SCUOLA DI DOTTORATO

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di / Department of

Statistics and Quantitative Methods

Dottorato di Ricerca in / PhD program   Statistics and Mathematical Finance   Ciclo / Cycle   XXXI

Curriculum in Statistics

CONTRIBUTIONS TO MODELLING VIA BAYESIAN NONPARAMETRIC MIXTURES

Cognome / Surname   Corradin   Nome / Name   Riccardo

Matricola / Registration number   727524

Tutore / Tutor: Prof. Riccardo Borgoni

Supervisor: Prof. Bernardo Nipoti

Coordinatore / Coordinator: Prof. Giorgio Vittadini

ANNO ACCADEMICO / ACADEMIC YEAR   2018/2019

UNIVERSITY OF MILANO BICOCCA

DOCTORAL THESIS

Contributions to modelling via Bayesian nonparametric mixtures

Author:
Riccardo CORRADIN

Supervisor:
Prof. Bernardo NIPOTI

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

in the

Department of Statistics and Quantitative Methods

31 October 2018

Abstract

Riccardo CORRADIN

Contributions to modelling via Bayesian nonparametric mixtures

Bayesian nonparametric mixtures are flexible models for density estimation and clustering, nowadays a standard tool in the toolbox of applied statisticians. The first proposal of such models was the Dirichlet process (DP) (Ferguson, 1973) mixture of Gaussian kernels by Lo (1984), contribution which paved the way to the definition of a wide variety of nonparametric mixture models. In recent years, increasing interest has been dedicated to the definition of mixture models based on nonparametric mixing measures that go beyond the DP. Among these measures, the Pitman-Yor process (PY) (Perman et al., 1992; Pitman, 1995) and, more in general, the class of Gibbs-type priors (see e.g. De Blasi et al., 2015) stand out for conveniently combining mathematical tractability, interpretability and modelling flexibility.

In this thesis we investigate three aspects of nonparametric mixture models, which, in turn, concern their modelling, computational and distributional properties. First, we study the effect of affine transformations of the data on posterior inference based on DP mixture models. Second, we propose a new efficient Markov chain Monte Carlo algorithm, named Importance Conditional Sampler (ICS), for sampling from the posterior distribution of PY mixture models and we show that, unlike state-of-the art competing algorithms, the efficiency of the proposed sampler is robust to the value of the parameters of the PY. Finally, we study some prior distributional properties of Gibbs-type priors and devise a simple strategy for the elicitation of the parameters of the prior process in Gibbs-type mixture models, based on available prior information on the size of the clusters underlying the data.

The thesis is organized as follows. The first chapter proposes a coincise review of the area of Bayesian nonparametric statistics, with focus on tools and models that will be considered in the following chapters. We first introduce the notions of exchangeability, exchangeable partitions and discrete random probability measures. We then focus on the DP and the PY case, main ingredients of second and third chapter, respectively. Finally, we briefly discuss the rationale behind the definition of more general classes of discrete nonparametric priors.

In the second chapter we propose a thorough study on the effect of invertible affine transformations of the data on the posterior distribution of DP mixture models, with particular attention to DP mixtures of Gaussian kernels (DPM-G). First, we provide an explicit result relating model parameters and transformations of the data. Second, we formalize the notion of asymptotic robustness of a model under affine transformations of the data and prove an asymptotic result which, by relying on the asymptotic consistency of DPM-G models, show that, under mild assumptions on the data-generating distribution, DPM-G are asymptotically robust.

The third chapter presents the ICS, a novel conditional sampling scheme for PY mixture models, based on a useful representation of the posterior distribution of a PY (Pitman, 1996) and on an importance sampling idea, similar in spirit to the augmentation step of the celebrated Algorithm 8 of Neal (2000). The proposed method conveniently combines the best features of state-of-the-art conditional and marginal methods for PY mixture models. Importantly, and unlike its most popular conditional competitors, the numerical efficiency of the ICS is robust to the specification of the parameters of the PY. The steps for implementing the ICS are described in detail and its performance is compared with popular competing algorithms. Finally, the ICS is used as a building block for devising a new efficient algorithm for the class of GM-dependent DP mixture models (Lijoi et al., 2014a; Lijoi et al., 2014b), for partially exchangeable data.

In the fourth chapter we study some distributional properties of Gibbs-type priors. The main result focuses on an exchangeable sample from a Gibbs-type prior and provides a conveniently simple description of the distribution of the size of the cluster where the $(m+1)$th observation is assigned to, given an unobserved sample of size m. The study of such distribution provides the tools for a simple, yet useful, strategy for prior elicitation of the parameters of a Gibbs-type prior, in the context of Gibbs-type mixture models.

The results in the last three chapters are supported by exhaustive simulation studies and illustrated by analysing astronomical datasets.

Contents

List of Figures

List of Tables

List of Abbreviations

BNP	**B**ayesian **N**on**p**arametric
CRM	**C**ompletely **R**andom **M**easure
DDP	**D**ependent **D**irichlet **P**rocess
DDPM	**D**ependent **D**irichlet **P**rocess **M**ixture model
DP	**D**irichlet **P**rocess
DPM	**D**irichlet **P**rocess **M**ixture model
DPM-G	DPM with **G**aussian kernel
EPPF	**E**xchangeable **P**artition **P**robability **F**unction
GM	**G**riffiths **M**ilne
GM-DDP	GM **D**ependent **D**irichlet **P**rocess
ICS	**I**mportance **C**onditional **S**ampler
hNRMI	**h**omogeneous **N**ormalized **R**andom **M**easure with **I**ndependent increments
MCMC	**M**onte **C**arlo **M**arkov **C**hain
MS	**M**arginal **S**ampler
NGG	**N**ormalized **G**eneralized **G**amma Process
NRMI	**N**ormalized **R**andom **M**easure with **I**ndependent increments
PY	**P**itman-**Y**or process
PYM	**P**itman-**Y**or **M**ixture model
PYM-G	PYM with **G**aussian kernel
RPM	**R**andom **P**robability **M**easure
RS	**R**etrospective **S**ampler
SS	**S**lice **S**ampler
VI	**V**ariation of **I**nformation

To Alice

Chapter 1

A concise introduction to Bayesian nonparametric statistics

He who loves practice without theory
is like the sailor who boards ship
without a rudder and compass and
never knows where he may cast.

Leonardo Da Vinci, *Inventor, scientist,*
mathematician, astronomer, physicist,
engineer, writer, etc.

1.1 Exchangeability and de Finetti's representation theorem

Let $X^{(\infty)}$ be an infinite sequence of observations, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let X_i, for $i \geq 1$, be the i-th element of the sequence, taking values on a measurable space $(\mathbb{X}, \mathscr{X})$, with $\mathbb{X}$ Polish space and $\mathscr{X}$ its Borel σ-field. The sequence $X^{(\infty)}$ is said *exchangeable* if, for any $n \geq 1$ and any permutation σ of $\{1, \dots, n\}$, we have that $(X_1, \dots, X_n)$ coincides in distribution with $(X_{\sigma(1)}, \dots, X_{\sigma(n)})$. Exchangeability is a probabilistic statement concerning homogeneity among the observations. It implies that, for every $n \geq 1$, the order in which the observations of a sample $\mathbf{X}^{(n)} = (X_1, \dots, X_n)$ are collected is irrelevant.
A fundamental result on exchangeability, known as *de Finetti's representation theorem*, was proved by de Finetti (1937). Let $\mathbb{X}^{(\infty)} = \mathbb{X} \times \mathbb{X} \times \dots$ and $\mathbb{M}_{\mathbb{X}}$ be the space of probability measures on $(\mathbb{X}, \mathscr{X})$.

Theorem 1 (de Finetti, 1937)**.** *The sequence $X^{(\infty)}$ is exchangeable if and only if there exists a probability measure Q on $\mathbb{M}_{\mathbb{X}}$ such that, for any $n \geq 1$, and $A = A_1 \times A_2 \times \cdots \times A_n \times \mathbb{X}^{(\infty)}$, with $A_i \in \mathcal{X}$, one has*

$$\mathbb{P}\left[X^{(\infty)} \in A\right] = \int_{\mathbb{M}_{\mathbb{X}}} \prod_{i=1}^{n} p(A_i) Q(\mathrm{d}p).$$

Theorem 1 states that an infinite sequence is exchangeable if and only if it is possible to represent it as a mixture of independent and identical distributed random variables. The measure Q is named *de Finetti measure* and plays the role of prior distribution in Bayesian statistics. De Finetti's result thus provides a neat justification for the use of prior distributions, which are not only convenient but are implied by the assumption of exchangeability of the observations. As a result, Theorem 1, sets solid theoretical grounds to Bayesian reasoning. Based on de Finetti's representation theorem, the exchangeability assumption can be represented in hierarchical form as

$$\begin{aligned} X_i|\tilde{p} &\overset{iid}{\sim} \tilde{p} \qquad i \geq 1 \\ \tilde{p} &\sim Q, \end{aligned} \tag{1.1}$$

where $\tilde{p}$ is a random probability measure and its distribution Q is a probability measure over the space $\mathbb{M}_{\mathbb{X}}$. The support of Q identifies two main classes of model and, more in general, two areas of research in Bayesian statistics. Specifically,

- if the distribution Q has finite dimensional support, then Model 1.1 is called *parametric*;
- if the distribution Q has infinite dimensional support, then Model 1.1 is called *nonparametric*.

Nonparametric models represent the main focus of this manuscript.

1.2 Exchangeable partitions and EPPF

In this section we introduce the concepts of partition, random partition and exchangeable random partition. Early studies in this area were motivated by population genetics applications (see, for example, Kingman, 1978). For more details on these topics, one can refer to Pitman (2006).

Let $\mathbb{N}_n = \{1, \dots, n\}$, we say that $\boldsymbol{\psi}_n = \{A_1, \dots, A_k\}$ is a *partition* of $\mathbb{N}_n$ if the sets $\{A_1, \dots, A_k\}$ are nonempty and such that $\mathbb{N}_n = \cup_i A_i$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$. Let $\mathcal{B}$ be the space of all partitions of $\mathbb{N}_n$ and $\mathscr{B}$ its discrete σ-field.

Definition 1. *A random partition is any measurable function $\Psi_n : (\Omega, \mathcal{F}, \mathsf{P}) \to (\mathcal{B}, \mathscr{B})$.*

The definition of Ψ_n induces a probability measure on the space of partitions $\mathcal{B}$.

Given a partition $\boldsymbol{\psi}_n = \{A_1, \dots, A_k\}$ of $\mathbb{N}_n$, the sequence of cardinalities of the blocks of $\boldsymbol{\psi}_n$, that is $\mathbf{c}_n = (n_1, \dots, n_k)$, with $n_i = |A_i|$, is named composition of n. A random partition Ψ_n induces a random composition $\mathcal{C}_n$. Note that the space of all compositions of n is the simplex

$$\triangle_n^k = \left\{ (n_1, \dots, n_k) \in \mathbb{N}^k \, : \, \sum_{j=1}^k n_j = n, n_j > 0 \right\}.$$

Let σ be a permutation of $\mathbb{N}_n$ and $\sigma(A) = \{\sigma(i) \, : \, i \in A\}$ (for example, see Figure 1.1).

Definition 2 (Exchangeable partition). *We say that a random partition Ψ_n is exchangeable if, given any permutation $\sigma : \mathbb{N}_n \to \mathbb{N}_n$, the distribution of Ψ_n is invariant with respect to σ, i.e. for any partition $\{A_1, \dots, A_k\}$ of $\mathbb{N}_n$ and for any permutation σ, the following holds*

$$P[\Psi_n = \{A_1, \dots, A_n\}] = P[\Psi_n = \{\sigma(A_1), \dots, \sigma(A_k)\}].$$

Equivalently, we say that a random partition Ψ_n is exchangeable if there exists a symmetric function $p_k^{(n)} : \mathcal{C}_n \to [0,1]$ such that, for every partition $\{A_1, \dots, A_k\}$ of $\mathbb{N}_n$

$$P[\Psi_n = \{A_1, \dots, A_k\}] = p_k^{(n)}(|A_1|, \dots, |A_k|) = p_k^{(n)}(n_1, \dots, n_k),$$

The function $p_k^{(n)}$ is named exchangeable partition probability function (EPPF) of Ψ_n.

The notion of EPPF was first studied by Pitman (1995). According to the previous definition, when a random partition is exchangeable, the probability associated to a particular realization depends only on the number of blocks and their size.

$$\begin{tikzcd}
\mathbb{N}_4 \arrow[rr, "\sigma"] & & \mathbb{N}_4 \\
A_1 = \begin{bmatrix}\{1\}\\\{2\}\end{bmatrix} \arrow[rr] \arrow[drr] & & \begin{bmatrix}\{1\}\\\{4\}\end{bmatrix} = \sigma(A_1) \\
A_2 = [\{3\}] \arrow[drr] & & [\{2\}] = \sigma(A_2) \\
A_3 = [\{4\}] \arrow[uurr, dashed] & & [\{3\}] = \sigma(A_3)
\end{tikzcd}$$

$$P(\{1,2\},\{3\},\{4\}) = p_3^{(4)}(2,1,1) = P(\{1,4\},\{2\},\{3\})$$

FIGURE 1.1: An example of the effect of a permutation σ on an exchangeable partition of $\mathbb{N}_4$.

An EPPF satisfies the following properties:

i) $p_1^{(1)} = 1$;

ii) for any $(n_1, \dots, n_k) \in \triangle_n^k$, with $n \geq 1$ and $1 \leq k \leq n$,

$$p_k^{(n)}(n_1, \dots n_k) = p_k^{(n)}(n_{\sigma(1)}, \dots, n_{\sigma(k)}),$$

where $\sigma(\cdot)$ is a permutation of $(1, \dots, k)$;

iii) for any $(n_1, \dots, n_k) \in \triangle_n^k$, with $n \geq 1$ and $1 \leq k \leq n$, the following addition rule holds

$$p_k^{(n)}(n_1, \dots, n_k) = \sum_{j=1}^{k} p_k^{(n+1)}(n_1, \dots, n_j + 1, \dots, n_k) + p_{k+1}^{(n+1)}(n_1, \dots, n_k, 1). \tag{1.2}$$

1.3 Partial exchangeability

When homogeneity among the observations is not realistic, the assumption of exchangeability might be too strong. In many applications, some notion of homogeneity can be reasonably assumed within subsets of the sample but not across the whole set of observations. When this is the case, one may resort to the weaker notion of partial exchangeability. Let $X_1^{(\infty)}, \dots, X_L^{(\infty)}$ be L infinite sequences of $\mathbb{X}$-valued random elements, defined on $(\Omega, \mathcal{F}, \mathbb{P})$, where, for every $l = 1, \dots, L$, $X_l^{(\infty)} := \{X_{1,l}, X_{2,l}, \dots\}$. The sequence $(X_1, \dots X_L)^{(\infty)}$ is termed *partially exchangeable* if, for any $n_1, \dots, n_L \geq 1$ and any permutation σ_l of $\{1, \dots, n_l\}$, with $l = 1, \dots, L$, we have that

$$(X_{1,1}, \dots, X_{n_1,1}, X_{1,2}, \dots, X_{n_2,2}, \dots, X_{1,L}, \dots, X_{n_L,L})$$

coincides in distribution with

$$(X_{\sigma_1(1),1}, \dots, X_{\sigma_1(n_1),1}, X_{\sigma_2(1),2}, \dots, X_{\sigma_2(n_2),2}, \dots, X_{\sigma_L(1),L}, \dots, X_{\sigma_L(n_L),L}).$$

Given two or more sequences of observations, partial exchangeability is thus a probabilistic statement on the homogeneity of the observations within, but not across, groups.
The de Finetti representation theorem, presented in Theorem 1, can be extended to the partial exchangeability case. Let $\mathbb{X}^\infty = \mathbb{X} \times \mathbb{X} \times \dots$ and $\mathbb{M}_{\mathbb{X}}$ the space of probability measures on $(\mathbb{X}, \mathscr{X})$.

Theorem 2 (de Finetti, 1938)**.** *The sequence $(X_1, \dots, X_L)^{(\infty)}$ is partially exchangeable if and only if there exists a probability measure Q on the L-fold product space $\mathbb{M}_{\mathbb{X}}^L$, such that, for any $n_l \geq 1$, and $A_{\bullet,l} = A_{1,l} \times A_{2,l} \times \cdots \times A_{n_l,l} \times \mathbb{X}^{(\infty)}$, with $A_{i_l,l} \in \mathcal{X}$ and $l = 1, \dots, L$, one has*

$$\mathbb{P}\left[X_1^{(\infty)} \in A_{\bullet,1}, \dots\ , X_L^{(\infty)} \in A_{\bullet,L}\right] \\ = \int_{\mathbb{M}_{\mathbb{X}}^L} \prod_{l=1}^{L} \prod_{i=1}^{n_l} p_l(A_{i_l,l}) Q(\mathrm{d}p_1, \dots, \mathrm{d}p_L).$$

Also in the partially exchangeable case we have an equivalent hierarchical representation, given by

$$\begin{aligned}(X_{i_1,1}, \dots, X_{i_L,L}) \mid \tilde{p}_1, \dots, \tilde{p}_L &\overset{\text{iid}}{\sim} \tilde{p}_1, \times \cdots \times \tilde{p}_L, \qquad 1 \leq i_l, \leq n_l,\ \forall l = 1, \dots, L \\ \tilde{p}_1, \dots, \tilde{p}_L &\sim Q\end{aligned}$$

where $\tilde{p}_1, \dots, \tilde{p}_L$ are random probability measures and Q is a probability measure over the l-fold product space $\mathbb{M}_{\mathbb{X}}^L$.

1.4 Distributions on the space of probability measures

A discrete random probability measure (RPM) on $(\mathbb{X}, \mathscr{X})$ can be defined as

$$\tilde{p} = \sum_{j=1}^{\infty} W_j \delta_{\tilde{X}_j}, \tag{1.3}$$

where $\{\tilde{X}_j\}_{j\geq 1}$ is a sequence of random atoms, i.i.d. from a diffuse probability distribution P_0 on $(\mathbb{X}, \mathscr{X})$, and $\{W_{\tilde{X}_1}, W_{\tilde{X}_2}, \dots\}$ is a sequence of random variables taking values in the infinite dimensional unit simplex

$$\triangle = \triangle_1^\infty = \left\{ \mathbf{w} = (w_1, w_2, \dots) : \forall j \in \mathbb{N}, w_j \geq 0, \sum_{j=1}^{\infty} w_j = 1 \right\}.$$

In this manuscript we will focus on *homogeneous* discrete RPMs, that is RPMs for which the distribution of the atoms $\{\tilde{X}_j\}_{j\geq 1}$ is independent of the distribution of the random jumps $\{W_j\}_{j\geq 1}$. A random probability measure defined as in Equation 1.3, induces a probability distribution on the space $(\mathbb{M}_{\mathbb{X}}, \mathscr{M}_{\mathbb{X}})$. In the next section we introduce a convenient constructive definition of the weights $\{W_j\}_{j\geq 1}$, named *stick-breaking* construction, that leads to the definition of a flexible class of discrete random probability measures. For more details on this or alternative definitions of discrete RPMs, one can refer, e.g., to Ghosal and Van Der Vaart (2017).

1.4.1 Definition through Stick-Breaking

The stick-breaking construction for defining the random weights $\{W_j\}_{j\geq 1}$ in (1.3) was first introduced, for the Dirichlet process (see Section 1.5), by Sethuraman (1994), and later extended so to define a much more general class of processes. To this end, a crucial contribution was provided by Ishwaran and James (2001).

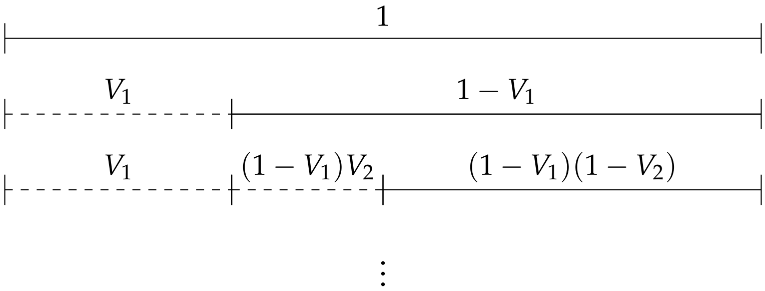

FIGURE 1.2: Graphical representation of the stick-breaking procedure.

The stick-breaking definition of the the weights $\{W_j\}_{j\geq 1}$ in (1.3) is nicely described by the following metaphor. A stick of unit length is broken, at a random point, into two bits, namely V_1 and $1 - V_1$. The weight W_1 is set equal to V_1, while the remaining part of length $1 - V_1$ is further split into two parts, of length $V_2(1 - V_1)$ and $(1 - V_2)(1 - V_1)$. The former is used to define W_2, the latter will be broken again so to the difne W_3 and so on (see Figure 1.2). In general, if $\{V_j\}_{j\geq 1}$ is a sequence of random variables such that $0 \leq V_i \leq 1$, then, for any $j \geq 1$ we can define the stick weight W_j as

$$W_j = V_j \prod_{l=1}^{j-1} (1 - V_l)$$

The following Lemma formalizes necessary and sufficient conditions that the sequence of random weights $\{V_j\}_{j\geq 1}$ must satisfy for the resulting sequence $\{W_j\}_{j\geq 1}$ to lie in $\triangle$ almost surely.

Lemma 1 (Ishwaran and James, 2001). *The sequence $\{W_j\}_{j\geq 1}$, defined by means the sequence $\{V_j\}_{j\geq 1}$, lies in $\triangle$ almost surely if and only if*

$$\mathbb{E}\left[\prod_{l=1}^{j} (1 - V_l)\right] \xrightarrow[j\to\infty]{} 0.$$

When the random variables in the sequence $\{V_j\}_{j\geq 1}$ are independent, then the previous condition is equivalent to

$$\sum_{l=1}^{\infty} \log \mathbb{E}[1 - V_l] = -\infty.$$

Moreover when $\{V_j\}_{j\geq 1}$ are i.i.d. random variables, it is sufficient that $\Pr(V_1 > 0) > 0$.

If the sequence $\{V_j\}_{j\geq 1}$ satisfies the conditions of Lemma 1, then corresponding sequence of weights $\{W_j\}_{j\geq 1}$ can be used to define a homogeneous discrete RPM on $(\mathbb{X}, \mathscr{X})$ as

$$\tilde{p} = \sum_{i=1}^{\infty} W_j \delta_{\tilde{X}_j}, \qquad W_j = V_j \prod_{l=1}^{j-1} (1 - V_l),$$

where $\{\tilde{X}_j\}_{j\geq 1}$ is a sequence of random atoms taking values on $\mathbb{X}$, and independent of $\{V_j\}_{j\geq 1}$.

In the seminal paper of Sethuraman (1994), the author considered the case $V_j \sim \mathrm{B}(1, \alpha(\mathbb{X}))$, with $\alpha(\cdot)$ is a measure over $(\mathbb{X}, \mathscr{X})$ and $\mathrm{B}(\cdot, \cdot)$ is the Beta distribution. Several extensions were proposed after his work (see, e.g., Ishwaran and James, 2001; Dunson and Park, 2008; Favaro et al., 2012a; Favaro et al., 2014; Favaro et al., 2016).

1.4.2 Predictive distribution

Let $\mathbf{X}^{(n)} = (X_1, \dots, X_n)$ be an exchangeable sample from $\tilde{p}$. Let $\tilde{p} \sim Q$, with Q de Finetti measure in (1.1) and $Q(\cdot \mid \mathbf{X}^{(n)})$ its posterior distribution. The distribution of the next observation, X_{n+1}, conditionally to the observed sample $X_1, \dots, X_n$, is named *predictive distribution* and it can be defined by the posterior expectation of $\tilde{p}$, as

$$P[X_{n+1} \in \mathrm{d}t \mid \mathbf{X}^{(n)}] = \int_{\mathbb{M}_{\mathbb{X}}} \tilde{p}(\mathrm{d}t) Q(\mathrm{d}\tilde{p} \mid \mathbf{X}^{(n)}), \tag{1.4}$$

where $\mathbb{M}_{\mathbb{X}}$ is the space of the probability measures with support $\mathbb{X}$. Due to the discreteness of $\tilde{p}$, with positive probability there will be ties in $\mathbf{X}^{(n)}$, thus inducing a partition of the sample. The predictive distribution in Equation 1.4 can then be represented in terms of EPPF of the possible partitions of $\mathbf{X}^{(n+1)}$, given the partition observed for $\mathbf{X}^{(n)}$. Let $X_1^*, \dots, X_k^*$ be the unique values in $\mathbf{X}^{(n)}$. The predictive distribution can be written as

$$P[X_{n+1} \in \mathrm{d}t | \mathbf{X}^{(n)}] \propto p_{k+1}^{(n+1)}(n_1, \dots, n_k, 1) P_0(\mathrm{d}t) + \sum_{j=1}^{k} p_k^{(n+1)}(n_1, \dots, n_j + 1, \dots, n_k) \delta_{X_j^*}(\mathrm{d}t), \tag{1.5}$$

where we recall that P_0 is a diffuse probability distribution on $(\mathbb{X}, \mathscr{X})$. An EPPF satisfies the addition rule in (1.2), then it is possible to normalize Equation 1.5, so to get the exact predictive distribution as

$$P[X_{n+1} \in \mathrm{d}t | \mathbf{X}^{(n)}] = \frac{p_{k+1}^{(n+1)}(n_1, \dots, n_k, 1)}{p_k^{(n)}(n_1, \dots, n_k)} P_0(\mathrm{d}t) + \sum_{j=1}^{k} \frac{p_k^{(n+1)}(n_1, \dots, n_j + 1, \dots, n_k)}{p_k^{(n)}(n_1, \dots, n_k)} \delta_{X_j^*}(\mathrm{d}t). \tag{1.6}$$

Blackwell and MacQueen (1973) first studied the predictive distribution for the Dirichlet process, introduced in the same year by Ferguson (1973). The authors proposed an intuitive interpretation in terms of Pólya urns.

1.5 Dirichlet process

While the nonparametric framework was already laid out in the 30's by Bruno de Finetti, it was only with the introduction of the Dirichlet process (DP) in the 70's, that the nonparametric approach could actually be implemented. The DP was introduced in the seminal paper of Ferguson (1973), and, due to its remarkable mathematical tractability, it still represents the default choice in many applications of Bayesian nonparametric models.

Definition 3 (Ferguson, 1973)**.** *A random measure $\tilde{p}$ on $(\mathbb{X}, \mathscr{X})$ is said to be a Dirichlet Process with base measure λ, DP(λ), if for every finite measurable partition $A_1, \dots, A_k$ of $\mathbb{X}$ we have*

$$(\tilde{p}(A_1), \dots, \tilde{p}(A_k)) \sim Dir(\lambda(A_1), \dots, \lambda(A_k)).$$

The process is equivalently parametrized by the total mass (also named prior precision or precision parameter) $\vartheta = \lambda(\mathbb{X})$ and $P_0 = \lambda/\lambda(\mathbb{X})$, the probability measure obtained by the normalization of λ (called center measure or base measure). We refer to the (ϑ, P_0) parametrization and denote it by DP(ϑ, P_0).

Definition through stick-breaking

The stick-breaking definition for the DP was introduced in the literature by Sethuraman (1994). In order to introduce it, we first define the one parameter *GEM* [1] distribution, as follows.

Definition 4. *Let $V_1, V_2, \dots \overset{\text{iid}}{\sim} Beta(1, \vartheta)$, with $\vartheta \in \mathbb{R}^+$, and define the weights $W_1 = V_1$, and for any $j \geq 1$, $W_j = V_j \prod_{k=1}^{j-1}(1 - W_k)$. The sequence $\{W_j\}_{j \geq 1}$ is said to follow a one-parameter GEM distribution, denoted by $\{W_j\}_{j \geq 1} \sim GEM(\vartheta)$.*

Note that for the sequence $\{W_j\}_{j\geq 1}$ the conditions of Lemma 1 hold. Let $\{W_j\}_{j\geq1} \sim GEM(\vartheta)$, $\tilde{X}_1, \tilde{X}_2, \dots \overset{\text{iid}}{\sim} P_0$, and assume that the sequences $\{W_j\}_{j\geq1}$ and $\{\tilde{X}_j\}_{j\geq1}$ are independent. Then,

$$\tilde{p} = \sum_{j=1}^{\infty} W_j \delta_{\tilde{X}_j},$$

is distributed as a DP with precision parameter ϑ and base measure P_0, that is $\tilde{p} \sim DP(\vartheta, P_0)$.

Predictive distribution

Blackwell and MacQueen (1973) first studied the predictive structure of the DP. Here we derive the same, by starting from the general representation (1.6), given in terms of the EPPF. The EPPF for a DP with precision parameter ϑ, is defined as follows

$$p_k^{(n)}(n_1, \dots, n_k) = \frac{\vartheta^k}{(\vartheta)_n} \prod_{j=1}^{k} (n_j - 1)!,$$

where $(a)_k := a(a+1)\cdots(a+k-1)$ is the rising factorial. Then, following Equation 1.6, we have that, given $\mathbf{X}^{(n)}$, the conditional probability for X_{n+1} to take a new value, is given by

$$\frac{p_{k+1}^{(n+1)}(n_1, \dots, n_k, 1)}{p_k^{(n)}(n_1, \dots, n_k)} = \frac{\dfrac{\vartheta^k}{(\vartheta)_{n+1}} \displaystyle\prod_{j=1}^{k+1}(n_j - 1)!}{\dfrac{\vartheta^k}{(\vartheta)_n} \displaystyle\prod_{j=1}^{k}(n_j - 1)!} = \frac{\vartheta}{\vartheta + n},$$

while the conditional probability that X_{n+1} coincides with X_j^*, value appeared n_j times in $\mathbf{X}^{(n)}$, is given by

[1] The GEM distribution is named after the studies of Griffiths, Engen and McClosky.

$$\frac{p_k^{(n+1)}(n_1,\dots,n_j+1,\dots,n_k)}{p_k^{(n)}(n_1,\dots,n_k)} = \frac{\dfrac{\vartheta^k}{(\vartheta)_{n+1}} n_j \displaystyle\prod_{j=1}^{k}(n_j-1)!}{\dfrac{\vartheta^k}{(\vartheta)_n}\displaystyle\prod_{j=1}^{k}(n_j-1)!} = \frac{n_j}{\vartheta+n}.$$

We can then write the predictive distribution for the DP as

$$P[X_{n+1}\in dt|X_1,\dots,X_n] = \frac{\vartheta}{\vartheta+n}P_0(dt) + \sum_{j=1}^{k}\frac{n_j}{\vartheta+n}\delta_{X_j^*}(dt).$$

1.5.1 Other properties

In this section we summarise some fundamental properties of the DP, as they will be useful in the rest of the manuscript.

Gamma CRM normalization

Let γ be a gamma completely random measure (CRM, see Appendix A for details), defined on $(\mathbb{X},\mathscr{X})$ and with Lévy intensity $\nu(\mathrm{d}s,\mathrm{d}x) = \exp\{-s\}/s\mathrm{d}s\,\vartheta P_0(\mathrm{d}x)$. The process γ has independent increments, that is, if $A_1,\dots A_k$ are measurable disjoint subsets of $\mathbb{X}$, then the random variables $\gamma(A_j)$'s are independent gammas, that is $\gamma(A_j)\overset{\text{ind}}{\sim} G(\vartheta P_0(A_j),1)$, with $j=1,\dots,k$. An alternative definition for the Dirichlet process is obtained by normalizing γ, thus defining a RPM on $(\mathbb{X},\mathscr{X})$ as

$$\tilde{p}(\cdot) = \frac{\gamma(\cdot)}{\gamma(\mathbb{X})}.$$

The definition of DP as the normalization of a gamma process was already given in Ferguson (1973). The connection between DP and gamma process is convenient as many known properties of the gamma process can be exploited to study the DP.

Expectation and Variance of the DP

In the next result we can derive the first two moments of the DP. For more details on the topic see, for example, Ghosal and Van Der Vaart (2017).

Proposition 1. *Let $\tilde{p}\sim PD(\vartheta,P_0)$, defined on $\mathbb{X}$, and $A\in\mathscr{X}$. Then*

$$\mathbb{E}[\tilde{p}(A)] = P_0(A)$$
$$Var[\tilde{p}(A)] = \frac{P_0(A)P_0(A^c)}{\vartheta+1}$$

Proof. By Definition 3, we have that $(\tilde{p}(A_1),\dots,\tilde{p}(A_k))\sim Dir(\vartheta P_0(A_1),\dots,\vartheta P_0(A_k))$, for any measurable partition $A_1,\dots,A_K$ of $\mathbb{X}$. This implies that $\tilde{p}(A)$ is a Beta random variable, specifically $\tilde{p}(A)\sim \mathrm{B}(\vartheta P_0(A),\vartheta P_0(A^c))$. It easily follows that

$$\mathbb{E}[\tilde{p}(A)] = \frac{\vartheta P_0(A)}{\vartheta P_0(A)+\vartheta P_0(A^c)} = P_0(A)$$

and
$$\mathrm{Var}[\tilde{p}(A)] = \frac{(\vartheta P_0(A))(\vartheta P_0(A^c))}{(\vartheta P_0(A) + \vartheta P_0(A^c))^2(\vartheta P_0(A) + \vartheta P_0(A^c) + 1)} = \frac{P_0(A)P_0(A^c)}{\vartheta + 1}$$

□

The base measure P_0, expected value of the DP, can then be interpreted as the prior guess in a Bayesian model. The last proposition also shows that, for any measurable A, the precision of $\tilde{p}(A)$ is a linear function of ϑ, thus motivating the name precision parameter.

Conjugacy

A peculiar property of the DP is its conjugacy (Ferguson, 1973). See also James et al. (2006).

Theorem 3 (Ferguson, 1973)**.** *Let $\tilde{p} \sim DP(\vartheta, P_0)$ and $X_1, \dots, X_n$ sampled independently from $\tilde{p}$. Then $\tilde{p}|X_1, \dots, X_n$ is a Dirichlet process characterized by total mass $\vartheta + n$ and base measure*

$$P_n = \frac{\vartheta}{\vartheta + n} P_0 + \frac{\vartheta}{\vartheta + n} \sum_{j=1}^{k} \frac{n_j}{n} \delta_{X_j^*},$$

where $X_1^*, \dots, X_k^*$ are the unique values of $X_1, \dots, X_n$, and $n_1, \dots, n_k$ the corresponding frequencies.

The previous result implies that, for any set $A \in \mathscr{X}$, the posterior mean of $\tilde{p}(A)$, given $X_1, \dots, X_n$, is given by

$$\mathbb{E}[\tilde{p}(A)|X_1, \dots, X_n] = \frac{\vartheta}{\vartheta + n} P_0(A) + \frac{\vartheta}{\vartheta + n} \sum_{j=1}^{k} \frac{n_j}{n} \delta_{X_j^*}(A),$$

a linear combination of the prior guess P_0 and the empirical distribution, both evaluated at A. Moreover, the variance goes to zero when the sample size n grows, indeed

$$\begin{aligned} \mathrm{Var}[\tilde{p}(A)|X_1, \dots, X_n] &= \frac{\mathbb{E}[\tilde{p}(A)|X_1, \dots, X_n]\mathbb{E}[\tilde{p}(A^c)|X_1, \dots, X_n]}{\vartheta + n + 1} \\ &\leq \frac{1}{4(\vartheta + n + 1)} = o(1). \end{aligned}$$

Marginal distribution

In the paper of Ferguson (1973) the author characterized the marginal distribution of a single value $X \mid \tilde{p} \sim \tilde{p}$, where $\tilde{p}$ is a DP, which give a simple strategy to sample a new atom X from $\tilde{p}$.

Proposition 2 (Ferguson, 1973). *Let $\tilde{p} \sim DP(\vartheta, P_0)$ and $X|\tilde{p} \sim \tilde{p}$. Then the marginal distribution of X corresponds to the base measure, $X \sim P_0$.*

The previous results is immediate by, given a set $A \in \mathscr{X}$,

$$\tilde{p}(A) = \mathbb{E}[\tilde{p}(A)|\tilde{p}] = \mathbb{E}[\tilde{p}(A)] = P_0(A).$$

As important consequence of the previous Proposition, we can sample a value from the marginal distribution of a generic X by sampling a value from the base measure.

Self-similarity

Another remarkable property of the DP is its self-similarity. The intuition is that the restriction of the DP to a subset of his support is still a DP, and, in addition, it is independent of the same process outside the restricted support.
Let $B \subseteq \mathbb{X}$ a measurable set and $\tilde{p}|_B$ denote the restriction of the random probability measure $\tilde{p}$ to B, that is $\tilde{p}|_B(\cdot) = \tilde{p}(\cdot \cap B)$.

Theorem 4. *Let $\tilde{p} \sim DP(\vartheta, P_0)$, $B \subseteq \mathbb{X}$ and $B^c = \mathbb{X} \setminus B$ be measurable sets. Then $\tilde{p}(B)$, $\tilde{p}(B^c)$, $\tilde{p}|_B$ and $\tilde{p}|_{B^c}$ are mutually independent. Moreover, $\tilde{p}|_B$ and $\tilde{p}|_{B^c}$ are still Dirichlet processes, with mass $\vartheta P_0(B)$ and $\vartheta P_0(B^c)$, respectively, and base measure $P_0|_B$ and $P_0|_{B^c}$, respectively.*

The proof exploits the representation of the DP as the normalization of a gamma process, and well-known properties of gamma random variables. For more details, one can refer to Ghosal and Van Der Vaart (2017).

Dirichlet random means

Let $\tilde{p}$ be a DP parametrized by the measure λ, that is $\tilde{p} \sim DP(\lambda)$. Let h be a real-valued measurable function defined on $\mathbb{X}$ such that

$$\int_{\mathbb{X}} \log(1 + |h(t)|)\lambda(dt) < \infty. \tag{1.7}$$

Let $\lambda_h = \lambda \circ h^{-1}$. Lijoi and Prünster (2009) showed that it is possible to characterize the random mean of h, with respect to a Dirichlet process distribution, as follows.

Proposition 3. *Let $\tilde{p} \sim DP(\lambda)$ and h be a real-valued measurable function, for which (1.7) holds. Then*

$$\int_{\mathbb{X}} h(t)\tilde{p}(dt) \overset{d}{=} \int_{\mathbb{X}} t\tilde{p}_h(dt),$$

where $\tilde{p}_h \sim DP(\lambda_h)$.

1.6 Pitman-Yor process

The Pitman-Yor (PY) process[2] is probably the most popular generalization of the DP process. This extension was first introduced by Perman et al. (1992), and later investigated by Pitman and Yor (1997) and Pitman (1995), among others. For an exhaustive presentation from a probabilistic perspective see Pitman (2006).
The distribution of a PY is characterized by a discount parameter $\sigma \in [0, 1)$, a strength parameter $\vartheta > -\sigma$, and a diffuse base measure P_0.
We introduce the PY process by generalizing the stick-breaking definition of the DP. To this end, we extend the definition of the GEM distribution (see Definition 4), to the two-parameter case.

Definition 5. *Let $\sigma \in [0, 1)$, $\vartheta > -\sigma$ and $V_1, V_2, \ldots$ be independent random variables such that for all $j = 1, 2, \ldots$, $V_j \overset{\text{ind}}{\sim} Beta(1 - \sigma, \vartheta + j\sigma)$. Define the weights $W_1 = V_1$, and, for $j \geq 2$, $W_j = V_j \prod_{k=1}^{j-1}(1 - W_k)$. The sequence $\{W_j\}_{j \geq 1}$ is said to follow a two-parameter GEM distribution, denoted by $\{W_j\}_{j \geq 1} \sim GEM(\vartheta, \sigma)$.*

[2] In the paper of Perman et al. (1992) the authors refer to the process as the two-parameter Poisson-Dirichlet process, the name Pitman-Yor process was introduced by Ishwaran and James (2001).

We observe that conditions of Lemma 1 hold for a sequence $\{W_j\}_{j\geq 1} \sim GEM(\vartheta,\sigma)$, and thus we use to define the sequence of random jumps of the RPM in Equation 1.3. Starting from the two-parameters GEM distribution, the PY can then be introduced as follows.

Definition 6. *Let $\sigma \in [0,1)$, $\theta > -\sigma$ and P_0 be a diffuse measure on $(\mathbb{X},\mathscr{X})$. Let $\{W_j\}_{j\geq 1} \sim GEM(\vartheta,\sigma)$ and $\tilde{X}_1, \tilde{X}_2, \ldots \overset{\text{iid}}{\sim} P_0$. Then the random probability measure*

$$\tilde{p} = \sum_{j=1}^{\infty} W_j \delta_{\tilde{X}_j}$$

is distributed as a Pitman-Yor process with parameters θ and σ and base measure P_0, $\tilde{p} \sim PY(\vartheta,\sigma,P_0)$.

The DP is recovered as a special case by setting the discount parameter σ equal to zero.

Predictive distribution

The EPPF of a PY with parameters σ and ϑ is defined as

$$p_k^{(n)}(n_1,\ldots,n_k) = \frac{\prod_{i=1}^{k-1}(\vartheta+i\sigma)}{(\vartheta+1)_{n-1}} \prod_{j=1}^{k}(1-\sigma)_{n_j-1},$$

where $(a)_k$ is the rising factorial. Then, following Equation 1.6, we have that, given $\mathbf{X}^{(n)}$, the conditional probability for X_{n+1} to take a new value, is given by

$$\frac{p_{k+1}^{(n+1)}(n_1,\ldots,n_k,1)}{p_k^{(n)}(n_1,\ldots,n_k)} = \frac{\dfrac{\prod_{i=1}^{k}(\vartheta+i\sigma)}{(\vartheta+1)_n}\displaystyle\prod_{j=1}^{k}(1-\sigma)_{n_j-1}}{\dfrac{\prod_{i=1}^{k-1}(\vartheta+i\sigma)}{(\vartheta+1)_{n-1}}\displaystyle\prod_{j=1}^{k}(1-\sigma)_{n_j-1}} = \frac{\vartheta+k\sigma}{\vartheta+n},$$

while the conditional probability that X_{n+1} coincides with X_j^*, value appeared n_j times in $\mathbf{X}^{(n)}$, is given by

$$\frac{p_k^{(n+1)}(n_1,\ldots,n_j+1,\ldots,n_k)}{p_k^{(n)}(n_1,\ldots,n_k,1)} = \frac{\dfrac{\prod_{i=1}^{k-1}(\vartheta+i\sigma)}{(\vartheta+1)_n}(1-\sigma)_{n_j}\displaystyle\prod_{l\neq j}^{k}(1-\sigma)_{n_l-1}}{\dfrac{\prod_{i=1}^{k-1}(\vartheta+i\sigma)}{(\vartheta+1)_{n-1}}\displaystyle\prod_{l=1}^{k}(1-\sigma)_{n_l-1}} = \frac{n_j-\sigma}{\vartheta+n}.$$

We can the write the predictive distribution for the PY process as

$$P[X_{n+1}\in dt|X_1,\ldots,X_n] = \frac{\vartheta+k\sigma}{\vartheta+n}P_0(dt) + \sum_{j=1}^{k}\frac{n_j-\sigma}{\vartheta+n}\delta_{X_j^*}(dt).$$

Posterior representation

Pitman (1996) provides a convenient representation of the posterior distribution of a PY.

Theorem 5. *(Corollary 20 in Pitman, 1996). Let $\mathbf{X}^{(n)}$ be a sample from $\tilde{p}$, with $\tilde{p} \sim PY(\vartheta,\sigma,P_0)$. Let $X_1^*,\ldots,X_k^*$ be the distinct values in $\mathbf{X}^{(n)}$, and $n_1,\ldots,n_k$ the corresponding frequencies. Then, conditionally on $\mathbf{X}^{(n)}$,*

$$\tilde{p} = \sum_{j=1}^{k} W_j \delta_{X_j^*} + W_{k+1}\tilde{p}_k$$

where $(W_1, \dots, W_k, W_{k+1}) \sim Dir(n_1 - \sigma, \dots, n_k - \sigma, \vartheta + k\sigma)$ *and* $\tilde{p}_k \sim PY(\vartheta + k\sigma, \sigma, P_0)$*, and* $\tilde{p}_k$ *is independent of* $(W_1, \dots, W_{k+1})$*.*

The previous result will be exploited in Chapter 3. A detailed proof of Theorem 5 can be found in the PhD dissertation of Carlton (1999).

1.7 Other generalizations of the DP

While the PY is arguably the most popular generalization of the DP, the problem of defining flexible classes of RPMs generalizing the DP, has attracted considerable attention in the last decade.

Extensions based on CRM

The study of CRMs (see Appendix A) and their normalization has been the focus of interesting contributions in recent Bayesian nonparametric literature. We have seen that a DP can be defined as the normalization of a gamma process (Ferguson, 1973). The same approach can be adopted more in general: a random probability measure can be defined by normalizing a CRM (Regazzini et al., 2003). Under the additional assumption that the jumps of the CRM are independent of their locations, we obtain RPMs named homogeneous normalized random measure with independent increments (hNRMI) (see Regazzini et al., 2003; James et al., 2009).
The use of hNRMs in Bayesian nonparametrics has grown considerably in the last decade. The more commonly adopted processes within this family are the normalized generalized gamma process (NGG) (Pitman, 2003; Lijoi et al., 2007c, see also Prünster, 2002, James, 2002, Lijoi and Prünster, 2003, Regazzini et al., 2003). The NGG includes, as special cases, the DP (Ferguson, 1973), the normalized σ-stable process (Kingman, 1975) and the normalized inverse Gaussian process (Lijoi et al., 2005b).

Extensions based on stick-breaking

In one of the main references for the definition of stick-breaking RPMs, Ishwaran and James (2001) were clear about the potential generality of this construction. In their paper, they introduced the stick-breaking representation as

$$W_1 = V_1, \qquad W_j = V_j \prod_{k=1}^{j-1} (1 - V_j)$$

where the random variables $V_j \sim Beta(a_j, b_j)$ are characterized by two sequences of parameters $\{a_j\}_{j\geq 1}$ and $\{b_j\}_{j\geq 1}$. With particular setting of the sequences $\{a_j\}_{j\geq 1}$ and $\{b_j\}_{j\geq 1}$, it is possible to reconstruct different families of process. We already discussed in the previous sections the Dirichlet process (Ferguson, 1973) and the Pitman-Yor process (Pitman and Yor, 1997), but the specification of $\{a_j\}_{j\geq 1}$ and $\{b_j\}_{j\geq 1}$ is known also for the Dirichlet-multinomial process (Muliere and Secchi, 1995), the m-spike model (Liu, 1996), the finite dimensional Dirichlet priors (Ishwaran and Zarepour, 2002) and the Beta two-parameter process (Ishwaran and Zarepour, 2000). The previous list reports some examples of processes that have a stick-breaking representation, but quoting Ishwaran and James (2001) *"more measures will eventually be recognized as being stick-breaking in nature"*, also without the requirement of Beta distributed stick breaks.

Extensions based on EPPFs

A key tool for understanding combinatorial properties of a RPM is represented by the EPPF, introduced in Section 1.2. We present here a classification of RPMs based on the form taken by their EPPF. This insightful result was proved in De Blasi et al. (2015) and will be useful in Chapter 4.
In a general framework the probability of getting a new value in a sampling sequence at the $(n+1)$-th step is a function of the number of already sampled values n, the number of distinct values k and their frequencies $n_1, \dots, n_k$, i.e.

$$\Pr[X_{n+1} = \text{"}new\text{"} \mid X_1, \dots, X_n] = q(n, k, n_1, \dots, n_k).$$

The previous relation comes from the fact that an EPPF is a function of the type $p_k^{(n)}(n_1, \dots, n_k) = g(n, k, n_1, \dots, n_k)$. Then it is possible to classify RPMs by their predictive distribution, as follows.

Proposition 4 (De Blasi et al., 2015)**.** *Let $\tilde{p}$ a random measure, with related EPPF $p_k^{(n)}(n_1, \dots, n_k)$. The following classification holds:*

(i) $q(n, k, n_1, \dots, n_k) = q(n)$ *if and only if $\tilde{p}$ is a Dirichlet process;*

(ii) $q(n, k, n_1, \dots, n_k) = q(n, k)$ *if and only if $\tilde{p}$ is a Gibb-type process;*

(iii) $q(n, k, n_1, \dots, n_k) = q(n, k, n_1, \dots, n_k)$ *otherwise.*

The family of Gibb-type priors has interesting properties as it generalizes the DP while mantaining a good degree of mathematical tractability: this is reflected by the fact that the probability of discovering a new value in the predictive distribution depends on the observed sample only through its sample size n and the number of distinct values k.

1.8 Definition of nonparametric priors

Due to their almost sure discreteness, the RPMs that we have considered in this chapter cannot be used as prior processes to model directly continuous fenomena.
Consider a homogeneous RPM defined as in Section 1.4. A first use of these processes consists in modelling phenomena with a discrete support. Let $\mathbf{X}^{(n)}$ be an exchangeable sample from $\tilde{p}$, where $\tilde{p}$ is a homogenous RPM. In the context of Bayesian nonparametric statistics, we refer to

$$\tilde{p} = \sum_{j=1}^{\infty} W_j \delta_{\tilde{X}_j}$$

as a *species sampling model*, name first introduced by Pitman (1995).
If instead we want to model continuous observations, we can follow the approach introduced by Lo (1984) for the DP, and use the discrete RPM as a mixing measure in a mixture model, thus obtaining an infinite mixture. That is, we consider a kernel function $k(x, \theta)$ on $\mathbb{X} \times \Theta$ and define the random density

$$\tilde{f}(x) = \int_{\Theta} k(x, \theta) \tilde{p}(\mathrm{d}\theta).$$

The random density $\tilde{f}$ lives in $\mathcal{F}$, the space of density functions on $\mathbb{X}$.
Finally, observe that while the use of a continuous kernel $k(x, \theta)$ serves the purpose of allowing discrete RPMs to be used for continuous fenomena, it is anyway possible to consider infinite

mixtures of discrete kernels, see for example Krnjajić et al. (2008), where a Poisson kernel is adopted.

1.9 Outline and main contributions

In this thesis we investigate three important aspects of nonparametric mixtures, about their modelling, computational and distributional properties. The chapters' order relies to the generality of the processes considered: in the next chapter, Chapter 2, we expound modelling properties of the DP. Chapter 3 describes computational aspects of the PY process. Finally Chapter 4 approach the Gibbs-type prior family, with a study on distributional properties of the size of the clusters underlying the data.

In details, Chapter 2 is a thorough study regarding the effect of invertible affine transformation of the data on the posterior distribution of Dirichlet process mixture models, with particular attention to DP mixtures of multivariate Gaussian kernels (DPM-G). We first provide an explicit specification of the parameters of a DPM-G model, based on the transformation considered. With the use of this particular specification, given the sampled data or the transformed data, the posterior distributions and the inference made with DPM-G models are equivalent. Then, after the formalization of asymptotic robustness of a model under affine transformation of the data, we prove a result which show that, under mild assumptions on the data-generating distribution, DPM-G models are asymptotically robust. The results are supported by an exhaustive simulation study on the effect of affine transformations.

Chapter 3 describes the importance conditional sampler, a novel sampling strategy for PY mixture models. The proposed method combines the best features of the conditional and marginal sampling strategies for PY mixture models, by the use of an useful representation of the posterior distribution of a PY (Pitman, 1996) and an importance sampling idea, similar in spirit to the augmentation step of the celebrated Algorithm 8 of Neal (2000). After an introduction to the modelling framework and the state-of-the-art of the main competitors, we expound the importance conditional sampler, with a detailed description of the steps for the implementation. We perform an exhaustive simulation study, where we compare our proposal with its main competitors. Finally the importance conditional sampler is used as a building block for devising a new efficient algorithm for the class of GM-dependent DP mixture models (Lijoi et al., 2014a; Lijoi et al., 2014b), for partially exchangeable data.

The last chapter, Chapter 4, propose an elicitation strategy for the parameters of a Gibbs-type prior, based on some distributional properties. We first introduce the Gibb-type priors and some introductory notions. Then we show the main contribution of the chapter, focuses on an exchangeable sample from a Gibbs-type prior and provides a simple and convenient description of the the size of the $(m+1)$th observation's cluster, given an unobserved sample of size m. We propose some results to characterize the cluster size distribution, based on the main result. We perform a study on the cluster size distribution for particular families of process, belong to the Gibbs-type priors: the DP, the PY process and the NGG. The study of such distribution provides the tools for a simple strategy for prior elicitation of the parameters for a Gibbs-type prior, in the context of Gibbs-type mixture models.

Chapter 2

Dirichlet process mixtures and affine transformation

> A process cannot be understood by stopping it. Understanding must move with the flow of the process, must join it and flow with it.
>
> Frank Herbert, *Author*

Based on:

Arbel, J., Corradin, R., Nipoti, B.
"Dirichlet process mixtures under affine transformations of the data"
Submitted (2018)

Arbel, J., Corradin, R., Lewandowski, M.
Discussion of "Bayesian Cluster Analysis: Point Estimation and Credible Balls"
Bayesian Analysis (2018)

A natural requirement for statistical methods for density estimation and clustering is for them to be robust under affine transformations of the data. Such a desideratum is exacerbated in multivariate problems where data components are incommensurable, that is not measured in the same physical unit, and for which, thus, the definition of a metric on the sample space requires the specification of constants relating units along different axes. As an illustrative example, consider astronomical data consisting of position and velocity of stars, thus living in the so-called phase-space: a metric on such a space can be defined by setting a dimensional constant to relate positions and velocities. In this setting, any sensible statistical procedure should be robust with respect to the specification of such a constant (Ascasibar and Binney, 2005; Maciejewski et al., 2009). This is specially important considering that often scarce to no a priori guidance about dimensional constants might be available, thus making the model calibration a daunting task. In this chapter we study how affine transformations of the data affect Bayesian posterior inference carried out based on Dirichlet process (DP) mixture models. While several kernels have been considered in the literature, including e.g. skew-normal (Canale and Scarpa, 2016), Weibull (Kottas, 2006), Poisson (Krnjajić et al., 2008), in this chapter we focus on the convenient and commonly adopted Gaussian specification first introduced by Lo (1984), and later extended by Müller et al. (1996) to the case of multivariate Gaussian kernel (in this chapter we let DPM-G denote a Dirichlet process mixture model with multivariate gaussian kernel). Although its properties have been thoroughly studied (see, e.g., Hjort et al., 2010), little attention has been dedicated to its robustness under data transformations (see Arbel and Nipoti, 2013). To the best of our knowledge, only Bean et al. (2016) and Shi et al. (2018) study the effect of data transformation under a DPM model. The goal of Bean et al. (2016) is to transform the sample so to facilitate the estimation of univariate densities on a new scale and thus to improve the performance of the methodology; Shi et al. (2018), instead, study the consistency properties of DPM models under affine data transformation, when studying the properties of the so-called low information omnibus prior for DPM models they introduce.

The class of DPM models is a very commonly used model framework, whose asymptotic properties have been studied by Wu and Ghosal (2010), Shen et al. (2013) and Canale and De Blasi (2017), among others. Our contribution to this area of research, presented in this chapter, focuses on the effect of applying an affine transformation to the data, in the context of DPM models, in terms of density estimation and inference. In Section 2.1 we describe the modelling framework and introduce the notation used throughout the chapter. In Section 2.2 we formalise the intuitive idea that a DPM-G model on a given dataset induces a DPM-G model on rescaled data and we provide the parameters mapping for the transformed DPM-G model. Section 2.3 is dedicated to a brief review of some relevant results on posterior consistency for DPM models. In Section 2.4 we introduce the notion of asymptotic robustness under affine transformations of the data and show that, under mild assumptions on the true data generating process, DPM-G models feature such robustness property. In Section 2.5 we discuss some recently introduced methodologies to get an estimated partition, in the context of clustering analysis, based on the output of a Markov chain Monte Carlo algorithm. The theoretical results presented in this chapter are supported by a simulation study, described in Section 2.6. Finally, in Section 2.7, we illustrate our findings by analysing an astronomical dataset and addressing a nontrivial classification problem arising in the astronomical study of globular clusters.

2.1 Modelling framework

Let $\mathbf{X}^{(n)} := (\mathbf{X}_1, \dots, \mathbf{X}_n)$ be a sample of size n of d-dimensional observations $\mathbf{X}_i := (X_{i,1}, \dots, X_{i,d})^\intercal$ defined on some probability space $(\Omega, \mathscr{A}, \mathbb{P})$ and taking values in $\mathbb{R}^d$, with probability density function f. Consider a situation where an affine transformation is applied on the data. An invertible affine transformation $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ is a map such that $g(\mathbf{x}) = \mathbf{C}\mathbf{x} + \mathbf{b}$ where $\mathbf{C}$ is an invertible matrix of dimension $d \times d$ and $\mathbf{b}$ a d-dimensional column vector. The nature of the transformation g is such that, if applied to a random vector $\mathbf{X}$ with probability density function f, it gives rise to a new random vector $g(\mathbf{X})$ with probability density function $f_g = |\det(\mathbf{C})|^{-1} f \circ g^{-1}$, where the constant $|\det(\mathbf{C})|^{-1}$ depends on the scaling factor of the transformation g and rescales the volume of the transformed function $f \circ g$, so that it integrates to one.

Let $k(\mathbf{x}; \boldsymbol{\theta})$ be a kernel function on $\mathbb{R}^d$ parameterized by $\boldsymbol{\theta} \in \Theta$, $\tilde{p}$ be a DP (see Section 1.5 for details) with parameters ϑ (precision parameter) and $P_0 := \mathbb{E}[\tilde{p}]$ (base measure), where P_0 is a distribution defined on Θ. A DPM model is defined as

$$\tilde{f}(\mathbf{x}) = \int_\Theta k(\mathbf{x}; \boldsymbol{\theta}) \mathrm{d}\tilde{p}(\boldsymbol{\theta}). \tag{2.1}$$

Henceforth we denote by $\mathscr{F}$ the space of all density functions with support on $\mathbb{R}^d$. A DPM model defined as in (2.1) defines a random density on $\mathscr{F}$, that is a probability distribution over the space $\mathscr{F}$. Due to the almost sure discreteness property of $\tilde{p}$, the random density $\tilde{f}$ can be equivalently rewritten as

$$\tilde{f}(\mathbf{x}) = \sum_{i=1}^{\infty} w_i k(\mathbf{x}; \tilde{\boldsymbol{\theta}}_i),$$

where the random atoms $\tilde{\boldsymbol{\theta}}_i$ are i.i.d. from P_0, and the random jumps W_i, independent of the atoms, admit the stick-breaking representation derived by Sethuraman (1994) (see Section 1.4.1).

We assume for the kernel a d-dimensional Gaussian distribution, with density function $\phi_d(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, provided that $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the column vector $\boldsymbol{\mu}$ and the matrix $\boldsymbol{\Sigma}$ represent, respectively, mean vector and covariance matrix of the Gaussian kernel. This specification defines the model referred to as d-dimensional location-scale DPM-G, which can be represented in hierarchical form as

$$\begin{aligned} \mathbf{X}_i \mid \boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) &\overset{\text{ind}}{\sim} \phi_d(\mathbf{x}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \\ \boldsymbol{\theta}_i \mid \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p}, \\ \tilde{p} &\sim DP(\vartheta, P_0). \end{aligned} \tag{2.2}$$

Although other specifications for the base measure can be considered (see, e.g., Görür and Rasmussen, 2010), we choose to work within the framework set forth by Müller et al., 1996 where P_0 is defined as the product of two independent distributions for the location parameter $\boldsymbol{\mu}$ and the scale parameter $\boldsymbol{\Sigma}$, namely a multivariate normal and an inverse-Wishart distribution, that is

$$P_0(\mathrm{d}\boldsymbol{\mu}, \mathrm{d}\boldsymbol{\Sigma}; \boldsymbol{\pi}) = N_d(\mathrm{d}\boldsymbol{\mu}; \mathbf{m}_0, \mathbf{B}_0) \times IW(\mathrm{d}\boldsymbol{\Sigma}; \nu_0, \mathbf{S}_0). \tag{2.3}$$

Note that this specification of the base measure conjugacy of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ does not hold jointly but only marginally. For the sake of compactness, we use the notation $\boldsymbol{\pi} := (\mathbf{m}_0, \mathbf{B}_0, \nu_0, \mathbf{S}_0)$ to denote the vector of hyperparameters characterising the base measure P_0. We denote by Π the prior distribution induced on $\mathscr{F}$ by the DPM-G model (2.2) with base measure (2.3).

2.2 DPM-G model and affine transformations of the data

Let $\tilde{f}_\pi$ be a DPM-G model defined as in (2.2), with base measure (2.3) and hyperparameters $\boldsymbol{\pi}$. The next result shows that, for any invertible affine transformation $g(\mathbf{x}) = \mathbf{C}\mathbf{x} + \mathbf{b}$ of the data, there exists an opportune specification $\boldsymbol{\pi}_g := (\mathbf{m}_0^{(g)}, \mathbf{B}_0^{(g)}, \nu_0^{(g)}, \mathbf{S}_0^{(g)})$ of the hyperparameters characterizing the base measure in (2.3), such that $\tilde{f}_{\pi_g} = |\det(\mathbf{C})|^{-1} \tilde{f}_\pi \circ g^{-1}$. That is, for every $\omega \in \Omega$ and given a random vector $\mathbf{X}$ distributed according to $\tilde{f}_\pi(\omega)$, we have that $\tilde{f}_{\pi_g}(\omega)$ is the random density of the transformed random vector $g(\mathbf{X})$. The intuition is that, considering that an affine transformation changes center and volume of the support of the distribution of the data, the hyperparameters $\boldsymbol{\pi}_g$ allow to change random density $\tilde{f}_{\pi_g}$ accordingly.

Proposition 5 (Arbel, C. and Nipoti). *Let $\tilde{f}_\pi$ be a location-scale DPM-G model defined as in (2.2), with base measure (2.3) and hyperparameters $\boldsymbol{\pi} = (\mathbf{m}_0, \mathbf{B}_0, \nu_0, \mathbf{S}_0)$. For any invertible affine transformation $g(\mathbf{x}) = \mathbf{C}\mathbf{x} + \mathbf{b}$, we have*

$$\tilde{f}_{\pi_g} = |\det(\mathbf{C})|^{-1} \tilde{f}_\pi \circ g^{-1},$$

where $\boldsymbol{\pi}_g := (\mathbf{C}\mathbf{m}_0 + \mathbf{b}, \mathbf{C}\mathbf{B}_0\mathbf{C}^\intercal, \nu_0, \mathbf{C}\mathbf{S}_0\mathbf{C}^\intercal)$.

The previous proposition follows from Proposition 3 (Lijoi and Prünster, 2009). For the sake of completeness we report here a direct proof of Proposition 5.

Proof. Model $\tilde{f}_\pi$ can be written as

$$\tilde{f}_\pi(\mathbf{x}) = \int (2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tilde{p}(\mathrm{d}\boldsymbol{\mu}, \mathrm{d}\boldsymbol{\Sigma}; \boldsymbol{\pi})$$

$$= \int (2\pi)^{-\frac{d}{2}} |\det(\mathbf{C})| \det(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\intercal)^{-\frac{1}{2}}$$
$$\times \exp\left\{-\frac{1}{2}(\mathbf{C}\mathbf{x}+\mathbf{b}-\mathbf{C}\boldsymbol{\mu}-\mathbf{b})^\intercal(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\intercal)^{-1}(\mathbf{C}\mathbf{x}+\mathbf{b}-\mathbf{C}\boldsymbol{\mu}-\mathbf{b})\right\} \tilde{p}(\mathrm{d}\boldsymbol{\mu}, \mathrm{d}\boldsymbol{\Sigma}; \boldsymbol{\pi}).$$

By performing the change of variables $\mathbf{S} = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\intercal$ and $\mathbf{m} = \mathbf{C}\boldsymbol{\mu} + \mathbf{b}$ and observing that, by standard properties of the inverse-Wishart and normal distributions,

1. $\boldsymbol{\Sigma} \sim IW(\nu_0, \mathbf{S}_0)$ implies $\mathbf{S} \sim IW(\nu_0, \mathbf{C}\mathbf{S}_0\mathbf{C}^\intercal)$,
2. $\boldsymbol{\mu} \sim N_d(\mathbf{m}_0, \mathbf{B}_0)$ implies $\mathbf{m} \sim N_d(\mathbf{C}\mathbf{m}_0 + \mathbf{b}, \mathbf{C}\mathbf{B}_0\mathbf{C}^\intercal)$,
3. $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ implies $\mathbf{C}\mathbf{X} + \mathbf{b} \sim N_d(\mathbf{m}, \mathbf{S})$,

we obtain

$$\tilde{f}_\pi(\mathbf{x}) = |\det(\mathbf{C})| \int (2\pi)^{-\frac{d}{2}} \det(\mathbf{S})^{-\frac{1}{2}}$$
$$\times \exp\left\{-\frac{1}{2}(\mathbf{C}\mathbf{x}+\mathbf{b}-\mathbf{m})^\intercal \mathbf{S}^{-1}(\mathbf{C}\mathbf{x}+\mathbf{b}-\mathbf{m})\right\} \tilde{p}(\mathrm{d}\mathbf{m}, \mathrm{d}\mathbf{S}; \boldsymbol{\pi}_g)$$
$$= |\det(\mathbf{C})| \tilde{f}_{\pi_g}(g(\mathbf{x})).$$

A simple reparametrisation leads to $\tilde{f}_{\pi_g} = |\det(\mathbf{C})|^{-1} \tilde{f}_\pi \circ g^{-1}$. All the identities in this proof are deterministic, that is they hold for every $\omega \in \Omega$. □

This result implies that, for any invertible affine transformation g, modelling the set of observations $\mathbf{X}^{(n)}$ with a DPM-G model (2.2), with base measure (2.3) and hyperparameters $\boldsymbol{\pi}$, is equivalent with assuming the same model with transformed hyperparameters $\boldsymbol{\pi}_g$, for the transformed observations $g(\mathbf{X})^{(n)} := (g(\mathbf{X}_1), \ldots, g(\mathbf{X}_n))$. As a by-product, the same posterior inference can be drawn conditionally on both the original and the transformed set of observations, as the conditional distribution of the random density $\tilde{f}_{\pi_g}$, given $g(\mathbf{X})^{(n)}$, coincides with the conditional distribution of $|\det(\mathbf{C})|^{-1} \tilde{f}_\pi \circ g^{-1}$, given $\mathbf{X}^{(n)}$. Proposition 5 thus provides a formal justification for the procedure of transforming data, e.g. via standardisation or normalisation, often adopted to achieve numerical efficiency: as long as the prior specification of the hyperparameters of a DPM-G model respects the condition of Proposition 5, transforming the data does not affect posterior inference.

The result in Proposition 5 is surely not surprising (see Lijoi and Prünster, 2009), nonetheless it provides an explicit and useful characterization of the hyperparameters π_g for the flexible class of DPM-G models considered here. In general, the elicitation of an honest prior, thus independent of the data, for the hyperparameters $\boldsymbol{\pi}$ of the base measure (2.3) of a DPM model is a difficult task. A popular practice, therefore, consists in setting the hyperparameters equal to some empirical estimates $\hat{\boldsymbol{\pi}}(\mathbf{X}^{(n)})$, by applying the so-called empirical Bayes approach (see, e.g., Lehmann and Casella, 2006). Recent investigations (Petrone et al., 2014; Donnet et al., 2018) provide a theoretical justification of this hybrid procedure by shedding light on its asymptotic properties. The next example shows that this procedure satisfies the assumptions of Proposition 5 and, thus, guarantees that posterior Bayesian inference, under an empirical Bayes approach, is not affected by affine transformations to the data.

Example 1 (Empirical Bayes approach)**.** A commonly used empirical Bayes approach for specifying the hyperparameters $\boldsymbol{\pi}$ of a DPM-G model, defined as in (2.2) and (2.3), consists in setting

$$\mathbf{m}_0 = \overline{\mathbf{X}}, \qquad \mathbf{B}_0 = \frac{1}{\gamma_1}\mathbf{S}^2_{\mathbf{X}}, \qquad \mathbf{S}_0 = \frac{\nu_0 - d - 1}{\gamma_2}\mathbf{S}^2_{\mathbf{X}}, \tag{2.4}$$

where $\overline{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i/n$ and $\mathbf{S}^2_{\mathbf{X}} = \sum_{i=1}^n(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^\intercal/(n-1)$ are the sample mean vector and the sample covariance matrix, respectively, and $\gamma_1, \gamma_2 > 0$, $\nu_0 > d+1$. This specification for the hyperparameters $\boldsymbol{\pi}$ has a straightforward interpretation. Namely, the parameter $\mathbf{m}_0$, mean of the prior guess distribution of $\boldsymbol{\mu}$, can be interpreted as the overall mean value and, in absence of available prior information, set equal to the observed sample mean. Similarly, the parameter $\mathbf{B}_0$, covariance matrix of the prior guess distribution of $\boldsymbol{\mu}$, is set equal to a penalised version of the sample covariance matrix $\mathbf{S}^2_{\mathbf{X}}$, where γ_1 takes on the interpretation of the size of the ideal prior sample upon which the prior guess on the distribution of $\boldsymbol{\mu}$ is based. Similarly, the hyperparameter $\mathbf{S}_0$ is set equal to a penalised version of the sample covariance matrix $\mathbf{S}^2_{\mathbf{X}}$, choice that corresponds to the prior guess that the covariance matrix of each component of the mixture coincides with a rescaled version of the sample covariance matrix. Specifically, $\mathbf{S}_0 = \mathbf{S}^2_{\mathbf{X}}(\nu_0 - d - 1)/\gamma_2$ follows by setting $\mathbb{E}[\boldsymbol{\Sigma}] = \mathbf{S}^2_{\mathbf{X}}/\gamma_2$ and observing that, by standard properties of the inverse-Wishart distribution, $\mathbb{E}[\boldsymbol{\Sigma}] = \mathbf{S}_0/(\nu_0 - d - 1)$. Finally the parameter ν_0 takes on the interpretation of the size of an ideal prior sample upon which the prior guess $\mathbf{S}_0$ is based. Next we focus on the setting of the hyperparameters $\boldsymbol{\pi}_g$, given the transformed observations $g(\mathbf{X})^{(n)}$. The same empirical Bayes procedure adopted in (2.4) leads to

$$\mathbf{m}_0^{(g)} = \overline{g(\mathbf{X})} = \mathbf{C}\overline{\mathbf{X}} + \mathbf{b}, \qquad \mathbf{B}_0^{(g)} = \frac{1}{\gamma_1}\mathbf{S}^2_{g(\mathbf{X})}, \qquad \mathbf{S}_0^{(g)} = \frac{\nu_0 - d - 1}{\gamma_2}\mathbf{S}^2_{g(\mathbf{X})}.$$

Observing that $\mathbf{S}^2_{g(\mathbf{X})} = \mathbf{C}\mathbf{S}^2_{\mathbf{X}}\mathbf{C}^\intercal$ and setting $\nu_0^{(g)} = \nu_0$ shows that the described empirical Bayes procedure corresponds to $\boldsymbol{\pi}_g = (\mathbf{C}\mathbf{m}_0 + \mathbf{b}, \mathbf{C}\mathbf{B}_0\mathbf{C}^\intercal, \nu_0, \mathbf{C}\mathbf{S}_0\mathbf{C}^\intercal)$ and, thus, by Proposition 5, $\tilde{f}_{\boldsymbol{\pi}_g} = |\det(\mathbf{C})|^{-1}\tilde{f}_{\boldsymbol{\pi}} \circ g^{-1}$. □

2.3 An introduction to posterior consistency

In the context of Bayesian nonparametric mixture models, due to the infinite dimensional nature of the mixing measure, the study of asymptotic properties, as the sample size grows up to ∞, is surely not trivial. Adopting the usual frequentist approach in the large n regime, here we work 'as if' the observations $\mathbf{X}^{(n)}$ were generated from a true and fixed generating process (see for instance Rousseau (2016)). We also assume that this data generating process admits a density function with respect to the Lebesgue measure, denoted by f^*, where $f^* \in \mathscr{F}$, with $\mathscr{F}$ space of all density function with suport $\mathbb{R}^d$. We already introduced the notation Π to denote the prior distribution of the random density $\tilde{f}$, induced on the space $\mathscr{F}$, by a DPM model defined as in (2.2) with base measure (2.3). Moreover, we denote by $\Pi(\cdot \mid \mathbf{X}^{(n)})$ the posterior distribution of the random density $\tilde{f}$, given a set of observations $\mathbf{X}^{(n)}$, and we focus on its behaviour when $n \to \infty$. We consider as metrics on $\mathscr{F}$ the Hellinger one, $d_H(f,g) = [\int(\sqrt{f} - \sqrt{g})^2]^{1/2}$, or the L_1 metric, $d_{L_1} = \int \mid f - g \mid$. Note that they induce an equivalent topology on the space $\mathscr{F}$, given by the fact that $d_H^2(f,g) \le d_{L_1}(f,g) \le 2d_H(f,g)$. Moreover, we denote by $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^d$.

By posterior consistency of $\Pi(\cdot \mid \mathbf{X}^{(n)})$ at the atom f^* of $\mathscr{F}$ we mean that the posterior distribution $\Pi(\cdot \mid \mathbf{X}^{(n)})$, when n goes to ∞, accumulates probability mass in a neighborhood of f^*, then for any $\varepsilon > 0$ we have

$$\Pi\left(\{f : \rho(f, f^*) > \varepsilon\} \mid \mathbf{X}^{(n)}\right) \to 0$$

in F_0^n-probability, where F_0^n is the n-product measure and F_0 is the probability measure associated to f^*, with ρ being the Hellinger metric or, equivalently, the L_1 metric. To study the behavior of the posterior distribution $\Pi(\cdot \mid \mathbf{X}^{(n)})$, one has first to establish some conditions on the support $\mathscr{F}$. We require that $\mathscr{F}$ is dense in the neighborhood of f^*, condition needed to study limit situations. About the density of the space $\mathscr{F}$, we consider the Kullback-Leibler divergence (KL, also known as relative entropy) to characterize the support. An useful interpretation of the Kullback-Leibler divergence between two functions, $KL(g \,||\, f) = d_{KL}(f, g)$, can be borrowed by the information theory, where this divergence is used to measure the information gained by using f instead of g. We say that Π satisfies the KL property at the point f^* if

$$\Pi\left(\left\{f : \int f \log(f/f^*) \leq \eta\right\}\right) > 0 \qquad \forall \eta > 0. \tag{2.5}$$

This property, in the context of DPM-G models, was first studied by Wu and Ghosal (2008), in the case of scalar covariance matrix, i.e. $\Sigma = \sigma^2 I$, for $\sigma^2 \in \mathbb{R}^+$ and I identity matrix. In a recent paper, Canale and De Blasi (2017) studied the non-scalar case, where Σ is defined on the space $\mathcal{S}_\Sigma$ of semidefinite positive matrices.

Lemma 2. *(Lemma 1 in Canale and De Blasi, 2017). Let Π be the prior on $\mathscr{F}$ induced by (2.2) with base measure (2.3), and $f^* \in \mathscr{F}$, true generating density of $\mathbf{X}^{(n)}$, satisfies the conditions*

- *A1. $0 < f^*(\mathbf{x}) < M$, for some constant M and for all $\mathbf{x} \in \mathbb{R}^d$,*
- *A2. $\left|\int f^*(\mathbf{x}) \log f^*(\mathbf{x}) \mathrm{d}\mathbf{x}\right| < \infty$,*
- *A3. $\exists\, \delta > 0$ such that $\int f^*(\mathbf{x}) \log\left(f^*(\mathbf{x})/\varphi_\delta(\mathbf{x})\right) \mathrm{d}\mathbf{x} < \infty$, where $\varphi_\delta(\mathbf{x}) = \inf_{\{\mathbf{t}:\, \|\mathbf{t}-\mathbf{x}\| < \delta\}} f^*(\mathbf{t})$,*
- *A4. for some $\eta > 0$, $\int \|\mathbf{x}\|^{2(1+\eta)} f^*(\mathbf{x}) \mathrm{d}\mathbf{x} < \infty$.*

Then Π satisfies the Kullback-Leibler property (2.5).

The previous assumptions, A1–A4, require some regularity for the true data-generating density f^*. Assumption A1 requires the density f^* to have full support. The second assumption requires the finiteness of the entropy, thus allowing the computation of the relative entropy. Assumption A3 refers to the local regularity of f^*, in terms of finite local relative entropy with respect to the local infimum. Finally, Assumption A4 requires the tails of f^* to be thin enough for some moment of order strictly larger than two to exist. We refer to Canale and De Blasi (2017) for a proof of Lemma 2.

Next we introduce some consistency results for DPM-G models which will represent the main ingredients for our study of the asymptotic robustness of the same class of models in the next section. While several authors considered the problem from a location-scale DPM models, we refer to the main results in the paper of Canale and De Blasi (2017), where they introduced a theorem with relaxed assumptions on the base measure specification. Let $\lambda_1(\mathbf{\Sigma}^{-1}), \ldots, \lambda_d(\mathbf{\Sigma}^{-1})$ be the sequence of eigenvalues of $\mathbf{\Sigma}^{-1}$, in increasing order. Henceforth we write $f(x) \lesssim g(x)$ to indicate that the inequality $f(x) \leq cg(x)$ holds for some constant c and for any x.

Theorem 6. *(Theorem 2 in Canale and De Blasi, 2017). Let $f^* \in \mathscr{F}$, true generating density of $\mathbf{X}^{(n)}$, satisfy the conditions of Lemma 2, and model $\mathbf{X}^{(n)}$ by means of a DPM-G model defined in (2.2). Suppose that the base measure P_0 has the product form $P_0(\mathrm{d}\boldsymbol{\mu}, \mathrm{d}\mathbf{\Sigma}) = P_{0,1}(\mathrm{d}\boldsymbol{\mu})P_{0,2}(\mathrm{d}\mathbf{\Sigma})$ and that $P_{0,1}$*

and $P_{0,2}$ satisfy the following conditions: for some positive constants c_1, c_2, c_3, $r > (d-1)/2$ and $\kappa > d(d-1)$,

B1. $P_{0,1}(\|\boldsymbol{\mu}\| > x) \lesssim x^{-2(r+1)}$,

B2. $P_{0,2}(\lambda_d(\boldsymbol{\Sigma}^{-1}) > x) \lesssim \exp\{-c_1 x^{c_2}\}$,

B3. $P_{0,2}\left(\lambda_1(\boldsymbol{\Sigma}^{-1}) < \frac{1}{x}\right) \lesssim x^{-c_3}$,

B4. $P_{0,2}\left(\frac{\lambda_d(\boldsymbol{\Sigma}^{-1})}{\lambda_1(\boldsymbol{\Sigma}^{-1})} > x\right) \lesssim x^{-\kappa}$,

all for any sufficiently large x. Then the posterior distribution $\Pi(\cdot|\mathbf{X}^{(n)})$ is consistent at f^*, that is, for every $\varepsilon > 0$,

$$\Pi\left(f : \rho(f, f^*) < \varepsilon \mid \mathbf{X}^{(n)}\right) \longrightarrow 1$$

as $n \to \infty$.

Conditions B1–B4 refer to the shape of the base measure. The first condition is about the tails of the location component distribution, and their weights. The second condition describes the behavior of the largest eigenvalue of $\boldsymbol{\Sigma}^{-1}$, thus requiring that the most dispersed component in the eigenvalue decomposition latent space does not explode. Condition B3 refers to the less dispersed component and its decreasing speed. The last condition B4 describes the ratio between the biggest and the smallest eigenvalues, preventing situations where the most disperse dimension is completely dominating the less dispersed one. We refer to Canale and De Blasi (2017) for details regarding the proof.
Theorem 6 proves that location-scale DPM-G models, under the described regularity conditions on the base measure, are consistent. In particular, the DPM-G model, defined as in (2.2) with base measure (2.3), satisfies the requirements of Theorem 6.

Lemma 3. *Conditions B1–B4 of Theorem 6 are satisfied by the multivariate normal / inverse-Wishart base measure (2.3) with $\nu_0 > (d+1)(2d-3)$.*

That is, if $f^* \in \mathscr{F}$ is the true generating density of $\mathbf{X}^{(n)}$, the previous results guarantee that, for $n \to \infty$, the posterior distribution accumulates probability mass in a small neighbourhood of f^*. Although the proof of Lemma 3 can be found in Canale and De Blasi (2017) (see Corollary 1, relying, in turn, on results by Shen et al. (2013)), we provide it here for the sake of completeness and in order to account for the slightly different prior specification considered in this thesis.

Proof. The proof is based on some results of Shen et al. (2013) combined with Canale and De Blasi (2017). We check, point-by-point, that the conditions of Theorem 6 are satisfied.

B1. Since $\boldsymbol{\mu} \sim N_d(\mathbf{m}_0, \mathbf{B}_0)$, then $\|\boldsymbol{\mu}\|^2 \sim \chi^2_d(\delta)$ where d is the dimension of $\boldsymbol{\mu}$ and $\delta = \|\mathbf{m}_0\|$ is the non-centrality parameter of the chi-squared distribution. Then, for sufficiently large x,

$$P_{0,1}\left(\|\boldsymbol{\mu}\|^2 > x\right) \leq \left(\frac{x}{d}\right)^{\frac{d}{2}} \exp\left\{\frac{d-x}{2}\right\} \lesssim x^{-2(r+1)},$$

which holds for $r > (d-1)/2$.

B2. We know that $\boldsymbol{\Sigma} \sim IW(\nu_0, \mathbf{S}_0)$ and we start by considering the case corresponding to $\mathbf{S}_0 = \mathbf{I}_d$, where $\mathbf{I}_d$ denotes the d-dimensional identity matrix. It is known that $\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}) \sim \chi^2_{\nu_0 d}$.

Thus, for sufficiently large x,

$$P_{0,2}\left(\lambda_d(\mathbf{\Sigma}^{-1}) > x\right) \le P_{0,2}\left(\mathrm{Tr}(\mathbf{\Sigma}^{-1}) > x\right)$$
$$\le \left(\frac{x}{\nu_0 d}\right)^{\frac{\nu_0 d}{2}} \exp\left\{\frac{\nu_0 d - x}{2}\right\} \lesssim \exp\left\{-c_1 x^{c_2}\right\},$$

for some positive constants c_1 and c_2. This result can be easily generalised to the case $\mathbf{S}_0 \neq \mathbf{I}_d$ since $IW(\mathrm{d}\mathbf{\Sigma}; \nu_0, \mathbf{S}_0) = \mathbf{S}_0^{-1} IW(\mathrm{d}\mathbf{\Sigma}; \nu_0, \mathbf{I}_d)$.

B3. We know that $\mathbf{\Sigma} \sim IW(\nu_0, \mathbf{S}_0)$ and we start by supposing that $\mathbf{S}_0 = \mathbf{I}_d$. The joint distribution of the eigenvalues $\boldsymbol{\lambda}\left(\mathbf{\Sigma}^{-1}\right)$ is known to be equal to

$$f_{\boldsymbol{\lambda}}(x_1, \dots, x_d) = c_{d,\nu_0} \exp\left\{-\sum_{j=1}^{d} \frac{x_j}{2}\right\} \prod_{j=1}^{d} x_j^{\frac{(\nu_0 - d + 1)}{2}} \prod_{j<k} (x_k - x_j),$$

for some normalising constant c_{d,ν_0}, if $(x_1, \dots, x_d) \in (0, \infty)^d$ is such that $x_1 \le \cdots \le x_d$, and equal to 0 otherwise. It is easy to verify that, on the support of $f_{\boldsymbol{\lambda}}$,

$$\prod_{j<k} (x_k - x_j) \le \prod_{j<k} x_k = \prod_{k=2}^{d} x_k^{k-1}.$$

The density function of $\lambda_1(\mathbf{\Sigma}^{-1})$ then becomes

$$f_{\lambda_1}(x_1) = \int \cdots \int f_{\boldsymbol{\lambda}}(x_1, \dots, x_d) \mathrm{d}x_2 \cdots \mathrm{d}x_d$$
$$\le c_{d,\nu_0} x_1^{\frac{\nu_0 - d + 1}{2}} \mathrm{e}^{-\frac{x_1}{2}} \prod_{k=2}^{d} \int_0^{\infty} x_k^{\frac{\nu_0 - d + 1}{2} + k - 1} \mathrm{e}^{-\frac{x_k}{2}} \mathrm{d}x_k$$
$$= c'_{d,\nu_0} x_1^{\frac{\nu_0 - d + 1}{2}} \exp\left\{-\frac{x_1}{2}\right\},$$

for some new normalising constant c'_{d,ν_0}. Then for any $x > 0$ we have

$$P_{0,2}\left(\lambda_1(\mathbf{\Sigma}^{-1}) < \frac{1}{x}\right) \le c'_{d,\nu_0} \int_0^{\frac{1}{x}} x_1^{\frac{\nu_0 - d + 1}{2}} dx_1 \lesssim x^{-c_3 x}$$

for some constant c_3 and sufficiently large x. Again, this result can be generalised to the case $\mathbf{S}_0 \neq \mathbf{I}_d$ since $IW(\mathrm{d}\mathbf{\Sigma}; \nu_0, \mathbf{S}_0) = \mathbf{S}_0^{-1} IW(\mathrm{d}\mathbf{\Sigma}; \nu_0, \mathbf{I}_d)$.

B4. We know that $\mathbf{\Sigma} \sim IW(\nu_0, \mathbf{S}_0)$ and we start by considering the case corresponding to $\mathbf{S}_0 = \mathbf{I}_d$. We define $Z(\mathbf{\Sigma}^{-1}) = \lambda_d(\mathbf{\Sigma}^{-1}) / \lambda_1(\mathbf{\Sigma}^{-1})$ and the function $q(\boldsymbol{\lambda}(\mathbf{\Sigma}^{-1})) = (\lambda_1(\mathbf{\Sigma}^{-1}), \dots, \lambda_{d-1}(\mathbf{\Sigma}^{-1}), Z(\mathbf{\Sigma}^{-1}))$. Let $J_{q^{-1}}$ denote the Jacobian of the inverse of the function q, and observe that

$$f_{\lambda_1, \dots, \lambda_{d-1}, Z}(x_1, \dots, x_{d-1}, z) = |J_{q^{-1}}| f_{\boldsymbol{\lambda}}(x_1, \dots, x_{d-1}, x_1 z).$$

Then, by marginalising with respect to the first $d-1$ components, we obtain

$$
\begin{aligned}
f_Z(z) &= \int \cdots \int |J_{q^{-1}}| f_\Lambda(x_1, \dots, x_{d-1}, x_1 z) \mathrm{d}x_1 \cdots \mathrm{d}x_{d-1} \\
&= \int \cdots \int c_{d,\nu_0} \exp\left\{ -\sum_{j=1}^{d-1} \frac{x_j}{2} - \frac{x_1 z}{2} \right\} \prod_{j=1}^{d-1} x_j^{\frac{\nu_0+1-d}{2}} (x_1 z)^{\frac{\nu_0+1-d}{2}} \\
&\quad \times \prod_{j<k\leq d-1} (x_k - x_j) \prod_{j=1}^{d-1} (x_1 z - x_j) x_1 \mathrm{d}x_1 \cdots \mathrm{d}x_{d-1} \\
&\leq \int \cdots \int c_{d,\nu_0} \exp\left\{ -\sum_{j=1}^{d-1} \frac{x_j}{2} - \frac{x_1 z}{2} \right\} \prod_{j=1}^{d-1} x_j^{\frac{\nu_0+1-d}{2}} (x_1 z)^{\frac{\nu_0+1-d}{2}} \\
&\quad \prod_{k=2}^{d-1} x_k^{k-1} \prod_{j=1}^{d-1} (x_1 z) x_1 \mathrm{d}x_1 \cdots \mathrm{d}x_{d-1} \\
\\
&= c'_{d,\nu_0} z^{(\nu_0+d-1)/2} \int \exp\left\{ -x_1 \left(\frac{z+1}{2} \right) \right\} x_1^{\nu_0+1} \mathrm{d}x_1 \\
&= c'_{d,\nu_0} (\nu_0+1)! \left(\frac{2}{z+1} \right)^{\nu_0+2} z^{(\nu_0+d-1)/2} \\
&= c''_{d,\nu_0} \frac{z^{(\nu_0+d-1)/2}}{(z+1)^{\nu_0+2}} \\
&\leq c''_{d,\nu_0} z^{-(\nu_0-d+5)/2},
\end{aligned}
$$

for some constants c_{d,ν_0}, c'_{d,ν_0} and c''_{d,ν_0}. Thus we have

$$
P_{0,2}\left(Z > x\right) = \int_x^\infty f_Z(z) \mathrm{d}z \leq c''_{d,\nu_0} \int_x^\infty z^{-(\nu_0-d+5)/2} \mathrm{d}z \lesssim x^{-\kappa},
$$

for sufficiently large x, where $\kappa = (\nu_0 - d + 3)/2 > d(d+1)$ by the assumption that $\nu_0 > (d+1)(2d-3)$.

□

2.4 Large *n* asymptotic robustness of DPM-G models

We are interested in the relation between a DPM-G specified conditionally to the data $\mathbf{X}^{(n)}$ and a DPM-G conditionally to the transformed data $g(\mathbf{X})^{(n)}$, with $g(\mathbf{x}) = \mathbf{C}\mathbf{x} + \mathbf{b}$ invertible affine transformation. Let Π be the prior distribution of the random density $\tilde{f}$, induced on the space $\mathscr{F}$, by a DPM model defined as in (2.2) with base measure (2.3). Moreover, we denote by $\Pi(\cdot \mid \mathbf{X}^{(n)})$ the posterior distribution given a set of observations $\mathbf{X}^{(n)}$ and by $\Pi(\cdot \mid g(\mathbf{X})^{(n)})$ the posterior distribution given a transformed set of observations $g(\mathbf{X})^{(n)}$. Finally, we use the notation $\Pi_2(\cdot \mid \mathbf{X}^{(n)})$ to refer to their joint posterior distribution on $\mathscr{F} \times \mathscr{F}$.
Let $f_g^* := |\det(\mathbf{C})|^{-1} f^* \circ g^{-1}$ be the true generating density corresponding to the transformed data. First we need to show that the KL property, introduced in (2.5), holds also for the true generating density of the transformed data. Next lemma shows that if f^* satisfies conditions A1–A4 of Lemma 2, the same are satisfied also by so f_g^*, for any invertible affine transformation g.

Lemma 4. *If conditions A1–A4 of Lemma 2 are satisfied by f^*, then for any invertible affine transformation $g(\mathbf{x}) = \mathbf{C}\mathbf{x} + \mathbf{b}$, they are also satisfied by f_g^*.*

Proof. We assume that f^* satisfies conditions A1–A4 of Theorem 7 and check that the same holds for f^*_g.

A1. Assume that $0 < f^*(\mathbf{x}) < M$ for every $\mathbf{x} \in \mathbb{R}^d$ and some $M > 0$. Then, for every $\mathbf{x} \in \mathbb{R}^d$, we have $f^*_g(\mathbf{x}) = |\det(\mathbf{C})|^{-1} f^*(g^{-1}(\mathbf{x}))$ which implies

$$0 < f^*_g(\mathbf{x}) < M' = |\det(\mathbf{C})|^{-1} M.$$

A2. Assume that f^* is such that $\left|\int f^*(\mathbf{x}) \log f^*(\mathbf{x}) d\mathbf{x}\right| < \infty$. Then, we have

$$\begin{aligned}
&\left|\int f^*_g(\mathbf{x}) \log f^*_g(\mathbf{x}) d\mathbf{x}\right| \\
&= \left|\int |\det(\mathbf{C})|^{-1} f^*(g^{-1}(\mathbf{x})) \log\left(|\det(\mathbf{C})|^{-1} f^*(g^{-1}(\mathbf{x}))\right) d\mathbf{x}\right| \\
&= |\det(\mathbf{C})|^{-1} \left|\int f^*(g^{-1}(\mathbf{x})) \log\left(|\det(\mathbf{C})|^{-1}\right) d\mathbf{x}\right. \\
&\qquad\qquad \left. + \int f^*(g^{-1}(\mathbf{x})) \log\left(f^*(g^{-1}(\mathbf{x}))\right) d\mathbf{x}\right| \\
&= \left|\int f^*(\mathbf{y}) \log\left(|\det(\mathbf{C})|^{-1}\right) d\mathbf{y} + \int f^*(\mathbf{y}) \log\left(f^*(\mathbf{y})\right) d\mathbf{y}\right| \\
&= \left|\log\left(|\det(\mathbf{C})|^{-1}\right) + \int f^*(\mathbf{y}) \log\left(f^*(\mathbf{y})\right) d\mathbf{y}\right| \\
&\leq \left|\log\left(|\det(\mathbf{C})|^{-1}\right)\right| + \left|\int f^*(\mathbf{y}) \log\left(f^*(\mathbf{y})\right) d\mathbf{y}\right| < \infty.
\end{aligned}$$

A3. Assume that f^* satisfies A3 with some δ'. Let $\delta = |\det(\mathbf{C})|^{-1}\delta'$ and observe that since g is invertible

$$\begin{aligned}
\varphi^{(g)}_\delta(g(\mathbf{y})) = \inf_{\{\mathbf{t}\,:\, \|\mathbf{t}-g(\mathbf{y})\|<\delta\}} f^*_g(\mathbf{t}) &= \inf_{\{\mathbf{s}\,:\, \|g(\mathbf{s})-g(\mathbf{y})\|<\delta\}} f^*_g(g(\mathbf{s})) \\
&= \inf_{\{\mathbf{s}\,:\, \|\mathbf{s}-\mathbf{y}\|<\delta'\}} f^*(\mathbf{s}) |\det(\mathbf{C})|^{-1} = \varphi_{\delta'}(\mathbf{y}) |\det(\mathbf{C})|^{-1}.
\end{aligned}$$

Then we have that

$$\begin{aligned}
\int f^*_g(\mathbf{x}) \log\left(\frac{f^*_g(\mathbf{x})}{\varphi^{(g)}_\delta(\mathbf{x})}\right) d\mathbf{x} &= \int f^*_g(g(\mathbf{y})) \log\left(\frac{f^*_g(g(\mathbf{y}))}{\varphi^{(g)}_\delta(g(\mathbf{y}))}\right) |\det(\mathbf{C})| d\mathbf{y} \\
&= \int f^*(\mathbf{y}) \log\left(\frac{|\det(\mathbf{C})|^{-1} f^*(\mathbf{y})}{|\det(\mathbf{C})|^{-1} \varphi_{\delta'}(\mathbf{y})}\right) d\mathbf{y} \\
&= \int f^*(\mathbf{y}) \log\left(\frac{f^*(\mathbf{y})}{\varphi_{\delta'}(\mathbf{y})}\right) d\mathbf{y} < \infty
\end{aligned}$$

where the last inequality holds by Assumption A3 on f^* with δ'. This finally shows that f^*_g satisfies Assumption A3 with δ.

A4. Observe that

$$\begin{aligned}\int \|\mathbf{x}\|^{2(1+\eta)} f_g^*(\mathbf{x})\mathrm{d}\mathbf{x} &= \int \|g(\mathbf{y})\|^{2(1+\eta)} f_g^*(g(\mathbf{y}))|\det(\mathbf{C})|\mathrm{d}\mathbf{y} \\ &= \int \|g(\mathbf{y})\|^{2(1+\eta)} f^*(\mathbf{y})\mathrm{d}\mathbf{y} \\ &\leq \int 2^{2(1+\eta)-1}\left(\|\mathbf{C}\mathbf{y}\|^{2(1+\eta)} + \|\mathbf{b}\|^{2(1+\eta)}\right) f^*(\mathbf{y})\mathrm{d}\mathbf{y},\end{aligned}$$

where the last inequality follows by combining triangular and Jensen's inequalities. Thus we can write

$$\begin{aligned}&\int \|\mathbf{x}\|^{2(1+\eta)} f_g^*(\mathbf{x})\mathrm{d}\mathbf{x} \\ &\quad\leq 2^{2(1+\eta)-1}\left(|\det(\mathbf{C})|^{2(1+\eta)}\int \|\mathbf{y}\|^{2(1+\eta)} f^*(\mathbf{y})\mathrm{d}\mathbf{y} + \|\mathbf{b}\|^{2(1+\eta)}\right) < \infty,\end{aligned}$$

where the last inequality follows by Assumption A4 on f^*.

□

The next result shows that, under mild conditions on the true generating distribution, the posterior joint distribution accumulates probability mass in the set

$$\{(f_1, f_2) \in \mathscr{F} \times \mathscr{F} \text{ s.t. } f_1 = |\det(\mathbf{C})| f_2 \circ g\},$$

when n goes to ∞. Henceforth, we will say that a DPM-G model (2.2) with base measure (2.3) is *asymptotically robust* to affine transformations of the data. Figure 2.1 displays an abstract representation of the region of $\mathscr{F} \times \mathscr{F}$ where the posterior joint distribution accumulates its probability mass, for large values of n.

Theorem 7. *Let $f^* \in \mathscr{F}$, true generating density of $\mathbf{X}^{(n)}$, satisfy the conditions*

A1. $0 < f^*(\mathbf{x}) < M$, for some constant M and for all $\mathbf{x} \in \mathbb{R}^d$,

A2. $\left|\int f^*(\mathbf{x}) \log f^*(\mathbf{x})\mathrm{d}\mathbf{x}\right| < \infty$,

A3. $\exists\, \delta > 0$ such that $\int f^*(\mathbf{x}) \log\left(f^*(\mathbf{x}) / \varphi_\delta(\mathbf{x})\right)\mathrm{d}\mathbf{x} < \infty$, where $\varphi_\delta(\mathbf{x}) = \inf_{\{\mathbf{t}\,:\,\|\mathbf{t}-\mathbf{x}\| < \delta\}} f^*(\mathbf{t})$,

A4. for some $\eta > 0$, $\int \|\mathbf{x}\|^{2(1+\eta)} f^*(\mathbf{x})\mathrm{d}\mathbf{x} < \infty$.

Let $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ be an invertible affine transformation and $\Pi_2(\cdot \mid \mathbf{X}^{(n)})$ be the joint posterior distribution induced by a DPM-G as (2.1) with base measure (2.3) where $\nu_0 > (d+1)(2d-3)$. Then, for any $\varepsilon > 0$,

$$\Pi_2((f_1, f_2) : \rho(f_1, |\det(\mathbf{C})| f_2 \circ g) < \varepsilon \mid \mathbf{X}^{(n)}) \longrightarrow 1$$

as $n \to \infty$.

Assumptions A1–A4 in Theorem 7 are the same of Lemma 2 and their interpretation was already discussed after Lemma 2. The proof of Theorem 7 follows by combining Theorem 6 with Lemma 3 and Lemma 4.

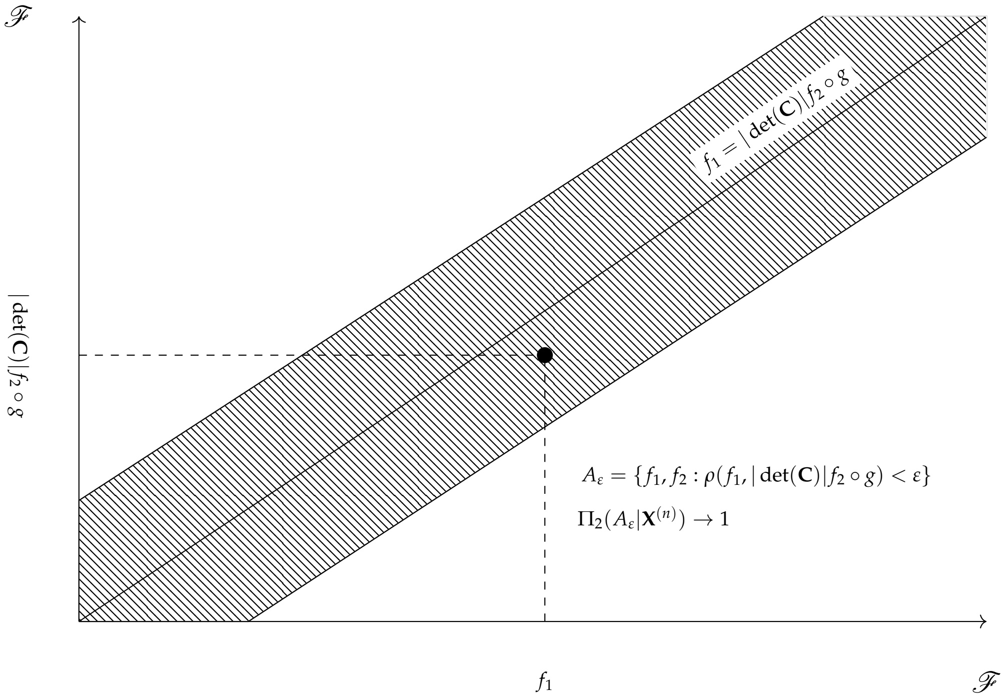
$\mathscr{F}$
$f_1 = |\det(\mathbf{C})|f_2 \circ g$
$|\det(\mathbf{C})|f_2 \circ g$
$A_\varepsilon = \{f_1, f_2 : \rho(f_1, |\det(\mathbf{C})|f_2 \circ g) < \varepsilon\}$
$\Pi_2(A_\varepsilon|\mathbf{X}^{(n)}) \to 1$
f_1
$\mathscr{F}$

FIGURE 2.1: Abstract graphical representation of the set in $\mathscr{F} \times \mathscr{F}$ where the posterior joint distribution accumulates its probability mass. In evidence a point on $\mathscr{F} \times \mathscr{F}$. Filled area: the subset of $\mathscr{F} \times \mathscr{F}$ where $\rho(f_1, |\, det(\mathbf{C}) \mid f_2 \circ g) < \varepsilon$, for a fixed $\varepsilon > 0$.

Proof of Theorem 7. By combining Lemma 3, Lemma 4 and Theorem 6, we have that for any $\epsilon > 0$,

$$\Pi\left(f : \rho(f, f^*) < {}^{\epsilon}/_{2} \mid \mathbf{X}^{(n)}\right) \longrightarrow 1,$$
$$\Pi\left(f : \rho(f, f_g^*) < {}^{\epsilon}/_{2} \mid g(\mathbf{X})^{(n)}\right) \longrightarrow 1,$$

as $n \to \infty$. We notice that the distance ρ is invariant with respect to change of variables and thus $\rho(|\det(\mathbf{C})|f_2 \circ g, f^*) = \rho(f_2, f_g^*)$. This, combined with the triangular inequality, leads to

$$\begin{aligned}
&\Pi_2((f_1, f_2) : \rho(f_1, |\det(\mathbf{C})|f_2 \circ g) < \epsilon \mid \mathbf{X}^{(n)}) \\
&\geq \Pi_2\left((f_1, f_2) : \rho(f_1, f^*) < {}^{\epsilon}/_{2}, \rho(f_2, f_g^*) < {}^{\epsilon}/_{2} \mid \mathbf{X}^{(n)}\right) \\
&\geq \Pi_2\left((f_1, f_2) : \rho(f_1, f^*) < {}^{\epsilon}/_{2} \mid \mathbf{X}^{(n)}\right) + \Pi_2\left((f_1, f_2) : \rho(f_2, f_g^*) < {}^{\epsilon}/_{2} \mid \mathbf{X}^{(n)}\right) - 1 \\
&= \Pi\left(f_1 : \rho(f_1, f^*) < {}^{\epsilon}/_{2} \mid \mathbf{X}^{(n)}\right) + \Pi\left(f_2 : \rho(f_2, f_g^*) < {}^{\epsilon}/_{2} \mid g(\mathbf{X})^{(n)}\right) - 1 \\
&\qquad \longrightarrow 1 + 1 - 1 = 1,
\end{aligned}$$

as $n \to \infty$. As a result, for $n \to \infty$,

$$\Pi_2((f_1, f_2) : \rho(f_1, |\det(\mathbf{C})|f_2 \circ g) < \varepsilon \mid \mathbf{X}^{(n)}) \longrightarrow 1. \qquad \square$$

2.5 Advances in partition estimation

A DPM model defined as in (2.2) implicitly defines a random clustering of the data, with two observations $\mathbf{X}_i$ and $\mathbf{X}_j$ belonging to the same cluster if and only if the corresponding parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ coincide. In recent years, particular attention has been dedicated to methods for defining point estimators in the space of partitions, given the output of a Markov chain Monte Carlo samplers (Dahl, 2006; Wade and Ghahramani, 2018; Rastelli and Friel, 2018). In this section we present the proposal of Wade and Ghahramani (2018) which will be exploited in the next sections.

Dealing with the estimation of a partition is a challenging problem: given a set $\mathbf{X}^{(n)}$, of n observations there are

$$\left\{ {n \atop k} \right\} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (k-j)^n$$

different ways to partition a set of $\mathbf{X}^{(n)}$ into k blocks, where $\left\{ {n \atop k} \right\}$ denotes a Stirling number of the second kind. Considering all the possible values for $k \in \{1, \dots, n\}$, there are

$$B_n = \sum_{k=1}^{n} \left\{ {n \atop k} \right\}$$

different ways to partition $\mathbf{X}^{(n)}$. Even for a small sample size n, the number B_n can get so large to make an exhaustive exploration of the space of all partitions an impossible task.

Let $\boldsymbol{\psi}_n$ be the random partition of $\mathbf{X}^{(n)}$ or, equivalently, a random partition of $\mathbb{N}_n = \{1, \dots, n\}$ (see Section 1.2), induced by a DPM model, and $p(\boldsymbol{\psi}_n \mid \mathbf{X}^{(n)})$ be its posterior distribution. Each time a realization from the posterior distribution of a DPM model is generated we obtain, as a by-product, also a realization $\boldsymbol{\psi}_n$ from its posterior distribution. Following Wade and Ghahramani (2018), we look for a point estimate of the partition which can be considered, in some sense, representative of the posterior distribution. From a decision theory point of view, one can consider a loss function $L(\boldsymbol{\psi}_n^*, \hat{\boldsymbol{\psi}}_n)$ which assesses the cost of estimating with $\hat{\boldsymbol{\psi}}_n$ an ideal true partition $\boldsymbol{\psi}_n^*$. Given that the true partition $\boldsymbol{\psi}_n^*$ is not available, we can average over all possible true partitions and consider the estimator

$$\overline{\boldsymbol{\psi}}_n = \arg\min_{\hat{\boldsymbol{\psi}}_n} \mathbb{E}\left[L(\boldsymbol{\psi}_n, \hat{\boldsymbol{\psi}}_n) \mid \mathbf{X}^{(n)}\right] = \arg\min_{\hat{\boldsymbol{\psi}}_n} \sum_{\boldsymbol{\psi}_n} L(\boldsymbol{\psi}_n, \hat{\boldsymbol{\psi}}_n) p(\boldsymbol{\psi}_n \mid \mathbf{X}^{(n)}). \tag{2.6}$$

A first simple choice for the loss function in equation (2.6) is the $0-1$ loss function, defined as $L_{0-1}(\boldsymbol{\psi}_n, \hat{\boldsymbol{\psi}}_n) = \mathbb{1}_{[\boldsymbol{\psi}_n \neq \hat{\boldsymbol{\psi}}_n]}$. Solving the problem described in equation (2.6) with L_{0-1} is equivalent to estimate the posterior mode of the partition probability distribution, that is

$$\overline{\boldsymbol{\psi}}_n = \arg\min_{\hat{\boldsymbol{\psi}}_n} \mathbb{E}\left[L_{0-1}(\boldsymbol{\psi}_n, \hat{\boldsymbol{\psi}}_n) \mid \mathbf{X}^{(n)}\right] = \arg\max_{\hat{\boldsymbol{\psi}}_n} p(\hat{\boldsymbol{\psi}}_n \mid \mathbf{X}^{(n)}).$$

The $0-1$ loss function is very simplistic as all the partitions which are different from the true partition are penalized in the same way. An important contribution in the study of more general loss functions was offered by Binder (1978). In the same paper Binder proposed the definition of a convenient loss function, hereafter named Binder loss function. Let k and $\hat{k}$ be the number of blocks in the partitions $\boldsymbol{\psi}_n$ and $\hat{\boldsymbol{\psi}}_n$, respectively. Moreover, we denote by n_{ij} the count of the observations which belong to the i-th block of $\boldsymbol{\psi}_n$ and to the j-th block of $\hat{\boldsymbol{\psi}}_n$. The Binder loss function is then defined as

$$L_B(\boldsymbol{\psi}_n, \hat{\boldsymbol{\psi}}_n) = \sum_{i<j} s_1 \mathbb{1}_{[\psi_i=\psi_j]} \mathbb{1}_{[\hat{\psi}_i \neq \hat{\psi}_j]} + s_2 \mathbb{1}_{[\psi_i \neq \psi_j]} \mathbb{1}_{[\hat{\psi}_i = \hat{\psi}_j]},$$

where s_1 and s_2 are penalizing weights. Setting $s_1 = s_2 = 1$ the Binder loss function becomes

$$L_B(\boldsymbol{\psi}_n, \hat{\boldsymbol{\psi}}_n) = \frac{1}{2}\left[\sum_{i=1}^{k} n_{i\cdot}^2 + \sum_{i=1}^{\hat{k}} n_{\cdot j}^2 - 2\sum_{i=1}^{k}\sum_{i=1}^{\hat{k}} n_{ij}^2\right],$$

with $n_{i\cdot} = \sum_{j=1}^{\hat{k}} n_{ij}$ and $n_{\cdot j} = \sum_{i=1}^{k} n_{ij}$. As investigated in Arbel et al. (2018a) with a simulation study, the partition estimated via Binder loss function tends to overestimate the number of blocks, when a subspace of the partition space larger than the one explored during the simulation is considered. This behaviour can be appreciated, for example, when the greedy search procedure of Wade and Ghahramani (2018), with the Binder loss function, is used.

Wade and Ghahramani (2018) propose to use a different loss function, based on the Variation of Information (VI), a function of the information, expressed in terms of entropy of the partitions $\boldsymbol{\psi}_n$ and $\hat{\boldsymbol{\psi}}_n$, and the information shared by $\boldsymbol{\psi}_n$ and $\hat{\boldsymbol{\psi}}_n$. The VI loss function was first introduced by in a paper of Meilă (2007), defined as

$$L_{VI}(\boldsymbol{\psi}_n, \hat{\boldsymbol{\psi}}_n) = \sum_{i=1}^{k} \frac{n_{i\cdot}}{n} \log_2\left(\frac{n_{i\cdot}}{n}\right) + \sum_{j=1}^{\hat{k}} \frac{n_{\cdot j}}{n} \log_2\left(\frac{n_{\cdot j}}{n}\right) - 2\sum_{i=1}^{k}\sum_{j=1}^{\hat{k}} \frac{n_{ij}}{n} \log_2\left(\frac{n_{ij}}{n}\right).$$

In the following sections we use the VI loss function and the relative estimated partition to assess a DPM-G model under affine transformations.

2.6 Simulation study

We performed a simulation study to provide empirical support to our results on the large n asymptotic robustness of a DPM-G model specified as in (2.2) with base measure (2.3), under affine transformations of the data. That is, given different affine transformations, we studied if the estimated densities and the related partitions become similar when the sample size grows, keeping the same specification for the prior model. We considered 15 different simulation scenarios. Specifically, we considered three different sample sizes, namely $n = 100$, $n = 300$ and $n = 1\,000$. Then, for each sample size, we generated a sample from a mixture of two Gaussian components, one being highly correlated and the other uncorrelated, defined as

$$\mathbf{X}^{(n)} \sim \frac{1}{2} N_2\left(\begin{bmatrix}-2\\-2\end{bmatrix}, \begin{bmatrix}1 & 0.85\\0.85 & 1\end{bmatrix}\right) + \frac{1}{2} N_2\left(\begin{bmatrix}2\\2\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right). \tag{2.7}$$

In order to test the robustness of the model under affine transformations of the data, we compressed or stretched the generated datasets by using five different constants, namely $c = 1/5$, $c = 1/2$, $c = 1$, $c = 2$ and $c = 5$. For each constant, we multiplied the simulated data by c, thus obtaining a transformed dataset $\mathbf{X}_c^{(n)} := c\mathbf{X}^{(n)}$. For each simulation scenario, namely $c \in \{1/5, 1/2, 1, 2, 5\}$, $n \in \{100, 300, 1\,000\}$, we generated 100 replicates. We then fitted a DPM-G model, specified as in (2.2) and (2.3), to each one of the 1 500 simulated datasets. In order to enhance the flexibility of the model, we completed its specification by setting a normal/inverse-Wishart prior distribution for the hyperparameters $(\mathbf{m}_0, \mathbf{B}_0)$ of the base measure (2.3). Namely, we set $\mathbf{B}_0 \sim IW(4, \text{diag}(\mathbf{15}))$ and $\mathbf{m}_0 \mid \mathbf{B}_0 \sim N(0, \mathbf{B}_0)$, specification chosen so that $\mathbb{E}[\boldsymbol{\mu}] = \mathbf{0}$ and to guarantee a prior guess on the location component $\boldsymbol{\mu}$ flat enough to cover the support of the non-transformed data. As for the scale component of the base measure (2.3), we set

$(\nu_0, \mathbf{S}_0) = (4, \text{diag}(\mathbf{1}))$. Finally, the mass parameter ϑ of the Dirichlet process was set equal to 1. See Appendix C for details on conjugate distributions.

Realisations of the mean of the posterior distribution were obtained by means of a Gibbs sampler relying on a Blackwell–McQueen Pólya urn scheme (see Müller et al., 1996), implemented in the `AFFINEpack` R package[1]. See Appendix B for details on the marginal approach for DPM models. For each replicate, posterior inference was drawn based on 5 000 iterations, obtained after discarding the first 2 500. Convergence of the chains was assessed by visually investigating traceplots referring to randomly selected replicates, which did not provide indication against it.

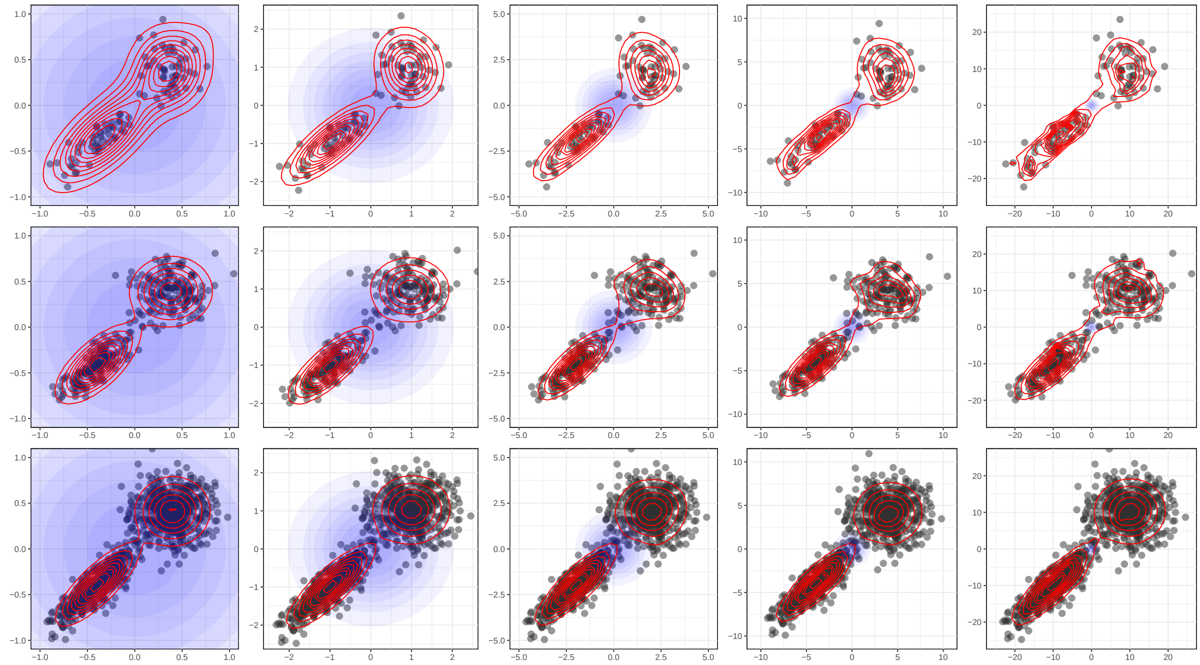

FIGURE 2.2: Simulation study. Based on a single replicate of the samples $\mathbf{X}^{(100)}$, $\mathbf{X}^{(300)}$ and $\mathbf{X}^{(1000)}$, scatter plot of the data (grey dots), contour plot of the estimated densities based on a DPM-G model (red curves) and contour plot for the expected prior density (blue filled curves). Left to right: rescaling constant $c = 1/5$, $c = 1/2$, $c = 1$, $c = 2$, $c = 5$. Top to bottom: sample size $n = 100$, $n = 300$, $n = 1\,000$.

Figure 2.2 shows, for every $n \in \{100, 300, 1\,000\}$ and $c \in \{1/5, 1/2, 1, 2, 5\}$, a contour plot of the estimated posterior densities. The difference between estimated densities, across different values of c, is apparent when $n = 100$, with the two extreme cases, namely $c = 1/5$ and $c = 5$, suggesting a different number of modes in the estimated density. For larger sample sizes, this difference is less evident and, when $n = 1\,000$, the contour plots are hardly distinguishable. These qualitative observations are in agreement with the large n asymptotic results of Theorem 7. The plots of Figure 2.2 refer to a single realisation of the samples $\mathbf{X}^{(100)}$, $\mathbf{X}^{(300)}$ and $\mathbf{X}^{(1000)}$ considered in the simulation study, although qualitatively similar results can be found in almost any replicate.

The findings drawn from a visual inspection of Figure 2.2 were confirmed by assessing the distance between estimated posterior densities. Specifically, for any considered sample size n and for any pair of values c_1 and c_2 taken by the constant c, we approximately evaluated the L^1 distance between the suitably rescaled estimated posterior densities obtained conditionally on

[1] The package is available at https://github.com/rcorradin/AFFINEpack and can be installed via devtools. For reproducibility, the code is available at https://github.com/rcorradin/Affine.

$\mathbf{X}_{c_1}^{(n)}$ and on $\mathbf{X}_{c_2}^{(n)}$. The results of such analysis are shown in Figure 2.3 and indicate that as the sample size grows, the difference in terms of L^1 distance strictly decreases.

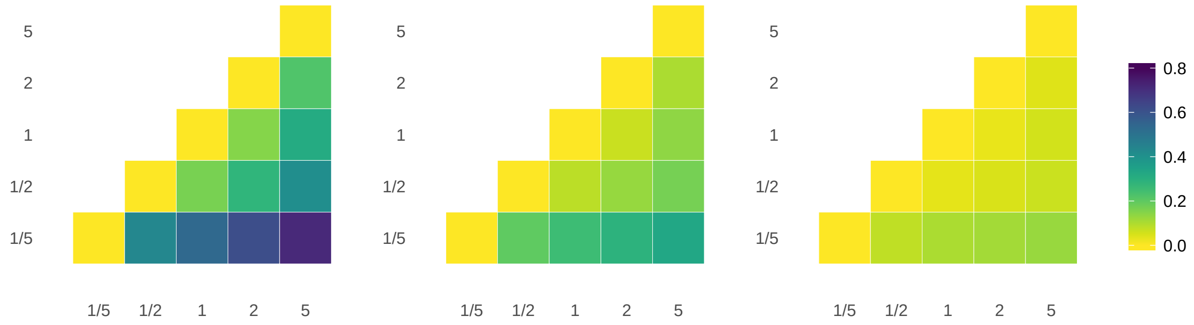

FIGURE 2.3: Simulation study. L^1 distance between suitably rescaled estimated densities after data transformations for different constants c_1 (X axis) and c_2 (Y axis), averaged over 100 replications. Left to right: sample size $n = 100$, $n = 300$ and $n = 1000$.

The posterior distribution of the random density induced by a DPM-G model provides interesting insight also on the clustering structure of the data. The second goal of the simulation study, thus, consisted in investigating the impact of the scaling factor c on the estimated number of groups in the partition induced on the data. To this end, for each considered n and c, we estimated $\hat{K}_n^{(\mathrm{VI})}$, the number of groups in the optimal partition estimated using a procedure introduced by Wade and Ghahramani, 2018 and based on the variation of information loss function. The average values for this quantity, over 100 replicates, are reported in Table 2.1. There appears to be a clear trend suggesting that a larger scaling constant c leads to a larger $\hat{K}_n^{(\mathrm{VI})}$: this finding is consistent with the fact that, if the data are stretched while the prior specification is kept unchanged, then we expect the estimated posterior density to need a larger number of Gaussian components to cover the support of the sample. For the purpose of this simulation study the main quantity of interest is the ratio between the estimated number of groups under any two distinct values c_1 and c_2 for the scaling constant c, that is $\hat{K}_{n,c_1}^{(\mathrm{VI})}/\hat{K}_{n,c_2}^{(\mathrm{VI})}$. The results presented in Table 2.1 clearly indicate that, as the sample size n becomes large, such ratios tend to approach 1. This suggests that the large n robustness property of the DPM-G model nicely translates to an equivalent notion of robustness in terms of the estimated number of groups $\hat{K}_n^{(\mathrm{VI})}$ in the data.

	$c = 1/5$	$c = 1/2$	$c = 1$	$c = 2$	$c = 5$
$n = 100$	1.81	2.04	2.84	5.96	10.52
$n = 300$	2.00	2.03	2.20	2.82	5.18
$n = 1000$	2.00	2.00	2.04	2.05	2.12

TABLE 2.1: Simulation study. Averages over 100 replicates for $\hat{K}_n^{(\mathrm{VI})}$, the number of clusters of the estimated partition estimated by means of Wade and Ghahramani (2018)'s variation of information method. Left to right: rescaling constant $c = 1/5$, $c = 1/2$, $c = 1$, $c = 2$, $c = 5$. Top to bottom: sample size $n = 100$, $n = 300$, $n = 1\,000$.

2.7 Application to the NGC 2419 globular cluster data

In this section we address an interesting astronomical classification problem, with data consisting of multiviariate observations with incommensurable components, and for which the robustness of DPM-G models, investigated in Sections 2.4 and 2.6 turns out to be useful. Specifically, we consider a dataset consisting of measurements on a set of 139 stars, possibly belonging to a globular cluster called NGC 2419 (see Ibata et al., 2011, for details on the globular cluster NGC 2419). Globular clusters are sets of stars orbiting some galactic center. The NGC 2419, showed in Figure 2.4, is one of the furthest known globular clusters in the Milky Way. For each star we observe a four-dimensional vector $(Y_1, Y_2, V, [\mathrm{Fe/H}])$, where (Y_1, Y_2) is a two-dimensional projection on the plane of the sky of the position of the star, V is its line of sight velocity and $[\mathrm{Fe/H}]$ its metallicity, a measure of the abundance of iron relative to hydrogen. Out of these four components, only Y_1 and Y_2 are measured in the same physical unit, while dimensional constants need to be specified in order to relate position, velocity and metallicity. A key question arising with these data consists in identifying the stars that, among the 139 observed, can be rightfully considered as belonging to NGC 2419: a correct classification would be pivotal in the study of the globular cluster dynamics. Astronomers expect the large majority of the observed stars to belong to the cluster: the remaining ones, called field stars or contaminants, are Milky Way stars, unrelated to the cluster, that happen to appear projected in the same region of the plane of the sky. In general the contaminants have different kinematic and chemical properties with respect to the cluster members. Considering the nature of the problem, this research question can be formalised as an unsupervised classification problem, the goal being the identification of the stars which belong to the largest cluster, which can be interpreted as the NGC 2419 globular cluster. Admittedly, the terms of such a classification problem are not limited to the considered dataset but, on the contrary, are ubiquitous in astronomy and, more in general, might arise in any field where data components are incommensurable.

FIGURE 2.4: An image of the remote Milky Way globular cluster NGC 2419 (about 300 000 light years away from the solar system). Picture by Bob Franke, with permission (www.bf-astro.com).

We fitted the DPM-G model, specified as in (2.2) and with base measure (2.3), to the NGC 2419 dataset described in Section 1. The ultimate goal of our analysis consists in classifying, by the use of an opportune density estimation, stars as belonging to the NGC 2419 globular cluster or as being contaminants: an accurate classification is crucial for the astronomers to study the dynamics of the globular cluster. Since the large majority of the stars in the dataset is expected to belong to the globular cluster, with only a few of them being contaminants, we will identify the globular cluster as the largest group in the estimated partition of the dataset.

Prior to any analysis, data were standardized component by component, the legitimacy of such procedure following from the robustness results of Theorem 7, then the estimated density were transformed back to the original scale. Hyperprior distributions were specified for the location parameter of the base measure (2.3) and on the DP mass parameter ϑ. Specifically, $\mathbf{B}_0 \sim IW(6, \mathrm{diag}(\mathbf{15}))$ and $\mathbf{m}_0 \mid \mathbf{B}_0 \sim N(0, \mathbf{B}_0)$, specification chosen to guarantee a prior guess on the location component $\boldsymbol{\mu}$ flat enough to cover the support of the data and centered at $\mathbf{0}$. In addition, the precision parameter ϑ was given a gamma prior distribution with parameters specified so to reflect the prior opinion of astronomers who would expect two distinct groups of stars in the dataset. Let $\tilde{p}$ be distributed as a DP with mass parameter θ. Following Pitman (2006), the expected number of unique values in a exchangeable sample of size n, distributed according to $\tilde{p}$, is, a priori, equal to

$$\mathbb{E}[K_n] = \sum_{i}^{n} \frac{\theta}{i - 1 + \theta}.$$

Here we assigned θ a hyperprior, specifically $\theta \sim Gamma(a_\theta, b_\theta)$, and we set $a_\theta = 1$ and $b_\theta = 5.26$, values chosen so that $\mathbb{E}[K_n] \simeq 2$. Finally, as far as the scale component of the base measure (2.3) is concerned, we set $(\nu_0, \mathbf{S}_0) = (26, \mathrm{diag}(\mathbf{21}))$, where the number of degrees of freedom $\nu_0 = 26$ of the inverse-Wishart distribution was chosen to guarantee the conditions of Theorem 7 and, in turn, the scale matrix $\mathbf{S}_0 = \mathrm{diag}(\mathbf{21})$ so that $\mathbb{E}[\boldsymbol{\Sigma}] = \mathrm{diag}(\mathbf{1})$. See Appendix C for details on the choice of conjugate distributions. Realisations of the mean of the posterior distribution were obtained by means of a Gibbs sampler relying on a Blackwell–McQueen Pólya urn scheme[2]. See Appendix B for details on the marginal approach for DPM models.

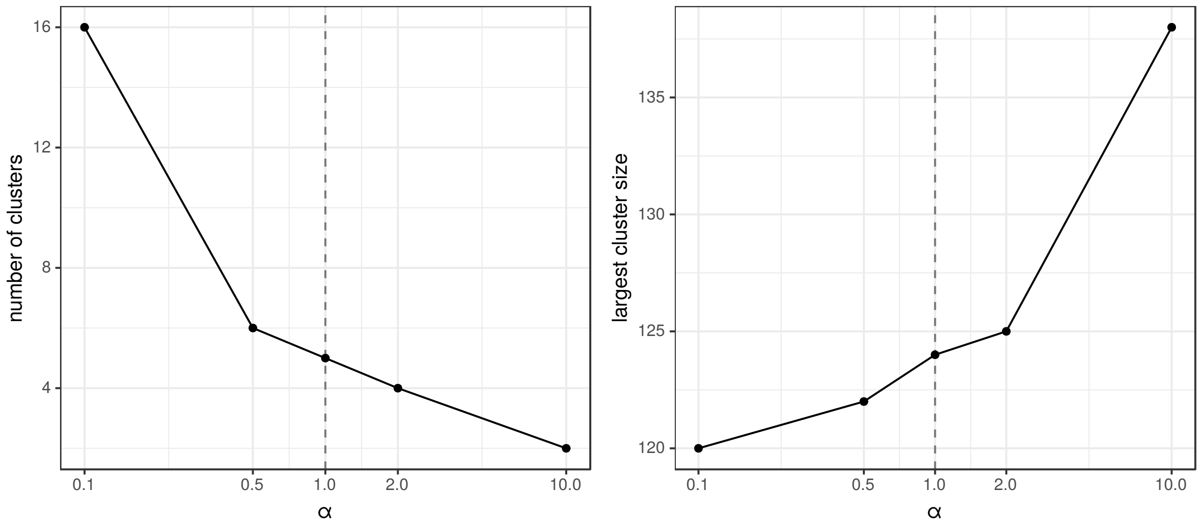

FIGURE 2.5: Analysis of the sensitivity to the values taken by the parameter α controlling the scale component in the base measure. Left plot: number of clusters. Right plot: size of the largest cluster. Optimal partitions were estimated by applying Wade and Ghahramani (2018)'s variation of information method.

[2] See footnote 1.

A difficult task arising when specifying a DPM-G model refers to the choice of appropriate values for the hyperparameters appearing in the base measure, as they might have an impact on posterior estimates, both in terms of density and clustering. While it was possible to set hyperprior distributions for the parameters of the location component, preserving the asymptotic result described in Theorem 7, the same could not be done for the scale component for which an arbitrary specification was chosen. Here we investigate the effect of such choice by performing a sensitivity analysis to evaluate the effect of the previous specification on the estimated clustering of the data. To this end we modified the model specification so that the expected value of the scale component, with $\nu_0 = 26$ and four dimensions, is equal to $\mathbb{E}[\Sigma] = \mathrm{diag}(\alpha)$. Then we set a grid of values $\alpha \in \{0.1, 0.2, 1, 5, 10\}$ and studied the effect of α on the optimal partition estimated by applying Wade and Ghahramani (2018)'s variation of information method, with a focus on the number of clusters and the size of the largest one. The effect of α on the estimated clustering shows a clear trend: as α grows, and thus the prior expectation of the scale component takes a large value, the number of clusters gets small and the size of the largest cluster increases; on the other hand, when α decreases, the model identifies more components with the largest being less populated. While any sensible statistical analysis should take this sensitivity into account, henceforth, for the sake of illustration, we set $\alpha = 1$, which corresponds to assuming $\mathbf{S}_0 = \mathrm{diag}(\mathbf{21})$.

In turn, posterior inference was drawn based on 20 000 iterations, after a burn-in period of 5 000 iterations. Convergence of the chains was assessed by visually investigating traceplots, which did not provide indication against it.

Figure 2.6 displays contour plots for the six two-dimensional projections of the estimated posterior density, while Figure 2.7 shows the scatter plots of the dataset with individual observations coloured according to their membership in the partition estimated based on the variation of information loss function (Wade and Ghahramani, 2018) and labeled as main group (grey circles) and other groups (coloured triangles). The estimated partition is composed of five groups. The largest one, identified as the globular cluster, consists of 124 stars. The remaining 15 stars are thus considered contaminants and are further divided into four groups, one composed by eight stars (group A), one containing five stars (group B) and two singletons (groups C and D). A visual investigation of Figure 2.7 suggests that stars in group A differ from those in the globular cluster in terms of metallicity and position, with the contaminants characterised by larger values for [Fe/H] and smaller values for Y_1 and Y_2. The stars in group B differ from the globular cluster in terms of velocity and metallicity, with the contaminants showing larger values for V and [Fe/H]). Finally, groups C and D are singletons, the first one being characterised by a high metallicity and an extremely small value for the velocity, the second one showing large values for both metallicity and location Y_1.

Our unsupervised statistical clustering can be compared to the clustering of Ibata et al. (2011) (described in their Figure 4) based on ad hoc physical considerations. Specifically, once the best fitting physical model, in the class of either Newtonian or Modified Newtonian Dynamics models, is detected, they use it in order to compute the average values of the physical variables describing the stars. Stars are then assigned to the globular cluster based on a comparison between their velocity and the average model velocity: those lying close enough are deemed to belong to the cluster, while the others are considered as potential contaminants. For the latter, the evidence of being contaminants is measured by evaluating how distant their metallicity is from the average model one. Two classifications are then proposed: the first one assigns to the globular cluster only the 118 stars for which the evidence seems strong, the second and less conservative strategy classifies as belonging to the globular cluster a total of 130 stars. Following this distinction and for the sake of simplicity, we summarise the results of Ibata et al. (2011)'s analysis, by devising three groups of stars:

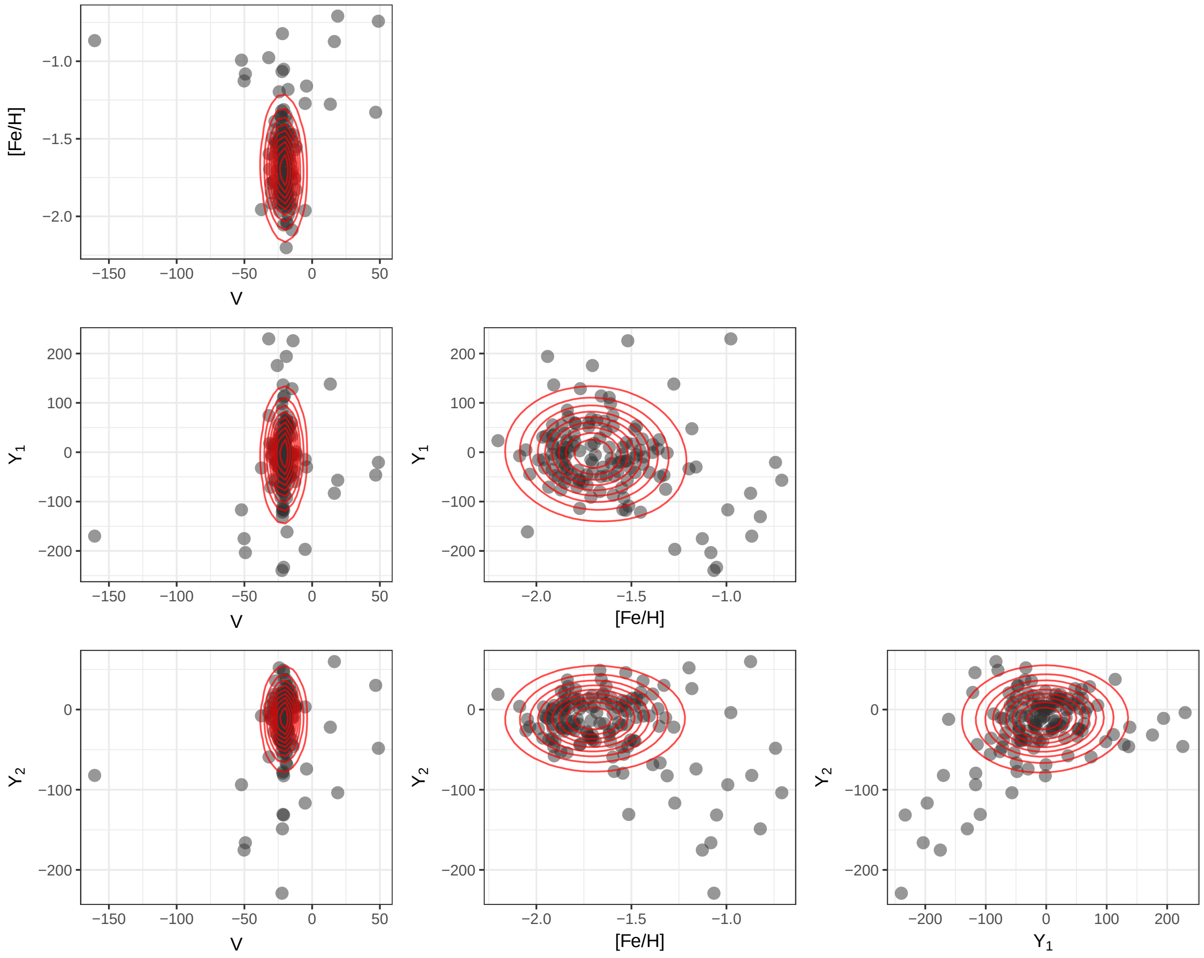

FIGURE 2.6: NGC 2419 data. Contour plots of the bivariate marginal densities estimated via DPM-G model.

- *globular cluster*: 118 stars deemed to belong to the globular cluster,
- *likely globular cluster*: 12 stars assigned to the globular cluster only when the less conservative procedure is adopted,
- *contaminants*: 9 stars with strong evidence of being contaminants.

For the purpose of comparison, we report in Table 2.2 the confusion matrix of the groups obtained via the DPG-G model against the groups detected by Ibata et al. All of the 124 stars belonging to the largest group of the partition estimated based on the DPM-G model belong to the groups identified as globular cluster or likely globular cluster by Ibata et al. At the same time, out of the nine stars classified as contaminants by Ibata et al. , the approach based on the DPM-G model assigns none to the globular cluster, three to group A, five stars to group B, which is composed only by stars considered contaminants in Ibata et al. , and the star of group C, which shows an extremely small value for the velocity variable. Finally, the group D contains only one star, which is not consider a contaminant in Ibata et al.

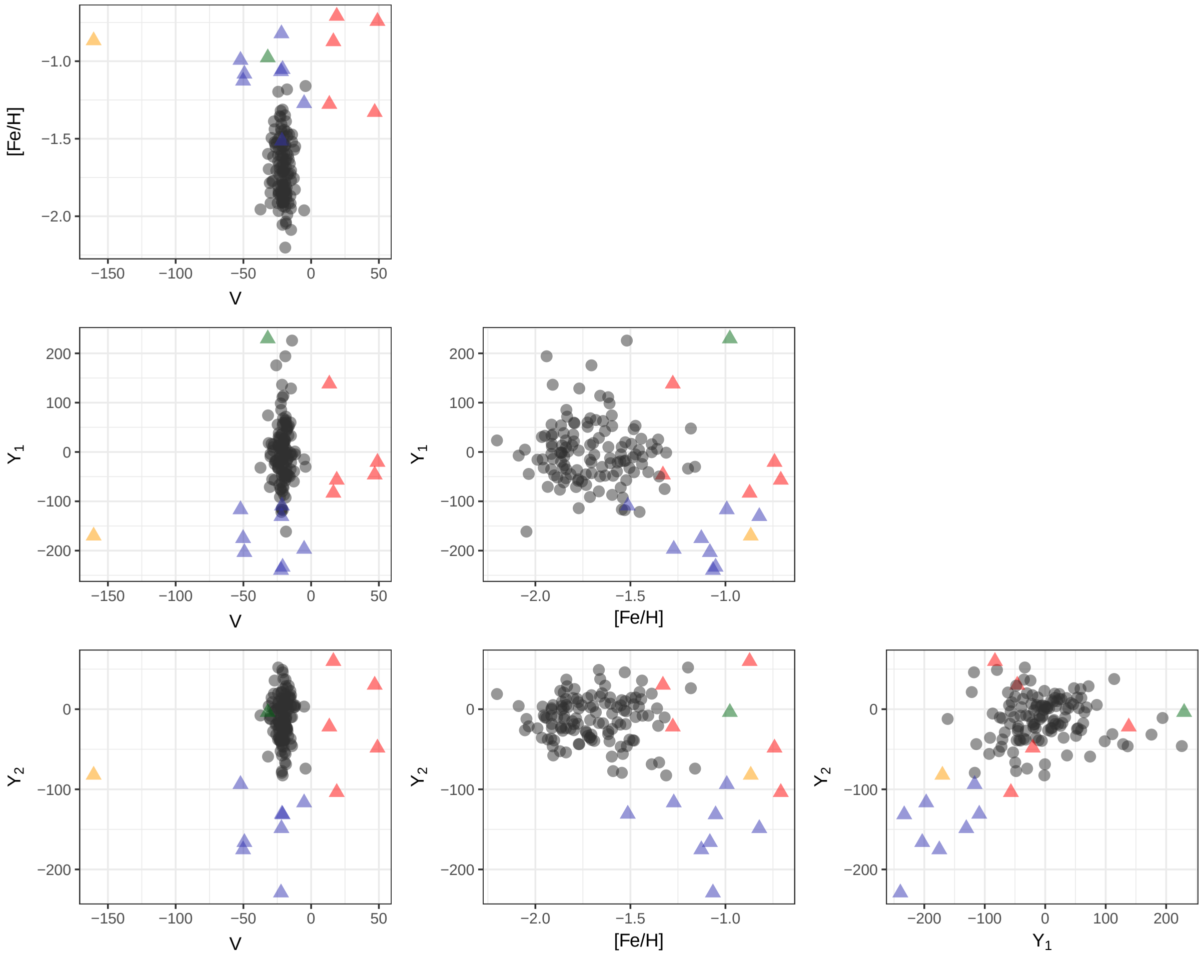

FIGURE 2.7: NGC 2419 data. Partition estimated via DPM-G models combined with Wade and Ghahramani (2018)'s variation of information method. Five groups are detected: the largest group (grey dots), group A (blue triangles), group B (red triangles), group C (one orange triangle), group D (one green triangle).

2.8 Conclusions

The purpose of this chapter was to investigate the behaviour of the multivariate DPM-G model when affine transformations are applied to the data. To this end we focused on the DPM-G model with independent normal and inverse-Wishart specification for the base measure.

Our investigation covered both the finite sample size and the asymptotic setting. Specifically, in Proposition 5, given any affine transformation g, an explicit model specification, depending on g, was derived so to ensure coherence between posterior inferences carried out based on a dataset or its transformation via g. We then considered a different setting where the specification of the model is assumed independent of the specific transformation g. In this case, we formalised the notion of asymptotic robustness of a model under transformations of the data and showed that mild conditions on the true data generating distributions are sufficient to ensure that the DPM-G model features such a property. Specifically, Theorem 7 shows that the posterior distributions obtained conditionally on a dataset or any affine transformation of it, become more and more similar as the sample size grows. Inference on densities and, as a by-product, on the clustering structure underlying the data, thus becomes increasingly less dependent on the affine transformation applied to the data, as the sample size grows to infinity.

			DPM-G groups				
			largest	*A*	*B*	*C*	*D*
		total	*124*	*8*	*5*	*1*	*1*
Ibata et al. groups	*globular cluster*	*118*	114	4	0	0	0
	likely globular cluster	*12*	10	1	0	0	1
	contaminants	*9*	0	3	5	1	0

TABLE 2.2: NGC 2419 data. Comparison between the groups identified by Ibata et al. (2011) and the groups estimated via DPM-G model.

As a special case, Theorem 7 implies that posterior inference based DPM-G models is asymptotically robust to data transformations commonly adopted for the sake of numerical efficiency, such as standardisation or normalisation. This observation is particularly relevant when dealing with the astronomical unsupervised clustering problem motivating this work.

Due to the lack of prior information on the dimensional constants relating different physical units, we resorted to a standardisation of each component of the data and chose an arbitrary model specification. Prior information was available in the form of the experts' prior opinion on the expected number of groups in the dataset and was used to elicit the hyperprior distribution for ϑ, the total mass parameter of the DP.

Chapter 3

Importance Conditional Sampler

A big computer, a complex algorithm
and a long time does not equal science.

Robert Gentleman, *Statistician*

Canale, A., Corradin, R., Nipoti, B.
Importance conditional sampler for Bayesian nonparametric mixtures.
In preparation.

Bayesian nonparametric mixtures have a central role in Bayesian nonparametric modelling. The first proposal of such models was the Dirichlet process (DP, Ferguson, 1973, see Section 1.5) mixture of Gaussian kernels, by Lo (1984), contribution which paved the way to the definition of a wide variety of nonparametric mixture models. Specifically, in recent years, increasing interest has been dedicated to the definition of mixture models based on nonparametric mixing measures more general than the DP (e.g. Nieto-Barajas et al., 2004; Lijoi et al., 2005b; Lijoi et al., 2005a; Lijoi et al., 2007c). Among these measures, the Pitman-Yor process (PY, Perman et al., 1992; Pitman, 1995, see Section 1.6) stands out for conveniently combining mathematical tractability, interpretability and modelling flexibility (see De Blasi et al., 2015).
Dealing with nonparametric mixtures involves the use of a random probability measure $\tilde{p}$ of infinite dimension (see Section 1.4 for details). Markov chain Monte Carlo (MCMC) sampling methods represent the gold standard for carrying out posterior inference based on nonparametric mixture models. Resorting to the terminology adopted by Papaspiliopoulos and Roberts (2008), existing MCMC sampling methods can be classified into marginal and conditional methods, the two classes being characterized by different ways to deal with the infinite-dimensional random probability measure $\tilde{p}$. While the marginal approach relies on the possibility of analytically marginalizing $\tilde{p}$ out, conditional methods work with finite dimensional summaries of $\tilde{p}$.
In this chapter we propose a novel algorithm for PY mixture models, named Importance Conditional Sampler, with the goal of combining the appealing features of conditional and marginal methods. Namely, we will show that, like existing marginal methods, the ICS *i)* has a simple and interpretable sampling scheme, reminiscent of the Blackwell-McQueen Pólya urn (Blackwell and MacQueen, 1973), and *ii)* requires the update of a bounded number of random elements per iteration of the algorithm. At the same time, being a conditional method, the ICS algorithm is such that *iii)* the step for allocating observations to different clusters is fully parallelizable, and *iv)* posterior uncertainty can be easily quantified. Our proposal exploits a convenient posterior representation of the PY process, proposed in Pitman (1996), combined with

an importance sampling idea, similar in spirit to the augmentation step of Algorithm 8 of Neal (2000). The name importance conditional sampler was thus chosen to reflect the conditional nature of the scheme and to stress the innovative use of importance sampling in this framework. The scheme characterizing the ICS can be naturally extended so to deal with partially exchangeable data, its simplicity and efficiency being preserved. To this end, we use the ICS as the building block of a new conditional algorithm for the class of GM-dependent DP mixture models (GM-DDP, see Lijoi et al., 2014a; Lijoi et al., 2014b; Griffin et al., 2013), for partially exchangeable data.

The rest of the chapter is ogranized as follows. In Section 3.1 we introduce the modelling framework and the notation used in the rest of the chapter. Section 3.2 introduces to the state-of-the-art of MCMC algorithms for PY mixture models and highlights some limitations of existing methods. Motivated by these, in Section 3.3, we introduce and illustrate our proposal. In Section 3.4 we provide the details of the implementation used for the competing algorithms that were considered in the simulation study presented in Section 3.5. In Section 3.6 we describe how the ICS scheme can be conveniently used to devise an efficient algorithm for GM-DDP mixture models. Finally, Section 3.7 is dedicated to some concluding remarks.

3.1 Modelling framework

Let $\mathbf{X}^{(n)} := (X_1, \dots, X_n)$ be a sample of size n of observations defined on some probability space $(\Omega, \mathscr{A}, \mathbb{P})$ and taking values in $\mathbb{X}$. Let $\mathscr{F}$ denote the space of all probability distributions on $\mathbb{X}$, then a Bayesian nonparametric mixture model is a random distribution taking values in $\mathscr{F}$, defined as

$$\tilde{f}(x) = \int_{\Theta} k(x;\theta) \mathrm{d}\tilde{p}(\theta), \tag{3.1}$$

where $k(x;\theta)$ is a kernel function defined on $\mathbb{X} \times \Theta$ and $\tilde{p}$ is a random mixing distribution. In this chapter we focus on $\tilde{p} \sim PY(\sigma, \vartheta; P_0)$, that is we assume that $\tilde{p}$ is a random probability measure distributed as a PY process with discount parameter $\sigma \in [0, 1)$, total strength parameter $\vartheta > -\sigma$, and diffuse base measure $P_0 \in \mathscr{F}$. The DP is obtained as a special case when $\sigma = 0$. Model (3.1) can be written in hierarchical form as

$$\begin{aligned} X_i \mid \theta_i &\overset{\text{ind}}{\sim} k(x_i;\theta_i) \qquad i = 1, \dots, n \\ \theta_i \mid \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p} \\ \tilde{p} &\sim Q, \end{aligned} \tag{3.2}$$

where Q denotes the distribution of $\tilde{p}$, known as de Finetti measure (see Section 1.1 for details). The joint distribution of $\boldsymbol{\theta} := (\theta_1, \dots, \theta_n)$ is characterized by the predictive distribution of the the PY, which, for any $i = 1, 2, \dots$, is given by

$$P[\theta_{i+1} \in \cdot \mid \theta_1, \dots, \theta_i] = \frac{\vartheta + k_i\sigma}{\vartheta + i} P_0(\cdot) + \sum_{j=1}^{k_i} \frac{n_j - \sigma}{\vartheta + i} \delta_{\theta_j^*}(\cdot), \tag{3.3}$$

where k_i is the number of distinct values θ_j^* appearing in the first i draws, and n_j, such that $\sum_{j=1}^{k_i} n_j = i$, is the number of θ_l, for $l \in \{1, \dots, i\}$, which coincide with θ_j^*.

Alternatively, the distribution of $\tilde{p}$ can be described by its stick-breaking representation, which, for the PY, was given by Perman et al. (1992). Specifically, $\tilde{p}$ can be thought of as an infinite sum of random jumps $\{W_j\}_{j=1}^{\infty}$ occurring at random locations $\{\tilde{\theta}_j\}_{j=1}^{\infty}$, that is

$$\tilde{p} = \sum_{j=1}^{\infty} W_j \delta_{\tilde{\theta}_j}, \tag{3.4}$$

where the distribution of the locations is independent of that one of the jumps and, while the $\tilde{\theta}_j$'s are independent and identically distributed from P_0, the distribution of the jumps is characterized by

$$W_1 = V_1, \quad W_j = V_j \prod_{l=1}^{j-1}(1 - V_l), \quad V_j \overset{\text{ind}}{\sim} \text{Beta}(1-\sigma, \vartheta + j\sigma). \tag{3.5}$$

The sequence $\{W_j\}_{j\geq 1}$ defined in (3.5) follows a two parameter GEM distribution, see Section 1.6 for more details.

3.2 State of the art

Marginal methods for nonparametric mixtures were first used in Escobar (1988) and Escobar and West (1995), contributions which focused on DP mixtures (DPM) of univariate Gaussian kernels. Extensions of such proposal include the works of Müller et al. (1996), Maceachern (1994), Maceachern and Müller (1998), Neal (2000) and Favaro and Teh (2013). It is worth noting that, while being the first class of MCMC methods for Bayesian nonparametric mixtures appeared in the literature, marginal methods are still routinely used in popular packages such as the `DPpackage` (Jara et al., 2011) library of `R`, the *de facto* software for many Bayesian nonparametric models.

Alternatively, conditional methods rely on the use of summaries, of finite and possibly random dimension, of realizations of $\tilde{p}$ from its stick-breaking representation. A first type of conditional approach can be found in Ishwaran and Zarepour (2000) and Ishwaran and James (2001), where a fixed truncation of the stick-breaking representation of a large class of random probability measures is considered and a bound for the introduced truncation error is provided. Muliere and Tardella (1998) (for the DP) and Arbel et al. (2018b) (for the PY) make the truncation level random so to make sure that the resulting error is smaller than a given threshold. On similar lines, Argiento et al. (2016a) and Argiento et al. (2016b) propose a random truncation for mixture models based, respectively, on the normalized generalized gamma process and, more in general, on normalized completely random measures (see Appendix A for details on completely random measures).

Exact solutions which avoid introducing truncation errors are the slice sampler of Walker (2007) and Kalli et al. (2011) and the retrospective sampler of Papaspiliopoulos and Roberts (2008). Although originally introduced for the case of DPM models, the schemes of slice sampler and retrospective sampler are naturally extended to a more general class of models admitting a stick-breaking representation, and thus, in particular, to the class of PY mixture (PYM) models. Finally, it is worth mentioning the Ferguson and Klass (1972) algorithm, a conditional method which can be conveniently used for normalized random measure mixture models (Barrios et al., 2013). Such a class of methods cannot be directly adopted for the PYM models and, nonetheless, relies on a truncation of the involved random measure (see Arbel and Prünster, 2017).

Marginal methods are appealing for their simplicity and for the fact that the number of random elements that must be drawn at each iteration of the sampler, i.e. the set of parameters $\boldsymbol{\theta}^{(n)}$, is fixed. Conditional methods advantageously exploit the conditional independence of the parameters θ_i, given $\tilde{p}$ (or a finite summary of it), and thus lead to algorithms where the update step of $\boldsymbol{\theta}^{(n)}$ is parallelizable. At the same time, the random truncation at the core of conditional

methods such as slice sampler and retrospective sampler, makes the number of random elements that must be drawn at each iteration of the algorithm, random and unbounded. While this last feature turns out not to be a problem when the DPM case is considered, the same makes the use of these algorithms for the more general case of PYM models problematic, specially when large values for the discount parameter are considered. Finally, it is important to observe that working without integrating $\tilde{p}$ out makes posterior uncertainty quantification much more natural in conditional methods, as they allow to sample trajectories of the posterior distribution, unlike marginal methods which produce realizations of its mean (see Arbel et al., 2016, for a discussion).
Next, we briefly describe the rationale of three algorithms which, among the ones mentioned, avoid introducing structural truncation errors, namely the marginal sampler, the slice sampler and the retrospective sampler. As for the marginal sampler and the slice sampler, more details on their implementation will be given in Section 3.4.

Marginal sampler

Inspired by the seminal work of Blackwell and MacQueen (1973), Escobar and West (1995) proposed a marginal algorithm for DPM models, where the mixing DP is integrated out. The same approach can be followed for the case of PYM models: after marginalizing with respect to the random probability measure $\tilde{p}$ in (3.2), a Gibbs sampler can be devised such that, in turn, every element of $\boldsymbol{\theta}^{(n)}$ is updated from its full conditional distribution

$$P[\theta_i \in \mathrm{d}t \mid \boldsymbol{\theta}_{(i)}, \mathbf{X}^{(n)}] \propto \frac{\vartheta + k_{(i)}\sigma}{\vartheta + n - 1} \int k(X_i, t) P_0(\mathrm{d}t) + \sum_{j=1}^{k_{(i)}} \frac{n_{(i),j}}{\vartheta + n - 1} k(X_i, \theta^*_{(i),j}) \delta_{\theta^*_{(i),j}}(\mathrm{d}t) \tag{3.6}$$

where $\boldsymbol{\theta}_{(i)} := (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n)$, $k_{(i)}$ is the number of distinct values in $\boldsymbol{\theta}_{(i)}$, $\theta^*_{(i),j}$ the j-th of such values and $n_{(i),j}$ its frequency. It is apparent that the structure of the full conditional (3.6) is reminiscent of (3.3).

□

Slice sampler

The core idea of the slice sampler of Walker (2007) and Kalli et al. (2011) consists in introducing an augmenting uniform random variable U, such that, conditionally on U, the number of jumps of $\tilde{p}$ that must be sampled is finite. The model in (3.1) can be written as

$$\tilde{f}(x) = \sum_{j=1}^{\infty} W_j k(x; \theta_j), \tag{3.7}$$

where the distribution of the W_j's is described by the stick-breaking construction in (3.5). Then, following Walker (2007), we consider a uniform random variable U and the joint density

$$\tilde{f}(x, u) = \sum_{j=1}^{\infty} \mathbb{I}_{[u < W_j]} k(x; \theta_j), \tag{3.8}$$

where U has the effect of *slicing* the distribution of the W_j's. It is apparent that, by marginalizing Equation 3.8 with respect to U, Equation 3.7 is recovered. Starting from the previous joint distribution, it is possible to devise a Gibbs-sampler involving the update of the augmenting

	$\sigma = 0$	$\sigma = 0.2$	$\sigma = 0.4$	$\sigma = 0.6$	$\sigma = 0.8$
$\vartheta = 0.1$	1.39	4.19	62.90	11497.19	94079.99
$\vartheta = 1$	6.36	20.77	168.13	19459.96	98391.76
$\vartheta = 10$	55.40	159.99	1371.14	43874.20	$> 10^5$
$\vartheta = 100$	542.60	1425.41	11921.01	86810.41	$> 10^5$

TABLE 3.1: Slice sampler: average number of jumps (out of 100 replicates) of the prior process that must be sampled in one iteration, for different specification of the strength parameter ϑ and the discount parameter σ.

	$\sigma = 0$	$\sigma = 0.2$	$\sigma = 0.4$	$\sigma = 0.6$	$\sigma = 0.8$
$\vartheta = 0.1$	1.42	4.44	60.76	9054.75	95545.97
$\vartheta = 1$	6.07	18.37	147.60	19309.82	99564.28
$\vartheta = 10$	53.96	146.54	2052.05	47614.75	$> 10^5$
$\vartheta = 100$	516.17	1459.22	11686.69	93930.31	$> 10^5$

TABLE 3.2: Retrospective sampler: average number of jumps (out of 100 replicates) of the prior process that must be sampled in one iteration, for different specification of the strength parameter ϑ and the discount parameter σ.

random variable. The convenience of the augmented model (3.8) lays in the fact that, conditionally to the value of U, the number of jumps of $\tilde{p}$ that must be sampled is random but finite.

□

Retrospective sampler

The retrospective sampler can be thought of as an adaptation of the well-known inverse cumulative distribution function method for sampling from the posterior distribution of non-parametric mixtures. Given the sequences $\{W_j\}_{j\geq 1}$ and $\{\tilde{\theta}_j\}_{j\geq 1}$ defining the RPM in (3.4), a realization from $\tilde{p}$ can be obtained by first we generating a uniformly distributed U, and then selecting the k-th value of $\{\tilde{\theta}_j\}_{j\geq 1}$ such that

$$\sum_{j=0}^{k-1} W_j < u \leq \sum_{j=1}^{k} W_j, \tag{3.9}$$

with the proviso $W_0 = 0$. Retrospective sampling simply exchanges the order of simulation between U and the pairs $\{W_j, \tilde{\theta}_j\}_{j\geq 1}$. Conditionally on a value sampled for U, if we need more W_j's than we currently have already sampled, we go back *retrospectively* until we have enough W_j's to satisfy the condition (3.9). Based on this idea, Papaspiliopoulos and Roberts (2008) derived an exact sampling scheme for the posterior distribution of a DPM model, which is naturally extended to the PYM model case.

□

While working more efficiently for the DPM model, both the slice sampler and the retrospective sampler might face serious computational issues when used to fit a PYM model with discount parameter σ deviating from 0. For both the algorithms, the number of jumps that must be sampled at each iteration can be so large that a practical implementation of the algorithms is not feasible. We empirically investigate this behaviour, by studying the number of jumps required, in the prior process, for different values of the strength parameter ϑ and the discount parameter

σ, with a fixed sample size $n = 100$. As shown in Tables 3.1 and 3.2, as the value of σ becomes large, the required number of jumps grows, exploding when $\sigma \geq 0.6$. Such behaviour can be understood by considering the stick-breaking representation of the PY process, described in Equation 3.5. Large values of the parameter σ make small jumps more likely to be sampled. As a result, a larger number of jumps will tend to be needed to satisfy the probabilistic conditions required by both the slice and the retrospective sampler. It is important to stress that, while the implementation of slice sampler and retrospective sampler for PYM is troublesome already for $n = 100$, the described issue is exacerbated for larger values of n. A similar behaviour of the PY has been recently highlighted in Arbel et al. (2018b), when studying the accuracy of the approximation obtained by truncating its stick-breaking representation.
Finally, the very similar results displayed in Table 3.2 and Table 3.1 are not surprising as the number of required jumps has the same distribution for the two algorithms. Indeed, it is easy to show that in both cases the number of jumps required *a priori* is given by

$$J^* = \min\left\{ J \geq 1 \,\text{s.t.} \sum_{j=1}^{J} W_j > B_n \right\}, \tag{3.10}$$

where B_n is a Beta random variable with parameters 1 and n. Due to the similar behaviour we decided to compare our proposal in the next sections only with one conditional sampler, the slice sampler.

3.3 The Importance Conditional Sampler

The random elements involved in a PYM model defined as in (3.2) are observations $\mathbf{X}^{(n)}$, parameters $\boldsymbol{\theta}$ and the PY random probability measure $\tilde{p}$. The joint distribution of $(\mathbf{X}^{(n)}, \boldsymbol{\theta}, \tilde{p})$ can be written as

$$p(\mathbf{X}^{(n)}, \boldsymbol{\theta}, \tilde{p}) = p(\mathbf{X}^{(n)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \tilde{p}) Q(\tilde{p}) = \prod_{i=1}^{n} k(X_i; \theta_i) \prod_{j=1}^{k_n} \tilde{p}(\mathrm{d}\theta_j^*)^{n_j} Q(\tilde{p}), \tag{3.11}$$

where $\boldsymbol{\theta}^* := (\theta_1^*, \ldots, \theta_{k_n}^*)$ is the vector of unique values in $\boldsymbol{\theta}$, with frequencies $(n_1, \ldots, n_{k_n})$ such that $\sum_{j=1}^{k_n} n_j = n$. In line of principle, the full conditional distributions of all random elements can be derived from (3.11) and used to devise a Gibbs sampler. Given that the vector $\mathbf{X}^{(n)}$, conditionally on $\boldsymbol{\theta}$, is independent of $\tilde{p}$, the only step of the Gibbs sampler which works conditionally on a realization of the infinite-dimensional $\tilde{p}$ is the full conditional of $\boldsymbol{\theta}$. The conditional distribution $p(\boldsymbol{\theta} \mid \mathbf{X}^{(n)}, \tilde{p})$ therefore will be the main focus of our attention: its study will allow us to identify a finite-dimensional summary of $\tilde{p}$, sufficient for the purpose of updating $\boldsymbol{\theta}$ from its full conditional distribution. As a result, as far as $\tilde{p}$ is concerned, only the update of its finite-dimensional summary will need to be included in the Gibbs-sampler, thus making the conditional strategy possible.
Our proposal exploits a convenient representation of the posterior distribution of a PY process, provided in Corollary 20 of Pitman (1996), introduced in Section 1.6, and for the sake of clarity reported in the next proposition with the notation adopted in this chapter.

Proposition 6. *(Corollary 20 in Pitman, 1996). Let $t_1, \ldots, t_n \mid \tilde{p} \sim \tilde{p}$ where $\tilde{p}$ is a $PY(\sigma, \vartheta; P_0)$, and denote by $(t_1^*, \ldots, t_{k_n}^*)$ and $(n_1, \ldots, n_{k_n})$ the set of k_n distinct values and corresponding frequencies in*

$(t_1, \dots, t_n)$. The conditional distribution of $\tilde{p}$, given $(t_1, \dots, t_n)$, coincides with the distribution of

$$p_0 \tilde{q}(\cdot) + \sum_{j=1}^{k_n} p_j \delta_{t_j^*}(\cdot),$$

where $(p_0, p_1, \dots, p_{k_n}) \sim \text{Dirichlet}(\vartheta + k_n\sigma, n_1 - \sigma, \dots, n_{k_n} - \sigma)$ and $\tilde{q} \sim PY(\sigma, \vartheta + k_n\sigma; P_0)$ is independent of $(p_0, p_1, \dots, p_{k_n})$.

To the best of our knowledge, Proposition 6 has not been exploited much for computationl purposes, with the only contribution we are aware of being Fall and Barat (2014).
In the context of mixture models we are considering here, Pitman's result implies that the full conditional distribution of $\tilde{p}$ coincides with the distribution of a mixture composed by a PY process $\tilde{q}$ with updated parameters, and a discrete random probability measure with k_n fixed jump points at $\mathbf{t} := (t_1^*, \dots, t_{k_n}^*)$. This means that, in the context of a Gibbs sampler, while, by conditional independence, the update of each parameter θ_i is done independently of the other parameters $(\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n)$, the distinct values $\boldsymbol{\theta}^*$ taken by the parameters at a given iteration, are carried on to the next iteration of the algorithm through $\tilde{p}$, in the form of fixed jump points $\mathbf{t}$. Specifically, if $\Theta^* := \Theta \setminus \{t_1, \dots, t_{k_n}\}$, then, for every $i = 1, \dots, n$, the full conditional distribution of the i-th parameter θ_i can be written as

$$P[\theta_i \in \mathrm{d}t \mid X_i, \tilde{p}] \propto p_0 k(X_i; t)\tilde{q}(\mathrm{d}t) + \sum_{j=1}^{k_n} p_j k(X_i; t_j^*)\delta_{t_j^*}(\mathrm{d}t), \tag{3.12}$$

where $\tilde{q}$ is the restriction of $\tilde{p}$ to Θ^*, $p_0 := \tilde{p}(\Theta^*)$ and, for every $j = 1, \dots, k_n$, $p_j := \tilde{p}(t_j^*)$. The full conditional in (3.12) is reminiscent of the Blackwell-MacQueen scheme (see Equation 3.6) characterizing the update of the parameters in marginal methods: the parameter θ_i can either coincide with one of the k_n fixed jump points of $\tilde{p}$ or take a new value from a distribution proportional to $k(X_i; t)\tilde{q}(\mathrm{d}t)$. The key observation at the basis of the ICS is that, for the purpose of updating the parameters $\boldsymbol{\theta}$, there is no need to know the whole realization of $\tilde{p}$ but it suffices to know the vector $\mathbf{t}$ of fixed jump points of $\tilde{p}$, the value $\mathbf{p} := (p_0, p_1, \dots, p_{k_n})$ taken by $\tilde{p}$ at the partition $(\Theta^*, t_1^*, \dots, t_{k_n}^*)$ of Θ, and how to sample from a distribution proportional to $k(X_i, t)\tilde{q}(\mathrm{d}t)$. As for the latter task, we adopt an importance sampling approach and we introduce auxiliary random variables, in the spirit of the augmentation step of Neal (2000)'s Algorithm 8, originally introduced to deal with a non-conjugate specification of the mixture model in the context of marginal methods. For more details on importance sampling one can refer to Agapiou et al. (2017) and references therein. Specifically, a vector $\mathbf{s} := (s_1, \dots, s_m)$, of arbitrary size $m \geq 1$, is generated from $\tilde{q}$ and then weighted by means of the kernel $k(X_i; \cdot)$. By almost sure discreteness of $\tilde{q}$, the generated vector will show ties with positive probability and thus will feature k_m distinct values $(s_1^*, \dots, s_{k_m}^*)$, with frequencies $(m_1, \dots, m_{k_m})$. The full conditional (3.12) can thus be rewritten as

$$P[\theta_i \in \mathrm{d}t \mid X_i, \tilde{p}] \propto p_0 \sum_{j=1}^{k_m} \frac{m_j}{m} k(X_i; s_j^*)\delta_{s_j^*}(\mathrm{d}t) + \sum_{j=1}^{k_n} p_j k(X_i; t_j^*)\delta_{t_j^*}(\mathrm{d}t). \tag{3.13}$$

From the last expression it is straightforward to identify $(\mathbf{s}, \mathbf{t}, \mathbf{p})$ as a finite dimensional summary of $\tilde{p}$, sufficient for the purpose of updating the parameters θ_i from their full conditionals. This means that, as far as $\tilde{p}$ is concerned, only its summary $(\mathbf{s}, \mathbf{t}, \mathbf{p})$ must be included in the updating steps of the Gibbs sampler. To this end, Proposition 6 provides the basis for the update of $(\mathbf{s}, \mathbf{t}, \mathbf{p})$. Indeed, conditionally on $\boldsymbol{\theta}$, the fixed jump points $\mathbf{t}$ coincide with the k_n distinct values appearing in $\boldsymbol{\theta}$, while the random vectors $\mathbf{p}$ and $\mathbf{s}$ are independent with

$\mathbf{p} \sim \text{Dirichlet}(\vartheta + \sigma k_n, n_1 - \sigma, \ldots, n_{k_n} - \sigma)$ and the joint distribution of $\mathbf{s}$ characterized by the predictive distribution of a $\text{PY}(\sigma, \vartheta + \sigma k_n; P_0)$, that is, for any $i = 0, 1, \ldots, m-1$,

$$P[s_{i+1} \in \mathrm{d}t \mid s_1, \ldots, s_i] = \frac{\vartheta + \sigma(k_n + k_i)}{\vartheta + k_n\sigma + i} P_0(\mathrm{d}t) + \sum_{j=1}^{k_i} \frac{n_j - \sigma}{\vartheta + \sigma k_n + i} \delta_{\mathbf{s}_j^*}(\mathrm{d}t), \tag{3.14}$$

where $(s_1^*, \ldots, s_{k_i}^*)$ is the vector of k_i distinct values appearing in $(s_1, \ldots, s_i)$, with corresponding frequencies $(n_1, \ldots, n_{k_i})$. Sampling $\mathbf{s}$ by means of (3.14) conveniently allows us to avoid the task of generating realizations of the infinite-dimensional random probability measure $\tilde{q}$.

We have now all the elements for devising a Gibbs sampler for posterior simulation of $(\mathbf{s}, \mathbf{t}, \mathbf{p})$: the proposed Importance Conditional Sampler (ICS) is summarised in Algorithm 1. In turn, a realization from the posterior distribution of $(\mathbf{s}, \mathbf{t}, \mathbf{p})$ defines an approximate realization $f(x)$ of the posterior distribution of the random density defined in (2.1), namely

$$f(x) = p_0 \sum_{l=1}^{k_m} \frac{m_l}{m} k(x; s_l^*) + \sum_{j=1}^{k_n} p_j k(x; t_j^*).$$

Algorithm 1 includes an additional reshuffling step, meant to improve the mixing of the algorithm and consisting in updating, at each iteration of the Gibbs sampler, the distinct values $\boldsymbol{\theta}^*$ from their full conditional distribution. Namely, for every $j = 1, \ldots, k_n$,

$$P[\theta_j^* \in \mathrm{d}t \mid \mathbf{X}^{(n)}] \propto P_0(\mathrm{d}t) \prod_{i \in C_j} k(X_i; t), \tag{3.15}$$

where $C_j = \{i \in \{1, \ldots, n\} : \theta_i = \theta_j^*\}$.

If the algorithm is ran for a total of R iterations, the first R_b of which are considered burn-in and discarded, then the posterior mean can be evaluated as

$$\hat{f}(x) = \frac{1}{R - R_b} \sum_{r=R_b+1}^{R} f^{(r)}(x),$$

where $f^{(r)}$ denotes the realization of $\tilde{f}$ obtained at the r-th iteration.

It is instructive to consider how the ICS works for the special case of DPM models (that is when $\sigma = 0$). In such case, the steps described in Algoritm 1 can be nicely interpreted by resorting to three fundamental properties characterizing the DP, namely conjugacy, self-similarity and availability of finite-dimensional distributions. More specifically, when $\sigma = 0$, step 4 of Algorithm 1 consists in generating the random weights $\mathbf{p}$ from a Dirichlet distribution of parameters $(\vartheta, n_1, \ldots, n_{k_n})$. This distribution for $\mathbf{p}$ follows by combining the conjugacy of the DP (Ferguson, 1973), for which $\tilde{p} \mid \boldsymbol{\theta} \sim DP(\vartheta; P_0 + \sum_{j=1}^{k_n} n_j \delta_{\theta_j^*})$, with the availability of finite-dimensional distributions of DP (Ferguson, 1973), which provides the distribution of $\mathbf{p}$, defined as the evaluation of the conditional distribution of $\tilde{p}$ on the partition of Θ induced by $\boldsymbol{\theta}$. Moreover, when $\sigma = 0$, according to the predictive distribution displayed in step 6 of Algorithm 1, the augmenting random variables $\mathbf{s}$ are exchangeable from $\tilde{q} \sim DP(\vartheta; P_0)$, with $\tilde{q}$ independent of $\mathbf{p}$. This nicely follows from the self-similarity of the DP (see, e.g., Ghosal, 2010) which implies that $\tilde{q} = \tilde{p}|_{\Theta^*}$ is independent of $\tilde{p}|_{\Theta \setminus \Theta^*}$, and thus of $\mathbf{p}$, and is distributed as a $\text{DP}(\vartheta P_0(\Theta^*); P_0|_{\Theta^*})$. As a by-product observe that, by diffuseness of P_0, we have $\tilde{q}$ has the same distribution of $\tilde{p}$, thus implying that, in the DP case, the auxiliary random variables $\mathbf{s}$ are generated from the prior model.

Algorithm 1: Importance conditional sampler for the Pitman-Yor mixture model

```
Set admissible initial values $\boldsymbol{\theta}^{(0)}$
for each iteration $r = 1, \dots, R$ do
    set $\mathbf{t}^{(r)} = \boldsymbol{\theta}^{(r-1)}$
    sample $\mathbf{p}^{(r)}$ from $\mathbf{p}^{(r)} \sim \text{Dirichlet}(\vartheta + k_n^{(r-1)}\sigma, n_1^{(r-1)} - \sigma, \dots, n_{k_n}^{(r-1)} - \sigma)$
    for each $i = 0, \dots, m-1$ do
        sample $s_{i+1}^{(r)}$ from
            $P[s_{i+1}^{(r)} \in \mathrm{d}t \mid s_1^{(r)}, \dots, s_i^{(r)}] = \frac{\vartheta + \sigma(k_n^{(r-1)} + k_i^{(r)})}{\vartheta + k_n^{(r-1)}\sigma + i} P_0(\mathrm{d}t) + \sum_{j=1}^{k_i^{(r)}} \frac{m_j^{(r)} - \sigma}{\vartheta + \sigma k_n^{(r-1)} + i} \delta_{s_j^{*(r)}}(\mathrm{d}t)$
    for each $i = 1, \dots, n$ do
        sample $\theta_i^{(r)}$ from
            $P[\theta_i^{(r)} = t \mid \cdots] \propto \begin{cases} p_0^{(r)} \frac{m_j^{(r)}}{m} k(X_i; s_j^{(r)}) & \text{if } t \in \{s_1^{*(r)}, \dots, s_{k_m^{(r)}}^{*(r)}\} \\ \tilde{p}_j^{(r)} k(X_i; t_j^{*(r)}) & \text{if } t \in \{t_1^{*(r)}, \dots, t_{k_m^{(r-1)}}^{*(r)}\} \\ 0 & \text{otherwise} \end{cases}$
    for each unique value $\theta_j^{*(r)}$ in $\boldsymbol{\theta}^{(r)}$ do
        update $\theta_j^{*(r)}$ from $P[\theta_j^{*(r)} \in \mathrm{d}t \mid \cdots] \propto P_0(\mathrm{d}t) \prod_{i \in C_j^{(r)}} k(X_i; t)$
end
```

3.4 Implementation of competing algorithms

In Section 3.5 the performance of the ICS will be compared with that one of the marginal sampler and the slice sampler. For the sake of simplicity, we considered all the samplers without introducing prior distributions on the hyerparameters, which are thus kept fixed. In this section, for the sake of completeness, we report the implementation of the marginal sampler and the slice sampler which were used in the study.

Algorithm 2: Marginal sampler

```
Set admissible initial values $\boldsymbol{\theta}^{(0)}$
for each iteration $r = 1, \dots, R$ do
    for each $i = 1, \dots, n$ do
        sample $\theta_i^{(r)}$ from
            $P[\theta_i^{(r)} = t \mid \dots] \propto \begin{cases} (n_j - \sigma) k(X_i; \theta_j^{*(r)}) & \text{if } \theta_j^{*(r)} \in \{\theta_1^{*(r)}, \dots, \theta_{k_m^{(r-1)}}^{*(r)}\} \\ (\vartheta + k\sigma) \int k(X_i, \theta) P_0(\mathrm{d}\theta) & \text{otherwise} \end{cases}$
    for each unique value $\theta_j^{*(r)}$ in $\boldsymbol{\theta}^{(r)}$ do
        update $\theta_j^{*(r)}$ from $P[\theta_j^{*(r)} \in \mathrm{d}t \mid \cdots] \propto P_0(\mathrm{d}t) \prod_{i \in C_j^{(r)}} k(X_i; t)$
end
```

The marginal sampler that we consider follows the steps of Müller et al. (1996), exception made for the use of hyperprior distributions. Algorithm 2 shows the pseudo-code of the marginal sampler's implementation.
For the implementation of the slice sampler, we followed the scheme presented in Kalli et al. (2011), with the weights $\xi_1, \xi_2, \dots$ introduced in their paper (see Kalli et al., 2011, for details on the weights) set equal to the PY weights, as in Walker (2007). Algorithm 3 shows the pseudo-code of the slice sampler's implementation.

Algorithm 3: Slice sampler for Pitman-Yor mixture model

```
Set admissible initial values $\boldsymbol{\theta}^{(0)}$
for each iteration $r = 1, \dots, R$ do
    set $\mathbf{t}^{(r)} = \boldsymbol{\theta}^{(r-1)}$
    for each $i = 1, \dots, n$ do
        Sample $u_i$ from the corresponding distribution
            $$u_i \sim Unif([0, w_i])$$
         where $w_i = w_j \, s.t. \, \theta_i = \theta_j^*$
    while $\sum_{j=1}^{k} w_j < 1 - u_i$, for any $i$ do
        Sample a new weight
            $$v_{k+1} \sim Beta(1-\sigma, \vartheta + (k+1)\sigma), \qquad w_{k+1} = v_{k+1} \prod_{l<k+1} (1 - v_l).$$
        Sample $t_{k+1}^{(r)} \sim P_0(\mathrm{d}\theta)$
        Set $k = k + 1$.
    for each $i = 1, \dots, n$ do
        sample $\theta_i^{(r)}$ from
            $$P[\theta_i^{(r)} = t \mid \cdots] \propto \begin{cases} \mathbb{1}_{[w_j > u_i]} k(X_i, t_j^{(r)}) & \text{if } t \in \{t_1^{*(r)}, \dots, t_k^{*(r)}\} \\ 0 & \text{otherwise} \end{cases}$$
    for each unique value $\theta_j^{*(r)}$ in $\boldsymbol{\theta}^{(r)}$ do
        update $\theta_j^{*(r)}$ from $P[\theta_j^{*(r)} \in \mathrm{d}t \mid \cdots] \propto P_0(\mathrm{d}t) \prod_{i \in C_j^{(r)}} k(X_i; t)$
        Sample the weight $w_j$ with
            $$v_j \sim Beta\left(1 - \sigma + n_j, \vartheta + j\sigma + n_j^+\right), \qquad w_j = v_j \prod_{l<j} (1 - v_l),$$
         where $n_j = \sum_{i=1}^{n} \mathbb{1}_{[\theta_i = \theta_j^*]}$ is the number of elements in the cluster $j$ and
         $n_j^+ = n - \sum_{l=1}^{j} n_j$.
end
```

A known problem, when dealing with the comparison of two or more methods, is the bias induced by different implementations (see Kriegel et al., 2017, for a discussion). In order to reduce to a minimum such distortion, we implemented all methods in a low level language, *C++* with the use of *RcppArmadillo* library, preserving, as far as possible, commonalities between the sub-routines of the samplers. All the samplers are implemented in the BNPmix R package [1]. More details are provided in Appendix D.

3.5 Simulation study

We want to investigate the performance of our proposal, the ICS, and compare it with the performance of the marginal sampler and the slice sampler. For the whole simulation study, simulated samples are generated from a mixture of two Gaussian distributions, namely

$$f^* = \frac{3}{4}N(-2.5, 1) + \frac{1}{4}N(2.5, 1). \tag{3.16}$$

The generated samples from the posterior distribution, are analysed by considering two quantities, aimed to assess their quality and the computational cost of the samplers. The first measure we consider is the effective sample size (ESS), defined as

$$\text{ESS} = \frac{R - R_b}{1 + \sum_{j=1}^{\infty} \rho_j(Y)}, \tag{3.17}$$

where R is the number of iterations in the MCMC chain, R_b the number of burn-in iterations, Y is a random quantity of interest and $\rho_j(\cdot)$ is the autocorrelation of lag j. In our studies we set Y equal to the number of active clusters in each iteration, that is the number of distinct values appearing in $\boldsymbol{\theta}$. The infinite sum in (3.17) is approximated by its truncation up to the $(R-R_b)/2$-th term, as it is standard practice. The ESS can be interpreted as the number of effective independent observation produced in the sampler, it is a measure of the quality of the produced sample.

A second measure we consider is the ratio between execution time T and the ESS, so to compare the computational cost of different samplers, in terms of the time required to get an independent realization. This is an indication of the effective time required to produce an independent sample from the posterior distribution.

For all the considered samplers a Normal-Inverse-Gamma base measure was considered, i.e. $P_0 = N(\mu; m_0, k_0\sigma) \times IG(\sigma; a_0, b_0)$. We set the base measure parameters as $m_0 = 0$, $k_0 = 6$, $a_0 = 2$ and $b_0 = 1$, specification such that the prior variance is equal to the actual variance of each component in the data-generating density (3.16), and the location component is rather flat on the support. We ran the model for $R = 1\,500$ of iterations, of which $R_b = 500$ are burn-in iterations. For the three samplers, we checked the convergence by visual investigation of the traceplots, which did not provide any evidence against it.

The first part of our study is focused on the ICS only and aims at investigating the effect on the generated posterior sample, of different choices for m, the number of values generated in the importance sampling step. To this end, we considered different scenarios, obtained by considering different sample sizes $n = \{100, 300, 1\,000\}$, different values for the parameter $m = \{1, 5, 10, 100\}$ and different values for the strength and discount parameters of the PY, $\vartheta = \{0.1, 1, 10\}$ and $\sigma = \{0, 0.2, 0.4, 0.6, 0.8\}$. The results are averaged over 10 replications.

Figure 3.1 shows the ESS for different choices of the parameter m, as a function of the discount parameter σ, for different choices of the strength parameter ϑ and the sample size n. The ESS does not appear to be affected by the choice of the discount parameter σ, while it is interesting to observe that, when the sample size increases, the ESS tends to decrease. What really matters in this study is the effect of m on the ESS, which is more apparent when we consider large

[1] The package is available at https://github.com/rcorradin/BNPmix and can be installed via devtools.

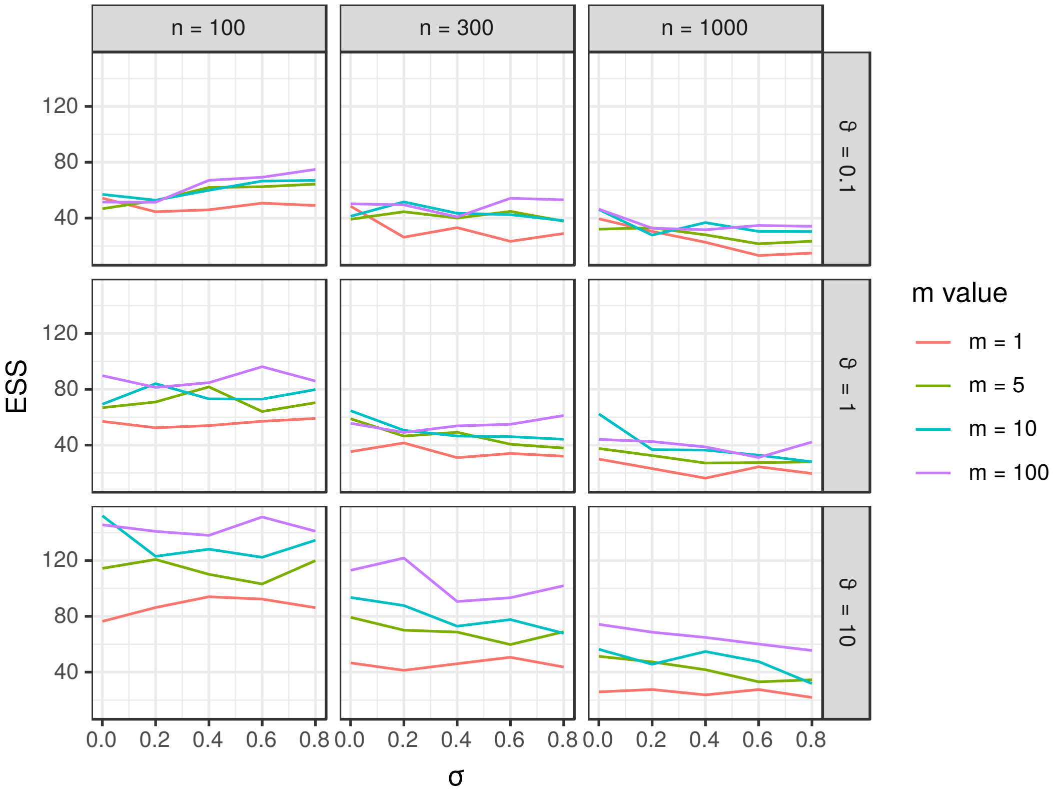

FIGURE 3.1: ICS: Effective sample size, for different values for the strength parameter $\vartheta = \{0.1, 1, 10\}$, different values of the discount parameter $\sigma = \{0, 0.2, 0.4, 0.6, 0.8\}$ and sample size $n = \{100, 300, 1\,000\}$.

values of the strength parameter ϑ. In general, and not surprisingly, larger values for m lead to larger ESS. More specifically, setting $m = 1$ seems to be a poor choice in terms of ESS. On the other end, if we compare the ESS for $m = 10$ and $m = 100$, it might be argued that the improvement obtained by setting $m = 100$ is marginal.

Figure 3.2 shows T/ESS, the ratio between execution time T and ESS, for the different scenarios considered, in log scale. The ratio appears to be affected only slightly by different choices of the discount parameter σ, exception made for $m = 100$ where σ has a stronger impact. Indeed, when $m = 100$, the time needed to sample an independent realization from the posterior distribution becomes clearly larger increase if the discount parameter grows. Again, what matters here is the effect of m on T/ESS. If we compare $m = 10$ and $m = 100$, it seems clear that setting $m = 100$ leads to a significantly larger ratio T/ESS. Based on these considerations, for the second part of the study and for the rest of the Chapter, we set $m = 10$, as it seems a good compromise when quality of the sample (ESS) and computational efficiency (T/ESS) are considered.

The second part of the study focuses on the comparison of ICS with the slice sampler and the marginal sampler. Also in this case, we compare different scenarios, obtained by considering different values for the strength parameter, $\vartheta = \{0.1, 1, 10\}$, for the discount parameter, $\sigma = \{0.0, 0.15, 0.3, 0.45, 0.6\}$, and different sample sizes $n = \{100, 300, 1\,000\}$. The results are averaged over 10 replications.

Figure 3.3 shows the ESS, for the different scenarios considered, as a function of the discount parameter σ. Note that the blue lines, corresponding to the ESS of the slice sampler, are available only for part of the range considered for σ: performing the simulation study in a reasonable

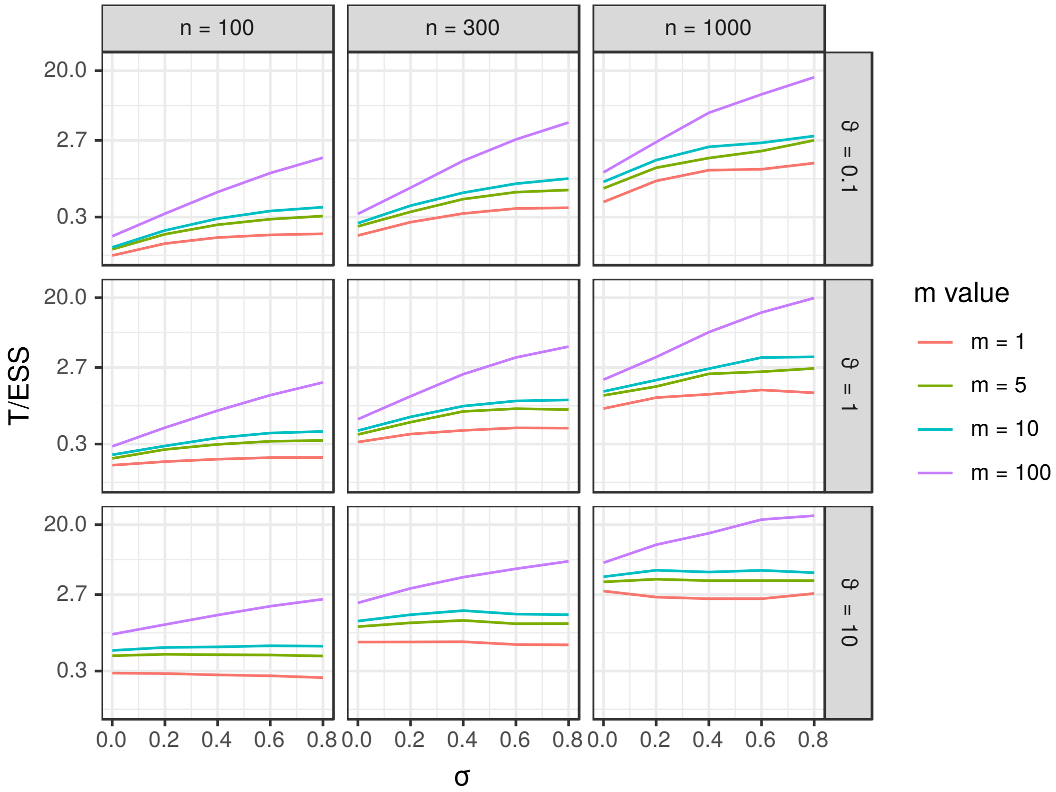

FIGURE 3.2: ICS: ratio between execution time and effective sample size, in log scale, for different values for the strength parameter $\vartheta = \{0.1, 1, 10\}$, different values of the discount parameter $\sigma = \{0, 0.2, 0.4, 0.6, 0.8\}$ and sample size $n = \{100, 300, 1\,000\}$.

time, with large values of the discount parameter σ, turned out to be infeasible. Compared with ICS and slice sampler, the marginal algorithm produces better samples, in terms of ESS, thing which is in particularly apparent for large values of the strength parameter ϑ. As far as the two conditional samplers are considered, the ICS generates better samples than the slice sampler.

Figure 3.4 shows the ratio between execution time T and ESS, in log scale, for the different scenarios and the three samplers considered. Although, again, the blue curves are available only for small values of σ, the plot clearly suggests that, in the slice sampler case, the considered ratio explodes when the discount parameter σ grows. The ICS appears to be rather constant with respect to the discount parameter, while the marginal sampler increases as σ becomes large. Overall, the time needed to sample an independent observation from the posterior distribution in the ICS is not greatly affected by the values of the discount and strength parameter. This last observation suggests the ICS might conveniently adopted, for example, when the model is endowed with hyperpriors for σ and ϑ. Finally and not surprisingly, for all the three samplers, the ratio T/ESS is increasing as a function of the sample size.

Due to the amount of time required to perform the previous simulation study, we considered only for the univariate case. Regarding the scalability of the algorithms, when the number of dimensions increases, the computational time required to evaluate the kernels grows: the differences measured for the different samplers in the univariate case are thus amplified.

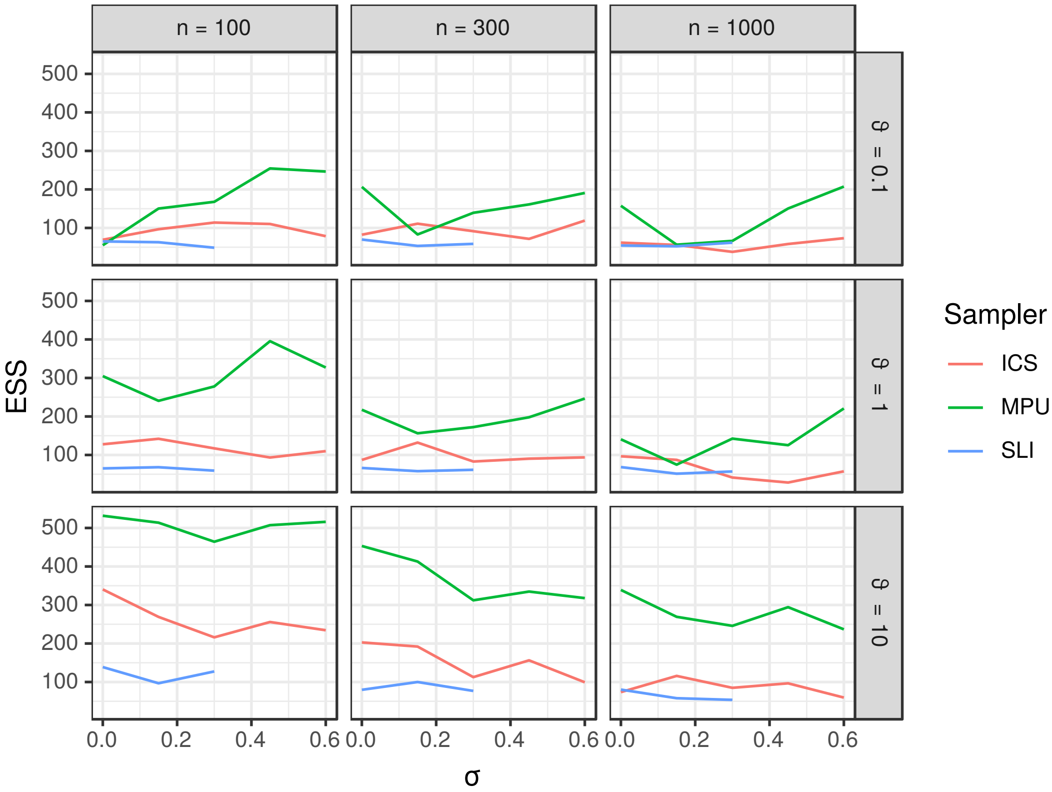

FIGURE 3.3: Effective sample size for the ICS, marginal sampler and slice sampler. Different values for the strength parameter $\vartheta = \{0.1, 1, 10\}$, different values of the discount parameter $\sigma = \{0, 0.15, 0.3, 0.45, 0.6\}$ and sample size $n = \{100, 300, 1\,000\}$.

3.6 ICS for dependent Dirichlet process mixtures

The promising results obtained in the simulation study of the previous section, make the ICS a good candidate for contexts where efficient samplers are needed in order to fit a Bayesian model based on the use of nonparametric mixtures. One of such examples is surely represented by the class of dependent Dirichlet process Maceachern (1999) and Maceachern (2000) for partially exchangeable data (see J. Foti and Williamson, 2015, for a review). Within this class of models, we consider a multivariate vector of GM-dependent Dirichlet processes (GM-DDP), as defined and studied in Lijoi et al. (2014a) and Lijoi et al. (2014b), inspired by the work of Griffiths and Milne (1978) on dependent and identically distributed Poisson random measures. For an allied approach see also Griffin et al. (2013). In this section we describe how the ICS can be used as a building block for devising a novel and efficient conditional algorithm for GM-DDP mixture models.

Let $\mu_0, \mu_1, \dots, \mu_L$ be independent gamma completely random measures (see Appendix A) with Lévy intensities respectively equal to $\nu_0, \nu_1, \dots, \nu_L$, where

$$\nu_0(\mathrm{d}s, \mathrm{d}x) = c(1-z)\frac{\mathrm{e}^{-s}}{s}\mathrm{d}s P_0(\mathrm{d}x),$$

$$\nu_l(\mathrm{d}s, \mathrm{d}x) = cz\frac{\mathrm{e}^{-s}}{s}\mathrm{d}s P_0(\mathrm{d}x), \qquad l = 1, \dots, L,$$

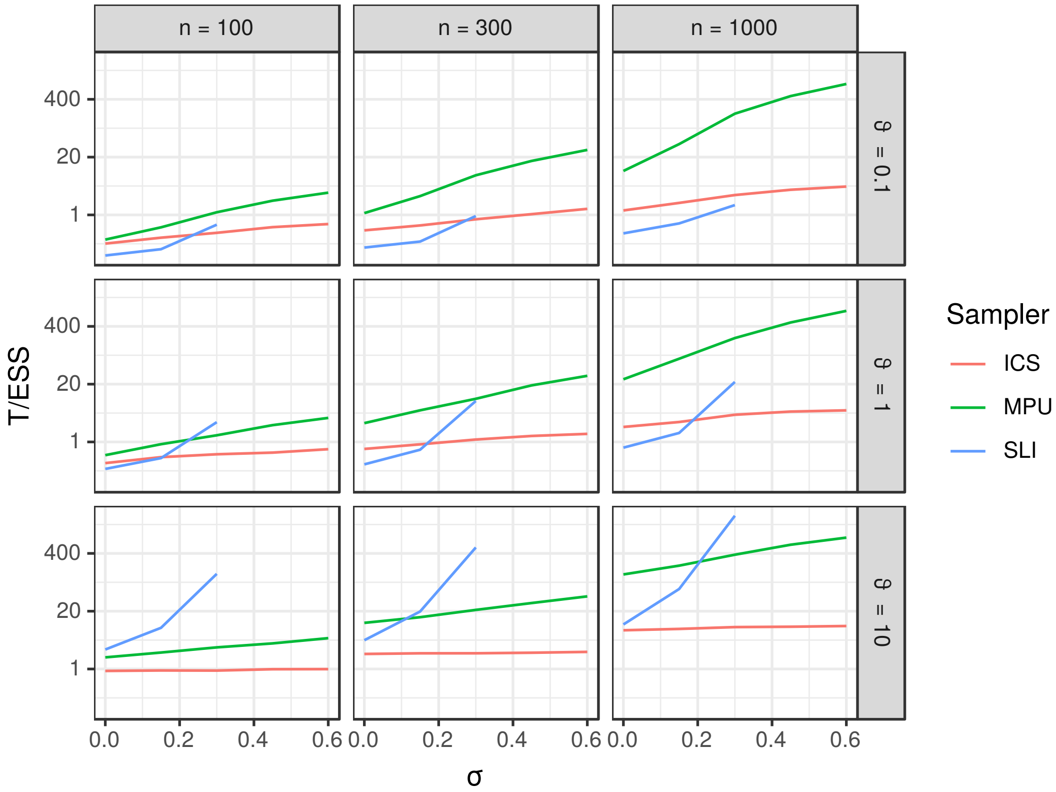

FIGURE 3.4: Ratio between execution time and effective sample size, in log scale, for the ICS, marginal sampler and slice sampler. Different values for the strength parameter $\vartheta = \{0.1, 1, 10\}$, different values of the discount parameter $\sigma = \{0, 0.15, 0.3, 0.45, 0.6\}$ and sample size $n = \{100, 300, 1\,000\}$.

and $z \in [0,1]$. A vector of GM-dependent gamma completely random measures $(\tilde{\mu}_1, \dots, \tilde{\mu}_L)$ is obtained by setting $\tilde{\mu}_l = \mu_l + \mu_0$ for $l = 1, \dots, L$. Marginally, the completely random measures $\tilde{\mu}_l$ are identically distributed with Lévy intensity $\nu(\mathrm{d}s, \mathrm{d}x) = c\mathrm{e}^{-s}/s\mathrm{d}sP_0(\mathrm{d}x)$. The use of the common term μ_0 in the definition of the $\tilde{\mu}_l$'s induce dependence across the components of $(\tilde{\mu}_1, \dots, \tilde{\mu}_L)$. A vector of identically distributed dependent Dirichlet processes is then obtained by normalizing each component of the vector $(\tilde{\mu}_1, \dots, \tilde{\mu}_L)$, thus obtaining $(\tilde{p}_1, \dots, \tilde{p}_L)$ where, for each component, we have

$$\begin{aligned}\tilde{p}_l &= \frac{\tilde{\mu}_l}{\tilde{\mu}_l(\mathbb{X})} \\ &= \frac{\mu_l + \mu_0}{\mu_l(\mathbb{X}) + \mu_0(\mathbb{X})} \\ &= \frac{\mu_l}{\mu_l(\mathbb{X})}\frac{\mu_l(\mathbb{X})}{\mu_l(\mathbb{X}) + \mu_0(\mathbb{X})} + \frac{\mu_0}{\mu_0(\mathbb{X})}\frac{\mu_0(\mathbb{X})}{\mu_l(\mathbb{X}) + \mu_0(\mathbb{X})} \\ &= p_l w_l + p_0(1 - w_l),\end{aligned}$$

where the $p_l := \mu_l/\mu_l(\mathbb{X})$, for $l = 0, 1, \dots, L$ and $w_l := \mu_l(\mathbb{X})/(\mu_l(\mathbb{X}) + \mu_0(\mathbb{X}))$. Notice that the random probability measures p_l's, for $l = 1, \dots, L$, are independent Dirichlet processes with base measure P_0 and total mass cz and, in turn, they are independent of p_0 which is a DP with total mass $c(1-z)$ and same base measure P_0. The vector of random weights $(w_1, \dots, w_L)$ takes values in $[0,1]^L$ and follows a multivariate beta distribution (Olkin and Liu, 2003). Moreover,

by standard properties of gamma random variables, it is easy to verify that $(p_0, p_1, \ldots, p_L)$ is independent of $(w_1, \ldots, w_L)$. The vector $(\tilde{p}_1, \ldots, \tilde{p}_L)$ is an L-dimensional vector of GM-DDP, where the components are identically distributed, with base measure P_0 and total mass c.
Let $(X_{1,1}, \ldots, X_{n_1,1}), (X_{1,2}, \ldots, X_{n_2,2}), \ldots, (X_{1,L}, \ldots, X_{n_L,L})$ be L sets of observations such that exchangeability holds within each set but not across different sets. We assume the following partially exchangeable mixture model (see Section 1.3) based on the vector $(\tilde{p}_1, \ldots, \tilde{p}_L)$ of GM-DDP.

$$\begin{aligned}(X_{i_1,1}, X_{i_2,2}, \ldots, X_{i_L,L}) \mid \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(L)} &\overset{\text{ind}}{\sim} \prod_{l=1}^{L} k(x_{i_l,l}; \theta_{i_l,l}) \\ \theta_{i,l} \mid (\tilde{p}_1, \ldots, \tilde{p}_L) &\overset{\text{iid}}{\sim} \tilde{p}_l \\ (\tilde{p}_1, \ldots, \tilde{p}_L) &\sim \text{GM-DDP}(c, z, P_0).\end{aligned} \tag{3.18}$$

By specifying a model as in (3.18) it follows

$$\begin{aligned}(w_1, \ldots, w_L) &\sim \text{mult-Beta}(cz, \ldots, cz) \\ (w_1, \ldots, w_L) &\perp\!\!\!\perp (p_0, p_1, \ldots, p_L).\end{aligned}$$

We want to exploit the ICS in this context, so to devise an efficient algorithm to fit the GM-DDP mixture model in (3.18). The main idea of our approach consists in working in two stages: first, once the observations are allocated to clusters belonging to either to the idiosyncratic process p_l's or to the common process p_0, we update the summaries for all the processes p_l, $l = 0, 1, \ldots, L$, that is

$$(\mathbf{s}_l, \mathbf{t}_l, \mathbf{p}_l), \qquad l = 0, \ldots, L,$$

as done in Section 3.3 for a single process. Second, for every $l = 1, \ldots, L$ and $1 \leq i_l \leq n_l$, we update $\theta_{i,l}$ from

$$\begin{aligned}&P[\theta_{i,l} \in \mathrm{d}t | X_{i,l}, \tilde{p}_l, \ldots] \\ &\quad \propto w_l \left(p_{0,l} \sum_{j=1}^{k_{m,l}} \frac{m_{j,l}}{m} k(X_{i,l}, s^*_{j,l}) + \sum_{j=1}^{k_{n,l}} p_{j,l} K(X_{i,l}, t^*_{j,l}) \delta_{t^*_{j,l}}(dt) \right) \\ &\quad + (1 - w_l) \left(p_{0,0} \sum_{j=1}^{k_{m,0}} \frac{m_{j,0}}{m} k(X_{i,l}, s^*_{j,0}) + \sum_{j=1}^{k_{n,0}} p_{j,0} K(X_{i,l}, t^*_{j,0}) \delta_{t^*_{j,0}}(dt) \right)\end{aligned} \tag{3.19}$$

A crucial step is the update of the weights $\mathbf{w} = (w_1, \ldots, w_L)$ in $[0, 1]^L$, a priori distributed as a multivariate beta distribution with parameters $(cz, \ldots, cz)$ and for which the full conditional is given by

$$P[\mathbf{w} = \boldsymbol{\omega} \mid c, z, \boldsymbol{\theta}, \ldots] \propto B(cz, \ldots, cz) \frac{\prod_{l=1}^{L} \omega_l^{cz + n_{l,l} - 1}}{(1 - \omega_l)^{cz - n_{l,0} + 1}} \left[1 + \sum_{l=1}^{L} \frac{\omega_l}{1 - \omega_l} \right]^{-Lcz}, \tag{3.20}$$

where $n_{l,l} = \sum_{i=1}^{n_l} \mathbb{I}_{[\theta_{i,l} \in p_l]}$ and $n_{l,0} = \sum_{i=1}^{n_l} \mathbb{I}_{[\theta_{i,l} \in p_0]}$. To this end, we adopted an importance sampling approach, consisting in sampling from an L-dimensional uniform proposal, and reweighting the sampled values by means of the full conditional (3.20).
The ICS for the GM-DDP mixture model is summarized in Algoritm 4.
For the sake of illustration, we used the ICS to fit a GM-DDP mixture model to analyze an astronomical dataset. We consider a subset of the data studied in Balogh et al. (2004), where

Algorithm 4: ICS for GM-dependent Dirichlet process mixture models

```
Set admissible initial values $\boldsymbol{\theta}^{(0)}$
for each iteration $r = 1, \dots, R$ do
    set $\mathbf{t}_0^{(r)} = \boldsymbol{\theta}_0^{(r-1)}$ where $\boldsymbol{\theta}_0^{(r-1)}$ are the common values
    sample $\mathbf{p}_0^{(r)}$ from $\mathbf{p}_0^{(r)} \sim \text{Dirichlet}(c(1-z), n_{1,0}^{(r-1)}, \dots, n_{k_0,0}^{(r-1)})$
    sample the sequence $\{s_{i,0}\}_{i=1}^{m}$ from a $DP(c(1-z), P_0)$
    for each urn $l = 1, \dots, L$ do
        set $\mathbf{t}_l^{(r)} = \boldsymbol{\theta}_l^{(r-1)}$ without including the common values
        sample $\mathbf{p}_l^{(r)}$ from $\mathbf{p}_l^{(r)} \sim \text{Dirichlet}(cz, n_{1,l}^{(r-1)}, \dots, n_{k_l,l}^{(r-1)})$
        sample the sequence $\{s_{i,l}\}_{i=1}^{m}$ from a $DP(cz, P_0)$
    sample $\mathbf{w}^{(r)}$ by the importance sampling step with weights as in (3.20)
    for each $i = 1, \dots, n_l;\ l = 1, \dots, L$ do
        sample $\theta_{i,l}^{(r)}$ from
            $$P[\theta_i^{(r)} = t \mid \cdots] \propto \begin{cases} w_l p_{0,l}^{(r)} \frac{m_{j,l}^{(r)}}{m} k(X_i; s_{j,l}^{(r)}) & \text{if } t \in \{s_{1,l}^{*(r)}, \dots, s_{k_l^{(r)},l}^{*(r)}\} \\ w_l p_{j,l}^{(r)} k(X_i; t_{j,l}^{*(r)}) & \text{if } t \in \{t_{1,l}^{*(r)}, \dots, t_{k_l^{(r-1)},l}^{*(r)}\} \\ (1-w_l) p_{0,0}^{(r)} \frac{m_{j,0}^{(r)}}{m} k(X_i; s_{j,0}^{(r)}) & \text{if } t \in \{s_{1,0}^{*(r)}, \dots, s_{k_0^{(r)},0}^{*(r)}\} \\ (1-w_l) \tilde{p}_{j,0}^{(r)} k(X_i; t_{j,0}^{*(r)}) & \text{if } t \in \{t_{1,0}^{*(r)}, \dots, t_{k_m^{(r-1)},0}^{*(r)}\} \\ 0 & \text{otherwise} \end{cases}$$
    for each unique value $\theta_{j,l}^{*(r)}$ in $\boldsymbol{\theta}_l^{(r)}$, $l = 0, \dots, L$ do
        update $\theta_{j,l}^{*(r)}$ from
            $$P[\theta_{j,l}^{*(r)} \in \mathrm{d}t \mid \cdots] \propto P_0(\mathrm{d}t) \prod_{\substack{\theta_{i,l} \in p_l \\ i \in C_{j,l}^{(r)}}} k(X_i; t)$$
end
```

the authors analyze the difference of ultraviolet and red filters ($U - R$) color distribution of a set of 24 346 galaxies, stratified by density (Mps) and luminosity (M_r) of the galaxy. Specifically, we consider the subset of $n = 6\,344$ observations with medium luminosity, i.e. $-20 < M_r < -19$. Following Balogh et al. (2004), the considered dataset is stratified into five different groups, by binning with respect to the density Mps. This leads to a dataset consisting of $L = 5$ groups of cardinality respectively equal to $n_1 = 478$, $n_2 = 1622$, $n_3 = 2515$, $n_4 = 1173$ and $n_5 = 556$. We modelled such data with a GM-DDP mixture model, specified with a Normal-Inverse-Gamma base measure, i.e. $P_0 = N(\mu; m_0, k_0\sigma) \times IG(\sigma; a_0, b_0)$. We specified the base measure's parameters as $m_0 = 2$, $k_0 = 5$, $a_0 = 4$ and $b_0 = 1$. For the sake of simplicity, we specified the mass parameter $c = 2$ and a non-informative prior weight $z = 0.5$ for the common process in the definition of the $\tilde{\mu}_l$'s. We ran the algorithm for 10 000 of iterations, after 2 500 burn-in iterations.

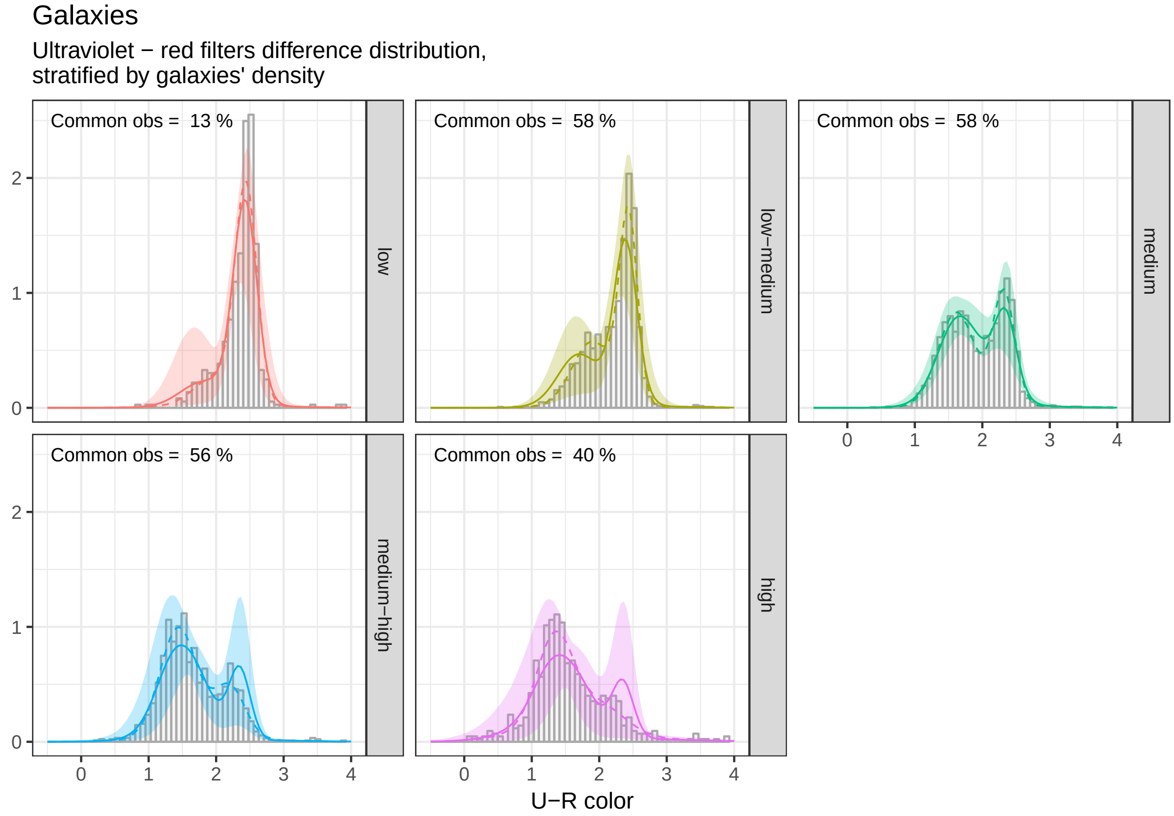

FIGURE 3.5: Galaxy colours data, GM-DDP mixture model. Solid lines: posterior mean for the GM-DDP mixture model; filled areas: 95% posterior credible bands; dashed lines: posterior mean for the marginal DPM model.

Figure 3.5 shows the results of the estimated posterior means for the five groups, obtained by fittin a GM-DDP mixture model (solid lines) and the equivalent estimates obtained by fitting five independent DPM models (dashed line), via ICS. The figure also shows 95% posterior credible intervals for the GM-DDP mixture model (filled area). A difference in terms of posterior means can be appreciated when the GM-DDP mixture and the independent DPM models are considered: the borrowing of information induced by the GM-dependence makes the bimodality of the distribution more apparent. For example in the fifth panel, the posterior mean is bimodal when dependence is exploited while it is not if the data are analyzed marginally.

3.7 Conclusions

We proposed a novel algorithm, the ICS, to sample from the posterior distribution of a PYM model, using a conditional approach. The method combines the appealing features of both marginal and conditional approaches: it is simple and interpretable, the number of random elements sampled in each step is bounded, the allocation of observations in clusters is not done sequentially and therefore is parallelize, and it allows easy quantification of the posterior uncertainty.

The implementation of the ICS, presented in Algorithm 1, is simple and reminiscent of the marginal approach. Interestingly, our proposal is the only conditional approach for PYM models we are aware of, which does not rely on the stick-breaking representation of the PY.

One of the reasons motivating us to study new strategies for sampling from the posterior of PYM models, was the need of a sampler whose efficiency is not affected by the value taken by the discount parameter σ. By means of a simulation study, we could see that the ICS meets this desideratum. Specifically, the ICS outperforms its competitors, as far as the ratio between execution time and effective sample size is considered.

Finally, the ICS is a good candidate for contexts where efficient samplers are needed in order to fit a Bayesian model based on the use of nonparametric mixtures. We consider a GM-DDP mixture model for partially exchangeable data and described how the ICS strategy can be adapted to this case.

Chapter 4

Elicitation of Gibbs-type priors

> Fairy tales lie just as much as statistics do, but sometimes you can find a grain of truth in them.
>
> Sergej Vasilievič Luk'janenko, *Author*

Corradin, R., Nipoti, B.
Elicitation of Gibbs-type priors via cluster size constraints.
In preparation.

The introduction of the Dirichlet process (DP) by Ferguson (1973) was a breakpoint in the history of Bayesian nonparametric statistics. Since the celebrated paper of Ferguson, the study of discrete random probability measures, generalizing the DP, has been an active area of research. A trade-off between flexibility and analytical tractability typically characterizes the classes of processes generalizing the DP, appeared in the literature. In this perspective, a central role has been played by the family of Gibbs-type priors, briefly introduced in Section 1.7, which can be considered as a convenient compromise between flexibility and tractability. Gibbs-type priors were first studied by Gnedin and Pitman (2006), and include, as special cases, commonly used nonparametric priors, such as the Dirichlet process (Ferguson, 1973), the Pitman-Yor process (PY, Perman et al., 1992), the normalized generalized gamma process (NGG, Lijoi et al., 2007c, see also Pitman, 2003, Prünster, 2002, James, 2002, Lijoi and Prünster, 2003, Regazzini et al., 2003) with its particular case, the inverse Gaussian process (Lijoi et al., 2005b), and the normalized σ-stable process (Kingman, 1975). For a recent review on the properties of Gibbs-type priors and their use in Bayesian nonparametric statistics, one can refer to De Blasi et al. (2015). Random probability measures in the family of Gibbs-type priors have been used to define nonparametric priors in a variety of fields. Just to mention a few examples, they have been exploited in mixture models (e.g., Ishwaran and James, 2001; Lijoi et al., 2007c), survival analysis (e.g., Jara et al., 2010), species sampling problems (e.g., Lijoi et al., 2007b; Lijoi et al., 2007a) and applications in linguistics and information retrieval (e.g., Teh, 2006; Teh and Jordan, 2010). In this chapter we focus on the study of some distributional properties of Gibbs-type priors and we show that our findings can be conveniently used for prior elicitation of the parameters of Gibbs-type priors in the mixture model framework. More specifically, under the assumption of observations being exchangeable under a Gibbs-type prior, we derive a novel (and surprisingly simple) expression for the distribution of the size of the cluster the $(n+1)$th observation will belong to, conditionally on an unobserved sample of size n. We will then show that this turns out to be convenient when one wants to incorporate into the model available prior information on the size of the clusters of the sample.

Section 4.1 gives a formal introduction to the class of Gibbs-type prior and their predictive distribution. In Section 4.2 we present our main result on the size of the of the $(n+1)$th observation's cluster. In Section 4.3 we investigate how the main result specializes when particular examples of Gibbs-type priors are considered, namely the Dirichlet process, the Pitman-Yor process and the normalized generalized gamma process. In Section 4.4 we describe a possible strategy for using the result of Section 4.2 to elicit the parameters of a Gibbs-type prior.

4.1 Gibbs-type prior and induced partitions

Let $\mathbb{X}$ be a Polish space, equipped with its Borel σ-field $\mathscr{X}$, and P_0 be a diffuse probability measure defined on the measurable space $(\mathbb{X}, \mathscr{X})$. We say that $\tilde{p}$ is a *species sampling model* (Pitman, 1995) if $\tilde{p}$ is a random probability measure of the form

$$\tilde{p} = \sum_{j=1}^{\infty} W_j \delta_{\tilde{X}_j},$$

where $\{W_j\}_{j\geq 1}$ is a sequence of nonnegative random weights, such that $\sum_{j=1}^{\infty} W_j = 1$ almost surely, $\{\tilde{X}_j\}_{j\geq 1}$ is a sequence of random atoms, independent and identically distributed from P_0, and weights and jumps are independent. Consider a sample $X_1, \ldots, X_n$ modelled as exchangeable with distribution governed by $\tilde{p}$, that is

$$\begin{aligned} X_j \, | \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p}, \qquad j = 1, \ldots n \\ \tilde{p} &\sim Q, \end{aligned} \tag{4.1}$$

where Q, the distribution of $\tilde{p}$, is a probability distribution over $\mathbb{M}_{\mathbb{X}}$, the space of all probability measures on $(\mathbb{X}, \mathscr{X})$. To get an interpretation of the previous quantities, in the species sampling context, a random atom $\tilde{X}_j$ could be thought of as a particular species, or as a label assigned to a particular species, and the corresponding random weight W_j as the proportion of the species X_j in the population. Let $\{X_j\}_{j\geq 1}$ be an exchangeable sequence governed by a species sampling model $\tilde{p}$, that is a sequence such that, for any $n \geq 1$ and $A_1, \ldots, A_n \in \mathscr{X}$,

$$P[X_1 \in A_1, \ldots, X_n \in A_n \mid \tilde{p}] = \prod_{j=1}^{n} \tilde{p}(A_j).$$

We refer to $\{X_j\}_{j\geq 1}$ as *species sampling sequence*. Due to the almost sure discreteness of $\tilde{p}$, there will be, almost surely, ties in a species sampling sequence.

A key quantity in the species sampling context, and more in general in Bayesian statistics, is the predictive distribution, that is the conditional distribution of the next observation X_{n+1}, given an observed sample $X_1, \ldots, X_n$ of size n (see Section 1.4). Recall that it can be defined as the posterior expectation of $\tilde{p}$, given $X_1, \ldots, X_n$, that is

$$P[X_{n+1} \in \mathrm{d}t \mid X^{(n)}] = \int_{\mathbb{M}_{\mathbb{X}}} \tilde{p}(\mathrm{d}t) Q(\mathrm{d}\tilde{p} \mid X^{(n)}), \tag{4.2}$$

where $Q(\cdot \mid X^{(n)})$ is the posterior distribution of $\tilde{p}$ and $\mathbb{M}_{\mathbb{X}}$ is the space of the probability measures with support $\mathbb{X}$.

Let $X_1^*, \dots, X_k^*$ be the k unique values in the sample $X_1, \dots, X_n$, and $\mathbf{n} = (n_1, \dots, n_k)$ be the corresponding frequencies. Moreover assume that $\tilde{p}$ is such that $\mathbb{E}[\tilde{p}] := P_0$, with P_0 diffuse distribution on $\mathbb{X}$. Then the predictive distribution in Equation 4.2 can be equivalently written as

$$P[X_{n+1} \in \mathrm{d}t \mid X^{(n)}] = g_0(n, k, \mathbf{n}) P_0(\mathrm{d}t) + \sum_{j=1}^{k} h_j(n, k, \mathbf{n}) \delta_{X_j^*}(\mathrm{d}t), \tag{4.3}$$

where $g_0(n, k, \mathbf{n})$ is the weight of the prior guess and $h_j(n, k, \mathbf{n})$ the weight of the j-th unique value X_j^* in the observed sample.

We focus our attention on the family of species sampling models for which the weight $g_0(n, k, \mathbf{n})$ in the predictive distribution (4.3), does not depend on $\mathbf{n}$. As anticipated in Section 1.7, the members of such a family of random probability measures are named Gibbs-type priors. In other terms, in Gibbs-type priors, the weight of the prior guess depends only on the size n of the observed sample and the number k of unique values in the sample.

Next, we recall two meaningful results that relate the form taken by the weights in the predictive distribution 4.3 with the PY process and, more in general, with Gibbs-type priors. The first result in this direction is a sufficientness postulate for the PY process, proved by Zabell (1997): the author lists a set of assumptions, sufficient to guarantee that a species sampling model is of PY type.

Proposition 7. *(Zabell, 1997) Let $\tilde{p}$ be an arbitrary species sampling model, with predictive probability as in Equation 4.3 and Ψ_n be the induced random partition on $\{1, \dots, n\}$. Consider the following assumptions:*

C1. $\Pr[\Psi_n = \psi_n] > 0$*, for any ψ_n partition of $\{1, \dots, n\}$;*

C2. $g_0(n, k, \mathbf{n}) = g_0(n, k)$*;*

C3. $h_j(n, k, \mathbf{n}) = h_j(n, n_j)$*, for any $j = 1, \dots, k$.*

If assumptions $C1 - C3$ hold, then there exist $\sigma \in [0, 1)$, $\vartheta > -\sigma$ and $c_n \geq 0$, such that:

i) if $k \geq 2$ then

$$g_0(n, k) = \frac{\vartheta + k\sigma}{\vartheta + n}, \qquad h_j(n, n_j) = \frac{n_j - \sigma}{\vartheta + n};$$

ii) if $k = 1$ then

$$g_0(n, k) = \frac{\vartheta + k\sigma}{\vartheta + n} - c_n, \qquad h_j(n, n_j) = \frac{n_j - \sigma}{\vartheta + n} + c_n.$$

Investigating the assumptions of Proposition 7 will help us understanding the properties of the PY process. Assumption C1 states that, a priori, all partitions have positive probability. Asumption C2 implies that the probability of sampling a new value, in the predictive distribution of a Pitman-Yor process, does not depend on the composition $\mathbf{n}$. From C3 it follows that the probability of observing a value X_j^* already appeared in the sample, is independent of the number k of already observed atoms and depends on $\mathbf{n}$ only through n_j.

Finally, c_n plays a penalization role when only one atom has appeared in the sample. When we set $c_n = 0$ for every $n \geq 1$, the predictive distribution becomes

$$\Pr[X_{n+1} \in \cdot \mid X^{(n)}] = \frac{\vartheta + k\sigma}{\vartheta + n} P_0(\cdot) + \frac{1}{\vartheta + n} \sum_{j=1}^{k} (n_j - \sigma) \delta_{X_j^*}(\cdot),$$

which corresponds to the predictive distribution of a PY process, given a sample $X^{(n)}$.
In a recent article Bacallado et al. (2017) proved an extension of the result in Proposition 7 to a more general case.

Proposition 8 (Proposition 1 in Bacallado et al., 2017)**.** *Let $\tilde{p}$ be an arbitrary species sampling model, with predictive probability as in Equation 4.3, allowing for either an infinite number of atoms or for a random number T of atoms, whit T supported in $\mathbb{N}$. Let Ψ_n be the random partition of $\{1, \dots, n\}$ induced by $\tilde{p}$. Consider the following assumptions:*

D1. $\Pr[\Psi_n = \psi_n] > 0$*, for any ψ_n partition of $\{1, \dots, n\}$;*

D2. $g_0(n, k, \mathbf{n}) = g_0(n, k)$*;*

D3. $h_j(n, k, \mathbf{n}) = h_j(n, k, n_j)$*, for any $j = 1, \dots, k$.*

If the assumptions $D1 - D3$ hold, then there exist a parameter $\sigma < 1$ and a collection of nonnegative weights $\{V_{n,k}, n \geq 1, 1 \leq k \leq n\}$, with $V_{1,1} = 1$ and satisfying the recursion rule $V_{n,k} = V_{n+1,k+1} + V_{n+1,k}(n - k\sigma)$, such that

$$g_0(n,k) = \frac{V_{n+1,k+1}}{V_{n,k}}, \qquad h_j(n,k,n_j) = \frac{V_{n+1,k}}{V_{n,k}}(n_j - \sigma), \tag{4.4}$$

for any $j = 1, \dots, k$.

A species sampling model $\tilde{p}$ for which holds (4.4), is named Gibbs-type prior, or Gibbs-type species sampling model, with discount parameter $\sigma < 1$.
Assumptions $D1 - D3$ are similar to assumptions $C1 - C3$, with the only difference involving $C3$ and $D3$, with $D3$ relaxing the assumption of the probability of an already observed atom being independent of k.
For the rest of this chapter, we will focus on the family of the Gibbs-type priors with discount parameter σ in $[0, 1)$, a family of species sampling models characterized by an infinite number of atoms.
Figures 4.1 and 4.2 show, with an example, two sampling schemes, the first one corresponding to the predictive distribution of a generic species sampling model, the second one to the predictive distribution of a Gibbs-type prior. Figure 4.2 nicely illustrates the two-stage sampling scheme characterizing Gibbs-type priors, where the next observation can either be a new species or an already observed one; conditionally on the fact that it is an already observed one, the probability of a specific unique value X_j^* depends only on its frequency n_j, the sample size n and the discount parameter σ.
Gibbs-type priors give rise to a conveniently simple EPPF (Pitman, 1995, see Section 1.2 for details). Next, we report the definition of Gibbs type random partitions.

Definition 7 (Gnedin and Pitman, 2006)**.** *An exchangeable random partition Ψ of the natural numbers is said to be of the Gibbs-type form if there exist a sequence $\{V_{n,k}\}_{n,k \geq 1}$, where $n, k \in \mathbb{N}$ and $1 \leq k \leq n$, such that for each composition $(n_1, \dots, n_k) \in \triangle_n^k$ the EPPF satisfies*

$$p_k^{(n)}(n_1, \dots, n_k) = V_{n,k} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1}, \tag{4.5}$$

where for $\{V_{n,k}\}$ holds the recursive identity $V_{n,k} = V_{n+1,k+1} + V_{n+1,k}(n - k\sigma)$, with the proviso that $V_{1,1} = 1$.

Note that the EPPF in Equation 4.5 satisfies the three EPPF properties stated in Section 1.2. As for the first property, since $V_{1,1} = 1$, it is immediate to check that $p_1^1(1) = V_{1,1}(1 - \sigma)_0 = 1$.

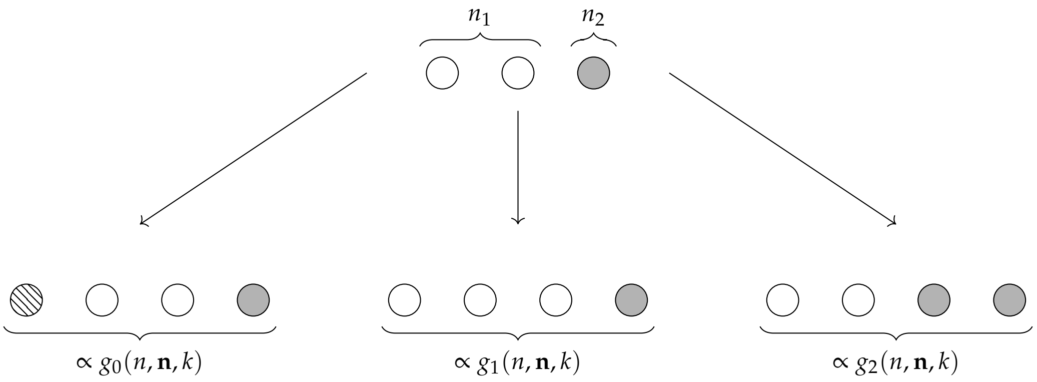
n_1
n_2
$\propto g_0(n, \mathbf{n}, k)$
$\propto g_1(n, \mathbf{n}, k)$
$\propto g_2(n, \mathbf{n}, k)$

FIGURE 4.1: A step of a sampling from the predictive distribution of a general species sampling model ($n = 3$, $k = 2$, $n_1 = 2$, $n_2 = 1$).

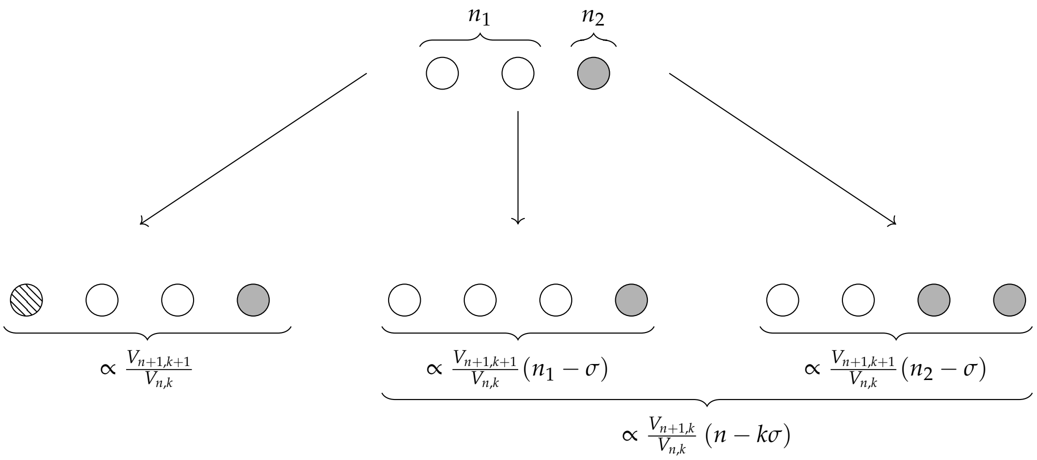
n_1
n_2
$\propto \frac{V_{n+1,k+1}}{V_{n,k}}$
$\propto \frac{V_{n+1,k+1}}{V_{n,k}}(n_1 - \sigma)$
$\propto \frac{V_{n+1,k+1}}{V_{n,k}}(n_2 - \sigma)$
$\propto \frac{V_{n+1,k}}{V_{n,k}}\,(n - k\sigma)$

FIGURE 4.2: A step of a sampling from the predictive distribution of a Gibbs-type prior ($n = 3$, $k = 2$, $n_1 = 2$, $n_2 = 1$).

Moreover, let $\rho : \mathbb{N}_k \to \mathbb{N}_k$ be a permutation of the first k natural numbers. Then the second property follows by observing that

$$p_k^{(n)}(n_{\rho(1)}, \dots, n_{\rho(k)}) = V_{n,k} \prod_{j=1}^{k} (1-\sigma)_{n_{\rho(j)}-1} = V_{n,k} \prod_{j=1}^{k} (1-\sigma)_{n_j-1} = p_k^{(n)}(n_1, \dots, n_k).$$

Finally, as for the third property, observe that, by applying the recursive identity $V_{n,k} = V_{n+1,k+1} + V_{n+1,k}(n - k\sigma)$ we get

$$\begin{aligned} V_{n,k} \prod_{r=1}^{k} (1-\sigma)_{n_r-1} &= \prod_{r=1}^{k} (1-\sigma)_{n_r-1} \left[V_{n+1,k+1} + V_{n+1,k}(n-k\sigma)\right] \\ &= \prod_{r=1}^{k+1} (1-\sigma)_{n_r-1} \left[V_{n+1,k+1} + V_{n+1,k} \sum_{j=1}^{k} (n_j - \sigma)\right] \\ &= V_{n+1,k+1} \prod_{j=1}^{k+1} (1-\sigma)_{n_j-1} + \sum_{j=1}^{k} V_{n+1,k}(n_j - \sigma) \prod_{r=1}^{k} (1-\sigma)_{n_r-1}, \end{aligned}$$

which is equivalent with

$$p_k^{(n)}(n_1,\dots,n_k) = p_{k+1}^{(n+1)}(n_1,\dots,n_k,1) + \sum_{j=1}^{k} p_k^{(n+1)}(\dots,n_j+1,\dots).$$

This proves the validity of the addition rule described in Section 1.2.

Starting from the EPPF function in Equation 4.5, we can derive the predictive distribution for a Gibbs-type prior. Recall that the predictive distribution can be expressed in terms of the EPPF as

$$\begin{aligned} P[X_{n+1} \in \mathrm{d}x \mid X_1,\dots,X_n] &= \frac{p_{k+1}^{(n+1)}(n_1,\dots,n_k,1)}{p_k^{(n)}(n_1,\dots,n_k,1)} P_0(\mathrm{d}x) \\ &+ \sum_{j=1}^{k} \frac{p_k^{(n+1)}(\dots,n_j+1,\dots)}{p_k^{(n)}(n_1,\dots,n_k,1)} \delta_{X_j^*}(\mathrm{d}x), \end{aligned} \tag{4.6}$$

with X_j^* j-th unique value. By substituting the EPPF function (4.5) in Equation 4.6 we obtain

$$P[X_{n+1} \in \mathrm{d}x \mid X_1,\dots,X_n] \propto \frac{V_{n+1,k+1}}{V_{n,k}} P_0(\mathrm{d}x) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{j=1}^{k} (n_j - \sigma)\delta_{X_j^*}(\mathrm{d}x). \tag{4.7}$$

When we add X_{n+1} to the sample, we can interpret $V_{n+1,k+1}/V_{n,k}$ as the probability of moving from a partition of n observations into k blocks to a partition of $n+1$ elements with $k+1$ blocks; similarly $V_{n+1,k}/V_{n,k}\sum_{j=1}^{k}(n_j-\sigma)$, that is $V_{n+1,k}/V_{n,k}(n-k\sigma)$, can be interpreted as the probability of moving from a partition of n observations into k blocks to a partition of $n+1$ observations with the same number k of blocks. The weights in (4.7) are of the form introduced in (4.4), then a Gibb-type partition leads to the predictive distribution of a Gibbs-type prior.

As noted in Lijoi et al. (2007c), the parameter σ regulates a reinforcement mechanism, controlling the concentration of the observations in different blocks of the partition: a large value of σ favors partitions with a large number of groups, with low frequencies in most of the groups and large abundances in a few groups.

Different choices for the sequence $\{V_{n,k}, n \geq 1, 1 \leq k \leq n\}$ lead to different families of processes. When $\sigma \in (0,1)$, following Pitman (2003) (see also Gnedin and Pitman, 2006, for details), it is possible to represent the non-negative weights $V_{n,k}$ in a convenient integral form, as described in the following result.

Proposition 9. *Let $\sigma \in (0,1)$. Then we can express the weights $V_{n,k}$ in Equation 4.7 via the integral representation*

$$V_{n,k} = \frac{\sigma^k}{\Gamma(n-k\sigma)} \int_0^\infty g(t)t^{-k\sigma} \int_0^1 s^{n-1-k\sigma} f_\sigma((1-s)t)\mathrm{d}s\mathrm{d}t,$$

for some function $g(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$, where $f_\sigma(\cdot)$ denotes the density function of a positive σ-stable random variable with Laplace transform $\exp\{-\lambda^\sigma\}$*, for any $\lambda > 0$.*

Thus we can parametrize a Gibbs-type prior by (σ, g, P_0). Particular choices for the function g lead to different processes, within the Gibbs-type family. For example, by setting $g(t) = g^{PY}(t;\sigma,\vartheta) := \sigma\Gamma(\vartheta)t^{-\vartheta}\Gamma(\vartheta/\sigma)$, for any choice of $\vartheta > -\sigma$, we recover the weights of the PY process, that is

$$V_{n,k}^{\mathrm{PY}} = \frac{1}{(\vartheta)_n} \prod_{i=0}^{k-1} (\vartheta + i\sigma). \tag{4.8}$$

The weights for the DP case, specified with total mass $\vartheta > 0$ and base measure P_0, are obtained by letting σ go to zero, which leads to

$$V_{n,k}^{\mathrm{DP}} = \frac{\vartheta^k}{(\vartheta)_n}. \tag{4.9}$$

By setting $g(t) = g^{\mathrm{NGG}}(t; \sigma, \beta) := \exp\{\beta - \beta^{1/\sigma} t\}$, for any $\beta > 0$, we obtain the weights for the NGG process , that is

$$V_{n,k}^{\mathrm{NGG}} = \frac{\mathrm{e}^{\beta} \sigma^{k-1}}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \beta^{\frac{i}{\sigma}} \Gamma\left(k - {}^{i}/_{\sigma}; \beta\right), \tag{4.10}$$

where $\Gamma(z; a) = \int_z^\infty s^{a-1} \mathrm{e}^{-s} \mathrm{d}s$ is the incomplete gamma function. Also in this case, the weights of the Dirichlet process are recovered by letting σ go to zero.

4.2 Cluster size distribution

In the species sampling framework, the predictive structure of the underlying random probability measure plays a key role. Given a sample of exchangeable observations $X_1, \ldots, X_n$, from (4.1), the object of study is the distribution of a set of future observations $X_{n+1}, \ldots, X_{n+m}$ and, when the latter set is not observed, the distribution of the $(n + m + 1)$th observation X_{n+m+1}. A related quantity of interest is the probability of detecting, at the $(n + m + 1)$th observation, species that appeared with any given frequency s in the enlarged, and only partially observed, sample of size $n + m$. The last quantity is known as *m-step s-discovery probability*.

These research questions were exhaustively addressed, for Gibbs-type prior models, in Lijoi et al. (2007b), Favaro et al. (2009) and Favaro et al. (2012b).

More in detail, Favaro et al. (2012b) introduced a Bayesian nonparametric estimator for the m-step s-discovery probability, under the assumption of a Gibbs-type prior model.

Our study aims at contributing to this area of research by providing new insight on the prior distribution of the m-step s-discovery probability, for Gibbs-type priors. In what follows, we set $n = 0$ and consider an unobserved sample of size m. We then denote by S_m the random number of times the species detected at the $(m + 1)$th observation, has appeared in the unobserved sample $X_1, \ldots, X_m$. Outside the species sampling metaphor, S_m can be seen as the random size of the cluster the observation X_{m+1} is assigned to, given that the sample $X_1, \ldots, X_m$ is unobserved.

The main result of our study consists in the derivation a novel and conveniently simple expression for the distribution of S_m. This simplicity of such an expression is the key ingredient for the study of distributional properties of S_m, which, as illustrated in Section 4.4, can be used for parameter elicitation in the context of mixture models with Gibbs-type random mixing measure.

Let $\tilde{p}$ be a random probability measure of Gibbs-type, with support $(\mathbb{X}, \mathscr{X})$, and such that $\mathbb{E}[\tilde{p}] = P_0$. Let $X_1, X_2, \ldots$ be an exchangeable sequence from $\tilde{p}$ and consider $X_1, \ldots, X_m$ unobserved. We define the cluster size distribution p_m as

$$p_m(s) := P[S_m = s] = \mathbb{E}\left[P\left[\sum_{i=1}^{m} \mathbb{1}_{[X_{m+1} = X_i]} = s \mid X_1, \ldots X_m \right] \right], \tag{4.11}$$

for $s \in \{0, 1, \dots, m\}$ and where the expectation in the last expression is with respect to $(X_1, \dots, X_m)$. Observe that, when $s = 0$, $p_m(s)$ coincides with the probability that X_{m+1} takes a new value, not appeared in the unobserved sample $X_1, \dots, X_m$. Next we present the main result of this chapter: it is remarkable that the expression we get for the distribution $p_m(s)$, defined in Equation 4.11, is in terms of only Gibbs weights $V_{j,1}$ with the second index set equal to 1.

Theorem 8. *Let $\tilde{p}$ be a Gibbs-type prior with $\sigma \in [0, 1)$, characterized by a sequence $\{V_{m,k} : m \geq 1, 1 \leq k \leq m\}$ satisfying the recursion identity $V_{m,k} = V_{m+1,k+1} + (m - k\sigma)V_{m+1,k}$. Then for any $s \in \{0, 1, \dots, m\}$ we have*

$$p_m(s) := \sum_{j=s}^{m} \binom{m}{j} \binom{j}{s} (-1)^{j-s} (1 - \sigma)_j V_{j+1,1}. \tag{4.12}$$

Proof. For every non-negative integer m and any $s \in \{0, 1, \dots, m\}$, we define

$$q_m(s) = \sum_{j=s}^{m} \binom{m-s}{j-s} (-1)^{j-s} (1 - \sigma)_j V_{j+1,1} \tag{4.13}$$

and observe that $p_m(s) = \binom{m}{s} q_m(s)$. It is then easy to verify that the triangular identity

$$q_m(s) = q_{m-1}(s) - q_m(s+1) \tag{4.14}$$

holds for any $m \geq 1$ and $s \in \{0, 1, \dots, m-1\}$. We then introduce the quantity $F_m(s)$ defined, for any non-negative integer m and any $s \in \{0, 1, \dots, m\}$, as

$$F_m(s) = (1 - \sigma)_s \sum_{k=0}^{m-s} V_{m+1,k+1} \left\{ \delta_{m,s} + (1 - \delta_{k,0}) \sum_{\pi \in \mathcal{P}_{m-s,k}} \prod_{j=1}^{k} (1 - \sigma)_{m_j - 1} \right\}, \tag{4.15}$$

where $\delta_{i,j}$ denotes the Kroneker delta function and $\mathcal{P}_{m-s,k}$ is the set of all partitions of $\{1, \dots, m - s\}$ into k groups. The proof is composed by the following three steps.

1. Proving that the weights $F_m(s)$ in (4.15) satisfy a triangular identity, analogous to (4.14). That is, for any $m \geq 1$ and $s \in \{0, 1, \dots, m-1\}$,

$$F_m(s) = F_{m-1}(s) - F_m(s+1). \tag{4.16}$$

2. Proving, by using (4.16), that, for any positive m and $s \in \{0, 1, \dots, m\}$, the weights in (4.13) and in (4.15) coincide, that is

$$q_m(s) = F_m(s). \tag{4.17}$$

3. Showing that

$$p_m(s) = \binom{m}{s} F_m(s). \tag{4.18}$$

Combining (4.17) and (4.18) completes the proof.

Step 1.
We denote by $\mathcal{P}_{m,k}$ the set of all partitions of $\{1, \dots, m\}$ into k groups and call ψ a generic

element of $\mathcal{P}_{m,k}$. We observe that

$$\sum_{\psi\in\mathcal{P}_{m,k}}\prod_{j=1}^{k}(1-\sigma)_{m_j-1} = (m-1-k\sigma)\sum_{\psi\in\mathcal{P}_{m-1,k}}\prod_{j=1}^{k}(1-\sigma)_{m_j-1} + \sum_{\psi\in\mathcal{P}_{m-1,k-1}}\prod_{j=1}^{k-1}(1-\sigma)_{m_j-1}$$

and therefore rewrite $F_m(s)$, for $s\in\{0,1,\dots,m-1\}$, as

$$\begin{aligned}
F_m(s) &= (1-\sigma)_s V_{m+1,m-s+1} \\
&\quad - (s+1-\sigma)(1-\sigma)_s \sum_{k=1}^{m-s-1} V_{m+1,k+1} \sum_{\psi\in\mathcal{P}_{m-s-1,k}}\prod_{j=1}^{k}(1-\sigma)_{m_j-1} \\
&\quad + (1-\sigma)_s \sum_{k=1}^{m-s-1} V_{m+1,k+1}(m-(k+1)\sigma) \sum_{\psi\in\mathcal{P}_{m-s-1,k}}\prod_{j=1}^{k}(1-\sigma)_{m_j-1} \\
&\quad + (1-\sigma)_s \sum_{k=2}^{m-s-1} V_{m+1,k+1} \sum_{\psi\in\mathcal{P}_{m-s-1,k-1}}\prod_{j=1}^{k-1}(1-\sigma)_{m_j-1} \\
&= (1-\sigma)_s V_{m+1,m-s+1} \\
&\quad - (1-\sigma)_{s+1} \sum_{k=1}^{m-(s+1)} V_{m+1,k+1} \sum_{\psi\in\mathcal{P}_{m-(s+1),k}}\prod_{j=1}^{k}(1-\sigma)_{m_j-1} \\
&\quad + (1-\sigma)_s \sum_{r=2}^{m-s} V_{m+1,r}(m-r\sigma) \sum_{\psi\in\mathcal{P}_{m-s-1,r-1}}\prod_{j=1}^{r-1}(1-\sigma)_{m_j-1} \\
&\quad + (1-\sigma)_s \sum_{k=2}^{m-s-1} V_{m+1,k+1} \sum_{\psi\in\mathcal{P}_{m-s-1,k-1}}\prod_{j=1}^{k-1}(1-\sigma)_{m_j-1} \\
&= (1-\sigma)_s V_{m+1,m-s+1} - F_m(s+1) + (1-\sigma)_s V_{m+1,m-s}(m-(m-s)\sigma) \\
&\quad + (1-\sigma)_s \sum_{k=2}^{m-s-1} (V_{m+1,k}(m-k\sigma) + V_{m+1,k+1}) \sum_{\psi\in\mathcal{P}_{m-s-1,k-1}}\prod_{j=1}^{k-1}(1-\sigma)_{m_j-1}.
\end{aligned}$$

Exploiting the recursion $\{V_{m,k} = V_{m+1,k+1} + (m-k\sigma)V_{m+1,k}\}$ twice, we get

$$\begin{aligned}
F_m(s) &= -F_m(s+1) + (1-\sigma)_s \sum_{k=2}^{m-s-1} V_{m,k} \sum_{\psi\in\mathcal{P}_{m-s-1,k-1}}\prod_{j=1}^{k-1}(1-\sigma)_{m_j-1} + (1-\sigma)_s V_{m,m-s} \\
&= -F_m(s+1) + (1-\sigma)_s \sum_{k=2}^{m-s} V_{m,k} \sum_{\psi\in\mathcal{P}_{m-s-1,k-1}}\prod_{j=1}^{k-1}(1-\sigma)_{m_j-1} \\
&= -F_m(s+1) + F_{m-1}(s).
\end{aligned}$$

Step 2.
By evaluating (4.13) and (4.15) it is immediate to verify that, for every non-negative m,

$$q_m(m) = (1-\sigma)_m V_{m+1,1} = F_m(m). \tag{4.19}$$

Next, we proceed by induction. For $m=1$, we know by (4.19) that $q_1(1) = F_1(1)$. Moreover, by evaluating (4.13) and (4.15) and exploiting $V_{m,k} = V_{m+1,k+1} + (m-k\sigma)V_{m+1,k}$, we check that

$q_1(0) = V_{2,2} = F_1(0)$ and conclude that, for $m = 1$, (4.17) holds for all $s \in \{0, 1, \ldots, m\}$.
Given a generic positive integer r, we assume that (4.17) holds true for $m = r$ and for every $s \in \{0, 1, \ldots, r\}$. The proof is concluded by showing that $q_{r+1}(s) = F_{r+1}(s)$ for every $s \in \{0, 1, \ldots, r+1\}$. We already know that $q_{r+1}(r+1) = F_{r+1}(r+1)$ by (4.19). When $s = r$, (4.17) allows to write $F_{r+1}(r) = F_r(r) - F_{r+1}(r+1)$. We observe that $q_r(r) = F_r(r)$ by assumption and we already checked that $q_{r+1}(r+1) = F_{r+1}(r+1)$. We conclude that $F_{r+1}(r) = q_r(r) - q_{r+1}(r+1)$, which is in turn equal to $q_{r+1}(r)$ by (4.14). The same argument can be repeated iteratively for all values of $s \in \{r-1, \ldots, 0\}$.
Step 3.
We call $\Psi_m = \{C_1, \ldots, C_k\}$ the random partition of $\{1, \ldots, m\}$ induced by the observations $X^{(m)}$. For any $C \subseteq \{1, \ldots, m\}$, we agree on the notation $\{X_{m+1} \in C\}$ to denote the event $\{X_{m+1} = X_i$ if $i \in C$ and $X_{m+1} \neq X_i$ if $i \in \{1, \ldots, m\} \setminus C\}$. Observe that, for any given subset $C \subseteq \{1, \ldots, m\}$ of size $|C| \in \{0, 1, \ldots, m\}$, we have

$$P[X_{m+1} \in C \mid X^{(m)}] = \begin{cases} \frac{V_{m+1,k+1}}{V_{m,k}} & \text{if } |C| = 0, \\ \frac{V_{m+1,k}}{V_{m,k}}(|C| - \sigma) & \text{if } |C| \neq 0 \text{ and } C \in \Psi_m, \\ 0 & \text{if } |C| \neq 0 \text{ and } C \notin \Psi_m. \end{cases} \tag{4.20}$$

Starting from (4.20) we can derive the marginal probability of $\{X_{m+1} \in C\}$ by marginalizing with respect to all possible partitions in $\mathcal{P}_{m,k}$. We consider three cases.

a. If $|C| = 0$, then

$$\begin{aligned} P[X_{m+1} \in C] &= \sum_{k=1}^{m} \sum_{\psi \in \mathcal{P}_{m,k}} P(X_{m+1} \in C \mid X^{(m)}) p_k^{(m)}(|C_1|, \ldots, |C_k|) \\ &= \sum_{k=1}^{m} V_{m+1,k+1} \sum_{\psi \in \mathcal{P}_{m,k}} \prod_{j=1}^{k} (1-\sigma)_{|C_j|-1}. \end{aligned}$$

b. If $|C| = m$, then

$$\begin{aligned} P[X_{m+1} \in C] &= \sum_{\psi \in \mathcal{P}_{m,1}} P(X_{m+1} \in C \mid X^{(m)}) p_1^{(m)}(m) \\ &= V_{m+1,1}(1-\sigma)_m. \end{aligned}$$

c. If $|C| \neq 0$ and $|C| \neq m$, then

$$\begin{aligned} P[X_{m+1} \in C] &= \sum_{k=1}^{m} \sum_{\psi \in \mathcal{P}_{m,k}} P(X_{m+1} \in C \mid X^{(m)}) p_k^{(m)}(|C_1|, \ldots, |C_k|) \\ &= \sum_{k=1}^{m} \sum_{\psi \in \mathcal{P}_{m,k} \,:\, C \in \psi} V_{m+1,k}(|C| - \sigma) \prod_{j=1}^{k} (1-\sigma)_{|C_j|-1} \\ &= (1-\sigma)_{|C|} \sum_{r=1}^{m-|C|} V_{m+1,r+1} \sum_{\psi \in \mathcal{P}_{m-|C|,r}} \prod_{j=1}^{r} (1-\sigma)_{|C_j|-1}. \end{aligned}$$

By combining a, b and c, we conclude that

$$\begin{aligned} P[X_{m+1} \in C] &= (1-\sigma)_{|C|} \sum_{k=0}^{m-|C|} V_{m+1,k+1} \left\{ \delta_{m,|C|} + (1-\delta_{k,0}) \sum_{\psi \in \mathcal{P}_{m-|C|,k}} \prod_{j=1}^{k} (1-\sigma)_{|C_j|-1} \right\} \\ &= F_m(|C|). \end{aligned} \tag{4.21}$$

Since, by definition $p_m(s) = \sum_{C:|C|=s} P(X_{m+1} \in C)$, we can combine (4.21) and (4.17) to show that

$$p_m(s) = \sum_{C:|C|=s} F_m(|C|) = \binom{m}{s} F_m(s).$$

□

Next, starting from the expression for the cluster size distribution p_m presented in Theorem 8, we study its moments.

Theorem 9. *Let S_m be a discrete random variable with distribution p_m defined as in Equation 4.12. Then, for every positive integer r, the r-th moment is defined as*

$$\mathbb{E}[S_m^r] = \sum_{j=1}^{m} m^{\underline{j}} (1-\sigma)_j \left\{ {r \atop j} \right\} V_{j+1,1}. \tag{4.22}$$

In Theorem 9, we denote by $m^{\underline{j}} = \prod_{s=0}^{j-1}(m-s)$ the falling factorial and by

$$\left\{ {r \atop j} \right\} = \frac{1}{j!} \sum_{s=0}^{j} s^r (-1)^{j-s} \binom{j}{s}$$

a Stirling number of the second kind, such that $\left\{ {r \atop j} \right\} = 0$ for any $j > r$.

Proof. Starting from the weights defined in Equation 4.12 we have.

$$\begin{aligned} \mathbb{E}[S_m^r] &= \sum_{s=0}^{n} s^r p_n(s) \\ &= \sum_{s=1}^{m} s^r \sum_{j=s}^{m} \binom{m}{j} \binom{j}{s} (-1)^{j-s} (1-\sigma)_j V_{j+1,1}, \end{aligned}$$

where the last term in the previous equation could be written, switching the sums, as

$$\begin{aligned} \mathbb{E}[S_m^r] &= \sum_{j=1}^{m} V_{j+1,1} (1-\sigma)_j \binom{m}{j} \sum_{s=1}^{j} s^r \binom{j}{s} (-1)^{j-s} \\ &= \sum_{j=1}^{m} V_{j+1,1} (1-\sigma)_j \frac{m!}{(m-j)!} \frac{1}{j!} \sum_{s=1}^{j} s^r \binom{j}{s} (-1)^{j-s} \\ &= \sum_{j=1}^{m} m^{\underline{j}} (1-\sigma)_j \left\{ {r \atop j} \right\} V_{j+1,1}, \end{aligned}$$

where the last term coincides with Equation 4.22. □

Starting from the previous result, in the next corollary, we display explicit expressions for the first two moments of p_m.

Corollary 1. *Let S be a discrete random variable defined as in Equation 4.12. Then the first two moments are*

$$\mathbb{E}[S_m] = m(1-\sigma)V_{2,1} \tag{4.23}$$

and

$$\mathbb{E}[S_m^2] = m(1-\sigma)V_{2,1} + m(m-1)(1-\sigma)(2-\sigma)V_{3,1}. \tag{4.24}$$

Proof. From Theorem 9, we have

$$\begin{aligned}\mathbb{E}[S_m] &= \sum_{j=1}^{m} m^{\underline{j}}(1-\sigma)_j \left\{ {1 \atop j} \right\} V_{j+1,1} \\ &= m(1-\sigma)V_{2,1} \sum_{s=0}^{1} s \binom{1}{s} (-1)^{j-s} \\ &= m(1-\sigma)V_{2,1},\end{aligned}$$

because $\left\{ {1 \atop j} \right\} = 0$ for any $j > 1$. For the second moment we have

$$\begin{aligned}\mathbb{E}[S_m^2] &= \sum_{j=1}^{m} m^{\underline{j}}(1-\sigma)_j \left\{ {1 \atop j} \right\} V_{j+1,1} \\ &= \sum_{j=1}^{2} m^{\underline{j}}(1-\sigma)_j \left\{ {1 \atop j} \right\} V_{j+1,1} \\ &= \mathbb{E}[S] + m(m-1)(1-\sigma)(2-\sigma)V_{3,1} \frac{1}{2!} \sum_{s=0}^{2} s \binom{1}{s} (-1)^{j-s} \\ &= m(1-\sigma)V_{2,1} + m(m-1)(1-\sigma)(2-\sigma)V_{3,1},\end{aligned}$$

because $\left\{ {2 \atop j} \right\} = 0$ for any $j > 2$.

□

In the next result we introduce a helpful strategy to evaluate, via Monte Carlo, the weights $p_m(s)$, for a generic Gibbs-type prior parametrized by (σ, g, P_0). Such a strategy is similar in spirit to the one presented in Arbel et al. (2017) and turns out to be very convenient when the weights $V_{j,1}$ appearing in (4.12) are not available in closed form. Let $S_{\sigma,1}$ denote a polynomially tilted positive σ-stable random variable with tilting parameter 1, with density function $p(x;\sigma) = \Gamma(1+\sigma)x^{-\sigma} f_\sigma(x)$ (see Devroye, 2009), and $B_{a,b}$ denote a beta random variable with parameters a and b.

Corollary 2. *Let $\sigma \in (0,1)$ and let $\tilde{p}$ be a Gibbs-type prior random probability measure, with parameter (σ, g, P_0). Given a species sampling sequence $\{X_j\}_{j \geq 1}$ modeled as in Equation 4.1, for any $m \geq 0$ and $s \in \{0, 1, \ldots, m\}$,*

$$p_m^g(s) = \frac{(1-\sigma)_s(\sigma)_{m-s}}{\Gamma(s+1)\Gamma(m-s+1)} \mathbb{E}\left[g\left(\frac{S_{\sigma,1}}{B_{m-s+\sigma,1-\sigma+s}} \right) \right], \tag{4.25}$$

where the random variables $S_{\sigma,1}$ and $B_{m-s+\sigma,1-\sigma+s}$ are independent.

Proof. Starting from Equation 4.12 of the weights $p_m(s)$ and substituting the integral representation for the $V_{j+1,1}$'s we have

$$\begin{aligned}
p^g_m(s) &= \sum_{j=s}^{m} \binom{m}{j}\binom{j}{s}(-1)^{j-s}(1-\sigma)_j V_{j+1,1}\\
&= \binom{m}{s}\frac{\sigma}{\Gamma(1-\sigma)}\int_0^\infty g(t)t^{-\sigma}\int_0^1 f_\sigma((1-q)t)q^{-\sigma}\left[\sum_{j=s}^{m}\binom{m-s}{j-s}(-1)^{j-s}q^j\right]dqdt\\
&= \binom{m}{s}\frac{\sigma}{\Gamma(1-\sigma)}\int_0^\infty\int_0^1 g(t)t^{-\sigma}f_\sigma((1-q)t)q^{s-\sigma}(1-q)^{m-s}dqdt\\
&= \binom{m}{s}\frac{\sigma}{\Gamma(1-\sigma)}\int_0^\infty\int_0^1 g(t)t^{-\sigma}f_\sigma(qt)(1-q)^{s-\sigma}q^{m-s}dqdt\\
&= \binom{m}{s}\frac{\sigma}{\Gamma(1-\sigma)}\int_0^\infty\int_0^1 g\left(\frac{w}{q}\right)(w/q)^{-\sigma}f_\sigma(w)(1-q)^{s-\sigma}q^{m-s-1}dqdt\\
&= C_{m,s,\sigma}\int_0^\infty\int_0^1 g\left(\frac{w}{q}\right)\Gamma(1+\sigma)w^{-\sigma}f_\sigma(w)\frac{(1-q)^{s-\sigma}q^{m-s+\sigma-1}}{B_{(m-s+\sigma,s-\sigma+1)}}dqdt\\
&= C_{m,s,\sigma}\mathbb{E}\left[g\left(\frac{S_{\sigma,1}}{B_{m-s+\sigma,1-\sigma+s}}\right)\right],
\end{aligned}$$

and denoting by $B_{(m-s+\sigma,s-\sigma+1)}$ the beta function, the normalizing constant $C_{m,s,\sigma}$ is equal to

$$\begin{aligned}
C_{m,s,\sigma} &= \binom{m}{s}\frac{\sigma B_{(m-s+\sigma,s-\sigma+1)}}{\Gamma(1-\sigma)\Gamma(1+\sigma)}\\
&= \binom{m}{s}\frac{\Gamma(m-s+\sigma)\Gamma(s-\sigma+1)}{\Gamma(1-\sigma)\Gamma(\sigma)\Gamma(m+1)}\\
&= \frac{(1-\sigma)_s(\sigma)_{m-s}}{\Gamma(s+1)\Gamma(m-s+1)},
\end{aligned}$$

where the last term is the constant in Equation 4.25.

□

Realizations of $p_m(s)$ can thus be obtained by sampling independently $B_{m-s+\sigma,s-\sigma+1}$ and $S_{\sigma,1}$. A sample from a polynomially tilted positive σ-stable distribution can be obtained by adopting the efficient strategy presented by Devroye (2009). It is particularly convenient to sample from the distribution $S_{\sigma,1}$ in Equation 4.25, due to the fact that it is independent of both m and s, and therefore, once σ is fixed, realizations of $S_{\sigma,1}$ can be stored and used to evaluate $p_m(s)$ for every $m \geq 1$ and $s \in \{0,1,\ldots,n\}$. The last observation is a convenient by-product of the aforementioned fact that the distribution p_m in (4.12) is expressed in terms of Gibbs weights $V_{j,1}$ with the second index fixed equal to 1.

4.3 Special cases: DP, PY and NGG

While in the previous section we studied distributional properties of the general class of Gibbs-type priors, here we consider how the same results specialize when popular examples within the Gibbs family are considered.

We start by combining the weights defined in Equation 4.9, Equation 4.8 and Equation 4.10, with the result in Theorem 8, to study the cluster size distribution for the DP, PY and NGG, respectively.

Corollary 3. *Let $\{X_n\}_{n\geq 1}$ be a sequence governed by a Gibbs-type prior $\tilde{p}$ and the distribution p_m be defined as in (4.11). Then the followings hold.*

i) If $\tilde{p}$ is a DP, then for any $m \geq 0$ and $s \in \{0, 1, \dots, m\}$

$$p_m^{DP}(s) = \frac{\vartheta(m-s+1)_s}{(\vartheta+m-s)_{s+1}}.$$

i) If $\tilde{p}$ is a PY process, then for any $m \geq 0$ and $s \in \{0, 1, \dots, m\}$

$$p_m^{PY}(s) = \binom{m}{s} \frac{(1-\sigma)_s(\vartheta+\sigma)_{m-s}}{(\vartheta+1)_m}. \tag{4.26}$$

i) If $\tilde{p}$ is a NGG process, then for any $m \geq 0$ and $s \in \{0, 1, \dots, m\}$

$$p_m^{NGG}(s) = \frac{e^{\beta}(1-\sigma)_s}{\Gamma(s+1)\Gamma(m-s+1)} \sum_{i=0}^{m} \binom{m}{i} (-1)^i \beta^{i/\sigma} (-i+\sigma)_{m-s} \Gamma\left(1 - {}^{i}/_{\sigma}; \beta\right). \tag{4.27}$$

Proof. We consider first the PY case, item *ii*) in Corollary 3. By plugging the expression for the weights in Equation 4.8 into the definition of $p_m(s)$ in Equation 4.12, we get

$$\begin{aligned}
p_m^{\mathrm{PY}}(s) &= \sum_{j=s}^{m} \binom{m}{j}\binom{j}{s} (-1)^{j-s}(1-\sigma)_j \frac{\vartheta}{(\vartheta)_{j+1}} \\
&= \binom{m}{s} \sum_{j=s}^{m} \binom{m-s}{j-s} (-1)^{j-s}(1-\sigma)_j \frac{\vartheta\Gamma(\vartheta)}{\Gamma(\vartheta+j+1)} \\
&= \binom{m}{s} \sum_{j=0}^{m-s} \binom{m-s}{j} (-1)^{j}(1-\sigma)_{j+s} \frac{\Gamma(\vartheta+1)}{\Gamma(\vartheta+s+j+1)} \\
&= \binom{m}{s} \frac{\Gamma(\vartheta+1)}{\Gamma(\vartheta+m+1)} \sum_{j=0}^{m-s} \binom{m-s}{j} (-1)^{j}(1-\sigma)_{j+s} \frac{\Gamma(\vartheta+1)}{\Gamma(\vartheta+s+j+1)} \\
&= \binom{m}{s} \frac{(1-\sigma)_s}{(\vartheta+1)_m} \sum_{j=0}^{m-s} \binom{m-s}{j} (-1)^{j}(1-\sigma+s)_j(\vartheta+s+j+1)_{(m-s)-j} \\
&= \binom{m}{s} \frac{(1-\sigma)_s(\vartheta+\sigma)_{m-s}}{(\vartheta+1)_m},
\end{aligned}$$

where the last term coincides with the result stated in Equation 4.26, and follow by observing that $(a+b)_k = \sum_{i=1}^{k} \binom{k}{i}(-1)^i(a+i)_{k-i}(-b)_i$, for any $k \in \mathbb{N}$.
The DP case, item *i*) in Corollary 3, is obtained by setting σ equal to zero in $p_m^{\mathrm{PY}}(s)$, that is

$$\begin{aligned}
p_m^{\mathrm{DP}}(s) &= p_m^{\mathrm{PY}}(s)\Big|_{\sigma=0} \\
&= \binom{m}{s} \frac{(1-\sigma)_s(\vartheta+\sigma)_{m-s}}{(\vartheta+1)_m}\Bigg|_{\sigma=0} \\
&= \binom{m}{s} \frac{(1)_s(\vartheta)_{m-s}}{(\vartheta+1)_m}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\Gamma(m+1)\Gamma(\vartheta+m-s)}{\Gamma(m-s+1)\Gamma(\vartheta+m+1)} \\
&= \vartheta \frac{(m-s+1)_s}{(\vartheta+m-s)_{s+1}}.
\end{aligned}
$$

For the NGG process we have

$$
\begin{aligned}
p_m^{\mathrm{NGG}}(s) &= \sum_{j=s}^{m} \binom{m}{j}\binom{j}{s}(-1)^{j-s}(1-\sigma)_j \frac{e^\beta}{\Gamma(j+1)} \sum_{i=0}^{j} \binom{j}{i}(-1)^i \beta^{i/\sigma}\Gamma(1-{}^{i}\!/\!{}_{\sigma};\beta) \\
&= \sum_{j=s}^{m}\sum_{i=0}^{s-1} \binom{m}{j}\binom{j}{s}(-1)^{j-s}(1-\sigma)_j \frac{e^\beta}{\Gamma(j+1)} \binom{j}{i}(-1)^i \beta^{i/\sigma}\Gamma(1-{}^{i}\!/\!{}_{\sigma};\beta) \\
&+ \sum_{j=s}^{m}\sum_{i=s}^{j} \binom{m}{j}\binom{j}{s}(-1)^{j-s}(1-\sigma)_j \frac{e^\beta}{\Gamma(j+1)} \binom{j}{i}(-1)^i \beta^{i/\sigma}\Gamma(1-{}^{i}\!/\!{}_{\sigma};\beta) \\
&= A + B.
\end{aligned}
$$

We focus on term A and observe that

$$
\begin{aligned}
A &= \sum_{j=s}^{m} \binom{m}{j}\binom{j}{s}(-1)^{j-s}(1-\sigma)_j \frac{e^\beta}{\Gamma(j+1)} \sum_{i=0}^{s-1} \binom{j}{i}(-1)^i \beta^{i/\sigma}\Gamma(1-{}^{i}\!/\!{}_{\sigma};\beta) \\
&= e^\beta \sum_{i=0}^{s-1}(-1)^i \beta^{i/\sigma}\Gamma(1-{}^{i}\!/\!{}_{\sigma};\beta) \sum_{j=s}^{m} \binom{m}{j}\binom{j}{s}\binom{j}{i}(-1)^{j-s}(1-\sigma)_j \frac{1}{\Gamma(j+1)} \\
&= e^\beta \sum_{i=0}^{s-1}(-1)^i \beta^{i/\sigma}\Gamma(1-{}^{i}\!/\!{}_{\sigma};\beta) \frac{\Gamma(m+1)\Gamma(s+1-\sigma)}{\Gamma(i+1)\Gamma(m-s+1)\Gamma(s+1)\Gamma(1-\sigma)\Gamma(s-i+1)} \\
&\times\ {}_2F_1(s-m, s+1-\sigma; s-i+1; 1) \\
&= \frac{e^\beta}{\Gamma(m-s+1)\Gamma(s+1)} \sum_{i=0}^{s-1}(-1)^i \beta^{i/\sigma}\Gamma(1-{}^{i}\!/\!{}_{\sigma};\beta) \\
&\times \frac{\Gamma(m+1)\Gamma(s+1-\sigma)(-i+\sigma)_{m-s}}{\Gamma(i+1)\Gamma(1-\sigma)\Gamma(s-i+1)(s-i+1)_{m-s}} \\
&= \frac{e^\beta(1-\sigma)_s}{\Gamma(m-s+1)\Gamma(s+1)} \sum_{i=0}^{s-1}(-1)^i \binom{m}{i} \beta^{i/\sigma}(-i+\sigma)_{m-s}\Gamma(1-{}^{i}\!/\!{}_{\sigma};\beta),
\end{aligned}
$$

where ${}_pF_q(a_1, \dots a_p; b_1, \dots, b_q; c)$ is the generalized hypergeometric function. In a similar way we work on B and obtain

$$
\begin{aligned}
B &= \sum_{j=s}^{m}\sum_{i=s}^{s} \binom{m}{j}\binom{j}{s}(-1)^{j-s}(1-\sigma)_j \frac{e^\beta}{\Gamma(j+1)} \binom{j}{i}(-1)^i \beta^{i/\sigma}\Gamma(1-{}^{i}\!/\!{}_{\sigma};\beta) \\
&= e^\beta \sum_{i=s}^{m}(-1)^i \beta^{i/\sigma}\Gamma(1-{}^{i}\!/\!{}_{\sigma};\beta) \sum_{j=i}^{m} \binom{m}{j}\binom{j}{s}\binom{j}{i}(-1)^{j-s}(1-\sigma)_j \frac{1}{\Gamma(j+1)}
\end{aligned}
$$

$$
\begin{aligned}
&= e^{\beta}\sum_{i=s}^{m}(-1)^{i}\beta^{i/\sigma}\Gamma(1-{}^{i}/_{\sigma};\beta)\frac{\Gamma(m+1)\Gamma(i+1-\sigma)(-1)^{i-s}}{\Gamma(i+1)\Gamma(m-i+1)\Gamma(s+1)\Gamma(1-\sigma)\Gamma(i-s+1)}\\
&\times {}_2F_1(i-m,i+1-\sigma;i-s+1;1)\\
&= \frac{e^{\beta}}{\Gamma(s+1)}\sum_{i=s}^{m}\binom{m}{i}(-1)^{i}\beta^{i/\sigma}\Gamma(1-{}^{i}/_{\sigma};\beta)\\
&\times (-1)^{i-s}\frac{\Gamma(i+1-\sigma)(-s+\sigma)_{m-i}}{\Gamma(1-\sigma)\Gamma(i-s+1)(i-s+1)_{m-i}}\\
&= \frac{e^{\beta}(1-\sigma)_s}{\Gamma(s+1)\Gamma(m-s+1)}\sum_{i=s}^{m}\binom{m}{i}(-1)^{i}\beta^{i/\sigma}(-i+\sigma)_{m-s}\Gamma(1-{}^{i}/_{\sigma};\beta)\\
&\times (-1)^{i-s}\frac{\Gamma(i+1-\sigma)\Gamma(-i+\sigma)}{\Gamma(1-\sigma+s)\Gamma(-s+\sigma)}\\
&= \frac{e^{\beta}(1-\sigma)_s}{\Gamma(s+1)\Gamma(m-s+1)}\sum_{i=s}^{m}\binom{m}{i}(-1)^{i}\beta^{i/\sigma}(-i+\sigma)_{m-s}\Gamma(1-{}^{i}/_{\sigma};\beta)
\end{aligned}
$$

Now, by combining A and B we have

$$
\begin{aligned}
p_m^{\mathrm{NGG}}(s) &= A+B\\
&= \frac{e^{\beta}(1-\sigma)_s}{\Gamma(m-s+1)\Gamma(s+1)}\sum_{i=0}^{s-1}\binom{m}{i}(-1)^{i}\beta^{i/\sigma}(-i+\sigma)_{m-s}\Gamma(1-{}^{i}/_{\sigma};\beta)\\
&+ \frac{e^{\beta}(1-\sigma)_s}{\Gamma(s+1)\Gamma(m-s+1)}\sum_{i=s}^{m}\binom{m}{i}(-1)^{i}\beta^{i/\sigma}(-i+\sigma)_{m-s}\Gamma(1-{}^{i}/_{\sigma};\beta)\\
&= \frac{e^{\beta}(1-\sigma)_s}{\Gamma(s+1)\Gamma(m-s+1)}\sum_{i=0}^{m}\binom{m}{i}(-1)^{i}\beta^{i/\sigma}(-i+\sigma)_{m-s}\Gamma(1-{}^{i}/_{\sigma};\beta)
\end{aligned}
$$

which coincides with the expression in Equation 4.27.

□

The results stated in Corollary 3 allows us to study the distribution of well known families of processes within the Gibbs family. Figure 4.3 displays the cluster size distribution p_m for a PY process, for different choices of the parameters σ and ϑ.
Observe that when $\vartheta = 1$ and $\sigma = 0$, case corresponding to the DP case with mass equal to one, the cluster size distribution is uniform over its support $\{0,\dots,m\}$. This means that, in this case, given an unobserved sample of size m, all the cluster sizes for the next observation X_{m+1}, are equally likely. For any fixed value σ, it can be appreciated that, when the parameter ϑ grows, the cluster size distribution concentrates its mass on small values; on the contrary, when the parameter ϑ becomes small, the cluster size distribution concentrates its mass on large values of its support.
The parameter σ plays an opposite role: for any fixed ϑ, large values of σ favour small cluster sizes, while small values of σ favour large cluster sizes.
Figure 4.4 shows the cluster size distribution for the NGG process, for different choices of the parameter $\beta = \{0.5, 1, 2\}$, different values of the discount parameter $\sigma = \{0.15, 0.3, 0.45, 0.6\}$, and $m = 10$. The role played by the parameters β and σ is analogous to the role of ϑ and σ in the PY process case. Specifically, for any fixed σ, when β becomes small, the mass in the cluster size distribution moves to large values of its support; when the parameter β grows, the cluster size distribution concentrates its mass on small values of its support. On the contrary, for any

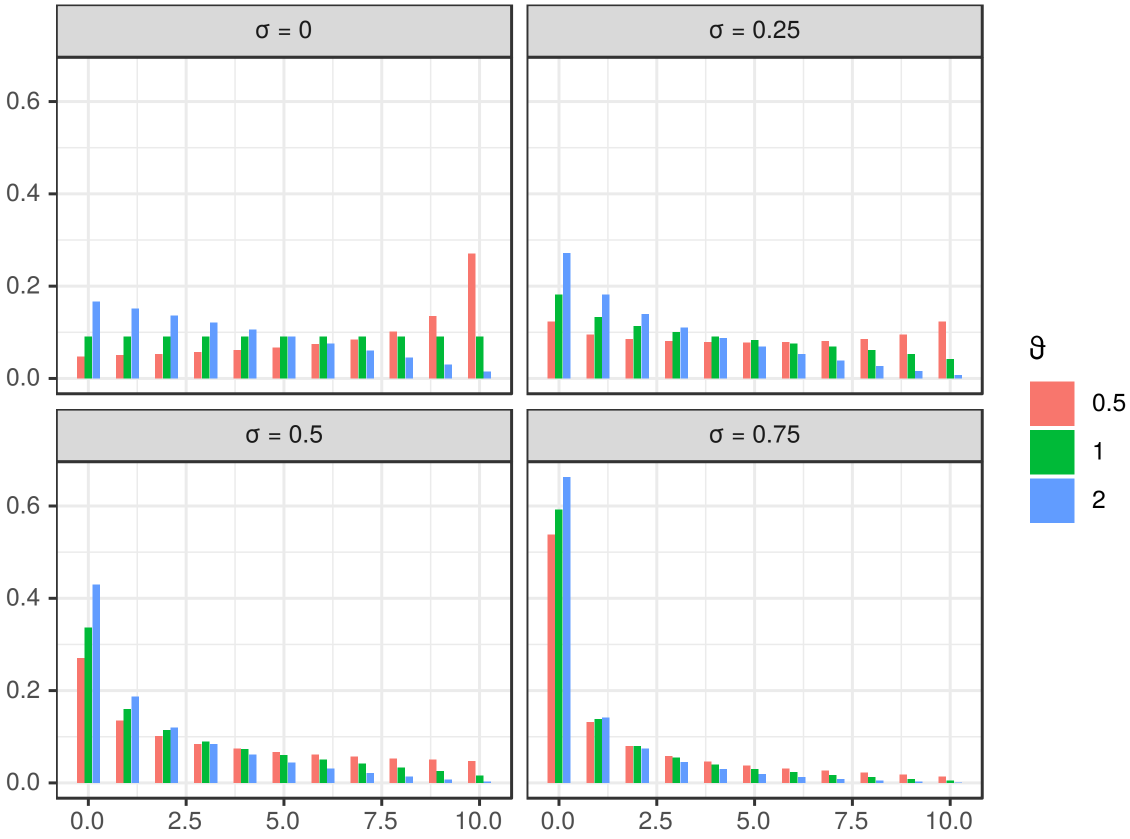

FIGURE 4.3: Cluster size distribution for the Pitman-Yor process, for different values of the strength parameter $\vartheta = \{0.5, 1, 2\}$, different values of the discount parameter $\sigma = \{0, 0.25, 0.5, 0.75\}$, and $m = 10$.

fixed β, if σ becomes large, the mass of the cluster size distribution moves from large values to small values of its support; when σ is small, large cluster sizes are favoured.
Starting from Corollary 1, we derive the expected values and the second moments of the cluster size distribution for the DP, PY and NGG cases.

Corollary 4. *Let $\tilde{p}$ be a Gibbs-type prior and S a discrete random variable with support $\{0, 1, \dots, m\}$ and distribution defined in Equation 4.12. Then the following results hold.*

i) If $\tilde{p}$ is a DP with mass parameter ϑ, then

$$\mathbb{E}[S_m] = \frac{m}{\vartheta + 1} \qquad \text{and} \qquad \mathbb{E}[S_m^2] = \mathbb{E}[S_m]\left(1 + 2\frac{m-1}{\vartheta + 2}\right)$$

ii) If $\tilde{p}$ is a PY process with discount parameter $\sigma \in [0, 1)$ and strength parameter $\vartheta > -\sigma$, then

$$\mathbb{E}[S_m] = \frac{m(1-\sigma)}{\vartheta + 1} \qquad \text{and} \qquad \mathbb{E}[S_m^2] = \mathbb{E}[S_m]\left(1 + \frac{(2-\sigma)(m-1)}{\vartheta + 2}\right)$$

iii) If $\tilde{p}$ is a NGG process with parameters $\sigma \in (0, 1)$ and $\beta > 0$, then

$$\mathbb{E}[S_m] = m(1-\sigma)e^{\beta}\left[\Gamma(1;\beta) - \beta^{1/\sigma}\Gamma\left(1 - {}^{1}/_{\sigma};\beta\right)\right]$$

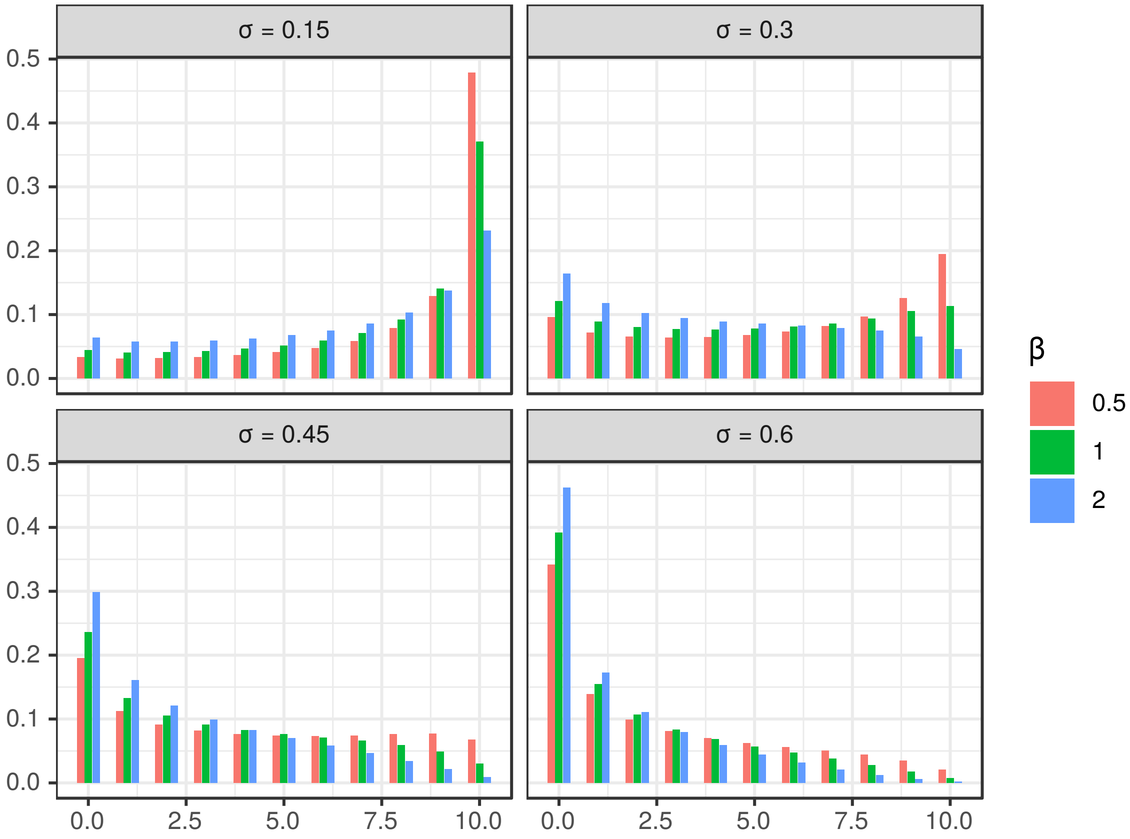

FIGURE 4.4: Cluster size distribution for the normalized generalized gamma process, for different values of $\beta = \{0.5, 1, 2\}$, different values of the discount parameter $\sigma = \{0.15, 0.3, 0.45, 0.6\}$, and $m = 10$.

and

$$\mathbb{E}[S_m^2] = m(1-\sigma)e^{\beta}\left\{\Gamma(1;\beta)\left[1+\frac{(m-1)(2-\sigma)}{2}\right]\right.$$
$$\left.-2\beta^{1/\sigma}\Gamma\left(1-{}^{1}\!/\!{}_{\sigma};\beta\right)+\frac{1}{2}\beta^{1/\sigma}\Gamma\left(1-{}^{2}\!/\!{}_{\sigma};\beta\right)\right\}$$

The proof of the previous Corollary is trivial, and can be obtained by substituting the process-specific weights $V_{j+1,1}$ in Equation 4.23 and Equation 4.24.

Next we study the behaviour of the expected values derived in Corollary 4, for different values of the parameters characterizing the process. Figure 4.5 displays the expected value curves for the PY process as a function of the discount parameter $\sigma \in [0, 1)$, for different choices of the strength parameter $\vartheta \in \{0.5, 1, 2\}$, and for $m = 10$.

Figure 4.5 shows a linear dependence between the expected value of the cluster size distribution and the discount parameter σ. When σ moves closer to the upper extreme of its support, the expected value decreases to zero. Figure 4.6 shows the behaviour of the variance of the cluster size distribution for the PY process, as function of the parameter σ, for the same conditions considered in Figure 4.5. The variance, as a function of the discount parameter σ, displays a quadratic shape, with a scaling effect controlled by the strength parameter ϑ.

Similar observations can be made for the normalized generalized gamma process.

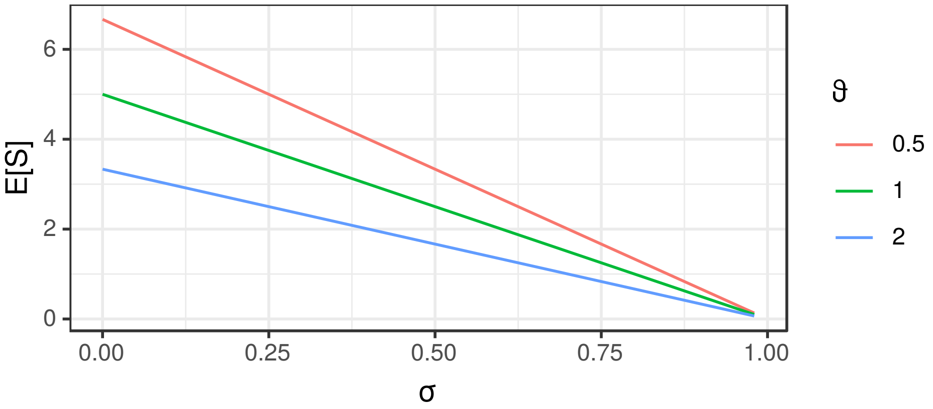

FIGURE 4.5: Expected value of the cluster size distribution for the Pitman-Yor process, as a function of the discount parameter σ ranging in $[0, 1)$, for different values of the strength $\vartheta \in \{0.5, 1, 2\}$, and $m = 10$.

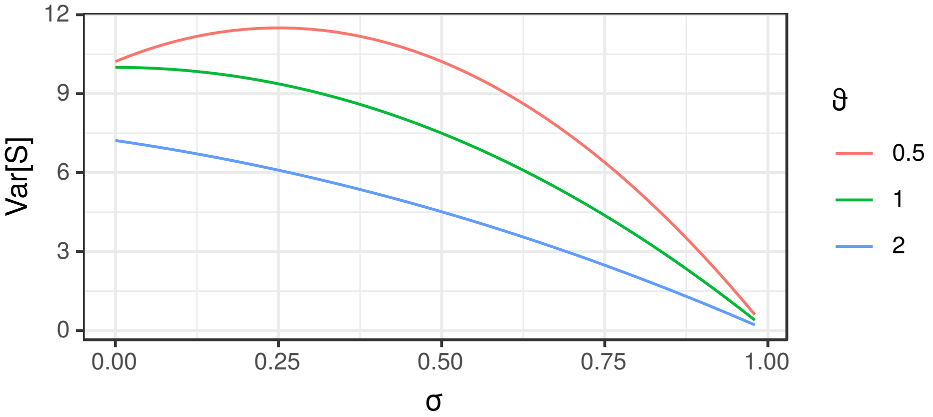

FIGURE 4.6: Variance of the cluster size distribution ffor the Pitman-Yor process, as a function of the discount parameter σ ranging in $[0, 1)$, for different values of the strength $\vartheta \in \{0.5, 1, 2\}$, and $m = 10$.

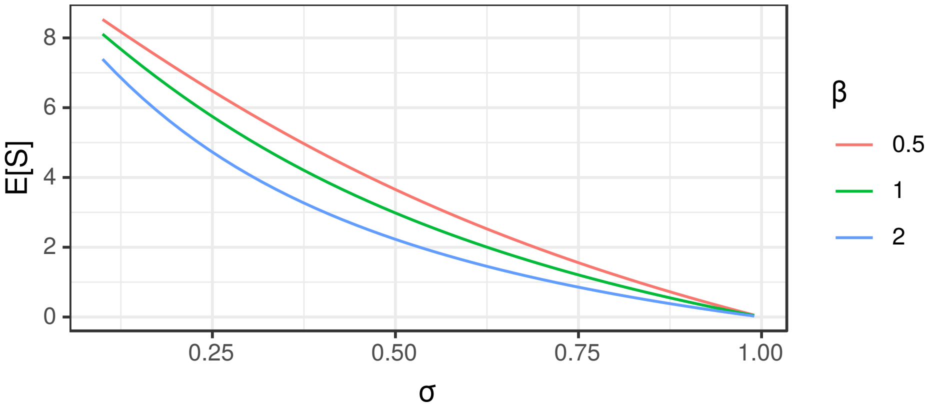

FIGURE 4.7: Expected value of the cluster size distribution for the normalized generalized gamma process, as a function of the discount parameter σ ranging in $(0, 1)$, for different values of the parameter $\beta \in \{0.5, 1, 2\}$ and $m = 10$.

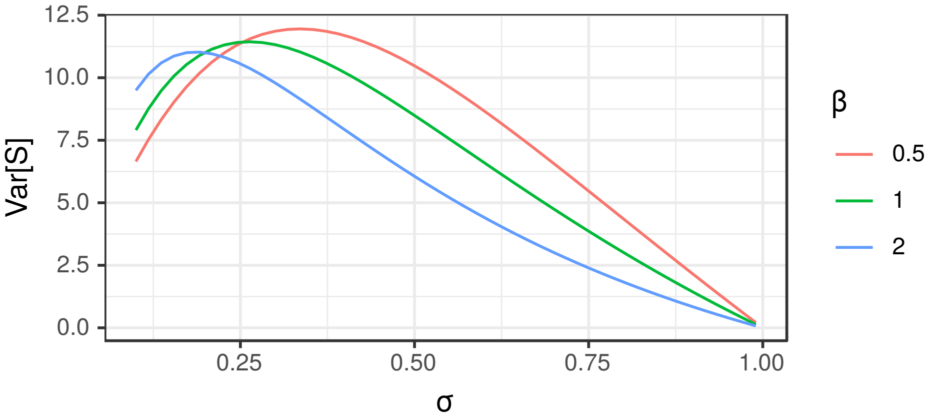

FIGURE 4.8: Variance of the cluster size distribution for the normalized generalized gamma, as a function of the discount parameter σ ranging in $(0,1)$, for different values of the parameter $\beta \in \{0.5, 1, 2\}$ and $m = 10$.

Figures 4.7 and 4.8 show, respectively, the expected value and the variance curves for the NGG process, as functions of the discount parameter $\sigma \in (0,1)$, for different choices of the parameter $\beta \in \{0.5, 1, 2\}$, and for $m = 10$. The expected value curves show a sub-linear shape and seem to approach linearity when β gets small. As for the variance, it interesting to observe that, at least for the considered scenarios, for small values of the discount parameter σ, the variance is an increasing function of β, while for large values of σ, the variance is decreasing in β.

4.4 Elicitation of a Gibbs-type prior parameters

A common problem arising when dealing with nonparametric models is their sensitivity to the parameters characterizing the prior process. Different parameter choices might lead to radically different posterior inferences and, as a result, to different statistical conclusions. As an explanatory example, we consider a dataset of size $n = 150$ generated from a mixture of two Gaussians and we fit to such data a PY mixture model with Gaussian kernel and Normal/Inverse-Gamma base measure, i.e. $P_0 = N(\mu; 0, 5\sigma) \times IG(\sigma; 2, 1)$.

		σ				
		0.0	0.2	0.4	0.6	0.8
ϑ	0.1	2.0	2.0	2.0	2.9	3.4
	0.5	2.0	2.1	2.3	3.0	3.8
	1	2.0	2.1	3.5	3.6	4.1
	5	4.5	7.3	7.2	7.3	8.0
	10	9.8	9.2	10.3	10.9	10.8
	25	16.2	17.8	18.5	18.3	18.1

TABLE 4.1: Number of clusters in the VI estimated partition (see Wade and Ghahramani, 2018, for details) for a PY mixture model. Left to right increasing the discount parameter, top to bottom increasing the strength parameter. Results averaged over 10 replications. Data simulated from a mixture of two Gaussian distribution, $\frac{1}{3}N(-2.5, 1) + \frac{2}{3}N(2.5, 1)$.

Table 4.1 shows the number of clusters of the VI estimated partition (see Wade and Ghahramani, 2018, for details). It is possible to appreciate that the estimated number of clusters is

very sensitive to the choice of the parameters. On top of that, it is important to stress that also the specification of the base measure, and in particular the parameters characterizing its scale component (here kept fixed across all the considered scenarios) might affect the estimated clustering. This aspect will be investigated in Section 4.4.1.
In this section we focus on the class of mixture models with Gibb-type mixing measure and investigate the elicitation of the parameter of the prior process when prior information on the clustering structure underlying the data is available.
Specifically, we want to use Theorem 8 to devise an elicitation strategy for the parameters of $\tilde{p}$, possibly restricting their support to a subset of plausible values. Suppose that a prior belief on the cluster size distribution is available, we aim at integrating such information in the specification of the Gibbs-type mixture model.
Here, we focus on the exemplifying case when the available prior knowledge refers to the prior probability that the $(m+1)$th observation X_{m+1} will be assigned to a cluster with size larger or equal than a given threshold τ_m, in the not yet observed sample $X_1, \dots, X_m$.
Formally, let S_m be a discrete random variable with distribution defined in Equation 4.11, taking values in $\{0, 1, \dots, m\}$. Observe that, for any threshold $\tau_m \in \{0, 1, \dots, m\}$ we have

$$\begin{aligned}
\Pr[S_m \geq \tau_m] &= \sum_{s=\tau_m}^{m} p_m(s) \\
&= \sum_{s=\tau_m}^{m} \sum_{j=s}^{m} \binom{m}{j} \binom{j}{s} (-1)^{j-s} (1-\sigma)_j V_{j+1,1} \\
&= \sum_{s=\tau_m}^{m} \sum_{j=s}^{m} \binom{m}{j} \binom{j}{s} (-1)^{j-s} (1-\sigma)_j V_{j+1,1} \\
&= \sum_{j=\tau_m}^{m} \binom{m}{j} (1-\sigma)_j V_{j+1,1} \sum_{s=\tau_m}^{j} \binom{j}{s} (-1)^{j-s} \\
&= \sum_{j=\tau_m}^{m} \binom{m}{j} \binom{j}{\tau_m} \frac{\tau_m}{j} (-1)^{j-\tau_m} (1-\sigma)_j V_{j+1,1}. \qquad (4.28)
\end{aligned}$$

Based on the last result, we can easily study the probability of any event of the type $\Pr[S_m \in A]$, with $A \subseteq \{0, 1, \dots, m\}$. Nonetheless, for the sake of illustration, here we focus only on events of the type $A = \{\tau_m, \tau_m + 1, \dots, m\}$. The last result can then be used, given a confidence level $p_{\tau_m} \in (0, 1)$, to restrict the parameter space to the subspace such that $\Pr[S \geq \tau_m] \lesseqgtr p_{\tau_m}$.
Given a threshold τ_m, we can study how the probability $\Pr[S_m \geq \tau_m]$ changes as a function of the parameters of a process. Figure 4.9 shows two surfaces corresponding to the probability $P[S_m \geq \tau_m]$ for a PY process, as a function of σ and ϑ, with $m = 30$. The left panel refers to a threshold $\tau_{30} = 10$, the right panel to $\tau_{30} = 20$. Observe that, for fixed values of σ and ϑ, a larger value of τ_m leads to a smaller value for $P[S_m \geq \tau_m]$.
Similar comments can be made for Figure 4.10, where the displayed surfaces correspond to $\Pr[S_m \geq \tau_m]$ for a NGG process, with $m = 30$: the left panel refers to a threshold $\tau_{30} = 10$, the right panel to $\tau_{30} = 20$. Also for the NGG process, when the value of τ_m increases, the surface of $P[S \geq \tau_m]$ assumes smaller values.

4.4.1 Application to the NGC 2419 globular cluster

We already introduced in Section 2.7 an astronomical dataset, referring a portion of sky with $m = 139$ stars, possibly belonging to a globular cluster called NGC 2419 (see Ibata et al.,

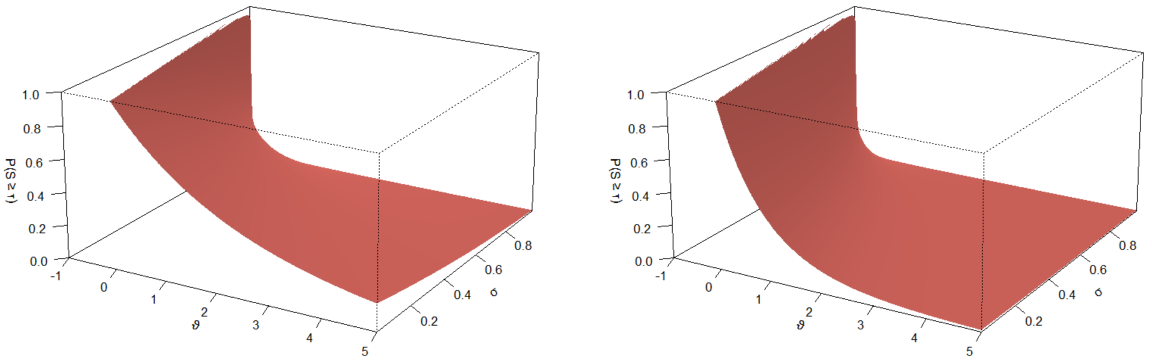

FIGURE 4.9: Probability $\Pr[S_m \geq \tau_m]$ for the Pitman-Yor case, as a function of $\sigma \in [0,1)$ and $\vartheta \in (-\sigma, \ldots, 5)$, for sample size $m = 30$, lower bounds $\tau_{30} = 10$ (left panel) and $\tau_{30} = 20$ (right panel).

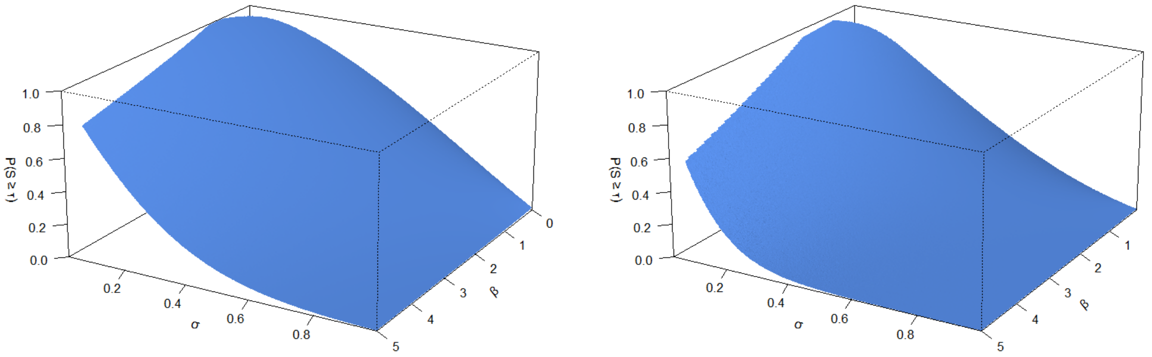

FIGURE 4.10: Probability $\Pr[S_m \geq \tau_m]$ for the normalized generalized gamma, as a function of $\sigma \in [0,1)$ and $\beta \in (0,10)$, for sample size $m = 30$, lower bounds $\tau_{30} = 10$ (left panel) and $\tau_{30} = 20$ (right panel).

2011, for details on the globular cluster). For each star we observe a four-dimensional vector $(Y_1, Y_2, V, [\text{Fe/H}])$, where (Y_1, Y_2) is a two-dimensional projection on the plane of the sky of the position of the star, V is its line of sight velocity and [Fe/H] its metallicity, a measure of the abundance of iron relative to hydrogen. A key question arising with these data consists in identifying the stars that, among the 139 observed, can be rightfully considered as belonging to NGC 2419.

Astronomers expect the large majority of the observed stars to belong to the globular cluster: a priori, they believe that, if an ideal $(m+1)$th star was to be observed in the same portion of sky, it would be very likely for it to belong to a large cluster, that is a cluster of size larger or equal than 100. We formalize this belief by setting the threshold $\tau_{139} = 100$ and the confidence level $p_{100} = 0.9$: we are then ready to implement the proposed elicitation strategy so to obtain a restricted support for the parameters σ and ϑ.

We considered a PY mixture model with a multivariate Gaussian kernel (PYM-G). Let $\phi_d(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$

be a multivariate Gaussian density function, and $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the column vector $\boldsymbol{\mu}$ and the matrix $\boldsymbol{\Sigma}$ represent, respectively, mean vector and covariance matrix of the Gaussian kernel. A PYM-G admits the following hierarchical representation:

$$\begin{gathered}\mathbf{X}_i \mid \boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \overset{\text{ind}}{\sim} \phi_d(\mathbf{x}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),\\ \boldsymbol{\theta}_i \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p},\\ \tilde{p} \sim PY(\vartheta, \sigma, P_0),\end{gathered}$$

where $\sigma \in [0, 1)$ and $\vartheta > -\sigma$.

We choose for the base measure P_0 the product of two distributions: the scale parameter is distributed as an inverse-Wishart distribution while, conditionally to the scale parameter, the location parameter is distributed as a normal distribution. The previous specification, namely a multivariate normal-inverse-Wishart distribution, that is

$$\begin{aligned}P_0(\mathrm{d}\boldsymbol{\mu}, \mathrm{d}\boldsymbol{\Sigma}; \boldsymbol{\pi}) &= P(\mathrm{d}\boldsymbol{\mu} \mid \boldsymbol{\Sigma}; \boldsymbol{\pi}) \times P(\mathrm{d}\boldsymbol{\Sigma}; \boldsymbol{\pi})\\ &= N_d(\mathrm{d}\boldsymbol{\mu}; \mathbf{m}_0, k_0\boldsymbol{\Sigma}) \times IW(\mathrm{d}\boldsymbol{\Sigma}; \nu_0, \mathbf{S}_0),\end{aligned}$$

is jointly conjugate to a Gaussian kernel. By the use of the result in Equation 4.28, we restrict the support of ϑ and σ to a subset of plausible values, based on the definition of the threshold $\tau_{139} = 100$ and the confidence level $p_{100} = 0.9$. We fitted the PYM-G model via the ICS algorithm introduced in Chapter 3, and implemented in the BNPmix R package[1]. By applying the proposed elicitation strategy, we were able to restrict the support of the parameters σ and ϑ to the set $C_{\sigma,\vartheta} := \{(\sigma, \vartheta) : \sigma \in (0, 1), \vartheta \in (-\sigma, \infty), \Pr[S_{139} \geq 100] \geq 0.9\}$, shown in Figure 4.11. We assigned a uniform hyperprior on $C_{\sigma,\vartheta}$ and introduced a griddy-Gibbs step (see Ritter and Tanner, 1992, for details) for the posterior update of σ and ϑ within the GIbbs sampler.

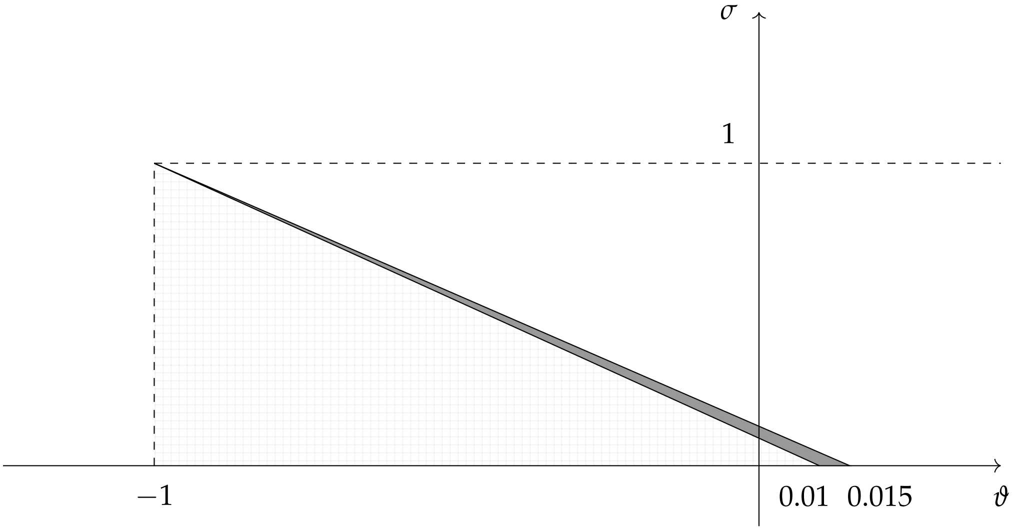

FIGURE 4.11: Restricted parameter space for a PYM-G with $n = 139$, $\tau_{139} = 100$ and $p_{100} = 0.9$, gray filled area.

Posterior inference was drawn based on 5 000 iterations, after 2 500 burn-in iterations. We specified the parameters in the base measure P_0 so that $\boldsymbol{\mu} \mid \boldsymbol{\Sigma} \sim N_4(0, 5\,\boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} \sim IW(\nu_0, \mathbf{S}_0)$, with $(\nu_0, \mathbf{S}_0) = (26, \mathrm{diag}(21))$. The particular choice for the parameters of the scale component

[1] The package is available at https://github.com/rcorradin/BNPmix and can be installed via devtools.

guarantees that $\mathbb{E}[\mathbf{\Sigma}] = \text{diag}(1)$, and $k_0 = 5$ is set so to ensure that the prior on the location component is not strongly concentrated around its mean value.

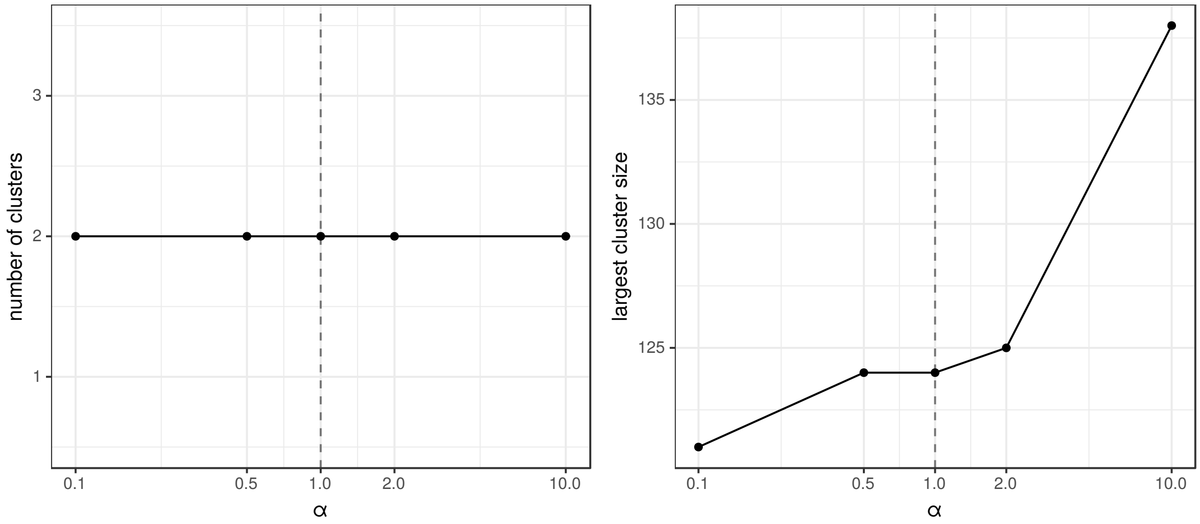

FIGURE 4.12: Analysis of the sensitivity to the values taken by the parameter α controlling the scale component in the base measure of the restricted PYM-G model. Left plot: number of clusters. Right plot: size of the largest cluster. Optimal partitions were estimated by applying Wade and Ghahramani (2018)'s variation of information method.

Different specifications of the base measure might lead to different posterior inference. As far as the clustering of the data is concerned, the parameters characterizing the scale component of the base measure are expected to be particularly relevant. In order to evaluate the effect of such a choice, we performed a sensitivity analysis for the specification of scale matrix characterizing the scale component of the base measure. Specifically, we set $\mathbf{\Sigma} \sim IW(\nu_0, \alpha \mathbf{S}_0)$, with $\nu_0 = 26$ and $\mathbf{S}_0 = \text{diag}(21)$, implying that $\mathbb{E}[\Sigma] = \text{diag}(\alpha)$, and considered a grid of values $\alpha \in \{0.1, 0.2, 1, 5, 10\}$. Figure 4.12 displays, as a function of α, the number of clusters and the size of the largest cluster in the optimal partition of the data obtained by applying Wade and Ghahramani (2018)'s variation of information method. It is interesting to observe that, at least for the range of values we considered, α seems to have no impact on the number of clusters in the optimal partition, which is constantly equal to 2. On the contrary, a clear trend can be appreciated by looking at the size of the largest cluster, with larger values of α leading to a more populated largest cluster. While any sensible statistical analysis should take this sensitivity into account, henceforth, for the sake of illustration, we set $\alpha = 1$, which corresponds to assuming $\mathbf{S}_0 = \text{diag}(\mathbf{21})$.

Figure 4.13 shows the scatter plots of the dataset with individual observations colored according to their membership in the partition estimated based on the variation of information loss function. The estimated partition is composed of two clusters and stars are either labeled as main group and other group. The largest group, identified as the set of stars belonging to the globular cluster, contains 124 stars (gray circles in Figure 4.13), size which is consistent with the prior belief expressed by the astronomers and incorporated into the prior model. The remaining 15 stars (red triangles in Figure 4.13) are thus considered contaminant.

Table 4.2 shows the comparison between the partitions estimated via PYM-G model, with the restricted parameters' space, the partition estimated via DPM-G model and the partition identified in Ibata et al. (2011) (for more details refer to Section 2.7). Interestingly, all the stars which, according to Ibata et al. (2011), belong to or are likely to belong to the globular cluster, fall in

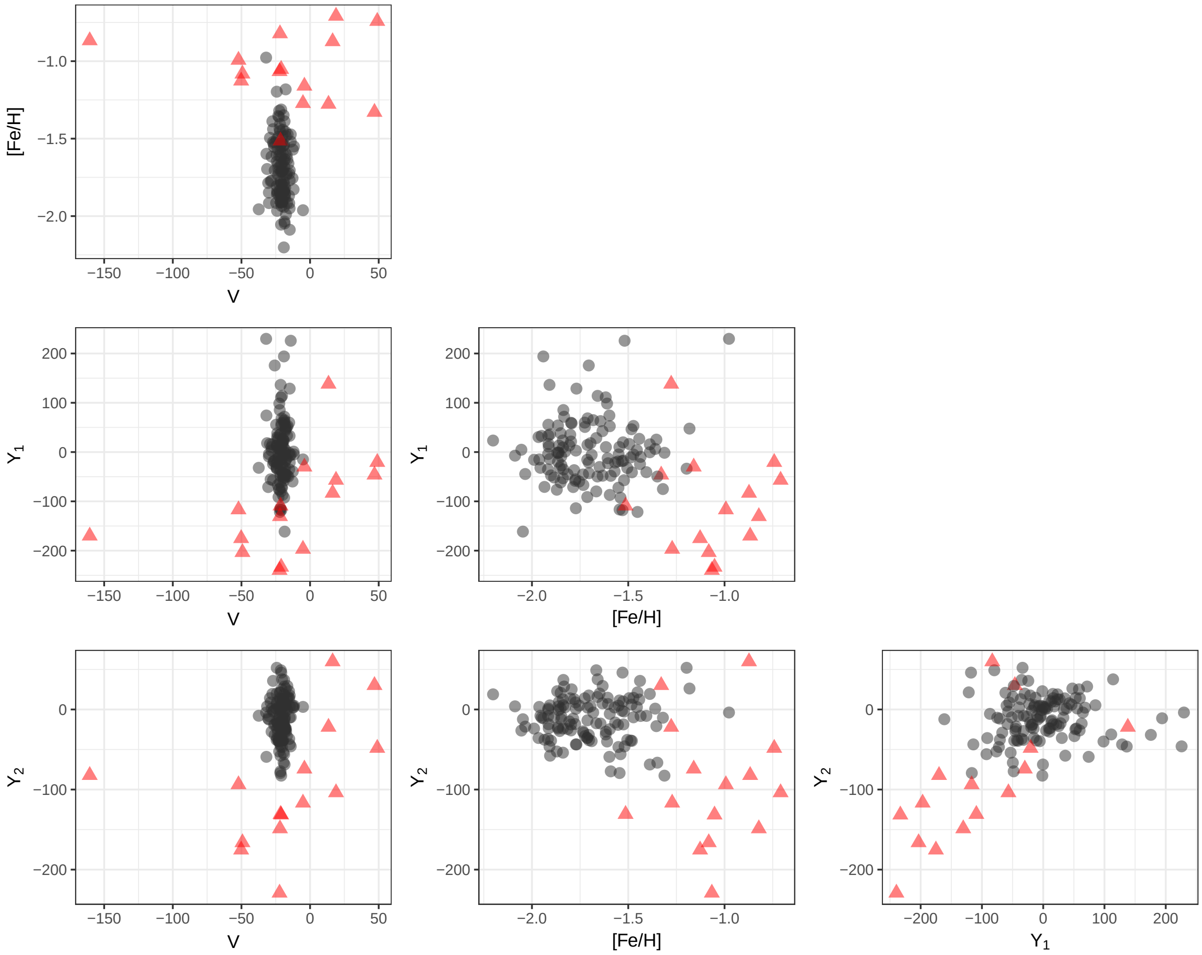

FIGURE 4.13: NGC 2419 data. Partition estimated via PYM-G model, with support for σ and ϑ restricted by the constraint $P[S_{139} \geq 100] \geq 0.9$, combined with Wade and Ghahramani (2018)'s variation of information method.

the largest cluster of the optimal partition estimated based on the restricted PYM-G model. Moreover, if we compare the optimal partition obtained by means of the restricted PYM-G model with that one obtained by means of a DPM-G model, it is possible to appreciate that the restricted PYM-G model nicely gathers into a single group all the stars which are deemed to contaminants, unlike the DPM-G which detected 4 distinct clusters of contaminants.

4.5 Conclusions

In this chapter we studied the distribution of the size of the cluster the $(m+1)$th observation is assigned to, given an unobserved sample of size m, in the context of Gibbs-type priors. Theorem 8, main result of the chapter, provides a conveniently simple expression for the cluster size distribution, which is a function of Gibbs weights only in the form $V_{j,1}$, with the second index equal to 1. We proposed a general representation for such distribution for the (g, σ, P_0) parametrization of Gibbs-type priors. Starting from Theorem 8, we studied the moments of the cluster size distribution, as showed in Corollary 1. We further investigated how the previous quantities specialize when three popular members of the Gibbs family are considered, namely DP, PY and NGG. Specifically, we obtained closed form expressions for the corresponding cluster size distributions in Corollary 3, and for the moments in Corollary 4. Finally, in Section 4.4,

			Ibata et al. groups			DPM-G groups				
			GC	*LGC*	*C*	*largest*	*A*	*B*	*C*	*D*
		total	*118*	*12*	*9*	*124*	*1*	*5*	*1*	*8*
restricted	*largest*	*124*	114	10	0	123	0	0	1	0
PYM-G groups	*A*	*15*	4	2	9	1	1	5	0	8

TABLE 4.2: NGC 2419 data. Comparison between the groups identified by the PYM-G restricted model with the partition described in Ibata et al. (2011) (GC globular cluster, LGC likely globular cluster and C contaminants) and the groups estimated via DPM-G model. PYM-G and DPM-G partitions estimated with Wade and Ghahramani (2018)'s variation of information method.

we considered mixture models with Gibbs-type mixing measure and proposed a convenient strategy for eliciting the parameters of the prior process, based on the distributional properties studied in this chapter. Our strategy is illustrated by analysing the globular cluster NGC 2419 dataset.

Chapter 5

Conclusions

> Every story is organic, and every story finds its own ending.
>
> Tom Coraghessan Boyle, *Author*

We presented a detailed investigation of three aspects concerning the modelling, computational and distributional properties of nonparametric mixtures.

Chapter 1 provides an introduction to the fundamentals of the Bayesian nonparametric area, which allowed us to introduce the main results needed to understand the concepts discussed in the rest of the manuscript.

In Chapter 2 we investigated two aspects of invertible affine transformation of the data within the context of Dirichlet process mixture models with Gaussian kernel (DPM-G). We first considered a finite sample size framework and derived an explicit expression which allows the parameters of a DPM-G model to be specified in such a way that the posterior inference obtained conditionally on transformed data is equivalent with that one obtained conditionally on the original data. The main contribution of Chapter 2 is the derivation of an asymptotic result which, under mild assumptions on the true data generating distribution, guarantees that the posterior distributions obtained, conditionally on a dataset or any affine transformation of it, become more and more similar as the sample size grows. We referred to such property as asymptotic robustness of DPM-G models to affine transformations of the data. In force of this asymptotic result, we carried out an analysis on an astronomical dataset of stars in the field of a globular cluster, by marginally standardizing the data and by choosing a noninformative prior specification for the standardized data.

Chapter 3 introduced a new strategy, named importance conditional sampler (ICS), to estimate Pitman-Yor mixture models, using a conditional approach. The proposed strategy does not rely on the stick-breaking representation of the Pitman-Yor process, thus avoiding the issues inherited from such representation by popular algorithms such as the slice and the retrospective sampler. The ICS was compared with popular algorithms for nonparametric mixture models, by performing an exhaustive simulation study. Unlike other conditional approaches, the efficiency of the ICS does not seem to be affected by the value taken by the discount parameter of the Pitman-Yor process. The same sampling idea was then adopted to efficiently carry out posterior inference based on flexible class of models for partially exchangeable data, obtained by normalizing dependent gamma completely random measures. The performance of the ICS scheme in this context was illustrated by analyzing an astronomical dataset referring to the color distribution of galaxies, stratified by their density.

In Chapter 4 we focused on species sampling problems and studied some prior properties of Gibbs-type priors with the purpose of devising a convenient strategy for parameter elicitation. Specifically we obtained a simple expression for the cluster size distribution of the next observation, given an unobserved exchangeable sample distributed according to a Gibbs-type prior. In addition, we could characterize the moments of such distribution and obtain explicit expression for special processes in the family of Gibbs-type priors. Finally we described a simple strategy for eliciting the parameters of Gibbs-type priors by exploiting our findings, and we illustrated such procedure by analyzing an astronomical dataset referring to stars in the field of a globular cluster. Based on the experts' opinion regarding the size of the largest cluster, we were able to restrict the process' parameter space for a Pitman-Yor mixture model to a subset of plausible values.

Appendix A

Completely Random Measure

Completely random measures (CRM) were first introduced by Kingman (1967). Thorough overviews can be found in Kingman (1993) and Daley and Vere-Jones (2008). For an insightful discussion on the role of CRMs in Bayesian nonparametric statistics, one can refers to the work of Lijoi and Prünster (2010). Let $(\mathbb{X}, \mathscr{X})$ be a measurable space, with $\mathbb{X}$ Polish space and $\mathscr{X} = \mathcal{B}(\mathbb{X})$ the corresponding Borel σ-algebra.

Definition 8. *We say that a measure μ on $(\mathbb{X}, \mathscr{X})$ is boundedly finite if $\mu(A) < \infty$ for any bounded set $A \in \mathscr{X}$.*

We denote by $\mathcal{M}_{\mathbb{X}}$ the space of boundedly finite measures over the measurable space $(\mathbb{X}, \mathscr{X})$, with $\mathscr{M}_{\mathbb{X}} = \mathcal{B}(\mathcal{M}_{\mathbb{X}})$ standing for the corresponding Borel σ-algebra.

Definition 9. *A measurable map μ from a probability space $(\Omega, \mathscr{B}, \mathbb{P})$ into $(\mathcal{M}_{\mathbb{X}}, \mathscr{M}_{\mathbb{X}})$ is named completely random measure (CRM) on $\mathbb{X}$ if the random variables defined by $\mu(A_1), \dots, \mu(A_k)$ are mutually independent, for any pairwise disjoint sets $A_1, \dots, A_k \in \mathbb{X}$.*

Kingman (1993) proved that a CRM is almost sure discrete, which implies that any realization of a CRM is a discrete measure with probability 1. A CRM μ defined on $(\mathbb{X}, \mathscr{X})$ can always be decomposed into two components, μ_c and μ_0:

$$\mu = \mu_c + \mu_0 = \sum_{i=1}^{\infty} W_i \delta_{X_i} + \sum_{i=1}^{M} J_i \delta_{x_i}, \tag{A.1}$$

where μ_c is still a completely random measure, with both random jumps W_i and random locations X_i, while μ_0 is a measure with random jumps J_i and fixed locations x_i. Moreover μ_c and μ_0 are independent. The first component μ_c in the representation described in Equation A.1 is also characterized by the *Lévy-Khintchine* representation:

$$\mathbb{E}\left[e^{-\int_{\mathbb{X}} f(x)\mu_c(dx)}\right] = \exp\left\{-\int\limits_{\mathbb{R}^+\times\mathbb{X}} [1 - e^{-sf(x)}]\nu(ds, dx)\right\}, \tag{A.2}$$

with $f : \mathbb{X} \to \mathbb{R}$ measurable function such that $\int |f| d\mu_c < \infty$ almost surely. A CRM as μ_c can be fully identified by the measure ν in Equation A.2, ν is named *Lévy intensity* of μ_c and it is a measure on $\mathbb{R}^+ \times \mathbb{X}$ such that

$$\int\limits_{B\times\mathbb{R}^+} \min\{s, 1\}\nu(ds, dx) < \infty, \tag{A.3}$$

for any set $B \in \mathscr{X}$. The measure ν contains all the information about the distribution of jumps and locations of the CRM μ_c. If the measure ν admits the following representation

$$\nu(ds, dx) = \rho(ds)\alpha(dx), \tag{A.4}$$

then jumps and locations have independent distributions and μ_c is called *homogeneous*.

Appendix B

Marginal distribution of DPM

Let $\tilde{p}$ follows a Dirichlet process distribution. Let $X_1, \dots, X_n$ be a n-size sample from a model with hierarchical specification

$$\begin{aligned} X_i \mid \theta_i &\sim k(X_i, \theta_i) \\ \theta_i \mid \tilde{p} &\sim \tilde{p} \\ \tilde{p} &\sim DP(\alpha, P_0). \end{aligned}$$

We want to marginalize out the distribution of $\tilde{p}$ in the joint distribution of $(\theta, \mathbf{X}^{(n)}, \tilde{p})$:

$$\mathbb{E}_{\tilde{p}}[\mathcal{L}(\boldsymbol{\theta}, \mathbf{X}^{(n)}, \tilde{p})] = \mathcal{L}(\boldsymbol{\theta}, \mathbf{X}^{(n)})$$

in the context of DPM, where $\tilde{p}$ is a DP, $\boldsymbol{\theta}$ a vector of atoms and $\mathbf{X}^{(n)}$ a vector of data. We have that

$$\begin{aligned} \mathbb{E}_{\tilde{p}}\left[\prod_{i=1}^{n} k(\mathbf{X}_i, \theta_i)\tilde{p}(\mathrm{d}\theta_i)\right] &= \mathbb{E}_{\tilde{p}}\left[\prod_{i=1}^{n} k(\mathbf{X}_i, \theta_i)\prod_{i=1}^{n} \tilde{p}(\mathrm{d}\theta_i)\right] \\ &= \mathbb{E}_{\tilde{p}}\left[\prod_{i=1}^{n} k(\mathbf{X}_i, \theta_i)\prod_{j=1}^{k} \tilde{p}(\mathrm{d}\theta_j^*)^{n_j}\right] = \mathbb{E}_{\tilde{p}}\left[\prod_{j=1}^{k} \tilde{p}(\mathrm{d}\theta_j^*)^{n_j}\prod_{i\in C_j} k(\mathbf{X}_i, \theta_j^*)\right] \end{aligned}$$

where $\boldsymbol{\theta}^*$ is the vector of unique values of $\boldsymbol{\theta}$ and n_j is the frequency of θ_j^* in $\boldsymbol{\theta}$, or equivalently the number of observation belong to the block in C_j of the induced partition. A DP could be expressed in term of normalization of a Gamma CRM μ (see A), so it is possible to write a DP as $\frac{\mu(d\theta_j^*)}{\mu(\Theta)}$, where Θ is the whole support of the process (i.e. the support of the base measure). Moreover the value $\mathbb{E}_\mu\left[e^{-\mu(t)}\right]$ with $\mu(t) = \int_\Theta t(x)\mu(\mathrm{d}x)$ is known. Then, following the previous equation, we have

$$\begin{aligned} &= \mathbb{E}_{\tilde{p}}\left[\prod_{j=1}^{k} \frac{\mu(\mathrm{d}\theta_j^*)^{n_j}}{\mu(\Theta)^{n_j}}\prod_{i\in C_j} k(\mathbf{X}_i, \theta_j^*)\right] \\ &= \mathbb{E}_{\mu}\left[\prod_{j=1}^{k} \mu(\mathrm{d}\theta_j^*)^{n_j}\prod_{i\in C_j} k(\mathbf{X}_i, \theta_j^*)\frac{1}{\Gamma(n)}\int_0^{+\infty} u^{n-1}e^{-u\mu(\mathbb{X})}\mathrm{d}u\right] \end{aligned}$$

$$= \frac{1}{\Gamma(n)} \int\limits_0^{+\infty} u^{n-1} \mathbb{E}_\mu \left[\prod_{j=1}^{k} \mu(\mathrm{d}\theta_j^*)^{n_j} e^{-u\mu(\mathbb{X})} \right] \prod_{i \in C_j} k(\mathbf{X}_i, \theta_j^*) \mathrm{d}u$$

Note that by the additive property of $\mu(\cdot)$, by setting $\Theta^* = \Theta \setminus \{\theta_1^*, \dots, \theta_k^*\}$, the measure μ evaluated over the entire support can be decomposed as $\mu(\Theta) = \mu(\Theta^*) + \mu(\mathrm{d}\theta_1^*) + \cdots + \mu(\mathrm{d}\theta_k^*)$. Back on the main equation, we have

$$= \frac{1}{\Gamma(n)} \int\limits_0^{+\infty} u^{n-1} \mathbb{E}_\mu \left[\left(\prod_{j=1}^{k} \mu(\mathrm{d}\theta_j^*)^{n_j} e^{-u\mu(\mathrm{d}\theta_j^*)} \right) e^{-u\mu(\Theta^*)} \right] \mathrm{d}u \prod_{i \in C_j} k(\mathbf{X}_i, \theta_j^*)$$

$$\overset{\substack{ind \\ inc}}{=} \frac{1}{\Gamma(n)} \int\limits_0^{+\infty} u^{n-1} \prod_{j=1}^{k} \mathbb{E}_\mu \left[\mu(\mathrm{d}\theta_j^*)^{n_j} e^{-u\mu(\mathrm{d}\theta_j^*)} \right] \underbrace{\mathbb{E}_\mu \left[e^{-u\mu(\Theta^*)} \right]}_{A} \mathrm{d}u \prod_{i \in C_j} k(\mathbf{X}_i, \theta_j^*)$$

We can rewrite A as $A = \mathbb{E}_\mu \left[e^{-u \int_{\mathbb{X}} \mathbb{1}_{[\mathbb{X}^*]}(x)\mu(dx)} \right] = \mathbb{E}_\mu \left[e^{- \int_{\mathbb{X}} u\mathbb{1}_{[\mathbb{X}^*]}(x)\mu(dx)} \right] = \mathbb{E} \left[e^{-\mu(u\mathbb{1}_{[\mathbb{X}^*]})} \right]$. Then, substituting the decomposition instead of A in the main equation, we have

$$= \frac{1}{\Gamma(n)} \int\limits_0^{+\infty} u^{n-1} \mathbb{E}_\mu \left[e^{-\mu(u\mathbb{1}_{[\Theta^*]})} \right] \prod_{j=1}^{k} \mathbb{E}_\mu \left[(-1)^{n_j} \frac{\partial^{n_j}}{\partial u^{n_j}} e^{-u\mu(\mathrm{d}\theta_j^*)} \right] \mathrm{d}u \prod_{i \in C_j} k(\mathbf{X}_i, \theta_j^*)$$

$$= \frac{1}{\Gamma(n)} \int\limits_0^{+\infty} u^{n-1} \mathbb{E}_\mu \left[e^{-\mu(u\mathbb{1}_{[\Theta^*]})} \right] \prod_{j=1}^{k} (-1)^{n_j} \frac{\partial^{n_j}}{\partial u^{n_j}} \mathbb{E}_\mu \left[e^{-\mu(u\mathbb{1}_{[\mathrm{d}\theta_j^*]})} \right] \mathrm{d}u \prod_{i \in C_j} k(\mathbf{X}_i, \theta_j^*)$$

The *Laplace transform* of a CRM is given by the *Lévy–Khintchine representation*. For the expected value of a Gamma CRM we have that

$$\begin{aligned} &\mathbb{E}_\mu \left[e^{-\mu(u\mathbb{1}_{[\Theta^*]})} \right] \\ &\quad = \exp \left\{ - \int_\Theta log \left(1 + u\mathbb{1}_{[\Theta^*]}(x) \right) \alpha(\mathrm{d}x) \right\} \\ &\quad = \exp \left\{ -\alpha(\Theta^*) \log(1+u) \right\} \\ &\quad = (1+u)^{-\alpha(\Theta^*)} \end{aligned}$$

and in the same way we can obtain $\mathbb{E}_\mu \left[e^{-\mu(u\mathbb{1}_{[\mathrm{d}\theta_j^*]})} \right] = (1+u)^{-\alpha(\mathrm{d}\theta_j^*)}$. Back to the main equation

$$= \frac{1}{\Gamma(n)} \int\limits_0^{+\infty} u^{n-1} (1+u)^{-\alpha(\Theta^*)} \prod_{j=1}^{k} (-1)^{n_j} \frac{\partial^{n_j}}{\partial u^{n_j}} (1+u)^{-\alpha(\mathrm{d}\theta_j^*)} \mathrm{d}u \prod_{i \in C_j} k(\mathbf{X}_i, \theta_j^*)$$

$$
\begin{aligned}
&= \frac{1}{\Gamma(n)} \int_0^{+\infty} u^{n-1}(1+u)^{-\alpha(\Theta^*)} \prod_{j=1}^{k} (-1)^{n_j}(-1)^{n_j} \\
&\quad \underbrace{\alpha(\mathrm{d}\theta_j^*)(\alpha(\mathrm{d}\theta_j^*)+1)\dots(\alpha(\mathrm{d}\theta_j^*)+n_j-1)(1+u)^{-\alpha(\mathrm{d}\theta_j^*)-n_j}}_{c\alpha(\mathrm{d}\theta_j^*)+c\alpha(\mathrm{d}\theta_j^*)^2+\cdots\simeq c\alpha(\mathrm{d}\theta_j^*)=\alpha(\mathrm{d}\theta_j^*)(n_J-1)!} \mathrm{d}u \prod_{i\in C_j} k(\mathbf{X}_i,\theta_j^*) \\
&= \frac{1}{\Gamma(n)} \int_0^{+\infty} u^{n-1}(1+u)^{-\alpha(\Theta^*)} \prod_{j=1}^{k} (1+u)^{-\alpha(\mathrm{d}\theta_j^*)-n_j} \alpha(\mathrm{d}\theta_j^*)(n_j-1)!\mathrm{d}u \prod_{i\in C_j} k(\mathbf{X}_i,\theta_j^*) \\
&= \frac{1}{\Gamma(n)} \int_0^{+\infty} u^{n-1}(1+u)^{-\alpha(\Theta)}(1+u)^{-n}\mathrm{d}u \prod_{j=1}^{k} \alpha(\mathrm{d}\theta_j^*)(n_j-1)! \prod_{i\in C_j} k(\mathbf{X}_i,\theta_j^*) \\
&= \frac{\prod_{j=1}^{k}\Gamma(n_j)}{\Gamma(n)} \prod_{j=1}^{k} \alpha(\mathrm{d}\theta_j^*) \prod_{i\in C_j} k(\mathbf{X}_i,\theta_j^*) \underbrace{\int_0^{+\infty} u^{n-1}(1+u)^{-\alpha(\Theta)-n}\mathrm{d}u}_{B}
\end{aligned}
$$

Setting $v = 1+u$, $u = v-1$, and with $l - \alpha(\mathbb{X}) - n < 1$, we can exploit B as

$$
\begin{aligned}
B &= \int_1^{+\infty} (v-1)^{n-1} v^{-\alpha(\Theta)-n}\mathrm{d}v \\
&= \int_1^{+\infty} \sum_{l=0}^{n-1} \binom{n-1}{l} v^l (-1)^{n-1-l} v^{-\alpha(\Theta)-n}\mathrm{d}v \\
&= \sum_{l=0}^{n-1} \binom{n-1}{l} (-1)^{n-1-l} \int_1^{+\infty} v^{l-\alpha(\Theta)-n}\mathrm{d}v \\
&= \sum_{l=0}^{n-1} \binom{n-1}{l} (-1)^{n-1-l} \frac{v^{l-\alpha(\Theta)-n}}{l-\alpha(\Theta)-n+1}\Bigg|_1^{+\infty} \\
&= \sum_{l=0}^{n-1} \binom{n-1}{l} (-1)^{n-1-l} (l-\alpha(\Theta)-n+1)^{-1} \\
&= \frac{(n-1)!(\alpha-1)!}{(n+\alpha-1)!} \\
&= \frac{\Gamma(n)\Gamma(\alpha)}{\Gamma(\alpha+n)}.
\end{aligned}
$$

Then back to the main equation we have

$$
\begin{aligned}
&= \frac{\prod_{j=1}^{k}\Gamma(n_j)\Gamma(n)\Gamma(\alpha)}{\Gamma(n)\Gamma(\alpha+n)} \prod_{j=1}^{k} \alpha(\mathrm{d}\theta_j^*) \prod_{i\in C_j} k(\mathbf{X}_i,\theta_j^*) \\
&= \prod_{j=1}^{k}\Gamma(n_j) \frac{1}{(\alpha)_n} \prod_{j=1}^{k} \alpha(\mathrm{d}\theta_j^*) \prod_{i\in C_j} k(\mathbf{X}_i,\theta_j^*) \\
&= \mathcal{L}(\boldsymbol{\theta}, \mathbf{X}^{(n)}),
\end{aligned}
$$

where $(\alpha)_n = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}$. Moreover let $\boldsymbol{\theta}_{(i)}$ be the vector $\boldsymbol{\theta} \setminus \theta_i$, we know that

$$\left(\theta_i|\boldsymbol{\theta}_{(i)}, \mathbf{X}^{(n)}\right) \propto \mathcal{L}\left(\mathbf{X}^{(n)}, \boldsymbol{\theta}_{(i)}\right),$$

then the marginal distribution of θ_i becomes

$$\left(\theta_i|\boldsymbol{\theta}_{(i)}, \mathbf{X}^{(n)}\right) \propto \delta_{\Theta^*}(\theta_i)\alpha(\mathrm{d}\theta_i)k(\mathbf{X}_i, \theta_i) + \sum_{j=1}^{k} \delta_{\theta_j}(\theta_i)n_j k(\mathbf{X}_i, \theta_j^*).$$

Appendix C

Posterior distributions

Let X and θ be two random variables, with density $p(Xx \mid \theta)$ and $p(\theta)$ respectively, where θ is a parameter which characterizes the distribution of X. Given an i.i.d. sample $X_1, \dots X_n$ from X, we are interested in the study of the distribution of $\theta \mid X_1, \dots X_n$, where, by the use of Bayes' rule, we have

$$p(\theta \mid \mathbf{X}) \propto p(\theta) p(X_1, \dots, X_n \mid \theta) = p(\theta) \prod_{i=1}^{n} p(X_i \mid \theta). \tag{C.1}$$

C.1 Univariate

Normal and Normal

Let $X \sim N(\mu, \sigma_X^2)$ and $\mu \sim N(m_0, \sigma_\mu^2)$. Suppose that we draw a sample $X_1, \dots, X_n$ i.i.d. to X, then by (C.1) we have

$$\begin{aligned} p(\mu \mid X_1, \dots X_n) &\propto \exp\left\{ -\frac{1}{2} \left[\frac{(\mu - m_0)^2}{\sigma_\mu^2} + \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma_X^2} \right] \right\} \\ &= \exp\left\{ -\frac{1}{2} \left[\frac{\mu^2 - 2\mu m_0 + m_0^2}{\sigma_\mu^2} + \sum_{i=1}^{n} \frac{X_i^2 - 2X_i\mu + \mu^2}{\sigma_X^2} \right] \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \left[\mu^2 \left(\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_X^2} \right) - 2\mu \left(\frac{m_0}{\sigma_\mu^2} + \frac{n\overline{X}}{\sigma_X^2} \right) \frac{\left(\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_X^2} \right)}{\left(\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_X^2} \right)} \right] \right\}, \end{aligned}$$

which identifies a Gaussian distribution, $\mu \mid X_1, \dots X_n \sim N(m_n, \sigma_{\mu,n}^2)$, with $m_n = \sigma_{\mu,n}^2 \left(\frac{m_0}{\sigma_\mu^2} + \frac{n\overline{X}}{\sigma_X^2} \right)$ and $\sigma_{\mu,n}^2 = \left(\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_X^2} \right)^{-1}$.

Normal and Normal-Inverse Gamma

Let $X \sim N(\mu, \sigma^2)$ and $(\mu, \sigma^2) \sim NIG(m_0, k_0, a_0, b_0)$, i.e. $\sigma^2 \sim IG(a_0, b_0)$ and $\mu \mid \sigma^2 \sim N(m_0, k_0\sigma^2)$. Suppose that we draw a sample $X_1, \dots, X_n$ i.i.d. to X. Recalling that

$$\sum_{i=1}^{n} (X_i - \mu)^2 = n(\mu - \overline{X})^2 + \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

and

$$k_0(\mu - m_0)^2 + n(\mu - \overline{X})^2 = (k_0 + n)(\mu - m_n)^2 + \frac{k_0 n}{k_0 + n}(\overline{X} - m_0)^2,$$

with $m_n = \frac{k_0 m_0 + n\overline{X}}{k_0 + n}$, then by (C.1) we can derive the posterior distribution for (μ, σ^2) as

$$\begin{aligned}
&p(\mu, \sigma^2 \mid X_1, \ldots, X_n) \\
&\quad\propto (\sigma^2)^{-\frac{1}{2}}(\sigma^2)^{-a_0 - \frac{n}{2} - 1} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(X_i - \mu)^2 + k_0(\mu - m_0)^2 + 2b_0\right]\right\} \\
&\quad\propto (\sigma^2)^{-\frac{1}{2}}(\sigma^2)^{-a_n - 1} \exp\left\{-\frac{1}{2\sigma^2}\left[k_n(\mu - m_n)^2 + \right.\right. \\
&\qquad\qquad \left.\left. 2b_0 + \sum_{i=1}^{n}(X_i - \overline{X})^2 + \frac{k_0 n}{k_n}(\overline{X} - m_0)^2\right]\right\},
\end{aligned}$$

where we identify the parameters for the posterior distribution

$$\begin{aligned}
k_n &= k_0 + n \\
m_n &= \frac{k_0 m_0 + n\overline{X}}{k_0 + n} \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}\left[\sum_{i=1}^{n}(X_i - \overline{X})^2 + \frac{k_0 n}{k_n}(\overline{X} - m_0)^2\right].
\end{aligned}$$

C.2 Multivariate

Normal and Normal

Let $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma_X)$ and $\boldsymbol{\mu} \sim N(\mathbf{m}_0, \Sigma_\mu)$ d-dimensional random vectors. Suppose that we draw a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ i.i.d. to $\mathbf{X}$, then by (C.1) we have

$$\begin{aligned}
p(\boldsymbol{\mu} \mid \mathbf{X}_1, \ldots \mathbf{X}_n) &\propto \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\mu} - \mathbf{m}_0)^\intercal \Sigma_\mu^{-1}(\boldsymbol{\mu} - \mathbf{m}_0) + \sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})^\intercal \Sigma_X^{-1}(\mathbf{X} - \boldsymbol{\mu})\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\mu}^\intercal \Sigma_\mu^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^\intercal \Sigma_\mu^{-1}\mathbf{m}_0 + n\boldsymbol{\mu}^\intercal \Sigma_X^{-1}\boldsymbol{\mu} - 2n\boldsymbol{\mu}\Sigma_X^{-1}\overline{\mathbf{X}}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\mu}^\intercal(\Sigma_\mu^{-1} + n\Sigma_X^{-1})\boldsymbol{\mu} - \right.\right. \\
&\qquad \left.\left. 2\boldsymbol{\mu}^\intercal\left(\Sigma_\mu^{-1}\mathbf{m}_0 + n\Sigma_X^{-1}\overline{\mathbf{X}}\right)\frac{(\Sigma_\mu^{-1} + n\Sigma_X^{-1})}{(\Sigma_\mu^{-1} + n\Sigma_X^{-1})}\right]\right\}
\end{aligned}$$

which identifies a Gaussian distribution, $\boldsymbol{\mu} \mid \mathbf{X}_1, \ldots \mathbf{X}_n \sim N(\mathbf{m}_n, \Sigma_{\mu,n})$, with $\mathbf{m}_n = \Sigma_{\mu,n}\left(\Sigma_\mu^{-1}\mathbf{m}_0 + n\Sigma_X^{-1}\overline{\mathbf{X}}\right)$ and $\Sigma_{\mu,n} = (\Sigma_\mu^{-1} + n\Sigma_X^{-1})^{-1}$.

Normal and Normal-Inverse Wishart

Let $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ and $(\boldsymbol{\mu}, \Sigma) \sim NIW(\mathbf{m}_0, k_0, \nu_0, S_0)$, i.e. $\Sigma \sim IW(\nu_0, S_0)$ and $\boldsymbol{\mu} \mid \Sigma \sim N(\mathbf{m}_0, k_0\Sigma)$, with $\mathbf{X}$ and $\boldsymbol{\mu}$ d-dimensional random vectors and Σ $d \times d$-dimensional random matrix. Suppose that we draw a sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ i.i.d. to $\mathbf{X}$. Recalling that

$$\sum_{i=1}^{n}(\mathbf{X}_i - \mu)(\mathbf{X}_i - \mu)^\intercal = n(\boldsymbol{\mu} - \overline{\mathbf{X}})(\boldsymbol{\mu} - \overline{\mathbf{X}})^\intercal + \sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^\intercal,$$

and

$$\begin{aligned} &k_0(\boldsymbol{\mu} - \mathbf{m}_0)(\boldsymbol{\mu} - \mathbf{m}_0)^\intercal + n(\boldsymbol{\mu} - \overline{\mathbf{X}})(\boldsymbol{\mu} - \overline{\mathbf{X}})^\intercal \\ &\qquad = (k_0 + n)(\boldsymbol{\mu} - \mathbf{m}_n)(\boldsymbol{\mu} - \mathbf{m}_n)^\intercal + \frac{k_0 n}{k_0 + n}(\overline{\mathbf{X}} - \mathbf{m}_0)(\overline{\mathbf{X}} - \mathbf{m}_0)^\intercal. \end{aligned}$$

Moreover we have

$$\begin{aligned} &\sum_{i=1}^{n}(\mathbf{X}_i - \mu)^\intercal \Sigma^{-1}(\mathbf{X}_i - \mu) \\ &\quad = \sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d}(\mathbf{X}_{ij} - \mu_j)^\intercal \Sigma_{jk}^{-1}(\mathbf{X}_{ik} - \mu_k) \\ &\quad = \sum_{j=1}^{d}\sum_{k=1}^{d}\Sigma_{jk}^{-1}\sum_{i=1}^{n}(\mathbf{X}_{ij} - \mu_j)(\mathbf{X}_{ik} - \mu_k)^\intercal \\ &\quad = \mathrm{Tr}\left[\Sigma_{jk}^{-1}\left(\sum_{i=1}^{n}(\mathbf{X}_{ij} - \mu_j)(\mathbf{X}_{ik} - \mu_k)^\intercal\right)\right] \\ &\quad = \mathrm{Tr}\left[\Sigma_{jk}^{-1}\left(\sum_{i=1}^{n}(\mathbf{X}_{ij} - \overline{\mathbf{X}})(\mathbf{X}_{ik} - \overline{\mathbf{X}})^\intercal + n(\overline{\mathbf{X}} - \boldsymbol{\mu})(\overline{\mathbf{X}} - \boldsymbol{\mu})^\intercal\right)\right]. \end{aligned}$$

Then the posterior distribution become

$$\begin{aligned} p(\boldsymbol{\mu}, \Sigma \mid \mathbf{X}_1, \dots, \mathbf{X}_n) &\propto k_0|\Sigma|^{-1}|\Sigma|^{-\frac{v_0+d+1}{2}}|\Sigma|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\left[\mathrm{Tr}(S_0\Sigma^{-1})\right.\right. \\ &\qquad \left.\left. + k_0(\boldsymbol{\mu} - \mathbf{m}_0)^\intercal\Sigma^{-1}(\boldsymbol{\mu} - \mathbf{m}_0) + \sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})^\intercal\Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu})\right]\right\} \\ &\propto k_0|\Sigma|^{-1}|\Sigma|^{-\frac{v_n+d+1}{2}}\exp\left\{-\frac{1}{2}\left[\mathrm{Tr}\left(\Sigma^{-1}\left(S_0 + \sum_{i=1}^{n}(\mathbf{X}_{ij} - \overline{\mathbf{X}})(\mathbf{X}_{ik} - \overline{\mathbf{X}})^\intercal +\right.\right.\right.\right. \\ &\qquad \left.\left.\left.\left.\frac{k_0 n}{k_0 + n}(\overline{\mathbf{X}} - \mathbf{m}_0)(\overline{\mathbf{X}} - \mathbf{m}_0)^\intercal\right)\right) + (k_0 + n)(\boldsymbol{\mu} - \mathbf{m}_n)^\intercal\Sigma^{-1}(\boldsymbol{\mu} - \mathbf{m}_n)\right]\right\}, \end{aligned}$$

where we identify a Normal-Inverse Wishart distribution with updated parameters

$$\begin{aligned}
k_n &= k_0 + n \\
\mathbf{m}_n &= \frac{k_0 \mathbf{m}_0 + n\overline{\mathbf{X}}}{k_0 + n} \\
\nu_n &= \nu_0 + n \\
S_n &= S_0 + \sum_{i=1}^{n} (\mathbf{X}_{ij} - \overline{\mathbf{X}})(\mathbf{X}_{ik} - \overline{\mathbf{X}})^\intercal + \frac{k_0 n}{k_0 + n} (\overline{\mathbf{X}} - \mathbf{m}_0)(\overline{\mathbf{X}} - \mathbf{m}_0)^\intercal
\end{aligned}$$

C.3 Dirichlet process mass in mixture models

Following the results presented in Escobar and West (1995), introducing an augmentation variable, it is possible to define a conjugate prior for the mass of a Dirichlet process. Let α be the total mass of a DP, let k be the number of blocks of a DP with mass α, given a sample of size n. Assuming a continuous distribution prior distribution for α, and an implied prior $p(k \mid n) = \mathbb{E}\left[p(k \mid \alpha, n)\right]$, we have that

$$p(k \mid \alpha, n) = c_n(k) n! \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \qquad k = 1, \ldots, n,$$

where $c_n(k)$ does not depend on α (see Antoniak, 1974). Except for the terms $c_n(k)n!$, which does not depend on α, n does not appear in the previous equation (see Escobar and West, 1995, for more details). According to (C.1) we have that $p(\alpha \mid k) \propto p(\alpha)p(k \mid \alpha)$. Now suppose that $\alpha \sim Gamma(a_0, b_0)$, parametrized with shape and rate parameters. Recalling that $\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} = \frac{(\alpha+n)\beta(\alpha+1,n)}{\alpha\Gamma(n)}$, with $\beta(\cdot,\cdot)$ the beta function, then for $k = 1, \ldots, n$ we have

$$\begin{aligned}
p(\alpha \mid k) &\propto p(\alpha)\alpha^{k-1}(\alpha + n)\beta(\alpha + 1, n) \\
&\propto p(\alpha)\alpha^{k-1}(\alpha + n) \int_0^1 x^\alpha (1 - x)^{n-1} \mathrm{d}x,
\end{aligned}$$

and the previous implies that $p(\alpha \mid k)$ is the marginal distribution of a joint distribution for (α, η), i.e. $p(\alpha, \eta \mid k) \propto p(\alpha)\alpha^{k-1}(\alpha + n)\eta^\alpha(1 - \eta)^{n-1}$. We can work with the respective conditional distributions, as follows. Let $p(\alpha) \sim Gamma(a_0, b_0)$ be the prior distribution for α. Then we have

$$\begin{aligned}
p(\alpha \mid \eta, k) &\propto \alpha^{a_0+k-2}(\alpha + n) \exp[-\alpha(b_0 - \log(\eta))] \\
&\propto \alpha^{a_0+k-1} \exp[-\alpha(b_0 - \log(\eta))] + n\alpha^{a_0+k-2} \exp[-\alpha(b_0 - \log(\eta))],
\end{aligned}$$

which identifies a mixture of two Gamma distributions, a $Gamma(a_0 + k, b - log(\eta))$ and a $Gamma(a_0 + k - 1, b - log(\eta))$, weighted by ω_η and $1 - \omega_\eta$ respectively, where ω_η is identified by the equation $\omega_\eta/(1 - \omega_\eta) = (a_0 + k - 1)/[n(b_0 - \log(\eta))]$. Finally, $p(\eta \mid \alpha, k) \propto \eta^\alpha(1 - \eta)^{n-1}$, with $0 < \eta < 1$, which implies that $\eta \sim Beta(\alpha + 1, n)$.

Appendix D

Code details

An issues which commonly appears when we are dealing with Bayesian nonparametric models is the time needed to estimate the models. We decide to implement in an efficient way all the routines used in the previous chapters. Considering that the main problem to deal with the BNP modelling is the presence of nested loops, which could be inefficient in high level languages as, for example, `R`, we decided to implement all the algorithms in a low level language, `C++`, to strongly reduce the computational time required for the model estimation.
We decided to base all the implementation of the routines on the use of *void functions*, functions which do not return any result argument and, eventually, point directly the objects in the memory to update them. This approach is particular suitable in the context of Monte Carlo Markov Chain (MCMC) methods, where the procedures are based on an updating rule, and the strategy of working directly with parameters updating instead of function which return arguments reduced the memory use and the computational time.
The `AFFINEpack` R package, available on `GitHub`, contains all the routines used in Chapter 2. The `BNPmix` R package, available on `GitHub`, contains all the routines used for Chapter 3 and Chapter 4. Focus the attention on the second one, the `BNPmix` package contains two main functions: `condMCMC` and `condMCMCmv`, the univariate and the multivariate functions respectively to estimate a Pitman-Yor mixture model.
Common arguments over the two functions are:

- data, a dataset (a vector for `condMCMC` and a matrix for `condMCMCmv`);
- grid, a grid to evaluate the estimated density;
- niter, number of total iterations;
- nburn, number of burn-in iterations;
- method, the sampler considered (`ICS` for importance conditional sampler, `SS` for slice sampler, `MS` for marginal sampler);
- process, `PY` or `DP`;
- mass, the total mass parameter ϑ;
- sigma_PY, the discount parameter σ, if 0 the DP routine is used;
- napprox, number of elements in the importance sampling step for the ICS.

Specific for the `condMCMC` function we have:

- m0, the mean value of the location component of the base measure;
- k0, the weight of the variance in the location component of the base measure;

- a0, shape parameter of the scale component of the base measure;
- b0, scale parameter of the scale component of the base measure.

Specific for the condMCMCmv function we have:

- m0, the mean vector of the location component of the base measure;
- k0, the weight of the covariance matrix in the location component of the base measure;
- n0, gdl of the scale component of the base measure;
- S0, scale matrix of the scale component of the base measure.

Both functions condMCMC and condMCMCmv return an S4 object modCond and modCondMv respectively. The returned object contains all the information regarding the estimation, as

- density, the estimated density;
- clust, the clusters matrix, one vector for each (niter - nburn) iteration;
- nnew, number of new clusters generated for each iteration;
- tot_time, execution time.

We extended also the plot function to the modCond and modCondMv classes.

Acknowledgements

I would like to say thanks to my supervisor. It is hard to find a good supervisor, it is hard to find a good friend, it is difficult if not impossible to find both in the same person.

I thank Julyan and Tony, this manuscript would not have been possible without them.

It is a pleasure to say thanks to my colleagues, in both the universities where I spent most of the time of my Ph.D., in Milano and in Dublin, for the stimulating chats and the pints. Thanks to all the people that I have met at the INRIA in Grenoble: it was just a short time together, but a good one.

I am grateful to my family, who supported me during my studies, with a special mention to my sister Lorella. Thanks to my mom, my uncles and all the support team. Thanks to the cousins' group, Alessandro, Francesca, Marco, Martina and Thomas for the good times together.

I had the pleasure to relocate many times during the last three years, over two different countries, and each time I found friends who are like second families. It is a honor to say thanks to Bianca, Lina and Roberto, who used to share with me great memories in Milano and Andrea C., Caroline, Denisa, Eleni, Elva, Marie and Pedro for making me feel like at home also in Ireland and Federico, for the recent good times back in Monza.

Thanks to Andrea B., Claudio, Guilherme, Francesco and Lucia for the time together in Dublin, it was a pleasure to enjoy the city with you.

Bibliography

[1] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. “Importance Sampling: Intrinsic Dimension and Computational Cost”. In: *Statist. Sci.* 32.3 (2017), pp. 405–431.

[2] C. E. Antoniak. “Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems”. In: *Ann. Statist.* 2.6 (1974), pp. 1152–1174.

[3] J. Arbel and B. Nipoti. “Discussion of “Bayesian Nonparametric Inference Why and How Comment”, by Müller and Mitra”. In: *Bayesian Analysis* 8.02 (2013), pp. 326–328.

[4] J. Arbel and I. Prünster. “A moment-matching Ferguson & Klass algorithm”. In: *Statistics and Computing* 27.1 (2017), pp. 3–17.

[5] J. Arbel, A. Lijoi, and B. Nipoti. “Full Bayesian inference with hazard mixture models”. In: *Computational Statistics & Data Analysis* 93 (2016), pp. 359–372.

[6] J. Arbel, S. Favaro, B. Nipoti, and Y. Whye Teh. “Bayesian nonparametric inference for discovery probabilities: Credible intervals and large sample asymptotics”. In: *Statistica Sinica* 27 (2017), pp. 839–858.

[7] J. Arbel, R. Corradin, and M. Lewandowski. “Discussion of “Bayesian Cluster Analysis: Point Estimation and Credible Balls”, by Wade and Ghahramani”. In: *Bayesian Analysis* 13.2 (2018), pp. 611–612.

[8] J. Arbel, P. De Blasi, and I. Prünster. “Stochastic approximations to the Pitman-Yor process”. In: (2018). arXiv: 1806.10867.

[9] R. Argiento, I. Bianchini, and A. Guglielmi. “A blocked Gibbs sampler for NGG-mixture models via a priori truncation”. In: *Statistics and Computing* 26.3 (2016), pp. 641–661.

[10] R. Argiento, I. Bianchini, and A. Guglielmi. “Posterior sampling from ε-approximation of normalized completely random measure mixtures”. In: *Electron. J. Statist.* 10.2 (2016), pp. 3516–3547.

[11] Y. Ascasibar and J. Binney. “Numerical estimation of densities”. In: *Monthly Notices of the Royal Astronomical Society* 356.3 (2005), pp. 872–882.

[12] S. Bacallado, M. Battiston, S. Favaro, and L. Trippa. “Sufficientness Postulates for Gibbs-Type Priors and Hierarchical Generalizations”. In: 32 (2017), pp. 487–500.

[13] M. L. Balogh, I. K. Baldry, R. Nichol, C. Miller, R. Bower, and K. Glazebrook. “The Bimodal Galaxy Color Distribution: Dependence on Luminosity and Environment”. In: *The Astrophysical Journal Letters* 615.2 (2004), p. L101.

[14] E. Barrios, A. Lijoi, L. E. Nieto-Barajas, I. Prünster, et al. “Modeling with normalized random measure mixture models”. In: *Statistical Science* 28.3 (2013), pp. 313–334.

[15] A. Bean, X. Xu, and S. MacEachern. “Transformations and Bayesian density estimation”. In: *Electronic Journal of Statistics* 10.2 (2016), pp. 3355–3373.

[16] D. A. Binder. “Bayesian cluster analysis”. In: *Biometrika* 65.1 (1978), pp. 31–38.

[17] D. Blackwell and J. B. MacQueen. "Ferguson Distributions Via Polya Urn Schemes". In: *The Annals of Statistics* 1.2 (1973), pp. 353–355.

[18] A. Canale and P. De Blasi. "Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation". In: *Bernoulli* 23.1 (2017), pp. 379–404.

[19] A. Canale and B. Scarpa. "Bayesian nonparametric location–scale–shape mixtures". In: *Test* 25.1 (2016), pp. 113–130.

[20] M. Carlton. "Applications of the Two-Parameter Poisson-Dirichlet Distribution". PhD thesis. UCLA Department of Statistics, 1999.

[21] D. B. Dahl. "Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model". In: *Bayesian Inference for Gene Expression and Proteomics*. Ed. by K.-A. Do, P. Müller, and M. Vannucci. Cambridge University Press, 2006, 201–218.

[22] D. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure*. Springer-Verlag New York, 2008.

[23] P. De Blasi, S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero. "Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process?" In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2 (2015), pp. 212–229.

[24] B. de Finetti. "La prévision : ses lois logiques, ses sources subjectives." In: *Annales de l'Institut Henri Poincaré* 7.1 (1937), pp. 1–68.

[25] B. de Finetti. "Sur la condition d'equivalence partielle." In: *Actualités scientiques et industrielles* 739 (1938), pp. 5–18.

[26] L. Devroye. "Random Variate Generation for Exponentially and Polynomially Tilted Stable Distributions". In: *ACM Trans. Model. Comput. Simul.* 19.4 (2009), 18:1–18:20.

[27] S. Donnet, V. Rivoirard, J. Rousseau, and C. Scricciolo. "Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures". In: *Bernoulli* 24.1 (2018), pp. 231–256.

[28] D. B. Dunson and J. Park. "Kernel stick-breaking processes". In: *Biometrika* 95.2 (2008), pp. 307–323.

[29] M. D. Escobar. "Estimating the means of several normal populations by nonparametric estimation of the distribution of the means". PhD thesis. Department of Statistics, Yale University, 1988.

[30] M. D. Escobar and M. West. "Bayesian density estimation and inference using mixtures". In: *Journal of the American Statistical Association* 90.430 (1995), pp. 577–588.

[31] M. D. Fall and É. Barat. "Gibbs sampling methods for Pitman-Yor mixture models". Technical report. 2014.

[32] S. Favaro, A. Lijoi, and I. Prünster. "On the stick-breaking representation of normalized inverse Gaussian priors". In: *Biometrika* 99.3 (2012), pp. 663–674.

[33] S. Favaro, A. Lijoi, C. Nava, B. Nipoti, I. Prünster, and Y. W. Teh. "On the Stick-Breaking Representation for Homogeneous NRMIs". In: *Bayesian Anal.* 11.3 (2016), pp. 697–724.

[34] S. Favaro and Y. W. Teh. "MCMC for Normalized Random Measure Mixture Models". In: *Statist. Sci.* 28.3 (2013), pp. 335–359.

[35] S. Favaro, A. Lijoi, R. H. Mena, and I. Prünster. "Bayesian Non-Parametric Inference for Species Variety with a Two-Parameter Poisson-Dirichlet Process Prior". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 71.5 (2009), pp. 993–1008.

[36] S. Favaro, A. Lijoi, and I. Prünster. “A New Estimator of the Discovery Probability”. In: *Biometrics* 68.4 (2012), pp. 1188–1196.

[37] S. Favaro, M. Lomeli, B. Nipoti, and Y. W. Teh. “On the stick-breaking representation of σ-stable Poisson-Kingman models”. In: *Electron. J. Statist.* 8.1 (2014), pp. 1063–1085.

[38] T. S. Ferguson. “A Bayesian analysis of some nonparametric problems”. In: *The Annals of Statistics* 1.2 (1973), pp. 209–230.

[39] T. S. Ferguson and M. J. Klass. “A representation of independent increment processes without Gaussian components”. In: *The Annals of Mathematical Statistics* 43.5 (1972), pp. 1634–1643.

[40] S. Ghosal. “The Dirichlet process, related priors and posterior asymptotics”. In: *Bayesian Nonparametrics*. Ed. by N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010, 35–79.

[41] S. Ghosal and A. Van Der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

[42] A. Gnedin and J. Pitman. “Exchangeable Gibbs partitions and Stirling triangles”. In: *Journal of Mathematical Sciences* 138.3 (2006), pp. 5674–5685.

[43] D. Görür and C. E. Rasmussen. “Dirichlet process gaussian mixture models: Choice of the base distribution”. In: *Journal of Computer Science and Technology* 25.4 (2010), pp. 653–664.

[44] J. E. Griffin, M. Kolossiatis, and M. F. Steel. “Comparing distributions by using dependent normalized random-measure mixtures”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75.3 (2013), pp. 499–529.

[45] R. Griffiths and R. Milne. “A class of bivariate Poisson processes”. In: *Journal of Multivariate Analysis* 8.3 (1978), pp. 380 –395.

[46] N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. *Bayesian nonparametrics*. Vol. 28. Cambridge University Press, 2010.

[47] R. Ibata, A. Sollima, C. Nipoti, M. Bellazzini, S. C. Chapman, and E. Dalessandro. “The globular cluster NGC 2419: a crucible for theories of gravity”. In: *ApJ* (2011).

[48] H. Ishwaran and L. F. James. “Gibbs Sampling Methods for Stick-Breaking Priors”. In: *Journal of the American Statistical Association* 96.453 (2001), pp. 161–173.

[49] H. Ishwaran and M. Zarepour. “Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-Parameter Process Hierarchical Models”. In: *Biometrika* 87.2 (2000), pp. 371–390.

[50] H. Ishwaran and M. Zarepour. “Exact and Approximate Sum Representations for the Dirichlet Process”. In: *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 30.2 (2002), pp. 269–283.

[51] N. J. Foti and S. Williamson. “A Survey of Non-Exchangeable Priors for Bayesian Nonparametric Models”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (2015).

[52] L. F. James. “Poisson Process Partition Calculus with applications to Exchangeable models and Bayesian Nonparametrics.” In: (2002). arXiv: math/0205093.

[53] L. F. James, A. Lijoi, and I. Prünster. “Conjugacy as a Distinctive Feature of the Dirichlet Process”. In: *Scandinavian Journal of Statistics* 33.1 (2006), pp. 105–120.

[54] L. F. James, A. Lijoi, and I. Prünster. "Posterior Analysis for Normalized Random Measures with Independent Increments". In: *Scandinavian Journal of Statistics* 36.1 (2009), pp. 76–97.

[55] A. Jara, E. Lesaffre, M. De Iorio, and F. Quintana. "Bayesian semiparametric inference for multivariate doubly-interval-censored data". In: *Ann. Appl. Stat.* 4.4 (2010), pp. 2126–2149.

[56] A. Jara, T. Hanson, F. Quintana, P. Müller, and G. Rosner. "DPpackage: Bayesian Semi- and Nonparametric Modeling in R". In: *Journal of Statistical Software* 40.5 (2011), pp. 1–30.

[57] M. Kalli, J. E. Griffin, and S. G. Walker. "Slice sampling mixture models". In: *Statistics and Computing* 21.1 (2011), pp. 93–105.

[58] J. F. C. Kingman. "Completely random measures". In: *Pacific J. Math.* 21.1 (1967), pp. 59–78.

[59] J. F. C. Kingman. "Random Discrete Distributions". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 37.1 (1975), pp. 1–22.

[60] J. F. C. Kingman. "Random Partitions in Population Genetics". In: *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 361.1704 (1978), pp. 1–20.

[61] J. F. C. Kingman. *Poisson Processes*. Oxford University Press, 1993.

[62] A. Kottas. "Nonparametric Bayesian survival analysis using mixtures of Weibull distributions". In: *Journal of Statistical Planning and Inference* 136.3 (2006), pp. 578–596.

[63] H. P. Kriegel, E. Schubert, and A. Zimek. "The (black) art of runtime evaluation: Are we comparing algorithms or implementations?" In: *Knowledge and Information Systems* 52.2 (2017), pp. 341–378.

[64] M. Krnjajić, A. Kottas, and D. Draper. "Parametric and nonparametric Bayesian model specification: A case study involving models for count data". In: *Computational Statistics & Data Analysis* 52.4 (2008), pp. 2110–2128.

[65] E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[66] A. Lijoi and I. Prünster. "On a normalized random measure with independent increments relevant to Bayesian nonparametric inference". In: *Proceedings of the 13th European Young Statisticians Meeting*. Bernoulli Society, 2003, pp. 123–134.

[67] A. Lijoi and I. Prünster. "Distributional properties of means of random probability measures". In: *Statist. Surv.* 3 (2009), pp. 47–95.

[68] A. Lijoi and I. Prünster. "Bayesian Nonparametrics". In: ed. by N. Hjort, C. Holmes, P. Müller, and S. Walker. Cambridge University Press, 2010. Chap. Models beyond the dirichlet process, pp. 80–136.

[69] A. Lijoi, R. H. Mena, and I. Prünster. "Bayesian Nonparametric Analysis for a Generalized Dirichlet Process Prior". In: *Statistical Inference for Stochastic Processes* 8.3 (2005), pp. 283–309.

[70] A. Lijoi, R. H. Mena, and I. Prünster. "Hierarchical Mixture Modeling With Normalized Inverse-Gaussian Priors". In: *Journal of the American Statistical Association* 100.472 (2005), pp. 1278–1291.

[71] A. Lijoi, R. H. Mena, and I. Prünster. "A Bayesian nonparametric method for prediction in EST analysis". In: *BMC Bioinformatics* 8.1 (2007), p. 339.

[72] A. Lijoi, R. H. Mena, and I. Prünster. "Bayesian Nonparametric Estimation of the Probability of Discovering New Species". In: *Biometrika* 94.4 (2007), pp. 769–786.

[73] A. Lijoi, R. H. Mena, and I. Prünster. "Controlling the reinforcement in Bayesian nonparametric mixture models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.4 (2007), pp. 715–740.

[74] A. Lijoi, B. Nipoti, and I. Prünster. "Bayesian inference with dependent normalized completely random measures". In: *Bernoulli* 20.3 (2014), pp. 1260–1291.

[75] A. Lijoi, B. Nipoti, and I. Prünster. "Dependent mixture models: Clustering and borrowing information". In: *Computational Statistics & Data Analysis* 71 (2014), pp. 417 –433.

[76] J. S. Liu. "Nonparametric hierarchical Bayes via sequential imputations". In: *Ann. Statist.* 24.3 (1996), pp. 911–930.

[77] A. Y. Lo. "On a class of Bayesian nonparametric estimates: I. Density estimates". In: *The Annals of Statistics* 12.1 (1984), pp. 351–357.

[78] S. N. Maceachern. "Estimating normal means with a conjugate style dirichlet process prior". In: *Communications in Statistics - Simulation and Computation* 23.3 (1994), pp. 727–741.

[79] S. N. Maceachern. "Dependent nonparametric processes." In: *ASA Proceedings of the Section on Bayesian Statistical Science*. Technical report. American Statistical Association, 1999.

[80] S. N. Maceachern. "Dependent Dirichlet processes." Technical report. 2000.

[81] S. N. Maceachern and P. Müller. "Estimating Mixture of Dirichlet Process Models". In: *Journal of Computational and Graphical Statistics* 7.2 (1998), pp. 223–238.

[82] M Maciejewski, S Colombi, C Alard, F Bouchet, and C Pichon. "Phase-space structures–I. A comparison of 6D density estimators". In: *Monthly Notices of the Royal Astronomical Society* 393.3 (2009), pp. 703–722.

[83] M. Meilă. "Comparing clusterings—an information based distance". In: *Journal of Multivariate Analysis* 98.5 (2007), pp. 873 –895.

[84] P. Muliere and P. Secchi. *A note on a proper Bayesian bootstrap*. Quaderni di Dipartimento. Dipartimento di economia politica e metodi quantitativi, Universita degli studi di Pavia, 1995.

[85] P. Muliere and L. Tardella. "Approximating Distributions of Random Functionals of Ferguson-Dirichlet Priors". In: *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 26.2 (1998), pp. 283–297.

[86] P. Müller, A. Erkanli, and M. West. "Bayesian curve fitting using multivariate normal mixtures". In: *Biometrika* 83.1 (1996), pp. 67–79.

[87] R. M. Neal. "Markov Chain Sampling Methods for Dirichlet Process Mixture Models". In: *Journal of Computational and Graphical Statistics* 9.2 (2000), pp. 249–265.

[88] L. E. Nieto-Barajas, I. Prünster, and S. G. Walker. "Normalized random measures driven by increasing additive processes". In: *Ann. Statist.* 32.6 (2004), pp. 2343–2360.

[89] I. Olkin and R. Liu. "A bivariate beta distribution". In: *Statistics & Probability Letters* 62.4 (2003), pp. 407 –412.

[90] O. Papaspiliopoulos and G. O. Roberts. "Retrospective Markov Chain Monte Carlo Methods for Dirichlet Process Hierarchical Models". In: *Biometrika* 95.1 (2008), pp. 169–186.

[91] M. Perman, J. Pitman, and M. Yor. "Size-biased sampling of Poisson point processes and excursions". In: *Probability Theory and Related Fields* 92.1 (1992), pp. 21–39.

[92] S. Petrone, J. Rousseau, and C. Scricciolo. "Bayes and empirical Bayes: do they merge?" In: *Biometrika* 101.2 (2014), pp. 285–302.

[93] J. Pitman. "Exchangeable and partially exchangeable random partitions". In: *Probability Theory and Related Fields* 102 (1995), pp. 145–158.

[94] J. Pitman. "Some Developments of the Blackwell-Macqueen URN Scheme". In: *Lecture Notes-Monograph Series* 30 (1996), pp. 245–267.

[95] J. Pitman. "Poisson-Kingman partitions". In: *Statistics and science: a Festschrift for Terry Speed*. Ed. by D. R. Goldstein. Vol. Volume 40. Lecture Notes–Monograph Series. Beachwood, OH: Institute of Mathematical Statistics, 2003, pp. 1–34.

[96] J. Pitman. *Combinatorial Stochastic Processes. Ecole d'Eté de Probabilités de Saint-Flour XXXII - 2002*. Ed. by J. Picard. Springer-Verlag Berlin Heidelberg, 2006.

[97] J. Pitman and M. Yor. "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator". In: *Ann. Probab.* 25.2 (1997), pp. 855–900.

[98] I. Prünster. "Random probability measures derived from increasing additive processes and their application to Bayesian statistics". PhD thesis. University of Pavia, 2002.

[99] R. Rastelli and N. Friel. "Optimal Bayesian estimators for latent variable cluster models". In: *Statistics and Computing* 28.6 (2018), pp. 1169–1186.

[100] E. Regazzini, A. Lijoi, and I. Prünster. "Distributional results for means of normalized random measures with independent increments". In: *Ann. Statist.* 31.2 (2003), pp. 560–585.

[101] C. Ritter and M. A. Tanner. "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler". In: *Journal of the American Statistical Association* 87.419 (1992), pp. 861–868.

[102] J. Rousseau. "On the frequentist properties of Bayesian nonparametric methods". In: *Annual Review of Statistics and Its Application* 3 (2016), pp. 211–231.

[103] J. Sethuraman. "A constructive definition of Dirichlet priors". In: *Statistica Sinica* 4 (1994), pp. 639–650.

[104] W. Shen, S. T. Tokdar, and S. Ghosal. "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures". In: *Biometrika* 100.3 (2013), pp. 623–640.

[105] Y. Shi, M. Martens, A. Banerjee, and P. Laud. "Low Information Omnibus (LIO) Priors for Dirichlet Process Mixture Models". In: *Bayesian Analysis* (2018).

[106] Y. W. Teh. "A Hierarchical Bayesian Language Model Based on Pitman-Yor Processes". In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. 2006, pp. 985–992.

[107] Y. W. Teh and M. I. Jordan. "Hierarchical Bayesian nonparametric models with applications". In: *Bayesian Nonparametrics*. Ed. by N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010, 158–207.

[108] S. Wade and Z. Ghahramani. "Bayesian Cluster Analysis: Point Estimation and Credible Balls". In: *Bayesian Anal.* 13.2 (2018), pp. 559–626.

[109] S. G. Walker. "Sampling the Dirichlet Mixture Model with Slices". In: *Communications in Statistics - Simulation and Computation* 36.1 (2007), pp. 45–54.

[110] Y. Wu and S. Ghosal. “Kullback Leibler property of kernel mixture priors in Bayesian density estimation”. In: *Electronic Journal of Statistics* 2 (2008), pp. 298–331.

[111] Y. Wu and S. Ghosal. “The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation”. In: *Journal of Multivariate Analysis* 101.10 (2010), pp. 2411–2419.

[112] S. Zabell. “The continuum of inductive methods revisited”. English. In: *The Cosmos of Science: Essays of Exploration*. Ed. by J. Earman and J. Norton. University of Pittsburgh Press, 1997, pp. 351–385.