A Gianluca,

cui chiedo perdono per non aver capito.

A mio fratello,

la persona più brillante che io conosca.

Ai sentieri di montagna,

quelli più impervi,

da cui tanto possiamo imparare

e a chi

con me

li ha percorsi.

Contents

Introduction

When publishing his seminal paper on Product Partition Models in 1990, Hartigan referred to the tragic accident of the Space Shuttle Challenger that happened in 1986. As he recalled, the National Academy of Engineering Committee, while reviewing NASA's safety procedures, recommended using more objective numerical probability assessments. Hartigan commented on the event by noting that the crucial choice for engineers in determining the probability of a launch failure was to decide how to make use of data coming from previous launches and how much importance give to each of the previous observations. One possibility was to regard each launch (and especially the upcoming one) as too far different and independent from the others, this way only attempting to trace through risks of failure in the various systems that must work for the launch to be a success. The usage of this approach would determine, in the first place, the waste of many useful data coming from former launches. Moreover, difficulties would arise from the fact that there are always many new factors present in different systems and circumstances of each launch, so that a great deal of guessing is necessary for constructing probability models for failure paths, with the consequence that the resulting probabilities can vary enormously with plausible changes and no single probability has much credibility. Another option, orthogonal to the previous one, could have been to consider all the launches as equally similar and important in the same way for determining the assessment of the failure probability of future launches. It is easy to see how (potentially higher) problems would have arisen by following this procedure. Indeed, the specific circumstances regarding each launch are unique and have to be considered in great detail. The third possible approach is the one

proposed by Hartigan. In the product partition approach, past launches of similarly complex systems are treated as *potentially relevant*, letting data have some say in determining whether or not they actually are. Each previous observation is treated as potentially part of a group which can be considered as relevant in assessing probabilities related to the upcoming event. The methodology proposed suggests an elegant way to determine how relevant these various programs are to a new launch by weighing all the various possible reference sets that can be used for computing the probabilities. In other words, the method consists of evaluating the probability of each lunch being similar to different possible previous observations and then in using those probabilities as weights assigned to the prediction that one would have is . This straightforward and intuitive strategy was formalised by Hartigan in his simplest form and then made more complex as well used as a building block for more composite structures by other authors, such as in the PPMx case we will introduce.

There is, though, a relevant feature that is likely to characterise real-world datasets which was not considered in Hartigan's work and has not been addressed in previous literature regarding such models. Indeed, even thinking about the mentioned case of assessing the probability of the Shuttle launch failure, there is a possible scenario which is very common in practice and corresponds to the case of new or past observation set being unavailable/missing. Indeed, datasets presenting unavailable observations are recurrent and often encountered in real data-driven research (Molenberghs et al., 2014). Generally speaking, missing data are problematic for both estimation and prediction but, while there is a broad literature dedicated to addressing the problem for parameter estimation and inference, the same attention has not been dedicated to methods for prediction. A recent contribute in this direction was given by Page et al. (2020). Their work, which lays as stepping stone for ours, is jointly based on developments of Hartigan's work and clustering nonparametric methods, aims at using those strategies in cases of missing datasets. This strategy will be at the core of our work and it will be explained in great details in what follows. Briefly speaking, it firstly consists in a prediction approach based on clustering, which will be in line with Hartigan's strategy

we just outlined. Importantly, the specific mechanism will allow to seamlessly accommodate datasets with missing entries. Finally, a cluster-specific posterior predictive will be used to make out of sample prediction. The scope of our work is to extend their strategy in order to better accommodate its usage with certain data structures.

The work is structured as follows. In the first section, we review the main concepts regarding product partition models and their extension to include covariates. We then review elements regarding the nonparametric framework and DP priors which will turn out to be important for our reasoning, also introducing the connection between PPM and DP mixture models. In Chapter 3, we present the main issue this work aims at dealing with, which is missingness of covariates in the dataset. Using the contents of Chapter 1 and Chapter 2 as building blocks, we present the strategy proposed by Page et al. (2020), which allows dealing with the problem in a seamless and elegant way. In Chapter 4, we move to enlarge the previously presented connection between nonparametric priors and PPM models. Then, we finally expand the model presented in Chapter 3 in order to make use of a broader class of priors.

A considerable part of the work consisted of developing the code for the model using Python3. The code was implemented entirely from scratch and is publicly available on GitHub[1].

We assume that the reader is familiar with a modicum of measure theory, Dirichlet distributions, and general knowledge of the Bayesian framework in the parametric case.

[1]https://github.com/beatrcecantoni

Chapter 1

Product Partition Models

1.1 Defining a PPM

In Hartigan's paper mentioned in the introduction (Hartigan, 1990), the class of Product Partition Models (PPM) was formalised. Let $[n] = \{1, ..., n\}$ denote a set of experimental units. A family ρ_n of subsets S_j of $[n]$ such that $S_j \cap S_i = \emptyset$ for $j \neq i$ and $\cup_{j=1}^{k} S_j = [n]$ is a partition and is denoted by $\rho_n = \{S_1, ..., S_k\}$.

Definition 1 A product partition model (PPM) is jointly characterised by:

- A distribution on partitions that satisfies the *product condition*, that is a distribution taking the form of

$$p(\rho_n = S_1, ..., S_k) = M \prod_{j=1}^{k} c(S_j) \tag{1.1}$$

  where $c(S_j)$ is a *cohesion function*, which is a function on each subset, and M is the normalizing constant $\sum_{p \in \mathcal{P}} \prod_{j=1}^{|\rho|} c(S_j)$, with $\mathcal{P}$ being the set of all possible partitions of $[n]$ into nonempy sets.

- A distribution on observations *given the partitions* that satisfies the *independence condition*, i.e. observations in different subsets are independent, so that they can be represented in a joint model that can be described in terms of independent sub-models

corresponding to clusters, that is

$$p(\mathbf{Y}|\rho_n) = \prod_{j=1}^{k} p_{S_j}(\mathbf{y}^*_{S_j}) \tag{1.2}$$

where $\mathbf{y}^*_{S_j}$ are the observations in cluster j.

This independent sampling model (i.e. the model that relates observed responses y_i to the subgroups) is often assumed to be such that exchangeability holds within clusters, this way allowing it to be written as independent sampling conditional on cluster-specific parameters θ_j

$$y_i|\theta, \rho_n \overset{ind}{\sim} p(y_i|\theta_j) \tag{1.3}$$

$$\theta_1, ..., \theta_k \overset{iid}{\sim} F_0 \tag{1.4}$$

$$p(\rho_n = S_1, ..., S_k) \propto \prod_{j=1}^{k} c(S_j) \tag{1.5}$$

If a PPM is chosen to model the data, the posterior distribution of ρ_n is again a product partition model, with cohesion functions given by $c(S_j)p_{S_j}(\mathbf{y}^*_{S_j})$. *Cohesion functions* are the crucial elements characterising the model, representing how strongly the elements in S_j are likely to actually be part of the same cluster. Different choices can be made regarding the form of $c(S_j)$ since, in principle, they can be any nonnegative function of S_j. In order to avoid indicating a prior for the cohesion of each possible cluster of any possible partition - that is, avoid determining a specific prior on each 2^n-1 nonempty subset of [n] - more general and systematically-given structures are usually chosen. A common strategy is to make use of the size of each cluster only, namely $|S_j|$, to calculate the probability of each partition. This choice makes the resulting model for ρ_n to be invariant under permutations of units in [n]. Importantly, this common choice can in some cases allow the implied distribution on partitions to be interpreted as implied by a nonparametric priors, a feature that will turn out to be useful for computational purposes. More about the importance of both aspects will

be explained in Chapter 2 and 3. For now, we just highlight how not every possible choice for cohesion functions which only depend on $|S_j|$ will allow respecting those properties. One example of a cohesion function violating the desided features is given by $c(S) = M(|S_j|)^2$ (see Müller et al., 2015). Relatedly, it is common to consider cohesion functions which generate a coherent family of models, that is, a class of models for which $p(\rho_n)$ arises from $p(\rho_{n+1})$ by marginalization of the $(n+1)st$ unit.

1.2 Product Partition Models with Covariates

Going now back to the Shuttle launch case mentioned in the introduction, recall that we want to determine which previous observations should be considered as relevant when making prediction about future events. Obviously, we could be interested in looking at *related data*, such as meteorological conditions, type of shuttle and so on. In other words, we could be interested in adding dependence on covariates to the model we just presented. Such a model is called PPMx. Intuitively, observed units which have similar covariates should be more likely to be clustered together. This goal is achieved by generalizing the PPM prior $p(\rho_n)$ to a model with $p(\rho_n|x)$, where x are the covariates observed for each y. More formally, a PPMX model is a Product Partition Model with the distribution on partitions formalised as

$$p(\rho_n = S_1, ..., S_k|\mathbf{X}) = M \prod_{j=1}^{k} c(S_j) g(\mathbf{x}_j^*) \tag{1.6}$$

where $\mathbf{x}_j^* = \{x_i : i \in S_j\}$ is the set of covariates arranged by clusters, $g(\mathbf{x}_j^*)$ is the so called *similarity function* and M is the normalization constant obtained as the reciprocal of the sum over all possible partitions of $[n]$.

1.2.1 Similarity function

Intuitively, low values of $g(x_j^*)$ should correspond to bad clusters while high values should indicate homogeneous covariate vectors. Formally, any nonnegative function g is a valid

similarity function. At the same time, it is important to choose functions which make the model computationally tractable. Müller et al. (2011) considered generic forms to construct similarity functions that meet those criteria. The most common choice is to introduce an auxiliary probability model whose main aim is to get a convenient expression for the similarity function. According to this strategy, x_j^* are considered as sampled from an hypothetical probability model with $q(x_i|\xi_j^*)$ and prior $q(\xi_j^*)$, that is

$$g(x_j^*) = \int \prod_{i=1}^{|S_j|} q(x_i|\xi_j^*) q(\xi_j^*) d\xi_j^* \tag{1.7}$$

In other words, we use the marginal model under an hypothetical $q(\cdot)$ to define $g(\mathbf{x}_j^*)$. A first but significant advantage of this strategy is that $g(x_j^*)$ will be easy to obtain when $q(x_i|\xi_j^*)$ and $q(\xi_j^*)$ are chosen to be conjugate. Using the Bayes formula, $g(x_j^*)$ can be expressed as

$$g(x_j^*) = \frac{\prod_{i=1}^{|S_j|} q(x_i|\xi_j^*) q(\xi_j^*)}{q(\xi_j^*|x_j^*)} \tag{1.8}$$

Some specific choices for particular data formats follows as given by Müller et al. (2015).

Continuous x_i: Defining ξ_j^* as (m_j, v_j), one possible choice is to use $q(x_i|\xi_j^*) = N(x_i|m_j, v_j)$ and $q(m_j, v_j)$ to be the conjugate normal inverse chi-square (or gamma) distribution. This is the strategy followed in the model we will later present.

Categoical x_i: For a categorical covariate with c levels, $x_i \in \{1, ..., c\}$ we can consider ξ_j^* as equal to $(\pi_1, ..., \pi_c)$, with $0 \leq \pi_r$ for all $r = 1, ..., c$ and $\sum_{r=1}^{c} \pi_r = 1$. Then use $q(x_i|\xi_j^*) = Multin(x_i|1, \xi_j^*)$ and $q(\xi_j^*) = Dir(\xi_j^*|\alpha)$ for some α.

Count covariates x_i: In this case it is convenient to assume $q(x_i|\xi_j^*) = Poiss(\xi_j^*)$ and $q(\xi_j^*)$ distributed as a gamma distribution.

Ordinal covariates x_i: For an ordinal covariate with c categories, Müller suggests to follow the strategy proposed in Johnson, Albert (2006). They define an auxiliary probability model for x by introducing a latent variable z, setting a cutoff $-\infty = \gamma_0 < \gamma_1 < \cdots \gamma_{c-1} < \gamma_c = \infty$ and then seting $x = r$ if $\gamma_{r-1} < z < \gamma_r$ and fixing the cutoffs as $\gamma_r = r - 1$ for $r = 1, ..., c$.

Then, as in the continuous case, let $\xi_j^*=(m_j, v_j)$ and use the auxiliary model $q(z|\xi_j^*) = N(z|m_j, v_j)$ and a normal-inverse chi-square (or gamma) distribution $q(\xi_j^*)$. Then $q(x_i = r|\xi_j^*) = q(\gamma_{r-1} < z \leq \gamma_r|\xi_j^*)$. Finally, if we consider $n_{jr}=\sum_{i\in S_j} I\{x_i = r\}$ to be the number of covariates with value r in cluster j, and letting ϕ_{m_j,v_j} be the $N(m_j, v_j)$ c.d.f., the similarity function is

$$g(x_j^*) = \int_{\mathbb{R}\curvearrowright\mathbb{R}^+} \prod_{r=1}^{c} \{\phi_{m_j,v_j}(\gamma_r) - \phi_{m_j,v_j}(\gamma_{r-1})\}^{n_{jr}} q(\xi_j^*) d\xi_j^* \tag{1.9}$$

Importantly, while in all previous cases the covariate x_i is considered to be univariate, the strategy can easily be extended to cases of higher dimensionality of the vector x_i, either by choosing multivariate distributions (when feasible) or simply using the method followed in our simulations, that is using

$$g(\mathbf{x}_j^*) = \prod_{l=1}^{q} g^l(x_{jl}^*) \tag{1.10}$$

where this last strategy accommodates also cases in which x_i is a q-dimensional vector of mixed types of covariates, with l being the index for the covariate's type and $g^\ell(\cdot)$ being the similarity function for the ℓ-th covariate.

The second but equivalently relevant advantage of using similarity functions of the form of (1.7) will be made more clear in section (2.2.3). For now, we just say it relates to the fact that, when a similarity function of this type is combined with a nonparametric prior for the partitions, the resulting joint model for data, parameters and random partitions maintains the same computational benefits we would have in a PPM with such priors but without covariates.

1.3 PPM and PPMx in the literature

Given their structure, which clearly divides experimental units into more homogeneous subgroups, product partition models have extensively been employed for clustering problems in a broad sense and in a variety of contests. The first application of PPM was proposed by Harti-

gan himself (1992) and concerns change points detection problems. In this case, partitions are restricted to be of *contiguous type*, that is $\rho_n = \{\{1, ..., i_1\}, \{\{i_1+1, ..., i_2\}, ..., \{i_{k-1}+1, ..., n\}\}$, where $1 \leq i_1 < i_2 < ... < i_{k-1} < n$. In particular, considering j to be signaling a break point, a possible choice for the form of cohesion functions is (Yao (1984)).

$$c(\{i+1, ..., j\}) = \begin{cases} p(1-p)^{j-i-1} & \text{if} \quad j < n \\ (1-p)^{j-i-1} & \text{if} \quad j = n \end{cases}$$

This approach can be applied in a number of cases. An example is text segmentation (Kehagias et al., 2004), which deals with the problem of dividing a text into homogeneous segments, such that each segment deals with a particular subject and different segments deal with different topics. More generally, product partitions models are used as a building block of many more complex modeling structures and, as we have seen, they have been extended to models able to involve the use of covariates. Moreover, for reasons which will be made clear in what follows, PPM have often been employed with nonparametric priors on the clustering partitions (Page et al. (2015), 2010 Lijoi, Prünster (2010)). For example, Page and Quintana employed this structure for making predictions via clustering functional curves. Their method, which they employed in the case of NBA's player performance curves and covariates, allows for a clustering strategy which incorporates both curve smoothness and subject-specific covariates' information. Another example is the strategy followed by Legramanti et al. (2021) in what they defined as the general extended stochastic block model (ESBM), which is employed to learn group structures among nodes in network data allowing for the inclusion of node attributes making use of a product partition model structure. Finally, a research direction that is worth to be mentioned is the one that conjugates models of the form of PPM with spatial statistical modeling, as done by Page, Quintana (2015). Indeed, when dealing with geostatistical or areal data, it is common to deal with functions or structures which describe the spatial structure and imply a spatial grouping such that observations near in space are assumed to behave similarly. In this context, product partition models can be extended to a spatial setting to develop spatial methods that explicitly model the partitioning of spatial

locations.

1.3.1 PPM for prediction

In all the cases we previously mentioned, PPM allow clustering the observed sample in a set of subgroups. Importantly for our aim and going back to the need of forecasting expressed by Hartigan, a PPM can be easily used to make predictions for out-of-sample observations, since "inference about particular future observations may then be made by first conditioning on the partition and then averaging over all partitions" (Hartigan, 1990). This is particularly valid in the case of PPM extended to integrate covariates. Moreover, we will see how this prediction strategy can easily allow making predictions in cases of missing covariates. Those tasks will be better addressed in section 2.3.3 and 4.3.

Chapter 2

Clustering and Prediction via Dirichlet Process

We just saw how Product Partition Models allow us to evaluate the probability of each partition to accurately describe the true allocation of observed units and how this method can naturally be applied for posterior clustering and cluster-based predictions for out-of-sample observations. Importantly, when the prior is assigned over the entire space of possible partitions of [n] observations, the number of clusters is not assigned *a priori*, but depends on the best fit for the observed data. This nice property is the one that also characterises nonparametric methods for clustering, in which the number of parameters is allowed to be infinite, this feature differentiating them from parametric models. Indeed, traditional parametric models for clustering use a fixed and finite number of parameters. Many problems arise as a consequence of this restrictive assumption. The first and most obvious one is the loss of flexibility, which can turn out to be of key importance when the number of clusters is in practice not known a priori. Moreover, parametric models can suffer from over or under-fitting problems in cases of misfit between the number of parameters and the size of the data set. This issue can seriously undermine model selection. As it was in terms of flexibility, the nonparametric framework can accommodate better solutions for model se-

lection. Indeed, by using a model with an unbounded number of parameters, underfitting is mitigated, while the Bayesian approach of computing or approximating the full posterior over parameters mitigates overfitting. Given the advantages of a nonparametric approach, Bayesian nonparametric strategies have for a long time encountered a major limitation, which is the tractability of posterior computation. This limitation has been overcome in the last decades, this way allowing for a significant growth of research in the field and to the rise of nonparametric models. In particular, as we will see, nonparametic priors do, under certain conditions, generate implied random partition of experimental units, this way allowing for using them as a building block of clustering models.

In this section, we start introducing preliminary concepts related to nonparametric Bayesian inference while introducing clustering method which make use of the most common nonparametric prior, the Dirichlet Process (DP). We also delve into computational strategies used for those models, which will turn out to be of key importance to implement the PPMx we previously introduced. We then move to highlight the connection between distributions induced by the DP, which we will then generalize in the following chapters, and a specific subclass of PPM and PPMx. We introduce general relevant theory concerning the Dirichlet Process, its usage for clustering and prediction, and computational methods. Importantly, as we will see, this method will turn out to be intrinsically related to the PPM method we described above.

2.1 Preliminaries: the Nonparametric setting and the Dirichlet Process

2.1.1 Rapresentation Theorem and Discrete Random Measures

We begin introducing the relevant elements which will allow us to state de Finetti's representation theorem in a way which is broad enough to embed the nonparametric case too.

As stated in the introduction, we assume that the reader is familiar with general notions of parametric Bayesian modeling and related relevant distributions, as well as with a modicum of measure theory.

Let $\mathbb{X}$ be a Polish space and $\mathcal{X}$ the Borel sigma algebra of subsets of $\mathbb{X}$.

Definition 1. Random Probability Measure A random probability measure is a random element G defined on a probability space $(\Omega, \mathbb{F}, \mathbb{P})$ and taking values in $(\mathcal{P}, \mathcal{B}(\mathcal{P}))$ where $\mathcal{P}$ is the set of all probability measures on $(\mathbb{X}, \mathcal{X})$ and $\mathcal{B}(\mathcal{P})$ is an appropriate sigma algebra on $\mathcal{P}$.

Definition 2. Prior Q A prior Q is the law of a Random Probability Measure G on a subset (or the entirety) of the space $\mathcal{P}$ of probability measures.

Theorem 1. de Finetti's Representation Theorem Suppose to have an (ideally) infinite sequence of observations $(X_n)_{n\geq 1}$ taking values in $\mathbb{X}$. This sequence is exchangeable if and only if, for any $n \geq 1$, there exists a unique distribution Q on $\mathcal{P}$ such that

$$X_i | G \overset{iid}{\sim} G \quad \text{for} \quad i = 1...n \tag{2.1}$$

$$G \sim Q \tag{2.2}$$

Hence, the theorem implies conditional independence of the observations given a random probability measure G such that $G \sim Q$. In other words,

$$\mathbb{P}[X_1 \in A_1, ..., X_n \in A_n] = \int_{\mathcal{P}} \prod_{i=1}^{n} G(A_i) Q(dG) \tag{2.3}$$

where the probability Q is uniquely determined.

If Q is degenerate on a subclass of $\mathcal{P}$ indexed by a finite dimensional parameter, we have a *parametric* model, otherwise, when the support of Q is infinite-dimensional one typically refers to a *nonparametric* inferential problem.

2.1.2 The Dirichlet Process

In the framework we just described, we still haven't made the prior Q explicit. The first prior to have been introduced was the Dirichlet Process (DP), which was formalised in the seminal paper by Ferguson (1973). DP priors and, more precisely, *DP mixture models* (also known as infinite mixture models), are the most common building blocks for nonparametric data modeling.

Two are the main concepts which should be given in order to intuitively introduce the definition. First, in line with what we have seen before, the DP is a distribution over probability measures. Hence, each draw from a Dirichlet process is itself a distribution and can be interpreted as a random probability measure. Second, it is called a *Dirichlet* process because it has Dirichlet distributed finite dimensional marginal distributions.

Definition 3. Dirichlet Process Let G be a random distribution, G_0 be a distribution over $\mathbb{X}$ and α be a positive real number. Then for any finite measurable partition $\{A_1, ..., A_r\}$ of $\mathbb{X}$, the vector $(G(A_1), ..., G(A_r))$ is random since G is random. We say G is *Dirichlet process distributed* with *base distribution* G_0 and *concentration parameter* α, if

$$(G(A_1), ..., G(A_r)) \sim Dir(\alpha G_0(A_1), ..., \alpha G_0(A_r)) \tag{2.4}$$

for every finite measurable partition $\{A_1, ..., Ar\}$ of $\mathbb{X}$ and is commonly written as

$$G \sim DP(\alpha, G_0), \tag{2.5}$$

By recalling that we are defining a distribution over distributions, we can give an intuitive interpretation to the elements characterising a DP. First of all, G_0 can be interpreted as the mean of the DP. Indeed, it can be shown that, for any measurable set $A \in \mathbb{X}$, we have that $E[G(A)] = G_0(A)$. Moreover, α can be interpreted as an inverse variance: $V[G(A)] = G_0(A)] = G_0(A)(1 - G_0(A))/(\alpha + 1)$, such that the larger α is, the smaller the variance and a drawing from the DP will be similar to its mean G_0 with higher probability. For this reason, we can say the concentration parameters reflects the prior strength around G_0. Intuitively,

we will have that $G(A) \to G_0(A)$ for any measurable A as $\alpha \to \infty$. Still, this does not imply that $G \to G_0$. This relates to two concepts which will turn out to be relevant for what follows. First, we will see how draws from a DP are discrete distributions almost surely (see section 2.1.5), even when G_0 is chosen to be smooth. Second, if smoothness is a concern, it is possible to extend the DP by combining G with continuous kernels in the way we will later see.

2.1.3 Posterior Distribution

Now that we specified a prior, we are interested in combining it with observed data and obtaining a posterior. Luckily, the conjugacy property of the DP allows us to analytically obtain an expression for the posterior. Let our sequence of independent draws from G be X_1,...,X_n. Moreover, let $\{A_1, ...A_r\}$ be a finite measurable partition of $\mathbb{X}$, and let n_k= $\#\{i : X_i \in A_k\}$ be the number of observed values in A_k. Considering the DP prior as expressed in (2.4), it can be showed that making use of the conjugacy between the Dirichlet and the multinomial distribution we have:

$$(G(A1), ..., G(Ar))|X_1, ..., X_n \sim Dir(\alpha G_0(A_1) + n_1, ..., \alpha G_0(A_r) + n_r) \tag{2.6}$$

Importantly, this expression will hold for all finite measurable partitions, this feature making the posterior distribution be a DP as well. Similarly to (2.5), the posterior can be expressed as

$$G|X_1, ..., X_n \sim DP(\alpha + n, \frac{\alpha}{\alpha + n} G_0 + \frac{n}{\alpha + n} \frac{\sum_{i=1}^{n} \delta_{X_i}}{n}) \tag{2.7}$$

that is, as a DP whose posterior base distribution is a weighted average between the prior base distribution and the empirical distribution $\frac{\sum_{i=1}^{n} \delta_{X_i}}{n}$, where the weight associated with the prior base distribution is proportional to α and the weight associated with the empirical distribution is proportional to n. Hence, taking $\alpha \to 0$, the predictive distribution will just be given by the empirical distribution (that is, the prior becomes uninformative) or, similarly, the posterior will be dominated by the empirical distribution when $n >> \alpha$. Having clear

the expression for the posterior, we can now move to the predictive one, and finally to the use of DP for clustering.

2.1.4 Predictive Distribution and Blackwell-MacQueen Urn Scheme

Suppose again to have a sequence of observations $X_1, ..., X_n$ and consider a potential observation X_{n+1}. We will have that

$$P(X_{n+1} \in A | X_1, ...X_n) = E[G(A)|X_1, ...X_n] \tag{2.8}$$

$$= \frac{1}{\alpha + n}(\alpha G_0(A) + \sum_{i=1}^{n} \delta_{X_i}(A)) \tag{2.9}$$

which can be rewritten marginalizing out G

$$X_{n+1}|X_1, ...X_n \sim \frac{1}{\alpha + n}(\alpha G_0 + \sum_{i=1}^{n} \delta_{X_i}) \tag{2.10}$$

hence, notice how the posterior base distribution of G given $X_1, ..., X_n$ is also the predictive distribution of X_{n+1}. The sequence of predictive distributions in the form of (2.10) for $X_1, X_2...$ is called the Blackwell and MacQueen urn scheme (Blackwell, MacQueen, 1973).

2.1.5 Discreteness of G

In this section, we want to show that draws from a DP are composed of weighted sum of point masses. This is possible by providing a constructive definition of the DP, namely the *stick-breaking* construction. The property of discreteness of G will turn out to be of key importance for clustering using DP. Consider the following elements

$$\beta_k \sim Beta(1, \alpha), \tag{2.11}$$

$$X_k^* \sim G_0 \tag{2.12}$$

$$\pi_k = \beta_k \prod_{l=1}^{k-1}(1 - \beta_k) \tag{2.13}$$

and construct $G = \sum_{k=1}^{\infty} \pi_k \delta_{X_k^*}$, then $G \sim DP(\alpha, G_0)$.

An intuitive interpretation of this process follows. Starting with a stick of length 1, we break it at β_1, assigning π_1 to be the length of the stick we just broke off and drawing from G_0 the associated value X_1. Following this process recursively, we can construct G as a DP. Hence, G, can be interpreted as a weighted sum of point masses, where the π are the weights and X^* are the associated values.

2.1.6 Clustering Property

Before moving to DP mixture models, we still have to observe a crucial property of the DP itself. Starting from the predictive distribution as expressed in (2.10), we can rewrite it considering the observed values of X_i as being grouped together when equal, as $X_1^*, ..., X_{k_n}^*$. We obtain as a predictive distribution the *generalised Polya urn scheme*

$$X_{n+1}|X_1, ..., X_n = \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^{k_n} n_j \delta_{X_i} \tag{2.14}$$

Some considerations follow, which will turn out to be useful for the purpose of this work. First of all, an important consequence of the predictive structure is that a sample will feature ties among the observations with positive probability. Indeed, looking at (2.14), we have that

$$\mathbb{P}(X_{n+1} = X_j) > 0 \quad \text{for all} \quad n_j > 1 \tag{2.15}$$

which implies that a sample from the DP will feature $k_n \leqslant n$ distinct values, which we can interpret as the number of different clusters. Now, we can note that the described process will induce a random - *random* since $X, ..., X_n$ are random - partition of the integers $\{1, ...n\}$, determined by grouping the X_i's with the same value. This is coherent with what is done in the stick-breaking process we just introduced, where each X_i is assigned to one weight π, which we can interpret here as the relative width of the cluster or the relative presence of each X_i.

Sampling scheme on the partition

Equation (2.14) for $X_1, ..., X_n$ induces a sampling scheme on the partition which is generally known as *Chinese restaurant process*[1], according to a metaphore which can be described as follows:

- the first customer arrives at the restaurant and sits at table 1 ($X_1 = X_1^*$)
- for n ≥ 1, the $(n+1)$th customer arrives at the restaurant and
  - sits at table j ($X_{n+1} = X_j^*$) with probability $\frac{n_j}{\alpha + n}$, which is proportional to the number of customers already sitting at the table
  - sits at a new table with $X_{n+1}^* \sim G_0$ with probability $\frac{\alpha}{\alpha + n}$.

Prior on clusters' dimension as induced by the DP

Together with the sampling scheme for the partition, we can also consider its distribution. Generally speaking, we can have that the distribution on partitions is given as a distribution on clusters' multiplicities, $\{n_1, ..., n_k\}$, as is this case. In particular, the distribution of the random partition obtained by grouping an exchangeable sequence by distinct types is called *Exchangeable Partition Probability Function* (EPPF) (Pitman, 1995). Since those probability functions will turn out to be of crucial importance for our reasoning, important considerations regarding their properties and role will be at the core of Chapter 4. For now, we just focus on deriving the distributions on $\{n_1, ..., n_k\}$ obtained from the predictive structure of the DP

[1]"I remember very clearly how that came about. I knew this way of constructing random permutations by inserting new elements into the cycles, and once you understood this you could read o things like the Logan–Shepp description of the asymptotic sizes of cycles, and so on. But I was struggling to find a good metaphor. In a conversation with Lester (Dubins) at our usual coffee shop, somehow we came up with the idea that the cycles should be circular tables where you can have any number of items, so we thought of customers in a restaurant, and the familiar instance of specifically circular tables was Chinese restaurants." Jim Pitman, from Aldous (2018)

by direct computation. First of all, recall that by exchangeability we do not need to consider the different orderings of $X^{(n)}$. so we can simply assume that

- the first n_1 observations are clients sitting at the same table, i.e. have an equal value for $X_1, ..., X_{n_1}$
- the following n_2 observations $X_{n_1}, ..., X_{n_2}$ are equal
- same rationale for n_3 until n_k

This way we will end up having

$$\begin{aligned} p(n_1, ..., n_k) = & \underbrace{1 \cdot \frac{1}{\alpha+1} \cdot \frac{2}{\alpha+2} \cdots \frac{n_1-1}{\alpha+n_1-1}}_{n_1 \cdot terms} \\ & \times \underbrace{\frac{\alpha}{\alpha+n_1} \frac{1}{\alpha+n_1+1} \cdots \frac{1}{\alpha+n_1+n_2-1}}_{n_2 \cdot terms} \\ \times \cdots \times & \underbrace{\frac{\alpha}{\alpha+\sum_{j<k} n_j} \frac{1}{\alpha \sum_{j<k} n_j + 1} \cdots \frac{n_k-1}{\alpha+\sum_{j<k} n_j - 1}}_{n_k terms} \end{aligned}$$

Where the first element of the first term represents the probability of the first client to sit at table, the second element represents the probability of the second client to sit at the same table for the fist one and so on. The last expression equals

$$= \frac{\alpha^k}{\alpha_{(n)}} \prod_{i=1}^{k} (n_i - 1)! \tag{2.16}$$

with $\alpha_{(n)}$ being the Pochhammer symbol, that is

$$\alpha_{(n)} = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)} = \alpha(\alpha+1) \cdots (\alpha+n-1) \tag{2.17}$$

Summing up, we can write the derived expression for the implied distribution over $\{n_1, ..., n_k\}$:

$$p(n_1, ..., n_k) = \frac{\alpha^k}{\alpha_{(n)}} \prod_{i=1}^{k} (n_i - 1)! \tag{2.18}$$

2.1.7 DP mixture models

As we have seen, the DP generates distributions that are discrete with probability one, this way arising conceptual problems for continuous density estimation. This limitation can be fixed by combining it with some continuous kernel and using the DP as the mixing measure in a mixture over some simple parametric forms. This way, the nonparametric nature of the Dirichlet process translates to mixture models with a countably infinite number of components. From now on, we partially change our notation in order to accommodate for a new "layer" to be introduced in the model. Indeed, we let the realizations of G to be named $\{\theta_1, ..., \theta_n\}$ and to represent a set of latent parameters, while we call $\{y_1, ..., y_n\}$ a set of observations, modeled as follows

$$y_i|\theta_i \overset{ind}{\sim} F(\theta_i) \quad \text{for} \quad i = 1, ..., n \tag{2.19}$$

$$\theta_i|G \overset{iid}{\sim} G \tag{2.20}$$

$$G \sim DP(\alpha, G_0) \tag{2.21}$$

As we have previously seen, G is discrete, hence multiple θ_i can take the same value simultaneusly, this way creating a cluster.

2.2 DP and PPM

We are now ready to reach the core of this section, that is, showing the link between the DP and a subclass of product partition models. This connection will turn out to be of crucial importance for computational purposes, and we will be able to make use of the strategies outlined in section 2.3 when making inference for our PPM(x).

2.2.1 PPM with DP-style cohesion functions

We start recalling that the crucial building block of a PPM is a prior over the partitions,

$$p(\rho_n = S_1, ..., S_k) \propto \prod_{j=1}^{k} c(S_j) \tag{2.22}$$

it is now straightfoward to note how, following the reasoning in the previous sections, we can use the distribution on $\{n_1, ..., n_k\}$ as a distribution on the partition. Note that the distribution (2.18) is in principle not a distribution on the partitions, but is more precisely a distribution over the number of ties and the corresonding absolute frequencies. At the same time, if we allow the cohesion function $c(S_j)$ to be just a function of the cardinality of S_j, what we need is precisely a prior on $\{n_1, ..., n_k\}$. This connection will be made more formal in chapter 4. For now, we just limit our selves to exploit this observation by using the prior implied by the DP as a prior for the PPM, that is

$$p(\rho_n) \propto \prod_{j=1}^{k} c(S_j) = \prod_{j=1}^{k} (|S_j|-1)! \tag{2.23}$$

for a complete model expressed as

$$y_i|\theta_1^*, ..., \theta_k^*, \rho_n \overset{ind}{\sim} p(y_i|\theta_j^*), \quad \text{for} \quad i \quad s.t. \quad \theta_i = \theta_j^* \tag{2.24}$$

$$\theta_1^*, ..., \theta_k^* \overset{iid}{\sim} G_0, \tag{2.25}$$

$$p(\rho_n) \propto \prod_{i=1}^{k} (|S_j|-1)! \tag{2.26}$$

2.2.2 PPM and DP and the implied prior on observations

Quintana, Iglesias (2003) clearly showed the DP mixture models are connected with a specific subclass of PPM having cohesion functions in the way we just showed. This equivalence can be shown by deriving the implied marginal distribution for the observations obtained in the two cases. Details about the full derivation are given in the appendix.

2.2.3 Extension of the equivalence to PPMX

In Chapter 1, we mentioned how a *similarity function* $g(\mathbf{x}^*)$ as implied by an auxiliary model, that is in the form $g(x_j^*) = \int \prod_{i=1} q(x_i|\xi_j^*)q(\xi_j^*)d\xi_j^*$, was a strategical choice. Indeed, by considering the covariates as sampled from such an hypothetical model and using at the same time DP-style cohesion functions as we just showed (we will eliminate this restriction in the following chapters), we will have that the resulting joint model for data, parameters and random partitions will be formally equal to the model we would obtain after marginalizing out the random probability measure in an hypotetical DP mixture model for an augmented vector $z_i = (y_i, x_i)$ with kernel $p(z|\theta_j^*, \xi_j^*) = p(y_i|\theta_j^*)q(x_i|\xi_j^*)$ and centering distribution in the DP prior $G_0(\theta^*, \xi^*) = p(\theta^*)q(\xi^*)$. In other words, the choice of $g(\mathbf{x}^*)$ as the marginal distribution of an auxiliary model allow us to apply to PPMx the same rationale we just saw and to benefit of the same advantages. The main benefit is a computational one, as explained in the following section.

2.3 Computational strategies

We have seen how the usage of certain cohesion functions allows for an interpretation of the PPM as induced after marginalizing out the random probability measure from a DP mixture model. This connection brings important and useful computational benefits, since we can make use of the strategies adopted to simulate posterior distributions of DP mixture models. Indeed, exact computation of posterior expectations for a Dirichlet process mixture model is infeasible when there are more than few observations. The use of a Dirichlet process mixture model was mainly made computationally feasible with the development of Markov Chain methods. In particular, there are two classes of algorithms, which differ according to the way G is handled. In one case, G is integrated out and the strategy followed consisted in constructing chains that allow to sample from an approximation of the posterior distribution of the parameters of each cluster and/or the mixture components associated with observations.

According to other strategies, simulating G is part of the algorithm. Additionally, in the case of a DP, there are algorithms which allow exact simulation even without sampling an infinite sequence, known as retrospective (Papaspiliopoulos, Roberts, 2008) and slice samplers (Walker, 2007). Here, we focus on methods following the first strategy and especially on methods expanding the sampling space to auxiliary parameters.

2.3.1 Gibbs sampling with conjugate priors

Remark 1. Gibbs Sampling Let us say we are interested in the distribution $p(\phi)$. MCMC methods consist in constructing a Markov Chain which is irreducible and invariant w.r.t. $p(\phi)$. Then, by the Ergodic theorem, it is possible to approximate $p(\phi)$ using the realizations of the chain, i.e. $\{\phi^0, \phi^1, ..., \phi^t\}$. In the specific case of Gibbs sampling, the distributions of interest typically regards more than one variable $\{\phi_1, ..., \phi_k\}$. In this case, the chain is constructed using the full (posterior) conditional of each variable, when available. Moreover, this structure can be generalised to cases in which it is not possible to sample from the full conditionals of each variable of interest. In this case, one possibility is to use a Metropolis-Hastings step can be used in order to substitute the full conditionals of those parameters, in the so called Metropolis within Gibbs.

As said, one possible strategy is to integrate out G and to deal with posterior simulation of the θ's (we are referring to the model as presented in section 2.1.7). The most intuitive approach is to use a Gibbs sampling and to repeatedly draw values for each θ_i (for $i = \{1, ..., n\}$) from its conditional distribution given θ_{-i} (all θ_j for $j \neq i$) and the data. The conditional prior reads

$$\theta_i|\theta_{-i}, y \sim \frac{\alpha}{n-1+\alpha} G_0 + \frac{1}{n-1+\alpha} \sum_{j \neq i} n_j \delta(\theta_j) \tag{2.27}$$

Combining the Polia-urn scheme with the likelihood $f(y_i, \theta_j)$, we would have:

$$\theta_i|\theta_{-i}, y_i \sim \alpha r_i G_i + \sum_{j \neq i} q_{i,j} \delta(\theta_j) \tag{2.28}$$

with G_i being the posterior distribution for θ based on the prior G_0 and the single observation y_i, with likelihood $f(y_i, \theta)$ and where

$$q_{i,j} = bf(y_i, \theta_j) \tag{2.29}$$

$$r_i = b \int f(y_i, \theta_j) dG_0(\theta) \tag{2.30}$$

and b such that $\sum_j q_{i,j} + \frac{r_i}{\alpha} = 1$. For this method to be feasible, it has to be possible to compute the integral defining r_i, but also to sample from G_i. This will generally be so when G_0 is the conjugate prior for the likelihood given by f. This algorithm, which is used by Escobar, West (1995) produces an ergodic chain but suffers of slow mixing (see Neal (2000)). The problem lies in the fact that the algorithm cannot change the θs for more than one observation simultaneously, since they are not updated by cluster but for each observation separately. This implies that changes to the θs values would require the passage trough a low probability intermediate state in which observations which correctly belong to the same cluster do not have the same θ, hence such changes can occur only rarely. The slow mixing problem is avoided by constructing a chain which allows to sample the θs - or more generally, each cluster parameters - simultaneously. In order to do that, instead of sampling for each θ_i for $i = \{1, ..., n\}$, we sample for a variable z_i which represents the observation's cluster allocation for $i = \{1, ..., n\}$. Once we have completed the iteration for all z_i, we will sample from the conditional for the parameters of each group. Notice that this implies than when cluster j is a singleton and we are currently sampling for the allocation of its observation, its cluster will be discarded given that it will appear as empty. Similarly from what we have seen before and again with the expressions in (2.29) and (2.30) with b being the appropriate normalizing constant, we will update the allocation using

$$z_i | z_{-i}, \theta, y_i \sim \alpha r_i G_i + \sum_{j=1}^{k} q_{i,j} \delta(\theta_j) \tag{2.31}$$

If a new cluster is selected, an associated value for θ_j is chosen by drawing it from G_i. Finally, at each iteration and once all observation have been allocated to an existing or

new cluster, we will end up with k different clusters and we will update each θ_j according to its conditional. This was the method used by Bush, MacEachern (1996) and allows to update each θ_j for all observations in a cluster simultaneously the slow mixing problem is avoided. Moreover, in many cases, we can integrate out the parameters θ and construct a Markov chain only for z_i. If this strategy solves the slow mixing problem, it is again only feasible if we can compute $\int f(y_i, \theta) dG_0(\theta)$ and sample new θ from G_i, which is generally feasible when conjugate priors are used, which is reasonably often not the case. West et al. (1994) suggested using a numerical quadrature or a Monte Carlo approximation to evaluate $\int f(y_i, \theta) dG_0(\theta)$, approximating the integral by an average over m values for θ drawn from G_0. Following this strategy, one can then approximate a draw from $p(\theta_j|y)$ by drawing from the m θ points with probabilities proportional to their likelihood. As elaborated by Müller, Quintana (2004), this approach turns out to be quite inaccurate in practice. Among other methods that can be used to overcome this limits in cases of non-conjugacy, a popular one was the one presented by Neal (2000). More generally, we move now to present strategies which make use of auxiliary variables.

2.3.2 Methods with auxiliary variables

An important strategy was proposed by MacEachern, Müller (1998). Their method makes use of a set of auxiliary parameters and is referred to as "no gaps" algorithm. It consist in an exact approach to handle non-conjugate priors via an augmented "no-gaps" model. However, this algorithm is characterised by a inefficiently low probability associated to new cluster. For this reason, Neal proposed a similar algorithm without this inefficiency, which is explained in the following subsection.

Remark 2. Gibbs sampling with auxiliary parameters. Let's say we are again interested in $p(\phi)$, now a single parameter, but we are not able to sample from it. A smart way to overcome this issue by making use of Gibbs sampling consists in using auxiliary variables $\mathbf{z}$ which form with ϕ an augmented model such that the implied distribution for ϕ doesn't

change. Then, for example, one can perform the Gibbs steps using $p(\phi|\mathbf{z})$ and $p(\mathbf{z}|\phi)$ and then discard the $\mathbf{z}$, remaining with ϕs, whose implied distribution will be the one of interest.

Neal Algorithm 8

In the case of DP mixture models, we can use this technique to update the partitions s_i without having to integrate w.r.t. G_0. First of all, we consider the s_i as the variables of interest, that is, we follow a strategy similar to the second proposed in the above section about Gibbs sampling. Then, we introduce some useful auxiliary variables and a smart augmented model. In our case the auxiliary variables can be considered as potential new clusters which are currently empty and, in particular, we focus on the potential parameters associated with those new clusters:

$$\{\theta_1^*, ..., \theta_{k^-}^*, \underbrace{\theta_{k^-+1}^*, ..., \theta_{k^-+m}^*}_{auxiliary}\} \tag{2.32}$$

where k^- is the number of non-empty clusters once y_i has been eliminated from its cluster. The advantage of those srategies is that the evaluation of integral can easily be replaced by simple likelihood evaluations based on the parameters associated with old clusters and the new potential parameters. Importantly, algorithms of this type have to be proposed in such a way that the augmented model implies a marginal distribution on $(\theta_1, ..., \theta_n)$ is not affected. The scheme introduced above will then be the following:

1. Sample from the conditional distribution of the parameters $\theta_{k^-+1}^*, ..., \theta_{k^-+m}^*$ associated to the auxiliary components given the current value of s_i and the rest of the state.

2. Perform a Gibbs sampling update for s_i using the conditional posterior distribution for s_i given the other s_j and m auxiliary components and their associated parameters implied by (33).

3. Discard all the parameters associated with clusters which are now not associated with any observation.

In the algorithm proposed by Müller and MacEachern, the augmented vector of parameters reads

$$\{\theta_1^*, ..., \theta_{k^-}^*, \underbrace{\theta_{k^-+1}^*, ..., \theta_n^*}_{auxiliary}\} \tag{2.33}$$

and the augmented model includes the constraint that there will be "no gaps" in the current allocation of the s_i, i.e. $n_i > 0$ for $j = 1, ..., k^-$ and $n_j = 0$ for $j = k^- + 1, ..., n$ Consider being at the step of updating s_i i.e. the cluster allocation for observation i, then their algorithm can be summarised as follows:

1. If $n_j > 1$, then

$$p(s_i = j | s_{-i}, \theta^{*-}, \mathbf{y}) \propto \begin{cases} n_j^- \quad p(y_i|\theta_j^*) & j = 1...k^- \\ \dfrac{\alpha}{k^- + 1} \quad p(y_i|\theta_j^*) & j = k^- + 1 \end{cases} \tag{2.34}$$

2. If $n_j = 1$ (the cluster is a singleton),
   with probability $\dfrac{k^-}{k^- + 1}$ leave s_i unchanged, otherwise remove s_j from the $j-th$ cluster and label $\theta_{k^-+1}^*$ as θ_j to comply with the no-gaps rule and update s_j according to point 1. above

The problem with this strategy is that, as noted by Neal (2000), the probability associated with selecting z_i to be a value different from all other z_j, that is to assign an observation to a new cluster with new parameters, is inefficiently low. Indeed, it is reduced from what one might expect by a factor of $k^- + 1$. This inefficiency has been solved with by using a different augmented model as presented again by Neal, according to what is know as "Neal algorithm 8".

Consider the augmented model q_n, such that:

$$q_n(\theta_n | \theta_{-n}, \theta_{k^-+1}^*, ..., \theta_{k^-+m}^*) \propto \sum_{j=1}^{k_n} n_{n,j} \delta_{\theta_j^*}(\theta_n) + \frac{\alpha}{m} \sum_{j=1}^{m} \delta_{\theta_{k^-+j}^*}(\theta_n) \tag{2.35}$$

As far as step 1 is concerned, the auxiliary parameters have to be sampled in the following way: if s_i=s_j for some $j \neq i$, the auxiliary parameters have no connection with the rest

of the state, or the observations, and are simply drawn independently from G_0. If s_i is a singleton, that is $s_i \neq s_j$ for all $j \neq i$, then it must be associated with one of the auxiliary parameters (for simplicity we assume it being the first one). Because of this, the parameter corresponding to the auxiliary cluster will in this case be set equal to the existing θ_i.

Step 2 will be performed using the conditional distribution that the augmented model q_n implies on s_i. Again, recall that the way step 1 is performed ensures that the implied marginal on $_i$, implied by the distribution on θ_n obtained marginalizing out w.r.t. $\theta^*_{k^-+1}, ..., \theta^*_{k^-+m}$ is the one we are interested in.

Being the observations exchangeable, we can assume for each {1,...,n} that we are updating c_i for the last observation and that the c_j for other observations have values in the set $\{1, ..., k^-\}$. Notice that for simplicity the number m of auxiliary parameters can be set equal to 1.

Once the partition has been updated, clusters parameters are also updated. In this case the chain has to be invariant w.r.t. (θ^*), so any update respecting this condition can be performed. If possible, the parameter will be updated using the full conditional, if not, a Metropolis step will be usually performed. The same will happen also for other parameters which are not cluster-specific (see the example below).

Algorithm 1: Neal Algorithm 8

Input: Observations vector $\mathbf{y}$; hyperparameters

Output:, estimated partition ρ_n; estimated parameters for each cluster (μ_j^*, σ_j^{2*}) and those common across clusters (μ_0, σ_0^2)

Initialization: randomly initialize the random partition $\rho_n^{(0)} = s_i, ...s_n$, the cluster specific parameters $\theta_1^{*(0)}, ..., \theta_k^{*(0)}$ and parameters common across clusters

for *i in* $\{1, ..., n\}$ **do**

  if $s_i = s_j$ *for some* $j \neq i$ **then**
    simulate $\theta_{k^-+1} \sim G_0$
  end
  else
    Set $\theta_{k^-+1} = \theta_{S_i}$
  end
  Assign i to a new cluster using as weights

$$p(s_i|s_{-i}, \theta^{*-}, \mathbf{y}) \propto \begin{cases} n_j^- & p(y_i|\theta_j^*) \quad \text{for} j = 1...k^- \\ \alpha & p(y_i|\theta_j^*) \quad \text{for} j = k^- + 1 \end{cases}$$

end

for *j in* $\{1, ..., k\}$ **do**
  Update θ_j
end

Update common parameters

2.3.3 Extending the computational strategy to DP-style PPMx

In the previous sections we saw, at first, how the PPM which certain cohesion functions can be interpreted as the model implied by a DP mixture model and then we reviewed the most common computational strategies for such a model. The same connection can be extended to PPMx with covariates. Again, it can be exploited for computational purposes. When covariates inform the partition, the allocation updating rule entering the algorithm in the training phase will be of the form:

$$p(s_i^{(t)} = j|-) \propto \begin{cases} \dfrac{c(\{S_j\{n+1\}\})}{c(S_j)} & \dfrac{g(x_j^* \cup \{x_i\})}{g(x_j^*)} \quad p(y_i|\theta_j^*) \quad \text{for} \quad j = 1...k^- \\ c(\{n+1\}) & g(\{x_i\}) \quad p(y_i|\theta_j^*) \quad \text{for} \quad j = k^- + 1 \end{cases}$$

which in the case of DP-style cohesion function will be

$$p(s_i^{(t)} = j|s_{-1}, \theta^{*-}, \mathbf{y}) \propto \begin{cases} n_j^- & \dfrac{g(x_j^* \cup \{x_i\})}{g(x_j^*)} \quad p(y_i|\theta_j^*) \quad \text{for} \quad j = 1...k^- \\ \alpha & g(\{x_i\}) \quad p(y_i|\theta_j^*) \quad \text{for} \quad j = k^- + 1 \end{cases}$$

As far as prediction is concerned, it can be performed online as part of the MCMC chain we just described, as suggested by Page et al. (2015). In particular, at each iteration, the posterior predictive distributions for out of sample observations is approximated by firstly evaluating the probability of it belonging to one of the existing clusters or to a new one as

$$p(s_i^{(t)} = j|-) \propto \begin{cases} n_j^- & \dfrac{g(x_j^* \cup \{x_i\})}{g(x_j^*)} \quad \text{for} \quad j = 1...k \\ \alpha & g(\{x_i\}) \quad \text{for} \quad j = k + 1 \end{cases} \tag{2.36}$$

which of course resemble the weights used for in-sample observations updates except for the likelihood term, which is clearly not available for datapoints which have not been observed yet. A prediction is then obtained by weighing over all clusters' predictions (using the previously obtained multinomial weights). Once the prediction has been obtained for each iteration, one carries out the typical Monte Carlo approximation to obtain an estimate of the posterior predictive. It is worth noting that, following this strategy, obtaining a punctual estimate for the parameters is not necessary and the updated parameters can be discarded, this way making the computation faster.

Chapter 3

Dealing with Missing Covariates

Datasets presenting unavailable observations are recurrent and often encountered in real data-driven research (Molenberghs et al., 2014). Generally speaking, missing data are problematic for both estimation and prediction. While there is a broad literature dedicated to address the problem for parameter estimation and inference, the same attention has not been dedicated to methods for prediction. The simplest strategy to follow for model training is the *complete-case approach*, which simply deletes units that present any missing observations. Of course, this option has a fair limited potential and is commonly not suitable. More complex strategies have been implemented, which in most cases rely on imputation of missing entries. The model at the core of our work tackles the issue from a complete different perspective. Here, we review the main concepts which are often encountered in literature regarding missing data cases and we also give a sense of the state of the art in this field of research.

Missing mechanisms

Literature regarding missing data commonly rely on some assumptions about the data mechanism which mainly refer to the taxonomy developed by Rubin (1976). According to this classification, missing data mechanisms can be classified as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR).

- Missingness completely at random refers to cases in which data are missing for reasons completely unrelated to either the observed or missing parts of the complete dataset. Often, evidence can be provided against the MCAR assumption through the association between missing data indicators and the data response vector.

- Under the plausible scenarios, the MCAR assumption does not hold. Following an easy example by Daniels, Hogan (2008) on a study concerning smoking cessation, it is realistic that within treatment group, participants observed to be heavier smokers are more likely to drop out (and their entries are more likely to be missing). A more realistic assumption for many longitudinal studies and especially clinical ones, is missing at random (MAR), which only requires missingness to be independent of other missing responses, conditionally on both observed responses and covariates. As a general rule (see again Daniels, Hogan, 2008) however, analyses under the weaker MAR assumption will provide valid inference when MCAR holds.

- Still, despite the MAR assumption being fairly general and allowing the missing data mechanism to be explained by data that are observed, there are cases in which it turns out to still be unrealistic. Those are cases in which the probability of missingness depends on the value of missing responses even while also conditioning on observed data. Those mechanisms are defined as missing not at random (MNAR). Considering again a smoking cessation study trial, there is empirical evidence that in smoking cessation studies, participants followed up after dropout are more likely than not to have experienced relapse. For example, in a simplified case in which the full data response consist of cessation outcomes at time 1 and time 2, while covariates are a treatment x and the missingness of data due to a drop out, the MNAR assumption will not hold. Indeed, the missing probability will depend on the availability observation cessation at time 1.

Starting from a complete dataset, it is possible to simulate and control missing mechanisms

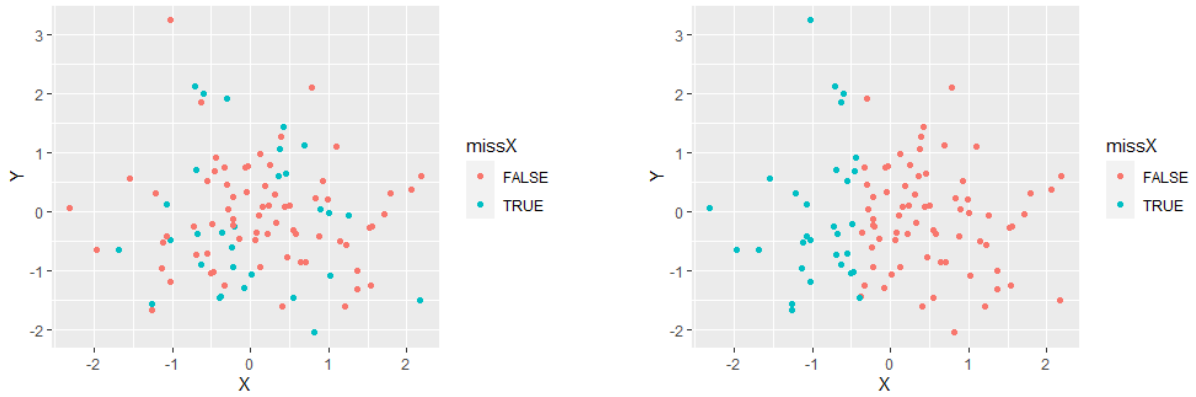


Figure 3.1: MCAR (left) and MNAR (right) mechanisms compared. Y is the dependent variable, while X is the explanatory variable. In both cases, the missing rate is 0.3

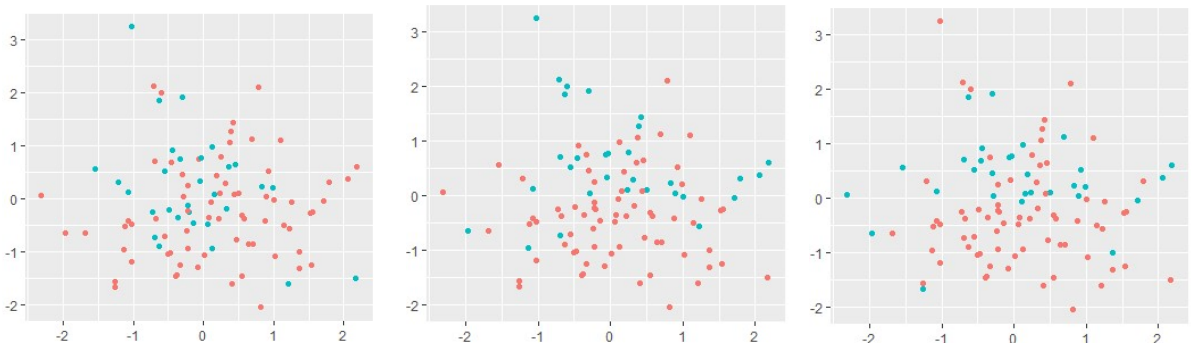


Figure 3.2: MAR mechanisms with different strength. When lower (left), the missing mechanism is similar to the MCAR case

by using the `mice` package on R. The model we will later propose is tested for those different mechanisms.

Imputation

One set of strategies consists of imputing missing observations. First of all, a poor imputation algorithm such as zero-imputation or mean imputation, which are easy to implement, can drastically reduce a model's prediction performance when the missing predictors are influential (Fletcher Mercaldo, Blume (2020)).

Conditional mean imputation and multiple imputation (see Rubin, 1988) are methods to probabilistically "fill" the missing values. The latter is typically used at the model training stage. It consists of drawing multiple values to fill the missing observations by using conditional distributions derived from the observed data and to use those placeholder values to fit the model. Through Rubin's rule it is possible to combine fits and thus build the placeholder values. This strategy, by leveraging partial information from incomplete data records, can significantly increase the imputation performance in cases of data being missing at random. This method has recently grown in popularity for model training, but also for observation prediction. However, as noted by Fletcher Mercaldo, Blume (2020), there are cases in which the user only has access to published parameters estimates and not to the original dataset. In those cases, this approach turns out to be limited. Indeed, in order to obtain real-time imputations which are coherent with the multiple imputation procedure followed for model training, the additional out-of-sample record should also be combined with the original data in order to properly determine the "fill"in missing predictors in accordance with the new observations. Those methods are limited not only for being dependent on having access to the original data and the imputation datasets, but they end up to demand significant computing power, to the point of being typically impractical in real world settings. Overall, as shown by (see Janssen et al. (2009) when dealing with clinical prediction modeling) these approaches degrade prediction performance when predictors are unavailable.

There are then other multiple imputation methods that worth to be mentioned. Importantly, a typical issue of methods handling missing covariates is the possibility to employ it for mixed data types. Some methods (like the one we will present) have the advantage to allow for being employed in such cases. One of those is the MICE method (multiple imputation by chained equations, van Buuren 2018), which can be used for the imputation of missing covariates one-at-a-time and employs conditionally specified models. The problem with this model lays in the fact that there is no theoretical guarantee that the conditionally specified models will produce a valid joint model for the covariates. A partial solution was introduced by

employing sequential Bayesian regression trees (BART) to impute missing covariates in the MICE framework. This method (Xu et al., 2016) solves the problem by producing valid joint distributions, but suffers from the fact that the order of the conditional models impacts the imputations. A similar strategy was followed by Burgette, Reiter (2010), who employed classification and regression trees (CART) to impute within a MICE type algorithm that allows for more flexibility in the conditional distributions. Similarly, Stekhoven, Bühlmann (2012) used random forests to do imputation. Following a different strategy, Storlie et al. (2019) presented a flexible yet complex Bayesian Nonparametric model which claims to address many of the shortcomings to standard approaches to missing predictors via jointly modeling mixed-types covariates while also including a smart variable selection component which improves the efficiency of the procedure. Some of those methods for multiple imputation are efficient but, again, they mainly focus on parameter inference. As stated, when prediction is tangentially considered, complications arise. For example, some procedures for imputation are structurally problematic when it is not possible to connect the covariates vector to the dependent variable (since in out-of sample prediction the latter is not available), this leading to significantly worsen the predictive efficacy of the method.

A final issue arises with the so called missing indicator approach (see Little, 1992). Indeed, under this approach, there is no clear way to handle the case of a new subject presenting missing pattern which is different from the ones present in the training dataset.

The method proposed by Page et al. (2015) and expanded here tackles the issue from a completely different perspective by allowing to avoid any type of imputation. One the one hand, it still requires access to the full training dataset (at least for covariates), but at the same time, being structured over Hartigan's model, naturally accommodates for prediction and allows to handle unobserved missing patterns.

3.1 Prediction with missing covariates using PPMx

This section, which serves as stepping stone for the model we will propose later on, mainly refers to the work of Page et al. (2020), whose main purpose was to extend the PPMx to a predictive model which accommodates the use of covariates' vectors with variable dimension. Following this strategy, it is possible to deal with covariates' missingness without the need to carry out any type of imputation. Indeed, when compared with the strategies we outlined above, this approach stems from a completely different perspective. A model based on the covariate-dependent approach we saw above is used in order to allocate the observation to a cluster. Then, a cluster-specific posterior predictive distribution makes the out-of-sample prediction mechanism straightforward. Notably, the model results being flexible in incorporating covariates and missing x_i in the *predictive* distribution too. Moreover, it allows to easily deal with missingness patterns which have not been observed among the observations in the training data set. Note that, in principle, $\mathbf{x}$ can appear both in the top level sampling model - like in a regression setting - and in the prior distribution on partitions. Following this strategy, missingness can be accommodated in both cases. In this implementation, for simplicity, we only consider the case in which the x_i only appear in the latter.

Denote by $\mathcal{O}_i$ the collection of covariate indices that are observed for subject i, such that the i^{th} subject's observed covariate vector can be now denoted as $\mathbf{x}_i^O = \{x_{i\ell} : \ell \in \mathcal{O}_i\}$ and the corresponding collection of covariates that belong to cluster j is $\mathbf{x}_j^{*O} = \{x_{i\ell} : \ell \in \mathcal{O}_i; i \in S_j\}$. Then, missing covariates can easily be accommodated in the PPMx by evaluating the similarity function for covariate p of cluster j based only on those subjects for which covariate p is observed. In other words, the similarity function used to evaluate the probability of each possible partition to be suitable for an observation will simply be based on those covariates which are measured, this way simply skipping over missing variables when evaluating the similarity function. This way, imputation is avoided, while complex interactions and nonlinear associations between covariates and responses are taken into account by the covariate-dependent cluster allocation mechanism. Denoting $C_{i\ell}$ as $\{\ell \in \mathcal{O}_i; i : i \in S_j\}$, we call $(\cdot)$ such a similarity

function, that is

$$g(\mathbf{x}_j^*) = \tilde{g}(\mathbf{x}_j^{*O}) \tag{3.1}$$

$$= \prod_{\ell=1}^{p} \tilde{g}(\mathbf{x}_j^{*O} = \{x_{i\ell} : \ell \in \mathcal{O}_i; i \in S_j\}) \tag{3.2}$$

$$= \prod_{\ell=1}^{p} \int \prod_{i \in C_{j\ell}} q(x_i|\xi_{j\ell}) q(\xi_{j\ell}) d\xi_{j\ell} \tag{3.3}$$

This similarity function can be used in both the training and testing phase. From a computational viewpoint, the methods needed to fit this model are unchanged with respect to the case with no missing observations, with the only difference being the skipping rule for missing covariates that has to be added to the algorithm and a matrix of indicators for the presence of covariates that must be carried along. In particular, the training allocation rule will be the following

$$p(s_i = j|s_{-i}, \mathbf{x}^O, \theta^*) \propto \begin{cases} \dfrac{c(S_j \cup \{y_i\})}{c(S_j)} \quad \dfrac{\tilde{g}(x_j^{*O} \cup \{x_n^O\})}{\tilde{g}(x_j^{*O})} \quad p(y_i|\theta_j^*) & \text{for} \quad j = 1, ..., k \\ c(\{y_i\}) \quad \tilde{g}(\{x_n^O\}) \quad p(y_i|\theta_j^*) & \text{for} \quad j = k+1 \end{cases} \tag{3.4}$$

Predicting Future Observations with Missing Entries

What is now relevant for us is the model's flexibility in capturing the role of covariates in the predictive distribution. Indeed, the similarity function for variable covariates vector dimension, $\tilde{g}(x_j^{*O})$, can be used also to accommodate incomplete covariates' vectors when making predictions for new individuals. The predictive allocation probability for the $(n+1)$st subject will be (with DP-style cohesion functions as above)

$$p(s_{n+1} = j|\rho_n, \mathbf{x}^O, \mathbf{x}_{n+1}^O) \propto \begin{cases} n_j \quad \dfrac{\tilde{g}(x_j^{*O} \cup \{x_{n+1}\})}{\tilde{g}(x_j^{*O})} & \text{for} j = 1...k_n \\ \alpha \quad g(\{x_{n+1}^O\}) & \text{for} j = k_n + 1 \end{cases} \tag{3.5}$$

such that any missing covariate for the $(n+1)st$ subject will be handled by simply skipping over it, this way making it always possible for the similarity function to be evaluated, even in

cases of missingness patterns which have not been observed yet. Moreover, this framework allows to allocate observations which have the extreame feature of having no covariates available. Indeed, in this case, the observation will be allocated to a cluster following a simple PPM. The way prediction was carried out within the algorithm follows the strategy suggested by Page et al. (2015). In particular, posterior predictive distributions are obtained online, in the sense that draws from this distribution are collected at each iteration within the MCMC algorithm. In particular, we use

$$p(y_{n+1}|\rho_n, \mathbf{x}^O, \mathbf{x}^O_{n+1}) = \sum_{j=1}^{k} w_j f_j(y_{n+i}; \theta_j) \tag{3.6}$$

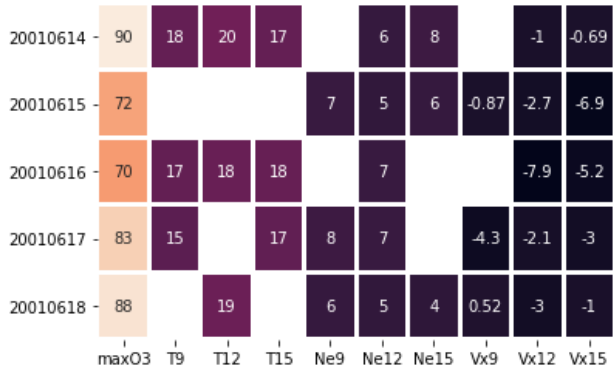
where the weights w_j correspond to the multinomial probabilities indicated above.

3.1.1 Implementation 1

The Ozone data set

In this section, in order to show the validity of the approach we introduced above, we use the dataset proposed by Page et al. (2020). The dataset is publicly available[1], and regards measurements of the maximum daily ozone in Rennes, together with relevant covariates. In particular, the covariates are temperature, nebulosity and projection of wind speed vectors and are all measured at three different moments of the day (respectively at 9:00, 12:00 and 15:00), for a total of nine continuous covariates.

Figure 3.3: A sample of five observations in the data set. White cells are shown when the corresponding data is missing.

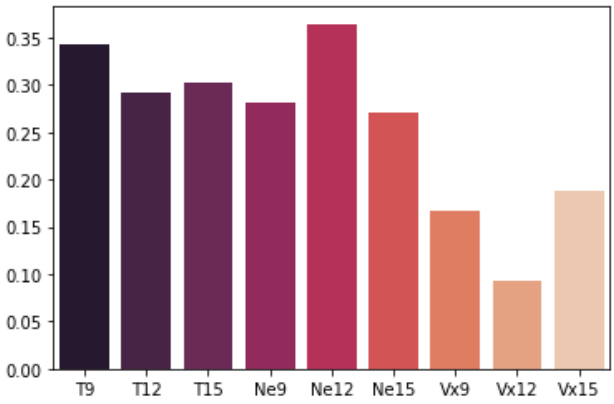


Since dealing with missing dependent variables is not the aim of this work, we remove the 16 observations which lack the value regarding the maximum daily ozone measurement. Figure 2 displays the percentage of missingness for each covariate in the resulting data set, which can vary significantly between different covariates.

The ozone data set particularly suits the aim of showing the flexibility of the proposed model by allowing us to show how the method can easily handle also cases in which the pattern of missingness has not been observed among individuals included in the training data set. Indeed, looking at Figure 3 we can notice that there is a number of missing patterns that

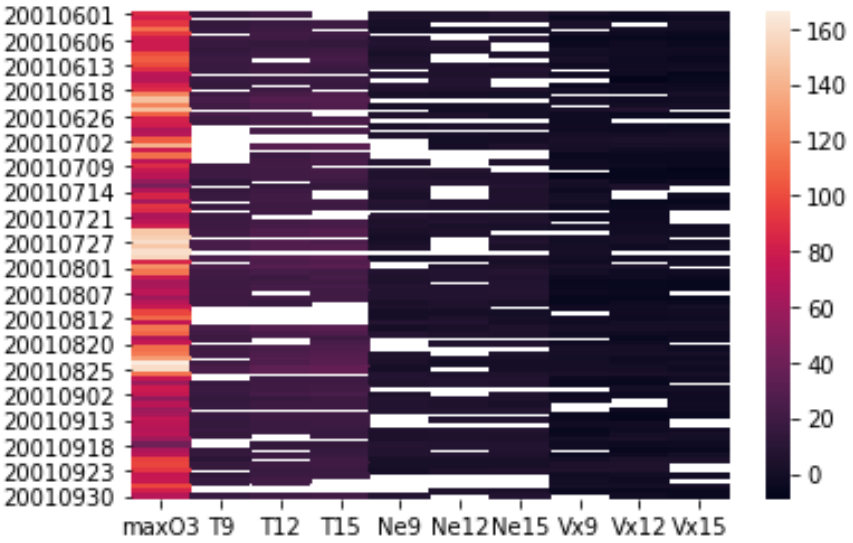
[1]https://github.com/njtierney/user2018-missing-data-tutorial/blob/master/ozoneNA.csv

Figure 3.4: Missing values by covariates



appear only one time. Moreover, only 14.6 % of observations are complete cases.

Figure 3.5: Missing pattern observed in the ozone data set for the response (maximum daily ozone measurements maxO3) and the nine covariates.



The model

We consider a hierarchical model to fit the maximum ozone observation data set.

$$
\begin{aligned}
y_i|\mu^*, \sigma^{2*}, c_i &\sim N(\mu^*_{c_i}, \sigma^{2*}_{c_i}) \quad \text{for} \quad i = 1, ..., n \\
(\mu^*_j, \sigma^*_j) &\sim N(\mu_0, \sigma^2_0) \times UN(0, a_\sigma) \quad \text{for} \quad j = 1, ..., k \\
(\mu_0, \sigma_0) &\sim N(m_0, v^2) \times UN(0, a_{\sigma_0}) \\
Pr(\rho_n = S_1, ..., S_k|\alpha, \mathbf{X}) &\propto \prod_{j=1}^{k} c(S_j|\alpha) g(\mathbf{x}^*_j)
\end{aligned}
$$

Again, note how covariates can in principle enter the mdel at the first level too. The uniform prior on cluster specific standard deviations follows what is suggested by Gelman (2006) as a choice when a Gaussian prior is used for the cluster-specific means. Moreover, prior parameters are set as m_0=0, v^2=100, a_0=10, a_{σ_0}=10.

Posterior inference

Inference is carried out by exploiting the connection between PPM and DP mixture models and following the approach introduced in section (2.3.2). More precisely, we randomly construct cross-validation data sets by shuffling the presented dataset. Each of them is randomly divided between 75 observations which are used to train the model and 21 observations used as testing data. We collect 5000 MCMC samples, to be used to fit the model after discarding the first 10000 as burn-in.

Algorithm 2: MCMC for Ozone Dataset (covariates in the PPMx only and DP-style cohesion functions)

Input: Observations vector **y** (training portion); Covariates matrix X (trining portion), hyperparameters m_0, v^2, a_σ, a_{σ_0},

Output: Estimated partition $S_1, ..., S_k$, Estimated parameters for each cluster (μ_j^*, σ_j^{2*}) and those common across clusters (μ_0, σ_0^2)

Initialization: Randomly initialize the partition $\rho_n^{(0)} = s_i, ...s_n$, i.e. $\{S_1^{(0)}, ..., S_k^{(0)}\}$; the cluster specific parameters $(\mu_j^{(0)*}, \sigma_j^{(0)2*})$ and the common parameters $(\mu_0^{(0)}, \sigma_0^{2(0)})$

for *t in T=26000* **do**

  for *i in* $\{1, ..., n\}$ **do**

    if $s_i = s_j$ *for some* $j \neq i$ **then**

      simulate $(\mu_{k^-+1}, \sigma^2_{k^-+1}) \sim G_0$, where G_0 corresponding to the prior on (μ_j^*, σ_j^{2*})

    end

    else

      Set $(\mu_{k^-+1}, \sigma^2_{k^-+1}) = (\mu^*_{s_i}, \sigma^{2*}_{s_i})$

    end

    Randomly assign i to a cluster using as weights (recall we are using DP-style cohesion functions)

    $$p(s_i^{(t)} | s_1^{(t)}, ..., s_{i-1}^{(t)}, s_{i+1}^{(t-1)}, ..., s_n^{(t-1)} \theta^{*-}, \mathbf{y}, X)$$
    $$\propto \begin{cases} n_j^- & \dfrac{g(x_j^* \cup \{x_i\})}{g(x_j^*)} \quad p(y_i|\theta_j^*) \quad for j = 1...k^- \\ \alpha & g(\{x_i\}) \quad p(y_i|\theta_j^*) \quad for j = k^- + 1 \end{cases}$$

  end

  for *j in* $\{1, ..., k\}$ *(the existing clusters at that iteration)* **do**

    update $(\mu_j^{*(t)}, \sigma_j^{2*(t)})$ using the conditional or a metropolis step

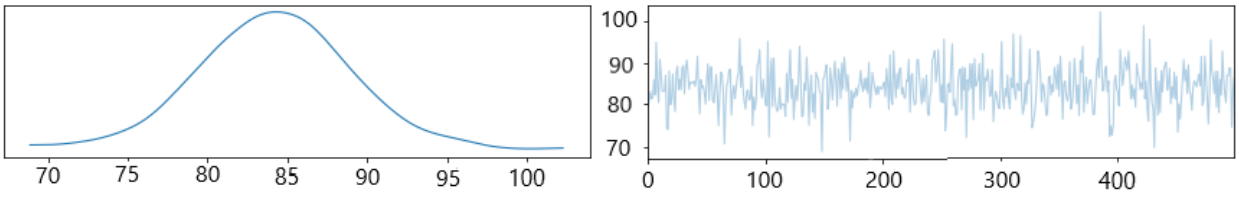
  end

  Update common parameters $(\mu_0^{(t)}, \sigma_0^{2(t)})$ using the conditional or a metropolis step

end

Results

Figure 3.6: Approximated posterior distribution for a prediction of an out-of-sample observation (left). Last 500 iterations of the chain (right)



The predictive performance seems to be in line with the results presented by Page et al. (2020). In the plots, we show the result regarding the prediction performed in one of the trials for one out-of sample observation. The chain seems to mix well and to well explore the space. Figure 3.6 (left) shows the approximated posterior distribution, while figure 3.6 (right) shows the last 500 points of the chain. As said, prediction is carried out online, and the collected points are collected to obtain an estimate using the common MCMC rationale. On the bottom of the page, figure 3.7 is an autocorrelation plot which shows that the chain fast reaches stationarity. Page et al. also compared their algorithm with more famous methods for missing covariates. In particular, they found out that their approach outperforms methods involving imputation, such as the one performed using the package by Gelman and Hill (2011 Su et al. (2011)), but is not as much competitive as BART when few covariates are in place.

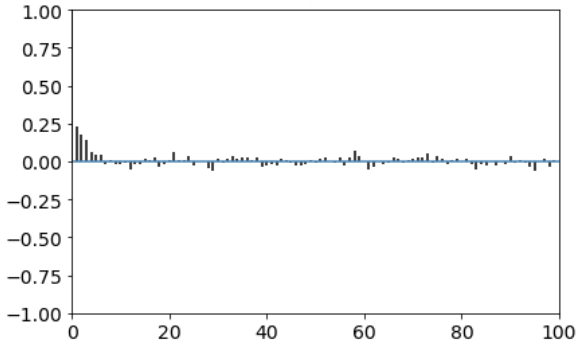


Figure 3.7: Autocorrelation plot

Chapter 4

Generalising to a broader class of priors

As we have seen, the DP induces a prior on the number of clusters and on their frequencies that can be easily employed as a building block for a PPMx which can, in turn, be used to flexibly and easily accommodate the need to make out-of-sample predictions with data sets presenting missing covariates according to the strategy proposed by Page et al. (2020). In what follows, we make our contribution by expanding this strategy to other nonparametric priors by exploiting a connection between PPM and a subclass and nonparametric priors which was clearly formulated by Lijoi, Mena and Prünster (2007c). We begin introducing this more general link and a broader class of priors called *Gibbs-type priors*.

4.1 A broader class of nonparametric priors

4.1.1 Discrete random probability measures

First of all, recall that our interest lies in priors with a clustering property, hence we want to focus on *discrete* random probability measures. Priors of this type can be grouped by using as grouping rule the type of strategy used for their constructions, with the Dirichlet process being a special case of many broader categories (for an extensive review see Rodriguez, Müller, 2013). Examples include priors obtained via:

1. stick-breaking, such as the Pitman-Yor process;
2. finite-dimensional distributions, such as the normalised inverse-Gamma process;
3. normalisation of independent increments, as in the case of normalised completely random measure;
4. (also by) the predictive structure, as in the case of species sampling models;
5. the EPPF, which are at the base of the construction of Gibbs-type priors

4.1.2 Species Sampling Models

Focusing on what is relevant for this work, we start introducing *species sampling models*. Introduced by Pitman (1996), species sampling models extend the DP to a broader class. As we have seen, part of the central role played by the DP as nonparametric prior is due to its marginalization property and to the tractability of the Polya urn scheme. Pitman proposed species sampling priors as generalizations of DP which enjoy much more flexibility but still preserve similar marginalisation properties, this way allowing to exploit the same computational strategies. As explained by Pitman, those features are especially appropriate for problems of *species sampling*. Following his metaphor, we can imagine to have a random sample $\theta_1, \theta_2, ...\theta_n$ drawn from a large population of individuals of various species, with θ_i representing the species of the ith individual sampled. According to this metaphor, the space on which θs are defined can be interpreted as a set of labels or tags referring to the various species. Moreover, if we imagine each label to be drawn from a diffuse distribution, the species will be coded by different tags almost surely. Now we can look at the Dirichlet Process and see how it can accommodate this description. In order to formally define the class introduced by Pitman, we start recalling the Blackwell-Macqueen scheme and noting how we can express the weights as functions of only the sample size,

$$p(\theta_{n+1}|\theta_1,...,\theta_n) \propto \begin{cases} \delta_{\theta_j^*}(\theta_{n+1}) & w.prob \quad \dfrac{n_j}{n+\alpha} \equiv p_j(\mathbf{n}) \\ G_0\theta_j^*(\theta_{n+1}) & w.prob \quad \dfrac{\alpha}{n+\alpha} \equiv p_{k+1}(\mathbf{n}) \end{cases} \tag{4.1}$$

Hence, the posterior predictive can be expressed as

$$\theta_{n+1}|\theta_1,...\theta_n \sim p_{k+1}(\mathbf{n}) + \sum_{j=1}^{k} p_j(\mathbf{n})\delta_{\theta_j^*} \tag{4.2}$$

where, again, $p_j(\mathbf{n})$ are weights which only depends on the cluster size n_j. This is a general predictive scheme that is common to all species sampling models priors. More formally, from Pitman (1995)

Definition 4. Species Sampling Sequence and PPF

Let $(\theta_1,...,\theta_n)$ be an exchangeable sequence of random variables and call a rule specifying both the distribution of θ_1 and the conditional distribution of θ_{n+1} given $\theta_1,...,\theta_n$ for each $n = 1, 2...$ a *prediction rule*. Moreover, consider θ_1 to be distributed according to a distribution G_0, which is commonly assumed to be nonatomic (diffuse). Then the sequence $(\theta_i,...,\theta_n)$ is a Species Sampling Sequence if it is subject to a *prediction rule* of the form

$$\begin{aligned} \mathbb{P}(\theta_1 \in \cdot) &= G_0(\cdot) \\ \mathbb{P}(\theta_{n+1} \in \cdot|\theta^{(n)}, k_n = k) &= p_{k+1}(\mathbf{n}) + \sum_{j=1}^{k} p_j(\mathbf{n})\delta_{\theta_j^*} \end{aligned}$$

where $\mathbf{n} = (\mathbf{n}_1, \mathbf{n}_2...)$ is the random vector of counts of various species observed in the sample. Moreover, the sequence of functions p_j for $j = 1, 2...$ is called a sequence of Predictive Probability Function (PPF), defined on the space of the vector of counts species counts, whose range is identified with the countable set $\mathbb{N}^* = \cup_{k=1}^{\infty}\mathbb{N}^k$. Determining the weights of the predictive distributions, any PPF needs to satisfy

$$p_j(\mathbf{n}) \geq 0; \qquad \sum_{j=1}^{k+1} p_j(\mathbf{n}) = 1 \tag{4.3}$$

for all $\mathbf{n} \in \mathbb{N}^*$. Note that considering as a *putative* PPF any sequence of functions p_j for $j = 1, 2...$ defined on $\mathbb{N}^*$ which satisfies (4.3), not all putative PPFs are PPFs, since (4.3) does not guarantee exchangeability of the sequence of random variables (for more details about PPFs see Lee et al., 2013).

4.1.3 Exchangeable Partition Probability Functions

Recall that our goal is to generalise the clustering and predictive approach presented in section 2.2. Hence, for reasons which will turn out to be clearer in the following sections, we shall now focus attention on a class of species sampling. In Chapter 2, when dealing with the expressions characterising the Dirichlet process, we saw how the DP was implicitly inducing a specific distribution over partitions, which turned out to be related with a particular specification of the cohesion functions in a product partition model. Here, we start by introducing more formally the concept of distributions over partitions, to then move to a class of species sampling models which are specifically defined in terms of the joint distribution on partitions they induce.

Distributions over random partitions known as *Exchangeable partition probability functions* (EPPF) have been introduced by Pitman (1995) even though, as recalled by Pitman himself, they rose as a natural but foundamental developements of earlier work presented by Kingman(1978a, 1978b 1982) regarding the concept of a *partition structure* and Kingman's representation, holding for *exchangeable* partitions. Here, we decided to follow Pitman's narrative in introducing EPPFs and to start by defining tools valid for a broader class of partitions, called *partially exchangeable* in line with the concept of partial exchangeability due to de Finetti (1938). As a preliminary concept, let $[n] = \{1, ..., n\}$ and recall that a partition of $[n]$ is an unordered collection of disjoint non-empty subsets of $[n]$, say $\{S_i\}$, with $\cup_i S_i = [n]$. moreover, a *random partition* of $[n]$ is a random variable ρ_n with values in the finite set of all partitions of $[n]$.

Definition 5. Partially exchangeable partition and probability function
Let $\mathbb{N}^* = \cup_{k=1}^{\infty}\mathbb{N}^k$, the set of finite sequences of positive integers. Denote a generic element of $\mathbb{N}^*$ by $\mathbf{n}= (n_1, ..., n_k)$. Call a random partition ρ_n of $[n]$ *partially exchangeable* (PE) if for every partition $\{S_i, ..., S_k\}$ of $[n]$, **where the** $S_1, ..., S_k$ **are in order of appearance**

$$\mathbb{P}(\rho = \{S_1, ..., S_k\}) = p(|S_1|, ..., |S_k|) \tag{4.4}$$

for some function $p(\mathbf{n}) = p(n_1, ..., n_k)$ defined for $\mathbf{n} \in \mathbb{N}^*$ with $\sum_{i=1}^{k}|S_k|= n$. Then call $p(\mathbf{n})$ a *partially exchangeable probability function.*

Combining this definition with the case in which $p(\mathbf{n})$ is symmetric we finally arrive at defining a EPPF.

Definition 6. Exchangeable partition and probability function (EPPF) A random partition ρ_n of {1,...,n} is *exchangeable* if and only if it is partially exchangeable with a PEPPF $p(\mathbf{n})$ which only depends on group sizes and not on the groups' order, i.e. it is invariant under permutation of its elementsfor every permutation σ of $[n]$. Using the notation we used above we have

$$\mathbb{P}(\rho = \{\sigma(A_1), ..., \sigma(A_k)\}) = p(|A_1|, ..., |A_k|) \tag{4.5}$$

Moreover, $p(\mathbf{n})$ is called *Exchangeable partition probability function* (EPPF)

4.1.4 Gibbs-type priors

Characterisation of Gibbs-type priors via EPPF

Now that we clarified the role of an EPPF as a distribution over partitions, we are ready to introduce the class of nonparametric prior which will turn out to be relevant for our work. We start recalling the probability on partitions induced by the DP. As already pointed out, it only depends on clusters' cardinality and is a simmetric function of its arguments. Indeed, it is an EPPF

$$p(\mathbf{n}) = \frac{\alpha^k}{\alpha_{(n)}}\prod_{j=1}^{k}(n_j - 1)! \tag{4.6}$$

Note now how it can be rewritten using $V_{n,k} = \dfrac{\alpha^k}{\alpha_{(n)}}$ as well as the fact that $\prod_{j=1}^{k}(n_j - 1)! \equiv \prod_{j=1}^{k} 1_{(n_j-1)}$ and rewrite it as

$$p(\mathbf{n}) = \underbrace{V_{n,k}} \quad \underbrace{\prod_{j=1}^{k} 1_{(n_j-1)}} \tag{4.7}$$

where the two underbraced elements can be seen as separate building blocks. More generally, we can generalise this tructure by giving the definition on a Gibbs-type priors as introduced by Gnedin, Pitman (2006).

Definition 7. Gibbs-type priors Let $\{V_{n,k}\}$ be a collection of coefficients satisfying the recursive equation

$$V_{1,1} = 1, \quad V_{n,k} = (n - \sigma k)V_{n+1,k} + V_{n+1,k+1}, \quad 1 \leq k \leq n \tag{4.8}$$

A random probability measure is said to be of Gibbs type if it has EPPF of the form

$$p(\mathbf{n}) = V_{n,k} \prod_{j=1}^{k} (1 - \sigma)_{(n_j-1)}, \tag{4.9}$$

with $\sigma \leq 1$ and $n = \sum_{j=1}^{k} n_j$.

Insightful analysis of theoretical aspects regarding this class of priors can be found in Lijoi et al. (2008b) and Favaro et al. (2013a). Their results on posterior distributions enable the employment of those priors in more applied works. With this regard, the potential usage of Gibbs type priors have been investigated in different fields, such as economics (Lijoi et al., 2016), functional data analysis (Canale et al., 2017), genomics (Lijoi et al.2007b, Lijoi et al. 2008, Favaro et al. 2009), epidemiology (Paganin et al., 2021) and generally speaking in problems regarding species sampling models (Favaro et al., 2013b). Moreover, for the reasons which will become clearer later, they are particularly suitable to be employed for topic modeling and network analysis.

4.1.5 A generalised connection between nonparametric priors and Product Partition Models

We are now ready to finally highlight the crucial connection between this class and product partition models. Product partition models are at the core of this work, but we recall their structure for the sake of simplicity

$$y_i|\theta_1^*, ..., \theta_k^*, \rho_n \overset{ind}{\sim} p(y_i|\theta_j^*) \quad \text{for} \quad j = 1, ..., k \tag{4.10}$$

$$\theta_j^* \overset{iid}{\sim} G_0 \tag{4.11}$$

$$p(\rho_n = S_1, ..., S_k) \propto \prod_{j=1}^{k} c(S_j) \tag{4.12}$$

With the DP, we saw how it was possible to use the distribution it induced on partitions as building-block for (4.12). In that case, the cohesion function expressing the distribution over partition were functions only depending on cardinalities, this way making the partition exchangeable according to the definition we saw above. Similarly, we could decide to use a prior on the partition in the model as in (4.12) to be of the form expressed in (4.9), which is the general expression for EPPF for Gibbs-type priors. What is crucial to point out is that, as firstly pointed out by Lijoi et al. (2007c) who rephrased a result formulated by Gnedin, Pitman (2006), EPPF that are suitable for the prior structure needed for an (exchangeable) PPM *will precisely be* of Gibbs-type.

Proposition 1. The *exchangeable* random partition ρ_n has distribution of the form

$$V_{n,k} \prod_{j=1}^{k} c(n_j)$$

for any $n = 1, 2, 3...$ and $1 \leq k \leq n$ if and only if

$$c(n_j) = (1 - \sigma)_{n_j - 1}$$

for some $\sigma \in [-\infty, 1]$ and also $V_{n,k} = (n - \sigma k)V_{n+1,k} + V_{n+1,k+1}$, with $(1 - \sigma)_{n_j - 1} = 1$ when $\sigma = -\infty$ and that ρ_n reduces to the singleton partition when $\sigma = 1$.

In other words, if we restrict our choice on the cohesion functions among $c(S_j)$ to only depend on the cardinality of clusters, the prior on partitions will be given by a suitable EPPF which must be of the form

$$p(\rho_n = S_1, ..., S_k) \propto \prod_{j=1}^{k} c(S_j) = \prod_{j=1}^{k} (1-\sigma)_{|S|-1} \tag{4.13}$$

This way, the resulting link between PPM and nonparametric priors is not limited to the DP framework, but allows for the more flexibility given by the role played by the parameter σ. We delve into the more flexibility of this framework and to its advantages in the following section.

4.2 The need to go beyond the Dirichlet

As just outlined, it is possible to extend the use of nonparametric priors for PPM to a broader class, called Gibbs-type priors. But what is the advantage of such a choice? As always, we want to obtain a model which accommodates well different data features. The use of a broader class of priors allows us to gain flexibility and to better adapt our prior beliefes to different data structures. Indeed, DP's assumptions and implied characteristics of data are often restrictive.

4.2.1 Limitations of the DP

We start introducing some limits characterizing a prior specified using a DP.

The growth of number of clusters

The first important feature is the growth rate of the number of clusters. Denoting by K_n the number of distinct values $\{\theta_1^*, ..., \theta_k^*\}$ generated from some discrete random measure G, it can be shown (Korwar, Hollander, 1973) that, under a DP specification, K_n grows logarithmically

to ∞. In particular

$$P \sim DP(\alpha, G_0) \Longrightarrow K_n \sim \alpha \log(n) \tag{4.14}$$

Hence, such a behaviour is unaffected by the features of the sample $\theta^{(n)}$, a characteristic that seems pretty restrictive for applications. Moreover, a logarithmic growth is not a realistic assumption for many real-world problems which enjoy the so-called power law distribution. (see Mitzenmacher (2004) for an extensive review and Teh (2006) on language modeling).

Definition 8. Power law distribution A distribution p on positive integers $k \in \mathbb{N}^*$ is a power-law if $\exists$ c, a > 0 such that $p(k) = ck^{-a}$

Denoting by $K_{l,n}$ the multiplicities, i.e. $K_{l,n} = \sum_{j=1}^{K_n} \mathbb{1}(n_j = l)$, we say that the exchangeable sequence $\{\theta_1, ..., \theta_n\}$ exhibits a power-law behavious if it induces distributions on K_n and on $K_{l,n}$, $1 \leq l \leq n$, which are asymptotically power-law. By extension, we say that an exchangeable sequence $\theta_1, ..., \theta_n$ exhibits a power-law behaviour if it induces distributions on K_n and on $K_{l,n}$, $1 \leq$ l $\leq$ n which are asymptotically power law. This is the case of Gibbs-type processes such as the Pitman-Yor process.

The predictive distribution

Recall that the aim of this work is to propose a flexible method to allow for out-of-sample prediction in the presence of missing data. Hence, an important characteristic we want to focus on when reviewing this class of nonparametric prior is its predictive distribution. In line with the notions introduced in section 4.1, we can express the predictive distribution using the EPPF

$$P(\theta_{n+1} \in \cdot | \theta^{(n)}) = \frac{p_{k+1}^{(n+1)}(n_1, ..., n_k, 1)}{p_k^{(n)}(n_1, ..., n_k)} G_0(\cdot) + \sum_{j=1}^{k} \frac{p_k^{(n+1)}(n_1, ..., n_j + 1, ..., n_k)}{p_k^{(n)}(n_1, ..., n_k)} \delta_{\theta_j^*}(\cdot) \tag{4.15}$$

And by the definition of Gibbs-type priors we gave, we will have

$$P(\theta_{n+1} \in \cdot | \theta^{(n)}) = \frac{V_{n+1,k+1}}{V_{n,k}} G_0(\cdot) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{j=1}^{k} (n_j - \sigma) \delta_{\theta_j^*}(\cdot) \tag{4.16}$$

this expression allows us to appreciate an operational limitation of the DP. We can look at (4.17) in order to have an intuitive sense of the cluster allocation and creation mechanism. Given the sample $\{\theta_1, ..., \theta_n\}$, the first step consists of allocating the mass between a new value θ^*_{n+1} sampled from G_0 and the set of observed *species* $\{\theta^*_1, ..., \theta^*_k\}$ (more about this step in the next subsection). In the second step, if we condition on θ_{n+1} being one of the previously observed one, we will have that it depends on two elements, the size $n_j$but also the parameter σ. As far as the size n_j is concerned, we have that, as in the DP, the more often a past observation is detected, the higher the probability of re-observing it. At the same time, diffently from the DP, there is a parallel mechanism driven by σ. In particular, the ratio of the probabilities assigned to any pair of (θ^*_i, θ^*_j) is given by $(n_i - \sigma)(n_j - \sigma)$. Hence, for $\sigma > 0$, a reinforcing mechanism, such that if $n_i > n_j$ the ratio is an increasing function of σ, is in place. In other words, the "richer gets richer" mechanism tends to reinforce. On the other hand, if $\sigma < 0$, the reinforcing mechanism works in the opposite way. Finally, if $\sigma \to 0$, we are back to the DP, where σ does not enter in the allocation procedure.

The probability of observing a new cluster

Related to the growth rate of the number of clusters, there is another key element characterising a RPM: the probability of observing a new distinct value which has not been observed yet (hence, according to the species metaphore, a new specie), i.e.

$$P(\theta_{n+1} = \text{'new'} | \theta_1, ..., \theta_n) \tag{4.17}$$

having an EPPF, we simply apply the definition of conditional probability to get

$$P(\theta_{n+1} = \text{ 'new'} | \theta_1, ..., \theta_n) = \frac{p^{(n+1)}_{k+1}(n_1, ..., n_k, 1)}{p^{(n}_k(n_1, ..., n_k)} \tag{4.18}$$

A useful and interesting classification of species sampling models and Gibbs-type priors can be given in terms of the implied structure of the probability of generating a new value, as done by De Blasi et al. (2015).

Proposition 2. Let G be a species sampling model. Consider Θ as being a finite-dimensional parameter possibly entering the specification of G (such as α in the previously presented version of the DP) and K_n to be again the number of distinct values. Then the following classification in terms of the structure of the probability of generating a new value holds:

1. $\mathrm{P}(\theta_{n+1} = \text{'new'} \mid \theta^{(n)}) = f(n, \Theta)$ *if and only if* G *is a Dirichlet process*;
2. $\mathrm{P}(X_{n+1} = \text{'new'} \mid \theta^{(n)}) = f(n, k, \Theta)$ *if and only if* G *is of Gibbs-type*;
3. $\mathrm{P}(\theta_{n+1} = \text{'new'} \mid \theta^{(n)}) = f(n, k, n_1, ..., n_k\Theta)$ *otherwise*

This characterization allows us to better appreciate the properties of Gibbs-type priors. Indeed, this category of priors seems to well balance analytical tractability and flexibility, making the learning mechanism for the probability of observing a new species or cluster also sensible to the number of species already observed, k. Looking at the first category, it seems too restrictive to let the probability of generating new values depend solely on the sample size n and on the parameter α. On the other hand, the general case gives rise to several analytical issues and typically complicated expressions. For a more extensive analysis regarding the problem of evaluating the probability of discovering a certain number of new species in a new sample of population units, see also Lijoi et al. (2007a). This theoretical result regarding the class of Gibbs type priors allows for its exploitation in applications which have at their core the discover of new elements (or species). As an example, see Battiston et al. (2018) for an application in a multi-armed bandit context.

4.2.2 Pitman-Yor

What we just briefly outlined are some elements which constrain the potential use of the DP and are overcame by the use of the more general class of Gibbs-type priors. We now introduce one specific nonparametric prior of such a family, the Pitman-Yor process. Recall

the Dirichlet process' predictive distribution (Polya Urn):

$$\theta_{n+1}|\theta_1, ..., \theta_n \sim \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{j=1}^{n} n_j \theta_j^* \tag{4.19}$$

and recall how one of the limitations of the DP concerns its inadequacy to allow K_n to follow a power-law. Hence, we would like to have a process which opens many more new clusters and at the same time prevents many of the old to grow and few of them to grow significantly. This can be achieved by taking the predictive and modifying the weights as follows: on one side, we want to discount the weights associated to old clusters, and this can be achieved by decreasing the probability n_j by $\frac{-\sigma}{\alpha + n}$ ($\sigma < 1$). This way, the *richer gets richer* property is counterbalanced, but still dominating for clusters with high n_j. On the other hand, the weight α associated to a new cluster can be increased by $\frac{\sigma k_n}{\alpha + n}$. Those adjustments lead to the predictive:

$$\theta_{n+1}|\theta_1, ..., \theta_n = \frac{\alpha + \sigma k_n}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{j=1}^{k_n} (n_j - \sigma) \delta_{\theta_j^*}^* \tag{4.20}$$

corresponding to a sampling scheme of the form

$$p(\theta_{n+1} = \theta_j^* | \theta_1^*, ..., \theta_k^*) \propto \begin{cases} n_j - \sigma & \text{for} \quad j = 1, ..., k \\ \alpha + k\sigma & \text{for} \quad j = k + 1 \end{cases} \tag{4.21}$$

Clearly, if $\sigma \to 0$, one recovers the DP. For completeness and in line with what we have seen before, we can also recover the EPPF of this process, which is

$$p(\rho_n) = \frac{\sigma^k (1 + \alpha/\sigma)_{(k-1)}}{(1 + \alpha)_{(n-1)}} \prod_{j=1}^{k} (1 - \sigma)_{(n_j - 1)} \tag{4.22}$$

Figure 1 shows the power-law behaviour of the Pitman-Yor process and how this depends on σ and α. We see that α controls the overall number of unique observations, while σ controls the asymptotic growth of the number of unique observations. In the last two plots, one can see the proportion of observations appearing only once. We can see that the asymptotic behaviour depends on σ but not on α, with larger σ's producing more rare observations.

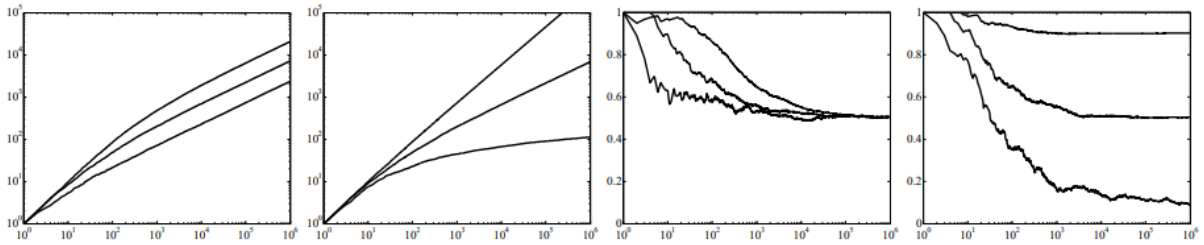


Figure 4.1: First plot: number of unique values as a function of the number of draws in a log-log scale, with $\sigma = 0.5$ and $\alpha = 1$ (bottom), $\alpha = 10$ (middle) and 100 (top). Second plot: $\alpha = 10$ and $\sigma = 0$ (bottom), $\alpha = 0.4$ (middle) and 0.9 (top). Third figure: proportion of values appearing only once, as a function of the number of draws, with values such as in the first pictures. Last figure: same, with values such as in the second picture

4.3 Missing data within the more general framework

We can now finally combine together all the elements we outlined in the previous sections. Indeed, we can proceed at first by combining more flexible priors (in this case a Pitman-Yor prior) within the PPMx framework, according to the result we present in section 4.1. Then, we can expand the previously presented strategy to handle data set missing covariates. This way, our rule for allocation in Neal's algorithm for training, also expanded to allow for missing entries by making use of $\tilde{g}(x_j^{*O})$, will be

$$p(s_n = j | s_{-i}, \mathbf{x}^O, \theta^*) \propto \begin{cases} (n_j^- - \sigma) & \dfrac{g(x_j^* \cup \{x_{n+1}\})}{g(x_j^*)} \quad p(x_i|\theta_j^*) \quad \text{for} \quad j = 1...k \\ (\alpha + k\alpha) & g(\{x_{n+1}\}) \quad p(x_i|\theta_j^*) \quad \text{for} \quad j = k+1 \end{cases} \tag{4.23}$$

from which we can easily obtain the prediction rule as we did in the case of DP-style priors.

4.4 Implemenation 2 - Simulation study

This second implementation aims at proving the effectiveness of our proposal. In this case, in order to better validate the model's performance under different circumstances, we make use of a simulated dataset. In particular, we simulate a dataset which shows a power law-distribution for the number of clusters. The dataset contains 200 observations, with a varying number of covariates. The covariates, which are simulated to be similar within the same cluster, are normalised. Moreover, in order to better allow for a comparison between our strategy and the one proposed in Chapter 3, we simulate a dataset characterised by a variability among observations which resembles the variability in the Ozone dataset. We first simulate the entire dataset and then delete some covariates according to missing patterns, such as shown is figure 4.2. Once data have been simulated and the missing entries have been deleted, we perform 10 different trials, shuffling the dataset before dividing it in training and testing observations in order to cross validate our model. To assess its validity, we consider both its ability to fit observed data and, more importantly, its predictive accuracy. For the latter, we measure the performance using a mean squared prediction error defined as

$$MPSE = \frac{1}{N_t} \sum_{i=1}^{N_t} (y_i - \hat{y}_i)^2$$

where N_t is the number of testing observations and $\hat{y}_i = E(y_i|\mathbf{y})$ is obtained through the chain as explained above. As a metric to evaluate the goodness of fit, we calculate a simple mean squared error (MSE).

4.4.1 Performance evaluation

Page et al. (2020) compared their algorithm employing DP-style cohesion function with common methods involving imputation, showing that it outperforms them, especially in cases of relatively high number of covariates. We take their approach as a benchmark to validate our proposal. Both strategies are tested on our simulated dataset. Figure 4.2 shows the comparison between our model and the model with DP-style cohesion functions when

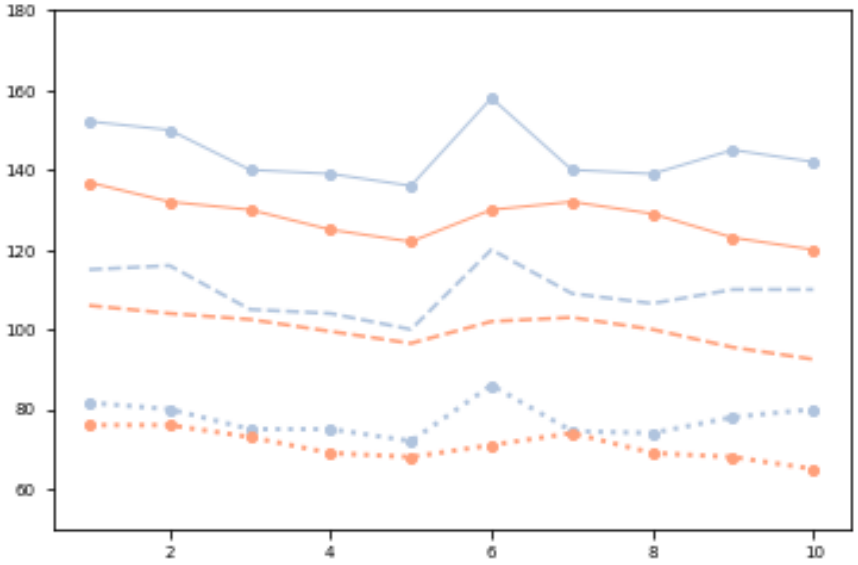


Figure 4.2: MSE and MPSE obtained by using both strategies. Orange lines refer to the model with PY-style cohesion functions and blue lines refer to DP-style cohesion functions. Continuous lines refer to MPSE, that is the performance in forecasting out-of-sample observations. Dotted lines in the lower part of the chart refer to the fitting performance of the model (MSE for in-sample observations). The two lines in-between refer to the overall performance

a power-law dataset is in used. Results are reported for 10 trials. In each of them, 3000 iterations were carried out and then the first 2000 were discarded. Predictions for each observation are averaged over iterations. In all our trials, predictions based on Gibbs-type cohesion functions seem to out perform the original model. Indeed, the use of a different prior seems to better reflect the underlying structure of the dataset from a conceptual point of view, this translating into a better prediction performance. In this case, the missing entries are deleted under a missing at random (MAR) pattern.

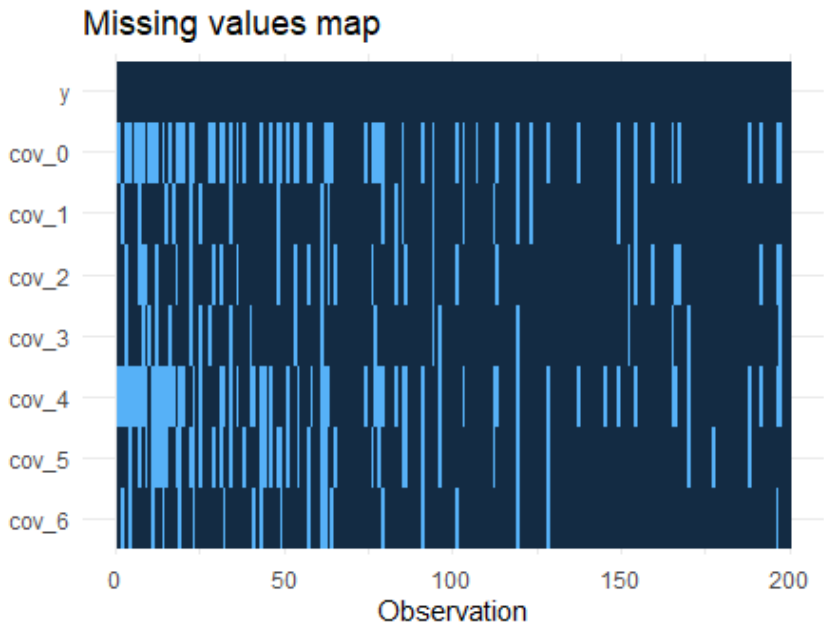
4.4.2 Testing for number of covariates of missingness mechanisms

We test the model's performance under different assumptions regarding the number of covariates informing the partition and the missingness pattern for the data.

Testing for different missingness mechanisms

As far as the model's performance when handling different missingness mechanisms is concerned, we decided to test its behaviour under MCAR, MAR and MNAR assumptions as introduced in Chapter 3. In all the trials, we consider a model with 7 covariates, this number being chosen as the median case with respect to the simulations with different number of covariates shown below. What we find is that, as far as the percentage of missingness being the same, there is no significance evidence of the model performing better in one of those cases. This result could well reflect the flexibility of this model, which easily accommodates out-of-sample prediction even when the missing pattern of the new observation has not been observed yet. When non-completely-random mechanisms such as MNAR (see figure 4.3) are in place, the performance varies depending on which clusters being the ones more affected. Overall, if data are systematically missing for observations which would naturally take part to the same clusters, grouping those observations under the PPMX framework could be harder, this way worsening the predictive performance for such observations.Under each assumption, 10 trials are performed, each corresponding to a different missing pattern being generated. Means obtained averaging over the trials are reported in the Appendix, together with a graphical representation of MCAR and MAR in our dataset. Overall, we believe that the variability observed in the outcome suggests that there is need of further investigation. We also notice how the performance differs between MSE and MPSE. Still, our strategy seems to outperform its DP-style counterpart under all assumptions.

Figure 4.3: Missing pattern after amputation has been carried out starting from the full dataset. In this case, data are missing under the not at random assumption (MNAR). In partcular, the missing rate has been chosen to be equal to 0.3, 0.1, 0.15, 0.1, 03, 01 for each covariate respectively



Testing for the number of covariates informing the partition

We also evaluated our model by looking at its performance under different lengths of covariates' vector. Starting with a full dataset with 12 covariates, we generate missing entries using a MCAR mechanism and we then test the model by sequentially deleting columns of the dataset. Intuitively, the learning process of the model benefits from more covariates contributing to inform the partition. Again, using the mode with DP-style cohesion funcyions as benchmark, our strategy ameliorates its performance. The magnitude of the increment varies, but seems to also depend on which columns are selected to remain part of the dataset when the number of covariates is sistematically decreased (i.e. on the missingness percentage of the covariates remaining in the dataset and on their informative power).

Conclusion and further directions

We started from an extension of the PPMx which is able to accommodate missing values without having to implement any type of imputation. We then presented our contribution, a model which naturally expands that strategy and aims at better accomodating certain data clustering structures, such as datasets with power-law distributed number of clusters. Repeatedly tested on both in-sample and out-of sample observations, we conclude that the predictive performance increases. This approach seems to benefit from a better conceptual understanding of the underlying data structure, this way allowing for the opening of new research directions in fields where data shares the same latent pattern. Indeed, this data structure does not only characterise our simple simulated dataset but, as already pointed out, it is a feature shared among datasets in different fields. Power laws appear widely in physics, biology, earth and planetary sciences, economics and finance, computer science, demography and the social sciences. Hence, future research directions regard the use of the presented method as a starting point for predictive strategies dealing with such data. In particular, one possible research direction concerns the use of this strategy in problems regarding missing data in topic modeling, such as in downloaded texts or self-reported surveys. Moreover, the flexibility of our methodology in allowing for the use of mixed data types seems to be potentially useful in cases of network data, which also present power-law distributions in many of their features

Bibliography

Aldous David. A Conversation with Jim Pitman // Statistical Science. 2018. 33, 3. 458–467.

Barry Daniel, Hartigan John A. Product partition models for change point problems // The Annals of Statistics. 1992. 260–279.

Battiston Marco, Favaro Stefano, Teh Yee Whye. Multi-armed bandit for species discovery: a Bayesian nonparametric approach // Journal of the American Statistical Association. 2018. 113, 521. 455–466.

Blackwell David, MacQueen James. Ferguson distributions via Pólya urn schemes // The annals of statistics. 1973. 1, 2. 353–355.

Burgette Lane F, Reiter Jerome P. Multiple imputation for missing data via sequential regression trees // American journal of epidemiology. 2010. 172, 9. 1070–1076.

Bush Christopher A, MacEachern Steven N. A semiparametric Bayesian model for randomised block designs // Biometrika. 1996. 83, 2. 275–285.

Canale Antonio, Lijoi Antonio, Nipoti Bernardo, Prünster Igor. On the Pitman–Yor process with spike and slab base measure // Biometrika. 2017. 104, 3. 681–697.

Daniels Michael J, Hogan Joseph W. Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis. (2008).

De Blasi Pierpaolo, Favaro Stefano, Lijoi Antonio, Mena Ramses H., Prunster Igor, Ruggiero Matteo. Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process? // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015. 37, 2. 212–229.

De Finetti, Bruno . Sur la condition d' Equivalence partielle. 1938.

Escobar Michael D, West Mike. Bayesian density estimation and inference using mixtures // Journal of the american statistical association. 1995. 90, 430. 577–588.

Favaro Stefano, Lijoi Antonio, Mena Ramsés H, Prünster Igor. Bayesian non-parametric inference for species variety with a two-parameter Poisson–Dirichlet process prior // Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2009. 71, 5. 993–1008.

Favaro Stefano, Lijoi Antonio, Pruenster Igor, others . Conditional formulae for Gibbs-type exchangeable random partitions // Annals of Applied Probability. 2013a. 23, 5. 1721–1754.

Favaro Stefano, Lijoi Antonio, Prünster Igor. A new estimator of the discovery probability // Biometrics. 2013b. 68, 4. 1188–1196.

Ferguson Thomas S. A Bayesian analysis of some nonparametric problems // The annals of statistics. 1973. 209–230.

Fletcher Mercaldo Sarah, Blume Jeffrey D. Missing data and prediction: the pattern submodel // Biostatistics. 2020. 21, 2. 236–252.

Gelman Andrew. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper) // Bayesian analysis. 2006. 1, 3. 515–534.

Gnedin Alexander, Pitman Jim. Exchangeable Gibbs partitions and Stirling triangles // Journal of Mathematical sciences. 2006. 138, 3. 5674–5685.

Hartigan John A. Partition models // Communications in statistics-Theory and methods. 1990. 19, 8.

Janssen Kristel JM, Vergouwe Yvonne, Donders A Rogier T, Harrell Jr Frank E, Chen Qingxia, Grobbee Diederick E, Moons Karel GM. Dealing with missing predictor values when applying clinical prediction models // Clinical chemistry. 2009. 55, 5. 994–1001.

Johnson Valen E, Albert James H. Ordinal data modeling. 2006.

Kehagias Ath, Nicolaou A, Petridis V, Fragkou P. Text segmentation by product partition models and dynamic programming // Mathematical and Computer Modelling. 2004. 39, 2-3. 209–217.

Kingman John FC. Random partitions in population genetics // Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences. 1978a. 361, 1704. 1–20.

Kingman John FC. The representation of partition structures // Journal of the London Mathematical Society. 1978b. 2, 2. 374–380.

Kingman John FC. The coalescent // Stochastic processes and their applications. 1982. 13, 3. 235–248.

Korwar Ramesh M, Hollander Myles. Contributions to the theory of Dirichlet processes // The Annals of Probability. 1973. 1, 4. 705–711.

Lee Jaeyong, Quintana Fernando, Müller Peter, Trippa Lorenzo. Defining Predictive Probability Functions for Species Sampling Models // Statistical science : a review journal of the Institute of Mathematical Statistics. 2013. 28. 209–222.

Legramanti Sirio, Rigon Tommaso, Durante Daniele, Dunson David B. Extended Stochastic Block Models with Application to Criminal Networks. 2021).

Lijoi Antonio, Mena Ramsés H, Prünster Igor. Bayesian nonparametric estimation of the probability of discovering new species // Biometrika. 2007a. 94, 4. 769–786.

Lijoi Antonio, Mena Ramsés H, Prünster Igor. A Bayesian nonparametric method for prediction in EST analysis // BMC bioinformatics. 2007b. 8, 1. 1–10.

Lijoi Antonio, Mena Ramsés H, Prünster Igor. Controlling the reinforcement in Bayesian non-parametric mixture models // Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2007c. 69, 4. 715–740.

Lijoi Antonio, Mena Ramses H, Prünster Igor. A Bayesian nonparametric approach for comparing clustering structures in EST libraries // Journal of Computational Biology. 2008a. 15, 10. 1315–1327.

Lijoi Antonio, Muliere Pietro, Prünster Igor, Taddei Filippo. Innovation, growth and aggregate volatility from a Bayesian nonparametric perspective // Electronic Journal of Statistics. 2016. 10, 2. 2179–2203.

Lijoi Antonio, Prünster Igor. Models beyond the Dirichlet process // Bayesian nonparametrics. 2010. 28, 80. 342.

Lijoi Antonio, Prünster Igor, Walker Stephen G. Bayesian nonparametric estimators derived from conditional Gibbs structures // Annals of applied probability. 2008b. 18, 4. 1519–1547.

Little Roderick JA. Regression with missing X's: a review // Journal of the American statistical association. 1992. 87, 420. 1227–1237.

MacEachern Steven N, Müller Peter. Estimating mixture of Dirichlet process models // Journal of Computational and Graphical Statistics. 1998. 7, 2. 223–238.

Mitzenmacher Michael. A brief history of generative models for power law and lognormal distributions // Internet mathematics. 2004. 1, 2. 226–251.

Molenberghs Geert, Fitzmaurice Garrett, Kenward Michael G, Tsiatis Anastasios, Verbeke Geert. Handbook of missing data methodology. (2014).

Müller Peter, Quintana Fernando, Rosner Gary L. A product partition model with regression on covariates // Journal of Computational and Graphical Statistics. 2011. 20, 1. 260–278.

Müller Peter, Quintana Fernando A. Nonparametric Bayesian data analysis // Statistical science. 2004. 95–110.

Müller Peter, Quintana Fernando Andrés, Jara Alejandro, Hanson Tim. Bayesian nonparametric data analysis. 2015).

Neal Radford M. Markov chain sampling methods for Dirichlet process mixture models // Journal of computational and graphical statistics. 2000. 9, 2. 249–265.

Paganin Sally, Herring Amy H, Olshan Andrew F, Dunson David B. Centered Partition Processes: Informative Priors for Clustering // Bayesian Analysis. 2021.

Page Garritt, Quintana Fernando. Spatial Product Partition Models // Bayesian Analysis. 04 2015. 11.

Page Garritt L., Quintana Fernando A., Müller Peter. Clustering and Prediction with Variable Dimension Covariates. 2020.

Page Garritt L, Quintana Fernando A, others . Predictions based on the clustering of heterogeneous functions via shape and subject-specific covariates // Bayesian Analysis. 2015. 10, 2. 379–410.

Papaspiliopoulos Omiros, Roberts Gareth O. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models // Biometrika. 2008. 95, 1. 169–186.

Pitman Jim. Exchangeable and partially exchangeable random partitions // Probability theory and related fields. 1995. 102, 2. 145–158.

Pitman Jim. Some developments of the Blackwell-MacQueen urn scheme // Lecture Notes-Monograph Series. 1996. 245–267.

Quintana Fernando A, Iglesias Pilar L. Bayesian clustering and product partition models // Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2003. 65, 2. 557–574.

Rodriguez Abel, Müller Peter. Nonparametric Bayesian Inference // NSF-CBMS Regional Conference Series in Probability and Statistics. 2013. 9. i–110.

Rubin Donald B. Inference and missing data // Biometrika. (1976). 63, 3. 581–592.

Rubin Donald B. An overview of multiple imputation // Proceedings of the survey research methods section of the American statistical association. 1988. 79–84.

Stekhoven Daniel J, Bühlmann Peter. MissForest—non-parametric missing value imputation for mixed-type data // Bioinformatics. 2012. 28, 1. 112–118.

Storlie Curtis B, Therneau Terry M, Carter Rickey E, Chia Nicholas, Bergquist John R, Huddleston Jeanne M, Romero-Brufau Santiago. Prediction and inference with missing data in patient alert systems // Journal of the American Statistical Association. 2019.

Su Yu-Sung, Gelman Andrew E, Hill Jennifer, Yajima Masanao. Multiple imputation with diagnostics in R: Opening windows into the black box // Journal of Statistical Software. 2011.

Teh Yee Whye. A hierarchical Bayesian language model based on Pitman-Yor processes // Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. 2006). 985–992.

Van Buuren Stef. Flexible imputation of missing data. 2018.

Walker Stephen G. Sampling the Dirichlet mixture model with slices // Communications in Statistics—Simulation and Computation. 2007. 36, 1. 45–54.

West Mike, Müller Peter, Escobar Michael D. Hierarchical Priors and Mixture Models, with Applications in Regression and Density Estimation // Aspects of Uncertainty: A Tribute to D. V. Lindley. 1994. 363–386.

Xu Dandan, Daniels Michael J, Winterstein Almut G. Sequential BART for imputation of missing covariates // Biostatistics. 2016. 17, 3. 589–602.

Yao Yi-Ching. Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches // The Annals of Statistics. 1984. 1434–1447.

Appendix A

Auxiliary model for similarity functions

Consider the auxiliary model

$$g(\mathbf{x}_j^*) = \int \prod_{i \in S_h} p(x_i|\xi_j)p(\xi_j)d\xi_j$$

In cases of missing covariates though, applying this strategy is not feasible, but we have to calculate the similarity function for each covariate separately and to apply the following formula

$$g(\mathbf{x}_j) = \prod_{l=1}^{d} g_l(\mathbf{x}_{jl}) \tag{A.1}$$

as suggested by Page et al. (2020). Hence, instead of a Normal-Inverse-Wishart conjugate model, we opt for a Normal-Gamma model with $\xi_j = (\mu_j, \sigma_j^2)$ and $\lambda_j = \frac{1}{\sigma_j^2}$ and we will have

$$p(\mathbf{x}_j|\mu_j, \lambda_j) = \frac{1}{(2\pi)^{n/2}} \lambda^{n/2} exp\left(-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \tag{A.2}$$

with conjugate prior

$$NG(\mu, \lambda|\mu_0, \kappa_0, \alpha_0, \beta_0) = N(\mu|\mu_0, (\kappa_0\lambda)^{-1})Ga(\lambda|\alpha_0, \beta_0) \tag{A.3}$$

$$= \frac{1}{Z_{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0)}\lambda^{1/2}exp\left(-\frac{\kappa_0\lambda}{2}(\mu-\mu_0)^2\right)\lambda^{\alpha_0-1}e^{-\lambda\beta_0} \tag{A.4}$$

$$= \frac{1}{Z_{NG}}\lambda^{\alpha-1/2}exp\left(-\frac{\lambda}{2}[\kappa_0(\mu-\mu_0)^2+2\beta_0]\right) \tag{A.5}$$

$$= \frac{\Gamma(\alpha_0)}{\beta_0^{\alpha_0}}\left(\frac{2\pi}{\kappa_0}\right)^{\frac{1}{2}} \tag{A.6}$$

which has as corresponding posterior

$$p(\mu, \lambda|\mathbf{x}) = NG(\mu, \gamma|\mu_n, \kappa_n, \alpha_n, \beta_n) \tag{A.7}$$

$$\mu_n = \frac{\kappa_0\mu_0 + n\overline{x}}{\kappa_0 + n} \tag{A.8}$$

$$\kappa_n = \kappa_0 + n \tag{A.9}$$

$$\alpha_n = \alpha_0 + n/2 \tag{A.10}$$

$$\beta_n = \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \overline{x})^2 + \frac{\kappa_0 n(\overline{x}-\mu_0)^2}{2(\kappa_0+n)} \tag{A.11}$$

and following the same rationale we follow before to obtain the marginal likelihood for observations, we can write:

$$p(\mu_j, \lambda_j|\mathbf{x}_j) = \frac{1}{p(\mathbf{x}_j)}\frac{1}{Z_0}NG'(\mu_j, \lambda_j|\cdot)\left(\frac{1}{2\pi}\right)^{n_j/2}\prod_{i}^{n_j}N'(x_i|\mu_j, \lambda_j) \tag{A.12}$$

$$= \frac{1}{Z_n}NG'(\mu_j, \lambda_j|\cdot) \tag{A.13}$$

where N' and NG' are the unnormalised distributions and Z_0 and Z_n being the corresponding normalizing constant. Then we will have

$$p(\mathbf{x}_j) = \frac{Z_n}{Z_0}(2\pi)^{n_j/2} = \frac{\Gamma(\alpha_{n_j})}{\Gamma(\alpha_0)}\left(\frac{\kappa_0}{\kappa_{n_j}}\right)^{1/2}\frac{1}{(2\pi)^{n/2}} \tag{A.14}$$

Appendix B

PPM and DP induced marginal distribution

In section 2.2, we mentioned how the connection between PPM and DP mixture models can be made more clear by deriving the implied marginal distribution for the observations obtained in the two cases. In order to show this equivalence, we start recalling the distribution on partitions of $\{\theta_1, ..., \theta_n\}$ as derived in (2.18) and expressed in terms of $|S_j|$ that is

$$p(n_1, ..., n_k) = \frac{\alpha^k}{\alpha_{(n)}} \prod_{j=1}^{k} (|S_j|-1)! \tag{B.1}$$

By the law of total probability we get

$$p(\theta) := p(\theta_1, ..., \theta_n) = \sum_{\rho_n \in P_n} p(\theta|\rho_n) p(\rho_n) \tag{B.2}$$

that is

$$p(\theta) = \sum_{\rho_n \in P_n} \frac{\alpha^k}{\alpha_{(n)}} \prod_{j=1}^{k} (|S_j|-1)!\, p(\theta|\rho_n) \tag{B.3}$$

Conditional on ρ_n, we can denote $(\theta_{j,1}, ..., \theta_{j,|S_j|})$ the observations into the group j. Notice that $\{\theta_{j,i}$ are all equal to a value sampled from G_0 and that the groups are independent among each other, this implying that we can sample the first observation in each group

independently from G_0 and assign the same values to the others, that is

$$p(\theta|\rho_n) = \prod_{j=1}^{k} G_0(\theta_{j,1}) \prod_{i=2}^{|S_j|} \delta_{\theta_{j,i}}(\theta_{j,i}) \tag{B.4}$$

$$p(\theta) = \sum_{\rho \in \mathcal{P}} \frac{1}{\prod_{l=1}^{n}(\alpha + l - 1)} \prod_{j=1}^{k} \alpha(|S_j|-1)!\, G_0(\theta_j, 1) \prod_{i=2}^{|S_j|} \delta_{\theta_{j,1}}(\theta_{j,i}) \tag{B.5}$$

which, using the notation used in (1) and (3), is equivalent to

$$p(\theta) \propto \sum_{\rho \in \mathcal{P}} \prod_{j=1}^{K} \alpha(|S_j|-1)!\, p_{S_j}(\theta_j^*) \tag{B.6}$$

$$p(\theta, \mathbf{y}) \propto \sum_{\rho \in \mathcal{P}} \prod_{j=1}^{k} \alpha(|S_j|-1)!\, p(\mathbf{y}|\theta_j^*) p_{S_j}(\theta_j^*) \tag{B.7}$$

$$p(\theta, \mathbf{y}) \propto \sum_{\rho \in \mathcal{P}} \prod_{j=1}^{k} \prod_{i=1}^{|S_j|} \alpha(|S_j|-1)!\, p(y_i|\theta_j^*) p_{S_j}(\theta_j^*) \tag{B.8}$$

$$p(\mathbf{y}) \propto \sum_{\rho \in \mathcal{P}} \prod_{j=1}^{K} \alpha(|S_j|-1)! \int \prod_{i=1}^{|S_j|} p(y_i|\theta) dG_0(\theta) \tag{B.9}$$

which can be brought back to the marginal implied in a PPM with c as in (3). Indeed, going back to the hierarchical PPM, the implied marginal on the observations corresponds to

$$p(\mathbf{y}) = \sum_{\rho \in \mathcal{P}} \prod_{j=1}^{k} \prod_{i=1}^{|S_j|} c(S_j) \int \prod_{i=1}^{|S_j|} p(y_i|\theta) dG_0(\theta) \tag{B.10}$$

such that by choosing a *cohesion function* as the one suggested in (B.1) one obtains exactly the marginal as in (B.9). Hence, for this choice of c, the integrated-out nonparametric model can be seen as a special case of a PPM.

Appendix C

Parameters update

Here we specify the formulas for updating clusters' parameters and parameters common to all clusters, obtained via simple algebra. When the full posterior was available in close form, we made use of a Gibbs sampling step, as for clusters' and common mean.
For common mean

$$p(\mu_0|\mu_j, m_0, v, \mu_0, \sigma_0, \mathbf{y}) = p(\mu_0|\mu_j, m_0, v, \sigma_0) \propto N\left(\frac{v^2 \sum_{j=1}^{k} \mu_j + \sigma_0^2 m_0}{nv^2 + \sigma_0^2}; \frac{v^2\sigma_0^2}{nv^2 + \sigma_0^2}\right)$$

For cluster mean

$$p(\mu_j|m_0, v, \mu_0, \sigma_j, \sigma_0, \mathbf{y}) = p(\mu_j|\mu_0, \sigma_j, \sigma_0, \mathbf{y}) \propto N\left(\frac{\sigma_0^2 \sum_{i=1}^{n_j} y_i + \sigma_j^2 \mu_0}{n_j\sigma_0^2 + \sigma_j^2}; \frac{\sigma_0^2\sigma_j^2}{n_j\sigma_0^2 + \sigma_j^2}\right)$$

For clusters' variance and common variance we use a Metropolis step with a Gaussian proposal.

Appendix D

Missingness patterns

The simulated model tested in Chapter 4 was tested for different missingness patterns as introduced in chapter 3. Figure D.1 reports the means obtained averaging over 10 trials with the same dataset. Figure D.2 shows the specific simulated missingness when testing the model with 7 covariates. The figure on the left shows the MCAR case, while the picture on the right shows the MAR case. In both cases the missing shares are respectively 0.3, 0.1, 0.15, 0.1, 0.3, 0.2, 0.1.

	MCAR	MAR	MNAR
MSE	52.88	52.68	53.00
MPSE	137.22	136.55	138.53
MSE(DP)	56.77	53.21	59.95
MPSE(DP)	143.99	145.42	142.25

Figure D.1

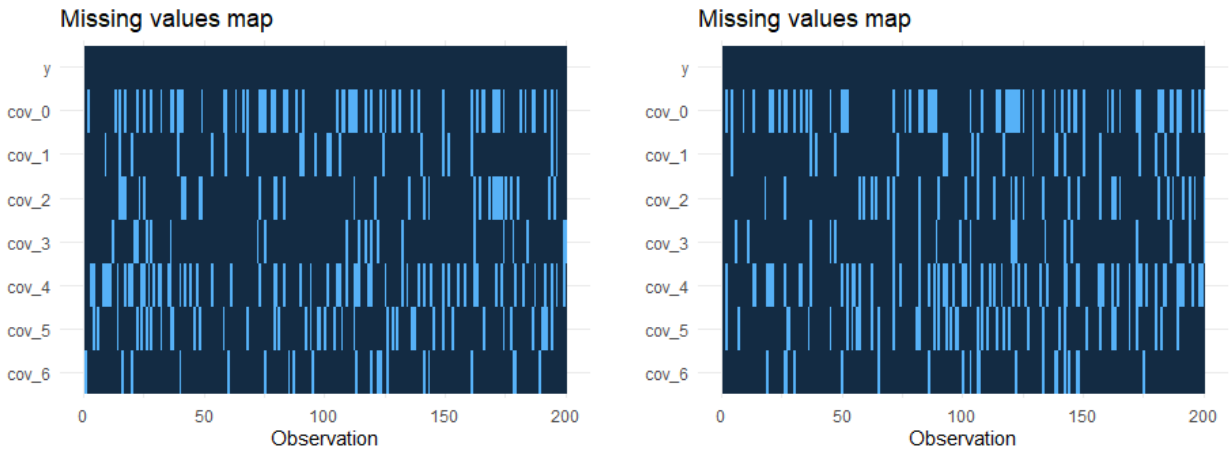

Missing values map
y
cov_0
cov_1
cov_2
cov_3
cov_4
cov_5
cov_6
0
50
100
150
200
Observation
Missing values map
y
cov_0
cov_1
cov_2
cov_3
cov_4
cov_5
cov_6
0
50
100
150
200
Observation

Figure D.2