Density Regression using Repulsive Distributions

José Quinlan
Departamento de Estadística
Pontificia Universidad Católica de Chile
jjquinla@mat.uc.cl

Fernando A. Quintana
Departamento de Estadística
Pontificia Universidad Católica de Chile
quintana@mat.uc.cl

Garritt L. Page
Department of Statistics
Brigham Young University
page@stat.byu.edu

June 12, 2017

Abstract

Flexible regression is a traditional motivation for the development of nonparametric Bayesian models. A popular approach for this involves a joint model for responses and covariates, from which the desired result arises by conditioning on the covariates. Many such models involve the convolution of a continuous kernel with some discrete random probability measure defined as an infinite mixture of i.i.d. atoms. Following this strategy, we propose a flexible model that involves the concept of repulsion between atoms. We show that this results in a more parsimonious representation of the regression than the i.i.d. counterpart. The key aspect is that repulsion discourages mixture components that are near each other, thus favoring parsimony. We show that the conditional model retains the repulsive features, thus facilitating interpretation of the resulting flexible regression, and with little or no sacrifice of model fit compared to the infinite mixture case. We show the utility of the methodology by way of a small simulation study and an application to a well known dataset.

Key Words: Dependent Gaussian mixture models, regression estimation, Gibbs measures, point processes.

1 Introduction

In Bayesian parametric and nonparametric hierarchical models, a common assumption is that parameters (atoms) in the latent level of a hierarchy are assumed to be mutually independent and originating from a common distribution. This is true for finite and infinite mixture models where latent parameters correspond to component (cluster) locations; see, for example, Frühwirth-Schnatter (2006) and Hjort et al. (2010). A consequence of the independence assumption is the creation of redundant clusters in the sense that cluster centers can be very close in space. Often times this can result in making the models unnecessarily complex, leading to overfitting. This in turn can negatively impact out-of-sample prediction and model interpretability. To counteract this, Quinlan et al. (2017) defined a class of probability distributions whose coordinates are encouraged to be mutually separated (a property referred to as *repulsion*). Further, they showed how the class of repulsive distributions can be employed to model component location parameters in a finite Bayesian Gaussian mixture when carrying out density estimation. Therefore, cluster locations are not modeled independently, but rather are encouraged to repel each other producing a more parsimonious mixture model. The key component of the probability law that produces repulsion is that small relative distances between the centers of the mixture components are penalized by way of a single parameter that controls the strength of the repulsion. The class of distributions proposed by Quinlan et al. (2017) is based on (finite) Gibbs Point Processes (Illian et al. 2008).

Although not considered in Quinlan et al. (2017), it is common in many studies for researchers to collect additional covariate information on each experimental unit or subject. This is the case in the well known application that we consider in Section 4. These data consist of duration times of Old Faithful geyser eruptions and the waiting time until the next eruption occurs. Interest lies in being able to learn how time until the next eruption influences eruption duration. There are a number of methods developed in Bayesian nonparametrics that are available to model such data. Among them are approaches classified as nonparametric residual distributions, nonparametric mean functions, and fully nonparametric regression which is often times referred to as Bayesian density regression (Dunson et al.

2007). For more details and references, we direct the reader to Chapter 4 of Müller et al. (2015). A motivation for considering these methods is the desire to flexibly accommodate arbitrarily shaped mean curves associated with the distribution of the response given a covariate. However, flexibility comes at a cost as it is common that a large number of clusters are created to carry this out, many of which are redundant. We extend the notion of a joint repulsive distribution to incorporate the information contained in covariates, which we use here to assist the formation of clusters. Even though other approaches for explicit repulsive distributions exist (see Petralia et al. 2012; Fúquene et al. 2016; Xie and Xu 2017), they have not yet extended them to include additional covariate information. We also mention that alternative approaches to introduce repulsion in mixture models involve the introduction of determinantal point processes Xu et al. (2016), approach that was recently extended in Bianchini et al. (2017), also including covariates in the prior distribution of cluster assignment. Their approach differs from ours though in the way the repulsion is modeled, and how the covariates are used to define this repulsion.

Because directly making the class of repulsive distributions covariate dependent renders computation intractable, our approach is similar to what was done in Quinlan et al. (2017). Specifically we incorporate covariates in a finite Bayesian Gaussian mixture model by modeling responses and covariates jointly, and then model component centers (e.g. locations) with a repulsive distribution.

The remainder of this article is organized as follows: in Section 2 we provide a concise description of dependent Gaussian finite mixture models for continuous covariates and discuss a novel Bayesian approach in which repulsion is introduced at a latent level in the joint distribution for responses and covariates (location parameters). We also provide guidance (similar to that found in Quinlan et al. 2017) to calibrate the hyperparameters of our model. Details associated with posterior sampling are also provided in this section. Sections 3 and 4 illustrate the performance of our proposal applied to synthetic and real data sets.

Computational strategies are provided in Appendix A.

2 Covariate Dependent Repulsive Gaussian Mixture Model (RGMMx)

Consider $n \in \mathbb{N}$ experimental units where on each the orderer pairs $(\boldsymbol{y}_1, \boldsymbol{x}_1), \ldots, (\boldsymbol{y}_n, \boldsymbol{x}_n)$ are recorded (to simplify the notation we use $[m] = \{1, \ldots, m\}$ with $m \in \mathbb{N}$). In this case, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{R}^d$ are the responses and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ are the corresponding subject-specific covariates. The motivation for considering $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is that they provide additional information regarding the distribution of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. It is common to assume mutual independence among the responses. However, assuming that responses are identically distributed is not tenable because of their dependence on covariates. The challenge is to model how the covariates guide the evolution of the mean response. Taking on a nonparametric Bayes approach permits modeling this covariate dependent evolution in a very flexible way.

The fundamental idea in the context of nonparametric Bayesian regression models is to estimate the average behavior of a response variable $\boldsymbol{y} \in \mathbb{R}^d$ as an (unknown) function of available covariates $\boldsymbol{x} \in \mathbb{R}^p$ for some $d, p \in \mathbb{N}$. Müller and Quintana (2004) provide a nice overview that details a number of possible Bayesian nonparametric regression approaches. Among these is the approach of Müller et al. (1996) which we adopt. With the flexibility of Gaussian mixture models to emulate smooth densities accurately in mind, they propose a statistical model that reduces regression estimation to a density estimation problem. Specifically they treat $\boldsymbol{u} = (\boldsymbol{y}, \boldsymbol{x}) \in \mathbb{R}^d \times \mathbb{R}^p$ as a random vector generated by a Dirichlet process Gaussian mixture model (DPM). The joint distribution is used to estimate the regression mean curve through the implied conditional distribution $(\boldsymbol{y} \mid \boldsymbol{x})$. To make these ideas concrete, consider the hierarchical model

$$\boldsymbol{u}_i \mid (\boldsymbol{\theta}_i, \boldsymbol{\Lambda}_i) \overset{ind.}{\sim} \mathrm{N}_{d+p}(\boldsymbol{\theta}_i, \boldsymbol{\Lambda}_i) \tag{1}$$

$$(\boldsymbol{\theta}_i, \boldsymbol{\Lambda}_i) \mid H \overset{i.i.d.}{\sim} H \tag{2}$$

$$H \mid \alpha, H_0 \sim \mathrm{DP}(\alpha, H_0), \tag{3}$$

where the $\mathrm{DP}(\alpha, H_0)$ denotes a Dirichlet process with base measure H_0 (which is often times selected to be the conjugate Normal-Inverse-Wishart) and dispersion parameter $\alpha \in (0, \infty)$. The base measure H_0 may itself have additional hyperparameters, with their corresponding hyperpriors. Being that $\mathrm{DP}(\alpha, H_0)$ is a discrete random probability measure (see Sethuraman 1994), Müller et al. (1996) show that the posterior predictive conditional density for $(\boldsymbol{y} \mid \boldsymbol{x})$ takes the form of a locally weighted mixture of linear regressions, also known as weight dependent Dirichlet process (WDDP). For more technical and computational details see Müller et al. (2015) and Jara et al. (2011).

In order to capture flexible mean structures the above proposal tends to produce a large number of covariate dependent clusters, many of which are redundant. To make the models more parsimonious, we propose a straightforward method similar to WDDP that introduces repulsion in the location parameters of the joint distribution for $\boldsymbol{u} = (\boldsymbol{y}, \boldsymbol{x})$. Thus, we consider $\boldsymbol{x}$ as a random quantity lying in $\mathbb{R}^p$ and therefore the stochastic behavior of $\boldsymbol{u}$ can be modeled on the product space $\mathbb{R}^d \times \mathbb{R}^p = \mathbb{R}^{d+p}$. Instead of employing a DPM for $\boldsymbol{u}$, we use the following finite Gaussian mixture model (see Quinlan et al. 2017, for justifications behind this model choice for density estimation):

$$\boldsymbol{u} \mid \boldsymbol{\pi}_{k,1}, \boldsymbol{\theta}_{k,d+p}, \boldsymbol{\Lambda}_{k,d+p} \sim \sum_{j=1}^{k} \pi_j \mathrm{N}_{d+p}(\boldsymbol{u}; \boldsymbol{\theta}_j, \boldsymbol{\Lambda}_j), \tag{4}$$

where $\boldsymbol{\pi}_{k,1} = (\pi_1, \ldots, \pi_k) \in \Delta_{k-1}$, $\boldsymbol{\theta}_{k,d+p} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) \in \mathbb{R}_k^{d+p} = \prod_{j=1}^{k} \mathbb{R}^{d+p}$ and $\boldsymbol{\Lambda}_{k,d+p} =$

$(\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_k) \in \mathbb{S}_k^{d+p} = \prod_{j=1}^k \mathbb{S}^{d+p}$. Here Δ_{k-1} is the standard $(k-1)$-simplex ($\Delta_0 = \{1\}$) and $\mathbb{S}^d$ is the space of real, symmetric and positive-definite matrices of dimension $d \times d$. Since $\boldsymbol{u}$ follows a finite Gaussian mixture model, the conditional distribution $(\boldsymbol{y} \mid \boldsymbol{x})$ also corresponds to a finite Gaussian mixture model. To see this, let $\boldsymbol{u}$ be modeled as in (4) and consider the following partitioned $\boldsymbol{\theta}_j$ and $\boldsymbol{\Lambda}_j$:

$$\boldsymbol{\theta}_j = \begin{pmatrix} \boldsymbol{\theta}_j^{\boldsymbol{y}} \\ \boldsymbol{\theta}_j^{\boldsymbol{x}} \end{pmatrix}, \qquad \boldsymbol{\Lambda}_j = \begin{pmatrix} \boldsymbol{\Lambda}_j^{\boldsymbol{yy}} & \boldsymbol{\Lambda}_j^{\boldsymbol{yx}} \\ \boldsymbol{\Lambda}_j^{\boldsymbol{xy}} & \boldsymbol{\Lambda}_j^{\boldsymbol{xx}} \end{pmatrix}.$$

Using standard properties of the Gaussian distribution it can be shown that the conditional distribution implied by (4) corresponds to the following weighted Gaussian regression

$$\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\pi}_{k,1}, \boldsymbol{\theta}_{k,d+p}, \boldsymbol{\Lambda}_{k,d+p} \overset{ind.}{\sim} \sum_{j=1}^{k} \pi_j(\boldsymbol{x}) \mathrm{N}_d(\boldsymbol{y}; \boldsymbol{\theta}_j(\boldsymbol{x}), \boldsymbol{\Lambda}_j(\boldsymbol{x})), \tag{5}$$

where $\pi_j(\boldsymbol{x})$, $\boldsymbol{\theta}_j(\boldsymbol{x})$ and $\boldsymbol{\Lambda}_j(\boldsymbol{x})$ have the following forms:

$$\pi_j(\boldsymbol{x}) \propto \pi_j \mathrm{N}_p(\boldsymbol{x}; \boldsymbol{\theta}_j^{\boldsymbol{x}}, \boldsymbol{\Lambda}_j^{\boldsymbol{xx}}) \tag{6}$$

$$\boldsymbol{\theta}_j(\boldsymbol{x}) = \boldsymbol{\theta}_j^{\boldsymbol{y}} + \boldsymbol{\Lambda}_j^{\boldsymbol{yx}} (\boldsymbol{\Lambda}_j^{\boldsymbol{xx}})^{-1} (\boldsymbol{x}_i - \boldsymbol{\theta}_j^{\boldsymbol{x}}) \tag{7}$$

$$\boldsymbol{\Lambda}_j(\boldsymbol{x}) = \boldsymbol{\Lambda}_j^{\boldsymbol{yy}} - \boldsymbol{\Lambda}_j^{\boldsymbol{yx}} (\boldsymbol{\Lambda}_j^{\boldsymbol{xx}})^{-1} \boldsymbol{\Lambda}_j^{\boldsymbol{xy}}. \tag{8}$$

To complete the Bayesian model, we need priors for $\boldsymbol{\pi}_{k,1}$, $\boldsymbol{\theta}_{k,d+p}$ and $\boldsymbol{\Lambda}_{k,d+p}$. It is common to assume mutual independent conjugate priors for these parameters (e.g., $\boldsymbol{\theta}_j \overset{i.i.d.}{\sim} \mathrm{N}_{d+p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \in \mathbb{R}^{d+p}$ and $\boldsymbol{\Sigma} \in \mathbb{S}^{d+p}$). Although this approach generates flexible structures that capture non-linear patterns for the mean response, as mentioned previously, the independence assumption can encourage the creation of more mixture components than actually needed to get a satisfactory fit. To avoid this, we propose modeling $\boldsymbol{\theta}_{k,d+p}$ with the repulsive distribution found in Quinlan et al. (2017). The repulsive feature is naturally inherited by

$(\boldsymbol{y} \mid \boldsymbol{x})$, encouraging more parsimonious models. Although the definition of the repulsive distribution is discussed at length in Quinlan et al. (2017), for the sake of completeness we provide it here.

Definition 2.1. Let $\boldsymbol{\theta}_{k,d+p} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) \in \mathbb{R}_k^{d+p} = \prod_{j=1}^{k} \mathbb{R}^{d+p}$, $\boldsymbol{\mu} \in \mathbb{R}^{d+p}$, $\boldsymbol{\Sigma} \in \mathbb{S}^{d+p}$, and and $\tau \in (0, \infty)$. A density for $\boldsymbol{\theta}_{k,d+p}$ that incorporates the repulsive property is

$$\mathrm{NRep}_{k,d+p}(\boldsymbol{\theta}_{k,d+p}) \propto \left\{ \prod_{j=1}^{k} \mathrm{N}_{d+p}(\boldsymbol{\theta}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \right\} \mathrm{R_C}(\boldsymbol{\theta}_{k,d+p}), \tag{9}$$

$$\mathrm{R_C}(\boldsymbol{\theta}_{k,d+p}) = \prod_{r<s}^{k} [1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)\}], \tag{10}$$

where $d, p, k \in \mathbb{N}$ with $k \geq 2$.

In what follows we use $\boldsymbol{\theta}_{k,d+p} \sim \mathrm{NRep}_{k,d+p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$ to denote that $\boldsymbol{\theta}_{k,d+p}$ follows the probability distributed in (9) and (10). Note that the parameters of this distribution are $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and τ. (10) introduces dependence between $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ by penalizing small relative distances through the expression $0.5\tau^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)$. The parameter τ controls the strength of the repulsion: as $\tau \to 0^+$, (9) converges functionally to an i.i.d. model for $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ with each following a common Gaussian distribution $\mathrm{N}_{d+p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Note further that this probability density equals 0 when $\boldsymbol{\theta}_r = \boldsymbol{\theta}_s$ for some $r \neq s$. Because of the repulsion, the implied conditional distribution $(\boldsymbol{y} \mid \boldsymbol{x})$ tends to fit flexible regression curves by using information from a small number of active clusters.

Notice that the joint likelihood derived from (4) involves an expansion into k^n terms, which is computationally expensive. An approach that simplifies the previous problem is based on introducing mutually independent auxiliary variables $z_1, \ldots, z_n \in [k]$, the mixture component indicators, such that $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ are conditionally independent given $z_1, \ldots, z_n$. The auxiliary variables can be thought of cluster labels: $\boldsymbol{u}_i$ is generated by the jth cluster if and only if $z_i = j$. Notice that from the following hierarchical stochastic formulation

$$\boldsymbol{u}_i \mid z_i, \boldsymbol{\theta}_{k,d+p}, \boldsymbol{\Lambda}_{k,d+p} \overset{ind.}{\sim} \mathrm{N}_{d+p}(\boldsymbol{u}_i; \boldsymbol{\theta}_{z_i}, \boldsymbol{\Lambda}_{z_i}) \tag{11}$$

$$z_i \mid \boldsymbol{\pi}_{k,1} \overset{i.i.d.}{\sim} \mathbb{P}(z_i = j) = \pi_j, \tag{12}$$

model (4) is recovered after marginalizing over each z_i in the joint distribution defined by (11) and (12).

Under this framework, the covariate dependent Bayesian repulsive Gaussian mixture model is completely specified by (11) and (12) with the following prior distributions (mutually independent):

$$\boldsymbol{\pi}_{k,1} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{k,1}) : \boldsymbol{\alpha}_{k,1} \in (0,\infty)^k \tag{13}$$

$$\boldsymbol{\theta}_{k,d+p} \sim \mathrm{NRep}_{k,d+p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau) : \boldsymbol{\mu} \in \mathbb{R}^{d+p}, \boldsymbol{\Sigma} \in \mathbb{S}^{d+p}, \tau \in (0,\infty) \tag{14}$$

$$\boldsymbol{\Lambda}_j \overset{i.i.d.}{\sim} \mathrm{IW}_{d+p}(\boldsymbol{\Psi}, \nu) : \boldsymbol{\Psi} \in \mathbb{S}^{d+p}, \nu \in (0,\infty). \tag{15}$$

From now on, we refer to the conditional model derived from (11)–(15) as RGMMx.

2.1 Parameter Calibration

One advantage of starting with (4) and then inducing (5) is that we can exploit essentially the same recommendations described in Quinlan et al. (2017) to elicit the hyperparameters in (13)–(15). For completeness, we provide details.

We suggest standardizing the response and covariates, which makes selecting values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ straightforward. This technique, which is suggested in Gelman et al. (2014), justifies the assignment of $\boldsymbol{\mu} = \mathbf{0}_{d+p}$ and $\boldsymbol{\Sigma} = \mathbf{I}_{d+p}$ where $\mathbf{0}_{d+p} \in \mathbb{R}^{d+p}$ is the vector whose entries are all equal 0 and $\mathbf{I}_{d+p}$ is the identity matrix of dimension $(d+p) \times (d+p)$. The same authors suggest that fixing $\boldsymbol{\alpha}_{k,1} = k^{-1}\mathbf{1}_{d+p}$ produces a weakly informative prior for the weights when the number of mixture components is relatively high. Here, $\mathbf{1}_{d+p} \in \mathbb{R}^{d+p}$ is a vector with entries equal 1. As for ν and $\boldsymbol{\Psi}$, their values are critical since misspecification can result in masking the repulsion effect: large variances can produce an overlap between mixture components, even though their location parameters are well-separated by the presence of repulsion. We suggest fixing $\nu = p + d + 4$ and $\boldsymbol{\Psi} = 3\psi\mathbf{I}_{d+p}$ with $\psi \in (0, \infty)$. This choice guarantees that $\boldsymbol{\Lambda}_j$ is centered at $\psi\mathbf{I}_{d+p}$ and has finite variance that is controlled by ψ.

Finally, specifying a value for τ is of principal interest because it guides the repulsion between the centers of each mixture components. In this case, we propose the following criterion: for fixed values $u, p \in (0, 1)$ choose τ such that

$$\mathbb{P}\left\{G \leq -2\log(1-u)\tau\right\} = p, \qquad G \sim \mathrm{G}(d/2 + p/2, 1/2). \tag{16}$$

After creating a grid of points in $(0, \infty)$ it is straightforward to find τ that satisfies (16). For more details about the motivation behind (16) see Quinlan et al. (2017).

2.2 Computation

In this section we describe the sampling mechanism to obtain posterior samples from RGMMx. Although we are not interested in making inference on parameters in (4), these realizations can be used directly to sample from the induced conditional distribution (5).

As a starting point, the posterior sampling procedure for $\boldsymbol{\pi}_{k,1}$, $\boldsymbol{\theta}_{k,d+p}$ and $\boldsymbol{\Lambda}_{k,d+p}$ is simply a Gibbs sampler. Due to conjugacy, the full conditional distributions for $\boldsymbol{\pi}_{k,1}$ and $\boldsymbol{\Lambda}_{k,d+p}$ have known closed forms and are easy to sample from. Unfortunately, this is not the case for $\boldsymbol{\theta}_{k,d+p}$. However, the fact that the coordinates of $\boldsymbol{\theta}_{k,d+p} \sim \mathrm{NRep}_{k,d+p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$ are exchangeable implies that all the associated full conditional distributions share the same functional form. Moreover, evaluating the respective densities has a low computational cost. Because of this, we incorporate a Metropolis–Hastings step inside the Gibbs Sampler. To illustrate the procedure we need the full conditional distribution $(\boldsymbol{\theta}_{k,d+p} \mid \cdots)$ which is given by

$$f(\boldsymbol{\theta}_{k,d+p} \mid \cdots) \propto \left\{ \prod_{j=1}^{k} \mathrm{N}_{d+p}(\boldsymbol{u}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right\} \prod_{r<s}^{k} [1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_s)\}],$$

where $\boldsymbol{\mu}_j = \boldsymbol{\Sigma}_j(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\Lambda}_j^{-1}\boldsymbol{s}_j)$, $\boldsymbol{s}_j = \sum_{i=1}^{n} \mathbb{I}_{\{j\}}(z_i)\boldsymbol{u}_i$, $\boldsymbol{\Sigma}_j = (\boldsymbol{\Sigma}^{-1} + n_j\boldsymbol{\Lambda}_j^{-1})^{-1}$ and $n_j = \mathrm{card}\{i \in [n] : z_i = j\}$. Based on this result, the full conditional distributions $(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}, \cdots)$ for $j \in [k]$ and $\boldsymbol{\theta}_{-j} = (\boldsymbol{\theta}_l : l \neq j) \in \mathbb{R}_{k-1}^{d+p}$ have corresponding densities

$$f(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}, \cdots) \propto \mathrm{N}_{d+p}(\boldsymbol{\theta}_j; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \prod_{l \neq j}^{k} \left[1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_l)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_l)\}\right].$$

The following pseudo-code could then be used to sample from $(\boldsymbol{\theta}_{k,d+p} \mid \cdots)$ using $f(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}, \cdots)$ through a random walk Metropolis–Hastings step inside the Gibbs Sampler:

```
1. Let $\boldsymbol{\theta}_{k,d+p}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_k^{(0)}) \in \mathbb{R}_k^{d+p}$ be the actual state for $\boldsymbol{\theta}_{k,d+p}$.

2. For $j = 1, \ldots, k$:

    (a) Draw a candidate $\boldsymbol{\theta}_j^{(1)}$ generated from $\mathrm{N}_{d+p}(\boldsymbol{\theta}_j^{(0)}, \boldsymbol{\Gamma}_j)$ with $\boldsymbol{\Gamma}_j \in \mathbb{S}^{d+p}$.

    (b) Set $\boldsymbol{\theta}_j^{(0)} = \boldsymbol{\theta}_j^{(1)}$ with probability $\min(1, \beta_j)$, where

        $\beta_j = \frac{\mathrm{N}_{d+p}(\boldsymbol{\theta}_j^{(1)}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\mathrm{N}_{d+p}(\boldsymbol{\theta}_j^{(0)}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \prod_{l \neq j}^{k} \left[ \frac{1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j^{(1)} - \boldsymbol{\theta}_l^{(0)})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_j^{(1)} - \boldsymbol{\theta}_l^{(0)})\}}{1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j^{(0)} - \boldsymbol{\theta}_l^{(0)})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_j^{(0)} - \boldsymbol{\theta}_l^{(0)})\}} \right].$
```

Since $\boldsymbol{\Gamma}_j$ controls the variability of the generated candidates, care must be taken when selecting it. An approach that seems to works well in practice is to fix

$$\boldsymbol{\Gamma}_j = \frac{1}{B} \sum_{t=1}^{B} \left\{ \boldsymbol{\Sigma}^{-1} + n_j^{(t)}(\boldsymbol{\Lambda}_j^{(t)})^{-1} \right\}^{-1} : n_j^{(t)} = \mathrm{card}\{i \in [n] : z_i^{(t)} = j\},$$

where $t \in [B]$ is the t-th iteration of the burn-in phase of length $B \in \mathbb{N}$. In Appendix A we provide the complete pseudo-code associated with the RGMMx model to obtain posterior samples for the parameters that appear in (4). We also discuss how they can be used to estimate regression densities and regression curves. We refer to the algorithm in Appendix A as Algorithm RGMMx.

3 Simulation Study

As a means to explore the proposed methodology we take on the simulation situation of Dunson et al. (2007). More specifically we generate data sets from the following density

$$f_0(y \mid x) = \exp(-2x)\mathrm{N}(y; x, 0, 01) + \{1 - \exp(-2x)\}\mathrm{N}(y; x^4, 0, 05) : y \in \mathbb{R},$$

where $x \in (0, 1)$ is taken as the covariate. This Gaussian regression mixture has weights that vary smoothly in x, different variances for each mixture component, and a non-linear mean in the second component. The mean regression curve m_0 associated with f_0 is

$$m_0(x) = \exp(-2x)x + \{1 - \exp(-2x)\}x^4 : x \in (0, 1). \tag{17}$$

Notice that by construction f_0 can take on a diverse number of shapes, ranging from unimodal (symmetric or asymmetric) to bimodal densities.

In this small simulation study we focus on proof of concept associated with the RGMMx by exploring how values of τ influence the repulsiveness of our methodology and ultimately the number of estimated clusters and goodness-of-fit. We do this by fitting the RGMMx to data generated using different values for τ, namely, 0.01, 0.1, 1 and 10. We evaluate goodness-of-fit by using the following metrics:

- Log Pseudo Marginal Likelihood (LPML) for the joint model (Christensen et al. 2011) which is a model fit metric that takes into account model complexity. We calculate the LPML by first estimating all the corresponding conditional predictive ordinates (Gelfand et al. 1992) using the method in Chen et al. (2000).

- L_1-metric between the estimated mean regression curve and m_0.

To see how τ influences the posterior distribution associated with the number of clusters, we include the following numeric indicators:

- Posterior average number of occupied components, i.e. $n_j = \text{card}\{i \in [n] : z_i = j\} > 0$.
- Standard deviation of the posterior average number of occupied components.

We generate 100 data sets, each of size 500, by first generating x from $\mathrm{U}(0,1)$ and for each x generating a realization y using $f_0(y \mid x)$. For each value of τ we fit RGMMx to data by collecting 5000 MCMC iterates after discarding the first 5000 as burn-in and thinning by 25. The rest of the prior parameter values in (13)–(15) are set as follows:

- $k = 10$, $d = 1$, $p = 1$, $\boldsymbol{\alpha}_{k,1} = 10^{-1}\mathbf{1}_{10}$, $\boldsymbol{\mu} = \mathbf{0}_2$, $\boldsymbol{\Sigma} = \mathbf{I}_2$, $\boldsymbol{\Psi} = 3^{-1}\mathbf{I}_2$ and $\nu = 6$.

In Figure 1 we provide the results of each metric for each value of τ considered by way of side-by-side box-plots. Interestingly, LPML is not a monotonic function of τ. There is an initial increase and then a decrease of LPML as τ increases (and thus the number of clusters decreases). The L_1-metric does not seem to be influenced by τ. As expected, increasing τ results in stronger repulsion and therefore less clusters. In addition, stronger repulsion produces less variability in the number of components. This makes sense because as the strength of repulsion increases, there are less locations at which cluster centers can exist.

4 Data Illustration

Azzalini and Bowman (1990) analized a data set concerning the eruptions from the Old Faithful geyser in Yellowstone National Park, Wyoming. These were continuous measurements from August 1 to August 15, 1985. The recorded data represents time eruptions

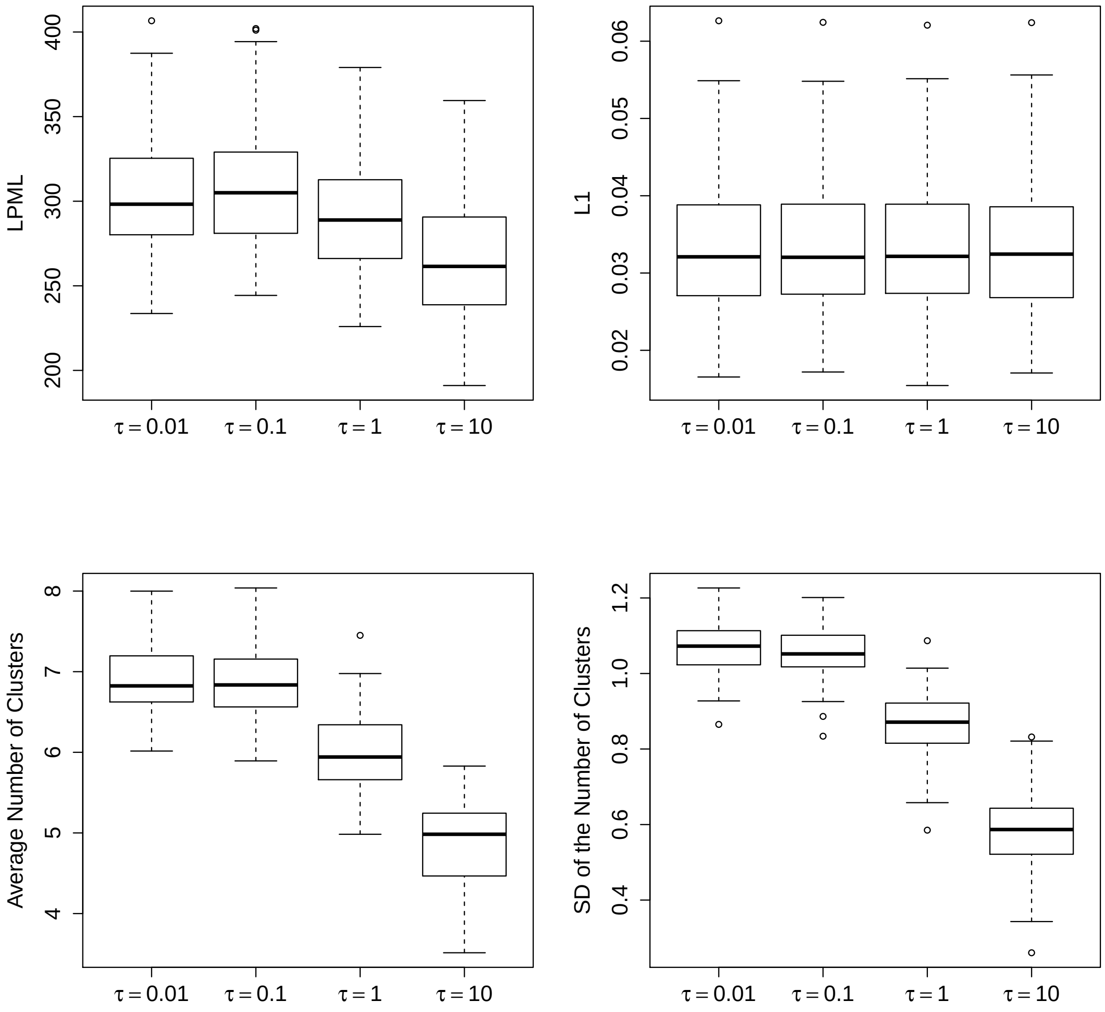

Figure 1: Boxplots that display LPML, L_1-metric, the average number of occupied mixture components, and the average standard deviation associated with the distribution of occupied mixture components for each value of τ.

(*duration*) and waiting times for each eruption (*waiting*), both in minutes. In this illustration, the first (last) variable is considered as the covariate (response). We removed 78 out of the 299 observations that were collected at night (coded as 2, 3 or 4 minutes) and originally described as “short”, “medium” or “long”. The original data can be found in `R` under the name *geyser* (MASS library).

We implemented two WDDP and two RGMMx versions to compare the respective regression mean curves for *waiting* in terms of *duration*. For each procedure we report the LPML as a measure of goodness-of-fit, a brief summary regarding the average number of occupied components, and posterior distribution associated with the number of clusters. We standardized the data before fitting the above models. Our main aim here is to assess the effect that the prior specification on the repulsion parameter has on the reported inference. Specific details now follow:

1. RGMMx: We coded Algorithm RGMMx in `Fortran` to generate posterior draws from this model. For both model specifications (referred to as RGMMx1 and RGMMx2) we collected 5000 MCMC iterates after discarding the first 10000 as burn-in and thinning by 20. We use the same prior parameter values in (13) to (15) for both models. Specifically, we use: $k = 10$, $d = 1$, $p = 1$, $\boldsymbol{\alpha}_{k,1} = 10^{-1}\mathbf{1}_{10}$, $\boldsymbol{\mu} = \mathbf{0}_2$, $\boldsymbol{\Sigma} = \mathbf{I}_2$, $\boldsymbol{\Psi} = 3^{-1}\mathbf{I}_2$ and $\nu = 6$. The respective values for τ, selected by the calibration criterion from Subsection 2.1, are provided below.

   (a) RGMMx1: $\tau = 0.2$ with $(u, p) = (0.999, 0.5)$.

   (b) RGMMx2: $\tau = 4.6$ with $(u, p) = (0.5, 0.8)$.

   We emphasize the fact that the setting $\tau = 0.2$ in RGMMx1 produces a fairly weak repulsive behavior when modeling the response and covariate jointly. The motivation behind this selection is to avoid underfitting as large values of τ could result in mixture models with a small number of occupied components. This would have serious repercussions regarding the quality of model fit and flexibility in estimating the regression curve. On the other hand, overfitting can be partially avoided by fixing the number of components to a reasonable value k that is not too large. In addition, since (10) models repulsion *softly* (see Ogata and Tanemura 1981), the strength of repulsion (i.e., the value of τ) would need to be very small for the active number of components to be large. As for $\tau = 4.6$ in RGMMx2, this value forces the number of occupied components to be smaller than RGMMx1.

2. WDDP: The baseline distribution H_0 in (3) that we use is the conjugate Normal-Inverse-Wishart

$$H_0(\boldsymbol{\theta}, \boldsymbol{\Lambda}) = \mathrm{N}_{d+p}(\boldsymbol{\theta}; \boldsymbol{m}_1, k_0^{-1}\boldsymbol{\Lambda})\, \mathrm{IW}_{d+p}(\boldsymbol{\Lambda}; \boldsymbol{\Psi}_1, \nu_1) : \nu_1 \in (0, \infty). \tag{18}$$

To complete the model specification given by (1)–(3) with (18), the following independent hyperpriors are assumed:

$$\alpha \mid a_0, b_0 \sim \mathrm{G}(a_0, b_0) : a_0, b_0 \in (0, \infty) \tag{19}$$

$$\boldsymbol{m}_1 \mid \boldsymbol{m}_2, \mathbf{S}_2 \sim \mathrm{N}_{d+p}(\boldsymbol{m}_2, \mathbf{S}_2) : \boldsymbol{m}_2 \in \mathbb{R}^{d+p}, \mathbf{S}_2 \in \mathbb{S}^{d+p} \tag{20}$$

$$k_0 \mid \tau_1, \tau_2 \sim \mathrm{G}(\tau_1/2, \tau_2/2) : \tau_1, \tau_2 \in (0, \infty) \tag{21}$$

$$\boldsymbol{\Psi}_1 \mid \boldsymbol{\Psi}_2, \nu_2 \sim \mathrm{IW}_{d+p}(\boldsymbol{\Psi}_2, \nu_2) : \boldsymbol{\Psi}_2 \in \mathbb{S}^{d+p}, \nu_2 \in (0, \infty). \tag{22}$$

As a comparison, we considered the `R` function `DPcdensity` available in `DPpackage` (Jara et al. 2011). Decisions on hyperprior parameter values were guided by Escobar and West (1995). For each of the following model specifications, named WDDP1 and WDDP2, we collected 5000 MCMC iterates after discarding the first 5000 as burn-in and thinning by 3. The respective prior hyperparameter values in (19)–(22) are provided below.

   (a) WDDP1: $d = 1$, $p = 1$, $a_0 = 10$, $b_0 = 1$, $\nu_1 = 4$, $\nu_2 = 4$, $\boldsymbol{m}_2 = \mathbf{0}_2$, $\mathbf{S}_2 = \mathbf{I}_2$, $\boldsymbol{\Psi}_2 = \mathbf{I}_2$, $\tau_1 = 6.01$ and $\tau_2 = 2.01$.

   (b) WDDP2: $d = 1$, $p = 1$, $a_0 = 2$, $b_0 = 4$, $\nu_1 = 4$, $\nu_2 = 4$, $\boldsymbol{m}_2 = \mathbf{0}_2$, $\mathbf{S}_2 = \mathbf{I}_2$,

$\mathbf{\Psi}_2 = \mathbf{I}_2$, $\tau_1 = 2.01$ and $\tau_2 = 1.01$.

These values make WDDP1 and RGMMx1 (WDDP2 and RGMMx2) "similar" in terms of the distribution for the number of occupied components *a priori*: more (less) spread around a high (small) average value.

Model	LPML	Mean (Clusters)	SD (Clusters)
RGMMx1	-195.14	6.15	0.97
RGMMx2	-208.62	4.90	0.74
WDDP1	-185.82	11.51	3.11
WDDP2	-226.73	5.64	1.39

Table 1: Summary statistics related to model fit and the number of clusters for Geyser data based on WDDP and RGMMx.

Table 1 and Figure 2 show that for a small number of clusters RGMMx tends to fit the Geyser data better than WDDP. On the other hand, WDDP is able to fit the data slightly better than RGMMx when the number of clusters increases. This slight increase in the LPML value comes at a substantial model complexity cost as the number of active mixture components is doubled. Figure 3 shows that there is no appreciable difference between the mean regression curves, but the estimated 95% point-wise credible bands under RGMMx are slightly wider than those under WDDP. Figure 4 displays an estimated partition for each procedure using Dahl's least squares method (Dahl 2006). Notice that partitions associated with model specifications that produce the highest LPML values (i.e., panels "b" (RGMMx2) and "d" (WDDP2)) agree, inducing four well-formed groups and one isolated point. Panel "a" (RGMMx1) reveals that a small repulsion effect allows grouping data into a moderate number of clusters that are fairly separated. Different is the case in panel "c" (WDDP1), where the number of clusters is quite high and some of these overlap.

Figures 5 and 6 provide a few estimated conditional densities for four different values of time eruptions (2, 3, 4 and 4.5), and 95% point-wise credible bands. As expected, all plots labeled with “a” (RGMMx1) exhibit similar densities to those labeled with “c” (WDDP1). The same scenario occurs between plots labeled with “b” (RGMMx2) and “d” (WDDP2). However, there are slight differences in the width of the point-wise credible bands with WDDP1 and WDDP2 being slightly narrower.

The take home message from this application is that the RGMMx, being a parametric model, is a simple, parsimonious alternative to WDDP in being able to capture flexible regression curves.

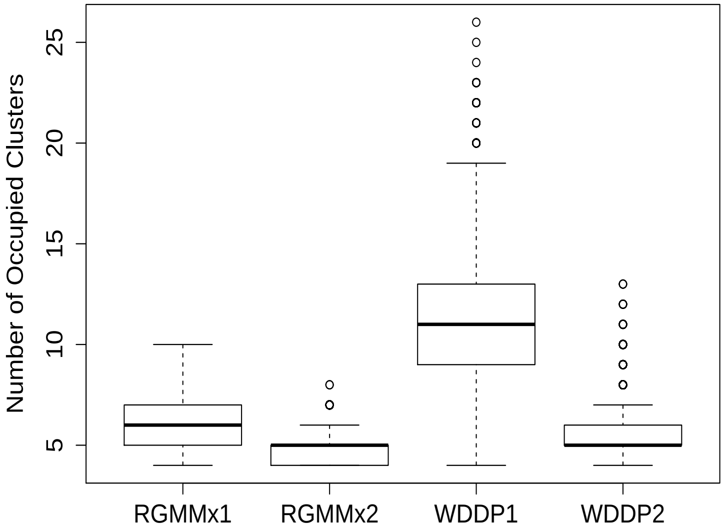

Figure 2: Side-by-side box-plots of the posterior distribution for the active number of clusters associated with the Geyser data.

5 Discussion and Future Work

This article contains extensions to repulsive mixture modeling that are completely methodological. Using a similar approach to Müller et al. (1996) we propose a finite Bayesian

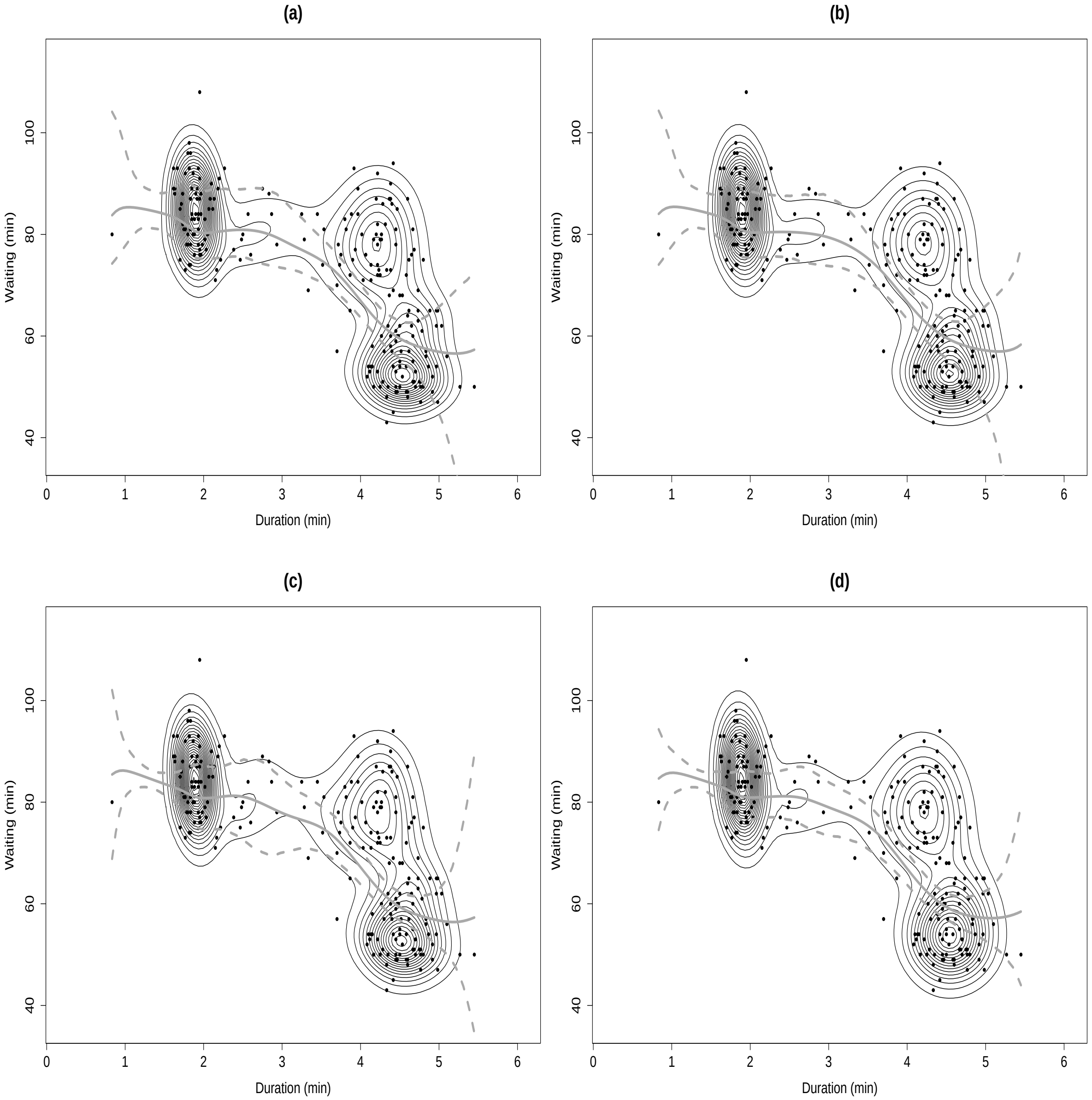

Figure 3: Estimated regression curve (gray solid) for Geyser data under (a) RGMMx1, (b) RGMMx2, (c) WDDP1 and (d) WDDP2. In each scenario, the gray dashed curves correspond to 95% point-wise credible intervals.

mixture of Gaussian distributions to jointly model responses and (continuous) covariates where the associated location parameters are driven by a probability distribution which encourages them to repel each other. The joint model naturally induces a conditional distri-

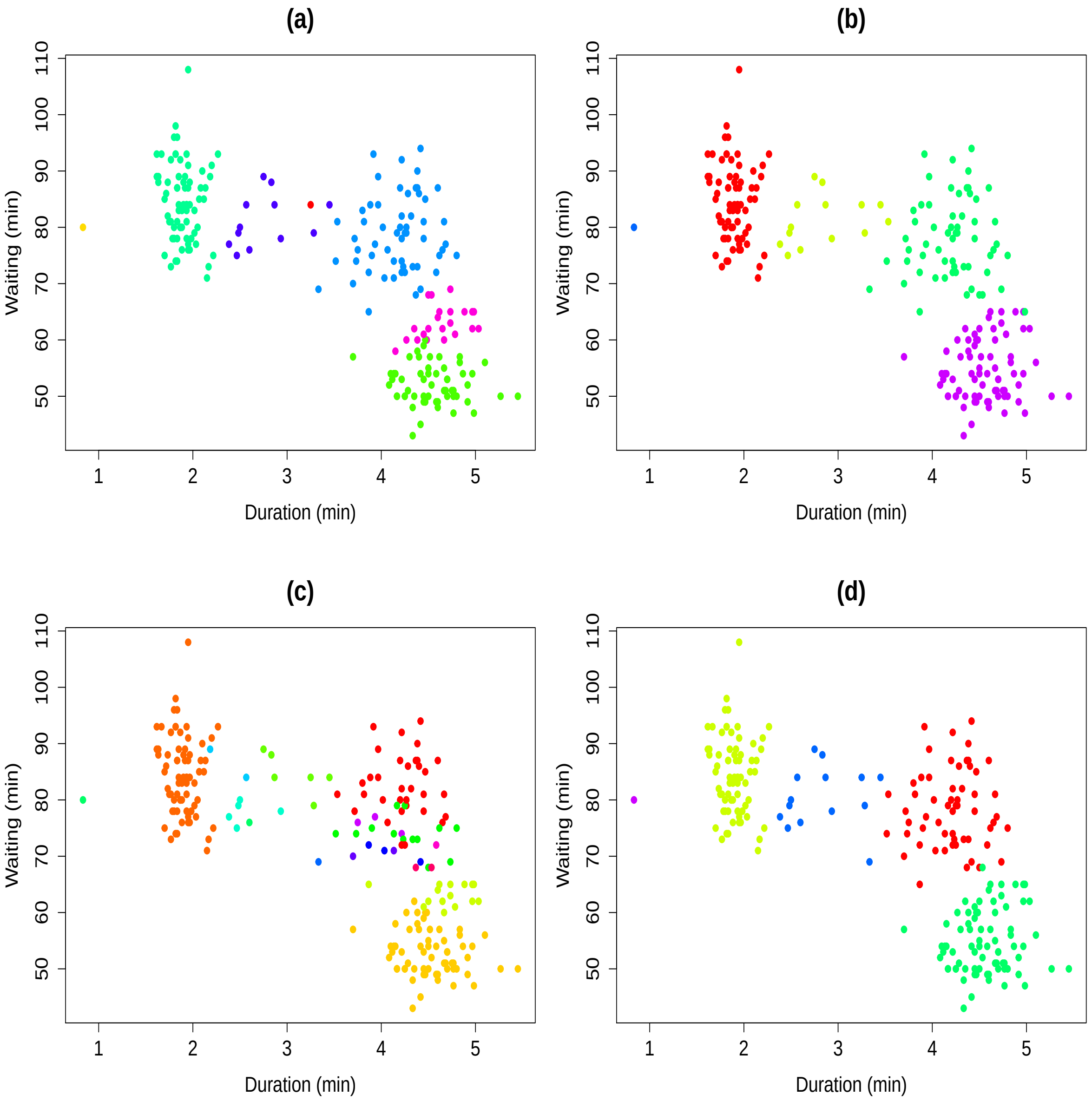

Figure 4: Estimated partitions using Dahl's least squares clustering algorithm for (a) RGMMx1, (b) RGMMx2, (c) WDDP1 and (d) WDDP2.

bution, which is a weighted mixture of Gaussian regressions that inherits the repulsion effect. An important consequence of this is that the conditional distribution allows estimation of regression curves in a flexible and parsimonious way, i.e. using a reduced number of mixture components at almost no cost in terms of goodness-of-fit. It is worth noting that a down-

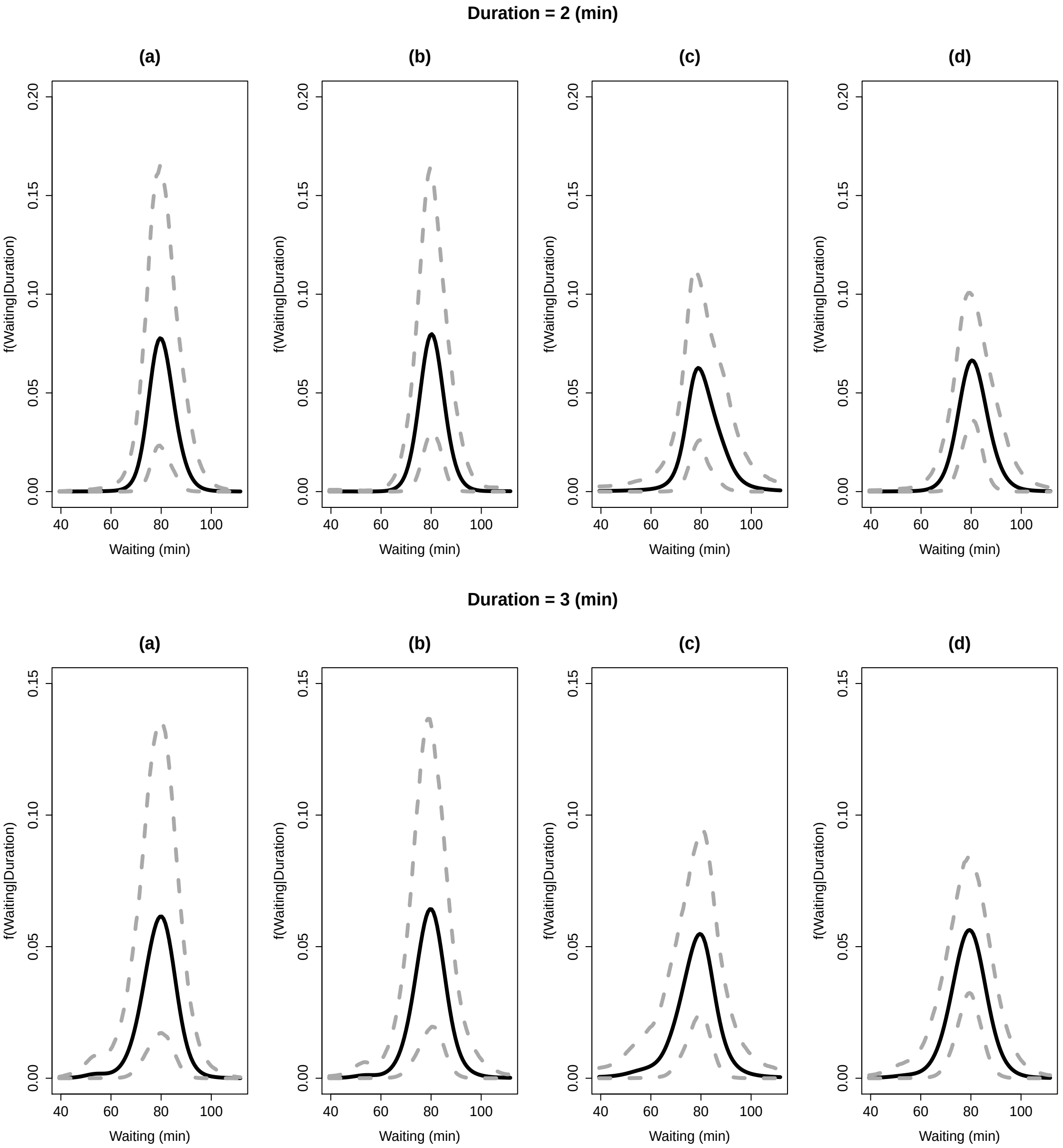

Figure 5: Estimated conditional densities (black solid) for Geyser data under (a) RGMMx1, (b) RGMMx2, (c) WDDP1 and (d) WDDP2. In each scenario, the gray dashed curves correspond to 95% point-wise credible intervals. Here, the selected time eruptions (*duration*) are 2 and 3 minutes.

side of both methodologies is the curse of dimensionality. That is, when responses and/or covariates lie on high dimensional Euclidean spaces, computation becomes very expensive.

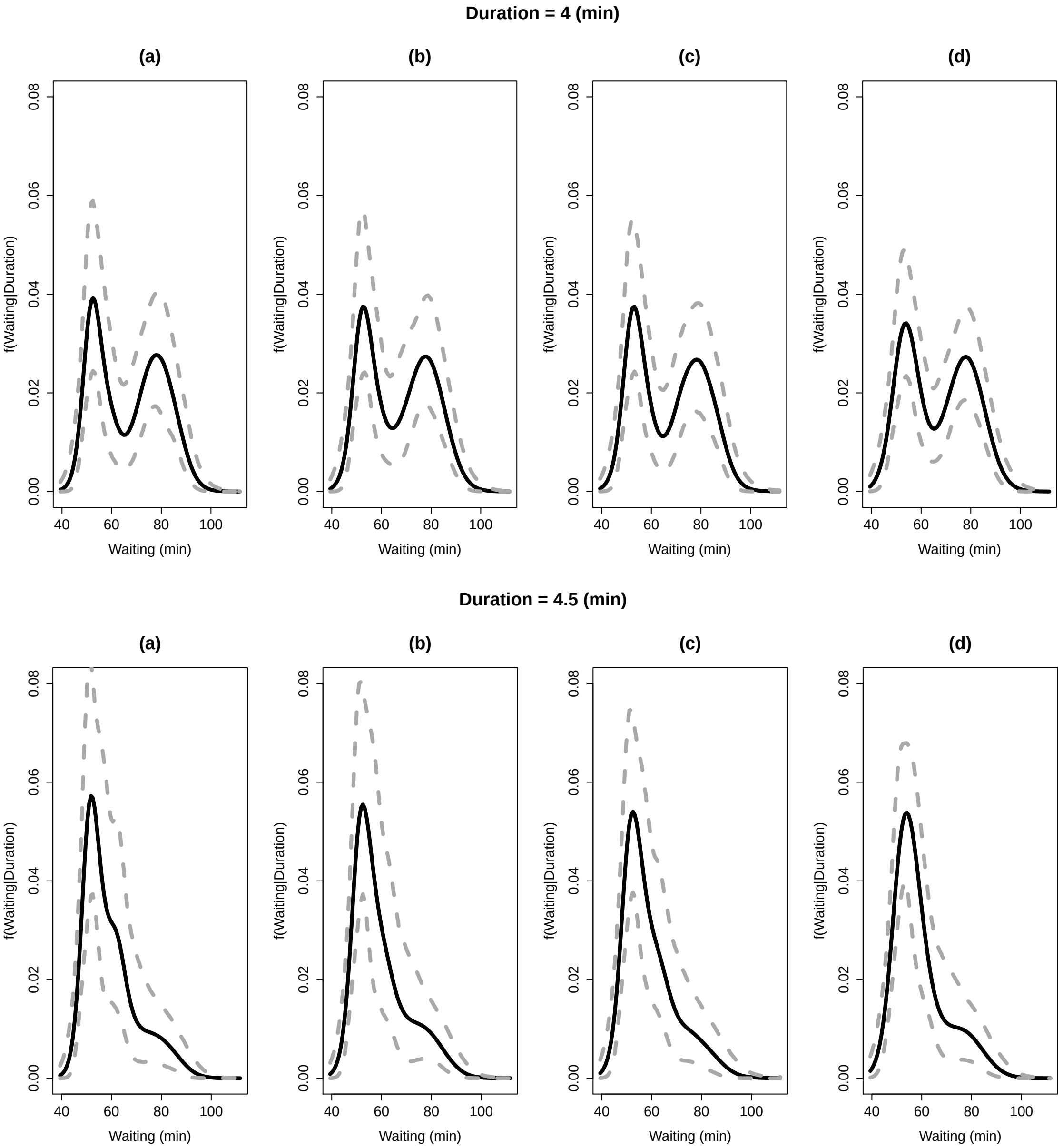

Figure 6: Estimated conditional densities (black solid) for Geyser data under (a) RGMMx1, (b) RGMMx2, (c) WDDP1 and (d) WDDP2. In each scenario, the gray dashed curves correspond to 95% point-wise credible intervals. Here, the selected time eruptions (*duration*) are 4 and 4.5 minutes.

Future research will be dedicated to studying the topological support associated with RGMMx to determine the class of regression curves that can be approximated (a study that is similar to what was done in Barrientos et al. 2012). Additionally, since τ seems to influence model fit and predictions it would be natural to treat it as an unknown and assign it a prior distribution. Doing this however will come at a formidable computational cost because it can be immediately shown that the posterior distribution of τ is doubly intractable. Finally, we would like to develop a method that avoids focusing on the joint distribution of a response and covariate and instead incorporates repulsion directly in the conditional distribution. One possible way of carrying this out is to employ a probit stick-breaking prior (Rodriguez and Dunson 2011) for the mixture weights, modeling the centers of each component by linear regressions. Repulsion would then influence the vector of regression coefficients producing well separated clusters of linear regressions. This type of model has also the advantage of not being restricted to the case of continuous covariates.

Acknowledgments

José Quinlan gratefully recognizes the financial support provided by CONICYT through Doctoral Scholarship Grant 21120153 and Fondecyt Grants 11121131 and 1141057. Fernando A. Quintana was supported by Fondecyt Grant 1141057 and Garritt L. Page was partially supported by Fondecyt Grant 11121131.

References

Azzalini, A. and Bowman, A. W. (1990), "A Look at Some Data on the Old Faithful Geyser," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 39, 357–365.

Barrientos, A. F., Jara, A., and Quintana, F. A. (2012), "On the Support of MacEachern's Dependent Dirichlet Processes and Extensions," *Bayesian Anal.*, 7, 277–310.

Bianchini, I., Guglielmi, A., and Quintana, F. A. (2017), "Determinantal point process mixtures via spectral density approach," *ArXiv e-prints.*

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000), *Monte Carlo Methods in Bayesian Computation*, Springer New York.

Christensen, R., Johnson, W., Branscum, A. J., and Hanson, T. (2011), *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*, CRC Press.

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. Vannucci, M., Do, K. A., and Müller, P., Cambridge University Press, pp. 201–218.

Dunson, D. B., Pillai, N., and Park, J.-H. (2007), "Bayesian density regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 163–183.

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

Frühwirth-Schnatter, S. (2006), *Finite mixture and Markov switching models*, Springer Series in Statistics, Springer, New York.

Fúquene, J., Steel, M., and Rossell, D. (2016), "On choosing mixture components via non-local priors," *ArXiv e-prints*.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992), "Model determination using predictive distributions with implementation via sampling-based methods," Tech. rep., DTIC Document.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2014), *Bayesian Data Analysis*, London: Chapman and Hall/CRC, 3rd ed.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010), *Bayesian Nonparametrics*, vol. 28, Cambridge University Press.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008), *Statistical analysis and modelling of spatial point patterns*, Statistics in Practice, John Wiley & Sons, Ltd., Chichester.

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011), "DPpackage: Bayesian Semi- and Nonparametric Modeling in R," *Journal of Statistical Software*, 40, 1–30.

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*, 95–110.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015), *Bayesian Nonparametric Data Analysis*, Springer.

Ogata, Y. and Tanemura, M. (1981), "Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure," *Annals of the Institute of Statistical Mathematics*, 33, 315–338.

Petralia, F., Rao, V., and Dunson, D. B. (2012), "Repulsive Mixtures," in *Advances in Neural Information Processing Systems 25*, eds. Pereira, F., Burges, C., Bottou, L., and Weinberger, K., Curran Associates, Inc., pp. 1889–1897.

Quinlan, J. J., Quintana, F. A., and Page, G. L. (2017), "Density Estimation using Repulsive Distributions," *submitted.*

Rodriguez, A. and Dunson, D. B. (2011), "Nonparametric Bayesian models through probit stick-breaking processes," *Bayesian analysis (Online)*, 6.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 639–650.

Xie, F. and Xu, Y. (2017), "Bayesian Repulsive Gaussian Mixture Model," *ArXiv e-prints.*

Xu, Y., Müller, P., and Telesca, D. (2016), "Bayesian Inference for Latent Biological Structure with Determinantal Point Processes (DPP)," *Biometrics*, 72, 955–964.

A Algorithm RGMMx

In what follows we describe the Gibbs Sampler for the RGMMx in its entirety. Let $B, S, T \in \mathbb{N}$ be the total number of iterations during the burn-in, the number of collected iterates, and the thinning, respectively.

- (Start) Choose initial values $z_i^{(0)} : i \in [n]$, $\boldsymbol{\pi}_{k,1}^{(0)}$ and $\boldsymbol{\theta}_{k,d+p}^{(0)}, \boldsymbol{\Lambda}_{k,d+p}^{(0)} : j \in [k]$. Set $\boldsymbol{\Gamma}_j = \mathbf{O}_{d+p} : j \in [k]$, where $\mathbf{O}_{d+p}$ is the null matrix of dimension $(d+p) \times (d+p)$.
- (Burn-in phase) For $t = 0, \ldots, B-1$:

  1. $(z_i^{(t+1)} \mid \cdots) \sim \mathbb{P}(z_i^{(t+1)} = j) = \pi_j^{(t,i)}$ independently for each $i \in [n]$, where,

  $$\pi_j^{(t,i)} = \left\{ \sum_{l=1}^{k} \pi_l^{(t)} \mathrm{N}_{d+p}(\boldsymbol{y}_i; \boldsymbol{\theta}_l^{(t)}, \boldsymbol{\Lambda}_l^{(t)}) \right\}^{-1} \pi_j^{(t)} \mathrm{N}_{d+p}(\boldsymbol{y}_i; \boldsymbol{\theta}_j^{(t)}, \boldsymbol{\Lambda}_j^{(t)}) : j \in [k].$$

  2. $(\boldsymbol{\pi}_k^{(t+1)} \mid \cdots) \sim \mathrm{Dir}(\boldsymbol{\alpha}_{k,1}^{t})$, where

  $$\begin{aligned} \boldsymbol{\alpha}_{k,1}^{(t)} &= (\alpha_1 + n_1^{(t+1)}, \ldots, \alpha_k + n_k^{(t+1)}) \\ n_j^{(t+1)} &= \mathrm{card}\{i \in [n] : z_i^{(t+1)} = j\} : j \in [k]. \end{aligned}$$

3. For $j = 1, \dots, k$:

   3.1. Generate $\boldsymbol{\theta}_j^{(\star)}$ from $\mathrm{N}_{d+p}(\boldsymbol{\theta}_j^{(t)}, \boldsymbol{\Omega}_j^{(t)})$, where

   $$\boldsymbol{\Omega}_j^{(t)} = \{\boldsymbol{\Sigma}^{-1} + n_j^{(t+1)}(\boldsymbol{\Lambda}_j^{(t)})^{-1}\}^{-1}.$$

   3.2. Update $\boldsymbol{\theta}_j^{(t)} \to \boldsymbol{\theta}_j^{(t+1)} = \boldsymbol{\theta}_j^{(\star)}$ with probability $\min(1, \beta_j)$, where

   $$\beta_j = \frac{\mathrm{N}_{d+p}(\boldsymbol{\theta}_j^{(\star)}; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}{\mathrm{N}_{d+p}(\boldsymbol{\theta}_j^{(t)}; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})} \prod_{l \neq j}^{k} \left[\frac{1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j^{(\star)} - \boldsymbol{\theta}_l^{(t)})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_j^{(\star)} - \boldsymbol{\theta}_l^{(t)})\}}{1 - \exp\{-0.5\tau^{-1}(\boldsymbol{\theta}_j^{(t)} - \boldsymbol{\theta}_l^{(t)})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_j^{(t)} - \boldsymbol{\theta}_l^{(t)})\}} \right].$$

   In the above expression for β_j

   $$\begin{aligned} \boldsymbol{\Sigma}_j^{(t)} &= \{\boldsymbol{\Sigma}^{-1} + n_j^{(t+1)}(\boldsymbol{\Lambda}_j^{(t)})^{-1}\}^{-1} \\ \boldsymbol{\mu}_j^{(t)} &= \boldsymbol{\Sigma}_j^{(t)}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + (\boldsymbol{\Lambda}_j^{(t)})^{-1}\boldsymbol{s}_j^{(t)}\} : \boldsymbol{s}_j^{(t)} = \sum_{i=1}^{n} \mathbb{I}_{\{j\}}(z_i^{(t+1)})\boldsymbol{y}_i. \end{aligned}$$

   Otherwise, set $\boldsymbol{\theta}_j^{(t+1)} = \boldsymbol{\theta}_j^{(t)}$.

   3.3. Update $\boldsymbol{\Gamma}_j \to \boldsymbol{\Gamma}_j + B^{-1}\boldsymbol{\Omega}_j^{(t)}$.

4. $(\boldsymbol{\Lambda}_j^{(t+1)} \mid \cdots) \sim \mathrm{IW}_{d+p}(\boldsymbol{\Psi}_j^{(t)}, \nu_j^{(t)})$ independently for each $j \in [k]$, where $\nu_j^{(t)} = \nu + n_j^{(t+1)}$ and

   $$\boldsymbol{\Psi}_j^{(t)} = \boldsymbol{\Psi} + \sum_{i=1}^{n} \mathbb{I}_{\{j\}}(z_i^{(t+1)})(\boldsymbol{y}_i - \boldsymbol{\theta}_j^{(t+1)})(\boldsymbol{y}_i - \boldsymbol{\theta}_j^{(t+1)})^\top.$$

- (Save samples) For $t = B, \dots, ST + B - 1$: Repeat steps 1, 2 and 4 of the burn-in phase. As for step 3, ignore 3.3, maintain 3.2 and replace 3.1 with

  3.1. Generate a candidate $\boldsymbol{\theta}_j^\star$ from $\mathrm{N}_{d+p}(\boldsymbol{\theta}_j^{(t)}, \boldsymbol{\Gamma}_j)$.

  Finally, save the generated samples every Tth iteration.

- (Posterior conditional predictive estimates) With the T saved samples, compute

  $$\begin{aligned} f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{u}_1, \dots, \boldsymbol{u}_n) &\approx \frac{1}{T} \sum_{t=1}^{T} \left\{ \sum_{j=1}^{k} \pi_j^{(t)}(\boldsymbol{x}) \mathrm{N}_{d+p}(\boldsymbol{u}; \boldsymbol{\theta}_j^{(t)}, \boldsymbol{\Lambda}_j^{(t)}) \right\} \\ \mathrm{E}[\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{u}_1, \dots, \boldsymbol{u}_n] &\approx \frac{1}{T} \sum_{t=1}^{T} \left\{ \sum_{j=1}^{k} m_j^{(t)}(\boldsymbol{x}) \pi_j^{(t)}(\boldsymbol{x}) \mathrm{N}_p(\boldsymbol{x}; (\boldsymbol{\theta}_j^{\boldsymbol{x}})^{(t)}, (\boldsymbol{\Lambda}_j^{\boldsymbol{xx}})^{(t)}) \right\} \end{aligned}$$

where $\boldsymbol{u} = (\boldsymbol{y}, \boldsymbol{x})$ and

$$\pi_j^{(t)}(\boldsymbol{x}) = \left[\frac{1}{T} \sum_{s=1}^{T} \left\{ \sum_{l=1}^{k} \pi_l^{(s)} \mathrm{N}_p(\boldsymbol{x}; (\boldsymbol{\theta}_l^{\boldsymbol{x}})^{(s)}, (\boldsymbol{\Lambda}_l^{\boldsymbol{xx}})^{(s)}) \right\} \right]^{-1} \pi_j^{(t)}$$

$$m_j^{(t)}(\boldsymbol{x}) = (\boldsymbol{\theta}_j^{\boldsymbol{y}})^{(t)} + (\boldsymbol{\Lambda}_j^{\boldsymbol{yx}})^{(t)} \{(\boldsymbol{\Lambda}_j^{\boldsymbol{xx}})^{(t)}\}^{-1} (\boldsymbol{x} - (\boldsymbol{\theta}_j^{\boldsymbol{x}})^{(t)}).$$