Semiparametric Bayesian Regression via Potts Model

Alejandro Murua\textsuperscript{a} and Fernando A. Quintana\textsuperscript{b}

\textsuperscript{a}Département de Mathématiques et de Statistique, Université de Montréal, Montréal, Canada; \textsuperscript{b}Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Santiago, Chile

**ABSTRACT**

We consider Bayesian nonparametric regression through random partition models. Our approach involves the construction of a covariance-dependent prior distribution on partitions of individuals. Our goal is to use covariate information to improve predictive inference. To do so, we propose a prior on partitions based on the Potts clustering model associated with the observed covariates. This drives by covariate proximity both the formation of clusters, and the prior predictive distribution. The resulting prior model is flexible enough to support many different types of likelihood models. We focus on the discussion of nonparametric regression. Implementation details are discussed for the specific case of multivariate multiple linear regression. The proposed model performs well in terms of model fitting and prediction when compared to other alternative nonparametric regression approaches. We illustrate the methodology with an application to the health status of nations at the turn of the 21st century. Supplementary materials are available online.

1. Introduction

Clustering is an ubiquitous problem in statistical theory and practice. Many methods and algorithms are available in the literature. A clustering procedure consists of constructing a partition of a set of individuals or experimental units into nonempty, disjoint, and exhaustive subsets. Usually, the number of subsets (the size of the partition) is not known beforehand. We focus on probability models for partitions (also referred to as random partition models), avoiding purely algorithmic methods. Specifically, we build a prior probability model for partitions that is indexed by covariates and has the property of encouraging coclustering of individuals with close covariate values.

Defining probability models for partitions has been a topic of active research in the past decade in the Nonparametric Bayes (NPB) community. The main reason comes from the NPB model itself. Consider any hierarchical model of the form \( y_i \mid \theta_i \sim p(y_i \mid \theta_i) \) and \( \theta_i \mid F_i \sim F_i \), for \( i \in [n] = \{1, \ldots, n\} \). When \( F_i \) is drawn from a discrete random probability measure, a prior distribution on partitions is induced. The discreteness of \( F_i \) implies the existence of ties among \( \theta_1, \ldots, \theta_n \). A partition is obtained by grouping together those elements in \( [n] \) that have identical \( \theta \) values. See, for example, Quintana (2006).

A natural extension of the random partition model consists of incorporating covariate values in its definition. The motivation for this comes in part from the NPB community’s interest in defining classes of covariate-dependent random probability measures \( \{F_i : x \in X\} \) indexed by sets of predictors/covariates \( x \in X \). The seminal work by MacEachern (1999, 2000) proposed a collection of dependent random probability measures with marginal distributions given by the Dirichlet process (DP) (Ferguson 1973). This general idea has been extended and applied to the construction of several types of random probability measures, producing what may be generically termed NPB regression. This includes, for instance, the notion of density regression (Dunson, Pillai, and Park 2007; Tokdar, Zhu, and Ghosh 2010). When the \( F_i \) distributions are discrete, a collection of random partition models indexed by the set of available covariate vectors \( x^n = (x_1, \ldots, x_n) \) emerges just as before. We note however, that the induced probability distribution may not be easily recognizable.

Exploiting the connection between product partition models (PPM) (Hartigan 1990; Barry and Hartigan 1992) and the DP (see, e.g., Quintana 2006), a covariate-dependent extension was proposed by Müller, Quintana, and Rosner (2011). The usual PPM involves a prior distribution on a partition \( \rho_n = (S_1, \ldots, S_k) \) of \( [n] \) as \( p(\rho_n) \propto \prod_{j=1}^k c(S_j) \), where \( c(S) \) (the cohesion function) describes how strongly the user thinks the elements of \( S \) are to co-cluster a priori. The covariate-dependent version (termed PPMx) alters this prior distribution as \( p(\rho_n \mid x^n) \propto \prod_{j=1}^k c(S_j) g(x^n_j) \), where \( x^n_j = [x_i : i \in S_j] \) is the set of all covariates for individuals in cluster \( j = 1, \ldots, k_n \), and \( g(x^n_j) \) is the similarity of cluster \( j \). Large similarity values are meant to represent homogeneous covariate values in the corresponding cluster. Some alternative extensions to build covariate-dependent random partition models explicitly can be found in Dahl (2008), Park and Dunson (2010), and Argiento, Cremaschi, and Guglielmi (2014).

Our motivation is similar to that of Müller, Quintana, and Rosner (2011), but our approach differs substantially from theirs in the way the prior distribution on partitions is constructed. Our approach is based on Potts model clustering, introduced

**CONTACT**

Fernando A. Quintana \textsuperscript{a}quintana@mat.uc.cl Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Santiago, Chile.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/jjCGS.

© 2017 American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America.
to the statistical community by Murua, Stanberry, and Stuetzle (2008), but developed earlier in the Physics literature (Blatt, Wiseman, and Domany 1996); see, for example, Reichardt and Bornholt (2004), Ott et al. (2005) and references therein. Specifically, our prior distribution on partitions is induced by applying the Potts model to the set of covariate values. Note that the Potts model formalizes our notion of prior coclustering encouraged by proximity. The Potts model has been successfully applied to recover cluster structures with various shapes. See, for example, Murua and Wicker (2014).

The article is organized as follows. Section 2 introduces the notation, and discusses the general modeling strategy, the prior on partitions, and a general posterior sampling algorithm. Section 3 leads the discussion to the realm of nonparametric multiple regression. Specifically, we consider a model with cluster-specific multivariate multiple regression parameters. The flexibility of the model is assessed and contrasted with competing alternatives in a simulation study described in Section 4. In Section 5 we apply our method to the analysis of the health status of nations at the turn of the 21st century. Final remarks and comments are given in Section 6. Proofs of technical results are given in a supplementary document.

2. The Modeling Approach

Suppose that each individual $i \in [n]$ in a given sample is associated with a $p$-dimensional vector $y_i$ of responses of interest and a $q$-dimensional covariate vector $x_i$. Suppose also that the interest lies in studying the relationship between $y_i$ and $x_i$, and in particular, in predicting the response $y_{n+1}$ associated with a covariate vector $x_{n+1}$ of a future individual. Let $p(y_i \mid x_i, \psi_i)$ be a likelihood model stating the relationship between the $i$th response and the associated covariate vector. For instance, the likelihood could be a multivariate multiple regression model $N_p(x_i, \theta, \Sigma_i)$. In this case $\psi_i$ would be given by the pair $(\theta_i, \Sigma_i)$. The specific nature of $p(y_i \mid x_i, \psi_i)$ will not be needed until later when we discuss implementation issues.

We assume a random partition model with a hierarchical structure for these data:

$$
\begin{align*}
 y_1, \ldots, y_n &\mid \rho_n, \psi_1, \ldots, \psi_n \sim p(y_i \mid x_i, \psi_i) \\
 \psi_1, \ldots, \psi_n &\sim p(\psi) \quad \text{and} \quad \rho_n \sim p(\rho_n \mid x^n).
\end{align*}
$$

(1)

Here $\rho_n$ is a partition of $[n]$ into $k_n$ subsets. Also, $s_1, \ldots, s_n$ are cluster membership indicators such that $s_i = j$ if the $i$th individual belongs to the $j$th cluster. In addition $\psi_j = \psi_n^s$ for all $i \in [n]$. Model (1) groups in cluster $j$ those individuals having identical parameter value $\psi_j$. Individuals within this cluster are conditionally iid given $\psi_j$.

The clustering distribution in (1) is generically denoted by $p(\rho_n \mid x^n)$. We clarify that the “$x^n$” notation refers not to an actual conditioning but to an indexing of the resulting distribution by the vector of available covariates $x^n$. We refer to models that give rise to this type of partitions as dependent random partition models. A review of such models was given by Müller and Quintana (2010). In defining $p(\rho_n \mid x^n)$, we follow the general principles laid out by Müller, Quintana, and Rosner (2011): two individuals $i_1$ and $i_2$ with similar covariate vectors $x_{i_1}$ and $x_{i_2}$ should be more likely to cocluster a priori. We adopt here a prior defined by the Potts model which we describe next.

2.1 The Potts Model Prior

To facilitate the upcoming development, we view the data structure as a graph $G = (Y^n, E)$ with $n$ nodes $y^n = (y_1, \ldots, y_n)$ and a set of edges $E$. We also consider the associated graph $(x^n, E)$ whose nodes are formed by the corresponding covariate vectors $x^n = (x_1, \ldots, x_n)$. These covariates will be used to index the random partition models as follows. Consider the Potts model applied to the graph $(x^n, E)$. We assume that the probability distribution $p(\rho_n \mid x^n)$ for partitions arises from the random cluster model (Sokal 1997) associated with the Potts model. This implies the existence of auxiliary binary variables, the so-called bonds $b = \{b_{ij}\}$, so that

$$
 p(\rho_n \mid x^n) = \sum_{b \in \rho_n} p(b \mid x^n).
$$

(2)

Here, the notation $b \Rightarrow \rho_n$ means that the partition $\rho_n$ arises from the connected components of the graph associated to edges given by $b$. To explain how (2) arises we introduce the following notation. Let $r > 1$ be an integer indicating the number of label colors for the individuals in the data graph. The value of $r$ is one of the defining parameters of the Potts model. Let $\zeta_{ij}$ be a color label indicator with $\zeta_{ij} = 1$ if $x_i$ and $x_j$ are neighbors in the graph $(x^n, E)$, that is, if there is an edge connecting them, and $0$ otherwise. Let $\delta_{ij} = 1$ if $x_i$ and $x_j$ have the same label, and $0$ otherwise. In general, we use values of $r$ in the range $10 \leq r \leq 30$. These are the recommended values in the literature (Blatt, Wiseman, and Domany 1996; Stanberry, Murua, and Cordes 2008). We stress that the color labels are not the clusters labels, they only serve to segment the graph. They are auxiliary variables that split and/or merge the connected components of the graph. Figure 1 displays an example of the coloring (labeling) and the creation of bonds in a data graph. The figure also illustrates the Swendsen-Wang algorithm described below. Let $K(\cdot, \cdot)$ be a Mercer kernel (Giroliami 2002). The similarities between pairs of covariate vectors are defined by $\kappa_{ij} = K(x_i, x_j)$. As usual, we will assume that $K$ is a function of the distances $\|x_i - x_j\|_2$, of the form $\kappa_{ij} = \kappa_{ij}(\sigma) = K(\|x_i - x_j\|_2 / \sigma)$, where $\sigma > 0$ is a bandwidth parameter. The kernel function is not limited to real-valued vector components, and can be easily modified to account for discrete variables. To avoid problems with different scales or variances in the data, we standardize the variables $x^n$. An alternative would be to consider variable-dependent bandwidth parameters in the kernel. But this would unnecessarily complicate the model. The choice of the bandwidth parameter is not straightforward. Blatt, Wiseman, and Domany (1996) suggested using the square-root of the mean of the squared distances. Murua and Wicker (2014) estimated it with a Wang-Landau-type stochastic algorithm (Wang and Landau 2001; Atchade and Liu 2010). However, they also suggested a faster estimate obtained from the mode of an informative empirical prior density of the bandwidth. In our experiments, we adopted this latter strategy.

The Potts model assumes the following density on the set of labels $\{\zeta_{ij}\}$ given $x^n$, or equivalently, the $\{\kappa_{ij}(\sigma)\}$ “distances”:

$$
 p(\{\zeta_{ij}\} \mid x^n) = \frac{1}{Z} \exp \left\{ -\beta \sum_{d_{ij}=1} \kappa_{ij}(\sigma) (1 - \delta_{ij}) \right\}.
$$

(3)
Here $\beta$ is the inverse of the so-called temperature parameter, and $Z$ is the normalizing constant which depends on $\beta$, $\sigma$ and $x^n$. As with the bandwidth parameter, $\beta$ needs to be estimated. In principle, it could be possible to make formal posterior inference on $\beta$ and $\sigma$. But in practice this is difficult. The main reason stems from the fact that the intractable normalizing constant $Z$ in (3), depends on these parameters. As a compromise, our approach considers setting the pair $(\beta,\sigma)$ to the mode of a data-driven prior as suggested in the work of Murua and Wicker (2014). To work in practice with the Potts model (3), Swendsen and Wang (1987) proposed an algorithm based on augmenting the model by introducing a set of latent variables, the bonds $b$. The bond $b_{ij} = 1$ is said to be frozen if $b_{ij} = 0$ and $a_{ij} = \delta_{ij} = 1$, that is, the points $x_i$ and $x_j$ are neighbors and have the same label. Otherwise, the bond is not frozen. The bond $b_{ij}$ becomes frozen with probability $p_{ij} = 1 - \exp(-\beta \kappa_{ij}(\sigma))$. The joint distribution of the labels and bonds is given by

$$p\left((z_{ij}), (b_{ij}) \mid x^n\right) = \frac{Z}{\sum_{\alpha_{ij}} \left[1 - p_{ij}(1 - b_{ij}) + p_{ij} b_{ij} \delta_{ij}\right]}.$$ (4)

This is known as the Fortuin-Kasteleyn-Swendsen-Wang model (Sokal 1997). The distribution in (3) follows by marginalizing out $b$ in (4). And $p(\mathbf{b} \mid x^n)$, needed in (2), follows by marginalizing out $\{z_{ij}\}$.

Swendsen and Wang (1987) used (4) to propose the following Gibbs sampling strategy for generating samples from the Potts model (and the associated random cluster model $p(\mathbf{b} \mid x^n)$):

Step 1. Given the labels $\{z_{ij}\}$, each bond $b_{ij}$ becomes frozen independently of the others with probability $p_{ij}$ if $a_{ij} = \delta_{ij} = 1$. Otherwise, $b_{ij}$ is set to 0.

Step 2. Given the bonds $\{b_{ij}\}$, each connected subset is assigned the same color label. The color label assignments are drawn uniformly at random and independently from each other.

It is important to note in Step 2 that all vertices in a connected component already share the same color assignment. By assigning a new color to the vertices in the connected component, neighboring vertices with the same color assignment become eligible to join the connected component in the next iteration (merging). Similarly, those neighboring vertices that were not joined by bonds but had the same color as the connected component vertices, become ineligible to join the component in the next iteration (splitting). Figure 1 displays an illustration of the Swendsen-Wang algorithm.

### 2.2 Posterior Simulation in the General Case

Our general modeling approach is expressed as (1) together with (2). The set of parameters $\Psi = (\psi_1, \ldots, \psi_n)$ may be alternatively represented as $(\rho_n, \psi_{\rho_n})$, where $\psi_{\rho_n} = (\psi^1_n, \ldots, \psi^n_\rho)$ and $k_\rho$ is the size of partition $\rho_n$. The corresponding posterior distribution of $\Psi$ is a crucial component for our posterior predictive inference goal.

Consider a hierarchical model with a discrete random probability measure as prior for the unknown distribution function. Random partition distributions arising from these type of models have several advantages. One of them is that there is typically an efficient posterior simulation algorithm that can be used or adapted to get samples for the posterior of partitions. Unfortunately, this is not the case for Potts model prior, and a different strategy is needed. Note first that

$$p(\psi \mid Y^n, x^n) = p(\rho_n, \psi_{\rho_n} \mid Y^n, x^n) = p(\psi_{\rho_n} \mid \rho_n, Y^n, x^n) p(Y^n \mid \rho_n, x^n) \times p(\rho_n \mid x^n) / p(Y^n \mid x^n).$$

Note that $p(Y^n \mid x^n) = \sum_{\rho_n} p(Y^n \mid \rho_n, x^n) p(\rho_n \mid x^n)$, where $p(\rho_n \mid x^n)$ (given in (2)) is intractable even for moderate values of $n$. Therefore, we use the Swendsen-Wang algorithm (Swendsen and Wang 1987) described in Section 2.1 as part of our MCMC approach to the target distribution $p(\psi \mid Y^n, x^n)$. Also note that changing the partition, say from $\rho_n$ to $\rho'_n$, involves changing the dimension of the parameters $\Psi$. Thus, it is necessary to resort to transdimensional MCMC. We choose to follow the framework given by Besag (2004), rather than the reversible jumps method proposed by Green (1995). Besag’s approach avoids the specification of a proposal bijection between the different dimensions. This is done by embedding the problem into a very-high-dimensional space. The idea is the following, where we drop the subindex $n$ for partitions to simplify the notation.

Let $\Omega = \{\rho(1), \rho(2), \ldots, \rho(B_\rho)\}$ be the collection of all possible partitions of $n$ nodes, where $B_\rho$ is the corresponding Bell number. Denote by $\Psi = (\psi_{\rho(1)}, \psi_{\rho(2)}, \ldots, \psi_{\rho(B_\rho)})$ the collection of all parameters associated to the elements of $\Omega$, and by $S_\rho$ the parameter space associated to the partition $\rho \in \Omega$. Also, for every $\rho \in \Omega$, let $p_0(\psi_{\rho} \mid \rho, Y^n, x^n)$ be any distribution on $S_\rho$. Define the target distribution as

$$\pi(\rho, \psi \mid Y^n, x^n) = p(\rho, \psi_{\rho} \mid Y^n, x^n) \prod_{\rho \neq \rho'} p_0(\psi_{\rho'} \mid R, Y^n, x^n).$$

It is easy to see that the marginal of $(\rho, \psi_{\rho})$ under this density is exactly $p(\rho, \psi_{\rho} \mid Y^n, x^n)$. The problem is thus reduced to sampling from $\pi(\rho, \psi \mid Y^n, x^n)$. Besag (2004) considered two different types of moves during the sampling: (a) from the current state $(\rho, \psi_{\rho})$, propose an updated vector of parameters $\psi_{\rho'}$ and leave the remaining $\psi_{\rho'}$’s unchanged; and (b) from the current state $(\rho, \psi_{\rho})$, propose for some $\rho' \neq \rho$, the update $(\psi_{\rho'}, \psi_{\rho'})$. The first update is generated with a proposal $q(\psi_{\rho'} \mid \psi_{\rho}, Y^n, x^n) p(\rho' \mid x^n) q_{\rho\rho'}(\psi_{\rho'}, \psi_{\rho'})$. The first
The discussion so far is generic in the sense that no particular form was assumed for the likelihood model $p(y_i | x_i, \psi_i)$. Naturally, details for particular applications are specific to whatever case is being considered. Thus, from now on we consider the special case of multivariate linear regression models. Our main goal here is to develop predictive inference for such regression models with multiple responses and predictors. The response vectors are assumed $p$-dimensional: $y_i \in \mathbb{R}^p$, and the corresponding covariates are $x_i \in \mathbb{R}^q$. The parameters are then $\psi_i = (\theta_i, \Sigma_i)$, where $\theta_i \in \mathbb{R}^{qp}$ and $\Sigma_i \in \mathbb{M}_{q \times q}$ are subject-specific regression coefficients and covariance matrices. Here, $\mathbb{M}_{q \times q}$ denotes the collection of matrices with $q$ rows and $q$ columns. The partition-based regression model has thus likelihood derived from (1). Note that we can write $\theta_i = (\theta_i^1, \ldots, \theta_i^p)$, where each $\theta_i^j \in \mathbb{R}^q$. Assume that

$$y_i | x_i, \psi_i \sim N_p (I_p \otimes x_i^\prime \theta_i, \Sigma_i),$$

where $I_p$ represents the identity matrix of dimension $a$ and $A \otimes B$ is the Kronecker product between two matrices $A$ and $B$. The individual likelihood corresponds to a $p$-dimensional multivariate normal with mean vector formed by the individual linear combinations $x_i' \theta_i$ for $\ell = 1, \ldots, q$, and covariance matrix $\Sigma_i$. The resulting density $p(y_i | x_i, \theta_i, \Sigma_i)$ is given by

$$(2\pi)^{-\frac{q}{2}} | \Sigma_i |^{-\frac{1}{2}} \exp \left\{ - \frac{1}{2} (y_i - I_p \otimes x_i' \theta_i)' \Sigma_i^{-1} (y_i - I_p \otimes x_i' \theta_i) \right\}.$$

(7)

The complete likelihood from (1) becomes

$$p(Y^n | x^n, \rho_n, \psi) = \prod_{j=1}^{k_n} p(y_j | x_j, \psi_j).$$

(8)

The model specification is completed by defining suitable priors on the partition $\rho_n$ and the parameters $\psi_i$. We assume that

$$\rho_n | x^n \sim \text{Potts model}, \quad (\theta_i, \Sigma_i) \sim \text{NIW}(\mu, \tau, v, \Lambda),$$

(9)

where the Potts model was stated in (2). Here, $\Lambda \in \mathbb{M}_{p \times p}$ and $\tau \in \mathbb{M}_{q \times q}$ are positive-definite matrices and $\mu \in \mathbb{R}^q$. Also, by $(\theta, \Sigma) \sim \text{NIW}(\mu, \tau, v, \Lambda)$ we mean the multivariate normal–inverse-Wishart distribution such that $\Sigma \sim IW(v, \Lambda)$ and $\theta \sim N_{qp}(\mu, \Sigma \otimes \tau^{-1})$, that is,

$$p(\theta | \Sigma) \propto |\Sigma|^{-(v+p+1)/2} \exp \left\{ - (1/2) \text{trace} \left( \Lambda \Sigma^{-1} \right) \right\},$$

$$p(\theta | \Sigma) \propto |\Sigma|^{-(q/2)} \tau^{p/2} \exp \left\{ - (1/2) \text{trace} \left( \theta - \mu \right)' \Sigma^{-1} \text{tr} \left( \theta - \mu \right) \right\}.$$

The regression model defined in (6)–(9) splits the sample into clusters of internally homogeneous individuals, as probabilistically described by the Potts model. In other words, the cluster formation is driven by the bonds, whose distribution is in turn determined by distances between the covariate vectors $|x_i|$. As a result, the model favors, a priori, clusters formed by individuals with similar covariate values. The model then implies common multiple multivariate regression parameters for individuals in the same cluster.

### 3.1 Posterior Simulation in the Regression Setting

To implement the Metropolis-Hastings posterior simulation scheme outlined in Section 2.2 and for posterior predictive inference we require explicit evaluation of (5). Note first that due to the multiplicative structure of the likelihood (8), we can factor the distribution is driven by the bonds, whose distribution is in turn a result, the model favors, a priori, clusters formed by individual with similar covariate values. The model then implies common multiple multivariate regression parameters for individuals in the same cluster.

$$\text{trace} \left( \Sigma_i^{-1} \sum_{i \in S_j} (y_i - I_p \otimes x_i' \theta_j) \right)$$

(10)
for \( j = 1, \ldots, k_n \). Let \( Y_j \in \mathbb{M}_{n \times p} \) and \( X_j \in \mathbb{M}_{n \times q} \) be the matrices whose rows are formed by the \( y_i \) and \( x_i \) vectors in \( y^*_j \) and \( x^*_j \), respectively. The proofs of the three results that follow can be found in the supplementary material.

**Lemma 1.** The least squares estimator of \( \theta^*_j \) is given by \( \hat{\theta}^*_j = \text{vec} \left( (X^*_j X_j)^{-1} X^*_j Y_j \right) \), for \( j = 1, \ldots, k_n \), where vec(\( A \)) denotes the vector formed by stacking the columns of the matrix \( A \). Moreover, the likelihood (10) can be written as

\[
(2\pi)^p |\Sigma_j|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \left( \Sigma_j^{-1} (X^*_j X_j) \right) (\theta^*_j - \hat{\theta}^*_j) \right\},
\]

where \( W_j = Y_j' (I_{n_j} - X_j (X_j X_j)^{-1} X_j') Y_j \).

The posterior distribution of \( (\theta^*_j, \Sigma_j^*) \) can be decomposed as

\[
P \left( \theta^*_j, \Sigma_j^* \mid y^*_j, x^*_j \right) = P \left( \theta^*_j \mid y^*_j, x^*_j \right) P \left( \Sigma^*_j \mid y^*_j, x^*_j \right).
\]

**Lemma 2.** For \( j = 1, \ldots, k_n \):

\[
\theta^*_j \mid y^*_j, x^*_j \sim N_{np}(\mu^{(j)}, \tau^{(j)}),
\]

where

\[
\mu^{(j)} = I_p \otimes (X^*_j X_j + \tau)^{-1} X^*_j \hat{\theta}^*_j + I_p \otimes (X^*_j X_j + \tau)^{-1} \mu_{mxs},
\]

\[
\tau^{(j)} = (X^*_j X_j + \tau)^{-1} (X^*_j X_j \hat{\theta}^*_{j,mxs} + \tau \mu_{mxs}),
\]

where for \( a \in \mathbb{R}^{np} \), \( a_{mxs} \) denotes the \( q \times p \) matrix derived by sequentially dividing \( a \) in \( p \) columns of dimension \( q \). In addition, \( \Sigma^*_j \mid y^*_j, x^*_j \sim IW(\nu + n_j, A + V_j) \), where \( V_j = (Y_j - X_j \mu_{mxs})' (Y_j - X_j \mu_{mxs}) + (\mu^{(j)}_{mxs} - \mu_{mxs})' \tau (\mu^{(j)}_{mxs} - \mu_{mxs}) \).

The results from Lemmas 1 and 2 are useful for posterior simulation. However, to evaluate (5) we need the marginal distribution \( p(Y^n \mid x^n, \rho_n) \), which follows from the following result.

**Lemma 3.** The marginal density of \( y^*_j, p(y^*_j \mid x^*_j) \), is given by

\[
\left( \frac{\Gamma_p \left( \frac{v + n_j}{2} \right)}{\pi^{n_j} \Gamma_p \left( \frac{v}{2} \right)} |A + V_j|^{|v + n_j|/2} \right) \times \left[ (2\pi)^{q(p-1)/2} \left( \frac{|\tau|}{|X^*_j X_j + \tau|} \right)^{p/2} \right],
\]

where \( V_j \) is given in the statement of Lemma 2, and \( \Gamma_p \) denotes the \( p \)-variate Gamma function, \( \Gamma_p(\xi) = \pi^{p(p-1)/4} \prod_{j=1}^p \Gamma(\xi + (1 - j)/2) \).

A direct consequence of Lemma 3 is the expression

\[
P \left( Y^n \mid x^n, \rho \right) = \frac{k_n}{\pi^{n_j} \Gamma_p \left( \frac{v + n_j}{2} \right)} \left| A + V_j \right|^{|v + n_j|/2} \times \left( \frac{|\tau|}{|X^*_j X_j + \tau|} \right)^{p/2},
\]

where \( V_{\ell} = (Y_{\ell} - X_{\ell} \mu_{m_{\ell,mxs}})(Y_{\ell} - X_{\ell} \mu_{m_{\ell,mxs}}) + (\mu^{(\ell)}_{mxs} - \mu_{mxs})' \tau (\mu^{(\ell)}_{mxs} - \mu_{mxs}) \), and \( \mu_{m_{\ell,mxs}} \) denotes the \( q \times p \) matrix derived by sequentially dividing \( \mu^{(\ell)}_{mxs} \) in \( p \) columns of dimension \( q \). Recall that \( \mu_{mxs} \) denotes the \( q \times p \) matrix derived by sequentially dividing \( \mu \) in \( p \) columns of dimension \( q \), and

\[
\mu^{(\ell)} = I_p \otimes (X_{\ell} X_{\ell} + \tau)^{-1} X_{\ell} X_{\ell} \hat{\theta} + I_p \otimes (X_{\ell} X_{\ell} + \tau)^{-1} \mu_{mxs}.
\]

where \( \hat{\theta} + \mu_{mxs} \) is the least squares estimator of \( \theta_{\ell} + \mu_{mxs} \) in \( X_{\ell} X_{\ell} + \tau \) is the least squares estimator of \( \theta_{\ell} + \mu_{mxs} \) in \( X_{\ell} X_{\ell} + \tau \).

In summary, our computational strategy for posterior simulation can be expressed as a Metropolis-within-Gibbs algorithm. Each scan consists of the following steps:

1. **Partition:** Let \( \rho_n(0) \) denote the currently imputed partition.
2. **Propose a new partition \( \rho_n(1) \) by applying steps 1 and 2 in Section 2.1.
3. **Accept \( \rho_n(1) \) with probability \( \min \{ p(Y^n \mid \rho_n(1), x^n) / p(Y^n \mid \rho_n(0), x^n) \} \), which can be evaluated in log-scale using (15).

**Parameters:** If \( k_n \) is the number of clusters of the currently imputed partition \( \rho_n = (\nu_1, \ldots, \nu_{k_n}) \), then for \( j = 1, \ldots, k_n \) draw new parameter values

\[
\Sigma^*_j \mid y^*_j, x^*_j \sim IW(\nu + n_j, A + V_j)
\]

and

\[
\theta^*_j \mid y^*_j, x^*_j \sim N_{qp}(\mu^{(j)}, \tau^{(j)}),
\]

where the corresponding distributions are specified in Lemma 2.

### 3.2 Posterior Predictive Inference

Another key element of our inferential effort is the implementation of posterior predictive inference. Let \( y_{n+1} \) denote a new individual and let \( x_{n+1} \) denote its associated covariate vector. The goal here is to find an expression for the predictive distribution \( P(Y_{n+1} \mid Y^n, x^n, x_{n+1}) \). This is done by conditioning on the partitions associated to the model. Let \( \rho_{n+1} \) be a partition of the
complete data \{\{y^n, y_{n+1}\}, \{x^n, x_{n+1}\}\}. We have
\[
P( y_{n+1} \mid y^n, x^n, x_{n+1}, \rho_{n+1} )
= P( y^n, y_{n+1} \mid x^n, x_{n+1}, \rho_{n+1} ) / P( y^n \mid x^n, x_{n+1}, \rho_{n+1} )
= |A + V_{\tilde{\ell}}(y_{n+1}, \rho_{n+1})|^{-(v+n_{\tilde{\ell}})/2}
/ \int_{W} |A + V_{\tilde{\ell}}(w, \rho_{n+1})|^{-(v+n_{\tilde{\ell}})/2} dw,
\]
where \(\tilde{\ell}\) denotes the index of the cluster \(S_{\tilde{\ell}}\) in \(\rho_{n+1}\) that includes the new individual \(n+1\). We have used the notation \(V_{\tilde{\ell}}(w, \rho_{n+1})\) to stress the dependency of \(V_{\tilde{\ell}}\) on \(\rho_{n+1}\) and \(w\). Note that the denominator of the above formula (which plays the role of a normalizing constant) depends on \(\rho_{n+1}\) as well. The posterior predictive distribution is given by
\[
P( y_{n+1} \mid y^n, x^n, x_{n+1} )
= \sum_{\rho_{n+1} \in A(\rho_{n+1})} |A + V_{\tilde{\ell}}(y_{n+1}, \rho_{n+1})|^{-(v+n_{\tilde{\ell}})/2}
/ \int_{W} |A + V_{\tilde{\ell}}(w, \rho_{n+1})|^{-(v+n_{\tilde{\ell}})/2} dw
\times p( \rho_{n+1} \mid y^n, x^n, x_{n+1}).
\]

Note that we have added a new point to the partition, so the \(n_j\)'s denote now cluster sizes associated with \(\rho_{n+1}\). Next, let \(\rho_{n}\) denote a posterior distribution of the data. Let \(A(\rho_{n+1})\) denote the collection of all partitions \(\rho_{n}\) giving rise to \(\rho_{n+1}\), by generating appropriate bonds \(b_{n+1}\) from \(x_{n+1}\) in the graph. Inversely, one can think of \(\rho_{n}\) as a partition of the original data that arises by marginalizing out all the bonds from \(x_{n+1}\). We have
\[
P( \rho_{n+1} \mid y^n, x^n, x_{n+1} )
= \sum_{\rho_{n} \in A(\rho_{n+1})} p( \rho_{n} \mid y^n, x^n ) p( \rho_{n+1} \mid \rho_{n}, x^n, x_{n+1} )
= \sum_{\rho_{n} \in A(\rho_{n+1})} p( \rho_{n} \mid y^n, x^n ) \sum_{b_{n+1} \rightarrow x_{n+1} \rho_{n+1}} p( b_{n+1} \mid \rho_{n}, x^n, x_{n+1} ).
\]

Now let \([\rho_{n}, b_{n+1}]\) denote a partition of \(x^n \cup \{x_{n+1}\}\) generated from \(\rho_{n}\) by the bonds \(b_{n+1}\) associated to \(x_{n+1}\). The above equation allows us to write the posterior predictive distribution \(P(y_{n+1} \mid y^n, x^n, x_{n+1})\) as
\[
\sum_{\rho_{n}} \sum_{b_{n+1}} p(\rho_{n} \mid y^n, x^n) p(b_{n+1} \mid \rho_{n}, x^n, x_{n+1})
\times \frac{|A + V_{\tilde{\ell}}(y_{n+1}, [\rho_{n}, b_{n+1}])|^{-(v+n_{\tilde{\ell}})/2}}{\int_{W} |A + V_{\tilde{\ell}}(w, [\rho_{n}, b_{n+1}])|^{-(v+n_{\tilde{\ell}})/2} dw}.
\]

Therefore, given a posterior sample of partitions \([\rho_{n,j} : j = 1, \ldots, M]\) and bonds \([b_{j+1} : j = 1, \ldots, M]\), we can estimate the posterior prediction by means of
\[
\hat{P}( y_{n+1} \mid y^n, x^n, x_{n+1} )
\defeq \frac{\sum_{j=1}^{M} |A + V_{\tilde{\ell}}(y_{n+1}, [\rho_{n,j}, b_{j+1}])|^{-(v+n_{\tilde{\ell}})/2}}{\int_{W} |A + V_{\tilde{\ell}}(w, [\rho_{n,j}, b_{j+1}])|^{-(v+n_{\tilde{\ell}})/2} dw}.
\]

In practice, the application of the predictive inference approach discussed above requires a method to deal with the integrals in the denominator of (16). We show that such integrals may be well approximated by expressions that can be resolved analytically.

**Proposition 1.** Let \(V_{\tilde{\ell}}(\rho_{n}) = V_{\tilde{\ell}}\) be as defined in Lemma 2 (the dependency on the partition \(\rho_{n}\) involving only the data is stressed). Let \((w, x_{n+1})\) be a new individual in cluster \(\tilde{\ell}\). Then the updated mean for this cluster is \(\mu_{\tilde{\ell}}^{(p)} = (n_{\tilde{\ell}} \mu_{\tilde{\ell}}^{(j)} + w)/(n_{\tilde{\ell}} + 1)\). Suppose that the mean \(\mu_{\tilde{\ell}}^{(p)} \approx \mu_{\tilde{\ell}}^{(j)}\), that is, suppose that either \(n_{\tilde{\ell}}\) is large or that \(w \approx \mu_{\tilde{\ell}}^{(j)}\). Then we have the approximation
\[
\int_{W} |A + V_{\tilde{\ell}}(w, [\rho_{n}, b_{n+1}])|^{-(v+n_{\tilde{\ell}})/2} dw
\approx \pi r^{2} \frac{\Gamma\left(\frac{v+n_{\tilde{\ell}}}{2}\right)}{(\Gamma\left(\frac{v}{2}\right) + \Gamma\left(\frac{n_{\tilde{\ell}}}{2}\right))}
\times \left[1 + (y_{n+1} - x_{n+1}^{\mu_{\tilde{\ell}}^{(j)}})^{\mu_{\tilde{\ell}}^{(j)}}\right]^{-1/2}
\times \left[(y_{n+1} + x_{n+1}^{\mu_{\tilde{\ell}}^{(j)}})^{\mu_{\tilde{\ell}}^{(j)}}\right]^{-(v+n_{\tilde{\ell}})/2}.
\]

The proof of Proposition 1 is given in the Supplementary material. The result is useful to generate a Monte Carlo approximation to the posterior predictive distribution of a new individual. Given a posterior sample of partitions \([\rho_{n,j} : j = 1, \ldots, M]\) and bonds \([b_{j+1} : j = 1, \ldots, M]\), compute
\[
\hat{P}( y_{n+1} \mid y^n, x^n, x_{n+1} ) \approx \sum_{j=1}^{M} \frac{\Gamma\left(\frac{v+n_{\tilde{\ell}}}{2}\right)}{(\Gamma\left(\frac{v}{2}\right) + \Gamma\left(\frac{n_{\tilde{\ell}}}{2}\right))}
\times \left[1 + (y_{n+1} - x_{n+1}^{\mu_{\tilde{\ell}}^{(j)}})^{\mu_{\tilde{\ell}}^{(j)}}\right]^{-1/2}
\times \left[(y_{n+1} + x_{n+1}^{\mu_{\tilde{\ell}}^{(j)}})^{\mu_{\tilde{\ell}}^{(j)}}\right]^{-(v+n_{\tilde{\ell}})/2}.
\]

Here, we have used the determinant identity \(|A + bb'\| = |A|(1 + b' A^{-1} b)|\), which holds for any invertible \(p \times p\) matrix \(A\) and any \(p\)-dimensional vector \(b\). Approximation (17) is implemented as an extension to the posterior simulation algorithm laid out in Section 3.1. An extra step is added to the algorithm to sample the additional bonds required to include \(x_{n+1}\) in the augmented partition \(\rho_{n+1}\). Expression (17) is evaluated after collecting a sample of size \(M\) of such partitions and bonds.

### 3.3 Posterior Inference on Partitions

The posterior distribution for the number of clusters can be easily obtained from MCMC samples, as each of these involves a partition of the data. To produce a meaningful posterior estimate consisting of a single partition, we use the proposal by Murua, Stanberry, and Stuetzle (2008). The method starts by finding the posterior mode of the number of clusters, and then builds a consensus clustering from all the MCMC sample partitions with this given number of clusters. The proportion of times a given pair of individuals appears together in the same cluster within a given partition is calculated for every pair of neighbors of the original data graph. The consensus clustering is obtained as the connected components of the graph whose edges consist of those neighbors with a positive proportion. In general, one could prune the edges with an associated proportion smaller than certain predetermined threshold (e.g., 0.5). Alternatively, Murua and Wicker (2014) set the threshold to the mode of the posterior density associated with the probability of having two given neighbors in the same cluster. We have adopted this latter strategy. We also have tried the least-squares clustering
(LSCLUST) procedure of Dahl (2006). This method is also based on pairwise proportions, but now computed using the entire set of posterior MCMC samples. The estimated partition is chosen as the minimizer of the sums of squares of the component-wise differences between these proportions and the individual co-clustering indicators among sampled partitions. For more details see Dahl (2006). In all the examples we tried, LSCLUST and consensus clustering gave the same results.

4. Simulation Studies

In this section we study through simulations the performance of the Potts regression model and compare it with that of three other alternative methods: FLEXMIX, as implemented in the homonymous R package (Leisch 2004; Grün and Leisch 2007), the PPMx approach by Müller, Quintana, and Rosner (2011), and Mixtures of Experts (McLachlan and Peel 2001). For the PPMx and mixtures of experts approaches we keep the same likelihood model (6) but change the proposed Potts model prior to a dependent PPM model in the former case or to a dependent mixture in the latter case. We describe three different simulation scenarios. Two of them relate directly to the multiple regression model in Section 3. In the last scenario we investigate the robustness and performance of the Potts regression model in a setting that is totally different from the one that motivated it. First we concentrate on the two simulation scenarios involving multiple regression.

Scenario I: The odd shape cluster dataset was conceived to see if the methods considered here can recover the arbitrary cluster structure of the covariates. It consists of four two-dimensional covariate groups and one-dimensional response variable. Groups 1 and 2 were generated as uniform draws around a couple of arcs; Groups 3 and 4 are Gaussian clumps. The groups can be seen in the left panel of Figure 2. The arcs were generated with 1000 points; the clumps were generated with 500 points each. The response variable Y corresponds to Gaussian values with means equal to group-dependent linear combinations of Y1 and Y2. The group-dependent linear combinations are 3Y1 + 2Y2, Y1 + 3Y2, −3Y1, and −3Y1 − 2Y2, respectively. The corresponding regression standard deviations are 0.03, 0.05, 0.04, and 0.05, respectively. The densities associated to Y for each group are displayed in the right panel of Figure 2.

Scenario II: The five-copula cluster dataset was conceived to study the performance of our method on a multivariate response data similar to the one studied in the application discussed in Section 5. It was generated using a mixture of five six-dimensional copulas as covariates and a two-dimensional mixture-component-dependent response variable. The copulas were generated from a multivariate Gaussian copula with compound symmetry and marginals set to normal, exponential, gamma, normal, gamma, and exponential. The components differ in the marginal distribution parameters. Details on the method used to generate the samples are given in the supplementary material.

To fit the models, we randomly generated 100 subsamples of size 200 from each of these datasets. We applied our procedure to each sample, and computed the associated root mean square error (RMSE). We also created an independent sample of size 100 to estimate the goodness of fit of the models fitted with each of the 100 training datasets. We report both the RMSE associated with the training data and the RMSE associated with the test data. Note that the RMSE may be computed in two different ways when using the Potts regression model. First, the posterior mean response for each covariate value can be estimated as the mean of the estimated posterior predictive density in (17). Alternatively, we can compute the mean of the corresponding sample values (associated with the covariates) generated from the MCMC output. In what follows we consider both methods. We ran our method with 20,000 burn-in iterations, but only kept 5000 samples after burn-in to do the data analysis. From preliminary computational experiments, we judged this simulation size to be sufficient for the target posterior and predictive inference.

For each compared procedure and each dataset used in the simulation, we computed the average of the 100 RMSE. The results are shown in Table 1 for the odd-shape-clusters dataset, and in Table 2 for the five-copula one. We also computed the adjusted Rand index (ARI) (Rand 1971; Hubert and Arabie

Table 1. Odd shape clusters. The ARI columns show the mean, upper and lower quartiles of the ARI computed using 100 samples of size 200. The numbers in parenthesis associated with the Mixture of Experts refer to the number of hidden units in the mixture. The darker cells correspond either to the lowest RMSE for training or testing, and to the highest ARI with respect to the covariates clustering structure.

| Method                         | RMSE Training | ARI (Mean) | ARI (LQ) | ARI (UQ) |
|-------------------------------|---------------|------------|----------|----------|
| Potts Regression (Density)     | 0.398         | 0.698      | 0.565    | 0.823    |
| Potts Regression (Means)       | **0.240**     | 0.698      | 0.565    | 0.823    |
| FLEXMIX                       | 1.987         | 0.515      | 0.438    | 0.537    |
| PPMx                          | 0.835         | 0.615      | 0.468    | 0.859    |
| Mixture of Experts (10)        | 0.581         | 0.768      | 0.582    | 0.893    |
| Mixture of Experts (6)         | 0.951         | **0.151**  | 0.764    | 0.801    |
| Mixture of Experts (5)         | 1.037         | 0.693      | 0.400    | 0.872    |
| Mixture of Experts (4)         | 1.000         | 0.663      | 0.403    | 0.866    |
| Mixture of Experts (3)         | 1.203         | 0.570      | 0.379    | 0.684    |

Figure 2. Odd shape data: Four groups of covariates (left panel) and the corresponding simulated response variable densities (right panel).
between the estimated and true data partitions. High ARI values indicate good agreement between the partitions. The ARI takes its highest value of 1.0 when the partitions being compared are identical.

The mixture of experts greatly overfits the data. Its performance is very dependent on the number of hidden units chosen to fit the data. However, it seems to recover well the clustering structure in the covariates. An explanation for this behavior might be that the mixture of experts considers too many parameters. This makes it very flexible so as to capture the clustering structure of the covariates. However, those many parameters also overfit the data, yielding poor predictions. The PPMx-type prior shows a somewhat better performance. It achieves a good partition estimation. However, some of the overfitting trend is also present, although not to the extent seen in the mixtures of experts models. On the other hand, FLEXMIX performs poorly on these datasets. We stress that none of the datasets in this simulation study fits exactly the assumptions of the models tested. It appears that FLEXMIX does not generalize easily to complex datasets.

We also consider a third scenario, given by the survival data analysis application described by Müller, Quintana, and Rosner (2011). In this case, we need to change the likelihood model to just a univariate normal distribution that does not involve covariate dependence. To compare Potts regression with PPMx, we use the same simulated data described in Table 1 of Müller, Quintana, and Rosner (2011). The competing models are defined as the univariate likelihood model described above, and either the Potts regression or PPMx priors. We find that PPMx has a very slightly better performance, and that density-based estimation of the mean performs somewhat worse than the other alternatives. However, as in the previous scenarios, the Potts regression model yields better results for the prediction cases. Due to lack of space, further details for this scenario are found in the supplementary material.

In summary, the Potts regression model seems to be the most reliable method for local regression among the methods compared in this simulation study. It shows the best performance in terms of testing (predictive) RMSE while doing a reasonable job in revealing the cluster structure in the covariates. Note that among the methods compared, the current implementation of FLEXMIX does not support multiple responses.

5. Data Illustration

On the eve of the 21st century, nations showed a wide variation in their health status. A lot of research has focused on developing policies that could directly affect health outcomes (Hegyvary, Berry, and Murua 2008; Wilkinson 1996; Collins 2003; Deaton 2002; Gatrell 2002; Poland et al. 1998; Dollar 2001). For illustrative purposes only, we consider here a sample from the cross-sectional analysis carried out by Hegyvary, Berry, and Murua (2008). It consists of 134 member countries of the United Nations, representing a wide range of possible economic and demographic conditions in the world. Data on diverse factors were gathered by Hegyvary, Berry, and Murua (2008) from the World Bank, the United Nations, the U.S. Central Intelligence Agency, and others. Out of the 18 factors considered by Hegyvary, Berry, and Murua (2008), we selected six factors as covariates: Secondary Education (SE), Human Resources (HR), Population Density (PDn), Population Dependency (PDp), Birth Rate (BR), and the level of Democracy (DL); and two factors as our response variable (the same two chosen by Hegyvary, Berry, and Murua (2008)): (1) Adjusted life expectancy (ALE), which is the average of the life expectancy and health adjusted life expectancy variables, and (2) Child mortality (CM), which corresponds to the under-five mortality rate. The choice of factors in our study was mainly driven by the availability of complete data (that is, no factors with missing values were selected). All factors were normalized to have mean zero and unit variance. More details on this dataset can be found in Hegyvary, Berry, and Murua (2008).

The Potts regression model yielded a RMSE of 0.266, and favored seven clusters. The three largest ones contain 45 (Cluster 4), 38 (Cluster 2), and 32 (Cluster 3) countries, respectively. The other clusters are small, containing 8 (Cluster 5), 5 (Cluster 6), and 3 (Clusters 1 an 7) countries. Cluster 4 contains most of the richest countries in the world which show moderate to high ALE values, and very low CM values. Cluster 2 contains the poorest countries. These presented very high CM values at the turn of the century. Cluster 3 contains most of the developing countries in Latin America and Asia. These presented mid-range values for CM and high ALE values. The list of countries making up each cluster is given in the supplementary material. The clusters on the ALE - CM space are displayed in the top image of Figure 3. For comparison purposes we ran a mixture of experts model on these data, using the same general approach as in Section 4. The best model included 4 experts, and yielded a RMSE of 0.426. The mixture of experts model chose two clusters. Basically, Clusters 2

| Table 2. Five copula clusters with the 2D response. |
|--------------------------------------------------|
| RMSE | Testing | Training | ARI (Mean) | ARI (LQ) | ARI (UQ) |
| Potts Regression (Density) | 36.426 | 19.789 | 0.695 | 0.610 | 0.780 |
| Potts Regression (Means) | 35.271 | 4.593 | 0.695 | 0.610 | 0.780 |
| PPMx | 40.957 | 0.245 | 0.992 | 0.994 | 1.000 |
| Mixture of Experts (10) | 47.280 | 0.223 | 0.985 | 0.986 | 1.000 |
| Mixture of Experts (6) | 47.444 | 0.271 | 0.963 | 0.976 | 1.000 |
| Mixture of Experts (5) | 47.487 | 0.284 | 0.955 | 0.975 | 1.000 |
| Mixture of Experts (4) | 47.357 | 0.900 | 0.770 | 0.744 | 0.797 |
| Mixture of Experts (3) | 46.515 | 2.386 | 0.545 | 0.474 | 0.609 |

Figure 3. Health data. The seven clusters favored by the Potts regression model.
and 3 are merged in the mixture of experts (see the bottom image in Figure 3). We also tried the PPMx approach, which gave a RMSE of 0.343 and 5 clusters according to the LSCLUST procedure. The results yielded by the Potts regression model are more in agreement with those found by Hegyvary, Berry, and Murua (2008). However, our results hint at a larger Cluster 4 than the one found in their work. In fact, our model puts together the ex-soviet countries with the richest countries. At the turn of the century, the relation between the health status of the ex-soviet countries with the demographic, education, and health system factors considered in this study was comparable to that of the richest western countries.

Of the six covariates considered, we took a closer look at BR and HR. These are variables that policy makers could and may want to control to improve the health status of a population. Figure 4 displays contours of the posterior predictive density associated with different levels of BR and HR while keeping the other four variables fixed at their overall observed median values. Each panel of this figure shows contours in ALE-CM space associated with every combination of the observed quantiles 0.1, 0.25, 0.50, 0.75, and 0.90 of the BR and HR variables. Observe that the “median” state of the health status of these 134 countries is right in the middle (the variables have been standardized). Not surprisingly, the overall pattern is that lower BR values and larger amounts of HR are associated with better health. But there are exceptions. For instance, while most countries in Cluster 4 showed a low birth rate, Israel had a relatively high birth rate and showed very good health outcomes. Note also that the best health outcomes are not necessarily associated with the highest amount of HR (see the top-left panels in column 2 and rows 1 and 2). For example, The United Kingdom, Cyprus, Greece, Portugal, and Japan presented very good health outcomes, even though their HR levels were among the lowest in Cluster 4. HR levels do make a difference in countries with a high birth rate, specially for lowering child mortality rates. Take, for instance, Cluster 3 which is composed of countries in the mid range of health outcomes. Examination of the corresponding posterior predictive density (contour plots are shown in the supplementary material) reveals that moving only a single variable, BR or HR, would have little impact in health outcomes. A noticeable change in health outcomes requires a simultaneous movement in both variables. This implies an interaction effect between these variables, suggesting that policy makers should not concentrate in single factors, but on several of them simultaneously. Contours for the other two large groups, Clusters 2 and 4, show no major changes, with “realistic” within-cluster changes in the levels of BR or HR. This is to be expected since these countries are located at either low or high extreme levels for both variables.

6. Discussion

The main purpose of this work is the use of Potts model as a covariate-dependent prior on partitions. The discussion focused on a particular example (multivariate multiple linear regression) as a concrete illustration, but the approach is quite general and can be extended to be used with literally any likelihood model of interest. This would also require extending our computational approach to the specific model assumptions adopted.

In our examples, the covariates were all of continuous type, but this is not to be seen as a restriction of the method. Any type of covariate can be used, as the kernel function $K(\cdot, \cdot)$ can be easily modified to account for such changes. It is important though to introduce some notion of standardization in the kernel values so as to mitigate the effect of different scales and/or variabilities in the covariate values.

Our simulation and data analysis results suggest that the Potts prior model has a superior performance in the cases we tried, with both synthetic and real data. The only exception we found to this trend is where the covariate dependence was restricted to the prior on partitions, in which case the PPMx and Potts regression have almost identical performance in the simulation.

A limitation of the algorithm outlined in Section 2.2 is that it requires analytical marginalization of $\psi_\rho$ to derive $p(Y^n | \rho', x^d)$, which is needed to compute (5). In the case
of local multiple regression, this equation can be calculated explicitly. We show that a clever approximation to the posterior predictive density makes the procedure very efficient and highly competitive with other flexible regression methods. When \( p(\psi_{\varphi} \mid \rho', Y^*, x^*) \) is intractable, one may need to consider other Metropolis-Hastings proposals. However, as noted in Section 2.2, no evident simplification of the acceptance ratio arises.

**Acknowledgments**

The authors thank all the reviewers for their valuable comments which helped improve the manuscript. The first author was partially supported by the NSERC grant 327689-06. The second author was partially funded by grant FONDECYT 1141057. The authors would like to thank Professor Devon Berry (College of Nursing and Health, Wright State University), and Professor Emerita Sue Hegyvary (School of Nursing, University of Washington) for facilitating us the data on health outcomes.

**References**

Argiento, R., Cremaschi, A., and Guglielmi, A. (2014), “A ‘Density-Based’ Algorithm for Cluster Analysis Using Species Sampling Gaussian Mixture Models,” *Journal of Computational and Graphical Statistics*, 23, 1126–1142. [265]

Atchade, Y. F., and Liu, J. S. (2010), “The Wang-Landau Algorithm in General State Spaces: Applications and Convergence Analysis,” *Statistica Sinica*, 20, 209–233. [266]

Barry, D., and Hartigan, J. A. (1992), “Product Partition Models for Change Point Problems,” *The Annals of Statistics*, 20, 260–279. [265]

Besag, J. (2004), “Markov Chain Monte Carlo Methods for Statistical Inference,” Technical Report, Department of Statistics, University of Washington, Seattle, Washington. [267]

Blatt, M., Wiseman, S., and Domany, E. (1996), “Superparamagnetic Clustering of Data,” *Physical Review Letters*, 76, 3251–3254. [266]

Collins, T. (2003), “Globalization, Global Health, and Access to Healthcare,” *International Journal of Health Planning and Management*, 18, 97–104. [272]

Dahl, D. B. (2006), “Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model,” in *Bayesian Inference for Gene Expression and Proteomics*, eds. K.-A. Do, P. Müller, and M. Vannucci, New York: Cambridge University Press. (2008), “Distance-Based Probability Distribution for Set Partitions With Applications to Bayesian Nonparametrics,” in *JSM Proceedings*, Section on Bayesian Statistical Science, Alexandria, VA: American Statistical Association. [265]

Deaton, A. (2002), “Policy Implications of the Gradient of Health and Wealth,” *Health Affairs*, 21, 13–30. [272]

Dollar, D. (2001), “Is Globalization Good for Your Health?” *Bulletin of the World Health Organization*, 79, 827–833. [272]

Dunson, D. B., Pillai, N., and Park, J. -H. (2007), “Bayesian Density Regression,” *Journal of the Royal Statistical Society, Series B*, 69, 163–183. [265]

Ferguson, T. S. (1973), “A Bayesian Analysis of Some Nonparametric Problems,” *Annals of Statistics*, 1, 209–230. [265]

Gatrell, A. C. (2002), *Geographies of Health*, Oxford: Blackwell. [272]

Girolami, M. (2002), “Mercer Kernel-Based Clustering in Feature Space,” *Neural Networks, IEEE Transactions on*, 13, 780–784. [266]

Green, P. J. (1995), “Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination,” *Biometrika*, 82, 711–732. [267]

Grün, B., and Leisch, F. (2007), “Fitting Finite Mixtures of Generalized Linear Regressions in R,” *Computational Statistics & Data Analysis*, 51, 5247–5252. [271]

Hartigan, J. A. (1990), “Partition Models,” *Communications in Statistics, Part A—Theory and Methods*, 19, 2745–2756. [265]

Hegyvary, S., Berry, D., and Murua, A. (2008), “Country Clustering to Evaluate Global Health Outcomes,” *Journal of Public Health Policy*, 29, 319–339. [272]

Hubert, L., and Arabie, P. (1985), “Comparing Partitions,” *Journal of Classification*, 2, 193–218. [272]

Leisch, F. (2004), “FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R,” *Journal of Statistical Software*, 11, 1–18. [271]

Lodewyckx, T., Kim, W., Lee, M., Tuerlinckx, E., Kuppens, P., and Wagenmakers, E. (2011), “A Tutorial on Bayes Factor Estimation With the Product Space Method,” *Journal of Mathematical Psychology*, 55, 331–347. [268]

MacEachern, S. N. (1999), “Dependent Nonparametric Processes,” in *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association. [265]

——— (2000), “Dependent Dirichlet processes,” *Technical Report*, Department of Statistics, The Ohio State University. [265]

McLachlan, G. J., and Peel, D. (2001), *Finite Mixture Models*, New York: Wiley. [271]

Müller, P., and Quintana, F. (2010), “Random Partition Models With Regression on Covariates,” *Journal of Statistical Planning and Inference*, 140, 2801–2808. [266]

Müller, P., Quintana, F. A., and Rosner, G. L. (2011), “A Product Partition Model with Regression on Covariates,” *Journal of Computational and Graphical Statistics*, 20, 260–278. [265,266,271,272]

Murua, A., Stanberry, L., and Stuetzel, W. (2008), “On Potts Model Clustering, Kernel K-Means, and Density Estimation,” *Journal of Computational and Graphical Statistics*, 17, 629–658. [266,270]

Murua, A., and Wicker, N. (2014), “The Conditional-Potts Clustering Model,” *Journal of Computational and Graphical Statistics*, 23, 717–739. [266,270]

Ott, T., Kern, A., Steeb, W. -H., and Stoop, R. (2005), “Sequential Clustering: Tracking Down the Most Natural Clusters,” *Journal of Statistical Mechanics: Theory and Experiment*, P11014. [266]

Park, J. -H., and Dunson, D. B. (2010), “Bayesian Generalized Product Partition Model,” *Statistics Sinica*, 20, 1203–1226. [265]

Poland, B., Coburn, D., Robertson, A. and Eakin, J. (1998), “Wealth, Equity and Health Care: A Critique of a ‘Population Health’ Perspective on the Determinants of Health,” *Social Science & Medicine*, 46, 785–798. [272]

Quintana, F. A. (2006), “A Predictive View of Bayesian Clustering,” *Journal of Statistical Planning and Inference*, 136, 2407–2429. [265]

Rand, W. M. (1971), “Objective Criteria for the Evaluation of Clustering Methods,” *Journal of the American Statistical Association*, 66, 846–850. [271]

Reichardt, J., and Bornholdt, S. (2004), “Detecting Fuzzy Community Structures in Complex Networks With a Potts Model,” *Physical Review Letters*, 93, 218701. [266]

Sokal, A. (1997), “Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms,” in *Functional Integration (Cargèse, 1996)*, Plenum, New York, vol. 361 of NATO Advanced Science Institutes Series B: *Physics*, pp. 131–192. [266,267]

Stanberry, L., Murua, A., and Cordes, D. (2008), “Functional Connectivity Mapping Using the Ferromagnetic Potts Spin Model,” *Human Brain Mapping*, 29, 422–440. [266]

Swendsen, R. H., and Wang, J.-S. (1987), “Nonuniversal Critical Dynamics in Monte Carlo Simulations,” *Physical Review Letters*, 58, 86–88. [267]

Tokdar, S. T., Zhu, Y. M. , and Ghosh, J. K. (2010), “Bayesian Density Regression With Logistic Gaussian Process and Subspace Projection,” *Bayesian Analysis*, 5, 319–344. [265]

Wang, F., and Landau, D. P. (2001), “Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States,” *Physical Review Letters*, 86, 2050–2053. [266]

Wilkinson, R. G. (1996), *Unhealthy Societies: The Afflictions of Inequality*, London: Routledge. [272]