Università degli studi di Milano–Bicocca
Scuola di Economia e Statistica

Corso di Laurea Magistrale in
Scienze Statistiche ed Economiche

The Gnedin model: recent developments and future perspectives

Relatore: Prof. Federico Camerlenghi
Corelatore: Dott. Tommaso Rigon

Tesi di laurea di:
Anna Petranzan
Matricola N. 858541

Anno Accademico 2023/2024

Contents

Ringraziamenti

Ringrazio il prof. Camerlenghi e il prof. Rigon per avermi guidato in questo lavoro di tesi e avermi sostenuta in questi mesi di preparazione al dottorato. Vi ringrazio per il lavoro preciso e accurato che abbiamo svolto.

Introduction

Central to Bayesian nonparametrics is the specification of prior distributions over infinite-dimensional spaces and the construction of probability measures over these spaces, enabling the modelling of unknown underlying structures using data. The mathematical demands are higher since defining well-structured probability distributions on potentially infinite-dimensional spaces is more difficult. Additionally, eliciting a prior in such a large space is a significant challenge. However, these solutions are increasingly valuable for solving real-world problems across a growing range of applications. For instance, genomic applications have introduced challenging inferential problems due to their unique characteristics, such as dealing with very large populations containing numerous distinct species where only a small portion of the population has been sampled. Species sampling problems are a significant area of application. These problems involve drawing samples from a population of individuals, which may belong to a possibly infinite number of species. When the number of species in the population is large, it is reasonable to assume it is infinite. Such problems are fundamental in ecological and biological studies, addressing issues like species richness evaluation, sampling experiment design, and rare species estimation. The main goal is to estimate the number of new species in a future sample, given an existing sample from the same population. To achieve this, Bayesian nonparametric models are employed, particularly discrete random probability measures like the Dirichlet process (Ferguson (1973)) and the two-parameter Poisson–Dirichlet process (Pitman & Yor (1997)). More generally, the category of normalized random measures (Regazzini et al. (2003)), driven by Gibbs-type priors—a generalization of well-known processes such as the Dirichlet or Pitman-Yor processes—is frequently employed.

This work focuses on a specific case of Gibbs-type priors, the Gnedin model, introduced by Gnedin (2010). The thesis is organized into three chapters. The first chapter introduces Gibbs-type priors (De Blasi et al. (2015)), which provide

a framework with characterisation based on their predictive structure, making them suitable for rigorous mathematical analysis. These priors enable Bayesian nonparametric models to balance flexibility and regularization, which is crucial for achieving robust and reliable inference, especially when the true underlying data-generating process is complex and unknown. While the Dirichlet process serves as a fundamental building block in Bayesian nonparametrics, Gibbs-type priors, encompass a broader class of prior distributions.

The second chapter is the core of this thesis, providing a detailed description of the Gnedin model. Some quantities within this model still lack precise formulation and definition. In this thesis, we aim to derive new results. These will include the prior expected value of distinct species and the posterior expected value of new species discovered and not observed in the initial sample. The posterior distribution of the random probability measure associated with the Gnedin model will also be formulated. The prior expected values are fundamental for creating a model-based rarefaction curve. This curve visually illustrates the relationship between the number of individuals randomly sampled and the corresponding number of distinct species observed. This tool offers insights into the adequacy of the sample size and allows for comparisons of species richness between samples with different sequencing volumes. The collection of posterior expected values as the sample size changes represents a model-based extrapolation of the accumulation curve, a useful tool for predicting the number of species. Additionally, consider a population of animals, plants, or similar entities. A fundamental problem is predicting how many new species will be observed in a future sample. We aim to make inferences by analytically deriving this statistic of interest and finding its probability law, leveraging the general achievements of De Blasi et al. (2015). This approach allows us to address several application problems, such as species richness estimation or prediction for rare species. These new findings are presented in this chapter.

Another possible application of this specific prior distribution is as a latent structure in a mixture model, useful for clustering and density estimation. In the Gnedin model, the posterior distribution for the associated random probability measure has not been previously discussed. In the second chapter, we build on the work of Argiento & De Iorio (2022) on Normalized Independent Finite Poisson Processes, of which Gibbs-type priors are a special case, to formalize the posterior distribution. We also provide an equivalent, simpler representation of this distribution.

In the final chapter, we focus on applying this model to species sampling problems. The new quantities defined in the previous chapter open up new possibilities for the analysis and application of this model. We present and analyze the results obtained using both simulated data and commonly used datasets, which are well-suited for evaluating the performance of different processes.

Chapter 1

A particular type of priors for Bayesian non-parametric methods

The Gnedin model is an example of a broader class of prior distributions known as Gibbs-type priors, which also includes well-known processes such as the Dirichlet process and the Pitman-Yor process. This chapter will discuss the characteristics and construction of these priors in detail. It will begin with a brief introduction to the fundamental concepts of Bayesian nonparametric statistics, followed by an overview of Gibbs-type priors. Then, the chapter will describe the Dirichlet and Pitman-Yor processes. Finally, it will explore species sampling problems, where the Gnedin model is applied in the context of this thesis work.

1.1 Bayesian Nonparametrics Statistics

In the Bayesian parametric setting, we start by considering a sequence of data points $X_1, X_2, \ldots$ and make the assumption that these data points are independent and identically distributed (i.i.d.) given a certain parameter which is typically drawn from a prior distribution. In contrast, the Bayesian nonparametric setting assumes that the data are conditionally i.i.d. given a random probability measure $\tilde{p}$. This random measure $\tilde{p}$ is not restricted to belong to a parametric family of distributions but can be any probability measure on the sample space $\mathbb{X}$. The flexibility of $\tilde{p}$ allows it to adapt to the complexity of the data, making Bayesian nonparametrics particularly useful in situations where the underlying distribution is unknown or cannot be adequately captured by a finite number of parameters. The central challenge in Bayesian nonparametrics lies in how to properly define the random probability measure $\tilde{p}$ and its associated prior distribution. Defining a prior in this context means specifying a probability distribution on the space of

all possible probability measures on $\mathbb{X}$, $\mathbb{P}_{\mathbb{X}}$. Each point in $\mathbb{P}_{\mathbb{X}}$ corresponds to a complete probability distribution on $\mathbb{X}$, and thus the prior distribution over $\mathbb{P}_{\mathbb{X}}$ reflects our uncertainty about which probability distribution from this space best represents the underlying data-generating process.

1.2 Gibbs-type priors

The Bayesian Nonparametrics field has mainly focused on proposing and studying classes of random probability measures that act as nonparametric priors. Some of these classes include the Dirichlet process (Ferguson (1973)) as a special case, which is fundamental in the field. However, when moving beyond the Dirichlet process, there is a trade-off between generality and tractability, both analytically and computationally. The two-parameter Poisson-Dirichlet process (Pitman & Yor (1997)) is likely the most successful proposal in this regard. It is possible to identify a large class of priors, which embeds the Pitman-Yor process as a special case and the Dirichlet process, too. Such a class is given by Gibbs–type priors, introduced in Gnedin & Pitman (2005). With references to De Blasi et al. (2015), in this section we will now introduce this particular general class of priors.

1.2.1 Discrete random probability measures

Consider an infinite exchangeable sequence $(X_n)_{n\geqslant 1}$, where each element X_i can be interpreted as the observed species labels in a set $\mathbb{X}$. We assume that $(X_n)_{n\geqslant 1}$ is exchangeable, i.e., the order in which the observations are recorded is irrelevant. Moreover, $\mathbb{P}_{\mathbb{X}}$ is the set of all probability measures on $\mathbb{X}$. According to the *de Finetti theorem* (de Finetti (1937)) this is equivalent to the existence of a random probability measure $\tilde{p}$ on $\mathbb{X}$ such that:

$$\begin{aligned} X_i \mid \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p} && i = 1, \dots, n, \\ \tilde{p} &\sim P, \end{aligned} \tag{1.1}$$

for any $n \geqslant 1$ where P is said the *de Finetti measure* and it is defined on $\mathbb{P}_{\mathbb{X}}$.

In Bayesian Nonparametrics, priors that assign probability one to discrete distributions are called discrete nonparametric priors. Any random probability

measure associated with a discrete prior can be represented as

$$\tilde{p} = \sum_{j=1}^{\infty} \tilde{p}_j \delta_{X_j}, \tag{1.2}$$

where δ_y represents the Dirac delta measure at y, $(\tilde{p}_j)_{j\geqslant 1}$ is a sequence of probability weights such that $\sum_{j\geqslant 1} \tilde{p}_j = 1$ almost surely and $(X_j)_{j\geqslant 1}$ is a sequence of $\mathbb{X}$-valued random variables. We assume that X_j's are independent and identically distributed (i.i.d.) from P, a non-atomic probability measure on $\mathbb{X}$, and that $(X_j)_{j\geqslant 1}$ and $(\tilde{p}_j)_{j\geqslant 1}$ are independent. The class of the Equation (1.2) is said to be a *proper species sampling model* (Pitman (1996)) and in particular Gibbs-type priors are notable species sampling models.

1.2.2 Excheangeable partition probability function

The random probability measure can be characterized by a random partition Π_n induced by a sample $\mathbf{X}_n = (X_1, \ldots, X_n)$ of size n. Due to the discrete nature of $\tilde{p}$, there will be identical values among $X_1, \ldots, X_n$ with positive probability, containing a total of $K_n = k$ distinct values with labels $X_1^*, \ldots, X_k^*$ with frequencies $n_1, \ldots, n_k$ such that $\sum_{j=1}^{k} n_j = n$. The ties among the observations $X_1, \ldots, X_n$ induce a random partition $\Psi_n = \{C_1, \ldots, C_k\}$ of the indices $\{1, \ldots, n\}$, where $C_j = \{i : X_i = X_j^*\}$ and $n_j = |C_j|$. In Gibbs-type priors, the law of the partition is such that:

$$\Pi_n(n_1, ..., n_k) = \mathbb{P}(\Psi_n = \{C_1, \ldots, C_k\}) = V_{n,k} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1}, \tag{1.3}$$

where $(a)_n = a(a+1)\ldots(a+n-1)$ denotes a rising factorial, also known as the Pochhammer symbol, $\sigma < 1$ is the discount parameter, and $V_{n,k}$'s are non-negative weights that satisfy the forward recursive equation

$$V_{n,k} = (n - \sigma k) V_{n+1,k} + V_{n+1,k+1},$$

for any $n \geqslant 1$ and $1 \leqslant k \leqslant n$. The function (1.3) is called Exchangeable Partition Probability Function (EPPF) and it characterises the Gibbs-type prior and so the distribution of the random probability measure $\tilde{p}$. Specifically, the weights form the basis of any priors of the Gibbs-type. It is always possible to express Gibbs partitions as a mixture with respect to the parameters of the associated process.

Moreover, the EPPF satisfies the following consistency relation:

$$\Pi_n(n_1,...,n_k) = \Pi_{n+1}(n_1,...,n_k,1) + \sum_{j=1}^{k} \Pi_{n+1}(n_1,...,n_j+1,\ldots,n_k).$$

1.2.3 Predictive distributions

Given a sample $X_1,\ldots,X_n$ generated from (1.1), the one-step ahead predictive distribution coincides with the posterior expected value of $\tilde{p}$:

$$\mathbb{P}(X_{n+1} \in \cdot \mid X_1,\ldots,X_n) = \int_{\mathbb{P}_{\mathbb{X}}} p(\cdot) Q(dp \mid X_1,\ldots,X_n)$$

where $Q(\cdot \mid X_1,\ldots,X_n)$ is the posterior distribution of $\tilde{p}$.

When choosing and assessing specific predictive models, it is important to take into account the probability of encountering a new, unique value that has not been already included in the existing sample $X_1,\ldots,X_n$. In other words, we need to consider:

$$\mathbb{P}(X_{n+1} = \text{"new"} \mid X_1,\ldots,X_n).$$

The random partition Π_n allows to derive directly the predictive distribution (Lijoi et al. (2007)), given by:

$$\begin{aligned}\mathbb{P}(X_{n+1} \in \cdot \mid X_1,\ldots,X_n) &= \frac{\Pi_{n+1}(n_1,...,n_k,1)}{\Pi_n(n_1,...,n_k)} P(\cdot) \\ &\quad + \sum_{j=1}^{k} \frac{\Pi_{n+1}(n_1,\ldots,n_j+1,\ldots,n_k)}{\Pi_n(n_1,...,n_k)} \delta_{X_j^*}(\cdot) \\ &= \frac{V_{n+1,k+1}}{V_{n,k}} P(\cdot) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{j=1}^{k} (n_j - \sigma)\delta_{X_j^*}(\cdot), \qquad (1.4)\end{aligned}$$

which is helpful to address posterior inference. In this way, we can predict the label for new observations. The predictive distribution is a linear convex combination of the prior guess P at the shape $\tilde{p}$ and of the weighted empirical distribution $\hat{P}_n = (n-k\sigma)^{-1}\sum_{j=1}^{k}(n_j-\sigma)\delta_{X_j^*}$. The mechanism of allocating the predictive mass among "new" and previously observed data can be split into two stages. Given a sample $X_1,\ldots,X_n$, the first step involves allocating the probability mass between a new value X_{k+1}^* sampled from P and the set of observed values $\{X_1^*,\ldots,X_k^*\}$. This initial step depends only on n and k, and not on the frequencies $n_1,\ldots,n_k$. The second step is as follows: conditionally, if X_{n+1}

is a new value, it is sampled from the base measure P. However, if X_{n+1} matches one of the previously observed values X_j^* for $j = 1, \ldots, k$, the probabilities of these coincidences are determined by the size n_j of each cluster and σ. Thus, while the frequencies n_j do not influence the probability of allocating a predicted value between "new" and "old", they do play a significant role if the predicted value coincides with a previously observed one: the more frequently a past observation has been detected, the higher the probability of re-observing it. Additionally, σ plays an interesting role in weighting the empirical measure, as for $\sigma > 0$, a reinforcement mechanism driven by σ occurs for those having higher frequencies, which represents an appealing feature in certain inferential contexts. If $\sigma < 0$, the reinforcement mechanism works inversely, meaning that the probabilities of coincidence are less than proportional to the cluster size. Moreover, the parameter σ also determines the rate at which the number of clusters, K_n increases as the sample size n increases. The larger σ, the faster the rate of increase of K_n or, in other words, the more new values are generated.

1.3 The Dirichlet process

The simplest nonparametric prior is the Dirichlet process, which was first defined by Ferguson (1973) and has been presented by Müller et al. (2015) and Orbanz (2014).

1.3.1 The Dirichlet distribution

Before introducing the Dirichlet process, it is necessary to review the Dirichlet distribution. As a multivariate generalization of the Beta distribution, the Dirichlet distribution can also be derived from the Gamma distribution.

Definition 1.1. Let $\mathbf{Z} = (Z_1, \ldots, Z_k)$ be a vector with k components, where $Z_i \geqslant 0$ for $i = 1, 2, \ldots, k$ and $\sum_{i=1}^{k} Z_i = 1$. Also, let $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_k]$, where $\alpha_i > 0$ for each i. Then the Dirichlet probability density function is

$$f(\mathbf{z}) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} z_i^{\alpha_i - 1},$$

where $\alpha_0 = \sum_{i=1}^{k} \alpha_i$. We denote this distribution by $\mathrm{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_k)$. The Dirichlet distribution is a distribution with k positive parameters $\boldsymbol{\alpha}$ with respect to a k-dimensional space.

The probability density function of the Dirichlet distribution for k random variables is a $k-1$ dimensional probability simplex that exists in a k-dimensional space. It can be demonstrated that the marginal distribution of Z_i is

$$\text{Beta}\left(\alpha_i, \sum_{j=1}^{k} \alpha_j - \alpha_i\right).$$

Gamma construction

We report here a useful result that will be exploited in the course of the thesis. The Dirichlet distribution can be constructed using Gamma distributions, which is a useful approach because it simplifies sampling from a Dirichlet distribution. Suppose we want to construct a Dirichlet distribution $\text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_k)$ with parameters $\alpha_1, \alpha_2, \ldots, \alpha_k$ all strictly positive. For each parameter α_i, generate a random variable Y_i that follows a Gamma distribution with shape parameter α_i and scale parameter 1:

$$Y_i \sim \text{Gamma}(\alpha_i, 1) \quad \text{for } i = 1, 2, \ldots, k.$$

Then, calculate the sum of the k generated Gamma variables:

$$S = \sum_{i=1}^{k} Y_i.$$

and normalize each Gamma variable Y_i by dividing by the sum S to obtain the random variable Z_i:

$$Z_i = \frac{Y_i}{S} \quad \text{for } i = 1, 2, \ldots, k.$$

The obtained variables $(Z_1, \ldots, Z_k)$ follow a Dirichlet distribution with parameters $\alpha_1, \alpha_2, \ldots, \alpha_k$.

1.3.2 Distributional form

An initial definition of the Dirichlet Process is given here.

Definition 1.2. Given a measurable space $(\mathbb{X}, \mathcal{A})$ a random distribution $\tilde{p}$ is said to follow a Dirichlet process prior with a base probability measure P and mass parameter or concentration α if

$$(\tilde{p}(A_1), \ldots, \tilde{p}(A_k)) \sim \text{Dir}(\alpha P(A_1), \ldots, \alpha P(A_k))$$

considering any arbitrary measurable partition $\{A_1, \ldots, A_k\}$ of $\mathbb{X}$. We will write $\tilde{p} \sim DP(\alpha, P)$.

For this definition to be valid, we must assume that P assigns positive mass to any set of the partition $A_1, \ldots, A_k$ of $\mathbb{X}$. For cases in which P attributes zero mass, see van der Vaart & Ghosal (2017).

1.3.3 Stick-breaking construction

An alternative definition is based on the *stick-breaking* construction provided by Sethuraman (1994).

Definition 1.3. Given a mass parameter α and a continuous distribution over a generic space $\mathbb{X}$, known as base measure, P. Considering two different independent sequences of random variables of atoms $(X_j)_{j\geqslant 1}$ and of weights $(\tilde{p}_j)_{j\geqslant 1}$ defined as:

$$X_j \overset{iid}{\sim} P$$

and

$$\tilde{p}_j = v_j \prod_{k=1}^{j-1}(1 - v_k) \text{ with } v_j \overset{iid}{\sim} Beta(1, \alpha).$$

If the sum of weights is equal to one almost surely, we obtain a random probability measure defined as follows:

$$\tilde{p}(\cdot) = \sum_{j=1}^{\infty} \tilde{p}_j \delta_{X_j}(\cdot)$$

is said a Dirichlet process with parameters P and α, $DP(\alpha, P)$.

It is important to note that this definition guarantees that $\tilde{p}$ is discrete, even if P is a continuous distribution. The term *stick-breaking* refers to the construction of the weights $\tilde{p}_j$ because we can think of the interval as a stick form in which pieces $(1 - v_k)$ are repeatedly broken off. Consider a sequence of random variables $v_1, v_2, \ldots$ in $[0, 1]$ which tells us how to break a stick of length 1. Each random variable is sampled from a Beta distribution. The construction proceeds as follows. Consider a stick of length 1 and break it into two pieces of length v_1 and $1 - v_1$, putting $\tilde{p}_1 := v_1$. The remaining stick of length $1 - v_1$ is again broken into two pieces of relative lengths v_2 and $1 - v_2$. Hence we set $\tilde{p}_2 := (1 - v_1) \cdot v_2$. The remaining stick has length $(1 - v_1) \cdot (1 - v_2)$. Iterating such a procedure, we define

the following infinite sequence of weights:

$$\tilde{p}_1 = v_1, \quad \ldots, \quad \tilde{p}_j = v_j \prod_{k=1}^{j-1} (1 - v_k),$$

and $(v_j)_{j \geqslant 1}$ such that $v_j \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$. We can demonstrate that these weights sum up to 1. Given this construction, the distribution of $(\tilde{p}_j)_{j \in \mathbb{N}}$ is usually called the $\text{GEM}(\alpha)$ distribution, from the names of its authors Griffiths (1979), Engen (1978), McCloskey (1965).

1.3.4 The Chinese restaurant process

The Chinese Restaurant Process (CRP) is an alternative representation of the Dirichlet process, illustrating how data points cluster. Suppose we have a sequence of data points $X_1, X_2, \ldots, X_n$ generated from $\tilde{p} \sim \text{DP}(\alpha, P)$. The CRP describes the probability of each data point joining an existing cluster or generating a new cluster.

The CRP with concentration α is the distribution on partitions that we obtain if we choose a Dirichlet process with parameters P and α as the distribution of a random probability measure. The choice of the base measure P does not affect the partition, and the CRP hence has only a single parameter.

This representation is useful for introducing predictive distribution. The CRP is a metaphor for how customers (the observations) are seated at tables (clusters) in a restaurant. When a new customer arrives, they either join an existing table with a probability proportional to the number of customers already seated there, or they start a new table with a probability proportional to the concentration parameter α.

Theorem 1.1. *Given a sample $X_1, \ldots, X_n$ generated from a $\tilde{p} \sim \text{DP}(\alpha, P)$, the predictive distribution for a new data point X_{n+1} can be expressed as:*

$$\mathbb{P}(X_{n+1} \in \cdot \mid X_1, X_2, \ldots, X_n) = \frac{\alpha}{\alpha + n} \cdot P(\cdot) + \sum_{i=1}^{n} \frac{1}{\alpha + n} \delta_{X_i}(\cdot)$$

where α is the mass parameter, P the base measure, and δ_{X_i} the Dirac delta measure at X_i.

This means that there is a probability of $\alpha/(\alpha + n)$ for the new data point X_{n+1} to be drawn from the base measure P, indicating the potential for a new

cluster. With a probability of $n_i/(\alpha+n)$, the new data point is the same as one of the existing data points X_i, demonstrating the clustering behaviour.

1.3.5 The Pólya urn scheme

The Pólya urn scheme offers a visual representation to help understand the predictive distribution of new samples under the Dirichlet process. The sample $(X_1,\dots,X_n)$ can be generated as follows.

1. At the first step, generate a draw X_1 from $P(\cdot)/\alpha$;
2. Generate X_2 conditionally on X_1 using the predictive distribution, in particular

$$X_2 \mid X_1 \sim \begin{cases} \delta_{X_1}(\cdot) & \text{with probability} \quad 1/(\alpha+1) \\ \frac{P(\cdot)}{\alpha} & \text{with probability} \quad \alpha/(\alpha+1) \end{cases}$$

$\vdots$

n. Generate X_n conditionally in $(X_1,\dots,X_{n-1})$ using again the predictive distribution

$$X_n \mid X_1 \dots, X_{n-1} \sim \begin{cases} \delta_{X_1}(\cdot) & \text{with probability} \quad 1/(\alpha+n-1) \\ \vdots & \\ \delta_{X_{n-1}}(\cdot) & \text{with probability} \quad 1/(\alpha+n-1) \\ \frac{P(\cdot)}{\alpha} & \text{with probability} \quad \alpha/(\alpha+n-1) \end{cases}$$

This process can be explained using an urn model. We will assume, for simplicity, that α is an integer, but the same process can be applied in a general framework. Imagine an urn containing α black balls, and let us also assume that the distribution $P(\cdot)/\alpha$ is uniform on the interval $(0,1)$. Each x in the interval $(0,1)$ represents a different color, distinct from the black color. We now describe the sampling procedure through the urn model. At step 1, draw a ball from the urn. If the colour is black (and at step 1 this is true), we return the black ball in the urn with an additional ball of a colour chosen uniformly at random in $(0,1)$, (we call it X_1). At step n of the procedure, we draw a ball from the urn and if the colour is not black, then we return the ball in the urn with an additional ball of the same

colour. On the other hand, if the colour is black, we return the black ball in the urn with an additional ball of a new colour chosen uniformly at random in $(0, 1)$.

Note that, according to such a procedure, at step 1 we select a new colour with probability 1. At step 2, there are $\alpha + 1$ balls, of which α are black and one is coloured. Then, the probability that the new ball belongs to a new colour equals $\alpha/(\alpha + 1)$, while the probability that the new ball is equal to X_1 coincides with $1/(\alpha + 1)$. At the nth step, we have $\alpha + (n - 1)$ balls, more precisely α black balls and $n - 1$ coloured balls. X_n is a new colour with probability $\alpha/(\alpha + n - 1)$, X_n is old and equals colour X_j, for $j = 1, \ldots, n - 1$, with probability $1/(\alpha + n - 1)$. Hence, the urn scheme gives us the same predictive distribution as the Dirichlet process.

1.3.6 Properties

Since the Dirichlet process places a distribution on the random measure $\tilde{p}$, the quantity $\tilde{p}(A)$ for any $A \subset \mathbb{X}$, where $\mathbb{X}$ is a generic space, is a random variable. Then we have:

$$E[\tilde{p}(A)] = P(A),$$
$$Var(\tilde{p}(A)) = \frac{P(A)\,(1 - P(A))}{\alpha + 1}.$$

The distribution P represents the expected shape of the random measure $\tilde{p}$ and α determines how variable the realisations around the prior guess P are. From the distributional form of the Dirichlet process and for the characterization of the Dirichlet distribution we can say that the marginal distribution of $\tilde{p}(A)$ is $\mathrm{Beta}\,(\alpha P(A), \alpha(1 - P(A)))$. Recalling that the $\mathrm{Beta}(a, b)$ distribution has an expected value of $a/(a + b)$ and a variance of $ab/((a + b)^2 \cdot (a + b + 1))$, we can demonstrate the properties of the Dirichlet process.

A further particularly useful characteristic of the Dirichlet process is its conjugacy under i.i.d. sampling. If $X_1, X_2, \ldots, X_n$ is an i.i.d. sample with $X_i \mid \tilde{p} \sim \tilde{p}$ and $\tilde{p} \sim DP(\alpha, P)$ then:

$$\tilde{p} \mid X_1, \ldots, X_n \sim DP\left(\alpha + n, \frac{\alpha P + \sum_{i=1}^{n} \delta_{X_i}}{\alpha + n}\right).$$

The posterior mean is

$$E[\tilde{p} \mid X_1, \ldots, X_n] = \frac{\alpha P + \sum_{i=1}^{n} \delta_{X_i}}{\alpha + n}$$

which can be interpreted as a weighted average between the base measure P with weight α and the empirical distribution $\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ with weight n. This is the same result as in Theorem 1.1.

Another important property of the Dirichlet process is the distribution of the number of distinct clusters, K_n, in a sample of size n. Given n data points drawn from a Dirichlet process $DP(\alpha, P)$, the number of distinct clusters K_n is a random variable that represents the number of unique values among these n points. The expected number of distinct clusters is given by:

$$E[K_n \mid \alpha] = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1}. \tag{1.5}$$

This summation can be approximated using the harmonic series, which is known to grow logarithmically. Specifically, the harmonic series $H_n = \sum_{i=1}^{n} \frac{1}{i}$ is approximately $\ln(n) + \gamma$, where γ is the Euler-Mascheroni constant. Therefore, the expected number of distinct clusters is:

$$E[K_n] \approx \alpha \sum_{i=1}^{n} \frac{1}{i} \approx \alpha(\ln(n) + \gamma).$$

This expression shows that the expected number of distinct clusters increases logarithmically with the sample size n, and is influenced by the concentration parameter α. Higher values of α lead to more clusters, as the process promotes more diverse partitions of the data.

1.4 The Pitman-Yor process

Sometimes the Dirichlet process can be too restrictive. When we generate n observations, $x_1, x_2, \ldots, x_n$, from a Dirichlet process, we have already observed that the number of distinct clusters in this data, K_n, grows logarithmically. It means that the clusters generated by the Dirichlet process are, on average, very similar in terms of numerosity. However, for many real-world problems, this kind of logarithmic growth is not a realistic assumption.

The Pitman-Yor process, $PY(\sigma, \alpha, P)$, follows a different distribution of the number of distinct clusters K_n. This process was originally defined in Perman (1990).Pitman & Yor (1997) worked a lot on this process, so it was later named after Ishwaran & James (2001). This process is characterised by two parameters,

not just one as in the Dirichlet process: the mass parameter α and the discount parameter σ. The Dirichlet process is a special case of the Pitman-Yor one. It is a Pitman-Yor process with a null discount parameter.

Definition 1.4. Given the base measure P on $\mathbb{X}$, a discount parameter $\sigma \in [0, 1)$ and a mass parameter $\alpha > -\sigma$. Considering two different independent sequences of random variables of atoms $(\tilde{p}_j)_{j \geqslant 1}$ and of weights $(X_j)_{j \geqslant 1}$ defined as:

$$X_j \overset{\text{iid}}{\sim} P$$

and

$$\tilde{p}_j = v_j \prod_{k=1}^{j-1} (1 - v_k) \quad \text{with} \quad v_j \overset{\text{ind}}{\sim} \text{Beta}(1 - \sigma, \alpha + j \cdot \sigma).$$

Then the random probability measure

$$\tilde{p}(\cdot) = \sum_{j=1}^{\infty} \tilde{p}_j \delta_{X_j}$$

is said Pitman-Yor process.

1.4.1 Properties

Let $\tilde{p} \sim PY(\alpha, \sigma, P)$ on a generic space $\mathbb{X}$ and $A \subset \mathbb{X}$, a measurable subset of $\mathbb{X}$, then:

$$E[\tilde{p}(A)] = P(A),$$

$$Var(\tilde{p}(A)) = P(A) \cdot (1 - P(A)) \cdot \frac{1 - \sigma}{\alpha + 1}.$$

The expected value of the Pitman-Yor process, similar to the Dirichlet case, is equal to the base measure evaluated on A. In this case, the variability of the process is determined by two parameters: α and σ. The variability decreases as α increases and for values of σ close to one.

1.4.2 Predictive distribution

Let $X_i \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}$ with $\tilde{p} \sim PY(\alpha, \sigma, P)$. The predictive distribution of X_{n+1} given $X_1, \ldots, X_n$ is:

$$\mathbb{P}(X_{n+1} \in A \mid X_1, \ldots, X_n) = \frac{\alpha + k \cdot \sigma}{n + \alpha} \cdot P(A) + \sum_{i=1}^{k} \frac{n_i - \sigma}{\alpha + n} \cdot \delta_{X_i^*}(A),$$

where $\{X_1^*, \ldots, X_k^*\}$, with $k \leqslant n$, denote the unique values in the set $\{X_1, \ldots, X_n\}$.

From this expression, it follows that the joint distribution of $X_1, \ldots, X_n$ can be defined from the generalisation of Polya's urn scheme, such that: X_1 is sampled from P and $X_{n+1} \mid X_1 = x_1, \ldots, X_n = x_n$ is a new value sampled by P with probability $(\alpha + k \cdot \sigma)/(\alpha + n)$, or a value that has already been observed X_i^* with probability $(n_i - \sigma)/(\alpha + n)$.

In a Pitman-Yor process, the probability of observing a new observation does not just depend on the concentration parameter α, but also on the penalty parameter σ and the number of previously generated distinct observations. The penalty coefficient σ affects the probability of generating new clusters and the expansion of existing ones. When both σ and α are larger, the probability of generating new distinct clusters increases while reducing the probability of populating existing clusters. Unlike the Dirichlet process in this case we tend to have many sparsely populated clusters and only a few highly populated clusters.

1.5 More considerations on the Gibbs-type priors

The EPPF is available in closed form for both Dirichlet and Pitman-Yor processes. For the Dirichlet process with mass parameter α is equal to:

$$\Pi_n(n_1, ..., n_k) = \frac{\alpha^k}{(\alpha)_n} \prod_{i=1}^{n} (n_i - 1)!$$

for any $n \geqslant 1$ and where $\alpha^k/(\alpha)_n$ is the $V_{n,k}(\alpha)$ term. For the PY process with mass parameter α and discount parameter σ, it coincides with:

$$\Pi_n(n_1, ..., n_k) = \frac{\prod_{i=1}^{k-1} (\alpha + i\sigma)}{(\alpha + 1)_{n-1}} \prod_{i=1}^{k} (1 - \sigma)_{n_i - 1}.$$

Gibbs-type priors can also be specified within other notable classes beyond those previously discussed. These include the normalized inverse Gaussian process Lijoi et al. (2005) and the normalized generalized Gamma process Lijoi et al. (2007).

1.6 Applications of Gibbs-type priors

Discrete nonparametric priors, such as Gibbs-type priors, are well-suited for addressing inferential issues in species sampling problems and mixture modelling. Species sampling problems refer to a broad class of statistical problems in which samples are assumed to be drawn from a population of individuals belonging to a possibly infinite number of species, suggesting that if the number of species in the population is large, then it is reasonable to assume that it is infinite. They are useful in ecological and biological studies and consent to address several issues, including the evaluation of species richness, the design of sampling experiments, and the estimation of rare species variety. The main objective is to estimate the number of new species in a further sample of size m having previously observed a sample of size n.

A discrete random probability measure is a powerful tool for describing the composition of a population with different species and specific proportions. The random proportions are represented as $\tilde{p}_j$. The observed species labels are denoted by the X_n's, hence the terminology species sampling model. Additionally, a sequence $(X_n)_{n\geqslant 1}$ is labelled as a species sampling sequence when it is exchangeable and satisfies the condition (1.1), with $\tilde{p}$ being a species sampling model. In several statistical applications, one typically observes a sample of species labels $X_1, \ldots, X_n$ and then plans further sampling $X_{n+1}, \ldots, X_{n+m}$ based on estimates of various quantities of interest. Examples of such quantities include the number of new distinct species that will be detected in a new sample of size m; the number of species with a given frequency, or with a frequency below a certain threshold, in $X_1, \ldots, X_{n+m}$; and the probability that the $(n+m+1)$-th draw will consist of a species with positive frequency in $X_1, \ldots, X_{n+m}$. These estimates provide measures of overall and rare species diversity, which are of interest in fields such as biology, ecology, and linguistics, among others.

Moreover, discrete nonparametric priors are fundamental components for hierarchical mixture models. These models are commonly applied in density estimation, clustering, and more intricate dependent structures. In a univariate scenario, where $f(\cdot \mid x)$ represents a density function on $\mathbb{R}$ for any x, we can

define:

$$\begin{aligned} Y_i \mid X_i &\overset{\text{ind}}{\sim} f(\cdot \mid X_i) \quad i = 1, \dots, n \\ X_i \mid \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p} \quad i = 1, \dots, n \\ \tilde{p} &\sim Q. \end{aligned} \tag{1.6}$$

We can use this model for both density estimation and clustering. The sequence of latent exchangeable random elements $(X_n)_{n \geqslant 1}$ and the unobserved number of distinct values K_n among $X_1, \dots, X_n$, where it is the number of clusters into which the observations $Y_1, \dots, Y_n$ can be grouped, are crucial. Posterior inferences for K_n are very important, and the specification of a Gibbs-type prior $\tilde{p}$ in equation (1.6) allows for an effective detection of the number of clusters that have generated the data.

In this thesis, the focus will be on species sampling problems. Therefore, this aspect will now be explored in greater detail.

1.6.1 Prediction in species sampling problems

Gibbs-type priors are a powerful tool for addressing prediction and estimation in species sampling problems, particularly when observations come from a population consisting of individuals from various types or species. An important applied problem is the estimation of the overall species diversity, specifically by predicting the number $K_m^{(n)} = K_{n+m} - K_n$ of "new" distinct species that will be observed in an additional sample of size m, given a sample of size n that has already been observed with an in-sample richness denoted by K_n. Here, $K_m^{(n)}$ represents the out-of-sample richness.

The a priori distribution of K_n induced by a Gibbs-type prior is provided by De Blasi et al. (2015) and it has a simple form:

$$\mathbb{P}(K_n = k) = V_{n,k} \frac{\mathcal{C}(n, k; \sigma)}{\sigma^k}, \tag{1.7}$$

where $\mathcal{C}(n, k; \sigma)$ denotes a generalised factorial coefficient (Charalambides (2005)) and it is equal to:

$$\mathcal{C}(n, k; \sigma) = \frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} (-i\sigma)_n.$$

We obtain this distribution marginalising Equation (1.3).

Given the prior distribution, we can obtain the a priori expected values $\mathbb{E}[K_1], \ldots, \mathbb{E}[K_n]$, which can be interpreted as a model-based rarefaction curve (Zito et al. (2023)). The rarefaction curve is a visual tool often used in ecology and biodiversity research to estimate species richness. It shows how the number of observed species changes as more individuals are sampled. This tool helps researchers understand how species richness increases with more sampling and also allows for comparison of species richness across different habitats or communities while taking into account differences in sample size. In the Dirichlet process case, the rarefaction curve is given by the Equation (1.5).

One of the advantages of using Gibbs-type priors with σ belonging to the interval $[0, 1)$ is that they are particularly suited when the population is composed of a large number of unknown species and the observed sample $X_1, \ldots, X_n$ contains only a small fraction of the species in the population. Note that $K_n = k$ is a sufficient statistic for predictions. Based on the EPPF, an explicit expression for the distribution of new observed distinct clusters in a new additional sample, $K_m^{(n)}$, conditionally on the information provided by $X_1, \ldots, X_n$, is defined by De Blasi et al. (2015) and it is as follows:

$$\mathbb{P}(K_m^{(n)} = j \mid X_1, \ldots, X_n) = \frac{V_{n+m,k+j}}{V_{n,k}} \cdot \frac{\mathcal{C}(m, j; \sigma, -n + k\sigma)}{\sigma^j} \tag{1.8}$$

where $X_1, \ldots, X_n$ are partitioned into $K_n = k$ clusters with frequencies $n_1, \ldots, n_k$ and $\mathcal{C}(m, j; \sigma, -n + k\sigma)$ is the non-central generalised factorial coefficient:

$$\mathcal{C}(m, j; \sigma, -n + k\sigma) = (j!)^{-1} \sum_{r=0}^{j} (-1)^r \binom{j}{r} (n - \sigma(r + k))_m.$$

The collection $\mathbb{E}[K_{n+1} \mid X_1, \ldots, X_n], \ldots, \mathbb{E}[K_{n+m} \mid X_1, \ldots, X_n]$ represents a model-based extrapolation curve. Considering a Dirichlet process, we have:

$$\mathbb{E}[K_{n+m} \mid \alpha, X_1, \ldots, X_n] = k + \sum_{i=1}^{m} \frac{\alpha}{\alpha + n + i - 1}.$$

The Bayesian nonparametric estimator for the number of distinct species considering a squared loss function is equal to the posterior expected values:

$$\hat{K}_m^{(n)} = \mathbb{E}[K_m^{(n)} \mid X_1, \ldots, X_n]. \tag{1.9}$$

Considering a Pitman-Yor process with parameters (α, σ) the posterior distribution is:

$$\mathbb{P}(K_m^{(n)} = j \mid X_1, \dots, X_n) = \frac{(\alpha/\sigma + k)_j}{(\alpha + n)_m} \, \mathcal{C}(m, j; \sigma, -n + k\sigma).$$

As consequence, Equation (1.9) becomes:

$$\hat{K}_m^{(n)} = \left(k + \frac{\alpha}{\sigma}\right) \left(\frac{(\alpha + n + \sigma)_m}{(\alpha + n)_m} - 1 \right).$$

The main advantage of this formulation is that it is explicit and can be exactly evaluated even when the size m of the additional sample is large compared to the size of the basic sample n.

Rarefaction and extrapolation curves are useful for understanding biodiversity, but summarising biodiversity with a single number may be challenging. The concept of richness, defined as the asymptotic value of an accumulation curve $\lim_{n\to\infty} K_n$, is not a good measure of biodiversity. Richness diverges when $\sigma \geqslant 0$ regardless of the observed data, but remains finite when $\sigma < 0$. Even though it could be tempting to stay away from models in the $\sigma \geqslant 0$ regime, there are good reasons not to. Reliable richness estimation usually occurs when accumulation curves stabilise, indicating saturation. In contrast, the estimation of richness in the early stages is often imprecise and can be compared to a random guess. Furthermore, models with $\sigma \geqslant 0$ often perform well in predicting future values $K_m^{(n)}$ in rapidly growing curves, compared to models with $\sigma \leqslant 0$. Estimating the total number of distinct values present in a given area is also possible, even in models where $\sigma \geqslant 0$. Models that currently exist focus on the $\sigma \geqslant 0$ case. This discussion highlights the need for a broader and more informative concept of diversity, called σ-diversity, introduced by Pitman (2003).

Chapter 2

The Gnedin Model

In 2010, Alexander Gnedin introduced a new model that can be seen as a special case of the Gibbs-type priors family, like the Dirichlet and the Pitman-Yor processes. This chapter is focused on the definition and investigation of the Gnedin model. We want to demonstrate some new results, like the prior expected value of distinct species and the posterior expected value of new species discovered and not observed in the initial sample. or the posterior distribution of the random probability measure of the Gnedin model.

2.1 Model definition

The innovative model introduced by Gnedin is a two-parameter species sampling model, provided with a straightforward update mechanism (Gnedin (2010)). Each partition is defined by a Dirichlet sampling model with a random number of components, H, representing the total distinct values in the entire population. This model is a specific type of Gibbs-type prior where σ is equal to -1, and the characterising parameter γ belongs to the interval $(0,1)$. We have seen in the previous chapter that Gibbs-type priors are characterized by their associated Exchangeable Partition Probability Function (EPPF). For this model, the weights also have a closed-form expression, as defined in Section 6 of Gnedin (2010). Given this result, we can specify the EPPF in Equation (1.3) and provide a definition for the Gnedin model.

Definition 2.1. Given a sample of n observations with k distinct values among it, the weights $V_{n,k}$ of the Gnedin model are as follows:

$$V_{n,k} = \frac{(k-1)!\,(1-\gamma)_{k-1}\,(\gamma)_{n-k}}{(n-1)!\,(1+\gamma)_{n-1}}. \tag{2.1}$$

If $\sigma = -1$ and $\gamma \in (0,1)$ the EPPF for the Gnedin model is defined as:

$$\Pi_n(n_1,\dots,n_k) = V_{n,k}\prod_{j=1}^{k} n_j! \tag{2.2}$$

where $(n_1,\dots,n_k)$ are the counts of the k distinct values in the sample such that $\sum_{j=1}^{k} n_j = n$ and the weights $V_{n,k}$ as in (2.1).

We can demonstrate that when $\sigma = -1$, the product in Equation (1.3) simplifies as follows:

$$\prod_{j=1}^{k}(1-\sigma)_{n_j-1} = \prod_{j=1}^{k}(1-(-1))_{n_j-1} = \prod_{j=1}^{k}(2)_{n_j-1}.$$

Next, we express the rising factorial $(2)_{n_j-1}$ using the gamma function:

$$(2)_{n_j-1} = \frac{\Gamma(2+n_j-1)}{\Gamma(2)}.$$

Substituting this into the product, we get:

$$\prod_{j=1}^{k}(2)_{n_j-1} = \prod_{j=1}^{k}\frac{\Gamma(2+n_j-1)}{\Gamma(2)} = \prod_{j=1}^{k} n_j!.$$

Thus, the product simplifies to $\prod_{j=1}^{k} n_j!$, confirming that Equation (2.2) is indeed satisfied.

Since we have defined the model, we can determine the random probability measure associated with the Gnedin model as in Section 5 of Gnedin (2010).

Remark 2.1. Given a distribution P on $\mathbb{X}$, termed base measure, a parameter $\sigma = -1$ and a parameter $\gamma \in (0,1)$ and considering two different independent sequences of weights $(\pi_h)_{h\geqslant 1}$ and atoms $(X_h)_{h\geqslant 1}$, defined as:

$$X_h \overset{\text{iid}}{\sim} P$$
$$(\pi_1,\dots,\pi_H) \mid H \sim \text{Dir}_H(1,\dots,1)$$

the random measure $\tilde{p}$ in de Finetti representation of $X_1, X_2, \dots$ is as follows:

$$X_h \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p} \tag{2.3}$$
$$\tilde{p} = \sum_{h=1}^{H} \pi_h \delta_{X_h}.$$

An important consideration is that the random number of components in the entire population, H, has a highly treatable and heavy-tailed prior distribution. It is defined in Gnedin (2010) [Equation (9)] as follows:

$$\mathbb{P}(H = h) = \frac{\gamma\,(1-\gamma)_{h-1}}{h!}. \tag{2.4}$$

2.1.1 Posterior distribution for the true number of distinct species

From the prior distribution of H, we can obtain a closed-form expression for the posterior distribution, as provided by Gnedin (2010) [Equation (10)].

Theorem 2.1. *Given a sample* $(X_1, \ldots, X_n)$ *with* $K_n = k \leqslant n$ *distinct values among it, the posterior distribution for the number of components* H *is as follows:*

$$\mathbb{P}(H = h \mid X_1, \ldots, X_n) = \frac{(n-1)!}{(k-1)!(h+n-1)} \prod_{i=1}^{k-1}(h-i) \prod_{j=1}^{k}(\gamma+n-j) \times \prod_{l=k}^{h-1}(l-\gamma). \tag{2.5}$$

Moreover, the expected value associated with the posterior distribution has an important role. It is defined as follows:

$$\mathbb{E}[H \mid X_1, \ldots, X_n] = \sum_{h=1}^{\infty} \mathbb{P}(H = h \mid X_1, \ldots, X_n) \cdot h. \tag{2.6}$$

We will show in the next chapter that given a sufficiently large value for h, the truncated posterior expected value for H, $\mathbb{E}[H \mid X_1, \ldots, X_n]$ converges to the total number of distinct species in the population, H, also known as species richness in the ecology problems. This provides a strong framework for predicting species diversity in large samples.

2.1.2 Predictive distribution

In Section 1.2.3 we have defined the general form of the predictive distribution in Equation (1.4). With the EPPF, we can easily derive this distribution as defined by Gnedin (2010) in Section 5. In particular, the probability of being a new sample from $P(\cdot)$, is

$$P_{\text{new}} = \frac{V_{n+1,k+1}}{V_{n,k}}$$

and the probability of being a value that has already been observed X_j^* is

$$P_{old}^{(j)} = \frac{V_{n+1,k}}{V_{n,k}}(n_j - \sigma) \qquad j = 1, \dots, k.$$

For the Gnedin model, considering the definition of the EPPF and its associated weights provided in Equation (2.1), with fixed $\sigma = -1$, the probabilities become

$$P_{new} = \frac{k(k-\gamma)}{n(\gamma+n)} \qquad P_{old}^{(j)} = \frac{\gamma+n-k}{n(\gamma+n)}(n_j+1). \tag{2.7}$$

Hence, substituting these probabilities into the generic predictive distribution for the Gibbs-type priors in Equation (1.4), it becomes:

$$P(A \mid X_1, \dots, X_n) = \frac{k(k-\gamma)}{n(\gamma+n)}P(A) + \sum_{j=1}^{k}\frac{(\gamma+n-k)}{n(\gamma+n)}(n_j+1)\delta_{X_j^*}(A)$$

where $(X_1^*, \dots, X_k^*)$ with $k \leqslant n$ are the distinct values among $(X_1, \dots, X_n)$ and A is a measurable set.

2.2 Posterior distribution of the Gnedin model

Among the quantities not yet detailed for the Gnedin model is the posterior distribution of the random probability measure $\tilde{p}$. The prior and posterior expected values of the number of unique species K_n within the sample will be defined later in the chapter. We want to determine and use this posterior distribution for future applications, like mixture models.

2.2.1 Hierarchical representation of the posterior distribution

We have previously observed that the Gnedin model can be interpreted as a Dirichlet process with a random number of components. This insight is particularly useful because it ensures that the conjugacy property (see Section 1.3.6) still holds, conditionally on the total number of components. This implies that the posterior distribution of weights remains a Dirichlet distribution:

$$(\pi_1, \dots, \pi_H) \mid H, X_1, \dots, X_n \sim Dir_H(1+n_1, \dots, 1+n_H). \tag{2.8}$$

We have two methods for obtaining the posterior distribution, and we can show that the final result is identical using both ways. The first solution

uses a hierarchical representation. Considering the Gnedin model defined as Equation (2.3), this method can be summarized in two steps: firstly, we sample $H \mid X_1, \ldots, X_n$ from the posterior distribution in Equation (2.5) and then given H, we sample $(\pi_1, \ldots, \pi_h) \mid H, X_1, \ldots, X_n$ from the Dirichlet distribution in Equation (2.8).

2.2.2 Latent variable representation of the posterior distribution

The second approach for obtaining the posterior distribution exploits the results of Argiento & De Iorio (2022) on Normalised Independent Finite Point Processes (NIFPPs), which include Gibbs-type priors as a special case. This work provides a detailed description of the posterior distribution of NIFPPs. Starting from these results, we adapt and specify them for the Gnedin model. We will first define the NIFPPs in the next section and then specify the posterior distribution of these processes and, in particular, the Gnedin model.

Normalized Independent Finite Poisson Processes

Firstly, we introduce the concept of Independent Finite Point Processes (IFPP) and their normalized version

Definition 2.2. Let $\nu(\cdot)$ be a density on $\mathbb{R}$ and $\mathbb{P}(H = h)$, $h = 0, 1, \ldots$ be a probability mass function. X is an Independent Finite Point Process (IFPP), denoted as $X \sim \text{IFPP}(\nu, \mathbb{P}(H = h))$, if its Janossy density (see Appendix B) can be written as:

$$j(\xi_1, \ldots, \xi_h) = h!\, \mathbb{P}(H = h) \prod_{j=1}^{h} \nu(\xi_j)$$

where:

i. $\xi_1, \ldots, \xi_h$ are the locations of the points;

ii. $\mathbb{P}(H = h)$ is the probability mass function that gives the probability of having h points;

iii. $\nu(\xi_j)$ is the density function evaluated at the location ξ_j.

Definition 2.3. Let $\mathcal{P} = \{(S_1, \tau_1), \ldots, (S_H, \tau_H)\} \sim \text{IFPP}(\nu, \mathbb{P}(H = h), P)$, with $\mathbb{P}(H = 0) = 0$. A normalized independent finite point process with parameters ν,

$\mathbb{P}(H = h)$, and P, is a discrete probability measure on $\mathbb{X}$ defined by

$$\tilde{p}(A) = \sum_{h=1}^{H} \pi_h \delta_{\tau_h}(A) \stackrel{d}{=} \sum_{h=1}^{H} \frac{S_h}{T} \delta_{\tau_h}(A), \tag{2.9}$$

where $T = \sum_{h=1}^{H} S_h$ and A denotes a measurable set of $\mathbb{X}$. We will write $\tilde{p} \sim \text{NIFPP}(\nu, \mathbb{P}(H = h), P)$.

For example, suppose ν, which represents the distribution of the random weights S_h, is a Gamma$(\tau, 1)$ density with a shape parameter τ greater than 0 and rate 1. Then, the NIFPP is a finite Dirichlet process, as in Equation (2.9). Conditionally on $H > 0$, the jump sizes $(\pi_1, \ldots, \pi_H)$ of $\tilde{p}$ are a sample from the H-dimensional $\text{Dir}_H(\tau, \ldots, \tau)$ distribution. The same result can also be interpreted as a Gibbs-type prior with a negative parameter. For more details, see De Blasi et al. (2015).

Having defined these quantities, we will now review the work of Argiento & De Iorio (2022) to establish the posterior distribution. Finally, we will provide proof of the equivalence of the two posterior representations.

Given the observations $(X_h)_{h \geqslant 1}$ such that

$$X_h \mid \tilde{p} \stackrel{iid}{\sim} \tilde{p}.$$

Assume that the prior distribution for $\tilde{p}$ is a NIFFP:

$$\tilde{p} = \sum_{h=1}^{H} \frac{S_h}{T} \delta_{Z_h} \tag{2.10}$$

where:

i. the atoms $Z_1, \ldots, Z_H$ are distributed as $Z_i \stackrel{iid}{\sim} P$;

ii. the random weights $(S_1, \ldots, S_H)$ are such that

$$S_h \sim \text{Gamma}(\tau, 1) \qquad h = 1, \ldots, H$$

with shape parameter τ and scale parameter 1;

iii. $T = \sum_{h=1}^{H} S_h$, this choice is necessary because $\tilde{p}$ must be a probability, so the weights must be normalized.

The model we have depicted above is equal to the one defined in Equation (2.3). However, in this case, we have characterized the weights differently, using the construction of the Dirichlet distribution via a Gamma distribution as described in Section 1.3.1. It follows that $(S_1/T, \cdots, S_H/T) \sim \text{Dir}_H(\tau, \cdots, \tau)$. Given the prior distribution for H as in Equation (2.4), we aim to demonstrate that this construction leads to the Gnedin model. To achieve this, we will consider the Exchangeable Partition Probability Function (EPPF) associated with the model we have just defined, because, as we have said before, EPPF is characterising the model.

Theorem 2.2. *Consider the statistical model* $X_i \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}$, *where* $\tilde{p}$ *is the random probability measure in* (2.10). *Assume that the parameter* τ *is equal to* 1. *The EPPF of a sample of size* n *equals that obtained by considering the random probability measure* $\tilde{p}$ *constructed from Dirichlet weights and defined in Equation* (2.2).

Proof. We aim to prove that the EPPFs of both representations are equal. The EPPF associated with the random probability measure in Equation (2.10) is given by:

$$\Pi_n(n_1, \ldots, n_k) = V_{n,k} \cdot \prod_{h=1}^{H} \frac{\Gamma(1+n_h)}{\Gamma(1)},$$

where

$$V_{n,k} = \sum_{h=0}^{+\infty} \frac{(h+k)!}{h!} \mathbb{P}(H = h+k) \frac{\Gamma(k+h)}{\Gamma(k+h+n)}.$$

Substituting the expression for $\mathbb{P}(H = h+k)$, we have:

$$V_{n,k} = \sum_{h=0}^{+\infty} \frac{(h+k)!}{h!} \cdot \frac{\gamma \cdot (1-\gamma)_{h+k-1}}{(h+k)!} \cdot \frac{\Gamma(k+h)}{\Gamma(k+h+n)}.$$

Now, simplifying the terms and explicitly writing the Pochhammer symbol $(a)_n = \Gamma(a+n)/\Gamma(a)$:

$$V_{n,k} = \sum_{h=0}^{+\infty} \frac{1}{h!} \cdot \frac{\Gamma(h+k-\gamma)}{\Gamma(1-\gamma)} \cdot \gamma \cdot \frac{\Gamma(k+h)}{\Gamma(k+h+n)}.$$

This simplifies further to:

$$V_{n,k} = \frac{\gamma}{\Gamma(1-\gamma)} \cdot \sum_{h=0}^{+\infty} \frac{\Gamma(h+k-\gamma)}{\Gamma(h+1)} \cdot \frac{\Gamma(k+h)}{\Gamma(k+h+n)}.$$

Rewrite the terms of the summation using the definition of the Pochhamer symbol.

$$\sum_{h=0}^{+\infty} \frac{\Gamma(h+k-\gamma)}{\Gamma(h+1)} \frac{\Gamma(k+h)}{\Gamma((k+h)+n)} = \sum_{h=0}^{+\infty} \frac{(k-\gamma)_h \, \Gamma(k-\gamma)}{h!} \frac{\Gamma(k+h)}{\Gamma(k+h+n)}$$
$$= \sum_{h=0}^{+\infty} \frac{(k-\gamma)_h \, \Gamma(k-\gamma)}{h!} \frac{(k)_h \, \Gamma(k)}{(k+n)_h \, \Gamma(k+n)}.$$

To simplify the expression, we first recall the definition of the hypergeometric function ${}_2F_1$ given in Equation (A.2). Using this, we can identify the constant factor and reduce the summation to a hypergeometric function form.

$$\begin{aligned}
\sum_{h=0}^{+\infty} \frac{\Gamma(h+k-\gamma)}{\Gamma(h+1)} \frac{\Gamma(k+h)}{\Gamma((k+h)+n)} &= \frac{\Gamma(k-\gamma)\,\Gamma(k)}{\Gamma(k+n)} \sum_{h=0}^{+\infty} \frac{(k-\gamma)_h}{h!} \frac{(k)_h}{(k+n)_h} \\
&= \frac{\Gamma(k-\gamma)\,\Gamma(k)}{\Gamma(k+n)} \sum_{h=0}^{+\infty} \frac{(1)^h}{h!} \frac{(k-\gamma)_h \, (k)_h}{(k+n)_h} \\
&= \frac{\Gamma(k-\gamma)\,\Gamma(k)}{\Gamma(k+n)} \cdot {}_2F_1(k-\gamma, k, k+n, 1) \\
&= \frac{\Gamma(k-\gamma)\,\Gamma(k)}{\Gamma(k+n)} \cdot \frac{\Gamma(k+n)\Gamma(k+n-k-k+\gamma)}{\Gamma(k+n-k)\Gamma(k+n-k+\gamma)} \\
&= \frac{\Gamma(k-\gamma)\,\Gamma(k)}{\Gamma(k+n)} \cdot \frac{\Gamma(k+n)\,\Gamma(n-k+\gamma)}{\Gamma(n)\,\Gamma(n-\gamma)}.
\end{aligned}$$

As consequence, the weights $V_{n,k}$ results as follows:

$$\begin{aligned}
V_{n,k} &= \frac{\gamma}{\Gamma(1-\gamma)} \cdot \frac{\Gamma(k-\gamma)\Gamma(k)}{\Gamma(k+n)} \cdot \frac{\Gamma(k+n)\Gamma(n-k+\gamma)}{\Gamma(n)\Gamma(n-\gamma)} \\
&= \gamma \cdot (1-\gamma)_{k-1} \cdot (k-1)! \cdot \frac{\Gamma(n-k+\gamma)}{\Gamma(n)\Gamma(n+\gamma)} \\
&= \gamma \cdot (1-\gamma)_{k-1} \cdot (k-1)! \cdot \frac{(\gamma)_{n-k} \cdot \Gamma(\gamma)}{(n-1)! \cdot \Gamma(n+\gamma)} \\
&= (1-\gamma)_{k-1} \cdot (k-1)! \cdot \frac{(\gamma)_{n-k} \cdot \Gamma(\gamma+1)}{(n-1)! \cdot \Gamma(n-1+\gamma+1)} \\
&= (1-\gamma)_{k-1} \cdot (k-1)! \cdot \frac{(\gamma)_{n-k}}{(n-1)! \cdot (\gamma+1)_{n-1}}.
\end{aligned}$$

This result matches the weights $V_{n,k}$ for the Gnedin model as specified in Equation (2.1). The EPPF becomes

$$\Pi_n(n_1, \ldots, n_k) = V_{n,k} \cdot \prod_{h=1}^{H} \frac{\Gamma(1+n_j)}{\Gamma(1)} = V_{n,k} \cdot \prod_{h=1}^{H} n_j!$$

the same as the one defined above in Equation (2.2). It proves we can use the results from Argiento & De Iorio (2022). □

Before determining the posterior distribution, we first reformulate the prior distribution of $\tilde{p}$ in Equation (2.10) as follows:

$$\mu = \sum_{h=1}^{H} S_h \, \delta_{X_h},$$

where S_h and X_h are defined accordingly. Consequently,

$$\tilde{p} = \frac{\mu}{\mu(\mathbb{X})},$$

where $\mathbb{X}$ denotes the entire space and $\mu(\mathbb{X}) = T$. We will now proceed to characterize the posterior distribution of μ and, as a consequence, the posterior distribution of $\tilde{p}$. As shown in Example 4.3 of Argiento & De Iorio (2022), to determine the required distribution, we introduce a suitable latent variable U_n, such that the density distribution is as follows:

$$f_{U_n}(u \mid \boldsymbol{x}) \propto \frac{u^{n-1}}{\Gamma(n)} \cdot \left\{ \sum_{h=0}^{+\infty} \frac{(h+k)!}{h!} \cdot (\Psi(u))^h \cdot \mathbb{P}(H = h+k) \right\} \cdot \prod_{j=1}^{k} \varkappa(n_j, u), \quad (2.11)$$

where

$$\Psi(u) = \frac{1}{(1+u)^1},$$

$$\varkappa(n_j, u) = \frac{1}{(1+u)^{n_j+1}} \cdot n_j!.$$

Using the prior distribution of H in Equation (2.4), the summation term in the density distribution of U_n can be simplified as:

$$\begin{aligned}&\sum_{h=0}^{+\infty} \frac{(h+k)!}{h!} \cdot \frac{1}{(1+u)^h} \cdot \gamma \cdot \frac{(1-\gamma)_{h+k-1}}{(h+k)!} \\ =&\frac{\gamma}{\Gamma(1-\gamma)} \sum_{h=0}^{+\infty} \frac{1}{h!} \cdot \frac{1}{(1+u)^h} \cdot \Gamma(h+k-\gamma).\end{aligned}$$

Use the definition of Gamma function: $\Gamma(z) = \int_0^{+\infty} e^{-t}\, t^{z-1}\, dt$ into the expression of the weights $V_{n,k}$ and it results as:

$$\begin{aligned}V_{n,k} =&\frac{\gamma}{\Gamma(1-\gamma)} \sum_{h=0}^{+\infty} \frac{1}{h!} \cdot \frac{1}{(1+u)^h} \int_0^{+\infty} e^{-x} \cdot x^{h+k-\gamma-1}\, dx \\ =&\frac{\gamma}{\Gamma(1-\gamma)} \int_0^{+\infty} e^{-x} \cdot x^{k-\gamma-1} \sum_{h=0}^{+\infty} \frac{\left(\frac{x}{1+u}\right)^h}{h!}\, dx \\ =&\frac{\gamma}{\Gamma(1-\gamma)} \int_0^{+\infty} e^{-x} \cdot x^{k-\gamma-1} \cdot e^{\frac{x}{1+u}}\, dx \\ =&\frac{\gamma}{\Gamma(1-\gamma)} \int_0^{+\infty} e^{-x\left(1-\frac{1}{1+u}\right)} \cdot x^{k-\gamma-1}\, dx \\ =&\frac{\gamma}{\Gamma(1-\gamma)} \int_0^{+\infty} e^{-x\cdot\frac{u}{1+u}} \cdot x^{k-\gamma-1}\, dx = \frac{\gamma}{\Gamma(1-\gamma)} \cdot \frac{\Gamma(k-\gamma)}{\left(\frac{u}{1+u}\right)^{k-\gamma}}.\end{aligned}$$

Then, the kernel of the density distribution of U_n is:

$$f_{U_n}(u \mid \boldsymbol{x}) \propto \frac{u^{n-1}}{\Gamma(n)} \left(\frac{u}{1+u}\right)^{\gamma-k} \gamma(1-\gamma)_{k-1} \cdot \prod_{j=1}^{k} \frac{1}{(u+1)^{n_j+1}} \cdot n_j!.$$

We can derive the distribution of μ as a special case of Theorem 2 in Argiento & De Iorio (2022).

Proposition 2.1. *Conditional on the observed data $X_1, \ldots, X_n$ and latent variable U_n the posterior distribution of the random variable μ is the sum of two independent components:*

$$\mu \mid X_1, \ldots, X_n, U_n \overset{d}{=} \mu^{(a)} + \mu^{(na)}.$$

Proof. To prove this equivalence in distribution, we must demonstrate that the two components, $\mu^{(a)}$ and $\mu^{(na)}$, are independent. Additionally, we need to analyze these two processes to ensure that they satisfy the conditions stipulated

by the theorem of Argiento & De Iorio (2022). The first term, $\mu^{(a)}$, is the allocated jumps process defined as follows:

$$\mu^{(a)} = \sum_{j=1}^{K_n} S_j^{(a)} \cdot \delta_{X_j^*}.$$

Here, K_n is the number of distinct clusters identified in the sample $(X_1, \ldots, X_n)$. For each cluster j, $S_j^{(a)}$ represents the size (or weight) of the jump at location X_j^*. The notation $\delta_{X_j^*}$ is a Dirac delta function. The density function of the weight $S_j^{(a)}$ is given as:

$$f_{S_j^{(a)}}(s) \propto e^{-us} s^{n_j} \phi(s)$$

where n_j is the number of observations in cluster j. Since $\phi(s)$ follows a Gamma distribution with scale and shape parameters both equal to 1, it is equivalent to an exponential distribution with a rate 1 and so we have:

$$\phi(s) = e^{-s}.$$

Substituting this result into the density function, we get:

$$f_{S_j^{(a)}}(s) \propto e^{-us} s^{n_j} e^{-s} = e^{-s(u+1)} s^{n_j}.$$

This is the density function of a Gamma distribution with shape parameter $n_j + 1$ and rate parameter $u + 1$. Therefore,

$$S_j^{(a)} \sim \mathrm{Gamma}(n_j + 1, u + 1).$$

We have demonstrated that the distribution of S_j aligns with the conditions specified in the theorem by Argiento & De Iorio (2022). The second term is the non-allocated jumps process, $\mu^{(na)}$, which is defined as:

$$\mu^{(na)} = \sum_{m=1}^{M^*} s_m^* \delta_{\tau_m^*}.$$

Here, M^* is the number of non-allocated jumps, s_m^* is the size of the m-th non-allocated jump, and τ_m^* is the location of the m-th non-allocated jump, with τ_m^* i.i.d. according to the base measure P:

$$\tau_m^* \overset{\text{iid}}{\sim} P.$$

The distribution of M^* is given by:

$$\mathbb{P}(M^* = m) \propto \frac{(m+k)!}{m!} \cdot (\Psi(u))^m \cdot \mathbb{P}(H = m+k) \tag{2.12}$$

where

$$\Psi(u) = \frac{1}{(1+u)}.$$

The size of each non-allocated jump, s^*_m, follows a modified distribution:

$$\phi^*(s) = e^{-us} \cdot \phi(s).$$

Since $\phi(s)$ is $Gamma(1,1)$ or equivalently $Exp(1)$, we have:

$$\phi^*(s) = e^{-us}e^{-s} = e^{-s(u+1)}.$$

Thus, $s^*_m \sim \text{Exp}(u+1)$, and the non-allocated jumps process $\mu^{(na)}$ is an independent finite point process, as required. Given the definition and construction of $\mu^{(a)}$ and $\mu^{(na)}$, and their conditional distributions given U_n and M^*, these two processes are independent. Consequently, the random measure μ is distributed as the sum of two independent components: the allocated jumps process $\mu^{(a)}$ and the non-allocated jumps process $\mu^{(na)}$. This completes the proof. □

Lastly, the total number of jumps H given the data $X_1, \ldots, X_n$ and the latent variable U_n, is decomposed into the number of allocated jumps K_n and the number of non-allocated jumps M^*, where:

$$H \mid X_1, \ldots, X_n, U_n = K_n + M^*.$$

Here, K_n is the number of distinct components observed in the data, and M^* is the number of additional predicted but not yet observed clusters. Summarizing, this second strategy involves the following steps: sample $U_n \mid X_1, \ldots, X_n$. Given an observation for the latent variable, sample $H \mid X_1, \ldots, X_n, U_n$. Then, derive $\mu \mid X_1, \ldots, X_n, U_n, H$. Given these steps, the posterior distribution for $\tilde{p}$ can be obtained as described previously.

2.2.3 Equivalence of the two posterior representations

In this section, we want to show the equivalence of the two methods. We want to demonstrate that by marginalizing the latent variable, U_n, we obtain the

hierarchical representation. The advantage of applying this representation is that it is easier to use, as sampling from the density of the latent variables is not straightforward. We first determine the normalizing constant of the density function of U_n as follows:

$$\int_0^{+\infty} f_{U_n}(u \mid \boldsymbol{x})du = \int_0^{+\infty} \frac{u^{n-1}}{\Gamma(n)}\gamma \left(\frac{u}{1+u}\right)^{\gamma-k} (1-\gamma)_{k-1} \frac{1}{(u+1)^{n+k}} \prod_{j=1}^{k} n_j!\, du$$
$$= \frac{\gamma\,\Gamma(k-\gamma)}{\Gamma(n)\Gamma(1-\gamma)} \prod_{j=1}^{k} n_j! \cdot \int_0^{+\infty} \frac{u^{n-1-k+\gamma}}{(u+1)^{n+\gamma}}\, du.$$

Let focus on the last term

$$\int_0^{+\infty} \frac{u^{n-1-k+\gamma}}{(u+1)^{n+\gamma}}\, du = \int_0^{+\infty} \left(\frac{u}{u+1}\right)^{n+\gamma} u^{-1-k}\, du.$$

To simplify this, we perform a change of variable. Let $\frac{u}{1+u} = x$, i.e. $u = \frac{x}{1-x}$. Substituting these into the integral, we have:

$$\int_0^1 x^{n+\gamma} \left(\frac{x}{1-x}\right)^{-1-k} \frac{1}{(1-x^2)}\, dx = \int_0^1 x^{n+\gamma-1-k}(1-x)^{k-1}\, dx = \mathcal{B}(n+\gamma-k, k)$$

where $\mathcal{B}(z_1, z_2) = \int_0^1 x^{z_1-1}(1-x)^{z_2-1}\, dx$ is the Beta function. The density function of U_n can be expressed as follows:

$$f_{U_n}(u \mid \boldsymbol{x}) = \frac{u^{n-1}}{\Gamma(n)}\,\gamma \left(\frac{u}{1+u}\right)^{\gamma-k} (1-\gamma)_{k-1}\, \frac{1}{(u+1)^{n+k}} \prod_{j=1}^{k} n_j!$$
$$\times \frac{1}{\frac{\gamma\,\Gamma(k-\gamma)}{\Gamma(n)\Gamma(1-\gamma)} \prod_{j=1}^{k} n_j!\, \mathcal{B}(n+\gamma-k, k)}.$$

We focus on the non-allocated jumps M^*. We want to show that by integrating U_n from the normalized posterior distribution of M^*, we can obtain the posterior distribution of H defined by Gnedin (2010) and in this thesis reported in Equation (2.5). We have already defined the distribution of M^* in Equation (2.12). In this context, we use the equivalent notation $\mathbb{P}(M^* = m \mid U_n)$ to highlight the connection with U_n. Marginalizing out the variable U_n from $\mathbb{P}(M^* = m \mid U_n)$ means:

$$\int_0^{+\infty} \mathbb{P}(M^* = m \mid U_n)\, f_{U_n}(u \mid \boldsymbol{x})\, du. \tag{2.13}$$

To make the calculations easier, we rewrite the density function of U_n in Equation (2.11) considering the distribution $\mathbb{P}(M^* = m)$ and the integral becomes:

$$\begin{aligned}
&\int_0^{+\infty} \mathbb{P}(M^* = m \mid U_n)\, \frac{u^{n-1}}{\Gamma(n)} \sum_{m=1}^{M^*} \mathbb{P}(M^* = m) \frac{1}{(u+1)^{n+k}} \\
&\qquad \times \prod_{j=1}^{k} n_j! \, \frac{1}{\int_0^{+\infty} f_{U_n}(u \mid \boldsymbol{x})\, du}\, du \\
&= \int_0^{+\infty} \frac{(m+k)!}{m!} \left(\frac{1}{1+u}\right)^m \frac{\gamma(1-\gamma)_{m+k-1}}{(m+k)!} \frac{u^{n-1}}{\Gamma(n)} \frac{1}{(u+1)^{n+k}} \\
&\qquad \times \prod_{j=1}^{k} n_j! \, \frac{1}{\frac{\gamma\,\Gamma(k-\gamma)}{\Gamma(n)\Gamma(1-\gamma)} \prod_{j=1}^{k} n_j!\, \mathcal{B}(n+\gamma-k, k)}\, du \\
&= \frac{1}{m!}\,(1-\gamma)_{m+k-1} \cdot \frac{\Gamma(1-\gamma)}{\Gamma(k-\gamma)\,\mathcal{B}(n+\gamma-k,k)} \int_0^{+\infty} \frac{u^{n-1}}{(u+1)^{m+k+n}}\, du \\
&= \frac{(1-\gamma)_{m+k-1}}{(1-\gamma)_{k-1}} \frac{1}{\mathcal{B}(n+\gamma-k,k)} \frac{1}{m!} \int_0^{+\infty} \frac{u^{n-1}}{(u+1)^{m+k+n}}\, du.
\end{aligned}$$

Let us focus on the last term

$$\int_0^{+\infty} \frac{u^{n-1}}{(u+1)^{m+k+n+1-1}}\, du = \int_0^{+\infty} \left(\frac{u}{1+u}\right)^{n-1} \frac{1}{(u+1)^{m+k+1}}\, du.$$

By exploiting the change of variables $\frac{u}{1+u} = x$, i.e. $u = \frac{x}{1-x}$ and $\frac{du}{dx} = \frac{1}{(1-x)^2}$, we obtain:

$$\int_0^1 x^{n-1}(1-x)^{m+k+1} \frac{1}{(1-x)^2}\, dx = \int_0^1 x^{n-1}(1-x)^{m+k-1}\, dx = \mathcal{B}(n, m+k).$$

The initial integral in Equation (2.13) is therefore:

$$\begin{aligned}
&\frac{(1-\gamma)_{m+k-1}}{(1-\gamma)_{k-1}} \frac{1}{\mathcal{B}(n+\gamma-k)} \frac{1}{m!} \mathcal{B}(n, m+k) \\
&= \frac{\Gamma(m+k+\cancel{-1}+\cancel{1}-\gamma)}{\Gamma(\cancel{1}-\gamma+k\cancel{-1})} \frac{1}{m!} \frac{\mathcal{B}(n, m+k)}{\mathcal{B}(n+\gamma-k, k)}.
\end{aligned}$$

A key property of the Beta function is its close relationship to the Gamma function: $\mathcal{B}(z_1, z_2) = (\Gamma(z_1)\Gamma(z_2))/\Gamma(z_1 + z_2)$. We use now this property:

$$
\begin{aligned}
&= \frac{\Gamma(m+k-\gamma)}{\Gamma(k-\gamma)} \frac{1}{m!} \frac{\Gamma(n)\,\Gamma(m+k)}{\Gamma(n+m+k)} \frac{\Gamma(n+\gamma)}{\Gamma(n+\gamma-k)\,\Gamma(k)} \\
&= \frac{\Gamma(n)}{\Gamma(n+m+k)\,\Gamma(k)} \frac{\Gamma(m+k)}{m!} \frac{\Gamma(m+k-\gamma)}{\Gamma(k-\gamma)} \frac{\Gamma(n+\gamma)}{\Gamma(n+\gamma-k)} \\
&= \frac{(n-1)!}{(n+m+k-1)!\,(k-1)!} \frac{\Gamma(m+k)}{\Gamma(m+1)} \frac{\Gamma(m+k-\gamma)}{\Gamma(k-\gamma)} \frac{\Gamma(n+\gamma)}{\Gamma(n+\gamma-k)}.
\end{aligned}
$$

Let us assume $m + k = h$, the expression is rewritten as:

$$
\begin{aligned}
&\int_0^{+\infty} \mathbb{P}(M^* = m \mid U_n)\, f_{U_n}(u \mid \boldsymbol{x})\, du = \frac{(n-1)!}{(n+h-1)!(k-1)!} \\
&\times (m+k-1)\ldots(m+1)\cdot(m+k-1)\ldots(k-\gamma)(n+\gamma-1)\ldots(n+\gamma-k) \\
&= \frac{(n-1)!}{(n+h-1)!(k-1)!)} \prod_{i=1}^{k-1}(h-i) \prod_{l=k}^{h-1}(l-\gamma) \prod_{j=1}^{k}(n+\gamma-j).
\end{aligned}
$$

We obtain the same distribution as in Equation (2.5). This result is very important because allows us to demonstrate that by marginalizing the latent variable, U_n, we obtain the hierarchical representation, and so the same posterior distribution of H.

Remark 2.2. The posterior distribution we just defined is essential for the use of discrete nonparametric priors, especially the Gnedin model, in hierarchical mixture models. In this thesis, we have focused exclusively on species sampling problems, for which we will provide some applied results in the next chapter. However, the definition of this new quantity opens up new and interesting prospects. Currently, the only way to perform posterior inference with the Gnedin model in mixture models is through the predictive distribution. For more details, see Moya & Walker (2024).

2.3 Expected values of distinct species: prior and posterior distributions

In the context of species sampling, this model has never been used before, and the necessary quantities have not yet been formalized. The objective of this thesis is to formalise the expected values for the number of unique species in a sample,

both a priori and a posteriori. This will allow us to verify the validity of this new model in this specific context.

2.3.1 Prior distribution of the number of distinct species in a sample

In Section 1.6.1, we have already defined the prior distribution of the number of distinct species K_n induced by a Gibbs-type prior in Equation (1.7). When σ is equal to -1, the generalized factorial coefficient in the prior distribution is given by:

$$\mathcal{C}(n, k; -1) = (-1)^k \binom{n-1}{k-1} \frac{n!}{k!}.$$

In particular, the normalization constant

$$d_{n,k} = \binom{n-1}{k-1} \frac{n!}{k!}$$

is known as the Lah number (Charalambides (2005)). For the Gnedin model, the prior distribution for the number of distinct species is then specified as follows:

$$\mathbb{P}(K_n = k) = V_{n,k} \cdot d_{n,k}. \tag{2.14}$$

All these results are provided by Gnedin (2010). Based on these, here we derive an important new result: the prior expected value for the number of distinct species K_n. As previously mentioned, this is a very useful result because it underlies the rarefaction curve and until now it had not yet been defined.

Theorem 2.3. *For the Gnedin model, the expected value for the number of distinct species K_n in a generic sample of size n is as follows:*

$$E[K_n] = \frac{n!}{(1+\gamma)_{n-1}}$$

where $\gamma \in (0, 1)$.

Proof. We are going to calculate the following quantity: $E[K_n]$

$$E[K_n] = \sum_{k=1}^{n} \mathbb{P}(K_n = k) \cdot k.$$

We begin by substituting the probability function defined in Equation (2.14) and simplifying the expression:

$$\begin{aligned}\mathbb{E}[K_n] &= \sum_{k=1}^{n} \frac{(k-1)!(1-\gamma)_{k-1}(\gamma)_{n-k}}{(n-1)!(1+\gamma)_{n-1}} \cdot \frac{(n-1)!}{(k-1)!(n-k)!} \cdot \frac{n!}{k!} \cdot k \\ &= \sum_{k=1}^{n} \frac{n!(1-\gamma)_{k-1}(\gamma)_{n-k}}{(1+\gamma)_{n-1}(k-1)!(n-k)!}.\end{aligned}$$

Next, we change the index of summation:

$$\mathbb{E}[K_n] = \sum_{k=0}^{n-1} \frac{n!(1-\gamma)_k(\gamma)_{n-k-1}}{(1+\gamma)_{n-1}k!(n-k-1)!}.$$

We then gather the constant terms and focus on the remaining terms of the summation:

$$\begin{aligned}\mathbb{E}[K_n] =& \frac{n}{(1+\gamma)_{n-1}} \cdot \sum_{k=0}^{n-1} \frac{(n-1)!}{k!(n-k-1)!} \cdot (1-\gamma)_k \cdot (\gamma)_{n-k-1} \\ =& \frac{n}{(1+\gamma)_{n-1}} \sum_{k=0}^{n-1} \binom{n-1}{k} \cdot (1-\gamma)_k \cdot (\gamma)_{n-k-1}.\end{aligned}$$

Applying Vandermonde's identity:

$$(a_1+a_2)_q = \sum_{i=0}^{q} \binom{q}{i} \cdot (a_1)_i \cdot (a_2)_{q-i},$$

we simplify the summation term and we obtain:

$$\mathbb{E}[K_n] = \frac{n}{(1+\gamma)_{n-1}} \cdot (1-\gamma+\gamma)_{n-1} = \frac{n}{(1+\gamma)_{n-1}} \cdot \frac{\Gamma(n)}{\Gamma(1)} = \frac{n!}{(1+\gamma)_{n-1}}.$$

This concludes the proof, we have obtained the prior expected value for the number of distinct species. □

2.3.2 Posterior distribution of the number of new distinct species discovered

We now look for the posterior distribution of the number of new distinct species discovered, as provided by Gnedin (2010). In Equation (1.8), we have defined the distribution of new observed distinct species $K_m^{(n)}$ in an additional sample of size m, conditionally on the information provided by $X_1, \ldots, X_n$, for the Gibbs-type

priors. As explained in Charalambides (2005), we know that when σ is equal to -1, the non-central generalized factorial coefficient coincides with the non-central Lah number, except for a difference in sign. Thus, it holds that:

$$\mathcal{C}(n,k;-1,r) = |\,\mathcal{L}(n,k,r)\,| = (-1)^n \mathcal{L}(n,k,r) = \frac{n!}{k!}\binom{n-r-1}{k-r-1}(-1)^k.$$

As a consequence, for the non-central generalized factorial coefficient in the distribution of $K_m^{(n)}$, the following equality holds:

$$\mathcal{C}(m,j;-1,-n-k) = \mathcal{L}(m,j,-n-k),$$

and the posterior distribution becomes:

$$\mathbb{P}(K_m^{(n)} = j \mid X_1,\dots,X_n) = \frac{V_{n+m,k+j}}{V_{n,k}} \cdot \frac{m!}{j!}\binom{m+n+k-1}{j+n+k-1}.$$

We want to define the posterior distribution of the number of distinct species in a sample for the Gnedin model, where the weights are defined as in Equation (2.1). Let us examine the ratio between the weights in the posterior distribution just provided:

$$\begin{aligned}
\frac{V_{n+m,k+j}}{V_{n,k}} &= \frac{(k+j-1)!\,(n-1)!}{(k-1)!\,(n+m-1)!} \cdot \frac{(1-\gamma)_{k+j-1}}{(1-\gamma)_{k-1}} \cdot \frac{(1+\gamma)_{n-1}}{(1+\gamma)_{n+m-1}} \cdot \frac{(\gamma)_{n+m-k-j}}{(\gamma)_{n-k}} \\
&= \frac{(k+j-1)!\,(n-1)!}{(k-1)!\,(n+m-1)!} \cdot \frac{\Gamma(1-\gamma+k+j-1)\,\Gamma(1-\gamma)}{\Gamma(1-\gamma)\,\Gamma(1-\gamma+k-1)} \\
&\quad \times \frac{\Gamma(1+\gamma+n-1)\,\Gamma(1+\gamma)}{\Gamma(1+\gamma)\,\Gamma(1+\gamma+n+m-1)} \cdot \frac{\Gamma(\gamma+n+m-k-j)\,\Gamma(\gamma)}{\Gamma(\gamma+n-k)\,\Gamma(\gamma)} \\
&= \frac{(k+j-1)!\,(n-1)!}{(k-1)!\,(n+m-1)!} \cdot \frac{(k-\gamma)_j\,(\gamma+n-k)_{m-j}}{(\gamma+n)_m}.
\end{aligned}$$

By formulating this ratio more effectively, in the Gnedin model, the posterior distribution of the number of distinct species in a sample can be expressed as:

$$\begin{aligned}
\mathbb{P}(K_m^{(n)} = j \mid X_1,\dots,X_n) = \frac{(k+j-1)!\,(n-1)!}{(k-1)!\,(n+m-1)!} \cdot \frac{(k-\gamma)_j\,(\gamma+n-k)_{m-j}}{(\gamma+n)_m} \\
\times \frac{m!}{j!}\binom{m+n+k-1}{j+n+k-1}.
\end{aligned} \tag{2.15}$$

As in the a priori case, it is essential to derive the expected value associated with this posterior distribution. In the posterior case as well, we prove the following result.

Theorem 2.4. *Given a sample* $X_1, \dots, X_n$ *of size* n *with* k *distinct value among it and considering an additional sample of size* m*, the posterior distribution for the number of distinct species* $K_m^{(n)}$ *is as in Equation* (2.15)*. The associated expected value is as follows:*

$$\begin{aligned} \mathbb{E}[K_m^{(n)} \mid X_1, \dots, X_n] = m\,(k-\gamma)\,(k+1)\,\frac{\Gamma(n)\,(n+m)_k\,(\gamma+n-k)_{m-1}}{(n+m)_m} \\ \times {}_3F_2(-(m+1), k+1, k-\gamma+1; k+n+1, -(\gamma+n-k+m-2); 1) \end{aligned}$$

where ${}_pF_q(a_1, \dots a_p; b_1, \dots, b_q; z)$ *is the generalized hypergeometric function. See* (A.1)*.*

Proof. We proceed by demonstrating that

$$\mathbb{E}[K_m^{(n)} \mid X_1, \dots, X_n] = \sum_{j=1}^{m} \mathbb{P}(K_m^{(n)} = j \mid X_1, \dots, X_n) \cdot j.$$

Firstly, substitute the probability function of $K_m^{(n)}$, defined in Equation (2.15) and explicit the binomial term:

$$\begin{aligned} \mathbb{E}[K_m^{(n)} \mid X_1, \dots, X_n] = \sum_{j=1}^{m} \frac{(k+j-1)!\,(n-1)!}{(k-1)!\,(n+m-1)!} \cdot \frac{(k-\gamma)\,j\,(\gamma+n-k)_{m-j}}{(n+\gamma)_m} \cdot \frac{m!}{j!} \\ \times \frac{(n-1)!\,m!\,(n+m+k-1)!}{(j+n+k-1)!\,(m-j)!} \cdot j. \end{aligned}$$

Then, change the index of the summation:

$$\begin{aligned} \mathbb{E}[K_m^{(n)} \mid X_1, \dots, X_n] = \sum_{j=0}^{m-1} \frac{(k+j)!}{j!\,(m-j-1)!} \cdot \frac{1}{(j+n+k)!} \cdot \frac{(k-\gamma)_{j+1}\,(\gamma+n-k)_{m-j-1}}{(n+\gamma)_m} \\ \times \frac{(n-1)!\,m!\,(n+m+k-1)!}{(n+m-1)!\,(k-1)!}. \end{aligned}$$

Rewrite the expression collecting the constant terms and focus on the summation term of $E[K_m^{(n)} \mid X_1, \dots, X_n]$, as follows:

$$\begin{aligned}
E[K_m^{(n)} \mid X_1, \dots, X_n] &= \frac{(n-1)!\, m!\, (n+m+k-1)!}{(n+m-1)!\, (k-1)!\, (n+\gamma)_m} \\
&\quad \times \sum_{j=0}^{m-1} \frac{(k+j)!}{j!\, (m-j-1)!} \cdot \frac{(\gamma+n-k)_{m-j}(k-\gamma)_{j+1}}{(j+n+k)!} \\
&= \frac{(n-1)!\, m\, (n+m+k-1)!}{(n+m-1)!\, (k-1)!\, (n+\gamma)_m} \cdot \sum_{j=0}^{m-1} \frac{(k+j)!\, (m-1)!}{j!\, (m-j-1)!} \cdot \frac{(\gamma+n-k)_{m-j}(k-\gamma)_{j+1}}{(j+n+k)!} \\
&= \frac{(n-1)!\, m\, (n+m+k-1)!}{(n+m-1)!\, (k-1)!\, (n+\gamma)_m} \cdot \sum_{j=0}^{m-1} \binom{m-1}{j} \frac{\Gamma(k+j+1)}{\Gamma(k+n+j+1)} \\
&\quad \times \frac{\Gamma(k-\gamma+j+1)}{\Gamma(k-\gamma)} (\gamma+n-k)_{m-j-1}.
\end{aligned}$$

It is important to note this equivalence:

$$\frac{\Gamma(k-\gamma+j+1)}{\Gamma(k-\gamma)} = \frac{\Gamma(k-\gamma+j+1)}{\Gamma(k-\gamma+1)} \frac{\Gamma(k-\gamma+1)}{\Gamma(k-\gamma)} = (k-\gamma+1)_j \cdot (k-\gamma),$$

due to use it in the equation above:

$$\begin{aligned}
E[K_m^{(n)} \mid X_1, \dots, X_n] = &\frac{(n-1)!\, m\, (n+m+k-1)!\, (k-\gamma)}{(n+m-1)!\, (k-1)!\, (n+\gamma)_m} \sum_{j=0}^{m-1} \binom{m-1}{j} \\
&\times \frac{\Gamma(k+j+1)}{\Gamma(k+n+j+1)} (k-\gamma+1)_j\, (\gamma+n-k)_{m-j-1}.
\end{aligned} \tag{2.16}$$

Focusing on the summation, it should be noted that the binomial term can be written as:

$$\begin{aligned}
\binom{m-1}{j} &= \frac{(m-1)!}{j!\, (m-1-j)!} = \frac{(m-1)\cdots(m-j-1+1)}{j!} \\
&= \frac{(-1)^j(-(m-1))\cdots(-(m-j))}{j!} \\
&= \frac{(-1)^j\, (-(m-1))_j}{j!}.
\end{aligned}$$

Recalling that $\Gamma(a+n) = (a)_n \cdot \Gamma(a)$, the summation in Equation (2.16) becomes

$$\sum_{j=0}^{m-1} \frac{(-1)^j(-(m-1))_j}{j!} \frac{(k+j)_j\,\Gamma(k+1)}{(k+n+1)_j\,\Gamma(k+n+1)}(k-\gamma+1)_j\,(\gamma+n-k)_{m-j-1}$$
$$= \frac{\Gamma(k+1)}{\Gamma(k+n+1)} \sum_{j=0}^{m-1} \frac{(-1)^j(-(m-1))_j}{j!}(k+1)_j(k-\gamma+1)_j(\gamma+n-k)_{m-j-1} \coloneqq R(\gamma).$$

Let us analyze the last term in $R(\gamma)$, specifically $(\gamma+n-k)_{m-j-1}$:

$$(\gamma+n-k)_{m-j-1} = \frac{\Gamma(\gamma+n-k+m-1-j)}{\Gamma(\gamma+n-k)}.$$

Let us set $\gamma+n-k=b$, so the expression simplifies to:

$$\frac{\Gamma(b+m-1-j)}{\Gamma(b)} = \frac{(b+m-1-j-1)\cdots b\cdot\Gamma(b)}{\Gamma(b)} = (b+m-1-j-1)\cdots b.$$

To further simplify, we multiply and divide this expression by the same factor and we obtain the following expression:

$$\begin{aligned}
&\frac{(b+m)(b+m-1)\cdots(b+m-1-j)}{(b+m)(b+m-1)\cdots(b+m-1-j)} \cdot \frac{\Gamma(b+m-1-j)}{\Gamma(b)} \\
=&\frac{\Gamma(b+m+1)}{(b+m)(b+m-1)\cdots(b+m-(j+1))\cdot\Gamma(b)} \\
=&\frac{\Gamma(b+m+1)}{\Gamma(b)} \cdot \frac{(-1)^j}{(-(b+m))(-(b+m)+1)\cdots(-(b+m-2)+j-1)} \\
=&\frac{\Gamma(b+m+1)}{\Gamma(b)} \cdot \frac{(-1)^j}{(b+m))(b+m-1)} \cdot \frac{1}{(-(b+m-2))_j} \\
=&\frac{\Gamma(b+m-1)\cdot(-1)^j}{\Gamma(b)\cdot(-(\gamma+n-k+m-2))_j}.
\end{aligned}$$

Substituting this into $R(\gamma)$ and using the definition of the generalized hypergeometric function ${}_3F_2$, we obtain:

$$R(\gamma) = \frac{\Gamma(k+1)}{\gamma(k+n-1)} \sum_{j=0}^{m-1} \frac{1}{j!} \cdot \frac{(-(m-1))_j\cdot(k+1)_j\cdot(k-\gamma+1)_j\cdot\Gamma(b+m-1)}{\Gamma(b)\cdot(-(\gamma+n-k+m-2))_j}.$$

This simplifies to:

$$R(\gamma) = \frac{\Gamma(k+1)}{\gamma(k+n-1)} \cdot \frac{\Gamma(b+m-1)}{\Gamma(b)} \cdot \sum_{j=0}^{m-1} \frac{1}{j!} \cdot \frac{(-(m-1))_j\cdot(k+1)_j\cdot(k-\gamma+1)_j}{(-(\gamma+n-k+m-2))_j}.$$

Recognizing the sum as a generalized hypergeometric function, we get:

$$R(\gamma) = \frac{\Gamma(k+1)}{\gamma(k+n-1)} \cdot \frac{\Gamma(b+m-1)}{\Gamma(b)} \times {}_3F_2\left(-(m-1), k+1, k-\gamma+1; k+n+1, -(\gamma+n-k+m-2); 1\right).$$

Finally, substituting this expression for $R(\gamma)$ into the expected value, we obtain:

$$E[K_m^{(n)} \mid X_1, \ldots, X_n] = \frac{(n-1)! \cdot m \cdot (n+m+k-1)!}{(n+m-1)! \cdot (k-1)! \cdot (n+\gamma)_m} \times \frac{\Gamma(k+1)}{\gamma(k+n-1)} \cdot \frac{\Gamma(b+m-1)}{\Gamma(b)} \times {}_3F_2\left(-(m-1), k+1, k-\gamma+1; k+n+1, -(\gamma+n-k+m-2); 1\right).$$

Recalling that b equals $\gamma + n - k$ and using the definition of the Pochhammer symbol, we get

$$E[K_m^{(n)} \mid X_1, \ldots, X_n] = m\,(k-\gamma)\,(k+1)\frac{\Gamma(n)\,(n+m)_k\,(\gamma+n-k)_{m-1}}{(n+m)_m} \times {}_3F_2(-(m-1), k+1, k-\gamma+1; k+n+1, -(\gamma+n-k+m-2); 1).$$

Therefore, the theorem is proved. □

Chapter 3

Applications of the Gnedin model to species sampling

The definition of new quantities reported in the previous chapter has opened up new possible analyses and considerations regarding the application of this model. In this chapter, we will present and analyze the results obtained by considering both simulated data and commonly used datasets that are well-suited for evaluating the performance of different processes.

3.1 Simulation studies

In this first section, we report the results obtained by considering simulated data that follow both the theoretical distribution, i.e. when the weights π_h of the random probability measure $\tilde{p}$ follow a Dirichlet distribution, and when this distribution is not respected. We want to assess the goodness of the model both in ideal and non-ideal situations. For each simulated dataset, we will evaluate the empirical a priori expected value curve and the rarefaction curve. We will then evaluate the predictive performance of the Gnedin model by studying the extrapolation curve and the asymptote to which it converges.

3.1.1 Data generation

The first dataset is generated through simulation based on the Gnedin model's definition. Fixed the number of distinct components in the entire population H, we sample the weights

$$(\pi_1, \ldots, \pi_H) \overset{\text{iid}}{\sim} \text{Dir}_H(1, \ldots, 1).$$

Fixed a sample size n, we sample with replacement n values from the integer vector containing values from 1 to H, i.e., the vector of atoms Z_h as $h = 1, \dots, H$, to obtain a sample $X_1, \dots, X_n$. This construction can indeed be extended beyond the theoretical model. It is possible to consider generic weights, which are not necessarily distributed according to a Dirichlet distribution. For example, the weights could be sampled from a Weibull distribution with a shape parameter equal to 2 and a scale parameter equal to 1, or one could randomly select H integer values and normalize them by their sum to obtain the weights. In this way, it is possible to evaluate the performance of the Gnedin model under conditions that deviate from the theoretically required ones.

3.1.2 Prior quantities of interest

First, we focus on analysing the model in its a priori component by estimating the rarefaction curve and empirically approximating the prior expected value of the number of distinct species K_n. Remember that in the previous chapter, in Theorem 2.3 we defined the a priori expected value of the number of distinct species. The empirical estimation of this quantity consists of obtaining an estimate of the parameter γ that characterises the Gnedin model and substituting it into the formula. The first attempt is to estimate the γ parameter using a maximum likelihood estimator. Recalling the EPPF formula for the Gnedin model in Equation (2.2) and the associated weights $V_{n,k}$ in Equation (2.1), the log-likelihood function results as:

$$\begin{aligned}\log(\Pi_n(n_1, \dots, n_k)) = &\log(k-1)! + \log(1-\gamma)_{k-1} + \log(\gamma)_{n-k} - \\ &- \log(n-1)! - \log(1+\gamma)_{n-1} + \sum_{j=1}^{k} \log n_j!.\end{aligned}$$

Using a numerical optimisation algorithm, we obtain the maximum of this function, $\hat{\gamma}_{ML}$.

We perform a sort of model checking to assess how well the expected value aligns with the observed data. To do this, we estimate the rarefaction curve. The estimation of this curve is obtained by making several permutations of the available sample $X_1, \dots, X_n$. For each sample size from 1 to n we estimate the number of distinct species within each permutation. Then we average the values obtained by keeping the sample size fixed. This procedure can be easily implemented by using a function within the *Vegan* package (RDocumentation

(2024)), known as *rarefy*. This function takes as input the community data, a matrix-like object or a vector and the sample size. The results obtained are the same as we see in Figure 3.1.

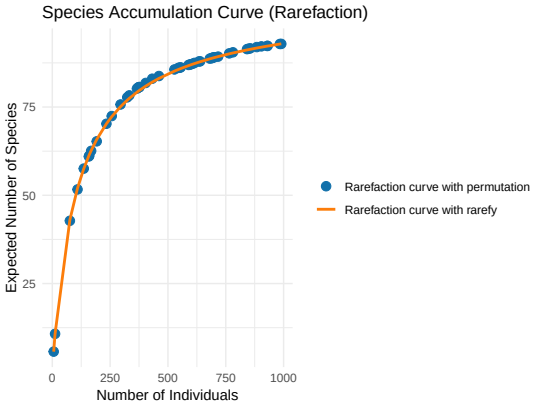


Figure 3.1: Comparison of estimation methods for the rarefaction curve considering simulated data with $(\pi_1, \ldots, \pi_H) \sim \mathrm{Dir}_H(1, \ldots, 1)$.

However, the comparison between the rarefaction curve and the empirical expected value of the number of distinct species does not provide good results. Indeed there is no coincidence in the final points of the two curves. The results are reported in Figure 3.2.

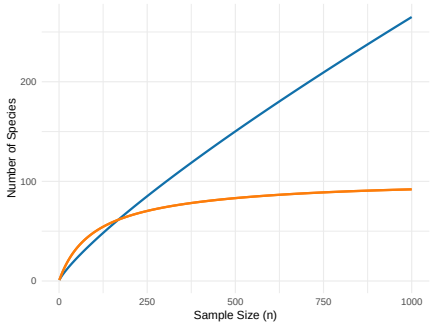


Figure 3.2: Simulated data with $(\pi_1, \ldots, \pi_H) \sim \mathrm{Dir}_H(1, \ldots, 1)$. Comparison between the rarefaction curve (orange) and the empirical expected value with $\hat{\gamma}_{ML}$ (blue).

The second estimation procedure for the parameter γ is the method of moments. We equal the a priori expected value $E[K_n]$ of the number of distinct species to the value of distinct species k in the observed sample and derive γ by solving the equation:

$$\hat{\gamma}_{MOM} \quad \text{w.r.t.} \quad E[K_n] = k.$$

Figure 3.3 shows how the curves change. In this second case, the results are improved: the final points of the rarefaction curve and that of the empirical expected value coincide.

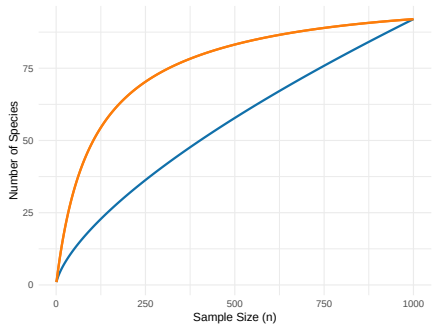


Figure 3.3: Simulated data with $(\pi_1, \ldots, \pi_H) \sim \mathrm{Dir}_H(1, \ldots, 1)$. Comparison between the rarefaction curve (orange) and the empirical expected value with $\hat{\gamma}_{MOM}$ (blue).

A mismatch between the rarefaction curve and the empirical curve of the a priori expected value of the number of distinct species K_n leads us to conclude that the model does not fit the data well a priori. However, the prior distribution for the number of distinct species is a heavy-tailed distribution, which makes it possible to sample huge values that make the expected value explosive. This study is focused on evaluating the predictive abilities of the model, and thus we will continue with our analysis.

3.1.3 Posterior inference and prediction

This thesis aims to evaluate the predictive performance of the Gnedin model. Therefore, we focus now on estimating the extrapolation curve using a Monte Carlo procedure. Before describing the implemented Monte Carlo algorithm, it is necessary to recall the probability of sampling a new value P_{new} in the predictive distribution of the Gnedin model in Equation (2.7). Given a sample $X_1, \ldots, X_n$ with k distinct values within it and fixing the size m of the additional sample, we proceed as follows. For a specified number of simulations, sample m times from a Bernoulli distribution with parameter P_{new}, where the sample size increases from $n+1$ to $n+m$. The value of k is updated at each iteration. A generated value of 1 indicates the presence of a new species in the additional sample, instead, a value of 0 indicates that the species found is already known and present in the observed sample. The m values generated from the Bernoulli distribution are iteratively summed to k and then these sums are averaged over all the final

values obtained from the different simulations for each sample size from $n+1$ to $n+m$.

Algorithm 1: Monte Carlo algorithm for extrapolation curve

Load: n, m, k, N_{sim}
Initialize: Matrix $K_m^{(n)}$ with dimensions $N_{sim} \times m$ with NA values
Set $K_m^{(n)}[\,,1] \leftarrow k$;
for $j \leftarrow 1$ *to* N_{sim} **do**
  for $i \leftarrow (n+1)$ *to* $(n+m)$ **do**
    Sample r from $\mathrm{Bernoulli}(P_{new})$;
    $K_m^{(n)}[j, i-(n-1)] \leftarrow K_m^{(n)}[j, i-n] + r$;
Result: A vector containing means of columns of $K_m^{(n)}$.

It is important to note that the extrapolation curve conditional on the data is not a direct continuation of the rarefaction curve, instead, the slope changes between the two curves. This change is crucial because it ensures the possibility of convergence. However, the mismatch between the a priori curves and the initial path of the rarefaction curve does not preclude the model from having good forecasting abilities.

Since the true value for H is unknown, we want to determine whether the value to which the extrapolation curve converges is a good estimate for the overall species variety. We use the limit of the a posteriori expected value of H as an estimate. This quantity is essential because it can be interpreted as a point estimate of richness, a fundamental measure of biodiversity which evaluates the number of distinct species. However, there persists the problem with the estimation: we have decided to adopt the following procedure that simplifies the code. As is already known, in Gnedin (2010) the posterior distribution for H in Equation (2.5) can also be written as:

$$\mathbb{P}(H = h \mid X_1, \ldots, X_n) = \frac{\Gamma(n)\,\Gamma(h)\,\Gamma(h-\gamma)\,\Gamma(n+\gamma)}{\Gamma(n+h)\,\Gamma(h-k+1)\,\Gamma(k-\gamma)\,\Gamma(n+\gamma-k)}. \tag{3.1}$$

From this result, the associated expected value can be found by evaluating the expression in Equation (2.6). The challenge lies in calculating this quantity because it involves an infinite sum. Several approaches can be taken. One approach is to truncate the posterior expected value, which involves evaluating the value for a fixed and arbitrarily large h. This means that for larger sample

sizes, the probability of sampling a new value, P_{new}, is nearly zero. Another procedure may be to evaluate the posterior expected value not of H, but of the number of distinct species $K_m^{(n)}$ as the sample size m of the additional sample diverges. This strategy requires calculating the hypergeometric function as determined in Theorem 2.4, but that is more difficult to implement. Therefore, we opt for the first method. For each dataset, we fix a sufficiently large value of h to ensure convergence, and once established, we calculate the expected value with the Equation (2.6) using the formulation for the posterior distribution of H as in Equation (3.1) and then the asymptote is identified. This allows us to understand better the predictive power of the Gnedin model. The excellent predictive ability of such a model can be best appreciated with simulated theoretical data, where the true value of the number of distinct species in the entire population H is known.

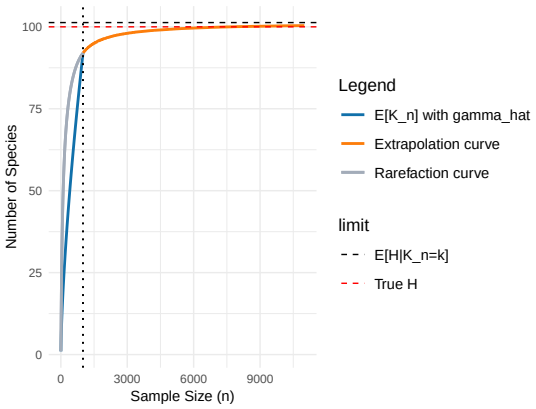
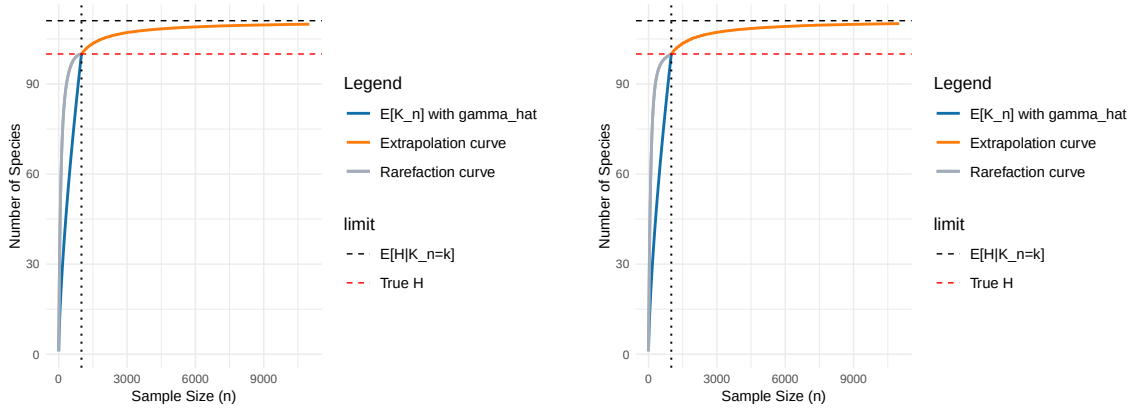


Figure 3.4: Extrapolation curve for simulated data with $(\pi_1, \ldots, \pi_H) \sim \mathrm{Dir}_H(1, \ldots, 1)$.

The value to which the extrapolation curve converges aligns with the estimated posterior expected value of H and differs only slightly from the true value of H.

Despite the excellent results for the future, some challenges remain. For example, when we simulate data by sampling weights π_h from a Weibull distribution with arbitrarily chosen parameters (shape equal to 2, and rate equal to 1) using a random sampling function in R, or by generating π_h by normalizing arbitrarily chosen integer values, we observe weaknesses in the model.



(a) Extrapolation curve for simulated data with weights as normalized integers.

(b) Extrapolation curve for simulated data with $\pi_h \sim Weibull(2, 1)$ for $h = 1, \dots, H$.

Figure 3.5: Extrapolation curves.

When the weights π_h are specified manually or are derived from a distribution different from a Dirichlet, there is a sudden change in the slope from the rarefaction curve to the extrapolation curve. This is a sign that the model is likely not correctly specified as it is. In the future, improving the model to perform well even when the weights are not the theoretical ones would be interesting. For example, one possible solution could be introducing a new parameter within the model.

Finally, we want to focus on the posterior distribution of H given to us in Gnedin (2010). This is a very useful measure of richness, and our Bayesian approach allows us also to assess the associated uncertainty. The estimation of the distribution is done by evaluating the function on a finite grid of values. Such values are chosen so that the sum of the probabilities obtained is 1. Again we consider the posterior distribution for H formulated as in Equation (3.1).

Since it has never been visualized, we want to understand the shape of this distribution for example to determine if it is symmetric. Without visualising it, we would not be able to tell whether the distribution is strongly skewed or not. We expect to have validation of the estimated posterior expected value of H by obtaining a distribution roughly centred around this value or if the distribution is skewed with the highest probability assigned to a value close to that of the estimated expected value of the number of distinct species H. The results are reported in Figure 3.6 and 3.7.

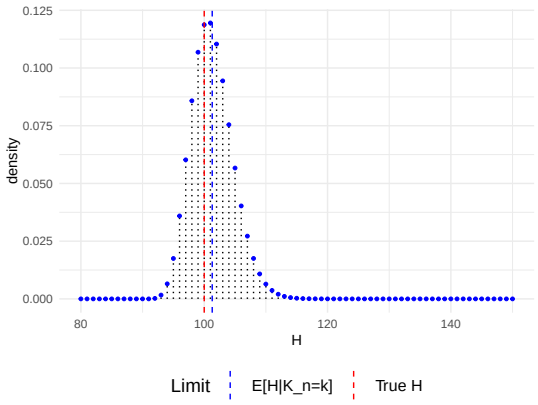


Figure 3.6: Estimation of the posterior distribution of H when $(\pi_1, \ldots, \pi_H) \sim \mathrm{Dir}_H(1, \ldots, 1)$. The true value of H is in red, and in blue is the estimated expected value of H.

In this scenario, the distribution is symmetric and centred around the estimated expected value of H, indicating a good match between the model and the data.

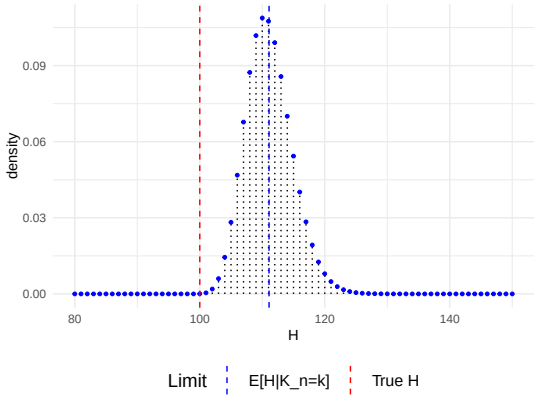


Figure 3.7: Estimation of the posterior distribution of H when π_h are normalized integers. The true value of H is in red and in blue is the estimated expected value of H.

This figure highlights the poor performance of the model when the weights π_h are not derived from the theoretical distribution. The results are quite similar, whether the weights are sampled from a Weibull distribution or obtained by normalizing random integers. Thus, only the distribution for normalized integers is presented here as an example.

3.2 Real data applications

In this section, we report the results obtained by applying what was previously described for the simulated data to data belonging to two distinct datasets: the Dune and Butterfly datasets.

3.2.1 Dune dataset

The Dune dataset is a community dataset with variables representing different species and data showing the abundance of each species in each of the selected sites. The dataset comes from a 1982 research project conducted on the Dutch island of Terschelling. The research aimed to study the connection between vegetation and management in dune meadows. The data was collected using an ordinal scale and out of a total of 80 sites, 20 were randomly chosen for analysis to account for overall variability. The dataset includes data on 30 recorded species and is part of the *Vegan* package of R. The species names are abbreviated to eight characters (4+4), such as *Agrostol* for *Agrostis stolonifera*. For our research, the subdivision into sites for species abundance is unnecessary. Therefore, we only account for the total number of times each species was observed, regardless of the site where the observation occurred. For more details, see RDocumentation (2024).

	Achimill	Agrostol	Airaprae	Alopgeni
1	16.00	48.00	5.00	36.00

Table 3.1: Here is an example of the first four species from the dataset.

3.2.2 Butterfly dataset

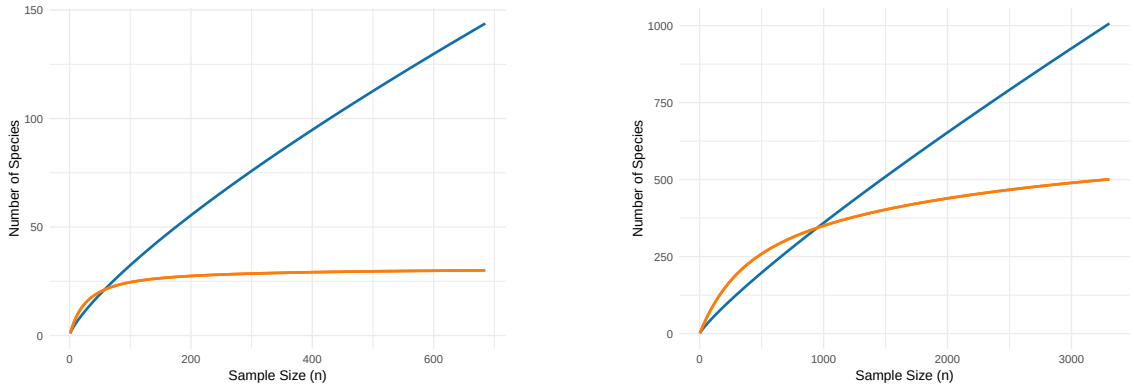
During World War II, the naturalist Alexander Corbet spent two years trapping butterflies in Malaysia. He found 118 rare species, that he caught only one of each. The dataset consists in Table 3.2, which presents the number y of species observed at each trapping frequency x over the two years. 74 species were trapped twice, and 44 species three times and so on.

x	1	2	3	4	5	6	7	8	9	10	11	12
y	118	74	44	24	29	22	20	19	20	15	12	14
x	13	14	15	16	17	18	19	20	21	22	23	24
y	6	12	6	9	9	6	10	10	11	5	3	3

Table 3.2: Butterfly dataset.

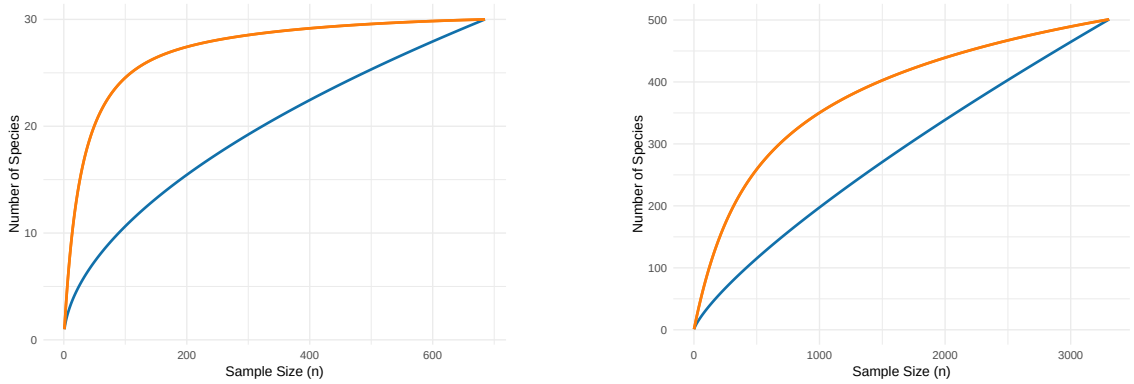
3.2.3 Analysis and Results from the Dataset

First of all, we focused on the a priori results. Again, in order to have a good comparison between the empirical curve of the a priori expected value and the rarefaction curve, we had to use a MOM estimator for the parameter γ as we can see from the Figures 3.8 and 3.9.



(a) Dune dataset. **(b)** Butterfly dataset.

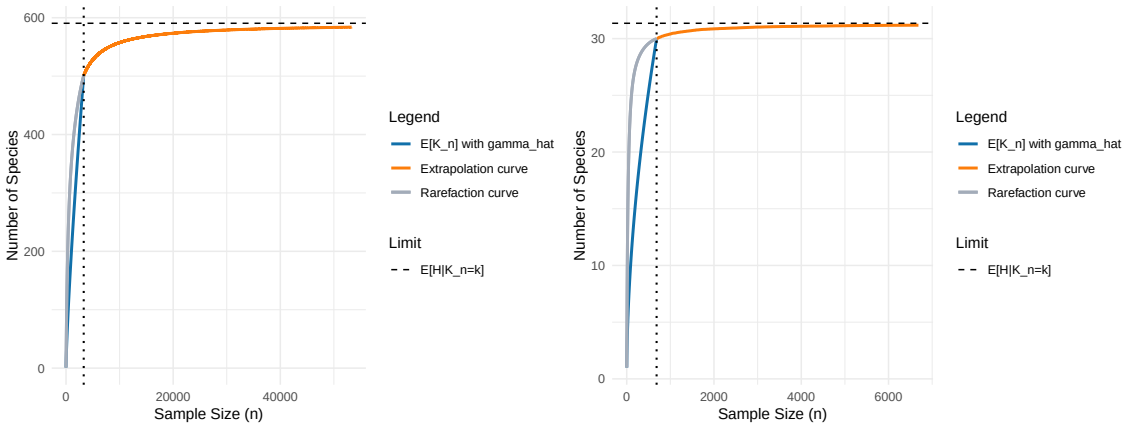
Figure 3.8: Comparison between the rarefaction curve (orange) and the empirical expected value with $\hat{\gamma}_{\mathrm{ML}}$ (blue).



(a) Dune dataset. **(b)** Butterfly dataset.

Figure 3.9: Comparison between the rarefaction curve (orange) and the empirical expected value with $\hat{\gamma}_{\mathrm{MOM}}$ (blue).

Then we focus on the posterior inference. First, we calculated the corresponding extrapolation curves and verified their consistency with the previously obtained results. Once again, the curves converged to the estimate of the posterior expected value for H. However, since the data are not simulated in this case, we cannot compare them with the true value of the number of distinct species in the entire population H.



(a) Dune dataset. **(b)** Butterfly dataset.

Figure 3.10: Extrapolation curves.

Finally, we estimated the posterior distribution of H.

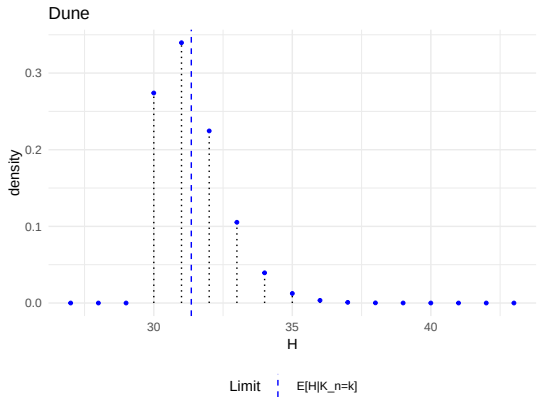


Figure 3.11: Estimation of the posterior distribution of H for Dune dataset.

The distribution for the Dune dataset shows less symmetry compared to the other scenarios. This is likely due to the limited support for H, with only a few values present. Consequently, the symmetry is less pronounced and the distribution is not as well-centered.

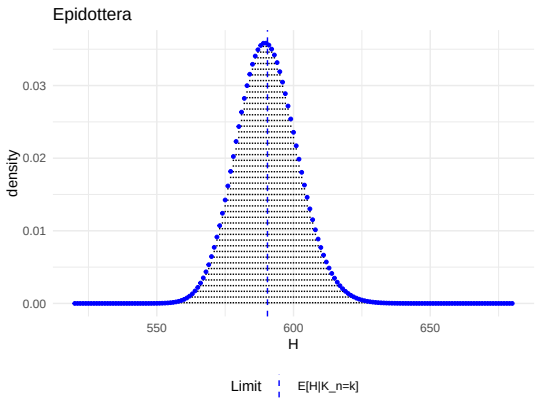


Figure 3.12: Estimation of the posterior distribution of H for Butterfly dataset.

For the Butterfly dataset, the distribution is highly symmetric, exceeding what is observed in other cases. This is due to the large number of values required to cover the entire probability, ensuring a well-distributed and balanced posterior.

Appendix A

Hypergeometric functions

Definition A.1. The generalized hypergeometric function ${}_pF_q$ is defined as a hypergeometric series:

$$ {}_pF_q(a_1, a_2, \ldots, a_p; b_1, b_2, \ldots, b_q; z) = \sum_{n=0}^{\infty} \frac{(a_1)_n (a_2)_n \cdots (a_p)_n}{(b_1)_n (b_2)_n \cdots (b_q)_n} \frac{z^n}{n!} \tag{A.1} $$

where: $(a_i)_n$ for $i = 1, \ldots, p$ and $(b_j)_n$ for $j = 1, \ldots, q$ are the Pochhammer symbols, representing the rising factorials:

$$ (a_i)_n = a_i(a_i + 1)(a_i + 2) \cdots (a_i + n - 1) \quad \text{with} \quad (a_i)_0 = 1. $$

When $p = 2$ and $q = 1$, the generalized hypergeometric function reduces to the ordinary hypergeometric function ${}_2F_1(a, b; c; z)$, also called *Gauss's hypergeometric function*.

Definition A.2. The hypergeometric function is denoted as ${}_2F_1(a, b; c; z)$ and is defined by the series:

$$ {}_2F_1(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!} \tag{A.2} $$

where $(q)_n$ is the Pochhammer symbol, representing the rising factorial:

$$ (q)_n = q(q + 1)(q + 2) \cdots (q + n - 1) \quad \text{with} \quad (q)_0 = 1. $$

A special hypergeometric identity includes Gauss's hypergeometric theorem. Indeed, when $z = 1$, the sum that characterizes the hypergeometric function is

finite, and it reduces as follows:

$$ {}_2F_1(a, b; c; 1) = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} \quad \text{if} \quad c > a + b. \tag{A.3} $$

Appendix B

Janossy density

The distribution of point processes whose realizations are almost surely finite can be described as follows. A finite point process X consists of a finite collection of points $X_1, \ldots, X_n$ within $\mathbb{R}^n$, where both the number of points n and the locations $X_1, \ldots, X_n$ are random variables. This can be represented as:

$$X = \{X_1, \ldots, X_n\}.$$

This means that a finite point process is defined by a discrete probability distribution p_n over $\mathbb{N}$, which describes the distribution of the number of points n in X, along with a family of joint probability density functions (PDFs) of the form:

$$\mathcal{S} = \{\pi_n(x_1, \ldots, x_n)\}_{n \in \mathbb{N}^+}$$

which describes how the points in X are distributed in $\mathbb{R}^n$. Since X is a set, it remains unchanged under any permutation of the points $X_1, \ldots, X_n$. This symmetry implies that the joint PDFs in $\mathcal{S}$ must be permutation invariant, meaning:

$$\pi_n(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) = \pi_n(x_1, \ldots, x_n)$$

for any permutation σ of the indices.

Definition B.1. The Janossy density of order n is defined as:

$$j_n(\{x_1, \ldots, x_n\}) \coloneqq \begin{cases} n!\, \pi_n(x_1, \ldots, x_n)\, p_n, & \text{if } n > 0 \\ p_0, & \text{if } n = 0. \end{cases} \tag{B.1}$$

It provides a way to describe the probability distribution of finding a specific number of points at particular locations in a given space.

In an infinitesimal sense, $j_n(x_1, \ldots, x_n)\, dx_1 \ldots dx_n$ represents the probability of finding exactly n points, with one point located in each of the infinitesimal regions centered at $x_1, \ldots, x_n$. For $n = 0$, Equation (B.1) is typically interpreted as $j_0(\emptyset) = p_0$.

Bibliography

Argiento, R. & De Iorio, M. (2022). Is infinity that far? A Bayesian nonparametric perspective of finite mixture. *Ann. Statist.* **50(5)**, 2641–2663.

Charalambides, C. A. (2005). *Combinatorial methods in discrete distributions*. John Wiley & Sons.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R., Prünster, I. & Ruggiero, M. (2015). Are Gibbs-type priors the natural generalization of the Dirichlet process? *IEEE Transactions Pattern Analysis and Machine Intelligence* **37(2)**, 212–229.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré* **7**, 1–68.

Engen, S. (1978). *Stochastic abundance models*. Monographs on Statistics and Applied Probability. Springer Dordrecht.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist* **1(2)**, 209–230.

Gnedin, A. (2010). A species sampling model with finitely many types. *Elect. Comm in Probab.* **15**, 79–88.

Gnedin, A. & Pitman, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences* **138**, 5674–5685.

Griffiths, R. C. (1979). Exact sampling distributions from the infinite neutral alleles mode. *Advances in Applied Probability* **11(2)**, 326–354.

Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96(453)**, 161–173.

Lijoi, A., Mena, R. H. & Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association* **100(472)**, 1278–1291.

Lijoi, A., Mena, R. H. & Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **69**, 715–740.

McCloskey, J. W. T. (1965). *A model for the distribution of individuals by species in an environment*. Ph.D. thesis, Michigan State University.

Moya, B. & Walker, S. G. (2024). Full uncertainty analysis for Bayesian nonparametric mixture. *Computational statistics and data analysis* **189**, 107838.

Müller, P., Quintana, F. A., Jara, A. & Hanson, T. (2015). *Bayesian nonparametric data analysis*, vol. 1. Springer.

Orbanz, P. (2014). Lecture notes on Bayesian Nonparametrics. *Journal of Mathematical Psychology* .

Perman, M. (1990). *Random Discrete Distributions Derived from Subordinators*. Ph.D. thesis, University of California, Berkeley.

Pitman, J. (1996). Some developments of the Blackwell-Macqueen urn scheme. *Lecture Notes-Monograph Series* **30**, 245–267.

Pitman, J. (2003). Poisson-Kingman partitions. *Lecture Notes-Monograph Series* **40**, 1–34.

Pitman, J. & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25(2)**, 855–900.

RDocumentation (2024). RDocumentation. Package Vegan. https://https://cran.r-project.org/web/packages/vegan/vegan.pdf.

Regazzini, E., Lijoi, A. & Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *Ann. Statist* **31(2)**, 560–585.

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

van der Vaart, A. & Ghosal, S. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

Zito, A., Rigon, T., Ovaskainen, O. & Dunson, D. B. (2023). Bayesian modelling of sequential discoveries. *Journal of the American Statistical Association* **188(544)**, 2521–2532.