PhD THESIS DECLARATION

The undersigned

SURNAME Sporysheva

FIRST NAME Polina
PhD Registration Number 1465731

Thesis title: Extensions of Species Sampling Models

PhD in Statistics

Cycle 25

Candidate's tutor Sonia Petrone
Year of thesis defence 2014

DECLARES

Under her responsibility:

1) that, according to Italian Republic Presidential Decree no. 445, 28th December 2000, mendacious declarations, falsifying records and the use of false records are punishable under the Italian penal code and related special laws. Should any of the above prove true, all benefits included in this declaration and those of the temporary embargo are automatically forfeited from the beginning;

2) that the University has the obligation, according to art. 6, par. 11, Ministerial Decree no. 224, 30th April 1999, to keep a copy of the thesis on deposit at the "Bilioteche Nazionali Centrali" (Italian National Libraries) in Rome and Florence, where consultation will be permitted, unless there is a temporary embargo protecting the rights of external bodies and the industrial/commercial exploitation of the thesis;

3) that the Bocconi Library will file the thesis in its "Archivio istituzionale ad accesso aperto" (institutional registry) which permits online consultation of the complete text (except in cases of a temporary embargo);

4) that, in order to file the thesis at the Bocconi Library, the University requires that the thesis be submitted online by the candidate in unalterable format to Società NORMADEC (acting on behalf of the University), and that NORMADEC will indicate in each footnote the following information:

   - thesis Extensions of Species Sampling Models;
   - by Sporysheva Polina;
   - defended at Università Commerciale "Luigi Bocconi" – Milano in 2014;
   - the thesis is protected by the regulations governing copyright (Italian law no. 633, 22th April 1941 and subsequent modifications). The exception is the right of Università Commerciale "Luigi Bocconi" to reproduce the same for research and teaching purposes, quoting the source;

5) that the copy of the thesis submitted online to NORMADEC is identical to the copies handed in/sent to the members of the Thesis Board and to any other paper or digital copy deposited at the University offices, and, as a consequence, the University is absolved from any responsibility regarding errors, inaccuracy or omissions in the contents

of the thesis;

6) that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (Italian law, no. 633, 22nd April 1941 and subsequent integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil, administrative or penal, and shall be exempt from any requests or claims from third parties;

7) that the PhD thesis is not the result of work included in the regulations governing industrial property, was not produced as part of projects financed by public or private bodies with restrictions on the diffusion of the results, and is not subject to patent or protection registrations, and therefore not subject to an embargo;

Date 10/11/2013

SURNAME Sporysheva
FIRST NAME Polina

Contents

Acknowledgements

First of all, I would like to thank my supervisor Sonia Petrone for her patience, enthusiasm, enormous knowledge and continuous support, both professional and personal. Her guidance helped me during the research and writing of this thesis.

I also would like to offer my special thanks to my co-advisor Sandra Fortini for her enthusiastic encouragement, valuable suggestions and for generously providing guidance on the technical aspects of this thesis.

My sincere thanks also to Pietro Rigo. Even a brief meeting with him was fruitful and affirmative.

I am very glad that I spent these years at Bocconi University. In my daily work I have been cordially supported by a friendly and cheerful group of fellow students. I am grateful to all of them for the stimulating discussions and for all the fun we have had together in the last four years.

Finally, I thank my wonderful family and Max for their love and unwavering belief in me. Without them, I would not be the person I am today.

Abstract

The thesis studies extensions of species sampling models with focus on implications in Bayesian nonparametrics. The interest in Bayesian nonparametrics has increased considerably in the last decades and arises in a variety of frameworks. In particular, the need of defining probability laws for vectors of random measures, appears in such problems as multivariate density estimation, nonparametric regression, Bayesian nonparametric inference for time series, two-sample problems and others. Most of the current proposals are based on specifying dependence at the level of the random measures, while the corresponding bivariate prediction rules become very complicated and intractable. On the other hand (in a univariate case) a powerful way to construct a nonparametric prior is the predictive approach, based on a sequence of predictive distributions. Species sampling models are a successful example of such construction. Thus, the idea of this thesis is to explore how one can construct a bivariate nonparametric prior starting from an appropriate and fairly tractable bivariate prediction rule, rather than specifying the dependence between random measures. We define the concept of bivariate random partitions, based on which we propose a class of bivariate species sampling models, characterized by a bivariate prediction rule. We specify the conditions on this prediction rule that guarantee partial exchangeability. We show that partial exchangeability and simplicity of the bivariate prediction rule strongly limit the class of bivariate processes. At the same time, in some applications the assumption of stationarity implied by partial exchangeability can be too restrictive. Motivated by this and based on the notion of partial conditional identity in distribution, we propose a generalization of bivariate species sampling models, that may have a tractable prediction rule. Although the proposed models are not partially exchangeable, they preserve most properties associated with partial exchangeability. Examples of such constructions will be discussed.

Chapter 1

Introduction and species sampling models

1.1 Motivation and outline of the thesis

Bayesian nonparametrics is a relatively young field where prior probability laws are defined on an infinite-dimensional parameter space. Species sampling problems that arise in ecology, biology and population genetics, are connected to Bayesian nonparametrics, since a population of species can be modeled as a discrete random measure. Many models have been proposed, starting from the model for species abundances due to Fisher et al. [1943] and followed by a class of species sampling models introduced by Pitman [1996a]. The latter, is a general class of discrete random probability measures in which random weights correspond to unknown proportions and are independent of labels. An example of such models is the Dirichlet Process, a cornerstone in Bayesian nonparametrics.

The species sampling theory is an interesting topic of research in a diversity of contexts. It involves combinatorics, Kingman's theory of partition structure, concepts of exchangeable random partitions and the predictive approach. The aim of this thesis is to study extensions of species sampling models in several directions.

The first direction is the generalization of the theory to the multivariate case. In real-life applications, it is possible to have a non-homogeneous species population. For example, in biology or fishery, samples might come from different parts of a lake, with different physical and chemical water properties. For such problems, the notion of exchangeability, that involves an assumption about complete symmetry among all observations, is too restrictive and partial exchangeability should be used. The interest is to define model for sampling species from two (or more) dependent populations. We focus on the bivariate case, but multivariate extension can be envisaged. Therefore, the aim is to construct a bivariate species theory that can model dependent random measures. In Bayesian nonparametrics, a similar interest arises in a variety of frameworks such as, e.g., multivariate density estimation, nonparametric regression, inference on time series data and two-sample problems. Although a lot of such models have been proposed, almost all of them have a very complicated prediction rule. Thus, the first objective of the thesis is to use a predictive approach for constructing a bivariate prior. It is not an easy task but we formalize the framework and give a contribution towards how one can generate a partially exchangeable sequence from a tractable bivariate prediction rule.

Another direction consists of weakening the assumption of stationarity in species sampling models. The assumption of stationarity is implied by exchangeability or partial exchangeability, and can be too restrictive in some applications. For example, if species are sampled from a population subject to some perturbation that destroys the stationarity before returning to some equilibrium. A notion of conditional identity in distribution can be considered, roughly speaking, as exchangeability without stationarity. The concept was hinted by Kallenberg [1988], then developed and studied by Berti et al. [2004]. In particular, a conditionally identically distributed sequence is asymptotically exchangeable and, as for any exchangeable sequence, its empirical and predictive measures converge to the same limit. Thus, the second objective of the thesis is to investigate an extension of partial exchangeability, that asymptotically, is still partially exchangeable. To this purpose, we develop the notion of partial conditional identity in distribution, and explore the corresponding extension of bivariate species sampling models.

The thesis is organized in the following way. In Chapter 1 we give a review of the main theory of exchangeability, partial exchangeability and predictive constructions of exchangeable laws. We discuss the theory of exchangeable random partitions introduced by Kingman [1978b] and Aldous [1985], who provided the building blocks for the species sampling theory. We specify the most important feature of species sampling models namely, that they constitute a class of random discrete measures defined through the predictive approach, i.e., through the conditional distribution of the next observation given the distinct values observed in the past, together with their abundances.

In Chapter 2 we propose the theory of bivariate species sampling models as a bivariate extension of species sampling models. We start by defining the concept of partially exchangeable bivariate random partitions, and give an extension of the theory of random partitions. The resulting theory is crucial for specifying a class of bivariate species sampling models through the predictive approach. We specify the conditions on the bivariate prediction rule that guarantee partial exchangeability of the corresponding model, observing that it is not easy to have a tractable and manageable prediction rule. Indeed, if we let the probability that the next observation is of the i-th observed species be proportional to a function of the frequency of that species in the bivariate sample, then the model reduces to a Dirichlet Process for a homogeneous population of species.

In Chapter 3 of the thesis, we focus on bivariate species sampling models that are not partially exchangeable. We start from the notion of conditional identity in distribution (Berti et al. [2004]), which generalizes the notion of exchangeability, and we propose the concept of partially conditional identity in distribution. The predictive and empirical measures of a sequence of such type have the same limit and, under some additional conditions, the sequence is asymptotically partially exchangeable. This allows us to construct a generalization of bivariate species sampling models which may have a tractable prediction rule, while preserving most of the properties of the bivariate species sampling models, and which can be used in cases where the assumption of exchangeability fails. We show that, at least in principle, one can use a partially conditionally identically distributed sequence as an approximation of a partially exchangeable sequence, while the limits of its empirical measure can be used to construct a novel prior for a vector of dependent random measures.

In Chapter 4 we discuss examples of partially c.i.d. structure proposed in Chapter 3. The first example, binary partially c.i.d. sequences, generalizes well-known reinforced urn models and can be used for explaining and predicting physical or behavioral phenomena in different areas. The second example belongs to the generalized bivariate species sampling model. We analyze its properties and specify a particular model for unseen species problems where the main focus is to study the appearance of a new species. Such problems arise in ecology, biology, language modeling (vocabulary studies) and other areas. Based on a sample from some population of species, one has to solve different predictive issues concerning the composition of the population. The proposed structure has a quite simple prediction rule, it is very flexible, and can accommodate power-law tails. In order to illustrate the good performance of the model we present an application for text analysis, however the model can be straightforwardly applied to other problems of species discovery. In the last section of Chapter 4 we state some problems for further research.

1.2 Exchangeability and predictive approach

Exchangeability plays a fundamental role in Bayesian statistics. This type of dependence assumption is used in modeling, as widely as the assumption of independence and identity in distribution (i.i.d.) is used in classical statistics. Roughly speaking the assumption of exchangeability means that the order of the observations does not matter. More precisely, let us consider an infinite sequence $(X_1, X_2, ...)$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with each X_i taking values in a complete and separable (Polish) metric space $\mathbb{X}$ endowed with the Borel σ-algebra $\mathcal{X}$.

Definition 1.2.1. *The sequence* $(X_1, X_2, ...)$ *is infinitely exchangeable if,* *for any $n \geqslant 1$ and any permutation σ of the indices $1, ..., n$, the joint probability distribution of the random vector $(X_1, ..., X_n)$ is the same as the joint probability distribution of $(X_{\sigma(1)}, ..., X_{\sigma(n)})$.*

Note that any i.i.d. sequence is also exchangeable, but the converse is not true, since elements of an exchangeable sequence are identically distributed but not necessarily independent.

The most important benefit of exchangeability is the celebrated representation theorem formulated by de Finetti [1937]. Recall that if we consider the set $\mathcal{P}_{\mathbb{X}}$ of all probability measures on $(\mathbb{X}, \mathcal{X})$, then $\mathcal{P}_{\mathbb{X}}$ itself, endowed with the Borel σ-field $\mathcal{B}(\mathcal{P}_{\mathbb{X}})$ generated by the topology of weak convergence, is Polish (Parthasarathy [1967]). We call a *random probability measure* on $\mathbb{X}$, a measurable function defined on some probability space $(\Omega, \mathcal{F}, Q)$ with values in $(\mathcal{P}_{\mathbb{X}}, \mathcal{B}(\mathcal{P}_{\mathbb{X}}))$.

Theorem 1.2.2. *(de Finetti's representation theorem) The infinite sequence $(X_1, X_2, ...)$ is exchangeable if and only if there exists a probability measure μ on $(\mathcal{P}_{\mathbb{X}}, \mathcal{B}(\mathcal{P}_{\mathbb{X}}))$ such that for any $n \geqslant 1$,*

$$\mathbb{P}\left[X_1 \in A_1, ..., X_n \in A_n\right] = \int_{\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} F(A_i)\mu(dF),$$

for any $A_1, ..., A_n$ in $\mathcal{X}$. Moreover, the sequence of empirical distributions $\hat{F}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(\cdot)$ converges weakly $\mathbb{P}$-a.s. to a random distribution function $\tilde{F}$ that is distributed according to μ. The probability measure μ is unique and is called the de Finetti measure of the sequence $(X_1, X_2, ...)$.

The random measure $\tilde{F}$ in the representation theorem is referred by Aldous [1985] as the *directing random measure.* Due to the representation theorem, exchangeability plays a an essential role in Bayesian inference, as it provides the fundamental justification for the *hypothetical approach*, that is, an exchangeable sequence can be represented as a mixture of i.i.d. sequences of random variables:

$$\begin{aligned} X_i | \, \tilde{F} &\overset{iid}{\sim} \tilde{F} \text{ (the statistical model)}, \\ \tilde{F} &\sim \mu \text{ (the prior)}. \end{aligned}$$

We call such a sequence of random variables, $(X_1, X_2, ...)$, a sample from $\tilde{F}$. In this thesis we focus on a nonparametric framework where the prior probability law μ is typically defined on an infinite-dimensional parameter space.

A basic problem that arises in Bayesian nonparametrics is how to construct nonparametric priors with full support through a *predictive approach*, where the emphasis is on the probability law of the observed quantities $(X_1, X_2, ...)$ and one can, at least in principle, characterize the prior through the sequence of predictive distributions. Denote the regular conditional (or predictive) distribution of X_{n+1} given $X_1, ..., X_n$ by

$$F_n(\cdot) = \mathbb{P}(X_{n+1} \in \cdot | X_1, ..., X_n).$$

It can be shown that the sequence of predictive distributions $(F_1, F_2, ...)$ converges almost surely to the $\mathbb{P}$-a.s. limit of the sequence of empirical distributions $\tilde{F}$ (see Fortini et al. [2000] for a general proof). This result is based on de Finetti's work in the case of $0 - 1$ exchangeable random variables. A stronger result, a Glivenko-Cantelli theorem for exchangeable sequences, was given by Berti and Rigo [1997] for $\mathbb{X} = \mathbb{R}$. This theorem establishes that $sup_x|F_n(x) - \hat{F}_n(x)| \to 0$ $\mathbb{P}$-a.s., which implies that $F_n \to \tilde{F}$ weakly $\mathbb{P}$-a.s. See Fortini and Petrone [2012] for a detailed review of a predictive construction of priors in Bayesian nonparametrics.

Based on these results, the de Finetti measure of the sequence may be obtained as the limiting probability law of the sequence of predictive measures.

Theorem 1.2.3. *(de Finetti's representation theorem in terms of the predictive distributions) Let $(X_1, X_2, ...)$ be an exchangeable sequence, and let $X_1 \sim F_0$ and F_n be the predictive distribution of $X_{n+1}|X_1, ..., X_n$, for $n \geqslant 1$. Then the sequence $(F_1, F_2, ...)$ converges weakly to a random distribution function $\tilde{F}$, $\mathbb{P}$-a.s. Furthermore for any $n \geqslant 1$,*

$$\mathbb{P}\left[X_1 \in A_1, ..., X_n \in A_n\right] = \int_{\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} F(A_i) \mu(dF),$$

where μ is the probability distribution of $\tilde{F}$.

According to this theorem one can construct an exchangeable probability law by specifying the sequence of predictive distributions. From the Ionescu-Tulcea theorem, it follows that there exists a unique probability measure on $\mathbb{X}^\infty$, such that F_0 is the distribution of the first coordinate and $(F_1, F_2, ...)$ is the sequence of predictive distributions. Of course, not any sequence of predictive distributions corresponds to an exchangeable law. Fortini et al. [2000] specify necessary and sufficient conditions.

In the present work, we focus on a class of nonparametric models, namely species sampling models, that are defined through the predictive approach.

1.3 Species sampling models

Species Sampling Models are a class of mainly discrete random measures introduced by Pitman [1996a]. Many discrete random measures widely used in the Bayesian nonparametric approach belong to this class. Although discreteness may be considered as a disadvantage, there are cases where the discreteness of a prior is a positive feature. For example, in species sampling problems in epidemiology, ecology or population genetics, one considers a sequence of random variables $(X_1, X_2, ...)$ as a random sample from a large population of species, where X_i represents the species of the i-th individual sampled. The space $\mathbb{X}$ should be considered as some set of tags used to label distinct species. Even though it is not the only interpretation for these models, in what follows we use the "species" terminology to help the intuition. Species sampling models theory is based on two aspects: the exchangeability and the predictive approach.

Consider the following prediction rule:

$$\begin{aligned}
&\mathbb{P}(X_1 \in \cdot) = \nu(\cdot), \text{ and for } n \geqslant 1, \\
&\mathbb{P}(X_{n+1} \in \cdot \mid X_1, ..., X_n, K_n = k) = \sum_{i=1}^{k} p_i(\mathbf{n}) \delta_{X_i^\star}(\cdot) + p_{k+1}(\mathbf{n}) \nu(\cdot),
\end{aligned} \tag{1.1}$$

where K_n is the number of distinct species $(X_1^\star, ..., X_k^\star)$ that appeared among the observations $(X_1, ..., X_n)$; $\mathbf{n} = (n_1, ..., n_k)$ is the random vector of counts of the species observed; and ν is some diffuse distribution. Given $K_n = k$, the sequence of weights $(p_i,\ i = 1, .., k+1)$ in (1.1) are called *predictive weights* and should be understood in the following way: given the sample $(X_1, ..., X_n)$, the next observation X_{n+1} can be either equal to the j-th observed species, with probability $p_j(\mathbf{n})$, or it can correspond to a new class, with probability $p_{k+1}(\mathbf{n})$. The assumption that labels for new classes are generated from a diffuse distribution is needed in order to guarantee that labels are $\mathbb{P}$-a.s. distinct. It is worth noting that in general the prediction rule (1.1) does not generate an exchangeable sequence. For example, a sequence subject to a prediction rule with predictive weights $p_i(n_1, ..., n_k) = \frac{1}{k+1}$, for $i = 1..k+1$, is not exchangeable. Species sampling models theory is focused on prediction rules that generate exchangeable sequences.

Definition 1.3.1. *The sequence $(X_1, X_2, ...)$ is called a species sampling sequence if it is exchangeable and has a prediction rule of the form* (1.1).

Recall, that we defined $F_n(\cdot) = \mathbb{P}(X_{n+1} \in \cdot | X_1, ..., X_n)$ be the regular conditional (or predictive) distribution of X_{n+1} given $X_1, ..., X_n$.

Theorem 1.3.2. *(Pitman [1996a]) Suppose $(X_1, X_2, ...)$ is a species sampling sequence with a prediction rule* (1.1). *Then*

- *The sequence of predictive distributions $(F_1, F_2, ...)$ converges in total variation norm $\mathbb{P}$-a.s. to the random distribution*

$$F(\cdot) = \sum_i \tilde{P}_i \delta_{X_i^\star}(\cdot) + (1 - \sum_i \tilde{P}_i)\nu(\cdot), \tag{1.2}$$

  where $\tilde{P}_i$ is the relative frequency of the i-th species to appear, i.e.,

$$\tilde{P}_i = \lim_{n\to\infty} \frac{n_i}{n} \ \ \mathbb{P} - a.s.$$

  The distinct values $(X_1^\star, X_2^\star, ...)$ are i.i.d.(ν), and they are independent of the $(\tilde{P}_1, \tilde{P}_2, ...)$.

- *$(X_1, X_2, ...)$ is a sample from F.*

The species sampling model appears, then, as a directing random measure of some species sampling sequence.

Definition 1.3.3. *A random distribution F of the form* (1.2) *is called a species sampling model.*

Note that the random distribution of the form (1.2) has a probability mass $\tilde{P}_i$ at $X_i^\star$ for each i such that $\tilde{P}_i > 0$, and the rest of its mass is distributed proportionally to ν. A species sampling models is called proper if $\sum_i \tilde{P}_i = 1$ $\mathbb{P}$-a.s. In this case F is $\mathbb{P}$-a.s. discrete. It is easily verified that a species sampling model is proper if and only if $\mathbb{P}\left(\tilde{P}_1 > 0\right) = 1$. In this thesis we only consider proper models, as we are interested in discrete random measures.

In order to construct a species sampling sequence one should choose a prediction rule that generates an exchangeable sequence. Specifying a set of predictive weights, such that $p_i(\mathbf{n}) \geqslant 0$ and $\sum_{i=1}^{k(n)+1} p_i(\mathbf{n}) = 1$, determines the probability law of a sequence of random variables $(X_1, X_2, ...)$ via the prediction rule (1.1), but not all of such sequences are exchangeable. As we mentioned before, the prediction rule (1.1) itself does not guarantee exchangeability and additional conditions on the predictive weights are required. Before specifying these conditions, let us note that the predictive weights are defined as functions of the vectors of counts of the various species observed in a sample, i.e., they are connected with random partitions. Therefore, in the following section we recall the theory of exchangeable random partitions, which was introduced by Kingman [1978a], Kingman [1978b] and Aldous [1985] and is closely related to the notion of exchangeability.

1.3.1 Exchangeable random partitions

Following notations induced by Pitman [1995], a *partition* of $N_n = \{1, ..., n\}$ is an unordered collection of disjoint, non-empty subsets of N_n, say $\{A_i\}$, with $\cup_i A_i = N_n$. The A_i will be called *classes* of the partition. For example, a possible partition of N_6 is

$$\{(1,2,5)(3,6)(4)\}.$$

A *random partition* of N_n is a random variable Π_n with values in the finite set of all possible partitions of N_n.

For $N = \{1,2,...\}$ a *random partition* of N is a sequence $\Pi = (\Pi_1, \Pi_2, ...)$ of random partitions of N_n defined on a common probability space, such that for $m < n$ the restriction of Π_n to N_m is Π_m. More precisely, given a partition $\{A_i\}$ of N_n, the restriction of the partition $\{A_i\}$ to N_m for $m < n$, is the partition of N_m whose classes are the non-empty members of $\{A_i \cap N_m\}$. Each permutation σ of N acts on subsets $B \subset \{1,2,...\}$ by $\sigma(B) = \{\sigma(i) : i \in B\}$, and so it acts on partitions by $\sigma(A_1, A_2, ...) = (\sigma(A_1), \sigma(A_2), ...)$. Let us now define an exchangeable random partition, that is, a random partition such that we can permute elements within classes and between classes.

Definition 1.3.4. *A random partition Π_n of N_n is exchangeable if for any permutation σ of indices $1, ..., n$*

$$\mathbb{P}(\Pi_n = \{A_1, ..., A_k\}) = \mathbb{P}(\Pi_n = \{\sigma(A_1), ..., \sigma(A_k)\}).$$

A random partition $\Pi = (\Pi_1, \Pi_2, ...)$ of N is exchangeable if Π_n is an exchangeable partition for every n.

The notion of exchangeable partition is closely related with the notion of exchangeability. First of all, it is easy to see that any exchangeable sequence $(X_1, X_2, ...)$ generates an exchangeable partition $\Pi(X_1, X_2, ...)$ via the equivalence relation

$$i \sim j \iff X_i = X_j. \tag{1.3}$$

Moreover, Kingman has proved the converse, that starting from any exchangeable partition one can construct an exchangeable sequence (see Aldous [1985] for a short proof).

Proposition 1.3.5. *(Kingman's representation) For every exchangeable random partition* Π *of* N*, there exists an exchangeable sequence (not necessarily unique),* $(X_1, X_2, ...)$*, directed by some random measure* F*, such that* $\Pi \overset{d}{=} \Pi(X_1, X_2, ...)$ *and the distribution of sizes of classes of* Π *is determined by the joint distribution of the sizes the atoms of* F*.*

Recall that, according to de Finetti's theorem, any infinite exchangeable sequence can be represented as a mixture of i.i.d. sequences. For exchangeable partitions of N, the role of the i.i.d. sequences is played by the "paint-box processes", that is the partitions generated via the equivalence relation (1.3), from sequences of i.i.d. random variables. The analogue of de Finetti's representation follows from Kingman's representation, and so, an exchangeable partition can be represented as a mixture of independent paintbox processes. More formally, let ϕ be a discrete distribution on $[0, 1]$ and $(X_1, X_2, ...)$ be i.i.d. according to ϕ. Let $\Pi = (X_1, X_2, ...)$ be a partition generated from $(X_1, X_2, ...)$ through the equivalence relation (1.3). In this construction the distribution of Π_n depends only on the sizes of the atoms of ϕ. If the ranked weights of ϕ are defined by a sequence $\rho = \{P_{(1)}, P_{(2)}, ...\}$, then call the partition Π above a *paintbox(ρ) process.* By construction, any paintbox process is an exchangeable partition of N and its distribution depends only on the sizes of the atoms of ϕ.

Note that, when we talk about a sequence, we are assuming some order of appearance of the elements, first X_1 then X_2 and so on. So far we defined a random partition as an unordered set of classes. In order to study further connections between exchangeable sequences and exchangeable partitions we need to specify some order on the classes of partitions, the most natural being is the order of appearance.

Definition 1.3.6. *A partition* $\{A_1, ..., A_k\}$ *of* N_n *has classes in order of appearance if* $1 \in A_1$ *and the first element of* $N_n \setminus (A_1 \cup ... \cup A_{i-1})$*, for each* $2 \leqslant i \leqslant k$*, belongs to* A_i*.*

This is a very natural ordering as we build classes in a partition according to the order in which such classes appear in the sequence. With this order, an equivalent definition of an exchangeable partition can be given. Let $N^\star = \cup_{k=1}^{\infty} N_k$ be the set of finite sequences of positive integers, denote also $\Sigma(\mathbf{n}) = \sum_{i=1}^{k} n_i$, for $\mathbf{n} = (n_1, ..., n_k) \in N^\star$, and $N_n^\star = \{\mathbf{n} \in N^\star : \Sigma(\mathbf{n}) = n\}$. Pitman [1995] gave the following characterization of exchangeable partitions with classes in order of appearance.

Proposition 1.3.7. *A random partition Π_n of N_n is exchangeable if and only if for every partition $\{A_1, ..., A_k\}$ of N_n, where $A_1, ..., A_k$ are in order of appearance,*

$$\mathbb{P}(\Pi_n = \{A_1, ..., A_k\}) = \mathbf{p}(\#(A_1), ..., \#(A_k)),$$

for some symmetric function $\mathbf{p}(\mathbf{n}) = \mathbf{p}(n_1, ..., n_k)$ defined for $\mathbf{n} \in N^\star$with $\sum(\mathbf{n}) = n$, where $(n_1, ..., n_k)$ is a vector of size of classes of the partition. Then, $\mathbf{p}(\mathbf{n})$ is called an exchangeable partition probability function (EPPF).

EPPFs play a crucial role in species sampling theory as they characterize the distribution of an exchangeable partition generated by a species sampling sequence. Using this characterization Pitman [1996a] specified the condition on predictive weights $(p_i,\ i = 1, 2, \ldots)$ that guarantees exchangeability of a sequence with a prediction rule (1.1).

Theorem 1.3.8. *(Pitman [1996a]) Given a diffuse probability distribution ν and a sequence of predictive weights, let $(X_1, X_2, ...)$ be a sequence of random variables with prediction rule* (1.1)*. Then $(X_1, X_2, ...)$ is exchangeable if and only if there exists a non-negative, symmetric function $\mathbf{p}$ defined on $N^\star$ such that*

$$p_i(n_1, ..., n_k) = \frac{\mathbf{p}(\mathbf{n}^{i+})}{\mathbf{p}(\mathbf{n})} = \frac{\mathbf{p}(\mathrm{n_1, ..., n_i + 1, ..., n_k})}{\mathbf{p}(n_1, ..., n_i, ..., n_k)} \; for\ 1 \leq i \leq k+1. \quad (1.4)$$

Then, $(X_1, X_2, ...)$ is a sample from F and the EPPF of $(X_1, X_2, ...)$ is the unique non-negative symmetric function $\mathbf{p}$ that satisfies (1.4) *and* $\mathbf{p}(1) = 1$*.*

In other words, a predictive scheme of the form (1.1) will generate an exchangeable sequence if and only if the predictive weights are obtained from some EPPF. In particular, as predictive weights are uniquely defined by the EPPF through (1.4), one can construct a species sampling sequence and, therefore a species sampling model, starting from any EPPF. In general, however, specifying a symmetric function that satisfies (1.4) is not simple. Let us think about this issue from a different perspective, taking into account that EPPFs are connected with the random weights of species sampling models.

Suppose we have a proper species sampling model, i.e., a discrete random measure $F(\cdot) = \sum_i P_i \delta_{X_i^\star}(\cdot)$ with random weights $P_i \geqslant 0$, $\sum_i P_i = 1$ $\mathbb{P}$-a.s., which are independent of the random labels $(X_1^\star, X_2^\star, ...) \overset{i.i.d.}{\sim} \nu$. The question is, how can we construct an EPPF that corresponds to the behavior of the exchangeable random partition generated by a sample from this species sampling model? In principle, knowing the joint distribution of the random weights $(P_1, P_2, ...)$, the corresponding EPPF can be computed as

$$\mathbf{p}(n_1, ..., n_k) = \sum_{i_1 \neq ... \neq i_k} \mathbb{E}\left[P_{i_1}^{n_1} \cdots P_{i_k}^{n_k}\right], \tag{1.5}$$

where the sum is taken over all sequences of distinct positive integers. In general, this formula is quite complicated, but if we assume an additional condition on the random weights $(P_1, P_2, ...)$, the formula becomes much simpler.

Recall that a *size-biased permutation* of the weights of some random measure is a natural analogue of the order of appearance for classes of partitions, i.e., it is the random order in which species appear in an exchangeable sample from the random measure. Roughly speaking, it is a condition for some kind of symmetry. We call the sequence $(P_1, P_2, ...)$ *invariant under size-biased permutation* if it has the same finite dimensional distributions under a size-biased permutation σ, i.e., $(P_1, P_2, ...) \overset{d}{=} (P_{\sigma(1)}, P_{\sigma(2)}, ...)$. A more detailed review of size-biased permutations is given by Donnelly and Joyce [1989]. Random discrete distributions that are invariant under size-biased permutations are discussed in Pitman [1996b]. Applying some results of Pitman [1995], one may conclude that, if the set of random weights of the species sampling model is invariant under size-biased permutations, the expression of the EPPF corresponding to the given random measure is more easily obtained.

Proposition 1.3.9. *(Pitman [1995]) Let* $F(\cdot) = \sum_i P_i \delta_{X_i^\star}(\cdot)$ *be a species sampling model. The EPPF function that corresponds to the exchangeable partition generated via equivalence relation* (1.3)*, from a sample from* F*, can be computed as:*

$$\mathbf{p}(n_1, ..., n_k) = \mathbb{E}\left[\left(\prod_{i=1}^{k} \tilde{P}_i^{n_i - 1}\right) \prod_{i=1}^{k-1} \left(1 - \sum_{j=1}^{i} \tilde{P}_j\right)\right], \tag{1.6}$$

where $\left(\tilde{P}_1, \tilde{P}_2, ...\right)$ *is a size-biased permutation of* $(P_1, P_2, ...)$*.*

If the random weights are not invariant under size-biased permutations, applying (1.6) is still possible, but becomes a tricky task. One way to get a set of random weights $(P_1, P_2, ...)$ that is invariant under size-biased permutations is to use a *stick-breaking* construction, where we assume that there exist a set of random variables $(U_1, U_2, ...)$ with values in $[0, 1]$ such that

$$P_i = (1 - U_i) \prod_{j=1}^{i-1} U_j, \; i \geqslant 1. \tag{1.7}$$

McCloskey [1965] and Perman et al. [1992] give conditions for the set of random weights constructed by the stick-breaking approach to be invariant under size-biased permutations.

Theorem 1.3.10. *(McCloskey) Let* $(P_1, P_2, ...)$ *be such that* $P_n \geqslant 0$*,* $\sum P_n = 1$*,* $P_1 < 1$ *and* $P_n = (1 - U_1)(1 - U_2)...(1 - U_{n-1})U_n$ *for* $n \geqslant 1$*, where* $(U_1, U_2, ...)$ *are i.i.d. with values in* $[0, 1]$*. Then* $(P_1, P_2, ...)$ *is invariant under size biased permutations if and only if the common distribution of the* U_i *is Beta(*$1, \theta$*) for some* $0 \leqslant \theta < \infty$*.*

Theorem 1.3.11. *(Perman, Pitman, Yor) Let* $(P_1, P_2, ...)$ *be such that* $P_n \geqslant 0$*,* $\sum P_n = 1$*,* $P_1 < 1$ *and* $P_n = (1 - U_1)(1 - U_2)...(1 - U_{n-1})U_n$ *for* $n \geqslant 1$*, where* $(U_1, U_2, ...)$ *are independent random variables (but not necessarily identically distributed) with values in* $[0, 1]$*. Then* $(P_1, P_2, ...)$ *is invariant under size biased permutations if and only if* $P_n > 0$ $\mathbb{P}$*-a.s. for all* n*, in which case the distribution of* U_n *is Beta*$(1 - \sigma, \theta + n\sigma)$ *for every* $n = 1, 2, ...$*, for some* $0 \leqslant \sigma < 1$*,* $\theta > -\sigma$*.*

Thus, there are only two ways to construct invariant under size-biased permutation random weights by using the stick-breaking construction with independent $(U_1, U_2, ...)$. One of such examples is considered in Favaro et al. where they obtain a stick-breaking representation of a normalized inverse Gaussian process. For other models we should assume $(U_1, U_2, ...)$ to be dependent.

In order to summarize the main relations in species sampling theory we plot a diagram (Table 1.1). Each number on the diagram corresponds to some fact or definition discussed before.

1. Any exchangeable sequence creates an exchangeable partition through the equivalence relation (1.3).

2. For any exchangeable partition, one can construct an exchangeable sequence such that the sequence of ranked weights $\left(P_{(1)}, P_{(2)}, ...\right)$ of its directing random measure coincides with the sequence of ranked limiting relative sizes of the classes of this partition (Kingman's representation).

3. A random partition with classes in order of appearance is exchangeable if and only if its distribution can be expressed through a symmetric function $\mathbf{p}$, called EPPF (Proposition 1.3.7).

4. If $(X_1, X_2, ...)$ has a prediction rule of form (1.1), then $(X_1, X_2, ...)$ is exchangeable if and only if the predictive weights can be expressed through some EPPF (Theorem 1.3.8).

5. If the sequence $(X_1, X_2, ...)$ is exchangeable and has a prediction rule of form (1.1), then it is a species sampling sequence (Definition 1.3.1).

6. If the prediction rule of form (1.1) generates an exchangeable sequence, then a sequence of predictive distributions converges $\mathbb{P}$-a.s. in total variation norm to a species sampling model (Theorem 1.3.2).

7. The EPPF function $\mathbf{p}$ can be computed through size-biased permutation of random weights of the species sampling model (Proposition 1.3.9).

Table 1.1: Scheme of relationships in species sampling theory.

$$\begin{tikzcd}
 & & \text{in order of appearance } \curvearrowright & & \textbf{EPPF}\text{: symmetric function } \mathbf{p} \curvearrowright \\
 & & \Pi_n = [A_1, .., A_k] \arrow[rr, Leftrightarrow, "(3)\ \mathit{def}", "\mathit{exchangeable}"'] \arrow[dd, shift left, "(2)\ \text{Kingman}"] & & \mathbb{P}\left(\Pi_n = [A_1, .., A_k]\right) = \mathbf{p}(n_1, .., n_k) \\
\textbf{SSS} & & & & \\
(5)\ (X_1, X_2, ...) & & (X_1, ..., X_n) \arrow[ll, Rightarrow] \arrow[uu, shift left, "(1)\ \overset{i \sim j}{\underset{X_i = X_j}{\Leftrightarrow}}"] \arrow[rr, Leftrightarrow, "(4)\ \mathit{Pitman}", "\mathit{exchangeable}"'] & & \textbf{PPF: } p_i(n_1, ..., n_k) = \frac{\mathbf{p}(n_1, ..., n_i + 1, ..., n_k)}{\mathbf{p}(n_1, ...n_k)} \\
\textbf{SSM} & & & & \\
(6)\ F(\cdot) \arrow[d, equal] & & \mathbb{P}\left(X_{n+1} \in \cdot \mid X_1, ..., X_n, k\right) = \sum_{i=1}^{k} p_i(n_1, .., n_k)\delta_{X_i^\star}(\cdot) + p_{k+1}(n_1, ..., n_k)\nu(\cdot) \arrow[ll, Rightarrow] \arrow[uu, Rightarrow] & & \\
\sum_{i=1}^{\infty} P_i^\star \delta_{X_i^\star}(\cdot) & & & & \\
 & & (7)\ \textbf{EPPF: } \mathbf{p}(n_1, .., n_k) = \mathbb{E}\left[\prod_{i=1}^{k} P_1^{\star\, n_1 - 1} \prod_{i=1}^{k-1}(1 - \sum_{j=1}^{i} P_j^\star)\right] & &
\end{tikzcd}$$

Many popular discrete random measures belong to the class of species sampling models. Lijoi and Prünster [2010] give a rich and detailed review. One class of models is given by homogeneous normalized random measures with independent increments. Another class is that of Gibbs-type random probability measures. They both fall into the more general class of Poisson-Kingman models. Can we construct EPPFs for all these species sampling models? Theoretically there exists an EPPF for an exchangeable sample from any species sampling model, but in practice, although a large number of discrete random measures of species sampling model type are known, only few of them have a simple EPPF and therefore, an interpretable prediction rule. For others it is too difficult to evaluate the EPPF explicitly. For example, an EPPF for normalized random measures is specified in Lijoi and Prünster [2010] (formula 3.32) but, apart from few cases, no closed form expressions are available, even if drawing samples from the predictive is still possible by conditioning on some set of latent variables. Pitman [2003] gives a general expression for the EPPF of Poisson-Kingman's models, but it is again too difficult to evaluate it for a particular model.

Let us consider two widely used examples of species sampling models that have simple EPPFs and tractable prediction rules.

Example 1: Dirichlet Process

A cornerstone in Bayesian nonparametrics is the Dirichlet Process introduced by Ferguson [1973]. Since the focus of this thesis is on the predictive approach, here we discuss only the predictive characterization of the Dirichlet Process. By Blackwell and MacQueen [1973] a prediction rule that generates an exchangeable sample from a Dirichlet Process with parameter $\theta > 0$ and base measure ν has the form:

$$\begin{aligned}
&\mathbb{P}(X_1 \in \cdot) = \nu(\cdot), \text{ and for } n \geqslant 1, \\
&\mathbb{P}(X_{n+1} \in \cdot \mid X_1, ..., X_n, K_n = k) = \sum_{i=1}^{k} \frac{n_i}{n+\theta}\delta_{X_i^\star}(\cdot) + \frac{\theta}{n+\theta}\nu(\cdot).
\end{aligned}$$

Thus, the probability that the next observation will be of the i-th observed species only depends on number of previous observations of such type. This characterization of the Dirichlet distribution was first discovered in the 1920s by the English philosopher W. E. Johnson. Similar characterization of the Dirichlet Prior was provided by Regazzini [1978] and Lo [1991]. This form of predictive is a natural way of thinking nonparametrically and that makes this prediction rule very attractive for biological sampling.

The corresponding EPPF can be found through (1.6) if we remember that, according to Sethuraman [1994], the set of random weights for a Dirichlet Process can be constructed using the stick-breaking approach

$$P_i = W_i \prod_{j=1}^{i-1} (1 - W_j),$$

where $W_i \overset{i.i.d.}{\sim} Beta(1, \theta)$. This distribution is called Griffiths-Engen-McCloskey, or GEM, distribution and it is invariant under size-biased permutations. Applying (1.6),

$$\mathbf{p}(n_1, ..., n_k) = \mathbb{E}\left[\left(\prod_{i=1}^{k} P_i^{n_i - 1} \right) \prod_{i=1}^{k-1} \left(1 - \sum_{j=1}^{i} P_j \right) \right] = \frac{\theta^k \prod_{i=1}^{k} (n_i - 1)!}{(\theta + 1)...(\theta + n - 1)}.$$

This formula is also called the *Ewen's sampling formula* (Ewens [1972]).

A particular feature of the Dirichlet Process is that a class that has been frequently observed has more chances to be observed again than a class that has been observed only rarely. This property is called “The rich get richer” property and it implies that in the resulting sequence there are only a few number of large classes and many small classes. Such behavior may be undesirable, leading to a generalization, known as the two-parameter Poisson Dirichlet Process.

Example 2: Two-parameter Poisson Dirichlet Process

The two-parameter Poisson Dirichlet Process, sometimes also called Pitman-Yor process, was proposed and analyzed in Perman [1990] and Perman et al. [1992]. Including an additional parameter σ makes this process more flexible and capable of accommodating power-law tails than the Dirichlet Process.

The prediction rule of this two-parameter process process with parameters $\theta > 0$, $0 < \sigma < 1$ and base measure ν is

$$\mathbb{P}(X_1 \in \cdot) = \nu(\cdot) \text{ and for } n \geqslant 1,$$

$$\mathbb{P}(X_{n+1} \in \cdot \mid X_1, ..., X_n, K_n = k) = \sum_{i=1}^{k} \frac{n_i - \sigma}{n + \theta} \delta_{X_i^\star}(\cdot) + \frac{\theta + k\sigma}{n + \theta} \nu(\cdot).$$

Here, the probability that the next observation joins the i-th observed species is smaller than for a Dirichlet Process, while the probability of observing a new class is instead larger. The additional parameter σ increases the probability of observing a new class.

Using stick-breaking approach, the set of random weights can be constructed as

$$P_i = W_i \prod_{j=1}^{i-1} (1 - W_j),$$

where $W_i \overset{ind}{\sim} Beta(1 - \sigma, \theta + i\sigma)$. As before, the resulting probability law of the random weights is invariant under size-biased permutations and through (1.6) we can get the EPPF:

$$\mathbf{p}(n_1, ..., n_k) = \frac{\theta(\theta + \sigma) \ldots (\theta + (k-1)\sigma) \prod_{i=1}^{k} (1 - \theta)_{n_i}}{(\theta + 1)...(\theta + n - 1)},$$

where $(a)_m = a(a+1)...(a+m-1)$.

1.4 Partial exchangeability and dependent random measures

This thesis is focused on species sampling from two (or more) populations, and exchangeability, which involves an assumption about complete symmetry among all observations, is generally too restrictive. It seems reasonable to assume that we can permute observations within each group, but not between groups. Such generalization of exchangeability is called partial exchangeability. There are slightly different notions of partial exchangeability in the literature, we will use partial exchangeability in the sense of de Finetti (de Finetti (1938); the English translation of this work can be found in Carnap and Jeffrey [1980]).

Definition 1.4.1. *An infinite sequence* $(X_1, X_2, ..., Y_1, Y_2, ...)$ *of* $\mathbb{X}$*-valued random variables is called partially exchangeable if, for any* n, $m \geqslant 1$ *and any permutations,* σ *of* $1, ..., n$ *and* π *of* $1, ..., m$,

$$(X_1, ..., X_n, Y_1, ..., Y_m) \overset{d}{=} (X_{\sigma(1)}, ..., X_{\sigma(n)}, Y_{\pi(1)}, ..., Y_{\pi(m)}).$$

Note that marginally $(X_1, X_2, ...)$ and $(Y_1, Y_2, ...)$ are both exchangeable but two exchangeable sequences are not necessarily partially exchangeable. An additional condition is needed, which will become clear after specifying the analogue of de Finetti's representation theorem for partially exchangeable sequences.

Theorem 1.4.2. *The sequence* $(X_1, X_2, ..., Y_1, Y_2, ...)$ *with values in a Polish space* $\mathbb{X}$*, is partially exchangeable if and only if there exists a probability measure* τ *on* $\mathcal{P}_{\mathbb{X}} \times \mathcal{P}_{\mathbb{X}}$*, the set of all probability measures on* $(\mathbb{X} \times \mathbb{X}, \mathcal{X} \times \mathcal{X})$*, such that for any* n, $m \geqslant 1$ *and any* $A_1, ..., A_n, B_1, ..., B_m$ *in* $\mathcal{X}$*,*

$$\begin{gathered}\mathbb{P}\left[X_1 \in A_1, ..., X_n \in A_n, Y_1 \in B_1, ..., Y_m \in B_m\right] = \\ = \int_{\mathcal{P}_{\mathbb{X}} \times \mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} F^X(A_i) \prod_{j=1}^{m} F^Y(B_j) \tau(dF^X, dF^Y).\end{gathered}$$

The probability measure τ *is called the de Finetti measure of the sequence* $(X_1, X_2, ..., Y_1, Y_2, ...)$*, it is unique and arises as the* $\mathbb{P}$*-a.s. limiting distribution of the sequence of empirical distributions*

$$(\tilde{F}_n^X, \tilde{F}_m^Y) = \left(\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \frac{1}{m}\sum_{i=1}^{m} \delta_{Y_i}\right).$$

In other words, a partially exchangeable sequence can be represented as a mixture of two independent i.i.d. sequences:

$$\begin{aligned}(X_i, Y_j)\,|\tilde{F}^X, \tilde{F}^Y &\overset{i.i.d.}{\sim} \tilde{F}^X \times \tilde{F}^Y; \\ (\tilde{F}^X, \tilde{F}^Y) &\sim \tau.\end{aligned}$$

Thus, the condition required for two exchangeable sequences to be partially exchangeable is: they should also be conditionally independent given the vector of their directing random measures.

Partial exchangeability is a natural assumption in species sampling problems, where we have two dependent, not homogeneous populations of species. In this setting, the random measures $(\tilde{F}^X, \tilde{F}^Y)$ correspond to the unknown species abundances in the populations. Specifying a nonparametric prior τ that provides dependence between these two random measures is not an easy task. At the same time, the problem is relevant and has been widely investigated since Bayesian nonparametric methods have become increasingly popular in empirical studies. Some review of methods for combining inference across related nonparametric models can be found in Müller et al. [2004] and Kolossiatis et al. [2013].

One possible solution is to link separate nonparametric models at the level of hyperparameters only, i.e., submodels are conditionally independent given hyperparameters. For example, Cifarelli and Regazzini [1978] proposed mixtures of products of Dirichlet Process, where the base measure of a Dirichlet Process prior for each submodel includes a regression on covariates that are specific to the submodel. These methods have recently been applied to problems in biostatistics, econometrics and survival analysis. A similar idea was proposed in Mallick and Walker [1997]. Although straightforward, these strategies are limited to learning about features that can be represented in hyperparameters. Alternative approaches are based on introducing dependence using stick-breaking construction proposed by Sethuraman [1994]. MacEachern [1999] proposed a general class of Dependent Dirichlet Processes with dependence imposed through correlations across the atoms and the point masses in Sethuraman's representation of the Dirichlet Process model mentioned before (equation (1.7)). Similarly, Griffin and Steel [2006] allow the nonparametric distribution to depend on covariates through an ordering of the random variables building the weights in the stick-breaking representation. Although this construction is widely used in many areas due to its attractiveness from a computational point of view, a drawback of the stick-breaking construction is the difficulty of studying the distributional properties of the nonparametric distributions, due to their analytical intractability. At the same time, there are no prediction rules available for these models. In this thesis we are interested on the predictive approach. Therefore, we focus on existing models, which allow for a predictive representation.

A widely used approach is to define dependence through the base measures of the Dirichlet Processes. Walker and Muliere [2003] construct Bivariate Dirichlet Processes, where the dependence between two conditionally independent Dirichlet Process priors is defined through a set of latent variables. Roughly speaking, for some set of latent variables $(Z_1, ..., Z_n)$, they consider two Dirichlet Processes with base measures $\nu + \sum_{i=1}^{n} \delta_{Z_i}$, where ν is a diffuse distribution. On one hand these latent variables help to specify the dependence, on the other hand, as the modified base measure is not diffuse, analyzing the corresponding random partitions becomes a very complicated task. The same problem arises for the Hierarchical Dirichlet Process proposed by Teh et al.. They consider Dirichlet Processes with base measures that are realizations of another Dirichlet Process. As before, the base measure is not diffuse and we cannot say if one observation is a new one or old. Of course, one could introduce a different set of latent variables, but integrating them out is not an easy problem, and the corresponding predictive rule becomes intractable.

Another way to define dependence arises from the use of mixtures of distributions. For example, the single-ϵ model was originally proposed by Müller et al. [2004] and was later developed and generalized by Kolossiatis et al. [2013]. They assume

$$F_X^\star = \epsilon\, F_0 + (1-\epsilon)F_1 \text{ and } F_Y^\star = \epsilon\, F_0 + (1-\epsilon)F_2,$$

where F_0, F_1, F_2 are some specified Dirichlet Processes and ϵ is a random variable with values in $[0, 1]$. The marginal distributions of $F_X^\star$, $F_Y^\star$ can be shown to be Dirichlet Processes and dependence is obtained through the common component F_0. The prediction rule of this processes can be defined only conditionally on a set of binary indicators depending on whether the observation is drawn from the common F_0 or not. Although Markov chain Monte Carlo methods can be used for exploring properties of these distributions, the predictive rule is complicated and intractable. Another example of such approach, proposed by Lijoi et al. [2011], presents the same problem. They consider a class of normalized dependent complete random measures, where dependence is defined through the additive representation of the underlying Poisson random measures $\tilde{N}_1$ and $\tilde{N}_2$:

$$\tilde{N}_1 = M_0 + M_1 \text{ and } \tilde{N}_2 = M_0 + M_2,$$

where M_0, M_1, M_2 are independent Cox processes. Again, a set of latent variables is involved in the specification of the predictive rule. Integrating them out is computationally possible but the predictive rule is too complicated.

An alternative model, which avoids the introduction of latent variables is the vector of two parameter Poisson Dirichlet Processes proposed by Leisen and Lijoi [2011], where dependence is induced by a Levy copula acting on the respective marginal intensities. Although an expression for the bivariate EPPF is provided, the predictive distributions are very complicated. Thus, to the best of our knowledge, there is no bivariate model that has a tractable prediction rule. The first aim of this thesis is to use a predictive approach and investigate to what extent one can generate a partially exchangeable sequence from a relatively simple bivariate prediction rule.

Chapter 2

Bivariate species sampling models

2.1 Introduction

In different areas of science, such as ecology, epidemiology and genetics, there is interest on analyzing species abundances within some region of interest. In the previous chapter we discussed a class of models, called species sampling models, that is widely used for such purposes. However, in many applications the region of interest is not homogeneous. For example, in biology, samples might come from different parts of a lake that have diverse physical and chemical properties of water. Then one is interested in a multivariate extention of this theory to species sampling from two (or more) dependent populations. We focus on the bivariate case, but multivariate extension can be envisaged.

In order to construct dependent random measures we use a predictive approach assuming partial exchangeability, since the notion of exchangeability, which implies complete symmetry among all observations, becomes too restrictive in such problems. We investigate to what extent one can generate a partially exchangeable sequence from a tractable bivariate prediction rule. Since bivariate partitions play a crucial role in specifying a bivariate prediction rule, we start by defining the concept of bivariate random partitions and partially exchangeable bivariate random partitions.

Specifying an order on classes of bivariate partitions allows us to define the distribution of random partially exchangeable bivariate partitions through some symmetric function. Then, we define a class of bivariate species sampling models and specify the conditions on the bivariate prediction rule that guarantee partial exchangeability of the corresponding process. We will see that the conditions for having a tractable and manageable prediction rule and partial exchangeability are quite restrictive. Indeed, if we assume that the probability of the next observation to be of the i-th observed species is proportional to a function of the frequency of that species in the bivariate sample, then the model reduces to a Dirichlet Process for a homogeneous population.

2.2 Bivariate partitions

In Chapter 1, we have seen that random partitions play a crucial role in specifying the prediction rule for species sampling models. If one has a sequence $X_{1:n} = (X_1, ..., X_n)$, then a random partition can be created using the equivalence relation (1.3). This relation is natural and straightforward. In this chapter we would like to generalize the species sampling theory for a bivariate case, thus a bivariate prediction rule should be specified. As we assume that the population is not homogeneous, we should also define a bivariate random partition. Suppose we have a bivariate sequence $(X_{1:n}, Y_{1:m}) = (X_1, ..., X_n, Y_1, ..., Y_m)$. How can one determine the bivariate partition $\Pi_{n,m}$ generated by this sequence? For a bivariate sequence there is no a unique way to define a bivariate partition, since shared values can be treated in different ways. Roughly speaking, there are three possibilities: the first one is to do not take into account shared values in the two sequences; the second one is to assume that some classes are shared, but some are specific and can not be shared; the third one is to assume that all values can be shared. We will discuss all three approaches and choose the one that is more appropriate for our purposes.

I. The simplest way to create a bivariate random partition from a bivariate sequence $(X_{1:n}, Y_{1:m})$ is to consider a vector of two univariate random partitions, created independently from $(X_1, ..., X_n)$ and $(Y_1, ..., Y_m)$ through the univariate equivalence relation (1.3), i.e.,

$$\Pi_{n,m} = (\Pi(X_{1:n}), \Pi(Y_{1:m})) = [(A_1, ..., A_k), (B_1, ..., B_r)],$$

where k and r are the number of distinct species in $(X_1, ..., X_n)$ and $(Y_1, ..., Y_m)$, correspondingly. Even if partitions Π_n and Π_m are created independently, they will be dependent if $(X_1, ..., X_n)$ and $(Y_1, ..., Y_m)$ are dependent. Note that the classes in two partitions that have the same index i, for example A_1 and B_1, can correspond to distinct values, as this construction does not take into account which classes can be shared. Such bivariate partition will be partially exchangeable if it is symmetric under any permutations σ of $1, ..., n$ and τ of $1, ..., m$, i.e.,

$$\begin{aligned}&\mathbb{P}\left(\Pi_{n,m} = [(A_1, ..., A_k), (B_1, ..., B_r)]\right) = \\ &= \mathbb{P}\left(\Pi_{n,m} = [(\sigma(A_1), ..., \sigma(A_k)), (\tau(B_1), ..., \tau(B_r))]\right),\end{aligned}$$

where each permutation σ of N acts on subsets $C \subset \{1, 2, ...\}$ by $\sigma(C) = \{\sigma(i) : i \in C\}$. An order on the classes of such bivariate random partitions can be defined independently for the two partitions. This structure is suitable for use in predictive rules, as it is consistent for adding new elements and all results from species sampling theory can be generalized in this framework. However, such construction does not connect classes in two partitions that correspond to the same value, while it seems desirable to use this information.

II. The second way to generate a bivariate random partition from a bivariate sequence $(X_{1:n}, Y_{1:m})$ is to separate common classes, observed in both groups, from specific classes observed only in one of the groups. Such approach is used in Lijoi et al. [2011] and Kolossiatis et al. [2013]. Suppose that among $(X_{1:n}, Y_{1:m})$, d values were shared, k specific values appeared in the first group, r specific values appeared in the second. Such partition can be described through 3 sub-partitions:

$$\Pi_{n,m} = \Pi(X_{1:n}, Y_{1:m}) = [(A_1, ..., A_k), (B_1, ..., B_r), (A_1^c, B_1^c), ..., (A_d^c, B_d^c)],$$

where the first sub-partition, $(A_1, ..., A_k)$, contains specific classes for the first group and each class A_i consists of indices belonging to this specific class. The second sub-partition, $(B_1, ..., B_r)$, contains specific classes for the second group and each class B_j consists of indices of elements in $(Y_{1:m})$ that belong to this specific class. The sub-partition, $(A_1^c, B_1^c), ..., (A_d^c, B_d^c)$, contains shared classes that consist on two sets of indices. It is clear that $\{\cup_{i=1}^p A_i\} \cup \left\{\cup_{i=1}^d A_i^c\right\} = \{1, ..., n\}$ and $\{\cup_{i=1}^r B_i\} \cup \left\{\cup_{i=1}^d B_i^c\right\} = \{1, ..., m\}$.

Separation of classes into specific and common allows us to connect shared classes. However, if some value has been observed in two sequences, we can be sure that this value is of common type, while if some value has been observed only in one sequence, we do not know if it is a specific or a common value. Thus, latent variables should be involved in order to tell us if the class is truly shared or specific. The result will be a conditional bivariate partition and the prediction rule will be also defined conditionally on latents variables. Integrating them out is a possible but complicated task.

At the same time, in species sampling problems, observations are sampled sequentially in accordance with the advent of new observations. The bivariate partition structure defined above is not appropriate for us, as it is not consistent under addition of new observations, i.e., it may happen that at time n some class is considered as specific, since it has been observed only in one group, but at the next moment we may observe such type also in the second group and the class has to be redefined as common. Remembering also, that an order of appearance was needed in one-dimensional case, it is not clear how to define an order on such classes if they can change type.

III. The last approach is to assume that all values can be shared. Such approach is used in Leisen and Lijoi [2011]. Suppose that p distinct values have been discovered among $(X_{1:n}, Y_{1:m})$, then we can construct a bivariate partition by defining for each value $i = 1, ..., p$ a pair of classes (A_i, B_i), where A_i consists of indices of elements in $(X_{1:n})$ that belong to class i and B_i consists of indices of elements in $(Y_{1:m})$ that belong to the same class. It can happen that one element in this pair is empty if the value was observed only in one of the sequences. We require $\cup_{i=1}^p A_i = \{1, ..., n\}$ and

$\cup_{j=1}^{p} B_j = \{1, ..., m\}$. Then, the bivariate partition is

$$\Pi(X_{1:n}, Y_{1:m}) = \{(A_1, B_1), ..., (A_p, B_p)\}.$$

Such bivariate partition is consistent under addition of new observations, and abundances of species of the same type are connected. We will use this approach. Formalizing,

Definition 2.2.1. *Suppose we have a bivariate sequence* $(X_{1:n}, Y_{1:m})$, *a bivariate partition* $\Pi(X_{1:n}, Y_{1:m}) = \{(A_1, B_1), ..., (A_p, B_p)\}$ *can be defined through three steps:*

1. *Generate a partition* $\Pi_n = (A_1, ..., A_k)$ *from the first sequence* $(X_{1:n})$ *via the equivalence relation* (1.3)*;*

2. *Generate a partition* $\Pi'_m = (B_1, ..., B_l)$ *from the second sequence* $(Y_{1:m})$ *via the equivalence relation* (1.3)*;*

3. *Connect the same classes, i.e., for* $i = 1...k$ *and* $j = 1...l$, $A_i \sim B_j$ *if and only if* $X_s = Y_t$ *for* $s \in A_i$, $t \in B_j$. *If for some* i, *there is no* j *such that* $A_i \sim B_j$ *then connect* A_i *with an empty set. If for some* j, *there is no* i *such that* $A_i \sim B_j$ *then connect* B_j *with an empty set.*

Now we are ready to define a bivariate random partition structure.

2.2.1 Random bivariate partitions

Denote $N_n \times N_m = \{1, ...n; 1, ..., m\}$ and $N \times N = \{1, 2, ...; 1, 2, ...\}$.

Definition 2.2.2. *A bivariate partition (bi-partition) of* $N_n \times N_m$ *is an unordered set of pairs* $\{(A_i, B_i)_i\}$, *called bi-classes (A-biclass and B-biclass), where* $\{A_i\}$ *are disjoint, possibly empty, subsets of* N_n, $\{B_i\}$ *are disjoint, possibly empty, subsets of* N_m, *and for every* i *either* A_i *or* B_i *is not empty. We also require that* $\cup_i A_i = N_n$, $\cup_i B_i = N_m$.

Given a bi-partition $\{(A_i, B_i)_i\}$ of $N_n \times N_m$, for $s < n$, $t < m$, the restriction of $\{(A_i, B_i)_i\}$ to $N_s \times N_t$ is a bi-partition of $N_s \times N_t$ whose bi-classes are non-empty members of $(A_i \cap N_s,\ B_i \cap N_t)$.

Definition 2.2.3. *A random bi-partition of $N_n \times N_m$ is a random variable Π_{nm} with values in the finite set of all bi-partitions of $N_n \times N_m$.*

Definition 2.2.4. *A random bi-partition of $N \times N$ is a sequence of random bi-partitions of $N_n \times N_m$, $\Pi = (\Pi_{nm})_{n,m \geqslant 1}$, defined on a common probability space such that for $s < n$, $t < m$ the restriction of Π_{nm} to $N_s \times N_t$ is Π_{st}.*

A pair of permutations, σ of $1, ..., n$ and τ of $1, ..., m$, act on a bi-class (A, B) by $(\sigma(A), \tau(B)) = (\{\sigma(i) : i \in A\}, \{\tau(j) : j \in B\})$.

Definition 2.2.5. Π_{nm} *is a partially exchangeable random bi-partition, if for any permutations, σ of $1, ..., n$ and τ of $1, ..., m$,*

$$\mathbb{P}\left[\Pi_{nm} = (A_i, B_i)_i\right] = \mathbb{P}\left[\Pi_{nm} = (\sigma(A_i), \tau(B_i))_i\right].$$

A sequence $\Pi = (\Pi_{nm})_{n,m \geqslant 1}$ is partially exchangeable if it is partially exchangeable for every (n, m).

It is worth noting, that Pitman [1995] used the notion of "partially exchangeable random partition", but it referred to a univariate partition, and by partial exchangeability he assumed that one can permute elements in each class but cannot permute elements between classes. Our partially exchangeable random bi-partition is different from his definition.

Proposition 2.2.6. *A partially exchangeable sequence $(X_1, X_2, ..., Y_1, Y_2, ...)$ defines a partially exchangeable random bi-partition via the rule defined in definition 2.2.1.*

Proof. Follows trivially from the definition. □

The following proposition gives an analogue of Kingman's representation (Proposition 1.3.5 in Chapter 1).

Proposition 2.2.7. *For every partially exchangeable random bi-partition Π of $N \times N$, there exists a partially exchangeable sequence (not necessarily unique), $(X_1, X_2, ..., Y_1, Y_2, ...)$, with directing random measures $\left(F^X, F^Y\right)$, such that $\Pi \overset{d}{=} \Pi\left(X_1, X_2, ..., Y_1, Y_2, ...\right)$. Moreover, the ranked sequence of* *limiting sizes of A-biclasses of Π_{nm} corresponds to the ranked sequence of sizes of the atoms of F^X, and the ranked sequence of limiting sizes of B-biclasses of Π_{nm} corresponds to the ranked sequence of sizes of the atoms of F^Y.*

Proof. Suppose for some (n, m) we have a partially exchangeable random bi-partition $\Pi_{nm} = \{(A_1, B_1), ..., (A_k, B_k)\}$. A bi-partition can be generated from a sequence by combining together elements that have the same values. Thus, the opposite procedure is to create a label for each bi-class.

Let $\xi_1, \xi_2, ... \overset{i.i.d.}{\sim} U[0, 1]$ independent of Π_{nm}. Then the values of ξ_i are $\mathbb{P}$-a.s. distinct. Define $C_i^X = r$ if $i \in A_r$, for $i = 1, ..., n$ and analogously define $C_j^Y = l$ if $j \in B_l$, for $j = 1, ..., m$. Now, define $X_i = \xi_{C_i^X}$ and $Y_j = \xi_{C_j^Y}$. Due to the construction, $X_{1:n}$ and $Y_{1:m}$ are partially exchangeable, as they are functions of a partially exchangeable bi-partition that is independent of $(\xi_1, \xi_2, ...)$.

Since $X_{1:n}$ and $Y_{1:m}$ are partially exchangeable, there exists a vector of directing random measures (F^X, F^Y). The rest of the proof follows from Kingman's representation (Proposition 1.3.5) applied to the two sequences separately.

□

In particular, Kingman's representation characterizes a random partition as a mixture of paintbox processes. A similar fact follows from the proposition above, but a bivariate paintbox process should be defined.

Definition 2.2.8. *If $(X_1, X_2, \dots) \overset{i.i.d}{\sim} \mu$ independent of $(Y_1, Y_2, \dots) \overset{i.i.d}{\sim} \mu$, then a bi-partition $\Pi = \Pi(X_1, X_2, ..., Y_1, Y_2, ...)$ is called a bivariate paintbox process.*

The following is a consequence of Proposition 2.2.7.

Corollary 2.2.9. *A partially exchangeable random bi-partition is a mixture of bivariate paintbox processes.*

2.2.2 Random bivariate partitions in order of appearance

Recall that the order of appearance defined on the classes of a random partition allows us to characterize the distribution of an exchangeable random partition by an EPPF, i.e., a symmetric function $\mathbf{p}$ (Definition 1.3.7). In order to extend this characterization to the bivariate case, a notion of bivariate order is needed. However, it is not trivial, as there is no natural order on pairs.

One can use the order of appearance, or "historical" order, but then "the history" is required to be known. Knowing $(X_{1:n}, Y_{1:m})$ we cannot say which was the "historical" order of appearance of the corresponding classes. Such order can be different if we first observe $(X_{1:n})$ then $(Y_{1:m})$ or instead first $(Y_{1:m})$ then $(X_{1:n})$. Thus, some additional information is needed.

One can note that, as we are dealing with partially exchangeable bi-partitions, this order is really irrelevant. This is true, because of technical reasons, however, expressing partial exchangeability of a bi-partition through some symmetric function $\mathbf{f}$, we still need some way of coding bi-classes, which need not be the formal order. We therefore propose a way of indexing bi-classes of a bi-partition. For example, let us consider a bi-partition of $N_4 \times N_7$ with unordered classes:

$$\Pi_{4,7} = \{(1,2;\, 4), (4;\, 5,6), (3;\, 1,2,3), (\emptyset;\, 7)\}.$$

The first class observed in the first group is $(1,2;\, 4)$ and we assign an index 1 to this class, i.e., $(1,2;\, 4)_1$. We assign index 2 to the second class observed in the first group, i.e., $(3;\, 1,2,3)_2$ and index 3 to the third class, $(4;\, 5,6)_3$. Now we can add extra indices to each class according to its order of appearance in the second group. Thus, the first class observed in the second group is $(3;\, 1,2,3)_2$ and we also assign index 1 to this class, i.e., $(3;\, 1,2,3)_{2,1}$, and we do the same for the rest of the classes. If some of A_i or B_i is empty we set the index 0, thus $(\emptyset;\, 7)$ has index $(0,4)$. Then, we can rewrite the unordered partition as $\Pi_{4,7} = \{(1,2;\, 4)_{1,2}, (4;\, 5,6)_{3,3}, (3;\, 1,2,3)_{2,1}, (\emptyset;\, 7)_{0,4}\}$.

Although this way of coding is quite complicated, it allows us to specify permutations on sizes of bi-classes and to give a characterization of a partially exchangeable random bi-partition.

Definition 2.2.10. *Define* $\Pi_{nm} = \{(A_{ij}, B_{ij})_{ij}\}$ *to be a bi-partition with classes in a bivariate order of appearance if for every* (A_{ij}, B_{ij}), i *is the order of the appearance of this bi-class among non-empty A-classes* ($i = 0$ *if* $A_i = \emptyset$) *and* j *is the order of appearance of this bi-class among non-empty B-classes* ($j = 0$ *if* $B_j = \emptyset$).

Under the bi-variate order of appearance it is easy to show that a characterization of partially exchangeable random bi-partition can be given.

Proposition 2.2.11. *A random bi-partition* Π_{nm} *of* $N_n \times N_m$ *is partially exchangeable if and only if*

$$\mathbb{P}\left(\Pi_{nm} = \{(A_{ij}, B_{ij})_{ij}\}\right) = \mathbf{f}\left((n_{ij}, m_{ij})_{ij}\right) = \mathbf{f}(\boldsymbol{n}, \boldsymbol{m}),$$

for some function $\mathbf{f}$ *that is bi-symmetric in the following sense: for every pair of permutations* σ *of* $1, ..., n$ *and* τ *of* $1, ..., m$,

$$\mathbf{f}(\mathbf{n}, \mathbf{m}) = \mathbf{f}\left((n_{ij}, m_{ij})_{ij}\right) = \mathbf{f}\left((n_{\sigma(i)\tau(j)}, m_{\sigma(i)\tau(j)})_{\sigma(i)\tau(j)}\right), \qquad (2.1)$$

where $\sigma(0) = \tau(0) = 0$ *and* $(\boldsymbol{n}, \boldsymbol{m}) = (n_{ij}, m_{ij})_{ij}$ *is a sequence of sizes of bi-classes* $(A_{ij}, B_{ij})_{ij}$. *We call* $\mathbf{f}$ *a bivariate EPPF.*

Formally, the function $\mathbf{f}$ in (2.1) should be called a Partially Exchangeable Bivariate Partition Probability Function (PE-BPPF) but for simplicity we will call it a bivariate EPPF.

Denote by $(\mathbf{n}^{\mathbf{i}+}, \mathbf{m})$ the updated vector of sizes of bi-classes after increasing by one the class observed on the i-th place in the first group, i.e. $n_{ij} + 1$. Analogously, $(\mathbf{n}, \mathbf{m}^{\mathbf{j}+})$ denotes the increasing $m_{ij} + 1$. Let $K_{n,m}$ be the number of bi-classes in $(\mathbf{n}, \mathbf{m})$. From the definition above and the addition rule of probability, a bivariate EPPF must satisfy: $\mathbf{f}(0, 1) = \mathbf{f}(1, 0) = 1$ and

$$\mathbf{f}(\mathbf{n}, \mathbf{m}) = \sum_{i=1}^{K_{n,m}+1} \mathbf{f}(\mathbf{n}^{\mathbf{i}+}, \mathbf{m}) = \sum_{i=1}^{K_{n,m}+1} \mathbf{f}(\mathbf{n}, \mathbf{m}^{\mathbf{j}+}).$$

2.3 Bivariate species sampling models

After defining the concept of bi-partitions we can return to species sampling problems. Recall that a species sampling sequence was defined as an exchangeable sequence having a predictive rule of the form (1.1). We are using a predictive approach, and so bivariate prediction rule that allows sharing values should be specified. Without loss of generality we will assume that the first observed value is X_1. Then, let us consider a bivariate prediction rule that satisfies the following constraint

$$\begin{aligned} &\mathbb{P}\left(X_1 \in \cdot\right) = \nu(\cdot), \;\; \mathbb{P}\left(Y_1 \in \cdot\right) = \nu(\cdot), \text{ and for } n \geqslant 1, \\ &F^X_{nn}(\cdot) = \mathbb{P}\left(X_{n+1} \in \cdot \mid X_{1:n}, Y_{1:n}, K_{nn} = k\right) = \\ &\qquad = \sum_{ij} p_{ij}(\mathbf{n},\mathbf{n})\delta_{Z_{ij}}(\cdot) + p_{k+1,0}(\mathbf{n},\mathbf{n})\nu(\cdot), \\ &F^Y_{nn}(\cdot) = \mathbb{P}\left(Y_{n+1} \in \cdot \mid X_{1:n}, Y_{1:n}, K_{nn} = k\right) = \\ &\qquad = \sum_{ij} p'_{ij}(\mathbf{n},\mathbf{n})\delta_{Z_{ij}}(\cdot) + p'_{0,k+1}(\mathbf{n},\mathbf{n})\nu(\cdot), \end{aligned} \tag{2.2}$$

where $K_{n,n}$ is the number of distinct species $(Z_{ij})_{ij}$ observed in the bivariate sample $(X_{1:n}, Y_{1:n})$, $(\mathbf{n},\mathbf{n}) = \left(n_{ij}, n'_{ij}\right)_{ij}$ is a vector of sizes of bi-classes of the bivariate partition $\Pi(X_{1:n}, Y_{1:n})$ with classes in bivariate order of appearance, and ν is a diffuse distribution. The functions $(p_{ij}, p'_{ij})_{ij}$ and $p_{k+1,0}$, $p'_{0,k+1}$ will be called *bivariate predictive weights* and should be understood in the following way: given the sample $(X_{1:n}, Y_{1:n})$, the next observation X_{n+1} can be either as the ij-th observed species with probability $p_{ij}(\mathbf{n},\mathbf{n})$, or corresponds to a new class with probability $p_{k+1,0}(\mathbf{n},\mathbf{n})$. Similarly, for Y_{n+1}. As before the assumption that labels for new classes are generated from a diffuse distribution is needed in order to guarantee that labels are $\mathbb{P}$-a.s. distinct. Note also, that in general the prediction rule that satisfies (2.2), does not generate a partially exchangeable sequence.

Definition 2.3.1. *We say that $(X_1, X_2, ...; Y_1, Y_2, ...)$ is a bivariate species sampling sequence if $(X_1, X_2, ...; Y_1, Y_2, ...)$ is a partially exchangeable sequence with prediction rule of the form* (2.2)*.*

Theorem 2.3.2. *If $(X_1, X_2, \dots, Y_1, Y_2, ...)$ is a partially exchangeable sequence with a prediction rule satisfying (2.2), then*

- *The sequence of predictive distributions $\left(F_n^X, F_n^Y\right)_{n\geqslant 1}$ converges $\mathbb{P}$-a.s. in total variation norm to the directing random measures (F^X, F^Y) of $(X_1, X_2, \dots, Y_1, Y_2, ...)$, furthermore*

$$\begin{aligned} F^X(\cdot) &= \sum_{ij} \tilde{P}_{ij}\delta_{Z_{ij}}(\cdot) + (1 - \sum_{i} \tilde{P}_{ij})\nu(\cdot), \\ F^Y(\cdot) &= \sum_{ij} \tilde{R}_{ij}\delta_{Z_{ij}}(\cdot) + (1 - \sum_{j} \tilde{R}_{ij})\nu(\cdot), \end{aligned} \tag{2.3}$$

  and $\tilde{P}_{ij}$ is the relative frequency of the i-th species to appear in the first sequence, i.e., $\tilde{P}_{ij} = \lim_{n\to\infty} \frac{n_{ij}}{n}$ $\mathbb{P}$-a.s.; $\tilde{R}_{ij}$ is the relative frequency of the j-th species to appear in the second sequence, i.e., $\tilde{R}_{ij} = \lim_{n\to\infty} \frac{n'_{ij}}{n}$ $\mathbb{P}$-a.s.; $(Z_{ij})_{ij} \overset{i.i.d.}{\sim} \nu$ independently of $(\tilde{P}_{ij}, \tilde{R}_{ij})_{ij}$;

- *(X_n, Y_n) is a sample from $\left(F^X, F^Y\right)$.*

Proof. The result follows from partial exchangeability and the theory of partially exchangeable bivariate random partitions. □

The vector $\left(F^X, F^Y\right)$ defined in (2.3) will be called a *bivariate species sampling model.*

Theorem 2.3.3. *For every pair of bivariate EPPF $\mathbf{f}$, and diffuse probability distribution ν, there is a unique distribution for a bivariate species sampling sequence $(X_1, X_2, \dots, Y_1, Y_2, ...)$ such that $\mathbf{f}$ is a bivariate EPPF of $(X_1, X_2, \dots, Y_1, Y_2, ...)$ and ν is the distribution of the first elements in the group.*

Proof. For each n, the bivariate EPPF specifies the probability of the partition induced by $(X_{1:n}, Y_{1:n})$, while the set of labels $(Z_1, Z_2, ...)$ is generated independently from the diffuse distribution ν. Thus, the joint distribution of $(X_{1:n}, Y_{1:n})$ is determined for every n by $(\mathbf{f}, \nu)$, which proves the uniqueness. Given a pair $(\mathbf{f}, \nu)$, we can construct a sequence $(X_{1:n}, Y_{1:n})$ by assigning for the bi-classes of the bi-partition defined by $\mathbf{f}$, the i.i.d. (ν) values $(Z_1, Z_2, ...)$. □

Theorem 2.3.4. *Given a diffuse probability distribution ν and a sequence of bivariate predictive weights, let $(X_1, X_2, \dots, Y_1, Y_2, \dots)$ has the prediction rule that satisfies* (2.2)*. Then $(X_1, X_2, \dots, Y_1, Y_2, \dots)$ is partially exchangeable if and only if there exists a non-negative, bi-symmetric function $\mathbf{f}$ such that, for $1 \geqslant i \geq K_{nn} + 1$,*

$$\begin{aligned} p_{ij}(\mathbf{n}, \mathbf{n}) &= \frac{\mathbf{f}(\mathbf{n}^{\mathbf{i}+}, \mathbf{n})}{\mathbf{f}(\mathbf{n}, \mathbf{n})} \text{ and } p_{k+1,0}(\mathbf{n}, \mathbf{n}) = \frac{\mathbf{f}\left((\mathbf{n}, \mathbf{n}), (1, 0)\right)}{\mathbf{f}(\mathbf{n}, \mathbf{n})}, \\ p'_{ij}(\mathbf{n}, \mathbf{n}) &= \frac{\mathbf{f}(\mathbf{n}, \mathbf{n}^{\mathbf{j}+})}{\mathbf{f}(\mathbf{n}, \mathbf{n})} \text{ and } p'_{0,k+1}(\mathbf{n}, \mathbf{n}) = \frac{\mathbf{f}\left((\mathbf{n}, \mathbf{n}), (0, 1)\right)}{\mathbf{f}(\mathbf{n}, \mathbf{n})}. \end{aligned} \tag{2.4}$$

Then, the bivariate EPPF of $(X_1, X_2, \dots, Y_1, Y_2, \dots)$ is the unique non-negative bi-symmetric function $\mathbf{f}$, such that the condition (2.4) *holds and $\mathbf{f}(0, 1) = \mathbf{f}(1, 0) = 1$.*

Proof. The result follows from Theorem 2.3.2 and Theorem 2.3.3. □

One can think about bivariate species sampling models as a general class of two dependent discrete random measures. Many models proposed in the literature (we briefly discuss them at the end of Chapter 1) belong to this class. At the same time, characterizing a bivariate prediction rule that generates a partially exchangeable sample from the bivariate species sampling model is not always possible. For example, for the Dependent Dirichlet Processes introduced by MacEachern [1999] or the Spatial Normalized Gamma Process due to Rao and Teh [2009], the bivariate prediction rules are not specified.

Some models allow a predictive characterization but only conditionally on a set of latent variables. After integrating them out, one can get the bivariate prediction rule. Examples of such models are the Bivariate Dirichlet Process of Walker and Muliere [2003], the Hierarchical Dirichlet Process of Teh et al., the dependent normalized complete random measures of Leisen and Lijoi [2011], the single-ϵ model that was originally proposed by Müller et al. [2004] and later developed by Kolossiatis et al. [2013], among others.

Another example of bivariate species sampling model is a vector of two-parameter Poisson-Dirichlet processes proposed by Leisen and Lijoi

[2011]. The distribution of the corresponding bivariate random partition is computed, but the distinguishing characteristic of this model is the fact that such function is bi-symmetric (in the sense defined in Proposition 2.2.11) only if there are no shared values. At the same time, the function is quite complicated and no explicit prediction rules are specified.

Although a lot of examples of bivariate species sampling models can be found in the literature, the corresponding bivariate prediction rules are quite complicated and intractable. To overcome this problem, instead of defining dependence on the level of random measures, let us think about tractable bivariate prediction rule.

Example 1

Two independent Poisson Dirichlet Processes constitute a trivial example:

$$\mathbb{P}\left(X_{n+1} \in \cdot \mid X_1, ..., X_n\right) = \sum_{i=1}^{k} \frac{n_i}{n+\theta_1}\delta_{X_i^\star}(\cdot) + \frac{\theta_1}{n+\theta_1}\nu(\cdot),$$

$$\mathbb{P}\left(Y_{n+1} \in \cdot \mid Y_1, ..., Y_n\right) = \sum_{j=1}^{r} \frac{n'_j}{n+\theta_2}\delta_{Y_j^\star}(\cdot) + \frac{\theta_2}{n+\theta_2}\nu(\cdot).$$

It is clear that a sequence generated by such prediction rule will be partially exchangeable but the structure is too simple.

Example 2

Thinking about species sampling from two populations, it seems natural to assume that the probability of the next observation to be of the i-th observed type is proportional to some function of n_i and n'_i, that is, the numbers of such observations appearing previously in the first and the second groups. Thus, it is natural to think about a prediction rule that satisfy the following constraint:

$$\begin{aligned}\mathbb{P}\left(X_{n+1} \in \cdot \mid X_{1:n}, Y_{1:n}\right) = & \sum_{ij} \frac{g(n_{ij}, n'_{ij})}{\sum_{ij} g(n_{ij}, n'_{ij}) + \theta_1} \delta_{Z_{ij}}(\cdot) + \\ & + \frac{\theta_1}{\sum_{ij} g(n_{ij}, n'_{ij}) + \theta_1} \nu(\cdot), \\ \mathbb{P}\left(Y_{n+1} \in \cdot \mid X_{1:n}, Y_{1:n}\right) = & \sum_{ij} \frac{g'(n_{ij}, n'_{ij})}{\sum_{ij} g'(n_{ij}, n'_{ij}) + \theta_2} \delta_{Z_{ij}}(\cdot) + \\ & + \frac{\theta_2}{\sum_{ij} g'(n_{ij}, n'_{ij}) + \theta_2} \nu(\cdot),\end{aligned}$$

for some functions g and g'. We call such rule a *natural bivariate prediction rule*. We know that a bivariate EPPF specifies a set of bivariate predictive weights in a unique way according to (2.4), but the converse is not true. In order to determine whether a sample from a natural bivariate prediction rule is partially exchangeable, we analyze when a sequence of bivariate predictive weights defines a bivariate EPPF.

2.4 When do bivariate predictive weights define a bivariate EPPF?

Let us give a brief review of the matter. In the 1920's, the Cambridge philosopher William Ernest Johnson discovered a characterization of the Dirichlet Distribution that was later called ”Johnson's sufficiency postulate” by Good [1965]. According to this postulate, the conditional probability of observing the i-th observed species only depends on the number of observations of the same type observed before. Similar characterization of the Dirichlet Prior are provided by Regazzini [1978] and Lo [1991]. This form of predictive is a natural way of thinking nonparametrically.

Recently, Lee et al. [2013] investigated the nature of predictive weights in species sampling problems and gave necessary and sufficient conditions for predictive weights to define an EPPF. Roughly speaking, in order to guarantee the existence of the EPPF, the predictive functions should define such EPPF in a unique way, and some condition of symmetry should be satisfied. More precisely, they noted that the an expression $\mathbf{p}(\mathbf{n^{j+}}) = p_j(\mathbf{n})\mathbf{p}(\mathbf{n})$ does not define an EPPF, $\mathbf{p}$, in a unique way. If, for example, in the univariate case, we consider partitions of N_3 with sizes of classes given by $\mathbf{n} = (2,1)$, then there are two possible partitions with such class sizes, $\{(1,3)(2)\}$ and $\{(1,2)(3)\}$. The EPPF can then be equal either to $p_2(1)p_1(1,1)$ or to $p_1(1)p_2(2)$. Thus, the first condition on the predictive weights is a condition of uniqueness, $p_i(\mathbf{n})p_j(\mathbf{n^{i+}}) = p_j(\mathbf{n})p_i(\mathbf{n^{j+}})$. But this condition is not enough for the existence of a bivariate EPPF; another necessary condition needed is the symmetry of the function $\mathbf{p}$.

In the bivariate case, the bivariate EPPF can be also defined sequentially. Suppose, we have a sequence of bivariate predictive weights $(p_{ij}, p'_{ij})_{ij}$, such that $p_{ij} > 0$, $p'_{ij} > 0$, and $\sum_{i=1}^{k+1} p_{ij} = 1$, $\sum_{j=1}^{k+1} p'_{ij} = 1$. For simplicity of writing, we skip the indexes ij that correspond to the bivariate order of appearance and use indexes $i = 1, ..., k+1$. It is possible to rewrite everything formally using indexes ij and the same results will follow. Thus, suppose we have a sequence of bivariate predictive weights $(p_i, p'_i)_{i=1}^{k+1}$ such that for $i = 1, ..., k+1$, $p_i > 0$, $p'_i > 0$, and $\sum_{i=1}^{k+1} p_i = 1$, $\sum_{j=1}^{k+1} p'_j = 1$. Then, the bivariate EPPF can be defined sequentially by

$$\begin{aligned} &\mathbf{f}(1,0) = \mathbf{f}(0,1) = 1, \\ &\mathbf{f}(\mathbf{n^{i+}}, \mathbf{n}) = \mathbf{f}(\mathbf{n}, \mathbf{n}) p_i(\mathbf{n}, \mathbf{n}), \\ &\mathbf{f}(\mathbf{n}, \mathbf{n^{j+}}) = \mathbf{f}(\mathbf{n}, \mathbf{n}) p'_j(\mathbf{n}, \mathbf{n}), \end{aligned} \tag{2.5}$$

for all bi-partitions $(\mathbf{n}, \mathbf{n})$ and $i, j = 1...k+1$. This formula will define $\mathbf{f}$ in a unique way if and only if

1. The probability of adding an observation of the i-th type in the first group and then adding an observation of the j-th type in the first group is equal to the probability of adding an observation of the j-th type in the first group and then adding an observation of the i-th type in the first group.

2. The probability of adding an observation of the i-th type in the second group and then adding an observation of the j-th type in the

second group is equal to the probability of adding an observation of the j-th type in the second group and then adding an observation of the i-th type in the second group.

3. The probability of adding an observation of the i-th type in the first group and then adding an observation of the j-th type in the second group is equal to the probability of adding an observation of the j-th type in the first group and then adding an observation of the i-th type in the second group.

Formally, these conditions are formulated in the following lemma.

Lemma 2.4.1. *Given the bivariate predictive weights* $(p_i, p'_i)_{i=1}^{k+1}$*, the bivariate EPPF* $\mathbf{f}$ *in* (2.5) *is uniquely defined if and only if*

1. $p_i(\mathbf{n}, \mathbf{n})p_j(\mathbf{n}^{\mathbf{i}+}, \mathbf{n}) = p_j(\mathbf{n}, \mathbf{n})p_i(\mathbf{n}^{\mathbf{j}+}, \mathbf{n})$*;*
2. $p'_i(\mathbf{n}, \mathbf{n})p'_j(\mathbf{n}^{\mathbf{i}+}, \mathbf{n}) = p'_j(\mathbf{n}, \mathbf{n})p'_i(\mathbf{n}^{\mathbf{j}+}, \mathbf{n})$*;*
3. $p_i(\mathbf{n}, \mathbf{n})p'_j(\mathbf{n}^{\mathbf{i}+}, \mathbf{n}) = p'_j(\mathbf{n}, \mathbf{n})p_i(\mathbf{n}, \mathbf{n}^{\mathbf{i}+})$*.*

Proof. Suppose we have two partitions with the same vector of sizes of bi-classes. Then, each partition can be obtained from another one by a series of permutations of two neighboring observations. The condition of the lemma guarantees uniqueness under permutations of two neighboring observations. □

Assume, that in species sampling from two populations, the probability that the next observation is of the i-th observed species is proportional to some function of the frequencies of such species in a bivariate sample.

Theorem 2.4.2. *Suppose bivariate predictive weights* $(p_i, p'_i)_{i=1}^{k+1}$ *satisfy conditions (1-3) of Lemma 2.4.1 and*

$$p_i(\mathbf{n}, \mathbf{n}) \propto \begin{cases} g(n_i, n'_i), \ i = 1 \ldots k; \\ \theta_1, \ i = k+1; \end{cases}$$

$$p'_i(\mathbf{n}, \mathbf{n}) \propto \begin{cases} g'(n_i, n'_i), \ i = 1 \ldots k; \\ \theta_2, \ i = k+1. \end{cases}$$

Then $\theta_1 = \theta_2$ *and* $g(n_i, n_i') = g'(n_i, n_i') = a(n_i + n_i')$.

Proof. Let us check the 1-st condition of Lemma 2.4.1:

$$\frac{g(n_i, n_i')}{\sum_{l=1}^k g(n_l, n_l') + \theta_1} \frac{g(n_j, n_j')}{\sum_{l \neq i} g(n_l, n_l') + g(n_i + 1, n_i') + \theta_1} =$$
$$= \frac{g(n_j, n_j')}{\sum_{l=1}^k g(n_l, n_l') + \theta_1} \frac{g(n_i, n_i')}{\sum_{l \neq j} g(n_l, n_l') + g(n_j + 1, n_j') + \theta_1}.$$

This equality holds if and only if

$$g(n_i + 1, n_i') - g(n_i, n_i') = g(n_j + 1, n_j') - g(n_j, n_j').$$

At the same time

$$\frac{g(n_i, n_i')}{\sum_{l=1}^k g(n_l, n_l') + \theta_1} \frac{\theta_1}{\sum_{l=1}^k g(n_l, n_l') + \theta_1} =$$
$$= \frac{\theta_1}{\sum_{l=1}^k g(n_l, n_l') + \theta_1} \frac{g(n_i, n_i')}{\sum_{l=1}^k g(n_l, n_l') + g(1, 0) + \theta_1}.$$

This equality will hold if and only if

$$g(n_i + 1, n_i') - g(n_i, n_i') = g(1, 0).$$

This means that $g(n_i, n_i') = g(1, 0)n_i + h(n_i')$ for some function $h(n_i')$. Analyzing in the same way the 2nd condition of Lemma 2.4.1 we get that $g'(n_i, n_i') = g'(0, 1)n_i + h'(n_i')$ for some function $h'(n_i')$. Let us consider now the 3rd condition of the Lemma. It means

$$\frac{g(n_i, n_i')}{\sum_{l=1}^k g(n_l, n_l') + \theta_1} \frac{\theta_2}{\sum_{l \neq i} g'(n_l, n_l') + g'(n_i + 1, n_i') + \theta_2} =$$
$$= \frac{\theta_2}{\sum_{l \neq i} g'(n_l, n_l') + \theta_2} \frac{g(n_i, n_i')}{\sum_{l=1}^k g(n_l, n_l') + g(0, 1) + \theta_1},$$

which implies

$$g'(n_i + 1, n_i') - g\prime(n_i, n_i') = g'(1, 0) \frac{\sum_{l=1}^k g'(n_l, n_l') + \theta_2}{\sum_{l=1}^k g(n_l, n_l') + \theta_1}.$$

Analogously

$$g(n_i+1, n_i') - g(n_i, n_i') = g(0,1)\frac{\sum_{l=1}^{k} g(n_l, n_l') + \theta_1}{\sum_{l=1}^{k} g'(n_l, n_l') + \theta_2}.$$

Since we want the functions g and g' to not depend on the partition, then $(\sum_{l=1}^{k} g'(n_l, n_l') + \theta_2)$ should be equal to $(\sum_{l=1}^{k} g(n_l, n_l') + \theta_1)$ for any n and n. This means that $g(n_i, n_i') = g'(n_i, n_i')$ for any n_i, n_i', and $\theta_1 = \theta_2 = \theta$. Then

$$g(n_i+1, n_i') - g(n_i, n_i') = g(0,1) \text{ and } g'(n_i+1, n_i') - g'(n_i, n_i') = g'(1,0).$$

Combining these results with the previous, we obtain that

$$g(n_i, n_i') = g'(n_i, n_i') = an_i + bn_i',$$

where $a = g(0,1) = g(0,1) = a$ and $g'(1,0) = g'(0,1) = b$. Using this information let us check the last part of the 3rd condition:

$$\frac{an_i + bn_i'}{an + bn' + \theta}\frac{an_j + bn_j'}{a(n+1) + bn' + \theta} = \frac{an_j + bn_j'}{an + bn' + \theta}\frac{an_i + bn_i'}{an + b(n'+1) + \theta},$$

thus $a = b$. And $g(n_i, n_i') = g'(n_i, n_i') = a(n_i + n_i')$. The proof is completed. □

It is easy to show that the bivariate prediction rule with bivariate predictive weights of the form $p_i(\mathbf{n}, \mathbf{n}) = a(n_i + n_i')$ and $g_i(\mathbf{n}, \mathbf{n}) = a(n_i + n_i')$ generates a partially exchangeable sequence if and only if $a = 1$.

We see that conditions for having both, a tractable, manageable prediction rule and partial exchangeability, are quite restrictive. Indeed, if we assume that the probability of the next observation belonging to the i-th observed species is proportional to a function of the frequency of such species in the bivariate sample, then the model reduces to a Dirichlet Process for a homogeneous population of species.

Chapter 3

Partially conditionally identically distributed sequences

3.1 Introduction and motivation

In Chapter 2 we discussed bivariate species sampling processes for a non homogeneous region of interest. These processes were assumed to be partially exchangeable which, in turn, implies stationarity. However, in some applications the assumption of stationarity can be too restrictive, for example, if species are sampled from a population subject to some perturbation that suppresses the stationarity before returning to some equilibrium. What model could we use then? Kallenberg [1988] noted that stationarity and conditional identity in distribution, together, imply exchangeability. Berti et al. [2004] specified and widely investigated conditional identity in distribution as a type of dependence that generalizes the notion of exchangeability. Indeed, if $(X_n)_{n\geqslant 1} = (X_1, X_2, ...)$ is a stationary and conditionally identically distributed sequence, then the joint distribution of $(X_k)_{k\geqslant n}$ does not depend on n, and converges weakly to an exchangeable law. The concept of conditional identity in distribution is a generalization of exchangeability which preserves some properties, namely, asymptotic exchangeability and the fact the sequences of predictive and empirical measures converge to the same limit.

The aim of this chapter is to investigate an extension of partial exchangeability in the same sense that conditional identity in distribution generalizes exchangeability. We start by reminding the basic ideas and main properties of conditional identity in distribution and then propose the novel concept of partial conditional identity in distribution. Combining the new type of dependence with bivariate prediction rules allows us to construct a generalization of bivariate species sampling models, which may have a tractable prediction rule, preserve most of the properties of bivariate species sampling models and can be used in cases where the assumption of partial exchangeability fails.

3.2 Conditionally identically distributed sequences (c.i.d.)

Berti et al. [2004] consider a sequence of random variables such that, given the past, all future elements are conditionally identically distributed. Formally, let $(X_n)_{n\geqslant 1}$ be defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with each X_n taking values in a complete and separable (Polish) metric space $\mathbb{X}$ endowed with the Borel σ-algebra $\mathcal{X}$. Assume the sequence $(X_n)_{n\geqslant 1}$ is adopted to a filtration $\mathcal{G} = (\mathcal{G}_n)_{n\geqslant 0}$.

Definition 3.2.1. *A sequence $(X_n)_{n\geqslant 1}$ is called conditionally identically distributed with respect to $\mathcal{G}$, abbreviated as $\mathcal{G}$-c.i.d., whenever*

$$\mathbb{E}\left[f(X_k)|\mathcal{G}_n\right] = \mathbb{E}\left[f(X_{n+1})|\mathcal{G}_n\right] \quad \mathbb{P}\text{-}a.s.\ , \tag{3.1}$$

for all $k > n \geqslant 0$ and all bounded measurable $f : \mathbb{X} \to \mathbb{R}$.

This means that, for each $n \geqslant 0$, all future observations $(X_k)_{k>n}$ are identically distributed given the past $\mathcal{G}_n$. In particular, the single random variables X_n are identically distributed.

Condition (3.1) is equivalent to the following:

$$(\mathbb{E}\left[f(X_{n+1})|\mathcal{G}_n\right])_{n\geqslant 0} \text{ is a } \mathcal{G}\text{-martingale,} \tag{3.2}$$

for every bounded measurable $f : \mathbb{X} \to \mathbb{R}$. The advantage of condition (3.2) is that it highlights the martingale-behavior of a predictive measure and it is generally easier to verify. Other ways to reformulate (3.1), in particular, in terms of stopping times, can be found in Berti et al. [2004].

A particular case of $\mathcal{G}$-c.i.d. is obtained when $\mathcal{G}$ is the natural filtration, i.e., $\mathcal{G} = \mathcal{G}^X$, where $\mathcal{G}_0^X = \{\emptyset, \Omega\}$ and $\mathcal{G}_n^X = \sigma(X_1, ..., X_n)$. In this case, the filtration is omitted from the notation and $(X_n)_{n\geqslant 1}$ is simply called *c.i.d.* or *natural c.i.d.* Clearly, if $(X_n)_{n\geqslant 1}$ is $\mathcal{G}$-c.i.d., then it is also c.i.d. In the special case of a natural filtration, conditions (3.1) and (3.2) above reduce to

$$(X_1, ..., X_n, X_{n+2}) \overset{d}{=} (X_1, ..., X_n, X_{n+1}), \ n \in Z_+. \tag{3.3}$$

An exchangeable sequence satisfies (3.3) so it is c.i.d., but a c.i.d. sequence is not necessarily exchangeable. The lost property is stationarity. Recall that a sequence $(X_n)_{n\geqslant 1}$ is called *stationary* if $(X_1, X_2, ...) \overset{d}{=} (X_2, X_3, ...)$. Then one can informally say that the 'natural c.i.d.' property is equivalent to exchangeability without stationarity.

Proposition 3.2.2. *(Kallenberg [1988]) If $(X_n)_{n\geqslant 1}$ is a stationary sequence of random variables satisfying (3.3), then $(X_n)_{n\geqslant 1}$ is exchangeable.*

In particular, this means that order is important for c.i.d. sequences. This is the first main distinction between the concepts of c.i.d. and exchangeability. Before discussing another link between exchangeability and c.i.d. remind that, according to Kingman,

Definition 3.2.3. *A sequence of random variables $(X_n)_{n\geqslant 1}$ is called asymptotically exchangeable if*

$$(X_{n+1}, X_{n+2}, ...) \overset{d}{\to} (Z_1, Z_2, ...) \ \textit{as } n \to \infty,$$

for some exchangeable sequence $(Z_n)_{n\geqslant 1}$.

Then, the following property comes from the results of Aldous [1985].

Proposition 3.2.4. *(Berti et al. [2004]) Any $\mathcal{G}$-c.i.d. sequence is asymptotically exchangeable.*

Roughly speaking, even if we remove stationarity from exchangeability, the sequence is approximately exchangeable for large n. The nature of this phenomenon is based on the martingale condition (3.2). Let $\mathcal{H}$ be the class of measurable functions $f : \mathbb{X} \to \mathbb{R}$ such that $\mathbb{E}\left[|\, f(X_1)\,|\right] < \infty$. Then, by Doob's convergence theorem, it can be shown (Lemma 2.1 in Berti et al. [2004]) that for each $f \in \mathcal{H}$ and any $\mathcal{G}$-c.i.d. sequence there exists an integrable random variable V_f, such that

$$\begin{aligned}&\mathbb{E}\left[f(X_{n+1})|\mathcal{G}_n\right] \to V_f \quad \mathbb{P}\text{-a.s. and in } \; L^1,\\ &\mathbb{E}\left[V_f|\mathcal{G}_n\right] = \mathbb{E}\left[f(X_{n+1})|\mathcal{G}_n\right] \;\; \mathbb{P}\text{-a.s. for every } n \geqslant 0.\end{aligned}$$

Moreover, from this result and Lemma 2.4 in Berti et al. [2004], follows the existence of a random measure α on $\mathbb{X}$ such that, for all bounded continuous functions f on $\mathbb{X}$,

$$V_f = \int f(t)\alpha(dt) \;\; \mathbb{P}\text{-a.s.}$$

If we consider the case when f is an indicator and $\mathcal{G}$ is the natural filtration, these results mean that, for any c.i.d. sequence $(X_n)_{n\geqslant 1}$, there exists a random measure α on $\mathbb{X}$ such that

$$\mathbb{P}\left(X_{n+1} \in \cdot|X_1, ..., X_n\right) \overset{weakly}{\to} \alpha(\cdot) \;\; \mathbb{P}\text{-a.s.},$$

and the exchangeable limit law is directed by α. Following the usual terminology of Aldous [1985], Berti et al. [2004] call this measure the *directing random measure* of $(X_n)_{n\geqslant 1}$. Knowing that α exists, one can construct an exchangeable sequence $(Z_n)_{n\geqslant 1}$, which is directed by α and thus asymptotically exchangeable.

Recall that exchangeable sequences also have by the nice property that their empirical and predictive measures converge to the same limit. Berti et al. [2004] (Theorem 2.2) prove that a Strong Law of Large Numbers holds for natural c.i.d. sequences. In particular, they showed the following.

Theorem 3.2.5. *Let $(X_n)_{n\geqslant 1}$ be c.i.d., then $\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}(\cdot) \to \alpha(\cdot)$ weakly $\mathbb{P}$-a.s.*

So, for a c.i.d. sequence, the predictive and empirical measures have the same limit, as in the case for an exchangeable sequence. It seems encouraging that the concept of c.i.d. shares so many properties with the concept of exchangeability. A natural question is whether there exists some analogues of the representation theorem. However, the random measure α exists only in the limit and it is difficult to characterize such limit explicitly. Thus, no useful representation theorems are available for c.i.d. sequences and this is the second main distinction between the notions of c.i.d. and exchangeability.

The following theorem provides a characterization of exchangeability in terms of conditional identity in distribution.

Theorem 3.2.6. *(Berti et al. [2004], Theorem 2.8) The following statements are equivalent:*

1. *$(X_n)_{n\geqslant 1}$ is exchangeable;*

2. *For any (finite) permutation τ of $\{1, 2, ...\}$, $\left(X_{\tau(1)}, X_{\tau(2)}, ...\right)$ is c.i.d.*

An example of a $\mathcal{G}$-c.i.d sequence, modified Polya urns, was given by Berti et al. [2004]. Suppose one has an urn that initially contains $w > 0$ white, and $r > 0$ red balls. Suppose also, that at each time $n \geqslant 1$, someone draws a ball and then replaces it together with d_n balls of the same color. If X_n is the indicator that a white ball is drawn at time n, then

$$\mathbb{E}\left(X_1\right) = \frac{w}{w+r}, \text{ and for } n \geqslant 1,$$

$$\mathbb{E}\left(X_{n+1} | X_{1:n}, d_{1:n}\right) = \frac{w + \sum_{i=1}^{n} d_i X_i}{w + r + \sum_{i=1}^{n} d_i} \;\; \mathbb{P}\text{-a.s. for all}\, n \geqslant 1,$$

where $(d_n)_{n\geqslant 1}$ is a sequence of random variables such that d_n is independent of $(X_i, d_j :\; i \leqslant n,\, j < n)$ for all $n \geqslant 1$.

While the sequence $(X_n)_{n\geqslant 1}$ generated from a modified Polya urn is c.i.d., it is not exchangeable, apart from some particular cases. For example, if $d_n = d$ for all n and some fixed integer d, this urn reduces to the classical Polya urn that generates an exchangeable sequence.

3.2.1 Prediction rules that generate c.i.d. sequences

In Chapter 1 we discussed species sampling sequences (Pitman [1996a]). The name “species sampling” is based on the idea that we can think about a sequence of random variables $(X_n)_{n\geqslant 1}$ as a sample from a large population of species. Then, X_n represents the species of the n-th individual sampled. We remind some definitions for the reader’s convenience.

Definition 3.2.7. *A sequence $(X_n)_{n\geqslant 1}$ is called a species sampling sequence if it is exchangeable and has the following prediction rule:*

$$\begin{aligned} &\mathbb{P}(X_1 \in \cdot) = \nu(\cdot) \text{ and, for } n \geqslant 1, \\ &F_n(\cdot) = \mathbb{P}(X_{n+1} \in \cdot \mid X_1, ..., X_n, K_n = k) = \\ &= \sum_{i=1}^{k} p_i(\mathbf{n})\delta_{X_i^\star}(\cdot) + p_{k+1}(\mathbf{n})\nu(\cdot), \end{aligned} \tag{3.4}$$

where K_n is the number of distinct species $(X_1^\star, ..., X_k^\star)$ among the observations $(X_1, ..., X_n)$, $\mathbf{n} = (n_1, ..., n_k)$ is the random vector of counts of the distinct species and ν is some diffuse distribution.

Given $K_n = k$, the sequence of predictive weights $(p_i,\ i = 1, .., k + 1)$ should be understood in the following way: given the sample $(X_1, ..., X_n)$, the next observation X_{n+1} can be either equal to the i-th observed species, with probability $p_i(\mathbf{n})$, or it may correspond to a new class, with probability $p_{k+1}(\mathbf{n})$. A more detailed review of species sampling sequences is given in Section 1.3 of Chapter 1.

Thus, exchangeability combined with a prediction rule of the form (3.4) results in a species sampling sequence. It is natural to combine some prediction rule with the concept of c.i.d., which generalizes exchangeability. This has been done by Bassetti et al. [2010]. In particular, they defined a so called generalized Ottawa sequence, which is c.i.d. It is worth noting that this is not the only example of c.i.d. sequences proposed in the literature, other models can be found in Bassetti et al. [2010].

Definition 3.2.8. *We say that a sequence $(X_n)_{n\geqslant 1}$ is a generalized Ottawa sequence if there exists a sequence of random variables $(Z_n)_{n\geqslant 1}$, with values* *in a Polish space endowed with its Borel σ-fields $(S, \mathcal{S})$, such that for each $n \geqslant 1$*

- *The regular conditional distribution of X_{n+1} given the filtration $\mathcal{G}_n = \mathcal{G}_n^X \vee \mathcal{G}_n^Z = \sigma(X_1, ..., X_n) \vee \sigma(Z_1, ..., Z_n)$ is*

$$\mathbb{P}\left[X_{n+1} \in \cdot | \mathcal{G}_n\right] = \sum_{i=1}^{n} p_{n,i}\left(z_1, ..., z_n\right) \delta_{X_i}(\cdot) + r_n\left(z_1, ..., z_n\right) \nu(\cdot), \quad (3.5)$$

  where X_{n+1} and $(Z_{n+1}, Z_{n+2}, ...)$ are conditionally independent given $\mathcal{G}_n$;

- *The functions r_n and $(p_{n,i})_{i=1}^n$ do not depend on the partition induced by $(X_1, ..., X_n)$, are strictly positive and*

$$r_n(Z_1, ..., Z_n) \geqslant r_{n+1}(Z_1, ..., Z_{n+1})\, \mathbb{P}\text{-}a.s.; \quad (3.6)$$

- *For each $(Z_1, ..., Z_n) \in S^n$ and $i = 1, ..., n-1$, the functions $p_{n,i}$ satisfy,*

$$\begin{aligned} p_{n,i}(z_1, ..., z_n) &= \frac{r_n(z_1, ..., z_n)}{r_{n-1}(z_1, ..., z_{n-1})} p_{n-1,i}(z_1, ..., z_{n-1}), \\ p_{n,n}(z_1, ..., z_n) &= 1 - \frac{r_n(z_1, ..., z_n)}{r_{n-1}(z_1, ..., z_{n-1})}, \end{aligned} \quad (3.7)$$

  with $r_0 = 1$.

Condition (3.7) ensures that condition (3.2) holds, i.e., that the predictive measure is a martingale, while condition (3.6) assures the positiveness of the weights. Thus, under (3.6) and (3.7), it can be shown that any generalized Ottawa sequence is c.i.d. with respect to the filtration $\mathcal{G} = \left(\mathcal{G}_n^X \vee \mathcal{G}_\infty^Z\right)_{n \geqslant 0}$, where $\mathcal{G}_\infty^Z = \vee_{n \geqslant 0} \mathcal{G}_n^Z$.

An advantage of this structure is that one can construct a generalized Ottawa sequence using any sequence of decreasing positive functions $(r_n)_{n \geqslant 0}$ with values on $[0, 1]$. All other weights will be of the form (3.7). This means that a generalized Ottawa sequence is defined through the sequence of probabilities of observing a new class at each time n. Knowing the behavior of the new species appearance in some real process, one could model such processes by defining the appropriate sequence of $(r_n)_{n\geqslant 0}$.

If we compare the expressions of the rules in (3.4) and (3.5), we can appreciate where they differ. First of all, in (3.4) the sum is taken from 1 to k, the number of distinct classes. Thus, each $p_i(\mathbf{n})$ corresponds to the i-th observed class for $i = 1, ..., k$ and depends on the vector of sizes of classes of the partition Π_n generated from the sample $(X_1, ..., X_n)$. While in (3.5), the sum is taken from 1 to n, so the weight $p_{n,i}(z_1, ..., z_n)$ corresponds to the individual observation X_i. Another distinction is that these weights do not depend on the partition given $(z_1, ..., z_n)$. The connection between the predictive weights $(p_i)_{i=1}^{k+1}$ in (3.4) and the weights $(p_{n,i})_{i=1}^{n}$ in (3.5) is then:

$$p_i = \sum_{i:X_i=X_i^\star} p_{n,i}.$$

Although this may appear just a matter of notation, the distinction arises from the fact that in (3.4) the data are exchangeable, so the order does not matter and we can combine all data of the same type into groups. In (3.5) order matters and it is more convenient to split the weights. For the same reason, analyzing the random partition, Π_n, induced by a generalized Ottawa sequence $(X_1, .., X_n)$, becomes a complicated task.

Finally, a more crucial distinction is that a set of latent variables $(Z_n)_{n\geqslant 1}$ is involved in the definition of the generalized Ottawa sequence. Indeed, (3.5) is a conditional prediction rule, and if we want to compare it with the prediction rule that characterizes a species sampling sequence, we should integrate out the latent variables $(Z_n)_{n\geqslant 1}$. If $(X_n)_{n\geqslant 1}$ is a generalized Ottawa sequence, then its prediction rule is

$$\begin{aligned}\mathbb{P}\left[X_{n+1} \in \cdot | X_1, .., X_n\right] = \sum_{i=1}^{n} \mathbb{E}\left[p_{n,i}\left(Z_1, .., Z_n\right) | X_1, .., X_n\right] \delta_{X_i}(\cdot) + \\ + \mathbb{E}\left[r_n\left(Z_1, ..., Z_n\right) | X_1, ..., X_n\right] \nu(\cdot).\end{aligned} \tag{3.8}$$

Reminding that any $\mathcal{G}$-c.i.d. sequence is also c.i.d., we can conclude that any generalized Ottawa sequence is also c.i.d. and the prediction rule (3.8), obtained from (3.5) with the restrictions (3.7) and (3.6), generates a c.i.d.

sequence. Although simulations can be used, computing the predictive weights in (3.8) can be complicated, and it is not possible to specify the predictive weights only through the sequence of decreasing positive functions $(r_n)_{n\geqslant 0}$. In the following section we will consider some examples of generalized Ottawa sequences.

3.2.2 Examples of generalized Ottawa sequences

The Beta-GOS process

This process (Costa et al. [2013]) is an example of generalized Ottawa sequence, where the set of additional random variables $(Z_n)_{n\geqslant 1}$ is defined as a sequence of independent random variables with Beta distribution, i.e., $Z_n \sim Beta(\alpha_n, \beta_n)$ independently, and the weights are given by

$$p_{n,i}(Z_1, ..., Z_n) = (1 - Z_i) \textstyle\prod_{j=i+1}^{n} Z_j, \quad r_n(Z_1, ..., Z_n) = \prod_{i=1}^{n} Z_i \ .$$

As we discussed before, computing the prediction distribution requires integration with respect to the latent variables $(Z_n)_{n\geqslant 1}$, and this is analytically complicated.

Generalized Ottawa sequence without latent variables

In order to compare species sampling sequences and generalized Ottawa sequences, we considered a particular case of generalized Ottawa sequence that is free of latent variables, that is, the functions r_n and $(p_{n,i})_{i=1}^{n}$ in (3.5) are not random. The prediction rule in (3.5) then reduces to

$$\mathbb{P}\left[X_{n+1} \in \cdot | X_1, ..., X_n\right] = \sum_{i=1}^{n} p_{n,i}\delta_{X_i}(\cdot) + r_n\nu(\cdot), \tag{3.9}$$

and conditions (3.6) and (3.7) become

$$\begin{aligned}
&r_n \geqslant r_{n+1} \text{ for } n \geqslant 0,\\
&p_{n,i} = \frac{r_n}{r_{n-1}} p_{n-1,i} \text{ for } i = 1, ..., n-1,\\
&p_{n,n} = 1 - \frac{r_n}{r_{n-1}}.
\end{aligned}$$

We call such sequence a *simple generalized Ottawa sequence* and it is c.i.d. with respect to the natural filtration. It is interesting to check when a simple generalized Ottawa sequence is exchangeable and, therefore, a species sampling sequence. In the following, we show that the only process that generates an exchangeable simple generalized Ottawa sequence is the Dirichlet Process.

Proposition 3.2.9. *Let $(X_n)_{n\geqslant 1}$ be a simple generalized Ottawa sequence. Then $(X_n)_{n\geqslant 1}$ is exchangeable if and only if*

$$\begin{aligned} p_{ni} &= \frac{1}{1+\theta} \, for\, i = 1, ..., n, \\ r_n &= \frac{\theta}{n+\theta}, \end{aligned}$$

where $\theta = p_1/r_1$.

Proof. $\Leftarrow$ Trivial.

$\Rightarrow$ First, let us prove that, if for any fixed n, $(X_1, .., X_n)$ is exchangeable with prediction rule (3.9), then $p_{ni} = p_{nj}$ for all $i, j = 1, ..., n$. Consider a set of $(n-1)$ partitions:

$$\begin{aligned} \Pi_n^1 &= (1)(2)...(n-2)(n-1,n), \\ \Pi_n^2 &= (1)(2)...(n-2,n)(n-1), \\ &\vdots \\ \Pi_n^{n-2} &= (1)(2,n)...(n-2)(n-1), \\ \Pi_n^{n-1} &= (1,n)(2)...(n-2)(n-1). \end{aligned}$$

This means that first $(n-1)$ observations are different but the last one belongs to one of the previously observed classes. By assumption, $(X_1, .., X_n)$ is exchangeable. In particular, this implies

$$\mathbb{P}\left(\Pi_n^1\right) = \mathbb{P}\left(\Pi_n^2\right) = ... = \mathbb{P}\left(\Pi_n^{n-1}\right).$$

Since weights do not depend on the partition, we obtain:

$$r_1 r_2 ... r_{n-2} p_{n-1,n-1} = r_1 r_2 ... r_{n-2} p_{n-1,n-2} = ... = r_1 r_2 ... r_{n-2} p_{n-1,1},$$

and so, $p_{n-1,i} = p_{n-1,j}$ for all $i, j = 1, ..., n-1$.

Denote $p_{ni} = p_n$ as now it does not depend on i. Then, condition (3.7) can be rewritten as

$$p_n = \frac{p_1}{r_1} r_n \ \text{ and } \ r_n = \frac{r_1 r_{n-1}}{p_1 r_{n-1} + r_1}.$$

Thus $p_n = p_1/(n-(n-1)r_1)$ and $r_n = r_1/(n-(n-1)r_1)$. Letting $\theta = \frac{r_1}{1-r_1}$, or equivalently, $p_1 = \frac{1}{1+\theta}$, we obtain

$$p_n = \frac{1}{n+\theta} \ \text{ and } \ r_n = \frac{\theta}{n+\theta}.$$

□

Thus, there exists a unique simple generalized Ottawa process that is exchangeable, and it has a Dirichlet Process de Finetti measure, since the Dirichlet Process with scale parameter θ is characterized by the prediction rule

$$\mathbb{P}\left[X_{n+1} \in \cdot | X_1, ..., X_n\right] = \sum_{i=1}^{n} \frac{1}{n+\theta} \delta_{X_i}(\cdot) + \frac{\theta}{n+\theta} \nu(\cdot). \tag{3.10}$$

Generalized Ottawa sequence that depends on the number of classes observed

For some problems, such as predicting the number of new classes in a future sample, it is natural to assume that the predictive weights depend on the data. A widely used example of process that allows such dependence and generates a species sampling sequence is the two-parameter Poisson Dirichlet Process (θ, σ) (Section 1.3). This process is no longer an example of a simple generalized Ottawa process, since the corresponding prediction rule is given by

$$\mathbb{P}\left[X_{n+1} \in \cdot | X_1, ..., X_n\right] = \sum_{i=1}^{n} \frac{1-\sigma/n_i}{n+\theta} \delta_{X_i}(\cdot) + \frac{\theta + k\sigma}{n+\theta} \nu(\cdot),$$

where n_i is the size of the class to which X_i belongs, and k is the number of classes observed in the sample. Thus, the weights depend on the observed $(n_1, ..., n_k)$ and are therefore random.

We would like to extend the prediction rule (3.9) by allowing the predictive weights to depend on the partition. Theoretically, we can specify the predictive weights to be dependent on whole vector of frequencies $(n_1, ..., n_k)$, but the condition that guarantees that such rule generates a c.i.d. sequence is too complicated and, therefore, not convenient. The simplest type of dependence on the data that can be used, is a dependence on the number of distinct classes. Thus, if K_n is the number of distinct classes observed in the sample of size n, then let the prediction rule be

$$\mathbb{P}\left[X_{n+1} \in \cdot | X_1, ..., X_n\right] = \sum_{i=1}^{n} p_{n,i}(K_n)\delta_{X_i}(\cdot) + r_n(K_n)\nu(\cdot). \tag{3.11}$$

Condition (3.7) of definition 3.2.8, which ensures the conditional identity in distribution of the sequence sampled form this rule can be expressed in the following way:

Proposition 3.2.10. *The prediction rule* (3.11) *generates a c.i.d. sequence if and only if*

$$\begin{aligned} p_{n,i}(K_n) = p_{n+1,i}(K_n)\left(1 - r_n(K_n)\right) + p_{n+1,i}(K_n+1)r_n(K_n) + \\ + p_{n+1,n+1}(K_n)p_{ni}(K_n) \textit{ for } i = 1, ..., n, \\ r_n(K_n) = r_{n+1}(K_n)\left(1 - r_n(K_n)\right) + p_{n+1,n+1}(K_n+1)r_n(K_n) + \\ + r_{n+1}(K_n+1)r_n(K_n). \end{aligned} \tag{3.12}$$

Proof. In order to show that the martingale condition (3.2) is satisfied we have to verify

$$\mathbb{E}\left[\mathbb{E}\left(f(X_{n+2})|\mathcal{G}_{n+1}\right)|\mathcal{G}_n\right] = \mathbb{E}\left[f(X_{n+1})|\mathcal{G}_n\right],$$

where $\mathcal{G}_n = \sigma(X_1, ..., X_n)$. The left hand side can be rewritten as

$$\begin{aligned} &\mathbb{E}\left[\mathbb{E}\left(f(X_{n+2})|\mathcal{G}_{n+1}\right)|\mathcal{G}_n\right] = \\ &= \mathbb{E}\left[\sum_{i=1}^{n+1} p_{n+1,i}(K_{n+1})f(X_i) + r_n(K_{n+1})\mathbb{E}\left[f(X_1)\right]|\mathcal{G}_n\right] = \\ &= \sum_{i=1}^{n} f(X_i)\mathbb{E}\left[p_{n+1,i}(K_{n+1})|K_n\right] + \mathbb{E}\left[f(X_{n+1})p_{n+1,n+1}(K_{n+1})|K_n\right] + \\ &+ \mathbb{E}\left[r_{n+1}(K_{n+1})|K_n\right]\mathbb{E}\left[f(X_1)\right]. \end{aligned}$$

As we know, $K_{n+1} = K_n + 1$ with probability $r_n(K_n)$ or $K_{n+1} = K_n$ with probability $1 - r(K_n)$. Then

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left(f(X_{n+2})|\mathcal{G}_{n+1}\right)|\mathcal{G}_n\right] = \\
&= \sum_{i=1}^{n} f(X_i)\left(p_{n+1,i}(K_n)(1 - r_n(K_n)) + p_{n+1,i}(K_n + 1)r_n(K_n)\right) + \\
&+ \sum_{i=1}^{n} f(X_i)\left(p_{n+1,n+1}(K_n)p_{n,i}(K_n)\right) + \mathbb{E}\left[f(X_1)\right]\left(p_{n+1,n+1}(K_n + 1)r_n(K_n)\right) + \\
&+ \mathbb{E}\left[f(X_1)\right]\left(r_{n+1}(K_n)(1 - r_n(K_n)) + r_{n+1}(K_n + 1)r_n(K_n)\right).
\end{aligned}$$

At the same time, the right hand side can be rewritten as

$$\mathbb{E}\left[f(X_{n+1})|\mathcal{G}_n\right] = \sum_{i=1}^{n} f(X_i)p_{n,i}(K_n) + \mathbb{E}\left[f(X_1)\right] r_n(K_n).$$

The expression of left is equal to the one on the right if and only if (3.12) holds. □

3.3 Partially c.i.d. sequences

Conditional identity in distribution was proposed as a generalization of exchangeability when there is no stationarity. In this section we generalize the notion of c.i.d. in the same sense that partial exchangeability generalizes exchangeability. The basic concepts regarding partial exchangeability in de Finetti's sense were given in Section 1.4.

The first problem we have to deal with is defining an order between two sequences $(X_n)_{n\geqslant 1}$ and $(Y_n)_{n\geqslant 1}$. Under partial exchangeability, order is not important, but for our construction the order does matter. Hereinafter, we assume that observations are coming in pairs:

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, \quad ... \tag{3.13}$$

and all further statements and definitions are given with respect to this order. We will see later that it is not relevant if X_n was actually observed earlier than Y_n or not, the important thing is to observe X_n and Y_n before X_{n+1} or Y_{n+1}. Theoretically, one can generalize the order (3.13), for example as

$$X_1, ..., X_{n_1}, Y_1, ..., Y_{m_1}, X_{n_1+1}, ..., X_{n_1+n_2}, Y_{m_1+1}, ..., Y_{m_1+m_2},$$

where $n_1, n_2, ...$ and $m_1, m_2, ...$ are either fixed or random. But then, theoretical properties of the sequences should be reanalyzed with respect to the new order.

Suppose we have two sequences of random variables $(X_n)_{n\geqslant 1}$ and $(Y_n)_{n\geqslant 1}$ that are defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, each variable takes values in a complete and separable (Polish) metric space $\mathbb{X}$ endowed with the Borel σ-algebra $\mathcal{X}$, and is adapted to a filtration $\mathcal{G} = (\mathcal{G}_n)_{n\geqslant 1}$.

Definition 3.3.1. *A sequence $(X_n, Y_n)_{n\geqslant 1}$ is said to be partially conditionally identically distributed with respect to a filtration $\mathcal{G}$, abbreviated as partially $\mathcal{G}$-c.i.d., if*

$$\begin{aligned} \mathbb{E}\left[f(X_k)|\mathcal{G}_n\right] &= \mathbb{E}\left[f(X_{n+1})|\mathcal{G}_n\right] \quad \mathbb{P}\text{-}a.s., \\ \mathbb{E}\left[f(Y_k)|\mathcal{G}_n\right] &= \mathbb{E}\left[f(Y_{n+1})|\mathcal{G}_n\right] \quad \mathbb{P}\text{-}a.s., \end{aligned} \tag{3.14}$$

for all $k > n \geqslant 0$ and all bounded measurable $f : \mathbb{X} \to \mathbb{R}$.

Condition (3.14) implies that for each $n \geqslant 0$, all future observations $(X_k)_{k>n}$ are identically distributed given the past $\mathcal{G}_n$ and all future $(Y_k)_{k>n}$ are identically distributed given the same past $\mathcal{G}_n$. In other words, we consider two sequences of random variables that are both separately $\mathcal{G}$-c.i.d. with respect to the same joint filtration $\mathcal{G}$. An important consequence is that all properties of $\mathcal{G}$-c.i.d. will hold for $(X_n)_{n\geqslant 1}$ and $(Y_n)_{n\geqslant 1}$ separately. Thus, being $\mathcal{G}$-c.i.d., marginally both $(X_n)_{n\geqslant 1}$ and $(Y_n)_{n\geqslant 1}$ are identically distributed, but these two marginal distributions can be different. The equivalent martingale condition is, in this case,

$$\begin{aligned} &\left(\mathbb{E}\left[f(X_{n+1})|\mathcal{G}_n\right]\right)_{n\geqslant 0} \text{ is a } \mathcal{G} - \text{martingale}, \\ &\left(\mathbb{E}\left[f(Y_{n+1})|\mathcal{G}_n\right]\right)_{n\geqslant 0} \text{ is a } \mathcal{G} - \text{martingale}, \end{aligned} \tag{3.15}$$

for every bounded measurable $f : \mathbb{X} \to \mathbb{R}$. As before, condition (3.15) specifies the martingale behavior of the predictive measures and will be convenient because it is generally easier to verify.

We define the partial natural filtration as $\mathcal{G} = \mathcal{G}^{XY} = \mathcal{G}^{X} \vee \mathcal{G}^{Y}$, where $\mathcal{G}_0^{XY} = \{\emptyset, \Omega\}$ and $\mathcal{G}_n^{XY} = \sigma(X_1, ..., X_n, Y_1, ..., Y_n)$. As before, in this case we omit the filtration and $(X_n, Y_n)_{n\geqslant 1}$ is just called *partially c.i.d.* or *natural partially c.i.d.*. Note, that a partially $\mathcal{G}$-c.i.d. sequence $(X_n, Y_n)_{n\geqslant 1}$ is also partially c.i.d. with respect to the natural filtration.

In the particular case of a natural partially c.i.d. sequence, conditions (3.14) and (3.15) reduce to

$$\begin{matrix} \begin{pmatrix} X_1, ..., X_n, X_{n+2} \\ Y_1, ..., Y_n \end{pmatrix} \overset{d}{=} \begin{pmatrix} X_1, ..., X_n, X_{n+1} \\ Y_1, ..., Y_n \end{pmatrix} \\ \begin{pmatrix} X_1, ..., X_n \\ Y_1, ..., Y_n, Y_{n+2} \end{pmatrix} \overset{d}{=} \begin{pmatrix} X_1, ..., X_n \\ Y_1, ..., Y_n, Y_{n+1} \end{pmatrix} \end{matrix} \quad for\, all\, n \geqslant 0. \tag{3.16}$$

A partially exchangeable sequence satisfies (3.16), so it is partially c.i.d. Moreover, a partially exchangeable sequence is partially c.i.d. for any order between $(X_n)_{n\geqslant 1}$ and $(Y_n)_{n\geqslant 1}$, not only for the order (3.13).

Notice that partially $\mathcal{G}$-c.i.d. sequences are defined through two sequences which are both separately $\mathcal{G}$-c.i.d. Thus, the existence of the directing random measures for $(X_n)_{n\geqslant 1}$ and $(Y_n)_{\geqslant 1}$ follows from the results of Berti et al. [2004] discussed in Section 3.3. That is, as before, for any partially $\mathcal{G}$-c.i.d. sequence $(X_n, Y_n)_{n\geqslant 1}$ and $f \in \mathcal{H}$, the class of measurable functions $f : \mathbb{X} \to \mathbb{R}$ such that $\mathbb{E}\left[|f(X_1)|\right] < \infty$, $\mathbb{E}\left[|f(Y_1)|\right] < \infty$, there exist integrable random variables denoted by V_f and W_f, such that

$$\begin{aligned} &\mathbb{E}\left[f(X_{n+1})|\mathcal{G}_n\right] \to V_f \;\; \mathbb{P}\text{-a.s. and in } L^1, \\ &\mathbb{E}\left[V_f|\mathcal{G}_n\right] = \mathbb{E}\left[f(X_{n+1})|\mathcal{G}_n\right] \;\; \mathbb{P}\text{-a.s. for every } n \geqslant 0, \\ &\mathbb{E}\left[f(Y_{n+1})|\mathcal{G}_n\right] \to W_f \;\; \mathbb{P}\text{-a.s. and in } L^1, \\ &\mathbb{E}\left[W_f|\mathcal{G}_n\right] = \mathbb{E}\left[f(Y_{n+1})|\mathcal{G}_n\right] \;\; \mathbb{P}\text{-a.s. for every } n \geqslant 0. \end{aligned} \tag{3.17}$$

Furthermore, these results, together with Lemma 2.4 in Berti et al. [2004], imply the existence of a vector of random measures (α, β) on $(\mathbb{X} \times \mathbb{X})$ such that, in the particular case when f is an indicator and $\mathcal{G}$ is a partial natural filtration,

$$\begin{aligned} &\mathbb{P}\left(X_{n+1} \in \cdot | X_{1:n}, Y_{1:n}\right) \overset{weakly}{\to} \alpha(\cdot)\ \mathbb{P}\text{-a.s.}, \\ &\mathbb{P}\left(Y_{n+1} \in \cdot | X_{1:n}, Y_{1:n}\right) \overset{weakly}{\to} \beta(\cdot)\ \mathbb{P}\text{-a.s.} \end{aligned} \tag{3.18}$$

We call the vector of random measures (α, β) the *directing random measure* of $(X_n, Y_n)_{n\geqslant 1}$. As for the univariate case, no representation theorems are available and the distribution of (α, β) cannot be characterized explicitly.

We have seen that for a natural c.i.d. process, the sequence of predictive measures converges to the same limit as the sequence of empirical measures. Partially c.i.d. sequences also preserve this property.

Theorem 3.3.2. *Let* $(X_n, Y_n)_{n\geqslant 1}$ *be a partially c.i.d. sequence. Then*

$$\begin{aligned} \tfrac{1}{n}\textstyle\sum_{i=1}^n \delta_{X_i}(\cdot) &\overset{weakly}{\to} \alpha(\cdot)\ \ \mathbb{P}\text{-}a.s., \\ \tfrac{1}{n}\textstyle\sum_{i=1}^n \delta_{Y_i}(\cdot) &\overset{weakly}{\to} \beta(\cdot)\ \ \mathbb{P}\text{-}a.s., \end{aligned}$$

where (α, β) *are as in* (3.18)*.*

Proof. Follows from the Strong Law of Large Numbers for c.i.d. (Theorem 2.2 in Berti et al. [2004]). □

Recalling that, in the case of the natural filtration, exchangeability is equivalent to conditional identity in distribution plus stationarity (Kallenberg [1988]), we want to examine what additional property is required for a partial c.i.d. sequence to be partially exchangeable.

As we mentioned before, a partially $\mathcal{G}$-c.i.d. sequence $(X_n, Y_n)_{n\geqslant 1}$ can be considered as two $\mathcal{G}$-c.i.d. sequences, with no constraint on their joint distribution. The same cannot be said in the case of partial exchangeability. If, for example, we consider two exchangeable sequences $(X_n)_{n\geqslant 1}$ and $(Y_n)_{n\geqslant 1}$, where $X_i = Y_i$ for all i, then these sequences are not partially exchangeable. According to Aldous [1985], two exchangeable sequences $(X_n)_{n\geqslant 1}$ and $(Y_n)_{n\geqslant 1}$ are also partially exchangeable if and only if they are conditionally independent given the vector of their directing random measures, say (α, β), i.e.,

$$(X_n)_{n\geqslant 1} \text{ and } (Y_n)_{n\geqslant 1} \text{ are independent } | \ (\alpha, \beta). \tag{3.19}$$

Thus, the conditions required for partially c.i.d. sequence to be also partially exchangeable are stationarity and condition (3.19).

Proposition 3.3.3. *The following statements are equivalent:*

- $(X_n, Y_n)_{n\geqslant 1}$ *is partially c.i.d., with both* $(X_n)_{n\geqslant 1}$ *and* $(Y_n)_{n\geqslant 1}$ *stationary, and conditionally independent given the vector of corresponding directing random measures* (α, β) *(i.e. condition* (3.19) *holds);*
- $(X_n, Y_n)_{n\geqslant 1}$ *is partially exchangeable.*

Proof. Trivial. □

Condition (3.19) is also required for asymptotic partial exchangeability. The following definition extends the notion of asymptotic exchangeability.

Definition 3.3.4. *A sequence* $(X_n, Y_n)_{n\geqslant 1}$ *is called asymptotically partially exchangeable if there exists a partially exchangeable sequence* $(Z_n, Z_n')_{n\geqslant 1}$ *such that*

$$(X_{n+1}, X_{n+2}, ..., Y_{n+1}, Y_{n+2}, ...) \overset{d}{\to} (Z_1, Z_2, ..., Z_1', Z_2', ...) \text{ as } n \to \infty. \tag{3.20}$$

Proposition 3.3.5. *If* $(X_n, Y_n)_{n\geqslant 1}$ *is partially* $\mathcal{G}$*-c.i.d. and condition (3.19) holds, then* $(X_n, Y_n)_{n\geqslant 1}$ *is asymptotically partially exchangeable.*

Proof. Both $(X_n)_{n\geqslant 1}$ and $(Y_n)_{n\geqslant 1}$ are separately c.i.d., therefore they are asymptotically exchangeable. Using their directing random measures α and β one can construct a partially exchangeable sequence $(Z_n, Z_n')_{n\geqslant 1}$ directed by these measures and such that

$$(X_{n+1}, .., X_{n+k}) \overset{d}{\to} (Z_1, ..., Z_k) \text{ and } (Y_{n+1}, .., Y_{n+k}) \overset{d}{\to} (Z_1', ..., Z_k'), \tag{3.21}$$

$(Z_n)_{n\geqslant 1}$ and $\mathcal{F} = \sigma(X_1, X_2, ...)$ are conditionally independent given α, $(Z_n')_{n\geqslant 1}$ and $\mathcal{G} = \sigma(Y_1, Y_2, ...)$ are conditionally independent given β. So, for any function f, the following is true:

$$\begin{aligned}&\mathbb{E}\left[f\left(X_{n+1}, ..., X_{n+k}, Y_{n+1}, ..., Y_{n+k}\right)\right] = \\ &\mathbb{E}\left[\mathbb{E}\left[f\left(X_{n+1}, ..., X_{n+k}, Y_{n+1}, ..., Y_{n+k}\right)|\alpha,\ \beta\right]\right] \overset{\text{as } n\to\infty}{\to} \\ &\to \mathbb{E}\left[\mathbb{E}\left[f\left(Z_1, ..., Z_k, Z_1', ..., Z_k'\right)|\alpha,\ \beta\right]\right] = \mathbb{E}\left[f\left(Z_1, ..., Z_k, Z_1', ..., Z_k'\right)\right].\end{aligned}$$

The convergence, guaranteed by conditions 3.19 and (3.21), implies

$$(X_{n+1}, X_{n+2}, ..., Y_{n+1}, Y_{n+2}, ...) \xrightarrow{d} (Z_1, Z_2, ..., Z_1', Z_2') \; as \; n \to \infty.$$

Therefore, the sequences are asymptotically partially exchangeable. □

In this way, the exchangeable asymptotic limit laws of a partially $\mathcal{G}$-c.i.d. process are directed by the random measures α and β respectively.

We have shown that the partially c.i.d. structure can be considered as a promising generalization which shares many properties of partial exchangeability. Moreover, together with condition (3.19), a partially c.i.d. sequence is also asymptotically partially exchangeable.

Example

The modified Polya urn defined in Berti et al. [2004] (and mentioned in Section 3.2) generates a c.i.d. sequence. We would like now to construct a bivariate version of it. Suppose that we have two urns, each of them initially containing $w > 0$ white and $r > 0$ red balls. At time n, one ball is drawn independently from each urn, recorded and placed back, say ball-1 and ball-2. Then, the first urn is reinforced with b_n balls of the same color as ball-1 and b_n' balls of the same type as ball-2. The second urn is reinforced with d_n balls of the same color as ball-1 and d_n' balls of the same type as ball-2. If we denote X_n to be an indicator of the event that white ball was drawn from the first urn and Y_n to be an indicator that white ball was drawn from the second urn, then $\mathbb{E}[X_1] = \mathbb{E}[Y_1] = w/(w + r)$ and

$$\mathbb{E}[X_{n+1}|\mathcal{G}_n] = \frac{w + \sum_{i=1}^n b_i X_i + \sum_{i=1}^n b_i' Y_i}{w + r + \sum d_i + \sum d_i'} \; \mathbb{P}\text{-a.s. for all } n \geqslant 1,$$

$$\mathbb{E}[Y_{n+1}|\mathcal{G}_n] = \frac{w + \sum_{i=1}^n d_i X_i + \sum_{i=1}^n d_i' Y_i}{w + r + \sum b_i + \sum b_i'} \; \mathbb{P}\text{-a.s. for all } n \geqslant 1,$$

where $\mathcal{G}_n = \sigma(X_{1:n}, Y_{1:n})$; and $(b_n, b_n')_{n\geqslant 1}$, $(d_n, d_n')_{n\geqslant 1}$ are sequences of random variables such that, for every n, (b_n, b_n') and (d_n, d_n') are independent of $\sigma(X_i, Y_i, b_j, b_j', d_j, d_j')$ for $i \leqslant n$, $j < n$.

It turns out that this scheme generates a partially c.i.d. sequence if and only if $d_n = b_n$ and $d_n' = b_n'$ for all n. This result is in some sense similar to the results in Section 2.4, where we wanted to specify a relatively simple structure (where predictive weights were functions of frequencies) and the model reduced to a Dirichlet Process for a homogeneous population of species. We are not going to discuss here other examples of partially c.i.d. sequences, as they will be widely discussed in Chapter 4.

We would like to note that the ordering (3.13) was chosen after considering other alternatives, such as defining two sequences to be c.i.d. with respect to different filtrations, for example, defining a prediction for Y_{n+1} conditional on $(X_{1:n+1}, Y_{1:n})$. Some of the results in this section can be replicated for different orderings, but the calculations become more complicated, while the ordering (3.13) that we propose here allows us to construct a quite natural generalization of partial exchangeability.

3.3.1 Limit theorem for partially c.i.d. sequences

In this subsection we show that, for a partially c.i.d. sequence, the vector of differences between predictive and empirical measures converges stably, under suitable conditions, to a mixture of Gaussians. Let us start with a definition of stable convergence.

Definition 3.3.6. *If $(X_n, Y_n)_{n\geqslant 1}$ is a sequence of vectors of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ converging in distribution to a vector of random variables (X, Y), we say that the convergence is stable if for all continuity points (x, y) of (X, Y) and all events $E \in \mathcal{F}$, the limit*

$$\mathop{lim}_{n\to\infty} \mathbb{P}\left(\{X_n \leqslant x,\, Y_n \leqslant y\} \cap E\right) = Q_{XY}(E),$$

exists, and $Q_{XY}(E) \to \mathbb{P}(E)$ as $(x, y) \to (\infty, \infty)$.

We require the following version of the martingale central limit theorem given in Hall and Heyde [1980], Theorem 3.2. Let $\{Y_{nk} : n \geqslant 1, k = 1, \ldots, k_n\}$ be an array of real, square-integrable random variables, where $k_n \to \infty$, and for all n, let $\mathcal{F}_{n0} \in \mathcal{F}_{n1} \in \cdots \in \mathcal{F}_{nk_n} \in \mathcal{A}$ be σ-fields with $\mathcal{F}_{n0} = \{\emptyset, \Omega\}$. If

1. $\sigma(Y_{nk}) \subset \mathcal{F}_{nk}$, $\mathbb{E}[Y_{nk}|\mathcal{F}_{n,k-1}] = 0$ $\mathbb{P}$-a.s., $\mathcal{F}_{nk} \in \mathcal{F}_{n+1,k}$;

2. $max_{1\leqslant k\leqslant k_n}|Y_{nk} \to 0$ in probability, $sup_n\mathbb{E}[max_{1\leqslant k\leqslant k_n}Y_{nk}^2] < \infty$;

3. $\sum_{k=1}^{k_n} Y_{nk}^2 \to L$ in probability for some random variable L,

then $\sum_{k=1}^{k_n} Y_{nk}$ converges stably to $\mathcal{N}(0, L)$.

Let $(X_n)_{n\geqslant 1}$ and $(Y_n)_{n\geqslant 1}$ be two sequences of random variables defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, taking values in a measurable space $(\mathbb{X}, \mathcal{X})$, and adapted to a filtration $\mathcal{G} = (\mathcal{G}_n)_{n\geqslant 0}$. For real measurable functions f and g such that, for all n, $\mathbb{E}|f(X_n)| + \mathbb{E}|g(Y_n)| < \infty$, define

$$M_n = f(X_n) - n\mathbb{E}\left[f(X_{n+1})|\mathcal{G}_n\right] + (n - 1\mathbb{E}\left[f(X_n)|\mathcal{G}_{n-1}\right]),$$
$$L_n = g(Y_n) - n\mathbb{E}\left[g(Y_{n+1})|\mathcal{G}_n\right] + (n - 1\mathbb{E}\left[g(Y_n)|\mathcal{G}_{n-1}\right]).$$

Define also

$$C_n = \frac{1}{\sqrt{n}}\left(f(X_1) + ... + f(X_n) - n\mathbb{E}[f(X_{n+1})|\mathcal{G}_n]\right),$$
$$D_n = \frac{1}{\sqrt{n}}\left(g(Y_1) + ... + f(Y_n) - n\mathbb{E}[g(Y_{n+1})|\mathcal{G}_n]\right).$$

Theorem 3.3.7. *Suppose $(X_n, Y_n)_n$ are partially $\mathcal{G}$-c.i.d. for some filtration $\mathcal{G}$. If*

$$\mathbb{E}f(X_1)^2 + \mathbb{E}g(Y_1)^2 + \underset{n}{sup}\mathbb{E}C_n^2 + \underset{n}{sup}\mathbb{E}D_n^2 < \infty \textit{ and}$$
$$\frac{1}{n}\sum_{i=1}^{n} M_i^2 \overset{a.s.}{\to} U,\ \frac{1}{n}\sum_{i=1}^{n} L_i^2 \overset{a.s.}{\to} V,\ \frac{1}{n}\sum_{i=1}^{n} M_iL_i \overset{a.s.}{\to} Z,$$

then $(C_n, D_n) \to \mathcal{N}(0, \Sigma)$ stably, where Σ is the random covariance matrix

$$\Sigma = \begin{bmatrix} U & Z \\ Z & V \end{bmatrix}.$$

Proof. (This proof has been suggested by P.Rigo and P.Berti) Let J_1 and J_2 be the limits of empirical measures, i.e. random variables such that $\frac{1}{n}\sum_{i=1}^{n} f(X_i) \overset{a.s.}{\to} J_1$ and $\frac{1}{n}\sum_{i=1}^{n} g(Y_i) \overset{a.s.}{\to} J_2$. Given $(a, b) \in \mathbb{R}^2$, for each $n \geqslant 1$ and $k = 1, ..., n$, define $\mathcal{F}_{n,k} = \mathcal{G}_k$ and

$$W_n = a\sqrt{n}\left(\tfrac{1}{n}\sum_{i=1}^{n} f(X_i) - J_1\right) + b\sqrt{n}\left(\tfrac{1}{n}\sum_{i=1}^{n} g(Y_i) - J_2\right),$$
$$Y_{n,k} = \mathbb{E}\left[W_n|\mathcal{G}_k\right] - \mathbb{E}\left[W_n|\mathcal{G}_{k-1}\right].$$

Then, condition (1) above holds and it is possible to show that

$$\sum_{k=1}^{n} Y_{n,k}^2 = \frac{1}{n}\sum_{k=1}^{n}(aM_k + bL_k)^2 \to a^2U + b^2V + 2abZ \;\; \mathbb{P}\text{-a.s.}$$

Since

$$Y_{nn}^2 = \sum_{k=1}^{n}(aM_k + bL_k)^2 - \frac{n-1}{n}\sum_{k=1}^{n-1}(aM_k + bL_k)^2 \to 0 \;\; \mathbb{P}\text{-a.s.},$$

then $\underset{1\leqslant k\leqslant n}{max} Y_{n,k}^2 \overset{a.s.}{\to} 0$. At the same time,

$$\mathbb{E}[Y_{nk}^2] \leqslant \sum_{k=1}^{n}\mathbb{E}[(\mathbb{E}\left[W_n|\mathcal{G}_k\right] - \mathbb{E}\left[W_n|\mathcal{G}_{k-1}\right])^2] = \mathbb{E}[C_n^2] + \mathbb{E}[D_n^2] < \infty.$$

Thus, conditions (2) and (3) are satisfied and the central limit theorem for martingales by Hall and Heyde [1980] yields

$$aC_n + bD_n = \mathbb{E}\left[W_n|\mathcal{G}_n\right] = \sum k = 1nY_{n,k} \overset{stably}{\to} \mathcal{N}\left(0, a^2U + b^2V + 2abZ\right).$$

Thus, the theorem is proved. □

3.3.2 Generalized bivariate species sampling sequences

In section 3.2.1 we discussed the problem of species sampling. According to Pitman, a species sampling sequence is an exchangeable sequence with a prediction rule of the form (1.1). However, it may be that the population of species is not homogeneous. This idea was explored in Chapter 2, where we defined a bivariate species sampling sequence as a partially exchangeable sequence sampled from a bivariate prediction rule (2.2). Due to the partial exchangeability property, these processes were stationarity. Nevertheless, for some applications the assumption of stationarity can be too restrictive. It is natural to combine a bivariate prediction rule with the concept of partial c.i.d. that was defined in the previous section, and which generalizes partial exchangeability.

Let us recall the notation introduced in Chapter 2. We denote by Π_{nn} a random bi-partition of $N_n \times N_n$, and by $\Pi(X_{1:n}, Y_{1:n})$ a random bi-partition created from a sequence $(X_1, ..., X_n; Y_1, ..., Y_n)$ by the rule specified in Definition 2.2.1.

Definition 3.3.8. *We say that a sequence $(X_n, Y_n)_{n\geqslant 1}$ of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, with values in $\mathbb{X}$, is a generalized bivariate species sampling sequence if it is partially c.i.d. and has a bivariate prediction rule that satisfies the following constraint:*

$$
\begin{aligned}
&X_1 \sim \nu,\ Y_1 \sim \nu \textit{ and, for } n \geqslant 1,\\
&F_n^X(\cdot) = \mathbb{P}\left[X_{n+1} \in \cdot|\mathcal{G}_n\right] = \sum_{i=1}^{n} p_{n,i}\left(\Pi_{nn}\right)\delta_{X_i}(\cdot) + \\
&\qquad\qquad\qquad\qquad\qquad + \sum_{i=1}^{n} p'_{n,i}\left(\Pi_{nn}\right)\delta_{Y_i}(\cdot) + r_n\left(\Pi_{nn}\right)\nu(\cdot),\\
&F_n^Y(\cdot) = \mathbb{P}\left[Y_{n+1} \in \cdot|\mathcal{G}_n\right] = \sum_{i=1}^{n} g_{n,i}\left(\Pi_{nn}\right)\delta_{X_i}(\cdot) + \\
&\qquad\qquad\qquad\qquad\qquad + \sum_{i=1}^{n} g'_{n,i}\left(\Pi_{nn}\right)\delta_{Y_i}(\cdot) + h_n\left(\Pi_{nn}\right)\nu(\cdot),
\end{aligned}
\tag{3.22}
$$

where $\mathcal{G}_n = \mathcal{G}_n^X \vee \mathcal{G}_n^Y = \sigma(X_1, ..., X_n) \vee \sigma(Y_1, ..., Y_n)$, the weights $p_{n,i}(\Pi_{nn})$, $p'_{n,i}(\Pi_{nn})$ and $r_n(\Pi_{nn})$, $g_{n,i}(\Pi_{nn})$, $g'_{n,i}(\Pi_{nn})$, $h_n(\Pi_{nn})$ are suitable measurable functions defined on random partitions Π_{nn} with values in $[0, 1]$, and ν is a diffuse measure on $\mathbb{X}$.

The idea behind this prediction rule is the usual: given the past observations $(X_{1:n}, Y_{1:n})$, the next observation X_{n+1} could be either equal to one of the previous $X_{1:n}$, or equal to one of the previous $Y_{1:n}$, or a new value. By “new” we mean a value that has not been observed neither among $(X_1, ..., X_n)$ nor among $(Y_1, ..., Y_n)$. The prediction Y_{n+1} is treated analogously. Predictive weights specify the probabilities of each of these events. In the particular case when $p'_{n,i} = g_{n,i} = 0$, for all n and $i = 1, ..., n$, the sequences are not allowed to share values, but the processes $(X_n)_{n\geqslant 1}$ and $(Y_n)_{n\geqslant 1}$ can be dependent, as other weights depend on the whole partition.

It is worth noting that a univariate generalized species sampling sequence was proposed by Bassetti et al. [2010], but their model is different, as it involves a set of latent variables. We, instead, would like to avoid the latent variables, since integrating them out is generally analytically complex, and we are searching for a reasonably simple prediction rule.

Example: In Section 2.3, Chapter 2, we discussed the natural example and we observed that, apart from the trivial case, this model does not generate a partially exchangeable sequence. Here, we would like to check if a particular example of this natural model generates a partially c.i.d. sequence. Let us consider a sequence with a bivariate prediction rule that satisfies the following constraint:

$$\begin{aligned}\mathbb{P}\left(X_{n+1}\in\cdot|X_{1:n},Y_{1:n}\right)&=\sum_{i=1}^{n}\frac{a}{\theta+an+a'n}\delta_{X_i}(\cdot)+\\&+\sum_{i=1}^{n}\frac{a'}{\theta+an+a'n}\delta_{Y_i}(\cdot)+\frac{\theta}{\theta+an+a'n}\nu(\cdot),\\\mathbb{P}\left(Y_{n+1}\in\cdot|X_{1:n},Y_{1:n}\right)&=\sum_{i=1}^{n}\frac{b}{\theta'+bn+b'n}\delta_{X_i}(\cdot)+\\&+\sum_{i=1}^{n}\frac{b'}{\theta'+bn+b'n}\delta_{Y_i}(\cdot)+\frac{\theta'}{\theta'+bn+b'n}\nu(\cdot).\end{aligned}\tag{3.23}$$

The following proposition establishes when the bivariate rule (3.23) generates a partially c.i.d. sequence.

Proposition 3.3.9. *The bivariate rule* (3.23) *generates a partially c.i.d. sequence if and only if*

$$\begin{aligned}\frac{a}{\theta+an+a'n}&=\frac{b}{\theta'+bn+b'n},\\\frac{a'}{\theta+an+a'n}&=\frac{b'}{\theta'+bn+b'n}.\end{aligned}$$

Proof. In order to prove the theorem we should check the martingale condition (3.15). Thus, the first condition we should check is

$$\mathbb{E}\left[\mathbb{E}\left[f(X_{n+2})|\mathcal{G}_{n+1}\right]|\mathcal{G}_n\right]=\mathbb{E}\left[f(X_{n+1})|\mathcal{G}_n\right].$$

Let us denote

$$p_n = \frac{a}{\theta + an + a'n}, \quad p_n' = \frac{a'}{\theta + an + a'n}, \quad r_n = \frac{\theta}{\theta + an + a'n},$$
$$g_n = \frac{b}{\theta + bn + b'n}, \quad g_n' = \frac{b'}{\theta + bn + b'n}, \quad h_n = \frac{\theta'}{\theta + bn + b'n}.$$

Then, we can rewrite the left hand side as

$$p_{n+1}\left(\sum_{i=1}^{n} f(X_i) + \mathbb{E}\left[f(X_{n+1})|\mathcal{G}_n\right]\right) + p_{n+1}'\left(\sum_{i=1}^{n} f(Y_i) + \mathbb{E}\left[f(Y_{n+1})|\mathcal{G}_n\right]\right) +$$
$$+ r_{n+1}\mathbb{E}\left[f(X_1)\right] = (p_{n+1} + p_{n+1}p_n + p_{n+1}'g_n)\sum_{i=1}^{n} f(X_i) +$$
$$+ (p_{n+1}' + p_{n+1}p_n' + p_{n+1}'g_n')\sum_{i=1}^{n} f(Y_i) + (r_{n+1} + p_{n+1}r_n + p_{n+1}'h_n)\,\mathbb{E}\left[f(X_1)\right].$$

While, the right hand side can be rewritten as $p_n \sum_{i=1}^{n} f(X_i) + p_n' \sum_{i=1}^{n} f(Y_i) + r_n\mathbb{E}\left[f(X_1)\right]$. Thus, we should check when

$$p_{n+1} = p_n(1 - p_{n+1}) - p_{n+1}'g_n,$$
$$p_{n+1}' = p_n'(1 - p_{n+1}) + p_{n+1}'g_n',$$
$$r_{n+1} = r_n(1 - p_{n+1}) + p_{n+1}'h_n.$$

From the first and second equalities we can get that

$$\frac{a}{\theta + an + a'n} = \frac{b}{\theta' + bn + b'n}, \quad \frac{a'}{\theta + an + a'n} = \frac{b'}{\theta' + bn + b'n}.$$

The same results follow from the second martingale condition. □

This result recalls a previous one which established the conditions this prediction rule to generate a partially exchangeable sequence. In that case, we saw that the predictive weights should be of the form $p_{n,i} = p_{n,i}' = g_{n,i} = g_{n,i}' = 1/(\theta + n + n)$, for $i = 1, ..., n$ and for every n. Here, weights a and a' can be different, but the predictive distribution of X_{n+1} coincides with the predictive distribution of Y_{n+1}.

Further examples are provided in Chapter 4. In particular, in Section 4.2 we propose an extension of prediction rule (3.23) that generates a partially c.i.d. sequence such that prediction rules are different for the two sequences.

Chapter 4

Partially c.i.d. sequences: examples and inference

In Chapter 3 we proposed a new type of bivariate dependence, extending the notion of conditional identity in distribution, and defined a class of generalized bivariate species sampling sequences. In this chapter we discuss some examples, investigate their features and present inference results. Section 4.1 provides an example of binary partially c.i.d. structures, which generalizes well-known reinforced urn models and can be used for explaining and predicting physical or behavioral phenomena in different areas. In Section 4.2 we define a simple generalized bivariate species sampling sequence for species sampling problems in which investigating the appearance of a new species is a main task. Such problems arise, for example, in ecology, biology and language modeling (vocabulary studies). The proposed structure has a simple prediction rule, is quite flexible to fit, and accommodates power-law tails. In the last section we state some problems for further research.

4.1 Binary partially c.i.d. sequences

The first example of a partially c.i.d. structure that we consider is a binary partially c.i.d. sequence. Although this type of sequence is not an example of generalized bivariate species sampling sequences, since only two possible values are allowed, it does constitute an interesting model which gives a bivariate generalization of well-known reinforced urn models. A particular property of the proposed model is that it allows two reinforcements for each urn. We start by recalling the notion of reinforced urns and its applications.

The idea of the Polya urn was originally proposed by Eggenberger and Pólya [1923]. Assume that we have an urn containing some initial number of balls of two colors, let say white and black. At time $n = 1, 2, ...$, we randomly sample a ball from the urn, observe its color and put it back along with an additional ball of the same color. Repeating this procedure sequentially, we obtain a set of colors that can be seen as an infinite sequence of Bernoulli random variables whose probability law has been reinforced. This self-reinforcement property can be expressed as “the rich get richer”. The crucial benefit of the urn model is that an infinite sequence of colors sampled from a Polya urn is infinitely exchangeable. Moreover, from de Finetti’s representation theorem, it follows that any infinite sequence of exchangeable Bernoulli random variables is reinforced. This scheme has a lot of applications in different areas and has also been enriched and generalized in many different ways. For example, Hill et al. [1980] proposed a generalized urn process where the conditional distribution of the next color, given the past, is Bernoulli($f(W_n)$), where W_n denoting the proportion of white balls in the urn at time n, and $f : [0, 1] \to [0, 1]$ is some function. The Polya urn is a particular case of this model, with $f(x) = x$. But, apart from few cases, the sequence of colors sampled from a generalized urn is not exchangeable.

Another generalization of the Polya urn scheme is a randomly reinforced urn, where the urn’s composition is reinforced with a random number of balls of the same color as that of the ball extracted. The infinite sequence of colors sampled from this urn is not exchangeable, but under some conditions on the random variables involved in the reinforcement, it is asymptotically exchangeable in the sense specified in Definition 3.2.3. In particular, this means that the sequence is not exchangeable at the beginning, but asymptotically, the tail of the sequence is exchangeable. Such model was discussed in Berti et al. [2004] as we mentioned in Section 3.2, May et al. [2005], Muliere et al. [2006].

Li et al. [1996] and Durham et al. [1998] discussed a useful application of reinforced models in randomized response-adaptive design for clinical trials. The two colors, in this case, correspond to two treatments. The idea is to assign a patient to a treatment according to the color of the sampled ball, and reinforce the urn with a ball of the same color if the treatment was successful. Durham et al. [1998] showed that under such construction, the probability of selecting the superior treatment tends to one. This feature makes the model very attractive for clinical trials. A review of asymptotically optimal response-adaptive designs for allocation of the best treatment is given in Flournoy et al. [2012]. Random processes with reinforcement are also widely used to explain and predict physical or behavioral phenomena in different areas. Pemantle [2007] gives a very detailed review of such applications. The path-dependent processes are also discussed in Brian Arthur et al. [1987], with applications in industrial location theory, chemical kinetics and the evolution of technological structure in economy.

Expanding the idea from one process to several dependent processes, one can imagine many applications for bivariate reinforced structures. Almost all of applications mentioned above can be generalized. For example, in the problem of random limiting market shares, discussed in Brian Arthur et al. [1987], it is assumed that two technologies come into the market and each new customer decides which of the two to buy, according to the decisions of the previous customers; the problem still makes sense in the case of several dependent markets. The problem of how customer behavior is learned, and how future behavior is influenced by rewards, is another example that can be interesting in the bivariate case. We can conclude that bivariate reinforced urns may be useful in different areas.

A system of reinforced urns has been proposed by Paganoni and Secchi [2004], in which resulting sequences of colors are asymptotically exchangeable. Assume we have two urns and, initially, the first urn contains B_0^I balls of color 1 and B_0^I balls of color 0 while the second urn contains B_0^{II} balls of color 1 and W_0^{II} balls of color 0. At time $n = 1, 2, ...$, we randomly sample a ball from the first urn and a ball from the second urn. Call X_n the color generated by the first urn and Y_n the color generated by the second urn. We observe X_n, Y_n and put them back. Additionally, we add a random number $r(I, X_n, Y_n, M_n) \geqslant 0$ of balls of the same color as X_n into the first urn, and random number $r(II, X_n, Y_n, M_n) \geqslant 0$ of balls of the same color as Y_n into the second urn. Given the past realizations, $\mathcal{G}_{n-1} = \sigma(X_{1:n-1}, Y_{1:n-1}, M_{1:n-1})$, the first reinforcement $r(I, X_n, Y_n, M_n)$, is assumed to be conditionally independent of X_n and the second reinforcement, $r(II, X_n, Y_n, M_n)$, is assumed to be conditionally independent of Y_n, where $(M_n)_{n\geqslant 1}$ is some set of i.i.d. random variables such that M_n is independent of $\mathcal{G}_{n-1}$ and of (X_n, Y_n).

The corresponding prediction rule is: $X_1 \sim Bernoulli(Z_0^I)$, $Y_1 \sim Bernoulli(Z_0^{II})$, where

$$Z_0^I = \frac{B_0^I}{B_0^I + W_0^I},\ Z_0^{II} = \frac{B_0^{II}}{B_0^{II} + W_0^{II}}, \tag{4.1}$$

and for $n \geqslant 1$, $X_{n+1}|\mathcal{G}_n \sim Bernoulli(Z_n^I)$, $Y_{n+1}|\mathcal{G}_n \sim Bernoulli(Z_n^{II})$, where

$$\begin{aligned}
&Z_n^I = \frac{B_n^I}{B_n^I + W_n^I},\ Z_n^{II} = \frac{B_n^{II}}{B_n^{II} + W_n^{II}},\\
&B_n^I = B_{n-1}^I + X_n r(I, X_n, Y_n, M_n),\\
&W_n^I = W_{n-1}^I + (1 - X_n) r(I, X_n, Y_n, M_n),\\
&B_n^{II} = B_{n-1}^{II} + Y_n r(II, X_n, Y_n, M_n),\\
&W_n^{II} = W_{n-1}^{II} + (1 - Y_n) r(II, X_n, Y_n, M_n).
\end{aligned} \tag{4.2}$$

Paganoni and Secchi [2004] show that the processes $(Z_n^I)_n \geqslant 1$ and $(Z_n^{II})_{n\geqslant 1}$ are martingales with respect to the filtration $\mathcal{G} = (\mathcal{G}_n)_{n\geqslant 0}$, which implies that the sequence $(X_n, Y_n)_{n\geqslant 0}$ is partially c.i.d. with respect to the same filtration.

Let us give a closer look at the rule defined by (4.2). If a ball of color 1 is sampled from the first urn, then we reinforce this urn with balls of color 1. But we cannot reinforce the first urn with balls of color 2 if, at the same time n, a ball of color 2 was sampled from the second urn. This model allows only one reinforcement per urn, while in some applications the fact that the event of interest has happened in the first group should reinforce the probability of this event in second group. It can be useful to allow for the possibility of two reinforcements per urn, depending on both balls sampled at time n.

Recall that in Section 3.3 we tried to construct such bivariate version of the modified urn model by considering

$$B_n^I = B_0^I + \sum_{i=1}^n d_i X_i + \sum_{i=1}^n d_i' Y_i,$$
$$B_n^{II} = B_0^{II} + \sum_{i=1}^n b_i X_i + \sum_{i=1}^n b_i' Y_i,$$

which seems to be a natural generalization of the reinforced urn model. Yet, it turned out that the constructed urns generate a partially c.i.d. sequence if and only if the prediction rules for the two urns are exactly the same. So, we asked ourselves a different question - what form of bivariate prediction rule, which allows two reinforcements per urn, is needed in order to generate a partially c.i.d. sequence.

Let us now consider two urns, and suppose $(X_{1:n}, Y_{1:n})$ is a bivariate sequence of random variables with values in $\{0, 1\}$, generated according to the following rule:

$$\begin{aligned} (X_{n+1}|X_{1:n}, Y_{1:n}) &= \begin{cases} 1 & \text{with probability } \ p_n(X_{1:n}, Y_{1:n}); \\ 0 & \text{with probability } \ 1 - p_n(X_{1:n}, Y_{1:n}); \end{cases} \\ (Y_{n+1}|X_{1:n}, Y_{1:n}) &= \begin{cases} 1 & \text{with probability } \ g_n(X_{1:n}, Y_{1:n}); \\ 0 & \text{with probability } \ 1 - g_n(X_{1:n}, Y_{1:n}). \end{cases} \end{aligned} \tag{4.3}$$

The predictive functions p_n, g_n depend on both frequencies of the events in the samples from each of the two urns. One can say that the rule (4.3) generates dependent sequences of Bernoulli random variables with reinforced laws.

In order to give conditions under which the bivariate prediction rule (4.3) generates a partially c.i.d. sequence we should check the crucial condition (3.15), that can be written as:

$$\begin{aligned} \mathbb{E}\left[p_{n+1}(X_{1:n+1}, Y_{1:n+1}) | X_{1:n}, Y_{1:n}\right] = p_n(X_{1:n}, Y_{1:n}), \\ \mathbb{E}\left[g_{n+1}(X_{1:n+1}, Y_{1:n+1}) | X_{1:n}, Y_{1:n}\right] = g_n(X_{1:n}, Y_{1:n}). \end{aligned} \tag{4.4}$$

Let us consider an example of such structure. We assume that the predictive weights are

$$\begin{aligned} p_n(X_{1:n}, Y_{1:n}) = \tfrac{p_1(K_n)+p_2(K'_n)}{f(X_{1:n-1}, Y_{1:n-1})}, \\ g_n(X_{1:n}, Y_{1:n}) = \tfrac{g_1(K_n)+g_2(K'_n)}{h(X_{1:n-1}, Y_{1:n-1})}, \end{aligned} \tag{4.5}$$

where $K_n = \sum_{i=1}^n X_i$ and $K'_n = \sum_{i=1}^n Y_i$. If $(X_i = 1)$ represents some event, then K_n tells how many times this event has happened before in the first group, and similarly for the second group. It is worth noting, that the denominators depend only on $(X_{1:n-1}, Y_{1:n-1})$.

The main difference with the interacting reinforced urns proposed by Paganoni and Secchi [2004], is the fact that we can reinforce the first urn with black balls if a black ball is sampled from the second urn, regardless the color of the ball sampled from the first urn. The following proposition can be formulated.

Proposition 4.1.1. *I. A sequence $(X_n, Y_n)_{n\geqslant 1}$ sampled from a prediction rule of the form* (4.3)*, assuming* (4.5)*, is a partially c.i.d. sequences with respect to the natural filtration $\mathcal{G}_n = \sigma\{X_{1:n}, Y_{1,n}\}$ if and only if*

$$\begin{aligned} f(X_{1:n+1}, Y_{1:n+1}) &= f(X_{1:n}, Y_{1:n}) + (p_1(K_n+1) - p_1(K_n)) + \\ &+ (p_2(K'_n+1) - p_2(K'_n)) \frac{f(X_{1:n}, Y_{1:n})\,(g_1(K_n) + g_2(K'_n))}{h(X_{1:n}, Y_{1:n})\,(p_1(K_n) + p_2(K'_n))}, \\ h(X_{1:n+1}, Y_{1:n+1}) &= h(X_{1:n}, Y_{1:n}) + (g_2(K'_n+1) - g_2(K'_n)) + \\ &+ (g_1(K_n+1) - g_1(K_n)) \frac{h(X_{1:n}, Y_{1:n})\,(p_1(K_n) + p_2(K'_n))}{f(X_{1:n}, Y_{1:n})\,(g_1(K_n) + g_2(K'_n))}. \end{aligned}$$

II. If a sequence $(X_n, Y_n)_n$ with prediction rule of the form (4.3) *is partially c.i.d., then $(Z_n^I = \mathbb{P}\left[X_{n+1} = 1 | \mathcal{G}_n\right])_{n\geqslant 1}$ converges a.s to a random element $Z_\infty^I \in [0,1]$ and $(Z_n^{II} = \mathbb{P}\left[Y_{n+1} = 1 | \mathcal{G}_n\right])_{n\geqslant 1}$ converges a.s to a random element $Z_\infty^{II} \in [0,1]$. These Z_∞^I and Z_∞^{II} are the limits of the relative frequencies $\frac{K_n^I}{n}$ and $\frac{K_n^{II}}{n}$, respectively.*

Proof. I. We have to check conditions (4.4). Consider the first condition:

$$\mathbb{E}\left[\frac{p_1(K_{n+1})+p_2(K'_{n+1})}{f(X_{1:n},Y_{1:n})}|X_{1:n},Y_{1:n}\right]=\mathbb{E}\left[\frac{p_1(K_{n+1})+p_2(K'_{n+1})}{f(X_{1:n},Y_{1:n})}|K_n,K'_n\right]=$$
$$=\frac{p_1(K_n+1)+p_2(K'_n+1)}{f(X_{1:n},Y_{1:n})}\frac{p_1(K_n)+p_2(K'_n)}{f(X_{1:n-1},Y_{1:n-1})}\frac{g_1(K_n)+g_2(K'_n)}{h(X_{1:n-1},Y_{1:n-1})}+$$
$$+\frac{p_1(K_n)+p_2(K'_n+1)}{f(X_{1:n},Y_{1:n})}\left(1-\frac{p_1(K_n)+p_2(K'_n)}{f(X_{1:n-1},Y_{1:n-1})}\right)\frac{g_1(K_n)+g_2(K'_n)}{h(X_{1:n-1},Y_{1:n-1})}+$$
$$+\frac{p_1(K_n+1)+p_2(K'_n)}{f(X_{1:n},Y_{1:n})}\frac{p_1(K_n)+p_2(K'_n)}{f(X_{1:n-1},Y_{1:n-1})}\left(1-\frac{g_1(K_n)+g_2(K'_n)}{h(X_{1:n-1},Y_{1:n-1})}\right)+$$
$$+\frac{p_1(K_n)+p_2(K'_n)}{f(X_{1:n},Y_{1:n})}\left(1-\frac{p_1(K_n)+p_2(K'_n)}{f(X_{1:n-1},Y_{1:n-1})}\right)\left(1-\frac{g_1(K_n)+g_2(K'_n)}{h(X_{1:n-1},Y_{1:n-1})}\right),$$

should be equal to $\frac{p_1(K_n)+p_2(K'_n)}{f(X_{1:n-1},Y_{1:n-1})}$. Solving this equation we get

$$f(X_{1:n},Y_{1:n})=f(X_{1:n-1},Y_{1:n-1})+(p_1(K_n+1)-p_1(K_n))+$$
$$+(p_2(K'_n+1)-p_2(K'_n))\frac{f(X_{1:n-1},Y_{1:n-1})\,(g_1(K_n)+g_2(K'_n))}{h(X_{1:n-1},Y_{1:n-1})\,(p_1(K_n)+p_2(K'_n))}.$$

Applying the same procedure to the second condition, we obtain an analogous expression for $h(X_{1:n},Y_{1:n})$.

II. Follows immediately from previous results for partially c.i.d. sequences. □

4.2 Generalized bivariate species sampling sequence for species discovery

Unseen species problems constitute a large class of problems that arise in ecology, biology, language modeling (vocabulary studies) and other areas. Suppose that a sample of size n from some population of species has been observed. Based on this sample, one has to solve different predictive issues concerning the composition of the population. One of the problems is to estimate the *coverage* C_n of the sample, that is, the proportion of the population that is represented in the given sample of size n. In other words, C_n is the sum of the probabilities of the observed classes or, equivalently, the probability that the next observation will not be of a new type. Although the coverage is not a parameter of the population, it helps to understand the redundancy of the observed sample.

Another problem is the prediction of $K_m^{(n)}$, the number of new species that will be discovered in an additional sample of size m, given the observed sample of size n. For m going to infinity, the problem turns into the problem of richness estimation. This analysis helps to understand the effectiveness of further sampling, and to estimate the size of the additional effort needed to reach a certain amount of discovered species.

Finally, we consider the problem of estimating the *discovery rate* $D_m^{(n)}$, that is, the probability that, given the sample of size n, the $(n+m+1)$-th observation will be of a new type, without actually observing the additional m-size sample. As a function of m, the discovery rate predicts the evolution of the probability of observing new species. Clearly, $D_{n+1}^{(n)} = 1 - C_n$.

These problems have a lot of applications in different fields, and so a lot of models have been proposed after the first works by Fisher et al. [1943], Good [1953], Good and Toulmin [1956]. Comprehensive reviews can be found in Bunge and Fitzpatrick [1993] and Colwell and Coddington [1994].

The first model in a Bayesian nonparametric framework was proposed by Tiwari and Tripathi [1989], who suggested the use of an exchangeable sequence, with a Dirichlet Process prior with scale parameter θ. Under this model, the estimated coverage, expected number of new species and discovery rate are:

$$\hat{C}_n = \frac{n}{\theta + n}, \quad \mathbb{E}\left[K_m^{(n)}\right] = \sum_{i=n}^{n+m} \frac{\theta}{\theta + i}, \quad D_m^{(n)} = \frac{\theta}{\theta + n + m}.$$

Unfortunately, this model is not appropriate for many applications, as the discovery rate, as a function of m, goes to zero too fast.

In recent works by Lijoi et al. [2007] and Favaro et al. [2009], the same estimators were calculated for a two-parameter Poisson Dirichlet prior with parameters (θ, σ). In this case, given the observed sample of size n with k

distinct classes,

$$\begin{aligned}\hat{C}_n &= 1 - \frac{\theta + k\sigma}{\theta + n},\\ \mathbb{E}\left[K_m^{n,k} | K_n = k\right] &= \left(k + \frac{\theta}{\sigma}\right)\left(\frac{(\theta+n+\sigma)_m}{(\theta+n+1)_m} - 1\right),\\ \hat{D}_m^{(n,k)} &= \frac{\theta + k\sigma}{\theta + n}\frac{(\theta+n+\sigma)_m}{(\theta+n+1)_m},\end{aligned} \tag{4.6}$$

where $(a)_m = (a)(a+1)...(a+m-1)$. These estimators are more attractive, as the discovery rate depends on the observed number of classes and does not go to zero so fast. Another nice advantage is that they can be easily computed for any size of n and m. This is important because in some problems, for example in genomics, n and m can be very large.

In a problem of bivariate species sampling, we assume that two populations are dependent, so observed values can be shared by them, and samples from one population could also provide information about the other population. We are not only interested in univariate quantities referring to each population, as specified before, but also in quantities that specifically arise in the bivariate case, such as $K_{m,m}^{(n,n)}$, the amount of new species in a complete future sample of size m, or the number of common species shared by the two populations.

Using independent univariate species sampling models, such as two-parameter Poisson-Dirichlet priors, might give good univariate estimates, but it would clearly overestimate $K_{m,m}^{(n,n)}$, as it cannot take into account the shared species. However, as discussed in Chapter 2, all known bivariate species sampling models are too complicated to be used efficiently and fast for such type of problems. Furthermore, the assumption of partial exchangeability, which implies stationarity, may be too restrictive in applications where data are collected over time and the populations have some form of evolution or temporary disequilibrium before returning to a stationary situation.

The generalized bivariate species sampling sequence defined in Section 3.3.2 is mainly constructed through the functions r_n and h_n, which correspond to the behavior of the new class discovery. This suggests the utility of this model for problems where investigating the appearance of a new classes is a main task. We propose a bivariate species sampling process that generates partially c.i.d. sequences and has a simple prediction rule which allows flexibility in defining the discovery rate.

4.2.1 Simple generalized bivariate species sampling sequence

The model that we propose is a particular case of the generalized bivariate species sampling sequences (Definition 3.3.8), where we assume that the predictive weights in the prediction rules do not depend on the observed data. More precisely, let $(X_n, Y_n)_{n\geqslant 1}$ be two sequences of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, with values in $\mathbb{X}$. Notice that, unlike the sequences considered in the previous section, these are not binary.

Definition 4.2.1. *We say that a generalized bivariate species sampling sequence $(X_n, Y_n)_{n\geqslant 1}$ is a simple generalized bivariate species sampling sequence if it is partially c.i.d. with bivariate prediction rule that satisfies the following constraint:*

$$\begin{aligned}
&X_1 \sim \nu,\ Y_1 \sim \nu,\ \textit{and for}\ n \geqslant 1,\\
&\mathbb{P}\left[X_{n+1} \in \cdot | X_{1:n}, Y_{1:n}\right] = \sum_{i=1}^{n} p_{n,i}\delta_{X_i}(\cdot) + \sum_{i=1}^{n} p'_{n,i}\delta_{Y_i}(\cdot) + r_n\nu(\cdot),\\
&\mathbb{P}\left[Y_{n+1} \in \cdot | X_{1:n}, Y_{1:n}\right] = \sum_{i=1}^{n} g_{n,i}\delta_{X_i}(\cdot) + \sum_{i=1}^{n} g'_{n,i}\delta_{Y_i}(\cdot) + h_n\nu(\cdot),
\end{aligned} \tag{4.7}$$

where functions $p_{n,i}$, $p'_{n,i}$, r_n, $g_{n,i}$, $g'_{n,i}$ and h_n do not depend on the bipartition Π_{nn}.

Although the proposed model is a special case of generalized bivariate species sampling sequence, it is more flexible than the model we discussed at the end of Section 3.3.2. Clearly, simple generalized bivariate species sampling sequences are not partially exchangeable, apart from one particular case.

Proposition 4.2.2. *The prediction rule* (4.7) *generates a partially exchangeable sequence* $(X_n, Y_n)_{n\geqslant}$ *if and only if*

$$p_{n,i} = p'_{n,i} = g_{n,i} = g'_{n,i} = \frac{\theta}{\theta + n + n} \, for \, i = 1, ..., n.$$

At the same time, not any prediction rule of the form (4.7) generates a partially c.i.d. sequence. The next result provides necessary and sufficient conditions.

Theorem 4.2.3. *A prediction rule* (4.7) *generates a partially c.i.d. sequence with respect to the natural filtration* $\mathcal{G} = \sigma(X_{1:n}, Y_{1:n})_{n\geqslant 0}$ *if and only if the functions* $p_{n,i}$, $p'_{n,i}$ *and* $g_{n,i}$, $g'_{n,i}$ *are positive and satisfy*

$$\begin{aligned}
&p_{n,i} = p_{n-1,i}(1 - p_{n,n}) - g_{n-1,i}p'_{n,n} && for \, i = 1, ..., n-1,\\
&p'_{n,i} = p'_{n-1,i}(1 - p_{n,n}) - g'_{n-1,i}p'_{n,n} && for \, i = 1, ..., n-1,\\
&r_n = r_{n-1}(1 - p_{n,n}) - h_{n-1}p'_{n,n}, &&\\
&g_{n,i} = g_{n-1,i}(1 - g'_{n,n}) - p_{n-1,i}g_{n,n} && for \, i = 1, ..., n-1,\\
&g'_{n,i} = g'_{n-1,i}(1 - g'_{n,n}) - p_{n-1,i}g_{n,n} && for \, i = 1, ..., n-1,\\
&h_n = h_{n-1}(1 - g'_{n,n}) - r_{n-1}g_{n,n}, &&
\end{aligned} \tag{4.8}$$

with $r_0 = 1$, $h_0 = 1$.

We would like to underline that the predictive weights are not allowed to depend on the observed bi-partition. Without this restriction, conditions (4.8) become intractable.

Proof. In order to prove the theorem, we have to understand when the martingale condition (3.15) holds. Let us start by checking when

$$\mathbb{E}\left(\mathbb{E}\left[f(X_{n+2}|\mathcal{G}_{n+1})\right]|\mathcal{G}_n\right) = \mathbb{E}\left[f(X_{n+1}|\mathcal{G}_n)\right].$$

The expression on the left can be rewritten as

$$\mathbb{E}\left(\mathbb{E}\left[f(X_{n+2}|\mathcal{G}_{n+1})\right]|\mathcal{G}_n\right) = \sum_{i=1}^{n} p_{n+1,i}f(X_i) + p_{n+1,n+1}\mathbb{E}\left[f(X_n+1)|\mathcal{G}_n\right] +$$
$$+\sum_{i=1}^{n} p'_{n+1,i}f(Y_i) + p'_{n+1,n+1}\mathbb{E}\left[f(Y_n+1)|\mathcal{G}_n\right] + r_{n+1}\mathbb{E}f(X_1) =$$
$$= \sum_{i=1}^{n} p_{n+1,i}f(X_i) + \sum_{i=1}^{n} p'_{n+1,i}f(Y_i) + r_{n+1}\mathbb{E}f(X_1) +$$
$$+ p_{n+1,n+1}\left(\sum_{i=1}^{n} p_{n,i}f(X_i) + \sum_{i=1}^{n} p'_{n,i}f(Y_i) + r_n\mathbb{E}f(X_1)\right) +$$
$$+ p'_{n+1,n+1}\left(\sum_{i=1}^{n} g_{n,i}f(X_i) + \sum_{i=1}^{n} g'_{n,i}f(Y_i) + h_n\mathbb{E}f(X_1)\right) =$$
$$= \sum_{i=1}^{n}\left(p_{n+1,i} + p_{n+1,n+1}p_{n,i} + p'_{n+1,n+1}g_{n,i}\right) f(X_i) +$$
$$+ \sum_{i=1}^{n}\left(p'_{n+1,i} + p_{n+1,n+1}p'_{n,i} + p'_{n+1,n+1}g'_{n,i}\right) f(Y_i) +$$
$$+ \left(r_{n+1} + p_{n+1,n+1}r_n + p'_{n+1,n+1}h_n\right)\mathbb{E}f(X_1).$$

This sum will be equal to

$$\mathbb{E}\left[f(X_{n+1}|\mathcal{G}_n)\right] = \sum_{i=1}^{n} p_{n,i}f(X_i) + \sum_{i=1}^{n} p'_{n,i}f(Y_i) + r_n\mathbb{E}f(X_1),$$

for any function f, if and only if

$$p_{n+1,i} + p_{n+1,n+1}p_{n,i} + p'_{n+1,n+1}g_{n,i} = p_{n,i} \text{ for } i = 1,...,n-1,$$
$$p'_{n+1,i} + p_{n+1,n+1}p'_{n,i} + p'_{n+1,n+1}g'_{n,i} = p'_{n,i} \text{ for } i = 1,...,n-1,$$
$$r_{n+1} + p_{n+1,n+1}r_n + p'_{n+1,n+1}h_n = r_n.$$

The same procedure is applied for the second equation,

$$\mathbb{E}\left(\mathbb{E}\left[f(Y_{n+2}|\mathcal{G}_{n+1})\right]|\mathcal{G}_n\right) = \mathbb{E}\left[f(Y_{n+1}|\mathcal{G}_n)\right].$$

Note, that these rules ensure that the weights sum to one. It can be easily checked by induction: for $n = 1$ it is true. If we assume it is true for $n-1$, then

$$\begin{aligned}
&\sum_{i=1}^{n} p_{n,i} + \sum_{i=1}^{n} p'_{n,i} + r_n = \\
&= p_{n,n} + p'_{n,n} + \sum_{i=1}^{n-1} \left(p_{n-1,i}(1-p_{n,n}) + g_{n-1,i}p'_{n,n}\right) + \\
&+ \sum_{i=1}^{n-1} \left(p'_{n-1,i}(1-p_{n,n}) + g'_{n-1,i}p'_{n,n}\right) + r_{n-1}(1-p_{n,n}) - h_{n-1}p'_{n,n} = \\
&= (1-p_{n,n}) \left(\sum_{i=1}^{n-1} p_{n-1,i} + \sum_{i=1}^{n-1} p'_{n-1,i} + r_{n-1} \right) + \\
&+ p'_{n,n} \left(\sum_{i=1}^{n-1} g_{n-1,i} + \sum_{i=1}^{n-1} g'_{n-1,i} + h_{n-1} \right) + p_{n,n} + p'_{n,n} = 1.
\end{aligned}$$

□

Condition (4.8) is more complicated than the condition (3.7) specified in Bassetti et al. [2010] for generalized Ottawa sequences, as it allows the weights for a group to depend on the weights for the other groups. Due to this complexity, we have to define the weights sequentially.

We are discussing a problem of species discovery, so we propose to construct weights that satisfy condition (4.8) by choosing a decreasing set of positive functions $(r_n)_{n\geqslant 0}$, $(h_n)_{n\geqslant 0}$, that correspond to the probabilities of observing new species. Then, we can specify $p_{n,n}$, $p'_{n,n}$ in such a way that they will satisfy $r_n = r_{n-1}(1-p_{n,n}) - h_{n-1}p'_{n,n}$, and specify $g_{n,n}$ and $g'_{n,n}$ such that they satisfy $h_n = h_{n-1}(1-g'_{n,n}) - r_{n-1}g_{n,n}$. One should also check that the resulting weights are positive. Let us consider some extreme cases:

- No shared values are allowed, i.e., $p'_{ni} = g_{ni} = 0$. Then, condition (4.8) becomes:

$$\begin{aligned}
&p_{n,i} = p_{n-1,i}(1-p_{n,n}) \quad \text{for } i = 1, ..., n-1, \\
&r_n = r_{n-1}(1-p_{n,n}), \\
&g'_{n,i} = g'_{n-1,i}(1-g'_{n,n}) \quad \text{for } i = 1, ..., n-1, \\
&h_n = h_{n-1}(1-g'_{n,n}),
\end{aligned}$$

which corresponds to two independent simple generalized Ottawa processes (discussed in Section 3.2.2).

- The first sequence can borrow values, while the second cannot, i.e., $g_{n,i} = 0$. Then, the prediction takes the form

$$\mathbb{P}\left[X_{n+1} \in \cdot | X_{1:n}, Y_{1:n}\right] = \sum_{i=1}^{n} p_{n,i} \delta_{X_i}(\cdot) + \sum_{i=1}^{n} p'_{n,i} \delta_{Y_i}(\cdot) + r_n \mu(\cdot),$$
$$\mathbb{P}\left[Y_{n+1} \in \cdot | X_{1:n}, Y_{1:n}\right] = \sum_{i=1}^{n} g'_{n,i} \delta_{Y_i}(\cdot) + h_n \mu(\cdot),$$

and condition (4.8) becomes:

$$\begin{aligned}
&p_{n,i} = p_{n-1,i}(1 - p_{n,n}) \quad \text{for } i = 1, ..., n-1,\\
&p'_{n,i} = p'_{n-1,i}(1 - p_{n,n}) - g_{n-1,i} p'_{n,n} \quad \text{for } i = 1, ..., n-1,\\
&r_n = r_{n-1}(1 - p_{n,n}) - h_{n-1} p'_{n,n},\\
&g'_{n,i} = g'_{n-1,i}(1 - g_{n,n}) \quad \text{for } i = 1, ..., n-1,\\
&h_n = h_{n-1}(1 - g'_{n,n}).
\end{aligned}$$

This means that the second sequence is again a simple generalized Ottawa sequence, while the first one can share values and, therefore, is not of the generalized Ottawa sequence type.

Depending on the type of problem, one can choose different structures. Returning to the problem of bivariate sampling, recall that we will estimate the following characteristics: C_n^I and C_n^{II}, coverages of the first and second groups respectively; the expected number of new classes to be discovered in the complete sample of size m, given the complete observed sample of size n; and the discovery rates $D_m^{(n)I}$ and $D_m^{(n)II}$ for each group. Assuming that the sequence $(X_{1:n}, Y_{1:n})$ is a simple generalized bivariate species sampling sequence, some estimators are immediately identified:

- The coverages are given by

$$\hat{C}_n^I = 1 - r_n \ \ and \ \ \hat{C}_n^{II} = 1 - h_n. \tag{4.9}$$

- The discovery rates $D_m^{I,n}$, $D_m^{II,n}$, i.e., the probabilities that the $(n+m+1)$th observation will yield a new value, without observing the m intermediate records, in the first and second group, respectively are

$$\hat{D}_m^{I,n} = r_{n+m} \ \ and \ \ \hat{D}_m^{II,n} = h_{n+m}. \tag{4.10}$$

However, the expected number of new classes in the next complete sample, $E_{m,m}^{(n,n)} = \mathbb{E}\left(K_{m,m}^{(n,n)}\right)$, is not so easy to calculate, and some additional analysis is needed in order to understand the distribution of $K_{m,m}^{(n,n)}$ for simple generalized bivariate species sampling sequences. We shall do this in the following section, for the general case.

4.2.2 Clustering properties of generalized bivariate species sampling sequences

Let $(X_n, Y_n)_{n\geqslant 1}$ be a generalized bivariate species sampling sequence (Definition 3.3.8) where, without loss of generality, we assume X_i to be always observed earlier than Y_i. Thus, we can combine the two sequences in single sequence, say $(Z_n)_{n\geqslant 1}$, in the following way:

$$(X_1, Y_1, X_2, Y_2, ...) \text{ is recoded into } \ (Z_1, Z_2, Z_3, Z_4, ...)\,,$$

where $Z_i = X_{(i+1)/2}$ if i is odd, $Z_i = Y_{i/2}$ if i is even. Let $(\Pi_n)_{n\geqslant 1}$ be a random partition generated from $(Z_{1:n})_{n\geqslant 1}$ and denote by $(K_n)_{n\geqslant 1}$ the length of this partition. In other words, K_{2n} is the number of distinct classes that has been observed among $(X_{1:n}, Y_{1:n})$. Thus, $(K_n)_{n\geqslant 1}$ is a sequence of random variables with values in $\mathbb{Z}^+$ that characterizes the clustering properties of $(Z_n)_{n\geqslant 1}$. Note that the behavior of $(K_n)_{n\geqslant 1}$ depends only on the functions r_n, h_n involved in the prediction rule (4.7). We assume that $K_0 = 0$, $K_1 = 1$, $K_2 = 2$, since for a bivariate generalized species sampling sequence the first two elements are drawn independently from a diffuse distribution ν.

We define an additional set of binary random variables $U_j = K_j - K_{j-1}$ for $j \geqslant 1$, so that $K_{2n} = \sum_{j=1}^{2n} U_j$. Therefore, the probability law of

U_j given the observed data (recall that for generalized bivariate species sampling sequence, r_n and h_n can depend on the partition) is

$$U_j = \begin{cases} 1 & \text{with probability } S_{j-1}; \\ 0 & \text{with probability } 1 - S_{j-1}, \end{cases}$$

where

$$S_j = \mathbb{P}\left(Z_j \text{ belongs to new class} | Z_{1:j-1}\right) = \begin{cases} r_{(j+1)/2}, \text{ if } j \text{ is odd;} \\ h_{j/2}, \text{ if } j \text{ is even.} \end{cases}$$

The joint distribution of $(U_i)_{i=1}^n$ is then given, for $(e_1, .., e_n) \in \{0, 1\}^n$, by

$$\mathbb{P}\left(U_1 = e_1, U_{2=1}, ..., U_n = e_n | r_{1:n}, h_{1:n}\right) = \prod_{i=3}^{2n} \left[S_i^{e_i}(1 - S_i)^{1-e_i}\right].$$

The distribution of K_n is obtained as

$$\mathbb{P}\left(K_{n+2} = k + 2\right) = \mathbb{P}\left(\sum_{j=3}^{n+2} U_j = k\right) = \sum_{\underline{e}} \mathbb{E}\left(\prod_{i=3}^{n+2} \left[S_i^{e_i}(1 - S_i)^{1-e_i}\right]\right),$$

where the summation is taken over the set of all sequences $\underline{e} = (e_1, .., e_n) \in \{0, 1\}^n$ such that $\sum_{i=1}^n e_i = k$.

The main formulas are the same as for generalized Ottawa sequences, therefore, according to Costa et al. [2013],

$$\mathbb{E}\left((K_{n+2} - 2)^l\right) = \mathbb{E}\left((\sum_{j=3}^{n+2} U_j)^l\right) = \sum_{m=1}^{n \wedge l} m! Stirling(k, m) \phi(n, m), \quad (4.11)$$

where $Stirling(l, m)$ is a Stirling number of the second type, i.e.

$$Stirling(l, m) = \frac{l!}{m!} \sum_{\{n_i > 0 : \sum n_i = l\}} \frac{1}{n_1! ... n_m!},$$

and

$$\phi(n, m) = \sum_{1 \leqslant l_1 < ... < l_m \leqslant n} \left(S_{l_1} ... S_{l_m}\right),$$

$$\mathbb{E}\left(e^{-tK_{n+2}}\right) = e^{-t} \sum_{m=0}^{n} (e^{-t} - 1)^m \phi(n, m),$$

with $\phi(n,0)=1$.

In such a way, the expected value of $(K_n)_{n\geqslant 1}$ for a simple generalized bivariate species sampling sequence can be expressed as:

$$\mathbb{E}\left(K_{n+2}-2\right)=1!S(1,1)\phi(n,1)=\mathbb{E}\left(\sum_{i=1}^{n-1} r_i\right)+\mathbb{E}\left(\sum_{i=1}^{n-1} h_i\right),$$

so that

$$\mathbb{E}\left(K_{n+2}\right)=2+\sum_{i=1}^{n-1}\left(r_i\right)+\sum_{i=1}^{n-1}\left(h_i\right).$$

This implies that the bivariate partition generated by a simple generalized bivariate species sampling sequence has independent increments, i.e., the events that the observations are old or new are independent, despite the fact that the process is defined through two dependent sequences. Here, we follow the terminology induced by Nacu [2006]. In particular, Nacu has shown that the only exchangeable partition with independent increments is a partition induced by a Dirichlet Process prior. Independence, in our case, can be explained as a consequence of the fact that the predictive functions in (4.7) are not allowed to depend on the observed partition.

Thus, the expected number of new classes in an additional sample of size m, for a simple generalized bivariate species sampling sequence is

$$E_{m,m}^{n,n}=\mathbb{E}\left[K_{m,m}^{(n,n)}\right]=\sum_{i=n}^{n+m-1} r_i+\sum_{i=n}^{n+m-1} h_i.$$

Therefore, using an appropriate set of functions, $(r_n)_{n\geqslant 1}$ and $(h_n)_{n\geqslant 1}$, one could easily model any type of behavior of a new class discovery.

4.2.3 Example: Model-1

As we have discussed before, one can construct a simple generalized bivariate species sampling sequence using a set of functions, $(r_n)_{n\geqslant 1}$ and $(h_n)_{n\geqslant 1}$, that do not depend on the observed partition. A natural idea is to use

$$r_n=\frac{\theta_1}{\theta_1+n} \text{ and } h_n=\frac{\theta_2}{\theta_2+n},$$

as for usual Dirichlet Processes (3.10) with parameters $\theta_1, \theta_2 > 0$. But then, r_n and h_n, as functions of n, would go to zero too fast as n grows, and so would the corresponding discovery rates. This suggests the introduction of some additional parameters that will preserve the discovery rate from fast decreasing. Given parameters $s, t \in [0,1]$ we define

$$r_n = \tfrac{\theta_1}{n^s+\theta_1} \quad \text{and} \quad h_n = \tfrac{\theta_2}{n^t+\theta_2}, \tag{4.12}$$

making the functions $(r_n)_{n\geqslant 1}$ and $(h_n)_{n\geqslant 1}$ more flexible. This example can be regarded as an extension of the Dirichlet Process, but at the same time, asymptotically, it has some connections with the two-parameter Poisson Dirichlet process.

Recall that the two-parameter Poisson-Dirichlet process was constructed as a generalization of the Dirichlet process, to accommodate power-law tails. In other words, the second parameter σ was introduced in order to make the tails more flexible and keep them from decreasing too fast. Indeed, the asymptotic behavior of the total number of classes in Dirichlet and Poisson-Dirichlet Processes are very different. Under a usual one-parameter Dirichlet prior $\frac{K_n}{log(n)} \to \theta$ $\mathbb{P}$-a.s. for $n \to \infty$, while for a two-parameter Poisson Dirichlet Process $\frac{K_n}{n^\sigma} \to \mathcal{L}$ $\mathbb{P}$-a.s. for $n \to \infty$, for some random variable $\mathcal{L}$, the distribution of which is given in Pitman [2006].

An interesting fact was noted by Bassetti et al. [2010], namely if one considers a generalized Ottawa sequence and specifies $r_n = \frac{\theta}{\theta+n^s}$ for some $s \in [0,1]$, then the asymptotic behavior of the number of new classes under this model is given by

$$\frac{K_n}{n^{1-s}} \underset{n\to\infty}{\longrightarrow} \frac{\theta}{1-s} \ \mathbb{P}\text{-a.s.}$$

The same speed of convergence is obtained under a two-parameter Poisson-Dirichlet prior with $\sigma = 1 - s$. The difference is that, for the example of generalized Ottawa sequence, the limit is not random. Following Aoki [2008], we say that K_n has the self-averaging property when its coefficient of variation tends to zero as the sample size becomes very large.

The two-parameter Poisson-Dirichlet instead, is an example of a non self-averaging K_n, i.e., its coefficient of variation does not decrease to zero even for a large sample size. This also means that the behavior of K_n depends on the sample and so, however large, a sample from such process is never a good description of the whole ensemble. We conclude that the form of $(r_n)_{n\geqslant 1}$ and $(h_n)_{1\geqslant 1}$ proposed in (4.12) is an interesting self-averaging extension of a two-parameter Poisson Dirichlet process.

Let us now specify the rest of the weights. It is worth noting that the relationship between $p_{n,n}$ and $p'_{n,n}$, $g_{n,n}$ and $g'_{n,n}$ affects on connections between $(p_{n,i})_{i=1}^n$ and $(p'_{n,i})_{i=1}^n$, $(g_{n,i})_{i=1}^n$ and $(g'_{n,i})_{i=1}^n$. In a general case this affect is not evident. We define this relation as

$$\begin{aligned} p'_{n,n} &= \alpha_n p_{nn}, \text{ where } \alpha_n = 1/n^s, \text{ so } p_{n,n} = \tfrac{r_{n-1}-r_n}{r_{n-1}-\alpha_n h_{n-1}}; \\ g_{n,n} &= \beta_n g'_{n,n}, \text{ where } \beta_n = 1/n^t, \text{ so } g'_{n,n} = \tfrac{h_{n-1}-h_n}{h_{n-1}-\beta_n r_{n-1}}. \end{aligned}$$

For $i = 1, ..., n-1$, the weights are defined through

$$\begin{aligned} p_{n,i} &= p_{n-1,i}(1 - p_{n,n}) - g_{n-1,i}p'_{n,n}, \\ p'_{n,i} &= p'_{n-1,i}(1 - p_{n,n}) - g'_{n-1,i}p'_{n,n}, \\ g_{n,i} &= g_{n-1,i}(1 - g'_{n,n}) - p_{n-1,i}g_{n,n}, \\ g'_{n,i} &= g'_{n-1,i}(1 - g'_{n,n}) - p_{n-1,i}g_{n,n}. \end{aligned}$$

It is clear that for $s = t = 1$, the random variable K_n^I has the same distribution as the random variable corresponding to the number of classes in a sample of size n from a usual Dirichlet Process prior with parameter θ_1. The same can be said about K_n^{II}, with parameter θ_2.

We will call Model-1 the model with prediction rule of the form (4.7) and weights defined as above, together with an additional condition on the parameters θ_1, θ_2, ξ_1, ξ_2, which we specify in the next subsection, to guarantee the positiveness of the weights.

Applying previous results for Model-1, the estimators for the usual characteristics of interest of a sample and the population are as follows:

- From expression (4.9), the estimated coverages become

$$\hat{C}_n^I = 1 - \frac{\theta_1}{\theta_1 + (n-1)^s} \; and \; \hat{C}_n^{II} = 1 - \frac{\theta_2}{\theta_2 + (n-1)^t}. \qquad (4.13)$$

- The expected number of additional new classes in a new sample of size $2m$ is
$$E_{m,m}^{(n,n)} = \sum_{i=n}^{n+m-1} \frac{\theta_1}{\theta_1 + i^s} + \sum_{i=n}^{n+m-1} \frac{\theta_2}{\theta_2 + i^t}. \tag{4.14}$$

- From expression (4.10), the estimated discovery rates $\hat{D}_m^{I,n}$, $\hat{D}_m^{II,n}$, i.e., the probability that the $(n+m+1)$-th observation will yield a new value, without observing m intermediate records, in the first group and second groups, respectively, are
$$D_m^{I,n} = \frac{\theta_1}{\theta_1 + (n+m)^s} \ \ and \ \ D_m^{II,n} = \frac{\theta_2}{\theta_2 + (n+m)^t}. \tag{4.15}$$

4.2.4 The condition on the positivity of the parameters

In order to be sure that the proposed weights can be used in the prediction rule, we should check the positiveness of r_n, h_n, $(p_{n,i}, p'_{n,i})_{i=1}^n$ and $(g_{n,i}, g'_{n,i})_{i=1}^n$. From expression (4.12), it is clear that r_n and h_n are always positive if θ_1, $\theta_2 > 0$ and $0 \leqslant s, t \leqslant 1$. Moreover $r_n > r_{n+1}$ and $h_n > h_{n+1}$ for every $n \geqslant 0$. Recall that
$$p_{n,n} = \tfrac{r_{n-1} - r_n}{r_{n-1} - \alpha_n h_{n-1}} \text{ and } g'_{n,n} = \tfrac{h_{n-1} - h_n}{h_{n-1} - \beta_n r_{n-1}},$$
where $\alpha_n = 1/n^s$ and $\beta_n = 1/n^t$. Then, $p_{n,n}$ and $g'_{n,n}$ are positive if and only if, for every n,
$$\alpha_n h_{n-1} < r_{n-1} < \frac{1}{\beta_n} h_{n-1}.$$
Therefore, the following condition should hold:
$$\frac{1}{n^s} < \frac{\theta_1(\theta_2 + (n-1)^t)}{\theta_2(\theta_1 + (n-1)^s)} < n^t. \tag{4.16}$$

The positiveness of $p'_{n,n} = \alpha_n p_{n,n}$ and $g_{n,n} = \beta_n g'_{n,n}$ follows from the positiveness of $(p_{n,n})_{n \geqslant 1}$ and $(g'_{n,n})_{n \geqslant 1}$. Let us consider other weights $(p_{n,i}, p'_{n,i})_{i=1}^{n-1}$ and $(g_{n,i}, g'_{n,i})_{i=1}^{n-1}$. It is clear, that $p_{n,i} > p'_{n,i}$ and $g'_{n,i} > g_{n,i}$ for $i = 1, .., n-1$. Thus, it is enough to show the positiveness of $(p'_{n,i})_{i=1}^{n-1}$ and $(g_{n,i})_{i=1}^{n-1}$.

First, $p'_{n+1,n}$ is positive, if and only if

$$p'_{n,n} > g'_{n,n}\frac{p'_{n+1,n+1}}{1-p_{n+1,n+1}}.$$

We denote $a_1^I = p'_{n+1,n+1}$ and $a_1^{II} = 1 - p_{n+1,n+1}$.

At the same time, $p'_{n+2,n}$ is positive if and only if

$$p'_{n,n} > g'_{n,n}\frac{a_1^I(1-p_{n+2,n+2}) + p'_{n+2,n+2}(1-g'_{n+1,n+1})}{a_1^{II}(1-p_{n+2,n+2}) + p'_{n+2,n+2}(g_{n+1,n+1})} = g'_{n,n}\frac{a_2^I}{a_2^{II}}.$$

Repeating this procedure k times one can get that $p'_{n+k,n+k}$ is positive if and only if $p'_{n,n} > g'_{n,n}a_k^I/a_k^{II}$, where

$$\begin{aligned} a_k^I &= a_{k-1}^I(1-p_{n+k,n+k})+ \\ &+ p'_{n+k,n+k}\left(b_{k-1}^I(1-g'_{n+k-1,n+k-1}) + g_{n+k-1,n+k-1}a_{k-2}^I\right), \\ a_k^{II} &= a_{k-1}^{II}(1-p_{n+k,n+k})+ \\ &+ p'_{n+k,n+k}\left(b_{k-1}^{II}(1-g'_{n+k-1,n+k-1}) + g_{n+k-1,n+k-1}a_{k-2}^{II}\right), \end{aligned}$$

and

$$\begin{aligned} b_k^I &= \left(a_{k-1}^I - a_{k-2}^I(1-p_{n+k-1,n+k-1})\right)/p'_{n+k-1,n+k-1}, \\ b_k^{II} &= \left(a_{k-1}^{II} - a_{k-2}^{II}(1-p_{n+k-1,n+k-1})\right)/p'_{n+k-1,n+k-1}. \end{aligned}$$

Thus, positiveness of $(p'_{n+i,n})_{i\geqslant 1}$ is guaranteed if

$$p'_{n,n} > g'_{n,n}max\left(a_1, a_2, ...\right),$$

and this should hold for every $n \geqslant 1$. Analogous condition is required for $(g_{n,i})_{i=1}^{n-1}$.

The role of the parameters in Model-1

It is clear that different values of the parameters θ_1, θ_2, s and t lead to different behaviors of the model. Roughly speaking, if θ_1 is small and s is close to one, the probability of observing a new species in the first group will be small and will decrease fast as n grows. If instead, θ_1 is large and s is close to zero, chances to observe new classes in this group will be large and decrease slowly. At the same time, if s and t are close to zero, $p_{n,n}$ is close $p'_{n,n}$, $g_{n,n}$ is close $g'_{n,n}$. It implies that, for $i = 1, ..., n-1$, $p_{n,i}$ and $p'_{n,i}$ are close, $g_{n,i}$ and $g'_{n,i}$ are close. If, instead, s and t are close to one, then $p_{n,i}$ is much larger than $p'_{n,i}$, $g'_{n,i}$ is much larger than $g_{n,i}$, for all $i = 1, ..., n$. In this subsection we illustrate the joint properties of the data generated from Model-1 for a case where the two groups have a different behavior. In particular, we are interested in the behavior of new classes appearance and of the weights $(p_{n,i})_{i=1}^n$, $(p'_{n,i})_{i=1}^n$ and $(g_{n,i})_{i=1}^n$, $(g'_{n,i})_{i=1}^n$, since for these weights we cannot obtain exact formulas and have to define them sequentially.

We fixed parameters $\theta_1 = 6$, $\theta_2 = 3$, $s = 0.4$, $t = 0.7$ and simulated two groups of data of size 1000. We expect the first group to have more distinct classes than the second one. Figure 4.1 illustrates the whole sample with observations from the first group in red and observations from the second group in blue. Given the stated ordering of the sequence $(X_n, Y_n)_{\geqslant 1}$, observations were simulated one by one – one from the first group, then one from the second and so on. However, in Figure 4.1 we first plot the first group, then the second. Each point on this plot represents an observation, with the x-coordinate corresponding to the index of the observation and the y-coordinate corresponding to the cluster (in order of appearance) to which the observation belongs. Therefore, observations on the same level on the plot are of the same type. For example, the first class was very frequently observed in both groups. In total, 437 distinct classes were observed. As expected, in the first group new classes are generated more frequently, while observations in the second group tend to join classes that have been already observed.

Figure 4.2 shows cluster sizes of the corresponding bi-partition, that is, each point in the plot represents a cluster, the x-coordinate corresponding to the number of observations of such type in the first group, the y-coordinate corresponding to the number of observations of such type in the second group. For instance, if the y-coordinate is equal to zero it means that this class has not been observed in the second group. Again, in the first group there are several classes which have not been observed

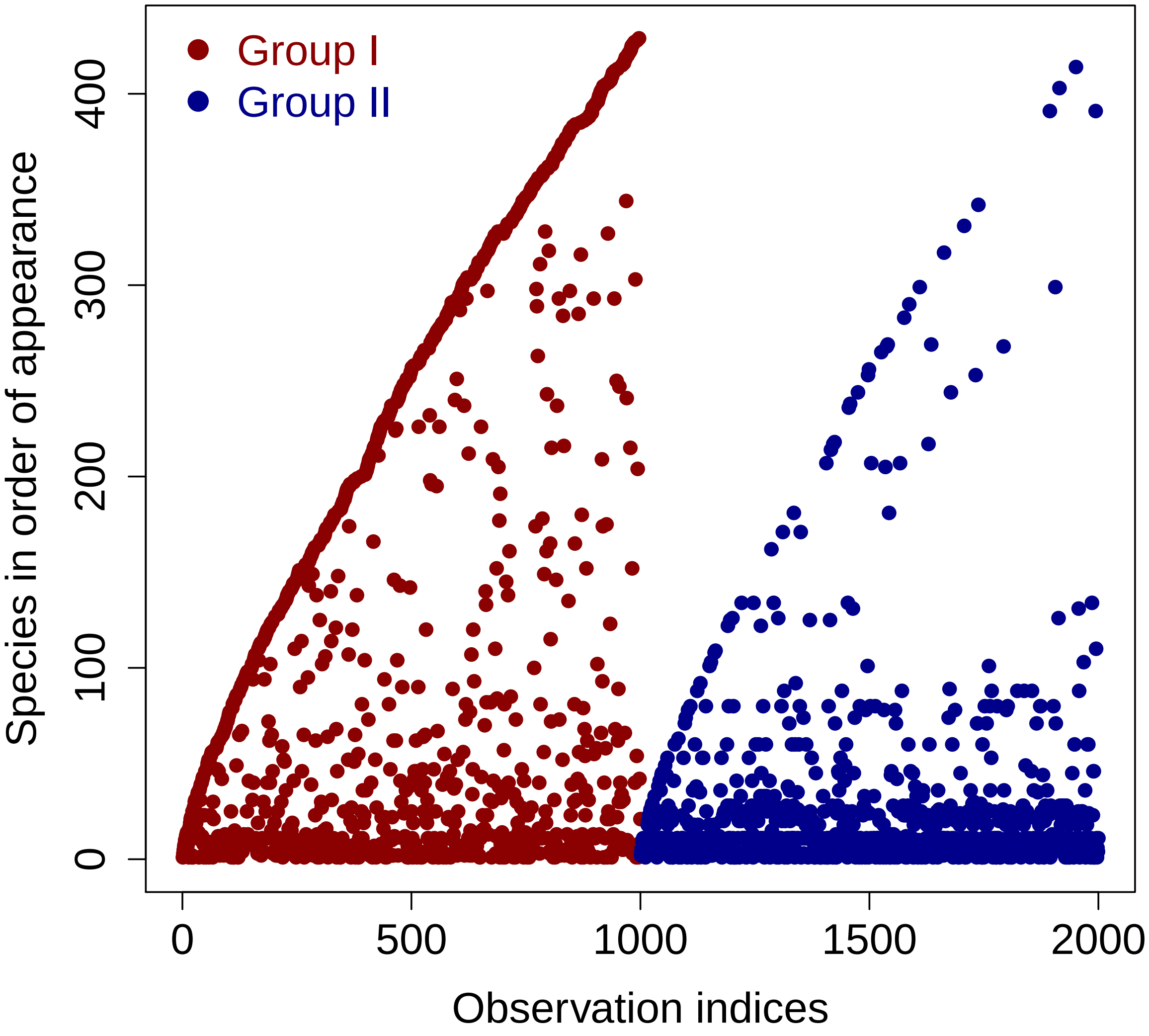


Figure 4.1: Bivariate sample of size $n = 1000$ from Model-1 with parameters $\theta_1 = 6$, $s = 0.4$, $\theta_2 = 3$, $t = 0.7$. Each point corresponds to one observation.

in the second group. Figure 4.3 shows the realization of K_n^I, the number of distinct species at time n in the first group, and K_n^{II}, the number of distinct species at time n in the second group, as functions of n. Again, as expected, many more distinct species are generated in the first group than in the second.

The simulation study illustrates that, although these two samples are

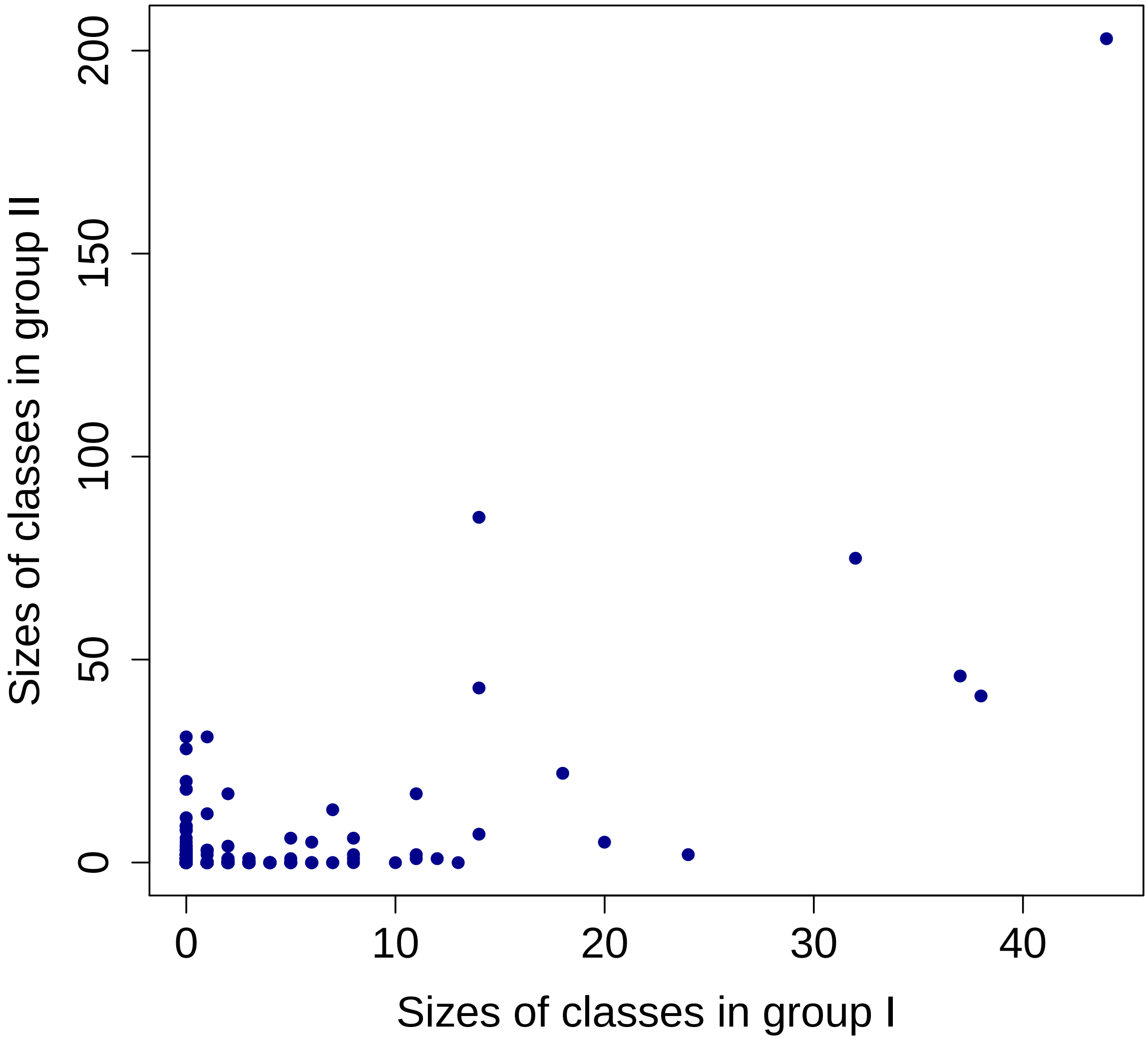


Figure 4.2: Cluster sizes in the bi-partition generated from the simulated data.

allowed to share values, the properties of the corresponding bi-partition are mainly based on the parameters that define $(r_n)_{n\geqslant 0}$ and $(h_n)_{n\geqslant 0}$.

Now, we describe the behavior of the other weights, $(p_{n,i})_{i=1}^n$, $(p'_{n,i})_{i=1}^n$ and $(g_{n,i})_{i=1}^n$, $(g'_{n,i})_{i=1}^n$ based on simulation, since we cannot obtain explicit formulas for them. In Figure 4.4a we plot the probabilities that X_{n+1} will join X_2 and Y_2, respectively, for $n = 1, .., 1000$. The behavior of these weights is quite similar, they are both monotonically decreasing with the

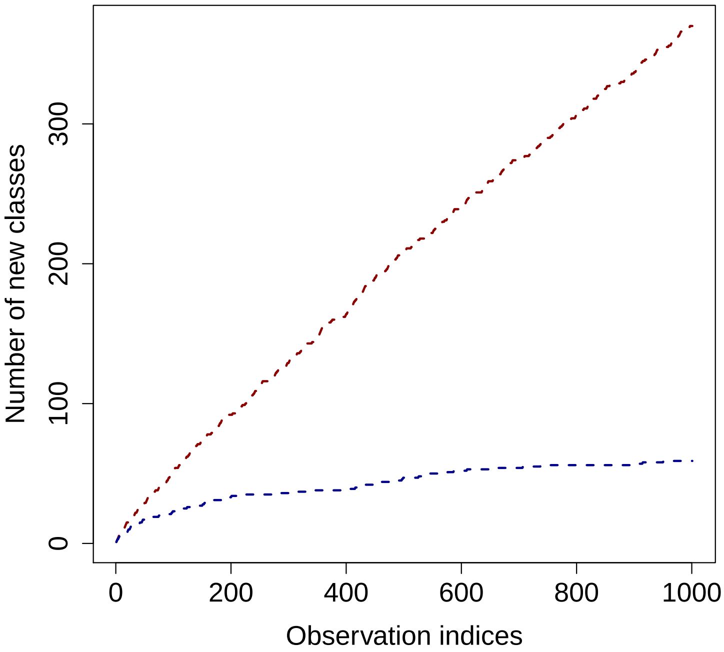
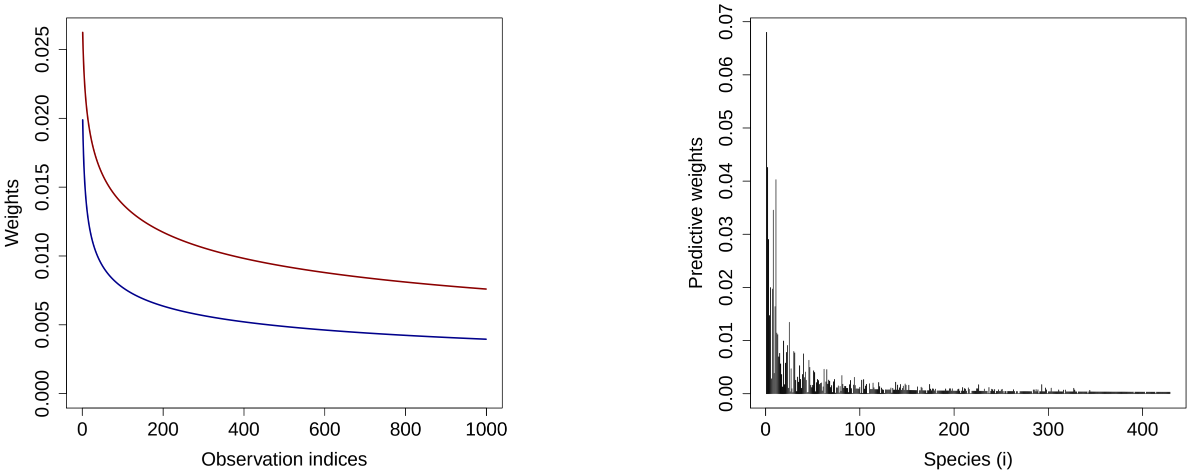


Figure 4.3: Behavior of the number of distinct species in group I (in red), and group II (in blue). Simulated data.



(a) Behavior of the weights $p_{n,2}$ (in red), the probability of joining X_2, and $p'_{n,2}$ (in blue), the probability of joining Y_2, as functions of n.

(b) Predictive weights corresponding to the probability that X_{n+1} is equal to one of the i-th observed species.

Figure 4.4: Behavior of predictive weights.

same speed. Recall that these weights do not depend on the realization of the data. In order to analyze the predictive probability that X_{1001} will be equal to one of the i-th observed clusters, we should take for each $i = 1, ..., 437$, the total number of classes, the sum over all the weights that correspond to cluster i in the first and second groups. We plot the results in Figure 4.4b. The first vertical line in the plot represents the predictive probability that X_{1001} is equal to the first observed species, and analogously for each of the 437 observed species. We can observe the general tendency of recent species to be more rare than old ones.

4.2.5 Inference on the parameters of Model-1

Suppose that two samples of size n from some non homogeneous population of species have been observed, and assume that the data are "physically" came from Model-1. In order to solve the different predictive issues concerning the composition of the population, then the parameters involved in Model-1 should be estimated. Let us denote by (Θ_1, Θ_2) the set of parameters, i.e., $\Theta_1 = (\theta_1, s)$, $\Theta_2 = (\theta_2, t)$. We start with a classical maximum likelihood approach, followed by a slightly different and much faster approach that also gives good estimators if the sample size is relatively large.

Suppose we have a bivariate sample $(X_{1:n}, Y_{1:n})$ from Model-1. According to the prediction rule (4.7), the distinct values in the sample are generated from some diffuse distribution ν, independently on the weights. Analyzing ν is not of interest in a problem of species sampling, since it only provides a set of labels. Consequently, we can ignore the exact values of $(X_{1:n}, Y_{1:n})$ and work instead with the random bi-partition $\Pi_{n,n}$ generated from $(X_{1:n}, Y_{1:n})$ according to the rule specified in Definition 2.2.1. We want to maximize the likelihood function of the bi-partition generated by $(X_{1:n}, Y_{1:n})$, i.e.,

$$\left(\hat{\Theta}_1, \hat{\Theta}_2\right) = \underset{(\Theta_1, \Theta_2)}{argmax} \left(\mathbb{P}\left[\Pi^{nn} = \left\{(A_i, B_i)_{i=1}^{k}\right\} \mid (\Theta_1, \Theta_2)\right]\right).$$

Note that for an exchangeable partition, the vector of sizes of the classes of a partition provides sufficient information about the whole partition, since the order does not matter. However, in our case, order matters, and $p_{n,i}$ and $p_{n,j}$ can be different even if they correspond to the same class. Moreover, if some class has been observed more than once in the whole sample then, when a new observation from the same class is observed, we cannot tell which previous observation from the same class the new observation is replacing.

To clarify this, let us consider a simple example. Denote by C_i^X the index of the class that holds X_i and by C_i^Y the index of the class that holds Y_i. Suppose that in the first group we have observed classes $C_{1:3}^X = (1, 2, 1)$ and in the second group classes $C_{1:3}^Y = (2, 3, 2)$. In total, 3 distinct classes have been observed. The probability of observing such event is

$$\begin{aligned}
&\mathbb{P}\begin{pmatrix} (1,\ 2,\ 1) \\ (2,\ 3,\ 2) \end{pmatrix} = \mathbb{P}(C_1^X = 1)\mathbb{P}(C_1^Y = 2) \\
&\mathbb{P}(C_2^X = 2 | C_1^X = 1, C_1^Y = 2)\,\mathbb{P}(C_2^Y = 3 | C_1^X = 1, C_1^Y = 2) \\
&\mathbb{P}(C_3^X = 1 | C_1^X = 1, C_1^Y = 2, C_2^X = 2, C_2^Y = 3) \\
&\mathbb{P}(C_3^Y = 2 | C_1^X = 1, C_1^Y = 2, C_2^X = 2, C_2^Y = 3).
\end{aligned}$$

Note that, Y_3 turned out to be of the same type as X_2 and Y_1. So, in order to calculate the desired probability, we should sum over all the weights corresponding to the previous observations of type 2, i.e.,

$$\mathbb{P}\begin{pmatrix} (1,\ 2,\ 1) \\ (2,\ 3,\ 2) \end{pmatrix} = 1 \cdot 1 \cdot p'_{1,1} \cdot h_1 \cdot p_{2,1} \cdot (g'_{2,1} + g_{2,2}).$$

Therefore, the function that we need to maximize is

$$\begin{aligned}
&\mathbb{P}\left[\Pi^{nn} = \left\{(A_i, B_i)_{i=1}^k\right\} | (\Theta_1, \Theta_2)\right] = \\
&= \prod_{i=2}^{n} \left[r_i^{I(X_i\, new)} \left(\sum_{j \in A_{C_i^X} \cap \{1..i-1\}} p_{ij} + \sum_{j \in B_{C_i^X} \cap \{1..i-1\}} p'_{ij} \right) \right] \\
&\prod_{i=2}^{n} \left[r_i^{I(Y_i\, new)} \left(\sum_{j \in A_{C_i^Y} \cap \{1..i-1\}} g_{ij} + \sum_{j \in B_{C_i^Y} \cap \{1..i-1\}} g'_{ij} \right) \right].
\end{aligned} \tag{4.17}$$

The weights $\left(p_{n,i},\, p'_{n,i}\, g_{n,i}\, g'_{n,i}\right)_{i=1}^{n}$ that appear in above expression do not have an explicit form and have to be defined sequentially. This makes the numerical maximization procedure cumbersome for large n (more than 5 hours were required for $n = 2000$).

At the same time, for a problem of species discovery, the coverages, discovery rates and expected numbers of new classes, depend only on the functions r_n and h_n, so one may be more interested in estimating r_n and h_n than all other weights. With this in mind, we propose another estimator based on simplified data. The corresponding likelihood has simpler analytical form and can be easily computed even for large n.

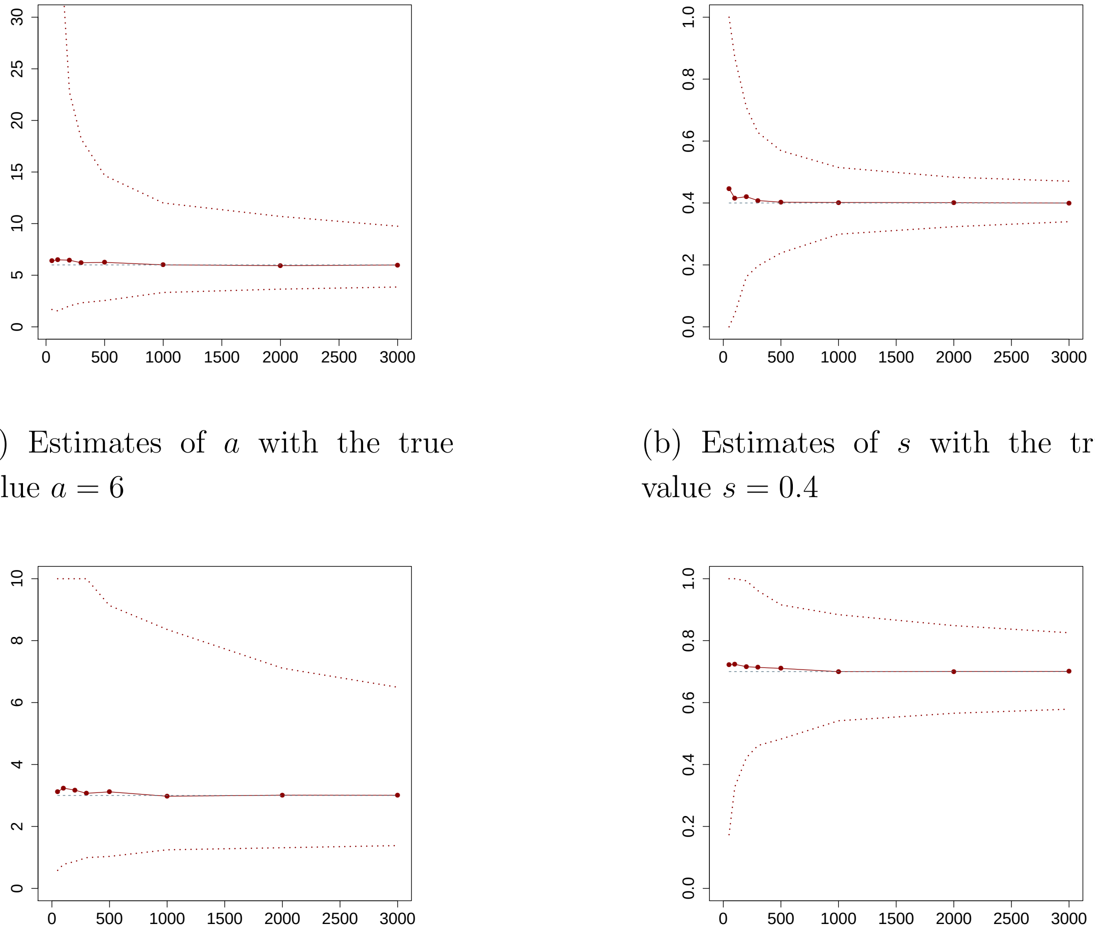
Based on bivariate sample $(X_{1:n}, Y_{1:n})$, we define a new set of random variables $(B^X_{1:n}, B^Y_{1:n})$ which are binary and indicate whether observation was old or new. That is, $B^X_i = 1$ if X_i has not been observed among $(X_{1:i-1}, Y_{1:i-1})$, and $B^Y_i = 1$ if Y_i has not been observed among $(X_{1:i-1}, Y_{1:i-1})$. Our idea is that the simplified data is asymptotically sufficient for a good estimation of the parameters of interest. Denote the number of new classes that are first observed in the first group by $K^I_n = \sum_{i=1}^n B^X_i$, and similarly for $K^{II}_n = \sum_{i=1}^n B^Y_i$. Clearly, $K_{n,n} = K^I_n + K^{II}_n$ is the total number of distinct classes in the sample. Then

$$\begin{aligned}&\mathbb{P}\left(B^X_1, ..., B^X_n, B^Y_1, ..., B^Y_n \,|\, (\Theta_1, \Theta_2)\right) = \\ &= \frac{\theta_1^{K^I_n - 1} \prod_{i:B^X_i=0}(i-1)^s}{(\theta_1+1^s)...(\theta_1+(n-1)^s)} \frac{\theta_2^{K^{II}_n - 1} \prod_{i:B^Y_i=0}(i-1)^t}{(\theta_2+1^t)...(\theta_2+(n-1)^t)}.\end{aligned} \tag{4.18}$$

The new estimators then become

$$\begin{aligned}\hat{\Theta}_1 &= \underset{\theta_1,s}{argmax}\left(\frac{\theta_1^{K^I_n - 1} \prod_{i:B^X_i=0}(i-1)^s}{(\theta_1+1^s)...(\theta_1+(n-1)^s)}\right), \\ \hat{\Theta}_2 &= \underset{\theta_2,t}{argmax}\left(\frac{\theta_2^{K^{II}_n - 1} \prod_{i:B^Y_i=0}(i-1)^t}{(\theta_2+1^t)...(\theta_2+(n-1)^t)}\right).\end{aligned} \tag{4.19}$$

We analyze the properties of the new estimators via simulations. For this we generated 500 samples of size 3000 from Model-1 with parameters $a_0 = 6$,$b_0 = 3$, $s_0 = 0.4$, $t_0 = 0.7$. For each sample we computed the incomplete maximum likelihood (4.19). It took less than one minute to estimate the parameters from each sample, while maximizing the complete likelihood would have taken hours. Thus, we obtained 500 estimations and took the 0.05, 0.50 and 0.95 quantiles. The results, presented in Figure 4.5, suggest that, asymptotically, these estimators are unbiased.



(a) Estimates of a with the true value $a = 6$

(b) Estimates of s with the true value $s = 0.4$

(c) Estimates of b with the true value $b = 3$

(d) Estimates of t with the true value $t = 0.7$

Figure 4.5: The true parameters of Model-1 compared to the estimates using the incomplete likelihood, with 90% quintiles. The gray dashed lines are the true values.

In Table 4.1 we also compared the value of the complete loglikelihood evaluated at the true parameters with the value of the complete loglikelihood evaluated at the parameters estimated by our proposed method.

Table 4.1: True likelihood and estimated likelihood

n	True Loglikelihood	Estimated Loglikelihood
50	-193.9369	-202.9014
100	-413.4797	-428.1927
200	-996.6889	-995.7221
300	-1630.9391	-1631.4429
400	-2268.4959	-2268.9500
500	-2951.2850	-2953.1099
600	-3656.8922	-3660.2970
700	-4340.4405	-4342.0585
800	-5068.1288	-5068.8504
900	-5746.7054	-5748.5179
1000	-6473.4486	-6476.8212
2000	-13871.1048	-13876.6271
3000	-21935.0407	-21943.6686

These graphs and table strongly suggest that, for a large sample size, the simpler estimator based on the incomplete likelihood does not loose relevant information.

4.2.6 Model-1 for analyzing texts

In this subsection we apply Model-1 to a real data example. As discussed before, one of the features of Model-1 is the behavior of the tails, which makes this model appropriate for data with power-law tails behavior. Word frequencies in natural language are an example of such data. We consider two texts and study the number of new words that will appear in the future. Notice that, in text-analysis, assumption regarding the order (3.13) does not make much practical sense, and the assumption of non-stationarity is also usually not needed; an assumption of partial exchangeability is generally more appropriate. We use this example to illustrate the performance of the proposed model, but we are aware that more appropriate problems should be addressed. However, one may imagine texts written over a relevant period of time, where indeed the assumption of stationarity implied by partial exchangeability might be too restrictive.

We considered two texts written by William Shakespeare – ”The Two Noble Kinsmen” and ”Romeo and Juliet”. The first one, according to a legend, was co-written with John Fletcher, while the authorship of the second one refers to Shakespeare only. Therefore, we assume that the underlying processes of new word generation are different for the two texts. In this example, the role of new species is played by new words. By “new” word we mean that this particular form of the word has not appeared before. For example, words ’do’ and ’does’ will correspond to two distinct species. We analyzed 7000 words from one text and 7000 from another, made predictions about the number of new words in a future sample of size 8000 (the next 7001-15000 words) and compared our prediction with the true observations.

Figure 4.6 represents the two observed texts, with words from the first group in red, words from the second group in blue. Each point on this plot corresponds to a word, its x-coordinate is the index of the word while the y-coordinate is the cluster (in order of appearance) to which the word belongs. So, observations on the same level on the plot correspond the same words. In total, 3621 distinct words are observed in the sample, of these, 640 words are shared, 1541 words are only observed in the first text, and 1440 words are observed only in the second text. It seems that the first process generates more new classes, but the difference is not big.

Figure 4.7 represents the bi-partition generated by these texts. Each point represents a species, i.e., a cluster corresponding to a distinct word. The x-coordinate indicates how many such words were observed in ”The Two Noble Kinsmen”, the y-coordinate indicates how many such words were observed in ”Romeo and Juliet”. This graph consistent with our guess,namely, that the first text, ”The Two Noble Kinsmen”, is slightly reacher of new words.

In order to estimate the model parameters we used the incomplete

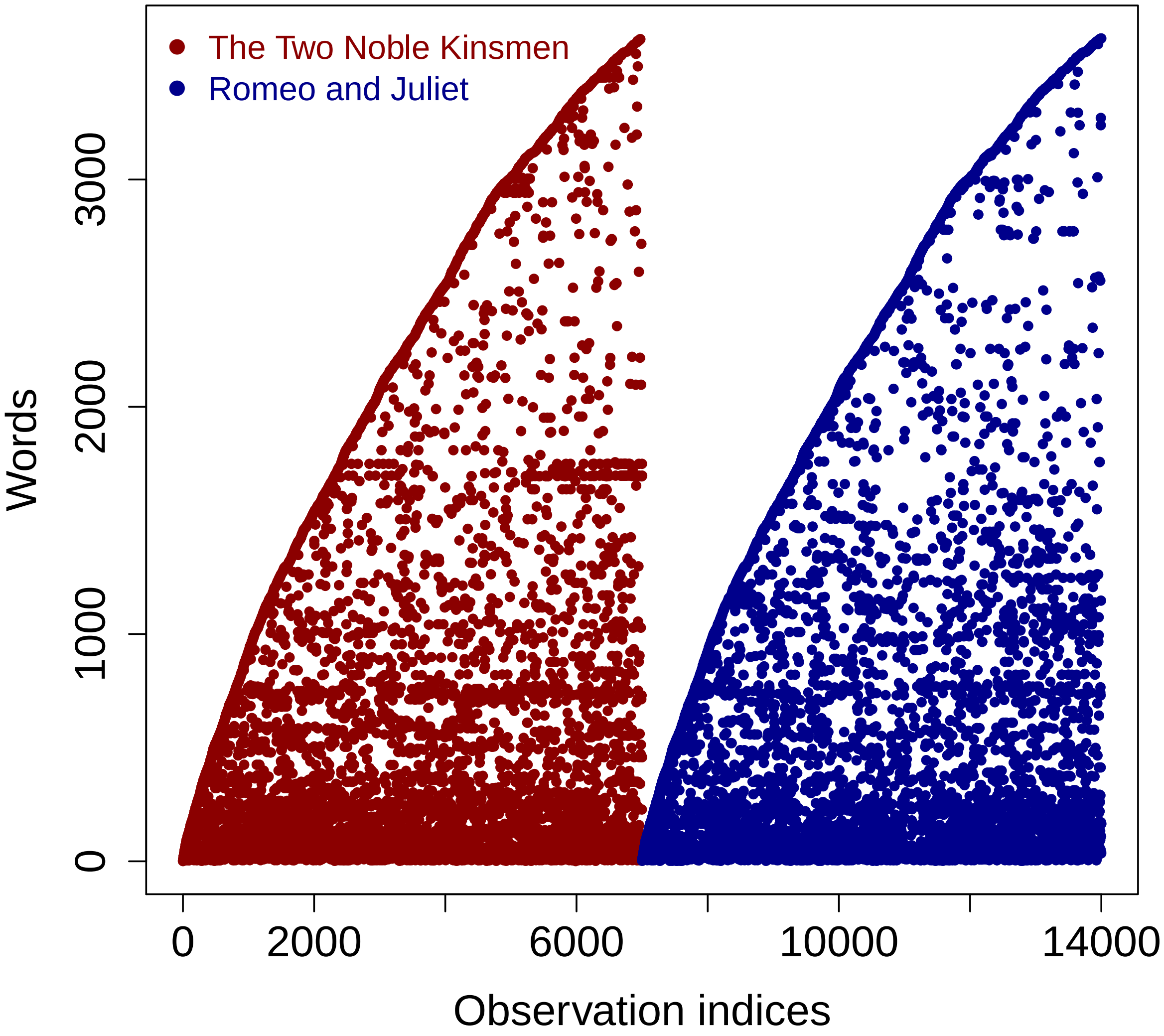


Figure 4.6: Distinct words in ”The Two Noble Kinsmen” and ”Romeo and Juliet”

estimator described before, since the sample size is large (7000 words from each text). The estimated parameters are: $\hat{\theta}_1 = 40.403$, $\hat{\theta}_2 = 6.897$, $\hat{s} = 0.608$ and $\hat{t} = 0.388$. Based on these estimates we can estimate coverages using formula (4.13):

$$\hat{C}^{I}_{7000} = 1 - \frac{\theta_1}{\theta_1 + (7000-1)^s} = 0.844,$$
$$\hat{C}^{II}_{7000} = 1 - \frac{\theta_2}{\theta_2 + (7000-1)^t} = 0.819.$$

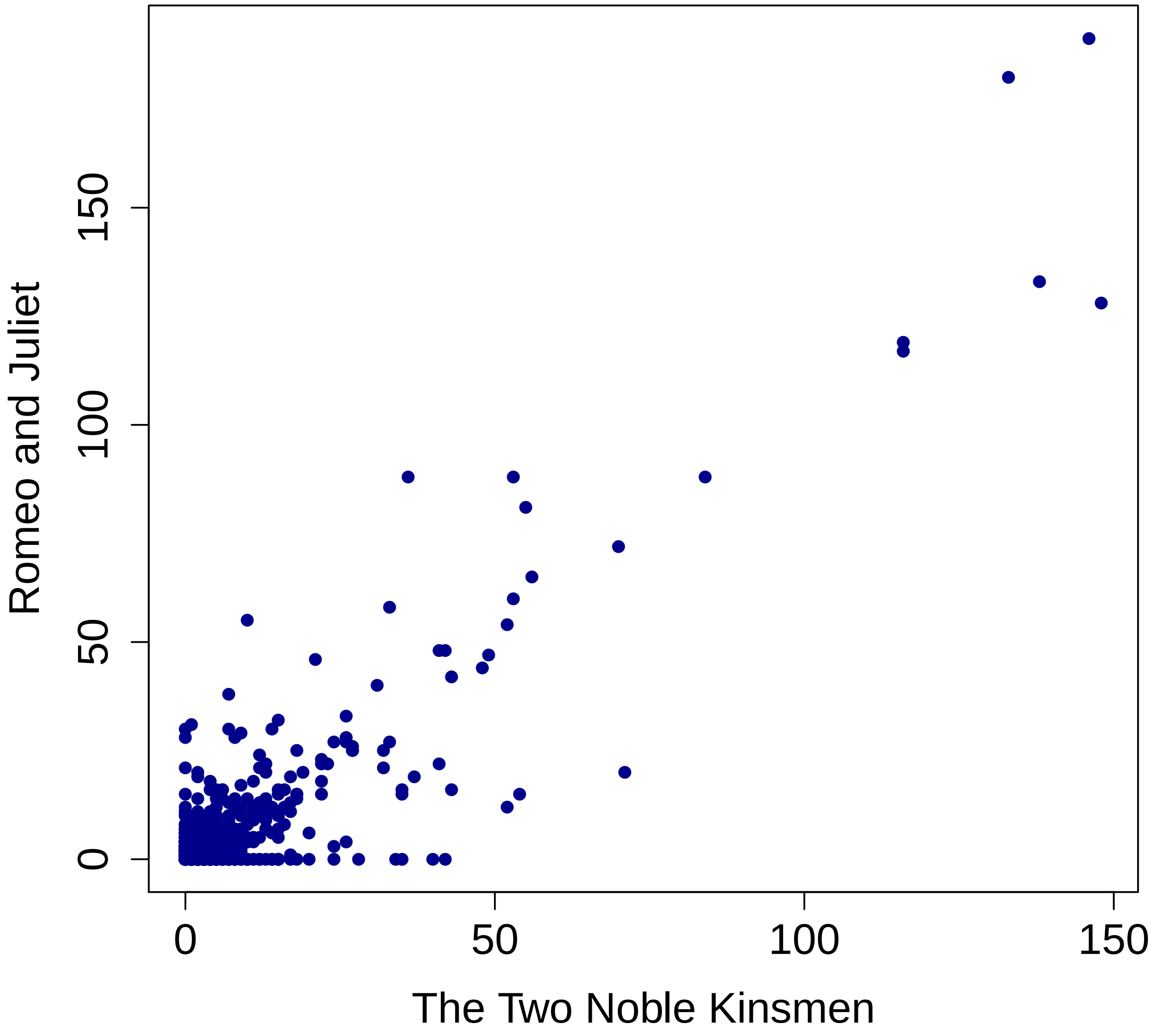
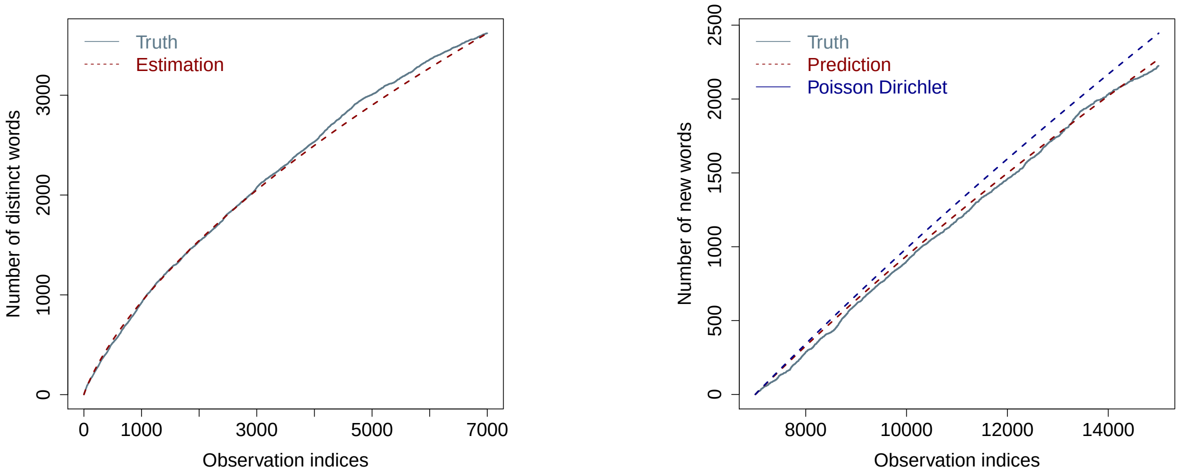


Figure 4.7: Sizes of classes of the bi-partition generated by ”The Two Noble Kinsmen” and ”Romeo and Juliet”. Each point represents a distinct word.

This means that $84,4\%$ of the whole population of words was presented in ”The Two Noble Kinsmen” and $81,9\%$ of the whole population of words was presented in ”Romeo and Juliet”. It also confirms our guess that the fact that two persons were writing the first novel could increase number of distinct words used.

In Figure 4.8a we can see the estimated number of new words, together with the real number of new words at each time $n = 1, ..., 7000$ for both



(a) Estimated number of distinct words, $\hat{K}_{n,n}$.

(b) Predicted number of new words, $K_{m,m}^{(7000,7000)}$.

Figure 4.8: Estimated and predicted number of words observed at time n in comparison with the truth.

texts. Our estimation fits the truth quite well. In Figure 4.8b we compare the predicted number of new words according to expression (4.14), with the true number of new words.

We also make a comparison of our model with other proposed models. As there are not so many models for such type of data present in the literature, we combined both texts and estimated the joint coverage, discovery rate and expected number of new words in a next sample, using a two-parameter Poisson Dirichlet prior, as described in Favaro et al. [2009]. The estimated parameters of the two-parameter Poisson-Dirichlet process are: $\hat{\theta} = 1.3$ and $\hat{\sigma} = 0.68$. We can see that our estimator gives a more precise estimation of the next sample.

We also use a Bayesian approach, taking into account our previous results, $\hat{\theta}_1 = 40.403$, $\hat{\theta}_2 = 6.897$, $\hat{s} = 0.608$ and $\hat{t} = 0.388$. We used the following priors on these parameters: $\theta_1 \sim \Gamma(10, 4)$, $\theta_2 \sim \Gamma(3, 2)$, $s \sim B(3, 2)$ and $t \sim B(2, 3)$. Samples from the posterior distributions were computed via Metropolis within Gibbs algorithm. For these computations, we also used an incomplete likelihood considering sample size is large.

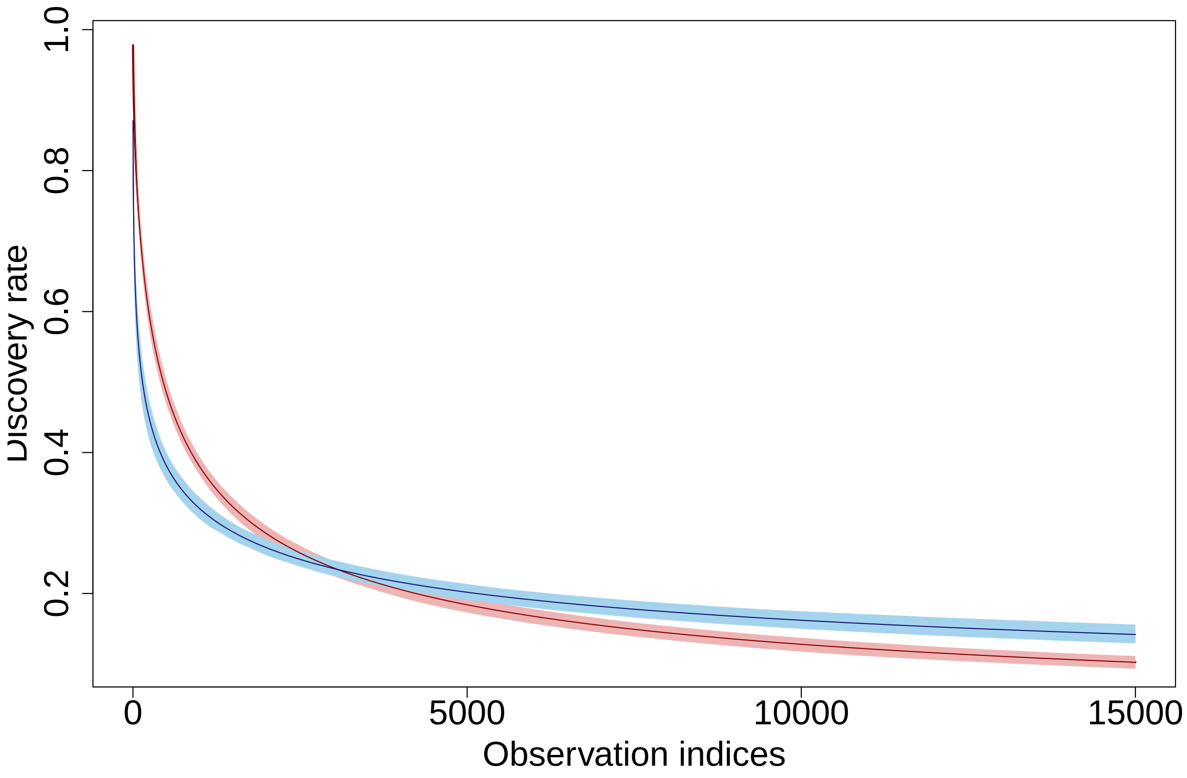
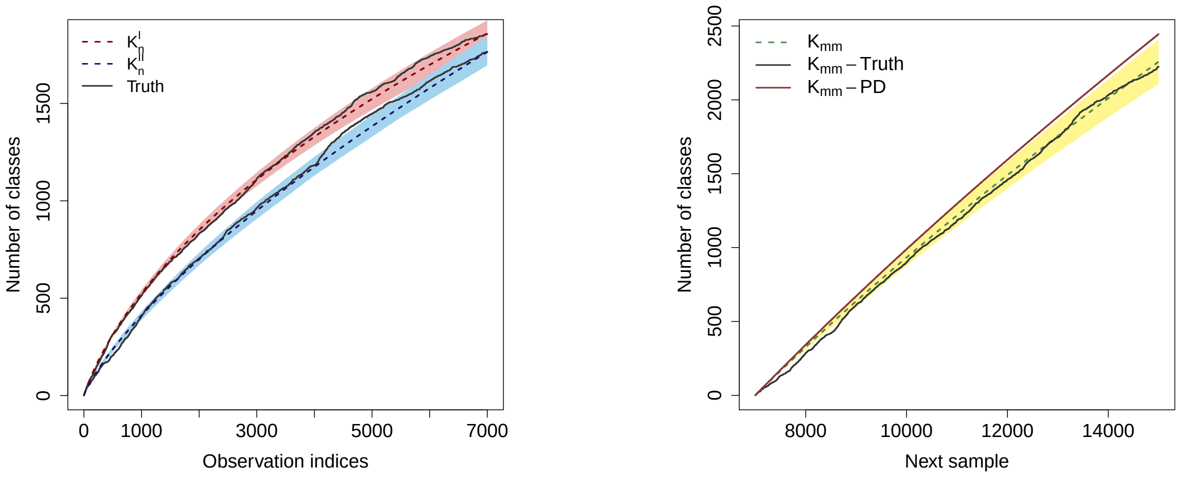


Figure 4.9: Estimated discovery rates for ”The Two Noble Kinsmen” (in red) and ”Romeo and Juliet” (in blue) with 95% credible intervals.



(a) Observed sample

(b) Next sample

Figure 4.10: Estimated number of distinct words, $K_{n,n}$, for $n = 1, ..., 1000$, and prediction for a next sample, $K_{m,m}^{(7000,7000)}$ for $m = 7001, ..., 15000$, with 95% credible bounds

Figure 4.9 shows the estimated discovery rates for the two texts with 95% credible regions. The evolution of the discovery rates is different for the two texts. At the beginning the first text, ”The Two Noble Kinsmen”, has higher probabilities of generating new words than ”Romeo and Juliet”, while after 4000 words the behavior changes. Figure 4.10 shows the estimated $K_{n,n}$ and predicted $K_{m,m}^{(n,n)}$ obtained from posterior samples, with 95% credible intervals. As before, we used a prediction obtained with a two-parameter Poisson Dirichlet prior based on a joint of both texts.

We conclude that the proposed Model-1 works well for discovering new words and can be also applied to other problems of species discovery.

4.3 Future work

In this section we briefly mention some ideas for further research.

Hierarchical partially c.i.d. models

Bayesian Nonparametric priors are widely used in hierarchical models for clustering, curve estimation and analyzing many types of dependence between data, including dependence arising from time and space. So one can use a generalized bivariate species sampling process as a prior in a hierarchical model. For example, one can consider $(X_{1:n}, Y_{1:n})$, where

$$
\begin{aligned}
&X_i|\nu_i \overset{ind}{\sim} \mathcal{N}\left(\nu_i, \tau^2\right), \\
&Y_i|\mu_i \overset{ind}{\sim} \mathcal{N}\left(\mu_i, \tau^2\right), \\
&(\nu_{1:n}, \mu_{1:n}) \sim \text{Model-1 defined before.}
\end{aligned}
$$

It is possible to show that $(X_{1:n}, Y_{1:n})$ is also partially c.i.d. and so, asymptotically, partially exchangeable. Clustering in this model depends on time. For estimation parameters one can use the MCMC sampling scheme for non-exchangeable processes proposed by Blei and Frazier [2011].

Partially c.i.d. with respect to different filtrations

We defined a partially c.i.d. sequence as a vector of two sequences that are both c.i.d. with respect to the same filtration $\mathcal{G}$. One can also consider two sequences that are c.i.d. with respect to different filtrations. For example, $(X_n)_{n\geqslant 1}$ may be c.i.d. with respect to $(X_{1:n}, Y_{1:n})_{n\geqslant 1}$ while $(Y_n)_{n\geqslant 1}$ is c.i.d. with respect to $(X_{1:n+1}, Y_{1:n})_{n\geqslant 1}$. Other type of filtrations could be considered depending on the application at hand.

Generalized bivariate species sampling sequences of Good-Toulmin type for species discovery

Good and Toulmin [1956] propose an estimator that is now widely used in problems of species discovery. According to this estimator, the probability that the next observation will be of a new type is equal to the proportion of classes observed only once to the whole sample, i.e. $p = \frac{m_1}{n}$. Then one can consider a bivariate prediction rule

$$\mathbb{P}\left[X_{n+1} \in \cdot | X_{1:n}, Y_{1:n}\right] = \sum_{i=1}^{n} p_{n,i}\delta_{X_i}(\cdot) + \sum_{i=1}^{n} p'_{n,i}\delta_{Y_i}(\cdot) + r_n\nu(\cdot),$$
$$\mathbb{P}\left[Y_{n+1} \in \cdot | X_{1:n}, Y_{1:n}\right] = \sum_{i=1}^{n} g_{n,i}\delta_{X_i}(\cdot) + \sum_{i=1}^{n} g'_{n,i}\delta_{Y_i}(\cdot) + h_n\nu(\cdot).$$

where $r_n = \frac{m_1^x}{n}$ and $h_n = \frac{m_1^y}{n}$. The conditions under which this prediction rule generates a partially c.i.d. sequence can be computed by checking condition (3.15).

Partially c.i.d. with respect to a more general order

An analogous analysis to the one done in this thesis can be pursued for different possible orderings.

Bibliography

D. Aldous. Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII 1983*, 1117:1–198, 1985.

M. Aoki. Thermodynamic limits of macroeconomic or financial models: One-and two-parameter Poisson–Dirichlet models. *Journal of Economic Dynamics and Control*, 32:66–84, 2008.

F. Bassetti, I. Crimaldi, and F. Leisen. Conditionally identically distributed species sampling sequences. *Advances in Applied Probability*, 42:433–459, 2010.

P. Berti and P. Rigo. A Glivenko-Cantelli theorem for exchangeable random variables. *Statistics & Probability Letters*, 32:385–391, 1997.

P. Berti, L. Pratelli, and P. Rigo. Limit theorems for a class of identically distributed random variables. *The Annals of Probability*, 32:2029–2052, 2004.

D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.

D. M. Blei and P. I. Frazier. Distance dependent Chinese restaurant processes. *The Journal of Machine Learning Research*, 12:2461–2488, 2011.

W. Brian Arthur, Y. M. Ermoliev, and Y. M. Kaniovski. Path-dependent processes and the emergence of macro-structure. *European Journal of Operational Research*, 30:294–303, 1987.

J. Bunge and M. Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88:364–373, 1993.

R. Carnap and R. C. Jeffrey. *Studies in Inductive Logic and Probability*. University of California Press, 1980.

D. Cifarelli and E. Regazzini. Problemi statistici non parametrici in condizioni di scambiabilita parziale e impiego di medie associative. Technical report, Quaderni dell'Istituto di Matematica Finanziaria, University Torino, 1978.

R. K. Colwell and J. A. Coddington. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London: Series B*, 345:101–118, 1994.

T. Costa, M. Guindani, F. Bassetti, F. Leisen, and E. Airoldi. Generalized species sampling priors with latent beta reinforcements. *arXiv preprint arXiv:1012.0866v3*, 2013.

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

P. Donnelly and P. Joyce. Continuity and weak convergence of ranked and size-biased permutations on the infinite simplex. *Stochastic Processes and their Applications*, 31:89–103, 1989.

S. D. Durham, N. Flournoy, and W. Li. A sequential design for maximizing the probability of a favourable response. *Canadian Journal of Statistics*, 26:479–495, 1998.

F. Eggenberger and G. Pólya. Über die statistik verketteter vorgänge. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 3:279–289, 1923.

W. J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3:87–112, 1972.

S. Favaro, A. Lijoi, and I. Pruenster. On the stick-breaking representation of normalized inverse Gaussian priors. *Biometrika*, 99:663–674.

S. Favaro, A. Lijoi, R. H. Mena, and I. Prünster. Bayesian non-parametric inference for species variety with a two-parameter Poisson–Dirichlet process prior. *Journal of the Royal Statistical Society: Series B*, 71:993–1008, 2009.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.

R. A. Fisher, A. S. Corbet, and C. B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 12:42–58, 1943.

N. Flournoy, C. May, and P. Secchi. Asymptotically optimal response-adaptive designs for allocating the best treatment: an overview. *International Statistical Review*, 80:293–305, 2012.

S. Fortini and S. Petrone. Predictive construction of priors in Bayesian nonparametrics. *Brazilian Journal of Probability and Statistics*, 26:423–449, 2012.

S. Fortini, L. Ladelli, and E. Regazzini. Exchangeability, predictive distributions and parametric models. *Sankhyā: The Indian Journal of Statistics, Series A*, 62:86–109, 2000.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264, 1953.

I. J. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* M.I.T. Press, 1965.

I. J. Good and G. H. Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43:45–63, 1956.

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

P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. New York: Academic Press, 1980.

B. M. Hill, D. Lane, and W. Sudderth. A strong law for some generalized urn processes. *The Annals of Probability*, 8:214–226, 1980.

O. Kallenberg. Spreading and predictable sampling in exchangeable sequences and processes. *The Annals of Probability*, 16:508–534, 1988.

J. F. C. Kingman. Random partitions in population genetics. *Proceedings of the Royal Society of London: Series A*, 361:1–20, 1978a.

J. F. C. Kingman. The representation of partition structures. *Journal of the London Mathematical Society*, 2:374–380, 1978b.

M. Kolossiatis, J. E. Griffin, and M. F. J. Steel. On Bayesian nonparametric modelling of two correlated distributions. *Statistics and Computing*, 23: 1–15, 2013.

J. Lee, F. A. Quintana, P. Müller, and L. Trippa. Defining predictive probability functions for species sampling models. *Statistical Science*, 28:209–222, 2013.

F. Leisen and A. Lijoi. Vectors of two-parameter Poisson–Dirichlet processes. *Journal of Multivariate Analysis*, 102:482–495, 2011.

W. Li, S. D. Durham, and N. Flournoy. Randomized Pólya urn designs. *Proceedings of the Biometric Section of the American Statistical Association*, pages 166–170, 1996.

A. Lijoi and I. Prünster. Models beyond the Dirichlet process. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics*, pages 80–136. Cambridge University Press, Cambridge, 2010.

A. Lijoi, R. H. Mena, and I. Prünster. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94:769–786, 2007.

A. Lijoi, B. Nipoti, and I. Prünster. Bayesian inference with dependent normalized completely random measures. *Carlo Alberto Notebooks*, 224, 2011.

A. Y. Lo. A characterization of the Dirichlet process. *Statistics & Probability Letters*, 12:185–187, 1991.

S. N. MacEachern. Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pages 50–55. Alexandria, VA: American Statistical Association, 1999.

B. K. Mallick and S. G. Walker. Combining information from several experiments with nonparametric priors. *Biometrika*, 84:697–706, 1997.

C. May, A. M. Paganoni, and P. Secchi. On a two-color generalized Pólya urn. *Metron-International Journal of Statistics*, 63:115–134, 2005.

J. W. McCloskey. *A model for the distribution of individuals by species in an environment*. PhD thesis, Michigan State University, 1965.

P. Muliere, A. M. Paganoni, and P. Secchi. A randomly reinforced urn. *Journal of Statistical Planning and Inference*, 136:1853–1874, 2006.

P. Müller, F. Quintana, and G. Rosner. A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B*, 66:735–749, 2004.

S. Nacu. Increments of random partitions. *Combinatorics, Probability and Computing*, 15:589–595, 2006.

A. M. Paganoni and P. Secchi. Interacting reinforced-urn systems. *Advances in Applied Probability*, 36:791–804, 2004.

K. R. Parthasarathy. *Probability Measures on Metric Spaces.* Academic Press, New York, 1967.

R. Pemantle. A survey of random processes with reinforcement. *Probability Surveys*, 4:1–79, 2007.

M. Perman. *Random discrete distributions derived from subordinators.* PhD thesis, University of California, Berkeley, 1990.

M. Perman, J. Pitman, and M. Yor. Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92: 21–39, 1992.

J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102:145–158, 1995.

J. Pitman. Some developments of the Blackwell-MacQueen urn scheme. In T. S. Ferguson, L. S. Shapley, and J. B. MacQueen, editors, *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, volume 30, pages 245–267. IMS, Hayward, CA, 1996a.

J. Pitman. Random discrete distributions invariant under size-biased permutation. *Advances in Applied Probability*, 28:525–539, 1996b.

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

J. Pitman. *Combinatorial Stochastic Processes*, volume 1875. Springer-Verlag, 2006.

V. Rao and Y.W. Teh. Spatial normalized gamma processes. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1554–1562, 2009.

E. Regazzini. Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilita. *Giornale dell'Instituto Italiano degli Attuari*, 41:77–89, 1978.

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

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.

R. C. Tiwari and R. C. Tripathi. Nonparametric Bayes estimation of the probability of discovering a new species. *Communications in Statistics-Theory and Methods*, 18:877–895, 1989.

S. Walker and P. Muliere. A bivariate Dirichlet process. *Statistics & Probability Letters*, 64:1–7, 2003.