UNIVERSITY COLLEGE LONDON

DOCTORAL THESIS

Covariate Dependent Random Measures with Applications in Biostatistics

---

Author:
William BARCELLA

Supervisor:
Prof. Maria DE IORIO

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Department of Statistical Science
Faculty of Mathematical and Physical Sciences

April, 2017

Declaration of Authorship

I, William BARCELLA, declare that this thesis titled "Covariate Dependent Random Measures with Applications in Biostatistics" and the work presented in it are my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:

Date:

"A few observation and much reasoning lead to error; many observations and a little reasoning to truth."

Alexis Carrel

Abstract

Covariate Dependent Random Measures with Applications in Biostatistics

by William BARCELLA

In Bayesian nonparametrics, the specification of suitable (for practical purposes) stochastic processes whose realisations are discrete probability measures plays a crucial role. Recently, real world applications have motivated the extension of these stochastic processes to incorporate covariate information in the realisations with the aim of constructing infinite mixture models having weights and/or component-specific parameters which depend on covariates. This work presents four different modelling strategies motivated by practical problems involving stochastic processes over covariate dependent random measures. After presenting the main concepts in Bayesian nonparametrics and reviewing relevant literature, we develop two Bayesian models which are extensions of augmented response mixture models. In particular, we construct a semi-parametric non-linear regression model for zero-inflated discrete distributions and propose techniques to perform variable selection in cluster-specific regression models. The third contribution presents a generalisation of Dirichlet Process for random probability measures to include covariate information via Beta regression. Properties of this new stochastic process are discussed and two illustrations are presented for dealing with spatially correlated observations and grouped longitudinal data. The last part of the thesis proposes a modelling strategy for time-evolving correlated binary vectors, which relies on latent variables. The distribution of these latent variables is assumed to be a convolution of Gaussian kernels with covariate dependent random probability measures. These four modelling strategies are motivated by datasets that come from medical studies involving lower urinary tract symptoms and acute lymphoblastic leukaemia as well as from publicly available data about primary schools evaluations in London.

Acknowledgements

Many people have contributed to this PhD thesis and have supported me during these years.

First and foremost, I would like to express my gratitude to my PhD advisor Maria De Iorio for her continuous support and for pushing me toward ambitious goals. I am very thankful for all the time she has dedicated to guide my research and for her sincere suggestions on my work and on my decisions. I am also thankful to my co-advisor Gianluca Baio: he has been of great help in many difficult moments. It has been a real pleasure to work with them.

I would like to thank my collaborators: James Malone-Lee, Stefano Favaro, Gary Rosner. Their inputs and questions have greatly shaped this thesis. Their passion for scientific research has strongly affected my approach toward problems.

These years at UCL have been special also because of my PhD-mates at the Statistical Science department, especially Peter Kenny, James Pitkin, Rodrigo Targino, Samuel Linvingstone and Elefteria Kotti. I am thankful to them for the many interesting discussions about statistical problems, career and, most importantly, for their friendship.

Special thanks go to Giacomo Zanella and Simon Lunagomez, for being great statisticians and even better friends. I owe them my gratitude for inspiring my curiosity and for their support and help.

During these years I have met many friends which have made London to be *home*. I am immensely grateful for this to Pietro, Barbara, Giulia, Lorenzo, Marco, Michela, Marco and Tommaso.

I am thankful to my family: Gianluigi, Milena, Roberto, Daniele. They have always supported me, even without understanding my choices, with unconditional love and enthusiasm.

Finally, I would like to express my deepest gratitude to Beatrice for she has always reminded me that my value is and will always be infinitely more than all my achievements and all mistakes. Her love and consolation have been the only light in the hardest moments. This work is dedicated to her.

Contents

List of Figures

List of Tables

List of Abbreviations

ALL	**A**cute **L**ymphoblastic **L**eukaemia
BMU	**B**lackwell-**M**acQueen **U**rn
BNP	**B**ayesian **N**on-**P**arametrics
BNP-ZIP	**B**ayesian **N**on-**P**arametric **Z**ero-**I**nflated **P**oisson
CART	**C**lassification **A**nd **R**egression **T**rees
CRM	**C**ompletely **R**andom **M**easure
CRP	**C**hinese **R**estaurant **P**rocess
DDP	**D**ependent **D**irichlet **P**rocess
DGDP	**D**ependent **G**eneralised **D**irichlet **P**rocess
DP-GLM	**D**irichlet **P**rocess (mixture of) **G**eneralised **L**inear **M**odels
DP	**D**irichlet **P**rocess
DPM	**D**irichlet **P**rocess **M**ixture
EDPM	**E**nriched **D**irichlet **P**rocess **M**ixture
GDD	**G**eneralised **D**irichlet **D**istribution
GDP	**G**eneralised **D**irichlet **P**rocess
GLM	**G**eneralised **L**inear **M**odel
GPPM	**G**eneralised **P**roduct **P**artition **M**odel
HME	**H**ierarchical **M**ixture (of) **E**xperts
LR	**L**ow **R**isk (group)
LUTS	**L**ower **U**rinary **T**ract **S**ymptoms
MCMC	**M**arkov **C**hain **M**onte **C**arlo
PPM	**P**roduct **P**artition **M**odel
PPMx	**P**roduct **P**artition **M**odel (with) Covariates
PR	**P**rofile **R**egression
PSBF	**PS**eudo **B**ayes **F**actor
PYP	**P**itman-**Y**or **P**rocess
RDPM	**R**estricted **D**irichlet **P**rocess **M**ixture
RPM	**R**andom **P**artition **M**odel
RPMS	**R**andom **P**artition **M**odel with covariate **S**election
RPMx	**R**andom **P**artition **M**odel (with) Covariates
SAP	**S**chool **A**ction **P**lus
SEN	**S**pecial **E**ducation **N**eed
SHR	**S**tandard/**H**igh **R**isk (group)
SSP	**S**pike (and) **S**lab **P**rior
SUR	**S**eemingly **U**nrelated **R**egression
UTI	**U**rinary **T**ract **I**nfection

WBC	**W**hite **B**lood **C**ell
WMDP	**W**eighted **M**ixture (of) **D**irichlet **P**rocesses
ZINB	**Z**ero-**I**nflated **N**egative **B**inomial
ZIP	**Z**ero-**I**nflated **P**oisson

To Beatrice.

Chapter 1

Bayesian nonparametrics and covariate dependent random measures

Bayesian nonparametrics is a famous and still developing field within Bayesian statistics. Its success is based on the great flexibility and adaptability to a wide variety of practical applications. The focus of this chapter is to introduce the main elements of this methodology and in particular the Dirichlet Process Mixture (DPM) models. The second part of the chapter introduces the objective of this work which is related to the extension of traditional DPM models to include covariate information and to use the resulting models for answering applied questions. After reviewing the main contributions from the literature, we present the outline of the work. Part of the material included in this chapter is based on the work in Barcella et al. [2017].

1.1 Introduction

The objective of this thesis is to discuss new practical developments of stochastic process priors over probability measures that could incorporate information contained in covariates. This field has recently become prominent, both in theory and applications, within the area of *Bayesian nonparametrics* (BNP). In the current section we introduce the main elements of BNP modelling and how covariate dependent random probability measures arise and the role they play, particularly from a practical point of view.

We introduce BNP models starting from briefly describing the main aspects of Bayesian inferential procedures. Consider a set of observations $y_1, \ldots, y_n$

which are realisations of random variables taking values in a measurable space called *sample space*. We assume that they are *independent and identically distributed* (*iid*) according to the probability density (or mass) function $p(y \mid \theta)$, which is indexed by some unknown quantity θ. $p(y \mid \theta)$ is often referred to as *model* (or sampling model).

The main target of statistical inference is to draw conclusions about θ, or functionals of θ. There can be various perspectives to approach this problem. In this work we concentrate exclusively on *Bayesian inference*. In this approach all unknown quantities are treated as random variables. As such, the Bayesian perspective is to look at θ as a random variable, so we write

$$y_1, \ldots, y_n \mid \theta \overset{iid}{\sim} p(y \mid \theta). \tag{1.1.1}$$

In order to perform an inference, Bayesian approach requires to complete the model in (1.1.1) with a *prior distribution*. This is the assumed distribution for θ before running the experiment that generated the observations. We write

$$\theta \sim p(\theta). \tag{1.1.2}$$

The choice of the prior distribution should incorporate all available knowledge on the parameter. This process, usually called *prior elicitation* (or *specification)*, can be driven by a number of procedures, from which we distinguish different areas in Bayesian statistics. A discussion about techniques for prior elicitation is in Chapter 3 in Robert [2007]. The need in Bayesian setting of a prior distribution and its elicitation are the most common criticisms to the Bayesian inferential method (see for a discussion Gelman [2008]). In fact, while all statistical inferential procedures assume a probabilistic model for the observations, only Bayesian inference requires the extra assumption of the parameter's distribution.

In the Bayesian paradigm, inference on θ involves estimating the *posterior distribution* of θ, defined as the distribution of θ given the information contained in the data, *i.e.* $p(\theta \mid y_1, \ldots, y_n)$. This distribution can be obtained by using *Bayes' theorem* as

$$p(\theta \mid y_1, \ldots, y_n) = \frac{p(y_1, \ldots, y_n \mid \theta) p(\theta)}{p(y_1, \ldots, y_n)},$$

where $p(y_1, \ldots, y_n \mid \theta)$ is referred to as *likelihood* or *joint density* of the observations and in light of (1.1.1) is calculated as

$$p(y_1, \ldots, y_n \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta).$$

Instead, $p(y_1, \ldots, y_n)$ is called *prior predictive* or *marginal likelihood* and is calculated as:

$$p(y_1, \ldots, y_n) = \int_{\Theta} p(y_1, \ldots, y_n \mid \theta) p(\theta) \mathrm{d}\theta,$$

where Θ denotes the support of θ (also called parameter space). More generally, a model can be defined in terms of a collection of probability measures over the sample space indexed by some parameter and the procedure for obtaining the posterior distribution described above applies only if the model is dominated[1]. Nevertheless, there might be ways of defining posterior distributions, consistently for all points in the sample space, which do not rely on Bayes' theorem.

Depending on the characteristics of Θ, we distinguish between two types of Bayesian models[2]: *parametric* and *nonparametric* models. In particular, a parametric model is characterised by Θ being (a subset of) a vector space of finite dimension (see Definition 1.1.7 in Robert [2007]). Parametric models are the most frequent in Bayesian literature and they can be suitable in a large number of situations. However, they may be over-restrictive in certain circumstances. In fact, parametric models fix the complexity of the model *a priori*, while sometimes it would be appealing to let the model increase in complexity as new observations become available. In order to achieve the latter feature an elegant strategy is to relax the condition of having the parameter space to be a finite vector space and to consider Θ to be an infinite vector space. Models characterised by such parameter spaces are called nonparametric. Reviews of these models can be found in Ghosh and Ramamoorthi [2003], Müller and Quintana [2004] and Hjort et al. [2010]. We call BNP the area of Bayesian inference dealing with nonparametric models.

[1] Let us define a model as the collection $\mathcal{P}$ of probability measures $P : \mathcal{B} \to [0, 1]$, where $\mathcal{B}$ is the σ-algebra generated by the subsets of the sample space. The model is said to be *dominated* if there exists a σ-finite measure $\mu : \mathcal{B} \to [0, +\infty)$ such that P is absolutely continuous with respect to μ, for all $P \in \mathcal{P}$.

[2] Following the terminology in Robert [2007], we call *Bayesian model* the set of the model (distribution) on the observations and the prior distribution.

The main consequence of having an infinite dimensional parameter space is that we need to specify a prior distribution over an infinite collection of random variables, so prior distributions become *stochastic process priors*. Great efforts have been spent in BNP literature to specify stochastic process priors which could be useful for practical problems and for which *stochastic process posteriors* could be derived analytically (or at least approximated). Based on which stochastic process is employed in the BNP model, we can distinguish between two main areas. The first one uses *Gaussian processes* (Rasmussen and Williams [2006]), while the second one employs *Completely Random Measures* (CRM, Kingman [1967], Lijoi and Prünster [2010]). CRM include a variety of objects, the most famous being the *Gamma process* and *Beta process*. In this work we deal prominently with a process called *Dirichlet Process*, which is obtainable normalising a Gamma process. Interestingly, realisations from a Dirichlet Process are discrete probability measures. A brief review about Dirichlet Process and its properties is presented in this chapter.

In the present work, we focus on methods which additionally index stochastic processes over random probability measures to some covariate space.

1.2 Dirichlet Process

Dirichlet Process (DP) is commonly described as a distribution over distributions, which are defined on a measurable space $(\Theta, \mathcal{A})$, where $\mathcal{A}$ is a σ-algebra of subsets of Θ. This stochastic process was first introduced by Ferguson [1973], who discussed also several of its properties, such as the conjugacy and discreteness of the realisations. Before moving to a formal definition we underline that a realisation from DP is neither a scalar nor a vector or a matrix, but is a probability measure.

We introduce here the concept of partition as it will become necessary for the definition of the DP. The collection of subsets $(S_1, \ldots, S_k)$ represents a partition of the set S if for every j and j' (which are taken to be different) we have that $S_j \cap S_{j'} = \emptyset$ and $\bigcup_{j=1}^{k} S_j = S$. Note that k can be finite or infinite.

Definition 1.2.1. ***(Dirichlet Process).*** *Let us consider a measurable space $(\Theta, \mathcal{A})$, a positive scalar α and a diffuse probability measure G_0 on $(\Theta, \mathcal{A})$. We call DP with parameters α and G_0 the stochastic process whose realisations are random probability*

measures with the following property. Taking G to be a realisation of a DP, we have that for every partition $(S_1, \ldots, S_k)$ of Θ the random vector $(G(S_1), \ldots, G(S_k))$ follows a Dirichlet distribution with parameters $(\alpha G_0(S_1), \ldots, \alpha G_0(S_k))$.

The quantity α is called *precision*, while G_0 is the *centre measure* (also called base measure). A compact way to write is

$$G \sim \text{DP}(\alpha, G_0).$$

Although not required from the original definition of the DP, we consider G_0 to be a diffuse probability measure and we highlight alternative interpretation of G_0 when required.

From Definition 1.2.1, the DP has Dirichlet distributed marginals. In this sense, the DP can be viewed also as the infinite generalisation of the Dirichlet distribution (see Griffiths and Ghahramani [2011]). The definition above does not provide a way to construct the DP, but it states the properties for a process to be a DP. In the same work Ferguson [1973] proved the existence of this object referring to the Kolmogorov's consistency theorem (Kolmogorov [1933]). Another proof for the existence of the DP was provided by Blackwell [1973].

A number of the properties of DP were presented by Ferguson [1973] and we recall here some of them. Let consider $G \sim \text{DP}(\alpha, G_0)$ on $(\Theta, \mathcal{A})$.

Under the topology of pointwise convergence, the support of the DP includes the set of all measures absolutely continuous with respect to G_0. Furthermore, the support of each realisation G is the same as G_0, *i.e.* Θ. If we take the random quantity $G(A)$, where A is a measurable set in $\mathcal{A}$, it follows that

$$\begin{aligned} \mathbb{E}[G(A)] &= G_0(A) \\ \mathbb{V}[G(A)] &= \frac{G_0(A)(1 - G_0(A))}{1 + \alpha}. \end{aligned}$$

From the latter we can notice that α determines the *distance* between G_0 and G.

Ferguson also derived the posterior of the DP. Consider the following hierarchical model

$$\begin{aligned} \theta_1, \ldots, \theta_n \mid G &\overset{iid}{\sim} G \\ G &\sim \text{DP}(\alpha, G_0), \end{aligned}$$

the distribution of $G \mid \theta_1, \ldots, \theta_n$ is still a DP. Indeed, we have that

$$G \mid \theta_1, \ldots, \theta_n \sim \text{DP}\left(\alpha + n, \frac{\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}\right),$$

where δ_{θ_i} is the Dirac measure that places a unitary mass of probability in correspondence of location θ_i. This is often referred to as the conjugacy property of the DP and it can be proved as an extension of the conjugacy property of Dirichlet distribution prior for the parameters of a Multinomial sampling model.

An additional property, which has already been mentioned in this chapter, is the almost sure discreteness of DP samples. This was proved by Ferguson [1973], Blackwell [1973] and Blackwell and MacQueen [1973]. The discreteness of DP samples is the property that makes this process ideal to deal with mixture models and, consequently, with model-based clustering. The idea stems from that if G, distributed as a DP, is discrete, this implies a positive probability of ties in a sequence $\theta_1, \ldots, \theta_n$ sampled from G. This in turn defines a partition of the set $N = \{1, \ldots, n\}$.

The first proof of the discreteness of the DP realisations provided by Ferguson [1973] relies on the representation of the DP as normalised realisations of a Gamma process[3]. The latter are discrete since they can be constructed from realisations of a Poisson Process, which is also discrete (see Kingman [1967] for details). Equivalently, Jordan and Teh [2014] showed that a Gamma Process is a CRM and since all CRM are discrete (Kingman [1967]), then normalising a CRM results in a discrete probability measure.

In next sections of this chapter we will present alternative representations of the DP presented in Blackwell and MacQueen [1973] and in Sethuraman [1994] and we will highlight the links among them and with Definition 1.2.1.

1.2.1 DP and Blackwell-MacQueen urn

A proof for the discreteness of the DP realisations was also provided by Blackwell and MacQueen [1973] and it relies on the representation of the DP samples through the use of a Pólya urn scheme. In particular, the authors extended the

[3]This can be regarded as the extension to processes of a known property of Dirichlet distributed random variables which can be represented normalising a finite set of Gamma random variables (Ferguson [1973]).

classical version of the Pólya urn and they introduced a generalised urn that takes the name Blackwell-MacQueen urn.

A review of Pólya urn schemes is in Mahmoud [2008]. In a traditional Pólya urn we assume to have a certain number of red balls and green balls. The process starts extracting a ball at random from the urn. At this point if the ball is, say, green we replace it into the urn and we add also one more green ball. The same happens for a red ball extracted.

Blackwell and MacQueen modified the scheme above replacing the colour of the balls with values from a probability measure G_0. Then, they constructed the sequence $\theta_1, \theta_2, \ldots$ according the to generalised Pólya urn in Blackwell and MacQueen [1973] (Blackwell-MacQueen Urn, BMU) as follows

$$\begin{aligned} \theta_1 &\sim G_0 \\ \theta_2 \mid \theta_1 &\sim \frac{\alpha}{\alpha+1} G_0 + \frac{1}{\alpha+1}\delta_{\theta_1} \\ &\ldots \\ \theta_n \mid \theta_1, \ldots, \theta_{n-1} &\sim \frac{\alpha}{\alpha+n-1} G_0 + \frac{1}{\alpha+n-1}\sum_{i=1}^{n-1}\delta_{\theta_i}. \end{aligned} \tag{1.2.1}$$

Blackwell and MacQueen [1973] proved three main results that we list below:

- for n that tends to ∞ the probability distribution in (1.2.1) converges almost surely to a discrete probability measure, that we call G;
- G is distributed as a DP(α, G_0);
- considering G from above we have that

$$\theta_1, \ldots, \theta_n \mid G \overset{iid}{\sim} G.$$

In order to understand the implications of the last point we present one of the most important theorem in Bayesian statistics: the *De Finetti's theorem* (de Finetti [1931]). However, before we need to introduce the concept of *exchangeability*.

Definition 1.2.2. ***(Exchangeability).*** *Let us consider a sequence of random elements $\theta_1, \theta_2, \ldots$ taking value on Θ. This is said to be infinitely exchangeable if the joint*

distribution of any finite subsequence is invariant to permutations of the indices. Thus, defining σ to be a permutation, n to be an arbitrary integer and $A_1, \ldots, A_n$ to be subsets of Θ, if $\theta_1, \theta_2, \ldots$ is infinitely exchangeable we have that

$$\Pr[\theta_1 \in A_1, \ldots, \theta_n \in A_n] = \Pr[\theta_{\sigma(1)} \in A_1, \ldots, \theta_{\sigma(n)} \in A_n].$$

An example of infinitely exchangeable sequence is $\theta_1, \theta_2, \ldots$ from (1.2.1). De Finetti's theorem links the concept of *exchangeability* with the concept of *independent and identically distributed*. Indeed, considering the sequence in Definition 1.2.2, De Finetti proved that this sequence is infinitely exchangeable if and only if the joint distribution of any subsequence composed by n elements can be written as

$$\Pr[\theta_1 \in A_1, \ldots, \theta_n \in A_n] = \int \prod_{i=1}^{n} G(A_i) \mathrm{d}P(G), \tag{1.2.2}$$

where G is a random probability measure, that is commonly known as the *directing random measure*, distributed according to P, called *De Finetti's mixing measure*. Expression (1.2.2) states that, conditioning on G, the random elements $\theta_1, \ldots, \theta_n$ are *iid*.

Recalling now the results in Blackwell and MacQueen [1973] if we take G to be the limiting distribution of (1.2.1), then we know that G is the the directing random measure for the sequence constructed using the BMU. In addition, we also know that P, *i.e.* the law of G, is the DP. From (1.2.1), we deduce that the joint distribution of a sequence $\theta_1, \ldots, \theta_n$ is

$$\Pr[\theta_1 \in A_1, \ldots, \theta_n \in A_n] = \prod_{i=1}^{n} \frac{\alpha G_0(A_i) + \sum_{l<i} \delta_{\theta_l}(A_i)}{\alpha + i - 1},$$

where $\delta_{\theta_l}(A_i)$ is equal 1 if θ_l belongs to A_i.

The discreteness of the process in (1.2.1) has two main implications: (i) the sequence $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ reduces to the sequence of its unique values $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_k^*)$, with $k \leq n$, (ii) the vector $\boldsymbol{s} = (s_1, \ldots, s_n)$ with $s_i \in \{1, \ldots, k\}$, which associates each $i \in \{1, \ldots, n\}$ to a component of the vector $\boldsymbol{\theta}^*$, defines a partition of the set $\{1, \ldots, n\}$.

The BMU is connected with another well known process called Chinese Restaurant Process (CRP, Aldous [1985]), which is obtained integrating out $\boldsymbol{\theta}^*$

from the BMU. The CRP takes its name from the following metaphor. Consider a Chinese restaurant with no customers inside. As soon as a customer, $i = 1$, enters into the restaurant a table is prepared by the waiters with probability 1. We denote this table with $j = 1$. If we now consider a second customer, $i = 2$, entering the restaurant, he has the choice of sitting around table $j = 1$ with the customer $i = 1$ or asking for a new table. The probabilities for the two choices are assumed to be proportional to the number of people sitting at $j = 1$, which we denote n_1, and α respectively. If we discover that eventually the second customer asked for a new table, $j = 2$, then the third costumer, $i = 3$, will have the choice of sitting at $j = 1$, $j = 2$ or open a new table with probability proportional to n_1, n_2 and α. Iterating this process generates a partition of the customers based on which table they sit at. Ignoring some steps, when the $(n + 1)$-th customer enters the restaurant he will sit at one of the k (with $k \leq n$) tables already occupied with probabilities proportional to $n_1, \ldots, n_k$ or to a new table with probability proportional to α. More formally we have that

$$\Pr[i = n + 1 \text{ assigned to } j \mid \rho_n] = \begin{cases} \frac{n_j}{\alpha+n}, & \text{for } j = 1, \ldots, k \\ \frac{\alpha}{\alpha+n}, & \text{for } j = k + 1 \end{cases}, \tag{1.2.3}$$

where ρ_n is a partition of $N = \{1, \ldots, n\}$.

From the last equation it is already evident that the probability to be assigned to a specific table does not depend neither on the names of the tables nor on the names of the customers, but only on the number of costumers sitting at each of the tables. Following the terminology in Pitman [1996] we call the resulting partition *exchangeable*. Such partitions imply, for example, that the probability for two customers of sitting together is constant and equal to $\frac{1}{\alpha+1}$ (this is easily computed considering $i = 1$ and $i = 2$).

From (1.2.3), we can compute the general formula of the probability of a generic partition of n items under the CRP as

$$p(\rho_n) = \frac{\alpha^k}{\alpha_{(n)}} \prod_{j=1}^{k} (n_j - 1)! \tag{1.2.4}$$

where $\alpha_{(n)} = \alpha(\alpha + 1) \ldots (\alpha + n - 1)$. This distribution belongs to the class of the Exchangeable Partition Probability Functions (EPPF, see Pitman [1996]).

Comparing (1.2.3) and (1.2.1), we can see that BMU and CRP are similar . In

particular, the difference is that BMU assigns a random value sampled from G_0 to each group of observations belonging to the partition distributed as a CRP. The vector containing all the group-specific values is $\boldsymbol{\theta}^*$ defined above.

We present additional results in terms of the partition implied by sampling from realisations of a DP, using the BMU. These results were discussed by Korwar and Hollander [1973] and Antoniak [1974]. Consider a sequence $\theta_1, \theta_2, \ldots$ from a BMU with parameters α and G_0, then the following properties hold:

- the expected number of ties a priori in the subsequence $\theta_1, \ldots, \theta_n$ is:

$$\mathbb{E}[k] = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1},$$

  which approximates to the quantity $\alpha \log\left(\frac{\alpha+n}{\alpha}\right)$ for large n and diverges for n that tends to ∞;

- the distribution of the number of ties, k, in the sequence $\theta_1, \ldots, \theta_n$ is

$$p(k) = c(n,k) n!\, \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)},$$

  where $c(n,k)$ is a constant obtainable from the recursion for Stirling numbers and $\Gamma(\cdot)$ indicates the Gamma function.

In this section we have introduced a representation of the DP samples that exploits a generalisation of the Pólya urn called BMU. Thanks to this representation we have been able to investigate in details the effects of the discreteness of the DP samples. However, so far we have not presented a constructive definition of the DP which will be presented in the next section.

1.2.2 A constructive definition of the DP

A constructive definition of the DP was presented by Sethuraman [1994]. Let us consider the following random discrete measure

$$G = \sum_{h=1}^{\infty} w_h \delta_{\theta_h}, \tag{1.2.5}$$

where both the locations of the atoms $\theta_1, \theta_2, \ldots$ in Θ and the weights are random and independently generated. In particular, when $\sum_h w_h = 1$ then G is a random probability measure associated with Θ. A possible procedure for generating the weights in (1.2.5) such that they sum up one is called *stick-breaking* (see for a review Ishwaran and James [2001]). In a stick-breaking process, we have

$$\begin{aligned} v_h &\sim \text{Beta}(v_h \mid a_h, b_h), \text{ for } h = 1, 2, \ldots \\ w_1 &= v_1 \\ w_h &= v_h \prod_{r<h} (1 - v_r), \text{ for } h = 2, 3, \ldots . \end{aligned} \tag{1.2.6}$$

In words, this consists in breaking a stick of length equal to one at a random point, saving one part and breaking at random the second part, iteratively. The locations attached to each weight are instead randomly sampled from a diffuse distribution G_0. The specification of the parameters a_h, b_h defines a number of different processes with different characteristics.

Sethuraman [1994] showed that if $a_h = 1$ and $b_h = \alpha$ for all $h = 1, 2, \ldots$ and, independently, $\theta_h \overset{iid}{\sim} G_0$ for all $h = 1, 2, \ldots$, the resulting measure is distributed as a DP(α, G_0). In other words, if G is a realisation of a DP(α, G_0) than G can be constructed by the stick-breaking procedure above with randomly assigned locations from G_0.

An interesting connection can be established between the stick-breaking definition of the DP and the BMU. In fact, it turns out that the probability of belonging to the first cluster characterised by the location θ_1 under (1.2.6), *i.e.* $w_1 \sim \text{Beta}(1, \alpha)$, is the same as the one implied by the BMU. The intuition behind the proof is now presented for w_1 (a more formal proof is presented by Jordan and Teh [2014]). Given the first iteration of the BMU which gives θ_1, the probability that n subsequent observations are assigned to the same θ_1 is equal to $\frac{n!}{(\alpha+1)_{(n)}}$ (which is equivalent to the probability of a partition with one block and cardinality $n + 1$ under the CRP). Let B_i for $i = 2, 3, \ldots$ denote a sequence of binary indicators such that $B_i = 1$ if θ_i takes an existing value in $\theta_1, \ldots, \theta_{i-1}$, whereas $B_i = 0$ if θ_i is a new as yet unobserved value in $\theta_1, \ldots, \theta_{i-1}$. We can rewrite the probability of having the first $n + 1$ observations in the same cluster in terms of B_i. Let us assume that $B_1, \ldots, B_n$ are independent Bernoulli random

variables, given the parameter ξ. We have that:

$$\Pr[B_2 = 1, \ldots, B_{n+1} = 1] = \int \xi^n \mathrm{d}Q(\xi)$$

The right hand side of last equation is the n-th moment of the quantity ξ. Thus, the problem at this point is to find a $Q(\xi)$ which has the n-th moment equal to $\frac{n!}{(\alpha+1)_{(n)}}$. It can be shown that the distribution satisfying the latter is the Beta with parameters 1 and α. An equivalent procedure applies to the probabilities of being assigned to other clusters, conditioning on not being assigned to θ_1.

In Figure 1.1 different samples (left panels) from three different DP's with the relative cumulative distribution functions (right panels) are shown. They all have the same G_0, which we set equal to the standard Normal distribution, but with $\alpha = \{1, 10, 100\}$, respectively. This picture clarifies the meaning of α as well as the role of the centre measure. In fact, G_0 determines the locations of the atoms of (1.2.5), while large values of α lead to weights distributed almost uniformly among the atoms. Differently, small values of α assign most of the probability to few atoms.

1.2.3 Pitman-Yor Process

The expected number of clusters in the partition induced by the DP depends on the precision parameter α and on the number n of samples generated from the DP distributed random measure G and grows approximately as $\log(n)$. This means that the expected number of clusters grows slowly with n. However, real-world applications often show faster rates. So, Pitman and Yor [1997] modified the construction of the DP in order to obtain implied partitions of the observations in which the expected number of clusters grows as a power law.

The resulting process is usually called Pitman-Yor Process (PYP) and its realisations are almost surely discrete random measures, whose weights can be constructed using the following stick-breaking process

$$\begin{aligned} v_h &\sim \text{Beta}(v_h \mid 1-d, \alpha + dh), \text{ for } h = 1, 2, \ldots \\ w_1 &= v_1 \\ w_h &= v_h \prod_{r<h} (1 - v_r), \text{ for } h = 1, 2, \ldots, \end{aligned}$$

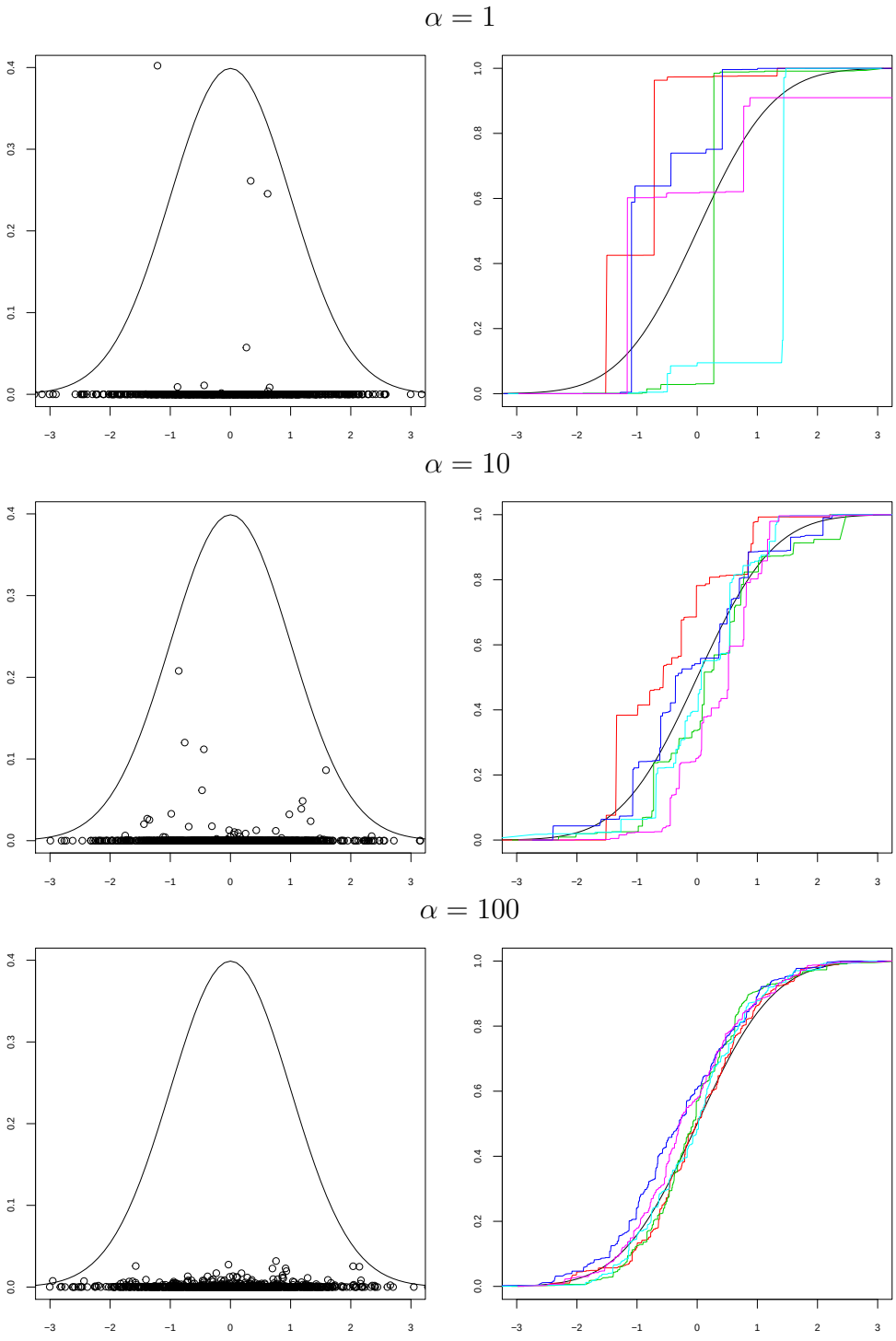


FIGURE 1.1: Examples of discrete probability measures (left panels) sampled from DP with center measure $G_0 = N(0, 1)$ (continuous black line) and $\alpha = 1$, 10, and 100 (from top to bottom). The cumulative distribution functions of five samples from the same DP are shown on the right panels.

where $d \in [0, 1)$ and $\alpha \in (-d, +\infty)$. Trivially, for $d = 0$ PYP becomes a DP with precision equal to α. It can be shown that the sequence of the weights decays slowly under a PYP, compared to a DP (where the weights decay exponentially quickly), then favouring the generation of new clusters.

The latter observation is also confirmed by looking at the process over partition induced by the PYP. This results in the following generalisation of the CRP in (1.2.3):

$$\Pr[i = n+1 \text{ assigned to } j \mid \rho_n] = \begin{cases} \frac{n_j - d}{\alpha + n}, & \text{for } j = 1, \dots, k \\ \frac{\alpha + dk}{\alpha + n}, & \text{for } j = k+1 \end{cases}, \tag{1.2.7}$$

where ρ_n is a partition of $N = \{1, \dots, n\}$.

Finally, taking the probabilities in (1.2.7) it can be shown that under a PYP with parameters d and α the expected number of clusters is

$$\mathbb{E}[k] = \frac{\Gamma(n + \alpha + d)\Gamma(\alpha + 1)}{d\Gamma(\alpha + d)\Gamma(n + \alpha)} - \frac{\alpha}{d},$$

which approximates to $\Gamma(\alpha + 1)/(d\Gamma(\alpha + d))n^d$.

1.3 Dirichlet Process Mixture Models

The aggregating property of DP makes it particularly effective to deal with clustering problems. In fact, arguably the most famous application of the DP is the Dirichlet Process Mixture (DPM) model (Lo [1984], Escobar and West [1995]), a class of models that can be expressed hierarchically as follows:

$$\begin{aligned} y_1, \dots, y_n \mid \theta_1, \dots, \theta_n &\overset{ind}{\sim} p(y_i \mid \theta_i) \\ \theta_1, \dots, \theta_n \mid G &\overset{iid}{\sim} G \\ G &\sim \text{DP}(\alpha, G_0). \end{aligned} \tag{1.3.1}$$

We write $\overset{ind}{\sim}$ to say *independent distributed*. This model assumes individual-level parameters θ_i for $i = 1, \dots, n$, but the discreteness of the DP distributed prior G implies that the vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ can be rewritten in terms of its unique values $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_k^*)$ and the assignment vector $\boldsymbol{s} = (s_1, \dots, s_n)$ (using the

notation introduced for the BMU in Section 1.2.1). The latter defines a partition of the observations whose sets can be interpreted as clusters of individuals.

In Figure 1.2 we display three sets of (truncated) infinite mixtures of Normal densities (one per panel) generated after sampling five independent realisations of G. We set the latter to be distributed as a DP with $\alpha = \{1, 10, 100\}$ respectively in each panel. As we would expect, $\alpha = 1$ gives *a priori* large probability to a small number of components of the mixture. On the contrary, $\alpha = 100$ gives high probability to a large number of components.

In order to highlight that (1.3.1) defines an infinite mixture model we write an equivalent representation of the DPM model given by

$$\begin{aligned} y_1, \ldots, y_n \mid G &\overset{iid}{\sim} p(y \mid G) \\ p(y \mid G) &= \int p(y \mid \theta) \mathrm{d}G(\theta) \\ G &\sim \mathrm{DP}(\alpha, G_0). \end{aligned} \tag{1.3.2}$$

Recalling the discrete nature of the DP samples as well as its representation in (1.2.5), we can rewrite the sampling model as an infinite mixture model:

$$y_1, \ldots, y_n \mid G \overset{iid}{\sim} \sum_{h=1}^{\infty} w_h p(y \mid \theta_h).$$

Alternatively, a DPM can be specified using the BMU and CRP, *i.e.* integrating out G from the joint distribution of $\theta_1, \ldots, \theta_n$. In particular, exploiting the probability over partitions of the indexes $\{1, \ldots, n\}$ in (1.2.4) implied by a DP distributed random measure, we can rewrite (1.3.1) as a Random Partition Model (RPM, see Lau and Green [2007] for details). An RPM is characterised by within-cluster-submodels and by a prior distribution on the partition. So, DPM in (1.3.1) is equivalent to

$$p(\rho_n, \boldsymbol{y}, \boldsymbol{\theta^*}) \propto \prod_{j=1}^{k} \left\{ \prod_{i \in S_j} \left[p(y_i \mid \theta_j^*) \right] g_0(\theta_j^*) \alpha (n_j - 1)! \right\}, \tag{1.3.3}$$

where g_0 is the density associated with the distribution G_0, while $S_j = \{i : s_i = j, \text{ for } i = 1, \ldots, n\}$. Compared to (1.3.1), this is the joint density model, with

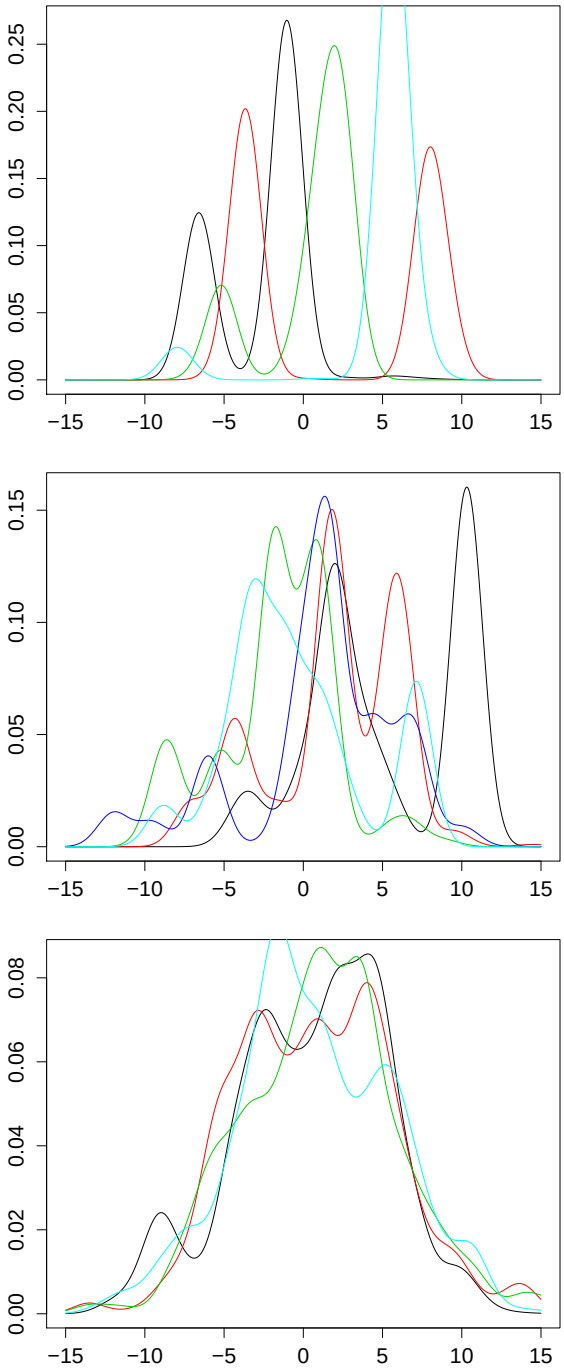


FIGURE 1.2: DPM of Normal densities referring to (1.3.1). Top panel displays five densities corresponding to five independent realisations of $G \sim \text{DP}(1, G_0)$. In middle and bottom panels, G is distributed as a DP with $\alpha = 10$ and $\alpha = 100$ respectively.

G integrated out and parameterised in terms of the partition of the observations and the unique values among individual parameters. The term $\alpha(n_j - 1)!$ is called *cohesion function* for the j-th group of observations and is often denoted by $c(S_j)$. Since $p(\rho_n)$ can be seen as the product of the cohesion functions for each of the groups, this links the DPM with a specific type of RPM called Product Partition Model (PPM, Barry and Hartigan [1992], Hartigan [1990]), characterised in the same way.

Extensions of the model in (1.3.1) and (1.3.3) can be obtained by employing more general classes of prior distributions for G (see for example Chapter 4). For detailed review see Lijoi and Prünster [2010].

Using (1.2.1), it is possible to specify the conditional posterior distribution of θ_i for the model in (1.3.3) as follows:

$$p(\theta_i \mid \boldsymbol{\theta}_{-i}, \boldsymbol{y}) \propto \sum_{l \neq i} p(y_i \mid \theta_i)\delta_{\theta_l}(\theta_i) + \alpha \int p(y_i \mid \theta)\mathrm{d}G_0(\theta)g_0(\theta_i \mid y_i), \tag{1.3.4}$$

where $\boldsymbol{\theta}_{-i}$ is the vector obtained from $\boldsymbol{\theta}$ after removing its i-th component and $g_0(\theta_i \mid y_i)$ is equal

$$g_0(\theta_i \mid y_i) = \frac{p(y_i \mid \theta_i)g_0(\theta_i)}{\int p(y_i \mid \theta)\mathrm{d}G_0(\theta)}.$$

The latter can be regarded as the posterior distribution of θ_i when s_i is different from all other indicators in $\boldsymbol{s}$.

1.3.1 Computational aspects of DPM models

Posterior inference in DPM models is often performed using Markov Chain Monte Carlo (MCMC) algorithms (for an introductory review on MCMC methods see Andrieu et al. [2003]). Posterior computations involving DPM models include the challenging step of either sampling $G \mid \boldsymbol{y}$ or $\rho_n \mid \boldsymbol{y}$. Both these posterior distributions present interesting challenges which have been largely discussed in the literature. In particular, the posterior of the random measure G is composed by the sum of an infinite collection of locations and weights, whereas the posterior of the partition of the observation ρ_n is a distribution with discrete support, where the number of possible partitions of the set $\{1, \ldots, n\}$

grows very fast at rate $\mathcal{O}(n^n)$[4].

We list below the main MCMC algorithms for posterior inference in DPM models. Exploiting the BMU versions of DPM model in (1.3.4), efficient algorithms were proposed by MacEachern and Müller [1998]. Given that G can be integrated out from the model, a Gibbs algorithm resamples the partition ρ_n from its full conditional removing one point at time and assigning it to a cluster at random according to the posterior version of the probability in (1.2.4). A split and merge method was proposed by Jain and Neal [2004] as a solution for the recurrent problem of MCMC approaches for partitions to remain blocked in configurations with high probability due to the fact that transition states are often characterised by very low probabilities. Alternatively, without integrating out G, the posterior distribution of interest, can be estimated through the slice sampler introduced by Walker [2007] or by the retrospective sampler of Papaspiliopoulos and Roberts [2008]. Approximated methods for sampling the posterior $G \mid \boldsymbol{y}$ were proposed by Ishwaran and Zarepour [2000] and Ishwaran and James [2001]. These consist in truncating the infinite mixture model implied by the DPM model in (1.3.2) and use the estimation techniques typical of the finite dimensional mixture models. A general review of available MCMC methods for DPM models is presented by Neal [2000].

1.4 Covariate dependent random measures and DPM

Recently, BNP literature has been increasingly focusing on DPM models that can vary flexibly across covariates. This need is motivated by several inferential problems such as density regression, covariate dependent clustering, nonlinear regression and classification. This can be achieved by enriching the structure of DPM in (1.3.1) in two ways: (i) specifying a covariate dependent prior distribution for the partition of the observations ρ_n in (1.3.3) or (ii) allowing the random measure G in (1.3.1) to depend on covariates. The resulting models are

[4]The number of partitions of a set containing n items is the *Bell number*, Bell(n), which is defined by the following recurrence relation:

$$\begin{aligned} \text{Bell}(0) &= 0 \\ \text{Bell}(n+1) &= \sum_{i=0}^{n} \binom{n}{i} \text{Bell}(i). \end{aligned}$$

commonly referred to as the Random Partition Models with Covariates (RPMx, Müller and Quintana [2010] and Dunson [2010]) and have been successfully applied to a wide range of real-data problems, including epidemiology (Park and Dunson [2010]), survival analysis (Müller et al. [2011]), genomics (Papathomas et al. [2012]), pharmacokinetics/pharmacodynamics (Müller and Rosner [1997]) and finance (Griffin and Steel [2006]).

When an RPMx is specified through the distribution of ρ_n, the idea is to preserve the product partition structure of the DPM, which often offers computational advantages. Considering a matrix of covariates $\boldsymbol{X}$ with n rows and D columns and letting $\boldsymbol{x}_i$ denote the i-th row, the idea is to write a prior distribution over the partition of the observations as

$$p(\rho_n \mid \boldsymbol{X}) \propto \prod_{j=1}^{k} c(S_j, \boldsymbol{X}_j^{\rho_n}), \tag{1.4.1}$$

where, for $j = 1, \dots k$, $\boldsymbol{X}_j^{\rho_n}$ is the subset of the rows of $\boldsymbol{X}$ associated with cluster j of the partition ρ_n. The covariate dependent cohesion function $c(S_j, \boldsymbol{X}_j^{\rho_n})$ is designed to assume higher values for *similar* covariates in the sub-matrix $\boldsymbol{X}_j^{\rho_n}$.

Alternatively, including covariate information within G requires the specification of a stochastic process prior over random measures which are indexed by covariate values. The most common idea is to start from (1.2.5) and to specify covariate dependent stochastic processes for the locations and for the weights (MacEachern [1999] and MacEachern [2000]). Then, recalling (1.3.1) the resulting model of the observations would be

$$y_1, \dots, y_n \mid G_{x_1}, \dots, G_{x_n} \sim \sum_{h=1}^{\infty} w_h(x_i) p(y \mid \theta_h(x_i)), \tag{1.4.2}$$

where both the weights and the locations of the mixture components are indexed by the covariates.

Although the strategies above are similar (at least in the objective) it is not always easy to link them, *i.e.* it is not always possible to derive a covariate dependent prior for ρ_n integrating out a covariate dependent random measure from the joint distribution of the parameters.

In the next sections we review the main contributions for specifying RPMx, both based on covariate dependent prior distribution of ρ_n or G.

1.4.1 Augmented Response Models

The most common strategy to include information about $\boldsymbol{X}$ into the partition model in a DPM framework has been to treat each covariate as a random variable, *i.e.* by specifying a suitable probability model. Müller et al. [1996] were the first to introduce this idea within the DPM framework. In their work they considered an augmented model defined on $\boldsymbol{Z} = (\boldsymbol{y}, \boldsymbol{X})$ and their objective was to estimate the smooth function $g(\boldsymbol{X}) = E(\boldsymbol{y} \mid \boldsymbol{X})$. They approached the problem by modelling $\boldsymbol{Z}$ as a DPM of $(R + D)$-dimensional distributions, where R is the dimension of the response variable (usually $R = 1$). Let $\boldsymbol{\Lambda}^*$ be the matrix containing the unique parameters for the k clusters, $(\Lambda_1^*, \ldots, \Lambda_k^*)$. Considering now a new observation $\tilde{\boldsymbol{z}} = (\tilde{y}, \tilde{\boldsymbol{x}})$, its predictive distribution can be derived as:

$$p(\tilde{y}, \tilde{\boldsymbol{x}} \mid \boldsymbol{\Lambda}^*) \propto \sum_{j=1}^{k} n_j p(\tilde{y}, \tilde{\boldsymbol{x}} \mid \Lambda_j^*) + \alpha \int p(\tilde{y}, \tilde{\boldsymbol{x}} \mid \Lambda) dG_0(\Lambda).$$

Assuming uncertainty about the realised value of $\tilde{\boldsymbol{x}}$, which might be a reasonable and necessary assumption when $\tilde{\boldsymbol{x}}$ is measured with error or not exactly known in real applications, allows us to rearrange the latter equation as

$$p(\tilde{y} \mid \tilde{\boldsymbol{x}}, \boldsymbol{\Lambda}^*) \propto \sum_{j=1}^{k} n_j p(\tilde{\boldsymbol{x}} \mid \Lambda_j^*) p(\tilde{y} \mid \tilde{\boldsymbol{x}}, \Lambda_j^*) + \alpha \int p(\tilde{y} \mid \tilde{\boldsymbol{x}}, \Lambda) p(\tilde{\boldsymbol{x}} \mid \Lambda) dG_0(\Lambda),$$

using Bayes' theorem. The quantity $n_j p(\tilde{\boldsymbol{x}} \mid \Lambda_j^*)$ depends on the cardinality of group j and on a measure of how likely it is that the new observation will be clustered in group j, based on the value of its covariates. The latter is the likelihood of the observed $\tilde{\boldsymbol{x}}$. The smooth function $g(\boldsymbol{X})$ is then estimated by taking the expectation with respect to $p(\tilde{y} \mid \tilde{\boldsymbol{x}}, \boldsymbol{\Lambda}^*)$. Muller et al. described in details the case where $\boldsymbol{Z}$ is a mixture of multivariate Gaussian distributions, which leads to simplified calculations for $g(\boldsymbol{X})$.

A similar approach was adopted by Müller et al. [2011]. They originally proposed a modification of a PPM, the PPMx (PPM with covariates), to incorporate measures of similarity among the covariates within each cluster employing the

following structure for the prior of the partition of the observations:

$$p(\rho_n \mid \boldsymbol{X}) \propto \prod_{j=1}^{k} c(S_j) f(\boldsymbol{X}_j^{\rho_n}), \tag{1.4.3}$$

where $f(\cdot)$, called *similarity function*, is an *ad hoc* function that takes large values for highly similar values of the covariates. The authors proposed as a default choice to specify $f(\cdot)$ a probability density (or mass) function. They showed under mild conditions that $f(\boldsymbol{X}_j^{\rho_n})$ can be seen as the likelihood of the covariates belonging to cluster j, from which the cluster specific parameters have been integrated out. Given the cluster specific parameters for the covariates, the joint probability of a PPMx is:

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}^*, \boldsymbol{\zeta}_1^*, \ldots, \boldsymbol{\zeta}_D^*, \rho_n) \propto \\ \prod_{j=1}^{k} \prod_{i \in S_j} \left[p(y_i \mid \theta_j^*, \boldsymbol{x}_i) f(\boldsymbol{x}_i \mid \zeta_{j1}^*, \ldots, \zeta_{jD}^*) \right] p(\theta_j^*) f(\zeta_{j1}^*, \ldots, \zeta_{jD}^*) c(S_j), \end{aligned} \tag{1.4.4}$$

where $\boldsymbol{\theta}^*$ and $\boldsymbol{\zeta}_1^*, \ldots, \boldsymbol{\zeta}_D^*$ include the unique values of the parameters of the distribution of the response and of the covariates for the k clusters, respectively. Expression (1.4.4) shows that the PPMx is a generalisation of the methodology proposed in Müller et al. [1996]. Taking $c(S_j)$ in (1.4.4) to be the cohesion function implied by the DP and the covariates to be random variables with distribution $p(\boldsymbol{x}_i \mid \zeta_{j1}^*, \ldots, \zeta_{jD}^*)$ (thus allowing the similarity function to be a valid probability density for the covariates), the PPMx simply reduces to a DPM on the joint distribution of the response and the covariates representable by the following hierarchy:

$$\begin{aligned} y_1, \ldots, y_n \mid \boldsymbol{X}, \boldsymbol{\theta} &\overset{ind}{\sim} p(y_i \mid \boldsymbol{x}_i, \theta_i) \\ \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \mid \boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n &\overset{ind}{\sim} p(\boldsymbol{x}_i \mid \boldsymbol{\zeta}_i) \\ (\theta_1, \boldsymbol{\zeta}_1), \ldots, (\theta_n, \boldsymbol{\zeta}_n) \mid G &\overset{iid}{\sim} G \\ G &\sim \text{DP}(\alpha, G_0), \end{aligned} \tag{1.4.5}$$

with $G_0 = G_{0\theta} \times G_{0\zeta}$ (where $\times$ denotes the product measure), $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{iD})$. Both (1.4.4) and (1.4.5) define a PPMx, in which $\boldsymbol{\theta}$ and

$\boldsymbol{\zeta}$ are assumed a priori locally independent but globally dependent. Therefore, every DPM can be represented as a PPMx, but the reverse is not always true. For this relation to hold, it is necessary that $p(y_i, \boldsymbol{x}_i \mid \theta_i, \boldsymbol{\zeta}_i) = p(y_i \mid \theta_i, \boldsymbol{x}_i)p(\boldsymbol{x}_i \mid \boldsymbol{\zeta}_i)$. In this perspective the PPMx generalises the work by Müller et al. [1996] allowing for the possibility of user-specific models for the covariates (via the similarity function).

Alternatively, Park and Dunson [2010] proposed the Generalised Product Partition Model (GPPM). The authors discussed how to incorporate covariate information in the conditional prior distribution in (1.2.1). This results in a generalised BMU scheme from which they derived a covariate dependent version of the PPM showing the same joint model in (1.4.4).

Within the PPMx framework in (1.4.4), the sampling model $p(y_i \mid \theta_j^*, \boldsymbol{x}_i)$ does not necessarily need to be a linear regression. Hannah et al. [2011] extended (1.4.4) to the broader Generalised Linear Model (GLM) framework through the appropriate specification of $p(y_i \mid \theta_j^*, \boldsymbol{x}_i)$. This generalisation allows the user to handle different types of data. They referred to this model as DP-GLM (see also Shahbaba and Neal [2009]). A parametric version, *i.e.* with a finite number of mixture components, of the DP-GLM constitutes a particular case of the Hierarchical Mixture of Experts (HME) model introduced by Jordan and Jacobs [1994] and specified in a Bayesian framework by Bishop and Svenskn [2002].

Profile Regression (PR; Molitor et al. [2010]) is another prominent example of augmented response models. In the original formulation this model handles a binary outcome $\boldsymbol{y} = (y_1, \ldots, y_n)$ which is common in epidemiological applications, but the model is easily generalised to different types of response variable. The PR model consists of two sub-models. The first one is the model for the response:

$$y_i \mid p_i \sim \text{Bernoulli}(p_i),$$

with a logistic regression on the mean p_i:

$$\log\left(\frac{p_i}{1-p_i}\right) = \theta_i + \boldsymbol{\kappa}\boldsymbol{w}_i. \tag{1.4.6}$$

$\boldsymbol{w}_i$ is a set of confounding variables with coefficients $\boldsymbol{\kappa}$, while θ_i is an individual random intercept.

The second sub-model is a mixture model on the covariates, such that conditioning on the cluster assignment vector, the probability of a specific covariate *profile* becomes:

$$\boldsymbol{x}_i \mid \boldsymbol{\zeta}_i \sim p(\boldsymbol{x}_i \mid \boldsymbol{\zeta}_i). \tag{1.4.7}$$

When $\boldsymbol{x}_i$ is a vector with D components, we can model each component independently and we can treat $\boldsymbol{\zeta}_i$ as a vector containing the parameters for each component of the profile, *i.e.* $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{iD})$.

In order to consistently estimate the posterior distribution of the partition and the cluster-specific parameters, the authors proposed to model jointly the random intercepts in (1.4.6) and the parameters of the covariates sub-model in (1.4.7) according to an unknown distribution G, which follows a DP with parameter α and G_0, with G_0 being the product measure of $G_{0\theta}$ and $G_{0\zeta}$. Expressing the joint model in terms of the implied partition and the cluster-specific parameters, the PR can be equivalently represented as the PPMx in (1.4.4).

For the augmented response class of models, `R` packages are available for the PPMx (`https://www.ma.utexas.edu/users/pmueller/prog.html#PPMx`) and for PR (`http://cran.r-project.org/web/packages/PReMiuM`).

1.4.2 Dependent Dirichlet Process

An alternative way to include covariate information in DPM is to allow the weights and/or the locations in the stick-breaking construction of the DP in (1.2.5) to depend on covariates. In particular, this can be represented in the following way:

$$\begin{aligned} G_x &= \sum_{h=1}^{\infty} w_h(\boldsymbol{x}) \delta_{\theta_h(\boldsymbol{x})} \\ w_h(\boldsymbol{x}) &= v_h(\boldsymbol{x}) \prod_{r<h} \left[1 - v_r(\boldsymbol{x})\right], \end{aligned} \tag{1.4.8}$$

under the constraint that $\sum_{h=1}^{\infty} w_h(\boldsymbol{x}) = 1$. $w_h(\cdot)$ is a function of the covariates. In this context $\boldsymbol{x}$ represents a point in some covariate space $\mathcal{X}$ and $v_h(\boldsymbol{x})$ is a realisation of a Beta distribution with parameters equal to 1 and $\alpha(x)$, the latter being the (positive) realisation of a stochastic process indexed at $x \in \mathcal{X}$. The model defined in (1.4.8) is a particular case of the Dependent Dirichlet Process (DDP, MacEachern [1999]). Each G_x is still marginally a DP for each $\boldsymbol{x}$. In

its original formulation, the DDP model allows for both covariate dependent weights as well as for covariate dependent locations, as in (1.4.8). In many applications the original formulation of the DDP has been reduced to accommodate covariate dependent locations only (examples are De Iorio et al. [2009, 2004], Gelfand et al. [2005], Duan et al. [2007], among others). However, in terms of random partition models the version of (1.4.8) including only covariate dependent weights (or the one including additionally covariate dependent locations) presents the most relevant construction (see Müller and Quintana [2010]). In this case the specification of a distribution for $w_h(\boldsymbol{x})$ is central, as it determines the structure of the dependence between the covariates and the weights, and consequently the way in which the covariate profiles inform the clustering structure.

Although assuming that the G_x are marginally (for each level of the covariates) a Dirichlet process or other known processes can be convenient (*e.g.* Griffin and Steel [2006], Griffin and Steel [2010] and Chung and Dunson [2011]), several authors have preferred to replace $v_h(\boldsymbol{x})$ and to employ a model $\Phi_h(\boldsymbol{x})$ in order to allow for more flexible stick-breaking processes. The resulting processes do not belong to DDP anymore. Examples include the Kernel stick-breaking (*i.e.* when $\Phi_h(\cdot)$ is a user-defined function with codomain in $(0, 1)$ which often captures the distance of the covariates from centroids) in Reich and Fuentes [2007] and Dunson and Park [2008], the Probit stick-breaking (*i.e.* $\Phi_h(\cdot)$ is the cumulative distribution function of a Normal density, whose input can be a function of the covariates or alternatively a spatial process indexed to the covariates) in Rodriguez et al. [2009], Chung and Dunson [2009], Rodriguez and Dunson [2011] and Arbel et al. [2016] and the Logistic stick-breaking (*i.e.* $\Phi_h(\cdot)$ is a Logit function, whose argument is a function of the covariates) in Ren et al. [2011] among others. See Foti and Williamson [2015] for a review. Similar approaches which however do not rely on a stick-breaking construction are in Karabatsos et al. [2012] and in Antoniano-Villalobos et al. [2014]. The choice of the distribution for $w_h(\boldsymbol{x})$ determines the DDP (or a dependent stick-breaking), which can then be used as mixing measure in a hierarchical model leading to (1.4.2). Note that it is also possible to assume an extra linear regression sampling model for y, *i.e.* $p(y_i \mid \boldsymbol{x}, \theta_h)$ instead of $p(y_i \mid \theta_h)$.

A related approach is the Weighed Mixture of DP (WMDP) by Dunson et al.

[2007], which can be thought of as a finite mixture of DP distributed components, one for each covariate level. The weights of this mixture are specified as functions of the covariates. The resulting random measures maintain covariate independent locations and can be used conveniently to specify an infinite mixture model with covariate dependent weights.

Alternative solutions introduce dependence within the more general construction of the discrete random measures based on Poisson random measures (Kingman [1967]). See for instance the works of Müller et al. [2004], Griffin and Leisen [2014], and Lijoi et al. [2014].

Many of DDP models can be fitted by the software `Bayesian Regression` by Karabatsos [2015, 2016] available at `http://georgek.people.uic.edu/BayesSoftware.html`.

1.4.3 Other Methods

In this section we briefly present two other methods that can be used to specify covariate dependent DPM.

The first one is the Restricted DPM (RDPM) model introduced by Wade et al. [2013]. The authors modified the usual structure of the DPM models by imposing restrictions to the distribution of the partition of the observations to follow the covariate proximity. For example, let us consider n instances of a univariate covariate, $x_1, \ldots, x_n$ and the permutation of $1, \ldots, n$ given by ordering increasingly the covariate values, namely $\sigma_x(1), \ldots, \sigma_x(n)$. The RDPM restricts the prior probability over the partition of the observations implied by a DPM and considers only the partitions for which $s_{\sigma_x}(1) \leq \ldots \leq s_{\sigma_x}(n)$. It can be shown that this construction satisfies the Ewens sampling law (Ewens [1972]) for the probability of the cluster frequencies. This same law is satisfied by partitions implied by the BMU in (1.2.1). This class of models is appealing because it does not assume any distribution on the covariates when accounting for the covariate similarity. The authors showed how to perform posterior inference in the RDPM through efficient MCMC algorithms. The mixing properties of the MCMC scheme are improved by restricting the support of the random partition.

A second alternative is represented by the Enriched Dirichlet Process Mixture (EDPM) model described in Wade et al. [2014]. The strength of this method

consists in its ability to create nested partitions (*i.e.* partitions within sets of a partition). To this end, the authors specified a DPM model for the response variable, setting a DP prior on the parameters of the sampling model for y. A DP prior, dependent on the parameters of the response, is used for the parameters of the sampling model on the covariates. This construction leads to a nested clustering structure of the observations: a first level of clustering is at the response level, whereas a second level is obtained within the clusters formed at the first stage according to a DPM model on the covariates.

1.4.4 Remarks

Covariate dependent Dirichlet process mixture models have been increasingly used in practice, especially when the objective is to specify flexible regression models. The main motivation underlying the use of such models is to improve predictions, in comparison to other possible nonparametric cluster-wise regression models. The latter has been demonstrated in simulation for augmented response models in Cruz-Marcelo et al. [2013]. The improvement in predictions is the result of substituting the traditional mixture weights in DPM models, which depend on the cardinalities of each cluster, with some function of the covariates. In this way the relation between covariates and response is studied within clusters of observations, whose assignment probabilities vary across the covariate space.

The review of covariate dependent DP presented in this section shows that there are mainly two strategies for specifying such models in the context of DP. The first way consists of modelling jointly the response and the covariates as a DPM of multivariate distributions. The main advantage of using this technique is its computational simplicity. In fact, for all types of covariates the main model remains a DPM, which has computational advantages allowing the use of the efficient algorithms introduced by MacEachern and Müller [1998] and Neal [2000] for posterior inference. For these models it is also possible to integrate out the variability on the mixing measure so that the conditional prior distributions on the parameters of the mixture model can be expressed as a modified Blackwell-MacQueen urn which includes the covariates (see Park and Dunson [2010]). On the other hand, the main disadvantage of this strategy is related to the fact that for high dimensional covariate space the likelihood of

the augmented response variables becomes dominated by the portion relative to the covariates and consequently the response does not inform effectively the clustering.

The second technique relies on modifying the stick-breaking process through which the weights (and/or the locations) of the traditional DPM models are constructed to include covariates. All contributions to this field can be divided between those that assume DPM models for each level of the covariates and those which do not. In the first case the stick-breaking procedure at each covariate level has to involve a sequence of Beta$(1, \alpha)$ random variables. This may be a limitation in incorporating complicated covariate dependencies in the weights, thus stick-breaking procedures which involve link functions that map some regression of the covariates into the $(0, 1)$ set have progressively been employed. Once a convenient link function is found, a variety of types of dependence can be accommodated in the weights, which is the main advantage of these techniques. However, this kind of models often leads to poor inference when few observations are available for each covariate level (even more so in presence of continuous covariates). Furthermore, posterior inference may require more sophisticated algorithms (as the slice sampler by Walker [2007] or retrospective sampler by Papaspiliopoulos and Roberts [2008]) or truncation of the infinite mixture to some fixed level for allowing the use of the blocked Gibbs sampler by Ishwaran and James [2001].

1.5 Outline of the thesis

In the following chapters we present four contributions to the literature of covariate dependent random measures which have been inspired by different applied problems. Although we will often refer to the specific questions while motivating and constructing each contribution, we believe that the methodological insights and the modelling strategies described in the sequel of this work can be useful in different fields requiring no or little adaptation, depending on the features of the specific problem. These contributions can be classified, using the terminology of the previous section, into the group of augmented response models and DDP mixture models.

We begin in Chapter 2 by dealing with the problem of specifying flexible regression models for count data showing an out of pattern number of zeros.

The motivation for this method comes from urology and is connected with the need of a diagnostic tool which could assist the clinicians while assessing the presence of infections of the lower urinary tract.

In Chapter 3 we investigate a way for inducing variable selection in cluster-wise linear regressions. In particular, the objective is to divide a simple linear regression problem in strata involving groups of observations which can be considered *similar* and to perform cluster-specific variable selection. The motivation comes from the analysis of which symptoms may be relevant predictors for different degrees of infection of the lower urinary tract.

The third contribution in Chapter 4 involves an extension of the DDP to allow for a better control of the induced partition of the observations while including covariate information. The use of the resulting process, called Dependent Generalised Dirichlet Process, is then exemplified on two data sets involving the evaluation of the risk of developing osteonecrosis as a consequence of the treatments for leukaemia and the study of the determinants of the performance of London primary schools.

The last contribution in Chapter 5 introduces a method based on covariate dependent random measures to model vectors of correlated binary variables evolving over time. The motivation for this comes from a data set containing profiles of symptoms recorded after a sequence of attendance visits in which we want to study possible interactions among the symptoms and their persistency.

The work is concluded in Chapter 6 with a summary of the main findings and a discussion of possible new research directions.

Chapter 2

A Bayesian nonparametric model for zero-inflated data

Lower Urinary Tract Symptoms (LUTS) affect a significant proportion of the population and often lead to a reduced quality of life. LUTS overlap across a wide variety of diseases, which makes the diagnostic process extremely complicated. In this chapter we focus on the relation between LUTS and Urinary Tract Infection (UTI). The latter is detected through the number of White Blood Cells (WBC) in a sample of urine: WBC≥ 1 indicates UTI and high levels may indicate complications. The objective of this work is to provide the clinicians with a tool for supporting the diagnostic process, deepening our understanding of LUTS and UTI. We analyse data recording both LUTS profile and WBC count for each patient. We propose to model the WBC using a random partition model in which we specify a prior distribution over the partition of the patients which incorporates information contained in the LUTS profile. Then, within each cluster of patients, the WBC counts are assumed to be generated by a zero-inflated Poisson distribution. The results of the predictive distribution allows identifying the symptoms configuration most associated with the presence of UTI as well as with severe infections. The material included in this chapter is based on the work of Barcella et al. [2016a].

2.1 Introduction

Lower Urinary Tract Symptoms (LUTS) define a group of symptoms that comprises urgency, pain, stress incontinence and voiding problems. They particularly affect elderly population with 40% of the men and 28% of the women with age between 70 and 79 years (Irwin et al. [2006]) suffering from them. This group of symptoms is related to a number of diseases (from neurological

pathologies to anxiety and stress) which are not directly identified by disjoint groups of LUTS, making the diagnostic process complicated. Often LUTS indicate the presence of Urinary Tract Infection (UTI), a condition that may lead to chronic problems when not readily diagnosed and consequently require time consuming and expensive treatments.

Given the difficulty in interpreting LUTS, specific exams are commonly employed in order to assess the presence of the infection. The published data show that the best biological indicator of UTI available is pyuria (≥ 1 White Blood Cell count (WBC) μl^{-1}) detected by microscopy of a fresh unspun, unstained specimen of urine (Khasriya et al. [2010], Kupelian et al. [2013]). In the presence of symptoms, any pyuria (≥ 1 WBC μl^{-1}) correlates with other independent inflammatory and microbiological markers distinguishing patients from controls (Gill et al. [2015], Khasriya et al. [2010], Kupelian et al. [2013]). This procedure allows counting the WBC, but on the other hand it can only be performed in specific laboratories, requiring time to return the results as well as representing a consistent cost for the health system. Therefore, it is common practice to use dipsticks for examining urine samples, which can reveal the presence of WBC which in turn indicates UTI. Dipsticks can be used by non-specialised clinicians and deliver a result in few instants. However, Khasriya et al. [2010] investigated the diagnostic power of dipstick urinalysis and identified deficiencies. Thus, an infection can be present much earlier than being diagnosed using a dipstick increasing significantly the risk of chronicity.

For all these reasons, it is valuable to study the relation between LUTS and UTI from a statistical point of view, in order to provide tools for assisting the clinicians during the diagnostic process. This is the broad objective of this chapter.

The starting point of our analysis is a dataset containing information about patients affected by LUTS for which the counts of the WBC from the microanalysis have been recorded together with the symptoms profiles. The latter are vectors of binary indicators which indicate the presence of the symptoms. The WBC counts in the dataset are zero more than 50% of the time, *i.e.* more than half of the patients do not show microscopic evidence of UTI. We thus propose an approach to model the relation between the WBC counts (response) and the LUTS profiles (covariates), which extends nonparametrically the well known

class of the zero-inflated distributions (Neelon et al. [2010]). This class of distributions has been extensively employed in a number of applications: it involves the specification of a parameter that regulates the inflation of the probability for a specific outcome which could not be modelled according to standard distributions.

Specifically, we propose a Bayesian RPM (Lau and Green [2007]) in which the covariates are used jointly with the response to inform the clustering structure of the observations which has been a priori assumed to follow a CRP (Aldous [1985]). For a review about random partition models with covariates see Section 1.4. Within each cluster we treat the WBC counts as *iid* random variables distributed according to a Zero-Inflated Poisson (ZIP) distribution with cluster-specific parameters. In this way we assume the covariates to affect the response only through the clustering structure. The latter assumption can be relaxed to allow also the mean of the Poisson component to depend on the covariates. We call the resulting model Bayesian Nonparametric ZIP model (BNP-ZIP).

BNP-ZIP allows associating different combinations of the covariates with different probabilities of having UTI (*i.e.* WBC$\geq$1) as well as with different levels of severity of the infections. The results of the study highlight the importance of the voiding class of symptoms for both the probability of being diagnosed with UTI and also its level of severity (which increases with the number of WBC in the urine). Differently, the urgency and stress incontinence symptoms have low probability of being associated with UTI when they appear alone or combined. We also believe that the predictive distributions which depend on the covariates may represent a useful tool for supporting the clinicians in the diagnostic process.

The rest of the chapter is organised as follows. In Section 2.2 we introduce and discuss the zero-inflated models, while in Section 2.3 we describe our nonparametric approach. Section 2.4 presents the analysis of the LUTS dataset. We conclude the chapter with a discussion of the results in Section 2.5.

2.2 Models with zero-inflated (or deflated) distributions

Count data with out-of-pattern number of zeros are common in numerous real world applications. Modelling such data without accounting for the excess of zeros may lead to biased estimates of the parameters. The common approach to deal with this problem involves the use of mixture models in which a distribution over counts (*e.g.* Poisson distribution, Negative Binomial distribution, etc.) is mixed with a Dirac measure located at zero. The most famous approaches include Hurdle models (Mullahy [1986]) and Zero-Inflated models (Lambert [1992]). The first type of models specifies a mixture of a point mass at zero and a zero truncated distribution for the non-zero observations. Differently, Zero-Inflated models mix a standard distribution with the Dirac measure and consequently model the inflation of the probability of the zero outcomes.

In this chapter we focus prominently on ZIP distributions. ZIP models can be extended to incorporate covariate information through regressions using convenient link functions on both the mixing probability and the mean parameter of the Poisson distribution. In order to account for the heterogeneity of the patients, random effect models are also employed (Hall [2000], Leann Long et al. [2015], Agarwal et al. [2002]). Random effects can either be assigned individually to each observation or to clusters of observations. The latter approach is more parsimonious in the number of parameters to be estimated but, when the clustering structure of the observations is not known a priori, it is often problematic to determine the number of clusters and their compositions in order to assign effectively the random effects avoiding problems of overfitting.

This motivates our proposed approach. Placing a prior distribution over the partition of the observations allows learning from the data the clustering structure and capturing patient heterogeneity within the data.

2.3 Bayesian Nonparametric ZIP model

We present in this section a nonparametric model capable of dealing with observations having excess of zeros and accounting for clustering of individuals. We

briefly show some properties of the model and we discuss MCMC algorithms for posterior and predictive inference.

2.3.1 Random Partition Zero-Inflated Poisson model

It is often of interest to model response variables within clusters. This allows us to account for possible patterns within the data as well as for highly dispersed observations and outliers. A common assumption is to consider the observations within each cluster as *iid* from a distribution having cluster-specific parameters. However, when the data are not naturally in clusters, it is also convenient to learn the clustering structure from the data. Thus, the strategy employed in this chapter consists in specifying a convenient prior over the partition of the patients and fitting independent models within each cluster. This modelling strategy belongs to the class of RPM (Lau and Green [2007]).

We recall the notation for partitions introduced in the previous chapter, where ρ_n indicates the partition of n items in clusters $\{S_1, \ldots, S_k\}$ and $\boldsymbol{s} = (s_1, \ldots, s_n)$ denotes the cluster assignment vector. Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a collection of variables presenting an out-of-pattern number of zero observations. We assume the following joint model for the components of $\boldsymbol{y}$:

$$\boldsymbol{y} \mid \rho_n, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^* \sim \prod_{j=1}^{k} \prod_{i \in S_j} [(1 - \mu_j^*)\delta_0(y_i) + \mu_j^* \text{Poisson}(y_i \mid \lambda_j^*)], \tag{2.3.1}$$

where $\mu_j^* \in (0, 1)$ and $\lambda_j^* \in (0, +\infty)$ are cluster-specific parameters, while $\delta_0(y_i)$ is the Dirac measure which places a unitary mass of probability at $y_i = 0$.

Within each cluster, the model in (2.3.1) is a mixture between two distributions: the first one is a point mass located at 0 and the second one is a Poisson distribution with cluster-specific mean equal to λ_j^*. The model above implies that $\Pr[y_i = 0 \mid s_i, \mu_{s_i}^*, \lambda_{s_i}^*] = 1 - \mu_{s_i}^* + \mu_{s_i}^* \exp(-\lambda_{s_i}^*)$ and consequently that the probabilities of all other outcomes different from 0 follow a rescaled Poisson distribution. The role of the parameter μ_j^* is crucial since it determines the inflation level for the probability of the 0 outcome. Note that under a conventional Poisson distribution with mean λ we have $\Pr[y_i = 0 \mid \lambda] = \exp(-\lambda)$.

An alternative distribution on the counts may be employed in (2.3.1) instead of the Poisson. A common example is represented by the Negative Binomial, which having two parameters can account for over-dispersed observations within each cluster. A useful parameterisation of the Negative Binomial that can be employed in this context is the one involving mean and dispersion parameters. Assuming these parameters together with the parameter controlling the zero-inflation to be cluster-specific allows writing an equivalent random partition Zero-Inflated NB (ZINB) model.

2.3.2 Prior partition model

A RPM requires the specification of a prior distribution over ρ_n. A common choice in BNP is to use the distribution over partitions implied by the CRP (Aldous [1985]) which have been introduced in Section 1.2.1 and we report here:

$$p(\rho_n \mid \alpha) \propto \prod_{j=1}^{k} \alpha (n_j - 1)! \,, \tag{2.3.2}$$

where α is a positive scalar parameter and n_j is the cardinality of cluster S_j. The distribution above implies that also k is random taking value in $\{1, \ldots, n\}$.

2.3.3 Clustering with covariates information

When covariates are available, it can be convenient to modify the CRP prior for the partition of the observations in (2.3.2) in order to include clustering information contained within the covariates. This is equivalent to assume higher prior probability for two individuals having the same (or similar) covariate profile to co-cluster. The specification of a distribution over the partition of the observations which could include covariates information has recently received remarkable attention in RPM literature and various solutions have been discussed in Section 1.4.

In this chapter we opt for specifying a model for the covariates in order to construct a covariate dependent model on the partition of the observations. In the previous chapter this strategy was referred to as *augmented response model* and it is one of the most common in practice for its computational tractability. It has been introduced by Müller et al. [1996] and extensions have been presented

by Shahbaba and Neal [2009], Park and Dunson [2010], Molitor et al. [2010], Müller et al. [2011], Hannah et al. [2011]. Let us consider a matrix of binary covariates $\boldsymbol{X}$ with n rows and D columns and denote with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iD})$ a generic row of $\boldsymbol{X}$. Similarly to $\boldsymbol{y}$, we assume clusters of rows of $\boldsymbol{X}$ to be generated by the same distribution. We use $\boldsymbol{\zeta}_j^* = (\zeta_{j1}^*, \ldots, \zeta_{jD}^*)$ to denote the cluster-specific parameters for the model of the covariates and we write

$$\boldsymbol{X} \mid \rho_n, \boldsymbol{Z}^* \sim \prod_{j=1}^{k} \prod_{i \in S_j} \prod_{d=1}^{D} \text{Bernoulli}(x_{id} \mid \zeta_{jd}^*), \tag{2.3.3}$$

where $\boldsymbol{Z}^* = (\boldsymbol{\zeta}_1^*, \ldots, \boldsymbol{\zeta}_k^*)$.

The formulation proposed above allows (2.3.2) to be rewritten as the conditional probability of the partition given the covariates, which is

$$p(\rho_n \mid \alpha, \boldsymbol{X}, \boldsymbol{Z}^*) \propto \prod_{j=1}^{k} \left(\alpha(n_j - 1)! \prod_{i \in S_j} \prod_{d=1}^{D} \text{Bernoulli}(x_{id} \mid \zeta_{jd}^*) \right). \tag{2.3.4}$$

We adopt the latter to be the prior over the random partition of the observations. The second part in the distribution above represents the likelihood of the covariates within cluster S_j which takes larger values in clusters having *similar* covariates. This corrects the probability of the partition implied by the CRP favouring clusters containing homogeneous covariate patterns.

An advantage of the proposed model on the partition of the observations is the flexibility with respect to the covariate type. In (2.3.4), modifying the model on the covariates with other suitable distributions allows the user to include in the partition information from different (or mixed) covariate types. On the other hand, the main disadvantage of this formulation arises when a large number of covariates is included in the model. In this situation, the clustering information contained in the covariates tends to dominate the partition which becomes insensitive to the clustering patterns contained in the outcome. A possible solution to this problem has been presented by Wade et al. [2014].

2.3.4 Joint probability model

We call the resulting model Bayesian Nonparametric ZIP model (BNP-ZIP), which can be summarised by the following joint probability model

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{X}, \rho_n, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*, \boldsymbol{Z}^*, \alpha) = \\ p(\boldsymbol{y} \mid \rho_n, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) p(\boldsymbol{X} \mid \rho_n, \boldsymbol{Z}^*) p(\rho_n \mid \alpha) p(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^*, \boldsymbol{Z}^*) p(\alpha), \end{aligned} \tag{2.3.5}$$

where $(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ are independent from $\boldsymbol{Z}^*$. We derive $p(\boldsymbol{y}, \rho_n, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*, \boldsymbol{Z}^*, \alpha \mid \boldsymbol{X})$ from (2.3.5), which gives an RPM with covariate dependent partition.

An important aspect of the proposed formulation is that the joint model in (2.3.5), when $p(\rho_n \mid \alpha)$ is as in (2.3.2), corresponds to the joint model under a DPM model in which a DP prior is specified for the parameters of the response and the covariates. Specifically, the joint model in (2.3.5) can be rewritten as the following hierarchical model:

$$\begin{aligned} y_i, \mid \mu_i, \lambda_i &\overset{ind}{\sim} (1-\mu_i)\delta_0(y_i) + \mu_i \text{Poisson}(y_i \mid \lambda_i) \\ \boldsymbol{x}_i \mid \boldsymbol{\zeta}_i &\overset{ind}{\sim} \prod_{d=1}^{D} \text{Bernoulli}(x_{id} \mid \zeta_{id}) \\ (\mu_i, \lambda_i, \boldsymbol{\zeta}_i) \mid G &\overset{iid}{\sim} G = \sum_{h=1}^{\infty} v_h \prod_{r<h} (1-v_r) \delta_{(\mu_h, \lambda_h, \boldsymbol{\zeta}_h)} \\ v_h \mid \alpha &\overset{iid}{\sim} \text{Beta}(v_h \mid 1, \alpha) \\ (\mu_h, \lambda_h, \boldsymbol{\zeta}_h) &\sim p(\mu_h, \lambda_h, \boldsymbol{\zeta}_h) = p(\mu_h, \lambda_h) p(\boldsymbol{\zeta}_h) \\ \alpha &\sim p(\alpha). \end{aligned} \tag{2.3.6}$$

The random quantity G in the model above has been constructed using the stick-breaking procedure and it has been proved by Sethuraman [1994] to be DP distributed. Details about the relationship between DP constructed by stick-breaking and the CRP can be found in Section 1.2.2, while the corresponding connection between DPM and RPM is presented in Section 1.3 (see also Quintana and Iglesias [2003]). The equivalence between the BNP-ZIP and a DPM model is very useful when performing posterior inference.

2.3.5 Prior specification

The model described in (2.3.5) is completed by specifying the hyperprior distributions for the parameters. We assume independent prior distributions for the cluster specific parameters

$$\begin{aligned}\mu_j^* &\sim \text{Beta}(\mu_j^* \mid a_\mu, b_\mu)\\ \lambda_j^* &\sim \text{Gamma}(\lambda_j^* \mid a_\lambda, b_\lambda)\\ \boldsymbol{\zeta}_j^* &\sim \prod_{d=1}^{D} \text{Beta}(\zeta_{jd}^* \mid a_\zeta, b_\zeta).\end{aligned}$$

We also assume a prior distribution for the parameter α of the distribution of ρ_n. This parameter takes positive real values, thus we employ a Gamma prior distribution with parameters a_α and b_α. However, a prior distribution over subsets of its real support can also be employed when in (2.3.6) the random distribution G is replaced with its truncated version (up to H mixture components)

$$G_H = \sum_{h=1}^{H} v_h \prod_{r<h}(1 - v_r)\delta_{(\mu_h, \lambda_h, \boldsymbol{\zeta}_h)},$$

with $v_H = 1$, for computational reasons. A detailed discussion about the approximation of G with G_H has been presented in Ishwaran and James [2002]. When G_H is employed Ohlssen et al. [2007] discussed the choice of a Uniform prior distribution for α.

2.3.6 Posterior inference and MCMC

The model described above is a joint RPM model on the response and the covariates. The connection between the proposed RPM and the DPM model highlighted above is convenient since it allows using available efficient MCMC algorithms developed for DPM models for sampling from the posterior distributions. A review of these algorithms can be found in Neal [2000].

In a Gibbs fashion, the posterior inference can be divided in three main stages. In the first stage we resample ρ_n from its full conditional, whereas in the second one we resample the cluster-specific parameters of the response and the covariates from their full conditional distributions and finally we resample

α from its full conditional distribution. The first stage can be performed using the Algorithm 8 in Neal [2000], or alternatively through the blocked Gibbs sampler proposed by Ishwaran and James [2001]. The cluster-specific parameters are resampled independently across clusters. A Metropolis-within-Gibbs step can be designed for the parameters of the response, while known full conditional distributions are available for the parameters of the covariate model. The resampling of α can be performed using a Metropolis-within-Gibbs step. Alternatively, imposing a Gamma prior on α leads to tractable full conditional distribution as discussed in Escobar and West [1995].

Posterior inference for BNP-ZIP can be also performed using `WinBUGS` (Lunn et al. [2000]), `JAGS` (Plummer et al. [2003]) or `Stan` (Carpenter et al. [2015]) softwares for Bayesian inference. `JAGS` code is provided in Appendix A. All these softwares implement a truncated version of the DPM model to perform inference.

Posterior predictive inference is a key aspect of BNP-ZIP. Using the distribution in (2.3.2) for the partition allows the model to grow in complexity when new observations arise adding clusters to the partition. Furthermore, enriching (2.3.2) with the information of the covariates, as showed in (2.3.3), encourages observations with similar covariates to be assigned to the same cluster and hence to predict similar responses. In a standard statistical problem the response of a new individual is unknown and needs to be evaluated, while the covariates are available. Denoting with $\tilde{\boldsymbol{x}}$ and $\tilde{y}$ respectively the covariates and the response for a new individual, the predictive distribution $p(\tilde{y} \mid \boldsymbol{y}, \boldsymbol{X}, \tilde{\boldsymbol{x}})$ can be evaluated within the MCMC scheme assigning the new individual to a cluster given the available information ($\tilde{\boldsymbol{x}}$ included) and sampling from the distribution in (2.3.1) using the parameters $\mu^*_{\tilde{s}}$ and $\lambda^*_{\tilde{s}}$, where $\tilde{s}$ is the cluster allocation for the new observation with covariates $\tilde{\boldsymbol{x}}$. Note that if $\tilde{s}$ indicates a new cluster the two parameters are sampled from their prior distributions. Details of this procedure are presented in Müller and Quintana [2010].

2.4 Data Analysis: Lower Urinary Tract Symptoms

In this section we present the analysis of the LUTS data using the BNP-ZIP model. After a detailed presentation of the data, we describe the results in terms

of clustering of the patients and of predictive inference. We also highlight the medical implications of the results.

2.4.1 Data

In this study we consider $n = 1424$ patients at the first visit attendance at the *Lower Urinary Tract Service Clinic* (Whittington Hospital, London, UK). All patients are female over 18 years of age. For each of them the result of the microanalysis of a sample of urine has been recorded in terms of the WBC count. Presence of WBC in the urine (regardless of the quantity) indicates the presence of UTI (Kupelian et al. [2013]). It is worth noticing that a large number of WBC is also the sign of a high degree of inflammation and thus can be somehow treated as an indicator of the severity of the infection. The empirical distribution of WBC count is strongly positively skewed: this is due to the fact that over 50% of the counts is equal to 0. Moreover the WBC counts different from 0 are highly dispersed, ranging from 1 to 3840.

For each of the patients a profile of LUTS has been recorded. Each profile contains information about four different types of symptoms: urgency symptoms, sudden urge to urinate; pain symptoms, pain while urinating or when the bladder is full; stress incontinence symptoms, episodes of incontinence caused by stressing the bladder; and voiding symptoms, problems in voiding the bladder. We recorded the profiles by binary vectors with 4 components, each taking value equal 1 when the corresponding category of symptoms is activated and zero otherwise. On average patients have between 2 and 3 categories activated, and there are 66 patients that do not show any symptom. 163 patients suffer for all four categories of symptoms.

2.4.2 Prior settings

For the analysis of the data described above we set the hyperparameters $a_\mu = b_\mu = a_\zeta = b_\zeta = 1$, implying minimal prior information. Also the hyperparameters a_λ and b_λ are set equal to 1. We adopt the blocked Gibbs sampler to sample from the full conditional distribution of the partition, approximating the complexity of the model up to a certain number of possible occupied clusters. We consider $H = 70$ as maximum number of clusters and we also set the

hyperparameters a_α and b_α equal to 1. The truncation of the Dirichlet process has been discussed by several authors. Following the strategy in Ishwaran and James [2002], the adopted truncation level leads to negligible approximation error (given the levels of α explored by the Gibbs sampler). A different practical approach to determine H has been discussed by Ohlssen et al. [2007], who employ also a Uniform prior for α on the set $(0, 10)$. This allows setting a priori the largest possible approximation error.

We initialise the MCMC chain taking random starting points from the prior distributions. We save 20 000 samples after a burnin period of 10 000 interactions. The convergence of the MCMC chain to the posterior distribution has been assessed by trace plots and computing sample autocorrelations and effective sample sizes.

2.4.3 Results

In this section we present the results obtained fitting the BNP-ZIP on the LUTS data set. We recall that the objective is to identify the categories of symptoms most associated with infection, *i.e.* with a count of WBC larger than 0. Furthermore, we want to assess which categoris of LUTS indicate a high level of WBC, which is then related to the severity of the UTI.

Clustering output

The starting point of our analysis consists in investigating the posterior distribution of the partition of the observations, *i.e.* $p(\rho_n \mid \boldsymbol{y}, \boldsymbol{X})$.

In order to investigate the composition of the clusters in terms of patients we compute the posterior probabilities for all pairs of observations to be assigned to the same cluster. These probabilities can be computed using the samples from $p(\rho_n \mid \boldsymbol{y}, \boldsymbol{X})$ of the MCMC algorithm. With the aim of highlighting the patterns that lead to the clustering structure, we plot the probabilities of co-clustering ordering the observations according to different criteria. Figure 2.1 shows the probabilities of co-clustering, ordering the patients for increasing values of WBC (left panel) and grouping the observations in terms of observed combinations of the covariate profiles (right panel).

In the left panel, blocks of observations with large probability of co-clustering are clearly visible along the diagonal of the plot. These blocks correspond to

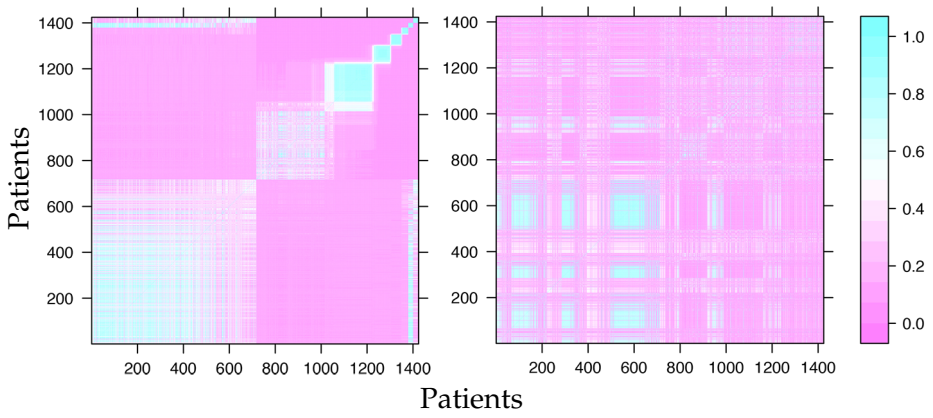


FIGURE 2.1: Level-plots of the probabilities of co-clustering of the patients ordered by increasing value of the response (*left panel*) and combinations of activated covariates (*right panel*).

groups of observations having similar responses. Evidently, these do not mix with other groups, indicating quite distinct clusters of patients. An interesting exception is represented by the first block that represents the patients with response equal to 0. In our data set the patients with WBC equal to 0 are 717. This block mixes with the blocks on the top right corner, which are those with the largest value of WBC, underlying the difficulty of the diagnostic process for UTI: similar patients (in terms of symptoms) may have a severe infections (using the number of WBC for evaluating the severity of UTI) or no UTI.

The right panel in Figure 2.1 displays the co-clustering probabilities rearranging the patients by different combinations of the covariates. Specifically, each covariate profile is composed by four binary indicators which imply 16 different combinations of covariates (all observed in the dataset).

Table 2.1 indexes the different combinations of the covariates following the order in which they appear in the right panel of Figure 2.1. Moreover, it gives the exact positions of the groups of patients characterised by the same covariate profile on the level-plot. Looking at the areas with high probability of co-clustering in the right panel of Figure 2.1 we notice that patients having the first seven combinations of covariates tend to co-cluster with the patients presenting the same symptoms and also with some of the patients having only one category activated. From the left panel of the same figure we know that ordering the patients based on the value of WBC highlights distinct clusters. Therefore, finding high co-clustering probabilities for these indexes of symptoms implies

TABLE 2.1: Combinations of the symptoms. The columns *From* and *To* identify the positions of the groups of patients sharing the same combination of covariates in Figure 2.1 (right panel)

Index	Urgency	Pain	Incontinence	Voiding	From	To
1	0	0	0	0	1	66
2	1	0	0	0	67	220
3	0	1	0	0	221	285
4	0	0	1	0	286	362
5	0	0	0	1	363	394
6	1	1	0	0	395	494
7	1	0	1	0	495	712
8	1	0	0	1	713	785
9	0	1	1	0	786	797
10	0	1	0	1	798	917
11	0	0	1	1	918	936
12	1	1	1	0	937	990
13	1	1	0	1	991	1159
14	1	0	1	1	1160	1246
15	0	1	1	1	1247	1263
16	1	1	1	1	1264	1424

that these are likely to indicate specific mixture components. On the other hand, indexes from 8 to 16 show less evident clustering structure (with some exception for example for index 12). This indicates that the symptom configurations coded with these indexes may belong to different mixture components which are also connected with different values of the WBC.

We further characterise the clusters in terms of symptoms considering a point estimate of ρ_n, say $\hat{\rho}_n = \{\hat{S}_1, \ldots, \hat{S}_k\}$, and controlling which symptoms are activated for the different sets of $\hat{\rho}_n$. We estimate $\hat{\rho}_n$ minimising the Binder loss function (Binder [1978]) which has the following form

$$L(\hat{\boldsymbol{s}}, \boldsymbol{s}) = \sum_{i<i'} \left(\ell_1 I_{\{\hat{s}_i \neq \hat{s}_{i'}\}} I_{\{s_i = s_{i'}\}} + \ell_2 I_{\{\hat{s}_i = \hat{s}_{i'}\}} I_{\{s_i \neq s_{i'}\}} \right), \tag{2.4.1}$$

where $\hat{\boldsymbol{s}}$ is a proposed partition, while $\boldsymbol{s}$ indicates the true partition. The choice of the constants ℓ_1 and ℓ_2 allows us to express the preference for a small number of large clusters or for a large number of small clusters, respectively. In our application we set $\ell_1 = \ell_2 = 1$ penalising both terms equally. In this application the true partition is unknown, but its distribution can be approximated using

the draws from the posterior distribution of the membership indicators. The posterior expectation of (2.4.1) is

$$\mathbb{E}(L(\hat{\boldsymbol{s}}, \boldsymbol{s}) \mid \text{Data}) = \sum_{i<i'} \mid I_{\{\hat{s}_i = \hat{s}_{i'}\}} - \gamma_{ii'} \mid$$

where $\gamma_{ii'} = \mathbb{E}(I_{\{s_i = s_{i'}\}} \mid \text{Data})$ and it can be consistently estimated by the samples from $p(\rho_n \mid \boldsymbol{y}, \boldsymbol{X})$ approximated by the MCMC. The $\hat{\boldsymbol{s}}$ minimising the latter expectation is taken as point estimate of ρ_n. The `R` package `mcclust` is available for deriving $\hat{\boldsymbol{s}}$ from MCMC samples of the partition of the observations (`https://cran.r-project.org/web/packages/mcclust/`).

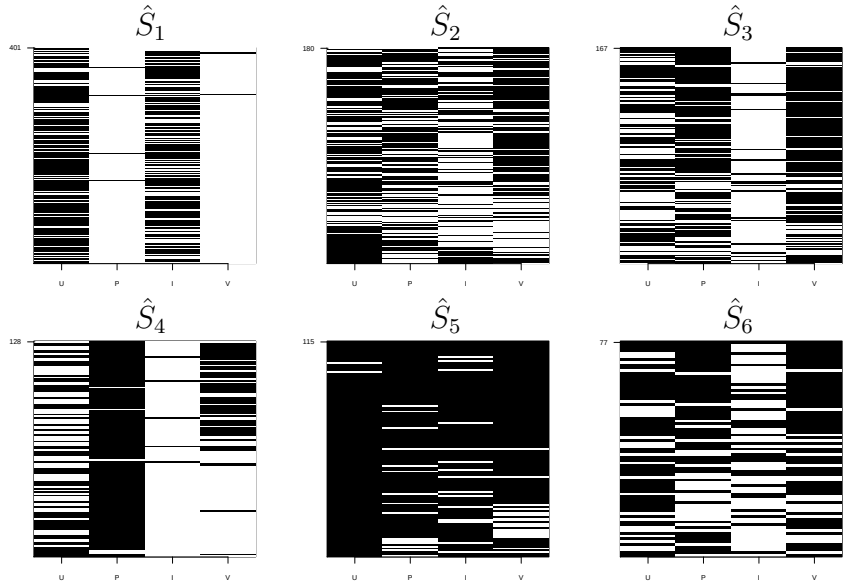


FIGURE 2.2: Symptom indicators (black) for the six largest clusters in $\hat{\rho}_n$, *i.e.* the partition estimated minimising the Binder loss function. For each panel (corresponding to a cluster), the x-axis is related to the symptoms, whereas y-axis shows the patients for each cluster. The number in the top-left corner of each panel corresponds to the cluster size.

In Figure 2.2 we display the composition (in terms of symptoms) of the six largest clusters, which contain 75% of the patients. Each panel corresponds to a cluster: each row of the plot corresponds to a patient while the x-axis represents the four symptoms. Black cells indicate activated symptoms. In most of the panels in Figure 2.2 a pattern is evident. For example cluster $\hat{S}_1$ (top-left panel), which corresponds to the largest estimated cluster, contains mainly

patients with urgency and stress incontinence symptoms. Other examples are $\hat{S}_4$ (bottom-left panel), which contains mainly patients with the pain symptoms activated, or $\hat{S}_5$ which shows patients with all symptoms activated. Recalling that each cluster is associated with similar covariates as well as with response values generated by the same distribution, finding a pattern in the covariates implies that particular symptoms, or combination of symptoms, are predictive of similar response levels.

In order to have a better understanding of how different combinations of symptoms relate with the response, in particular for those symptoms which have high uncertainty about the clustering assignment, we explore the predictive distribution of the response conditioning on symptoms combinations.

Predictive inference

Treating the symptom profiles as random in order to incorporate the covariate information in the partition of the observations has remarkable advantages in practice when the objective is to predict the level of WBC (the response), given the symptom profile $\tilde{\boldsymbol{x}}$. The BNP-ZIP will tend to assign the new patient to the cluster characterised by similar/equal symptoms combination, and thus predict a value of the response similar to the response of the patients in that cluster. This practical advantage has been widely discussed in the literature by Müller et al. [1996], Müller et al. [2011], Park and Dunson [2010], Hannah et al. [2011] (among the others) and in the review papers by Müller and Quintana [2010] and Cruz-Marcelo et al. [2013].

We analyse the predictive distribution $p(\tilde{y} \mid \boldsymbol{y}, \tilde{\boldsymbol{x}}, \boldsymbol{X})$ in order to gain some understanding about the relationship between the different covariates combinations and the presence and severity of UTI. In Figure 2.3, we plot the posterior predictive distribution of y, $p(\tilde{y} = 0 \mid \boldsymbol{y}, \boldsymbol{X}, \tilde{\boldsymbol{x}})$, for $\tilde{\boldsymbol{x}}$ equal to the different combinations of the covariates indexed according to Table 2.1. This is equivalent to the predictive distribution of not having UTI. This figure shows that the covariates with index 2,4 and 7 have posterior median probability of WBC equal to 0 close to 0.9 and with small dispersion. Moreover, Figure 2.2 seems to suggest that these covariate indexes often co-cluster (see top-left panel relative to $\hat{S}_1$). Also covariate index 1 has a similar median, but with larger dispersion. The

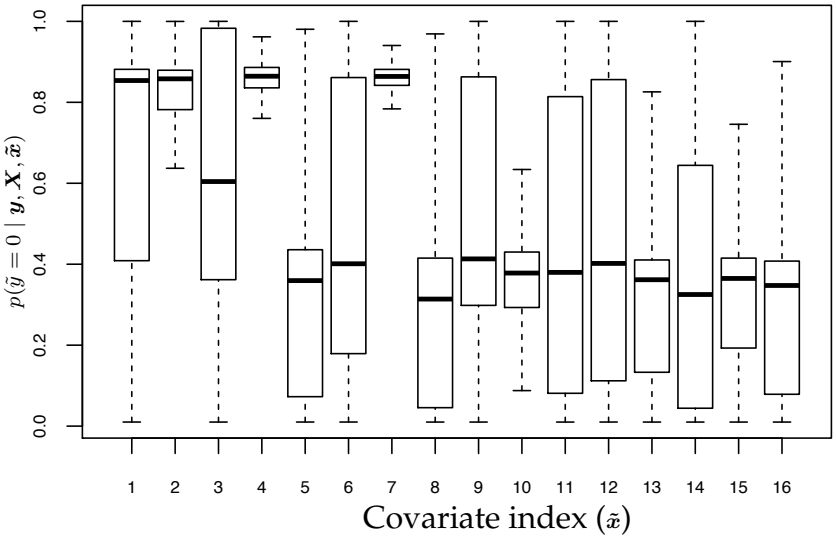


FIGURE 2.3: Posterior predictive distributions of the probability of WBC equal to 0, given the covariate indexes in Table 2.1. The black line in each box represents the median.

first three combinations of the covariates highlight that urgency and stress incontinence symptoms (or their combination) are associated with low probability of UTI, while index 1 corresponds to the configuration without symptoms. Other combinations present similar and very high median probability of having UTI. Interestingly, the profiles presenting voiding category activated have low medians for the probability of WBC equal to 0 and small dispersion. This is evident especially for index 10 and 13, which seem to often belong to the same cluster (see top and bottom right panels referring to $\hat{S}_3$ and $\hat{S}_6$ in Figure 2.2). Also pain symptoms seem connected with infection, although the respective distributions are right skewed or very dispersed (see box plots relative to the covariates indexed as 3, 6, 9 and 12).

In order to study the relation between the categories of symptoms and the severity of UTI, we compute the distribution of the third quartile of the predictive distribution for all the combinations of the covariates. While clinicians commonly agree that high levels of WBC are connected with complicated infections, the third quartile of the distribution of the WBC does not have *per se* a clinical interpretation. In fact, the choice of the third quartile has only a statistical interpretation. The distributions of these quantities for all symptoms combinations are displayed in Figure 2.4. The distributions displayed are of-

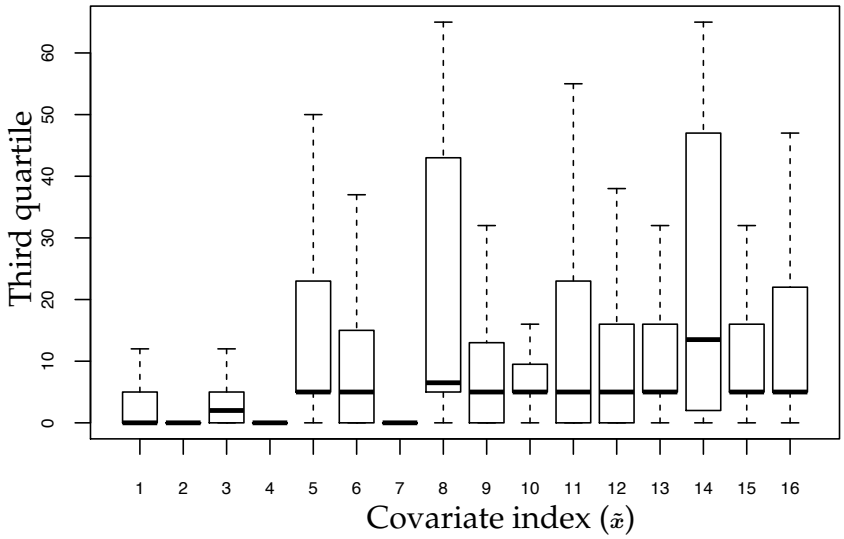


FIGURE 2.4: Distributions of the third quartile of the predictive distribution of WBC, given the covariate indexes in Table 2.1. The black line in each box represents the median.

ten right skewed with very long tail. The median of all distributions is smaller than 20. The largest median is associated with profile 14, which has also the second longest tail. Profiles 14 and 8 are characterised by the voiding category activated together with the urgency and stress incontinence categories, which confirms the results about the probability of having UTI. This suggests that not only voiding category indicates high probability of UTI, but also it indicates severe infection (when combined with urgency and stress incontinence problems).

We have performed a similar analysis using a ZINB within-cluster likelihood, which we call Bayesian Nonparametric ZINB model (BNP-ZINB). This has been done to check whether the Poisson assumption within each cluster could be too restrictive. We compared the BNP-ZIP with BNP-ZINB using the Brier score function as described in the next section. The results of this comparison show that BNP-ZIP produces more accurate prediction for the presence of infection, while it is comparable to the BNP-ZINB for predicting high values of WBC.

The results of the analysis are of considerable clinical importance. Most clinicians assume that pain is the primary symptom indicative of urinary infection. In fact, some doctors will not consider the diagnosis of UTI in the absence of

pain. Thus, the findings of this chapter suggest that the treatment of the infection should reverse this situation. Regrettable, urologists assume that the voiding symptoms are caused by a structural obstruction of the urethra and treat affected women by stretching the urethra. This procedure is unlikely to help infection and carries the risk of causing urinary incontinence. Instead, it is more likely that the voiding symptoms arise because of the inflammation induced by swelling of the urethra induced by the infection which in turn causes a relative obstruction to the urinary outflow.

Comparison with Related Methods

We evaluate the performance of the proposed method comparing it with a Bayesian ZIP model (see for example Neelon et al. [2010]) and with a DPM of Poisson distributions equivalent to the BNP-ZIP except for the absence of zero-inflating parameters in the likelihood. In order to perform the comparison we divide the the entire data set into a training set (which contains 80% of the records) and a test set, both maintaining the same proportion of covariate types as the whole data set. After fitting the model on the training set we evaluate predictive performance using the test set. We use the distribution of the Brier score (Brier [1950]) which is calculated as

$$\text{Brier}_{(q)} = \frac{1}{m}\sum_{i=1}^{m}\left(f_i^{(q)} - y_i^{(q)}\right)^2$$

where $y_i^{(q)}$ is equal to 1 if $y_i > q$ and 0 otherwise, $f_i^{(q)}$ is the probability to observe a response larger than q and m is the dimension of the test set. Small values of the Brier score function indicate good predictions. We consider $q = 0, 10, 45$, the latter two being the third quartile and mean of the WBC. The results show that the nonparametric methods outperform evidently the parametric ZIP. Instead, between the BNP-ZIP and DPM of Poisson distributions the differences are less evident (especially for the discretisation level equal to 10), but in favour of the proposed method. The same conclusions can be achieved also comparing the models in terms of Deviance Information Criterion (Spiegelhalter et al. [2002]). Although the predictive performances are similar, the main difference between a DPM of Poisson distributions and the BNP-ZIP is in the cluster composition

and consequently their interpretation, which we reckon more natural and connected to traditional ZIP regression models. In fact, in our model clusters with the same combination of covariates can accommodate both the excess of zeros and the non-zero counts. On the other hand, the DPM creates clusters with the mean of the Poisson very close to zero to accommodate the excess number of zeros, but it also yields extra clusters if for same combination of symptoms a significant number of high counts are observed. As a result, our model leads to a more parsimonious representation of the clustering structure.

Traditional methods that can be used for the analysis of WBC counts using the symptoms as predictors include Classification And Regression Trees (CART, Breiman et al. [1984]) and Random Forests (Breiman [2001]). These are likelihood-free methods which partition progressively the covariate space according to some decision rule in order to reduce the variability of the associated response variable within each partition set. We compare the partition obtained through these methods with the one estimated by the proposed technique. Both CART and Random Forests highlight the importance of voiding symptoms. For random forests this has been evaluated using the decrease in residual sum of squares in a cluster (or node) achievable splitting on a certain variable. The importance of voiding symptoms in the analysis with BNP-ZIP is evident by looking at the division between the groups of indexes in Figure 2.3 and 2.4.

Sensitivity analysis

The BNP-ZIP requires the specification of four pairs of hyperparameters, namely (a_α, b_α), (a_ζ, b_ζ), (a_μ, b_μ) and (a_λ, b_λ). We check the sensitivity of our model to different choices of the hyperparameters, focusing on the effects on cluster compositions. We propose two different checks. The first one consists of computing the absolute values of the difference (entry-wise) of the co-clustering probability matrices obtained with the values of the hyperparameters in Section 2.4.2 (used as reference values) and under alternative scenarios. We summarise the distribution of the entries of the upper-triangular matrix containing the absolute valued differences of the co-clustering probabilities using 95% credible intervals. The second method consists of estimating the mode of the number of clusters (ordered by size) which contains 95% of the patients under different choices of the hyperparameters.

We start considering the sensitivity of the proposed model to a_α and b_α, keeping the reference values for the other hyperparameters. We set different scenarios in order to have different values of prior expectation and variance of the number of clusters, *i.e.* $\mathbb{E}(k)$ and $\mathbb{V}(k)$ (formulae for approximating these quantities are presented in Jara et al. [2007]). The reference choice, $a_\alpha = b_\alpha = 1$, leads, for $n = 1424$, to $\mathbb{E}(k) \approx 7.84$ and $\mathbb{V}(k) \approx 44.57$, which we reckon to be a good trade-off between prior mean and prior variance (*e.g.* compare $\mathbb{E}(k)$ with the prior standard deviation of k). The scenarios considered are presented in Table 2.2.

TABLE 2.2: Results of the sensitivity analysis for different choices of hyperparameters. *Upper bound* refers to the upper bound of the 95% credible intervals of distribution of the absolute values of the differences of the co-clustering probabilities. *Mode* indicates the mode of the distribution of the number of clusters (ordered by size) which contain 95% of the patients.

Scenario	$\mathbb{E}(k)$	$\mathbb{V}(k)$	Upper bound	Mode
Reference	7.84	44.57	-	10
(i) $a_\alpha = 1, b_\alpha = 5$	2.51	3.59	0.0575	10
(ii) $a_\alpha = 3, b_\alpha = 1$	19.02	95.27	0.1465	14
(iii) $a_\alpha = 5, b_\alpha = 5$	7.84	13.87	0.0690	10
(iv) $a_\alpha = 3, b_\alpha = 2$	10.84	34.18	0.1975	15

The results presented in Table 2.2 show that the proposed model is robust to the choices of hyperparameters in scenario (i) and (iii) compared to the reference scenario. Differently, scenarios (ii) and (iv) are less robust. This suggests that increasing the prior expectation of the number of clusters impact the posterior inference in particular when this variations does not correspond to a relative increase in the variance of k.

We use the same strategy to assess the sensitivity of the model to the hyperparameters for the distributions of the cluster specific parameters. In addition to the choice adopted in this chapter, *i.e.* $a_\zeta = b_\zeta = a_\mu = b_\mu = a_\lambda = b_\lambda = 1$, we consider the scenarios in Table 2.3. The hyperparameters a_α and b_α are set equal to 1 in all scenarios above. Results for all scenarios (including the reference one) are in the same table. These show that under the stated criteria the BNP-ZIP is robust to different values of a_ζ, b_ζ, a_μ and b_μ. Differently, the clustering composition is slightly affected by the choice of a_λ, b_λ. In fact, the distribution of the differences of co-clustering probabilities shifts to higher values following

TABLE 2.3: Results of the sensitivity analysis for different choices of hyperparameters. *Upper bound* refers to 95% credible intervals of distribution of the absolute values of the differences of the co-clustering probabilities. *Mode* indicates the mode of the distribution of the number of clusters (ordered by size) which contain 95% of the patients.

Scenario	Upper bound	Mode
Reference	-	10
$(a_\zeta = 0.5, b_\zeta = 0.5)$, $(a_\mu = 1, b_\mu = 1)$, $(a_\lambda = 1, b_\lambda = 1)$	0.0510	10
$(a_\zeta = 1.5, b_\zeta = 1.5)$, $(a_\mu = 1, b_\mu = 1)$, $(a_\lambda = 1, b_\lambda = 1)$	0.0380	10
$(a_\zeta = 1, b_\zeta = 1)$, $(a_\mu = 1, b_\mu = 1)$, $(a_\lambda = 1, b_\lambda = 0.1)$	0.1485	13
$(a_\zeta = 1, b_\zeta = 1)$, $(a_\mu = 1, b_\mu = 1)$, $(a_\lambda = 0.1, b_\lambda = 0.1)$	0.0875	12
$(a_\zeta = 1, b_\zeta = 1)$, $(a_\mu = 0.5, b_\mu = 0.5)$, $(a_\lambda = 1, b_\lambda = 1)$	0.0610	10
$(a_\zeta = 1, b_\zeta = 1)$, $(a_\mu = 1.5, b_\mu = 1.5)$, $(a_\lambda = 1, b_\lambda = 1)$	0.0535	10

higher prior variances for λ_j^*. In the same way also the mode of the number of clusters containing the 95% of the patients increase.

2.5 Discussion

The present chapter proposes an approach for the study of LUTS and their relation with UTI. LUTS comprise a group of symptoms that can indicate a variety of diseases, however they are frequently associated with UTI. The latter is identified through the presence of WBC in the urine. Moreover, large WBC counts can also be connected with the severity of the infection and the degree of inflammation. Finally, UTI can become chronic if treatments for acute infections are not delivered promptly. For these reasons it is valuable to gain insight into the relationship between LUTS and UTI and to provide the clinicians with a tool capable of supporting the diagnostic process.

To this end, we propose a model for a dataset of patients affected by LUTS and for which both the symptoms profiles and WBC counts have been provided. More than half of the patients present WBC counts equal to 0, forcing a modelling strategy that could take this into account. Thus, we propose a ZIP model for the WBC with cluster-specific parameters. We employ a prior distribution on the possible partitions of the observations that includes also

clustering information within the covariates. Covariate information is incorporated by modelling the covariates as random and deriving the distribution of the partition given the covariates.

The proposed model strategy, called BNP-ZIP, builds on existing literatures about covariates dependent random partition models and zero-inflated (or deflated) distributions. BNP-ZIP allows estimating the probability of having UTI and the level of UTI (measured in number of WBC in the urine) given the patients symptoms. Thus, it identifies the combinations of covariates related with the largest probability of having UTI as well as those connected with the largest counts of WBC. Furthermore, the covariate dependent partition can model the over-dispersion in the data by including a larger number of clusters leading to robust estimates. BNP-ZIP can be specified also as a DPM of the response variable and the covariates jointly. This property, which has already been widely used in the Bayesian literature, simplifies posterior computations, allowing also the use of convenient MCMC samplers for numerical approximations.

The results show the importance of the urgency and stress incontinence symptoms. Patients with these symptoms activated are often clustered together and have probability close to 0.9 to have WBC equal to 0, which is equivalent to the absence of the infection. On the other hand voiding symptoms are highly related with a large probability of having UTI. Furthermore, a large number of WBC is predicted for the combinations of covariates including voiding category together with the urgency and stress incontinence symptoms, which underlines the importance of voiding symptoms in evaluating UTI. In fact, this has strong clinical impact since in clinical practice pain symptoms are generally considered related with infection and voiding symptoms are instead treated as consequences of structural obstruction of the urinary tract. In this sense, the estimated predictive distributions may offer an interesting tool for clinicians to support diagnosis.

Chapter 3

Variable selection in covariate dependent random partition models

*LUTS can indicate the presence of UTI, a condition that if it becomes chronic requires expensive and time consuming care as well as leading to reduced quality of life. Detecting the presence and gravity of an infection from the earliest symptoms is then highly valuable. Typically, WBC measured in a sample of urine is used to assess UTI. We consider clinical data from 1341 patients at their first visit in which UTI is diagnosed (*i.e. *WBC$\geq$ 1). In addition, for each patient, a clinical profile of 34 symptoms is recorded. In this paper we propose a BNP regression model based on the DP prior aimed at providing the clinicians with a meaningful clustering of the patients based on both the WBC (response variable) and possible patterns within the symptoms profiles (covariates). This is achieved by assuming a probability model for the symptoms as well as for the response variable. To identify the symptoms most associated to UTI, we specify a spike and slab centre measure for the regression coefficients: this induces dependence of symptom selection on cluster assignment. Posterior inference is performed through MCMC methods. The material included in this chapter is based on the works of Barcella et al. [2015] (model specification and data analysis) and Barcella et al. [2017] (literature review on variable selection).*

3.1 Introduction

In medical settings, individual level data are often collected for the relevant subjects on a variety of variables; these typically include background characteristics (*e.g.* sex, age, social circumstances) as well as information directly related to the interventions being applied (*e.g.* clinical measurements such as blood

pressure or the results of a particular test). This set up applies for both experimental and observational studies — perhaps even more so in the latter case, when data are often (although not always) collected using registries or administrative databases.

Arguably, the most common use of such data involves some form of regression analysis where the main "outcome" variable is related to (some of) the covariates (or "*profiles*") that have been collected. More specifically, clinicians may be interested in identifying suitable subgroups of patients presenting similar features; this categorisation can be used, for example, to suitably apply the optimal treatment for the (sub)population that will benefit the most. Alternatively, the focus may be on finding the covariates that best describe the variation in the outcome, for example in order to determine which symptoms should be measured to better characterise the chance that a new (as yet unobserved) patient is affected by a particular disease. The first of these tasks can be framed in the broader statistical problem of *clustering*, while the second one is an example of *model selection* (also called *variable selection*).

More interestingly, because complex and heterogeneous data are increasingly often collected and used for the analysis, a further connection between clustering and model selection can be considered, *i.e.* that one (set of) covariate(s) may be relevant in explaining the outcome variable for a subset of subjects, but not for others. In other words, the two tasks can be mixed in a more comprehensive analysis strategy to produce cluster-specific model selection.

For example, the dataset motivating this chapter includes records of LUTS and WBC for a number of patients affected by UTI (*i.e.* WBC$\geq$ 1). For each individual, the set of LUTS constitute the patient's profile, while WBC can be considered as an indicator of UTI, the actual condition being investigated; the clinical objective is to assess the potential relationship between the symptoms and the infection. Section 2.1 presents background information and relevant references about LUTS and UTI.

A standard approach to deal with these problems is to employ generalised linear models, including random effects (usually modelled using a Normal distribution) to account for heterogeneity between patients. This is clearly a restrictive assumption in many applications as often the distribution of the random effects is non-Normal, multi-modal, or perhaps skewed. In our analysis,

we move beyond the traditional parametric hierarchical models, in order to account for the known patient heterogeneity that cannot be described in a simple parametric model. This heterogeneity is a common feature of many biomedical data and assuming a parametric distribution or mis-specifying the underlying distribution would impose unreasonable constraints; this in turn may produce poor estimates of parameters of interest. It is therefore important to use non-parametric approaches to allow random effects to be drawn from a sufficiently large class of distributions. That is the modelling strategy we adopt in this chapter.

In order to take into account the heterogeneity among the patients, it is convenient to study the relationship between the covariates and the response within groups of patients having similar symptoms profiles and similar levels of WBC (*i.e.* in a clustering setting). In addition, it is crucial to evaluate which symptoms are explanatory of the level of WBC within each group (*i.e.* in a variable selection setting). The goal is to develop a method for assessing the relationship between a response variable (in our case WBC) and a set of covariates (the profile) within clusters of patients with similar characteristics, in order to make prediction about the response for a new patient. This will ultimately provide valuable information on the mechanisms of action of the underlying disease being investigated.

To this aim, we develop a modelling strategy based on BNP methods that allows us to accomplish both tasks at once. We propose a (potentially infinite) mixture of regression models to link the response with the covariates, where also the weights of the mixture can depend on the covariates. In this way, observations will be clustered based on the information contained in both the clinical profiles and the outcome variable. Within each cluster, variable selection is achieved employing *spike and slab* prior distributions that assign positive probability to the regression coefficients being equal to zero. The Bayesian framework allows us to perform both tasks simultaneously in a probabilistically sound way, so that clustering and variable selection inform each other. The results of the application on LUTS data show that our formulation leads to improved predictions, in comparison to other existing methods.

The rest of the chapter is organised as follows. In Sections 3.2 we review the variable selection techniques for covariate dependent clustering. Then in

Section 3.3 we introduce the details of our proposed approach and briefly explain how to perform posterior inference while in Section 3.5 we show how to summarise posterior inference output in a meaningful way. In Section 3.6 we present an application of our model to the LUTS dataset mentioned above. Finally, in Section 3.7 we discuss our results and draw conclusions. In addition, a simulation study is presented in Appendix B.2.

3.2 Covariate dependent clustering and variable selection

The two main topics of this chapter are *covariate dependent clustering* and *variable selection*. For the former, a review of the relevant literature on Bayesian nonparametric methods has been presented in Section 1.4. This has highlighted different ways of incorporating flexibly covariate information in a DPM (Lo [1984]) model or in the corresponding RPM.

Increasing research interest has been devoted to develop variable selection strategies in covariate dependent DPM models. Bayesian methods for variable selection have a long history and a variety of different techniques have been proposed to achieve this task (see O'Hara et al. [2009]). Within the regression framework, this corresponds to evaluate the uncertainty about the selection of covariates to include in the model. One of the most common ways to perform Bayesian variable selection in a regression framework consists in specifying prior distributions favouring shrinkage toward zero on the regression coefficients. Similarly, indicators can be included in the model to select which covariates are active. Alternatively, a prior distribution directly over the model structure can be specified. In this section we describe exclusively variable selection techniques proposed for covariate dependent DPM models and related models. We divide available tools for augmented response models and DDP.

3.2.1 Variable Selection for Augmented Response Models

Product Partition Model with Covariates (PPMx)

A variable selection strategy for the PPMx was proposed by Müller et al. [2011] and described in details by Quintana et al. [2015b]. Without loss of generality

we start our discussion by considering the PPMx from the RPM point of view. It is possible to rewrite the similarity function in (1.4.3) as the product of the similarity functions of each individual covariate, *i.e.* $f(\boldsymbol{X}_j^{\rho_n}) = \prod_{d=1}^{D} f(\boldsymbol{x}_{jd}^{\rho_n})$, where $\boldsymbol{x}_{jd}^{\rho_n}$ is the sub-vector of elements of column d of $\boldsymbol{X}$ which includes the elements corresponding to cluster j. Variable selection is then introduced employing binary indicators γ_{jd}^* for $j = 1, \dots k$ and $d = 1, \dots, D$ within the distribution of the partition:

$$p(\rho_n \mid \boldsymbol{X}, \boldsymbol{\gamma}) \propto \prod_{j=1}^{k} c(S_j) \prod_{d=1}^{D} f(\boldsymbol{x}_{jd}^{\rho_n})^{\gamma_{jd}^*}. \tag{3.2.1}$$

The presence of the binary indicators allows the probability of the partition to depend on a subset of covariates within each cluster. In fact, $\gamma_{jd}^* = 0$ eliminates the effect on the distribution of the partition of covariate d in cluster j. The authors described a local (cluster specific) summary of the importance of each covariate which requires the identification of the clusters, whose labels are arbitrary and suffer from the label switching problem. A global measure is also derived as the cluster-wise average of the posterior means of γ_{jd}^*, weighted by the relative cluster cardinalities.

In this setting, extra care is required for the specification of $f(\cdot)$. In order to perform variable selection, $f(\cdot)$ must always take values larger than 1 (otherwise excluding a covariate always increases the prior probability). The authors discuss convenient choices of $f(\cdot)$. The model is completed by introducing in the hierarchy a prior distribution for the indicators. In particular, the authors propose to use a Bernoulli prior distribution assuming a logistic link for the probability of success.

Another example of variable selection in PPMx framework is the work of Kunihama and Dunson [2014]. They propose a method for testing conditional independence of the response and a specific covariate given all the other covariates. This method is based on the conditional mutual information to measure the strength of the dependence and to select relevant covariates.

Profile Regression (PR)

Papathomas et al. [2012] investigated the problem of performing variable selection within the Profile Regression framework when all the covariates are categorical (see also Papathomas and Richardson [2014]). Let us recall that PR can

be decomposed into two sub-models: a model on the covariates and one on the response. These are linked by using a joint DP prior on the set of parameters common to both the submodels. In order to introduce variable selection we need to rewrite (1.4.7) in the following way:

$$\boldsymbol{x}_i \mid \zeta^*_{j1}, \dots \zeta^*_{jD} \sim \prod_{d=1}^{D} p(x_{id} \mid \zeta^*_{jd}).$$

Variable selection is then performed by replacing the distribution of each covariate with:

$$p^{VS}(x_{id} \mid \zeta^*_{jd}, \pi_d) = (1 - \pi_d)p(x_{id} \mid \zeta^*_{jd}) + \pi_d r_d(x_{id}), \tag{3.2.2}$$

where the superscript *VS* indicates that the implied probability has been modified to perform variable selection, $\pi_d \in (0, 1)$ is a continuous weight and $r_d(x_{id})$ indicates the proportion of times covariate d takes value x_{id}. From (3.2.2), the posterior distribution of π_d can be used to study the global importance of d-th covariate in terms of clustering. In this setting a Beta hyperprior distribution for each π_d or alternatively a mixture of a Beta distribution and Dirac measure (with Bernoulli distributed indicators) may be preferred to induce extra sparsity. The authors compared their approach that uses continuous weights to a version that employs cluster specific binary indicators for each covariate. The latter idea can be represented in the following way:

$$p^{BVS}(x_{id} \mid \zeta^*_{jd}, \boldsymbol{\gamma}^*_d) = p(x_{id} \mid \zeta^*_{jd})^{\gamma^*_{jd}} r_d(x_{id})^{(1-\gamma^*_{jd})},$$

where $\gamma^*_{jd} = 1$ indicates that covariate d is informative with respect to cluster j. This approach is a generalisation to Profile Regression of a solution proposed by Chung and Dunson [2009]. In contrast with the continuous case, the natural choice of prior distribution for each γ^*_{jd} is Bernoulli with mean distributed as a Beta distribution. Extra sparsity can be achieved substituting the latter Beta distribution with a mixture of a Beta distribution and Dirac measure (with Bernoulli distributed indicators). In this setting, a global summary of the importance of each variable can be obtained from the posterior distribution of the hyperparameters governing the distribution of the γ^*_{jd}'s. A local (cluster specific) measure based on γ^*_{jd}'s is not straightforward and requires an approach

similar to the one employed for PPMx.

The results presented by Papathomas et al. [2012] and obtained employing the extra sparsity alternative of both variable selection methods described above show a comparable performances of the two methods in terms of variable selection, although preference is given to continuous weights due to faster MCMC convergence.

An extension of the methods above was proposed by Liverani et al. [2015] to deal with continuous covariates. This consists in modifying (3.2.2) by substituting $r_d(x_{id})$ with a suitable summary statistics, for example the observed mean of the d-th covariate.

3.2.2 Variable Selection for DDP

To the best our knowledge, general variable selection strategies have not been implemented in the DDP framework. However, in the case of the dependent stick-breaking process Chung and Dunson [2009] showed how to perform covariate selection when the weights of the random probability measure are constructed by a Probit link stick-breaking. Recalling the stick-breaking procedure in (1.4.8) the following specification is proposed:

$$\begin{aligned} \psi_h(\boldsymbol{x}) &= \Phi\left(\nu_h(\boldsymbol{x})\right) \prod_{r<h} \left[1 - \Phi\left(\nu_r(\boldsymbol{x})\right)\right] \\ G_x &= \sum_{h=1}^{\infty} \psi_h(\boldsymbol{x}) \delta_{\theta_h}, \end{aligned} \tag{3.2.3}$$

where $\Phi(\cdot)$ is the standard Normal cumulative distribution functions and $\nu_k(\cdot)$ is a predictor which can be specified for example as $\nu_h(\boldsymbol{x}) = \boldsymbol{\xi}_h \boldsymbol{x}$. Variable selection is then achieved by introducing binary indicators:

$$\boldsymbol{\xi}_h \sim \prod_{d=1}^{D} p(\xi_{hd} \mid a_d)^{\gamma_{hd}} (\delta_0(\xi_{hd}))^{(1-\gamma_{hd})}, \tag{3.2.4}$$

where a_d denotes the covariate specific parameters of the distributions of ξ_{kd} for all k. Considering a regression sampling model $p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}_k^*)$, it is possible to link the results of the variable selection performed in (3.2.4) directly to the

parameters β_{kd} in the regression model for the response so that when $\gamma_{kd} = 0$ both β_{kd} and ξ_{kd} are set equal to 0.

3.2.3 Remarks

In this section we have reviewed the available methodologies for performing variable selection in covariate dependent random partition models. The idea behind these methods is to select the important covariates for their role in terms of clustering. This is because most RPMx include covariate information within the distribution of the partition of the observations. This is for example the case of the variable selection methods proposed for the PPMx or for PR, which in principle do not exclude the covariates from affecting the response variables (*e.g.* when a regression model is included in the sampling distribution), but they exclude the covariates which do not contribute in separating the observations into different clusters. The main limitation of this methods is that the clustering structure contained in the covariates may not be connected to a corresponding separation of response variables. This implies that selected covariates can be important for clustering the covariates, but not the response.

Furthermore, in mixture models the clusters of observations are directly connected to the mixture components. When variables are dependent, mixture components can be inferred to include this dependence. This strategy implies that covariates may be selected as important because highly dependent on other covariates and independent from the response, which is not an appealing property.

A more elaborate solution which links variable selection in terms of clustering and association with the response level has been presented by Chung and Dunson [2009]. This proposal employs common binary indicators for each covariate in both the sampling model and the model of the weights. This implies that if a covariate is excluded from the model of the weights is automatically excluded from the model of the response, potentially overcoming both of the above limitations. However, the results of the variable selection can be unstable given that covariates may be important for mixture components with negligible weight. This problem can be partially mitigated using a truncated version of the random measure. More details about these final considerations are reported in Chapter 6.

3.3 Random Partition Model with Covariate Selection

In this section we develop the Random Partition Model with Covariate Selection (RPMS) and briefly explain the MCMC algorithm employed to perform posterior inference.

3.3.1 Regression Model

We use a linear regression model to explain the relationship between the response and the covariates. Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ denote the response variable. Then, we assume

$$y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}_i, \lambda_i \sim \text{Normal}(y_i \mid \boldsymbol{x}_i \boldsymbol{\beta}_i', \lambda_i).$$

Here, we assume that $x_{id} \in \{0, 1\}$ for every $i = 1, \ldots, n$ and $d = 1, \ldots, D$. Thus, we can interpret $\boldsymbol{X}$ as the matrix containing the information about the presence of D symptoms for each of the n patients; these symptoms are assumed to have a potential effect on the response $\boldsymbol{y}$. We focus on binary covariates, because in clinical settings symptoms are often recorded as binary indicators (in fact, that is the case in our motivating example). Extensions to other type of covariates is however trivial.

The goal is to specify a prior structure that allows detecting a possible clustering structure based on symptoms profiles and then identifying which variables most influence (globally or in some clusters) the response variable.

3.3.2 Model on the Covariates and Prior Specification

To allow for covariate dependent clustering, we exploit ideas in Müller et al. [1996] assuming a probability model for the vectors of covariates:

$$\boldsymbol{x}_i \mid \zeta_1, \ldots, \zeta_D \sim \prod_{d=1}^{D} \text{Bernoulli}(x_{id} \mid \zeta_{id}).$$

In addition, we specify a joint DP prior distribution on $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{iD})$ and $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{iD})$:

$$\begin{aligned} (\boldsymbol{\beta}_1, \boldsymbol{\zeta}_1), \ldots, (\boldsymbol{\beta}_n, \boldsymbol{\zeta}_n) \mid G &\sim G \\ G &\sim \text{DP}(\alpha, G_0) \end{aligned} \tag{3.3.1}$$

where α is the precision parameter and G_0 is the centre measure of the process. Recalling Section 3.1, the DP in (3.3.1) assigns a positive probability for two observations i and i' to have the same values $(\boldsymbol{\beta}_i, \boldsymbol{\zeta}_i) = (\boldsymbol{\beta}_{i'}, \boldsymbol{\zeta}_{i'})$. We denote with $(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*) = ((\boldsymbol{\beta}_1^*, \boldsymbol{\zeta}_1^*), \ldots, (\boldsymbol{\beta}_k^*, \boldsymbol{\zeta}_k^*))$ the unique values for the parameters. This construction implies that observations are clustered on the basis of both their covariates profile and the relationship between covariates and responses.

The model is completed by specifying a conjugate Gamma prior on the regression precision assuming $\lambda_1 = \ldots = \lambda_n = \lambda$ and modelling $\lambda \sim \text{Gamma}(\lambda \mid a_\lambda, b_\lambda)$, as well as using a Gamma hyperprior on the concentration parameter of the DP (Escobar and West [1995]): $\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$. These are common prior choices as they enable easier computations.

The Spike and Slab Base Measure

The choice of the centre measure of the DP is crucial. We assume that $\boldsymbol{\beta}_i$ and the $\boldsymbol{\zeta}_i$ are independent in the centre measures. We choose a spike and slab distribution as centre measure for the regression coefficients to perform variable selection (see George and McCulloch [1993] and Malsiner-Walli and Wagner [2011] for a review on spike and slab distribution for variable selection). Thus we define:

$$G_0 = \prod_{d=1}^{D} \{[\pi_d \delta_0(\beta_{hd}) + (1 - \pi_d)\, \text{Normal}(\beta_{hd} \mid m_d, \tau_d)]\, \text{Beta}(\zeta_{hd} \mid a_\zeta, b_\zeta)\},$$

which is simply the product measure on the space of the regression coefficients and of the parameters defining the distribution of the covariates. The notation β_{hd} and ζ_{hd} highlights the fact that the centre measure is assumed to be the same across the observations. In G_0, the first term in the square brackets is the spike and slab distribution, where $\delta_0(\beta_{hd})$ is a Dirac measure that assigns probability

1 to the value 0. Thus a spike and slab distribution is a mixture of a point mass at 0 (in correspondence of which, $\beta_{hd} = 0$) and a Normal distribution, with weights given by π_d and $(1 - \pi_d)$, respectively.

A conjugate centre measure is employed for the covariate specific parameters ζ_{hd} for ease of computations. We assume the same hyperpriors for the parameters in G_0 as in Kim et al. [2009]. In particular we set a spike and slab hyperprior for each π_d:

$$\begin{aligned} \pi_1, \ldots, \pi_D \mid \omega_1, \ldots, \omega_D &\sim \prod_{d=1}^{D} \{(1-\omega_d)\delta_0(\pi_d) + \omega_d \, \text{Beta}(\pi_d \mid a_\pi, b_\pi)\} \\ \omega_1, \ldots, \omega_D &\sim \prod_{d=1}^{D} \text{Beta}(\omega_d \mid a_\omega, b_\omega) \end{aligned}$$

The latter solution has been proposed by Lucas et al. [2006] to induce extra sparsity on the regression coefficients, encouraging those associated with the covariates with no effect on the response variable to shrink toward zero. As shown in Kim et al. [2009], it is possible to integrate out π_d from the centre measure, obtaining:

$$G_0 = \prod_{d=1}^{D} \{[\omega_d r_\pi \delta_0(\beta_{hd}) + (1 - \omega_d r_\pi)\text{Normal}(\beta_{hd} \mid m_d, \tau_d)]\text{Beta}(\zeta_{hd} \mid a_\zeta, b_\zeta)\}$$

where $r_\pi = \frac{a_\pi}{(a_\pi + b_\pi)}$.

We set $m_1 = \ldots = m_D = 0$; in addition, we use a Gamma prior for the precision parameter of the Normal component of the spike and slab prior:

$$\tau_1, \ldots, \tau_D \sim \prod_{d=1}^{D} \text{Gamma}(\tau_d \mid a_\tau, b_\tau).$$

3.4 Posterior Inference

MCMC algorithms have been largely employed in similar settings for approximating posterior and predictive inference. Since our model can be rewritten using a DPM formulation on the response and the covariates jointly, efficient

Gibbs sampler schemes are available. We follow the auxiliary parameter algorithm proposed in Neal [2000]. This procedure updates first the vector of cluster allocations $\boldsymbol{s}$ and then separately all the cluster-specific parameters and the parameters that do not depend on the cluster allocation. The update of the partition $\boldsymbol{s}$ is performed after G has been integrated out. A detailed description of the algorithm is reported in Appendix B.1. We present below a summary:

(i) Update the membership indicator $\boldsymbol{s} = (s_1, \ldots, s_n)$ using the Gibbs sampling procedure for non-conjugate centre measure based on the auxiliary variable algorithm presented in Neal [2000].

(ii) Update the precision of the DP, α, exploiting the method introduced in Escobar and West [1995], setting $\alpha \sim \text{Gamma}(\alpha \mid a_\alpha, b_\alpha)$ a priori.

(iii) Update $\boldsymbol{\zeta}^* = (\boldsymbol{\zeta}_1^*, \ldots, \boldsymbol{\zeta}_k^*)$ from the full conditional distribution, given the new configuration of $\boldsymbol{s}$ in (i).

(iv) Update $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_k^*)$ from the full conditional posterior distribution, given the new configuration of $\boldsymbol{s}$ in (i).

(v) Update $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_D)$ from the full conditional distribution. To draw from this distribution we implement the algorithm described in Kim et al. [2009].

(vi) Update $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_D)$ from the full conditional distribution.

(vii) Update the precision of the regression λ from the full conditional distribution.

3.5 Summarising Posterior Output

The choice of a spike and slab centre measure implies that the coefficients β_{jd}^* have positive probability to be equal to zero. We propose here two ways of summarising the MCMC output that highlight the effect of using a spike and slab prior distribution. These two methods are then applied to the real data example in the following section.

In our framework a covariate can be explanatory for one cluster and not for another. Thus, a first method to analyse the results would be to compute

the probability that the $d-$th covariate has explanatory power in cluster j, *i.e.* $p(\beta^*_{jd} \neq 0)$, given a partition of the observations in clusters. The literature proposes a variety of methods for extracting a meaningful partition from the MCMC output (Dahl [2009], Molitor et al. [2010]). In our application we have decided to report the partition obtained by minimising the Binder loss function (Binder [1978]), which has been described in (2.4.1). Then, conditioning on the selected partition, we can compute the posterior distribution of the regression coefficients for each cluster, together with the probability of inclusion of a certain covariate, *i.e.* $1 - p(\beta^*_{jd} = 0 \mid \hat{\boldsymbol{s}})$, where $\hat{\boldsymbol{s}}$ is the Binder configuration.

A second way of summarising the posterior output from a variable selection perspective is based on predictive inference. Let us consider the situation in which a new patient enters the study with profile $\tilde{\boldsymbol{x}}$. Using the proposed approach, the posterior distribution of the regression coefficients depends on the structure of the patient's profile. This is due to the fact that the cluster allocation depends on it. In fact, in RPMS the predictive distribution of the cluster allocation is:

$$p(\tilde{s} \mid \tilde{\boldsymbol{x}}, \ldots) \propto \begin{cases} n_j \prod_{d=1}^{D} g_{jd}(\tilde{x}_d) & \text{for } j = 1, \ldots, k \\ \alpha \prod_{d=1}^{D} g_{0d}(\tilde{x}_d) & \text{for } j = k+1 \end{cases} \tag{3.5.1}$$

where $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \ldots, \tilde{x}_D)$ and $\tilde{s}$ are the profile for the new patient and its cluster allocation, respectively. In addition, $g_{jd}(\tilde{x}_d) = \zeta_{jd}^{*\tilde{x}_d}(1-\zeta^*_{jd})^{(1-\tilde{x}_d)}$ and $g_{0d}(\tilde{x}_d) = (\int_0^1 q^{(\tilde{x}_d + a_\zeta - 1)}(1-q)^{(b_\zeta - \tilde{x}_d)}\mathrm{d}q)/(\int_0^1 u^{(a_\zeta-1)}(1-u)^{b_\zeta-1}\mathrm{d}u)$ are the likelihood for the new observation to belong in cluster j and the prior predictive distribution of the new observation, respectively. The probability in (3.5.1) comes directly from the predictive scheme of the DP in (1.2.1). Hence, we focus on $p(\tilde{\beta}_d = 0 \mid \tilde{\boldsymbol{x}}, \boldsymbol{y}, \boldsymbol{X})$. This probability can be approximated using the MCMC samples. Moreover, it is possible to look at the predictive distribution of the response, namely $\tilde{y}$, that by construction depends on the variable selection.

Alternatively, we could compute the posterior probability of $p(\beta^*_{1d} = \ldots = \beta^*_{kd} = 0 \mid \cdot)$ to summarise the overall importance of the $d-$th covariate. This posterior probability can be approximated empirically by calculating the proportion of iterations in the MCMC run in which the regression coefficient for

the $d-$th covariate is equal to zero in all the clusters: $\beta^*_{1d} = \ldots = \beta^*_{kd} = 0$. However, this summary of the variable selection can be very sensitive to the number and cardinalities of the clusters.

3.6 Lower Urinary Tract Symptoms data

In this section we present the results of the application of the RPMS to the LUTS database. First, we briefly describe the database, then we give details of the choice of the hyperparameters and MCMC settings. We briefly introduce the competitor model and finally we report the results for clustering and variable selection.

3.6.1 Data

We consider data on 1341 patients extracted from the LUTS database collected at the LUTS clinic, Whittington Hospital Campus, University College London. The patients are women, affected by LUTS. We consider data at the first attendance visit.

For each patient, the presence of 34 LUTS has been recorded together with the WBC count in a sample of urine. Differently from the study in Chapter 2, we consider a more detailed characterisation of LUTS profile which was described in Khasriya et al. [2017]. These symptoms are stored as binary variables (1 indicates the presence of the symptoms and 0 the absence). We report the frequency distribution of the symptoms in the 1341 patients in Table 3.1. The symptoms can be grouped into the four categories described in Chapter 2: urgency symptoms (symptoms from 1 to 8), stress incontinence symptoms (9 to 14), voiding symptoms (15 to 21) and pain symptoms (22 to 34).

It is of clinical interest to investigate the relationship between LUTS and UTI, where the latter is measured by the number of WBC. In particular, a value of WBC ≥ 1 is indicative of the presence of infection, and high value of the WBC count indicate an high degree of inflammation and can be considered as a measure of the severity of the infection.

In this paper we focus only on patients with UTI (WBC≥ 1). We consider a logarithmic transformation of the WBC data and model the log-transformed

TABLE 3.1: Lists of the 34 symptoms with the frequency of occurrence.

Symptom	**Frequency**	**Symptom**	**Frequency**
1) Urgency incontinence	0.4146	18) Straining to void	0.0828
2) Latchkey urgency	0.4280	19) Terminal dribbling	0.1641
3) Latchkey incontinence	0.2304	20) Post void dribbling	0.0820
4) Waking urgency	0.5496	21) Double voiding	0.1193
5) Waking incontinence	0.2595	22) Suprapubic pain	0.1611
6) Running water urgency	0.2901	23) Filling bladder pain	0.2148
7) Running water incontinence	0.1365	24) Voiding bladder pain	0.0567
8) Premenstrual aggravation	0.0515	25) Post void bladder pain	0.0723
9) Exercise incontinence	0.1462	26) Pain fully relieved by voiding	0.0634
10) Laughing incontinence	0.1536	27) Pain partially relieved by voiding	0.1260
11) Passive incontinence	0.0783	28) Pain unrelieved by voiding	0.0164
12) Positional incontinence	0.0850	29) Loin pain	0.2081
13) Standing incontinence	0.0895	30) Iliac fossa pain	0.0895
14) Lifting incontinence	0.1104	31) Pain radiating to genitals	0.0865
15) Hesitancy	0.1797	32) Pain radiating to legs	0.0649
16) Reduced stream	0.1909	33) Dysuria	0.1484
17) Intermittent stream	0.1514	34) Urethral pain	0.0507

WBC using a Normal distribution. Figure 3.1 displays the kernel density estimation of the log-transformed WBC.

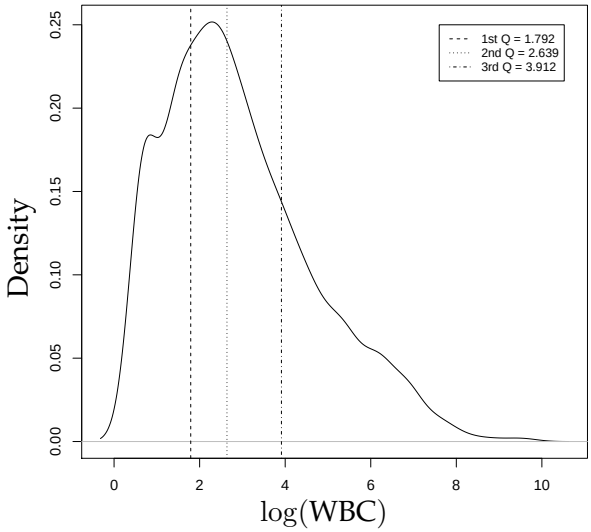


FIGURE 3.1: Kernel density estimate of the response variable log(WBC). The vertical lines correspond to the quartiles.

3.6.2 Prior Specification

The hyperparameters of the spike and slab prior in the centre measure are set as follows: $a_\pi = a_\omega = 1$, $b_\pi = b_\omega = 0.15$, $a_\tau = b_\tau = 1$ and $a_\zeta = b_\zeta = 1$. We note here that we set vague prior beliefs on the distribution of the parameters, except for the prior on π_d and ω_d to make computations more stable and encourage sparseness in the regression coefficients. The hyperparameters a_λ, b_λ for the precision λ in the regression density are set both to be equal to 1. Finally, for the prior on concentration parameter of the DP we use $a_\alpha = b_\alpha = 1$.

We run the MCMC sampler for 10 000 iterations with a burn-in period of 1 000 iterations. Details on the algorithm are given in Appendix B.1. To update the membership indicator, we use the auxiliary variable approach described in Neal [2000] that requires the choice of a tuning parameter M. In our experience, $M = 100$ gives a good trade-off between execution time and efficiency of the Gibbs sampler. The convergence of the chains is assessed by trace plots and by the Gelman and Rubin's convergence diagnostic (Gelman and Rubin [1992]), the latter for the parameters that do not depend on the cluster assignment.

3.6.3 The competitor model: SSP

In order to highlight the potential and advantages of RPMS, we compare its results with the model described in Kim et al. [2009], which we believe is the closest competitor. For simplicity, we refer to this model as SSP (Spike and Slab Prior). This assumes the same Normal specification for the WBC counts and a DP prior on the regression coefficients with a spike and slab centre measure for the regression coefficients.

The difference with our own specification consists in that the the SSP treats the covariates as given, instead of associated with a probability distribution. This implies that in the SSP the centre measure of the DP prior reduces to:

$$G_0 = \prod_{d=1}^{D} \{\pi_d \delta_0(\beta_{hd}) + (1 - \pi_d) N(\beta_{hd} \mid m_d, \tau_d)\}.$$

We follow the same strategy adopted for the RPMS of integrating out from each part of the centre measure the π_d. The model described in Kim et al. [2009] involves also the use of a DP prior on the precision in the regression model, but

for a fair comparison with the RPMS we use a version of the model without this further complication. Moreover, the results obtained by the SSP with or without the DP prior distribution on λ have not shown to be significantly different. In the MCMC algorithm, we use the same initial values, tuning parameters and number of iterations utilised for the RPMS to obtain the posterior distributions.

3.6.4 Clustering outputs

The proposed method, as explained above, employs a DP prior for the regression coefficients and for the parameters governing the profile distribution. The main consequence is that the implied clustering is influenced by both the distribution of the y and by possible patterns within the profiles.

Figure 3.2 reports the posterior distribution for k, *i.e.* the number of clusters, from the RPMS model. The configuration involving 14 clusters is clearly the one with the highest probability. This is the first significant difference with the SSP model that has a clear mode at $k = 1$. This is due to the fact that the SSP takes into account only the variability in the regression coefficients. In the RPMS the covariates contribute to inform the partition of the observations.

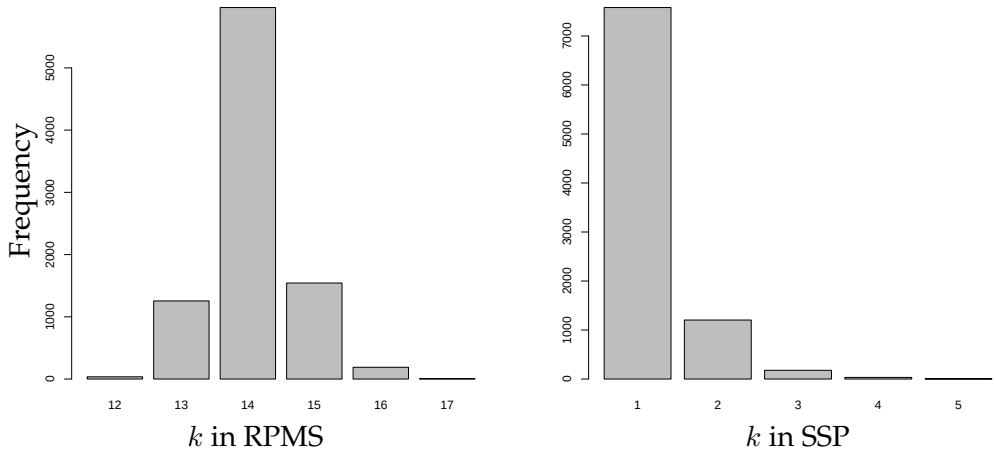


FIGURE 3.2: Posterior distribution of the number of clusters k for RPMS and for SSP models.

To summarise the posterior inference on the clustering we report the partition which minimises the Binder loss function (Binder [1978]) in (2.4.1). In our

case this leads to a configuration with 14 clusters, with the 9 largest clusters containing 92.5% of the observations.

Figure 3.3 displays the presence of the 34 symptoms (on the columns) for the patients assigned to each of the nine largest clusters. Recalling that the symptoms can be grouped into four main categories, *i.e.* urgency, stress incontinence, voiding and pain symptoms, we can see that the largest cluster contains patients with a small number of symptoms belonging to all the four categories. In the second largest cluster, almost all patients present the fourth class of symptoms and almost none the third one. A high frequency of the other urgency symptoms is also evident. The third largest cluster includes patients with a high frequency of pain symptoms together with urgency symptoms (even though with a lower probability). The fourth cluster is characterised by a high frequency of urgency symptoms; the fifth cluster by a high frequency of voiding symptoms; the sixth cluster by a high frequency of incontinence symptoms; the seventh cluster by a high frequency of urgency and incontinence symptoms; the eighth cluster by a high frequency of urgency and pain symptoms and the ninth cluster by a high frequency of urgency and voiding symptoms.

This distribution of the symptoms across the Binder configuration suggests that the symptoms classes are informative for the partition. Consequently, it is likely that each combination of symptoms has a particular effect on the WBC counts distribution: this is because cluster specific regression coefficients are associated to cluster specific probabilities of having the symptoms.

3.6.5 Variable selection outputs

Our proposed model performs simultaneously clustering and variable selection. It is of clinical interest to check which symptoms have a significant impact on the response variable. In our case, this means checking which symptoms are more likely to be predictive of the underlying severity of the infection.

In this section we will use the two ways of summarising the variable selection information produced by the RPMS that have been described in section 3.5. The first one is based on the Binder estimate of the clustering configuration, while the second focusses on the predictive distribution for a new patient. We first report the posterior probability of each symptom to be included in the model, conditional on the Binder estimate of the clustering configuration.

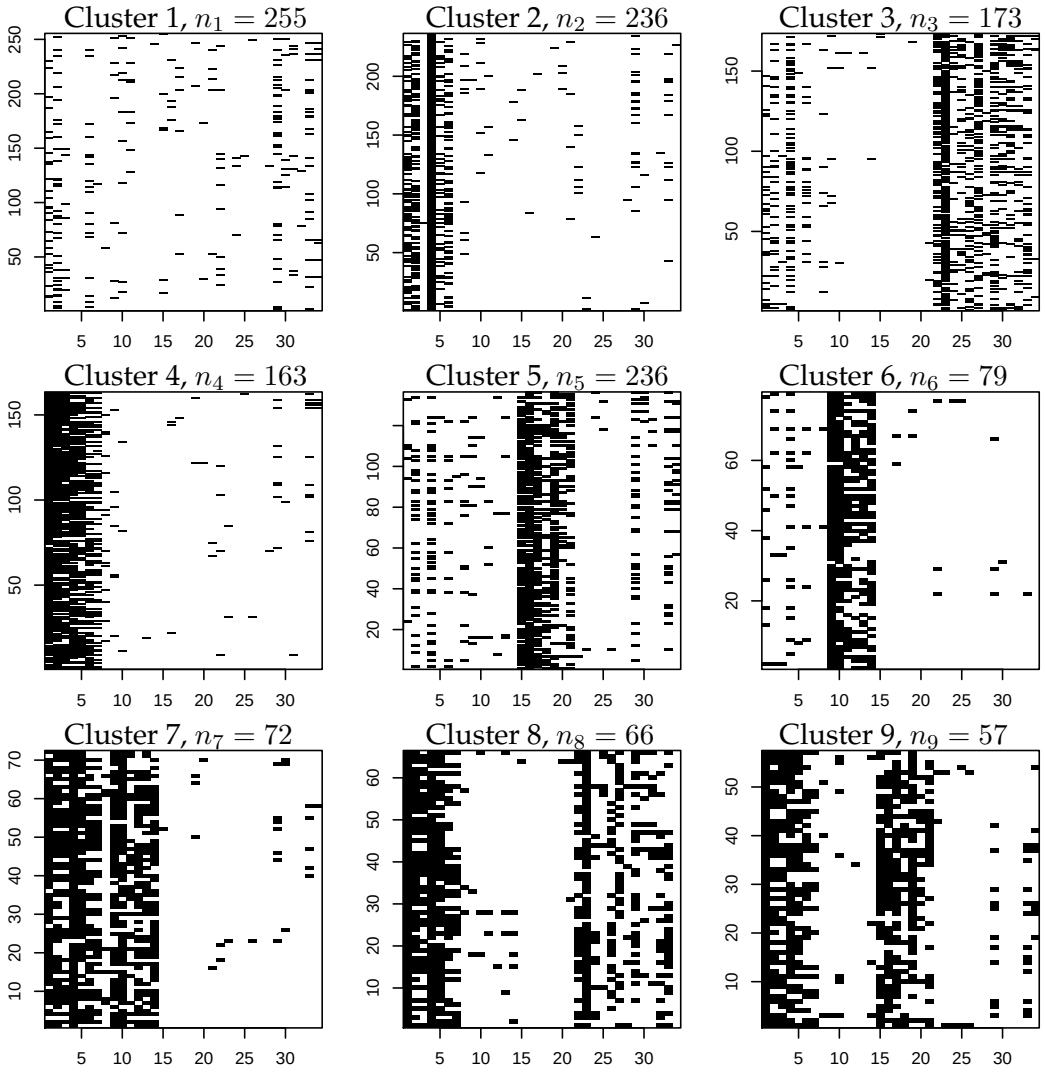


FIGURE 3.3: Symptom indicators (black) for the 9 biggest clusters of the partition obtained by minimising the Binder loss function. The horizontal axis of each panel (corresponding to a cluster) displays the index of the symptoms, whereas the index of the patients in each cluster is on the vertical axis. For each cluster the cardinality is also displayed.

Figure 3.4 displays the probability of inclusion, *i.e.* $1 - p(\beta^*_{jd} = 0 \mid \hat{\boldsymbol{s}}, \boldsymbol{y}, \boldsymbol{X})$ for the 9 largest clusters according to the Binder estimate ordered by size. For example, let us consider the fifth row (which refers to the fifth cluster). From Figure 3.3, we see that this cluster contains mainly the symptoms from 15 to 22 (cfr. the list in Table 3.1). Consequently, in Figure 3.4 the probability that

symptoms 15, 16, 17, 19 are included in the regression model is close to 0.9. On the contrary, for symptoms 18, 20, 21 and 22 the probability of being included is low. Figure 3.4 also suggests the importance of the symptoms in the urgency class and of *dysuria* and *loin pain* within the pain class.

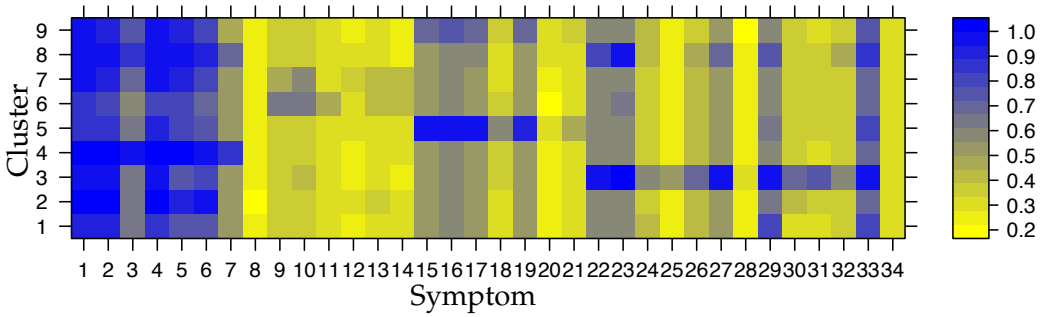


FIGURE 3.4: Probability of inclusion, *i.e.* $\beta \neq 0$, for each symptoms in the 9 biggest clusters of the partition estimated by minimising the Binder loss function.

The second way of presenting variable selection results is by considering predictive inference. Recall that for a new patient entering the study the distribution of the regression coefficients depends on her profile, *i.e.* $\tilde{\boldsymbol{x}}$, through the cluster assignment. This does not happen in the SSP, in which the predictive probability for the cluster assignment depends exclusively on the cardinality of the clusters.

To illustrate the last considerations, we take $\tilde{\boldsymbol{x}}$ including the presence of symptoms 1, 2 and 4 from Table 3.1. Figure 3.5 shows the density estimation of the posterior distribution of the regression coefficients related to the three symptoms in $\tilde{\boldsymbol{x}}$. We present the output from both the RPMS and the SSP. The evident differences between the distributions are due to the fact that in the SSP the regression coefficients do not depend on the individual's profile, which they do in the RPMS. Consequently, in the SSP the posterior distribution of the regression coefficients is the same for every (new) patient, while in the RPMS it can vary, depending on the covariates. Moreover, in the RPMS the spike and slab prior distribution can be seen as a within–cluster prior.

The different posterior distributions of the regression coefficients for the SSP and the RPMS have obviously an impact on the predictive distribution of $\tilde{y}$. Figure 3.6 displays the predictive distribution of the response given a profile

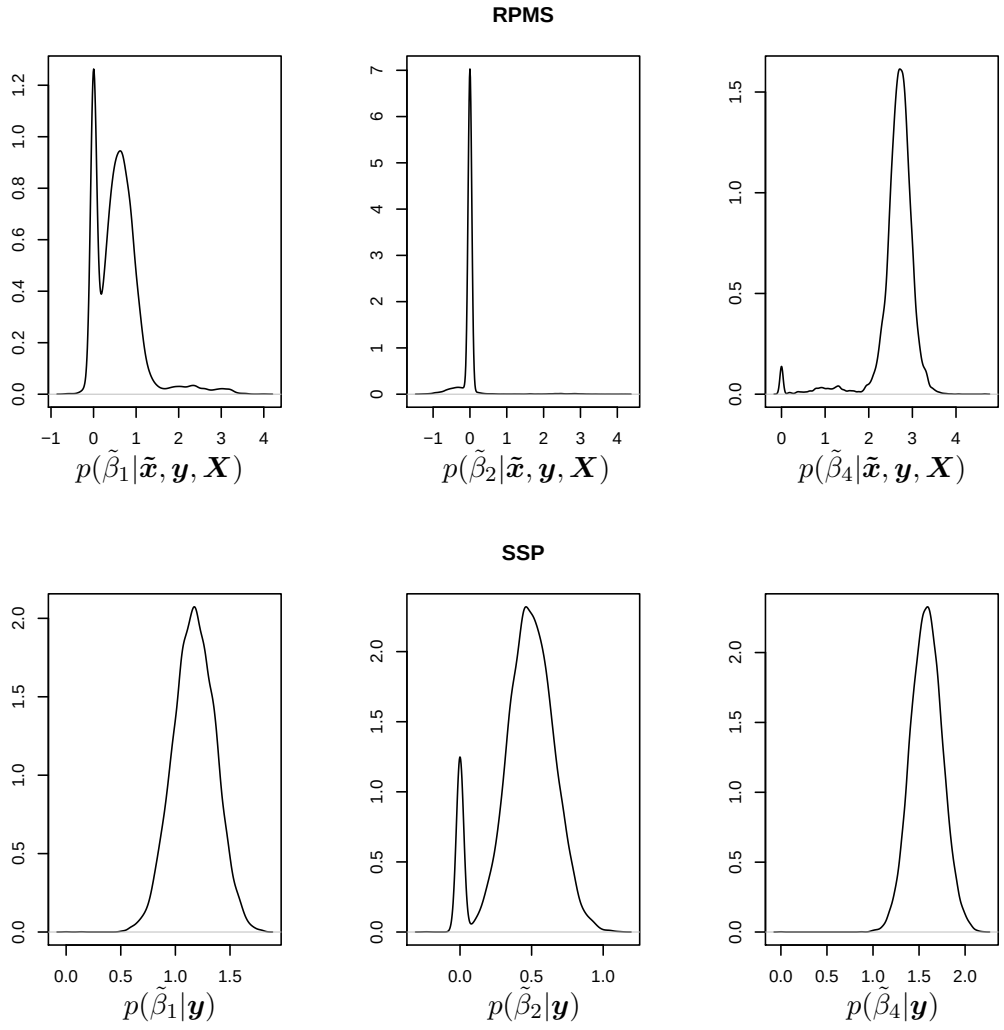
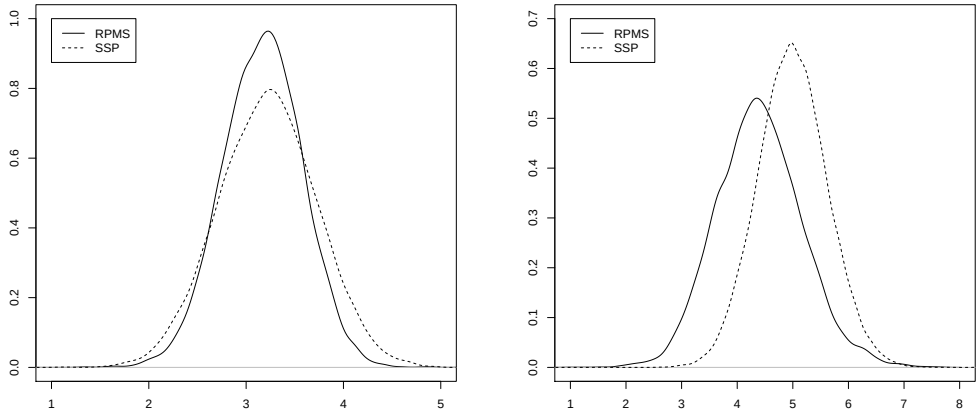


FIGURE 3.5: Kernel density estimation of the posterior distribution for $\tilde{\beta}_1$, $\tilde{\beta}_2$ and $\tilde{\beta}_4$ given the new profile $\tilde{\boldsymbol{x}}$ with $x_1 = x_2 = x_4 = 1$. The first row refers to the RPMS model, while the second row refers to the SSP model. For SSP the posterior distribution for the regression coefficients does not depend on $\boldsymbol{x}$.

$\tilde{\boldsymbol{x}}$ (we assume these are the same as those considered in Figure 3.5) and for a different profile $\tilde{\boldsymbol{x}}'$, which is characterised by a large number of symptoms: $x_1 = x_2 = x_3 = x_4 = x_5 = x_6 = x_7 = x_{22} = x_{23} = x_{27} = x_{28} = x_{32} = x_{33} = 1$.

In the first case, the distributions obtained from the SSP and the RPMS have

similar means, but the one estimated by the RPMS seems to have smaller variance. On the other hand for $\tilde{\boldsymbol{x}}'$ the predictive distributions seem more substantially different.



$p(\tilde{y}|\tilde{\boldsymbol{x}}, \boldsymbol{y}, \boldsymbol{X})$ $p(\tilde{y'}|\tilde{\boldsymbol{x}'}, \boldsymbol{y}, \boldsymbol{X})$

FIGURE 3.6: Kernel density estimation of the predictive distribution of $\tilde{y}$ given profile $\tilde{\boldsymbol{x}}$ with $x_1 = x_2 = x_4 = 1$ and of $\tilde{y'}$ given profile $\tilde{\boldsymbol{x}'}$ with $x_1 = x_2 = x_3 = x_4 = x_5 = x_6 = x_7 = x_{22} = x_{23} = x_{27} = x_{28} = x_{32} = x_{33} = 1$

In order to determine whether the proposed model leads to improved predictions we employ the Brier statistic (Brier [1950]), similarly to what has been done in Section 2.4. Each response variable has been discretised in correspondence of the quartiles of the empirical distribution of the $\log(\text{WBC})$ and the probability for an individual to have a response value larger than a given quartile has been calculated. We denote with $\text{Brier}_{(q)}$ the Brier statistic calculated using the q-th quartile as threshold. For all three thresholds the posterior expectation of $\text{Brier}_{(q)}$ is lower for the RPMS model, indicating better predictive power compared to SSP. A simulation study is included in Appendix B.2, which clarifies the reason for the improvement of predictions.

Evidence in favour of the hypothesis that $\text{WBC} \geq 1$ indicates the presence of UTI is given in Kupelian et al. [2013]. This recent result extends the analogous one by Dukes [1928], where $\text{WBC} \geq 10$ was considered. All patients in our study have $\text{WBC} \geq 1$ and thus it is very likely they are affected by UTI. However, if on the one hand a large number of WBC in a sample of urine will increase the confidence about the presence of UTI, on the other hand no work

has been published that describes the severity of infection in relation to higher values of WBC. Nevertheless, specialists consider fully reasonable to associate high degree of inflammation to large values of WBC.

Hence, discretising the response variable to make prediction is reasonable both to assess the likelihood of having an infection and to evaluate the general status of the disease (in terms, for example, of the degree of inflammation). Moreover, discretising the response variable transforms the problem into a classification one, which links our model to the very important area of risk prediction models. A review of these models can be found in Gerds et al. [2008], who highlight the most common techniques to perform model selection in this context.

3.7 Discussion

In this work we have proposed the RPMS, a DPM of Normal regressions with covariate dependent weights, capable of simultaneously performing clustering and variable selection. This is achieved employing a DPM on the joint distribution of the response and the covariates, together with spike and slab prior distributions on the regression coefficients within the clusters of the partition. The latter allows performing variable selection and therefore identifying covariates with high explanatory power on the response. The proposed method is designed to handle binary covariates, due the dataset motivating this work, even though it is straightforward to include other types of covariates (and also mixed types). Although we have presented the model for Normally distributed response data, it is possible to extend it to the generalised linear model framework.

The main feature of the model lies in the fact that the RPMS takes into account also possible patterns within the covariate space. The results of the analysis highlight the diagnostic power of the symptoms. On the other hand, the SSP focuses on the variability within the response. The results of the posterior inference show that in the partition generated by the RPMS the clusters are characterised by the presence of certain classes of symptoms or combinations of classes, revealing that these classes are informative and further investigation might lead to the identification of disease sub-types.

The results of the variable selection has been summarised in two ways: fixing a meaningful partition (we have opted for the Binder estimate) or fixing a specific profile in a predictive fashion. In the first case, the analysis of the posterior distribution of the regression coefficients conditional to the Binder partition shows the overall importance of the urgency symptoms, and the cluster-specific importance of certain particular symptoms. The second way to display the variable selection output is from a predictive perspective. We have assumed that a new patient's profile has been collected. The distribution of the regression coefficients for the new patient depends on the profile and this permits an individual-based assessment of the important symptoms that determine the distribution of WBC. The SSP's estimated posterior distributions of the regression coefficients instead do not depend on the patient profile. This difference allows the RPMS to achieve more accurate prediction of the WBC compared to the SSP.

We believe that the use of Bayesian nonparametric methods, although computationally more expensive, offers the flexibility necessary to capture the complexity of modern clinical data and consequently improved predictive power, especially in cases where the use of parametric approaches would impose unrealistic assumptions on the data generating process.

Chapter 4

Dependent generalised Dirichlet process priors

We propose a novel Bayesian nonparametric process prior for modelling collections of random discrete distributions. This process is defined by combining a Generalised Dirichlet Process with a suitable Beta regression framework that introduces dependence among the discrete random distributions. This strategy allows for covariate dependent clustering of the observations. Some advantages of the proposed approach include wide applicability, ease of interpretation and efficient MCMC algorithms. The methodology is illustrated through two real data applications involving acute lymphoblastic leukaemia and London primary schools quality evaluations. The material included in this chapter is based on the work in Barcella et al. [2016b].

4.1 Introduction

Very often real world applications involve observational data that are collected in groups or clusters. These can be characterised, for example, by spatial or temporal coordinates, as samples from the same experimental unit, or more generally by shared levels of covariates. In such settings, a common strategy is to model the data by introducing random effects to account for the correlation of the observations within each group. This approach allows robust estimation of the parameters shared by all clusters. Generalised linear mixed effects models are examples in the regression framework.

Common distributions for random effects are *e.g.* normal distributions or Student-t distributions, but these may be too restrictive in some circumstances. A variety of solutions have been presented as more flexible alternatives. Among

these proposals, nonparametric techniques, such as infinite mixture models, are gaining popularity. The most general proposals for random effects' distributions assume an infinite mixture model for groups of observations and introduce dependence among the parameters of the mixture models (*i.e.*, the weights and/or the locations). Each infinite mixture model is a convolution of a parametric density kernel with a discrete random probability measure that has (*a priori*) an infinite number of locations and weights. Thus, the problem of inducing dependence among the infinite mixture models can be rewritten in terms of the dependence among the discrete random probability measures indexed by the different groups or clusters of observations.

A seminal contribution in this field is the extension of the DP (Ferguson [1973]) called the DDP (MacEachern [1999] and MacEachern [2000]; see also Cifarelli and Regazzini [1978]). The DDP is constructed in such a way that each group of observations is distributed as a DP. The random effects distributions thereby become DPM (Lo [1984]) as in (1.3.2). Dependence among the different DP probability measures is induced by specifying convenient stochastic process priors indexed by the groups of observations, leading to group-specific weights and locations. One can specify such models by enriching the structure of the stick-breaking representation of the DP presented by Sethuraman [1994] reported in (1.2.5). A review of the most relevant contributions in this area is presented in Section 1.4.2.

In this chapter, we propose a novel approach that generalises the DDP of MacEachern [2000] by assuming that the discrete random measure associated with each group of observations is distributed according to a Generalised Dirichlet Process (GDP, Hjort [2000]) prior. The GDP employs a richer parametrisation compared to the usual DP and, for this reason, allows for more flexibility. The dependence among the different random measures is induced by specifying a convenient prior for the weights of the measures, while assuming the locations to be the same across all groups of observations (although alternatives with also covariate dependent locations can be easily specified). We call the resulting process the Dependent Generalised Dirichlet Process (DGDP).

The DGDP has a better control of the implicit partition of the observations defined by the different mixture components compared to the DDP case, in terms of the distributions of number and size of the clusters. The law of the partition induced by samples from a GDP can be derived analytically allowing

for a better interpretation of that quantity and an increased number of computational strategies compared to other processes where this cannot be derived. Furthermore, including the covariates within the weights of the process, the DGDP leads to improved predictive power compared to processes including covariate information only within the locations (Cruz-Marcelo et al. [2013]).

We illustrate the use of the DGDP in two real data applications, one related to the analysis of Acute Lymphoblastic Leukaemia (ALL), and the other one on London primary schools quality evaluations. In the first application, we model trajectories of triglycerides quantities as a function of the treatment administered to patients affected by leukaemia. The increase in triglycerides quantity given by leukaemia's treatments can lead to osteonecrosis. We estimate the distribution of the triglycerides using the DGDP indexed by different risk levels (high and low risk) of developing osteonecrosis. The results highlight the ability of the DGDP to capture the interaction between the exposure to treatment for leukaemia and the risk of developing osteonecrosis. In the second application we model the scores of the quality of primary schools in London. We use DGDP mixtures of continuous latent distributions indexed by different London boroughs in order to estimate flexibly the baseline probabilities for the different scores. We induce spatial dependence across boroughs through a convenient stochastic process prior within the DGDP. We then add linear regression components of the main school features with common coefficients across all mixture components: this allows us to estimate interpretable effects of the school features on the scores. In these illustrations, we use the DGDP in order to consider potential (latent) groupings of observations over and above groupings based on measured covariates including less restrictive prior distributions, while preserving computational simplicity.

The chapter is organised as follows. Section 4.2 reviews the main properties of the GDP and presents new results. In Section 4.3, we introduce the DGDP, and we present possible MCMC algorithms for posterior inference in Section 4.4. The real data examples analysed using the DGDP are in Section 4.5. We conclude with a discussion in Section 4.6. Proofs are deferred to Appendix C.

4.2 Generalised Dirichlet Process

4.2.1 Definition

Let us consider a measurable space $(\Theta, \mathcal{A})$ and an associated probability measure G. We say that G is distributed according to a Generalised Dirichlet Process (GDP, Hjort [2000], Ishwaran and James [2001]) with parameters $\boldsymbol{\phi} = \{\phi_h\}_{h=1}^{\infty}$ (with each element belonging to $\mathbb{R}^+$), $\boldsymbol{\mu} = \{\mu_h\}_{h=1}^{\infty}$ (with each element belonging to $(0,1)$), and center measure G_0 (a non-atomic probability measure on Θ) if it admits the following stick-breaking representation:

$$G = \sum_{h=1}^{\infty} w_h \delta_{\theta_h}, \tag{4.2.1}$$

where $\{\theta_h\}_{h=1}^{\infty} \overset{iid}{\sim} G_0$ and $\{w_h\}_{h=1}^{\infty}$ are constructed via the stick-breaking procedure. This involves a sequence of random variables $\{v_h\}_{h=1}^{\infty}$ taking values on $(0,1)$. Common practice is to assume that the random variables are Beta distributed, which we do here. We parameterise the Beta density function as

$$p(v_h \mid \phi_h, \mu_h) = \frac{\Gamma(\phi_h)}{\Gamma(\phi_h\mu_h)\Gamma(\phi_h(1-\mu_h))} v_h^{\phi_h\mu_h - 1}(1-v_h)^{\phi_h(1-\mu_h)-1},$$

where $\mu_h = \mathbb{E}[v_h]$ and $\phi_h = \mu_h(1-\mu_h)/\mathbb{V}[v_h] - 1$, with $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$ denoting the expectation and variance operators, respectively. Thus, we assume $\{v_h\}_{h=1}^{\infty} \overset{ind}{\sim} \text{Beta}(v_h \mid \phi_h\mu_h, \phi_h(1-\mu_h))$, independent from $\{\theta_h\}_{h=1}^{\infty}$, and we specify the infinite sequence of weights setting $w_1 = v_1$ and obtaining the other weights as

$$w_h = v_h \prod_{r<h}(1-v_h), \quad h = 2, 3, \ldots. \tag{4.2.2}$$

The resulting measure, G, is a proper random distribution function. Indeed it can be easily verified that $\sum_{h=1}^{\infty} \mathbb{E}[\log(1-v_h)] = -\infty$, which is a necessary and sufficient condition for $\sum_{h=1}^{\infty} w_h = 1$. (See Ishwaran and James [2001] for a detailed proof). We write $G \sim \text{GDP}(\boldsymbol{\phi}, \boldsymbol{\mu}, G_0)$.

A more parsimonious formulation of the GDP, described by Hjort [2000], assumes $\{\phi_h\}_{h=1}^{\infty} = \phi$ and $\{\mu_h\}_{h=1}^{\infty} = \mu$; we denote it as $\text{GDP}(\phi, \mu, G_0)$. Figure 4.1 depicts some realisations from a GDP with a standard Normal centre measure G_0 and different (constant) values for parameters ϕ and μ.

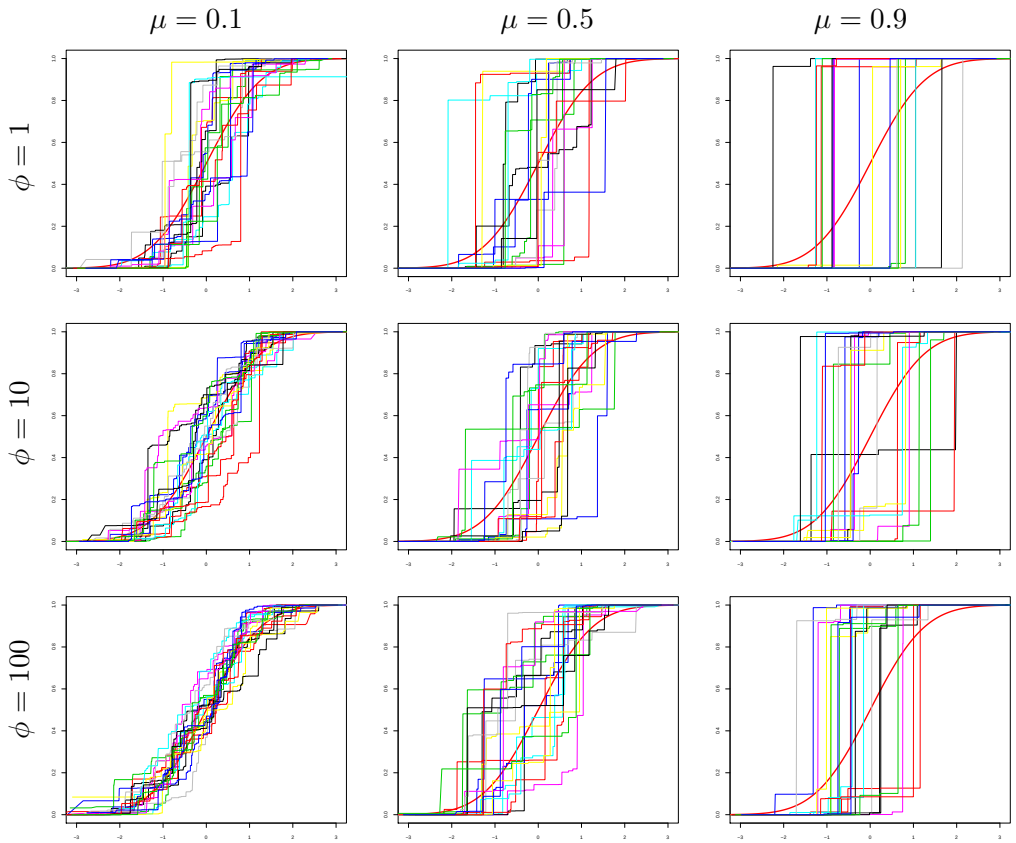


FIGURE 4.1: Cumulative distributions functions of samples obtained from a GDP with different combinations of the parameters and G_0 being a standard Normal distribution (red line in each panel).

As the name suggests, the GDP generalises the well-known DP (Ferguson [1973]), which can be specified by a GDP with $\{\phi_h = \mu_h^{-1}\}_{h=1}^{\infty}$ and $\{\mu_h\}_{h=1}^{\infty} = \mu$.

4.2.2 Moments

If $G \sim \text{GDP}(\boldsymbol{\phi}, \boldsymbol{\mu}, G_0)$, we have that

$$\mathbb{E}[G(A)] = G_0(A).$$

The variance of $G(A)$ is given by

$$\mathbb{V}[G(A)] = (1 - G_0(A))G_0(A)\mathbb{E}\left[\sum_{h=1}^{\infty} w_h^2\right]. \tag{4.2.3}$$

The expectation in the last equation cannot be computed explicitly, unless we consider the constant-parameter case GDP(ϕ, μ, G_0). With constant parameters, Hjort [2000] showed that

$$\mathbb{E}\left[\sum_{h=1}^{\infty} w_h^2\right] = \frac{\mathbb{E}[v^2]}{2\mathbb{E}[v] - \mathbb{E}[v^2]},$$

where v is a Beta random variable with parameters $(\phi\mu, \phi(1-\mu))$. Thus, $\mathbb{E}[v] = \mu$ and $\mathbb{E}[v^2] = \mu(1-\mu)/(\phi+1) + \mu^2$.

The third central moment of the random quantity $G(A)$ in the constant parameter case when $G \sim$ GDP(ϕ, μ, G_0) is

$$\mathbb{E}[(G(A) - G_0(A))^3] = G_0(A)(1 - G_0(A))(1 - 2G_0(A))\frac{\mathbb{E}[v^3]}{3(\mathbb{E}[v] - \mathbb{E}[v^2]) - \mathbb{E}[v^3]}.$$

As suggested by Hjort [2000], this allows to appreciate the extra flexibility introduced by having three parameters, *i.e.* ϕ, μ and G_0, compared to the DP case which employs only two, *i.e.* μ and G_0. In fact, after matching the first two moments of $G(A)$, setting a common G_0 and a specific value for ϕ, it can be shown that the ratio of the third central moments under a GDP and a DP ranges in a set of values that includes 1 (the case in which the GDP is a DP, *i.e.* $\phi = \mu^{-1}$), demonstrating the extra flexibility given by the additional parameter.

4.2.3 Distributional sampling properties

We now derive some properties of the GDP. Consider G as in (4.2.1). Because G is discrete, a sample $(\theta_1, \ldots, \theta_n)$ from G induces a random partition of the set $N = \{1, \ldots, n\}$ into k blocks with frequencies $(n_1, \ldots, n_k)$. We denote by $p(n_1, \ldots, n_k)$ the probability of any particular partition of $\{1, \ldots, n\}$, with k blocks and block-specific frequencies $(n_1, \ldots, n_k)$. In Definition 4 of Pitman [1995] this is referred to as the partially exchangeable partition probability function. Under the GDP with constant parameters, an application of Corollary 7 of

Pitman [1995] leads to an explicit expression for $p(n_1, \ldots, n_k)$, i.e.

$$\begin{aligned} & p(n_1, \ldots, n_k) \\ & = \mathbb{E}\left[\left(\prod_{j=1}^{k} w_j^{n_j - 1}\right) \prod_{r=1}^{k-1} \left(1 - \sum_{p=1}^{r} w_p\right)\right] \\ & = \frac{(\phi(1-\mu))^{k-1}}{(\phi)_{(n-1)}} \prod_{r=1}^{k-1} \frac{(\phi(1-\mu)+1)_{(\sum_{j=r+1}^{k} n_j - 1)}}{(\phi)_{(\sum_{j=r+1}^{k} n_j - 1)}} \prod_{j=1}^{k} (\phi\mu)_{(n_j - 1)}, \end{aligned} \tag{4.2.4}$$

where $(a)_{(b)} = a(a+1)\cdots(a+b-1)$ is the rising factorial number. We note that if we set $\phi = \mu^{-1}$, then the first product over i in (4.2.4) cancels, leading to the Ewens partition probability function (Ewens [1972]) induced by a sample drawn from a DP (Blackwell and MacQueen [1973]).

According to the theory of partially exchangeable random partitions developed in Pitman [1995], (4.2.4) characterises the predictive probabilities of the GDP with constant parameters. See Proposition 10 in Pitman [1995]. In particular, consider a sample of size n from a GDP(ϕ, μ, G_0) and assume that it induces a partition of $\{1, \ldots, n\}$ into k blocks, labelled by $\theta_1^*, \ldots, \theta_{K_n}^*$, with corresponding frequencies $(n_1, \ldots, n_k)$. Then

$$\Pr[\theta_{n+1} \notin \{\theta_1^*, \ldots, \theta_k^*\}] = \frac{\phi(1-\mu)}{\phi + n - 1} \prod_{i=1}^{k-1} \frac{\phi(1-\mu) + \sum_{j=i+1}^{k} n_j}{\phi + \sum_{j=i+1}^{k} n_j - 1} \tag{4.2.5}$$

and

$$\Pr[\theta_{n+1} = \theta_r^*] = \frac{\phi\mu + n_r - 1}{\phi + n - 1} \prod_{i=1}^{r} \frac{\phi(1-\mu) + \sum_{j=i+1}^{k} n_j}{\phi + \sum_{j=i+1}^{k} n_j - 1}. \tag{4.2.6}$$

for any $r = 1, \ldots, k$. Unfortunately, due to the cumbersome dependency on k and the frequencies n_i, the predictive probabilities (4.2.5) and (4.2.6) neither allow us to obtain moments of the distribution of k nor moments of the distribution of the number of blocks with certain frequencies.

We now determine the asymptotic behaviour of k as n grows. Using results in Karlin [1967], one can show that

$$\frac{k}{\log(n)} \to \frac{1}{\psi^{(0)}(\phi) - \psi^{(0)}(\phi(1-\mu))} \tag{4.2.7}$$

almost surely, as $n \to +\infty$. In (4.2.7), $\psi^{(0)}(x)$ denotes the polygamma function,

i.e., the first derivative of the logarithm of the Gamma function with respect to x. Details of the derivation of this result are in Appendix C.1. If $\phi = \mu^{-1}$, then the large n asymptotic result in (4.2.7) reduces to the well-known large n asymptotic behaviour of k under the assumption of the DP. Indeed, $\psi^{(0)}(1/\mu) - \psi^{(0)}(1/\mu - 1) = (1/\mu - 1)^{-1}$ and, hence, $k/\log(n) \to (1/\mu - 1)$ almost surely, as $n \to +\infty$.

The richer parameterisation of the GDP allows controlling simultaneously different important features of the partition (see Rodriguez and Dunson [2014]). For instance, fixing $\mathbb{E}(k)$, the parameters of GDP can control quantities such as the cardinality of the largest clusters, the average cluster size for different values or the number of clusters with cardinality equal one. This is in contrast with what happens using the DP, where the precision parameter governs all this quantities at once. Figure 4.2 and 4.3 show the extra flexibility of GDP compared to DP in terms of the probability of having clusters with cardinality equal one and expected size of the largest cluster for different values of the expected number of clusters. Similarly, the additional flexibility can be appreciated also by looking at the distribution of k under the GDP and the DP, after matching the first moment, as in Figure 4.4.

4.2.4 Truncated GDP

We next consider a modified version of (4.2.1) that includes a finite number H of atoms. We write:

$$G_H = \sum_{h=1}^{H} w_h \delta_{\theta_h}. \tag{4.2.8}$$

As in the infinite dimensional case, the locations are *iid* samples from G_0. The weights are constructed with the same stick-breaking procedure presented above, with the exception of the last weight, w_H, which is set to the value that makes the weights sum to 1. We denote this truncated process $\text{GDP}_H(\boldsymbol{\phi}, \boldsymbol{\mu}, G_0)$.

Truncated versions of the DP and other random probability measures have been employed in the literature, because they allow simplified computation when used as prior mixing distributions. Obviously, the use of a truncated process introduces an approximation error. The most common way to control this error was proposed by Ishwaran and James [2001] (Theorem 1) and has

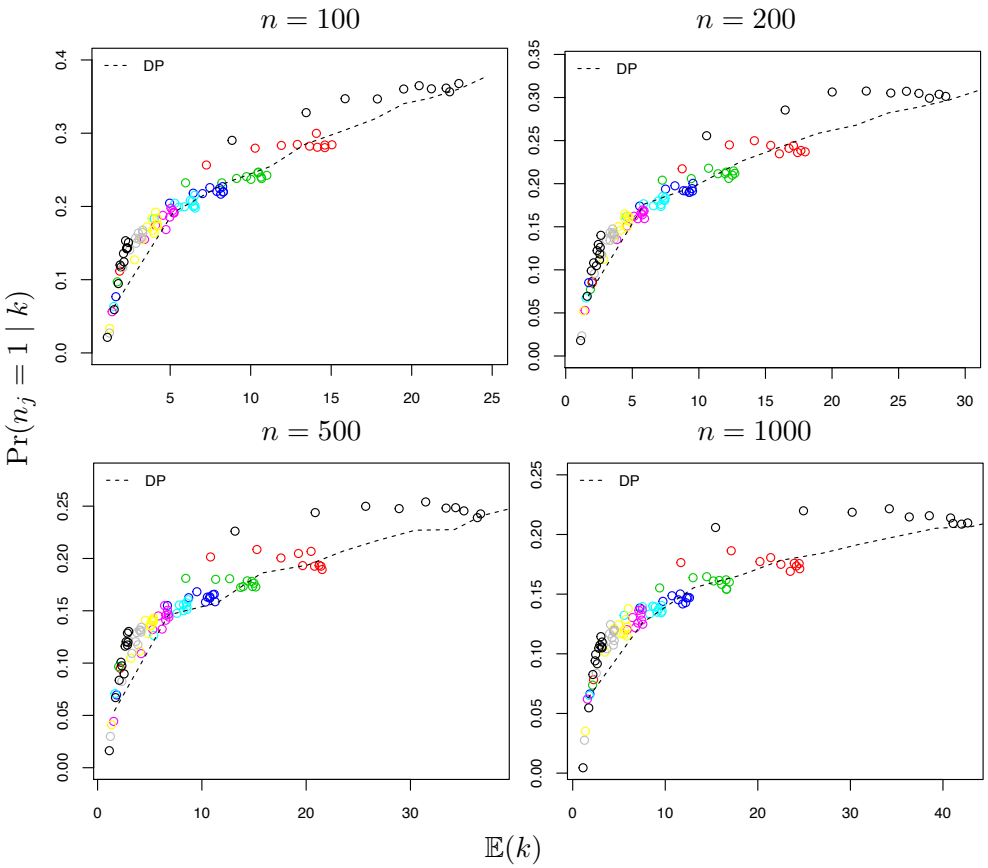


FIGURE 4.2: Probability of clusters with size one for different expected number of clusters in samples of size $n = \{100, 200, 500, 1000\}$ from a GDP. Dots are obtained by different combinations of the parameter values (colours correspond to different values of μ). Dashed lines are obtained using DP with different values of the precision parameters.

been adapted for many other processes. Consider the model

$$y_i \mid \theta_i \overset{ind}{\sim} f(y_i \mid \theta_i) \quad i = 1, \ldots, n$$

and compute the marginal density

$$k_H(\boldsymbol{y}) = \int \left(\prod_{i=1}^{n} \int f(y_i \mid \theta_i) \mathrm{d}G_H(\theta_i) \right) \mathrm{d}F(G_H),$$

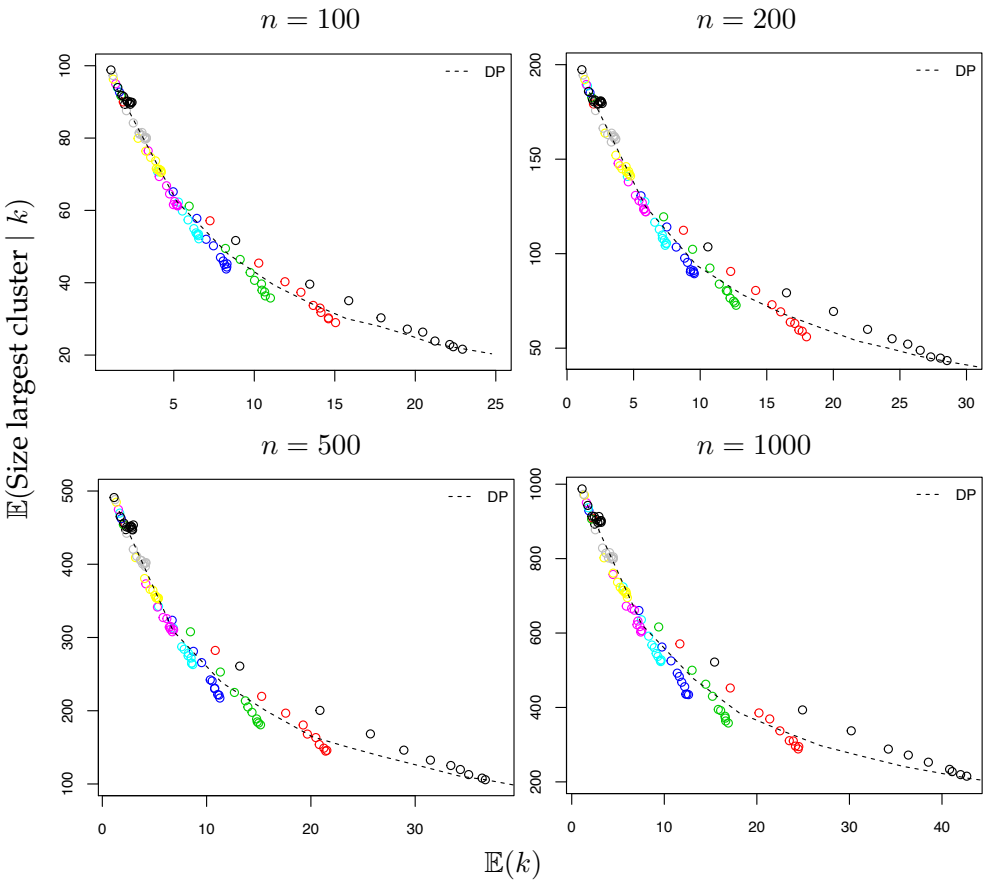


FIGURE 4.3: Expected size of the largest clusters for different expected number of clusters in samples of size $n = \{100, 200, 500, 1000\}$ from a GDP. Dots are obtained by different combinations of the parameter values (colours correspond to different values of μ). Dashed lines are obtained using DP with different values of the precision parameters.

where $F(\cdot)$ is the distribution of GDP$_H$. Then, the following result holds

$$\|k_H(\boldsymbol{y}) - k(\boldsymbol{y})\| \leq 4 \left(1 - \mathbb{E}\left[\left(\sum_{h=1}^{H-1} w_h \right)^n \right] \right),$$

where $\|\cdot\|$ denotes the L_1 norm, and $k(\boldsymbol{y})$ is the the marginal density computed under the GDP in (4.2.1). A proof of this result is in Ishwaran and James [2002]. The L_1 distance of the two marginal densities tends to 0 as H increases, as we might expect. The result above allows us to set an upper bound to the approximation error, leading to a specific number of components, H.

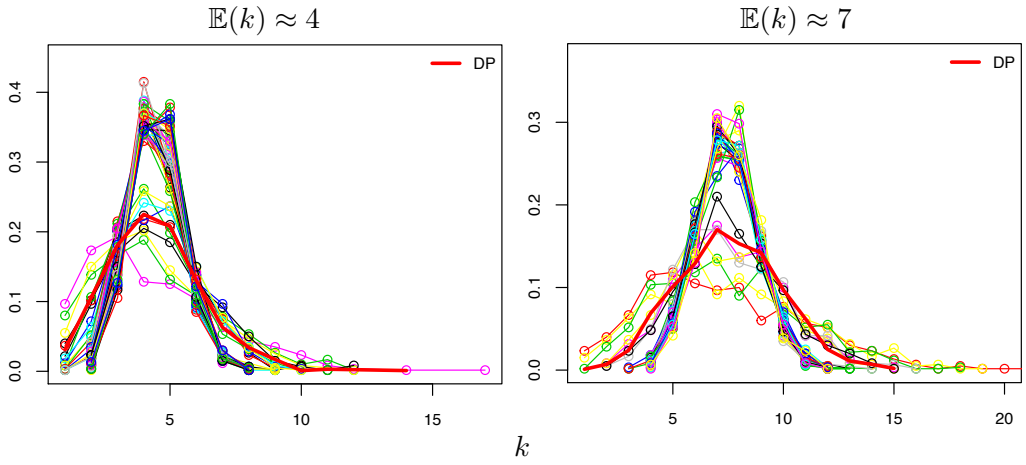


FIGURE 4.4: Distributions of the number of clusters k under DP (red line) and GDP (coloured lines), all having $\mathbb{E}(k) \approx 4$ (left panel) and $\mathbb{E}(k) \approx 7$ (right panel). The different colours correspond to different combinations of the parameters μ and ϕ.

In mixture models, it is common to truncate the mixing measure to a specific level for computational purposes. This is particularly true when the mixing measure is not distributed as a DP for which simple and efficient Gibbs samplers are available. When the mixing measure is a GDP (consequently, also a DP), however, the joint distribution of the truncated sequence of weights, namely $\boldsymbol{w} = (w_1, \ldots, w_H)$, has a known distribution. This distribution is the Generalised Dirichlet Distribution (GDD; Connor and Mosimann [1969], Ishwaran and Zarepour [2000]) and has the following density function,

$$
\begin{aligned}
&p(\boldsymbol{w} \mid \boldsymbol{\phi}, \boldsymbol{\mu}) \qquad (4.2.9)\\
&= \left(\prod_{h=1}^{H-1} \frac{\Gamma(\phi_h)}{\Gamma(\phi_h \mu_h)\Gamma(\phi_h(1-\mu_h))} \right) w_1^{\phi_1\mu_1 - 1} \cdots w_{H-1}^{\phi_{H-1}\mu_{H-1}-1} \\
&\quad \times w_H^{\phi_{H-1}(1-\mu_{H-1})-1} (1-\tilde{w}_1)^{\phi_1(1-\mu_1)-\phi_2} \cdots (1-\tilde{w}_{H-2})^{\phi_{H-2}(1-\mu_{H-2})-(\phi_{H-1})},
\end{aligned}
$$

with $\tilde{w}_k = \sum_{h=1}^{k} w_h$. This distribution is conjugate with the multinomial distribution, which simplifies the design of an algorithm for posterior sampling. The posterior distribution of $\boldsymbol{w}$ under the GDD can be sampled via stick-breaking with updated parameters. This leads to simple calculations in the case GDP_H with constant parameters when we want to sample from the posterior of ϕ and

μ, encouraging the use of parsimonious models. Discussion of this point continues in Section 4.4.

Other useful results may be obtained when considering a random truncation level for the GDP (see Muliere and Tardella [1998]). Consider the GDP with $(v_h)_{h\geq 1}$ being independent and identically distributed Beta random variables with parameters $(\phi\mu, \phi(1-\mu))$. We define the following random discrete probability measure

$$G_\varepsilon = \sum_{h=1}^{H_\varepsilon} w_h \delta_{\theta_h} + R_\varepsilon \delta_{\theta_0} \tag{4.2.10}$$

where the w_hs follow the usual stick-breaking construction, i.e., $w_h = v_h \prod_{r<h}(1-v_r)$;

$$H_\varepsilon = \inf\left\{ H \in \mathbb{N} : \sum_{h=1}^{H} w_h > 1-\varepsilon \right\};$$

$$R_\varepsilon = 1 - \sum_{h=1}^{H_\varepsilon} w_h = \prod_{h=1}^{H_\varepsilon}(1-v_h);$$

and θ_0 is a random variable with distribution G_0 that is independent of $(v_h)_{h\geq 1}$ and $(\theta_h)_{h\geq 1}$. Using similar arguments as for Lemma 2 in Muliere and Tardella [1998], one can verify that the truncated random discrete probability measure G_ε converges weakly to the random discrete probability measure G, with respect to the Prohorov metric, as $\varepsilon \to 0$. A fundamental role in the definition (4.2.10) is played by the truncation level H_ε. In fact, knowing the distribution of the random variable H_ε allows us to sample G_ε as close to G as we wish. In case G follows a GDP, it can be shown that H_ε converges in distribution to a Gaussian random variable, as ε tends to 0. This follows by a direct application of the Central Limit Theorem for renewal processes. An extended description of the latter result is in Appendix C.2.

4.3 Dependent GDP

Recalling the definition of the GDP in (4.2.1), a realisation from a GDP is an almost surely discrete probability measure. While the discreteness of G may seem unappealing, the use of such objects as random prior distributions is common

in BNP, such as when dealing with density estimation. The most famous example is the DPM, which results from convolving a density kernel parameterised by some quantity with a random prior distribution that is distributed according to a DP. One may adopt an equivalent strategy using a GDP. The resulting model is represented by the following hierarchy:

$$\begin{aligned} y_1, \ldots, y_n \mid G &\overset{iid}{\sim} \int f(y_i \mid \theta) \mathrm{d}G(\theta) \\ G \mid \boldsymbol{\phi}, \boldsymbol{\mu}, G_0 &\sim \text{GDP}(\boldsymbol{\phi}, \boldsymbol{\mu}, G_0), \end{aligned} \tag{4.3.1}$$

where the quantities $\boldsymbol{\phi} = \{\phi_h\}_{h=1}^{\infty}$ and $\boldsymbol{\mu} = \{\mu_h\}_{h=1}^{\infty}$ require the specification of suitable hyperprior distributions. According to the hierarchical formulation (4.3.1), the resulting sampling model is equivalent to an infinite mixture model with weights constructed as in (4.2.2).

Using a similar argument to the one presented in MacEachern [1999] and MacEachern [2000], the model in (4.3.1) can be enriched when covariates are available, assuming the observations are generated by a collection of infinite mixture models indexed by the covariate space and sharing hyperparameters. We achieve this result by modifying the GDP in such a way that the sequence $\{w_h\}_{h=1}^{\infty}$ is a function of the covariates. Given the parameterisation of the Beta distribution that we used as the prior for the sequence $\{v_h\}_{h=1}^{\infty}$, we can express the expectations of the latter quantities as functions of the covariates. We call the resulting process the Dependent GDP (DGDP). More specifically, for a generic point $x \in \mathcal{X}$, where $\mathcal{X}$ is the covariate space, a sample from DGDP is

$$G_x = \sum_{h=1}^{\infty} w_{h,x} \delta_{\theta_h},$$

where $\{\theta_h\}_{h=1}^{\infty} \overset{iid}{\sim} G_0$, $w_{1,x} = v_{1,x}$ and

$$w_{h,x} = v_{h,x} \prod_{r<h} (1 - v_{r,x}), \quad h = 2, 3, \ldots$$

Each $v_{h,x}$ is independently distributed following a Beta$(v_{h,x} \mid \phi_h \mu_h(x), \phi_h(1 - \mu_h(x)))$, where $\mu_h(\cdot)$ is a random mean function mapping into the set $(0, 1)$. Using the DGDP, the hierarchical model in (4.3.1) can be rewritten as

$$
\begin{aligned}
y_1, \ldots, y_n \mid G_{x_1}, \ldots, G_{x_n} &\overset{ind}{\sim} \int f(y_i \mid \theta) \mathrm{d}G_{x_i}(\theta) \\
G_{x_1}, \ldots, G_{x_n} \mid \boldsymbol{\phi}, \boldsymbol{\mu}(\cdot), G_0 &\overset{ind}{\sim} \text{DGDP}(\boldsymbol{\phi}, \boldsymbol{\mu}(x_i), G_0),
\end{aligned} \tag{4.3.2}
$$

where $\boldsymbol{\phi} = \{\phi_h\}_{h=1}^{\infty}$ and $\boldsymbol{\mu}(\cdot) = \{\mu_h(\cdot)\}_{h=1}^{\infty}$. In case $\mathcal{X}$ is a dense set, each y_i is associated with an individual random measure, *i.e.* G_{x_i}. If $\mathcal{X}$ is not dense, then there may be ties in the vector $(x_1, \ldots, x_n)$, which leads to ties in the corresponding random measures $(G_{x_1}, \ldots, G_{x_n})$, *i.e.* groups of observations having the same covariates share the same random measure. Furthermore, it is trivial to generalise the DGDP to the case with non-common location parameters, which can be obtained substituting G_0 with a stochastic process indexed by $x \in \mathcal{X}$.

A key aspect of the construction above is the infinite sequence of random functions $\{\mu_h(\cdot)\}_{h=1}^{\infty}$, which incorporates the dependence of the random measures on the covariates and the association between random measures indexed by different covariate values in $\mathcal{X}$. One way to evaluate the dependence between random distributions is by considering a measurable set $A \in \mathcal{A}$, two locations $x, x' \in \mathcal{X}$, and the covariance $\mathbb{C}[G_x(A), G_{x'}(A)]$. Considering location-specific mean functions and precisions, the covariance is equal to

$$
\mathbb{C}[G_x(A), G_{x'}(A)] = (1 - G_0(A))G_0(A)\mathbb{E}\left[\sum_{h=1}^{\infty} w_{h,x} w_{h,x'}\right],
$$

which converts to $\mathbb{V}[G_x(A)]$ when $x = x'$ (compare to (4.2.3)). Assuming a constant mean function and precision across locations simplifies the calculations, as it was the case with such an assumption for the moments of the GDP. In particular, considering $\{\mu_h(\cdot)\}_{h=1}^{\infty} = \mu(\cdot)$ and $\{\phi_h\}_{h=1}^{\infty} = \phi$ allows one to write

$$
\mathbb{E}\left[\sum_{h=1}^{\infty} w_{h,x} w_{h,x'}\right] = \frac{\mathbb{E}[v_x v_{x'}]}{\mathbb{E}[v_x] + \mathbb{E}[v_{x'}] - \mathbb{E}[v_x v_{x'}]},
$$

where v_x is a Beta random variable with parameters $(\phi\mu(x), \phi(1 - \mu(x)))$.

Hatjispyros et al. [2015] argued that another convenient way to learn about similarities among to dependent random measures is to look at the distance between the infinite mixture models induced by two random measures indexed

at two different locations in the covariate space. We apply this to two random measures distributed according to a DGDP. In particular, defining $f_x(y) = \int f(y \mid \theta)\mathrm{d}G_x(\theta)$ and $f_{x'}(y) = \int f(y \mid \theta)\mathrm{d}G_{x'}(\theta)$ to be two sampling mixture models indexed at $x, x' \in \mathcal{X}$, respectively, with $G_\cdot \sim \text{DGDP}(\boldsymbol{\phi}, \boldsymbol{\mu}(\cdot))$, the expected L_2-distance (denoted $\|\cdot\|_2$) between $f_x(y)$ and $f_{x'}(y)$ is given by

$$\mathbb{E}[\|f_x(y) - f_{x'}(y)\|_2] = (a-b)\mathbb{E}\left[\sum_{h=1}^{\infty}(w_{h,x} - w_{h,x'})^2\right],$$

where $a = \mathbb{E}\left[\int f(y \mid \theta_h)^2 dy\right]$ and $b = \mathbb{E}\left[\int f(y \mid \theta_h) f(y \mid \theta_j) dy\right]$. The latter equation shows that using covariate-dependent weights allows one to set mixture models to be arbitrarily close, despite the fact that the mixture models share common locations. This could be an argument in favour of using a stochastic process with only the weights indexed by the covariates.

Using the same approach we employed for calculating the moments of the GDP and assuming $\{\phi_h\}_{h=1}^{\infty} = \phi$ and $\{\mu_h(\cdot)\}_{h=1}^{\infty} = \mu(\cdot)$, we can write

$$\mathbb{E}\left[\sum_{h=1}^{\infty}(w_{h,x} - w_{h,x'})^2\right] = \frac{\mathbb{E}[v_x^2]}{2\mathbb{E}[v_x] - \mathbb{E}[v_x^2]} + \frac{\mathbb{E}[v_{x'}^2]}{2\mathbb{E}[v_{x'}] - \mathbb{E}[v_{x'}^2]} - \frac{2\mathbb{E}[v_x v_{x'}]}{\mathbb{E}[v_x] + \mathbb{E}[v_{x'}] - \mathbb{E}[v_x v_{x'}]}. \tag{4.3.3}$$

We can derive expressions for the latter expectations for different choices of $\mu(\cdot)$ and ϕ. We can gain some insight into the distance measure represented by the previous equation by assuming

$$\mu(x) = \frac{\exp(x\mu)}{1 + \exp(x\mu)},$$

which is the usual link function for logistic regression. We consider $x = \{0, 1\}$ for simplicity and evaluate the expectation in (4.3.3) for different values of μ and ϕ. The results are shown in Figure 4.5. We note that as ϕ tends to infinity, the DGDP defined for this example becomes the Logit stick-breaking process introduced in Ren et al. [2011]. Similarly, assuming $\mu(\cdot)$ to be a Probit regression and letting ϕ tend to infinity leads the DGDP to become the Probit stick-breaking process introduced in Rodriguez and Dunson [2011].

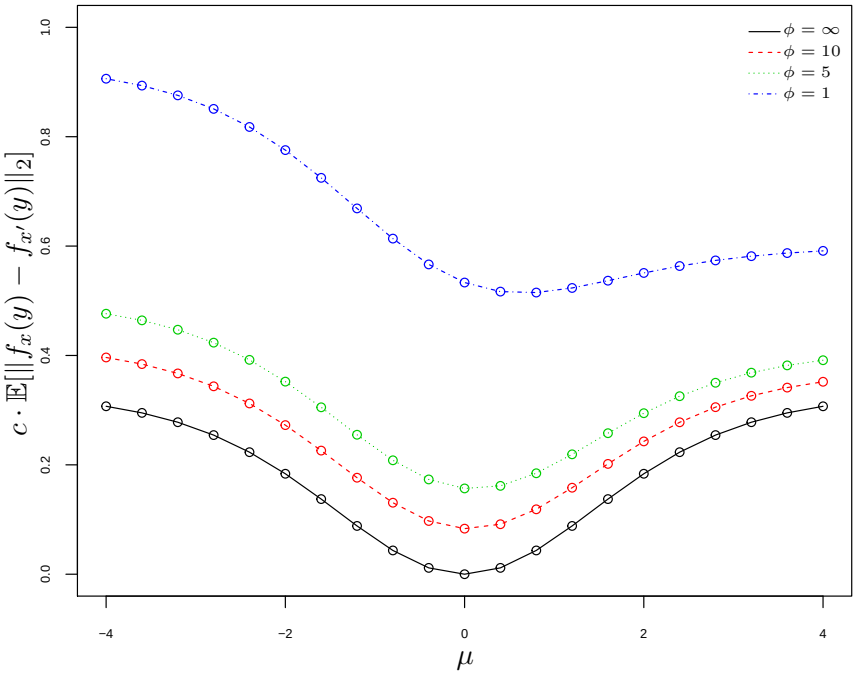


FIGURE 4.5: Expected L_2-distance between two mixture models corresponding to $x = \{0, 1\}$ generated according to a DGDP (up to a constant dependent on the center measure, c) for different values of ϕ and μ, the latter being the parameter of $\mu(x)$, a Logistic regression without intercept.

4.4 Posterior Inference via MCMC

In this section, we discuss posterior inference under the DGDP mixture model. We use the blocked Gibbs sampler presented by Ishwaran and James [2001], which provides an approximate inference based on the process in (4.2.8). An alternative approach, based on the slice sampler described by Walker [2007], is presented in Appendix C.3. In practice, the latter is equivalent to blocked Gibbs sampler but involves a random truncation of the GDP. Another possibility for carrying out posterior inference for the DGDP is via the retrospective sampler of Papaspiliopoulos and Roberts [2008]. Additionally, the result in Equation (4.2.4) makes it possible to design a Gibbs algorithm for sampling the underlying partition of the observations, if the DGDP has non-common locations across different levels of the covariates.

Consider the following parameterisation of the DGDP mixture model in (4.3.2)

$$\begin{aligned} y_i \mid \theta_i &\sim f(y_i \mid \theta_i) \\ \theta_i \mid G_{x_i} &\sim G_{x_i} = \sum_{h=1}^{\infty} w_{h,x} \delta_{\theta_h} \\ G_{x_i} \mid \phi, \mu(\cdot), G_0 &\sim \text{DGDP}(\phi, \mu(x_i), G_0) \\ \phi, \mu(\cdot), G_0 &\sim g(\phi) q(\mu(\cdot)) r(G_0), \end{aligned}$$

where, for $i = 1, \ldots, n$, y_i represents the response variable, x_i is a covariate, and θ_i is the individual-specific parameter. We also introduce the latent indicator s_i that takes positive integer values, with $s_i = h$ indicating $\theta_i = \theta_h$.

The main assumption for performing inference using the blocked Gibbs algorithm is that

$$G_x \approx G_{x,H} = \sum_{h=1}^{H} w_{h,x} \delta_{\theta_h},$$

where $G_{x,H}$ is constructed as the truncated GDP for each observed level x. Once the approximating process is assumed and an appropriate truncation level H is set, the following steps are necessary for posterior inference.

i) *Resample s_i, for $i = 1, \ldots, n$.* After the approximation of the process to a finite number of atoms, resampling is based on

$$\Pr[s_i = h \mid \cdot] \propto w_{h,x_i} f(y_i \mid \theta_h), \quad \text{for } h = 1, \ldots, H.$$

ii) *Resample θ_h, for $h = 1, \ldots, H$.* Given the latent assignment variables $s_1, \ldots, s_n$, we have a number of independent models characterised by model-specific parameters. Specifically, the full conditional densities can be written as

$$p(\theta_h \mid \cdot) \propto g_0(\theta_h) \prod_{\{i:s_i=h\}} f(y_i \mid \theta_h), \quad \text{for } h = 1, \ldots, H,$$

where $g_0(\cdot)$ is the density function of G_0. If the set $\{i : s_i = h\}$ is empty for any h, sample from G_0 in this step.

iii) *Resample $w_{h,x}$, for $h = 1, \ldots, H$ and all observed x in the data set.* This step

comes from the fact that for a specific covariate value x, the joint distribution $w_{1,x}, \ldots, w_{H,x}$ follows the GDD, which is conjugate to the multinomial distribution and can be sampled using a stick-breaking procedure with

$$v_{h,x} \mid \cdot \sim \text{Beta}\left(v_{h,x} \mid \phi\mu(x) + \sum_{\{i:x_i=x\}} (\mathbb{I}(s_i = h)), \right.$$
$$\left. \phi(1-\mu(x)) + \sum_{\{i:x_i=x\}} (\mathbb{I}(s_i > h))\right),$$

for $h = 1, \ldots, H$.

iv) *Resample* $\phi, \mu(\cdot)$. We consider resampling ϕ and $\mu(\cdot)$ when these are constant across locations. The extension to location-specific parameters is straightforward. This step of the algorithm depends largely on the type of problem considered, particularly for $\mu(\cdot)$. The step involves sampling from

$$p(\phi, \mu(\cdot) \mid \cdot) \propto \prod_x p(\boldsymbol{w}_x \mid \phi, \mu(\cdot)) p(\phi, \mu(\cdot)),$$

where the density of the weights associated with different locations comes from the GDD in equation (4.2.9), modified for constant parameters.

v) *Resample* G_0. This step depends on the nature of the problem under consideration. It is only needed when G_0 is not known or, more commonly, when its parameters are unknown.

4.5 Applications

4.5.1 Acute Lymphoblastic Leukaemia and Dyslipidemia

Childhood Acute Lymphoblastic Leukaemia (ALL) is a cancer that affects the production of blood cells. The bone marrow produces an excess of lymphoblasts, which are immature white blood cells. Children affected by ALL are currently treated with combinations of chemotherapies, and the drug regimens include a class of steroids called glucocorticoids, such as dexamethasone. While this therapy has improved cure rates for patients, it is associated with a number of

side effects. One adverse side effect is osteonecrosis, a disease that is associated with reduced blood flow to bones and joints, leading to bone cell death and possible fractures. The pathogenesis of osteonecrosis and its relationship with treatments for childhood ALL are described by Kawedia et al. [2011]. In particular, poor metabolism of the glucocorticoids included in the treatment of ALL may lead to this disease. The association between these steroids and the risk of osteonecrosis is thought to be through the glucocorticoid's effect on lipid levels. The effect leads to an increase in the size of lipocytes (fat cells) and subsequent marrow ischemia and apoptosis. These complications often result in bone necrosis, pain, and inability to use the joint.

Recent studies have shown that other drugs that are part of ALL therapy, such as asparaginase, may lead to osteonecrosis by a different mechanism than that of steroids. The objective of this analysis is to model the change of lipid measures over time (in particular triglycerides) during ALL therapy as a function of a biomarker of the pharmacological activity of asparaginase. We use albumin level as this biomarker, since higher asparaginase activity leads to lower albumin levels.

This study includes $n = 198$ ALL patients who have been classified by clinicians into two risk groups based on expected outcome. Children in the low-risk group (LR) have a better chance of cure than children in the standard/high-risk group (SHR). Factors at baseline that determine a patient's risk group are age (younger children tend to have better outcomes than older children), initial white blood cell (WBC) count (very high counts require more intensive treatment), sex (females have a somewhat greater chance of cure than males), race (Caucasian children tend to have better outcomes), and subtype of the disease, to name a few. The data set includes 93 ALL patients in the SHR group and 105 patients in the LR group.

Because the LR group tends to have a better prognosis than the SHR group, the treatment regimens for the risk groups differ. The SHR group receives more intensive therapy than the LR group. The different treatment regimens include different doses and schedules of dexamethasone and asparaginase, the two drugs that are associated with risk of osteonecrosis. The analysis considers each patient's measurements of triglycerides (mg/dL) and albumin (g/dL) from blood samples at baseline ($t = 0$), week 7 ($t = 7$), week 8 ($t = 8$), and week 12 ($t = T = 12$) of treatment. Patients received both drugs at the start of weeks

7 and 8 but not at baseline or week 12.

We denote the $\log_2$ transformation of the triglyceride level for the i-th patient at time t by $y_{i,t}$. We assume the following model for the triglyceride trajectories, $\boldsymbol{y}_i = (y_{i,0}, \ldots, y_{i,T})$,

$$\boldsymbol{y}_i \mid \boldsymbol{B}_i, \Omega_i \sim \text{MNormal}_T \left(\begin{array}{c|c} y_{i,0} & \boldsymbol{x}_{i,0}\boldsymbol{\beta}_{i,0} \\ \vdots & \vdots \quad , \Omega_i \\ y_{i,T} & \boldsymbol{x}_{i,T}\boldsymbol{\beta}_{i,T} \end{array} \right),$$

where $\text{MNormal}_T(\cdot \mid \cdot, \cdot)$ denotes the $T-$dimensional Normal distribution, $\boldsymbol{B}_i = (\boldsymbol{\beta}_{i,0}, \ldots, \boldsymbol{\beta}_{i,T})$ is a matrix of coefficients, and $\boldsymbol{X}_i = (\boldsymbol{x}_{i,0}, \ldots, \boldsymbol{x}_{i,T})$ is a matrix of time-dependent covariates that includes the measured albumin levels at different times, along with an intercept. Ω_i is the variance-covariance matrix, and we assume

$$\Omega_i = \sigma_i^2 H(\rho_i).$$

As in Quintana et al. [2015a], we specify the matrix $H(\rho_i)$ such that the covariance $\mathbb{C}[y_{i,t}, y_{i,s}] = \sigma_i^2 \rho_i^{|t-s|}$. This choice induces a correlation structure among the elements in $\boldsymbol{y}_i$ that is equivalent to one implied by an autoregressive model with time lag of one.

We account for possible heterogeneity between patients by assuming *a priori* that the trajectories come from a mixture of distributions. We also assume different but correlated mixing measures for patients belonging to the two risk groups (LR and SHR). This assumption allows us to control for information implied by being in a certain risk group, making more realistic the linear dependence of the triglyceride values on the albumin levels. The latter argument is similar to one described in Papageorgiou et al. [2015]. We formalise this assumption through the following hierarchical structure for the patient-specific parameters.

$$\begin{aligned} (\boldsymbol{B}_i, \sigma_i^2, \rho_i) \mid G_{\boldsymbol{z}_i} &\sim G_{\boldsymbol{z}_i} \\ G_{\boldsymbol{z}_i} \mid \phi, \mu(\cdot), G_0 &\sim \text{DGDP}(\phi, \mu(\boldsymbol{z}_i), G_0), \end{aligned}$$

where $\boldsymbol{z}_i = (1, z_i)$ and z_i is equal to 1 if the i-th patient belongs to the LR and 0 if SHR. We assume that the hypermean for the stick-breaking sequence is a

Logistic regression on z_i,

$$\mu(\boldsymbol{z}_i) = \frac{\exp(\boldsymbol{z}_i\boldsymbol{\eta})}{1+\exp(\boldsymbol{z}_i\boldsymbol{\eta})}.$$

The regression parameters in the hypermean function are multivariate Normal,

$$\boldsymbol{\eta} \sim \text{MNormal}_2(\boldsymbol{\eta} \mid \mathbf{0}_2, \sigma_\eta^2 I_2),$$

where $\mathbf{0}_N$ is a N-dimensional vector of zeros, and I_N denotes the identity matrix of dimension $N \times N$. Finally, we specify a gamma hyperprior distribution for the precision of the DGDP

$$\phi \sim \text{Gamma}(\phi \mid a_\phi, b_\phi)$$

and the following for the prior mean measure of the process,

$$G_0 = \text{U}(\sigma \mid a_\sigma, b_\sigma)\text{U}(\rho \mid 0, 1) \prod_{t=0}^{T} \text{MNormal}_T(\boldsymbol{\beta}_t \mid \mathbf{0}_2, \sigma_\beta^2 I_2).$$

We fix σ_β^2 and σ_γ^2 to 100; set σ_η^2, a_ϕ, and b_ϕ to 1; and let a_σ and b_σ equal 0 and 5, respectively. We run the blocked Gibbs sampler discussed in Section 4.4 with truncation level $H = 30$ and 50 000 iterations after a burnin period of 30 000, saving every tenth sample.

The expectations of the posterior predictive distributions for the triglycerides are depicted in Figure 4.6, showing different trajectories corresponding to different risk groups and different values of albumin at baseline and weeks 7, 8, and 12. Overall, the SHR-specific trajectories are higher than those corresponding to the LR patients with the same albumin levels. The predicted triglyceride values at each time point, as a function of albumin, indicate a negative relationship between albumin and triglyceride levels for both risk groups. This relationship suggests that a reduction in the asparaginase activity, which is in turn related to an increase in albumin level, leads to a reduction in triglycerides in both risk groups, with a stronger effect among the SHR patients.

The largest difference in the values of the triglyceride trajectories is observed between the $t = 7$ and $t = 8$, when patients receive both the glucocorticoid and asparaginase. Figure 4.7 shows marginal density estimates of the distributions

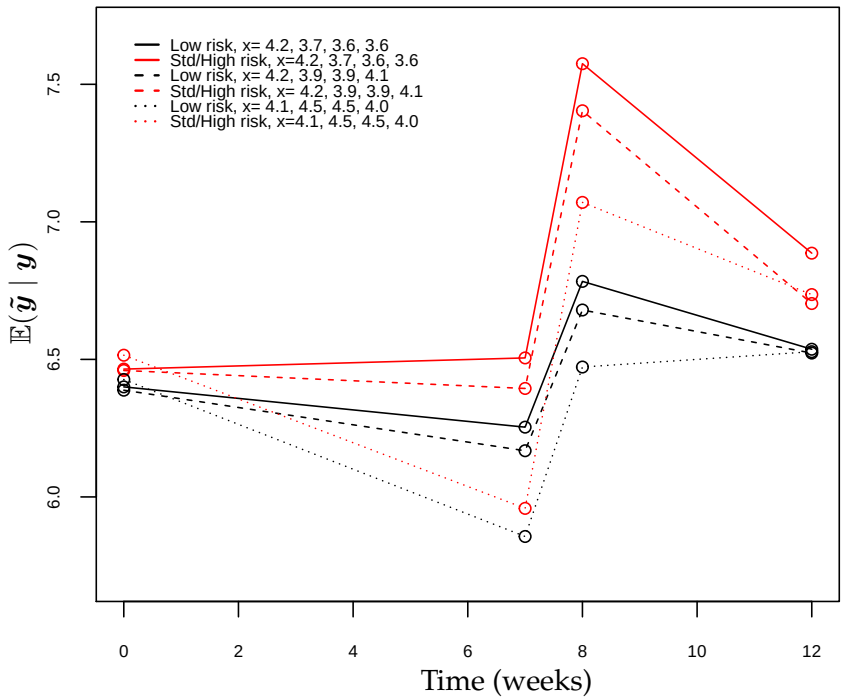


FIGURE 4.6: Posterior predictive mean triglycerides for the two risk groups and different albumin values at baseline and weeks 7, 8, and 12.

of triglycerides at week 8, where the different type of lines and colours correspond to the legend in Figure 4.6. Figure 4.7 shows the posterior predictive densities for week 8 triglycerides, which are mixtures with weights that vary across risk group. The curves corresponding to the SHR group assign high probability to a mixture component located around 9 $\log_2$(mg/dl). This component is centred at a relatively high value and leads to the differences seen in the expectations observed in Figure 4.6. The other apparent mixture component has a roughly equivalent location for both risk groups and is centred around 6.5 $\log_2$(mg/dl). This observation suggests that the risk group-specific differences in triglyceride values evident at week 8 are driven by a subset of the SHR patients, whereas the other SHR patients show similar triglyceride values as the LR risk group. An equivalent, although less evident, pattern can be seen in the marginal density distributions for the triglycerides at weeks $t = 7$ and 12.

In Figure 4.8, we show the posterior marginal densities for the effects of albumin on triglycerides at each of the four time points under analysis (i.e., the

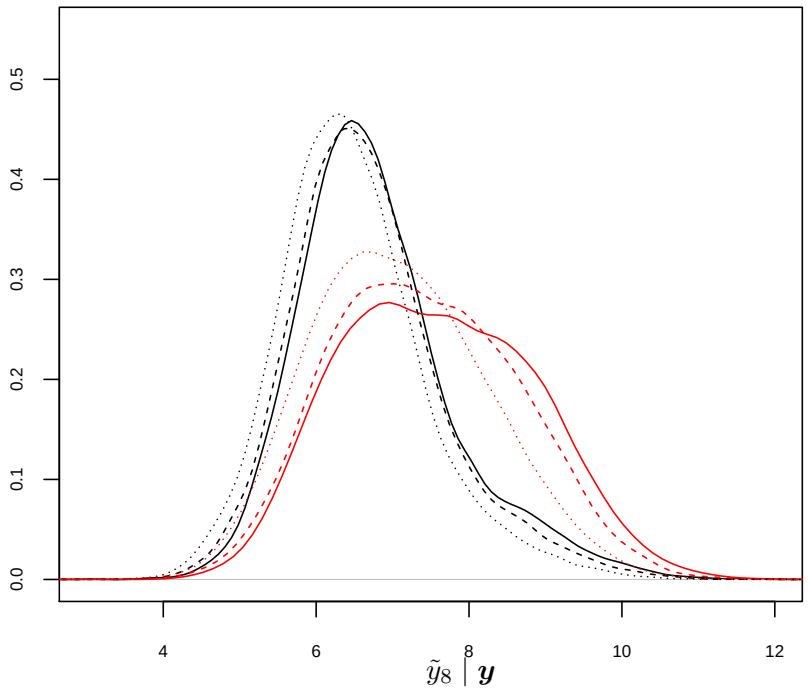


FIGURE 4.7: Marginal posterior predictive densities of triglycerides at week 8. The different lines correspond to the legend in Figure 4.6.

time and group-specific regression coefficients). While the relationship between albumin and triglycerides at baseline (top-left panel) seems similar for the risk groups, the densities diverge at $t = 7$. That is, after the start of treatment, a group of patients (mostly SHR patients) shows a stronger negative relation between albumin and triglycerides, while a number of other patients (mostly belonging to the LR group) exhibit a weaker negative effect of albumin on triglycerides. This pattern also appears at week 8, although the albumin effect is less negative than at week 7 for the majority of LR patients. At week 12, the majority of the mass corresponding to the albumin effect on triglycerides among the LR patients is centred a little to the right of zero. The effect for the SHR group at $t = 12$, however, remains bimodal, with the left-hand component remaining strongly negative and the right-hand component looking much like the density corresponding to the majority of the LR patients. These observations suggest that a subset of the SHR patients may be at higher risk of osteonecrosis, perhaps because of greater sensitivity to the drugs.

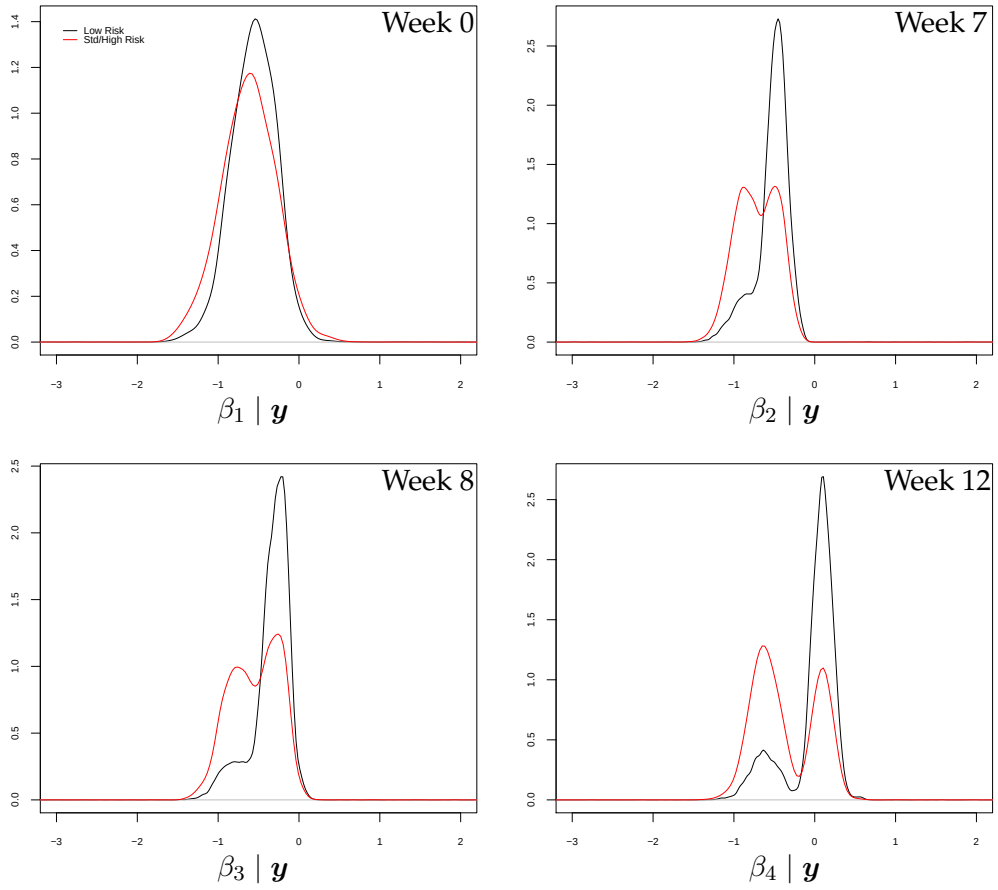


FIGURE 4.8: Posterior densities of the regression coefficients related to albumin at times $t = 0, 7, 8$, and 12. Red lines indicate the posterior densities for SHR patients, and black lines the posterior densities for LR patients.

We compare the performance of the DGDP mixture model described above for the data analysis in this section with those obtainable with related and more standard alternatives. The first competitor model is a parametric mixed-effect model, which is specified by a sampling model equivalent to the one in (4.5.1) where the individual effects, namely $\boldsymbol{B}_i$ and Ω_i, have been replaced with parameters shared by all patients. In addition, information regarding the risk groups is included via random effects within the model of the mean. Borrowing strength across groups is favoured by a suitable hierarchical structure. The second competitor is a DDP mixture model which is specified as the DGDP mixture model above, except for the distribution of the mixing weights which follows marginally (for each value of $\boldsymbol{z}$) a DP with precision parameter equal

$\alpha(\boldsymbol{z}) = \exp(\boldsymbol{z}\boldsymbol{\eta})$.

The competitor models are assessed using Pseudo Bayes Factor (PSBF, Geisser and Eddy [1979], Gelfand and Dey [1994]). When two models, M_l and M_r, are considered, the PSBF is defined as

$$\text{PSBF}(M_l, M_r) = \frac{\prod_{i=1}^{n} p_{M_l}(\boldsymbol{y}_i \mid \boldsymbol{Y}_{-i})}{\prod_{i=1}^{n} p_{M_r}(\boldsymbol{y}_i \mid \boldsymbol{Y}_{-i})},$$

where $p_{M_l}(\boldsymbol{y}_i \mid \boldsymbol{Y}_{-i})$ and $p_{M_r}(\boldsymbol{y}_i \mid \boldsymbol{Y}_{-i})$ are posterior predictive densities of $\boldsymbol{y}_i$ under M_l and M_r, including the information of $\boldsymbol{Y}_{-i}$, the matrix containing all observations except for $\boldsymbol{y}_i$. All these posterior predictive densities, often called conditional predictive ordinates, have been approximated using MCMC samples. PSBF is preferred to the common Bayes factor or posterior Bayes factor (Aitkin [1991]) because less sensitive to prior choices and more simple from a computational point of view.

The results of the comparison show evidence in favour of proposed DGDP mixture model against the two competitors models. In particular, log(PSBF) of the DGDP mixture model and the mixed-effect model is equal to 39.16. Instead, the same quantity calculated using the DGDP and DDP mixture models is equal to 10.85.

4.5.2 Ofsted Evaluation of London Primary Schools

In the UK, state maintained schools are periodically evaluated by the Office for Standards in Education, Children's Services and Skills (Ofsted), using publicly available data about each school as well as proper inspections in order to gather evidence about the quality of the schools. Ofsted assesses the overall quality of the teaching, achievements of the pupils and, more generally, the level of each school's environment. The main summary result of the Ofsted analysis is the *overall effectiveness*, which is a four-point ordinal scale where 1 indicates an outstanding quality, 2 a good quality, 3 the need of improvements and 4 inadequacy. More details about Ofsted, the inspections and evaluation methods can be found at `www.gov.uk/governemnt/publications/school-inspection-handbook-from-september-2015`.

In this analysis we focus on London primary schools and we control for

possible indicators of a school quality among the most common characteristics of the schools. These include the percentage of missed sessions (authorised and non authorised absences), the percentage of pupils with a Special Education Needs (SEN) statement or on School Action Plus (SAP), the percentage of pupils with English not as first language, the percentage of pupils eligible for free school meals, the pupils-teachers ratio and religious denomination. All data can be downloaded from `www.gov.uk`.

The most intuitive way to model similar observations includes ordinal regression of the Ofsted scores over the school features. This allows one to determine which among the school characteristics is a predictor of the school quality and the marginal impact of each predictor. Similar regression models involve the use of continuous latent variables and cut-offs that relate the latent variable to a discrete random variable and give the probability of each score (see Albert and Chib [1993] and Chib and Greenberg [1998]). The typical choice for the distribution of the latent variable is the Gaussian that often incorporates a regression model on the mean when covariates are available. However, the Normal distribution can be overly restrictive in many situations and for this reason mixture models are often employed (examples are Chib and Greenberg [2010],Gill and Casella [2009] and Kottas et al. [2005]). We follow the idea of employing a mixture model for the latent variable. More specifically, in our analysis we would like to model the observations as coming from Normal linear regressions that have random effects and variances distributed according to nonparametric mixtures that vary across boroughs, allowing for spatial dependence.

With these motivations we introduce a spatial version of DGDP and we use it as a prior for the intercept and variance of a Normal density kernel. We also include a regression model of the school-specific characteristics on the mean of the Normal kernel. This is a latent variable representing the school quality which, when discretised according to fixed thresholds, gives the likelihood of the Ofsted evaluations. Let us define the vector $y_1, \ldots, y_n$ containing ordered univariate observations corresponding to the Ofsted evaluation for each of the schools. Obviously, $y_i \in \{1, \ldots, C\}$, for $i = 1, \ldots, n$, with $\{1, \ldots, C\}$ being the

ordered set of possible scores. As mentioned previously, we model similar observations using a set of latent variables, $z_1, \ldots, z_n$, such that:

$$\begin{aligned} y_i = 1 \quad &\text{if} \quad z_i \leq \gamma_1, \\ y_i = c \quad &\text{if} \quad \gamma_{k-1} < z_i \leq \gamma_c, \text{ for } c = 2, \ldots, C-1 \\ y_i = C \quad &\text{if} \quad z_i > \gamma_{C-1}, \end{aligned}$$

with $\{\gamma_1, \ldots, \gamma_{C-1}\}$ being a set of cutoffs. In order to achieve our desiderata, we assume the latent variables to be distributed as a DGDP mixture of regressions as follows

$$\begin{aligned} z_i \mid G_{l_i} &\sim \int \text{Normal}(z_i \mid \beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}', \sigma^2) dG_{l_i}(\beta_0, \sigma^2) \\ G_{l_i} \mid \phi, \mu(\cdot), G_0 &\sim \text{DGDP}(\phi, \mu(l_i), G_0), \end{aligned}$$

where $l_i \in \{1, \ldots, L\}$ indicates the borough of the $i-$th school, while $\boldsymbol{x}_i$ is the collection of individual features for the same school. A similar idea for modelling the latent variable with covariate-dependent weights has been proposed by DeYoreo and Kottas [2014], where in order to include covariate information in the mixture weights the authors model jointly the covariates and the latent variable and derive the conditional distribution of the latent variable given the covariates. The idea presented here, however, is closer to the one in Gill and Casella [2009], but we specify a nonparametric prior also for the variance of the latent variable to allow more flexibility in the latent variable distribution.

In this setting, the choice of fixing the cut-offs $\gamma_1, \ldots, \gamma_{K-1}$ to arbitrary values allows identification of all parameters and does not prevent the approximation of all possible distributions of the latent variable. Details about both identifiability (in the sense of the likelihood) and the flexibility of mixture of latent distributions with fixed cutoffs can be found in DeYoreo and Kottas [2014] and Kottas et al. [2005]. In this chapter the robustness of the model to the choice of the cut-offs have been studied without finding relevant differences for the tested choices of values.

$\phi, \mu(\cdot), G_0$ represent the parameters of the DGDP. First, we assume the precision parameter of the DGDP follows

$$\phi \sim \text{Gamma}(\phi \mid a_\phi, b_\phi).$$

A crucial role is played by the $\mu(\cdot)$. This has to incorporate the spatial information within the process. We assume

$$\mu(l_i) = \frac{\exp(\alpha_1 + \alpha_2 \eta_{l_i})}{1 + \exp(\alpha_1 + \alpha_2 \eta_{l_i})},$$

which is the usual logistic regression function. We model $\boldsymbol{\eta} = (\eta_1 \ldots, \eta_L)$

$$\boldsymbol{\eta} \sim \text{MNormal}_L(\boldsymbol{\eta} \mid \mathbf{0}_L, \boldsymbol{M}_\eta^{-1}).$$

In the latter equation, $\mathbf{0}_L$ is a vector with components all equal to 0 and length L and $\boldsymbol{M}_\eta$ is a precision matrix that governs the spatial correlation. We assume $\boldsymbol{M}_\eta = \tau \boldsymbol{A} + \boldsymbol{I}_L$, with τ being a positive scalar, $\boldsymbol{I}_L$ is an identity matrix with dimension $L \times L$ and $\boldsymbol{A}$ is an $L \times L$ matrix with entries $a_{l,l}$ equal to the number of neighbours of borough l and $a_{l,l'}$ equal to 0 if boroughs l and l' are not neighbours and -1 otherwise. The idea of using a Gaussian Markov Random Field (Fernández and Green [2002]) follows a similar objective to the one described in Papageorgiou et al. [2015].

Finally, we assume the center probability measure, G_0, to be the product of a Normal distribution with mean 0 and variance $\sigma^2_{\beta_0}$ and a Gamma distribution with parameters equal to a_{σ^2} and b_{σ^2}. We also assume a Normal distribution with 0 mean and σ^2_β variance for each regression coefficient in $\boldsymbol{\beta}$.

For the data analysis, we set the following hyperparameters: $a_\phi = 1, b_\phi = 1, \sigma^2_\beta = \sigma^2_{\beta_0} = 100, a_{\sigma^2} = b_{\sigma^2} = 0.1$. We select $\alpha_1 = -1, \alpha_2 = 2.5$ and $\tau = 2$ from a grid of values minimising the misclassification prediction error. Although there are four possible Ofsted grades, we aggregate the worst two levels, so that in our example $C = 3$. We then fix $\gamma_1 = -1$ and $\gamma_2 = 1$. The number of school-specific features is $D = 6$ and the number of schools that have been considered in our analysis is $n = 1043$. The covariates entering the regression model of the latent variable have been centred (except for the indicator of religious denomination which is binary, equal to 1 when a religious denomination is present). We run the Blocked Gibbs sampler algorithm discussed in Section 4.4, using truncation level $H = 20$ and 100 000 iterations, after a burnin period of 10 000. We save every tenth sample from the MCMC.

We begin the analysis of the results by looking at the posterior distribution of β_0, which is the intercept of the latent variable distribution. This distribution,

depicted in Figure 4.9, is a collection of mixture distributions, whose weights are indexed to the 32 boroughs of London and are correlated. The intercept,

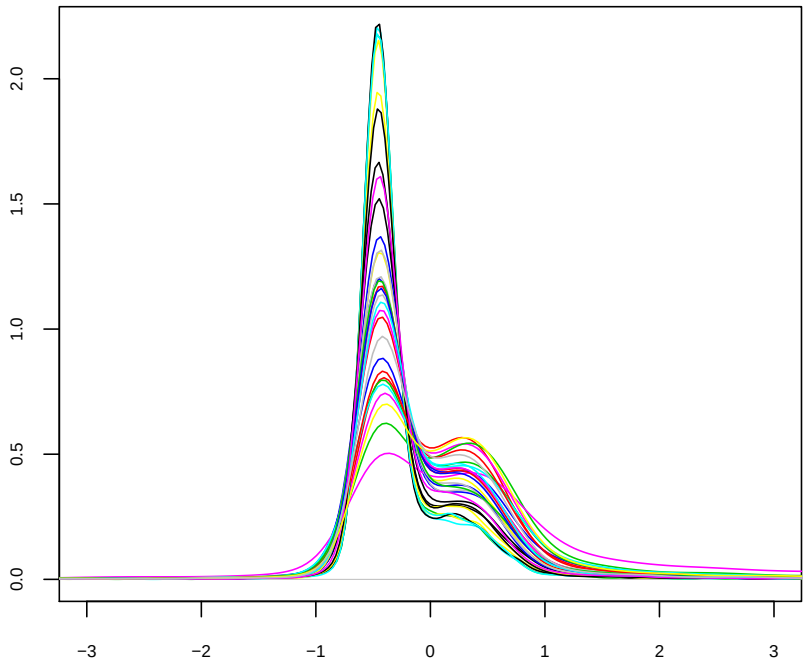


FIGURE 4.9: Posterior density of the β_0. Different colours correspond to different borough locations.

together with the variance of the latent variable (which has been modelled similarly by a collection of correlated mixture distributions), characterises the distribution of the latent variable and consequently the baseline probabilities of the different Ofsted evaluations in different boroughs.

In Figure 4.10, we show the probability of having poor, good, and outstanding quality schools across different boroughs. We set all covariates equal to their mean, except for the indicator of religious denomination which is set equal to zero (*i.e.* no religious denomination). This figure highlights the presence of three areas of London that show similar patterns. These are the boroughs belonging to inner London, and, among those of outer London, the western and eastern boroughs. The most evident pattern is that the western and inner London boroughs tend to have higher probability of having outstanding schools and consequently lower probability for poor quality schools. Furthermore, within the three areas defined above it also is possible to notice similar

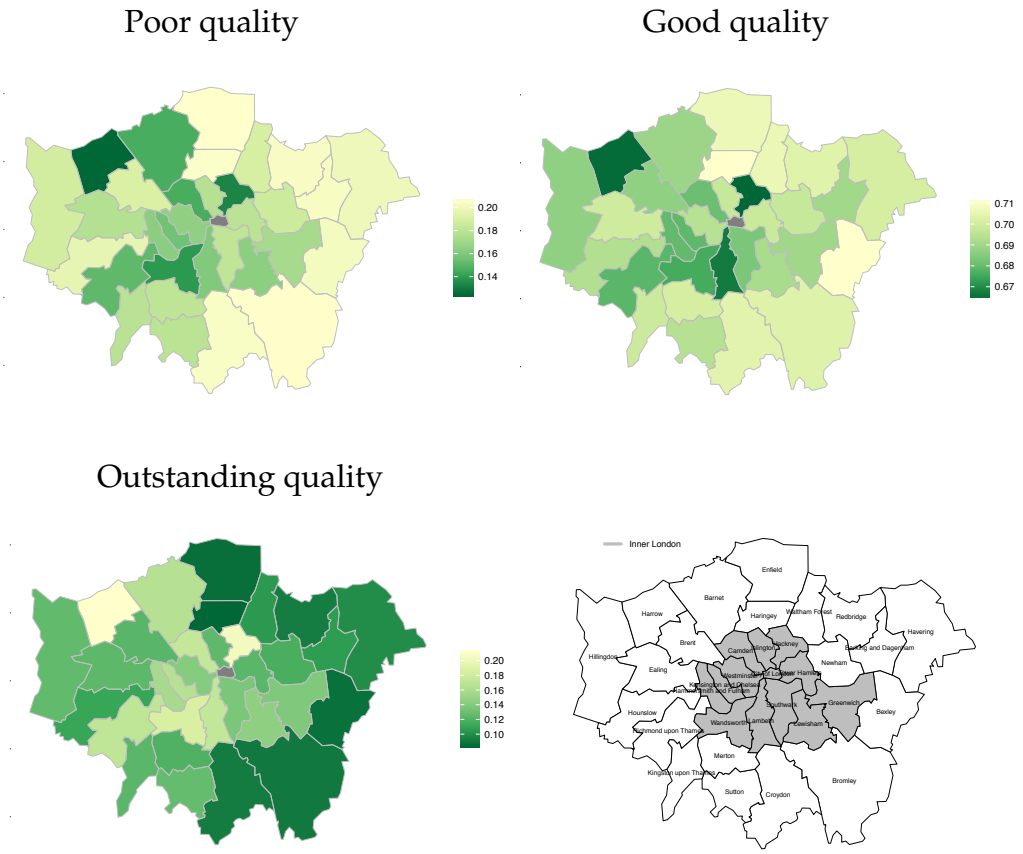


FIGURE 4.10: Posterior predictive probability for poor quality (top-left panel), good quality (top-right panel) and outstanding quality (bottom-left panel) for school with individual features set to observed mean values and no religious denomination.

probabilities among neighbouring boroughs (*e.g.* Croydon, Bromley and Bexley or Enfield and Haringey).

Once a flexible baseline probability is estimated, we check the effect of school features on the overall quality of the schools. These effects are estimated through the vector β, which captures systematic changes in the probabilities of observing different Ofsted scores explained by variations in the school-specific covariates. In Figure 4.11, we show the posterior distributions of the regression coefficients for the different covariates, together with the limits of 95% and 68% percent credible intervals (blue dashed lines and violet dashed lines, respectively). The covariate with the strongest effect is the percentage of pupils eligible for free school meals. The respective posterior distribution has mean around -0.1,

which implies a negative impact on the school quality. This variable is generally considered as an indicator of the social class of the pupils. Another variable that has a similar impact on the quality is the percentage of pupils which do not speak English as first language. This covariate negatively affects the probability of high Ofsted score, but the relative posterior distribution has larger variance and the 95% credible interval contains 0, although the 68% credible interval does not. Finally, having a religious denomination positively affects the quality, with a posterior mean around 0.1. Also for this coefficient the 95% confidence interval contains 0 and the 68% credible interval does not. All other variables have negligible effect on the outcome.

We compare the DGDP mixture model of this application with a similar DDP mixture model for which the precision parameter of the stick-breaking is specified as $\alpha(l_i) = \exp(\alpha_1 + \alpha_2 \eta_{l_i})$. For the competitor model we set $\tau = 1$, $\alpha_1 = -1$ and $\alpha_2 = 2.5$, using an equivalent strategy as the one employed for the DGDP. All other specifications are equivalent for both models. As for the previous examples, competitors models are assessed using PSBF. The result of log(PSBF)$= 2.33$ shows similar performance of the competitor models, but in favour of the proposed approach.

4.6 Discussion

In this chapter we introduce the DGDP, a stochastic process over discrete random probability measures. The DGDP has GDP distributed marginals. This process directly generalises the famous DDP, which instead has DP distributed marginals. The generalisation allows more flexibility at the marginal level, as well as better interpretability of the parameters. The DGDP can be constructed using sequences of correlated stick-breaking weights indexed by covariate levels. Random functions of the covariate levels can be included in the means of the Beta random variables included in the stick-breaking process. When Probit or Logit regression models are employed, the DGDP can be seen as a stochastic version of the Probit stick-breaking or Logit stick-breaking processes, respectively.

The first part of this chapter describes the main properties of the GDP and introduces new distributional properties of samples generated by realisations from a DGDP, along with results about random truncation of the process. In the

second part, we define the DGDP and present different criteria for assessing the strength of dependence between DGDP marginals that are indexed by different points in the covariate space. We discuss different MCMC algorithms for the posterior inference of DGDP mixture models and give details for two of them (one contained in Appendix C.3). The last part of the chapter illustrates two applications of the DGDP. First, we use the the DGDP for modelling longitudinal data to assess the effect of asparaginase activity on triglyceride levels when the former is used to treat patients affected by ALL. Inference is based on albumin levels, which served as surrogate for asparaginase activity. The second application includes the study of the effect of various school features on the Ofsted evaluation of effectiveness. We control for the borough in which each school is located and for the spatial dependence across neighbouring boroughs using the DGDP.

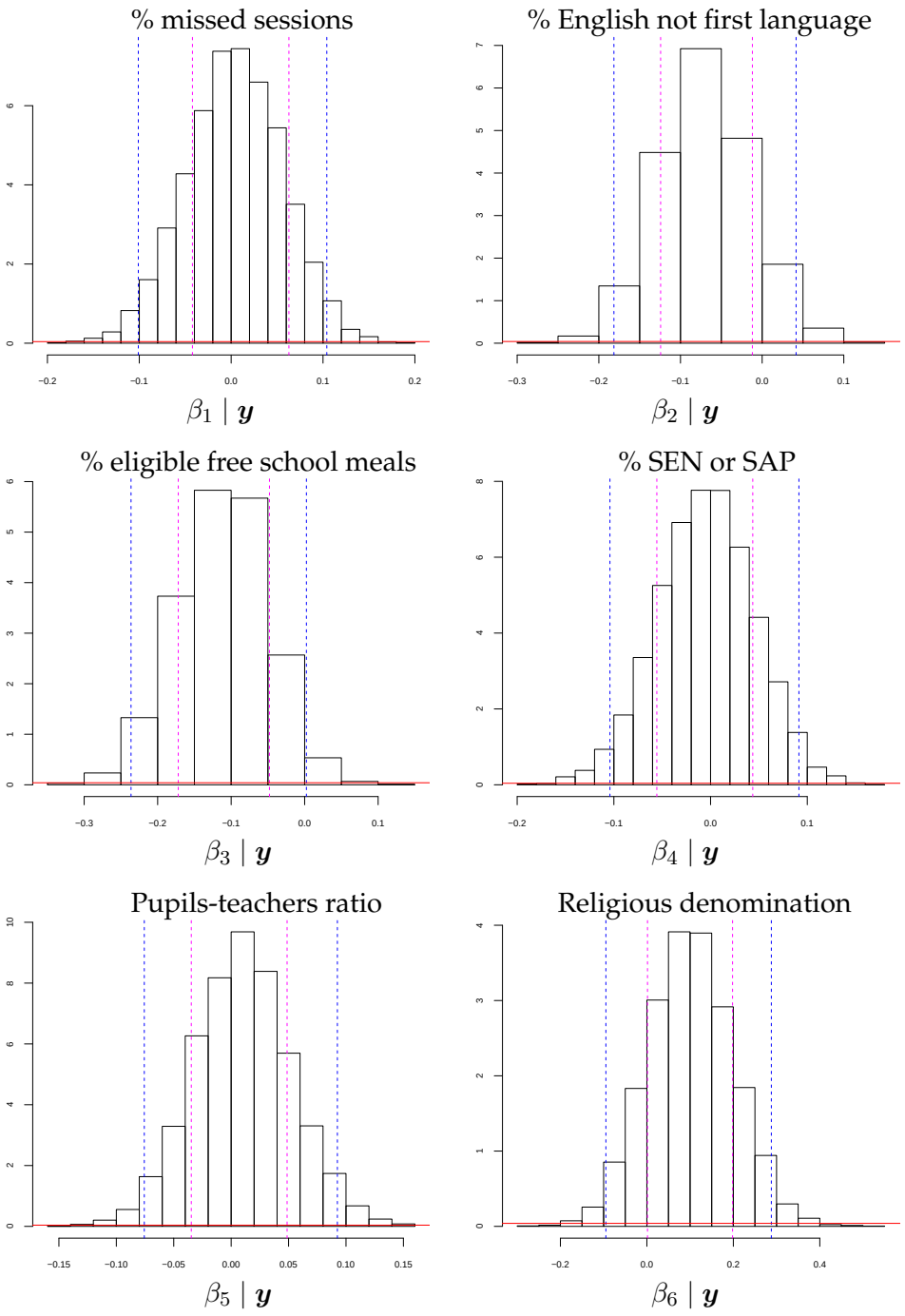


FIGURE 4.11: Posterior density of the regression coefficients relative to the percentage of missed sessions, percentage of pupils for which English is not the first language, percentage of pupils eligible for free school meals, percentage of pupils under SEN or SAP programs, pupils-teachers ratio and religious denomination. Blue dashed lines indicate 95% credible intervals, while violet dashed lines indicate 68% credible intervals.

Chapter 5

Dynamic nonparametric Probit model for correlated binary variables

We present a semi-parametric model for modelling time-evolving vectors of correlated binary variables. This relies on the introduction of continuous latent variables which are discretised to obtain the sampling model. We assume the distribution of the latent variables to be an infinite mixture of distributions with weights that vary across some covariate space and with mean and covariance matrix being component-specific. This distribution includes also an autoregressive term that captures the time evolution of the latent variables and therefore of the binary observations. The proposed method is motivated by the study of LUTS observed at subsequent attendance visits. In particular, we evaluate the temporal dependence among the symptoms controlling for the presence of UTI. The results show that the most recurrent symptoms are stress incontinence and voiding, which are also the most related with presence of pyuria, the best biomarker of infections. Furthermore, we observe that the correlation among symptoms changes over time. The pair of symptoms which appear to be the most correlated are pain and voiding. The material included in this chapter is based on the work in Barcella et al. [2016c].

5.1 Introduction

LUTS represent a group of signs which appear as the consequence of a number of possible diseases. These symptoms are commonly classified into four categories: urgency, stress incontinence, voiding and pain symptoms. LUTS affect

a very large proportion of the population, especially elderly people, and they contribute significantly to the costs of the health systems. In fact, the diseases that lead to the presence of LUTS can often become chronic, as such requiring expensive and time consuming treatments (see Section 2.1 for more details). A relevant example is represented by UTI.

In this chapter we investigate the temporal evolution of LUTS as recorded at subsequent clinic attendance visits. We do so accounting for covariates and for correlations among symptoms. In particular, we are interested in controlling for the presence of pyuria in the urine, which is the best biomarker for UTI, in order to obtain robust estimates of the parameters which govern the temporal evolution of the symptoms. For this purpose we analyse a dataset recording the presence of at least one symptom within the four categories of LUTS via binary indicators in 1015 patients at 4 different attendance visits (the full list of symptoms is in Table 3.1). Data have been collected at the *Lower Urinary Tract Service Clinic* (Whittington Hospital, London, UK). Furthermore, at each attendance visit indicators for the presence of pyuria have been also recorded together with the age of the patients.

From a statistical point of view, the task above can be framed in terms of modelling correlated binary variables, where the correlation is among symptoms. The problem of modelling correlated binary variables is frequent in applications and a number of different solutions have been proposed in the literature. One of the most common strategies involves the introduction of continuous latent variables, which are related to the binary variables via thresholds. The success of this class of models is given by the possibility of including complex structures in the latent variables whose distributions are chosen to facilitate posterior inference. Examples are represented by the Probit and Logit models (Albert and Chib [1993]), for which multivariate extension are available (Ashford and Sowden [1970], Chen [2004], Chib and Greenberg [1998], O'Brien and Dunson [2004]). In this chapter we focus on the Probit model, which involves Gaussian latent variables discretised at 0: positive and negative values of the latent variables correspond to 1 and 0, respectively, at binary level.

When covariates are available, they can be accommodated in the mean of the latent variables using a simple linear regression, which facilitates the interpretation of covariate effect. This simple structure can be generalised also when the latent variable is multivariate, *i.e.* when the objective is to model vectors of

binary variables. This can be achieved by imposing a seemingly unrelated regressions (SUR, Zellner [1962]) structure on the latent variables. SUR involves a set of univariate regression models with distinct parameters, but correlated error structure. Once again, time dependence can be incorporated in the model through autoregressive components within the latent variable distributions. An example for univariate binary time series is the work by Giardina et al. [2011].

Recently, assumptions on the latent variables distribution have been relaxed by introducing nonparametric distributions (see Jara et al. [2007]). These often involve DPM models, for which a review is presented in Section 1.3. DPM of latent distributions have been employed to model vectors of binary variables (Jara et al. [2007]) and univariate binary time series (Di Lucca et al. [2013]). In this setting covariates can be included as fixed regression effects (as shown in Jara et al. [2007]) in all mixture components favouring the interpretability of the regression coefficients or within the weights of the mixture components modelling jointly the latent variables and the covariates (see DeYoreo et al. [2015]). Other related solutions can be found for multicategorical discrete outcomes, *e.g.* Kottas et al. [2005] and DeYoreo and Kottas [2014]. A relevant contribution for this chapter is in DeYoreo and Kottas [2015], where a model for time series of univariate ordinal categorical variables is proposed including covariate information through an appropriate model and capturing the time evolution through a time-evolving version of the DDP.

In this work we propose a model for LUTS which employs latent variables, whose distributions are assumed to be semi-parametric. Risk factors for LUTS and a lagged latent components are included in the latent variables distribution via a SUR model, where the different levels of SUR are indexed to the different symptoms. Intercepts of the regressions and the covariance matrix included in the joint distribution of the error terms are assumed to be component-specific in an infinite mixture model where the mixture weights vary (and are correlated) across different pyuria states and for first visits and follow-ups. This produces the effect of a non-linear regression of the pyuria and visit indicators in the mean and covariance matrix of the latent variables. This also allows us to achieve more robust estimates for the other linear regression coefficients included in the model, especially for the autoregressive coefficients, in a fashion similar to what has been described by Papageorgiou et al. [2015]. The results of the data analysis highlight the different behaviours of the symptoms, both

in terms of correlations with other symptoms and in terms of recurrence. A relevant connection has been found between the presence of pyuria and the probability of observing voiding symptoms, which in turn is highly correlated with stress incontinence and pain symptoms. In addition, correlations between pairs of symptoms do not seem constant for all patients and in some case they are affected by the presence of pyuria.

The rest of the chapter is organised as follows. In Section 5.2 we present the detail of the proposed model, while in Section 5.3 we present a MCMC algorithm for sampling from the posterior distributions of the parameters. The results of the application on the LUTS data set are shown in Section 5.4. We conclude with a discussion in Section 5.5.

5.2 A semi-parametric model for binary variables

We consider a study involving N patients, for which a $D-$dimensional binary vector is recorded at T subsequent time points. We denote the collection of all binary records with $\boldsymbol{Y}$, an array having dimensions $N \times D \times T$. Let $y_{i,d,t}$ and $\boldsymbol{y}_{i,t}$ denote a single entry and a row of binary variables of $\boldsymbol{Y}$ for patient i at time t, respectively.

5.2.1 Dynamic multivariate Probit model

We model $\boldsymbol{Y}$ introducing an array of correlated continuous latent variables, which we denote with $\boldsymbol{Z}$, having entries $z_{i,d,t} \in \mathbb{R}$ and such that:

$$y_{i,d,t} = 1 \text{ if and only if } z_{i,d,t} \geq 0.$$

Given the condition above, we can write the likelihood of $\boldsymbol{Y}$ as

$$\Pr(\boldsymbol{Y} \mid \boldsymbol{\Theta}) = \int \prod_{\{y_{i,d,t}=1\}} \mathbb{I}_{[0,+\infty)}(z_{i,d,t}) \prod_{\{y_{i,d,t}=0\}} \mathbb{I}_{(-\infty,0)}(z_{i,d,t}) F(\mathrm{d}\boldsymbol{Z} \mid \boldsymbol{\Theta}),$$

where $F(\boldsymbol{Z} \mid \boldsymbol{\Theta})$ is the joint distribution of $\boldsymbol{Z}$, parameterised by $\boldsymbol{\Theta}$.

We recall that T represents the time dimension of $\boldsymbol{Y}$ and we assume for the density of $F(\boldsymbol{Z} \mid \boldsymbol{\Theta})$, which we denote by $f(\boldsymbol{Z} \mid \boldsymbol{\Theta})$, the following factorisation

$$f(\boldsymbol{Z} \mid \boldsymbol{\Theta}) = \prod_{i=1}^{N} \left\{ f(\boldsymbol{z}_{i,1} \mid \Theta_{i,1}) \prod_{t=2}^{T} f(\boldsymbol{z}_{i,t} \mid \boldsymbol{z}_{i,t-1}, \Theta_{i,t}) \right\}, \tag{5.2.1}$$

where $\boldsymbol{z}_{i,t} = (z_{i,1,t}, \ldots, z_{i,D,t})$, *i.e.* the row of $\boldsymbol{Z}$ corresponding to the $i-$th patient at time t. The latter equation imposes a Markov structure to the distribution of the latent variables, which directly determines the distribution of $\boldsymbol{Y}$.

A computationally convenient assumption is to assume $f(\cdot)$ to be a multivariate Normal distribution. This has a number of implications. First, at any given time t, the probability of observing $\boldsymbol{y}_{i,t}$ can be calculated as an integral under a multivariate Normal distribution, as in multivariate Probit models. Secondly, the Markov structure assumed in (5.2.1) can be easily accommodated within the model using appropriate autoregressive terms. Furthermore, covariates can be easily included in the model. Finally, the Normal assumption simplifies the calculations, allowing the use of standard algorithms such as the one proposed by Albert and Chib [1993] for posterior inference.

Let $\boldsymbol{X}$ denote an array of dimension $N \times P \times T$, containing records of P time-dependent covariates. We write $\boldsymbol{x}_{i,t}$ to indicate the row of $\boldsymbol{X}$ corresponding to the $i-$th patient at the $t-$th time. Recalling (5.2.1), we specify the following distribution for the latent variable at time 1:

$$f(\boldsymbol{z}_{i,1} \mid \Theta_{i,1} = (\boldsymbol{\alpha}_{i,1}, \Lambda, \Sigma_{i,1})) = \text{MNormal}_D(\boldsymbol{z}_{i,1} \mid \boldsymbol{\alpha}_{i,1} + \Lambda \boldsymbol{x}'_{i,1}, \Sigma_{i,1}), \tag{5.2.2}$$

where $\text{MNormal}_D(\cdot)$ is the $D-$dimensional Normal distribution, $\boldsymbol{\alpha}_{i,1}$ is a vector of intercepts of length equal to D and Λ is a $D \times P$ matrix of regression coefficients. Equivalently, we specify the following transition density for $t = 2, \ldots, T$:

$$f(\boldsymbol{z}_{i,t} \mid \boldsymbol{z}_{i,t-1}, \Theta_{i,t} = (\boldsymbol{\alpha}_{i,t}, \Lambda, \Gamma, \Sigma_{i,t})) = \text{MNormal}_D(\boldsymbol{z}_{i,t} \mid \boldsymbol{\alpha}_{i,t} + \Lambda \boldsymbol{x}'_{i,t} + \Gamma \boldsymbol{z}'_{i,t-1}, \Sigma_{i,t}), \tag{5.2.3}$$

where Γ is a $D \times D$ matrix containing the autoregressive coefficients.

The model described above is connected with SUR models in that at a certain time point, given all parameters and lagged latent components, the distributions above imply D distinct regressions on the means of the latent variables.

The latter are linked together via the error distributions governed by the covariance matrix $\Sigma_{i,t}$.

Λ and Γ are assumed to be constant across time. This simplifies interpretability of the coefficients. In particular, simple Bayesian hypothesis testing on temporal dependence among the binary variables can be performed and this is an important requirement for our motivating application. However, extensions including temporal dependent versions of the matrices of coefficients Λ_t and Γ_t can be easily specified.

5.2.2 Nonparametric prior model

The quantities $\boldsymbol{\alpha}_{i,t}$ and $\Sigma_{i,t}$ characterise the baseline probability of observing $\boldsymbol{y}_{i,t}$. We want to employ a prior distribution which is flexible enough to capture possible heterogeneity across patients at different time point.

Let us introduce an additional array of covariates, $\boldsymbol{U}$. This is an $N \times R \times T$ array, where $\boldsymbol{u}_{i,t}$ is a row of $\boldsymbol{U}$ which encodes the information about the $i-$th patient at time t, such as treatment arm, risk group and other common indicators which may evolve overtime. This is not uncommon in biostatistics where clinicians classify patients into different classes of risk for developing a specific disease based on clinical history and characteristics. We consider the case in which $\boldsymbol{U}$ contains binary indicators. Consequently, the $\boldsymbol{y}_{i,t}$'s are implicitly clustered according to the various combinations of the binary covariates contained in $\boldsymbol{u}_{i,t}$.

We include via a flexible model the information contained in $\boldsymbol{U}$ in the baseline probability of $\boldsymbol{y}_{i,t}$. Using an approach similar to the one employed in mixed-effect modelling, we assume each group defined by $\boldsymbol{U}$ to have a specific random prior distribution as follows

$$(\boldsymbol{\alpha}_{i,t}, \Sigma_{i,t}) \mid G_{\boldsymbol{u}_{i,t}} \sim G_{\boldsymbol{u}_{i,t}},$$

where $G_{\boldsymbol{u}_{i,t}}$ is a discrete distribution of the form

$$G_{\boldsymbol{u}_{i,t}} = \sum_{h=1}^{\infty} w_{h,\boldsymbol{u}_{i,t}} \delta_{\boldsymbol{\alpha}_h, \Sigma_h}, \tag{5.2.4}$$

where, for each $\boldsymbol{u}_{i,t}$, $\sum_h w_{h,\boldsymbol{u}_{i,t}} = 1$, and δ_a denotes a unit point mass at a. The use of such random prior distribution implies that the latent variable density can be written as

$$\begin{aligned} f(\boldsymbol{z}_{i,t} \quad | \quad & \boldsymbol{z}_{i,t-1}, \Lambda, \Gamma, G_{\boldsymbol{u}_{i,t}}) \\ = \quad & \int f(\boldsymbol{z}_{i,t} \mid \boldsymbol{z}_{i,t-1}, \Lambda, \Gamma, \boldsymbol{\alpha}_{i,t}, \Sigma_{i,t}) \mathrm{d}G_{\boldsymbol{u}_{i,t}}(\boldsymbol{\alpha}_{i,t}, \Sigma_{i,t}) \\ = \quad & \sum_{h=1}^{\infty} w_{h,\boldsymbol{u}_{i,t}} \text{MNormal}_D(\boldsymbol{z}_{i,t} \mid \boldsymbol{\alpha}_h + \Lambda \boldsymbol{x}'_{i,t} + \Gamma \boldsymbol{z}'_{i,t-1}, \Sigma_h), \end{aligned}$$

which is an infinite location-scale mixture of multivariate Normal distributions, having weights which vary according to the components in $\boldsymbol{u}_{i,t}$.

Specifying a prior of $G_{\boldsymbol{u}_{i,t}}$ is equivalent to find suitable prior distributions for the collection of the weights, $w_{h,\boldsymbol{u}_{i,t}}$, and for the locations, $\boldsymbol{\alpha}_h$ and Σ_h.

Starting from the weights, we need a stochastic process prior indexed at various levels of $\boldsymbol{u}_{i,t}$, whose realisations are distributions over an infinite dimensional simplex. Furthermore, it is desirable to borrow strength across the different groups of observations implied by $\boldsymbol{U}$. Different solutions for these tasks have been developed in the field of RPMx and a survey of the major alternatives is in Section 1.4.

In Chapter 4 we have introduced the Dependent Generalised Dirichlet Process (DGDP), a process over collections of distributions which can be considered as an extension of the DDP. We opt for using the same idea of the DGDP also in this chapter, but we discuss alternative solutions in the next paragraphs. DGDP assumes a particular stick-breaking process prior for the weights of (5.2.4) where $w_{1,\boldsymbol{u}_{i,t}} = v_{1,\boldsymbol{u}_{i,t}}$ and

$$w_{h,\boldsymbol{u}_{i,t}} = v_{h,\boldsymbol{u}_{i,t}} \prod_{l<h} (1 - v_{l,\boldsymbol{u}_{i,t}}), \text{ for } h = 2, 3, \ldots$$

with

$$v_{1,\boldsymbol{u}_{i,t}}, v_{2,\boldsymbol{u}_{i,t}}, \ldots \mid \boldsymbol{\mu}, \phi \sim \text{Beta}(v_{h,\boldsymbol{u}_{i,t}} \mid \phi \text{logit}^{-1}(\boldsymbol{\mu} \boldsymbol{u}_{i,t}), \phi(1 - \text{logit}^{-1}(\boldsymbol{\mu} \boldsymbol{u}_{i,t}))),$$

for all combination of $\boldsymbol{u}_{i,t}$. In the latter equation, ϕ is a positive parameter and $\boldsymbol{\mu}$ is vector of real parameters. The almost sure discreteness of DGDP realisations imposes a clustering structure of the attendance visits: observations sharing

the same value of $\boldsymbol{\alpha}_h^*$ and Σ_h^* can be interpreted as a cluster. Compared to DDP (which will in principle produce an equivalent clustering effect of attendance visits), the DGDP assumes extra flexibility to the distribution of the partition of the observations especially in terms of number and size of clusters as a consequence of the richer parameterisation of the GDP compared to the traditional DP (see Section 4.2.3).

Independently from the generation of the weights, the locations, which are shared among all $G_{\boldsymbol{u}_{i,t}}$ for different values of $\boldsymbol{u}_{i,t}$, are generated as follows. We first assume

$$\boldsymbol{\alpha}_h \sim \text{MNormal}_D(\boldsymbol{m}_D, \sigma_\alpha^2 I_D), \text{ for } h = 1, 2, \ldots$$

Σ_h^* requires likelihood identifiability conditions imposed by the thresholding of the latent variables. In order to avoid over-restrictive constraints on the covariance matrix, we follow the works by Jara et al. [2007] and Pourahmadi [1999] where the conditional variances are constrained. We write $\Sigma_h^{*-1} = L_h^{*\prime} I_D L_h^*$, where L_h^* is a lower triangular matrix, with ones on the diagonal and unconstrained values on the non-zero entries. Collecting these entries in the vector ν_h^*, which has $D_\nu = D(D-1)/2$ components, we assume

$$\nu_h \sim \text{MNormal}_{D_\nu}(\boldsymbol{m}_{D_\nu}, \sigma_\nu^2 I_{D_\nu}), \text{ for } h = 1, 2, \ldots$$

In this chapter we assume shared locations for all $\boldsymbol{u}_{i,t}$, following the argument discussed in Chapter 4. However, it is possible to extend this construction including also covariate dependent locations. This requires the use of suitable stochastic process priors indexed by the covariates.

5.2.3 Prior distribution specification

The model described above requires the specification of prior (and hyperprior) distributions for the remaining unknown parameters. Let $\lambda_{d,p}$ and $\gamma_{d,d'}$ denote a single entry in matrix Λ and Γ respectively. We assume

$$\lambda_{d,p} \overset{iid}{\sim} \text{Normal}(\lambda \mid m_\lambda, \sigma_\lambda^2)$$

$$\gamma_{d,d'} \overset{iid}{\sim} \text{Normal}(\gamma \mid m_\gamma, \sigma_\gamma^2).$$

The hyperprior distributions for the values of $\boldsymbol{\mu}$ and ϕ are set as follows

$$\boldsymbol{\mu} \sim \text{MNormal}_R(\boldsymbol{\mu} \mid m_\mu, \sigma_\mu^2 I_R)$$

$$\phi \sim \text{Gamma}(\phi \mid a_\phi, b_\phi),$$

where the latter is a Gamma distribution with expectation a_ϕ / b_ϕ.

5.3 Posterior inference

Posterior inference can be performed using MCMC methods. For the model described in Section 5.2, samples from posterior distribution of the parameters can be approximated using a Metropolis-within-Gibbs algorithm. We summarise below the main parts of the algorithm.

i) *Resample $\boldsymbol{Z}$ given all other parameters.* In the proposed model there are $N \times D \times T$ latent variables, which can be resampled sequentially starting at $t = 1$ using the algorithms for binary Probit models described in Albert and Chib [1993] and Holmes and Held [2006].

ii) *Resample Λ given all other parameters.* These parameters are shared by all $i = 1, \ldots, N$ and $t = 1, \ldots, T$ and can be resampled directly from the full conditionals.

iii) *Resample Γ given all other parameters.* These parameters are shared by all $i = 1, \ldots, N$ and $t = 2, \ldots, T$ and can be resampled directly from the full conditionals.

iv) *Resample $\{(\boldsymbol{\alpha}_{i,t}, \Sigma_{i,t}), \forall i, t\}$ given all other parameters.* This step is based on a finite truncation of the process in (5.2.4), that include only a large enough number H of atoms, which we denote with $G^H_{\boldsymbol{u}_{i,t}}$. For details on the finite approximation of (5.2.4) see Section 4.2.4. Then, the full conditional follows:

$$\Pr((\boldsymbol{\alpha}_{i,t}, \Sigma_{i,t}) = (\boldsymbol{\alpha}_h, \Sigma_h) \mid \ldots) \propto w_{h,\boldsymbol{u}_{i,t}} f(\boldsymbol{z}_{i,t} \mid \boldsymbol{\alpha}^*_h, \Sigma^*_h, \ldots), \text{ for } h = 1, \ldots, H.$$

v) *Resample $G^H_{\boldsymbol{u}_{i,t}}$ for all $\boldsymbol{u}_{i,t}$ given all other parameters.* This step consists of resampling the locations $\{(\boldsymbol{\alpha}_h, \Sigma_h), h = 1, \ldots, H\}$ and weights $\{w_{h,\boldsymbol{u}_{i,t}}, h =$

$1, \ldots, H$ and all $\boldsymbol{u}_{i,t}\}$. For location parameters, full conditionals can be derived using the likelihood of the observations associated to each location and the prior distribution of $\boldsymbol{\alpha}_h$ and Σ_h (for the latter update see Jara et al. [2007]). Resampling the weights uses the joint distribution of the weights (marginally for each $\boldsymbol{u}_{i,t}$) which follows a GDD (Connor and Mosimann [1969]). We augment the parameters space with indicators $s_{i,t}$ taking value in $\{1, \ldots, H\}$ telling to which location of $G^H_{\boldsymbol{u}_{i,t}}$ the observation i, t is assigned in step iv). This produces a conjugate update using a Multinomial distribution for $s_{i,t}$.

vi) *Resample $(\boldsymbol{\mu}, \phi)$ given all other parameters.* This step requires a Metropolis scheme for sampling from the following full conditional

$$p(\boldsymbol{\mu}, \phi \mid \ldots) \propto \prod_{\boldsymbol{u}} p(\boldsymbol{w}_{\boldsymbol{u}} \mid \boldsymbol{\mu}, \phi) p(\boldsymbol{\mu}, \phi),$$

where $\boldsymbol{w}_{\boldsymbol{u}} = (w_{1,\boldsymbol{u}}, \ldots, w_{H,\boldsymbol{u}})$ and $p(\boldsymbol{w}_{\boldsymbol{u}} \mid \boldsymbol{\mu}, \phi)$ is the GDD.

The algorithm described above is based on the truncation of the prior distribution in (5.2.4) up to H atoms. Different strategies have been discussed in the literature to determine a value of H in case the random distribution follows a DP or more general processes. Examples are the works of Ohlssen et al. [2007] and Ishwaran and Zarepour [2000]. A simple strategy that can be adapted to the case of DGDP consists in checking that the expectation and variance of the weight $w_{H,\boldsymbol{u}}$ are adequate for the values of $\boldsymbol{\mu}$ and ϕ explored by the MCMC.

An alternative approach for posterior sampling which does not involve an approximation of the nonparametric prior can be designed based on the slice sampling algorithm presented in Kalli et al. [2011] and Walker [2007] and described in Appendix C.3.

5.4 Application: Lower Urinary Tract Symptoms

In this section we present background information, exploratory analysis and results of the application of the model described in the previous sections for the study of the evolution of LUTS.

5.4.1 Dataset

The dataset employed in the analysis contains information on 1015 female patients affected by LUTS, who have attended at four subsequent visits the *Lower Urinary Tract Service Clinic* (Whittington Hospital, London, UK). At each attendance visit the following information is collected: the date of the visit, the age of the patient, the presence of urgency symptoms (binary), presence of pain symptoms (binary), presence of stress incontinence symptoms (binary), and the count of white blood cells (WBC) in sample of urine.

The most frequently observed symptom is urgency, which affects 72.32% of the patients at the first attendance visit and in 61.77% at follow-up visits. The least frequent symptom is stress incontinence, which is observed in 37.14% of first visits and in 34.93% of follow-ups. Pain and voiding symptoms are observed in 54.68% and 40.00% of the first visits and 44.83% and 34.93% of the follow-ups, respectively.

WBC counts are used to assess the presence of pyuria, which is considered the best biomarker for UTI (see Section 2.1). While evidence have been collected that relates WBC count larger than or equal to one to the presence of infection, the threshold of ten WBC in a urine sample is the sensitivity of common dipstick tests used to assess the presence of pyuria. Following these thresholds we generate two binary indicators from WBC and we refer to them as *mild pyuria* in the case of $1 \leq \text{WBC} \leq 9$, and to *severe pyuria* for $\text{WBC} \geq 10$. These indicators can be interpreted in terms of severity of the infection: in fact, large WBC counts indicate a high degree of inflammation which may lead to complications. Mild pyuria has been observed in 17.86% of the attendance visits, while severe pyuria has been recorded in 18.74% of the cases. Considering the first attendance visits exclusively, these become 18.62% and 21.97%, respectively.

The average number of days between attendance visits is approximately 91 (standard deviation equal to 143.22), where the shortest period is observed between the first two visits with an average of 85 days (standard deviation equal to 147.58) and the longest one is recorded between the last two visits with an average of 95 days (standard deviation equal to 135.07). Finally, the age of the patients at first attendance visits is on average 54, with a sample standard deviation of 17.27.

In terms of treatment regimes, all patients have been treated with a combination of antimuscarinic and bladder retaining after the first attendance visits until the fourth one. Furthermore, patients diagnosed with mild and severe pyuria have been treated with antibiotics.

5.4.2 Notation and choice of hyperparameters

We denote the array containing information on the presence of the symptoms as $\boldsymbol{Y}$. Given the information above this has dimension $1015 \times 4 \times 4$ (corresponding to the number of patients, the number of symptoms and the number of attendance visits, respectively), and contains binary observations. The symptoms appearing in a row $\boldsymbol{y}_{i,t}$ of $\boldsymbol{Y}$ are abbreviated as U, P, S, and V, standing for urgency, pain, stress incontinence and voiding symptoms, respectively. Consequently, in the following sections the indexes d and d' take value in $\{\mathrm{U}, \mathrm{P}, \mathrm{S}, \mathrm{V}\}$.

We denote with $\boldsymbol{X}$ the array containing information about the age of the patients and the time in days from the first attendance visits of all subsequent visits. In this application we only consider the age at the first attendance visits instead of the age when each attendance visit takes place: in fact, the latter can be derived using the time between visits. Therefore, $\boldsymbol{X}$ has dimension $1015 \times 2 \times 4$ corresponding to the number of patients, the number of covariates and the number of attendance visits. Across attendance visits, the first covariate (*Age*) remains constant, while the second one (*Days*) has entries equal to zero for the first attendance visits. For each attendance visit, the two covariates have been centred and rescaled to have mean and variance equal to 0 and 1, respectively.

The indicators for mild and severe pyuria together with an indicator for first attendance visits and intercepts have been included in $\boldsymbol{U}$, which is then an array with dimension $1015 \times 4 \times 4$. Consequently, considering the row vectors $\boldsymbol{u}_{i,t}$ of $\boldsymbol{U}$ and all possible values they can take, we can identify six groups of attendance visits which are all observed in our data set.

Both $\boldsymbol{X}$ and $\boldsymbol{U}$ are used as covariates, however their effect on the latent variables are different. In particular, the effect of $\boldsymbol{U}$ is non-linear via the intercepts $\boldsymbol{\alpha}$ and the covariance matrix $\boldsymbol{\Sigma}$. Differently, the entries of $\boldsymbol{X}$, similarly to the autoregressive terms, affect the latent variables linearly. The decision of which covariates to assign to $\boldsymbol{U}$ and to $\boldsymbol{X}$ has been driven by the specific application.

In particular, $\boldsymbol{U}$ induces a natural clustering of the observations based on the levels of the UTI and first attendance visits and follow-ups.

We set the following hyperparameter values: $\boldsymbol{m}_\alpha = \mathbf{0}_D$ (where $\mathbf{0}_a$ denotes a vector with a components all equal to zero); $\boldsymbol{m}_\nu = \mathbf{0}_{D_\nu}$; $m_\lambda = m_\gamma = m_\mu = 0$; $\sigma^2_\alpha = \sigma^2_\nu = \sigma^2_\lambda = \sigma^2_\gamma = 100, \sigma^2_\mu = 1$ and $a_\phi = b_\phi = 1$. We use the algorithm described in Section 5.3 for performing posterior inference. We initialise the algorithm drawing random values from the prior distribution of each parameter and we run the algorithm for 100 000 iterations, with a burning period of 20 000, and we save every tenth sample. Convergence of MCMC has been assessed using trace and autocorrelation plots.

5.4.3 Results

Before discussing the results of the application, we notice that the marginal probability of observing a certain symptom is equal to

$$\Pr(y_{i,d,t} = 1 \mid \alpha_{i,d,t}, \boldsymbol{\lambda}_d, \boldsymbol{\gamma}_d, \sigma_{i,d,t}, \boldsymbol{z}_{i,t-1}) = \Phi\left(\frac{\alpha_{i,d,t} + \boldsymbol{\lambda}_d \boldsymbol{x}'_{i,t} + \boldsymbol{\gamma}_d \boldsymbol{z}'_{i,t-1}}{\sigma_{i,d,t}}\right),$$

where $\boldsymbol{\lambda}_d$ and $\boldsymbol{\gamma}_d$ are the d-th rows of Λ and Γ, respectively. This shows that $\alpha_{i,d,t}$ controls the baseline probability of the symptom d at visit t, together with $\sigma_{i,d,t}$, the square root of the d-th diagonal component of $\Sigma_{i,t}$. When the covariates $\boldsymbol{x}_{i,t}$ and $\boldsymbol{z}_{i,t}$ have components equal to zero, the sign of $\alpha_{i,d,t}$ determines if the probability of observing the symptoms is larger than 0.5, which happens when the parameter is positive. The dependence among symptoms can be evaluated computing the conditional distributions of components of the latent vector, *i.e.* $p(z_{i,d,t} \mid \boldsymbol{z}_{i,-d,t}, \ldots)$, where $\boldsymbol{z}_{i,-d,t}$ is obtained removing $z_{i,d,t}$ from $\boldsymbol{z}_{i,d,t}$.

We discuss posterior inference for the parameters of the proposed model distinguishing between the individual effects, $\boldsymbol{\alpha}_{i,t}$ and $\Sigma_{i,t}$, and the shared effects (*i.e.* shared among all patients), Λ and Γ.

Individual effects

We begin by looking at the posterior distribution of the vector $\boldsymbol{\alpha}_{i,t}$. Its discrete prior distribution induces ties across different patients and times.

Figure 5.1 shows the posterior distribution of the intercepts for the patients at first attendance visits for different levels of pyuria. The difference between

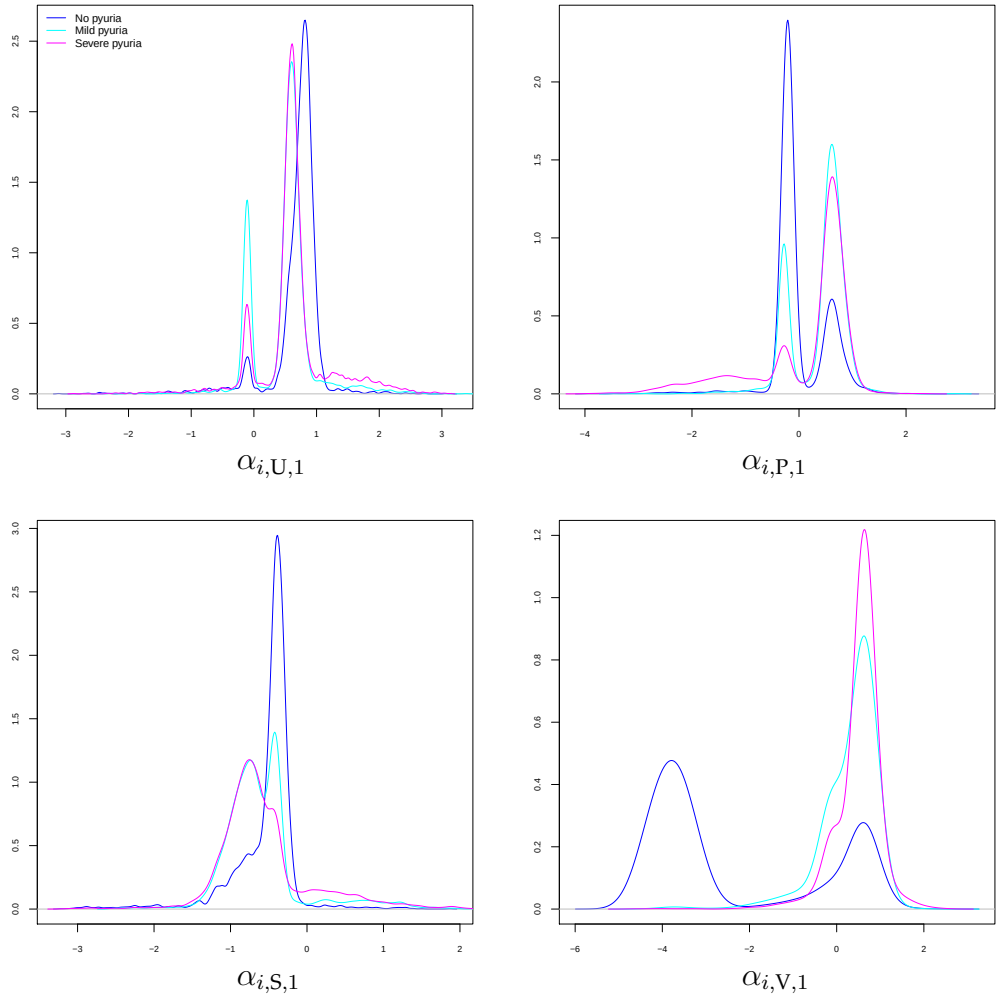


FIGURE 5.1: Posterior density of $\alpha_{i,d,t}$, $d = \{\mathrm{U}, \mathrm{P}, \mathrm{S}, \mathrm{V}\}$, when t is the first attendance visit, for different pyuria levels.

the posterior densities for the patients without pyuria and those with mild or severe pyuria is evident. On the other hand, densities corresponding to patients with mild or severe pyuria levels are similar. In particular, the presence of pyuria (either mild or sever) leads to higher posterior mean for the latent variables associated with pain and voiding symptoms. The opposite happens for urgency and stress incontinence symptoms, for which the presence of pyuria reduces the mean of the associated latent variables.

Recalling that the sign of the $\alpha_{i,d,t}$ controls the probability of a symptom being larger or smaller than 0.5, we notice that for pain and voiding symptoms

the presence of pyuria changes the sign of the posterior expectation of $\alpha_{i,d,t}$. The largest effect of pyuria is observed for voiding symptoms.

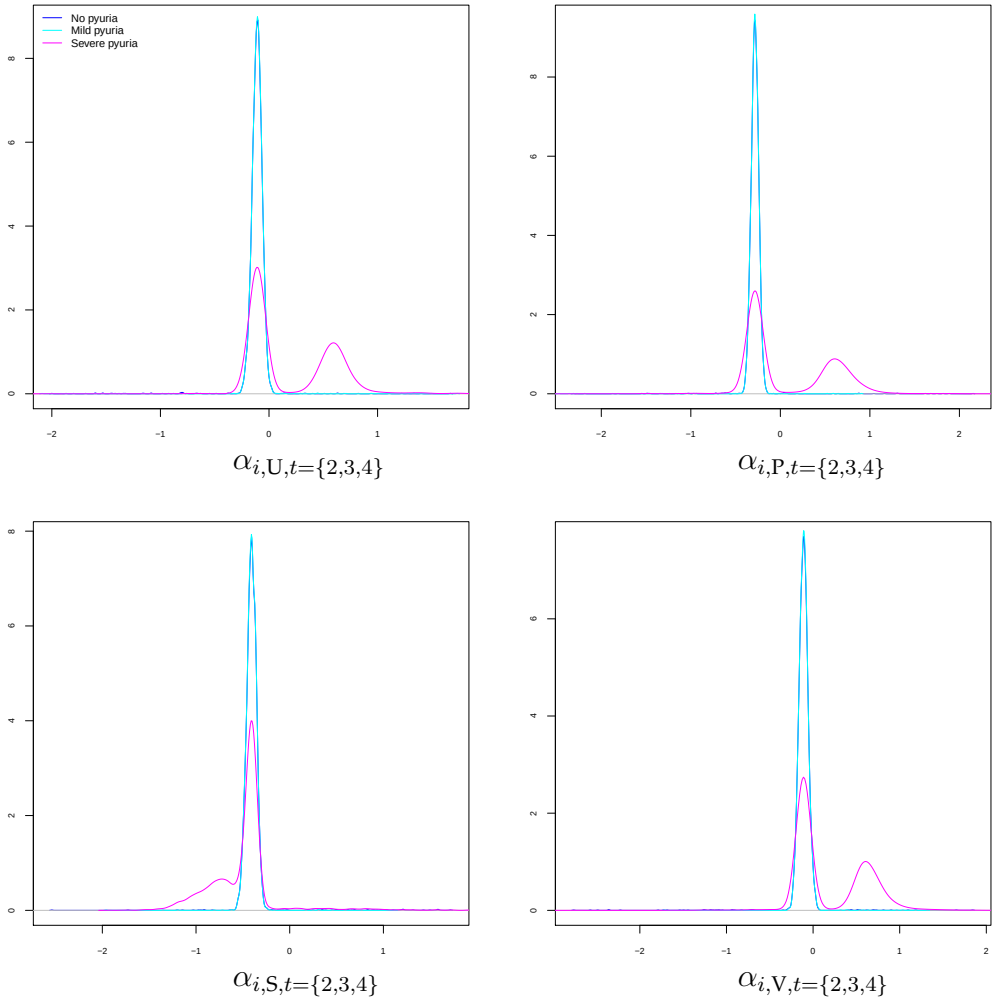


FIGURE 5.2: Posterior density of $\alpha_{i,d,t}$, $d = \{\mathrm{U}, \mathrm{P}, \mathrm{S}, \mathrm{V}\}$, and $t = \{2, 3, 4\}$ for different pyuria levels.

For all attendance visits after the first one the posterior densities are presented in Figure 5.2. Recall that patients after the first attendance visit have all been treated for LUTS. In this case the posterior expectations of the $\alpha_{i,d,t}$'s are all negative for the cases with no and mild pyuria, which show almost identical posterior distributions. Differently, the presence of severe pyuria strongly affects the posterior densities of $\alpha_{i,d,t}$: positive marginal posterior expectations are observed for urgency, pain and voiding symptoms.

The other individual coefficients estimated in the proposed formulation are the entries of the covariance matrix, *i.e.* $\Sigma_{i,t}$. Our focus is on computing the correlations' coefficients from $\Sigma_{i,t}$. Similarly to $\boldsymbol{\alpha}_{i,t}$, the posterior distributions of these coefficients are indexed by the values of $\boldsymbol{u}_{i,t}$. Given that we are including in our model four symptoms, we have six pairwise correlations, which vary across different pyuria levels and between first attendance visits and follow-ups. The posterior density estimates for each correlation are reported in Figure 5.3.

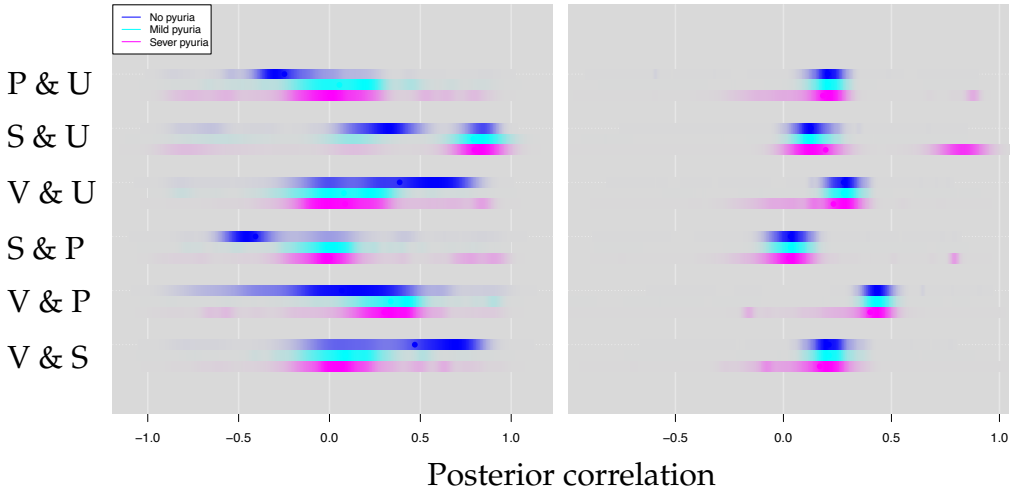


FIGURE 5.3: Posterior densities of the correlation between different pairs of latent variables (corresponding to different symptoms) at first attendance visit (left panel) and at follow-up visits (right panel), for different levels of pyuria.

We divide in two different panels the correlations estimated for first attendance visits (left panel) and for follow-up visits (right panel), while different colours correspond to different pyuria levels. On the y-axis, the correlations between symptoms are abbreviated using the letters indicating the symptoms.

The covariate that has the strongest impact on these posterior distributions is the indicator of first attendance visit. This can be seen by comparing the densities reported on the left panel with those on the right panel in Figure 5.3. Obviously, part of the difference may be due to the fact that correlations for follow-up visits are estimated using a larger number of observations compared to those at first attendance visits. In some case the posterior expectation of the correlations move toward zero after the first visit attendance (S & U and S & P).

In the other cases, we notice a positive increase of the correlation values (P & U and V & P).

The proposed model allows identifying clusters of attendance visits having different correlation patterns between pairs of symptoms. In addition, these patterns can vary across different pyuria levels. This is the case of S & P where a group of observations with no pyuria shows negative correlations at first attendance visit, or the case of S & U where a group of visits characterised by severe pyuria shows higher correlation values at follow-up attendance visits.

Shared effects

We now discuss the estimated posterior distributions of the parameters shared by all patients. The posterior distribution of Γ is parametric because we want to preserve simple interpretation of the marginal effects.

TABLE 5.1: Summary of the posterior distributions of $\gamma_{d,d'}$, with $d = \{U, P, S, V\}$ referring to the latent dependent variables of the d-th symptoms and $d' = \{U, P, S, V\}$ to the latent independent variables of the d'-th symptoms.

	mean	sd	2.5%	25%	50%	75%	97.5%
$\gamma_{U,U}$	0.74	0.04	0.67	0.72	0.74	0.76	0.82
$\gamma_{U,P}$	0.02	0.03	-0.04	-0.00	0.02	0.04	0.08
$\gamma_{U,S}$	0.10	0.02	0.06	0.09	0.10	0.12	0.15
$\gamma_{U,V}$	-0.09	0.02	-0.12	-0.10	-0.09	-0.07	-0.06
$\gamma_{P,U}$	-0.03	0.03	-0.09	-0.05	-0.03	-0.01	0.03
$\gamma_{P,P}$	0.70	0.03	0.65	0.69	0.70	0.72	0.76
$\gamma_{P,S}$	-0.06	0.02	-0.10	-0.08	-0.06	-0.04	-0.01
$\gamma_{P,V}$	0.01	0.01	-0.02	-0.00	0.01	0.02	0.04
$\gamma_{S,U}$	-0.03	0.03	-0.10	-0.05	-0.03	-0.01	0.04
$\gamma_{S,P}$	0.03	0.04	-0.04	0.00	0.03	0.05	0.10
$\gamma_{S,S}$	0.84	0.04	0.78	0.82	0.84	0.87	0.93
$\gamma_{S,V}$	0.05	0.02	0.01	0.03	0.05	0.06	0.08
$\gamma_{V,U}$	-0.13	0.04	-0.23	-0.16	-0.13	-0.10	-0.06
$\gamma_{V,P}$	-0.01	0.05	-0.10	-0.05	-0.02	0.01	0.09
$\gamma_{V,S}$	0.05	0.03	-0.01	0.03	0.05	0.06	0.10
$\gamma_{V,V}$	0.78	0.03	0.71	0.76	0.78	0.80	0.83

In Table 5.1, posterior summaries for each entry of the matrix Γ have been reported. These parameters capture the autoregressive effects of the latent variables for each symptom and are assumed to be time-invariant. We also assume that the latent variable of each symptom at a specific attendance visit is affected by the values of the latent variables of all symptoms at the previous attendance visit. In this way we want to control for the temporal interaction among symptoms.

We begin considering the autoregressive effects, *i.e.* the parameters governing the dependence between the latent variables at subsequent time points, namely $\gamma_{U,U}, \gamma_{P,P}, \gamma_{S,S}$ and $\gamma_{V,V}$. The posterior distributions of these parameters are concentrated on positive values and the largest posterior expectations are estimated for stress incontinence and voiding symptoms. The latter symptoms can be considered as the most recurrent ones. Pain symptoms instead appear to be the least recurrent having the smallest posterior mean. All posterior 95% credible intervals for these parameters do not contain zero.

On the contrary, cross-effects, *i.e.* the parameters governing the dependence between different symptoms at subsequent attendance visits, are often centred around zero (considering posterior 95% credible intervals). Exceptions are $\gamma_{U,V}$, $\gamma_{V,U}$, $\gamma_{S,P}$, which show a negative effect, and $\gamma_{U,S}$ and $\gamma_{S,V}$, which show positive effect.

TABLE 5.2: Summary of the posterior distributions of $\lambda_{d,m}$, with $d = \{U, P, S, V\}$ referring to the symptoms and $m = \{Age, Days\}$ to different covariates.

	mean	sd	2.5%	25%	50%	75%	97.5%
$\lambda_{U,\,Age}$	0.07	0.02	0.03	0.06	0.07	0.09	0.12
$\lambda_{P,\,Age}$	-0.13	0.02	-0.18	-0.15	-0.13	-0.12	-0.09
$\lambda_{S,\,Age}$	0.02	0.03	-0.04	-0.00	0.02	0.04	0.07
$\lambda_{V,\,Age}$	-0.00	0.03	-0.07	-0.02	-0.00	0.02	0.06
$\lambda_{U,\,Days}$	0.06	0.03	0.01	0.04	0.06	0.08	0.12
$\lambda_{P,\,Days}$	0.06	0.02	0.01	0.04	0.06	0.07	0.10
$\lambda_{S,\,Days}$	0.06	0.03	-0.00	0.04	0.06	0.08	0.11
$\lambda_{V,\,Days}$	0.05	0.04	-0.02	0.02	0.05	0.07	0.12

Table 5.2 reports posterior summaries for the regression coefficients of *Age* and *Days*. The former has a positive effect on the probability of observing urgency symptoms, while negative effect on the probability of observing pain

symptoms. Differently from these two cases, posterior 95% credible intervals for stress incontinence and voiding symptoms are centred around zero.

The period in days from the first visit attendance increases the probability of observing urgency and pain symptoms while it has no evident effect on stress incontinence and voiding symptoms as 95% credible intervals contain the value zero.

Predictive inference

We summarise the main results described in previous sections in Figure 5.4. This depicts different trajectories of the probability of observing the different

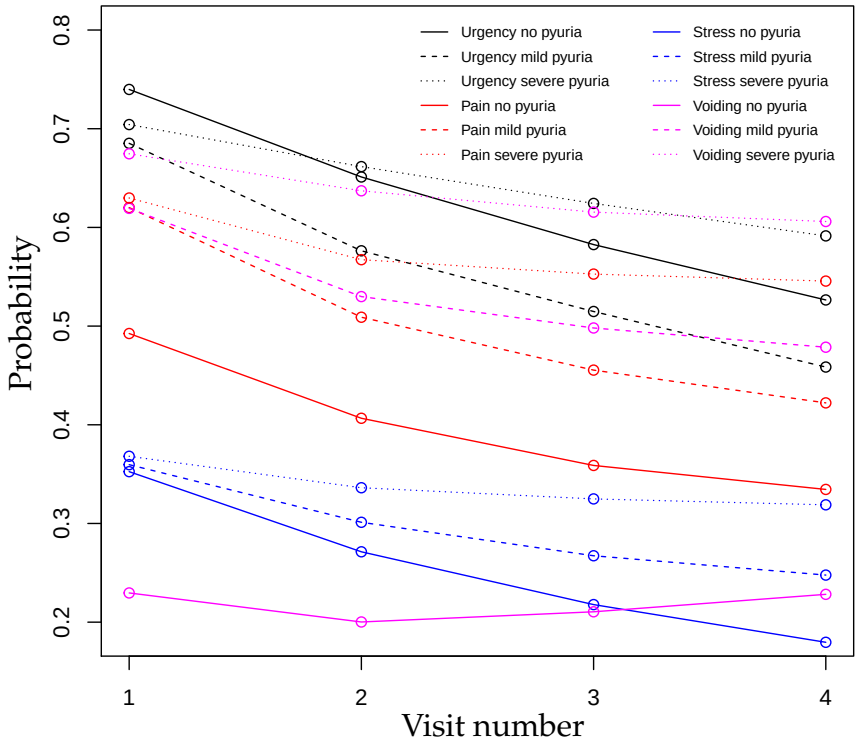


FIGURE 5.4: Marginal predictive probability for the four categories of symptoms at four subsequent attendance visits. We consider patients having for all four visits either no pyuria (solid lines) or mild pyuria (dashed lines) or severe pyuria (dotted lines). Colours correspond to different categories of symptoms: black lines correspond to urgency symptoms, red lines to pain, blue lines to stress incontinence and violet lines to voiding.

categories of symptoms over time, using marginal predictive distributions. We derive numerically these predictive distributions over the trajectories fixing *Age* equal to the mean observed value, as well as the mean value of *Days*. We then explore the probabilities for symptoms in three types of patients: no pyuria, mild pyuria and severe pyuria for all four the attendance visits.

The trajectories for all symptoms are decreasing over time, the only exception being voiding symptoms in patients with no pyuria where after a decrease in probability at the second attendance visit, the probability of observing symptoms slightly increase during the last two attendance visits. The group of symptoms which is mostly affected by the presence of pyuria is voiding, followed by pain. Both urgency and stress incontinence symptoms are less affected by the presence of pyuria. In particular, while in general both mild and severe pyuria increase the probability of observing stress incontinence symptoms, mild pyuria decreases the probability of observing urgency symptoms and severe pyuria instead initially decreases the probability of observing urgency symptoms compared to the case of no pyuria, but this probability increases over the subsequent attendance visits, relatively to no pyuria patients.

We conclude by saying that although the predictive distributions show decreasing trajectories for almost all symptoms, it is hard to conclude on the effectiveness of the treatments administered to the patients. This is because this study include a single treatment arm and it has only been designed to collect information about dependence among symptoms across time and the relation of symptoms with the degrees of infection.

5.5 Discussion

In this chapter we propose a method for modelling vectors of correlated binary variables evolving over time. The motivation for this work comes from a study of time evolving records of LUTS and connected risk factors. Similar to a traditional Probit model, we assume the binary variables to be distributed as a discretised version of continuous latent variables. We assumed the latent variables to be generated from an infinite mixture model, with weights that vary across a covariate space. This approach is different from the one proposed by DeYoreo and Kottas [2015], where the dependence from the covariates is obtained

assuming covariates to be random variables. Moreover, the time evolution of the symptoms is modelled using autoregressive and cross-effect terms.

This specification induces very flexible distributions for vectors of binary variables while allowing the user to maintain clear interpretation of the parameters of interest. Since the resulting model is a mixture of Gaussian distributions, covariates can be accommodated in the mean of each mixture component as regression terms. The same applies also for all autoregressive components.

At the latent variable level, the proposed model can be considered a non-parametric version of a mixed-effect model, where the group specific random effects are the mixing distributions and the groups are implied by different combinations of the binary covariates, similarly to ANOVA models. The choice of the stochastic process prior for the mixing distribution should reflect the type of information we want to include from covariates and the available prior information. In this work we choose the DGDP, for its flexibility (as demonstrated in Chapter 4) and the possibility of including an ANOVA model in the stick-breaking process of the weights. However, when the covariates are continuous the use of DGDP leads inevitably to overparameterised models, suggesting the use of alternative stochastic process priors. Examples are the Probit stick-breaking (Rodriguez and Dunson [2011]), Logit stick-breaking (Ren et al. [2011]) and Kernel stick-breaking (Dunson and Park [2008]) processes.

The results of the data analysis show that the different levels of pyuria strongly affect the probability of observing the symptoms, particularly voiding symptoms which seem to be the most probable group of symptoms to activate in case of infection. The effect of pyuria on the symptoms correlations is instead less clear, with some exception. In order to evaluate the temporal dependence among symptoms we summarise the posterior distribution of the autoregressive coefficients and we find that the most recurrent symptoms appear to be stress incontinence and voiding. We also evaluate cross-effects among symptoms occurring at subsequent times, finding strong interactions between urgency and voiding symptoms.

Finally, the proposed model formulation has been introduced for dealing with vectors of binary variables observed for a number of times which is equal for all individuals, as LUTS data set contains patients visited in four occasions.

However, the same model formulation can be easily extended to include sequences of binary vectors with heterogenous number of components across patients.

Chapter 6

Final remarks

We summarise the main contributions of this work to the literature of covariate dependent random measures and mixture modelling. We highlight for each of the previous chapters the adopted modelling strategy and we discuss the motivating applications and results. Then, we conclude by presenting open research questions for RPMx when mixed types of covariate are employed and when the objective is to assess relevant covariates. Finally, a possible extension to DGDP is briefly discussed.

6.1 Summary of the main contributions

This work focuses on covariate dependent random probability measures from a modelling perspective. The literature on this field has rapidly grown over the last two decades thanks to the development of suitable stochastic processes over random discrete probability measures indexed by some covariate space which could be used to specify infinite mixture models with covariate dependent mixing probabilities. The first contributions to this field can be found in MacEachern [1999] and [2000] (see also Cifarelli and Regazzini [1978]), which extended the DP to include covariate information in the precision parameter and in the centre measure, which could be then used to specify infinite mixture models. On the other hand, increasing research efforts have been focusing on the problem of including covariate information in a mixture model focusing on enriching the model over the partitions of the observations, *i.e.* when the random probability measure used to specify the mixture model is integrated out. This idea was first developed in the seminal paper of Müller et al. [1996], where the objective was to specify a flexible non-linear regression model using mixtures of distributions with covariate dependent weights.

In Chapter 1 we introduce the main ideas of Bayesian inference and BNP, reviewing in details the DP, which is the main stochastic process over discrete probability measures used in BNP, and its main application to mixture modelling. In the same chapter we also discuss the relevant literature about covariate dependent random measures and RPMx, highlighting the relation between these ideas.

Chapter 2 introduces the first contribution of this work which consists of an adaptation of RPMx to deal with zero-inflated observations, extending ZIP and ZINB linear regressions to the non-linear case. In particular, we specify a RPMx where observations are grouped in terms of both response and covariates. The sampling distribution within each cluster of the response is assumed to be a zero-inflated distribution. This modelling strategy is motivated by a data set containing counts of WBC in urine samples of patients suffering from LUTS. The aim is to infer the relation between LUTS and WBC counts and predict the WBC levels, which are the the best biomarker for UTI. So, the proposed strategy models jointly the LUTS and WBC levels as a DPM in order to flexibly estimate the conditional distribution of the response given the covariates.

Chapter 3 extends the class of RPMx to perform cluster specific variable selection. In particular, response and covariates are modelled jointly and the distribution employed for the response includes a linear regression of the covariates in the parameter governing the mean. We assume a spike and slab prior distribution for the regression coefficients within each cluster. The results of the variable selection are summarised by conditioning on a point estimate of the partition of the observations. The data set motivating this modelling strategy involves patients affected by UTI (*i.e.* WBC≥ 1) and suffering from LUTS. The idea is to stratify the patients in different groups containing individuals with similar levels of WBC and similar LUTS profiles and to investigate which symptoms are connected with the various degrees of the infection.

In the remaining chapters we focus on strategies to include covariates directly into the random probability measures. Chapter 4 is dedicated to introduce a novel stochastic process whose realisations are covariate dependent random probability measures which marginally (for each level of the covariates) follow a GDP. The latter is a generalisation of the DP that employs a richer parameterisation and shows appealing property in terms of implied partition of the observations. In particular, the parameters can control simultaneously

the size and the number of the clusters, resulting in a more flexible partition. The covariates are included within the GDP parameterising the Beta random variables of the stick-breaking process of the weights in terms of means and precisions and including regressions within those means. We apply the resulting process, namely DGDP, to a data set involving pediatric patients affected by ALL in order to study the effect of asparginase treatment to the risk of developing osteonecrosis. The latter is controlled by the levels of triglycerides during the continuation period of the treatment. We also apply the DGDP to assess the most relevant determinants of Ofsted evaluations of primary schools in London. In particular, we specified a DGDP that could capture borough specific effects (treated as confounders) as well as the dependence across neighbouring boroughs.

In Chapter 5 we extend the latent variables models for correlated binary variables (such as the Probit and Logit models), assuming the latent distribution to be a collection of infinite mixture of SUR models with Gaussian errors and mixture weights that could vary across a covariate space. We use SUR models in order to account for the time evolution of the latent variables including autoregressive terms. The cluster-specific parameters are assumed to be the intercepts of the latent variables and the covariance matrices. In this way, the use of covariate dependent random measures implicitly defines a dual regression, *i.e.* a regression on the parameters governing the mean and the covariance of the distribution. Indeed, the covariates could affect both the mean levels of the latent variables and the covariance. The proposed model is motivated by a data set containing LUTS observations along with covariates at four subsequent attendance visits. The objective is to improve the understanding about the evolution of LUTS through time, assessing which class of symptoms is more recalcitrant as well as investigating the correlation among symptoms at each attendance visit and across attendance visits.

6.2 Open research questions

In this section we outline some open research questions which can be relevant in deepening the understanding and applicability of covariate dependent mixture models.

6.2.1 RPMx with mixed covariates

All RPMx models proposed in the literature, included those presented in this work, can in principle include all types of covariates. When covariates are included in the prior distribution over the partition of observations this can be done by adapting the similarity function in (1.4.3). In particular, when the similarity function is chosen to be the joint density of the covariates in a cluster (*e.g.* in PR models), this requires the inclusion of suitable distributions within the likelihood.

Although the use of density functions to capture the similarity within the covariates is well established, the literature does not present practical considerations on which types of covariates tend to dominate the clustering structure of the observations. The latter point becomes even more relevant in light of one of the main drawbacks of the RPMx constructed by modelling the covariates, that is the likelihood of the joint model may be dominated by the information contained within the covariates becoming insensitive to the patterns in the response variable. In addition, the literature does not present a way to include ordinal discrete covariates. Although this should not be problematic, it includes an extra-level of complexity given that very often discrete distributions for ordinal categorical variables are represented by discretising continuous latent variables (often Gaussian).

When covariates are included directly into the random measure (in the construction of the weights), there is no reason to think that different types of covariates should affect differently the clustering of the observations. In this setting problems arise when covariates are continuous. Indeed, methods like DDP and DGDP lead to overparameterised models, where each observation is equipped by a unique random probability measure. Also the PSBP, which in principle does not suffer from the latter limitation, may encounter difficulties in estimating parameters in the parts of the covariate space where observations are sparse.

6.2.2 Variable selection in RPMx

Variable selection in covariate dependent mixture models is an increasing area of research, but the number of contributions is still relatively limited. For augmented response models solutions are proposed for PPMx and PR. In these

models the covariates are included within the prior probability over the partition of the observations, while the sampling models do not include covariates. The role of the covariates is in favouring *a priori* observations presenting *similar* covariates to be assigned to the same cluster, *i.e.* to be included under the same mixture component. This is often achieved by modelling jointly the response variable $\boldsymbol{y}$ and the covariates $\boldsymbol{X}$, reporting inference for the random quantity $\boldsymbol{y} \mid \boldsymbol{X}$, which represents a (non-linear) regression model. Advantages and drawbacks of modelling covariates jointly with the response are discussed in Chapter 1.

In such a setting, one would expect that the role of the covariate would be to identify clusters of observations having similar response values, while being characterised also by similar covariates. Equivalently, from a variable selection perspective, relevant covariates should be those relevant in separating the response values in different groups. However, the variable selection techniques introduced for these models select the covariates for their importance in identifying the clustering structure of the observations characterised by response and covariate values, this being a limitation of the available methodology.

We illustrate a consequence of the latter point with an example. Let consider a data set containing a univariate response variable and three binary covariates. We assume that the response, y_i, with $i = 1, \ldots, n$, is independent of the covariates and generated according to a Normal distribution with mean and variance equal to 5 and 1, respectively. Then, we assume one of the covariates, $x_{1,i}$, to be independent from the other covariates and generated according to a Bernoulli with mean equal to 0.5. The remaining covariates, $x_{2,i}$ and $x_{3,i}$, are instead dependent. In particular, we assume that $p(x_{i,3} = 1 \mid x_{i,2} = 1) = 0.8$ and $p(x_{i,3} = 1 \mid x_{i,2} = 0) = 0.3$. We generate $x_{i,2}$ according to a Bernoulli with mean equal 0.5. We fit the PR in (3.2.2) on the simulated data, including the modification of the likelihood in (1.4.7) in order to perform variable selection. The relevant posterior distributions for this discussion are in Figure 6.1. The posterior distribution of the partition of the observations indicates that the configuration of the partition of the observations with the highest posterior probability has two clusters (top-left panel). These are inferred by the model to capture the effect of the interaction among the covariates. Indeed, neither the response nor the covariates are generated from mixture of distributions, and additionally the

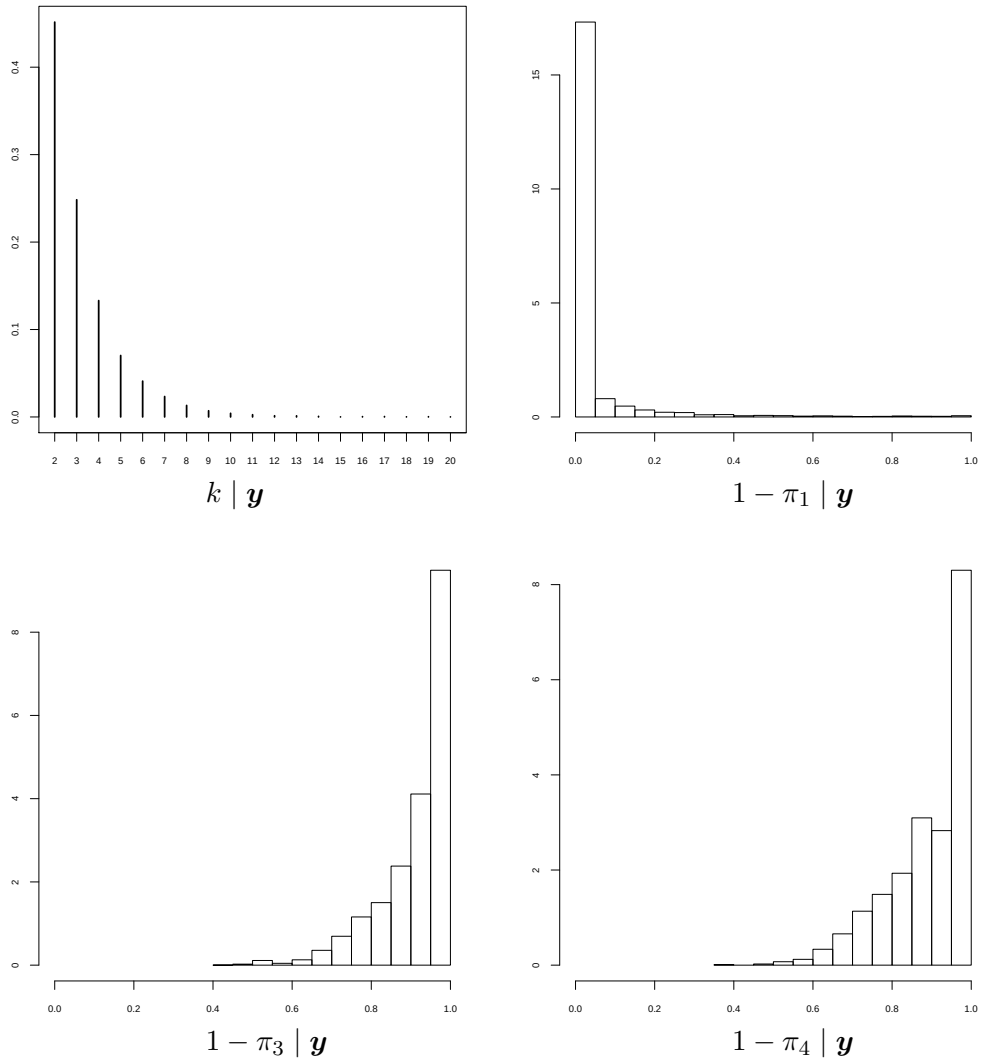


FIGURE 6.1: Posterior distributions of the number of clusters, k, (*top-left panel*) and of the inclusion probability of the covariates, $1 - \pi_1$ (*top-right panel*), $1 - \pi_2$ (*bottom-left panel*) and $1 - \pi_3$ (*bottom-right panel*).

response shows similar values in these clusters. The result of the variable selection, summarised by looking at the probabilities of covariate inclusion $1 - \pi_1$, $1 - \pi_2$ and $1 - \pi_3$, indicates that the second and third covariate are relevant for the model (bottom panels). However, we know that this happens because of the dependence between these two covariates.

An additional problem with these models arises when a regression model is included in each cluster specific distribution of the response. In fact, both

variable selection procedures for PR and PPMx remove the effect of the covariates on the partition of the observations, while the same covariates are still present in the model. This problem is circumvented in the modelling strategy proposed in Chapter 3. In that case, the outcome of the variable selection is evaluated only in terms of the linear effects of the covariates on the mean of the response in each cluster, after fixing a relevant configuration of the partition of the observations according to some criteria.

Both the problems described above have been addressed by the PSBP, *i.e.* when covariates are included in the weights of the random measure. For this method the variable section is induced by the use of latent indicators (see (3.2.4)) which link the results of the selection in terms of clustering and of cluster-wise linear regressions. However, results are often sensitive to the level of truncation. In particular, including a large number of components in the truncated process, which in principle should lead to a better approximation of the authentic nonparametric structure, the variable selection procedure tends to include a larger number of covariates. Chung and Dunson [2009] stated that this arises because covariates may become relevant for low probability components of the covariate dependent mixture model, but this is not a desirable effect.

6.2.3 Extension to the DGDP

The DGDP has been introduced in details in Chapter 4, and demonstrated in two real world applications. A possible extension is presented in Chapter 5 where we used a DGDP to model multiple longitudinal binary vectprs. This has been done by including an additional dimension in the covariates corresponding to the labels of each longitudinal series. For a similar applied problem, GDP was applied in Rodriguez and Dunson [2014] to extend the framework of the nested DP (Rodriguez et al. [2008]).

An additional extension of the GDP consists in letting the precision parameters to vary across the covariate space. This is common in the double-GLM framework. The stick-breaking process in (4.2.2) then becomes

$$v_{h,x} \sim \text{Beta}(v \mid \phi(x)\mu(x), \phi(x)(1-\mu(x))),$$

so that similarly to $\mu(x)$, also $\phi(x)$ is a random function of x with codomain being in the positive real numbers. A natural choice for $\phi(x)$ is $\exp(\phi_1 + \phi_2 x)$.

Appendix A

Appendix for Chapter 2

A.1 JAGS code for BNP-ZIP

We provide below the JAGS code for BNP-ZIP.

```
model{
  C <- 10000 # Zero-trick (see Neelon et al, 2010)

for(i in 1:N) {
    z[i] <- step(y[i] - 1) # Indicator function for y>0

    lambda[i] <- lamj[g[i]] # Parameter for Poisson distribution
    p.y[i] <- muj[g[i]] # Parameter for zero-inflation

##### ZIP model #####
    pz.y[i] <- p.y[i]*(1 - exp(-lambda[i])) # Probability of y>0

    ll[i] <- (1 - z[i]) * log(1 - pz.y[i]) + z[i] * (log(pz.y[i])
             + y[i] * log(lambda[i]) - lambda[i] - loggam(y[i] + 1)
             - log(1 - exp(-lambda[i]))) # Log-likelihood

    phs[i] <- -ll[i] + C  # Zero-trick

    zeros[i] ~ dpois(phs[i]) # "zero" is a vector with n zeros

##### Covariate model #####
    for(p in 1:P) {
      x[i,p] ~ dbern(phi[g[i],p])
    }
```

```
    g[i] ~ dcat(psi[]) # distribution over the cluster assignment
}


##### Within-cluster priors ######
    for(clus in 1:K) {
      muj[clus] ~ dbeta(1,1)I(0.01,0.99)
      lamj[clus] ~ dgamma(1,1)
        for(p in 1:P) {
          phi[clus,p] ~ dbeta(1,1)I(0.01,0.99)
        }
    }

##### Dirichlet Process Prior #####
  alpha ~ dgamma(1,1)
  for(clus in 1:(K - 1)) {
    V[clus] ~ dbeta(1,alpha)
  }
  psi[1] <- V[1]                                    # Stick breaking
  for(clus in 2:(K - 1)) {
    psi[clus] <- V[clus] * (1 - V[clus-1]) * psi[clus-1] / V[clus-1]
  }
  psi[K] <- 1 - sum(psi[1:(K - 1)])
}
```

This uses a trick to code the Zero-Inflated Poisson model which was employed in the `WinBUGS` code of Neelon et al. [2010] (available at `http://people.musc.edu/~brn200/winbugs/`).

Appendix B

Appendix for Chapter 3

B.1 Posterior inference for RPMS

In this appendix we present the details of the updating steps of the Gibbs sampler scheme adopted.

Membership Indicator

This step follows the updating Gibbs-type algorithm for DPM with non conjugate base measure in Neal [2000] called *auxiliary parameter* approach. Let first define s_i to be the membership indicator for the observation i and $\boldsymbol{s}_{(i)}$ to be the vector of the membership indicator for the n observations but from which s_i is removed. Let us also define k^- to be the number of clusters when i is removed, n_j^- for $j = 1, \ldots, k^-$ to be the cardinality of the clusters when i is removed. Thus the full conditional distribution for each indicator is:

$$p(s_i \mid \boldsymbol{s}_{(i)}, \boldsymbol{\beta}^*, \boldsymbol{\zeta}^*, \boldsymbol{X}, \boldsymbol{y}, \lambda)$$
$$\propto \begin{cases} n_j p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}_j^*, \lambda) \prod_{d=1}^{D} p(x_{id} \mid \zeta_{jd}^*) & j = 1, \ldots, k^- \\ \frac{\alpha}{M} p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}_m^*, \lambda) \prod_{d=1}^{D} p(x_{id} \mid \zeta_{md}^*) & j = k^- + 1 \text{ and } m = 1, \ldots, M \end{cases},$$

where $(\boldsymbol{\beta}_m^*, \boldsymbol{\zeta}_m^*)$ for $m = 1, \ldots, M$ are draws from the the base measure in (3.3.2).

Precision of the DP

In order to update the parameter α of the DP we need to introduce an additional parameter u such that $p(u \mid k, \alpha) = \text{Beta}(\alpha + 1, n)$ (see Escobar and West [1995] for detailed explanation). We can sample from the full conditional:

$$p(\alpha \mid u, k) = \xi \text{Gamma}(a_\alpha + k, b_\alpha - \log(u)) + (1 - \xi)\text{Gamma}(a_\alpha + k - 1, b_\alpha - \log(u)),$$

where $\xi = (a_\alpha + k - 1)/(\alpha_1 + k - 1 + nb_\alpha - n\log(u))$.

Covariate Parameters

For the update of the parameters of the covariates, we work separately for each of the D covariates within each of the k clusters. Thus the full conditional posterior distributions are:

$$p(\zeta^*_{jd} \mid \boldsymbol{x}^*_{jd}) \propto \prod_{i \in S_j} \text{Bernoulli}(x_{id} \mid \zeta^*_{jd}) \text{Beta}(\zeta^*_{jd} \mid a_\zeta, b_\zeta)$$

$$\zeta^*_{jd} \mid \cdot \sim \text{Beta}\left(\zeta^*_{jd} \mid a_\zeta + \sum_{i \in S_j} x_{id}, b_\zeta - \sum_{i \in S_j} x_{id} + n_j\right),$$

for $j = 1, \ldots, k$ and $d = 1, \ldots, D$.

Regression Coefficients

As for the case of the parameters of the covariates, we consider separately each of the D covariates and each of the k clusters. It follows that:

$$p(\beta^*_{jd} \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta}^*_{j(d)}) \propto \prod_{i \in S_j} \text{Normal}(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}^*_j, \lambda)[\omega_d r_\pi \delta_0(\beta^*_{jd}) + (1 - \omega_d r_\pi)\text{Normal}(\beta^*_{jd} \mid m_d, \tau_d)] =$$

$$= \prod_{i \in S_j} \omega_d r_\pi \delta_0(\beta^*_{jd}) N(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}^*_j, \lambda) + (1 - \omega_d r_\pi)\text{Normal}(\beta^*_{jd} \mid m_d, \tau_d)\text{Normal}(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}^*_j, \lambda),$$

$\boldsymbol{\beta}^*_{j(d)}$ is the vector $\boldsymbol{\beta}^*_j$ where the dth component is removed. The first part of the last equation will be 0 with some probability. Let us consider the second

part of that equation. This is proportional to

$$\exp\left\{-\frac{1}{2}\tau_d(\beta^*_{jd}-m_d)^2\right\}\exp\left\{-\frac{1}{2}\sum_{i\in S_j}\lambda(y_i-\boldsymbol{x}_{i(d)}\boldsymbol{\beta}^*_{j(d)}-x_{id}\beta^*_{jd})^2\right\}=$$

$$=\exp\left\{-\frac{1}{2}\left[\beta^{*2}_{jd}(\tau_d+\lambda x^2_{id})-2\beta^*_{jd}\left(m_d\tau_d+\lambda\sum_{i\in S_j}(x_{id}A_i)\right)\right]\right\},$$

with $A_i=(y_i-\boldsymbol{x}_{i(d)}\boldsymbol{\beta}^*_{j(d)})$, $\boldsymbol{x}_{i(d)}$ is the vector $\boldsymbol{x}_i$ where the dth is removed.

Thus, the full conditional probabilities for the Gibbs sampler are:

$$\beta^*_{jd}\mid\cdot=\begin{cases}0 & \text{w. p. } \theta_{jd}\\ \sim\text{Normal}\left(\frac{m_d\tau_d+\sum_{i\in S_j}(\lambda x_{id}A_i)}{\tau_d+\sum_{i\in S_j}(\lambda x^2_{id})},\tau_d+\sum_{i\in S_j}(\lambda x_{id})\right) & \text{w. p. } (1-\theta_{jd})\end{cases}.$$

Finally the weights $\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k)$ are:

$$\theta_{jd}=\frac{\omega_d r_\pi}{\omega_d r_\pi+(1-\omega_d r_\pi)C},$$

with C:

$$C=\sqrt{\left(\tau_d+\sum_{i\in S_j}(x^2_{id}\lambda)\right)^{-1}\tau_d}$$
$$\exp\left\{-\frac{1}{2}\tau_d m_d^2+\frac{1}{2}\left(\tau_d+\sum_{i\in S_j}(x^2_{id}\lambda)\right)^{-1}\cdot\left(m_d\tau_d+\sum_{i\in S_j}(\lambda x_{id}A_i)\right)\right\}.$$

Weights of the Spike and Slab Prior

Following what was done in Kim et al. [2009] let us call $r_d=\omega_d r_\pi$ and $r_\pi=a_\pi/(a_\pi+b_\pi)$.

$$p(r_d)=\text{Beta}\left(\frac{r_d}{r_\pi}\mid a_\pi,b_\pi\right)\frac{1}{r_\pi}=$$

$$\frac{1}{\mathrm{B}(a_\pi, b_\pi)} \left(\frac{1}{r_\pi}\right)^{a_\pi + b_\pi - 1} r_d^{a_\pi - 1}(r_\pi - r_d)^{b_\pi - 1}.$$

The full conditional is the following:

$$p(r_d \mid \boldsymbol{\beta}_d^*) \propto p(r_d) r_d^{\sum_j 1(\beta_{jd}^* = 0)} (1 - r_d)^{\sum_j 1(\beta_{jd}^* \neq 0)}.$$

This is an unknown distribution and a draw from it is obtainable computing the inverse of the cumulative distribution function over a grid of values. We select the point on the grid that gives the closest value of the inverse cumulative distribution function to a draw from a uniform distribution on $(0, 1)$.

Precision of the Base Measure

We update the precision of the normal part of the base measure considering separately each of the D covariates.

$$p(\tau_d \mid \boldsymbol{\beta}_j^*, a_\tau, b_\tau) \propto \mathrm{Gamma}(\tau_d \mid a_\tau, b_\tau) \prod_{j=1}^{k} [\omega_d r_\pi \delta_0(\beta_{jd}^*) + (1 - \pi_d r_\omega) N(\beta_{jd}^* \mid m_d, \tau_d)]$$

$$= \mathrm{Gamma}(\tau_d \mid a_\tau, b_\tau) \prod_{j=1}^{n_d^+} N(\beta_{jd}^+ \mid m_d, \tau_d),$$

where n_d^+ is the number of clusters that have non-zero coefficients in position d, whereas β_{jd}^+ for $j = 1, \ldots, n_d^+$ is the list of these non zero coefficients. Thus, it is possible to draw from the following known distribution:

$$\tau_d \mid \boldsymbol{y}, \boldsymbol{\beta}_j^*, a_\tau, b_\tau \sim \mathrm{Gamma}\left(\tau_d \mid a_\tau + \frac{n_d^+}{2}, b_\tau + \frac{1}{2}\sum_{j=1}^{n_d^+} (\beta_{jd}^+ - m_d)^2\right).$$

Precision of the Regression

The precision of the regression is updated in a conjugate form as it follows:

$$p(\lambda \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}^*, a_\lambda, b_\lambda) \propto \prod_{j=1}^{k} \prod_{i \in S_j} \text{Normal}(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}_j^*, \lambda) \text{Gamma}(\lambda \mid a_\lambda, b_\lambda)$$

$$\lambda \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}^*, a_\lambda, b_\lambda \sim \text{Gamma}\left(\lambda \mid n/2 + a_\lambda, \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i \beta_i)^2/2 + b_\lambda\right).$$

B.2 Simulation Study

We present a simulation study to show the performances of RPMS in terms of variable selection. Specifically, we discuss the estimated posterior distribution of the regression coefficients: $p(\beta_{\cdot d} \mid \boldsymbol{y}, \boldsymbol{X})$. We check the ability of the RPMS in identifying the coefficients which are different than zero within different clusters and we underline the advantages of modelling the covariates as random variables. We compare the results of the proposed model with those achievable using the SSP model.

Data Generation

We consider $n = 300$ observations. We consider a partition of the observations constituted by two clusters, *i.e.* $\rho_n = \{S_1, S_2\}$, and we assign half of the observations to each cluster (thus, $n_1 = n_2 = 150$). A matrix of covariates, $\boldsymbol{X}$, is generated assuming each entry $x_{id} \sim \text{Bernoulli}(x_{id} \mid \zeta_{id})$ for $i = 1, \ldots, n$ and $d = 1, 2, 3$, along with an intercept. The parameters of the covariate model are assigned according to the following table:

TABLE B.1: Cluster-specific parameters for the covariate model in the simulation study.

	ζ_{i1}	ζ_{i2}	ζ_{i3}
$i \in S_1$	0.9	0.2	0.5
$i \in S_2$	0.2	0.9	0.5

We generate a response vector with components $y_i \sim \text{Normal}(\boldsymbol{x}_i \boldsymbol{\beta}_{id}, \lambda = 1)$, for $i = 1, \ldots, n$. We set the intercept equal to zero in both clusters and we consider the following cluster-specific regression coefficients:

TABLE B.2: Cluster-specific parameters for the response model in the simulation study.

	β_{i1}	β_{i2}	β_{i3}
$i \in S_1$	5	0	8
$i \in S_2$	0	-5	0

Results for RPMS

In this section we present and discuss the performance of the RPMS in the simulation study designed in the previous section. We focus on the performance in terms of variable section. The latter can be assessed by displaying the posterior distribution of the regression coefficients. Given the fact that the RPMS models also the covariates, it is convenient to concentrate on the posterior distribution of the parameters given the observed combinations of the covariates within $\boldsymbol{X}$.

In our simulation $\boldsymbol{X}$ is a binary matrix with three columns and we observe eight possible combinations of values. The latter is illustrated in Table B.3 in which labels for different combinations are also given.

TABLE B.3: Observed combinations of values within the covariate matrix $\boldsymbol{X}$

	$x_{\cdot 1}$	$x_{\cdot 2}$	$x_{\cdot 3}$
l_1	1	0	1
l_2	1	1	1
l_3	1	0	0
l_4	1	1	0
l_5	0	0	1
l_6	0	0	0
l_7	0	1	0
l_8	0	1	1

In Figure B.1 we present the samples of the posterior distributions for the three regression coefficients for each combination of the covariates, *i.e.* $l_1, \ldots, l_8$. Using the proposed model the covariates inform the clustering estimation through their auxiliary model. In particular, observations with similar profile of covariates have a priori higher probability to co-cluster.

The latter has effects which become evident considering, for instance, the first row of panels in Figure B.1, corresponding to the posterior distributions associated with the covariates combination $l_1 = (1, 0, 1)$. This profile has a large

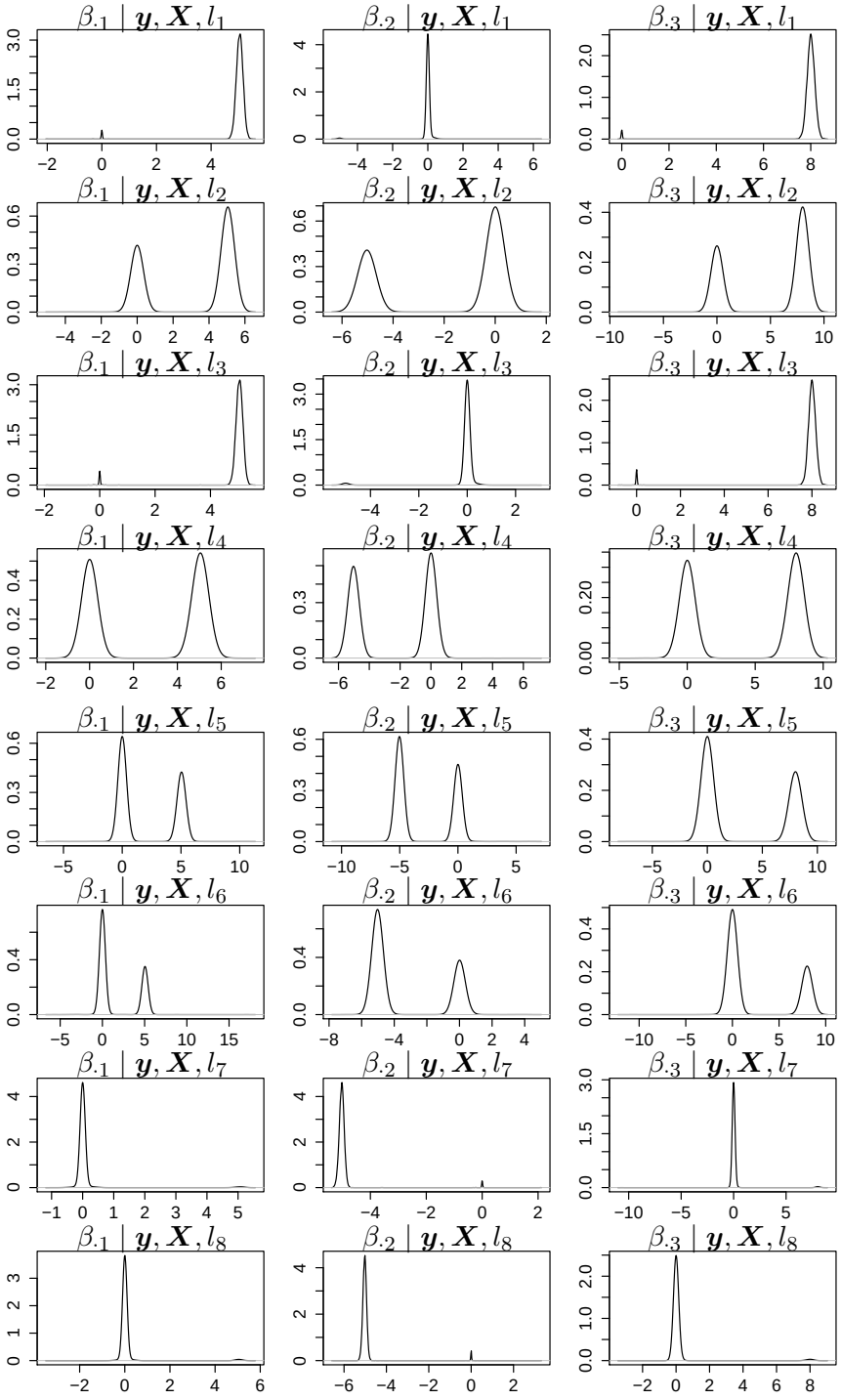


FIGURE B.1: Posterior densities of the regression coefficients for the RPMS for all combinations of the covariates $l_1, \ldots, l_8$.

probability to have been generated within cluster S_1, because the first covariate is activated while the second one is not, reflecting the first row of probabilities in Table B.1. For this reason the panel corresponding to the first regression coefficients is dominated by a distribution with mean 5, which is equal to the correct $\beta_{\cdot 1}$ for cluster S_1 in Table B.2. Equivalently, the second panel shows a spike in correspondence with $\beta_{\cdot 2} = 0$ and the third one a component with mean 8, reflecting the true values for $\beta_{\cdot 2}$ and $\beta_{\cdot 3}$, respectively, in cluster S_1.

Consequently, the possibility of modelling the covariates influences the variable selection within each cluster, resulting in a higher accuracy while selecting the important covariates. This is evident in Figure B.1, for instance taking the level $l_8 = (0, 1, 1)$. The panel corresponding to $\beta_{\cdot 3}$ shows a spike in correspondence of 0, which means that the third covariate is not important explaining the response for observations with profile l_8. However we know that the in general the third regression coefficient could be either 8 or 0. The strong evidence in favour of excluding the $x_{\cdot 3}$ is given by modelling the level l_8: this profile has a larger probability to have been generated in the cluster S_2.

Employing a spike and slab prior allows the user to estimate the probability for a regression coefficient to be exactly 0. The results relative to this simulation study are reported in Table B.4.

TABLE B.4: Empirical posterior probability of the regression coefficients to be equal to 0 for different combinations of the covariates.

	$p(\beta_{\cdot 1} = 0)$	$p(\beta_{\cdot 2} = 0)$	$p(\beta_{\cdot 3} = 0)$
l_1	0.01	0.93	0.01
l_2	0.36	0.59	0.37
l_3	0.02	0.92	0.02
l_4	0.45	0.50	0.46
l_5	0.56	0.39	0.57
l_6	0.64	0.32	0.65
l_7	0.94	0.01	0.96
l_8	0.93	0.02	0.95

In Table B.5 we record the observed frequency (induced by the data generating process) of the assignment of the different profiles to the two clusters. From this table, the profile l_1 is 98% of the times generated within cluster S_1 and consequently we expect the posterior of the regression coefficients given l_1 to be different than 0 for $\beta_{\cdot 1}$ and $\beta_{\cdot 3}$ whereas equal to 0 for $\beta_{\cdot 2}$. Indeed, this is

confirmed looking at the posterior probabilities in Table B.4, which show that the probability of $\beta_{\cdot 1}$ and $\beta_{\cdot 3}$ to be equal to 0 is equal to 0.01 and for $\beta_{\cdot 2}$ is 0.93.

TABLE B.5: Observed assignment of the different levels of covariates to clusters S_1 and S_2.

	$p(l_\cdot \in S_1)$	$p(l_\cdot \in S_2)$
l_1	0.98	0.02
l_2	0.63	0.37
l_3	1.00	0.00
l_4	0.52	0.48
l_5	0.44	0.56
l_6	0.40	0.60
l_7	0.00	1.00
l_8	0.00	1.00

Results for SSP

In order to highlight the advantages of using the RPMS, we compare the results presented in the previous section with the analogous analysis under the SSP model. The latter is similar to the RPMS, but it does not employ a model on the covariates. In other words the covariates do not contribute to the clustering (and hence to the variable selection). In this section we present the results of the SSP in terms of variable selection when applied to the same simulated data that we used in the previous section.

Also for SSP the variable selection output can be summarised exploring the posterior distribution of the regression coefficients, which are displayed in Figure B.2. Differently from the RPMS, this does not depend on the covariates information. This has two main consequences: (i) it presents more uncertainty in the variable selection and (ii) the posterior distributions are not robust when interactions within the covariates are present.

The first point can be exemplified considering the posterior distribution of $\beta_{\cdot 3}$. In Figure B.2 this corresponds to the right panel, which shows a two component mixture of a spike with location in 0 and probability 0.52 and a slab distribution with mean 8 and probability 0.48. Although this reflects the way in which we assign the regression coefficients to the observations, it can be a poor result once a new set of covariate becomes available. This is because we

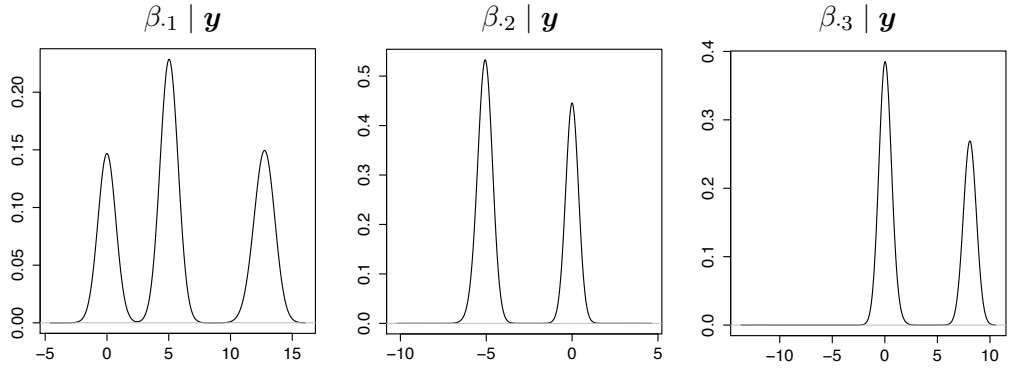


FIGURE B.2: Posterior densities of the regression coefficients for SSP.

assign, independently on the observed combination of new covariates, approximately 50% of the probability to $\beta_{\cdot 3} = 0$, leading to limited predictive power. On the other hand, in Figure B.1 the posterior distribution of $\beta_{\cdot 3}$ is displayed across different scenarios determined by different combinations of the covariates. This allows us to identify with less uncertainty when the variable selection is performed, leading to better predictions.

The second point can be visualised by looking at the left panel in Figure B.2, corresponding to the posterior distribution of $\beta_{\cdot 1}$. Surprisingly a third component with mean 13 is displayed, which is the sum of the effects of the covariates in cluster S_1. This bias in the posterior distribution is given by that SSP considers also the case in which $\beta_{\cdot 3}$ is set equal to 0 and the entire effect on the response is mistakenly attributed to $\beta_{\cdot 1}$. So, ignoring the interaction within the covariates (in fact if $x_{i1} = 1$ is likely that $\beta_{\cdot 3}$ is different from 0) leads to a bias which is not suffered by the RPMS. The latter, modelling the covariates, can identify profile-specific posterior distributions and thus it is able to assign the correct effect to the parameters.

Appendix C

Appendix for Chapter 4

C.1 Asymptotic behaviour of K_n

In this appendix, we study the large n asymptotic behaviour of k, which is the random number of ties in a sample of size n from G, assuming $G \sim \text{GDP}(\phi, \mu, G_0)$. This can be easily characterised by resorting to Karlin [1967]. In particular, k is the number of boxes occupied by n balls, with probability $P_i = v_i \prod_{1 \leq j \leq i-1}(1 - v_j)$ of a ball being in the box i. Here, v_i is a Beta random variables with parameters $(\phi\mu, \phi(1-\mu))$. Recall that Karlin [1967] considered non-random occupational probabilities, i.e., non-random P_i for any $i \geq 1$. From the strong law of large number, as $i \to +\infty$

$$-\frac{1}{i}\log(P_i) = -\frac{1}{i}\log(v_i) - \frac{1}{i}\sum_{j=1}^{i-1}\log(1-v_j) \to -\mathbb{E}[\log(1-v)], \qquad \text{(C.1.1)}$$

with $v \sim \text{Beta}(\phi\mu, \phi(1-\mu))$. We are using the fact that the values of $\phi\mu$ and $\phi(1-\mu)$, both greater than zero, are such that $\int_0^1 |\log(v)| p(v) \mathrm{d}v < +\infty$. We can compute explicitly the expectation in (C.1.1), i.e.

$$-\mathbb{E}[\log(1-v)] = \psi^{(0)}(\phi) - \psi^{(0)}(\phi(1-\mu))$$

as $i \to +\infty$, where $\psi^{(0)}(x)$ denotes the polygamma function, i.e. the first derivative of the logarithm of the Gamma function with respect to x. Then, see Section 2 in Karlin [1967], as $n \to +\infty$,

$$|\{i : P_i > n^{-1}\}| \overset{\text{a.s.}}{\sim} \frac{\log(n)}{\psi^{(0)}(\phi) - \psi^{(0)}(\phi(1-\mu))}.$$

Furthermore, by Theorem 8 in Karlin [1967], one has $k \,|\, (P_i)_{i\geq 1} \overset{\text{a.s.}}{\sim} [\psi^{(0)}(\phi) - \psi^{(0)}(\phi(1-\mu))]^{-1}\log(n)$ as $n \to +\infty$ and, hence, also $k \overset{\text{a.s.}}{\sim} [\psi^{(0)}(\phi) - \psi^{(0)}(\phi(1-\mu)]^{-1}\log(n)$ as $n \to +\infty$. That is,

$$\frac{k}{\log(n)} \to \frac{1}{\psi^{(0)}(\phi) - \psi^{(0)}(\phi(1-\mu))} \tag{C.1.2}$$

almost surely, as $n \to +\infty$.

C.2 Randomly truncated GDP

Consider the stochastic process defined in (4.2.10). When $\phi = \mu^{-1}$, Muliere and Tardella [1998] proved that N_ε is a Poisson random variable with parameter $-(\mu^{-1}-1)\log\varepsilon$. In general, the distribution of N_ε is not simple. Specifically,

$$\begin{aligned} N_\varepsilon &= \inf\left\{ n \in \mathbb{N} : \sum_{h=1}^{n} w_h > 1-\varepsilon \right\} \\ &= \inf\{n \in \mathbb{N} : R_\varepsilon < \varepsilon\} \\ &= \inf\{n \in \mathbb{N} : \log R_\varepsilon < \log\varepsilon\} \end{aligned}$$

is the number of arrivals at a time $-\log\varepsilon$ in a renewal process whose inter-arrival times have the same distribution as the random variable $w = -\log(1-v)$, where $v \sim \text{Beta}(\phi\mu, \phi(1-\mu))$. That is w has the following distribution

$$p(w) = \frac{\Gamma(\phi)}{\Gamma(\phi\mu)\Gamma(\phi(1-\mu))}(1-\mathrm{e}^{-w})^{\phi\mu-1}\mathrm{e}^{-\phi(1-\mu)w}\mathrm{d}w$$

with

$$\mathbb{E}[w] = \psi^{(0)}(\phi) - \psi^{(0)}(\phi(1-\mu))$$

and

$$\mathbb{V}[w] = \psi^{(1)}(\phi(1-\mu)) - \psi^{(1)}(\phi),$$

where $\psi^{(m)}(x)$ denotes the polygamma function, i.e. the derivative of order $(m+1)$ of the logarithm of the Gamma function with respect to x. Note that if $\phi = \mu^{-1}$, then W is a negative exponential random variable with parameter

$(\mu^{-1}-1)$ and, accordingly, N_ε becomes the number of arrivals at time $-\log\varepsilon$ for a Poisson process with rate $(\mu^{-1}-1)$.

Although the distribution of N_ε does not have a simple expression, we can say something about the distribution of N_ε for small ε. Indeed, according to the definition (4.2.10), we are interested in small ε so that G_ε is a good approximation of G. Let Z be a standard Gaussian random variable. By the central limit theorem for renewal processes, as $\varepsilon \to 0$

$$\frac{N_\varepsilon - \frac{-\log\varepsilon}{\mathbb{E}[w]}}{\left(-\frac{\mathbb{V}[w]\log\varepsilon}{(\mathbb{E}[w])^3}\right)^{1/2}} \to Z$$

i.e.,

$$N_\varepsilon \approx \left(-\frac{\mathbb{V}[w]\log\varepsilon}{(\mathbb{E}[w])^3}\right)^{1/2} Z - \frac{\log\varepsilon}{\mathbb{E}[w]}$$

for small ε. Note that if $\phi = \mu^{-1}$, then $\mathbb{E}[w] = (\mu^{-1}-1)^{-1}$ and $\mathbb{V}[w] = (\mu^{-1}-1)^{-2}$. One then recovers the well-known Gaussian approximation of a Poisson random variable with parameter $-(\mu^{-1}-1)\log\varepsilon$ as $\varepsilon \to 0$. Figure C.1 depicts some numerical illustrations of the distribution of N_ε for $\varepsilon = 0.1, 0.01, 0.001, 0.0001$, and for fixed $\phi = 1$ and $\mu = 0.5$.

C.3 Slice sampling for DGDP

In this appendix we present an algorithm for posterior inference that relies on an extension of the slice sampler presented by Walker [2007], namely the dependent slice-efficient sampler introduced by Kalli et al. [2011]. This algorithm employs most of the steps presented in Section 4.4, but it does not truncate deterministically G_x. Instead, we augment the parameter space with a uniformly distributed latent variables u_i $(i = 1, \ldots, n)$ in such a way that the joint distribution of the parameter and the latent variable becomes (for each $i = 1, \ldots, n$)

$$(\theta_i^*, u_i) \mid G_{x_i} \sim \sum_{h=1}^{\infty} \mathbb{I}(u_i < w_{h,x_i})\delta_{\theta_j}.$$

Thus, we replace steps i) and iii) of Section 4.4 with

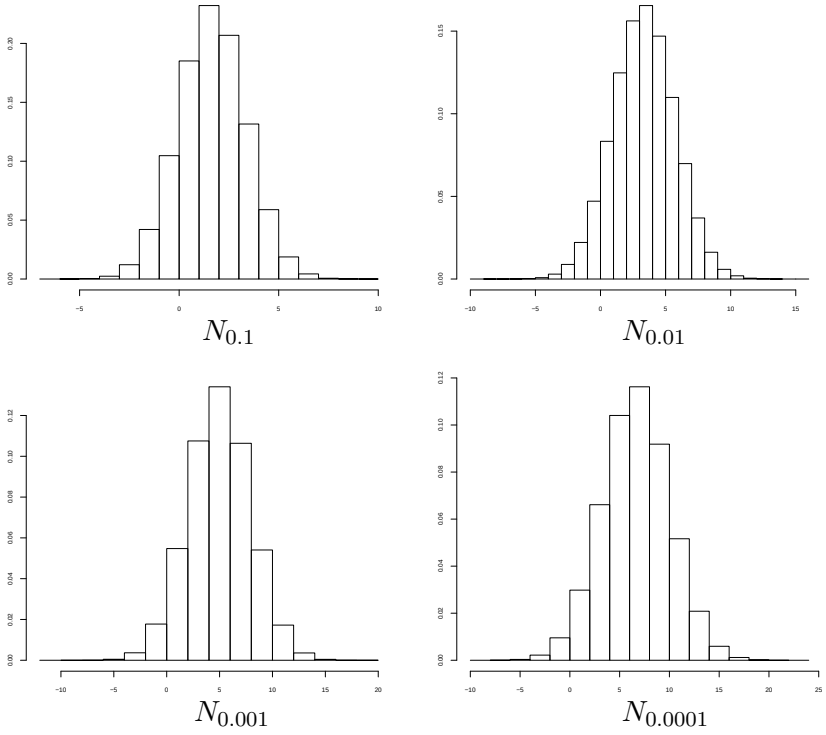


FIGURE C.1: Distribution of N_ε with ε equal to 0.1 (top-left), 0.01 (top-right), 0.001 (bottom-left) and 0.0001 (bottom-right), under a GDP with parameters $\phi = 1$ and $\mu = 0.5$.

i) *Resample s_i, for $i = 1, \ldots, n$.* This step replaces step i) in Section 4.4 by including the latent variables

$$\Pr(s_i = h \mid \cdot) \propto \mathbb{I}(u_i < w_{h,x_i}) f(y_i \mid \theta_h), \qquad \text{for } h = 1, \ldots, H^*$$

ii) *Resample the truncation level H^*, and simultaneously resample $w_{h,x}$ and θ_h.* As noticed by Walker [2007], introducing $u_1, \ldots, u_n$ allows considering at each iteration only a finite number atoms and weights of G_x, for all different values of x. In particular, we need only H^* atoms and weights for each G_x, where

$$H^* = \max\left(H_x^* : \min\left(H_x^* : \sum_{h=1}^{H_x^*} w_{h,x} \geq 1 - \min(u_1, \ldots, u_n) \right) \text{ for all } x \right).$$

This step is performed together with the generation of new θ_h and $w_{h,x}$

from their respective full conditionals (see step ii) and step iii) in Section 4.4).

In addition, we perform the following step

iii) *Resample u_i, for $i = 1, \ldots, n$.* This is a uniform update of the latent variable:

$$p(u_i \mid \cdot) \propto \mathbb{I}(0 < u_i < w_{s_i, x_i}).$$

Bibliography

Agarwal, D. K., Gelfand, A. E., and Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, 9(4):341–355.

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 53:111–142.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679.

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer Berlin Heidelberg.

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to mcmc for machine learning. *Machine Learning*, 50(1-2):5–43.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 1:1152–1174.

Antoniano-Villalobos, I., Wade, S., and Walker, S. G. (2014). A bayesian nonparametric regression model with normalized weights: A study of hippocampal atrophy in alzheimer's disease. *Journal of the American Statistical Association*, 109(506):477–490.

Arbel, J., Mengersen, K., , and Rousseau, J. (2016). Bayesian nonparametric dependent model for partially replicated data: the influence of fuel spills on species diversity. *The Annals of Applied Statistics*, In press.

Ashford, J. and Sowden, R. (1970). Multi-variate probit analysis. *Biometrics*, 26(3):535–546.

Barcella, W., De Iorio, M., and Baio, G. (2017). A comparative review of variable selection techniques for covariate dependent dirichlet process mixture models. *Canadian Journal of Statistics*, In press.

Barcella, W., De Iorio, M., Baio, G., and Malone-Lee, J. (2015). Variable selection in covariate dependent random partition models: an application to urinary tract infection. *Statistics in Medicine*, 35(8):1373–1389.

Barcella, W., De Iorio, M., Baio, G., Malone-Lee, J., et al. (2016a). A Bayesian nonparametric model for white blood cells in patients with lower urinary tract symptoms. *Electronic Journal of Statistics*, 10(2):3287–3309.

Barcella, W., De Iorio, M., Favaro, S., and Rosner, G. L. (2016b). Dependent generalised Dirichlet process priors for the analysis of acute lymphoblastic leukaemia. *Biostatistics*, In press.

Barcella, W., De Iorio, M., and Malone-Lee, J. (2016c). Modelling correlated binary variables: an application to lower urinary tract symptoms. Submitted.

Barry, D. and Hartigan, J. A. (1992). Product partition models for change point problems. *The Annals of Statistics*, 20(1):260–279.

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika*, 65(1):31–38.

Bishop, C. M. and Svenskn, M. (2002). Bayesian hierarchical mixtures of experts. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 57–64. Morgan Kaufmann Publishers Inc.

Blackwell, D. (1973). Discreteness of Ferguson selections. *The Annals of Statistics*, 1(2):356–358.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Carpenter, B., Gelman, A., and Hoffman, M. (2015). Stan: a probabilistic programming language. Submitted.

Chen, M.-H. (2004). *Skewed link models for categorical response data*, pages 131–151. Chapman and Hall/CRC.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2):347–361.

Chib, S. and Greenberg, E. (2010). Additive cubic spline regression with Dirichlet process mixture errors. *Journal of Econometrics*, 156(2):322–336.

Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104(488):1646–1660.

Chung, Y. and Dunson, D. B. (2011). The local Dirichlet process. *Annals of the Institute of Statistical Mathematics*, 63(1):59–80.

Cifarelli, D. and Regazzini, E. (1978). Problemi statistici non parametrici in condizioni di scambiabilità parziale e impiego di medie associative. Technical report, Quaderni Istituto Matematica Finanziaria dell'Università di Torino.

Connor, R. J. and Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206.

Cruz-Marcelo, A., Rosner, G. L., Müller, P., and Stewart, C. F. (2013). Effect on prediction when modeling covariates in Bayesian nonparametric models. *Journal of Statistical Theory and Practice*, 7(2):204–218.

Dahl, D. B. (2009). Modal clustering in a class of product partition models. *Bayesian Analysis*, 4(2):243–264.

de Finetti, B. (1931). Sul significato soggettivo della probabilità. *Fundamenta Mathematicae*, 17:298–329.

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, 65(3):762–771.

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, 99(465):205–215.

DeYoreo, M. and Kottas, A. (2014). Bayesian nonparametric modeling for multivariate ordinal regression. *arXiv preprint arXiv:1408.1027*.

DeYoreo, M. and Kottas, A. (2015). Modeling for dynamic ordinal regression relationships: an application to estimating maturity of rockfish in california. *arXiv preprint arXiv:1507.01242*.

DeYoreo, M., Kottas, A., et al. (2015). A fully nonparametric modeling approach to binary regression. *Bayesian Analysis*, 10(4):821–847.

Di Lucca, M. A., Guglielmi, A., Müller, P., and Quintana, F. A. (2013). A simple class of Bayesian nonparametric autoregression models. *Bayesian Analysis*, 8(1):63–88.

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). Generalized spatial Dirichlet process models. *Biometrika*, 94(4):809–825.

Dukes, C. (1928). Some observations on pyuria. *Proceedings of the Royal Society of Medicine*, 21(7):1179–1183.

Dunson, D. B. (2010). *Nonparametric Bayes applications to biostatistics.*, pages 223–273. In Hjort et al. [2010].

Dunson, D. B. and Park, J.-H. (2008). Kernel stick-breaking processes. *Biometrika*, 95(2):307–323.

Dunson, D. B., Pillai, N., and Park, J.-H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):163–183.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical population biology*, 3(1):87–112.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Fernández, C. and Green, P. J. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):805–826.

Foti, N. J. and Williamson, S. A. (2015). A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):359–371.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 56:501–514.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035.

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, 3(3):445–449.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

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

Gerds, T. A., Cai, T., and Schumacher, M. (2008). The performance of risk prediction models. *Biometrical Journal*, 50(4):457–479.

Ghosh, J. and Ramamoorthi, R. (2003). *Bayesian Nonparametrics*. Springer.

Giardina, F., Guglielmi, A., Quintana, F. A., and Ruggeri, F. (2011). Bayesian first order auto-regressive latent variable models for multiple binary sequences. *Statistical Modelling*, 11(6):471–488.

Gill, J. and Casella, G. (2009). Nonparametric priors for ordinal Bayesian social science models: specification and estimation. *Journal of the American Statistical Association*, 104(486):453–454.

Gill, K., Horsley, H., Kupelian, A. S., Baio, G., De Iorio, M., Sathiananamoorthy, S., Khasriya, R., Rohn, J. L., Wildman, S. S., and Malone-Lee, J. (2015). Urinary atp as an indicator of infection and inflammation of the urinary tract in patients with lower urinary tract symptoms. *BMC Urology*, 15(7):1–9.

Griffin, J. E. and Leisen, F. (2014). Compound random measures and their use in Bayesian nonparametrics. *arXiv preprint arXiv:1410.0611*.

Griffin, J. E. and Steel, M. F. (2010). Bayesian nonparametric modelling with the Dirichlet process regression smoother. *Statistica Sinica*, 20(4):1507.

Griffin, J. E. and Steel, M. J. (2006). Order-based dependent Dirichlet processes. *Journal of the American statistical Association*, 101(473):179–194.

Griffiths, T. L. and Ghahramani, Z. (2011). The indian buffet process: An introduction and review. *The Journal of Machine Learning Research*, 12:1185–1224.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, 56(4):1030–1039.

Hannah, L., Blei, D., and Powell, W. (2011). Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 1:1–33.

Hartigan, J. A. (1990). Partition models. *Communications in Statistics-Theory and Methods*, 19(8):2745–2756.

Hatjispyros, S. J., Nicoleris, T., and Walker, S. G. (2015). Dependent random density functions with common atoms and pairwise dependence. *arXiv preprint arXiv:1510.07153*.

Hjort, N. L. (2000). Bayesian analysis for a generalised Dirichlet process prior. Technical report, Matematisk Institutt, Universitetet i Oslo.

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

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168.

Irwin, D. E., Milsom, I., Hunskaar, S., Reilly, K., Kopp, Z., Herschorn, S., Coyne, K., Kelleher, C., Hampel, C., Artibani, W., and Abrams, P. (2006). Population-based survey of urinary incontinence, overactive bladder, and other lower urinary tract symptoms in five countries: results of the epic study. *European Urology*, 50(6):1306–1315.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Ishwaran, H. and James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures. *Journal of Computational and Graphical Statistics*, 11(3):508–532.

Ishwaran, H. and Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390.

Jain, S. and Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182.

Jara, A., García-Zattera, M. J., and Lesaffre, E. (2007). A Dirichlet process mixture model for the analysis of correlated binary responses. *Computational Statistics & Data Analysis*, 51(11):5402–5415.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2):181–214.

Jordan, M. I. and Teh, Y. W. (2014). *A Gentle Introduction to the Dirichlet Process, Beta Process and Bayesian Nonparametrics*. Unpublished.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105.

Karabatsos, G. (2015). A menu-driven software package for Bayesian regression analysis. *The ISBA Bulletin*, 22:13–16.

Karabatsos, G. (2016). A menu-driven software package of Bayesian nonparametric (and parametric) mixed models for regression analysis and density estimation. *Behavior Research Methods*, In press.

Karabatsos, G., Walker, S. G., et al. (2012). Adaptive-modal Bayesian nonparametric regression. *Electronic Journal of Statistics*, 6:2038–2068.

Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *Journal of Applied Mathematics and Mechanics*, 17(24):373–401.

Kawedia, J. D., Kaste, S. C., Pei, D., Panetta, J. C., Cai, X., Cheng, C., Neale, G., Howard, S. C., Evans, W. E., Pui, C.-H., et al. (2011). Pharmacokinetic, pharmacodynamic, and pharmacogenetic determinants of osteonecrosis in children with acute lymphoblastic leukemia. *Blood*, 117(8):2340–2347.

Khasriya, R., Barcella, W., Swamy, S., Gill, K., Kupelian, A. S., and Malone-Lee, J. (2017). A measure of the symptoms of chronic urinary tract infection in women that captures treatment effects. Submitted.

Khasriya, R., Khan, S., Lunawat, R., Bishara, S., Bignal, J., Malone-Lee, M., Ishii, H., O'Connor, D., Kelsey, M., and Malone-Lee, J. (2010). The inadequacy of urinary dipstick and microscopy as surrogate markers of urinary tract infection in urological outpatients with lower urinary tract symptoms without acute frequency and dysuria. *The Journal of Urology*, 183(5):1843–1847.

Kim, S., Dahl, D. B., and Vannucci, M. (2009). Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models. *Bayesian Analysis*, 4(4):707.

Kingman, J. (1967). Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78.

Kolmogorov, A. N. (1933). *Foundations of probability*. Nathan Morrison, Chelsea, New York.

Korwar, R. M. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *The Annals of Probability*, 1(4):705–711.

Kottas, A., Müller, P., and Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics*, 14(3):610–625.

Kunihama, T. and Dunson, D. B. (2014). Nonparametric Bayes inference on conditional independence. *arXiv preprint arXiv:1404.1429*.

Kupelian, A. S., Horsley, H., Khasriya, R., Amussah, R. T., Badiani, R., Courtney, A. M., Chandhyoke, N. S., Riaz, U., Savlani, K., Moledina, M., Montes, S., O'Connor, D., Visavadia, R., Kelsey, M., Rohn, J. L., and Malone-Lee, J. (2013). Discrediting microscopic pyuria and leucocyte esterase as diagnostic surrogates for infection in patients with lower urinary tract symptoms: results from a clinical and laboratory evaluation. *BJU International*, 112(2):231–238.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558.

Leann Long, D., Preisser, J. S., Herring, A. H., and Golin, C. E. (2015). A marginalized zero-inflated Poisson regression model with random effects. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 84(15):815–830.

Lijoi, A., Nipoti, B., Prünster, I., et al. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli*, 20(3):1260–1291.

Lijoi, A. and Prünster, I. (2010). *Models beyond the Dirichlet process*, pages 80–135. In Hjort et al. [2010].

Liverani, S., Hastie, D. I., Papathomas, M., and Richardson, S. (2015). PReMiuM: an R package for profile regression mixture models using Dirichlet processes. *Journal of Statistical Software*, 64(7):1–30.

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

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M. (2006). *Sparse statistical modelling in gene expression genomics*, volume 1, pages 155–176. Cambridge University Press, Cambridge.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, pages 50–55.

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

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238.

Mahmoud, H. (2008). *Pólya Urn Models*. Chapman and Hall/CRC.

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

Molitor, J., Papathomas, M., Jerrett, M., and Richardson, S. (2010). Bayesian profile regression with an application to the national survey of children's health. *Biostatistics*, 11:484–498.

Muliere, P. and Tardella, L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics*, 26(2):283–297.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83(1):67–79.

Müller, P. and Quintana, F. (2010). Random partition models with regression on covariates. *Journal of Statistical Planning and Inference*, 140(10):2801–2808.

Müller, P., Quintana, F., and Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):735–749.

Müller, P., Quintana, F., and Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278.

Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical science*, 19(1):95–110.

Müller, P. and Rosner, G. L. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association*, 92(440):1279–1292.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Neelon, B. H., O'Malley, A. J., and Normand, S.-L. T. (2010). A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling*, 10(4):421–439.

O'Brien, S. M. and Dunson, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics*, 60(3):739–746.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4(1):85–117.

Ohlssen, D. I., Sharples, L. D., and Spiegelhalter, D. J. (2007). Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Statistics in Medicine*, 26(9):2088–2112.

Papageorgiou, G., Richardson, S., and Best, N. (2015). Bayesian non-parametric models for spatially indexed data of mixed type. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(5):973–999.

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

Papathomas, M., Molitor, J., Hoggart, C., Hastie, D., and Richardson, S. (2012). Exploring data from genetic association studies using Bayesian variable selection and the Dirichlet process: application to searching for gene× gene patterns. *Genetic Epidemiology*, 36(6):663–674.

Papathomas, M. and Richardson, S. (2014). Exploring dependence between categorical variables: benefits and limitations of using variable selection within Bayesian clustering in relation to log-linear modelling with interaction terms. *arXiv preprint arXiv:1401.7214*.

Park, J.-H. and Dunson, D. B. (2010). Bayesian generalized product partition model. *Statistica Sinica*, 20(20):1203–1226.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. *Lecture Notes-Monograph Series*, pages 245–267.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900.

Plummer, M. et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, volume 124, page 125. Technische Universit at Wien.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):557–574.

Quintana, F. A., Johnson, W. O., Waetjen, E., and Gold, E. (2015a). Bayesian nonparametric longitudinal data analysis. *Journal of the American Statistical Association*, 111(515):1168–1181.

Quintana, F. A., Müller, P., and Papoila, A. L. (2015b). Cluster-specific variable selection for product partition models. *Scandinavian Journal of Statistics*, 42(4):1065–1077.

Rasmussen, C. E. and Williams, C. (2006). *Gaussian processes for machine learning. 2006*. MIT Press.

Reich, B. J. and Fuentes, M. (2007). A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *The Annals of Applied Statistics*, 1(1):249–264.

Ren, L., Du, L., Carin, L., and Dunson, D. (2011). Logistic stick-breaking process. *The Journal of Machine Learning Research*, 12:203–239.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Rodriguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian analysis*, 6(1):145–177.

Rodriguez, A. and Dunson, D. B. (2014). Functional clustering in nested designs: modeling variability in reproductive epidemiology studies. *The Annals of Applied Statistics*, 8(3):1416–1442.

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154.

Rodriguez, A., Dunson, D. B., and Taylor, J. (2009). Bayesian hierarchically weighted finite mixture models for samples of distributions. *Biostatistics*, 10(1):155–171.

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

Shahbaba, B. and Neal, R. (2009). Nonlinear models using Dirichlet process mixtures. *The Journal of Machine Learning Research*, 10:1829–1850.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Wade, S., Dunson, D. B., Petrone, S., and Trippa, L. (2014). Improving prediction from Dirichlet process mixtures via enrichment. *The Journal of Machine Learning Research*, 15(1):1041–1071.

Wade, S., Walker, S. G., and Petrone, S. (2013). A predictive study of Dirichlet process mixture models for curve fitting. *Scandinavian Journal of Statistics*, 41(3):580–605.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation*, 36(1):45–54.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348–368.