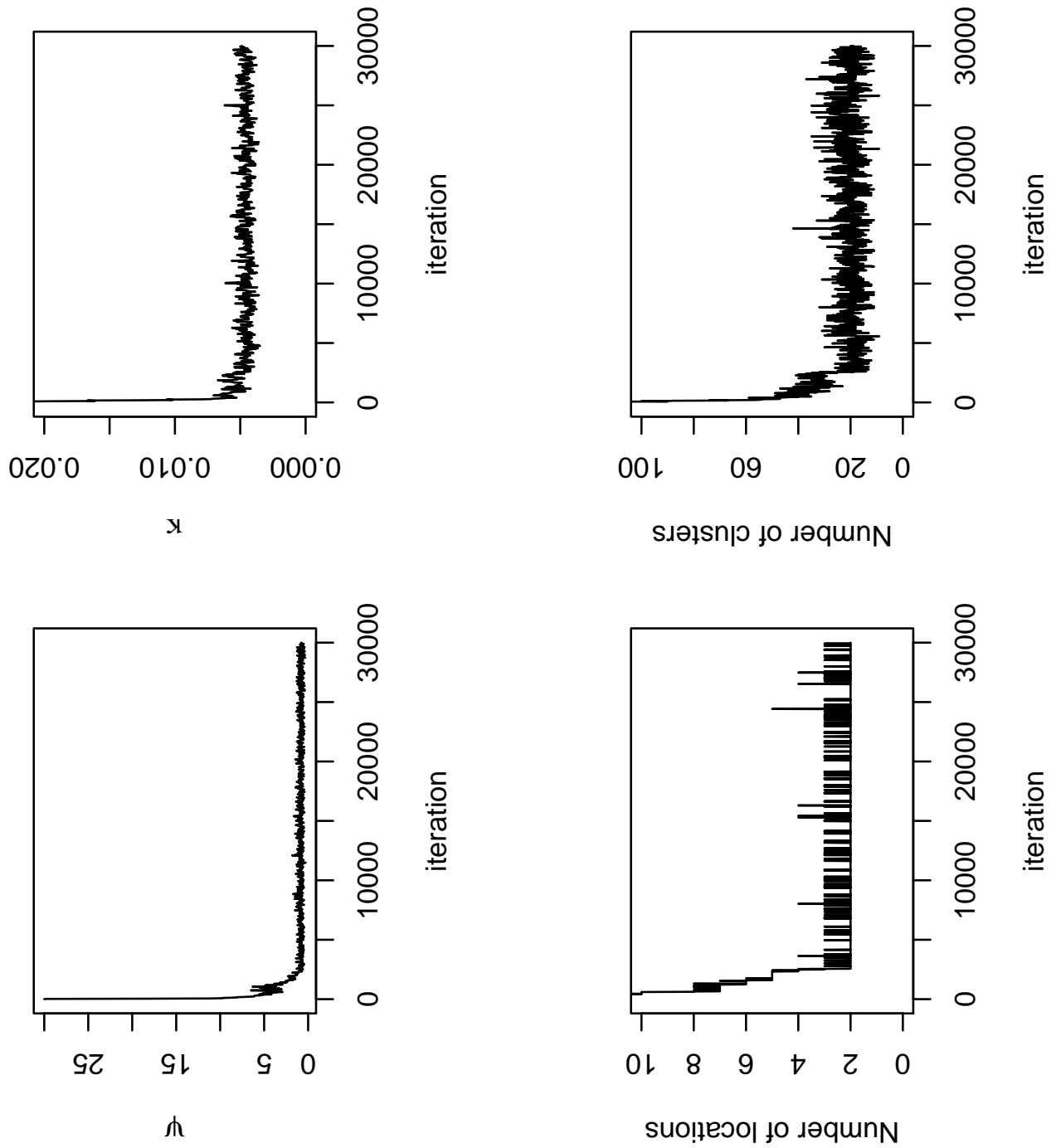


Figure 4: Trace plots for ψ , κ , the number of basis locations occupied, and the total number of clusters obtained for 30,000 MCMC iterations in the BMI application.

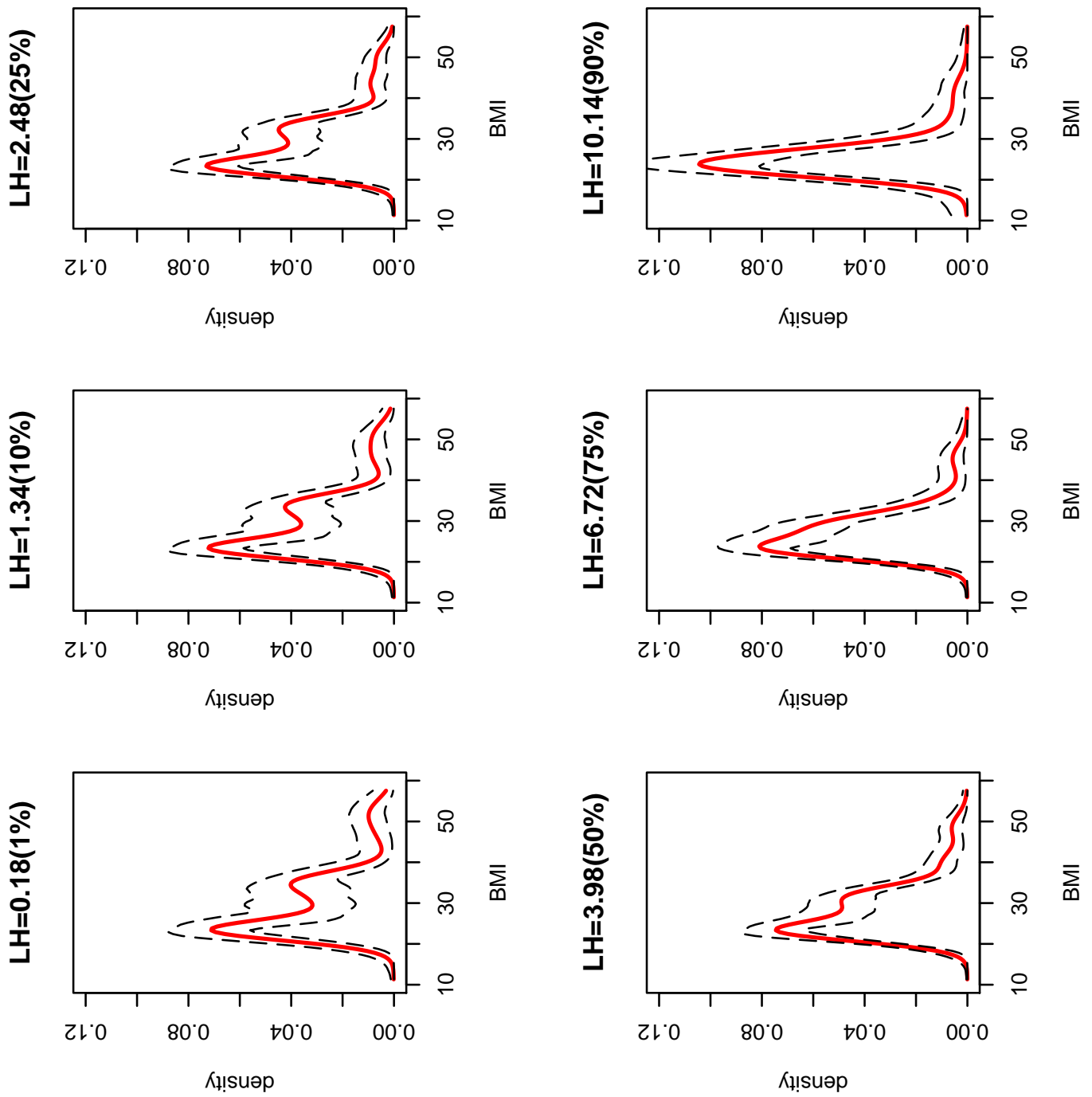


Figure 5: Predictive densities for body mass index (BMI) conditional on a range of values for luteinizing hormone (LH), with age fixed at the sample mean. Posterior predictive means (solid lines) and 99% pointwise credible intervals are shown.

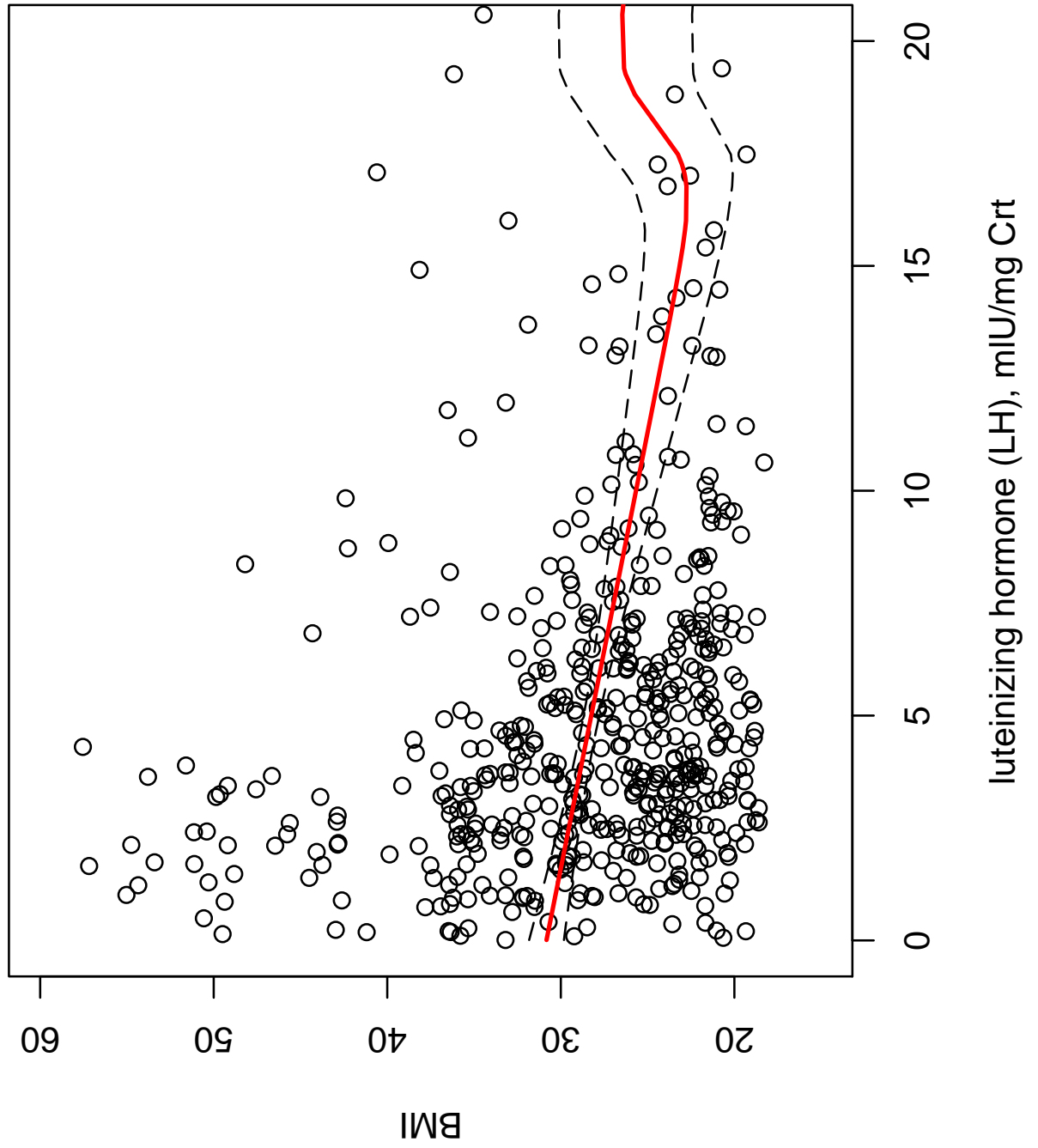


Figure 6: LH and BMI raw data, with the age-adjusted predictive mean curve (solid line) and 99% credible intervals (dashed lines).

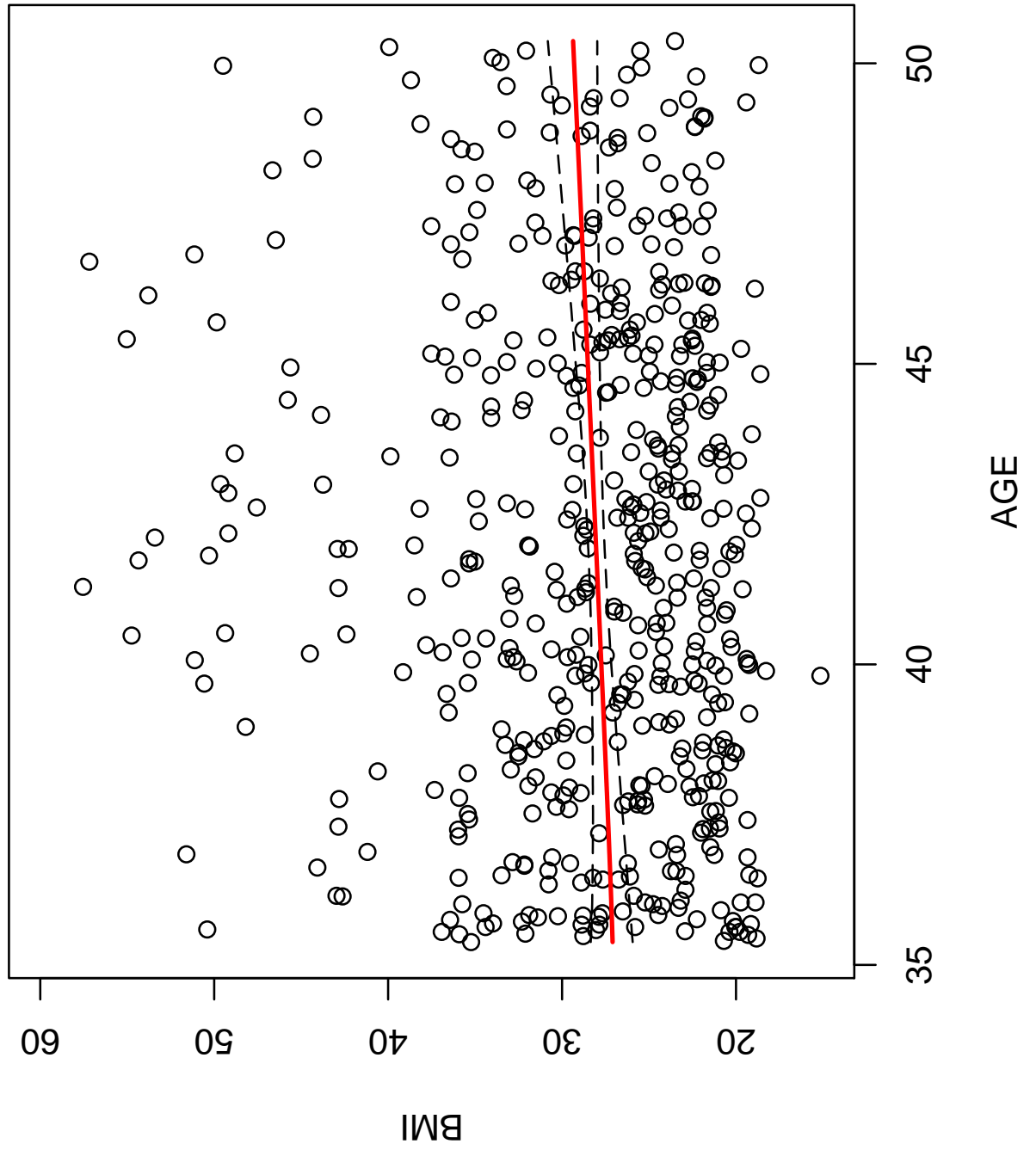


Figure 7: Age and BMI raw data, with the LH-adjusted predictive mean curve (solid line) and 99% credible intervals (dashed lines).