PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

BAYESIAN NONPARAMETRIC HYPOTHESIS TESTING

Luz Adriana Pereira Hoyos

Pontificia Universidad Católica de Chile
Faculty of Mathematics, Department of Statistics
Santiago de Chile, Chile
2019

Bayesian nonparametric hypothesis testing

by

Luz Adriana Pereira Hoyos

A dissertation submitted in partial fulfillment
of the requirements for the degree of:

Doctor of Philosophy

in

Statistics

We accept this thesis as
conforming to the required standard

Pontificia Universidad Católica de Chile
Facultad de Matemáticas, Departamento de Estadística
Santiago de Chile, Chile
2019

Dedication

To my husband, with love.

Acknowledgments

I am deeply grateful to...

Abstract

In this thesis, we propose novel Bayesian Nonparametric hypothesis testing procedures for correlated data. First, we develop and study a proposal for comparing the distributions of paired samples. Next, we propose and analyze a hypothesis testing procedure for longitudinal data analysis. Both proposals are based on a flexible model for the joint distribution of the observations. The flexibility is given by a mixture of Dirichlet processes. Besides, for setting up the hypothesis testing procedures, we use a hierarchical representation with a spike-slab prior specification for the base measure of the Dirichlet process and a prior specification on the space of models.

For the paired sample test, we use an appropriate parametrization for the kernel of the mixture to facilitate the comparisons and posterior inference. Consequently, the joint model allows us to derive the marginal distributions and test whether they differ or not. The procedure exploits the correlation between samples, relaxes the parametric assumptions, and detects possible differences throughout the entire distributions.

For the longitudinal data, we propose to use a mixture of Dependent Dirichlet Processes to capture the correlation between the repeated measurements. The weights of the mixture are built via a stick-breaking prior, that comes from a Markovian process evolving in time. The effect of the predictors is modeled by the underlying atoms. The proposal can provide an estimation of the density through the time for different levels of the predictors, and at the same time can identify the effect of the predictors, without assuming restrictive distributional assumptions.

We show the performance throughout the document of our proposals in illustrations with simulated and real data sets. Finally, we provide concluding remarks and discuss open problems.

Keywords: Spike-slab priors; Dirichlet Process; Shift function; Mixture model; Dependent samples; Time-dependent data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this Chapter, we give an account of some of the main concepts in statistics, that will be used to set up notation and basic ideas about Bayesian nonparametric inference. We start the discussion introducing concepts such as random phenomenon, probability space, random variables, statistical model, stochastic dependence, and exchangeability. We provide some examples and illustrations to describe the methods and approaches that will be used. Later, we introduce a general approach to perform hypothesis testing from a Bayesian point of view, as well as, some elements of the variable selection methods in the context of regression models, which together with the Dirichlet process prior will be the main tools to develop our new Bayesian hypothesis testing procedures. This Chapter ends describing the open problems that will be working up in the next chapters.

1.1 Basic concepts

In statistics, the main goal is to try to explain the behavior of a phenomenon surrounded by uncertainty. The uncertainty can be raised by epistemic causes or aleatory causes Mena (2015). The epistemic uncertainty is related to the lack of information and it could be reduced with the arrival of new information, whereas the aleatory uncertainty is related to the natural intrinsic variation of the phenomenon. Hence, the essential job in statistics is to find a *model* that describes the behavior of the phenomenon and that reduces the grade of uncertainty about the events of our interest. These events can be expressed in mathematical language through of a probability space. A probability space is a triple $(\Omega, \mathcal{F}, P)$, where Ω is called sample space and denotes the set of all possible outcomes of the phenomena or experiments at issue, $\mathcal{F}$ is a collection of subsets of Ω and constitutes a σ-$field$ or also called as σ-$algebra$. Finally, P is a probability measure ($P : \mathcal{F} \mapsto [0, 1]$) which satisfies $P(A) \geq 0$ for all events $A \in \mathcal{F}$ with $P(\Omega) = 1$ and

$$P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P\left(A_i\right), \tag{1-1}$$

for disjoint events $A_1, A_2, \ldots \in \mathcal{F}$.

Frequently, to define a probabilistic model on $\mathcal{F}$ the events $A_1, A_2, \ldots \in \mathcal{F}$ are converted into numerical quantities called random variables. Strictly speaking, a real value function Y defined on Ω is said to be a random variable if for every Borel set $A \subset \mathbb{R}$ we have $Y^{-1}\{\omega : Y(\omega) \in A\} \in \mathcal{F}$. If Y is a random variable, then Y induces a probability measure on $\mathbb{R}$ called distribution by setting the set function $P_Y(A) = P(Y^{-1}(A))$ for Borel sets A. The distribution of a random variable Y is usually described by its distribution function, $F_y(Y) = P_Y((-\infty, y]) = \mathbb{P}(Y \leq y)$ and satisfied the Theorem (1) (Durret 2010). We denote $\mathcal{Y}$ as the sample space of Y, i.e., the space on the data y lie.

Theorem 1. Any distribution function F has the fowolling propierties:

1. F is nondecreasing
2. $\lim_{y\to\infty} F(y) = 1 \quad \lim_{y\to-\infty} F(y) = 0$
3. F is right continuous, that is, $\lim_{x\downarrow y} F(x) = F(y)$
4. if $F(y-) = \lim_{x\uparrow y} F(x)$ then $F(y-) = P(Y < y)$
5. $P(Y = y) = F(y) - F(y-)$

When the distribution function $F_y(Y) = \mathbb{P}(Y \leq y)$ has the form

$$F(y) = \int_{-\infty}^{y} f(y)\, dy, \tag{1-2}$$

we say that Y has density function $f(y)$.

Definition 1. We define a statistical model as a family of probability distributions on a sample space $\mathcal{Y}$, indexed by values $\theta \in \Theta$, called parameters, that is, unknown quantities, which are the object of the inference. The statistical model is given by

$$\{P_\theta(y), y \in \mathcal{Y}, \theta \in \Theta\},$$

where each P_θ is a probability measure, and Θ is the parameter space. If Θ has finite dimension, $\boldsymbol{\Theta} \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$ and $d < \infty$, then the model is called parametric, and if $\boldsymbol{\Theta}$ has infinite dimension, it is called a nonparametric model.

Now, we discuss two ideas about the parameters that index a statistical model. First, the statistical model transfers to the parameters the uncertainty underlying to a random variable. To explain it, we consider a classical and didactic example "flipping a fair coin", which was taken from Mena (2015), to introduce the idea about how the uncertainty underlying to a random variable Y is transferred to θ.

In this example, we have two possible outcomes, "head" or "tail", then the sample space is given by $\Omega = \{head, tail\} = \{\omega_1, \omega_2\} = \{0, 1\}$ and the σ-*algebra* by $\mathcal{F} = \{\Omega, \{0\}, \{1\}, \emptyset\}$. Consider Y as a random variable that assigns 1 if the outcome is tail and 0 otherwise. Naturally, the Bernoulli model could be considered as a statistical model to Y. Therefore, our interest is on $P_Y(1) = \mathbb{P}(Y(\omega_1) = 1) = \theta$, i.e. a value $\theta \in [0, 1]$. In others words, we have transferred our interest in understanding the phenomenon's uncertainty to the parameter (Mena 2015).

Second, the parameter at a statistical model can be seen as a pattern that explains the data. To show this, let us present an example taken from Orbanz (2013) about a linear regression problem. In this problem, the data should show a clear linear trend and the line that we use to explain the data could be understood as the pattern that dominates to them. If θ is a linear function, the parameter space Θ is the set of linear functions on $\mathbb{R}$, therefore, given a θ the model describes how the dots scatter around the line. Since we consider a simple linear regression model we have a linear function on $\mathbb{R}^2$, that can be specified using two scalars, an offset, and a slope, thus the parameter space has finite dimension, on $\mathbb{R}^2$. However, if the data set not follows a linear pattern, we could think about a continuous regression function and reasonably smooth, i.e. the set of all twice continuously differentiable functions on $\mathbb{R}$, and in this case, the parameter space Θ is infinite dimensional and the statistical model is of nonparametric nature.

According to Mena (2015), only is possible get statistical learning about a data set if these are stochastically connected. This connection let us get a statistical model for our data. Usually, when we consider a random sample $y_1, \ldots, y_n$ as n observations of a random variable Y at a phenomenon, these observations are assumed as distinct realizations of a random variable and also could be considered as physically independent, that is, observations not measured in the same sample unit. Physically independent observations imply symmetry among $y_1, \ldots, y_n$, namely the joint probability distribution is not affected by the order in which the observations were sampled, thus physically independent is a particular case of symmetry or also called exchangeability. In a statistical model, we can have a set of observations physically independent or exchangeable but not stochastically independent, otherwise, the model does not have sense.

Definition 2. A finite set $\{Y_i\}_{i=1}^n$ of random variables is said to be *finite exchangeable* if

$$\{Y_1, \ldots, Y_n\} \stackrel{d}{=} \left\{Y_{\tau(1)}, \ldots, Y_{\tau(n)}\right\}, \tag{1-3}$$

for any permutation τ of $\{1, \ldots, n\}$. An infinite collection $\{Y_i\}_{i=1}^\infty$ is said to be *exchangeable* if every subcollection is exchangeable.

Now, we consider the following theorem called *de Finetti's representation theorem for binary variables.*

Theorem 2. (Diaconis 1977) An infinite sequence of $Y := \{0,1\}$-valued random variables, $\{Y_i\}_{i=1}^{\infty}$, is said to be exchangeable if and only if there exist a distribution q on $[0,1]$ such that for all $n \geq 1$

$$P\left(Y_1 = y_1, \ldots, Y_n = y_n\right) = \int_{[0,1]} \theta^{s_n} \left(1-\theta\right)^{n-s_n} q(d\theta), \tag{1-4}$$

where $s_n := \sum_{i=1}^{n} y_i$ denotes the number of successes. Furthermore, q is such that its cumulative distribution function is

$$q(t) = \lim_{n\to\infty} P\left(\frac{S_n}{n} \leq t\right). \tag{1-5}$$

Namely, $\frac{S_n}{n} \overset{a.s}{\to} \Theta$ with $\Theta := \lim_{n\to\infty} \frac{S_n}{n}$, the (strong-law) limiting relative frequency of 1s.

This theorem says that if we have an exchangeable (rather than independent) binary sequence $\{Y_i\}_{i=1}^{\infty}$ the probabilities $\theta's$ not are only frequencies of infinite numbers of observations, indeed θ could be thought like as random quantity by each realization of $Y_1, \ldots, Y_n$, with distribution function $q(t)$. A general version of the de Finetti's theorem is given by,

Theorem 3. (Hewitt & Savage 1955) Let $\mathbb{Y}$ be a Polish space endowed with is Borel σ-field $\mathcal{B}$ and $\mathcal{P}_{\mathbb{Y}}$ the space of all probability measures on $(\mathbb{Y}, \mathcal{B})$. An infinite sequence of $\mathbb{Y}$-valued random variables, $\{Y\}_{i=0}^{\infty}$ is exchangeable if and only if there exist Q on $(\mathcal{P}_{\mathbb{Y}}, \mathscr{P}_{\mathbb{Y}})$ such that,

$$P\left[Y_1 \in A_1, \ldots, Y_n \in A_n\right] = \int_{\mathscr{P}_{\mathbb{Y}}} \prod_{i=1}^{n} P\left(A_i\right) Q(dP), \tag{1-6}$$

for all $n \geq 1$ and $A_i \in \mathcal{B}$

Theorem 3 can be interpreted as that the exchangeability justifies the existence of a random parameter at a statistical model and hence of its prior distribution Q. This serves as a justification of the subjective view of the Bayesian statistics.

In Bayesian statistics, the parameter is considered as a random variable. The basic principle is that all forms of uncertainty should be random, hence in this context, θ is a random variable with values in the parameter space Θ. Since we are interested in to model how θ is distributed, we start for proposing a specific distribution Q given our prior knowledge about the phenomenon. This knowledge is updated according to the data using the Bayes rule. The distribution Q is called the prior distribution or de Finetti's measure. Thus, the model in this context is

given by

$$Y_1, Y_2, \ldots | \Theta \overset{ind}{\sim} P_\Theta,$$
$$\Theta \sim Q.$$

Our objective is then to determine the posterior distribution, i.e., the conditional distribution of Θ given the data,

$$Q\left[\Theta \in \cdot | Y_1 = y_1, \ldots, Y_n = y_n\right]. \tag{1-7}$$

As a consequence, we have that our grade of uncertainty about the stochastic behavior of $\{Y_i\}_{i=1}^{\infty}$ in the light of observations should decrease if we improve our knowledge about the parameter of the distribution. Indeed, if in the Theorem (3) we consider $\mathcal{P}_{\mathbb{Y}}$ as an infinite dimensional space, then the concept of Bayesian nonparametric takes sense, and as a natural consequence of (3), we have that now we learn about the whole infinite structure and not only on a finite dimensional parameter as in parametric case. Under this approach, the subjectivity in the selection of Q is not completely removed, instead is reduced to just the choice of de Finetti's measure Q.

1.2 Dirichlet process

As in Frequentist statistics, in Bayesian statistics, there are also nonparametric models. Such models have an infinite dimensional parameter space. In this setting, we need to choose a prior distribution on an infinite dimensional space. The most used prior F at the Bayesian nonparametric statistics is the Dirichlet process, which was introduced by Ferguson (1973). This process is denoted by DP (α, F_0), where F_0 is a distribution function and should be thought as a prior guess of F. M is called the mass or precision parameter since, it controls how tightly the prior is around F_0. The following definition was taken from Ferguson (1973),

Definition 3. Let $Z_1, Z_2 \ldots, Z_k$ be independent random variables with Z_j distributed Ga $(M_j, 1)$, where $M_j > 0$ for some $j, j = 1, 2, \ldots, k$. The Dirichlet distribution with parameter $(M_1, \ldots, M_k)$, denoted by Dir $(M_1, \ldots, M_k)$, is defined as the distribution of $(Y_1, \ldots, Y_k)$, where

$$Y_j = \frac{Z_j}{\sum_{i=1}^{k} Z_i} \qquad \text{for} j = 1, 2, \ldots, k. \tag{1-8}$$

This distribution is always singular with respect to Lebesgue measure in k-dimensional space since $Y_1 + Y_2 + \ldots + Y_k = 1$. In addition, if any $M_j = 0$, the corresponding Y_j is degenerated at zero. However, if $M_j > 0$ for all j, the $k-1$ dimensional distribution

of $(Y_1, \ldots, Y_{k-1})$ is absolutely continuous with density

$$f(y_1, \ldots, y_{k-1}|M_1, \ldots, M_k) = \frac{\Gamma(M_1 + \ldots + M_k)}{\Gamma(M_1) \cdots \Gamma(M_k)} \left(\prod_{j=1}^{k-1} y_j^{n_j - 1} \right) \times \left(1 - \sum_{j=1}^{k-1} y_j \right)^{M_k - 1} I_{\mathbb{S}}(y_1, \ldots, y_{k-1}), \quad (1\text{-}9)$$

where $\mathbb{S}$ is the simplex

$$\mathbb{S} = \left\{ (y_1, \ldots, y_{k-1}) : y_j \geq 0, \sum_{j=1}^{k-1} y_j \leq 1 \right\}. \quad (1\text{-}10)$$

For $k = 2$ the expression (1-10) reduces to the Beta distribution denoted by Beta(M_1, M_2).

The main properties of the Dirichlet distribution are:

1. If $(Y_1, \ldots, Y_k) \in \text{Dir}(M_1, \ldots, M_k)$ and $r_1, \ldots, r_l$ are integers such that $0 < r_1 < \ldots < r_l = k$, then,

$$\left(\sum_{i=1}^{r_1} Y_i, \sum_{i=r_1+1}^{r_2} Y_i, \ldots, \sum_{i=r_{(l+1)-1}}^{r_l} Y_i \right) \in \text{Dir} \left(\sum_{i=1}^{r_1} M_i, \sum_{i=r_1+1}^{r_2} M_i, \ldots, \sum_{i=r_{(l+1)-1}}^{r_l} M_i \right).$$

   In particular, the marginal distribution of each Y_i is Beta,

$$Y_i \in \text{Beta} \left(M_j, \left(\sum_{i=1}^{k} M_i \right) - M_j \right).$$

2. If $(Y_1, \ldots, Y_k) \in \text{Dir}(M_1, \ldots, M_k)$ then,

$$E\{Y_i\} = \frac{M_i}{M}, \quad (1\text{-}11)$$

$$E\{Y_i^2\} = \frac{M_i(M_i + 1)}{M(M + 1)}, \quad (1\text{-}12)$$

$$E\{Y_i Y_j\} = \frac{M_i M_j}{M(M + 1)} \qquad \text{for} \quad i \neq j, \quad (1\text{-}13)$$

   where $M = \sum_{i=1}^{k} M_i$.

3. If the prior distribution of $(Y_1, \ldots, Y_k)$ is Dir$(M_1, \ldots, M_k)$ and if

$$P\{X = j|Y_1, \ldots, Y_k\} = Y_j \quad a.s. \quad \text{for} \quad j = 1, \ldots, k,$$

then the posterior distribution of $(Y_1, \ldots, Y_k)$ given $X = j$ is

$$\text{Dir}\left(M_1^{(j)}, \ldots, M_k^{(j)}\right)$$

where $M_1^{(j)} = M_i, \text{if} \quad i \neq j$ or $M_1^{(j)} = M_j + 1, \text{if} \quad i = j$.

Definition 4. Let $M > 0$ and F_0 be a probability measure defined on Ω. A Dirichlet Process (DP) with parameters (M, F_0) is a random probability measure F defined on Ω which assigns probability $F(A)$ to very (measurable) set A such that for each (measurable) finite partition $\{A_1, \ldots, A_k\}$ of Ω the joint distribution of the $(F(A_1), \ldots, F(A_k))$ is the Dirichlet distribution with parameters

$$(MF_0(A_1), \ldots, MF_0(A_k)).$$

The Dirichlet process is commonly used in Bayesian nonparametric as a prior model because it is a distribution over distributions, each realization of the process is itself a distribution. The Dirichlet process has nature discrete, i.e., the random realizations of F are discrete with probability 1.

From Ferguson (1973) we have that if F is a random measure then $F(A)$ is a random variable to any $A \subset \Omega$ and thus $F(A) \sim \text{Beta}\{MF_0(A), M(1 - F_0(A))\}$. Such that,

$$E\{F(A)\} = F_0(A) \text{ and} \tag{1-14}$$

$$\text{Var}\{F(A)\} = \frac{F_0(A)(1 - F_0(A))}{M + 1}. \tag{1-15}$$

From (1-15) we can see that M controls the variability of realizations in the process, thus large values of M reduces the variability of realizations around F_0.

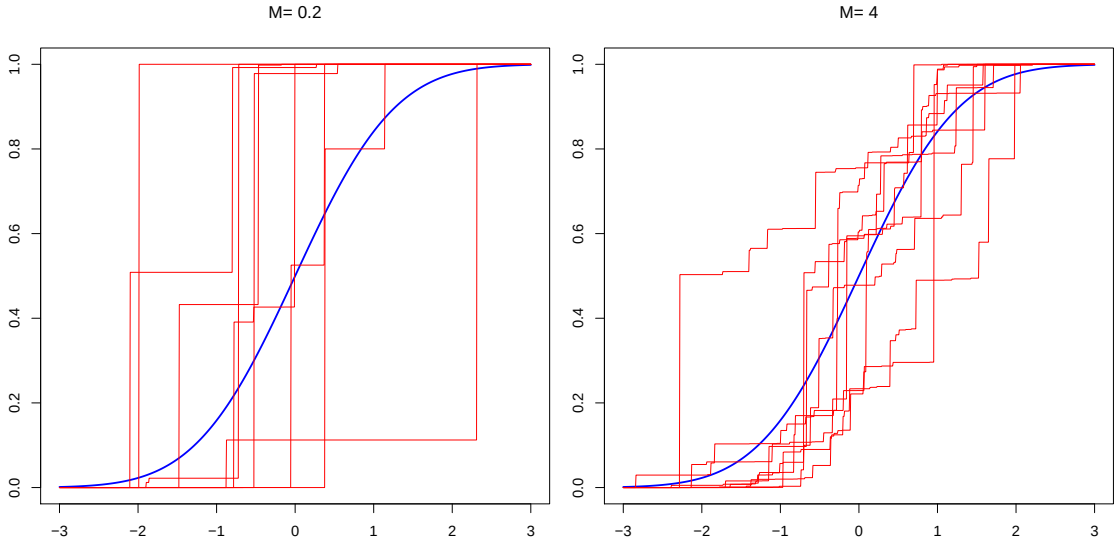
Figure **1-1** illustrates the role of the parameters in the Dirichlet process. Note, for example, when M is small the realizations of the process have bigger deviations from the base measure. When M is big the realizations are more concentrated around the base measure F_0.

Ferguson (1973) showed using Kolmogorov's consistency theorem that the Dirichlet process exists and in addition proved that is conjugate under i.i.d sampling, i.e., if $\theta_1, \ldots, \theta_n$ is an i.i.d sample with $\theta_i | F \sim F$ and $F \sim \text{DP}(M, F_0)$ then,

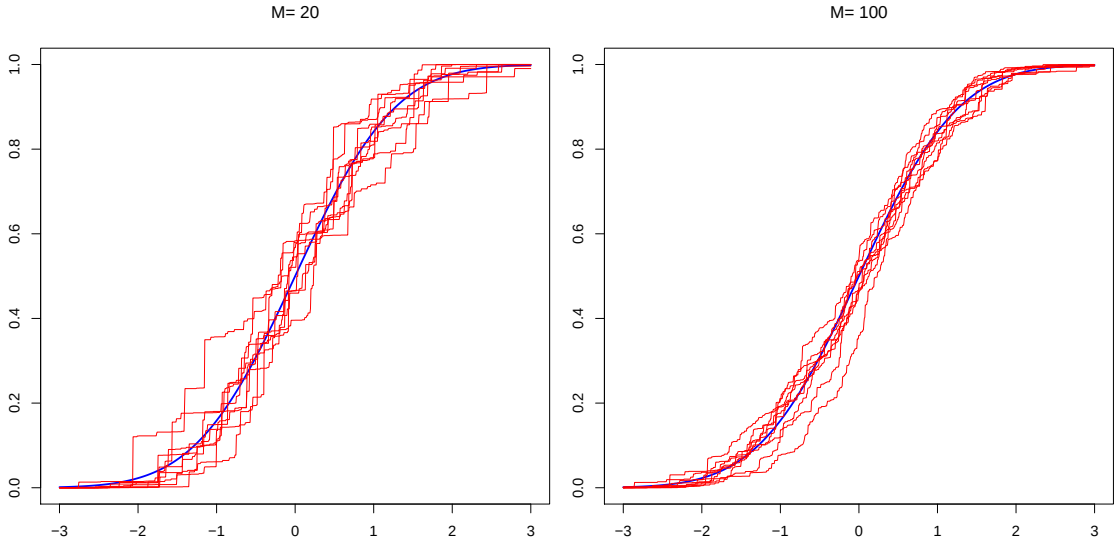
$$F|\theta_1, \ldots, \theta_n \sim \text{DP}\left(M + n, \frac{MF_0 + \sum_{i=1}^n \delta_{\theta_i}}{M + n}\right), \tag{1-16}$$

and the posterior mean is given by,

$$E(F|\theta_1, \ldots, \theta_n) = \frac{MF_0 + \sum_{i=1}^n \delta_{\theta_i}}{M + n}. \tag{1-17}$$



(a) (b)



(a) (b)

Figure **1-1**: Random distributions generated from the Dirichlet process prior with varying mass parameter M. For all cases, the baseline measure corresponds to a standard Gaussian distribution. Each case contains 10 independent realizations with a common value for M. Note that M controls both the variability of the realizations around F_0 and the relative size of the jumps.

The posterior mean is a weighted average between the base distribution F_0 and the empirical distribution function. Since the empirical distribution function is a consistent estimator if $\theta_i's$ are iid, is easy to show that when $n \longrightarrow \infty$, then $F(A)|\theta_1, \ldots, \theta_n \xrightarrow{p} F_T(A)$ for any measurable set A, being F_T the true distribution function, see e.g. Rodríguez & Müller (2013).

On the other hand, given the discrete nature of the Dirichlet process it can be represented as a weighted sum of point masses. Thus, if we have θ_j i.i.d with prior distribution F, which is a random measure such that $F \sim DP(M, F_0)$, then F can be writed as $F(\cdot) = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}(\cdot)$, where $w_1, w_2, \ldots$ are random weights and $\delta_\theta(\cdot)$ denotes the Dirac measure at θ_j. Sethuraman (1994) proposed a particular form to construct the weights w_j, known as the "stick-breaking" construction. This construction considers $w_1 = \nu_1$, and $w_j = \nu_j \prod_{k<j} \{1 - \nu_k\}$ with $\nu_j \sim \text{Beta}(1, M)$ i.i.d. This constructive definition has taken an important role in the developed of research on nonparametric mixture areas, especially, the Dirichlet process mixture model has evidenced an increasing use for modeling complex data, see e.g. Hennig et al. (2015), Airoldi et al. (2019).

1.2.1 Dirichlet mixture model

The Dirichlet process is a famous prior for a random probability measure, its realizations or trajectories are discrete with probability 1, and dependents of two parameters, the mass parameter M and the base measure F_0. Thus, if our random measure is discrete, then the Dirichlet process is an appropriate to model for it, but when the random measure has continuous nature the Dirichlet process is not an option. This limitation can be fixed by convolving the trajectories of the Dirichlet process with some continuous kernel, see, e.g. Müller et al. (2015). This strategy was proposed by Ferguson (1983) and Lo (1984) and later used by Escobar (1988, 1994) and Escobar & West (1995) among many others.

Suppose $\theta \in \Theta$ be the parameter in the statistical model for the random variable Y, where $f_\theta(\cdot)$ is a continuous density and Θ a finite dimensional parameter space, for instance, $f_\theta(\cdot)$ could be a Gaussian density with $\theta = (\mu, \sigma^2)$, and $\Theta = \mathbb{R} \times \mathbb{R}_+$. Thus, given a discrete distribution function on Θ, such that $\theta_i|F \overset{iid}{\sim} F$, where $F \sim DP(M, F_0)$, the random variable Y has a probability density function given by

$$f_F(y) = \int f_\theta(y) F(d\theta). \tag{1-18}$$

The model in (1-18) is known as Dirichlet process mixtures model and denoted by

DPM. This model can be written in a hierarchical way as follow,

$$\begin{array}{c} y_i|\theta_i \overset{ind}{\sim} f_{\theta_i}, \\ \theta_i|F \overset{iid}{\sim} F, \end{array} \tag{1-19}$$

where $F \sim DP(M, F_0)$, θ_i's are conditionally independent given F and the observations y_i's are conditionally independent of the other observations given θ_i.

In the next, section we present the mixture Gaussian model as a particular case from the model in (1-19), also we discussed a strategy for the posterior inference and ending the section with an illustration.

1.3 Density estimation with infinite mixture of Gaussian distributions

In this section, we focus our attention in Dirichlet process mixture with Gaussian kernel, namely, we consider $f_\theta(\cdot)$ as a Gaussian density, $\mathrm{N}(y \mid \theta)$. Then, from (1-18) we have that

$$f_F(y) = \int \mathrm{N}(y|\theta)\, F(d\theta), \tag{1-20}$$

where $\theta = (\mu, \sigma^2)$, $F \sim DP(M, F_0)$ with $M > 0$, and F_0 a distribution on $\mathbb{R} \times \mathbb{R}_+$.

Now, note that given the discrete nature of the Dirichlet process, (1-19) can be written as a weighted infinite sum of $f_{\theta_j}(\cdot)$, $j = 1, 2, \ldots$. Specifically, for a Gaussian kernel, we have the following expression,

$$\begin{aligned} f_F(y) &= \int \mathrm{N}(y|\theta) \sum_{j=1}^{\infty} w_j \delta_{\theta_j}(d\theta) \\ &= \sum_{j=1}^{\infty} w_j \int \mathrm{N}(y|\theta)\, \delta_{\theta_j}(d\theta) \\ &= \sum_{j=1}^{\infty} w_j \mathrm{N}(y \mid \theta_j) \end{aligned} \tag{1-21}$$

In (1-21) we have to draw the weights w_j and atoms $\theta_j = \left(\mu_j, \sigma_j^2\right)$ for estimating of $f_F(y)$. The w_j's are defined on the simplex space, in fact, $w_j > 0$ and $\sum_{j=1}^{\infty} w_j = 1$, $j = 1, 2, \ldots$. They can be obtained by using the stick–breaking representation. Here, obviously the essential problem is how to sample an infinite number of weights and atoms. In next section, we discuss a strategy for posterior inference in the model (1-21).

1.3.1 Posterior Inference

Several strategies have been proposed for sampling DPM models. The first Gibbs sampler for the DPM model was proposed by Escobar (1988, 1994). Many variations of the Escobar's algorithm have been developed, see for example MacEachern (1994), MacEachern & Müller (1998), Neal (2000). All these algorithms work in the solution of the integral in (1-20), which is a marginal integrate on F, i.e., over a random distribution function, and for this reason, they are called *marginal* methods. In this way, they left out the problem of sampling infinite dimensional parameters. However, some authors like Ishwaran & James (2001), Papaspiliopoulos & Roberts (2008) and Walker (2007), have proposed strategies to sampling from (1-21), but with a finite numbers of components in each iteration of a Markov Chain with a correct stationary distribution, for instance, Walker (2007) resorts to ideas from slice sampling to construct a Gibbs sampling algorithm based on four simple steps. These latter methods are called *conditional* methods.

The slice sampling method to DPM proposed by Walker (2007) uses an ingenious idea, his proposal introduces latent variables to transform (1-21) in a finite sum. Let us define U be a latent variable, such that,

$$f_{w,\theta}(y,u) = \sum_{j=1}^{\infty} \mathbf{1}(u < w_j)\, \mathbf{N}(y|\theta_j)\,. \tag{1-22}$$

We have changed the notation f_F by $f_{w,\theta}$ to indicate dependence of density function on w and $\theta = (\mu, \sigma^2)$. Note that, when we introduce a latent variable u we solve the problem of infinite terms at sum, because only a finite number of w_j satisfies the condition $(u < w_j)$. Walker (2007) introduced this variable without modifies the original density, since if we marginalize over u in (1-22) we return to (1-21). Alternatively, we can write (1-22) as

$$f_{w,\theta}(y,u) = \sum_{j=1}^{\infty} w_j \mathrm{Unif}(u|0,w_j)\, \mathrm{N}(y|\theta_j)\,, \tag{1-23}$$

where with probability w_j, Y and U are independent. Furthermore, Y and U are Gaussian and Uniform distributed, respectively. Hence, the marginal density for U is given by

$$f_{w,\theta}(u) = \sum_{j=1}^{\infty} w_j \mathrm{Unif}(u|0,w_j) = \sum_{j=1}^{\infty} \mathbf{1}(u < w_j)\,. \tag{1-24}$$

Equivalently, if we consider the set $A_w = \{j : w_j > u\}$ we can rewrite (1-22) as,

$$f_{w,\theta}(y,u) = \sum_{j \in A_w(u)} \mathrm{N}(y \mid \theta_j)\,, \tag{1-25}$$

where $A_w(u)$ is a finite set for all $u > 0$. Then the conditional density for y given u is

$$f_{w,\theta}(y|u) = \frac{1}{f_w(u)} \sum_{j \in A_w(u)} \mathbf{N}(y|\theta_j), \tag{1-26}$$

where $f_w(u) = \sum_j \mathbf{1}(u < w_j)$ is the marginal density for u, with $u \in (0, w^*)$ and where w^* is the largest w_j. Thus, given a latent variable u, we have a finite mixture model with equal weights, equal to $1/f_w(u)$.

On the other hand, since a DPM model induce the existence of clusters, Walker (2007) introduces an additional variable d_i to indicates if y_i belongs to cluster k, i.e. the new latent variable indicates which of these finite clusters provides the observation, hence joint density distribution of (y_i, u_i, d_i) is given by

$$f_{w,\theta}(y_i, u_i, d_i) = \mathbf{1}(u_i < w_{d_i}) \mathbf{N}(y|\theta_d). \tag{1-27}$$

For a sample random $y_1, \ldots, y_n$ our likelihood function will take the form

$$\prod_{i=1}^{n} \mathbf{1}(u < w_{d_i}) \mathbf{N}(y_i|\theta_{d_i}), \tag{1-28}$$

with $\theta_{d_i} = \left(\mu_{d_i}, \sigma^2_{d_i}\right)$. Kalli et al. (2011) summary at four steps the algorithm for the posterior inference in the augmented model model. The variables to sampler via the Gibbs algorithm are

$$\left\{\left(\mu_j, \sigma_j^2, \nu_j\right), j = 1, 2, \ldots; (d_i, u_i), i = 1, \ldots, n\right\}. \tag{1-29}$$

Next, we enumerate the Gibbs sampling steps.

1. Sampling μ_j and σ_j from $f\left(\mu_j, \sigma_j^2|\ldots\right) \propto p_0\left(\mu_j, \sigma_j^2\right) \prod_{d_i=j} \mathbf{N}\left(y_i|\mu_j, \sigma_j^2\right)$.

2. Drawing ν_j from $\text{Beta}(\nu_j|a_j, b_j)$, where

$$a_j = 1 + \sum_{i=1}^{n} \mathbf{1}(d_i = j),$$

and

$$b_j = M + \sum_{i=1}^{n} \mathbf{1}(d_i > j),$$

and compute the weights w_j's by "stick breaking" representation.

3. Sampling latent variable u from uniform distribution,

$$f(u_i|\ldots) \propto \mathbf{1}(0 < u_i < w_{di}).$$

4. Sampling a number k of clusters for which $w_k > u_i$ from $P(d_i = k|\ldots) \propto \mathbf{1}(k : w_k > u_i)\,\mathbf{N}(y_i|\mu_k, \sigma_k^2)$. To complete this stage, we consider k as a set $\{1, \ldots, N\}$ where $N = \max_i N_i$ and N_i is the largest integer l for which $w_l > u_i$, and we have to sample up to the integer N.

Finally, we include an additional step in order to update the mass parameter M. For this, we follow the methodology discussed by Escobar & West (1995). In consequence, we have for any $k = 1, \ldots, n$, that the conditional distribution for M is given by

$$f(M|k) \propto f(M)\, M^{k-1}\,(M+n)\,\beta(M+1, n), \tag{1-30}$$

where $\beta(\cdot,\cdot)$ is the usual beta function. Then, equivalently, (1-30) can be written as

$$f(M|k) \propto f(M)\, M^{k-1}\,(M+n)\int_0^1 x^{\alpha}\,(1-x)^{n-1}\,\mathrm{dx}. \tag{1-31}$$

This suggests that $f(M|k)$ is the marginal distribution from a joint for M and a continuous quantity η such that

$$f(M, \eta|k) \propto f(M)\, M^{k-1}\,(M+n)\,\eta^{M}\,(1-\eta)^{n-1}, \tag{1-32}$$

for $M > 0$ and $0 < \eta < 1$. From (1-32), we can compute the conditional posterior for M and for η, thus if we assume a Gamma prior distribution, $\mathrm{Ga}(c, d)$, for the mass parameter, M is drawn from a mixture of two gamma densities,

$$f(M|\eta, k) = \tau_\eta \mathrm{Ga}(c+k, d-\log(\eta)) + (1-\tau_\eta)\,\mathrm{Ga}(c+k-1, d-\log(\eta)) \tag{1-33}$$

with weights τ_η defined by ${}^{\tau_\eta}/_{(1-\tau_\eta)} = {}^{(c+k-1)}/_{(n(d-\log(\eta)))}$. From (1-32) the variable η has conditional density, $f(\eta|M, k) \propto \eta^{M}\,(1-\eta)^{n-1}$, $0 < \eta < 1$. Then, $f(\eta|M, k) \sim \mathrm{Beta}(M+1, n)$.

1.3.2 An Illustration

In this section, we provide illustrations of density estimation with infinite mixture using the Walker's algorithm. The data are simulated from variations of the model

$$Y_i \sim w_1 \mathrm{N}\left(y \mid \mu_1, \sigma_1^2\right) + w_2 \mathrm{N}\left(y \mid \mu_2, \sigma_2^2\right) + w_3 \mathrm{N}\left(y \mid \mu_3, \sigma_3^2\right). \tag{1-34}$$

The hyperparameters for the Gaussian prior distribution of μ_j, $j = 1, 2, 3$, was fixed at mean 0 and variance $1/s$, with $s = 0.1$. For the parameter precision $\lambda_j = 1/\sigma_j^2$, $i = 1, 2, 3$, we assume a Gamma prior distribution, $\mathrm{Ga}(\epsilon, \epsilon)$ with $\epsilon = 0.5$. We draw the atoms $\theta_j = (\mu_j, \sigma_j)$ from the conditionals

$$f(\mu_j|\ldots) = N\left(\frac{\xi_j \lambda_j}{m_j \lambda_j + s}, \frac{1}{m_j \lambda_j + s}\right), \tag{1-35}$$

$$f\left(\lambda_j|\ldots\right) = \text{Ga}\left(\epsilon + \frac{m_j}{2}, \epsilon + \sum_{i=1}^{n} \mathbf{1}\left(d_i = j\right)\left(y_i - \mu_j\right)^2\right), \tag{1-36}$$

where $\xi_j = \sum_{i=1}^{n} y_i \mathbf{1}\left(d_i = j\right)$ and $m_j = \sum_{i=1}^{n} \mathbf{1}\left(d_i = j\right)$. On the other hand, for the mass parameter M we fixed the hyperparameters for the gamma distribution at $c = 2$ and $d = 10$. The table (**1-1**) presents the simulation cases used in the illustration of the density estimation procedure.

Scenarios	ω_i	μ_1	σ_1	μ_2	σ_2	μ_3	σ_3
I	0.5 0.5 0.0	-2	1	-2	1		
II	0.7 0.3 0.0	-2	1	2	1		
III	0.25 0.25 0.50	-4	1	0	1	8	1
IV	0.05 0.15 0.80	-4	1	0	1	8	1

Table **1-1**: Parametrization of model (1-34), scenarios I to IV show the variations of the model used to illustrate the density estimation.

The figure **1-2** displays the result of the fit in each simulated scenario via a Dirichlet mixture model. We consider a sample size $n = 100$ for the four scenarios. The base measure for the Dirichlet process, F_0, was assumed as a Gaussian distribution.

Next, we introduce some aspects of hypothesis testing from a Bayesian point of view as well some elements of Bayesian variable selection in the regression model, which will be key ingredients in the setting of our procedure.

1.4 Hypothesis Testing and Model Selection

In this section, we follow some basic ideas exposed by Berger (1985) and Ghosh et al. (2007) for introducing the hypothesis testing procedures under a Bayesian approach of inference. For starting, suppose that a random variable Y, with density $f\left(y \mid \boldsymbol{\theta}\right)$ is observed with $\boldsymbol{\theta}$ an unknown element of the parameter space Θ, and that we are interested in testing

$$\begin{aligned} H_0 &: Y \text{ has density } f\left(y \mid \boldsymbol{\theta}\right) \text{ where } \boldsymbol{\theta} \in \Theta_0, \\ H_1 &: Y \text{ has density } f\left(y \mid \boldsymbol{\theta}\right) \text{ where } \boldsymbol{\theta} \in \Theta_1. \end{aligned} \tag{1-37}$$

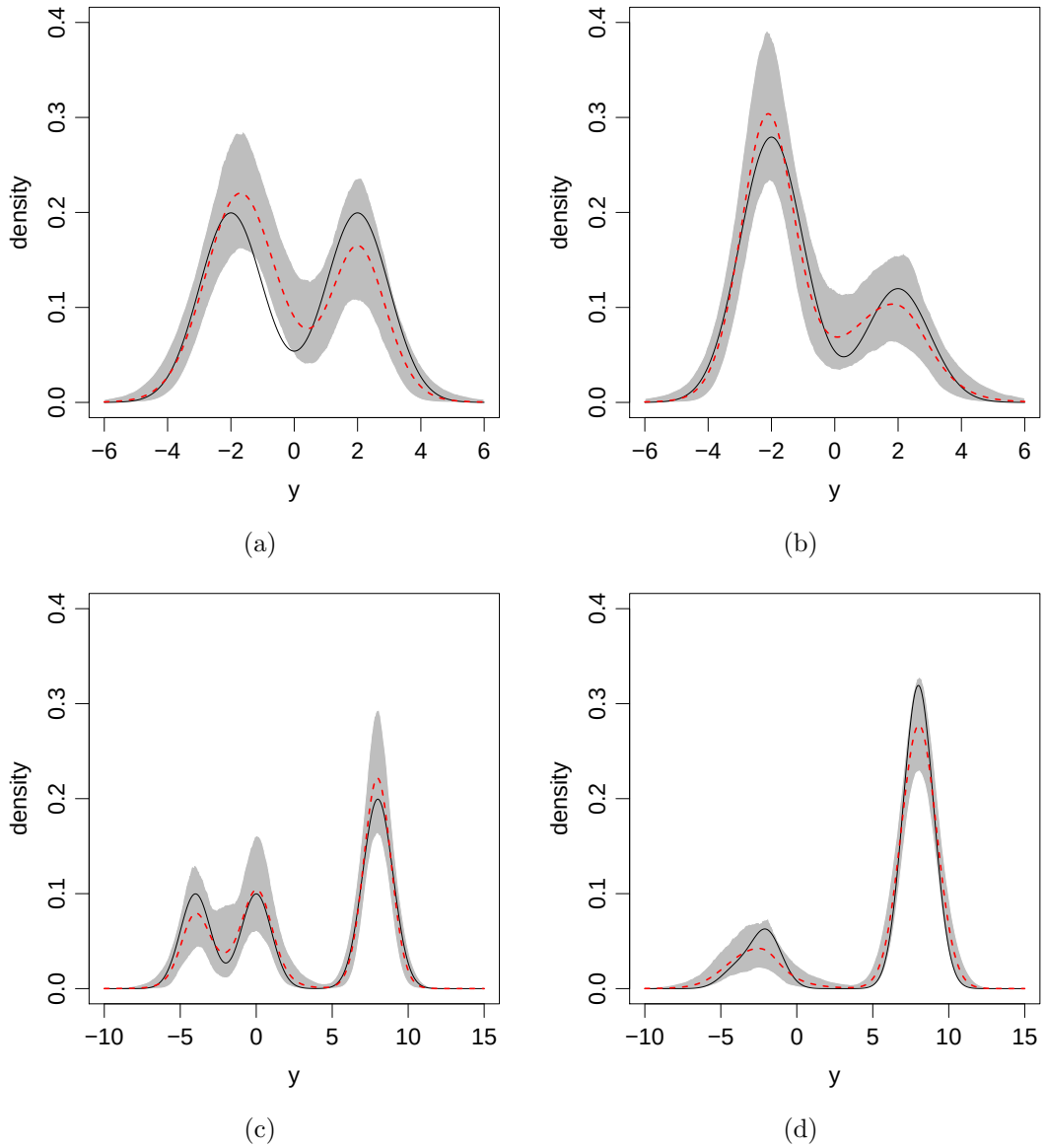


Figure **1-2**: Density estimation for scenarios I to IV. (a) Scenario I, (b) Scenario II, (c) Scenario III, (d) Scenario IV. The dotted red line is the posterior mean $\hat{f}_{w,\theta}$, while the black continuous line is the "true" density. The gray regions correspond to 0.95 point-wise credibility sets.

In classical hypothesis testing, this procedure is developed concerning two types of errors, the error type I and type II, which can be interpreted as the chance with an observed sample the test statistic leads to take a wrong decision. In Bayesian analysis, the goal is to compute the posterior probability for H_0 and H_1 and then takes a decision accordingly. To that end, we suppose that $g_i(\boldsymbol{\theta})$ is the prior distribution of θ_i, conditional on $H_i,\ i = 0, 1$, then we can decide between H_0 and H_1 using the quantity known as Bayes Factor (BF),

$$BF(\boldsymbol{y}) = \frac{h_0(\boldsymbol{y})}{h_1(\boldsymbol{y})}, \tag{1-38}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$ denotes a random sample from Y and

$$h_i = \int_{\Theta_i} f(\boldsymbol{y} \mid \boldsymbol{\theta})\, g_i(\boldsymbol{\theta})\, d\boldsymbol{\theta} \quad i = 0, 1 \tag{1-39}$$

If, further we consider $\pi_0 = P(\Theta_0)$ and $\pi_1 = 1 - P(\Theta_0)$ as the prior to H_0 and H_1, respectively, then the posterior odds ratio of H_0 relative to H_1,

$$\left(\frac{\pi_0}{1-\pi_0}\right) BF(\boldsymbol{y}), \tag{1-40}$$

could be used to decide about H_0. However, the former computations may not be easy to perform depending on the complexity of the model. A possible solution is to approximate the BF via the Bayesian Information Criterion (BIC) as proposed by Schwarz (1978) or to resort to other approximations in the MCMC posterior algorithm.

1.4.1 Bayesian Variable Selection in Regression Models

Variable selection is a traditional problem in Statistic, and especially in regression models. The goal is to select from a list of regressors those that have a significant effect on the response variable, and therefore which should be included in the model. In Bayesian variable selection, many methods have been proposed (Mitchell & Beauchamp 1988, George & McCulloch 1993, Chipman 1996, Geweke 1996, Kuo & Mallick 1998, Chipman et al. 2001, Ishwaran & Rao 2003, 2005), most of them based on the mixture of two components as the prior distribution for the parameters in the regression model. This prior is a mixture of a spike and slab distributions, the spike concentrated at zero and the slab comparably flat.

The Section follows the ideas exposed by Ishwaran & Rao (2005), Wagner & Malsiner-Walli (2011) to explain the role of the spike-slab prior in the variable selection procedure. To begin, let us consider the standard linear regression model,

given by

$$Y_i = \beta_0 + \beta_1 \mathrm{x}_{i,1} + \ldots + \beta_K \mathrm{x}_{i,k} + \epsilon_i, \quad i = 1, \ldots, n \tag{1-41}$$

with n independent responses Y_i and K-dimensional covariates $\mathrm{x}_i = (\mathrm{x}_{i,1}, \ldots, \mathrm{x}_{i,K})^T$, where $\epsilon_i \overset{\text{ind}}{\sim} \mathrm{N}(0, \sigma^2)$. The goal is to find the subset of parameters $(\beta_1, \ldots, \beta_q)^T$ nonzero. Here, the spike-slab distributions are used as prior on the parameters for being selected. To specify the spike-slab prior on the parameter vector $(\beta_1, \ldots, \beta_K)^T$ an indicator vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)$ is defined, where δ_K takes the value of 1 if β_K is allocated to the slab component and zero if it is assigned to the spike. Thus, if the parameters β_k are assumed independent a priori, then spike-slab prior for the parameter vector $\boldsymbol{\beta}$ can be specified as

$$p(\boldsymbol{\beta} \mid \boldsymbol{\delta}) = \prod_{j:\delta j'=1} p_{\text{slab}}(\beta_{\delta j'}) \times \prod_{j:\delta j=0} p_{\text{spike}}(\beta_{\delta j}), \tag{1-42}$$

where p_{slab} and p_{spike} denote the univariate slab and spike distribution, respectively, and $p(\delta_i = 1 \mid w_j) = w_j$ with $w_j \sim \text{Beta}(a_w, b_w)$. Two types of spike distributions are common in the Bayesian literature, absolutely continuous and a point mass at zero, so called Dirac spikes.

Absolutely continuous Spike

An absolutely continuous spike prior can be any continuous unimodal distribution with mode at zero (Wagner & Malsiner-Walli 2011). A popular version of the spike-slab prior, where this two components are specified from the same distribution was proposed by George & McCulloch (1993), here the prior distribution is a mixture of two Gaussian distributions

$$\beta_k \mid \zeta_k \overset{\text{ind}}{\sim} (1 - \zeta_k)\, \mathrm{N}\left(0, \tau_k^2\right) + \zeta_k \mathrm{N}\left(0, c_k^2 \tau_k^2\right) \quad k = 1, \ldots, K, \tag{1-43}$$

where $\tau_k^2 > 0$ is some suitably small value, and $c_k > 1$ is some suitably large value. In consequence, if $\zeta_k = 1$ then β_k will have a prior with large hypervariance and the values drawn in the posterior distribution will be large for β_k. The opposite occurs when $\zeta = 0$. As highlighted in George & McCulloch (1993, 1997), the choice of τ_k^2 and c_k are not an easy task. They suggest choice this hyperparameters of such that $\text{Var}_{spike(\beta_k)}/\text{Var}_{slab(\beta_k)} \ll 1$. Thus, if $\beta_k \sim \mathrm{N}(0, \tau_k^2)$, then β_k can be *safely* replaced by 0 (George & McCulloch 1993). Under this setting, usually, the random variables δ_k are taken as independent from a Bernoulli (w_k) where $0 < w_k < 1$. To deal with the laborious task of choose values for c_k and τ_k^2, Ishwaran & Rao (2000) proposed to relax the choosing of this couple of parameters through a continuous bimodal prior distribution for $\gamma_k = \zeta \tau_k^2$. As a result, they defined the following prior hierarchical

for β_k

$$\begin{aligned} \beta_k \mid \zeta_k, \tau_k^2 &\overset{\text{ind}}{\sim} \mathrm{N}\left(0, \zeta_k \tau_k^2\right), \quad k = 1, \ldots, K \\ \zeta_k \mid \nu, w &\overset{\text{i.i.d}}{\sim} (1-w)\delta_{\nu_0}(\cdot) + w\delta_1(\cdot), \\ \tau_k^{-2} \mid a_1, a_2 &\overset{\text{i.i.d}}{\sim} \mathrm{Gamma}(a_1, a_2), \\ w &\sim \mathrm{Uniform}[0,1]. \end{aligned} \tag{1-44}$$

In (1-44) ν_0 is a small value near zero and the hyperparameters a_1 and a_2 are chosen so that $\gamma_k = \zeta\tau_k^2$ has a continuous bimodal distribution with spike at ν_0 and a right-continuous tail. The ν_0 parameter is important because it allows shrinking the hypervariance and obtains posterior values for β_k near zero, while the right tail is used to identify nonzero β_k. In this hierarchical model the parameter w control the size of the model, since its value has a link with how likely is that ζ_k equals to 1 or ν_0 (Ishwaran & Rao 2003, 2005). The prior proposed in (1-44) is known as a Gaussian mixture of inverse Gamma distributions. Here, marginally both spike and slab components are student distributed, $p_{spike}(\beta_k) = t_{2a_1}(0, \nu_0 {}^{a_2}/_{a_1})$ and $p_{slab}(\beta_k) = t_{2a_1}(0, {}^{a_2}/_{a_1})$. Later, Ishwaran & Rao (2005), in order to control the reduction of the prior effect, as occur for any prior when the sample size increases, propose to introduce a rescaled spike-slab prior that involves replacing the responses Y_i with ones transformed by $\sqrt{n}$ factor. As a consequence, they have to include in the model a factor for adjusting the variance in the new data.

Dirac Spike

The Dirac spike is defined as $P(\beta_k \mid \zeta_k = 0) = \delta_0(\beta_k)$, thus, any mixture of a Dirac delta that put mass at zero with a slab continuous distribution, could be considered as a spike-slab prior distribution. Here, the prior distribution for β_k is given by,

$$\beta_k \mid \zeta_k \overset{\text{ind}}{\sim} (1-\zeta_k)\,\delta_0(\cdot) + \zeta_k \mathrm{N}\left(0, \tau_k^2\right) \quad k = 1, \ldots, K, \tag{1-45}$$

where $\delta_0(\cdot)$ denotes the Dirac-measure at zero and $\tau_k^2 > 0$. As before, ζ_k is a Bernoulli variable that takes values 0 or 1, with $P(\zeta_k = 1 \mid w) = w$ and $0 < w < 1$. Bayesian variable selection with spike-slab priors can be developed via MCMC methods Wagner & Malsiner-Walli (2011). In particular, to simulate from the Dirac spike it is necessary to compute the marginal likelihood, namely, integrating over the parameter involved in the selection, in each step of the Gibbs sampling. On the contrary, this is not necessary when the spike prior is an absolutely continuous distribution. However, under this last, an approximation of $P(\beta_k = 0 \mid \text{Data})$ is provided (George & McCulloch 1993).

Once the prior distribution on the parameters of the model has been defined, the coming step is to choose the prior distribution on the model space. In the

next section, we discuss the prior distribution on the model space to complete the background necessary to define a Bayesian hypothesis procedure properly.

1.4.2 Prior distribution on space of models

In this section, we follow the ideas exposed by Taylor-Rodríguez et al. (2016), we present priors distribution usually used in variable selection under a hierarchical condition in the model. The relevance of respecting the polynomial hierarchical among variables in model variable selection is explained by Peixoto (1990). A model accomplishes with a strong hierarchical condition if for any predictor in the model, every lower-order predictors associated with it is included in the model (Griepentrog et al. 1982, Peixoto 1987, 1990, McCullagh & Nelder 1989, Nelder 2000). Bayesian variable selection, in models for which the hierarchical structure is respected, consists in the comparison of models $\tilde{\mathsf{m}}$ in a model space $\mathcal{M}$ through their posterior probabilities, which are given by $p\left(\tilde{\mathsf{m}} \mid \boldsymbol{y}, \mathcal{M}\right) \propto \pi\left(\boldsymbol{y} \mid \tilde{\mathsf{m}}\right) \pi_{\tilde{\mathsf{m}}}\left(\tilde{\mathsf{m}} \mid \mathcal{M}\right)$. So that, these probabilities depend on the prior on the model space as well as on the prior on the model-specific parameters. Different prior structures on model spaces that accomplish the weak and the strong Hierarchical condition are developed in Taylor-Rodríguez et al. (2016). For instance, let us consider as predictors in a regression model the intercept, x_1, x_2 and the interaction $\mathrm{x}_1\mathrm{x}_2$. Then, a model that includes the intercept, x_1, and $\mathrm{x}_1\mathrm{x}_2$, is a model that satisfies the weak hierarchical condition, whereas if it also includes x_2, then it satisfies the strong hierarchical (SH) condition . In particular, for this thesis, our interest is on models that satisfy the strong hierarchical condition. Under the SH condition, the posterior concentrates on a single best model within the model space, while under Weak Hierarchical (WS) condition does not occur Taylor-Rodríguez et al. (2016).

Next, we provide five prior specifications on models space proposed by Taylor-Rodríguez et al. (2016). For starting, we introduce the essential elements and notation used in these definitions. Let us consider the polynomial regression model,

$$\boldsymbol{y} = \sum \beta_{(\alpha_1,\ldots,\alpha_k)} \prod_{j=1}^{k} \mathrm{x}_j^{\alpha_j} + \boldsymbol{\epsilon}, \tag{1-46}$$

where $\boldsymbol{y}$ is the vector of observations, $(\alpha_1, \ldots, \alpha_k) = \boldsymbol{\alpha} \in \mathbb{N}_0^k$, where $\mathbb{N}_0^k$ is the set of natural numbers including 0, $\boldsymbol{\epsilon} \sim \mathrm{N}\left(0, \sigma^2 \boldsymbol{I}\right)$. Thus, for $\mathrm{x}_1^2\mathrm{x}_2$ the vector $\boldsymbol{\alpha}$ is equal to $(2, 1)$, and the order of a term in the model is given by $\sum \alpha_j$, therefore the order for $\mathrm{x}_1^2\mathrm{x}_2$ is 3. In (1-46) we denote the base model by $\tilde{\mathsf{m}}_B$, which consist of terms that are not subject to selection and is nested in the full model, denoted by $\tilde{\mathsf{m}}_F$.

In order to define the prior probability for $\tilde{\mathsf{m}}$, Taylor-Rodríguez et al. (2016) used the assumptions of conditional independence and immediate inheritance, exposed by

Chipman (1996). In consequence, if two nodes, $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ have the same order j, then $\gamma_{\boldsymbol{\alpha}}$ and $\gamma_{\boldsymbol{\alpha}'}$ are assumed conditionally independent given $\boldsymbol{\gamma}^{<j}(\tilde{\mathsf{m}}) = \cup_{v=0}^{j-1}\boldsymbol{\gamma}^{v}(\tilde{\mathsf{m}})$, where $\boldsymbol{\gamma}_{\alpha}$ is an indicator function describing whether $\boldsymbol{\alpha}$ is included in $\tilde{\mathsf{m}}$. On the other side, the immediate inheritance is defined as the probability that the node $\boldsymbol{\alpha}$ with order j be included in the model $\tilde{\mathsf{m}}$ given that it contains all the lower-order predictors associated to it, this set is known as the parent set for $\boldsymbol{\gamma}$ and denoted by $\mathcal{P}(\boldsymbol{\alpha})$, formally, $\pi\left(\gamma_{\boldsymbol{\alpha}}(\tilde{\mathsf{m}}) = 1 \mid \boldsymbol{\gamma}^{<j}(\tilde{\mathsf{m}}), \mathcal{M}\right) = \pi\left(\gamma_{\boldsymbol{\alpha}}(\tilde{\mathsf{m}}) = 1 \mid \boldsymbol{\gamma}_{\mathcal{P}(\boldsymbol{\alpha})}(\tilde{\mathsf{m}}), \mathcal{M}\right)$. Under the assumptions of conditional independence and immediate inheritance, the probability of $\tilde{\mathsf{m}}$ is

$$\pi(\tilde{\mathsf{m}} \mid \pi_{\mathcal{M}}, \mathcal{M}) = \prod \pi_{\boldsymbol{\alpha}}(\tilde{\mathsf{m}})^{\gamma_{\boldsymbol{\alpha}}(\tilde{\mathsf{m}})} \left(1 - \pi_{\boldsymbol{\alpha}}(\tilde{\mathsf{m}})\right)^{(1-\gamma_{\boldsymbol{\alpha}}(\tilde{\mathsf{m}}))},$$

with $\pi_{\mathcal{M}} = \{\pi_{\boldsymbol{\alpha}}(\tilde{\mathsf{m}}) : \boldsymbol{\alpha} \in \Upsilon(\tilde{\mathsf{m}}_F), \tilde{\mathsf{m}} \in \mathcal{M}\}$ and where $\Upsilon(\tilde{\mathsf{m}})$ and $\mathcal{C}(\tilde{\mathsf{m}})$ are used to denote the sets $\Upsilon(\tilde{\mathsf{m}}) = \tilde{\mathsf{m}} \backslash \tilde{\mathsf{m}}_B$ and

$$\mathcal{C}(\tilde{\mathsf{m}}) = \{\boldsymbol{\alpha} \in \Upsilon(\tilde{\mathsf{m}}_F) \backslash \Upsilon(\tilde{\mathsf{m}}) : \tilde{\mathsf{m}} \cup \{\boldsymbol{\alpha}\} \text{ satisfies the SH condition}\}.$$

Then, under the SH condition, $\pi_{\boldsymbol{\alpha}}(\tilde{\mathsf{m}}) = \mathbf{0}$ if $\boldsymbol{\gamma}_{\mathcal{P}(\boldsymbol{\alpha})}(\tilde{\mathsf{m}}) = \mathbf{0}$. Below, we present the prior distribution proposed by Taylor-Rodríguez et al. (2016).

Hierarchical Uniform Prior (HUP)

The HUP assumes that nonzero probabilities on the space of models are all equals. Specifically, for a model $\tilde{\mathsf{m}} \in \mathcal{M}$ is assumed that the prior $\pi_{\alpha}(\tilde{\mathsf{m}}) = \pi$ for all $\boldsymbol{\alpha} \in \Upsilon(\tilde{\mathsf{m}}) \cup \mathcal{C}(\tilde{\mathsf{m}})$. Then, if we complete the Bayesian formulation with $\pi \sim \text{Beta}\,(a, b)$ we have

$$\pi^{HUP}\left(\tilde{\mathsf{m}} \mid \mathcal{M}, a, b\right) = B\left(|\Upsilon(\tilde{\mathsf{m}})| + a, |\mathcal{C}(\tilde{\mathsf{m}})| + b\right) / B(a, b),$$

where B is the beta function. The HUP assigns equal probabilities to all models for which the sets $\Upsilon(\tilde{\mathsf{m}})$ and $\mathcal{C}(\tilde{\mathsf{m}})$ have the same cardinality.

Hierarchical Independence Prior (HIP)

This prior assumes that all $\pi_{\alpha}(\tilde{\mathsf{m}})$ are different, such that $\pi_{\alpha}(\tilde{\mathsf{m}}) \sim \text{Beta}\,(a_{\alpha}, b_{\alpha})$, then the prior probability of $\tilde{\mathsf{m}}$ under the HIP is

$$\pi^{HIP}\left(\tilde{\mathsf{m}} \mid \mathcal{M}, \boldsymbol{a}, \boldsymbol{b}\right) = \left(\prod_{\boldsymbol{\alpha} \in \Upsilon(\tilde{\mathsf{m}})} \frac{a_{\alpha}}{a_{\alpha} + b_{\alpha}}\right) \left(\prod_{\boldsymbol{\alpha} \in \mathcal{C}(\tilde{\mathsf{m}})} \frac{b_{\alpha}}{a_{\alpha} + b_{\alpha}}\right),$$

where the product over the empty set is assumed equal to 1. Under the SH condition, the HIP with parameters $a_{\alpha} = b_{\alpha} = 1$ coincides with the Chipman's prior, where the conditional inclusion probability is 0.5 for each term.

Hierarchical Order Prior (HOP)

This prior assumes equality between the nonzero $\pi_{\boldsymbol{\alpha}}(\tilde{\mathsf{m}})$ with the same order and independence across the different orders. Define $\Upsilon_j(\tilde{\mathsf{m}}) = \{\boldsymbol{\alpha} \in \Upsilon(\tilde{\mathsf{m}}) : \text{order}(\boldsymbol{\alpha}) = j\}$ and $\mathcal{C}_j(\tilde{\mathsf{m}}) = \{\boldsymbol{\alpha} \in \mathcal{C}(\tilde{\mathsf{m}}) : \text{order}(\boldsymbol{\alpha}) = j\}$. Here, $\pi_{\boldsymbol{\alpha}}(\tilde{\mathsf{m}}) = \pi^{(j)}(\tilde{\mathsf{m}})$ for all $\boldsymbol{\alpha} \in \Upsilon_j(\tilde{\mathsf{m}}) \cup \mathcal{C}_j(\tilde{\mathsf{m}})$, and $\pi^{(j)}(\tilde{\mathsf{m}}) \sim \text{Beta}\,(a_j, b_j)$ provides a prior probability, then

$$\pi^{HOP}\left(\tilde{\mathsf{m}} \mid \mathcal{M}, \boldsymbol{a}, \boldsymbol{b}\right) = \prod_{j=\mathcal{J}_{\mathcal{M}}^{min}}^{\mathcal{J}_{\mathcal{M}}^{max}} \left(\frac{B\left(|\Upsilon_j(\tilde{\mathsf{m}})| + a_j, |\mathcal{C}_j(\tilde{\mathsf{m}})| + b_j\right)}{\left(B\left(a_j, b_j\right)\right)} \right),$$

if $a_j = b_j = 1$ we have the hierarchical version of the Scott & Berger (2010) multiplicity correction. Now, if we consider $a_j = 1$ and $b_j = |\Upsilon_j(\tilde{\mathsf{m}}) \cup \mathcal{C}_j(\tilde{\mathsf{m}})|$, then we have the hierarchical penalization introduced by Wilson et al. (2010). This parametrization produces a penalization more strong as the model becomes more complex.

Hierarchical Length Prior (HLP) and Hierarchical Type Prior (HTP)

These priors penalize by the number of nodes that $\boldsymbol{\alpha}$ has connection, namely, whose nodes with more connections in a model's graph should be penalized differently from nodes that have fewer connections. These priors are equivalent when maximum order in the $\tilde{\mathsf{m}}_F$ is 3. In particular, for the HLP, the set of nodes in $\Upsilon_j(\tilde{\mathsf{m}}) \cup \mathcal{C}_j(\tilde{\mathsf{m}})$ are given by

$$\left(\Upsilon_j(\tilde{\mathsf{m}}) \cup \mathcal{C}_j(\tilde{\mathsf{m}})\right)_\ell = \{\boldsymbol{\alpha} \in \Upsilon_j(\tilde{\mathsf{m}}) \cup \mathcal{C}_j(\tilde{\mathsf{m}}) : \text{length}(\boldsymbol{\alpha}) = \ell\},$$

which are assumed as independent across the different length groups but the nodes within a given $\left(\Upsilon_j(\tilde{\mathsf{m}}) \cup \mathcal{C}_j(\tilde{\mathsf{m}})\right)_\ell$ are exchangeable, namely, nodes of a given order with the same number of parents in $\tilde{\mathsf{m}}_F$ are assumed to be exchangeable, as long as their inclusion satisfies the hierarchical condition of $\mathcal{M}$. On the other hand, for the HTP, the nodes group is defined by

$$\left(\Upsilon_j(\tilde{\mathsf{m}}) \cup \mathcal{C}_j(\tilde{\mathsf{m}})\right)_t = \{\boldsymbol{\alpha} \in \Upsilon_j(\tilde{\mathsf{m}}) \cup \mathcal{C}_j(\tilde{\mathsf{m}}) : \text{type}(\boldsymbol{\alpha}) = t\},$$

thus the nodes that have the same type are assumed exchangeable.

As already mentioned, the elements presented in this chapter are the main ingredients in the development of the thesis. With these in mind, we propose novel nonparametric Bayesian hypothesis testing procedures. In Chapter 2, our proposal is focused on paired data, and we would like to compare the distributions of these dependent populations. We present the hierarchical model on which the hypothesis testing has been set up as well as a posterior inference algorithm. We provide a Monte Carlo study for assessing the performance of our methodology, at the ending of Chapter 2 we supply an example with real data. In Chapter 3, we deal with

the modeling of longitudinal data, where a set of independent experimental units are measured through the time, and the goal is to test whether some predictors have an effect on the response variable. We propose a flexible procedure based on a Dependent Dirichlet process, for modeling the correlation structure among the observations from the same experimental unit, and we use the theory of variable selection and the priors on the model space to set up the hypothesis testing procedure. Like in Chapter 2, we supply examples with simulated data for assessing the performance of our proposal, an example with real data is also included. Finally, in Chapter 4, we present a general discussion on the results from Chapter 2 and 3, as well as future works and open problems. It is worth mentioning that Chapters 2 and 3 are self-contained.

Chapter 2

A Bayesian nonparametric testing procedure for paired samples

2.1 Abstract

We propose a Bayesian hypothesis testing procedure for comparing the distributions of paired samples. The procedure is based on a flexible model for the joint distribution of both samples. The flexibility is given by a mixture of Dirichlet processes. Our proposal uses a spike-slab prior specification for the base measure of the Dirichlet process and a particular parametrization for the kernel of the mixture in order to facilitate comparisons and posterior inference. The joint model allows us to derive the marginal distributions and test whether they differ or not. The procedure exploits the correlation between samples, relaxes the parametric assumptions and detects possible differences throughout the entire distributions. A Monte Carlo simulation study comparing the performance of this strategy to other traditional alternatives is provided. Finally, we apply the proposed approach to spirometry data collected in the U.S. to investigate changes in pulmonary function in children and adolescents in response to air polluting factors.

Keywords: Spike-slab priors; Dirichlet Process; Shift function; Mixture model; Dependent samples.

2.2 Introduction

Bayesian nonparametric models have shown to be a powerful alternative to parametric statistical models (see, e.g., Ghosh & Ramamoorthi 2003, Müller & Quintana 2004, Hjort et al. 2010, Müller & Mitra 2013). The literature on hypothesis testing from the Bayesian nonparametric (BNP) standpoint is relatively new, and has focused on few, and very specific problems. Comparisons between two independent samples has received most of the attention; some proposals based on Pólya tree pri-

ors can be found in Ma & Wong (2011), Chen & Hanson (2014), Huang & Ghosh (2014), Holmes et al. (2015), and Soriano & Ma (2017). Proposals based on Dirichlet process priors are provided in the works of Borgwardt & Ghahramani (2009), Bhattacharya & Dunson (2012), Shang & Reilly (2017), and Al-Labadi & Zarepour (2017). BNP methods for dependent or paired samples include the works of Filippi et al. (2016), Filippi & Holmes (2017). Both of these works test for dependence between the two samples using Dirichlet processes and Pólya tree priors, respectively.

In the context of multiple hypothesis testing, some developments are found in the works of Gopalan & Berry (1998), Scott (2009), Kim et al. (2009), Cipolli III et al. (2016) and Gutiérrez et al. (2019) among others. Although the literature on BNP hypothesis testing is growing, there is still room to relax the constraining, and often times untenable assumptions predominant in the classical hypothesis testing literature, and propose more realistic procedures. A classical, and particularly important problem in a wide variety of scientific inquiries, is that of paired sample comparison, but is yet to be suitably formulated from the BNP standpoint, which constitutes the focus of this article.

In order to formalize the comparison of two dependent samples, let Y_{ij}, $i = 1, \ldots, n$ be random variables which represent the measurements for the i-th individual or sample unit at a time j, with $j = 1, 2$. For instance, the measurements could represent the responses of patients before and after the application of a treatment. Thus, for each individual, we have a bivariate vector $\mathbf{Y}_i = (Y_{i1}, Y_{i2})^t$, which follows a joint distribution $G(\cdot)$ with support in $\mathbb{R}^2$. Our aim is to develop a Bayesian nonparametric hypothesis testing procedure to perform comparisons between the marginal distributions of Y_{i1} and Y_{i2} denoted by $G_1(\cdot)$ and $G_2(\cdot)$, respectively. In particular, we would like to test the following hypotheses

$$H_0 : G_1(\cdot) = G_2(\cdot) \text{ vs. } H_1 : G_1(\cdot) \neq G_2(\cdot), \tag{2-1}$$

which synthesize the *paired sample* testing problem. Following a Bayesian approach, our goal is to assess the strength of the evidence in favor of both of these hypotheses by means of their posterior probabilities. Furthermore, if the evidence for H_1 is stronger, we want to visualize in which aspects (i.e., regions of the support) the distributions differ. For instance, if we consider a typical situation for paired sample testing, in which we would like to determine the effect of a treatment on a particular health outcome measurement, there exists the possibility that the effect varies across members of the population. In these situations the differences between the distributions could be reflected, as an increased dispersion or as changes in the symmetry of the distribution. Hence, traditional paired sample tests may prove ineffective, as many of these are exclusively aimed at detecting differences in location or scale of the distributions.

The most popular procedure for this type of data is the T-test for paired samples proposed by Student (1908). This test follows a parametric approach to infer about differences in the location parameters in G_1 and G_2 assuming a Gaussian distribution. In particular, the paired T-test bases the inferences on the differences $D_i = Y_{i2} - Y_{i1}$ for each subject. A popular alternative for testing differences in scale in paired samples is the Morgan-Pitman test (Morgan 1939, Pitman 1939). The Morgan-Pitman test is based on the correlation coefficient between two linear combinations of the variables Y_1 and Y_2, which are assumed to follow a bivariate Gaussian distribution. Among the nonparametric procedures available to compare paired samples are the Wilcoxon signed-rank test (Wilcoxon 1945) and its Bayesian version proposed by Benavoli et al. (2014). While the Wilcoxon test is also based upon the differences D_i, it relaxes the parametric assumptions required for the T-test. The method proposed by Benavoli et al. (2014) uses a Dirichlet process as a prior on the D_i's.

Asides from the Morgan-Pitman test, all of the alternatives mentioned above depend on the D_i's to compare paired samples; however, using the D_i's with this purpose has been shown to be problematic. For paired sample T-tests, Zimmerman (1997) found that the sign of the correlation between the measurements can impact the power of the test. In particular, positive correlations increase the probability of rejecting the null hypothesis when it is false. Conversely, negative correlations reduce such probability. This is because $\text{Var}(D) = \text{Var}(Y_2) + \text{Var}(Y_1) - 2\text{Cov}(Y_1, Y_2)$, thus tests based on the differences D lose power as the correlation between Y_1 and Y_2 approaches minus one, since the variance of D increases when the sign of $\text{Cov}(Y_1, Y_2)$ is negative. Having noticed this issue, Girón et al. (2003) developed an alternative to the paired T-test via a model selection approach. This strategy uses a hierarchical model to represent the dependence between the measurements for each individual under a bivariate Gaussian distribution.

Although the proposal in Girón et al. (2003) considers the dependence in the paired observations, it relies on strong parametric assumptions, such as normality and a positive correlation structure between the paired samples, restricting its applicability. In this work, we propose a novel BNP hypothesis testing procedure based on the joint distribution $G(\cdot)$ that explicitly accounts for the dependence between the samples, bypassing the restrictive assumption on the sign of the covariance structure. The inference on $G(\cdot)$ is based on an infinite mixture model (see, e.g., Ghosh & Ramamoorthi 2003, Müller & Quintana 2004, Hjort et al. 2010, Müller & Mitra 2013), where the mixing distribution follows a Dirichlet process (Ferguson 1973, 1974). This formulation provides flexibility in the estimation of $G(\cdot)$ and, consequently, in the marginal distributions $G_1(\cdot)$ and $G_2(\cdot)$. We consider a continuous version of the spike-slab prior (Mitchell & Beauchamp 1988) for the base

measure of the Dirichlet process and a conditionally independent parametrization of the bivariate kernel in the mixture. These choices define a Bayesian hypothesis testing procedure, which yields estimates for $\mathbb{P}(H_0 \mid \boldsymbol{Y})$ and $\mathbb{P}(H_1 \mid \boldsymbol{Y})$. In addition, due to the nonparametric nature of our procedure, we are able to detect differences across the entire distribution and not only in location and scale as in the traditional alternatives. Finally, whenever differences between the two marginal distributions are identified, these differences can be further investigated using the shift function (see, e.g., Doksum 1974, Doksum & Sievers 1976, Hollander & Korwar 1980, Wells & Tiwari 1989, Lu et al. 1994).

The remainder of the manuscript is organized as follows. In Section 2.3, we present the proposed paired-sample hypothesis testing procedure, providing details about the parametrization of the bivariate kernel and the definition of the prior on the random mixing distribution. In Section 2.4, we present a general outline for how to conduct posterior inference using the proposed methods, together with a strategy to visually compare differences between the distributions. A Monte Carlo simulation study is provided in Section 2.5, comparing the performance of our procedure against some traditional alternatives. In Section 2.6, we show an application on real data from a spirometry study. Finally, in Section 2.7, we present a discussion and concluding remarks.

2.3 The proposed hypothesis testing procedure

In this section, we develop the proposed methods to perform hypothesis testing of paired samples. Suppose that the random vectors $\boldsymbol{Y}_i$ $(1 \leq i \leq n)$ with support in $\mathbb{R}^2$, are independent and identically distributed (i.i.d.) from the joint distribution $G(\cdot)$. Assume that G is absolutely continuous with respect to the Lebesgue measure. Hence, $G(\cdot)$ has density $g(\cdot)$ which is specified by the following nonparametric Bayesian mixture model,

$$\begin{aligned} \boldsymbol{Y}_i \mid F &\overset{iid}{\sim} g(\cdot) := \int_{\Theta} \mathrm{K}\left(\cdot \mid \theta\right) dF(\theta), \quad i = 1, \ldots, n \\ F \mid H_\zeta &\sim \mathrm{DP}\left(M, F_{0|H_\zeta}\right), \\ H_\zeta &\sim \pi_{\mathcal{M}}. \end{aligned} \tag{2-2}$$

In (2-2) $\mathrm{K}\left(\cdot \mid \theta\right)$ is a continuous kernel and F is a random probability measure which follows a Dirichlet process, with parameters M and $F_{0|H_\zeta}$. M is called the mass parameter of the process, because it controls the variability of the realizations from the process around the base measure $F_{0|H_\zeta}$, which can be thought as a prior to F. Note that, the specification of F is conditional on the hypotheses H_ζ, where $\zeta \in \{0, 1\}$ and $\pi_{\mathcal{M}}$ is a prior on the space of hypotheses $\mathcal{M} = \{H_0, H_1\}$. A natural choice for

the prior $\pi_{\mathcal{M}}$ over the discrete space $\mathcal{M}$ is a Bernoulli distribution with parameter $\pi = P(H_1)$. We complete the specification for this part of the model by letting π follow a Beta$(1/2, 1/2)$ prior, which corresponds to the Jeffreys prior for proportions. This prior pushes most of its probability mass towards values near zero and one, making it suitable for testing.

2.3.1 Parametrization of $\mathrm{K}(\cdot \mid \theta)$

Our testing strategy is completed by specifying the kernel $\mathrm{K}(\cdot \mid \theta)$. Given the support of the observations, a natural choice for $\mathrm{K}(\cdot \mid \theta)$ is a bivariate Gaussian distribution $\mathrm{N}_2(\cdot \mid \theta)$, $\theta = (\mu, \Sigma)$. Considering the nature of the problem, the parametrization for the kernel must accommodate the hypothesis testing problem specified in (2-1). To this end, note that the bivariate Gaussian distribution can be expressed as,

$$\mathrm{N}_2(y_1, y_2 \mid \theta) = \int \mathrm{N}(y_1 \mid \theta, \phi)\mathrm{N}(y_2 \mid \theta, \phi)d\phi, \tag{2-3}$$

where ϕ is a nuisance parameter such that Y_1 and Y_2 are conditionally independent given ϕ. The formulation provided in (2-3) is widely used in mixed models (see, e.g. Eisenhart 1947, Henderson 1953, Fahrmeir et al. 2013), where the parameter ϕ represents a random effect. The random effects model yields marginals that only allow for positive correlations between the repeated measures (Fitzmaurice et al. 2009, p. 6). However, our formulation must capture both positive and negative correlations between the paired samples. In order to deal with this restriction, we propose an alternative straightforward conditionally independent formulation, that enables capturing both positive and negative correlations between the paired samples. The conditional model is given by

$$\begin{aligned} Y_{ij} &= \beta_1 + \beta_2 X_{ij} + Z_{ij}\delta_i + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, 2, \\ \epsilon_{i1} &\overset{iid}{\sim} \mathrm{N}\left(0, \sigma_1^2\right), \\ \epsilon_{i2} &\overset{iid}{\sim} \mathrm{N}\left(0, \sigma_1^2\sigma_2^2\right), \\ \delta_i &\overset{iid}{\sim} \mathrm{N}\left(0, \tau^2\right), \end{aligned} \tag{2-4}$$

In model (2-4) β_1 corresponds to the mean value of Y_1 and β_2 represents the mean shift of Y_2 in relation to Y_1. X_{ij} is a given known value, which takes the values 0 if the i-th observation was taken at time $j = 1$ and 1 if it was taken at time $j = 2$. The variable Z_{ij} is a latent variable that takes the values -1 or 1, with probability mass function given by

$$P(Z_{ij} = z_{ij}) = \begin{cases} \gamma_j & \text{if} \quad z_{ij} = -1, \\ 1 - \gamma_j & \text{if} \quad z_{ij} = 1, \end{cases}$$

where $\gamma_j \sim \text{Beta}(a, b)$ and the precision $1/\tau^2$ follows a Gamma distribution denoted by $\text{Ga}(1/\tau^2 \mid a_0, b_0)$ with $E(1/\tau^2) = \dfrac{a_0}{b_0}$ and $Var(1/\tau^2) = \dfrac{a_0}{b_0^2}$. This parametrization for the Gamma distribution will be used in the remainder of the manuscript. The random effects δ_i and the error terms $\epsilon_{i1}, \epsilon_{i2}$ are assumed to be mutually independent. The conditional model (2-4) follows the representation in (2-3) with $\phi = (\delta_i, Z_{i1}, Z_{i2})$.

Proposition 1. From the conditional model in (2-4) we have:

1. The marginal model is given by $\boldsymbol{Y} \sim \text{N}_2(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} \beta_1 \\ \beta_1 + \beta_2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma_1^2 + \tau^2 & (1-2\gamma_1)(1-2\gamma_2)\tau^2 \\ (1-2\gamma_1)(1-2\gamma_2)\tau^2 & \sigma_1^2\sigma_2^2 + \tau^2 \end{pmatrix}.$$

2. $-1 < \text{Corr}(Y_1, Y_2) < 1$.

Proof. The proof is given in the Appendix A. □

We adopt the parametrization of Proposition 1 for the bivariate Gaussian kernel because: 1) the parameters β_2 and σ_2^2 capture the possible differences between the 1st and 2nd measurements, making this formulation amenable to testing the hypotheses in (2-1) (see Remark 1); 2) the conditional representation of (2-4) facilitates posterior sampling given the simplicity of the structure provided by the hierarchical representation (see the details in Section 2.4); 3) positive and negative covariance values between the paired observations can be captured.

Remark 1. From model (2-2) and Proposition 1 we have that the joint and the marginals distributions can equivalently be expressed as

$$\begin{aligned} G(\cdot) &= \sum_{h \geq 1} w_h \Phi_2(\cdot \mid \mu_h, \Sigma_h), \\ G_1(\cdot) &= \sum_{h \geq 1} w_h \Phi\left(\cdot \mid \beta_{1h}, \sigma_{1h}^2 + \tau_h^2\right), \\ G_2(\cdot) &= \sum_{h \geq 1} w_h \Phi\left(\cdot \mid \beta_{1h} + \beta_{2h}, \sigma_{1h}^2\sigma_{2h}^2 + \tau_h^2\right), \end{aligned}$$

where $\Phi(\cdot)$ denotes the Gaussian cumulative density function, μ_h and Σ_h are defined as in Proposition 1.

Note that, the sign of the covariance in Σ is determined by γ_1 and γ_2, and the magnitude is mainly captured by τ^2. For instance, if γ_1 or γ_2 is close to 0.5, the covariance is close to zero. If γ_1 and γ_2 are less than 0.5, then the covariance is positive. Finally, if $\gamma_1 < 0.5$ and $\gamma_2 > 0.5$ (or viceversa) the covariance is negative. The previous parametrization allows us to define the hypothesis testing procedure properly, with a minimum cost in the correlation, which can not take the singleton values {-1} and {1}. However, as we will show in Section 2.4, such a restriction rarely poses a problem in practice.

2.3.2 Induced prior on the random measure F

Given the discrete nature of the Dirichlet process, F can be expressed as a weighted sum of point masses such that $F(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\theta_h}(\cdot)$, where $w_1, w_2, \ldots$ are random weights and $\theta_1, \theta_2, \ldots$ are i.i.d. random variables from the distribution $F_{0|H_\zeta}$. Here, $\delta_\theta(\cdot)$ denotes the Dirac measure at θ. In the above specification, the weights w_h's can be obtained via the "stick-breaking" construction proposed by Sethuraman (1994). Under this specification, the weights are given by $w_1 = \nu_1$ and $w_h = \nu_h \prod_{k<h}(1-\nu_k)$, where $\nu_h \overset{iid}{\sim} \text{Beta}(1, M)$. To build additional flexibility into the model, we assume a Gamma distribution for the parameter M denoted by $\text{Ga}(M \mid a_1, b_1)$.

Now, for the base measure $F_{0|H_\zeta}$ (explicitly dependent on H_ζ) we adopt a modified version of the continuous spike and slab formulation of Ishwaran & Rao (2005), which defines a hypothesis testing procedure. In particular, the base measure under H_0 is defined as

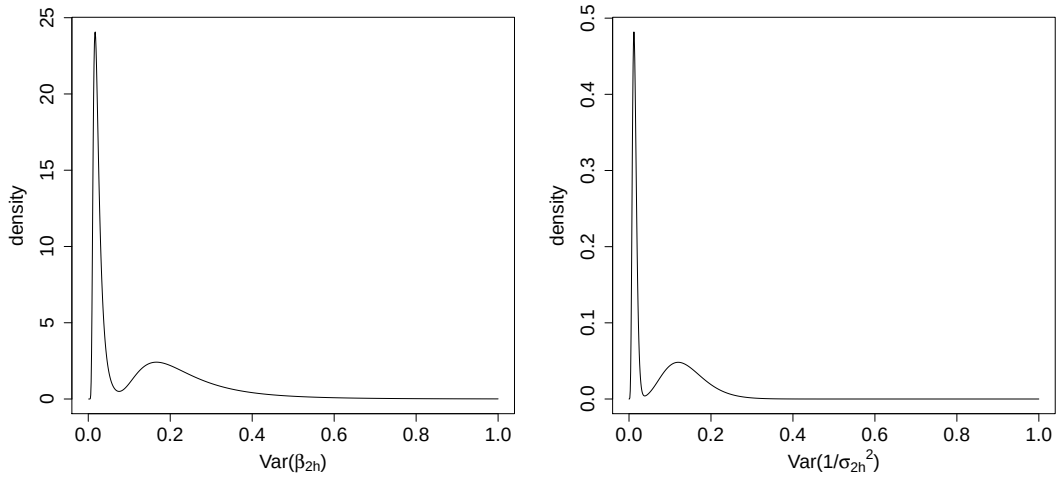
$$F_{0|H_\zeta} : \text{N}_2\left((\beta_{1h}, \beta_{2h})^T \mid \boldsymbol{\mu}_0 = (0,0)^T, \Sigma_0 = \text{diag}\left[\psi, \kappa\left(\mathbb{1}_{(\zeta=1)} + \nu_0 \mathbb{1}_{(\zeta=0)}\right)\right]\right) \times \quad (2\text{-}5)$$
$$\text{Ga}\left(1/\sigma_{1h}^2 \mid \epsilon, \epsilon\right) \text{Ga}\left(1/\sigma_{2h}^2 \mid s\left(\mathbb{1}_{(\zeta=1)} + \nu_0^{-1}\mathbb{1}_{(\zeta=0)}\right), s\left(\mathbb{1}_{(\zeta=1)} + \nu_0^{-1}\mathbb{1}_{(\zeta=0)}\right)\right),$$

which is completed with the specification of the hyperpriors on κ and s given by

$$\begin{aligned} \frac{1}{\kappa} &\sim \text{Ga}(a_2, b_2), \\ s &\sim \text{Ga}(a_3, b_3). \end{aligned} \quad (2\text{-}6)$$

The values a_i, b_i, $i = 2, 3$, are chosen so that the $\text{Var}(\beta_{2h})$ and $\text{Var}(1/\sigma_{2h}^2)$ have a continuous bimodal distribution with a spike at ν_0 and a right continuous tail as in Ishwaran & Rao (2005). The ν_0 value is chosen so that ν_0 is a positive near-zero value. Thus, if $\zeta = 0$ (supporting H_0), then the base measure corresponds to the spike, meaning that the atoms $\{\beta_{2h}\}_{h\geq 1}$ and $\{\sigma_{2h}^2\}_{h\geq 1}$ follow distributions with a small variance tightly concentrated about 0 and 1, respectively. On the contrary, when $\zeta = 1$ (supporting the alternative H_1) the base measure is the slab component of the density.

Following Ishwaran & Rao (2005), we fixed the hyperparameters in (2-6) at $a_2 = 5$, $b_2 = 1$, $a_3 = 5$, $b_3 = 50$, and $\nu_0 = 0.1$. Figure **2-1** shows the behavior of the spike-slab prior density for the hypervariance in (2-5) when fixing $\pi = P(\zeta = 1) = 0.5$. The remainder of the hyperparameters in (2-5) were fixed at $\psi = 10$ and $\epsilon = 0.1$, so that the prior for the parameters β_{1h} and σ_{1h}^2 are relatively uninformative. Exact computation of $P(H_0 \mid \text{Data})$ is unviable as it requires marginalizing over $\{\beta_{2h}\}_{h\geq 1}$ and $\{\sigma_{2h}^2\}_{h\geq 1}$ the infinite dimensional model in (2-2) (Geweke 1996, Smith & Kohn 1996, Malsiner-Walli et al. 2011)



(a) (b)

Figure **2-1**: Conditional density for the hypervariances $\varphi_1 = \text{Var}(\beta_{2h})$ and $\varphi_2 = \text{Var}\left(1/\sigma^2_{2h}\right)$: (a) $f_{\varphi_1}(\varphi_1 \mid \zeta, \nu_0, \pi) = \pi \text{ InvGa}\left(\varphi_1 \mid a_2, b_2\right) + {}^1\!/\!_{\nu_0} (1 - \pi) \text{ InvGa}\left({}^{\varphi_1}\!/\!_{\nu_0} \mid a_2, b_2\right)$, with $a_2 = 5$ and $b_2 = 1$. (b) $f_{\varphi_2}(\varphi_2 \mid \zeta, \nu_0, \pi) = \pi \text{ InvGa}\left({}^1\!/\!_{\varphi_2} \mid a_3, b_3\right) + {}^1\!/\!_{\nu_0} (1 - \pi) \text{ InvGa}\left({}^{\nu_0}\!/\!_{\varphi_2} \mid a_3, b_3\right)$, with $a_3 = 5$, $b_3 = 50$. In both cases we set (for illustration purposes) $\pi = 0.5$ and $\nu_0 = 0.1$.

2.4 Posterior inference

We now turn to a general description of the approach proposed to conduct posterior inference. First, we develop the Gibbs sampling algorithm used to obtain posterior draws from model (2-2). In the second part of this section, we supply a formal definition of the shift function, and a strategy to derive it from the Gibbs algorithm.

2.4.1 Gibbs algorithm

In order to fit Dirichlet process mixture (DPM) models, one must deal with the estimation of infinite-dimensional parameters. Several strategies exist to sample DPM models (e.g., Escobar 1988, 1994, MacEachern 1994, MacEachern & Müller 1998, Neal 2000, Walker 2007, Kalli et al. 2011); among them we consider the method of Walker (2007) for its efficiency due to the slice sampling step, which adapts the number of components in the mixture according to the complexity of the data, thus reducing the infinite dimensional space to a finite one. Given the discrete nature of the Dirichlet process, from Remark in 1, we have that the mixture model in (2-2) can be rewritten as the weighted infinite sum of continuous kernels given by

$$g\left(y\right) = \sum_{h \geq 1} w_h \text{N}_2\left(y \mid \theta_h\right), \tag{2-7}$$

where $\theta_h = (\mu_h, \Sigma_h)$. Walker (2007) defined an augmented model given by

$$g(y,u) = \sum_{h\geq 1} \mathbf{1}(u < w_h)\, \mathrm{N}_2(y \mid \theta_h), \tag{2-8}$$

where u is Unif$(0,1)$. The model in (2-8) is finite, because only a finite number of w_h's satisfy the condition $(u < w_h)$. Augmenting the likelihood with u does not alter the original density; in fact marginalizing Equation (2-8) over u leads back to Equation (2-7). Since the DPM model induces the existence of clusters, Walker (2007) introduces an additional membership latent variable denoted by d_i, $i = 1, \ldots, n$. This variable labels the cluster each observation is generated from, resulting in the augmented joint likelihood given by

$$g(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{d}) \propto \prod_{i=1}^{n} \mathbf{1}(u_i < w_{d_i})\, \mathrm{N}_2(y_i \mid \theta_{d_i}). \tag{2-9}$$

Combining the distributional assumptions provided in Section 2.3 with the augmented likelihood defined above, the sampling algorithm consists of the following Gibbs steps

Algorithm 1

[1] $p(\theta_h \mid \ldots) \propto f_{0|H_\zeta}(\theta_h) \prod_{\{i:d_i=h\}} \mathrm{N}_2(y_i \mid \theta_h)$,

[2] $p(\nu_h \mid \ldots) \propto \mathrm{Beta}\left(1 + \sum_{i=1}^{n} \mathbb{1}_{(d_i=h)}, M + \sum_{i=1}^{n} \mathbb{1}_{(d_i>h)}\right)$,

[3] $p(u_i \mid \ldots) \propto \mathbb{1}_{(0<u_i<w_{d_i})}$,

[4] $P(d_i = k \mid \ldots) \propto \mathbb{1}_{(k:w_k>u_i)} \mathrm{N}_2(y_i \mid \theta_k)$,

[5] $P(\zeta = 1 \mid \ldots) = \frac{\pi \prod_{h=1}^{N} \mathrm{N}_{(\beta_{2,h}|0,\kappa)} \mathrm{Ga}\left(1/\sigma^2_{2,h}|s,s\right)}{\pi \prod_{h=1}^{N} \mathrm{N}_{(\beta_{2,h}|0,\kappa)} \mathrm{Ga}\left(1/\sigma^2_{2,h}|s,s\right) + (1-\pi) \prod_{h=1}^{N} \mathrm{N}_{(\beta_{2,h}|0,\kappa\nu_0)} \mathrm{Ga}\left(1/\sigma^2_{2,h}|s\nu_0^{-1},s\nu_0^{-1}\right)}$,

[6] $P(\pi \mid \ldots) \propto \mathrm{Beta}\left(\frac{1}{2} + \mathbb{1}_{(\zeta=1)}, \frac{3}{2} - \mathbb{1}_{(\zeta=1)}\right)$,

[7] $P\left(\frac{1}{\kappa} \mid \ldots\right) \propto \mathrm{Ga}\left(a_2 + \frac{N}{2}, b_2 + \frac{\sum_{h=1}^{N} \beta^2_{2,h}}{2\left(\mathbb{1}_{(\zeta=1)} + \nu_0 \mathbb{1}_{(\zeta=0)}\right)}\right)$,

[8] $P(s \mid \ldots) \propto s^{a_3-1} \exp\Big\{ - b_3 s + Ns\left(\mathbb{1}_{(\zeta=1)} + \nu_0^{-1}\mathbb{1}_{(\zeta=0)}\right) \log\left(s\left(\mathbb{1}_{(\zeta=1)} + \nu_0^{-1}\mathbb{1}_{(\zeta=0)}\right)\right)$
$- N \log\left(\Gamma\left(s\left(\mathbb{1}_{(\zeta=1)} + \nu_0^{-1}\mathbb{1}_{(\zeta=0)}\right)\right)\right)$
$+ s\left(\mathbb{1}_{(\zeta=1)} + \nu_0^{-1}\mathbb{1}_{(\zeta=0)}\right)\left(\sum_{h=1}^{N} \log\left(1/\sigma^2_{2,h}\right) - 1/\sigma^2_{2,h}\right)\Big\}$.

A key element of Algorithm 1 is that the subscript h takes values in the set $\{1, \ldots, N\}$, where $N = \max_i\{N_i\}$ and N_i is the largest integer l for which $w_l > u_i$. Thus, at each iteration, once N is determined only a finite number of weights and atoms is to be sampled. Steps [2] to [8] in Algorithm 1 are relatively straightforward, in particular, in step [8] a Metroplis-Hasting step is necessary. Step [1] needs more computation due to the special parametrization of θ_h. To sample each element of θ_h, we resort to the conditional model in (2-4). The full conditionals for the elements of θ_h are drawn from a bivariate

Gaussian distribution for the parameters β_1 and β_2. The parameters $1/\sigma_1^2$ and $1/\sigma_2^2$ are sampled from their full conditional distributions, which are Gamma. The full conditional distributions for the latent parameters δ_i, $1/\tau^2$ and γ_j in model (2-4) are Normal, Gamma and Beta, respectively. Finally, the latent parameters Z_{ij} are sampled from a discrete full conditional distribution with support in the set $\{-1, 1\}$. Explicit forms for each full conditional distribution involved in the estimation of θ_h are provided in Appendix B. This Appendix also provides the mechanism used to update the mass parameter M, as suggested in (Escobar & West 1995).

We fix the hyperparameters for τ^2 in model (2-4) at $a_0 = b_0 = 0.01$, because these values yield a weakly informative prior. To promote values of γ_j close to zero or one, we fixed its hyperparameters to $a = b = 1/2$. As a consequence, the expression $(1 - 2\gamma_1)(1 - 2\gamma_2)$ is *a priori* close to -1 or 1; this expression determines the sign of the correlation. Furthermore, the magnitude of the correlation is mainly driven by the parameter τ^2.

From Algorithm 1, the posterior probability for the alternative hypothesis $P(H_1 \mid \boldsymbol{Y}) = P(\zeta = 1 \mid \boldsymbol{Y})$ can be approximated with

$$\mathbb{P}(H_1 \mid \boldsymbol{Y}) \approx \frac{1}{B} \sum_{\ell=1}^{B} \mathbf{1}_{\{\zeta^{(\ell)}=1\}}, \tag{2-10}$$

where $\zeta^{(\ell)}$, $\ell = 1, \ldots, B$, are samples from the full conditional distribution of Step [5]. Assuming a zero–one loss function, we select the most probable hypothesis.

2.4.2 Visualization of the differences

If the posterior evidence favors the alternative hypothesis, we would like to visualize in what aspects the distributions $G_1(\cdot)$ and $G_2(\cdot)$ differ. To do so we compute the *Shift function* as a measure of the difference between the two populations. This function was proposed by Doksum (1974) and Doksum & Sievers (1976). The idea behind the Doksum's proposal is to find a function $\Delta(\cdot)$, such that, $Y_1 + \Delta(Y_1)$ has the same distribution as Y_2. Formally, $\Delta(\cdot)$ is a function such that $G_1(Y_1) = G_2(Y_1 + \Delta(Y_1))$ or equivalently, $\Delta(Y_1) = G_2^{-1}\{G_1(Y_1)\} - Y_1$. Note that, under H_0, $\Delta(Y)$ is equal to 0, for all Y. Under H_1, $\Delta(Y)$ is different from 0 for some set $A := \{Y : \Delta(Y) \neq 0\}$. The set A provides information on what regions of the distribution are different.

Deriving the shift function using Algorithm 1 is immediate, since for each iteration of the Gibbs algorithm, we have posterior random realizations of $G_1^{(\ell)}$ and $G_2^{(\ell)}$, $\ell = 1, \ldots, B$. Defining the left inverse of G_2 as $G_2^{-1}(u) = \inf\{x : G_2(x) \geq u\}$, a random realization of $\Delta(Y)^{(\ell)}$ can be computed as

$$\Delta(Y)^{(\ell)} = \begin{cases} G_2^{-1(\ell)}\left\{G_1^{(\ell)}(Y)\right\} - Y & \text{if } \zeta^{(\ell)} = 1, \\ 0, \forall \quad Y & \text{if } \zeta^{(\ell)} = 0. \end{cases} \tag{2-11}$$

With the posterior realizations of the shift function, it is possible to compute some

functionals, as the sample posterior mean $\bar{\Delta}(Y)$ and a 95% credible set. The credible set is particularly useful to determine the set A, which can be visualized looking at the values of Y such that $\Delta(Y) \neq 0$.

2.5 Monte Carlo simulation study and illustrations

This section provides a Monte Carlo simulation study. To ease the interpretation, we assume that observations come from n subjects that were measured before and after the application of a treatment. Our goal is to evaluate the ability of the hypothesis testing procedure to detect the treatment effect, especially when the distributions differ. In the Monte Carlo study, we consider six scenarios, four of them illustrate global changes in the distributions before and after the treatment, the other two represent local changes. We provide further details on two of the scenarios considered to illustrate the entire inferential process. Data for all the scenarios were generated from variations of the following model

$$\boldsymbol{Y}_i \sim \omega_1 \mathrm{SN}_2(\boldsymbol{Y}_i \mid \boldsymbol{\mu}_1, \Sigma_1, \boldsymbol{\alpha}_1) + \omega_2 \mathrm{N}_2(\boldsymbol{Y}_i \mid \boldsymbol{\mu}_2, \Sigma_2) + \omega_3 \mathrm{N}_2(\boldsymbol{Y}_i \mid \boldsymbol{\mu}_3, \Sigma_3), \tag{2-12}$$

where $\mathrm{SN}_2(\boldsymbol{\mu}, \Sigma, \boldsymbol{\alpha})$ denotes a bivariate skew normal distribution with location $\boldsymbol{\mu}$, scale Σ and shape $\boldsymbol{\alpha}$, and again $\mathrm{N}_2(\boldsymbol{\mu}, \Sigma)$ denotes a bivariate normal distribution with location $\boldsymbol{\mu}$ and scale Σ.

Scenarios I and II were designed for assessing the performance of our BNP testing procedure versus the traditional alternatives when the Gaussian assumption is valid. In particular, scenario I represents an effect of the treatment in location (where T-test assumptions hold), while scenario II shows a treatment effect in the scale (where Morgan-Pitman test's assumptions hold). A global change in location arises when the effect of the treatment is the same across all individuals in the population. Scenarios III, IV and V, were designed to emulate situations where only a portion of the population is influenced by the treatment, or when the magnitude of the treatment effect varies across individuals. Finally, scenario VI was planned with the aim of generating samples beyond the mixture of Gaussian distributions and to emulate situations when the asymmetry of the distribution changes. Table **2-1** provides details for the settings considered under each scenario.

With each scenario we consider sample sizes $n_1 = 50$, $n_2 = 150$ and $n_3 = 300$. For each particular combination of scenario and sample size we generated 100 Monte Carlo replicates, for a total of 1,800 experiments. Model (2-2) was fitted via the Gibbs algorithm of Section 2.4.1 to each of the 1,800 datasets generated, with 10,000 iterations, a burn-in period of 2,000, and thinning the samples by keeping only every 8th draw of the sampled parameters. The model was implemented in the R Programming Language (R Core Team 2018). The hyperparameters or the prior distribution of the mass parameter M were fixed at $a_1 = 20$ an $b_1 = 1$ chosen after calibration. This choice performed well under all scenarios.

Scenarios	ω	$\boldsymbol{\mu}_1$	Σ_1	$\boldsymbol{\alpha}$	$\boldsymbol{\mu}_2$	Σ_2	$\boldsymbol{\mu}_3$	Σ_3	Grid of values
I (Global shift)	1.0 0.0 0.0	$\begin{bmatrix} 0 \\ \mu_2 = c \end{bmatrix}$	$\begin{bmatrix} 1 & -0.80 \\ -0.80 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$					$c \in \{0, 0.3, 0.6, 0.9, 1.6, 2.6, 3.6, 4.6, 5.6\}$
II (Global dispersion)	1.0 0.0 0.0	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.95 \\ 0.95 & \sigma_2^2 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$					$\sigma_2^2 \in \{1, 3, 5, 7, 9\}$
III (Mixture global shift)	0.50 0.50 0.00	$\begin{bmatrix} 0 \\ -\mu_2 = c \end{bmatrix}$	$\begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ \mu_2 = c \end{bmatrix}$	$\begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$			$c \in \{0, 1, 2, 3, 4\}$
IV (Local dispersion)	0.60 0.40 0.00	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 2 \\ 2 \end{bmatrix}$	$\begin{bmatrix} 4 & -3.6 \\ -3.6 & \sigma_2^2 \end{bmatrix}$			$\sigma_2^2 \in \{4, 12, 20, \ldots, 44\}$
V (Mixture local shift)	0.50 0.25 0.25	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ -\mu_2 = c \end{bmatrix}$	$\begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ \mu_2 = c \end{bmatrix}$	$\begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}$	$c \in \{0, 1.5, 3, \ldots, 7.5\}$
VI (Global asymmetry)	1.0 0.0 0.0	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ \alpha_2 \end{bmatrix}$					$\alpha_2 \in \{0, 1.5, 2, 4, 6, 8\}$

Table **2-1**: Parametrization of model (2-12) for scenarios I to VI in the Monte Carlo simulation study.

Additionally, for all datasets generated under the different scenarios we compute the following classical alternatives: T-test, the Wilcoxon signed-rank test together with its Bayesian version (Benavoli et al. 2014), and the Morgan-Pitman test (Morgan 1939, Pitman 1939), which is a traditional alternative for testing differences in scale in paired samples. The Morgan-Pitman test is based on the correlation coefficient between two linear combinations of the variables Y_1 and Y_2. To compare the performance of our proposal with the alternative tests, we explore the power to detect the alternative hypothesis. Considering that the power is a frequentist concept, we adapted our Bayesian test to get a measurement of the statistical power. In particular, we used a zero-one loss function, thus we select the alternative hypothesis when its posterior probability is bigger than 0.5 and report the proportion of times that the BNP test selected the correct hypothesis. For the classical alternatives, the null hypothesis was rejected considering a significance level of 5%. The test of Girón et al. (2003) was not included in the comparisons, because it is based on a Bivariate normal distribution, and it is only able to detect changes in location when the correlation between the measurements is positive, which is a particular case of our BNP test.

Figure **2-2** shows the power curves for scenarios I to III. In Scenario I, the performance of the classical location tests was quite good, as expected, given that in this case all of the assumptions required for the classical tests hold. Importantly, under this scenario our method's performance was comparable to that of the classical approaches, even with a sample size of 50. For scenario I, the Morgan-Pitman test behaves as expected given that it does not detect differences in location. In scenario II, our test has a similar performance as Morgan-Pitman test. However, the Morgan-Pitman test attained a higher proportion of false positives than the proposed method, especially with smaller sample sizes. In scenario III, the proposed BNP test and the Morgan-Pitman test do an outstanding job at detecting differences, with approximately equal performance for $n \geq 150$.

Figure **2-3** provides the power curves for scenarios IV through VI. In scenarios IV and V, all classical location tests were unable to detect the alternative hypothesis for any value in the corresponding grids. The performance of our proposal was very good, especially

with $n \geq 150$. The Morgan-Pitman test was able to detect the alternative hypothesis, which can be thought of as a change in scale due to the treatment. However, the true difference is due to the mixture specification of scenarios IV and V. In scenario VI, again the BNP testing procedure achieves robust results. Even for $n = 50$, the power for the BNP test is close to that of the Morgan-Pitman test. The Morgan-Pitman test shows very good results for all sample sizes in this scenario, although the true differences are in the symmetry and not only in the scale. The performance of the location tests, as expected, is weak; regardless of the sample size, the power was never close to 1.

To elaborate on the results that are derived from the proposed method, we provide posterior inference for scenarios IV and V from Table **2-1**. In particular, assuming a sample size of $n = 150$, we consider these scenarios since they pose a bigger challenge. Similar figures for scenarios I, II, III and VI are included in the Appendix. For scenario IV, we provide posterior inference for data generated assuming $\sigma_2^2 = 20$. For scenario V, the setting considered from the grid was $c = 4.5$. Figure **2-4** compares the true and estimated joint densities, the estimation corresponds to the posterior mean. The figure also includes the true and estimated marginal distributions along with their 95% credible sets (represented by the grey regions). The true marginal densities are represented by continuous black lines, and the estimated posterior mean for the measurements at time 1 is given by the red dotted line, while the posterior mean for the marginal at time 2 is represented by the blue dashed line. Finally, we include the true (continuous black line) and estimated shift functions (dashed line) along with the 95% point-wise credible sets.

As seen in Figure **2-4**, the estimation of the joint distribution and consequently the marginal distributions was quite good in both scenarios. The estimation of the shift function was also accurate. The shift function for scenario IV suggests that the difference between $G_1(\cdot)$ and $G_2(\cdot)$ is due to changes in the tails of the distribution. The shift function for scenario V indicates differences across the entire distribution.

2.6 An application to spirometry data

In this Section, we develop an application of our hypothesis testing procedure to a real data set. The application is in the context of spirometry studies. The data set generated by Dockery et al. (1983) contains information from a cohort of 13,379 U.S. children born on or after 1967 enrolled in the first or second grade in elementary school. The purpose of the study was to identify changes in the pulmonary function in children and adolescents associated to air pollution factors, which could influence lung function development. The study includes measurements of the forced expiratory volume in one second also known as FEV1. In this application we only consider a subset of the data publicly available from Fitzmaurice et al. (2010). This subset contains records from 228 girls who resided in Topeka, Kansas (USA). The measurements were obtained annually, each girl in the longitudinal study has between one and twelve records over time. In this application we only consider the two measurements from the first and third years. We fitted model (2-2) to the logarithms of FEV1 considering the same settings used in the Monte Carlo simulation

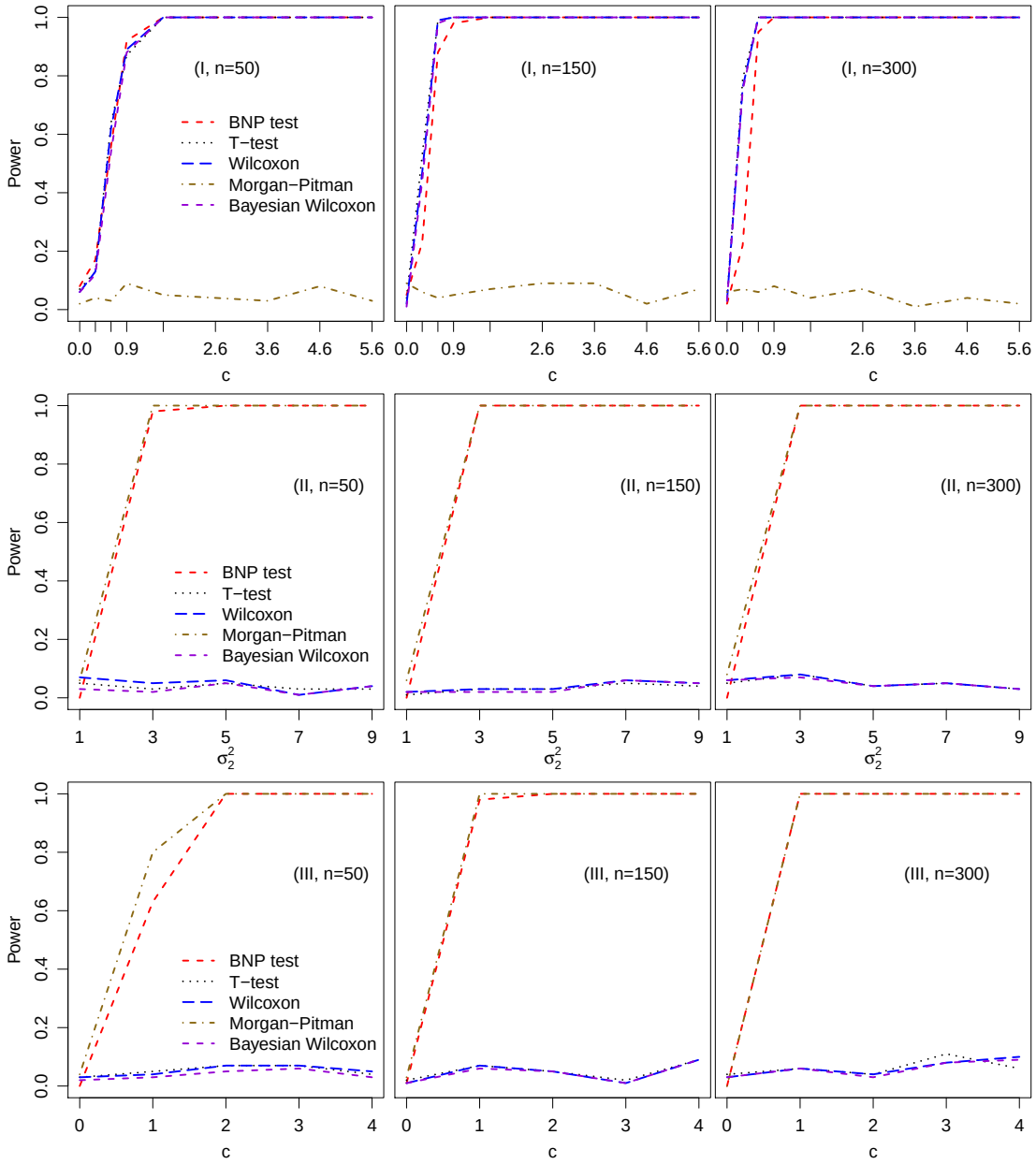


Figure **2-2**: Power to detect the alternative hypothesis in the Monte Carlo simulation study of Section 2.5. Top panel scenario I, middle panel scenario II and bottom panel scenario III with different sample sizes.

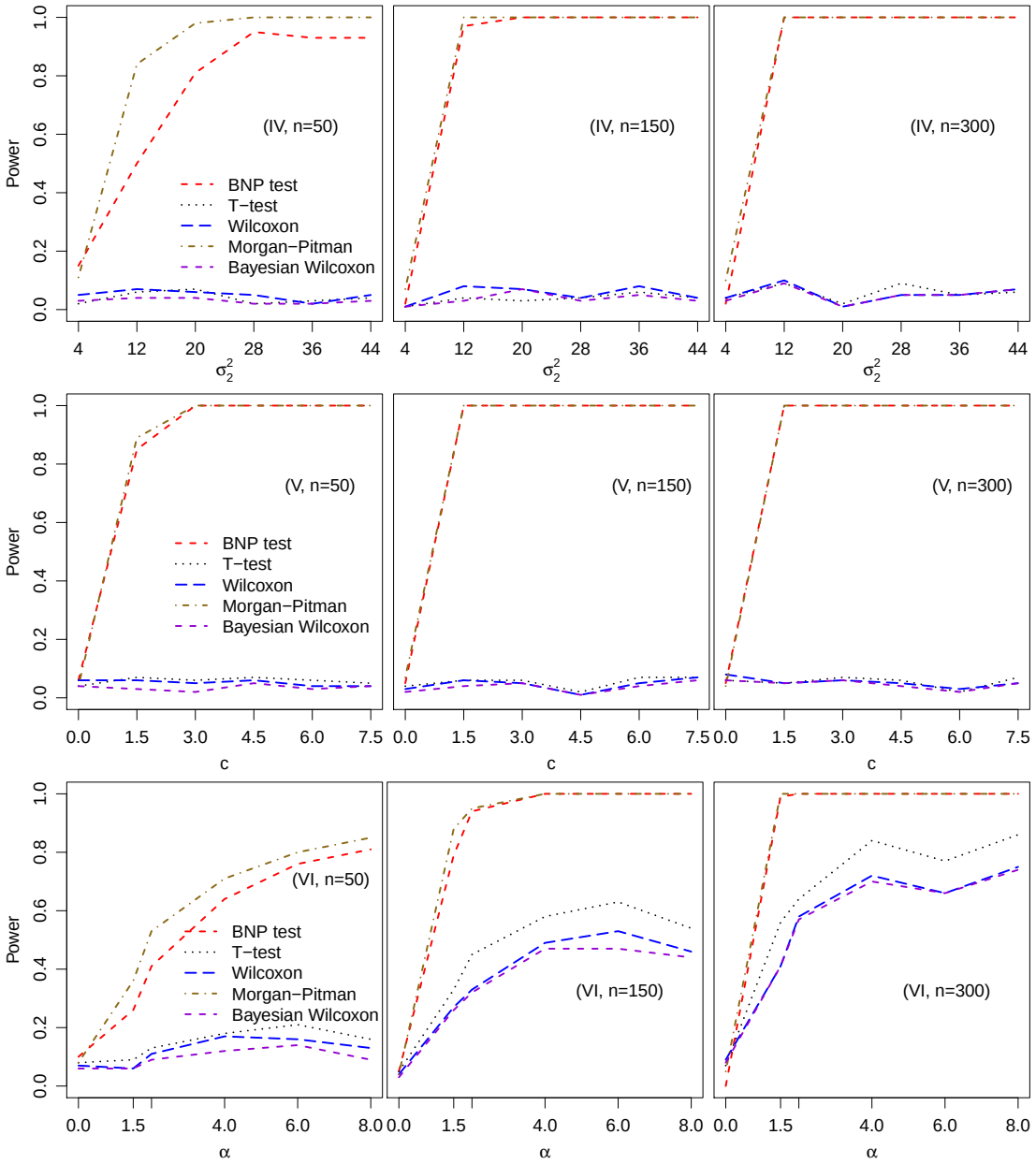


Figure **2-3**: Power to detect the alternative hypothesis in the Monte Carlo simulation study of Section 2.5. Top panel scenario IV, middle panel scenario V and bottom panel scenario VI with different sample sizes.

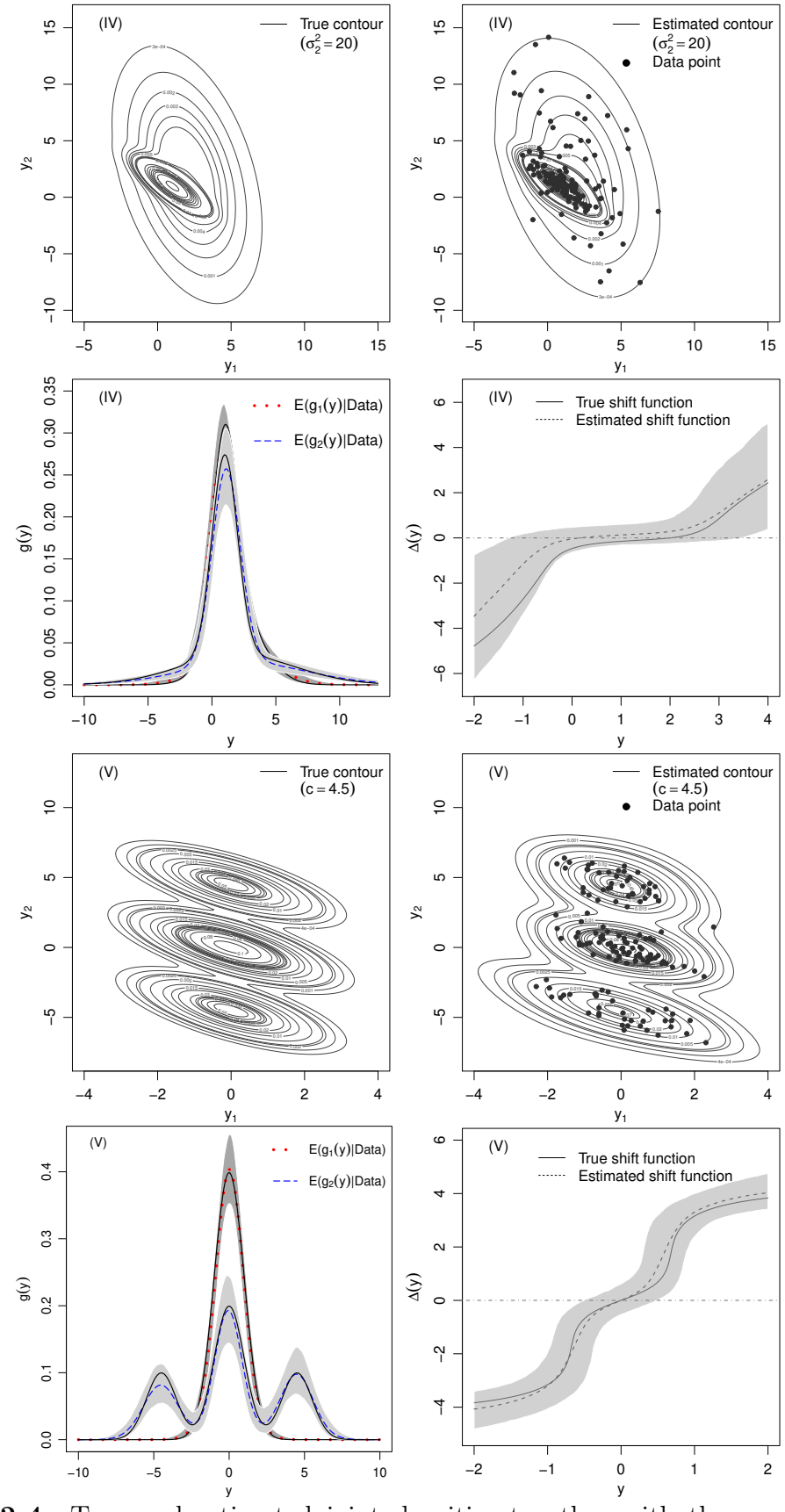


Figure **2-4**: True and estimated joint densities together with the corresponding true and estimated marginal densities and shift functions for scenarios IV and V of Section 2.5.

study. For this data set, the method took ~ 9 minutes on a standard computer (AMD A8-6410 processor running at 2.0 GHz) having the algorithm generate 10000 posterior draws.

The posterior probability for the alternative hypothesis was 0.737. Thus, the proposed BNP test supports the hypothesis that the marginal distributions are different. The classical tests also detected differences in the marginal distributions with a significance level of 5%. The T-test (p-value $<$ 2.2e-16), the Wilcoxon test (p-value $<$ 2.2e-16) and its Bayesian version ($\hat{P}\left(H_1 \mid \text{Data}\right) > 0.95$), all identified differences in location, while the Morgan-Pitman test (p-value $= 0.00042$) detected differences in scale. Thus, the conclusion with the classical test is that the girls experienced a change in location and variability of their pulmonary capacity.

Figure **2-5** shows the posterior inference obtained from the BNP test. The joint estimated density suggests a strong correlation between the paired measurements. From the estimated marginal and cumulative distributions together with the shift function, it is possible to conclude that the increment in FEV1 was not constant across girls in the sample. In fact, the girls with the largest expiratory capacity underwent a greater increment. This conclusion could not have been obtained from any of the classical alternatives.

2.7 Concluding remarks

We have proposed a procedure based on a BNP model for comparing the distributions of paired samples. The comparison of paired samples is a classical problem in statistics with important applications in many fields of science. The available tests are feature specific, in the sense, that they target changes in specific attributes of the distributions, such as location or scale. Additionally, some standard tests are only valid under restrictive parametric assumptions that are rarely fulfilled in practice.

The proposed BNP test bypasses the need for these restrictive parametric assumptions, and at the same time, exploits the dependence between the samples. Furthermore, the proposed strategy is able to detect differences across the entire distribution. We provide a simple heuristic to visualize the differences between the distributions using the shift function. As seen in the analysis of the spirometry data, the shift function confirms that the treatment has an effect that varies across members of the population.

Our approach yields consistently good results as shown in Section 2.5, in many cases outperforming the traditional tests. As expected, the classical tests were able to detect differences for the features for which they were designed, that is, location or scale, but may be misleading in many instances, pointing to differences in the erroneous features of the distributions. From the study in Section 2.5, we can conclude that for problems with small sample sizes, specific location or scale tests could be preferred, given that they target specific features of the distributions. Nevertheless, many current applications have sufficiently large data to make our approach the more appealing alternative. In

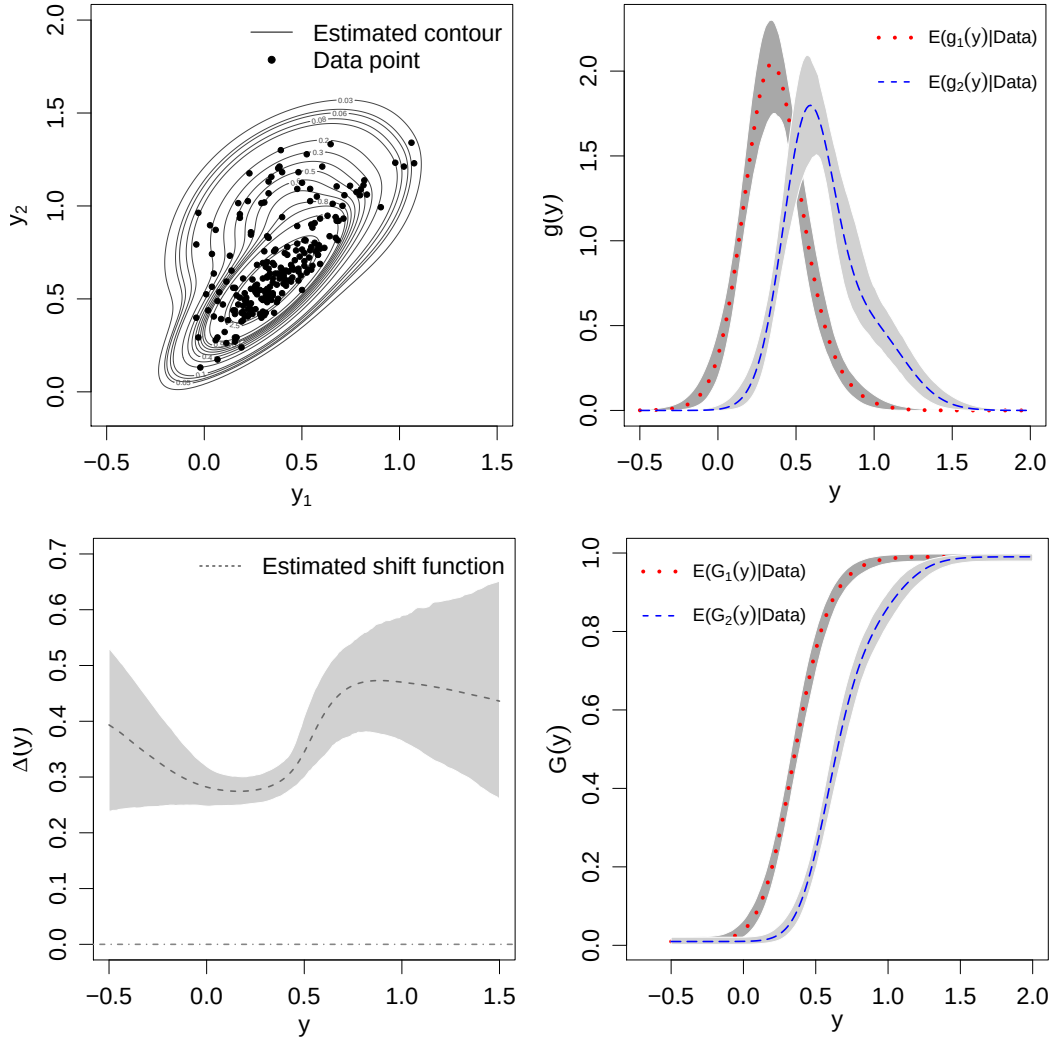


Figure **2-5**: Estimation of the joint and the marginal densities as well as of the shift function and cumulative distribution for the spirometry study of Section 2.6.

synthesis, the strategy developed in this article provides a hypothesis testing procedure that is simultaneously capable of producing smooth, uncertainty equipped estimates for the density. Additionally, the estimation of the shift function allows us to identify which parts of the population are effectively influenced by the treatment and in what amount.

2.8 Appendix A

Proof of proposition 1

1. From the Gaussian assumption for the random effects and the errors we have,

$$\begin{aligned} E\left[Y_{ij} \mid X_{ij} = x_{ij}\right] &= E\left[\beta_1 + \beta_2 x_{ij} + Z_{ij}\delta_i + \epsilon_{ij}\right] \\ &= \beta_1 + \beta_2 x_{ij} + E\left[Z_{ij}\delta_i\right] \\ &= \beta_1 + \beta_2 x_{ij} + E\left[Z_{ij}\right] E\left[\delta_i\right] \\ &= \beta_1 + \beta_2 x_{ij}, \end{aligned}$$

therefore, $E\left[Y_{i1} \mid X_{i1} = 0\right] = \beta_1$ and $E\left[Y_{i2} \mid X_{i2} = 1\right] = \beta_1 + \beta_2$. On the other hand,

$$\begin{aligned} \operatorname{Var}\left[Y_{ij} \mid X_{ij} = x_{ij}\right] &= \mathrm{E}\left[\operatorname{Var}\left(Y_{ij} \mid X_{ij} = x_{ij}, Z_{ij}, \delta_i\right)\right] + \\ &\quad \operatorname{Var}\left[\mathrm{E}\left(Y_{ij} \mid X_{ij} = x_{ij}, Z_{ij}, \delta_i\right)\right] \\ &= \mathrm{E}\left[\operatorname{Var}\left(\beta_1 + \beta_2 x_{ij} + Z_{ij}\delta_i + \epsilon_{ij} \mid X_{ij} = x_{ij}, Z_{ij}, \delta_i\right)\right] + \\ &\quad \operatorname{Var}\left[\mathrm{E}\left(\beta_1 + \beta_2 x_{ij} + Z_{ij}\delta_i + \epsilon_{ij} \mid X_{ij} = x_{ij}, Z_{ij}, \delta_i\right)\right] \\ &= \mathrm{E}\left[\operatorname{Var}\left(\epsilon_{ij}\right)\right] + \operatorname{Var}\left[Z_{ij}\delta_i\right] \\ &= \mathrm{E}\left[\operatorname{Var}\left(\epsilon_{ij}\right)\right] + \tau^2. \end{aligned}$$

Thus, if $j = 1$ then $\operatorname{Var}\left[Y_{i1} \mid X_{i1} = 0\right] = \sigma_1^2 + \tau^2$, and if $j = 2$ then $\operatorname{Var}\left[Y_{i2} \mid X_{i2} = 1\right] = \sigma_1^2\sigma_2^2 + \tau^2$. Finally,

$$\begin{aligned} \operatorname{Cov}\left[Y_{i1}, Y_{i2} \mid X_{ij} = x_{ij}\right] &= \mathrm{E}\left[\operatorname{Cov}\left(Y_{i1}, Y_{i2} \mid X_{ij} = x_{ij}, Z_{ij}, \delta_i\right)\right] \\ &\quad + \operatorname{Cov}\left[\mathrm{E}\left(Y_{i1} \mid X_{i1} = 0, Z_{ij}, \delta_i\right) \mathrm{E}\left(Y_{i2} \mid X_{i2} = 1, Z_{ij}, \delta_i\right)\right] \\ &= \operatorname{Cov}\left[Z_{i1}\delta_i, Z_{i2}\delta_i\right] \\ &= \mathrm{E}\left[\left(Z_{i1}\delta_i - \mathrm{E}\left(Z_{i1}\delta_i\right)\right)\left(Z_{i2}\delta_i - \mathrm{E}\left(Z_{i2}\delta_i\right)\right)\right] \\ &= \mathrm{E}\left[Z_{i1}Z_{i2}\delta_i^2\right] \\ &= \mathrm{E}\left[Z_{i1}\right] \mathrm{E}\left[Z_{i2}\right] \mathrm{E}\left[\delta_i^2\right] \\ &= \left(1 - 2\gamma_1\right)\left(1 - 2\gamma_2\right)\tau^2, \end{aligned}$$

where, $\tau^2 > 0$ and $\gamma_1, \gamma_2 \in (0, 1)$.

2. If $\gamma_1 > \frac{1}{2}$ and $\gamma_2 < \frac{1}{2}$ or $\gamma_1 < \frac{1}{2}$ and $\gamma_2 > \frac{1}{2}$ then $\left(1 - 2\gamma_1\right)\left(1 - 2\gamma_2\right) < 0$ and $\operatorname{Cov}\left[Y_{i1}, Y_{i2}\right] < 0$. If $\gamma_1 < \frac{1}{2}$ and $\gamma_2 < \frac{1}{2}$ or $\gamma_1 > \frac{1}{2}$ and $\gamma_2 > \frac{1}{2}$ then $\left(1 - 2\gamma_1\right)\left(1 - 2\gamma_2\right) > 0$ and $\operatorname{Cov}\left[Y_{i1}, Y_{i2}\right] > 0$. On other hand,

$$\left|\operatorname{Cov}(Y_{i1}, Y_{i2})\right| = \left|\left(1 - 2\gamma_1\right)\left(1 - 2\gamma_2\right)\right|\tau^2 < \tau^2, \tag{2-13}$$

and using the Cauchy-Schwarz inequality, we have

$$|\text{Corr}(Y_1, Y_2)| = \frac{|(1-2\gamma_1)(1-2\gamma_2)|\tau^2}{\sqrt{\left(\sigma_1^2+\tau^2\right)\left(\sigma_1^2\sigma_2^2+\tau^2\right)}} < \frac{\tau^2}{\sqrt{\left(\sigma_1^2+\tau^2\right)\left(\sigma_1^2\sigma_2^2+\tau^2\right)}} \leq 1.$$

In fact, we have two cases from the above expression, the first one is when we suppose that H_0 is true and we take the limit of τ^2 going to infinity, when $\tau^2 \longrightarrow \infty$, then we have that

$$\lim_{\tau^2\to\infty} \frac{|(1-2\gamma_1)(1-2\gamma_2)|\tau^2}{\left(\sigma_1^2+\tau^2\right)} = |(1-2\gamma_1)(1-2\gamma_2)| < 1. \tag{2-14}$$

In the second case, we suppose that H_1 is true and,

$$\lim_{\tau^2\to\infty} \frac{(1-2\gamma_1)(1-2\gamma_2)\tau^2}{\sqrt{\left(\sigma_1^2+\tau^2\right)\left(\sigma_1^2\sigma_2^2+\tau^2\right)}} = |(1-2\gamma_1)(1-2\gamma_2)| < 1. \tag{2-15}$$

Consequently, $|\text{Corr}(Y_1, Y_2)| < 1$.

2.9 Appendix B

In the Section 2.4, we have provided an algorithm of eight stages for the posterior inference of the BNP model. However, the Step [1] requires to clarify some details about the computation due to the special parametrization of θ_h. In the following, we supply details about the full conditional distributions used in the Step [1]. Additionally, we present the details for the updating of the mass parameter M of the Dirichlet Process.

Gibbs sampling for the posterior inference of θ_h.

According to parametrization in the Proposition 1, we develop a Gibbs sampling for updating $\theta_h = (\mu_h, \Sigma_h)$. Next, we present the expressions of the full conditional distributions for the parameters $\beta_1, \beta_2, \sigma_1^2, \sigma_2^2, \tau^2, \gamma_1, \gamma_2, Z_{i1}, Z_{i2}, \delta_i$.

1. Full conditional distributions for updating μ.

   Sample β_1, β_2 from

   $$\beta_1, \beta_2 | \ldots \sim \text{N}_2\left[\left(n\mathbf{X}^T\Sigma^{-1}\mathbf{X}+\Sigma_0^{-1}\right)^{-1}\left(\mathbf{X}^T\Sigma^{-1}\sum_{i=1}^{n}\boldsymbol{Y}_i+\Sigma_0^{-1}\mu_0\right), \left(n\mathbf{X}^T\Sigma^{-1}\mathbf{X}+\Sigma_0^{-1}\right)^{-1}\right],$$

   where $\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ is a design matrix and μ_0, Σ_0 are the hyper-parameters in the prior distribution defined in $F_{0|H_\zeta}$.

2. Full conditionals for updating Σ.

   (i) Sample σ_1^2 from

   $$\frac{1}{\sigma_1^2} \mid \ldots \sim \text{Ga}\left[n+\epsilon, \lambda_1\right],$$

where $\lambda_1 = \frac{\sum_{i=1}^{n}(y_{i1}-(\beta_1+Z_{i1}\delta_i))^2}{2} + \frac{\sum_{i=1}^{n}(y_{i2}-(\beta_1+\beta_2+Z_{i2}\delta_i))^2}{2\sigma_2^2} + \epsilon$.

(ii) Sample σ_2^2 from

$$\frac{1}{\sigma_2^2} \mid \ldots \sim \text{Ga}\left[\frac{n}{2} + s\left(\mathbb{1}_{(\zeta=1)} + \nu_0^{-1}\mathbb{1}_{(\zeta=0)}\right), \lambda_2\right],$$

where $\lambda_2 = \frac{\sum_{i=1}^{n}(y_{i2}-(\beta_1+\beta_2+Z_{i2}\delta_i))^2}{2\sigma_1^2} + s\left(\mathbb{1}_{(\zeta=1)} + \nu_0^{-1}\mathbb{1}_{(\zeta=0)}\right))$.

In (i)-(ii) ϵ and s correspond to the values defined in $F_{0|H_\zeta}$.

(iii) Sample δ_i from

$$\delta_i \mid \ldots \sim \text{N}\left[\frac{\frac{Z_{i1}(y_{i1}-\beta_1)}{\sigma_1^2} + \frac{Z_{i2}(y_{i2}-\beta_1-\beta_2))}{\sigma_1^2\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_1^2\sigma_2^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_1^2\sigma_2^2} + \frac{1}{\tau^2}}\right].$$

(iv) Sample $Z_{i1} \in \{-1, 1\}$ with probabilities

(a) $P\left(Z_{i1} = -1 \mid \ldots\right) \propto \gamma_1 \times \exp\left\{\frac{-1}{2\sigma_1^2}\left(y_{i1} - \beta_1 + \delta_i\right)^2\right\}$,

(b) $P\left(Z_{i1} = 1 \mid \ldots\right) \propto (1-\gamma_1) \times \exp\left\{\frac{-1}{2\sigma_1^2}\left(y_{i1} - \beta_1 - \delta_i\right)^2\right\}$.

(v) Sample $Z_{i2} \in \{-1, 1\}$ with probabilities

(a) $P\left(Z_{i2} = -1 \mid \ldots\right) \propto \gamma_2 \times \exp\left\{\frac{-1}{2\sigma_1^2\sigma_2^2}\left(y_{i2} - \beta_1 - \beta_2 + \delta_i\right)^2\right\}$,

(b) $P\left(Z_{i2} = 1 \mid \ldots\right) \propto (1-\gamma_2) \times \exp\left\{\frac{-1}{2\sigma_1^2\sigma_2^2}\left(y_{i2} - \beta_1 - \beta_2 - \delta_i\right)^2\right\}$.

(vi) Sample γ_j from

$$\gamma_j \mid \ldots \sim \text{Beta}\left(\frac{1}{a} + \sum_{i=1}^{n}\mathbf{1}_{\{Z_{ij}=-1\}}, \frac{1}{b} + n - \sum_{i=1}^{n}\mathbf{1}_{\{Z_{ij}=-1\}}\right).$$

(vii) Sample τ^2 from

$$\frac{1}{\tau^2} \mid \ldots \sim \text{Ga}\left[a_0 + \frac{n}{2}, b_0 + \frac{\sum_{i=1}^{n}\delta_i^2}{2}\right].$$

Updating of the mass parameter M.

To sample M, we follow the idea proposed by Escobar & West (1995). Thus, M can be sampled from a mixture of two gamma densities given by

$$\pi\left(M|\eta, k\right) = \tau_\eta \text{Ga}\left(a_1 + k, b_1 - \log\left(\eta\right)\right) + \left(1 - \tau_\eta\right)\text{Ga}\left(a_1 + k - 1, b_1 - \log\left(\eta\right)\right),$$

where for all $k > 1$, $\tau_\eta = {}^1\!/_{1+\frac{n(b_1-\log(\eta))}{a_1+k-1}}$, n is the sample size, a_1 and b_1 are the hyper-parameters for a prior Gamma distribution defined on M (see, Section 2.4.1), and η is a random continuous variable defined on the unit interval, such that $\pi\left(\eta|M, k\right) \sim \text{Beta}\left(M + 1, n\right)$.

2.10 Appendix C

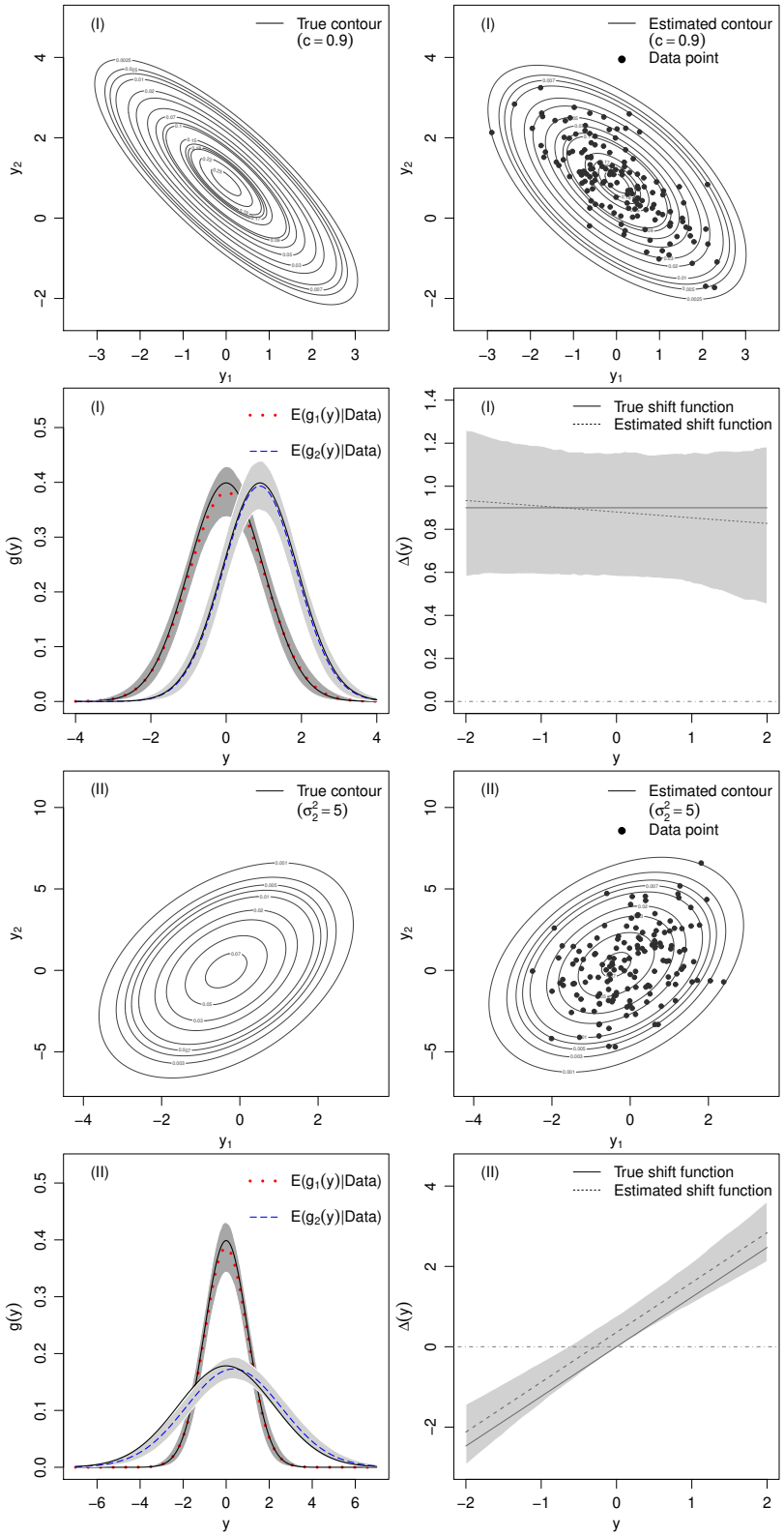


Figure **2-6**: The contour of the joint distribution, the marginal density distributions, and the shift function for scenarios I and II.

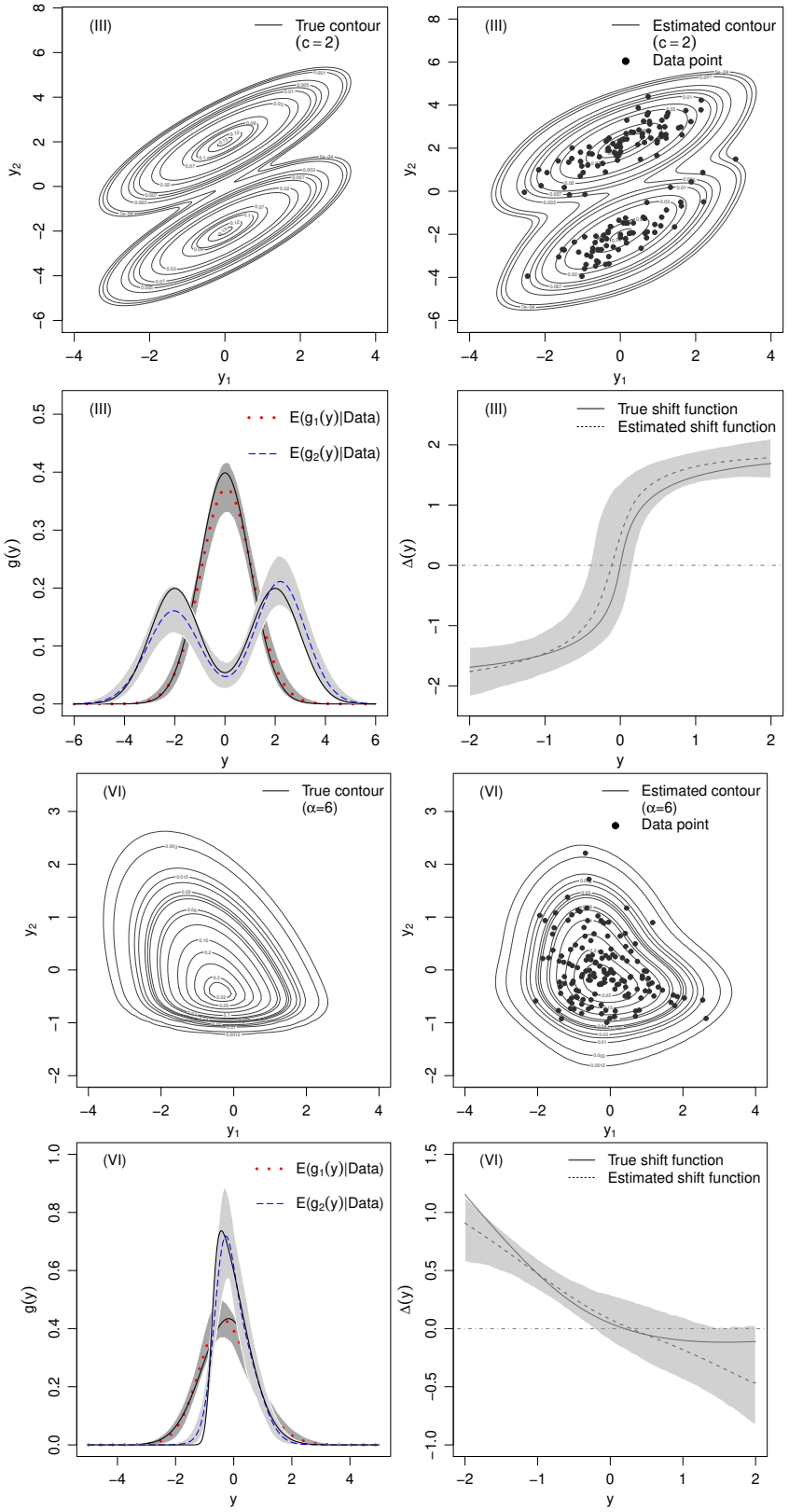


Figure **2-7**: The contour of the joint distribution, the marginal density distributions, and the shift function for scenarios III and VI.

Chapter 3

Bayesian nonparametric hypothesis testing for longitudinal data analysis

3.1 Abstract

We propose a Bayesian nonparametric hypothesis testing procedure to find the possible effect of predictors over the response variable in longitudinal data analysis. The method is highly flexible because it does not assume a particular covariance structure nor a distribution for the random effects, as usually done in longitudinal data analysis. The proposal models the behavior of the repeated measurements with a mixture of dependent Dirichlet processes. The weights of the mixture are built via a stick-breaking prior, that comes from a Markovian process evolving in time. The effect of the predictors is modeled by the underlying atoms. A hierarchical representation is used to define a hypothesis testing procedure for experimental designs that can include the effects of interactions. Illustrations with simulated and real data sets, as well as a comparison study, are also presented.

Keywords: Dependent Dirichlet process; Markov Process; Spike and slab prior; Time-dependent data.

3.2 Introduction

In longitudinal data analysis, experimental units are measured repeatedly over time. In this context, units are assumed to be mutually independent, and measurements for the same experimental unit are time-dependent. Suppose that our data consists of random realizations of the time-dependent variables y_{it_j} from m mutually independent experimental units, where $i = 1, \ldots, m$, $j = 1, \ldots, n_i$ and $t_1 < \ldots < t_j < \ldots < t_{n_i}$ are ordered-time indexes in the interval $[0, T]$ not necessarily equally spaced. Experimental units have a *monotone missing-data* pattern, meaning that all measurement times are the same across experimental units until dropout.

Assume that we observe time-fixed covariates, $\mathsf{x}_i \in \mathbb{R}^q$, for each unit sample, and that the final aim is to test hypotheses about possible effects of the covariates on the response variable. Any suitable model for this kind of data should capture the correlation among the repeated measurements within the same experimental unit. Simultaneously, it should detect meaningful treatment effects, by setting a formal hypotheses testing procedure. A simplistic approach for modeling longitudinal data, assumes exchangeability between the repeated measurements over time (de Finetti 1931, 1937). This amounts to say that the joint distribution, given the predictor x, is defined as

$$\mathbb{P}(Y_{t_1} \in A_1, \ldots, Y_{t_n} \in A_n \mid \mathsf{x}) = \int_{\mathcal{F}} F(A_1 \mid \mathsf{x}) \cdots F(A_n \mid \mathsf{x}) \mu(dF), \tag{3-1}$$

where μ is a probability measure characterizing the exchangeable sequence $\{Y_t : t \in [0, T]\}$, also known as the de Finetti's measure. In hierarchical terms

$$\begin{aligned} \mathbb{P}(Y_{t_1} \in A_1, \ldots, Y_{t_n} \in A_n \mid F, \mathsf{x}) &= \prod_{i=1}^{n} F(A_i \mid \mathsf{x}), \\ F &\sim \mu. \end{aligned} \tag{3-2}$$

In (3-2), given the distribution $F(\cdot)$, the random variables are conditionally independent. In the longitudinal data analysis literature, (3-1) and (3-2) are known as the marginal and conditional models, respectively. A simple and useful approach to model the joint distribution $\mathbb{P}(\cdot)$ is via a parametric restriction of μ. Typically, $F(\cdot)$ is given by a Gaussian distribution including random effects ϕ, fixed effects β, and a scale parameter σ. In this latter case, the uncertainty in ϕ is also modeled with a Gaussian prior with zero mean and unknown variance.

In this context, the covariate vector x is partitioned in two components, fixed and random effects, leading to the well-known mixed models (Laird & James 1982). Under this approach, the inference is based on the marginal model (Verbeke & Molenberghs 2009), which is obtained integrating the random effects with respect to the distribution μ. As a consequence, the research questions are answered performing the corresponding hypothesis testing on the fixed effects. A widely used procedure for contrasting hypotheses are based on t and F tests, which make use of the degrees of freedom, corresponding to the number of *independent observations*. This approach to longitudinal data analysis is highly sensitive to multivariate Gaussian assumption, and thus also the estimation of the underlying degrees of freedom (Weiss 2005). In practice, a way to deal with this is to approximate the t-distribution with the Gaussian distribution, which works for large data sets. That said, there are no guidelines to select the sample size (Luke 2017). There are methods to estimate the number of degrees of freedom (Satterthwaite 1941, Kenward & Roger 1997), however, these can lead to notable differences in the corresponding $p-$values as they exhibit a dependence on the significance level α and on the covariance matrix Weiss (2005).

The literature on classical approaches for longitudinal data analysis is vast, with the most popular approach based on variations and extensions of the hierarchical specification of (3-2), see e.g. Laird & James (1982), Liang & Zeger (1986), Breslow & Clayton

(1993), Azzalini (1994), Diggle et al. (2002), Huang & Zhou (2002), Hogan et al. (2004), Fitzmaurice et al. (2010), Chen & Zhong (2010), Liu (2015), among many others. Special emphasis is placed to capture complex behaviour in the mean or correlation structures, which is often done by modeling the fixed effects via splines, kernel estimators, (see, e.g. Silverman 1984, Heckman 1986, Speckman 1988, Zeger & Diggle 1994, Carroll et al. 1997, Stone et al. 1997, Silverman 1998, Zhang et al. 1998, Verbyla et al. 1999, Welsh et al. 2002, Carroll et al. 2004, Crainiceanu & Ruppert 2004, Chen & Jin 2005). On the other hand, for random effects, different correlation structures and methodologies for its estimation have been proposed, e.g., Jennrich & Schluchter (1986), Muñoz et al. (1992), Nuñez Antón & Zimmerman (2000), Zimmerman (2000), Wang & Hin (2010).

On the Bayesian counterpart, there is a strong emphasis on parametric models, Lopes et al. (2003), Li et al. (2010), Wang & Daniels (2013), Müller et al. (2014), Huang et al. (2014), Dahlin et al. (2016) and Castro et al. (2018), among others. Bayesian nonparametric constructions using Dirichlet processes under different parameterizations have been also studied in Kleinman & Ibrahim (1998), Müller et al. (2005), Kliethermes (2013), Savitsky & Paddock (2014), Shang (2016), Quintana et al. (2016) and Linero (2017). Related to hypothesis testing procedures Dahl & Newton (2007) propose a nonparametric Bayesian methodology for multiple hypothesis testing in random effects models. The procedure is used to test if the random effects are zero or not. They use a Dirichlet process with base measure a continuous distribution centered at zero. Latter, Kim et al. (2009) proposed to modify the base measure of the DP by a spike-slab distribution.

In this work we consider the time evolution of the longitudinal data by means of a Markovian Dependent Dirichlet processes (DDP) (MacEachern 1999, 2000, Gutiérrez et al. 2016). In particular, our proposal considers the correlation between repeated measurements and provides with a formal hypothesis testing procedure to find possible effects of the predictors on the response variable. We do this, by capturing the time-dynamics in the underlying stick-breaking weights, and the covariate effects, in the locations. An adhoc hierarchical specification will be used with the purpose of formally define a hypothesis testing procedure for the fixed effects. The prior on the hypothesis space borrows ideas from the model selection literature in order to control for multiple testing (Berger & Pericchi 1996, George 2000, Berger et al. 2001, Scott & Berger 2010). In particular, we use the prior specification to test simple effects, as well as, the effect of the interactions between the predictors proposed by Taylor-Rodríguez et al. (2016). Our proposal is a novel nonparametric methodology in the context of longitudinal data analysis as, in addition to testing the effect of predictors, it captures the correlation between the repeated measurements and provides a time-dependent density estimation for different levels of the covariates.

The remainder of the manuscript is organized as follows. In Section 3.3, we present the Bayesian nonparametric model for hypothesis testing in longitudinal data analysis. Here, we also provide details about the prior specification on the hypotheses space and discuss the framework of the Markov model use to capture the time-dependent measurements. The algorithm for posterior inference is deferred to Subsection 3.4, with some of its details

contained in the Appendix B 3.9. In Section 3.5, we illustrate and compare our proposal with simulated data. An application to a real data is found in Section 3.6. A discussion and some concluding remarks are given in Section 3.7.

3.3 A Bayesian nonparametric model for longitudinal data

In order to capture the time-dependence in the longitudinal data observations, y_{it_j}, and at the same time provides a hypothesis testing procedure, able to detect the effects of the predictors, we propose the following hierarchical model

$$\begin{aligned} \mathsf{y}_{it_j} \mid t_j, \mathsf{x}, \mathcal{P} &\overset{\text{ind}}{\sim} \int N(\cdot \mid \mu, \sigma^2) P_{t,\mathsf{x}}(d\mu, d\sigma^2), \\ P_{t,\mathsf{x}} \mid H_\gamma &\sim \pi_{DDP}(\cdot \mid H_\gamma), \\ H_\gamma &\sim \pi_{\mathcal{M}}, \end{aligned} \tag{3-3}$$

where $\mathcal{P} = \{P_{t,\mathsf{x}} : t \in [0,T], \mathsf{x} \in \mathbb{R}^q\}$, $\pi_{DDP}(\cdot \mid H_\gamma)$ is the prior induced by a Dependent Dirichlet Process under the hypothesis H_γ, and $\pi_{\mathcal{M}}$ is a prior distribution defined on the hypotheses space $\mathcal{M}$. Model (3-3) follows the general form of (3-2), but it goes beyond the exchangeable case, including a Markovian dependent structure in the second hierarchical level and a prior on the hypotheses space. The Markovian dependence induces a richer correlation structure on the observations compared to the exchangeable case. The details of each component of model (3-3) are given in the following sections.

3.3.1 The space of hypotheses

To keep simplicity in the model, we assume that the predictors only affect the locations of the mixture model (3-3) through a linear form given by

$$\mu(\mathsf{x}_i) = \mathsf{z}_i \boldsymbol{\beta}, \tag{3-4}$$

where z_i is a design vector and $\boldsymbol{\beta}$ is a parameter vector. For instance, if we have a predictor vector given by $\mathsf{x} = (x_1, x_2)^t$, where x_1 and x_2 are discrete, each one with two levels. In such a case, we need two dummy variables to properly specify the design matrix. Thus, let $z_1 = 1$ for level 1 of x_1 and zero in other case. Analogously, let z_2 be the dummy variable for x_2, then a row of the design matrix could be given by $\mathsf{z}_i = (1, z_1, z_2, z_1 z_2)$, where $z_1 z_2$ represents an interaction term. Consequently, the parameter vector is $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$, where β_0 is the intercept. Here, any effect of the predictors is captured by $\{\beta_j; j = 1, 2, 3\}$. As usual, any possible effect should be relative to the intercept, which is commonly called the reference cell constrain. In this particular example, the elements of the hypothesis space $\mathcal{M}$ can be represented by the following binary sequence $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$, where $\gamma_j \in \{0, 1\}$ and $\gamma_j = 1$ if $\beta_j \neq 0$. Thus, for instance the model where only β_2 is different to zero is represented by $\boldsymbol{\gamma} = (0, 1, 0)$. The sequence $\boldsymbol{\gamma} = (0, 0, 1)$, which represents the model with locations $\mu(\mathbf{x}) = \beta_0 + \beta_3 z_1 z_2$ is not in the model space, because it does not accomplish the Strong Heredity condition (SH)(Nelder 1998). A statistical model fulfills the Strong

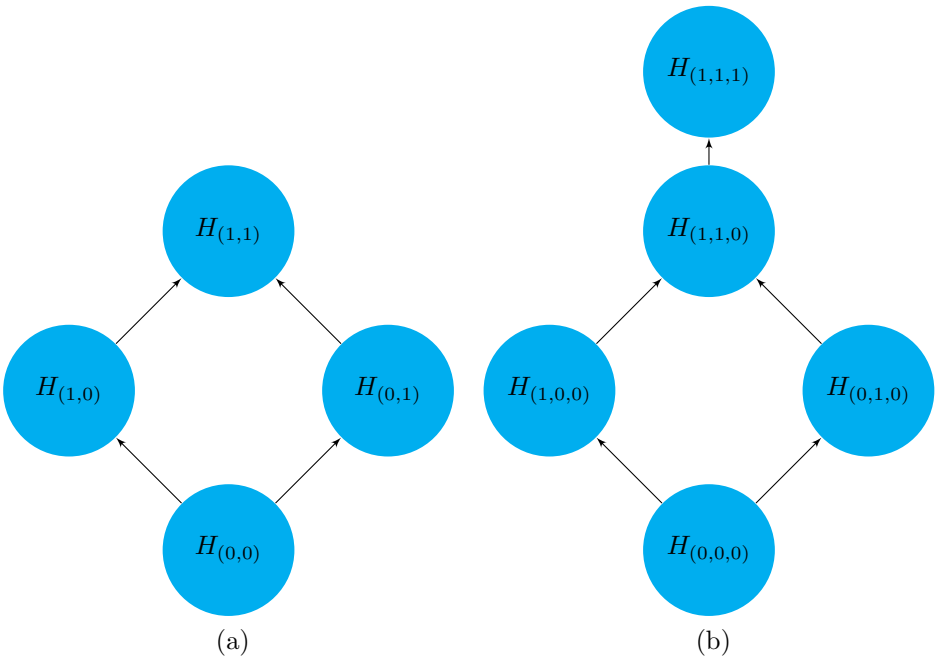


Figure **3-1**: Hasse diagram of the space of hypotheses $\mathcal{M}$ for a model with the main effects z_1, z_2 (a). Space of hypotheses $\mathcal{M}$ for a model with main effects z_1, z_2 and the interaction term $z_1 z_2$.

Heredity condition if for any predictor the very lower-order predictors associated with it are also in the model (Taylor-Rodríguez et al. 2016). In general, the model space considering interactions and satisfying the SH condition can be represented by the following set

$$\mathcal{M} := \{H_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} \in \{0,1\}^p \text{ such that } \boldsymbol{\gamma} \text{ satisfies the SH condition}\}. \tag{3-5}$$

The size of the space $\mathcal{M}$ is 2^p, for a model without interactions, where p is the number of columns in z minus one, while the size of the space of hypothesis for a model that includes interaction is

$$\sum_{j=0}^{p} \binom{p}{j} + \sum_{j=2}^{\ell} \sum_{\{k:k\geq j\}}^{\ell} \sum_{i\geq 1} \binom{\frac{p!}{j!(p-j)!}}{i},$$

where i takes values from 1 to the total number of terms with j-order in the model, and ℓ is the maximum order between the terms. Figure **3-1** shows the model space for the above examples, that is, with and without interactions when $p = 3$.

The specification of the binary vectors $\boldsymbol{\gamma}$ is relevant, because the definition of a prior on the hypotheses space is equivalent to the definition of a prior on the model space given by the binary vectors $\boldsymbol{\gamma}$. Therefore, the posterior probability of a particular hypothesis $H_{\boldsymbol{\gamma}}$ is equal to the posterior probability on $\boldsymbol{\gamma}$, that is $\pi(H_{\boldsymbol{\gamma}} \mid \boldsymbol{y}) = \pi(\boldsymbol{\gamma} \mid \boldsymbol{y})$.

3.3.2 Prior on the hypotheses space

In this section, we will discuss the choice of the prior distribution on the space $\mathcal{M}$. The choice of the prior distribution on the space of hypotheses requires careful consideration as seemingly innocuous alternatives can have undesirable consequences (Gutiérrez et al. 2019). As expected, the posterior inference is remarkably sensitive to this prior information for small and moderate sample sizes. For example, it has been a common practice to assume equal prior probabilities on all hypotheses; however, this seemingly non-informative alternative favors models of a particular level of complexity, making this choice inadequate (Taylor-Rodríguez et al. 2016). As a result, we would like a prior that assigns probability to the hypothesis according to the level of model complexity. Whit this in mind, we propose to use the Hierarchical Order Prior (HOP) proposed by Taylor-Rodríguez et al. (2016) for the space $\mathcal{M}$. This prior specification produces strong penalization as the model become more complex.

To define the hierarchical order prior structure, let $\mathcal{M}$ be the space that involves models, such that the order is less than or equal to ℓ and that satisfy the SH condition. In $\mathcal{M}$, the null hypothesis $H_{(0,0,\ldots,0)}$ represents the base model, denoted by $\tilde{\mathsf{m}}_{\mathsf{B}}$, that is, a model that only contains the intercept, while $H_{(1,1,\ldots,1)}$ is the full model denoted by $\tilde{\mathsf{m}}_{\mathsf{F}}$. In addition, let $\pi_j(\tilde{\mathsf{m}}) = \pi\left(\gamma_j(\tilde{\mathsf{m}}) = 1 \mid \gamma_{\mathcal{P}(j)}, \mathcal{M}\right)$ be the conditional inclusion probability for j-term in the model $\tilde{\mathsf{m}}$, where $\gamma_j(\tilde{\mathsf{m}})$ is the indicator function that describes whether j is included in $\tilde{\mathsf{m}}$ and $\gamma_{\mathcal{P}(j)}$ denotes the parent set for j, which contains all the lower-order predictors associated with the j-term. Under the HOP, for a ℓ-order term $j^{(\ell)}$, let us define the prior as $\pi_j(\tilde{\mathsf{m}}) = \pi^{(\ell)}(\tilde{\mathsf{m}})$ for all $j \in \Upsilon_{(\ell)}(\tilde{\mathsf{m}}) \cup \mathcal{C}_{(\ell)}(\tilde{\mathsf{m}})$, where $\Upsilon_{(\ell)}(\tilde{\mathsf{m}}) = \{j \in H_{\boldsymbol{\gamma}} : \text{order}(j) = \ell\}$ and $\mathcal{C}_{(\ell)}(\tilde{\mathsf{m}}) = \{j^{(\ell)} \in H_{\boldsymbol{\gamma}_{\tilde{\mathsf{m}}_{\mathsf{F}}}} : \tilde{\mathsf{m}} \cup j^{(\ell)} \text{satisfies the SH condition}\}$. If we assume that $\pi^{(\ell)}(\tilde{\mathsf{m}}) \sim \mathsf{Be}(a_\ell, b_\ell)$ with $a_\ell = 1$ and $b_\ell = |\Upsilon_{(\ell)}(\tilde{\mathsf{m}}) \cup \mathcal{C}_{(\ell)}(\tilde{\mathsf{m}})|$ for all ℓ, together with independence across the different orders, we have a prior distribution for the hypotheses space given by

$$\pi(H_{\boldsymbol{\gamma}} \mid \mathcal{M}, a_\ell, b_\ell) = \prod_{\ell=\mathcal{L}_{\mathcal{M}}^{\min}}^{\mathcal{L}_{\mathcal{M}}^{\max}} \left(\mathsf{Be}(|\Upsilon_{(\ell)}(\tilde{\mathsf{m}})| + a_\ell, |\mathcal{C}_{(\ell)}(\tilde{\mathsf{m}})| + b_\ell)\right) / \mathsf{Be}(a_\ell, b_\ell), \tag{3-6}$$

where $\mathcal{L}_{\mathcal{M}}^{\min}$ and $\mathcal{L}_{\mathcal{M}}^{\max}$ is the minimum and maximum ℓ-order, respectively, in $\tilde{\mathsf{m}}_{\mathsf{F}}$. Following Taylor-Rodríguez et al. (2016), we will use the parameterization $a_\ell = 1$ and $b_\ell = |\Upsilon_{(\ell)}(\tilde{\mathsf{m}}) \cup \mathcal{C}_{(\ell)}(\tilde{\mathsf{m}})|$ in $\pi^{(\ell)}(\tilde{\mathsf{m}})$, which produces a hierarchical version of the penalization proposed by Wilson et al. (2010).

Figure **3-2** shows the prior specification for the models described in Figure **3-1**, that is, a model in the space $\mathcal{M}$ considering only the main effects, and a model in the same space that includes also the interaction of the first order.

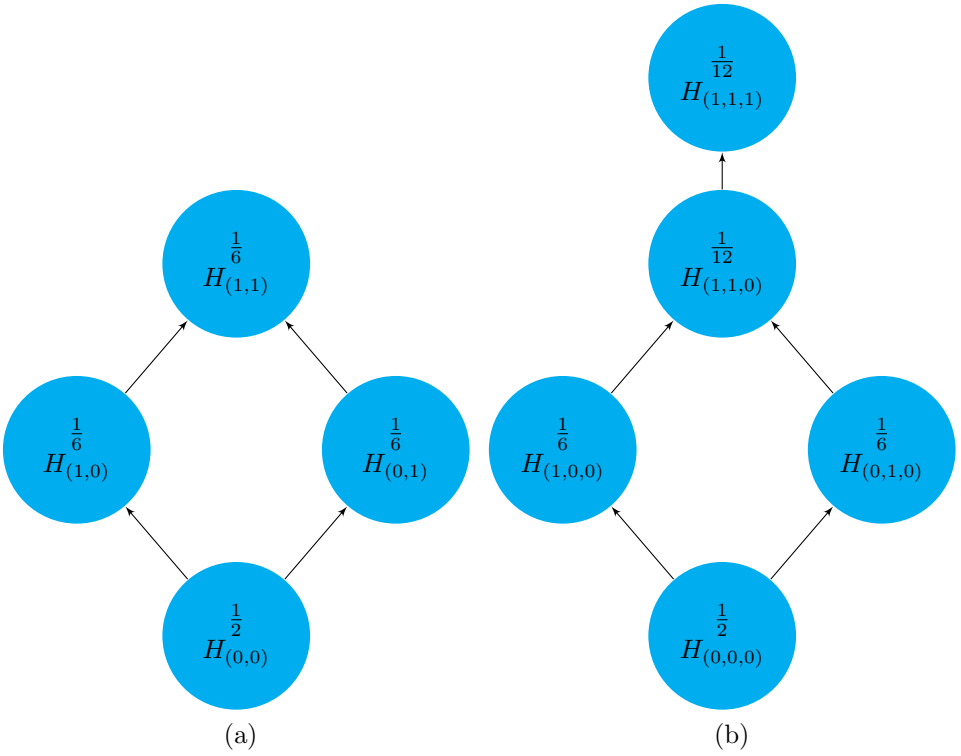


Figure **3-2**: Prior distribution defined in the space of hypotheses $\mathcal{M}$ for a model with the main effects z_1 and z_2 (a). Prior specification in the space $\mathcal{M}$ for a model with main effects z_1, z_2 and the interaction term $z_1 z_2$.

3.3.3 Prior distribution induced by the random process $\mathcal{P}$

The elements of the class $\mathcal{P}$ are given by

$$\mathsf{P}_{t,\mathsf{x}}(\cdot) = \sum_{l\geq 1} \mathsf{w}_l(t)\, \delta_{(\mu_l(\mathsf{x}),\sigma_l^2)}(\cdot), \tag{3-7}$$

where

$$\mathsf{w}_l(t) := \mathsf{v}_l(t) \prod_{k<l} (1-\mathsf{v}_k(t)). \tag{3-8}$$

In (3-7) and (3-8), $\mathsf{v} = (\mathsf{v}_\infty(t); t \geq 0)$ and $\theta = (\mu_\infty(\mathsf{x}), \sigma^2_\infty; \mathsf{x} \in \mathbb{R}^q)$ are independent collections of independent stochastic processes $\mathsf{v}_\infty(t) := \{\,(\mathsf{v}_l(t)) : t \in [0,T], l = 1,2,\ldots\}$ and $\theta_\infty(\mathsf{x}) := \{(\mu_l(\mathsf{x})), (\sigma_l^2) : \mathsf{x} \in \mathbb{R}^q, l = 1,2,\ldots\}$ each taking values on $(0,1)$ and $\mathbb{X} := \mathbb{R}\times\mathbb{R}_+$, respectively, and with the restriction that $\sum_l \mathsf{w}_l(t) = 1$ for each $t \in [0,T]$. The construction of equation (3-7) has been widely studied in the literature, some references are Gelfand & Kottas (2001), Griffin & Steel (2006), Rodríguez & Dunson (2011), Mena et al. (2011), Gutiérrez et al. (2016), among many others. A excellent trade off choice is the dependent process proposed by Gutiérrez et al. (2016), as it provides with a general and tractable model for time-evolving densities. Here, we add the covariate-dependence to this later approach. While a lot has been said about DDP constructions, to the best of our knowledge, little has been said about their use in longitudinal data analysis in order to detect the effect of the predictors using a hypothesis testing procedure. Here we look for this new venture.

A key aspect to build a time dependent density model is to select a suitable stochastic process for the sticks and/or the particles. In general, the desirable properties of a stochastic process for time dependent density estimation are: 1) ability to share information between the different times to assure that the resulting model is not over-parametrized, 2) flexibility to capture the changes in the density across the time, 3) simplicity to facilitate the posterior inference. A general class of stochastic processes that satisfies properties 1) to 3) is the strictly stationary Markovian processes. In fact, the Markov property offers a key advantage to the hour of infer from the data, because the learning process depend only of the last state. Additionally, the strictly stationary property reduces the number of parameters in the model, because the correlation structure, in this kind of processes, is habitually governed by a few parameters.

We propose to use a strictly stationary Markovian process for $\mathsf{v}_\infty(t) := \{\,(\mathsf{v}_l(t))_{t\geq 0} : l = 1,2,\ldots\}$. We denote the transition probabilities by $\mathsf{P}^{\mathsf{v}}_t(\mathsf{v}_t \mid \mathsf{v}_0)$ and the invariant distribution by π_{v}. Notice that, for instance, if π_{v} coincides with a Beta distribution, $\mathsf{Be}(1,M)$, thus the invariant distribution of G_t is a Dirichlet process centered at $\mathbb{E}[\mathsf{G}_t] = \mathsf{G}_0$, namely $\mathsf{G}_t \sim \mathcal{DP}(M\mathsf{G}_0)$ marginally.

In particular, we propose to use the following jump process for the sticks

$$\mathsf{P}^{\mathsf{v}}_t(\mathsf{v}_l(t_j) \in B \mid \mathsf{v}_l(t_{j-1})) \;=\; (1-\alpha^d_{\mathsf{v}})\pi_{\mathsf{v}}(B) + \alpha^d_{\mathsf{v}}\delta_{\mathsf{v}(t_{j-1})}(B), \tag{3-9}$$

where $\delta_x(B)$ is the Dirac measure at x, $d = t_j - t_{j-1}$, and α_{v} is a continuous parameter in the interval $(0, 1)$, whose prior distribution is assumed as $\mathsf{Be}(a_0, b_0)$, and whose hyperparmeters were fixed at $a_0 = b_0 = 1$.

Proposition 2. The correlation between the sticks induced by the jump process in (3-9) is given by $\mathbb{C}orr(\mathsf{v}_t, \mathsf{v}_{t+d}) = \alpha^d$.

Proposition (2) implies that the jump process for the sticks induces positive correlation, which is going down to zero as d is going to infinity. The proof is provided in the Appendix A 3.8.

Proposition 3. Under model (3-3) we have that

$$\mathbb{C}orr(Y_t, Y_{t+d}) = \frac{\left(2\mu_{\mathsf{v}} - \mu_{\mathsf{v}}^{(2)}\right)\left(\alpha^d \sigma_{\mathsf{v}} + \mu_{\mathsf{v}}^2\right)}{\mu_{\mathsf{v}}^{(2)}\left(2\mu_{\mathsf{v}} - \mu_{\mathsf{v}}^2 - \alpha^d \sigma_{\mathsf{v}}\right)}, \tag{3-10}$$

where $\mu_{\mathsf{v}} := \mathbb{E}[\mathsf{v}]$, $\mu_{\mathsf{v}}^{(2)} := \mathbb{E}[\mathsf{v}^2]$, and $\sigma_{\mathsf{v}} := \mathbb{V}ar[\mathsf{v}]$.

Proposition (3) provides a general expression for the correlation between two observations separated by a distance d. This expression is a function of the first and second moment of the sticks and of course, it is also a function of the correlation between the sticks. The proof is given in the Appendix A 3.8.

For the particular case when $\mathsf{v}_{lt} \overset{\text{iid}}{\sim} \text{Be}(1, M)$, we have that $\mu_{\mathsf{v}} = {}^{1}/_{1+M}$, $\mu_{\mathsf{v}}^{(2)} = {}^{2}/_{(1+M)(2+M)}$ and $\sigma_{\mathsf{v}} = {}^{M}/_{(1+M)^2(2+M)}$. Now, using (3-10), we have

$$\mathbb{C}orr(Y_t, Y_{t+d}) = \frac{(1+M)(2+M+\alpha_{\mathsf{v}}^d M)\alpha_{\mathsf{v}}^d}{(2+M)(1+2M) - \alpha_{\mathsf{v}}^d M}. \tag{3-11}$$

In (3-11) as $\alpha_{\mathsf{v}} \in (0, 1)$, if $d \longrightarrow \infty$, then $\text{Corr}(Y_t, Y_{t+d}) \longrightarrow 0$; conversely, if $d \longrightarrow 0$ then $\text{Corr}(Y_t, Y_{t+d}) \longrightarrow 1$.

To complete the prior specification of the elements of $\mathcal{P}$, we assumed Gamma prior distribution for the precision parameter M, that is, $\mathsf{Ga}(M \mid a_1, b_1)$, where a_1 and b_1 were fixed at 0.01.

3.3.4 Prior distribution induced by the atoms of the Gaussian kernel

Let $\theta_l = (\mu_l(\mathsf{x}), \sigma_l^2)_{l \geq 1}$ be the atoms for the mixture model (3-3), where $\mu_l(\mathsf{x})$ is the linear combination of (3-4) and σ_l^2 is an scalar greater than zero. We specify the base measure $\mathsf{P}_{0|\gamma}$ in the DDP conditionally on the hypothesis H_γ. The specification is given by a spike-slab prior (George & McCulloch 1993, Ishwaran & Rao 2005, Ročková & George 2016), which allow us to test the hypothesis about the effects of the predictors on the response variable. Formally, the base measure conditionally on H_γ or equivalently on

$\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)$ is defined as

$$\mathsf{P}_{0|\boldsymbol{\gamma}} : \mathsf{N}\left\{(\beta_0, \ldots, \beta_p)^T \mid \boldsymbol{\mu}_{0|\boldsymbol{\gamma}}, \Sigma_{0|\gamma}\right\} \mathrm{Ga}\left(\frac{1}{\sigma^2} \mid \epsilon, \epsilon\right), \tag{3-12}$$

where $\mathsf{N}\left(\cdot \mid \mu_{0|\gamma}, \Sigma_{0|\gamma}\right)$ denotes a multivariate normal distribution and $\mathrm{Ga}(\epsilon, \epsilon)$ denotes a gamma distribution. The hyper-parameter $\boldsymbol{\mu}_{0|\gamma}$ was fixed at $\mathbf{0}$ and we suppose that $\Sigma_{0|\gamma}$ is given by a diagonal matrix, such that

$$\Sigma_{0|\gamma} = \mathrm{diag}\left[\psi_0, \psi_1\left(\mathbb{1}_{(\gamma_1=1)} + \nu_0 \mathbb{1}_{(\gamma_1=0)}\right), \ldots, \psi_p\left(\mathbb{1}_{(\gamma_p=1)} + \nu_0 \mathbb{1}_{(\gamma_p=0)}\right)\right]. \tag{3-13}$$

The specification of the base measure is completed by the hyperprior on ψ_j

$$\frac{1}{\psi_j} \sim \mathrm{Ga}\left(a, b\right), \qquad j = 1, \ldots, p.$$

The values for a and b are choosen such that the prior for $\mathrm{Var}\left(\beta_j\right)$ is a continuous bimodal distribution, with spike at ν_0 and right continuous tail for the slab component as in George & McCulloch (1993), and Ishwaran & Rao (2005). In (3-13) the parameter ν_0 is known as the shrink parameter, and is chosen such that $\nu_0 \ll 1$. Note that, under the above parametrization when $\gamma_j = 0$ the prior is a spike distribution and the atoms $\{\beta_{j,l}\}_{l\geq 1}$ are close to zero, supporting H_0. Conversely, when $\gamma_j = 1$ the distribution is slab, and the atoms $\{\beta_{j,l}\}_{l\geq 1}$ are far from zero, which support the alternative hypothesis. Following Ishwaran & Rao (2005) we fixed $a = 5$, $b = 0.3$ and $\nu_0 = 0.1$. The rest of hyperparmeters ψ_0 and ϵ were fixed at 10 and 0.1, respectively, so that the prior distributions were uninformative.

Remark 2. From model (3-3) we have that the marginal distribution at time $t = t'$, is given by

$$f_{\mathsf{z},t=t'}\left(\mathsf{y}\right) = \sum_{l\geq 1} \mathsf{w}_l(t')\mathrm{N}(\mathsf{y} \mid \mathsf{z}\boldsymbol{\beta}_l, \sigma_l^2), \tag{3-14}$$

thus, the weights and scales are common to all predictors levels, and the possible effect of the predictors is captured by the locations β_j, $j = 1, \ldots, p$, as described in Section 3.3.

Indeed, with the above specification we provide an approximation of $P\left(H_0 \mid \text{Data}\right)$. An exact computation of $P\left(H_0 \mid \text{Data}\right)$ involves the use of a spike distribution concentrated exactly in zero, like as a Dirac measure. Unfortunately, posterior inference with the Dirac measure would require to draw $\boldsymbol{\gamma}$ from the marginal posterior $P(\boldsymbol{\gamma} \mid \mathsf{y})$ integrating over the infinite dimensional parameter $\{\beta_{j,l}\}_{l\geq 1}$ (Geweke 1996, Smith & Kohn 1996, Malsiner-Walli et al. 2011), which is unviable for the model in (3-3).

3.4 Posterior Inference

In this Section, we develop a general scheme of the posterior algorithm for the mixture model (3-3), the details are provided in the Appendix B 3.9. It is well known, that given

the discrete nature of the Dirichlet process, model (3-3) can be rewritten as a weighted infinite sum of continuous kernels

$$f_{t,\mathsf{x}}(\mathsf{y} \mid \mathsf{v}_\infty(t), \theta_\infty(\mathsf{x})) = \sum_{l\geq 1} \mathsf{w}_l(t)\mathrm{N}(\mathsf{y} \mid \theta_l(\mathsf{x})). \tag{3-15}$$

The model in (3-15) has potentially an infinite number of components. In practice, one must take a strategy to deal with the infinite-dimensional parameters. For this, our plan is to develop a Gibbs sampler algorithm with slice sampling steps as in Walker (2007). In this algorithm, we have to consider an augmented model given by

$$f_{t,\mathsf{x}}(\mathsf{y}, \mathsf{u}, \mathsf{s} \mid \mathsf{v}_\infty(t), \theta_\infty(\mathsf{x})) = \sum_{l\geq 1} \mathbb{1}(\mathsf{u} < \mathsf{w}_\mathsf{s}(t))\mathrm{N}(\mathsf{y} \mid \theta_\mathsf{s}(\mathsf{x})), \tag{3-16}$$

where s denotes the allocation variable of y and u is a uniform random variable in the interval $(0, \mathsf{w}_\mathsf{s})$. In model (3-16), we should highlight that it is a finite model, because only a limited number of $\mathsf{w}_\mathsf{s}(t)$ satisfies the condition that $(\mathsf{u} < \mathsf{w}_\mathsf{s}(t))$. This number depends on the data complexity, note also that the variable u is included without modifies the original density, because if we marginalize over u in (3-16) we return to model (3-15). Therefore, a crucial step in the Gibbs sampling is to determinate a set $\{1, \ldots, N\}$ over which s will take values, here N is $\max_i N_i$ and N_i is the largest integer h for which $\mathsf{w}_h > \mathsf{u}_i$. Thus, once N is determined the algorithm is reduced to sample a finite number of weights and atoms. Now, given that in model (3-3) the dependency among Y is modeled through the process $\mathsf{v}_\infty(t)$, then the observations y_{ij} are conditionally independents and the joint likelihood is given by

$$\mathcal{L}(\mathsf{Y}_{\widetilde{n}}, \mathsf{U}_{\widetilde{n}}, \mathsf{S}_{\widetilde{n}} \mid \mathsf{v}_\infty^{(n_i)}, \theta_\infty) = \prod_{i=1}^{m}\prod_{j=1}^{n_i} \mathbb{1}(\mathsf{u}_{ij} < \mathsf{w}_{\mathsf{s}_{ij}}(t_j))\mathrm{N}(\mathsf{y}_{\mathsf{ij}} \mid \mu_{\mathsf{s}_{ij}}(x), \sigma^2_{\mathsf{s}_{ij}}), \tag{3-17}$$

where $\mathsf{v}_\infty^{(n_i)} := \{\mathsf{v}_\infty(t_j)\}_{j=1}^{n_i}$ denotes an infinite collection of the stick-breaking components sample at times $(t_1, \ldots, t_{n_i})$, $\mathsf{Y}_{\widetilde{n}} = \left(\mathsf{y}_{1t_j}, \mathsf{y}_{2t_j}, \ldots, \mathsf{y}_{mt_j}\right)^T$ represents the vector of observations where y_{it_j} are the subvectors with the observations for each subject, namely, $\mathsf{y}_{it_j} = (y_{i1}, y_{i2}, \ldots, y_{in_i})$, with $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$. $\mathsf{U}_{\widetilde{n}} = \left(\mathsf{u}_{1t_j}, \mathsf{u}_{2t_j}, \ldots, \mathsf{u}_{mt_j}\right)^T$ denotes the vector of slice variables, here u_{it_j} represents the subvector $(u_{i1}, u_{i2}, \ldots, u_{in_i})$. Finally, $\mathsf{S}_{\widetilde{n}} = \left(\mathsf{s}_{1t_j}, \mathsf{s}_{2t_j}, \ldots, \mathsf{s}_{mt_j}\right)^T$ represents the vector of the membership variables, that contains the subvectors $\mathsf{s}_{it_j} = (s_{i1}, s_{i2}, \ldots, s_{in_i})$. We use the subscript $\widetilde{n}$ for indicating the total number of observations, namely, $\sum_{i=1}^{m} n_i$.

With all elements defined above, we present a straightforward Gibbs sampling algorithm to the posterior inference on the model in (3-3).

Algorithm [1]

The full conditionals distributions are given by:

1. $p\left(\boldsymbol{\beta}_l \mid \ldots\right) := \mathrm{N}_p\left\{\left(\Sigma_{0|\gamma}^{-1} + \mathbf{Z}_A^T\mathbf{Z}_A/\sigma_l^2\right)^{-1}\left(\Sigma_{0|\gamma}^{-1}\boldsymbol{\mu}_{0|\gamma} + \mathbf{Z}_A^T\mathbf{y}_A/\sigma_l^2\right), \left(\Sigma_{0|\gamma}^{-1} + \mathbf{Z}_A^T\mathbf{Z}_A/\sigma_l^2\right)^{-1}\right\}$,

2. $p\left(\frac{1}{\psi_j} \mid \ldots\right) := \mathrm{Ga}\left(a + \frac{N}{2}, b + \frac{\sum_{l=1}^N \beta_{jl}^2}{2\left(\mathbb{1}_{(\gamma_j=1)} + \nu_0\mathbb{1}_{(\gamma_j=0)}\right)}\right)$,

3. $p\left(\frac{1}{\sigma_l^2} \mid \cdots\right) := \mathrm{Ga}\left(\kappa + \frac{|A|}{2}, \kappa + \frac{SSR(\boldsymbol{\beta}_l)}{2}\right)$,

4. $p(\mathsf{u}_{ij} \mid \cdots) := \mathsf{U}(0, \mathsf{w}_{\mathsf{s}_{ij}})$,

5. $P(\mathsf{s}_{ij} = k \mid \cdots) \propto \mathbb{1}(k : \mathsf{w}_k(t_j) > \mathsf{u}_{ij})\,\mathrm{N}(\mathsf{y}_{ij} \mid \theta_k)$,

6. $p(\mathsf{v}_l(t_1) \mid \cdots) := q_{01}\mathrm{Be}(\mathsf{v}_l(t_1) \mid 1 + n_{i1}, M + \kappa_{i1}) + q_{02}\mathbb{1}_{\{\mathsf{v}_l(t_1)=\mathsf{v}_l(t_2)\}}$,

7. $p(\mathsf{v}_l(t_j) \mid \cdots) := q_{0j}\mathrm{Be}(\mathsf{v}_l(t_j) \mid 1 + n_{ij}, M + \kappa_{ij}) + q_{1j}\mathbb{1}_{\{\mathsf{v}_l(t_j)=\mathsf{v}_l(t_{j-1})\}} +$
$q_{2j}\mathbb{1}_{\{\mathsf{v}_l(t_j)=\mathsf{v}_l(t_{j+1})\}} + q_{3j}\mathbb{1}_{\{\mathsf{v}_l(t_j)=\mathsf{v}_l(t_{j+1})=\mathsf{v}_l(t_{j-1})\}}$, for $j \neq \{1, \tau\}$,

8. $p(\mathsf{v}_l(t_\tau) \mid \cdots) := q_{0\tau}\mathrm{Be}(\mathsf{v}_l(t_\tau) \mid 1 + n_{i\tau}, M + \kappa_{i\tau}) + q_{1\tau}\mathbb{1}_{\{\mathsf{v}_l(t_\tau)=\mathsf{v}_l(t_{\tau-1})\}}$,

9. $p(\alpha_\mathsf{v} \mid \cdots) \propto \mathrm{Be}\left(a_0, b_0\right) \times \prod_{l=1}^N\{p\left(\mathsf{v}_l\left(t_1\right) \mid \cdots\right) \times \prod_{j=2}^{T-1} p\left(\mathsf{v}_l\left(t_j\right) \mid \cdots\right) \times p\left(\mathsf{v}_l\left(t_T\right) \mid \cdots\right)\}$,

10. $$P\left(\boldsymbol{\gamma} = (0, 0, \ldots, 0) \mid \ldots\right) \propto \pi_{\mathcal{M}}\left(H_{(0,0,\ldots,0)}\right)\prod_{l=1}^N\prod_{j=1}^p \mathrm{N}\left(\beta_{jl} \mid 0, \nu_0\psi_j\right),$$
$$P\left(\boldsymbol{\gamma} = (1, 0, \ldots, 0) \mid \ldots\right) \propto \pi_{\mathcal{M}}\left(H_{(1,0,\ldots,0)}\right)\prod_{l=1}^N \mathrm{N}\left(\beta_{1l} \mid 0, \psi_1\right)\prod_{j=2}^p \mathrm{N}\left(\beta_{jl} \mid 0, \nu_0\psi_j\right),$$
$$\vdots$$
$$P\left(\boldsymbol{\gamma} = (1, 1, \ldots, 1) \mid \ldots\right) \propto \pi_{\mathcal{M}}\left(H_{(1,1,\ldots,1)}\right)\prod_{l=1}^N\prod_{j=1}^p \mathrm{N}\left(\beta_{jl} \mid 0, \psi_j\right).$$

In Algorithm [1], A denotes an index set that contains the indexes of the observations that belong to the cluster l, namely, $A = \{i, j : \mathsf{s}_{ij} = l\}$ and $|A|$ denotes the size of the set. Therefore, $\mathbf{y}_A$ is a subvector from the vector of observations $\mathsf{Y}_{\widetilde{n}}$ and $\mathbf{Z}_A$ a submatrix of the design matrix $\mathbf{Z}_{(m\times(p+1))}$. In step 2, $SSR(\boldsymbol{\beta}_l)$ corresponds to the expression $\mathbf{y}_A^T\mathbf{y}_A - 2\boldsymbol{\beta}_l^T\mathbf{Z}_A^T\mathbf{y}_A + \boldsymbol{\beta}_l^T\mathbf{Z}_A^T\mathbf{Z}_A\boldsymbol{\beta}_l$. Finally, in order to facilitate the readability of the manuscript, the details for sampling the sticks in the steps 6 to 8, together with the Metropolis-Hasting steps used at 9 and 10 are given in the Appendix B 3.9. This Appendix also provides the methodology used for the updating of the mass parameter M in the DDP.

3.4.1 Visualization of the differences

If the posterior evidence favors the alternative hypothesis, we would like to visualize the differences in the response between the levels of the predictor in a given time. Without loss of generality, if we suppose x_1 as a discrete predictor with two levels, from (3-14) we have that for $x_1 = 0$, the cumulative density function is denoted by $F_{t,0}(\cdot) = \sum_{l \geq l} \mathsf{w}_l(t)\Phi(\cdot \mid \beta_{0l}, \sigma_l^2)$ and for $x_1 = 1$, $F_{t,1}(\cdot) = \sum_{l \geq l} \mathsf{w}_l(t)\Phi(\cdot \mid \beta_{0l} + \beta_{1l}, \sigma_l^2)$, where $\Phi(\cdot)$ denotes the Gaussian cumulative density function. To visualize the differences between $F_{t,0}(\cdot)$ and $F_{t,1}(\cdot)$, we propose to compute the *Shift function* as a measure of the difference between the two populations at $t = t'$.

The *Shift function* was proposed by Doksum (1974) and Doksum & Sievers (1976). The idea behind the Doksum's proposal is to find a function $\Delta(\cdot)$, such that, $Y_{(0,t')} + \Delta\left(Y_{(0,t')}\right)$ has the same distribution as $Y_{(1,t')}$. Here, $Y_{(0,t')}$ and $Y_{(1,t')}$ denotes the responses at time t', with the level $x_1 = 0$ and $x_1 = 1$, respectively. Formally, $\Delta(\cdot)$ is a function such that $F_{t,0}\left(Y_{(0,t')}\right) = F_{t,1}\left(Y_{(1,t')} + \Delta\left(Y_{(1,t')}\right)\right)$ or equivalently, $\Delta\left(Y_{(0,t')}\right) = F_{t,1}^{-1}\left\{F_{t,0}\left(Y_{(0,t')}\right)\right\} - Y_{(1,t')}$. Note that, if $F_{t,0}(\cdot) = F_{t,1}(\cdot)$ then $\Delta(Y)$ is equal to 0, for all Y at $t = t'$. Conversely, $\Delta(Y)$ is different from 0 for some set $A := \{Y : \Delta(Y) \neq 0 \text{ at } t = t'\}$. The set A provides information on what regions of the distribution are different at the time $t = t'$.

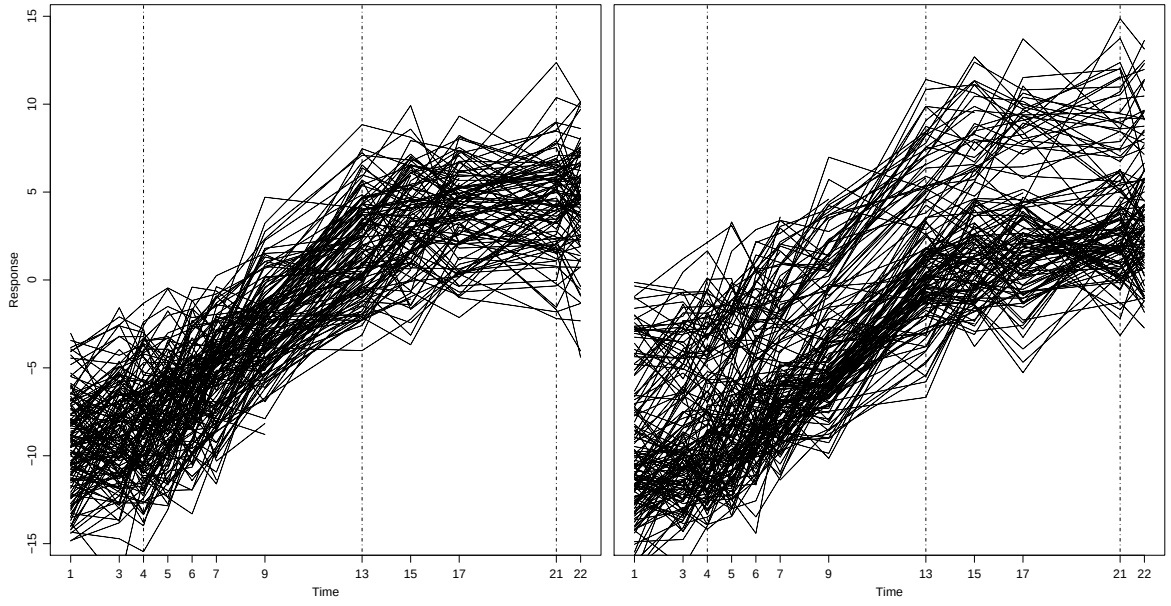
Deriving the shift function using Algorithm [1] is immediate, since for each iteration of the Gibbs algorithm, we have posterior random realizations of $F_{t,0}^{(\ell)}$ and $F_{t,1}^{(\ell)}$, $\ell = 1, \ldots, B$. Defining the left inverse of $F_{t,1}$ as $F_{t,1}^{-1}(u) = \inf\{q : F_{t,1}(q) \geq u\}$, a random realization of $\Delta(Y)^{(\ell)}$ at the time $t = t'$ can be computed as

$$\Delta(Y)^{(\ell)} = \begin{cases} F_{t,1}^{-1(\ell)}\left\{F_{t,0}^{(\ell)}(Y)\right\} - Y & \text{if } \gamma^{(\ell)} = 1, \\ 0, \forall \quad Y & \text{if } \gamma^{(\ell)} = 0. \end{cases} \tag{3-18}$$

With the posterior realizations of the shift function, it is possible to compute some functionals, as the sample posterior mean $\bar{\Delta}(Y)$ and a 95% credible set at each point in time t. The credible set is particularly useful to determine the set A, which can be visualized looking at the values of Y such that $\Delta(Y) \neq 0$.

3.5 Data illustration

In this section, we illustrate the use and performance of our model with simulated data. We simulated responses $\mathbf{y_i}$ for a total of $m = 300$ subjects. Each individual has a maximum of twelve measurements over the times, $t_1 = 1$, $t_2 = 3$, $t_3 = 4$, $t_4 = 5$, $t_5 = 6$, $t_6 = 7$, $t_7 = 9$, $t_8 = 13$, $t_9 = 15$, $t_{10} = 17$, $t_{11} = 20$, $t_{12} = 22$. Additionally, as usual in longitudinal studies, we suppose a percentage of the subjects are leaving the study as time progresses. The percentage of individuals leaving the study at each time are: 5% at t_6 and t_7, a 10% at t_8, 5% at t_9 and finally, a 2.5% at t_{10} and t_{11}, thus at t_{12} we have lost 30% of the subjects that began the study. The index i of the individuals that leave the study was selected randomly. We suppose that 153 individuals received the treatment and 147 are in the control group, then the covariate z_i is equal to 1 for individuals in the treatment



(a) (b)

Figure **3-3**: Profile plot of responses over the time. Consecutive observations within a subject are connected by a line. (a) Profiles of the control group, (b) Profiles of the treatment group. The vertical dashed lines indicate the points at which the density section was estimated in the Figure **3-4**.

group and 0 for individuals in the control group. Then, under this setup, we generated the responses for each individual from the following model

$$\mathbf{y}_i = \boldsymbol{\beta}_i + \boldsymbol{f}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, 300, \tag{3-19}$$

where $\mathbf{y}_i = \left(\mathsf{y}_{it_1}, \ldots, \mathsf{y}_{it_{n_i}}\right)^t$, $\boldsymbol{\beta}_i = (\beta_0 + \beta_1 \mathsf{z}_i) \times \mathbf{1}^t_{(1\times n_i)}$, z_i is the dichotomy variable defined above without dependence of the time. The vector $\boldsymbol{f}_i = \left(f\left(t_1\right), \ldots, f\left(t_{n_i}\right)\right)^t$ has components $f(t_j) = \left({}^{15}/_{1+\exp\{-(t_j-9)/3\}}\right) - 12$. In addition, $\boldsymbol{\epsilon}_i = \left(\epsilon_{i1}, \ldots, \epsilon_{it_{n_i}}\right)^t$ is distributed as a mixture of multivariate Gaussian distributions and its distribution is conditional on z_i as follows,

$$f\left(\boldsymbol{\epsilon}_i \mid \mathsf{z}_i\right) = \begin{cases} 0.4 \times \mathsf{N}_{ni}\left(-2 \times \mathbf{1}^t_{(1\times n_i)}, \Sigma\right) + 0.6 \times \mathsf{N}_{ni}\left(\frac{4}{3} \times \mathbf{1}^t_{(1\times n_i)}, \Sigma\right), & \text{if } \mathsf{z_i} = 0 \\ 0.7 \times \mathsf{N}_{ni}\left(-2 \times \mathbf{1}^t_{(1\times n_i)}, \Sigma\right) + 0.3 \times \mathsf{N}_{ni}\left(\frac{14}{3} \times \mathbf{1}^t_{(1\times n_i)}, \Sigma\right), & \text{if } \mathsf{z_i} = 1 \end{cases}$$

with $\mathbb{E}\left(\epsilon_{ij}\right) = 0 \quad \forall i, t_j$ and covariance matrix $\Sigma = [\sigma_{ij}]^{n_i}_{i,j=1}$, where $\sigma_{ij} = \lambda^2 \alpha^{|t_j - t_i|}$, $\lambda > 0$ and $\alpha \in (0, 1)$. Specifically, the model in 3-19 was set up with $\beta_0 = 1.5$, $\beta_1 = 0.2$ and for the autoregressive structure, we fixed λ at 5 and α in 0.6.

Figure **3-3** shows the profiles of the responses simulated for the subjects. The Model 3-3 was fitted via the Gibbs Algorithm [1] of Section 3.4, we run 40,000 iterations, burn-in a period of 8,000 and thinning the sample by keeping only each 32th draw of the sample parameters.

To get a basic idea of the quality of the estimations, we provide posterior inferences for some time points. Figure **3-4** presents the true and estimated densities at the sections that we choose in Figure **3-3**, the estimation corresponds to the posterior mean. The figures also include gray regions for representing the point-wise 95% credible intervals. The posterior probability estimate for our procedure to the hypotesis H_1, $\hat{P}(\gamma_1 = 1 \mid \ldots)$, was approximately 1. Additionally, Figure **3-5** shows the performance of our proposal for estimating the correlation structure, the estimation corresponds to the posterior mean of the correlation together with the point-wise 95% credible intervals. Finally, it is important to note that our BNP longitudinal test was able to detect small differences, in spite of it be close to 0 ($\beta_1 = 0.2$).

3.6 Application to real data sets

In this section, we develop an application of our model to real data. The dataset comes from an epidemiology study conducted in the Netherlands, in two different areas, rural and urban. We use a sub-sample of the data publicly available in Fitzmaurice et al. (2010). The data contains the measurements of the forced expiratory volume in one second (FEV1) at 133 subjects residents in the rural area. The individuals chosen were those greater than 36 years old at their entry to the study and whose smoking status did not change over time. The smoking status was divided into a current or former smoker. A subject that smokes at least one cigarette by day is classified as current, and otherwise as a former smoker. Each individual was measured between 1 to 7 times, they were measured every three years for up 21 years. Figure **3-6** shows the profiles of the 133 subjects for the forced expiratory volume in one second (FEV1).

We fitted model (3-3) to the FEV1 data using the same setup of Section 3.5 for the hyperparameters. We ran $10,000$ Monte Carlo iterations. The posterior probability for the alternative hypothesis was approximately 0.606, which indicates that there are differences in the forced expiratory volume (FEV1) between current smokers and former smokers. We also obtained the classical tests like as Wald Test and the F-test from the fitted of a linear mixed model with random intercept, in both cases the $p-$values (0.00453) suggest that exist differences in the FEV1. Thus, we can conclude that the condition of current smoker reduces pulmonary capacity.

Figure **3-7** and **3-8** show the estimations of the density for the forced expiratory volume in one second (FEV1) by time points. The red dashed line corresponds to posterior mean of the FEV1 for the current smoker group and the green dashed line for the former smoker group. In the figure the gray regions represents the 95% credibility intervals. Figure **3-8** also include the Shift function estimated by time.

In general, the shift function tendency is not so far from zero, which indicates that there are no marked differences between the FEV1 of a former and current smoker. The observed differences are smaller in both tails of the distributions. Furthermore, the magnitude of

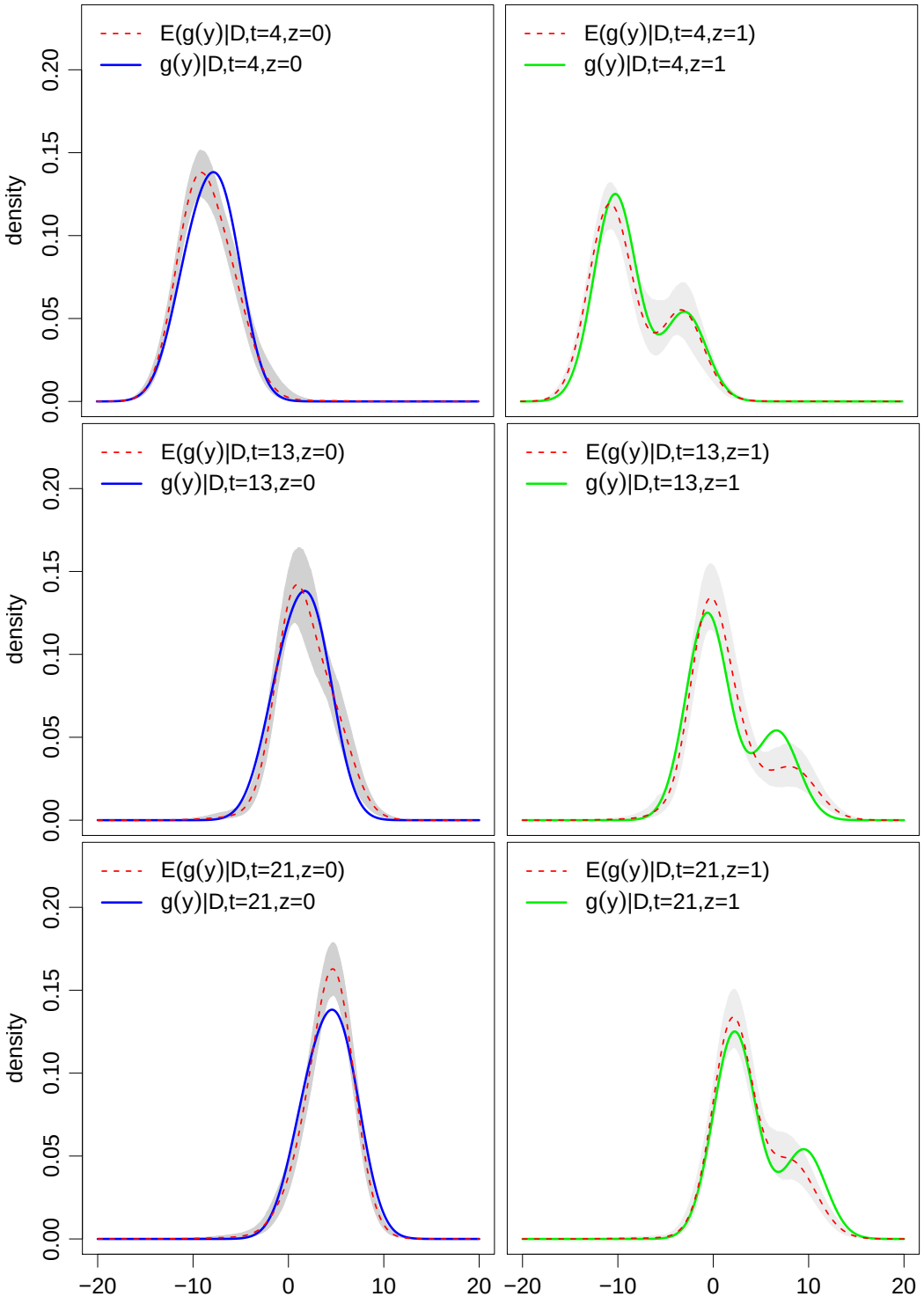


Figure **3-4**: True and estimated densities at the time points selected in Figure **3-3**. $g(\mathsf{y} \mid D, t, z = 0)$ denotes the density for the control group and $g(\mathsf{y} \mid D, t, z = 1)$ for the treatment group. The dashed red line is the DDP estimated and the grey regions represent the point-wise 95% credible intervals. The green and blue solid line represent the true density for the treatment and control group, respectively.

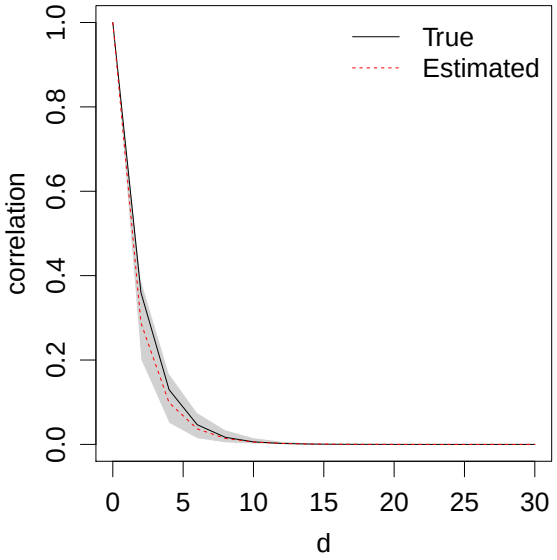
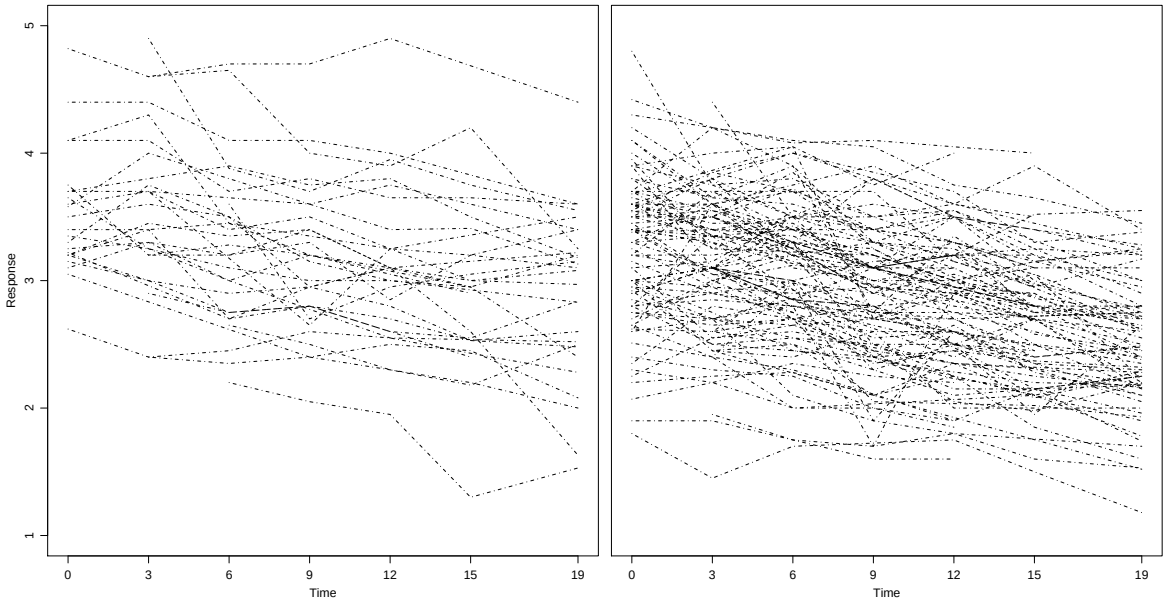


Figure **3-5**: Plot of the correlation. The dashed red line is the estimated correlation, the grey region represents the point-wise 95% credible intervals.



(a) (b)

Figure **3-6**: Profile plot of the forced expiratory volume in one second (FEV1). The observations within a subject are connected by a line. (a) Profiles of former smoker group, (b) Profiles of current smoker group.

the differences is decreasing over time. This last result could not be discovered using traditional methods.

3.7 Concluding remarks

Longitudinal data structures are common in different scientific fields as Biology, Agriculture, and Medicine. A key element in this kind of data is that the measurements are taken repetitively over time. Consequently, the data are physically and stochastically dependent through time. The reference model for this type of data is the mixed model, which, through the inclusion of specific-subject parameters allows the modeling of the inter-individual differences and induces correlation between the repeated measurements. Under this approach, the inference about the possible effect of the predictors on the response variable is based on the marginal model. The T-test, together with F-test, are frequently used to identify the effect of the predictors based on the fixed effects on the marginal model. Both tests are based on the Gaussian assumption on the errors and random effects of the model. Additionally, the degrees of freedom employed in both test are based on approximations (Satterthwaite 1941, Kenward & Roger 1997).

As in mixed models, our proposal is also based on a hierarchical specification. In such specification, the data are conditionally independent, given a collection of random measures. In a second hierarchy, a Markov process relates the measures through the stick-breaking construction. The second hierarchy is defined conditionally on a hypothesis, and the model is completed with a prior distribution on the hypotheses space, which penalizes more complex models. Our construction induces correlation on the observations as described in Proposition 2. When the sticks follow a Beta distribution with parameters 1 and M, we have that, marginally, our model is a mixture of Dirichlet processes. The correlation induced on the observations is a function of the parameters M, α, and the distance between the observations d. Then, when the $d = 0$ the correlation is 1 and when d goes to infinity, the correlation goes to zero. Thus, our proposal goes beyond the exchangeable case usually employed in mixture models.

The flexibility in our BNP specification makes our model reasonable in a variety of scenarios, including nonlinear behavior of the response across the time and different correlation structures. Such flexibility allows us to test the effect of the predictors even when the data does not follow a Gaussian distribution. The results with simulated data showed that our model was flexible enough to capture nonlinear tendencies. The procedure estimated well the correlation structure of the data and was able to detect the effect of the predictors. Because the procedure captures differences across the entire distribution, in the application to real data, we were capable of identifying the magnitude of the differences in the distribution between the current and former smokers for the forced expiratory volume in different moments of the time. In both, the illustration with simulated data and the application to real data, the BNP procedure was competitive compared to the standard tests used in mixed models. We are running extensive Monte Carlo simulations scenarios to evaluate the performance of our proposal, especially concerning the capacity to detect

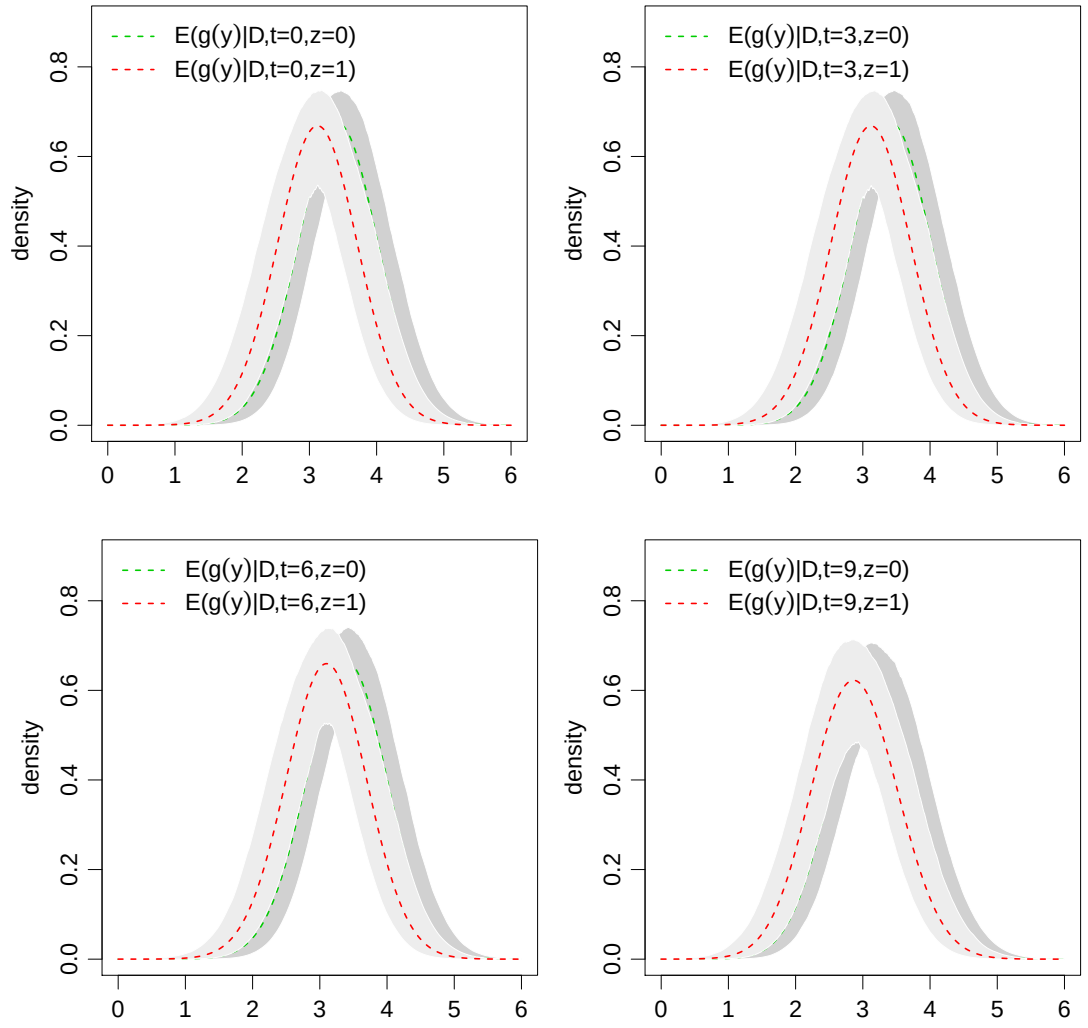


Figure **3-7**: Estimated densities of the FEV1 by time points. $E\left(g(\mathsf{y} \mid D, t, z=0)\right)$ denotes the posterior estimated density at the time t for the former smoker group and $E\left(g(\mathsf{y} \mid D, t, z=1)\right)$ for the current smoker group. The dashed red line is the DDP estimated and the grey regions represent the point-wise 95% credible intervals.

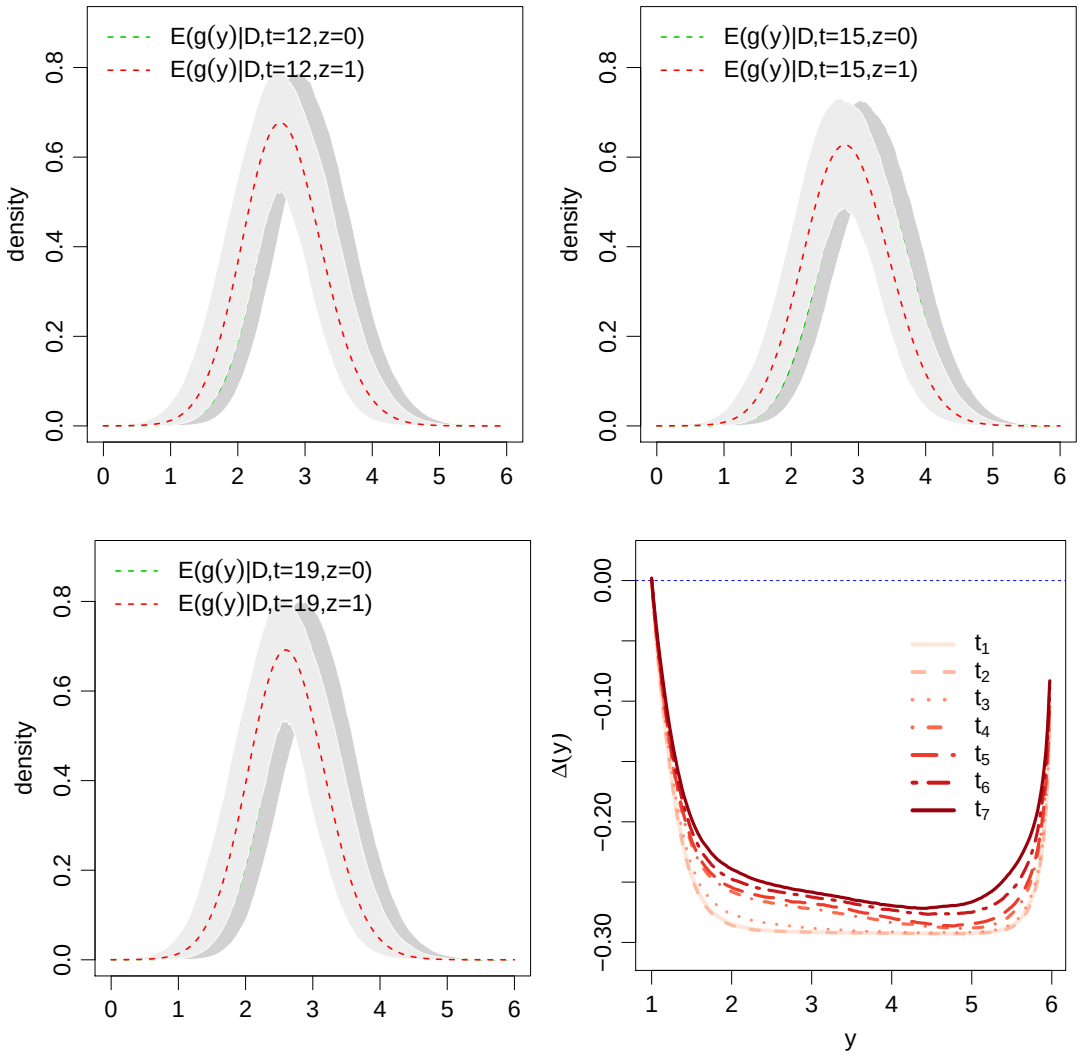


Figure **3-8**: Top panel and left bottom panel show the estimated densities of the FEV1 at the last three time points. $E\left(g(\mathsf{y} \mid D, t, z = 0)\right)$ denotes the posterior estimated density at the time t for the former smoker group and $E\left(g(\mathsf{y} \mid D, t, z = 1)\right)$ for the current smoker group. The dashed red line is the DDP estimated and the grey regions represent the point-wise 95% credible intervals. The right bottom panel presents the shift function by time, the color intensity represents the increment in the time, thus the dark red corresponds to differences in the density between current smoker and the former smoker by the last measure over the time, while the lightest red is associated to the differences in over first measure.

the effect of the predictors without restrictive distributional assumptions.

3.8 Appendix A

Proof of Proposition 2

Let $\mathbb{E}[\mathsf{v}_n] = \mathbb{E}[\mathsf{v}_{n+1}] := \mu$, $\mathbb{E}[\mathsf{v}_n^2] := \mu^{(2)}$ and $\mathbb{V}(\mathsf{v}_n) = \mathbb{V}(\mathsf{v}_{n+1}) := \sigma$, then

$$\mathsf{Cov}(\mathsf{v}_n, \mathsf{v}_{n+1}) = \mathbb{E}[\mathsf{v}_n \mathsf{v}_{n+1}] - \mathbb{E}[\mathsf{v}_n]\mathbb{E}[\mathsf{v}_{n+1}]$$

$$\begin{aligned}
\mathbb{E}[\mathsf{v}_n \mathsf{v}_{n+1}] &= \mathbb{E}[\mathbb{E}[\mathsf{v}_n \mathsf{v}_{n+1} \mid \mathsf{v}_n]] \\
&= \mathbb{E}[\mathsf{v}_n \mathbb{E}[\mathsf{v}_{n+1} \mid \mathsf{v}_n]] \\
&= \mathbb{E}[\mathsf{v}_n[(1-\alpha^d)\mu + \alpha^d \mathsf{v}_n]] \\
&= (1-\alpha^d)\mu\mathbb{E}[\mathsf{v}_n] + \alpha^d \mathbb{E}[\mathsf{v}_n^2] \\
&= (1-\alpha^d)\mu^2 + \alpha^d \mu^{(2)} \\
&= \alpha^d \sigma + \mu^2,
\end{aligned}$$

and

$$\begin{aligned}
\mathsf{Cov}(\mathsf{v}_n, \mathsf{v}_{n+1}) &= \alpha^d \sigma + \mu^2 - \mu^2 \\
&= \alpha^d \sigma
\end{aligned}$$

thus

$$\mathsf{Corr}(\mathsf{v}_n, \mathsf{v}_{n+1}) = \alpha^d.$$

Proof of Proposition 3

With the purpose of simplifies the notation, let $t' = t + d$, then, the covariance between two random measures Y_{it} and $Y_{it'}$ is given by

$$\begin{aligned}
\mathsf{Cov}[Y_{it}, Y_{it'}] &= \mathbb{E}\left[\mathsf{Cov}\left(Y_{it}, Y_{it'} \mid t_j, \mathsf{x}_{it_j}, \mathcal{P}\right)\right] + \mathsf{Cov}\left[\mathbb{E}\left(Y_{it} \mid t_j, \mathsf{x}_{it_j}, \mathcal{P}\right), \mathbb{E}\left(Y_{it'} \mid t_j, \mathsf{x}_{it_j}, \mathcal{P}\right)\right] \\
&= \mathbb{E}\left[0\right] + \mathsf{Cov}\left[\mathbb{E}\left(Y_{it} \mid t_j, \mathsf{x}_{it_j}, \mathcal{P}\right), \mathbb{E}\left(Y_{it'} \mid t_j, \mathsf{x}_{it_j}, \mathcal{P}\right)\right] \\
&= \mathbb{E}\left[\sum_{l=1}^{\infty} \mathsf{w}_{lt}\mu_l(\mathsf{x}) \sum_{l=1}^{\infty} \mathsf{w}_{lt'}\mu_l(\mathsf{x})\right] - \mathbb{E}\left[\sum_{l=1}^{\infty} \mathsf{w}_{lt}\mu_l(\mathsf{x})\right] \mathbb{E}\left[\sum_{l=1}^{\infty} \mathsf{w}_{lt'}\mu_l(\mathsf{x})\right],
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}\left[\sum_{l=1}^{\infty} \mathsf{w}_{lt}\mu_l(\mathsf{x})\right] &= \sum_{l=1}^{\infty} \mu(\mathsf{x})\,\mathbb{E}[\mathsf{w}_{lt}] \\
&= \sum_{l=1}^{\infty} \mu(\mathsf{x})\,\mathbb{E}\left[\mathsf{v}_{lt}\prod_{j=1}^{l-1}(1-\mathsf{v}_{jt})\right] \\
&= \sum_{l=1}^{\infty} \mu(\mathsf{x})\,\mathbb{E}\left[\mathsf{v}_{lt}\right]\prod_{j=1}^{l-1}\mathbb{E}(1-\mathsf{v}_{jt}) \\
&= \sum_{l=1}^{\infty} \mu(\mathsf{x})\,\mu_{\mathsf{v}}(1-\mu_{\mathsf{v}})^{l-1} \\
&= \sum_{l=0}^{\infty} \mu(\mathsf{x})\,\mu_{\mathsf{v}}(1-\mu_{\mathsf{v}})^{l} \\
&= \mu(\mathsf{x}).
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[\sum_{l=1}^{\infty} \mathsf{w}_{lt}\mu_l(\mathsf{x})\sum_{l=1}^{\infty} \mathsf{w}_{lt'}\mu_l(\mathsf{x})\right] &= \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\mathbb{E}[\mathsf{w}_{kt}\mathsf{w}_{lt'}\mu_k(\mathsf{x})\mu_l(\mathsf{x})] \\
&= \sum_{k=1}^{\infty}\mathbb{E}[\mathsf{w}_{kt}\mathsf{w}_{kt'}]\mathbb{E}[\mu_k(\mathsf{x})\mu_k(\mathsf{x})] + \sum_{k=1}^{\infty}\sum_{l=k+1}^{\infty}\mathbb{E}[\mathsf{w}_{kt}\mathsf{w}_{lt'}]\mathbb{E}[\mu_k(\mathsf{x})\mu_l(\mathsf{x})] \\
&+ \sum_{l=1}^{\infty}\sum_{k=l+1}^{\infty}\mathbb{E}[\mathsf{w}_{kt}\mathsf{w}_{lt'}]\mathbb{E}[\mu_k(\mathsf{x})\mu_l(\mathsf{x})].
\end{aligned}$$

For $l = k$, we have

$$\begin{aligned}
\mathbb{E}[\mathsf{w}_{kt}\mathsf{w}_{kt'}] &= \mathbb{E}\left[\mathsf{v}_{kt}\prod_{j=1}^{k-1}(1-\mathsf{v}_{jt})\mathsf{v}_{kt'}\prod_{j=1}^{k-1}(1-\mathsf{v}_{jt'})\right] \\
&= \mathbb{E}\left[\mathsf{v}_{kt}\mathsf{v}_{kt'}\prod_{j=1}^{k-1}(1-\mathsf{v}_{jt})(1-\mathsf{v}_{jt'})\right] \\
&= \mathbb{E}\left[\mathsf{v}_{kt}\mathsf{v}_{kt'}\prod_{j=1}^{k-1}(1-\mathsf{v}_{jt'}-\mathsf{v}_{jt}+\mathsf{v}_{jt}\mathsf{v}_{jt'})\right] \\
&= \mathbb{E}\left[\mathsf{v}_{kt}\mathsf{v}_{kt'}\right]\prod_{j=1}^{k-1}\mathbb{E}\left[1-\mathsf{v}_{jt'}-\mathsf{v}_{jt}+\mathsf{v}_{jt}\mathsf{v}_{jt'}\right] \\
&= \varphi_{tt'}\prod_{j=1}^{k-1}(1-2\mu_{\mathsf{v}}+\varphi_{tt'}) \\
&= \varphi_{tt'}(1-2\mu_{\mathsf{v}}+\varphi_{tt'})^{k-1}
\end{aligned}$$

where $\varphi_{tt'} := \mathbb{E}\left[\mathsf{v}_{kt}\mathsf{v}_{kt'}\right]$ and $\mu_{\mathsf{v}} := \mathbb{E}[\mathsf{v}_{kt}] = \mathbb{E}[\mathsf{v}_{kt'}]$. On the other hand, we have that

$$\begin{aligned}
\mathbb{E}[\mu_k(\mathsf{x})\mu_l(\mathsf{x})] &= \mathbb{V}\left[\mu_k(\mathsf{x})\right] + \mathbb{E}^2\left(\mu_k(\mathsf{x})\right) \\
&= \sigma_0^2(\mathsf{x}) + \mu^2(\mathsf{x}).
\end{aligned}$$

For $l > k$, we have

$$\begin{aligned}
\mathbb{E}[\mathsf{w}_{kt}\mathsf{w}_{lt'}] &= \mathbb{E}\left[\mathsf{v}_{kt}\mathsf{v}_{lt'}(1-\mathsf{v}_{kt'})\prod_{j=1}^{k-1}(1-\mathsf{v}_{jt})(1-\mathsf{v}_{jt'})\prod_{s=k+1}^{l-1}(1-\mathsf{v}_{st'})\right] \\
&= \mathbb{E}\left[(\mathsf{v}_{kt}\mathsf{v}_{lt'}-\mathsf{v}_{kt}\mathsf{v}_{lt}\mathsf{v}_{kt'})\prod_{j=1}^{k-1}(1-\mathsf{v}_{jt'}-\mathsf{v}_{jt}+\mathsf{v}_{jt}\mathsf{v}_{jt'})\prod_{s=k+1}^{l-1}(1-\mathsf{v}_{st'})\right] \\
&= (\mathbb{E}[\mathsf{v}_{lt'}]\mathbb{E}[\mathsf{v}_{kt}]-\mathbb{E}[\mathsf{v}_{lt}]\mathbb{E}[\mathsf{v}_{kt}\mathsf{v}_{kt'}])\prod_{j=1}^{k-1}\mathbb{E}[1-\mathsf{v}_{jt'}-\mathsf{v}_{jt}+\mathsf{v}_{jt}\mathsf{v}_{jt'}]\prod_{s=k+1}^{l-1}\mathbb{E}[1-\mathsf{v}_{st'}] \\
&= \mu_{\mathsf{v}}(\mu_{\mathsf{v}}-\varphi_{tt'})(1-2\mu_{\mathsf{v}}+\varphi_{tt'})^{k-1}(1-\mu_{\mathsf{v}})^{l-k-1}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\mu_k(\mathsf{x})\mu_l(\mathsf{x})] &= \mathbb{E}[\mu_k(\mathsf{x})]\mathbb{E}[\mu_l(\mathsf{x})] \\
&= \mu^2(\mathsf{x})
\end{aligned}$$

Finally for $l < k$ we obtain

$$\begin{aligned}
\mathbb{E}[\mathsf{w}_{kt}\mathsf{w}_{lt'}] &= \mathbb{E}\left[\mathsf{v}_{kt}\mathsf{v}_{lt'}(1-\mathsf{v}_{kt'})\prod_{j=1}^{l-1}(1-\mathsf{v}_{jt})(1-\mathsf{v}_{jt'})\prod_{s=l+1}^{k-1}(1-\mathsf{v}_{st'})\right] \\
&= \mathbb{E}\left[(\mathsf{v}_{kt}\mathsf{v}_{lt'}-\mathsf{v}_{kt}\mathsf{v}_{lt}\mathsf{v}_{kt'})\prod_{j=1}^{l-1}(1-\mathsf{v}_{jt'}-\mathsf{v}_{jt}+\mathsf{v}_{jt}\mathsf{v}_{jt'})\prod_{s=l+1}^{k-1}(1-\mathsf{v}_{st'})\right] \\
&= (\mathbb{E}[\mathsf{v}_{lt'}]\mathbb{E}[\mathsf{v}_{kt}]-\mathbb{E}[\mathsf{v}_{lt}]\mathbb{E}[\mathsf{v}_{kt}\mathsf{v}_{kt'}])\prod_{j=1}^{l-1}\mathbb{E}[1-\mathsf{v}_{jt'}-\mathsf{v}_{jt}+\mathsf{v}_{jt}\mathsf{v}_{jt'}]\prod_{s=l+1}^{k-1}\mathbb{E}[1-\mathsf{v}_{st'}] \\
&= \mu_{\mathsf{v}}(\mu_{\mathsf{v}}-\varphi_{tt'})(1-2\mu_{\mathsf{v}}+\varphi_{tt'})^{l-1}(1-\mu_{\mathsf{v}})^{k-l-1}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\mu_k(\mathsf{x})\mu_l(\mathsf{x})] &= \mathbb{E}[\mu_k(\mathsf{x})]\mathbb{E}[\mu_l(\mathsf{x})] \\
&= \mu^2(\mathsf{x})
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{E}\left[\sum_{l=1}^{\infty}\mathsf{w}_{lt}\mu_l(\mathsf{x})\sum_{l=1}^{\infty}\mathsf{w}_{lt'}\mu_l(\mathsf{x})\right] &= \sum_{k=1}^{\infty}\varphi_{tt'}(1-2\mu_{\mathsf{v}}+\varphi_{tt'})^{k-1}\left[\sigma_0^2(\mathsf{x})+\mu^2(\mathsf{x})\right] \\
&+ \sum_{k=1}^{\infty}\sum_{l=k+1}^{\infty}\mu^2(\mathsf{x})\mu_{\mathsf{v}}(\mu_{\mathsf{v}}-\varphi_{tt'})(1-2\mu_{\mathsf{v}}+\varphi_{tt'})^{k-1}(1-\mu_{\mathsf{v}})^{l-k-1} \\
&+ \sum_{l=1}^{\infty}\sum_{k=l+1}^{\infty}\mu^2(\mathsf{x})\mu_{\mathsf{v}}(\mu_{\mathsf{v}}-\varphi_{tt'})(1-2\mu_{\mathsf{v}}+\varphi_{tt'})^{l-1}(1-\mu_{\mathsf{v}})^{k-l-1} \\
&= \left[\sigma_0^2(\mathsf{x})+\mu^2(\mathsf{x})\right]\varphi_{tt'}\sum_{k=1}^{\infty}(1-2\mu_{\mathsf{v}}+\varphi_{tt'})^{k-1} \\
&+ \mu^2(\mathsf{x})\mu_{\mathsf{v}}(\mu_{\mathsf{v}}-\varphi_{tt'})\sum_{k=1}^{\infty}(1-2\mu_{\mathsf{v}}+\varphi_{tt'})^{k-1}\sum_{l=k+1}^{\infty}(1-\mu_{\mathsf{v}})^{l-k-1}+ \\
&+ \mu^2(\mathsf{x})\mu_{\mathsf{v}}(\mu_{\mathsf{v}}-\varphi_{tt'})\sum_{l=1}^{\infty}(1-2\mu_{\mathsf{v}}+\varphi_{tt'})^{l-1}\sum_{k=l+1}^{\infty}(1-\mu_{\mathsf{v}})^{k-l-1} \\
&= \frac{\left[\sigma_0^2(\mathsf{x})+\mu^2(\mathsf{x})\right]\varphi_{tt'}}{2\mu_{\mathsf{v}}-\varphi_{tt'}}+2\frac{\mu^2(\mathsf{x})(\mu_{\mathsf{v}}-\varphi_{tt'})}{2\mu_{\mathsf{v}}-\varphi_{tt'}}
\end{aligned}$$

thus, we have

$$\begin{aligned}
\mathsf{Cov}[Y_{it},Y_{it'}] &= \mathbb{E}\left[\sum_{l=1}^{\infty}\mathsf{w}_{lt}\mu_l(\mathsf{x})\sum_{l=1}^{\infty}\mathsf{w}_{lt'}\mu_l(\mathsf{x})\right]-\mathbb{E}\left[\sum_{l=1}^{\infty}\mathsf{w}_{lt}\mu_l(\mathsf{x})\right]\mathbb{E}\left[\sum_{l=1}^{\infty}\mathsf{w}_{lt'}\mu_l(\mathsf{x})\right] \\
&= \frac{\left[\sigma_0^2(\mathsf{x})+\mu^2(\mathsf{x})\right]\varphi_{tt'}}{2\mu_{\mathsf{v}}-\varphi_{tt'}}+2\frac{\mu^2(\mathsf{x})(\mu_{\mathsf{v}}-\varphi_{tt'})}{2\mu_{\mathsf{v}}-\varphi_{tt'}}-\mu^2(\mathsf{x}) \\
&= \frac{\sigma_0^2(\mathsf{x})\varphi_{tt'}}{2\mu_{\mathsf{v}}-\varphi_{tt'}}.
\end{aligned}$$

On the other hand, the variance of Y_{it} is given by

$$\begin{aligned}
\mathsf{Var}\left[Y_{it}\right] &= \mathsf{Cov}[Y_{it},Y_{it}] \\
&= \frac{\sigma_0^2(\mathsf{x})\varphi_{tt}}{2\mu_{\mathsf{v}}-\varphi_{tt}}
\end{aligned}$$

where $\varphi_{tt}=\mathbb{E}[\mathsf{v}_{kt}^2]:=\mu_{\mathsf{v}}^{(2)}$. Then,

$$\mathsf{Corr}[Y_{it},Y_{it'}] = \frac{\varphi_{tt'}(2\mu_{\mathsf{v}}-\varphi_{tt})}{\varphi_{tt}(2\mu_{\mathsf{v}}-\varphi_{tt'})}$$

If we consider the jump process for v_{kt} like in (3-9), then the correlation is given by

$$\begin{aligned}
\mathsf{Corr}[Y_{it},Y_{it'}] &= \frac{\varphi_{tt'}(2\mu_{\mathsf{v}}-\varphi_{tt})}{\varphi_{tt}(2\mu_{\mathsf{v}}-\varphi_{tt'})} \\
&= \frac{(2\mu_{\mathsf{v}}-\mu_{\mathsf{v}}^{(2)})\left(\alpha^d\sigma_{\mathsf{v}}+\mu_{\mathsf{v}}^2\right)}{\mu_{\mathsf{v}}^{(2)}(2\mu_{\mathsf{v}}-\alpha^d\sigma_v-\mu_{\mathsf{v}}^2)},
\end{aligned}$$

where $\mathsf{Var}[\mathsf{v}]=\sigma_{\mathsf{v}}$.

3.9 Appendix B

Gibbs sampling for the posterior inference

In this section we provide details on the Gibbs sampling for updating the parameters $\mathsf{v}_l(t_j)$, α_{v}, $\boldsymbol{\gamma}$ and M of the Algorithm 1.

1. Updating of the weights processes.

   We have that for $j \neq \{1, \tau\}$, where τ is the last point at the time with observations, the posterior distribution is given by

$$\begin{aligned} \mathcal{L}(\mathsf{v}_l(t_j) \mid \cdots) &\propto \mathsf{p}_{\mathsf{v}}(\mathsf{v}_l(t_{j+1}) \mid \mathsf{v}_l(t_j))\, \mathsf{p}_{\mathsf{v}}(\mathsf{v}_l(t_j) \mid \mathsf{v}_l(t_{j-1})) \\ &\times \mathsf{v}_l(t_j)^{n_{ij}} (1 - \mathsf{v}_l(t_j))^{\kappa_{ij}}, \end{aligned} \tag{3-20}$$

   where $n_{ij} = \sum_{i=1}^{m} \mathbb{1}\{\mathsf{s}_{ij} = l\}$ and $\kappa_{ij} = \sum_{i=1}^{m} \mathbb{1}(\mathsf{s}_{ij} > l)$. On the other hand, for $j = 1$ it is given by

$$\begin{aligned} \mathcal{L}(\mathsf{v}_l(t_1) \mid \cdots) &\propto \mathsf{p}_{\mathsf{v}}(\mathsf{v}_l(t_2) \mid \mathsf{v}_l(t_1))\, \mathsf{p}_{\mathsf{v}}(\mathsf{v}_l(t_1)) \\ &\times \mathsf{v}_l(t_1)^{n_{ij}} (1 - \mathsf{v}_l(t_1))^{\kappa_{ij}} \end{aligned} \tag{3-21}$$

   and finally for $j = \tau$ we have

$$\begin{aligned} \mathcal{L}(\mathsf{v}_l(t_\tau) \mid \cdots) &\propto \mathsf{p}_{\mathsf{v}}(\mathsf{v}_l(t_\tau) \mid \mathsf{v}_l(t_{\tau-1})) \\ &\times \mathsf{v}_l(t_\tau)^{n_{ij}} (1 - \mathsf{v}_l(t_\tau))^{\kappa_{ij}}. \end{aligned} \tag{3-22}$$

   In particular, if $\pi_{\mathsf{v}} = \mathsf{Be}(1, M)$ then (3-20) is given by

$$\begin{aligned} \mathcal{L}(\mathsf{v}_l(t_j) \mid \cdots) \propto\ & q_{0j}\mathrm{Be}(\mathsf{v}_l(t_j) \mid 1 + n_{ij}, M + \kappa_{ij}) + q_{1j}\mathbb{1}_{\{\mathsf{v}_l(t_j) = \mathsf{v}_l(t_{j-1})\}} + \\ & q_{2j}\mathbb{1}_{\{\mathsf{v}_l(t_j) = \mathsf{v}_l(t_{j+1})\}} + q_{3j}\mathbb{1}_{\{\mathsf{v}_l(t_j) = \mathsf{v}_l(t_{j+1}) = \mathsf{v}_l(t_{j-1})\}}, \end{aligned}$$

   where

$$\begin{aligned} q_{0j} &= (1 - \alpha_{\mathsf{v}}^d)^2 M \frac{\Gamma(1 + n_{ij})\Gamma(M + \kappa_{ij})}{\Gamma(1 + M + n_{ij} + \kappa_{ij})} \mathrm{Be}(\mathsf{v}_l(t_{j+1}) \mid 1, M), \\ q_{1j} &= \alpha_{\mathsf{v}}^d (1 - \alpha_{\mathsf{v}}^d) \mathsf{v}_l(t_{j-1})^{n_{ij}} (1 - \mathsf{v}_l(t_{j-1}))^{\kappa_{ij}} \mathrm{Be}(\mathsf{v}_l(t_{j+1}) \mid 1, M), \\ q_{2j} &= \alpha_{\mathsf{v}}^d (1 - \alpha_{\mathsf{v}}^d) M \frac{\Gamma(1 + n_{ij})\Gamma(M + \kappa_{ij})}{\Gamma(1 + M + n_{ij} + \kappa_{ij})} \mathrm{Be}(\mathsf{v}_l(t_{j+1}) \mid 1 + n_{ij}, M + \kappa_{ij}), \\ q_{3j} &= \alpha_{\mathsf{v}}^{2d} \mathsf{v}_l(t_{j-1})^{n_{ij}} (1 - \mathsf{v}_l(t_{j-1}))^{\kappa_{ij}}. \end{aligned}$$

   In the same way, the expression in (3-21) is given by

$$\mathcal{L}(\mathsf{v}_l(t_1) \mid \cdots) \propto q_{01}\mathrm{Be}(\mathsf{v}_l(t_1) \mid 1 + n_{i1}, M + \kappa_{i1}) + q_{02}\mathbb{1}_{\{\mathsf{v}_l(t_1) = \mathsf{v}_l(t_2)\}},$$

   where

$$\begin{aligned} q_{01} &= (1 - \alpha_{\mathsf{v}}^d) M \mathrm{Be}(\mathsf{v}_l(t_2) \mid 1, M), \\ q_{02} &= \alpha_{\mathsf{v}}^d \mathrm{Be}(\mathsf{v}_l(t_2) \mid 1 + n_{i1}, M + \kappa_{i1}). \end{aligned}$$

and finally (3-22) by

$$\mathcal{L}(\mathsf{v}_l(t_\tau) \mid \cdots) \quad \propto \quad q_{0\tau}\mathrm{Be}(\mathsf{v}_l(t_\tau) \mid 1+n_{i\tau}, M+\kappa_{i\tau}) + q_{1\tau}\mathbb{1}_{\{\mathsf{v}_l(t_\tau)=\mathsf{v}_l(t_{\tau-1})\}},$$

where

$$\begin{aligned} q_{0\tau} &= (1-\alpha_{\mathsf{v}}^d)M\frac{\Gamma(1+n_{i\tau})\Gamma(M+\kappa_{i\tau})}{\Gamma(1+M+n_{i\tau}+\kappa_{i\tau})}, \\ q_{1\tau} &= \alpha_{\mathsf{v}}^d \mathsf{v}_l(t_{\tau-1})^{n_{i\tau}}(1-\mathsf{v}_l(t_{\tau-1}))^{\kappa_{i\tau}}. \end{aligned}$$

2. Updating α_{v}.

   The posterior distribution for the correlation parameter between the sticks is given by

$$\mathcal{L}(\alpha_{\mathsf{v}} \mid \cdots) \propto \mathrm{Be}\left(a_2, b_2\right) \times \prod_{l=1}^{N}\{\mathcal{L}\left(\mathsf{v}_l\left(t_1\right) \mid \cdots\right) \times \prod_{j=2}^{T-1}\mathcal{L}\left(\mathsf{v}_l\left(t_j\right) \mid \cdots\right) \times \mathcal{L}\left(\mathsf{v}_l\left(t_T\right) \mid \cdots\right)\}$$

   and because this expression does not have a closed form, then a Metropolis-Hasting step is needed. We propose to use a truncated normal distribution as a proposal distribution for α_{v}, thus at iteration i, a candidate α_{v}^* is to propose from $\mathrm{N}\left(\alpha_{\mathsf{v}}^* \mid \alpha_{\mathsf{v}}^{(\mathsf{i}-1)}, c\right)\mathbb{1}_{[0,1]}$. Then, $\alpha_{\mathsf{v}} = \alpha_{\mathsf{v}}^*$ with probability min(r,1), where r is defined as

$$r = \frac{\mathcal{L}(\alpha_{\mathsf{v}}^* \mid \cdots)\mathrm{N}\left(\alpha_{\mathsf{v}}^{(\mathsf{i}-1)} \mid \alpha_{\mathsf{v}}^*, c\right)\mathbb{1}_{[0,1]}}{\mathcal{L}(\alpha_{\mathsf{v}}^{(\mathsf{i}-1)} \mid \cdots)\mathrm{N}\left(\alpha_{\mathsf{v}}^* \mid \alpha_{\mathsf{v}}^{(\mathsf{i}-1)}, c\right)\mathbb{1}_{[0,1]}}$$

3. Updating of vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_j)$.

   We use a Metropolis-Hastings step with probabilities of acceptation proportional to

$$\begin{aligned} P\left(\boldsymbol{\gamma} = (0,0,\ldots,0) \mid \ldots\right) &\propto \pi_{\mathcal{M}}\left(H_{(0,0,\ldots,0)}\right)\prod_{l=1}^{N}\prod_{j=1}^{p}\mathrm{N}\left(\beta_{jl} \mid 0, \nu_0\psi_j\right), \\ P\left(\boldsymbol{\gamma} = (1,0,\ldots,0) \mid \ldots\right) &\propto \pi_{\mathcal{M}}\left(H_{(1,0,\ldots,0)}\right)\prod_{l=1}^{N}\mathrm{N}\left(\beta_{1l} \mid 0, \psi_1\right)\prod_{j=2}^{p}\mathrm{N}\left(\beta_{jl} \mid 0, \nu_0\psi_j\right), \\ &\vdots \\ P\left(\boldsymbol{\gamma} = (1,1,\ldots,1) \mid \ldots\right) &\propto \pi_{\mathcal{M}}\left(H_{(1,1,\ldots,1)}\right)\prod_{l=1}^{N}\prod_{j=1}^{p}\mathrm{N}\left(\beta_{jl} \mid 0, \psi_j\right). \end{aligned}$$

   The proposed candidate γ^* for the MH step is a vector that is different of $\gamma^{(\mathsf{i}-1)}$ in one element only. For instance, if we have four covariables and at $\mathsf{i}-1$ step, $\gamma^{(\mathsf{i}-1)} = (0,1,0,1)$, then the candidate is sampled of the set

$$\{(1,1,0,1), (0,0,0,1), (0,1,1,1), (0,1,0,0)\},$$

   with acceptation probability $\phi' = {}^{P(\gamma^*|\ldots)}/_{P(\gamma^*|\ldots)+P(\gamma^{(\mathsf{i}-1)}|\ldots)}$, thus at iteration i if

$\mathsf{u} \leq \phi'$, then $\gamma^{(\mathsf{i})} = \gamma^*$ else $\gamma^{(\mathsf{i})} = \gamma^{(\mathsf{i}-1)}$, here $\mathsf{u} \sim \mathsf{U}(0,1)$.

4. The mass parameter M is updated as in Escobar & West (1995).
   Thus, M is sampled from a gamma mixture given by

$$\pi(M \mid N, \eta) = \phi \mathrm{Ga}\left(M \mid a_1 + N, b_1 - \log(\eta)\right) + (1-\phi)\mathrm{Ga}\left(M \mid a_1 + N - 1, b_1 - \log(\eta)\right)$$

   where for all $N > 1$, $\phi = 1/\left(1 + \frac{n(b_1 - \log(\eta))}{(a_1 + N - 1)}\right)$, n is the sample size, and η is a continuous random variable such that $\pi\left(\eta \mid M, N\right) \sim \mathrm{Be}\left(M+1, n\right)$.

Chapter 4

Concluding remarks and future directions

In this thesis, we have presented two novel procedures to perform hypothesis testing in data structures, which considers correlation. In Chapter 2, we proposed a procedure to compare the marginal distributions of paired samples, which is an essential problem in statistics. The proposal follows a Bayesian Nonparametric approach of inference. Thus, it was able to detect differences across the entire distribution. The hypothesis testing procedure showed results consistently good in the simulated scenarios, and was competitive compared to the traditional tests, even in situations where the assumptions for the conventional test were accomplished.

An interesting problem related to the results in Chapter 2 is to identify whether two samples are correlated or not (see, e.g. Filippi et al. 2016, Filippi & Holmes 2017). This can be developed exploiting the hierarchical structure and the special parametrization of the kernel to identify if the correlation is zero or not, which can be tested, for example, defining a spike-slab prior distribution for the parameter τ^2.

In Chapter 3, we proposed a procedure for testing the effect of predictors on the response variable in the context of longitudinal data analysis. The procedure was able to capture the correlation among the observations from the same individual, and at the same time, it was capable of detecting the effect of the predictors. In the illustration with real and simulated data, our method showed excellent performance for detecting the effects of predictors on the response variable, and the time evolution of the density was appropriately captured.

The problem modeled in Chapter 3 has received a lot of attention in the literature. There are many variations of statistical models for longitudinal data. Our proposal is flexible, but it just considers the case of monotone missing-data patterns. Specifically, we have assumed that an individual is measured from the first follow-up time and if a *dropout* takes place, then the individual is retired of the study definitively. Additionally, we suppose an *ignorable* mechanism to the missing. In fact, we have assumed a mechanism of missing completely at random (MCAR). The data are said to be missing completely at

random when the probability that responses are missing not dependent to either specific values. Bayesian nonparametric approaches when missingness is monotone can be found in Wang et al. (2010), Daniels & Linero (2015), Linero (2015). However, in longitudinal studies, given the dynamic to collect the observations, in practice is frequent the presence of non-ignorable and non-monotone missingness. Under non-monotone missing patterns, for instance, either a non-response can occur starting the follow-up, or in the middle of the follow-up period, just to mention a few situations among other more complex.

An engaging issue arises from Chapter 3 is the development of a hypothesis testing procedure in the presence of non-ignorable and non-monotone missing data in longitudinal studies. Thus, non-ignorable missing data implies to include a hierarchy for the missing data in the model, that could allow incorporating the missing information in the joint probability of the observed data. However, the treatment of non-monotone missing data is not necessarily straightforward. Its modeling exhibit essential challenges for the prior on the completed data space, and there are few works available in the Bayesian literature (see, e.g., Linero (2017)).

Bibliography

Airoldi, E., Blei, D., Erosheva, E. & Fienberg, S. (2019), *Handbook of Mixed Membership Models and Their Applications*, Chapman and Hall/CRC.

Al-Labadi, L. & Zarepour, M. (2017), 'Two-sample Kolmogorov-Smirnov test using a Bayesian nonparametric approach', *Mathematical Methods of Statistics* **26**(3), 212–225.

Azzalini, A. (1994), 'Logistic regression for autocorrelated data with application to repeated measures', *Biometrika* **81**(4), 767–775.

Benavoli, A., Corani, G., Mangili, F., Zaffalon, M. & Ruggeri, F. (2014), A Bayesian wilcoxon signed-rank test based on the dirichlet process, *in* 'International Conference on Machine Learning (ICML)', pp. 1026–1034.

Berger, J. (1985), *Statistical decision theory and Bayesian analysis; 2nd ed.*, Springer Series in Statistics, Springer, New York.

Berger, J. O. & Pericchi, L. R. (1996), 'The intrinsic Bayes factor for model selection and prediction', *Journal of the American Statistical Association* **91**(433), 109–122.

Berger, J., Pericchi, L. & Ghosh, J. (2001), Objective Bayesian methods for model selection: introduction and comparison, *in* 'In Model selection, volume 38 of IMS Lecture Notes Monogr.', Inst. Math. Statist., pp. 137–207.

Bhattacharya, A. & Dunson, D. (2012), 'Nonparametric Bayes classification and hypothesis testing on manifolds', *Journal of multivariate analysis* **111**, 1–19.

Borgwardt, K. & Ghahramani, Z. (2009), 'Bayesian two-sample tests'.
URL: *http://arxiv.org/abs/0906.4032*

Breslow, N. E. & Clayton, D. G. (1993), 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association* **88**(421), 9–25.

Carroll, R. J., Fan, J., Gijbels, I. & Wand, M. P. (1997), 'Generalized partially linear single-index models', *Journal of the American Statistical Association.* **92**(438), 477–489.

Carroll, R. J., Hall, P., Apanasovich, T. V. & Lin, X. (2004), 'Histospline method in nonparametric regression models with application to longitudinal/clustered data', *Statistical Sinica* **14**(3), 649–674.

Castro, L., Wang, W., Lachos, V., de Carvalho, V. & Bayes, C. (2018), 'Bayesian semiparametric modeling for HIV longitudinal data with censoring and skewness', *Statistical Methods in Medical Research* **28**(5), 1–20.

Chen, K. & Jin, Z. (2005), 'Local polynomial regression analysis for clustered data', *Biometrika* **92**(1), 59–74.

Chen, S. & Zhong, P. (2010), 'ANOVA for longitudinal data with missing values', *The Annals of Statistics* **38**(6), 3630–3659.

Chen, Y. & Hanson, T. E. (2014), 'Bayesian nonparametric k-sample tests for censored and uncensored data', *Computational Statistics & Data Analysis* **71**, 335–346.

Chipman, H. (1996), 'Bayesian variable selection with related predictors', *The Canadian Journal of Statistics* **24**(1), 17–36.

Chipman, H., George, E. & McCulloch, R. (2001), 'The practical implementation of Bayesian model selection', *Lecture Notes- ...* **38**.

Cipolli III, W., Hanson, T. & McLain, A. (2016), 'Bayesian nonparametric multiple testing', *Computational Statistics & Data Analysis* **101**, 64–79.

Crainiceanu, C. & Ruppert, D. (2004), 'Restricted likelihood ratio tests in nonparametric longitudinal models', *Statistica Sinica* **14**(3), 713–729.

Dahl, D. & Newton, M. (2007), 'Multiple hypothesis testing by clustering treatment effects', *Journal of the American Statistical Association* **102**(478), 517–526.

Dahlin, J., Robert, K. & Thomas, B. (2016), Bayesian inference for mixed effects models with heterogeneity, Technical report, Linköpings universitet, Sweden.

Daniels, M. J. & Linero, A. R. (2015), *Bayesian Nonparametrics for Missing Data in Longitudinal Clinical Trials*, Springer International Publishing, Cham, pp. 423–446.

de Finetti, B. (1931), *Funzione Caratteristica Di un Fenomeno Aleatorio*, 6. Memorie, Academia Nazionale del Linceo, pp. 251–299.

de Finetti, B. (1937), 'La Prévision: Ses Lois Logiques, Ses Sources Subjectives', *Annales de l'Institut Henri Poincaré* **17**, 1–68.

Diaconis, P. (1977), 'Finite forms of de Finetti's theorem on exchangeability', *Synthese* **36**, 271–281.

Diggle, P. J., Heagerty, P. J.and Liang, K. Y. & Zeger, S. L. (2002), *Analysis of Longitudinal Data*, Oxford: Oxford University Press.

Dockery, D., Berkey, C., Ware, J., Speizer, F. & Ferris, B. (1983), 'Distribution of FVC and FEV1 in children 6 to 11 years old', *American Review of Respiratory Disease* **128**(3), 405–412.

Doksum, K. (1974), 'Empirical probability plots and statistical inference for nonlinear models in the two-sample case', *The Annals of Statistics* **2**(2), 267–277.

Doksum, K. & Sievers, G. (1976), 'Plotting with confidence: Graphical comparison of two populations', *Biometrika* **63**(3), 421–434.

Durret, R. (2010), *Probability. Theory and examples. Fourth Edition.*, Cambridge Series in Statistical and Probabilistic Mathematics.

Eisenhart, C. (1947), 'The assumptions underlying the analysis of variance', *Biometrics* **3**(1), 1–21.

Escobar, M. (1988), Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Unpublished Ph.D. dissertation, Department of Statistics. Yale University.

Escobar, M. (1994), 'Estimating normal means with a Dirichlet process prior', *Journal of the American Statistical Association* **89**, 268–277.

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

Fahrmeir, L., Kneib, T., Lang, S. & Marx, B. (2013), *Regression. Models, methods and applications*, Springer.

Ferguson, T. (1973), 'A Bayesian analysis of some nonparametric problems', *The Annals of Statistics* **1**, 209–230.

Ferguson, T. (1983), 'Bayesian density estimation by mixtures of normal distribution', *Recent Advances in Statistics* pp. 287–302.

Ferguson, T. S. (1974), 'Prior distribution on the spaces of probability measures', *Annals of Statistics* **2**(4), 615–629.

Filippi, S. & Holmes, C. (2017), 'A Bayesian nonparametric approach to testing for dependence between random variables', *Bayesian Analysis* **12**(4), 919–938.

Filippi, S., Holmes, C. & Nieto-Barajas, L. (2016), 'Scalable Bayesian nonparametric measures for exploring pairwise dependence via Dirichlet Process Mixtures', *Electronic Journal of Statistics* **10**(2), 3338–3354.

Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenbergs, G. (2009), *Handbooks of modern statistical methods. Longitudinal data analysis*, Chapman & Hall/CRC.

Fitzmaurice, G., Laird, N. & Ware, J. (2010), *Applied Longitudinal Analysis*, Wiley Series in Probability and Statistics.

Gelfand, A. E. & Kottas, A. (2001), 'Nonparametric Bayesian modeling for stochastic order', *Annals of the Institute of Statistical Mathematics* **53**, 865–876.

George, E. (2000), 'The variable selection problem', *Journal of the American Statistical Association* **95**(452), 1304–1308.

George, E. I. & McCulloch, R. E. (1993), 'Variable selection via Gibbs sampling', *Journal of the American Statistical Association* **88**(423), 881–889.

George, E. I. & McCulloch, R. E. (1997), 'Approaches for Bayesian variable selection', *Statistica Sinica* **7**, 339–373.

Geweke, J. (1996), Variable selection and model comparison in regresssion, *in* 'In Bayesian Statistics 5', University Press, pp. 609–620.

Ghosh, J., Delampady, M. & Samanta, T. (2007), *An Introduction to Bayesian Analysis: Theory and Methods*, Springer.

Ghosh, J. K. & Ramamoorthi, R. V. (2003), *Bayesian nonparametrics*, Springer, New York, USA.

Girón, F., Martínez, M. & Moreno, E. (2003), 'Bayesian analysis of matched pairs', *Journal of Statistical Planning and Inference* **113**(1), 49–66.

Gopalan, R. & Berry, D. (1998), 'Bayesian multiple comparisons using Dirichlet Process priors', *Journal of the American Statistical Association* **93**(1), 1130–1139.

Griepentrog, G., Ryan, J. & Smith, L. (1982), 'Linear transformations of polynomial regression models', *The American Statistician* **36**(3a), 171–174.

Griffin, J. E. & Steel, M. F. J. (2006), 'Order-based dependent Dirichlet processes', *Journal of the American Statistical Association* **101**, 179–194.

Gutiérrez, L., Barrientos, A. F., González, J. & Taylor-Rodríguez, D. (2019), 'A Bayesian nonparametric multiple testing procedure for comparing several treatments against a control', *Bayesian Analysis* **14**(2), 649–675.

Gutiérrez, L., Mena, R. H. & Ruggiero, M. (2016), 'A time dependent bayesian nonparametric model for air quality analysis', *Computational Statistics and Data Analysis* **95**, 161–175.

Heckman, N. (1986), 'Spline smoothing in partial linear models', *Journal of the American Statistical Association* **48**(2), 244–248.

Henderson, C. R. (1953), 'Estimation of variance and covariance components', *Biometrics* **9**(2), 226–252.

Hennig, C., Meila, M., Murtagh, F. & Rocci, R. (2015), *Handbook of Cluster Analysis*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

Hewitt, E. & Savage, L. J. (1955), 'Symmetric measures on cartesian products', *Transactions of the American Mathematical Society* **80**(2), 470–501.

Hjort, N. L., Holmes, C., Müller, P. & Walker, S. (2010), *Bayesian nonparametrics*, Cambridge University Press, Cambridge, UK.

Hogan, J. W., Lin, X. & Herman, B. (2004), 'Mixtures of varying coefficient models for longitudinal data with discrete or continuous nonignorable dropout', *Biometrics* **60**(4), 854–864.

Hollander, M. & Korwar, R. (1980), *Nonparametric Bayesian Estimation of the Horizontal Distance Between Two Populations*, Defense Technical Information Center.

Holmes, C., Caron, F., Griffin, J. & Stephens, D. (2015), 'Two-sample bayesian nonparametric hypothesis testing', *Bayesian Analysis* **10**(2), 297–320.

Huang, J.and Wu, C. & Zhou, L. (2002), 'Varying-coefficient models and basis function approximation for the analysis of repeated measures', *Biometrika* **89**(4), 111–128.

Huang, L. & Ghosh, M. (2014), 'Two-sample hypothesis testing under Lehmann alternatives and Pólya tree priors', *Statistica Sinica* **24**(4), 1717–1733.

Huang, Y., Hu, X. & Dagne, G. (2014), 'Jointly modeling time-to-event and longitudinal data: a bayesian approach', *Stat Methods Appl* **23**, 95–121.

Ishwaran, H. & James, L. (2001), 'Gibbs sampling methods for stick-breaking priors', *Journal of the American Statistical Association* **96**, 161–173.

Ishwaran, H. & Rao, J. S. (2000), Bayesian nonparametric mcmc for large variable selection problems. unpublished manuscript.

Ishwaran, H. & Rao, J. S. (2003), 'Detecting differentially expressed genes in microarrays using bayesian model selection', *Journal of the American Statistical Association* **98**(462), 438–455.

Ishwaran, H. & Rao, J. S. (2005), 'Spike and slab variable selection: Frequentist and Bayesian strategies', *The Annals of Statistics* **33**(2), 730–773.

Jennrich, R. L. & Schluchter, M. D. (1986), 'Unbalanced repeated-measures models with structured covariance matrices', *Biometrics* **42**(4), 805–820.

Kalli, M., Griffin, J. E. & Walker, S. (2011), 'Slice sampling mixture models', *Statistics and Computing* **21**, 93–105.

Kenward, M. G. & Roger, J. H. (1997), 'Small sample inference for fixed effects from restricted maximum likelihood', *Biometrics* **53**(3), 983–997.

Kim, S., Dahl, D. B. & Vannucci, M. (2009), 'Spiked Dirichlet Process prior for Bayesian multiple hypothesis testing in random effects models', *Bayesian Analysis* **4**(4), 707–732.

Kleinman, K. & Ibrahim, J. (1998), 'A semi-parametric Bayesian approach to generalized linear mixed models', *Statistics in Medicine* **17**, 2579–2596.

Kliethermes, S. (2013), A Bayesian nonparametric approach to modeling longitudinal growth curves with non-normal outcomes, PhD thesis, University of Iowa, https://ir.uiowa.edu/etd/2546. https://doi.org/10.17077/etd.qfku6z5f.

Kuo, L. & Mallick, B. (1998), 'Variable selection for regression models', *Sankhyā: The Indian Journal of Statistics, Series B* **60**(1), 65–81.

Laird, N. & James, W. (1982), 'Random-effects models for longitudinal data', *Biometrics* **38**(4), 963–974.

Li, Y., Lin, X. & Müller, P. (2010), 'Bayesian inference in semiparametric mixed models for longitudinal data', *Biometrics* **66**, 70–78.

Liang, K. Y. & Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**(1), 13–22.

Linero, A. (2017), 'Bayesian nonparametric analysis of longitudinal studies in the presence of informative missingness', *Biometrika* **104**(2), 327–341.

Linero, A. R. (2015), Nonparametric Bayes: Inference Under Nonignorable Missingness, PhD thesis, University of Florida.

Liu, X. (2015), *Methods and Applications of Longitudinal Data Analysis*, Academic Press.

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

Lopes, H., Müller, P. & Rosner, G. (2003), 'Bayesian meta-analysis for longitudinal data models using multivariate mixture priors', *Biometrics* **59**, 66–75.

Lu, H. H. S., Wells, M. T. & Tiwari, R. C. (1994), 'Inference for shift functions in the two-sample problem with right-censored data: With applications', *Journal of the American Statistical Association* **89**(427), 1017–1026.

Luke, S. (2017), 'Evaluating significance in linear mixed-effects models in R', *Behavior Research Methods* **49**(4), 1494–1502.

Ma, L. & Wong, W. H. (2011), 'Coupling optional Pólya trees and the two sample problem', *Journal of the American Statistical Association* **106**(496), 1553–1565.

MacEachern, S. (1994), 'Estimating normal means with a conjugate style dirichlet process prior', *Communications in Statistics - Simulation and Computation* **23**, 727–741.

MacEachern, S. N. (1999), Dependent nonparametric processes, *in* 'ASA Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA', American Statistical Association.

MacEachern, S. N. (2000), Dependent dirichlet processes, Technical report, Department of Statistics, The Ohio State University.

MacEachern, S. N. & Müller, P. (1998), 'Estimating mixture of Dirichlet process models', *Journal of Computational and Graphical Statistics* **7**, 223–238.

Malsiner-Walli, E., Wagner, H. & Kepler, J. (2011), 'Comparing Spike and Slab priors for Bayesian variable selection', *Austrian Journal of Statistics* **40**(4), 241–262.

McCullagh, P. & Nelder, J. (1989), *Generalized linear models*, (2nd ed.) London:Chapman & H all.

Mena, R. (2015), Estimación de densidades dinámica vía el proceso de Dirichlet difuso, *in* 'Encuentro nacional de jóvenes Investigadores en matemáticas (ENJIM)'.

Mena, R. H., Ruggiero, M. & Walker, S. G. (2011), 'Geometric stick-breaking processes for continuous-time Bayesian nonparametric modeling', *Journal of Statistical Planning and Inference* **141**, 3217–3230.

Mitchell, T. & Beauchamp, J. (1988), 'Bayesian variable selection in linear regression', *Journal of the American Statistical Association* **83**(104), 1023–1032.

Morgan, W. A. (1939), 'A test for the significance of the difference between two variances in a sample from a normal bivariate population', *Biometrika* **31**(1/2), 13–19.

Muñoz, A., Carey, V., Schouten, J. P., Segal, M. & Rosner, B. (1992), 'Parametric family of correlation structures for the analysis of longitudinal data.', *Biometrics* **48**(3), 733–742.

Müller, P. & Mitra, R. (2013), 'Bayesian nonparametric inference–why and how', *Bayesian Analysis* **8**(2), 269–302.

Müller, P. & Quintana, F. A. (2004), 'Nonparametric Bayesian data analysis', *Statistical Science* **19**(1), 95–110.

Müller, P., Quintana, F., Jara, A. & Hanson, T. (2015), *Bayesian nonparametric data analysis*, Springer International Publishing.

Müller, P., Quintana, F., Rosner, G. & Maitland, M. (2014), 'Bayesian inference for longitudinal data with non-parametric treatment effects', *Biostatistics* **15**(2), 341–352.

Müller, P., Rosner, G., De lorio, M. & MacEachern, S. (2005), 'A Nonparametric Bayesian model for inference in related longitudinal studies', *Journal of the Royal Statistical Society. Series C, Applied Statistics* **54**, 611–626.

Neal, R. (2000), 'Markov chain sampling methods for dirichlet process mixture models', *Journal of Computational and Graphical Statistics* **9**, 249–265.

Nelder, J. (1998), 'The selection of terms in response-surface models - how strong is the weak-heredity principle?', *American Statistician* **52**(4), 315–318.

Nelder, J. (2000), 'Functional marginality and response-surface fitting', *Journal of Applied Statistics* **27**(1), 109–112.

Nuñez Antón, V. & Zimmerman, D. (2000), 'Modelling nonstationary longitudinal data', *Biometrics* **56**(3), 699–705.

Orbanz, P. (2013), Lecture notes on bayesian nonparametrics, Class notes for a phd level course on bayesian nonparametrics, Columbia University.

Papaspiliopoulos, O. & Roberts, G. (2008), 'Retrospective markov chain monte carlo methods for dirichlet process hierarchical models', *Biometrika* **95**, 169–186.

Peixoto, J. (1987), 'Hierarchical variable selection in polynomial regression models', *The American Statistician* **41**(4), 311–313.

Peixoto, J. (1990), 'A property of well-formulated polynomial regression models', *The American Statistician* **44**(1), 26–30.

Pitman, E. J. G. (1939), 'A note on normal correlation', *Biometrika* **31**(1/2), 9–12.

Quintana, F., Johnson, W., Waetjen, L. & Gold, E. (2016), 'Bayesian nonparametric longitudinal data analysis', *Journal of the American Statistical Association* **515**(115), 1168–1181.

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Ročková, V. & George, E. I. (2016), 'The Spike-and-Slab LASSO', *Journal of the American Statistical Association* .
URL: *https://doi.org/10.1080/01621459.2016.1260469*

Rodríguez, A. & Dunson, D. B. (2011), 'Nonparametric Bayesian models through probit stick-breaking processes', *Bayesian Analysis* **6**(1), 145–177.

Rodríguez, A. & Müller, P. (2013), 'Nonparametric Bayesian inference', *NSF-CBMS Regional Conference Series in Probability and Statistics* **9**, 1–110.

Satterthwaite, F. E. (1941), 'Synthesis of variance', *Psychometrika* **6**(5), 309–316.

Savitsky, T. & Paddock, S. (2014), 'Bayesian semi- and non-parametric models for longitudinal data with multiple membership effects in R', *Journal of Statistical Software* **57**(3), 1–35.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Scott, J. (2009), 'Nonparametric Bayesian multiple testing for longitudinal performance stratification', *The Annals of Applied Statistics* **3**(4), 1655–1674.

Scott, J. & Berger, J. (2010), 'Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem', *Annals of Statistics* **38**(5), 2587–2619.

Sethuraman, J. (1994), 'A constructive definition of Dirichlet priors', *Statistica Sinica* **4**, 639–650.

Shang, K. (2016), An Approach to Nonparametric Bayesian Analysis for High Dimensional Longitudinal Data Sets, PhD thesis, University of Minnesota.

Shang, K. & Reilly, C. (2017), 'Non parametric Bayesian analysis of the two-sample problem with censoring', *Communications in Statistics-Theory and Methods* **46**(15), 12008–12022.

Silverman, B. (1984), 'Spline smoothing: the equivalent variable kernel method', *Annals of Statistics* **12**(3), 898–916.

Silverman, B. (1998), 'Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion)', *Journal of the American Statistical Association* **93**(443), 961–976.

Smith, M. & Kohn, R. (1996), 'Nonparametric regression using Bayesian variable selection', *Journal of Econometrics* **75**(2), 241–262.

Soriano, J. & Ma, L. (2017), 'Probabilistic multi-resolution scanning for two-sample differences', *Journal of the Royal Statistical Society, Series B* **79**(2), 547–572.

Speckman, P. (1988), 'Kernel smoothing in partial linear models', *Journal of the Royal Statistical Society, Series B.* **50**(3), 413–436.

Stone, C., Hansen, M., Kooperberg, C. & Truong, Y. K. (1997), 'Polynomial splines and their tensor products in extended linear modeling (with discussion)', *Annals of Statistics* **25**(4), 1371–1470.

Student (1908), 'The probable error of a mean', *Biometrika* **6**(1), 1–25.

Taylor-Rodríguez, D., Womack, A. & Bliznyuk, N. (2016), 'Bayesian variable selection on model spaces constrained by heredity conditions', *Journal of Computational and Graphical Statistics* **25**(2), 515–535.

Verbeke, G. & Molenberghs, G. (2009), *Linear Mixed Models for Longitudinal Data*, Springer Series in Statistics.

Verbyla, A. P., Cullis, B. R., Kenward, M. G. & Welham, S. J. (1999), 'The analysis of designed experiments and longitudinal data using smoothing splines', *Journal of the Royal Statistical Society, Series B.* **48**(1), 269–311.

Wagner, H. & Malsiner-Walli, G. (2011), 'Comparing spike and slab priors for bayesian variable selection', *Austrian Journal of Statistics* **40**(4), 241–264.

Walker, S. (2007), 'Sampling the dirichlet mixture model with slices', *Communications in Statistics - Simulation and Computation* **36**, 45–54.

Wang, C., Daniels, M. J., Scharfstein, D. O. & Land, S. (2010), 'A Bayesian shrinkage model for incomplete longitudinal binary data with application to the breast cancer prevention trial', *Journal of the American Statistical Association* **105**(492), 1333–1346.

Wang, Y. & Daniels, M. (2013), 'Bayesian modeling of the dependence in longitudinal data via partial autocorrelations and marginal variances', *Journal of Multivariate Analysis* **116**, 130–140.

Wang, Y.-G. & Hin, L.-Y. (2010), 'Modeling strategies in longitudinal data analysis: Covariate, variance function and correlation structure selection', *Computational Statistics & Data Analysis* **54**(12), 3359–3370.

Weiss, R. (2005), *Modeling Longitudinal Data*, Springer Texts in Statistics.

Wells, M. T. & Tiwari, R. C. (1989), 'Bayesian quantile plots and statistical inference for nonlinear models in the two sample case with incomplete data', *Communications in Statistics - Theory and Methods* **18**(8), 2955–2964.

Welsh, A. H., Lin, X. & Carroll, R. J. (2002), 'Marginal longitudinal nonparametric regression: Locality and efficiency of spline and kernel methods', *Journal of the American Statistical Association* **97**(458), 482–493.

Wilcoxon, F. (1945), 'Individual comparisons by ranking methods', *Biometrics Bulletin* **1**(6), 80–83.

Wilson, M. A., Iversen, E. S., Clyde, M. A., Schmidler, S. C. & Schildkraut, J. (2010), 'Bayesian model search and multilevel inference for SNP association studies', *The annals of applied statistics* **4**(3), 1342–1364.

Zeger, S. L. & Diggle, P. J. (1994), 'Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters', *Biometrics* **50**(3), 689–699.

Zhang, D., Lin, X.and Raz, J. & Sowers, M. (1998), 'Semiparametric stochastic mixed models for longitudinal data', *Journal of the American Statistical Association* **93**(442), 710–719.

Zimmerman, D. (1997), 'Teacher's corner: A note on interpretation of the paired-samples t test', *Journal of Educational and Behavioral Statistics* **22**(3), 349–360.

Zimmerman, D. (2000), 'Viewing the correlation structure of longitudinal data through a PRISM', *The American Statistician* **54**(4), 310–318.