Universidad Nacional Autónoma de México

Programa de Maestría y Doctorado en Ciencias Matemáticas y de la Especialización en Estadística Aplicada

Hierarchies of dependent random measures

T E S I S

que para optar por el grado de:

Maestra en Ciencias Matemáticas

PRESENTA:

Aidée Violeta Enroth Ortiz

Director de tesis:
Dr. Ramsés Humberto Mena Chávez
I.I.M.A.S.

Ciudad de México, Agosto 2021

Acknowledgments

I want to express my deepest gratitude to all my loved ones, for always being there for me; to the director of this thesis, Ramsés, for his correct guidance and continuous support; to the Postgraduate Program of Mathematics and to all of my teachers, for their knowledge and dedication; finally to CONACyT and the PAPIIT IG100221 project, for the financial support provided during my studies and the completion of this thesis.

Agradecimientos

Quiero expresar mi profundo agradecimiento a todos mis seres queridos, por estar siempre ahí; al director de esta tesis, Ramsés, por su acertada orientación y apoyo continuo; al programa de Posgrado en Ciencias Matemáticas y a mis profesores, por sus conocimientos y su dedicación; finalmente a CONACyT y al proyecto PAPIIT IG100221, por el apoyo económico brindado durante mis estudios y la realización de este trabajo de tesis.

Contents

Introduction

A fundamental aim of statistics consists in the prediction of future outcomes of a sequence $(\mathrm{x}_i)_{i\geq 1}$, on the basis of a sample $(\mathrm{x}_i)_{i=1}^n$. In order to face this problem, the observations $\mathrm{x}_1, \mathrm{x}_2, \ldots$ are required to satisfy some symmetry condition that permits us to treat them as if they were analogous. In the Bayesian nonparametric setting, such symmetry corresponds to *exchangeability*. A sequence of random variables $(\mathrm{x}_i)_{i\geq 1}$, taking values on a Borel space $\mathbb{X}$ endowed with its Borel σ-algebra $\mathcal{X}$, is exchangeable if its finite-dimensional distributions are invariant under finite permutations. This means that $(\mathrm{x}_i)_{i=1}^n \stackrel{d}{=} \left(\mathrm{x}_{\rho(i)}\right)_{i=1}^n$ for every $n \geq 1$ and any permutation ρ of $\{1, \ldots, n\}$. Equivalently, exchangeable observations are invariant with respect to the order in which they were recorded. Bruno de Finetti's representation theorem states that a sequence is exchangeable if and only if there exists a probability measure Q that takes values on the space of all probability measures over $\mathbb{X}$, $\mathrm{P}_{\mathbb{X}}$ with its Borel σ-algebra $\mathcal{P}_{\mathbb{X}}$, such that

$$\mathbb{P}\left[\mathrm{x}_1 \in A_1, \ldots, \mathrm{x}_n \in A_n\right] = \int_{\mathrm{P}_{\mathbb{X}}} \prod_{i=1}^{n} \tilde{p}(A_i) Q(\mathrm{d}\tilde{\mathrm{p}})$$

for any collection of measurable sets $(A_i)_{i=1}^n$. This can be stated as that the sequence $(\mathrm{x}_i)_{i\geq 1}$ is conditionally i.i.d. given the random probability measure $\tilde{p}$, whose distribution is Q. If the support of Q is an infinite-dimensional subspace of $\mathrm{P}_{\mathbb{X}}$, then Q is termed a *nonparametric prior*, and prediction within an exchangeable setting with such priors has been extensively studied.

In a large variety of applications, such as when data are generated from different though related experiments or populations, exchangeability is not an appropriate assumption as usually there exist some degree of heterogeneity within the samples. The experiments identify m sequences $(\mathrm{x}_{1,j})_{j\geq 1}, \ldots, (\mathrm{x}_{m,j})_{j\geq 1}$ taking values on the same space $(\mathbb{X}, \mathcal{X})$ such that the homogeneity assumption of exchangeablity may hold within each experiment $(\mathrm{x}_{i,j})_{j\geq 1}$, though not necessarily across different $(\mathrm{x}_{i,j})_{j\geq 1}$ and $(\mathrm{x}_{k,j})_{j\geq 1}$, where $i \neq k$. In this case, a more general form of dependence is needed, such as *partial exchangeability*. Partial exchangeability was introduced in de Finetti (1937) as a need to cover these situations, as in the words of de Finetti himself

> "*To get from the case of exchangeability to other cases which are more general but still tractable, we must take up the case where we still encounter 'analogies' among the events under consideration, but without attaining the limiting case of exchangeability*"

An array $\mathbf{x} = \left((\mathrm{x}_{i,j})_{j\geq 1}\right)_{i=1}^{m}$ of m sequences of random variables is partially exchangeable if the joint law of the vector $\left((\mathrm{x}_{1,j})_{j=1}^{n_1}, \ldots, (\mathrm{x}_{m,j})_{j=1}^{n_m}\right)$ is invariant under permutations ρ_i of the indices $\{1, \ldots, n_i\}$ within each sample. This means that

$$\left((\mathrm{x}_{1,j})_{j=1}^{n_1}, \ldots, (\mathrm{x}_{m,j})_{j=1}^{n_m}\right) \stackrel{d}{=} \left(\left(\mathrm{x}_{1,\rho_1(j)}\right)_{j=1}^{n_1}, \ldots, \left(\mathrm{x}_{m,\rho_m(j)}\right)_{j=1}^{n_m}\right).$$

A corresponding representation theorem for partially exchangeable arrays holds true, so that we can characterize such arrays by the means of a probability measure Q_m that takes values on $(\mathrm{P}_{\mathbb{X}}^m, \mathcal{P}_{\mathbb{X}}^m)$ and such that

$$\mathbb{P}\left[\bigcap_{i=1}^{m}\left\{\mathbf{x}_i^{(n_i)} \in A_i\right\}\right] = \int_{\mathrm{P}_{\mathbb{X}}^m}\prod_{i=1}^{m}\tilde{p}_i^{(n_i)}(A_i)\mathrm{Q}_m(\mathrm{d}\tilde{p}_1, \ldots, \mathrm{d}\tilde{p}_m)$$

for any integers $n_i \geq 1$ and $A_i \in \mathcal{X}^{(n_i)}$. Q_m dictates the dependence between the vector of random probability measures $(\tilde{p}_1, \ldots, \tilde{p}_m)$ and since both exchangeability and complete independence occur as a limiting case of partial exchangeability, it is desirable to have a prior Q_m that can cover the full range of possible dependence structures.

As aforementioned, a large amount of literature on Bayesian nonparametric statistics has been developed under the assumption of exchangeability and this case is well understood. However, models for partially exchangeable data are still the subject of current literature. Nested models were studied in Camerlenghi et al. (2019a) and Camerlenghi (2015), where the random probability measures $\tilde{p}_1, \ldots, \tilde{p}_m$ are exchangeable themselves, and thus their dependence is dictated by $\tilde{p}_1, \ldots, \tilde{p}_m \,|\, \tilde{q} \sim \tilde{q}$, where $\tilde{q}$ is a random probability measure on $(\mathrm{P}_{\mathbb{X}}, \mathcal{P}_{\mathbb{X}})$. Additive structures such that $\tilde{p}_i := T(\tilde{\mu}_i + \tilde{\mu}_0)$, where $\tilde{\mu}_0$ is a common random probability measure and T is a suitable transformation such as normalization have been studied back to Müller et al. (2004). Here we will study a construction for Q_m based on hierarchical processes, meaning that Q_m will be expressed as

$$\begin{aligned} \tilde{p}_i \,|\, \tilde{p}_0 &\sim \mathrm{Q}_i(\tilde{p}_0) \quad \text{with } \mathbb{E}[\tilde{p}_i \,|\, \tilde{p}_0] = \tilde{p}_0 \text{ for } i = 1, \ldots, m \\ \tilde{p}_0 &\sim Q_0. \end{aligned}$$

This means that to enable the dependence across the m samples, each of the $\tilde{p}_i$'s will share the same random base measure $\tilde{p}_0$. The choice of $\mathrm{Q}_i(\tilde{p}_0)$ will be based on completely random measures, either it being through the normalization or a transformation of a completely random measure.

Organization of the document

The outline of the thesis is as follows. In Chapter 1 we recall some basics and notation, namely the concept of exchangeability and the representation theorem for exchangeable sequences (Sections 1.1 and 1.2). In Section 1.3 we study Poisson processes on general spaces and review some of their distributional properties, to the aim of constructing completely random measures based on these processes, as exposed in Section 1.4. In Chapter

2, we study nonparametric priors constructed from transformations of completely random measures and some of their key properties, such as the partition structure (2.1.2) and a posterior representation (2.1.3). Examples of these priors are exposed in Section 2.2 and 2.3. The notion of partial exchangeability is discussed at the beginning of Chapter 3, and its corresponding representation theorem. Hierarchical processes constructed from normalized completely random measures are studied in Section 3.2. As in the exchangeable framework, this construction implies that there will be ties within each sample and across different samples as well, so the partition induced by partially exchangeable arrays whose dependence structure is dictated by hierarchical NRMIs and the distribution of the number of blocks is studied in Sections 3.2.2 and 3.2.3 respectively. Sections 3.3 and 3.4 exhibit particular examples of hierarchical processes. In Chapter 4 we apply the theory developed in Chapter 3 to density estimation of m partially exchangeable sequences of data by the means of a Monte Carlo Markov Chain algorithm. We first make a brief review on mixture models based on the nonparametric priors studied in Chapter 2 (Sections 4.1.1 and 4.2.1) to then extend the methodology to partially exchangeable arrays (Section 4.2). Finally, we designed two small experiments to test the performance of several hierarchical processes on Section 4.3.

Preliminaries

The aim of this preliminary chapter is to lay the foundations of the framework in which we will be working throughout this thesis. Sections 1.1 and 1.2 explain the bases of Bayesian statistics according to the concept of exchangeability and the representation theorem for exchangeable sequences, a result that guarantees the existence of a random probability measure and the need for an initial distribution. Section 1.3 gives a brief introduction to Poisson processes and their properties, since they constitute a fundamental tool for the construction of random probability measures. This Section relies mainly on Kingman (1993). Finally, Section 1.4 deals with completely random measures, their atomic decomposition and some examples.

1.1 Exchangeability and random measures

According to Goldstein (2013), uncertainty can be categorized as either aleatory or epistemic: epistemic uncertainty is that which relates to our lack of knowledge, whereas aleatory uncertainty is inherent to the phenomenon under study. In statistical analysis, given data $\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$, aleatory and epistemic uncertainty are expressed through a parametric family of distributions $\{\mu_\theta : \theta \in \Theta\}$ and the parameter θ respectively. The classical or so called frequentist approach assumes that θ is an unknown but fixed quantity, in contrast with the Bayesian approach, that considers the parameter itself as a random variable with an initial probability distribution, called the *prior distribution*. The reason behind this assumption lies on the concept of exchangeability and the celebrated representation theorem for exchangeable sequences.

Before moving on, let us state that we will be working on a Polish space $\mathbb{X}$, meaning a complete and separable metric space, endowed with its Borel σ-algebra $\mathcal{X}$. We will denote as $\mathrm{M}_\mathbb{X}$ the space of all of boundedly finite measures on $(\mathbb{X}, \mathcal{X})$, that is $m(A) < \infty$ for $m \in \mathrm{M}_\mathbb{X}$ and any bounded set $A \in \mathcal{X}$. $\mathcal{M}_\mathbb{X}$ will be the smallest σ-algebra on $\mathrm{M}_\mathbb{X}$ such that the evaluation mappings $\pi_A : \mu \to \mu(A)$ are measurable for every $A \in \mathcal{X}$. Similarly, $\mathrm{P}_\mathbb{X}$ will be the space of all probability measures on $\mathbb{X}$ and $\mathcal{P}_\mathbb{X}$ its associated Borel σ-algebra. Note that $\mathrm{P}_\mathbb{X}$ is a measurable subset of $\mathrm{M}_\mathbb{X}$.

Definition 1.1. Let $\mathbb{X}$ be a Polish space endowed with its Borel σ-algebra $\mathcal{X}$. A finite collection of random variables $(\mathrm{x}_i)_{i=1}^n$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values on $(\mathbb{X}, \mathcal{X})$ is said to be *finitely exchangeable* if

$$(\mathrm{x}_1, \ldots, \mathrm{x}_n) \overset{d}{=} \left(\mathrm{x}_{\rho(1)}, \ldots, \mathrm{x}_{\rho(n)}\right),$$

for any permutation ρ of the indices $\{1, \ldots, n\}$. An infinite sequence of random variables $(\mathrm{x}_i)_{i\geq 1}$ is *exchangeable* if every subcollection is finitely exchangeable.

Example 1.1. Let $(\mathrm{x}_i)_{i=1}^n$ be an exchangeable sequence and r the correlation coefficient. If $n < \infty$, then $r \geq -\frac{1}{n-1}$ and if $n = \infty$, then $r \geq 0$ (Aldous (1985)). This means that, while independent and identically distributed random variables are clearly exchangeable, exchangeability does not imply independence.

In a Bayesian context, exchangeability replaces the heavy assumption of i.i.d. observations needed in the classical approach for a minimal assumption, in the sense that it reflects nothing more than symmetry (physical independence and sampling order invariance). Furthermore, the representation theorem for exchangeable sequences, first proved by de Finetti (1931) for dichotomic random variables and further extended to general spaces by Hewitt and Savage (1955), characterizes exchangeable sequences by a unique distribution Q on $\mathrm{P}_{\mathbb{X}}$, hence why exchangeability is inherently related to random measures. For an extensive account on exchangeable random elements see Gil-Leyva (2021).

Definition 1.2. Given two measurable spaces $(S, \mathcal{S})$ and $(T, \mathcal{T})$, a *kernel* from S to T is a function $\tilde{\mu} : S \times \mathcal{T} \to \mathbb{R}_+$ such that

1. For any fixed $s \in S$, the mapping $A \to \tilde{\mu}(s, A)$ is a measure on $(T, \mathcal{T})$.
2. For any fixed $A \in \mathcal{T}$, the mapping $s \to \tilde{\mu}(s, A)$ is a measurable function.

If $\tilde{\mu}(s, T) = 1 \ \forall s \in S$, then $\tilde{\mu}$ is a *probability kernel*.

Definition 1.3. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(\mathbb{X}, \mathcal{X})$ be a measurable space. A *random measure* over $(\mathbb{X}, \mathcal{X})$ is a kernel $\tilde{\mu} : \Omega \times \mathcal{X} \to \mathbb{R}_+$. Alternatively, a random measure can be thought as a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$, taking values on $(\mathrm{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$.

Definition 1.4. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(\mathbb{X}, \mathcal{X})$ be a measurable space. A *random probability measure* over $(\mathbb{X}, \mathcal{X})$ is a probability kernel $\tilde{\mu} : \Omega \times \mathcal{X} \to \mathbb{R}_+$, or a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$ that takes values on $(\mathrm{P}_{\mathbb{X}}, \mathcal{P}_{\mathbb{X}})$.

Theorem 1.1. *Let $\mathbb{X}$ be a Polish space and $\mathcal{X}$ its Borel σ-algebra. A sequence $(\mathrm{x}_i)_{i\geq 1}$ of $\mathbb{X}$-valued random variables is exchangeable if and only if there is a probability measure* Q *on* $\mathrm{P}_{\mathbb{X}}$ *such that for every $n \geq 1$ and $(A_i)_{i=1}^n \in \mathcal{X}$*

$$\mathbb{P}\left[\mathrm{x}_1 \in A_1, \ldots, \mathrm{x}_n \in A_n\right] = \int_{\mathrm{P}_{\mathbb{X}}} \prod_{i=1}^n \tilde{p}(A_i) Q(\mathrm{d}\tilde{p}).$$

Moreover, the distribution of $\tilde{p}$ is unique and for every $A \in \mathcal{X}$, the empirical distribution satisfies

$$P_n(A) := \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathrm{x}_i}(A) \overset{a.s.}{\to} \tilde{p}(A)$$

whenever $n \to \infty$. The random probability measure $\tilde{p}$ is known as the directing random measure and its distribution Q *is known as the de Finetti measure of* $(\mathrm{x}_i)_{i\geq 1}$.

The proof of Theorem 1.1 can be found at Appendix A. This theorem is better known as *de Finetti's representation theorem*, and an alternative way to rephrase it is as follows: there exists a random probability measure $\tilde{p}$, with $\tilde{p} \sim \mathrm{Q}$, such that

$$\mathbb{P}\left[\mathrm{x}_1 \in A_1, \ldots, \mathrm{x}_n \in A_n \,|\, \tilde{p}\right] = \prod_{i=1}^{n} \tilde{p}(A_i).$$

This means that, given $\tilde{p}$, $(\mathrm{x}_i)_{i\geq 1}$ are i.i.d. and distributed as $\tilde{p}$. Thus the representation theorem provides an answer to the questions of why we should use parameters and why we should put priors on them, since the product $\prod_{i=1}^{n} \tilde{p}(A_i)$ is formed as if it were a likelihood for $\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ conditional on a quantity $\tilde{p}$, with Q being a prior on $\tilde{p}$. With this in mind, we could think that an exchangeable sequence $(\mathrm{x}_i)_{i\geq 1}$ can be generated as a two step procedure:

1. Generate a random parameter value $\tilde{p} \sim \mathrm{Q}$, i.e. draw a probability distribution at random from the distribution Q.
2. Sample $(\mathrm{x}_i)_{i\geq 1}$ according to $\mathrm{x}_1, \mathrm{x}_2, \ldots \,|\, \tilde{p} \overset{\text{i.i.d}}{\sim} \tilde{p}$.

Alternatively, written in a hierarchical and somewhat more familiar way, exchangeability can be expressed hierarchically as

$$\begin{aligned} \mathrm{x}_i \,|\, \Theta &\overset{\text{i.i.d.}}{\sim} \Theta \\ \Theta &\sim \mathrm{Q}. \end{aligned}$$

1.2 The Bayesian approach

Hereinafter we will adopt the notation used in a Bayesian context and denote as $\mathrm{p}(\mathrm{x})$ the density or probability mass function of the random variable x and as $\mathrm{p}(\mathrm{x} \,|\, \mathrm{y})$ the conditional density or probability mass function of x given y.

Assume that $(\mathrm{x}_i)_{i\geq 1}$ is a presumably exchangeable sequence and let $\mathbf{x}^{(n)} = (\mathrm{x}_1, \ldots, \mathrm{x}_n)$. For the sake of exposition we assume the existence of a density for all of the necessary distributions, so that the Bayesian procedure can be summarized as follows.

1. Suppose that we have chosen a specific prior distribution Q on $\mathrm{P}_{\mathbb{X}}$ for the unknown parameter Θ. $\mathbf{x}^{(n)}$ can then be modeled as i.i.d., sampled from the conditional joint density of $\mathbf{x}^{(n)}$ given Θ

$$\mathrm{p}\left(\mathbf{x}^{(n)} \,|\, \Theta\right) = \prod_{i=1}^{n} \mathrm{p}(\mathrm{x}_i \,|\, \Theta).$$

2. The conditional distribution of Θ given $\mathbf{x}^{(n)}$ is called the *posterior distribution* of Θ and can be obtained by Bayes' theorem as

$$\mathrm{p}\left(\Theta \,|\, \mathbf{x}^{(n)}\right) \propto \mathrm{p}\left(\mathbf{x}^{(n)} \,|\, \Theta\right) \mathrm{p}(\Theta).$$

3. From the posterior distribution we can make inferences about future data via the *posterior predictive distribution*

$$\mathrm{p}\left(\mathrm{x}_{n+1} \,|\, \mathbf{x}^{(n)}\right) = \int \mathrm{p}\left(\mathrm{x}_{n+1} \,|\, \Theta\right) \mathrm{p}\left(\mathrm{d}\Theta \,|\, \mathbf{x}^{(n)}\right),$$

   or make inferences about Θ itself, such as the *posterior mean*

$$\mathbb{E}\left[\Theta \,|\, \mathbf{x}^{(n)}\right] = \int \Theta \mathrm{p}(\mathrm{d}\Theta \,|\, \mathbf{x}^{(n)}).$$

Note that the above procedure would be pointless if instead of exchangeability we were to assume complete independence, as in that case

$$\mathbb{P}\left[\mathrm{x}_{n+1} \in A_{n+1} \,|\, \mathrm{x}_1 \in A_1, \ldots, \mathrm{x}_n \in A_n\right] = \mathbb{P}\left[\mathrm{x}_{n+1} \in A_{n+1}\right]$$

for any $A_1, \ldots, A_{n+1} \in \mathcal{X}$, and therefore previous observations would not provide any information to update Θ. A more comprehensive study about the Bayesian procedure and exchangeability can be consulted in Schervish (1996).

1.2.1 Why going nonparametric?

Let us denote as $\mathbb{S}(\mathrm{Q}) \subseteq \mathrm{P}_{\mathbb{X}}$ the support of the prior distribution Q for Θ, that is the set on which Q concentrates all its mass. In a Bayesian model, $\mathbb{S}(\mathrm{Q})$ characterizes completely the possible observation models because, as Θ takes values in $\mathbb{S}(\mathrm{Q})$, it is necessarily the case that the sequence $(\mathrm{x}_i)_{i\geq 1}$ is generated by a distribution in $\mathbb{S}(\mathrm{Q})$ a.s. When $\mathbb{S}(\mathrm{Q})$ is a finite dimensional subspace of $\mathrm{P}_{\mathbb{X}}$, the resulting Bayesian model is called parametric, whereas whenever the support of Q is an infinite dimensional subspace of $\mathrm{P}_{\mathbb{X}}$, the model is *nonparametric*. In this setting, Q is referred to as a *nonparametric prior*.

There are some situations on which a parametric distribution does not suffice to capture accurately the uncertainty about Θ, let it be due to lack of information about the phenomenon or due to wanting a model that reflects as much as possible the information contained in the observations. In these cases it is more convenient to consider an infinitely-supported prior Q. To illustrate this assertion, suppose that μ_θ is a family of parametric

distributions indexed by the elements θ of the parameter space Θ. The parametric model $\mathrm{x} \,|\, \theta \sim \mu_\theta$ with $\theta \sim \mu$ for $\theta \in \Theta$ can be considered within a nonparametric Bayesian framework as $\mathrm{x} \,|\, \tilde{p} \sim \tilde{p}$ with $\tilde{p} \sim \mathrm{Q}$, where Q has property of $\mathrm{Q}(\{\mu_\theta \,:\, \theta \in \Theta\}) = 1$. Thus, the parametric model may be translated as having a very strong prior opinion, as it uses a prior that assigns probability one to a very small subset of all densities, as depicted in Figure 1.1.

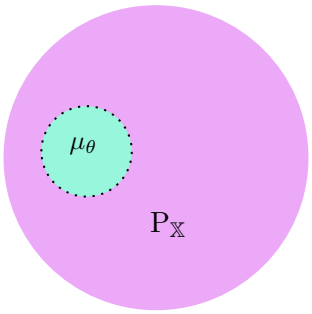


Figure 1.1

The main motivation behind Bayesian nonparametric statistics is to avoid restrictive parametric assumptions about the distribution that generates the data. The appellative nonparametric might be misleading as it gives the impression that are no parameters in the model when in fact, quite the opposite is true, as actually there are infinitely many. Nonparametric should be interpreted as *there is no need to predefine the dimensionality for* θ and might be more fittingly called hugely parametric. It is important to note that, while using infinite (or at least very high)-dimensional priors brings high flexibility, there ain't no such thing as free lunch. Avoiding parametric assumptions about Θ inherently entails the problem of studying and constructing random probability measures over this big parameter spaces. This is a difficult demand mathematically speaking but also computationally, since having such a large space on which inferences can be made requires that the posterior computations are tractable. This is why the use of Bayesian nonparametric models was slowed down during their early development, due to the lack of tools to manipulate the posterior distribution.

There are several procedures for the construction of nonparametric priors: through the specification of finite dimensional distributions $(\tilde{p}(A_1), \ldots, \tilde{p}(A_n))$ for $\{A_i\}_{i=1}^n \subseteq \mathcal{X}$ for $n \geq 1$, or direct methods, such as the one described in Pitman (1996) for constructing a large class of random probability measures called species sampling processes. Here we will focus on priors for exchangeable sequences based on completely random measures, to then extend the methodology to a more general setup.

1.3 Poisson processes

Poisson processes are a key to the probabilistic structure of completely random measures, as Kingman (1967) presented a way to represent completely random measures as a three component sum, one of them based on a Poisson process. In this section we will review some of the most important properties of Poisson processes.

Definition 1.5. Let $(\mathbb{X}, \mathcal{X})$ be a measurable space and let Υ be a random countable subset of $\mathbb{X}$. Let N be a random variable defined as

$$\mathrm{N}(A) = |\Upsilon \cap A| \quad \forall A \in \mathcal{X}.$$

Υ is *Poisson process* on $\mathbb{X}$ if

- For any disjoint measurable sets $A_1, \ldots, A_d \in \mathcal{X}$, $\mathrm{N}(A_1), \ldots, \mathrm{N}(A_d)$ are independent.
- $\mathrm{N}(A) \sim \text{Poisson}(\mu(A))$ where μ is called the *mean measure* and $0 \leq \mu(A) \leq \infty$.

For a measurable set A, $\mathrm{N}(A)$ is the count of the atoms of A that lie in Υ, as depicted in Figure 1.2. Alternatively N can be seen as $\mathrm{N}(\,\cdot\,) = \sum_{\upsilon \in \Upsilon} \delta_\upsilon(\,\cdot\,)$.

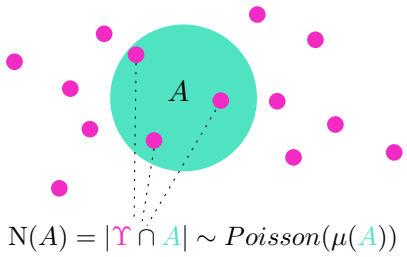


Figure 1.2: The dots are the Poisson process Υ with mean μ, and A is any measurable set.

N is a random measure, referred to as a *Poisson random measure*, since for a disjoint sequence of measurable sets $(B_j)_{j \geq 1}$

$$\mathrm{N}\left(\bigcup_{j=1}^{\infty} B_j\right) = \left|\Upsilon \cap \left(\bigcup_{j \geq 1} B_j\right)\right| = \left|\bigcup_{j=1}^{\infty} (\Upsilon \cap B_j)\right| = \sum_{j=1}^{\infty} |\Upsilon \cap B_j| = \sum_{j=1}^{\infty} \mathrm{N}(B_j),$$

and clearly $\mathrm{N}(\emptyset) = 0$.

Example 1.2. Let $\mathbb{X} = [0, \infty)$ and λ be the Lebesgue measure on $\mathbb{X}$. The Poisson random measure associated with the Poisson process Υ with mean measure λ is just the one-dimensional time-homogeneous Poisson process (a pure-birth Markov chain with birth rate one). Υ is the random set of the jump times of the process.

Remark. If μ is the mean measure of a Poisson process, then it is diffuse, i.e. $\mu(\{x\}) = 0$ for all $x \in \mathbb{X}$. This is because for $x \in \mathbb{X}$, by definition, $\mathbb{P}[\mathrm{N}(\{x\}) = 2] = \frac{\mu(\{x\})^2}{2} e^{-\mu(\{x\})} = 0$, which leads to the result.

A realization of a homogeneous Poisson process is shown in Figure 1.3.

1.3.1 Properties

There are a few operations under which the Poisson process is closed; these are summarized in this section. For easier understanding we omit some of the proofs and refer to Kingman (1993) for further details, whilst some other proofs are attached at Appendix A.

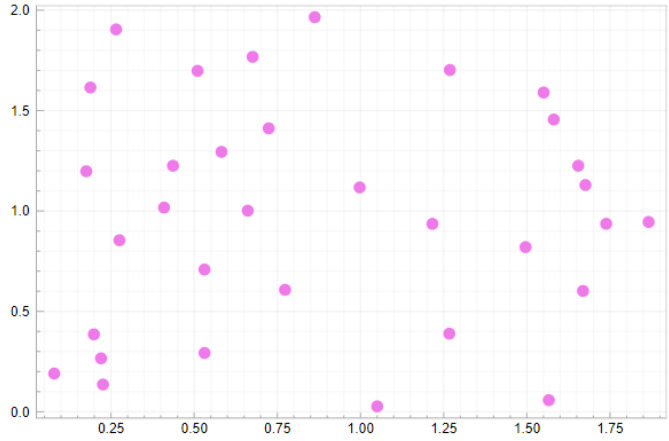


Figure 1.3: Homogeneous Poisson process on $[0,2] \times [0,2]$ *with rate* 10.

Theorem 1.2. *(Disjointness) Let* Υ_1 *and* Υ_2 *be independent Poisson processes with mean measure* μ_1 *and* μ_2 *respectively. If* A *is a measurable set such that* $\mu_1(A)$ *and* $\mu_2(A)$ *are finite, then* Υ_1 *and* Υ_2 *are a.s. disjoint on* A*, i.e.* $\mathbb{P}[\Upsilon_1 \cap \Upsilon_2 \cap A = \emptyset] = 1$.

This can be interpreted intuitively as follows: as the mean measure of a Poisson process is diffuse, the probability of assigning mass two to a point (the two Poisson processes both containing the same point) is zero.

Theorem 1.3. *(Superposition and restriction)*

(1) Let $\Upsilon_1, \Upsilon_2, \ldots$ *be a countable collection of independent Poisson processes on* $\mathbb{X}$ *and let* Υ_j *have mean measure* μ_j *for each* j*. Then* $\bigcup_{j\geq 1} \Upsilon_j$ *is a Poisson process with mean measure* $\mu = \sum_{j=1}^{\infty} \mu_j$.

(2) Let Υ *be a Poisson process on* $\mathbb{X}$ *with mean measure* μ*. For every* $B \in \mathcal{X}$*,* $\Upsilon \cap B$ *is a Poisson process on* $\mathbb{X}$ *with mean measure* $\mu_B(\cdot) = \mu(\cdot \cap B)$*. Equivalently,* $\Upsilon \cap B$ *can also be viewed as a Poisson process with state space* B *with mean measure given by the restriction of* μ *on* B.

The superposition theorem tells us that by joining a countable set of independent Poisson process results in another Poisson process with an updated mean measure, while the restriction theorem can be thought as an intersection operation.

Given a Poisson process, a natural question to ask is what happens if we take a transformation $h(v)$ for each point $v \in \Upsilon$. Interestingly, under certain conditions, a functional transformation of a Poisson process gives another Poisson process on the output space.

Theorem 1.4. *(Mapping) Let* Υ *be a Poisson process with state space* $\mathbb{X}$ *and* σ*-finite mean measure* μ*. Consider a measurable map* h *from* $\mathbb{X}$ *to another polish space* S*. If the pushforward measure* $\mu^*(\cdot) = \mu(h^{-1}(\cdot))$ *is diffuse, then* $h(\Upsilon) = \{h(v) : v \in \Upsilon\}$ *is a Poisson process on* S *with mean measure* μ^*.

All these properties are necessary to establish Campbell's theorem, since with it we will obtain the characteristic functional of the Poisson process, a very powerful tool for deriving further properties. Although Campbell's theorem contains a straightforward computation of the mean and variance of sums over Poisson processes, here we provide a partial version, which serves our purpose.

Theorem 1.5. *(Campbell's) Let Υ be a Poisson process on $\mathbb{X}$ with mean measure μ and let $f : \mathbb{X} \to \mathbb{R}$ be a measurable function. Let*

$$\Sigma = \sum_{v \in \Upsilon} f(v).$$

Σ is absolutely convergent in probability if and only if $\int_{\mathbb{X}} \min(|f(x)|, 1)\mu(\mathrm{d}x) < \infty$. If such condition holds, then

$$\mathbb{E}[e^{t\Sigma}] = exp\left(\int_{\mathbb{X}} \left(e^{tf(x)} - 1\right) \mu(\mathrm{d}x)\right)$$

for every $t \in \mathbb{R}$ such that the integral converges.

See Appendix A for a proof. Restricting Campbell's theorem to $f : \mathbb{X} \to \mathbb{R}^+$ and setting $t = -1$ we obtain the *characteristic functional*

$$\mathbb{E}\left[e^{-\Sigma}\right] = \exp\left\{-\int_{\mathbb{X}} \left(1 - e^{-f(x)}\right) \mu(\mathrm{d}x)\right\}. \tag{1.1}$$

Aptly named, the characteristic functional uniquely characterizes a Poisson process.

Now let Υ be a Poisson process on $\mathbb{X}$ and suppose that for each $v \in \Upsilon$, we assign a random variable $m_v \in T$, where $(T, \mathcal{T})$ is some measurable space. Additionally, the distribution of m_v may depend on v but not on the other points of Υ, and the random variables m_v are independent for different values of v. Note that each pair (v, m_v) can be regarded as a random variable taking values in the product space $\mathbb{X} \times T$, and consequently the whole set of pairs

$$\Upsilon^* := \{(v, m_v) : v \in \Upsilon\} \tag{1.2}$$

is random countable subset of $\mathbb{X} \times T$. We can think of the random variable m_v as a mark on each atom v, as shown Figure 1.4.

Definition 1.6. Let Υ be a Poisson process on $\mathbb{X}$ with mean measure μ and let $p : \mathbb{X} \times \mathcal{T} \to [0, 1]$ be a probability kernel. The random countable set Υ^* as defined in (1.2) is called a *marking of* Υ if its projection onto $\mathbb{X}$ is Υ and the conditional distribution of Υ^* given Υ makes the set of marks $\{m_v\}_{v \in \Upsilon}$ independent and distributed as $p(v, \cdot)$.

We will see next that Υ^* is indeed another Poisson process on $\mathbb{X} \times T$.

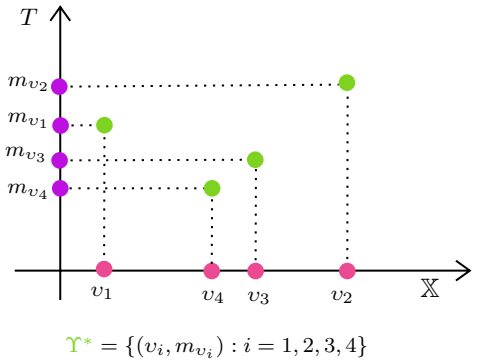


Figure 1.4: The green dots represent Υ^* on the product space $\mathbb{X} \times T$.

Theorem 1.6. *(Marking) The random subset Υ^* is a Poisson process on $\mathbb{X} \times T$ with mean measure μ^* given by*

$$\mu^*(A) = \iint_{(x,m)\in A} \mu(\mathrm{d}x)p(x, \mathrm{d}m).$$

A proof of Theorem 1.6 can be found at Appendix A. An immediate corollary is that the mark random variables $\{m_\upsilon\}_{\upsilon\in\Upsilon}$ themselves form a Poisson process on T: this is true because of the mapping theorem, as m_υ is a projection of Υ^* over T.

1.4 Completely random measures

As aforementioned, we will focus on choices of Q constructed from the transformation or the normalization of completely random measures, and that have almost sure discrete realizations. As these objects will be the building blocks of all nonparametric priors studied in this document, before describing how construct such priors and describe how to use them in Bayesian inference, we will dedicate this section to study of some of their most relevant structural properties.

Definition 1.7. If $\tilde{\mu}$ is a random measure such that, for any $d \geq 2$ and collection of pairwise disjoint sets $A_1, \ldots, A_d \in \mathcal{X}$, the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_d)$ are mutually independent, then $\tilde{\mu}$ is called a *completely random measure*.

In words, completely random measures assign independent masses to disjoint subsets. As we can think of a random probability measure $\tilde{\mu}$ on $(\mathbb{X}, \mathcal{X})$ as a stochastic process $\{\tilde{\mu}(A)\}_{A\in\mathcal{X}}$ indexed by measurable sets, a completely random measure generalizes the notion of independent increments that is familiar in the case in which $\mathbb{X} = \mathbb{R}$. We will dive into this idea further on.

1.4.1 Decomposition of CRMs

Definition 1.8. A completely random measure $\tilde{\mu}$ is said to be Σ-finite if there exist a partition $(S_i)_{i\geq 1} \subseteq \mathcal{X}$ such that $\mathbb{P}\left[\tilde{\mu}(S_i) < \infty\right] > 0 \ \forall i$.

Σ-finite completely random measures can be constructed from Poisson processes and admit a unique decomposition as the summation over three parts: a deterministic measure, a purely atomic measure with fixed atom locations and a discrete measure with random jumps and atoms.

Theorem 1.7. *Let $\tilde{\mu}$ be a Σ-finite completely random measure on $(\mathbb{X}, \mathcal{X})$. $\tilde{\mu}$ can be a.s. decomposed as*

$$\tilde{\mu} = \tilde{\mu}_d + \tilde{\mu}_f + \tilde{\mu}_r,$$

where

- *$\tilde{\mu}_d$ is a non-atomic, non-random measure.*
- *$\tilde{\mu}_f$ is an atomic measure with M fixed atoms $(z_i)_{i=1}^M$ and non-negative, random atom masses $(\mathrm{v}_i)_{i=1}^M$ independent of each other*

$$\tilde{\mu}_f = \sum_{i=1}^{M} \mathrm{v}_i \delta_{z_i},$$

  with $M \in \mathbb{N}_0 \cup \{\infty\}$.

- *$\tilde{\mu}_r$ is an atomic measure with random atom locations $(\zeta_i)_{i\geq 1}$ and random atom masses $(\mathrm{w}_i)_{i\geq 1}$*

$$\tilde{\mu}_r = \sum_{i\geq 1} \mathrm{w}_i \delta_{\zeta_i},$$

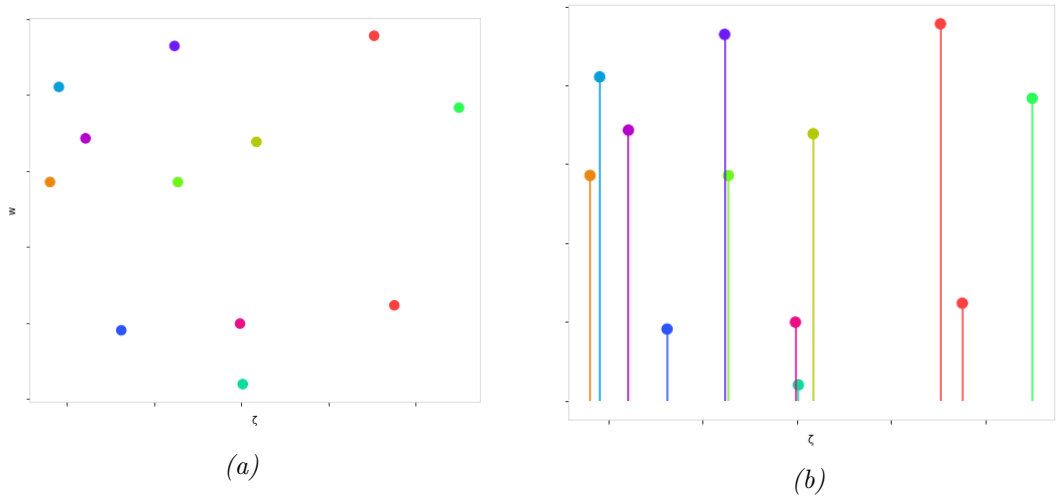
  where $\{(\zeta_i, \mathrm{w}_i)\}_{i\geq 1}$ come from a Poisson process Υ^* with mean measure ν that satisfies $\int_B \int (x \wedge 1)\, \nu(\mathrm{d}v, \mathrm{d}x)$ for $B \in \mathcal{X}$. We will refer to $(\zeta_i)_{i\geq 1}$ indistinctly as the atoms or locations of $\tilde{\mu}$ and to $(\mathrm{w}_i)_{i\geq 1}$ as the masses, weights or jumps of $\tilde{\mu}$.

Remark. Even though Υ^* looks like a marked Poisson process, this is not necessarily true because the projection of ν onto $\mathbb{X}$ need not be σ-finite.

Proof of Theorem 1.7 can be found at Appendix A. From now on we will assume that $\tilde{\mu}$ has no fixed points of discontinuity and no deterministic drift, i.e. we will be working with completely random measures that adopt the form $\tilde{\mu} = \tilde{\mu}_r$ so that its realizations are discrete a.s. An illustration of the basic construction for a completely random measure is given in Figure 1.5. The Poisson process on $\mathbb{R}_+ \times \mathbb{X}$ consists of a countable, and usually infinite, set of points in a product space, as shown in 1.5a. The resulting completely random measure is constructed by dropping lines from each point (w_i, ζ_i) down to $(0, \zeta_i)$, as in 1.5b.

Instead of dealing with $\tilde{\mu}$ itself, we will focus our attention on the measure ν, often called the *Lévy intensity measure*. It regulates the intensity of the jumps and their locations, and according to Kallenberg (2017), ν can be decomposed as

$$\nu(\mathrm{d}v, \mathrm{d}x) = \alpha(\mathrm{d}x)\rho_x(\mathrm{d}v),$$



(a) *(b)*

Figure 1.5

where α is a σ-finite measure on $(\mathbb{X}, \mathcal{X})$ and ρ_x is a kernel on $\mathbb{X} \times \mathbb{R}_+$. ρ_x controls the jump intensity and α determines the location of the jumps. If $\rho_x(\mathrm{d}v) \equiv \rho(\mathrm{d}v)$ for some measure ρ on $\mathbb{R}_+$, then the masses are independent of the locations and both $\tilde{\mu}$ and ν are said to be *homogeneous*. If not, then the masses depend on the locations and $\tilde{\mu}$ and ν are *non-homogeneous*. We will write $\tilde{\mu} \sim \mathrm{CRM}\,(\rho_x, \alpha)$ to refer to the distribution of a completely random measure whose Lévy intensity is $\nu(\mathrm{d}v, \mathrm{d}x) = \alpha(\mathrm{d}x)\rho_x(\mathrm{d}v)$. If in particular α is a finite measure with total mass θ, so that $\alpha(\mathrm{d}x) = \theta P_0(\mathrm{d}x)$, we will write $\tilde{\mu} \sim \mathrm{CRM}(\rho, \theta, P_0)$.

The intensity ν encodes all the information needed to generate a sample from a CRM. Ferguson and Klass (1972) provided a way of generating a realization of $\tilde{\mu}$ by sampling from the underlying Poisson process, specifically by sampling the weights in a decreasing order. In here, the authors proved that the distribution of the ordered weights $\mathrm{w}_{(1)} \geq \mathrm{w}_{(2)} \geq \cdots$ depends only on $\mathrm{w}_{(1)}$, namely the distribution of $\mathrm{w}_{(j)}$ given $\left(\mathrm{w}_{(i)}\right)_{i=1}^{j-1}$ equals the distribution of the largest weight $\mathrm{w}_{(1)}$, truncated from above. Algorithm 1.1 describes this procedure.

Algorithm 1.1: Sample of a CRM $\tilde{\mu}$ with Lévy intensity $\nu(\mathrm{d}v, \mathrm{d}x) = \alpha(\mathrm{d}x)\rho_x(\mathrm{d}v)$.

Sample the atoms $\zeta_i \overset{\text{i.i.d.}}{\sim} \alpha(\mathrm{d}x)$.

Sample the atom's masses according to

$$\mathbb{P}[\mathrm{w}_{(1)} \leq w_1] = \exp\left(-\int_{w_1}^{\infty} \rho_{\zeta_1}(\mathrm{d}w)\right) \text{ for } 0 < w_1$$

$$\mathbb{P}[\mathrm{w}_{(2)} \leq w_2 \,|\, \mathrm{w}_{(1)} = w_1] = \exp\left(-\int_{w_2}^{w_1} \rho_{\zeta_2}(\mathrm{d}w)\right) \text{ for } 0 < w_2 < w_1$$

$$\vdots$$

$$\mathbb{P}[\mathrm{w}_{(j)} = w_j \,|\, \mathrm{w}_{(1)} = w_1, \ldots, \mathrm{w}_{(j-1)} = w_{j-1}] = \exp\left(-\int_{w_{j-1}}^{w_1} \rho_{\zeta_j}(\mathrm{d}w)\right) \text{ for } 0 < w_j < \cdots < w_1$$

Just as random variables are characterized by their Laplace transform, completely random measures are characterized by their Laplace functional.

Definition 1.9. The *Laplace functional* of the completely random measure $\tilde{\mu}$ is given by

$$\mathbb{E}\left[e^{-\int_{\mathbb{X}} f(x)\tilde{\mu}(\mathrm{d}x)}\right] = \exp\left[-\int_{\mathbb{R}_+}\int_{\mathbb{X}}\left(1-e^{-vf(x)}\right)\nu(\mathrm{d}v,\mathrm{d}x)\right] \tag{1.3}$$

for any measurable function $f : \mathbb{X} \to \mathbb{R}_+$. In particular, choosing $f = t\mathbb{1}_A$ with $A \in \mathcal{X}$ and $t > 0$ yields the *Laplace transform* of the random variable $\tilde{\mu}(A)$. We will write $\mathbb{E}\left[e^{-\int_{\mathbb{X}} f(x)\tilde{\mu}(\mathrm{d}x)}\right] = \mathbb{E}[e^{-\tilde{\mu}(f)}]$.

Definition 1.10. The *Laplace exponent* of the CRM $\tilde{\mu}$ is given by

$$\psi(f) = \int_{\mathbb{R}_+}\int_{\mathbb{X}}\left(1-e^{-vf(x)}\right)\nu(\mathrm{d}v,\mathrm{d}x). \tag{1.4}$$

In particular, if $\tilde{\mu}$ is homogeneous and $\theta = \alpha(\mathbb{X}) < \infty$

$$\psi(f) = \theta\int_{\mathbb{R}_+}\left(1-e^{-vf}\right)\rho(\mathrm{d}v).$$

1.4.2 Increasing additive processes

Now a brief review on additive processes will be made to apply all the theory developed above to the case where $\mathbb{X} = \mathbb{R}$.

Definition 1.11. A stochastic process $(\xi_s)_{s\geq 0}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is an *additive process* if

- $\xi_0 = 0$ a.s.
- For any choice of $n \geq 1$ and $0 \leq s_1 < \cdots < s_n$, the random variables $\{\xi_{s_{j+1}} - \xi_{s_j}\}_{j=1}^{n-1}$ are independent (independent increments property).
- There is a measurable Ω_0 such that $\mathbb{P}[\Omega_0] = 1$ and for $\omega \in \Omega_0$, $\xi_s(\omega)$ is right-continuous in $s \geq 0$ and has left limits in $s > 0$ (càdlàg trajectories).
- $\lim_{h\to 0}\mathbb{P}\left[|\xi_{s+h} - \xi_s| \geq \epsilon\right] = 0$ for $s \geq 0$ and $\epsilon > 0$ (it is stochastically continuous).

It is called a *Lévy process* if, in addition, $(\xi_{s+t} - \xi_s) \overset{d}{=} \xi_t$ for all $s, t \geq 0$ (stationary increments property).

Definition 1.12. An additive process $(\xi_s)_{s\geq 0}$ with a.s. increasing sample paths is named an *increasing additive process*. If $(\xi_s)_{s\geq 0}$ is also a Lévy process, then it is termed a *subordinator*.

Completely random measures and subordinators relate to each other in the following way: let $\tilde{\mu}$ be a completely random measure on $\mathbb{R}$, bounded on finite sets a.s. and let ξ be its right continuous distribution function, defined as

$$\xi(s) = \begin{cases} \tilde{\mu}((0,s]) & \text{if } s \geq 0 \\ -\tilde{\mu}((s,0]) & \text{if } s < 0 \end{cases}$$

From the independence property of $\tilde{\mu}$ over disjoint sets, for a strictly increasing set of indexes $(s_i)_{i=1}^n$ the random variables

$$\xi(s_{i+1}) - \xi(s_i) = \tilde{\mu}\left((s_i, s_{i+1}]\right)$$

are independent. By construction ξ is increasing and right-continuous, meaning that the process $(\xi_s)_{s\in\mathbb{R}}$ defines an increasing additive process, where we adopted the notation $\xi(s) = \xi_s$. Therefore an IAP can be seen as the càdlàg distribution function induced by a completely random measure on $\mathbb{R}$. From Theorem 1.7 we know that ξ_s can be decomposed into the superposition of three independent processes, specifically an increasing deterministic process, a random increasing function that jumps at fixed discontinuities and a random component that can be described by a Poisson process on the half plane $\Upsilon^* = \{(\zeta, \mathrm{w}) : \mathrm{w} > 0\}$.

Now assume stationary increments so that $(\xi_s)_{s\in\mathbb{R}}$ is a subordinator. Stationarity then rules out the existence of fixed atoms, and the deterministic process is a constant multiple of the Lebesgue measure. The Poisson process in the representation of Theorem 1.7 inherits the stationary property from $(\xi_s)_{s\in\mathbb{R}}$. In the light of this, a subordinator has for $s \in \mathbb{R}$ the representation

$$\xi_s = \begin{cases} \beta s + \sum_{i\geq 1} \mathrm{w}_i \mathbb{1}_{(0,s]}(\zeta_i) & \text{for } t \geq 0 \\ \beta s - \sum_{i\geq 1} \mathrm{w}_i \mathbb{1}_{[s,0)}(\zeta_i) & \text{for } t < 0 \end{cases}$$

where $\beta \geq 0$ and Υ^* is a Poisson process on $\mathbb{X}^*$ that is invariant under translations on the x-axis. The mean measure must have the form $\nu(\mathrm{d}z)\mathrm{d}x$, where ν is a measure on $\mathbb{R}_+$ such that $\int (1 - e^{-z})\, \nu(\mathrm{d}z) < \infty$. Consequently we've found a characterization of subordinators based on the perspective of completely random measures. This result can be re-expressed in the same way as it appears in Bertoin (1996), considering the restriction of ξ to $\mathbb{R}_+$.

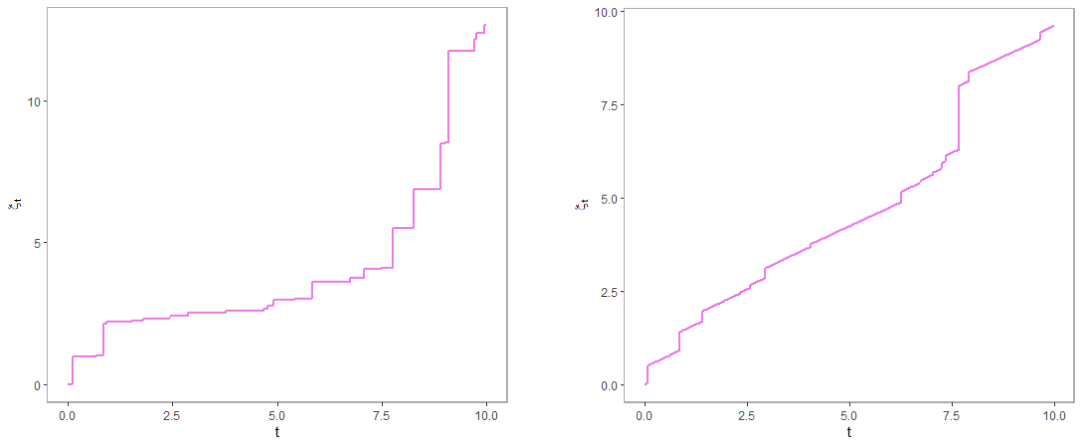
Theorem 1.8. *A subordinator $(\xi_s)_{s\geq 0}$ can be written as*

$$\xi_s = \beta s + \sum_{i\geq 1} \mathrm{w}_i \mathbb{1}_{\{\zeta_i \leq s\}},$$

for some constant $\beta \geq 0$ and where $\{(\zeta_i, \mathrm{w}_i)\}_{i\geq 1}$ is a countable set of points of a Poisson process with intensity $\nu(\mathrm{d}z)\mathrm{d}x$, where ν is a Lévy measure.

Subordinators can then be decomposed into two components: a deterministic drift component and a Poisson point process. Figure 1.6 shows the trajectories of two subordinators, one without drift and one with positive drift. Note that in both scenarios, if ξ_T is finite at a time T, its jumps partition the interval $[0, T)$.

The atoms of $\tilde{\mu}$ correspond to the jumps of ξ, and these occur at the projection Υ on the x-axis of Υ^*. By the mapping theorem, Υ is again a Poisson process with constant mean measure equal to $\nu(\mathbb{R}_+)$. Since for any $A \in \mathcal{B}(\mathbb{R})$ the random variable $\mathrm{N}(A)$ is Poisson distributed and counts how many jumps occur on A, if $\nu(\mathbb{R}_+) = \infty$, then $\mathrm{N}(A) = \infty$ a.s. and Υ is a dense set on $\mathbb{R}$, i.e. ξ is an *infinite activity process.*



(a) Subordinator without drift ($\beta = 0$). *(b) Subordinator with positive drift ($\beta > 0$).*

Figure 1.6

Finally, the Laplace transform (1.3) of subordinators has the form

$$\mathbb{E}\left[e^{-t\tilde{\mu}((0,s])}\right] = \mathbb{E}\left[e^{-t\xi_s}\right] = \exp\left[-s\left(\beta t + \int_{\mathbb{R}_+}(1-e^{tx})\nu(\mathrm{d}x)\right)\right].$$

Note that the above reasoning can be reverted: if $(\xi_s)_{s\geq 0}$ is a subordinator, $\tilde{\mu}$ defined as

$$\tilde{\mu}((s,t]) = \xi_t - \xi_s$$

is a completely random measure.

1.4.3 Further examples

Example 1.3. A *gamma random measure* $\tilde{\mu}$ is a homogeneous CRM with Lévy intensity given by

$$\nu(\mathrm{d}v, \mathrm{d}x) = \frac{e^{-v}}{v}\mathrm{d}v\alpha(\mathrm{d}x).$$

For $f : \mathbb{X} \to \mathbb{R}+$ measurable, using the series representation

$$\left(1 - e^{-vf(x)}\right) = \sum_{i\geq 1}\frac{(-1)^{i+1}(vf(x))^i}{i!},$$

we can compute

$$\int_{\mathbb{R}_+}\left(1-e^{-vf(x)}\right)e^{-v}v^{-1}\mathrm{d}v = \int_{\mathbb{R}^+}\sum_{i\geq 1}\frac{(-1)^{i+1}(vf(x))^i}{i!}e^{-v}v^{-1}\mathrm{d}v$$

$$
\begin{aligned}
&= \sum_{i\geq 1} \frac{(-1)^{i+1} f(x)^i}{i!} \int_{\mathbb{R}^+} v^{i-1} e^{-v} \mathrm{d}v \\
&= \sum_{i\geq 1} \frac{(-1)^{i+1} f(x)^i}{i!} \Gamma(i) \\
&= \log(1 + f(x)).
\end{aligned}
$$

The Laplace functional of $\tilde{\mu}$ is

$$
\mathbb{E}[e^{-\tilde{\mu}(f)}] = \exp\left[- \int_{\mathbb{X}} \log\left(1 + f(x)\right) \alpha(\mathrm{d}x) \right],
$$

as long as $\int \log(1+f)\alpha < \infty$. Taking $f = t\mathbb{1}_A$ with $A \in \mathcal{X}$ and $t > 0$ yields

$$
\mathbb{E}\left[e^{-t\tilde{\mu}(A)} \right] = (1+t)^{-\alpha(A)}.
$$

As a result $\tilde{\mu}(A) \sim \mathrm{Ga}(\alpha(A), 1)$.

Example 1.4. For $\sigma \in (0,1)$, a *σ-stable CRM* is a random measure $\tilde{\mu}_\sigma$ with Lévy intensity

$$
\nu(\mathrm{d}v, \mathrm{d}x) = \frac{\sigma v^{-1-\sigma}}{\Gamma(1-\sigma)} \mathrm{d}v \alpha(\mathrm{d}x).
$$

If $f : \mathbb{X} \to \mathbb{R}_+$ is measurable, then

$$
\begin{aligned}
\int_{\mathbb{R}_+} \left(1 - e^{-vf(x)}\right) v^{-1-\sigma} \mathrm{d}v &= -\left(1 - e^{-vf(x)}\right) v^{-\sigma} \sigma^{-1} \big|_0^\infty + \int_{\mathbb{R}_+} v^{-\sigma} f(x) e^{-vf(x)} \sigma^{-1} \mathrm{d}v \\
&= f(x)^\sigma \sigma^{-1} \int_{\mathbb{R}_+} u^{-\sigma} e^{-u} \mathrm{d}u \\
&= \Gamma(1-\sigma) \sigma^{-1} f(x)^\sigma,
\end{aligned}
$$

where the first equality follows by integrating by parts and the second one by the change of variable $u = vf(x)$. Therefore the Laplace functional of $\tilde{\mu}_\sigma$ is

$$
\mathbb{E}[e^{-\tilde{\mu}(f)}] = \exp\left[- \int_{\mathbb{X}} f(x)^\sigma \alpha(\mathrm{d}x) \right]
$$

for f such that $\int_{\mathbb{X}} f(x)^\sigma \alpha < \infty$. Again considering $f = t\mathbb{1}_A$ for any $t > 0$ and $A \in \mathcal{X}$, the Laplace transform of $\tilde{\mu}_\sigma(A)$ is given by

$$
\mathbb{E}\left[e^{-t\tilde{\mu}_\sigma(A)} \right] = e^{-t^\sigma \alpha(A)}.
$$

Consequently $\tilde{\mu}_\sigma(A)$ follows a positive stable distribution.

Example 1.5. A CRM characterized by a Lévy intensity of the form

$$
\rho(\mathrm{d}v)\alpha(\mathrm{d}x) = \frac{e^{-\kappa v} v^{-1-\gamma}}{\Gamma(1-\gamma)} \mathrm{d}v \theta P_0(\mathrm{d}x)
$$

where $\kappa \geq 0$ and $\gamma \in (0,1)$, is known as a *generalized Gamma CRM*, see Brix (1999).

Normalized random measures

Recall that from now on we will focus on choices of completely random measures $\tilde{\mu}$ without drift and without fixed atoms, that is $\tilde{\mu}(\,\cdot\,) = \sum_{i\geq 1} \mathrm{w}_i \delta_{\zeta_i}(\,\cdot\,)$ where $(\mathrm{w}_i)_{i\geq 1}$ are positive random jumps and $(\zeta_i)_{i\geq 1}$ random locations. Imposing certain conditions on the Lévy intensity that characterizes $\tilde{\mu}$ can ensure that the total mass, $\tilde{\mu}(\mathbb{X})$, is positive and finite a.s. and in such cases, Regazzini et al. (2003) introduced a way to construct random probability measures through the normalization of CRMs. In Section 2.1 we describe such random probability measures, called normalized random measures with independent increments (NRMI), an overview of their basic properties and definitions of important quantities related to them. This Section is based on the work of Regazzini et al. (2003), James et al. (2006) and James et al. (2009). Section 2.2 exposes concrete examples of NRMIs while Section 2.3 describes the Pitman-Yor process, which is not a NRMI but can be constructed as the normalization of a random measure that is not completely random.

2.1 Normalized random measures with independent increments

Definition 2.1. Let $\tilde{\mu}$ be a completely random measure on $(\mathbb{X}, \mathcal{X})$ with Lévy intensity ν such that $0 < \tilde{\mu}(\mathbb{X}) < \infty$ a.s. A *normalized random measure with independent increments* (NRMI) on $(\mathbb{X}, \mathcal{X})$ is a random probability measure defined as

$$\tilde{p}(A) := \frac{\tilde{\mu}(A)}{\tilde{\mu}(\mathbb{X})}, \tag{2.1}$$

for any $A \in \mathcal{X}$.

In terms of the Lévy intensity:

- $\tilde{\mu}(\mathbb{X}) < \infty$ a.s is equivalent to ask that the Laplace exponent $\psi(u)$ is finite for any $u > 0$. If ν is homogeneous then this is true whenever $\alpha(\mathrm{d}x)$ is a finite measure.
- $\tilde{\mu}(\mathbb{X}) > 0$ a.s. if $\nu(\mathbb{R}_+ \times \mathbb{X}) = \infty$. In the homogeneous setting $\rho(\mathbb{R}_+) = \infty$ ensures this, meaning that the completely random measure must have infinite activity.

The independent increment name comes from the fact that originally NRMIs were presented as reparameterized increasing additive procesess on $\mathbb{X} = \mathbb{R}$. Specifically, let α be a non-null finite measure on $\mathbb{R}$ with total mass θ and let A be its distribution, that is $A(x) = \alpha((-\infty, x])$ for $x \in \mathbb{R}$. Let $(\xi_t)_{t\geq 0}$ be an increasing additive process defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $0 < \xi_\theta < \infty$ a.s. As for an additive process, the marginal distribution of each ξ_t is infinitely divisible for all $t \geq 0$, the Lévy-Khintchine formula holds (see Theorem A.1 on Appendix A). In terms of the triplet $(a_\theta, \sigma_\theta, \nu_\theta)$, the condition $0 < \xi_\theta < \infty$ a.s. is satisfied if $\nu_\theta(\mathbb{R}_+) = \infty$. A NRMI is the random measure $\tilde{p}$ defined is terms of the time changed process $\left(\xi_{A(x)}\right)_{x\in\mathbb{R}}$ as

$$\tilde{p}((-\infty, x]) = \frac{\xi_{A(x)}}{\xi_\theta} \quad \text{for } x \in \mathbb{R}.$$

If in particular $(\xi_t)_{t\geq 0}$ is a subordinator without drift, $\xi_\theta > 0$ a.s. if ξ is a infinite activity process. From the subordinator decomposition of Theorem 1.8, we know that in this case the process ξ jumps infinitely often in any finite interval; the time change ξ_A can be seen as observing the process ξ over the finite interval $(0, \theta)$, with a time deformation governed by the measure α. $\left(\xi_{A(x)}\right)_{x\in\mathbb{R}}$ inherits all but the stationary property of ξ, i.e. it is an increasing additive process and, accordingly, can be seen as the distribution function induced by a CRM on $\mathbb{R}$. Normalizing its jumps yields a monotone increasing stochastic process on $[0, 1]$. In an abstract space, this procedure can be thought as to normalizing all the lines from Figure 1.5b so their sum is one.

Strictly speaking we should term the random probability measure in (2.1) a normalized CRM and it reduces to a NRMI when $\mathbb{X} = \mathbb{R}$, although we preserve the term NRMI on abstract spaces. We will be working mostly in the homogeneous case and further on α will be assumed to be a non-null finite measure of total mass θ. By defining

$$P_0(\,\cdot\,) := \frac{\alpha(\,\cdot\,)}{\theta},$$

we can set $\alpha(\mathrm{d}x) = \theta P_0(\mathrm{d}x)$. We will write NRMI$(\rho, \theta, P_0)$ to refer to the distribution of a homogeneous NRMI whose associated CRM has intensity $\rho(\mathrm{d}v)\theta P_0(\mathrm{d}x)$ and we will refer to P_0 as the *base measure*.

2.1.1 Moments and covariance

When the base measure P_0 is non-atomic, expressions for the moments of homogeneous NRMIs exist.

Proposition 2.1. If $\tilde{p} \sim$ NRMI(ρ, θ, P_0) with P_0 non-atomic and $A, B \in \mathcal{X}$, then

$$\begin{aligned}
\mathbb{E}\left[\tilde{p}(A)\right] &= P_0(A) \\
\operatorname{var}\left[\tilde{p}(A)\right] &= P_0(A)\left(1 - P_0(A)\right)\mathcal{I}_\theta \\
\operatorname{cov}\left(\tilde{p}(A), \tilde{p}(B)\right) &= \left[P_0(A \cap B) - P_0(A)P_0(B)\right]\mathcal{I}_\theta,
\end{aligned}$$

where $\Psi(u) := \frac{\psi(u)}{\theta}$ is the normalized Laplace exponent and $\mathcal{I}_\theta := -\theta \int_{\mathbb{R}_+} u e^{-\psi(u)} \frac{\mathrm{d}^2}{\mathrm{d}u^2} \Psi(u) \mathrm{d}u$.

Hence a homogeneous NRMI $\tilde{p}$ has a prior guess at the shape $P_0(\,\cdot\,) = \mathbb{E}[\tilde{p}(\,\cdot\,)]$ and the quantity $\mathcal{I}_\theta$ controls the dispersion around P_0.

2.1.2 Partition structure

From now on we will suppose that $(\mathrm{x}_i)_{i\geq 1}$ is a sequence of exchangeable random variables such that for $n \geq 1$

$$\begin{aligned} \mathrm{x}_i \,|\, \tilde{p} &\sim \tilde{p} \quad i = 1, \ldots, n \\ \tilde{p} &\sim \mathrm{NRMI}(\rho, \theta, P_0). \end{aligned} \tag{2.2}$$

Consider $\mathbf{x}^{(n)} = (\mathrm{x}_1, \ldots, \mathrm{x}_n)$ a sample of size n from the model (2.2). The a.s. discreteness of NRMIs implies that there will be duplicated values, also called *ties*, with positive probability among the observations. In this section we will see how the ties that appear in a sample of an NRMI generate a partition over the set $\{1, \ldots, n\}$, and to this aim we will first make a brief review on random partitions.

Definition 2.2. Given a finite set B, a *partition of B into k blocks* is an unordered collection of k non-empty disjoint sets $\{B_i\}_{i=1}^k$ such that $B_i \subseteq B$ for each $i = 1, \ldots, k$ and $\cup_{i=1}^k B_i = B$.

Let $[n] := \{1, 2, \ldots, n\}$. Denote as $\mathcal{P}_{[n]}^k$ the set of partitions of $[n]$ into k blocks and $\mathcal{P}_n := \bigcup_{k=1}^n \mathcal{P}_{[n]}^k$ the set of all possible partitions of $[n]$. Given a partition of B into k blocks, the vector of the sizes of each block $(|B_1|, \ldots, |B_k|)$ of the partition defines *a composition of n* into k parts, i.e. it is a sequence of k positive integers such that their sum is equal to n. Denote as $\mathcal{C}_n^k$ the set of compositions of n into k parts and $\mathcal{C}_n := \bigcup_{k=1}^n \mathcal{C}_n^k$ the set of all possible compositions of n. $\mathcal{P}_{[3]}$ is presented below as an example.

$$\mathcal{P}_{[3]} = \mathcal{P}_{[3]}^1 \cup \mathcal{P}_{[3]}^2 \cup \mathcal{P}_{[3]}^3$$

$$\mathcal{P}_{[3]}^1 = \Big\{ \{\{1,2,3\}\} \Big\}$$

$$\mathcal{P}_{[3]}^2 = \Big\{ \{\{1\},\{2,3\}\}, \{\{2\},\{1,3\}\}, \{\{3\},\{1,2\}\} \Big\}$$

$$\mathcal{P}_{[3]}^3 = \Big\{ \{\{1\},\{2\},\{3\}\} \Big\}$$

The number of different ways to partition the set $[n]$ into k blocks equals to the *Stirling numbers of the second kind*, which are defined in terms of the (n,k)*-th incomplete Bell polynomial* $B_{n,k}$ as follows

$$B_{n,k}(x) := \frac{n!}{k!} \sum_{(\mathrm{n}_1, \ldots, \mathrm{n}_k) \in \mathcal{C}_{[n]}^k} \prod_{i=1}^k \frac{x_{\mathrm{n}_i}}{\mathrm{n}_i!} \quad \text{for } x = (x_1, x_2, \ldots)$$

$$S_{n,k} := \left| \mathcal{P}_{[n]}^k \right| = B_{n,k}(\bar{1}) \quad \text{where } \bar{1} = (1, 1, \ldots).$$

Definition 2.3. A *random partition* Π_n is a random variable that takes values on $\mathcal{P}_{[n]}$.

A random object that is implicitly defined on a random partition is K_n, the number of blocks of Π_n. If we were to know $\mathbb{P}[\Pi_n = \pi]$ for all partitions $\pi \in \mathcal{P}_{[n]}$, the marginal distribution of K_n could be computed summing over all possible partitions, that is

$$\mathbb{P}[\mathrm{K}_n = k] = \sum_{\pi \in \mathcal{P}^k_{[n]}} \mathbb{P}[\Pi_n = \pi],$$

for $k = 1, \ldots, n$.

Definition 2.4. A random partition Π_n is called *exchangeable* if for every permutation $\rho : [n] \to [n]$ and $\pi \in \mathcal{P}_{[n]}$

$$\mathbb{P}\left[\Pi_n = \pi\right] = \mathbb{P}\left[\Pi_n = \rho(\pi)\right],$$

where $\rho(\pi)$ is the partition resulting from permuting each element according to ρ. A sequence of random partitions $(\Pi_n)_{n\geq 1}$, where Π_n is a random partition of $[n]$, is exchangeable if Π_n is exchangeable for every $n \geq 1$.

The exchangeability assumption is equivalent to saying that the distribution of the partition should depend only on the unordered sizes of the blocks. Therefore, there exists a function $\Phi_k^{(n)}$ of compositions that is symmetric in its arguments such that, for any specific partition assignment $\pi = \{B_1, \ldots, B_k\}$ and $\mathrm{n}_i := |B_i|$ for $i = 1, \ldots, k$, we have that

$$\mathbb{P}\left[\Pi_n = \pi\right] =: \Phi_k^{(n)}\left(\mathrm{n}_1, \ldots, \mathrm{n}_k\right).$$

The function $\Phi_k^{(n)}$ is called the *exchangeable partition probability function* (EPPF) (Pitman (1995)).

Definition 2.5. The sequence of random partitions $(\Pi_n)_{n\geq 1}$ is *consistent* if the projection of Π_n to $[m]$ is a.s. equal to Π_m for every $m < n$ and $n > 1$.

Consistency implies that each Π_n is the partition resulting from Π_{n+1} by discarding, from the latter, the integer $n+1$; we can think of this as if the indices arrive one at a time: first 1, then 2 up to n or beyond.

A consistent sequence $(\Pi_n)_{n\geq 1}$ can be regarded as a random element of the set $\mathcal{P}_{\mathbb{N}}$ of partitions of $\mathbb{N}$, equipped with the σ-algebra generated by the restriction maps from $\mathcal{P}_{\mathbb{N}}$ to $\mathcal{P}_{[n]}$ for all $n \geq 1$. On the other hand, an exchangeable sequence $(\Pi_n)_{n\geq 1}$ is consistent if its EPPF satisfy the following *addition rule*

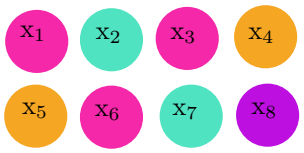
$$\Phi_k^{(n)}(\mathrm{n}_1, \ldots, \mathrm{n}_k) = \Phi_{k+1}^{(n+1)}(\mathrm{n}_1, \ldots, \mathrm{n}_k, 1) + \sum_{j=1}^{k} \Phi_k^{(n)}(\mathrm{n}_1, \ldots, \mathrm{n}_{j-1}, \mathrm{n}_{j+1}, \ldots, \mathrm{n}_k),$$

for each composition $(\mathrm{n}_1, \ldots, \mathrm{n}_k)$ of n. Thus the distribution of a consistent exchangeable random partition of $\mathbb{N}$ is determined by an EPPF, subject to the previous addition rule. For an exhaustive account on random partitions see Gil-Leyva (2016) and Pitman (2006).

In general, sampling from an exchangeable sequence with an a.s. discrete de Finetti measure naturally leads to look at the induced partition structure. In this particular case, for a given $n \geq 1$, a sample $\mathbf{x}^{(n)}$ from model (2.2) exhibits k unique values $\{\mathrm{x}_1^*, \ldots, \mathrm{x}_k^*\}$, each with frequency n_j such that $(\mathrm{n}_1, \ldots, \mathrm{n}_k)$ is a composition of n. The appearance of ties among $\mathbf{x}^{(n)}$ induce an equivalence relation on $[n]$ where

$$i \sim j \quad \Longleftrightarrow \quad \mathrm{x}_i = \mathrm{x}_j.$$

This equivalence relation, in turn, induces a partition Π_n of $[n]$. Due to the exchangeability of $(\mathrm{x}_1, \mathrm{x}_2, \ldots)$, the distribution of the random partition Π_n only depends on the number of unique values displayed in the sample and their frequencies. Hence we can characterize its distribution by an EPPF and moreover, the sequence of partitions are consistent, as they are generated by consecutive sampling. An example is shown in Figure 2.1, where the sample $\mathbf{x}^{(8)}$ induces a partition over $[8]$. Different colors represent different values, meaning that $k = 4$ and n_i $(i = 1, \ldots, 4)$ is the frequency of each color. Π_i is the partition generated up to stage i for $1 \leq i \leq 8$.



$\mathrm{n}_1 = 3 \quad \mathrm{n}_2 = 2 \quad \mathrm{n}_3 = 2 \quad \mathrm{n}_4 = 1$

$$\Pi_1 = \Big\{\{1\}\Big\} \qquad \Pi_5 = \Big\{\{1,3\},\{2\},\{4,5\}\Big\}$$
$$\Pi_2 = \Big\{\{1\},\{2\}\Big\} \qquad \Pi_6 = \Big\{\{1,3,6\},\{2\},\{4,5\}\Big\}$$
$$\Pi_3 = \Big\{\{1,3\},\{2\}\Big\} \qquad \Pi_7 = \Big\{\{1,3,6\},\{2,7\},\{4,5\}\Big\}$$
$$\Pi_4 = \Big\{\{1,3\},\{2\},\{4\}\Big\} \qquad \Pi_8 = \Big\{\{1,3,6\},\{2,7\},\{4,5\},\{8\}\Big\}$$

Figure 2.1: Partitions generated by consecutive sampling.

Although not always available in closed form, there is an expression for the EPPF of homogeneous NRMIs.

Proposition 2.2. Let $\tilde{p} \sim \mathrm{NRMI}(\rho, \theta, P_0)$. The EPPF of the partition induced by the sample $\mathbf{x}^{(n)}$ that exihibts k unique values equals

$$\Phi_k^{(n)}(\mathrm{n}_1, \ldots, \mathrm{n}_k) = \frac{\theta^k}{\Gamma(n)} \int_{\mathbb{R}_+} u^{n-1} e^{-\theta\Psi(u)} \left[\prod_{j=1}^{k} \tau_{\mathrm{n}_j}(u)\right] \mathrm{d}u, \tag{2.3}$$

where for every $m \geq 1$

$$\tau_m(u) := \int_{\mathbb{R}_+} v^m e^{-uv} \rho(\mathrm{d}v).$$

A proof of Proposition 2.2 can be found at Appendix A. Having the EPPF gives information about the clustering behavior of the prior process and moreover, the predictive distribution can be expressed in terms of the EPPF by noting that

$$\begin{aligned}\mathbb{P}\left[\mathrm{x}_{n+1} = \text{new} \,|\, \mathbf{x}^{(n)}\right] &= \mathbb{P}\left[\mathrm{x}_{n+1} \neq \mathrm{x}_j^* \,\forall j \in \{1,\ldots,k\} \,|\, \mathbf{x}^{(n)}\right] \\ &= \frac{\Phi_{k+1}^{(n+1)}(\mathrm{n}_1,\ldots,\mathrm{n}_k,1)}{\Phi_k^{(n)}(\mathrm{n}_1,\ldots,\mathrm{n}_k)} \qquad (2.4) \\ \mathbb{P}\left[\mathrm{x}_{n+1} = \text{old} \,|\, \mathbf{x}^{(n)}\right] &= \mathbb{P}\left[\mathrm{x}_{n+1} = \mathrm{x}_j^* \,|\, \mathbf{x}^{(n)}\right] \\ &= \frac{\Phi_k^{(n+1)}(\mathrm{n}_1,\ldots,\mathrm{n}_j+1,\ldots,\mathrm{n}_k)}{\Phi_k^{(n)}(\mathrm{n}_1,\ldots,\mathrm{n}_k)}. \qquad (2.5)\end{aligned}$$

Equation (2.4) means we would generate a new value $\mathrm{x}_{n+1} = \mathrm{x}_{k+1}^*$ that will naturally have frequency one. On the other hand, (2.5) just means that x_{n+1} equals a repeated value, so we augment the frequency of that chosen value by 1.

2.1.3 Posterior characterization

Although NRMIs are not conjugate, there exists a posterior characterization in terms of latent variable such that, conditional on that variable, the posterior of a NRMI coincides with the posterior of another NRMI with fixed points of discontinuity. This can be a seen as *conditional conjugacy*. Let $\mathrm{x}_1^*,\ldots,\mathrm{x}_k^*$ be the k unique values displayed on the sample $\mathbf{x}^{(n)}$ from model (2.2), with corresponding frequencies $\mathrm{n}_1,\ldots,\mathrm{n}_k$ such that $\sum_{j=1}^k \mathrm{n}_j = n$.

Assume that $\tilde{\mu}(\mathbb{X})$ is absolutely continuous with respect to the Lebesgue measure so that it admits a density g. Define the positive random variable U_n as $U_n := \frac{\Gamma_n}{\tilde{\mu}(\mathbb{X})}$, where $\Gamma_n \sim \mathrm{Ga}(n,1)$ independently of $\tilde{\mu}(\mathbb{X})$. For any $n \geq 1$, the density of U_n coincides with

$$f(u) = \frac{u^{n-1}}{\Gamma(n)} \int_{\mathbb{R}_+} t^n e^{-ut} g(t) \mathrm{d}t.$$

Furthermore, the conditional distribution of U_n given $\mathbf{x}^{(n)}$ admits a density function and it is given by

$$f_{U_n \,|\, \mathbf{x}^{(n)}}(u) \propto u^{n-1} e^{-\psi(u)} \prod_{j=1}^k \tau_{\mathrm{n}_j}(u),$$

where $\psi(u)$ is the Laplace exponent and $tau_m(u)$: is as in Proposition 2.2. The random variable U_n is the one that will make the posterior distribution tractable.

Theorem 2.1. *Consider $\mathbf{x}^{(n)}$ a sample from model (2.2), with P_0 non-atomic. The posterior distribution of the unnormalized CRM $\tilde{\mu}$ given $\mathbf{x}^{(n)}$ and U_n is*

$$\tilde{\mu} \,|\, \left(\mathbf{x}^{(n)}, U_n\right) \stackrel{d}{=} \tilde{\mu}^{(U_n)} + \sum_{j=1}^k J_j^{(U_n)} \delta_{\mathrm{x}_j^*}.$$

The posterior distribution of $\tilde{p}$ given $\mathbf{x}^{(n)}$ and U_n is

$$\tilde{p}\,|\,\left(\mathbf{x}^{(n)}, U_n\right) \stackrel{d}{=} w\frac{\tilde{\mu}^{(U_n)}}{\tilde{\mu}^{(U_n)}(\mathbb{X})} + (1-w)\,\frac{\sum_{j=1}^{k} J_i^{(U_n)}\delta_{\mathrm{x}_j^*}}{\sum_{j=1}^{k} J_j^{(U_n)}},$$

where

- *$\tilde{\mu}^{(U_n)}$ is a completely random measure with intensity $\nu^{(U_n)}(\mathrm{d}v, \mathrm{d}x) = e^{-U_n s}\rho(\mathrm{d}v)\alpha(\mathrm{d}x)$.*
- *The non-negative jump random variables $\left\{J_i^{(U_n)}\right\}_{i=1}^{k}$ are independent from each other and from $\tilde{\mu}^{(U_n)}$, each with density w.r.t. the Lebesgue measure proportional to $v^{\mathrm{n}_i}e^{-U_n v}\rho(\mathrm{d}v)$*
- *$w = \frac{\tilde{\mu}^{(U_n)}(\mathbb{X})}{\tilde{\mu}^{(U_n)}(\mathbb{X})+\sum_{j=1}^{k} J_j^{(U-n)}}$.*

Given the latent variable U_n, a posteriori $\tilde{\mu}$ is again a completely random measure with fixed points of discontinuity corresponding to the locations of the unique values $\{\mathrm{x}_j^*\}_{j=1}^{k}$ among the observations. The availability of the posterior distribution makes it then possible to determine the predictive distribution, since this equals to

$$\mathbb{P}\left[\mathrm{x}_{n+1} \in \cdot\,|\,\mathbf{x}^{(n)}\right] = \mathbb{E}\left[\tilde{p}(\,\cdot\,)\,|\,\mathbf{x}^{(n)}\right] = \int_{\mathbb{R}_+} \mathbb{E}\left[\tilde{p}(\,\cdot\,)\,|\,\mathbf{x}^{(n)}, U_n = u\right] f_{U_n\,|\,\mathbf{x}^{(n)}}\mathrm{d}u.$$

Proposition 2.3. If $\mathbf{x}^{(n)}$ is sampled according to model (2.2), the predictive distribution of observation x_{n+1} given $\mathbf{x}^{(n)}$ is given by

$$\mathbb{P}\left[\mathrm{x}_{n+1} \in \cdot\,|\,\mathbf{x}^{(n)}\right] = \omega_0^{(n)}P_0(\,\cdot\,) + \frac{1}{n}\sum_{j=1}^{k}\omega_j^{(n)}\delta_{\mathrm{x}_j^*}(\,\cdot\,),$$

where

$$\omega_0^{(n)} = \frac{1}{n}\int_{\mathbb{R}_+} u\tau_1(u) f_{U_n\,|\,\mathbf{x}^{(n)}}(u)\mathrm{d}u \quad \text{and} \quad \omega_j^{(n)} = \int_{\mathbb{R}_+} u\frac{\tau_{\mathrm{n}_j+1}(u)}{\tau_{\mathrm{n}_j}(u)} f_{U_n\,|\,\mathbf{x}^{(n)}}(u)\mathrm{d}u.$$

This resulting predictive distribution takes an intuitive form: it is a linear combination of the base measure P_0 (the prior guess) and a weighted version of the empirical distribution. Note that if the EPPF of the prior process is known, then the weights $\omega_0^{(n)}$ and $\omega_j^{(n)}$ coincide with those at (2.4) and (2.5) respectively.

2.2 Examples

Now it is time to apply the theory developed above to specific instances of Lévy intensities, as by now it is evident that this is the only thing needed to calculate all the expressions described in the previous sections for particular cases of NRMIs.

2.2.1 Dirichlet process

Nonparametric priors flourished after the pioneering paper of Ferguson (1973), in which the Dirichlet Process was introduced. This is arguably one the most popular models and it appears as a special case of a number of other more general models, as we shall see further on.

Definition 2.6. Let $(Z_i)_{i=1}^n$ be a finite sequence of independent random variables such that $Z_i \sim \text{Gamma}(a_i, 1)$, where $a_i > 0$ for $i = 1, \ldots, n$ and let Y_j be

$$Y_j = \frac{Z_j}{\sum_{i=1}^n Z_i}$$

for $j = 1, \ldots, n$. The distribution of the random probability vector $Y := (Y_1, \ldots, Y_n)$ is called the *Dirichlet distribution* of parameters $(a_1, \ldots, a_n)$ and we write $Y \sim \text{Dir}(a_1, \ldots, a_n)$. Its density with respect to the $n-1$-dimensional Lebesgue measure is

$$f(y) = \frac{\Gamma(\sum_{j=1}^n a_j)}{\prod_{j=1} \Gamma(a_j)} \left(\prod_{j=1}^{n-1} y_j^{a_j - 1} \right) \left(1 - \sum_{j=1}^{n-1} y_j \right)^{a_n - 1} \mathbb{1}_{\Delta_{n-1}}(y)$$

where Δ_{n-1} is the $n-1$-dimensional simplex $\{(y_1, \ldots, y_{n-1}) : y_j \geq 0, \sum_{j=1}^{n-1} y_j \leq 1\}$.

Ferguson proved that the Dirichlet distribution satisfies the Kolmogorov consistency conditions, so that Daniel-Kolmogorov's existence theorem ensures the existence of a random process whose finite-dimensional distributions are Dirichlet-distributed.

Definition 2.7. Let α be a non-null finite measure on a Polish space with its Borel σ-algebra $(\mathbb{X}, \mathcal{X})$. A random variable $\tilde{p}$ on $\text{P}_{\mathbb{X}}$ is said to be a *Dirichlet process* if for all measurable partitions $(A_i)_{i=1}^n$ of $\mathbb{X}$

$$(\tilde{p}(A_1), \ldots, \tilde{p}(A_n)) \sim \text{Dir}\left(\alpha(A_1), \ldots, \alpha(A_n)\right)$$

We write $\tilde{p} \sim \mathfrak{D}(\alpha)$.

Aside from the specification of the finite-dimensional distributions, Ferguson provided an alternative way to construct the Dirichlet process based on the gamma CRM from example 1.3, whose Lévy intensity given by

$$\nu(\mathrm{d}v, \mathrm{d}x) = \frac{e^{-v}}{v} \mathrm{d}v \alpha(\mathrm{d}x) = \frac{e^{-v}}{v} \mathrm{d}v \theta P_0(\mathrm{d}x)$$

with $0 < \theta < \infty$ and P_0 a probability measure over $(\mathbb{X}, \mathcal{X})$. Clearly $\rho(\mathbb{R}_+) = \infty$, so $0 < \tilde{\mu}(\mathbb{X}) < \infty$ a.s. and hence we can consider the NRMI

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

If $(A_i)_{i=1}^n$ is a partition of $\mathbb{X}$, we previously proved that $\{\tilde{\mu}(A_i)\}_{i=1}^n$ are independent random variables such that $\tilde{\mu}(A_i) \sim \text{Gamma}(\theta P_0(A_i), 1)$ for $i = 1, \ldots, n$. Noting that

$$\tilde{p}(A_j) = \frac{\tilde{\mu}(A_j)}{\tilde{\mu}(\mathbb{X})} = \frac{\tilde{\mu}(A_j)}{\sum_{i=1}^n \tilde{\mu}(A_i)},$$

we can conclude that this NRMI and the Dirichlet process of parameter θP_0 are equal in distribution, since clearly

$$(\tilde{p}(A_1), \ldots, \tilde{p}(A_n)) \sim \text{Dir}\left(\theta P_0(A_1), \ldots, \theta P_0(A_n)\right).$$

As $\psi(u) = \theta \log(1+u)$, $\Psi(u) = \log(1+u)$, one has that $\mathcal{I}_\theta = \frac{1}{1+\theta}$ and for $A, B \in \mathcal{X}$

$$\begin{aligned}
\mathbb{E}\left[\tilde{p}(A)\right] &= P_0(A) \\
\text{var}\left[\tilde{p}(A)\right] &= \frac{P_0(A)\left(1 - P_0(A)\right)}{1+\theta} \\
\text{cov}(\tilde{p}(A), \tilde{p}(B)) &= \frac{P_0(A \cap B) - P_0(A)P_0(B)}{1+\theta}.
\end{aligned}$$

The parameter θ controls the variability of the realizations around the expected shape of the random distribution, P_0. This is why θ is called the *precision parameter* and its effect is depicted in Figure 2.2. The baseline measure P_0 is a standard normal distribution and its cumulative distribution function is depicted as a thick black line. Each panel contains 12 realizations with a common value of θ. Note how θ controls both the variability around the base measure and the size of the jumps.

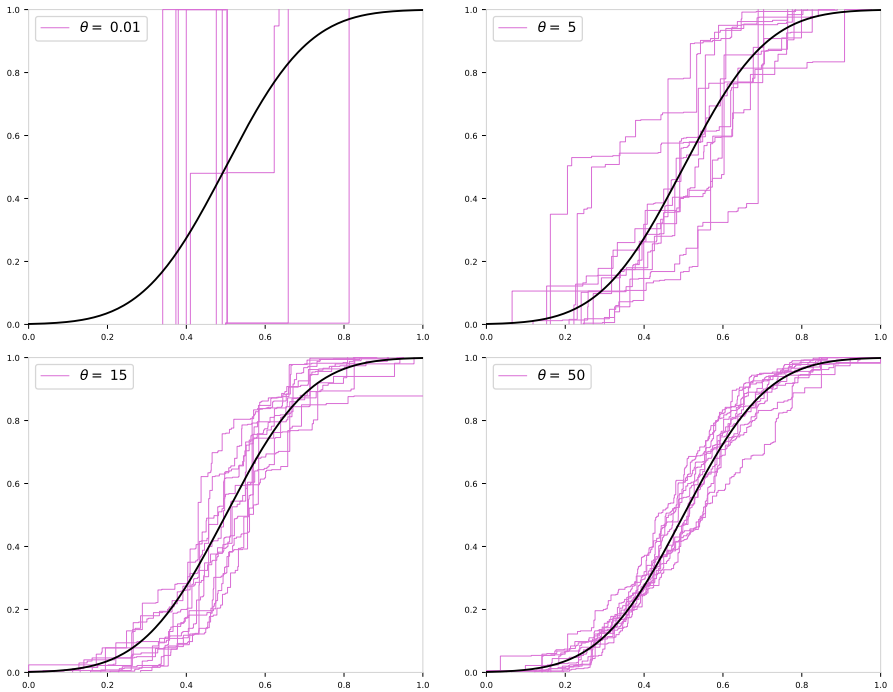


Figure 2.2: Empirical distribution functions generated from a Dirichlet process with varying θ, using the Pólya Urn sampling scheme.

Now assume that $\tilde{p} \sim \mathfrak{D}(\theta P_0)$ in (2.2) and that we have sampled $\mathbf{x}^{(n)}$. In this case the EPPF has a closed form, so we will recover the predictive distribution from it. First we need to compute $\tau_{\mathrm{n}_j}(u)$

$$\tau_{\mathrm{n}_j}(u) = \int_{\mathbb{R}_+} v^{\mathrm{n}_j} e^{-uv} e^{-v} v^{-1} \mathrm{d}v = \int_{\mathbb{R}_+} v^{\mathrm{n}_j - 1} e^{-(1+u)v} \mathrm{d}v = \frac{\Gamma(\mathrm{n}_j)}{(1+u)^{\mathrm{n}_j}}.$$

The EPPF is given by

$$\begin{aligned}
\Phi_k^{(n)}(\mathrm{n}_1, \ldots, \mathrm{n}_k) &= \frac{\theta^k}{\Gamma(n)} \int_{\mathbb{R}_+} u^{n-1} e^{-\psi(u)} \left[\prod_{j=1}^{k} \tau_{\mathrm{n}_j}(u) \right] \mathrm{d}u \\
&= \frac{\theta^k}{\Gamma(n)} \prod_{j=1}^{k} \Gamma(\mathrm{n}_j) \int_{\mathbb{R}_+} \frac{u^{n-1}}{(1+u)^{n+\theta}} \mathrm{d}u \\
&= \frac{\theta^k \Gamma(\theta)}{\Gamma(n+\theta)} \prod_{j=1}^{k} \Gamma(\mathrm{n}_j) \\
&= \frac{\theta^k}{(\theta)_n} \prod_{j=1}^{k} (\mathrm{n}_j - 1)!,
\end{aligned}$$

where $(\theta)_n = \frac{\Gamma(n+\theta)}{\Gamma(\theta)}$ is the *Pochhammer symbol*. This EPPF is known as *Ewens's sampling formula*. A straightforward computation yields

$$\begin{aligned}
\frac{\Phi_{k+1}^{(n+1)}(\mathrm{n}_1, \ldots, \mathrm{n}_k, 1)}{\Phi_k^{(n)}(\mathrm{n}_1, \ldots, \mathrm{n}_k)} &= \frac{\theta}{n+\theta} \\
\frac{\Phi_k^{(n+1)}(\mathrm{n}_1, \ldots, \mathrm{n}_j + 1, \ldots, \mathrm{n}_k)}{\Phi_k^{(n)}(\mathrm{n}_1, \ldots, \mathrm{n}_k)} &= \frac{\mathrm{n}_j}{n+\theta}.
\end{aligned}$$

It follows that

$$\mathbb{P}\left[\mathrm{x}_{n+1} \in \cdot \,|\, \mathbf{x}^{(n)} \right] = \frac{\theta}{n+\theta} P_0(\,\cdot\,) + \frac{1}{n+\theta} \sum_{j=1}^{k} \mathrm{n}_j \delta_{\mathrm{x}_j^*}(\,\cdot\,). \tag{2.6}$$

The sequence of predictive distributions described by (2.6) are known as the *Pólya urn scheme* and was studied in Blackwell and MacQueen (1973). Specifically, they observed that a Dirichlet process can be characterized by its predictive distribution, in the sense that as $n \to \infty$

$$\frac{\theta P_0(\,\cdot\,) + \sum_{i=1}^{n} \delta_{\mathrm{x}_i}(\,\cdot\,)}{\theta + n} \overset{\text{a.s.}}{\to} \mathrm{P} \quad \text{where } \mathrm{P} \sim \mathfrak{D}(\theta P_0).$$

This provides an easy sampling scheme, as described in Algorithm 2.1.

Algorithm 2.1: Sample $\mathbf{x}^{(n)}$ from a Dirichlet process.

Sample $\mathrm{x}_1 \sim P_0$.
for $i = 2, \ldots, n$ **do**
  Sample x_i according to

$$\mathrm{x}_i = \begin{cases} \mathrm{x}_j^* \text{ with probability } \frac{\mathrm{n}_j}{\theta+i-1}, \ j = 1, \ldots, k \\ \mathrm{x}_{k+1}^* \sim P_0 \text{ with probability } \frac{\theta}{\theta+i-1} \end{cases}$$

Now lets look at the number of blocks of the partition generated by the sample $\mathbf{x}^{(n)}$, K_n (which we've called above k, for the sake of simplicity). By summing over all possible partitions of $[n]$ of size k, for $k = 1, \ldots, n$, we can compute

$$\mathbb{P}[\mathrm{K}_n = k] = \sum_{\pi \in \mathcal{P}_{[n]}^k} \mathbb{P}[\Pi_n = \pi] \frac{\theta^k}{(\theta)_n} = \sum_{(\mathrm{n}_1, \ldots, \mathrm{n}_k)} \prod_{j=1}^{k} (\mathrm{n}_j - 1)! = \frac{\theta^k}{(\theta)_n} |\mathfrak{s}_{n,k}|,$$

where $|\mathfrak{s}_{n,k}| := B_{n,k}(x)$ with $x = (x_1, x_2, \ldots)$ and $x_i = (i-1)!$ is the *unsigned Stirling number of the first kind*. Alternatively, K_n can be expressed as a sum of indicator functions

$$\mathrm{K}_n = \sum_{j=1}^{n} \mathbb{1}_{A_j},$$

where A_j is the event that observation x_j falls into a new group. From the predictive distribution we see that $\mathbb{1}_{A_j} \sim \text{Bernoulli}\left(\frac{\theta}{\theta+j-1}\right)$, so that the expected number of blocks displayed on the first n observations is

$$\mathbb{E}[\mathrm{K}_n] = \sum_{j=1}^{n} \mathbb{E}[\mathbb{1}_{A_j}] = \sum_{j=1}^{n} \frac{\theta}{\theta + j - 1}.$$

Korwar and Hollander (1973) proved that the expected number of blocks grows logarithmically in the number of observations n, that is

$$\frac{\mathbb{E}[\mathrm{K}_n]}{\log(n)} \to \theta \quad \text{a.s. when } n \to \infty.$$

This slow growth makes sense because the larger n_j is, the higher the probability that the block x_j^* will grow. We would expect then to see large groups, which necessarily leads to the number of these being much smaller than n. θ directly controls the number of groups, as shown in Figure 2.3. Larger values of θ imply a larger number of groups a priori, as the distribution of K_n places the mode at higher values of k as θ grows.

The distribution over partitions induced by the Dirichlet process is often called the *Chinese restaurant process* due to the following metaphor. Imagine an initially empty Chinese restaurant with an unlimited number of tables, each with unlimited space for customers to seat. The first customer seats at table 1. The second customer decides either to sit with the first customer, or by herself at a new table and, in general, the $n+1$-th

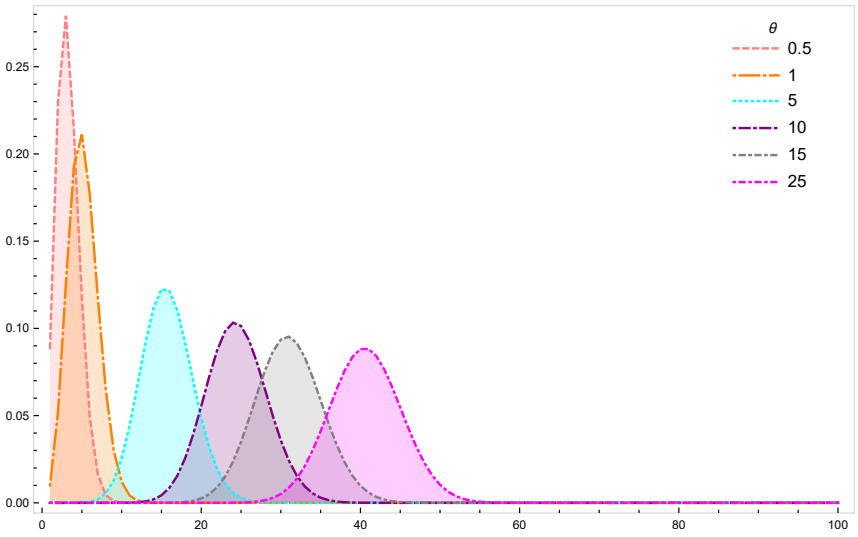


Figure 2.3: Distribution of K_{100} *for the Dirichlet process, with varying values of* θ*.*

customer either sits at an already occupied table j with probability $\frac{\mathrm{n}_j}{\theta+n}$, or sits at a new table with probability $\frac{\theta}{\mathrm{n}_j+n}$. Identifying customers with integers, once n customers have sat down the tables define a partition over $[n]$. An example is depicted in Figure 2.4. The white circles are tables and the blue dots represent customers seating at that table. If a 13-th customer were to arrive, she would choose a new table with probability $\frac{\theta}{11+\theta}$. If not, the occupied table that is more likely to be chosen is the first one.

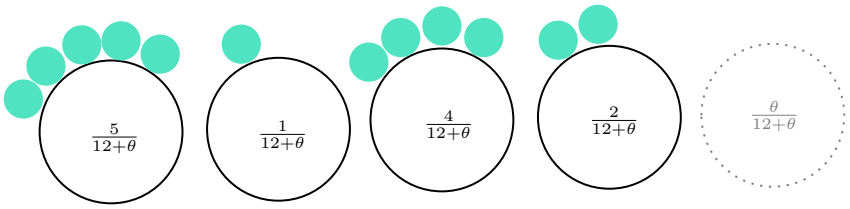


Figure 2.4: A possible seating arrangement after 12 *customers enter the restaurant in the CRP metaphor.*

2.2.2 Normalized σ-stable process

Recall the σ-stable CRM from example 1.4, whose Lévy intensity is

$$\nu(\mathrm{d}v, \mathrm{d}x) = \frac{\sigma v^{-1-\sigma}}{\Gamma(1-\sigma)} \mathrm{d}v \theta P_0(\mathrm{d}x), \tag{2.7}$$

for $\sigma \in (0,1)$ and $0 < \theta < \infty$. As $\rho(\mathbb{R}_+) = \infty$, we can normalize this CRM to obtain the *normalized σ-stable process*. This random probability measure was first studied in Kingman. (1975). We previously showed that $\psi(u) = \theta u^\sigma$, so $\Psi(u) = u^\sigma$, $\mathcal{I}_\theta = 1 - \sigma$ and for measurable sets A, B

$$\begin{aligned}
\mathbb{E}\left[\tilde{p}(A)\right] &= P_0(A) \\
\operatorname{var}\left[\tilde{p}(A)\right] &= P_0(A)\left(1 - P_0(A)\right)(1-\sigma) \\
\operatorname{cov}(\tilde{p}(A), \tilde{p}(B)) &= \left(P_0(A \cap B) - P_0(A)P_0(B)\right)(1-\sigma).
\end{aligned}$$

To compute the EPPF, note that $\tau_{\mathrm{n}_j}(u)$ equals

$$\begin{aligned}
\tau_{\mathrm{n}_j}(u) &= \int_{\mathbb{R}_+} v^{\mathrm{n}_j} e^{-uv} \frac{\sigma v^{-1-\sigma}}{\Gamma(1-\sigma)} \mathrm{d}v \\
&= \frac{\sigma}{\Gamma(1-\sigma)} \int_{\mathbb{R}_+} v^{\mathrm{n}_j - \sigma - 1} e^{-uv} \mathrm{d}v \\
&= \frac{\sigma \Gamma(\mathrm{n}_j - \sigma)}{\Gamma(1-\sigma) u^{\mathrm{n}_j - \sigma}}.
\end{aligned}$$

The EPPF is thus given by

$$\begin{aligned}
\Phi_k^{(n)}(\mathrm{n}_1, \ldots, \mathrm{n}_k) &= \frac{\theta^k}{\Gamma(n)} \int_{\mathbb{R}_+} u^{n-1} e^{-\psi(u)} \left[\prod_{j=1}^k \tau_{\mathrm{n}_j}(u)\right] \mathrm{d}u \\
&= \frac{\sigma^k \theta^k}{\Gamma(n)\Gamma(1-\sigma)} \prod_{j=1}^k \Gamma(\mathrm{n}_j - \sigma) \int_{\mathbb{R}_+} u^{\sigma k - 1} e^{-\theta u^\sigma} \mathrm{d}u \\
&= \frac{\sigma^k \theta^k}{\Gamma(n)\Gamma(1-\sigma)} \prod_{j=1}^k \Gamma(\mathrm{n}_j - \sigma) \frac{\Gamma(k)}{\sigma \theta^k} \\
&= \frac{\sigma^{k-1}\Gamma(k)}{\Gamma(n)} \prod_{j=1}^k (1-\sigma)_{\mathrm{n}_j - 1}.
\end{aligned}$$

A simple calculation leads to

$$\begin{aligned}
\frac{\Phi_{k+1}^{(n+1)}(\mathrm{n}_1, \ldots, \mathrm{n}_k, 1)}{\Phi_k^{(n)}(\mathrm{n}_1, \ldots, \mathrm{n}_k)} &= \frac{k\sigma}{n} \\
\frac{\Phi_k^{(n+1)}(\mathrm{n}_1, \ldots, \mathrm{n}_j + 1, \ldots, \mathrm{n}_k)}{\Phi_k^{(n)}(\mathrm{n}_1, \ldots, \mathrm{n}_k)} &= \frac{\mathrm{n}_j - \sigma}{n}.
\end{aligned}$$

Hence

$$\mathbb{P}\left[\mathrm{x}_{n+1} \in \cdot \,|\, \mathbf{x}^{(n)}\right] = \frac{k\sigma}{n} P_0(\,\cdot\,) + \frac{1}{n} \sum_{j=1}^k (\mathrm{n}_j - \sigma)\, \delta_{\mathrm{x}_j^*}(\,\cdot\,).$$

Analogously to the Dirichlet process, this sequence of predictive distributions can be used to generate a sampling scheme as follows.

Algorithm 2.2: Sample $\mathbf{x}^{(n)}$ from a σ-stable process.

Sample $\mathrm{x}_1 \sim P_0$.
for $i = 2, \ldots, n$ **do**
  Sample x_i according to

$$\mathrm{x}_i = \begin{cases} \mathrm{x}_j^* \text{ with probability } \frac{\mathrm{n}_j - \sigma}{i-1}, \ j = 1, \ldots, k \\ \mathrm{x}_{k+1}^* \sim P_0 \text{ with probability } \frac{k\sigma}{i-1} \end{cases}$$

The distribution of the number of groups equals to

$$\mathbb{P}[\mathrm{K}_n = k] = \frac{\sigma^{k-1}\Gamma(k)}{\Gamma(n)} \sum_{\pi \in \mathcal{P}_{[n]}^k} \prod_{j=1}^{k} (1-\sigma)_{\mathrm{n}_j - 1} = \frac{\sigma^{k-1}\Gamma(k)}{\Gamma(n)} S_{n,k}^{\sigma},$$

where $S_{n,k}^{\alpha,\beta}$ is the generalized Stirling number, defined as

$$S_{n,k}^{\alpha} := B_{n,k}(x),$$

with $x = (x_1, x_2, \ldots)$ and $x_j = (\beta - \alpha)_{(j-1)}$ for $\alpha, \beta \in \mathbb{R}$ and we've adopted the notation $S_{n,k}^{\alpha} = S_{n,k}^{\alpha,1}$. In this case, σ controls the flatness of the distribution of K_n instead of the location of the mode, as shown in Figure 2.5. Larger values of σ translate into more platykurtic distributions, meaning that the larger is σ, the flatter is the curve.

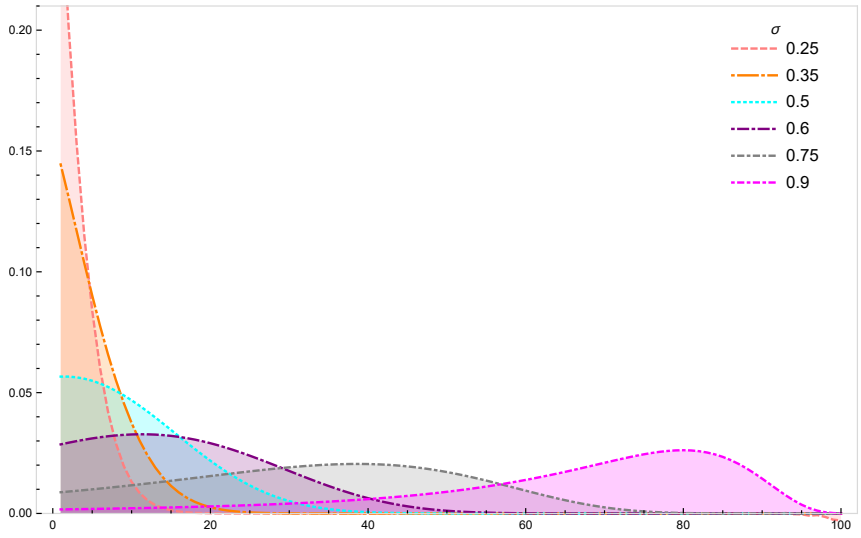


Figure 2.5: Distribution of K_{100} *for the normalized stable process, with varying values of* σ.

Finally, the expected number of distinct observations a priori is

$$\begin{aligned}\mathbb{E}[\mathrm{K}_n] &= \sum_{k=1}^{n} k\mathbb{P}[\mathrm{K}_n = k] \\ &= \sum_{k=1}^{n} k\frac{\sigma^{k-1}\Gamma(k)}{\Gamma(n)} S_{n,k}^{\sigma} \\ &= \frac{1}{\Gamma(n)\sigma}\sum_{k=1}^{n} S_{n,k}^{\sigma}\prod_{i=1}^{k}(\sigma + i\sigma),\end{aligned}$$

where we've written $k\sigma^{k-1}\Gamma(k) = \sigma^{k-1}k! = \frac{1}{\sigma}\prod_{i=1}^{k} i\sigma = \frac{1}{\sigma}\prod_{i=0}^{k-1}(\sigma + i\sigma)$. Using the following identity from Charalambides and Singh (1988) for the generalized Stirling numbers

$$(x)_n = \sum_{k=1}^{n} S_{n,k}^{\alpha}\prod_{i=0}^{k-1}(x + i\alpha) \tag{2.8}$$

leads to

$$\mathbb{E}[\mathrm{K}_n] = \frac{(\sigma)_n}{\sigma\Gamma(n)} = \frac{(\sigma+1)_{n-1}}{\Gamma(n)}.$$

Note that the total mass θ is not involved in either the EPPF or distribution or K_n, meaning that we could assume that $\theta = 1$.

2.3 Pitman-Yor process

Now it is time to cheat a little bit. The *Pitman-Yor process*, also known as the *two parameter Poisson-Dirichlet process*, introduced in Pitman and Yor (1997), cannot be constructed through the normalization of a CRM but it is derivable from a stable subordinator by a change of measure. This process is part of a class of models called the *Poisson-Kingman models*.

For $\sigma \in (0, 1)$, let $\tilde{\mu}_\sigma$ be a σ-stable CRM, with ρ_σ the jump part of the Lévy intensity and let T_σ be the total mass of $\tilde{\mu}_\sigma$, that is

$$T_\sigma := \tilde{\mu}_\sigma(\mathbb{X}) = \sum_{i\geq 1} \mathrm{w}_i\delta_{\zeta_i}(\mathbb{X}) = \sum_{i\geq 1}\mathrm{w}_i.$$

Up to this point we know that $0 < T_\sigma < \infty$ a.s. and that T_σ follows a positive stable distribution with density f_σ with respect to the Lebesgue measure.

A σ-stable Poisson-Kingman model is a generalization of the stable NRMI, obtained by suitably tilting (deforming) the distribution of the total mass T_σ. Let $\left(\mathrm{w}_{(i)}\right)_{i\geq 1}$ be the ranked jumps of $\tilde{\mu}_\sigma$, i.e. decreasing rearrangement of the jumps $\mathrm{w}_{(1)} \geq \mathrm{w}_{(2)} \geq \cdots$ and define the normalized jumps P_i as

$$P_i := \frac{\mathrm{w}_{(i)}}{T_\sigma} \quad \text{for } i \geq 1.$$

Definition 2.8. Consider the σ-stable CRM and its ordered normalized jumps just as above and let γ be a probability distribution over $\mathbb{R}_+$. Denote by $\mathrm{PK}(\rho_\sigma \,|\, t)$ the regular conditional distribution of the normalized ordered jumps $\left(P_{(i)}\right)_{i\geq 1}$ given $T_\sigma = t$. The distribution

$$\int_{\mathbb{R}_+} \mathrm{PK}(\rho_\sigma \,|\, t)\gamma(\mathrm{d}t)$$

over the simplex $\{(P_1, P_2, \ldots) : \ P_1 \geq P_2 \geq \cdots \geq 0 \text{ and } \sum_{i\geq 1} P_i = 1\}$ is named a *σ-stable Poisson-Kingman distribution* with mixing distribution γ.

A *σ-stable Poisson-Kingman model* with parameter γ is the almost surely discrete random probability measure $\tilde{p}_{\sigma,\gamma}$ given by

$$\tilde{p}_{\sigma,\gamma}(\,\cdot\,) = \sum_{i=1}^{\infty} P_i \delta_{\mathrm{x}_i}(\,\cdot\,),$$

where P_i follow a σ-stable Poisson-Kingman distribution. We will write $\tilde{p}_{\sigma,\gamma} \sim \mathrm{PK}(\rho_\sigma, \gamma)$.

Intuitively, this class of random probability distributions are σ-stable NRMIs conditional on the total mass T_σ and mixed with respect to some distribution γ on the positive real line. Although we will only be using the σ-stable Poisson-Kingman model, in general Poisson-Kingman models are defined just as above by substituting the σ-stable CRM with any homogeneous CRM. This can be consulted in Pitman (2003).

If we take $\gamma(\mathrm{d}t) = h(t)f_\sigma(t)\mathrm{d}t$ with $h : \mathbb{R}_+ \cup \{0\} \to \mathbb{R}_+$ a nonnegative measurable function, according to the definition of $\tilde{p}_{\sigma,\gamma}$ we can write

$$\tilde{p}_{\sigma,\gamma}(\,\cdot\,) = \frac{\tilde{\mu}_{\sigma,\gamma}(\,\cdot\,)}{T_{\sigma,\gamma}},$$

where $\tilde{\mu}_{\sigma,\gamma}(\,\cdot\,)$ is an a.s. discrete random measure on $(\mathrm{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$ with distribution $P_{\sigma,\gamma}$ and $T_{\sigma,\gamma} := \tilde{\mu}_{\sigma,\gamma}(\mathbb{X})$ is its total mass. In particular $T_{\sigma,\gamma}$ has density w.r.t. the Lebesgue measure equal to $f_{T_{\sigma,h}} = h(t)f_\sigma(t)$ and furthermore, $\tilde{\mu}_{\sigma,\gamma}$ is absolutely continuous with respect to the distribution P_σ of $\tilde{\mu}_\sigma$ and it satisfies

$$\frac{\mathrm{d}P_{\sigma,\gamma}(m)}{\mathrm{d}P_\sigma} = h(m(\mathbb{X})) \quad \text{ for } m \in \mathrm{M}_{\mathbb{X}}.$$

The aforementioned normalized σ-stable process can be recovered by choosing γ as the distribution of T, that is by setting $h(t) = 1$.

Pitman (2003) provides an expression for the EPPF of a σ-stable Poisson Kingman model conditioned on its total mass, stated in the next proposition.

Proposition 2.4. Let $\mathbf{x}^{(n)}$ be a sample from an exchangeable sequence such that

$$\mathrm{x}_i \,|\, \tilde{p} \sim \tilde{p}, \quad \tilde{p} \sim \mathrm{PK}(\rho_\sigma, \gamma).$$

The EPPF of the partition induced by $\mathbf{x}^{(n)}$, conditioned on the total mass T_σ, is given by

$$\Phi_k^{(n)}(\mathrm{n}_1,\ldots,\mathrm{n}_k \,|\, T_\sigma = t) = \frac{\sigma^k t^{-n}}{\Gamma(n-k\sigma) f_\sigma(t)} \int_0^t s^{n-k\sigma-1} f_\sigma(t-s)\mathrm{d}s \prod_{j=1}^k (1-\sigma)_{\mathrm{n}_j-1}\,.$$

Marginalizing over t yields the EPPF of the σ-stable PK model

$$\Phi_k^{(n)}(\mathrm{n}_1,\ldots,\mathrm{n}_k) = \frac{\sigma^k}{\Gamma(n-k\sigma)} \int_{\mathbb{R}_+} \frac{t^{-n}}{f_\sigma(t)} \int_0^t s^{n-k\sigma-1} f_\sigma(t-s)\mathrm{d}s\gamma(\mathrm{d}t) \prod_{j=1}^k (1-\sigma)_{\mathrm{n}_j-1}\,.$$

If the mixing distribution takes the form $\gamma(\mathrm{d}t) = h(t)f_\sigma(t)\mathrm{d}t$, then

$$\begin{aligned}
\int_{\mathbb{R}_+} \frac{t^{-n}}{f_\sigma(t)} \int_0^t s^{n-k\sigma-1} f_\sigma(t-s)\mathrm{d}s\gamma(\mathrm{d}t) &= \int_{\mathbb{R}_+}\int_0^t t^{-n} s^{n-k\sigma-1} f_\sigma(t-s)h(t)\mathrm{d}s\mathrm{d}t \\
&= \int_{\mathbb{R}_+}\int_s^\infty t^{-n} s^{n-k\sigma-1} f_\sigma(t-s)h(t)\mathrm{d}t\mathrm{d}s \\
&\quad \text{(by making } u = t-s \text{)} \\
&= \int_{\mathbb{R}_+}\int_{\mathbb{R}_+} (u+s)^{-n} s^{n-k\sigma-1} f_\sigma(u)h(u+s)\mathrm{d}s\mathrm{d}u.
\end{aligned}$$

For any $\sigma \in (0,1)$ and $\theta > -\sigma$, the *Pitman-Yor process* is a σ-stable Poisson-Kingman model with γ of the form

$$\gamma(\mathrm{d}t) = \frac{\Gamma(\theta+1)}{\Gamma\left(\frac{\theta}{\sigma}+1\right)} t^{-\theta} f_\sigma(t)\mathrm{d}t.$$

This means that

$$\begin{aligned}
\int_{\mathbb{R}_+} \frac{t^{-n}}{f_\sigma(t)} \int_0^t s^{n-k\sigma-1} f_\sigma(t-s)\mathrm{d}s\gamma(\mathrm{d}t) &\propto \int_{\mathbb{R}_+}\int_0^\infty (u+s)^{-n-\theta} s^{n-k\sigma-1} f_\sigma(u)\mathrm{d}s\mathrm{d}u \\
&= \int_{\mathbb{R}_+} f_\sigma(u)\left[\int_0^\infty (u+s)^{-n-\theta} s^{n-k\sigma-1}\mathrm{d}s\right]\mathrm{d}u \\
&= \frac{\Gamma(n-k\sigma)\Gamma(\theta+k\sigma)}{\Gamma(n+\theta)} \int_{\mathbb{R}_+} f_\sigma(u) u^{-\theta-k\sigma}\mathrm{d}u \\
&= \frac{\Gamma(n-k\sigma)\Gamma(\theta+k\sigma)}{\Gamma(n+\theta)} \frac{\Gamma\left(\frac{\theta}{\sigma}+k+1\right)}{\Gamma(\theta+k\sigma+1)},
\end{aligned}$$

where we've used that the r-th moment of a σ-stable random variable exists and equals $\frac{\Gamma(1-r/\sigma)}{\Gamma(1-r)}$ for $-\infty < r < \sigma$. Therefore the EPPF is

$$\begin{aligned}
\Phi_k^{(n)}(\mathrm{n}_1,\ldots,\mathrm{n}_k) &= \frac{\sigma^k}{\Gamma(n-k\sigma)}\int_{\mathbb{R}_+}\frac{t^{-n}}{f_\sigma(t)}\int_0^t s^{n-k\sigma-1}f_\sigma(t-s)\mathrm{d}s\gamma(\mathrm{d}t)\prod_{j=1}^k(1-\sigma)_{\mathrm{n}_j-1}\\
&= \sigma^k\frac{\Gamma(\theta+1)}{\Gamma(\theta+n)}\frac{\Gamma(\theta+k\sigma)}{\Gamma(\theta+k\sigma+1)}\frac{\Gamma\left(\frac{\theta}{\sigma}+k+1\right)}{\Gamma\left(\frac{\theta}{\sigma}+1\right)}\prod_{j=1}^k(1-\sigma)_{\mathrm{n}_j i-1}\\
&= \frac{\sigma^k}{(\theta+1)_{n-1}}\frac{1}{(\theta+k\sigma)}\prod_{i=1}^k\left(\frac{\theta}{\sigma}+i\right)\prod_{j=1}^k(1-\sigma)_{\mathrm{n}_j-1}\\
&= \frac{\prod_{i=1}^{k-1}(\theta+i\sigma)}{(\theta+1)_{n-1}}\prod_{j=1}^k(1-\sigma)_{\mathrm{n}_j-1}\,.
\end{aligned}$$

We will write $\mathcal{PY}(\sigma,\theta)$ to refer to the distribution of a Pitman-Yor process with parameters σ,θ. The evident similarities between this EPPF and the ones of the two previously studied processes is no coincidence, as the Pitman-Yor process encompasses as particular cases both the Dirichlet process and the normalized σ-stable process.

- The normalized σ-stable process is recovered by setting $\theta = 0$.
- When $\sigma \downarrow 0$ we recover the Dirichlet process.

Now computing the weights leads to the predictive distribution

$$\begin{aligned}
\frac{\Phi_{k+1}^{(n+1)}(\mathrm{n}_1,\ldots,\mathrm{n}_k,1)}{\Phi_k^{(n)}(\mathrm{n}_1,\ldots,\mathrm{n}_k)} &= \frac{\theta+k\sigma}{n+\theta}\\
\frac{\Phi_k^{(n+1)}(\mathrm{n}_1,\ldots,\mathrm{n}_j+1,\ldots,\mathrm{n}_k)}{\Phi_k^{(n)}(\mathrm{n}_1,\ldots,\mathrm{n}_k)} &= \frac{\mathrm{n}_j-\sigma}{n+\theta}\\
\mathbb{P}\left[\mathrm{x}_{n+1}\in\cdot\,|\,\mathbf{x}^{(n)}\right] &= \frac{\theta+k\sigma}{n+\theta}P_0(\,\cdot\,)+\frac{1}{n+\theta}\sum_{j=1}^k(\mathrm{n}_j-\sigma)\,\delta_{\mathrm{x}_j^*}(\,\cdot\,).
\end{aligned}$$

Here the probability of obtaining new values is monotonically increasing in k and the value of σ can be used to control strength of the dependence on k. If a new value enters the sample at stage $n+1$ (with frequency 1), then it gets assigned in the empirical part of the predictive distribution a mass proportional to $1-\sigma$ (less than its cluster size, unlike in the Dirichlet process) and consequently, a mass proportional to σ is added to the probability of generating a new value. That is, if $\mathrm{x}_{n+1} = \mathrm{x}_{k+1}^*$ then at stage $n+2$ we would have

$$\begin{aligned}
\mathbb{P}\left[\mathrm{x}_{n+2}\in\cdot\,|\,\mathbf{x}^{(n+1)}\right] &= \left[\frac{\theta+k\sigma}{n+1+\theta}+\frac{\sigma}{n+1+\theta}\right]P_0(\,\cdot\,)+\sum_{j=1}^k\frac{(\mathrm{n}_j-\sigma)}{n+1+\theta}\delta_{\mathrm{x}_j^*}(\,\cdot\,)\\
&\quad+\frac{1-\sigma}{n+1+\theta}\delta_{\mathrm{x}_{k+1}^*}(\,\cdot\,).
\end{aligned}$$

This means that if x_{n+1} is a new value, the probability of generating another new value increases by $\frac{\sigma}{n+\theta+1}$ even though it simultaneously decreases as a function of the sample size. However, once a new value has been generated, subsequent re-observations increase its mass by a factor proportional to 1. The tractability of the predictive distribution allows a simple sampling scheme as described in the next algorithm.

Algorithm 2.3: Sample $\mathbf{x}^{(n)}$ from $\mathcal{PY}(\sigma, \theta)$.

Sample $\mathrm{x}_1 \sim P_0$.
for $i = 2, \ldots, n$ **do**
  Sample x_i according to

$$\mathrm{x}_i = \begin{cases} \mathrm{x}_j^* \text{ with probability } \frac{\mathrm{n}_j - \sigma}{\theta + i - 1}, \ j = 1, \ldots, k \\ \mathrm{x}_{k+1}^* \sim P_0 \text{ with probability } \frac{\theta + k\sigma}{\theta + i - 1} \end{cases}$$

Just as the Dirichlet process has the Chinese restaurant process as a metaphor for the induced distribution over partitions, the Pitman-Yor process has a similar metaphor which is called the (σ, θ)-*seating plan*, as described in Pitman (2006): given σ, θ such that $0 < \sigma < 1$ and $\theta > -\sigma$, again picture a Chinese restaurant with infinite capacity for both customers and tables. If up to a certain point n customers have sat down across k different tables, with n_j customers at the j-th table, the next person that comes in either sits at an occupied table with probability $\frac{\mathrm{n}_j - \sigma}{n+\theta}$ or sits at a new table with probability $\frac{\theta + k\sigma}{n+\theta}$. An example is presented at Figure 2.6. Suppose that $n = 12$ customers have arrived and sat down across $k = 4$ tables; if a 13-th customer were to arrive, he or she would choose a new table with probability $\frac{\theta+4\sigma}{12+\theta}$ or sit at an already occupied table with probability as marked inside each circle.

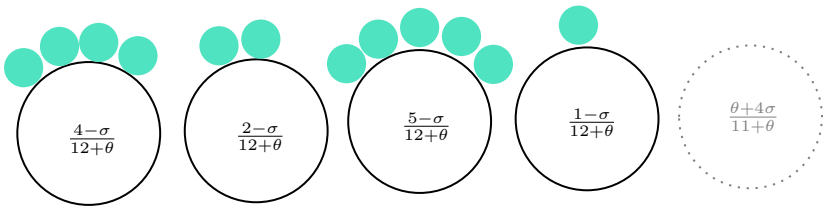


Figure 2.6: A possible seating arrangement after 12 *customers enter the restaurant in the* (σ, θ)*-seating plan.*

The distribution of the number of groups can be computed as follows.

$$\begin{aligned} \mathbb{P}[\mathrm{K}_n = k] &= \sum_{\pi \in \mathcal{P}_{[n]}^k} \frac{\prod_{i=1}^{k-1} (\theta + i\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{\mathrm{n}_j - 1} \\ &= \frac{\prod_{i=1}^{k-1} (\theta + i\sigma)}{(\theta+1)_{n-1}} \sum_{\pi \in \mathcal{P}_{[n]}^k} \prod_{j=1}^{k} (1-\sigma)_{\mathrm{n}_j - 1} \\ &= \frac{\prod_{i=1}^{k-1} (\theta + i\sigma)}{(\theta+1)_{n-1}} S_{n,k}^{\sigma}. \end{aligned}$$

Figure 2.7 shows the distribution of K_{100}. From here it is evident that the Pitman-Yor process allows more flexibility than the Dirichlet or Stable process, as the fine-tuning of the parameters (σ, θ) can yield a wide variety of shapes and placements of the mode of the a priori distribution of K_n.

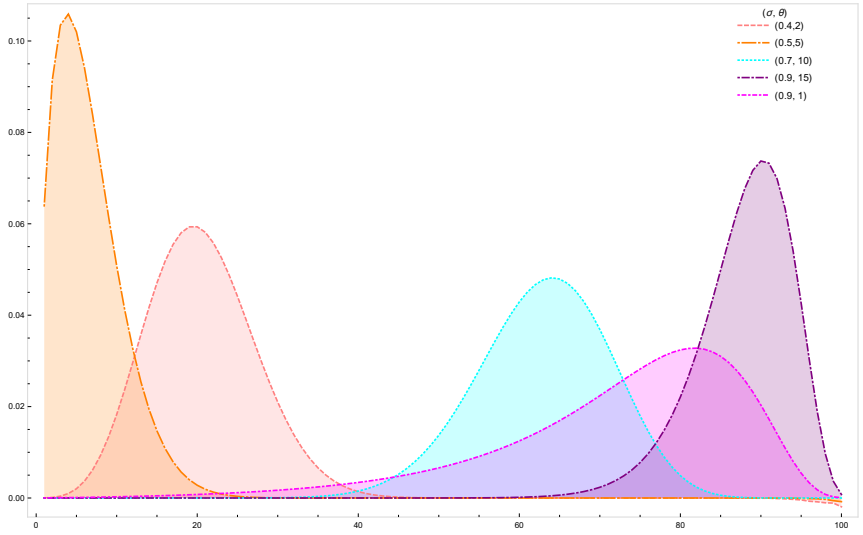


Figure 2.7: Distribution of K_{100} *for the Pitman-Yor process, with varying values of* (σ, θ).

The expected number of clusters a priori is

$$\begin{aligned}
\mathbb{E}[\mathrm{K}_n] &= \sum_{k=1}^{n} k\mathbb{P}[\mathrm{K}_n = k] = \sum_{k=1}^{n} kS_{n,k}^{\sigma} \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \\
&= \sum_{k=1}^{n} S_{n,k}^{\sigma} \frac{k\sigma \prod_{i=1}^{k-1}(\theta + i\sigma)}{\sigma(\theta+1)_{n-1}} \\
&= \sum_{k=1}^{n} S_{n,k}^{\sigma} \frac{(\theta + k\sigma)\prod_{i=1}^{k-1}(\theta + i\sigma)}{\sigma(\theta+1)_{n-1}} - \sum_{k=1}^{n} S_{n,k}^{\sigma} \frac{\theta \prod_{i=1}^{k-1}(\theta + i\sigma)}{\sigma(\theta+1)_{n-1}} \\
&= \sum_{k=1}^{n} S_{n,k}^{\sigma} \frac{\prod_{i=0}^{k-1}(\theta + \sigma + i\sigma)}{\sigma(\theta+1)_{n-1}} - \sum_{k=1}^{n} S_{n,k}^{\sigma} \frac{\prod_{i=0}^{k-1}(\theta + i\sigma)}{\sigma(\theta+1)_{(n-1)}},
\end{aligned}$$

where we have written $\prod_{i=1}^{k}(\theta + i\sigma) = \prod_{i=0}^{k-1}(\theta + \sigma + i\sigma)$. The identity presented in equation (2.8) allows us to conclude

$$\mathbb{E}[\mathrm{K}_n] = \frac{(\theta+\sigma)_n}{\sigma(\theta+1)_{(n-1)}} - \frac{(\theta)_n}{\sigma(\theta+1)_{(n-1)}} = \frac{(\theta+\sigma)_n}{\sigma(\theta+1)_{(n-1)}} - \frac{\theta}{\sigma}.$$

Regarding to its asymptotic behavior, Pitman (2003) proved that

$$\frac{\mathrm{K}_n}{n^{\sigma}} \to Y_{\frac{\theta}{\sigma}} \text{ a.s.},$$

where Y_q has a density given by $f(y) = \frac{\Gamma(q\sigma+1)}{\sigma\Gamma(q+1)} y^{q-1-\frac{1}{\sigma}} f_{\sigma}\left(y^{-\frac{1}{\sigma}}\right)$ for $q \geq 0$. This means that the number of blocks under a Pitman-Yor process increases at a higher rate than in the Dirichlet process.

Hierarchical processes

In a large variety of applied problems where data is generated from different although related studies, exchangeability is not an appropriate assumption as usually some degree of heterogeneity is present. For instance, consider an experiment in which n subjects are receiving a medical treatment and the sex of each participant is known. If we were to assume exchangeability, this would translate into assuming that the covariate male/female does not matter at all. Another example would be to consider multicenter trial studies, where subjects in different centers undergo different treatments. More often than not, these types of covariates are judged as potentially meaningful, so in situations in which we wish to acknowledge differences between groups of observations, a more complex form of symmetry is needed. In Section 3.1 we explain the concept of partial exchangeability, derive the corresponding representation theorem for partially exchangeable arrays and briefly describe some constructions for nonparametric priors for this setting. Section 3.2 is latter dedicated to the construction of such priors based on hierarchical structures of NRMIs and the study of their basic properties. In Section 3.3 we present examples of hierarchical processes based on a particular choice of the underlying CRM, and in Section 3.4 we extend the results for hierarchical NRMIs to hierarchies of Pitman-Yor processes. Most of the results displayed in this Chapter are based on Camerlenghi (2015), Camerlenghi et al. (2019b) and Bassetti et al. (2020).

3.1 Partial exchangeability

Whenever the data generating mechanism is such that there exist homogeneity within each experiment and heterogeneity across different experiments, e.g. when observations come in distinct groups with those in the same group being more similar than across groups, we find ourselves with partial exchangeability. This type of dependence was introduced de Finetti (1937).

Definition 3.1. Let $\{(\mathrm{x}_{i,j})_{j\geq 1} : i = 1, \ldots, m\}$ be m sequences of random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values on a Polish space $\mathbb{X}$ endowed with its Borel σ-algebra $\mathcal{X}$. Let $\mathbf{x}^{(i)} = (\mathrm{x}_{i,j})_{j\geq 1}$. The array of $\mathbb{X}$-valued random elements $\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\right)$ is *partially exchangeable* if for every $n_i \geq 1$ and for all permutations ρ_i of

$\{1, \ldots, n_i\}$, where $i = 1, \ldots, m$

$$\left((\mathrm{x}_{1,j})_{j=1}^{n_1}, (\mathrm{x}_{2,j})_{j=1}^{n_2}, \ldots, (\mathrm{x}_{m,j})_{j=1}^{n_m}\right) \overset{d}{=} \left((\mathrm{x}_{1,\rho_1(j)})_{j=1}^{n_1}, (\mathrm{x}_{2,\rho_2(j)})_{j=1}^{n_2}, \ldots, (\mathrm{x}_{m,\rho_m(j)})_{j=1}^{n_m}\right)$$

This means that exchangeability holds true within each of the m separate groups but not across them, as depicted in Figure 3.1.

$$(\mathrm{x}_{1,1}, \mathrm{x}_{1,2}, \mathrm{x}_{1,3}, \mathrm{x}_{2,1}, \mathrm{x}_{2,2}) \overset{d}{=} (\mathrm{x}_{1,2}, \mathrm{x}_{1,3}, \mathrm{x}_{1,1}, \mathrm{x}_{2,2}, \mathrm{x}_{2,1})$$

but

$$(\mathrm{x}_{1,1}, \mathrm{x}_{1,2}, \mathrm{x}_{1,3}, \mathrm{x}_{2,1}, \mathrm{x}_{2,2}) \overset{d}{\neq} (\mathrm{x}_{1,1}, \mathrm{x}_{2,1}, \mathrm{x}_{1,2}, \mathrm{x}_{2,2}, \mathrm{x}_{1,3})$$

Figure 3.1

A reasonable symmetry assumption for the first example of the medical treatment would be partial exchangeability, that is to consider all females as exchangeable with one another but not with males. On multicenter trials, partial exchangeability just entails that those that undergo treatments on the same facility are

Partial exchangeability was extensively studied in de Finetti (1937) and later in Diaconis and Freedman (1978) and Aldous (1981), and an analogue of Bruno de Finetti's representation theorem for partially exchangeable arrays is valid. Let us denote a sample of size $n_i \geq 1$ of $\mathbf{x}^{(i)}$ in a partially exchangeable array as

$$\mathbf{x}_i^{(n_i)} = (\mathrm{x}_{i,j})_{j=1}^{n_i}$$

where $n_i \geq 1$ for $i = 1, \ldots, m$.

Theorem 3.1. *Let $\{(\mathrm{x}_{i,j})_{j\geq 1} : i = 1, \ldots, m\}$ be as in definition 3.1. The array $\left(\mathbf{x}^{(i)}\right)_{i=1}^{m}$ is partially exchangeable if and only if there exist a probability measure Q_m on $(\mathrm{P}_{\mathbb{X}}^m, \mathcal{P}_{\mathbb{X}}^m)$ such that*

$$\mathbb{P}\left[\bigcap_{i=1}^{m}\left\{\mathbf{x}_i^{(n_i)} \in A_i\right\}\right] = \int_{\mathrm{P}_{\mathbb{X}}^m}\prod_{i=1}^{m}\tilde{p}_i^{(n_i)}(A_i)\mathrm{Q}_m(\mathrm{d}\tilde{p}_1, \ldots, \mathrm{d}\tilde{p}_m) \tag{3.1}$$

for any integers $n_i \geq 1$ and $A_i \in \mathcal{X}^{(n_i)}$, where $p^{(q)} = p \times \cdots \times p$ denotes the q-fold product measure on $\mathbb{X}^q$ for any $q \geq 1$. Q_m is called the de Finetti measure.

This could be expressed in terms of the random probability measure $\tilde{p} = \tilde{p}_1 \times \cdots \times \tilde{p}_m$ on $(\mathbb{X}^m, \mathcal{X}^m)$ such that $\tilde{p} \sim \mathrm{Q}_m$. In this case, (3.1) reads as

$$\mathbb{P}\left[\bigcap_{i=1}^{m}\left\{\mathbf{x}_i^{(n_i)} \in A_i\right\} \,\middle|\, \tilde{p}_1, \ldots, \tilde{p}_m\right] = \prod_{i=1}^{n_1}\tilde{p}_1(A_1)\cdots\prod_{i=1}^{n_m}\tilde{p}_m(A_m),$$

so that, conditional on $\tilde{p}$, the random variables are i.i.d. within the same sequence and independent across different sequences. Hierarchically

$$\begin{aligned}(\mathrm{x}_{1,j_1},\ldots,\mathrm{x}_{m,j_m}) \,\big|\, \tilde{p}_1,\ldots,\tilde{p}_m &\overset{\text{i.i.d.}}{\sim} \tilde{p}_1 \times \cdots \times \tilde{p}_m \quad (j_1,\ldots,j_m) \in \mathbb{N}^m \\ (\tilde{p}_1,\ldots,\tilde{p}_m) &\sim \mathrm{Q}_m.\end{aligned} \tag{3.2}$$

Q_m plays the role of a prior distribution for the vector of random probability measures $(\tilde{p}_1,\ldots,\tilde{p}_m)$, dictating the dependence across the different groups.

Maximal dependence among $(\tilde{p}_i)_{i=1}^m$ occurs when when Q_m degenerates onto the diagonal and $\tilde{p}_1 = \cdots = \tilde{p}_m = \tilde{p}$ a.s. This corresponds to full exchangeability, as in this case the whole collection $\{(\mathrm{x}_{i,j})_{j\geq 1} : i = 1,\ldots,m\}$ is exchangeable within each of the m groups but also across them, so we might treat them as if they were part of a unique, bigger group. On the other hand, independence occurs when the $\tilde{p}_i$'s are (unconditionally) independent with respect to Q_m. In this case, there is complete heterogeneity between the m groups and we might model each one of them separately. These two choices are extremes because the first one implies maximum borrowing of strength while the latter implies no borrowing of strength, as depicted in Figure 3.2. Going back to the example of the medical treatment, the first choice would pool together male and female patients and the second one would assume different effects of the treatment on males and females with independent priors. In most cases, the desired level of borrowing of strength lies in-between these two extremes. The specific structure of Q_m will determine both the presence and the intensity of the borrowing of strength between the m partially exchangeable sequences.

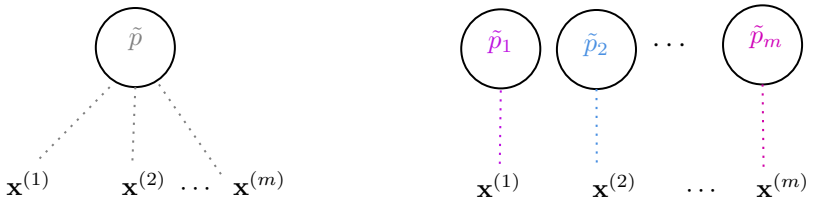


Figure 3.2: One common random probability measure $\tilde{p}$ for the m sequences versus m distinct random probability measures, independent across studies.

Nonparametric proposals for Q_m have been studied back to Cifarelli and Regazzini (1978), where the authors assumed that the random probability measures $\tilde{p}_i$ all have a hyperparameter α (with parametric form) that allows for some borrowing of strength even in the extreme cases, as shown in Figure 3.3. A natural next step was to consider a more complex hyperparameter, namely when α is itself a random probability measure, as it was the case in MacEachern (1999), where the Dependent Dirichlet Process was introduced.

We will focus on hierarchical constructions of Q_m and assume that the elements of the collection $\{\tilde{p}_1,\ldots,\tilde{p}_m\}$ are conditionally i.i.d. given another discrete random probability measure $\tilde{p}_0$. This choice of Q_m selects, with probability 1, vectors of discrete probability measures. One of the most popular methods to specify Q_m is the superposition of random probability measures: consider $(\tilde{p}_i)_{i=1}^m$ a collection of random probability measures, where each $\tilde{p}_i$ corresponds to the sequence $(\mathrm{x}_{i,j})_{j\geq 1}$ for $i = 1,\ldots,m$ and the base measure P_0 of each $\tilde{p}_i$ is no longer deterministic, but random. This would mean that the prior Q_m now

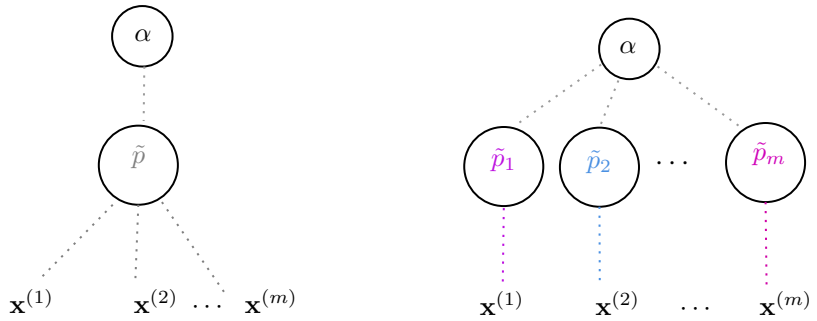


Figure 3.3: The common hyperparameter α, shared by the random probability measures $(\tilde{p}_i)_{i=1}^m$, allows for borrowing of strength even in the limiting cases.

takes the form

$$\begin{aligned} \tilde{p}_i \,|\, \tilde{p}_0 &\sim Q_i(\tilde{p}_0) \quad \text{with } \mathbb{E}[\tilde{p}_i \,|\, \tilde{p}_0] = \tilde{p}_0 \text{ for } i = 1, \ldots, m \\ \tilde{p}_0 &\sim Q_0. \end{aligned}$$

This type of constructions are known as *hierarchical processes*. General classes of vectors of normalized random measures were proposed in Camerlenghi et al. (2019b) and Camerlenghi (2015), in which a systematic investigation of the most relevant distributional properties for this choices of Q_m were studied.

Definition 3.2. Let $\tilde{\mu}$ be a a.s. discrete random measure defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $(\mathrm{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$. Furthermore, assume that $0 < \tilde{\mu}(\mathbb{X}) < \infty$ and consider the random probability measure

$$\tilde{p}(\,\cdot\,) := \frac{\tilde{\mu}(\,\cdot\,)}{\tilde{\mu}(\mathbb{X})} = \mathrm{NRM}(\mathrm{P}),$$

where $\mathbb{E}[\tilde{p}] = \mathrm{P}$ is a probability distribution on $(\mathbb{X}, \mathcal{X})$. With this in mind, consider m partially exchangeable sequences $\left(\mathbf{x}^{(i)}\right)_{i=1}^m$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $(\mathbb{X}, \mathcal{X})$, whose partially exchangeable dependence structure is dictated by

$$\begin{aligned} \tilde{p}_i \,|\, \tilde{p}_0 &\overset{\text{i.i.d.}}{\sim} \mathrm{NRM}(\tilde{p}_0) \quad \text{for } i = 1, \ldots, m \\ \tilde{p}_0 &\sim \mathrm{NRM}(P_0). \end{aligned} \tag{3.3}$$

We will refer to the vector of random probability measures $(\tilde{p}_1, \ldots, \tilde{p}_m)$ as a *vector of hierarchical normalized random measures*. If in particular $\tilde{\mu}_i$ and $\tilde{\mu}_0$ are completely random, then we will refer to $(\tilde{p}_1, \ldots, \tilde{p}_m)$ as a *vector of hierarchical NRMIs*, or HNRMI for short.

We will deal with two specifications of $\tilde{\mu}$ and $\tilde{\mu}_0$: when $\tilde{\mu}$ is a homogeneous completely random measure and when the distribution of $\tilde{\mu}$ is obtained by transforming the distribution of a CRM.

3.2 Hierarchical NRMIs

3.2.1 Covariance structure

Theorem 3.2. *Assume that*

$$\tilde{p}_i \,|\, \tilde{p}_0 \overset{i.i.d.}{\sim} NRMI(\rho, \theta, \tilde{p}_0), \quad \tilde{p}_0 \sim NRMI(\rho_0, \theta_0, P_0)$$

for $i = 1, \ldots, m$, where P_0 is non-atomic. Then, for any $A \in \mathcal{X}$ and $i \neq j$

$$corr\left(\tilde{p}_i(A), \tilde{p}_j(A)\right) = \left\{ 1 + \theta_0 \theta \frac{\int_{\mathbb{R}_+} u e^{-\theta \Psi(u)} \tau_2(u) \mathrm{d}u \int_{\mathbb{R}_+} u e^{-\theta_0 \Psi_0(u)} \tau_{1,0}^2(u) \mathrm{d}u}{\int_{\mathbb{R}_+} u e^{-\theta_0 \Psi_0(u)} \tau_{2,0}(u) \mathrm{d}u} \right\}^{-1}$$

where $\tau_m(u) = \int_{\mathbb{R}_+} v^m e^{-uv} \rho(\mathrm{d}v)$ and $\tau_{m,0}(u) = \int_{\mathbb{R}_+} v^m e^{-uv} \rho_0(\mathrm{d}v)$ for $m \geq 1$.

The proof can be found in Appendix B.

Remark. Note that the correlation coefficient does not depend on the choice of the set A, and it is always positive.

3.2.2 Partition structure

As the choice of the prior distribution Q_m in (3.3) selects a.s. vectors of discrete probability measures, there will be ties, with positive probability, within the same sample and across different samples as well. In the exchangeable framework, the partition structure is characterized by the EPPF; in a partially exchangeable context, an analogous object to the EPPF, termed a *partially exchangeable partition probability function* (pEPPF), serves the exact same purpose in this more general set up.

Suppose that $n_i \geq 1$ and that we have sampled $\mathbf{x}_i^{(n_i)}$ from $\mathbf{x}^{(i)}$ $(i = 1, \ldots, m)$. Let us denote as $\mathbf{x} = \left(\mathbf{x}_i^{(n_i)} : i = 1, \ldots, m\right)$ the m samples and let $N = \sum_{i=1}^m n_i$ be its total size. There will be k_i distinct values specific to sample $\mathbf{x}_i^{(n_i)}$ for $i = 1, \ldots, m$ and, additionally, k_0 distinct values shared across the m samples. Let $k = \sum_{i=0}^m k_i$ denote the total number of unique values across $\mathbf{x}$, so that we can codify the corresponding frequencies of each value as

$$\mathbf{n}_i := (n_{i,1}, \ldots, n_{i,k}) \quad \text{for } i = 1, \ldots, m,$$

with the constraint that $\sum_{j=1}^k n_{i,j} = n_i$. $\mathbf{x}$ induces a partition over $[N]$. An example is as depicted in Figure 3.4, where samples from two partially exchangeable sequences of sizes $n_1 = 4 = n_2$ are presented. Different colors represent distinct values. If we were to enumerate the elements of the sample in order of appearance, the partition of [8] induced by the sample is $\tilde{\pi}$.

The pEPPF is defined as

$$\mathbf{\Pi}_k^{(N)}(\mathbf{n}_1, \ldots, \mathbf{n}_m) = \mathbb{E}\left[\int_{\mathbb{X}^k} \prod_{j=1}^k \tilde{p}_1^{n_{1,j}}(\mathrm{d}x_j) \cdots \tilde{p}_m^{n_{m,j}}(\mathrm{d}x_j) \right].$$

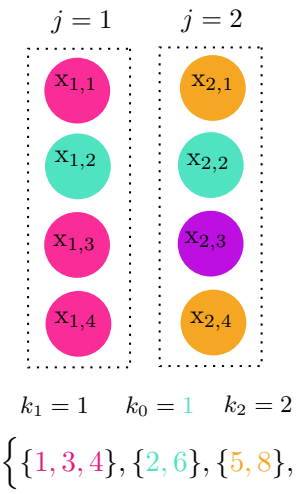


Figure 3.4: Partition over $[8]$ *induced by a sample of* 2 *partially exchangeable sequences.*

In the example of Figure 3.4, the pEPPF would be modelling the probability of the joint partition as the one showed in Figure 3.5.

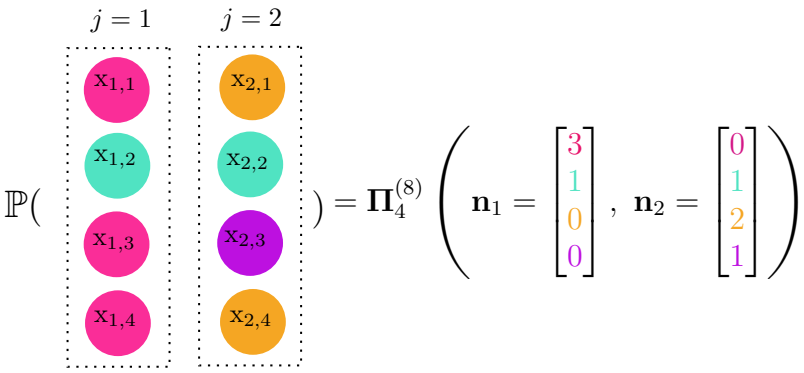


Figure 3.5

To study the partition generated by a sample of a partially exchangeable array, we first introduce a useful metaphor that serves as a generalization of the Chinese restaurant process, aptly named the *Chinese restaurant franchise* (CRF for short), introduced in Teh et al. (2006). The hierarchical structure implies that we have two levels: an observable level and a latent level, governed by $\tilde{p}_0$. Suppose that there are m restaurants that share an infinite global menu, whose dishes are generated by the top level base measure P_0. Each restaurant has an infinite capacity for both tables and customers. Each table serves the same dish, chosen by the first customer that sits, and the same dish can be served at different tables within the same restaurant or across different restaurants. The observable level is the dish arrangement among customers and the tables that partition customers in each restaurant i identify the latent level of the model.

Let $\mathrm{x}_{i,j}$ represent the label of the dish served in the i-th restaurant of the franchise to the j-th customer for $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$, with $n_{i,j}$ being the number of customers eating dish j at restaurant i. After N customers have arrived, there are k distinct dishes being served across the m restaurants. Define

$$\bar{n}_{\bullet j} = \sum_{i=1}^{m} n_{i,j} \quad j = 1, \ldots, k,$$

the quantity $\bar{n}_{\bullet j}$ is the total number of people eating dish j across the franchise. The $n_{i,j}$ customers eating dish j may be further partitioned into tables, so let $\ell_{i,j}$ be the number of tables in restaurant i serving dish j, whose range is $\{1, \ldots, n_{i,j}\}$ if dish j is served at restaurant i and 0 otherwise, and is upper bounded by the number of customers eating dish j in that restaurant. Let $\boldsymbol{\ell}_i$ be

$$\boldsymbol{\ell}_i := (\ell_{i,1}, \ldots, \ell_{i,k}),$$

and finally assume that the dot $\bullet$ represents marginal counts, so that

$$\begin{aligned}
\bar{\ell}_{\bullet j} &= \sum_{i=1}^{m} \ell_{i,j} \quad j = 1, \ldots, k \\
\bar{\ell}_{i\bullet} &= \sum_{j=1}^{k} \ell_{i,j} \quad i = 1, \ldots, m \\
|\boldsymbol{\ell}| &= \sum_{i=1}^{m} \sum_{j=1}^{k} \ell_{i,j}.
\end{aligned}$$

The total number of tables across the franchise serving dish j is $\bar{\ell}_{\bullet j}$, while $\bar{\ell}_{i\bullet}$ is the total number of tables in restaurant i and $|\boldsymbol{\ell}|$ is the total number of tables occupied across the m restaurants. Now we are going to augment the structure by introducing the quantity $q_{i,j,t}$ that represents the refined partition, and equals the frequency of customers at restaurant i eating dish j and sitting at table t, for $j = 1, \ldots, k$ and $t = 1, \ldots, \ell_{i,j}$. If

$$\mathbf{q}_{i,j} = (q_{i,j,1}, \ldots, q_{i,j,\ell_{i,j}}),$$

then $\mathbf{q}_{i,j}$ is the frequency vector of customers in restaurant i eating dish j at each of the $\ell_{i,j}$ tables. If $n_{i,j} = 0$ for some i and j, then $\mathbf{q}_{i,j} = (0, \ldots, 0)$. Note that by marginalizing over the tables we can recover the observed frequencies as $n_{i,j} = |\mathbf{q}_{i,j}| = \sum_{t=1}^{\ell_{i,j}} q_{i,j,t}$, while $q_{i\bullet t} = \sum_{j=1}^{k} q_{i,j,t}$ is the number of customers seated at table t in restaurant i. Figure 3.6 depicts two possible seating arrangements for the CRF metaphor based on the sample of the example depicted in Figure 3.4.

Theorem 3.3. *Suppose that the sequences* $\left(\mathbf{x}^{(i)}\right)_{i=1}^{m}$ *are partially exchangeable and governed by*

$$\tilde{p}_i \,|\, \tilde{p}_0 \overset{i.i.d.}{\sim} NRMI(\rho, \theta, \tilde{p}_0), \quad \tilde{p}_0 \sim NRMI(\rho_0, \theta_0, P_0)$$

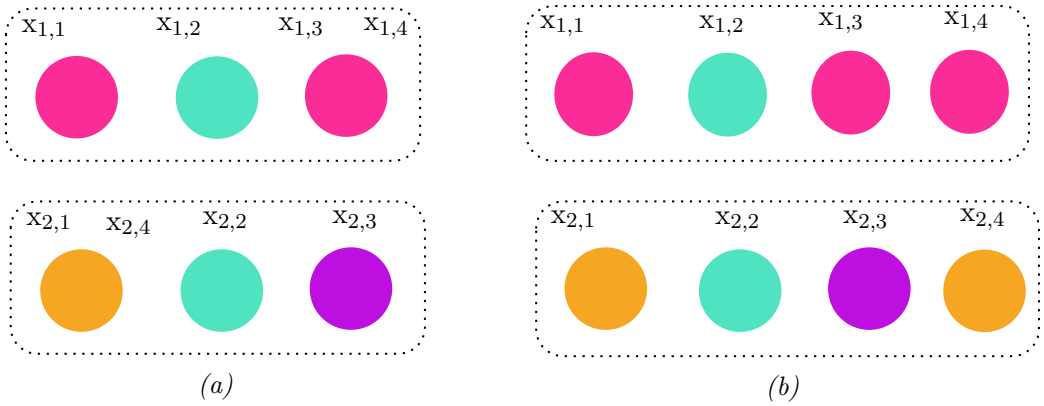


Figure 3.6: Two possible table configurations that could have yielded the sample from Figure 3.4.

with P_0 non-atomic. If we have sampled $\mathbf{x}$ with $n_i \geq 1$ and k distinct values are displayed, then the pEPPF of the partition induced over $[N]$ is

$$\begin{aligned}\boldsymbol{\Pi}_k^{(N)}(\mathbf{n}_1, \ldots, \mathbf{n}_m) = \sum_{\boldsymbol{\ell}} \sum_{\mathbf{q}} \Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\bar{\ell}_{\bullet 1}, \ldots, \bar{\ell}_{\bullet k}) \\ \times \prod_{i=1}^{m} \prod_{j=1}^{k} \frac{1}{\ell_{i,j}!} \binom{n_{i,j}}{q_{i,j,1} \cdots, q_{i,j,\ell_{i,j}}} \Phi_{\bar{\ell}_{i\bullet},i}^{(n_i)}(\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,k}),\end{aligned} \tag{3.4}$$

where $\Phi_{\cdot,0}^{(\cdot)}$ indicates the EPPF induced by an exchangeable sequence drawn from a NRMI of parameters (ρ_0, θ_0, P_0) and

$$\Phi_{\bar{\ell}_{i\bullet},i}^{(n_i)}(\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,k}) = \frac{\theta^{\bar{\ell}_{i\bullet}}}{\Gamma(n_i)} \int_{\mathbb{R}_+} u^{n_i - 1} e^{\theta \Psi(u)} \prod_{j=1}^{k} \prod_{t=1}^{\ell_{i,j}} \tau_{q_{i,j,t}}(u) \mathrm{d}u.$$

$\sum_{\mathbf{q}}$ is a sum over all partitions and $\sum_{\boldsymbol{\ell}}$ is a sum over all compatible table configurations, i.e. over $\ell_{i,j} \in \{1, \ldots, n_{i,j}\}$ with $\ell_{i,j} = 0$ if $n_{i,j} = 0$.

Proof can be found at Appendix B. If $n_{i,j} = 0$ for some i and j, since in this case $\mathbf{q}_{i,j} = (0, \ldots, 0)$, one has that

$$\Phi_{\bar{\ell}_{i\bullet},i}^{(n_i)}(\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,k}) = \Phi_{\bar{\ell}_{i\bullet},i}^{(n_i)}(\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,j-1}, \mathbf{q}_{i,j+1}, \ldots, \mathbf{q}_{i,k}).$$

The backbone of (3.4) is the product

$$\Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\bar{\ell}_{\bullet 1}, \ldots, \bar{\ell}_{\bullet k}) \prod_{i=1}^{m} \Phi_{\bar{\ell}_{i\bullet},i}^{(n_i)}(\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,k}),$$

which describes the random partition's structure acting on the two levels of the hierarchy: the samples (restaurants) and the whole collection of samples (the franchise). $\prod_{i=1}^{m} \Phi_{\bar{\ell}_{i\bullet},i}^{(n_i)}$ captures the former, by describing the probability that the n_i customers in restaurant i are partitioned into $\bar{\ell}_{i\bullet}$ tables, each one occupied by $q_{i,j,t}$ clients. The latter is identified by $\Phi_{k,0}^{(|\boldsymbol{\ell}|)}$, that can be interpreted as the probability that the overall $|\boldsymbol{\ell}|$ tables are partitioned

into k groups according the dishes being served. The specific structure for the arrangement of Figure 3.6b is shown in Figure 3.7. $\Phi^{(8)}_{4,0}$ partitions the eight tables into four groups, i.e. different colors, while each $\Phi^{(4)}_{4,i}$ partitions the four customers into 4 tables at each restaurant, divided with the corresponding dish frequencies dictated by $\mathbf{q}_{i,j}$.

$$\Phi^{(8)}_{4,0}(3,2,1,1)$$

$$\Phi^{(4)}_{4,1}\left(\underbrace{(1,1,1)}_{\mathbf{q}_{1,1}}, \underbrace{(1)}_{\mathbf{q}_{1,2}}\right) \qquad \Phi^{(4)}_{4,2}\left(\underbrace{(1)}_{\mathbf{q}_{2,2}}, \underbrace{(1,1)}_{\mathbf{q}_{2,3}}, \underbrace{(1)}_{\mathbf{q}_{2,4}}\right)$$

Figure 3.7: Backbone of the pEPPF based on the configuration displayed on Figure 3.6b.

3.2.3 Distribution of the number of groups

A natural issue to address once the pEPPF is known, is the distribution of the number K_N of different values out of the $N = \sum_{i=1}^m n_i$ partially exchangeable observations. To this aim let us introduce a collection of latent random variables $\{(\mathrm{t}_{i,j})_{j\geq 1} : i = 1,\dots,m\}$ such that

$$\begin{aligned} \mathrm{t}_{i,j} \,|\, \tilde{q}_i &\overset{\text{i.i.d.}}{\sim} \tilde{q}_i \\ \tilde{q}_i &\sim \mathrm{NRMI}(\theta,\rho,G), \end{aligned}$$

with G a diffuse probability measure. In terms of the Chinese restaurant franchise, $\mathrm{t}_{i,j}$ is the label of the table where the j-th customer of the i-th restaurant is seated. The distribution of K_N can be described by considering:

- Independent random variables K_{i,n_i} that equal, for each $i = 1,\dots,m$, the number of distinct values in $\mathbf{t}_i^{(n_i)} = (\mathrm{t}_{i,1},\dots,\mathrm{t}_{i,n_i})$.
- The random variable $\mathrm{K}_{0,t}$ that represents the number of distinct values out of the t exchangeable random elements generated from $\tilde{p}_0$.

Theorem 3.4. *Suppose that K_N is the number of different values in the m partially exchangeable samples $\left(\mathbf{x}_i^{(n_i)}\right)_{i=1}^m$, governed by a vector of hierarchical NRMIs. For any $k = 1,\dots,N$ one has that*

$$\mathbb{P}[\mathrm{K}_N = k] = \sum_{t=k}^{N} \mathbb{P}[\mathrm{K}_{0,t} = k]\,\mathbb{P}\left[\sum_{i=1}^{m} \mathrm{K}_{i,n_i} = t\right]. \tag{3.5}$$

A proof of Theorem 3.4 can be found in Appendix B. The probability distributions of $\mathrm{K}_{0,t}$ and of K_{i,n_i} are derived from their EPPFs and coincide with

$$\mathbb{P}[\mathrm{K}_{0,t} = k] = \frac{1}{k!} \sum_{(r_1,\dots,r_k)\in\mathcal{C}^k_{[t]}} \binom{t}{r_1\cdots r_k} \Phi^{(t)}_{k,0}(r_1,\dots,r_k) \tag{3.6}$$

for any $k \in \{1, \ldots, t\}$, and

$$\mathbb{P}[\mathrm{K}_{i,n_i} = \zeta] = \frac{1}{\zeta!} \sum_{(r_1,\ldots,r_\zeta) \in \mathcal{C}^{\zeta}_{[n_i]}} \binom{n_i}{r_1 \cdots r_\zeta} \Phi^{(n_i)}_{\zeta,i}(r_1, \ldots, r_\zeta) \tag{3.7}$$

for $\zeta \in \{1, \ldots, n_i\}$. In words, $\mathbb{P}\left[\sum_{i=1}^{m} \mathrm{K}_{i,n_i} = t\right]$ in (3.5) identifies the total number of tables on which the customers were partitioned, whereas $\mathbb{P}[\mathrm{K}_{0,t} = k]$ identifies the number of unique dishes assigned to these tables.

3.2.4 Posterior characterization

A similar characterization of the posterior distribution as the one in Proposition 2.1 is available for partially exchangeable sequences. Namely, the posterior distribution will be composed in two blocks, one concerning the root of the hierarchy (in terms of $\tilde{\mu}_0$) and the second one concerning the vector of random probability measures. Let $\mathrm{x}^*_1, \ldots, \mathrm{x}^*_k$ be the k unique values displayed across $\mathbf{x}$ and assume that U_0 is a positive random variable that, conditional on $\mathbf{x}$ and on the latent tables' labels $\mathbf{t} = \{\mathbf{t}_i^{(n_i)} : i = 1, \ldots, m\}$, admits a density with respect to the Lebesgue measure that satisfies

$$f_0(u \,|\, \mathbf{x}, \mathbf{t}) \propto u^{|\boldsymbol{\ell}|-1} e^{-\theta_0 \Psi_0(u)} \prod_{j=1}^{k} \tau_{\bar{\ell}_{\bullet j},0}(u).$$

Theorem 3.5. *Suppose that a sample* $\mathbf{x}$ *has been obtained from a partially exchangeable sequence, governed a vector of hierarchical NRMIs. Then*

$$\tilde{\mu}_0 \,|\, (\mathbf{x}, \mathbf{t}, U_0) \sim \eta^*_0 + \sum_{j=1}^{k} I_j \delta_{\mathrm{x}^*_j},$$

where η^*_0 *and* $\sum_{j=1}^{k} I_j \delta_{\mathrm{x}^*_j}$ *are independent and:*

- η^*_0 *is a CRM with intensity* $\nu_0(\mathrm{d}v, \mathrm{d}x) = e^{-U_0 x} \rho_0(v) \mathrm{d}v \theta_0 P_0(\mathrm{d}x)$.
- *The jump random variables* $(I_j)_{j=1}^{k}$ *are independent and nonnegative, each with density* $f_j(v \,|\, \mathbf{x}, \mathbf{t}) \propto v^{\bar{\ell}_{\bullet j}} e^{-vU_0} \rho_0(v)$.

Let $\mathbf{U} = (U_1, \ldots, U_m)$ be a vector of restaurant-specific r.v. whose components are conditionally independent, given $(\mathbf{x}, \mathbf{t})$, with density w.r.t. the Lebesgue measure

$$f_i(u \,|\, \mathbf{x}, \mathbf{t}) \propto u^{n_i - 1} e^{-\theta \Psi(u)} \prod_{j=1}^{k} \prod_{t=1}^{\ell_{i,j}} \tau_{q_{i,j,t}}(u).$$

Theorem 3.6. *Suppose that* $\mathbf{x}$ *is a sample of* m *partially exchangeable sequences, governed a vector of hierarchical NRMIs. Then*

$$(\tilde{\mu}_1, \ldots, \tilde{\mu}_m) \,|\, (\mathbf{x}, \mathbf{t}, \mathbf{U}, \tilde{\mu}_0) \sim (\tilde{\mu}^*_1, \ldots, \tilde{\mu}^*_m) + \left(\sum_{j=1}^{k} \sum_{t=1}^{\ell_{1,j}} J_{1,j,t} \delta_{X^*_j}, \ldots, \sum_{j=1}^{k} \sum_{t=1}^{\ell_{d,j}} J_{m,j,t} \delta_{\mathrm{x}^*_j} \right),$$

where the two summands on the right hand side are independent, $\sum_{t=1}^{\ell_{i,j}} J_{i,j,t} \equiv 0$ *if* $n_{i,j} = 0$ *and:*

- $(\tilde{\mu}_1^*, \ldots, \tilde{\mu}_m^*)$ *is a vector of hierarchical CRMs such that, conditional on* $\tilde{\mu}_0^* = \eta_0^* + \sum_{j=1}^{k} I_j \delta_{\mathrm{x}_j^*}$*, each* $\tilde{\mu}_i^*$ *has intensity* $\nu_i(\mathrm{d}v, \mathrm{d}x) = e^{-U_i v}\rho(v)\mathrm{d}v\theta\tilde{p}_0^*(\mathrm{d}x)$*, with* $\tilde{p}_0^*(\,\cdot\,) = \frac{\tilde{\mu}_0^*(\,\cdot\,)}{\tilde{\mu}_0^*(\mathbb{X})}$.

- *The jump random variables* $\{(J_{i,j,t})_{t=1}^{\ell_{i,j}} : i \in \{1, \ldots, m\},\, j \in \{1, \ldots, m\}\}$ *are independent and nonnegative, each with density* $f_{i,j,t}(v) \propto e^{-U_i v} v^{q_{i,j,t}} \rho(v)$ *when* $n_{i,j} \geq 1$*, whereas* $J_{i,j,t} = 0$ *a.s. if* $n_{i,j} = 0$.

The proofs of Theorems 3.5 and 3.6 can be found at Appendix B.

3.2.5 Sampling scheme

To construct a generative sampling scheme for hierarchical NRMIs, we will make use of a posterior characterization of $\mathbf{x}$ that relies on two latent variables. In the Chinese restaurant franchise metaphor, recall that the random variable $\mathrm{t}_{i,j}$ from Theorem 3.4 is the label of the table at which the j-th customer on restaurant i sits, and let $\mathrm{d}_{i,t}$ be the label of the dish served at table t in restaurant i. Let the random variables $\phi_1, \ldots, \phi_k$ denote the k unique dishes. To illustrate these new quantities, Figure 3.8 exhibits the latent structure of the arrangement of the example of Figure 3.6a: the dotted gray lines represent the indexes $\mathrm{t}_{i,j}$ that identify customers with tables at each restaurant i, whereas the dashed dark gray lines represent the indexes $\mathrm{d}_{i,t}$, which identify tables with dishes (colors) sampled from the top level RPM's base measure P_0.

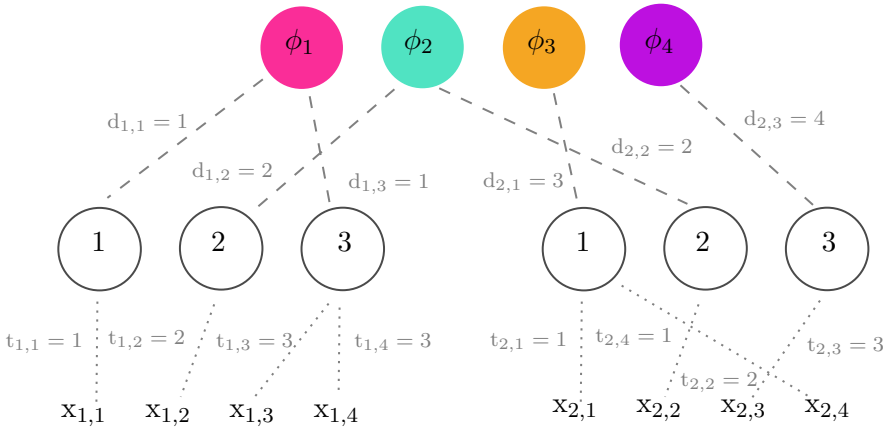


Figure 3.8: Latent random variables $\mathrm{t}_{i,j}$ *and* $\mathrm{d}_{i,t}$ *for the sample of Figure 3.6a.*

The random variables $\mathrm{t}_{i,j}$ and $\mathrm{d}_{i,t}$ allow us to record the clustering behavior of the hierarchical process at both levels of the hierarchy: $\mathrm{t}_{i,j}$ indicates the *local* cluster membership whereas $\mathrm{d}_{i,t}$ associates each of these clusters across the m groups with a *global* label sampled from P_0. More formally, given a random partition Π, let $\mathscr{C}_j(\Pi)$ be the random index of the block containing element j, that is $\mathscr{C}_j(\Pi) = c$ if j belongs to the c-th block of Π. Regarding a vector of hierarchical NRMIs, let $\widetilde{\Phi}_i$ indicate the EPPF induced by an

exchangeable sequence drawn from the process $\tilde{p}_i \,|\, \tilde{p}_0 \overset{\text{i.i.d.}}{\sim} \text{NRMI}(\rho, \theta, \tilde{p}_0)$, for $i = 1, \ldots, m$, and let Φ_0 be the EPPF associated with $\tilde{p}_0 \sim \text{NRMI}(\rho_0, \theta_0, P_0)$, just as in Theorem 3.3. Let $\bar{\pi}_1, \ldots, \bar{\pi}_m$ denote independent random partitions of $\mathbb{N}$, each with EPPF $\widetilde{\Phi}_i$, and let $\bar{\pi}_{i,n_i}$ be the restriction of $\bar{\pi}_i$ to $[n_i]$, which has a probability distribution $\Phi^{n_i}_{|\bar{\pi}_i|,i}$. Additionally, $\tilde{\pi}_0$ is a random partition of $\mathbb{N}$ such that, conditional on $(\tilde{\pi}_{1,n_1}, \ldots, \tilde{\pi}_{m,n_m})$, its restriction $\tilde{\pi}_{0,h}$ to $[h]$ has probability distribution $\Phi^{(h)}_{k,0}$, where h is the sum of the number of blocks in each partition $\tilde{\pi}_{i,n_i}$, i.e. $h = \sum_{i=1}^{m} |\tilde{\pi}_{i,n_i}|$. In terms of these partitions, the labels $\mathrm{t}_{i,j}$ and $\mathrm{d}_{i,t}$ can be described as

$$\mathrm{d}^*_{i,j} := \mathscr{C}_{\mathscr{D}(i,\mathrm{t}_{i,j})}(\tilde{\pi}_{0,h}) \qquad \mathscr{D}(i,t) := \sum_{r=1}^{i-1} |\tilde{\pi}_{k,n_r}| + t \qquad \mathrm{t}_{i,j} := \mathscr{C}_j(\tilde{\pi}_i),$$

where $\mathrm{d}^*_{i,j} = \mathrm{d}_{i,\mathrm{t}_{i,j}}$ for $\mathrm{d}_{i,t} := \mathscr{C}_{\mathscr{D}(i,t)}(\tilde{\pi}_{0,h})$. $\mathrm{t}_{i,j}$ has a straightforward interpretation: it merely indicates the block to which the j-th observation in the i-th partition belongs to. The quantity $\mathscr{D}(i,t)$ scrolls the index to the current block t in group i by summing the number of blocks $|\tilde{\pi}_{i,n_i}|$ in the previous $i-1$ partitions.

Consider again the example shown in Figure 3.6a. In Figure 3.9, we show the corresponding partitions $\tilde{\pi}_{i,n_i}$ of $[n_i]$ into $\bar{\ell}_{i\bullet}$ blocks, generated by the labels of the table on which each customer sits, and the partition $\tilde{\pi}_{0,|\boldsymbol{\ell}|}$ of $[|\boldsymbol{\ell}|]$ generated by the four different dish labels. The partitions $\tilde{\pi}_{i,n_i}$ act at a restaurant level and the partition $\tilde{\pi}_{0,6}$ acts within the 6 tables across the franchise. As there are no tables previous to restaurant 1, $\mathscr{D}(1, c_{1,j}) = t_{1,j}$ and therefore $\mathrm{d}^*_{1,j} = \mathscr{C}_{t_{1,j}}(\tilde{\pi}_{0,6})$. On the other hand, when $i = 2$ we have to take into account that there are $|\tilde{\pi}_{1,4}| = 3$ tables in restaurant 1, so that the first table in restaurant two actually corresponds to the fourth table served across the franchise and so on, meaning that $\mathscr{D}(2, \mathrm{t}_{2,j}) = 3 + \mathrm{t}_{2,j}$.

$$\tilde{\pi}_{0,6} = \Big\{ \underset{\phi_1}{\{1,3\}}, \underset{\phi_2}{\{2,5\}}, \underset{\phi_3}{\{4\}}, \underset{\phi_4}{\{6\}} \Big\}$$

$$\mathrm{d}^*_{1,j} = \begin{cases} \mathscr{C}_1(\tilde{\pi}_{0,6}) = 1 & j = 1 \\ \mathscr{C}_2(\tilde{\pi}_{0,6}) = 2 & j = 2 \\ \mathscr{C}_3(\tilde{\pi}_{0,6}) = 1 & j = 3, 4 \end{cases} \qquad \mathrm{d}^*_{2,j} = \begin{cases} \mathscr{C}_4(\tilde{\pi}_{0,6}) = 3 & j = 1, 4 \\ \mathscr{C}_5(\tilde{\pi}_{0,6}) = 2 & j = 2 \\ \mathscr{C}_6(\tilde{\pi}_{0,6}) = 4 & j = 4 \end{cases}$$

$$\tilde{\pi}_{1,4} = \Big\{ \underset{\phi_1}{\{1\}}, \underset{\phi_2}{\{2\}}, \underset{\phi_3}{\{3,4\}} \Big\} \qquad \tilde{\pi}_{2,4} = \Big\{ \underset{\phi_4}{\{1,4\}}, \underset{\phi_2}{\{2\}}, \underset{\phi_3}{\{3\}} \Big\}$$
$$\mathrm{t}_{1,1} = 1,\ \mathrm{t}_{1,2} = 2,\ \mathrm{t}_{1,3} = 3 = \mathrm{t}_{1,4} \qquad \mathrm{t}_{2,1} = 1 = \mathrm{t}_{2,4},\ \mathrm{t}_{2,2} = 2,\ \mathrm{t}_{2,3} = 3$$

Figure 3.9: Partitions generated by the sample of Figure 3.6a and their corresponding index random variables $\mathrm{t}_{i,j}$ *and* $\mathrm{d}^*_{i,j}$.

In the following proposition we will see that we can characterize the joint law of $\mathbf{x}$ in terms of these latent variables, following along the same lines of reasoning as in the example presented above.

Proposition 3.1. Let $\mathbf{x}$ be m samples of a partially exchangeable sequence, governed by a vector of hierarchical NRMIs and let $(\phi_n)_{n\geq 1}$ be a sequence of i.i.d. random variables with distribution P_0. Then $\mathbf{x} \overset{d}{=} \left[\phi_{\mathrm{d}^*_{i,j}} : j = 1, \ldots, n_i,\, i \in \{1, \ldots, m\}\right]$.

This corresponds to Proposition 4 in Bassetti et al. (2020), and it holds true for a broader class of hierarchical processes, hierarchical species sampling models, which include both hierarchical NRMIs and the hierarchical Pitman-Yor process. This proposition tells us that we can reconstruct the seating plan of Figure 3.6a by first sampling the dotted lines (customer-table assignments), the dashed lines (table-dish assignments) in Figure 3.8 and then assign each label (color) $\{\phi_n\}_{n=1}^4 \overset{\text{i.i.d.}}{\sim} P_0$ to each table, and therefore to each customer. To derive a sampling scheme, recall that if EPPF of an exchangeable sequence is known, given a particular instance of a partition of $[n]$ into k blocks, each of size n_j $(j = 1, \ldots, k)$, the probability of adding a new block containing $n+1$ is given by $\omega_0^{(n)}(\mathrm{n}_1, \ldots, \mathrm{n}_k)$ as in (2.4), while the probability of adding $n+1$ to the j-th block is given by $\omega_j^{(n)}(\mathrm{n}_1, \ldots, \mathrm{n}_k)$ as in (2.5). We will omit the values $(\mathrm{n}_1, \ldots, \mathrm{n}_k)$ unless it is necessary to make explicit the composition on which each one of them is evaluated at. Let $\omega_j^{(n)}$ and $\omega_0^{(n)}$ be the weights of the predictive distribution of the random partitions $\tilde{\pi}_i$ with EPPF $\tilde{\Phi}_i$, for $i = 1, \ldots, m$ and let $\widetilde{\omega}_j^{(n)}$ and $\widetilde{\omega}_0^{(n)}$ defined in an analogous way for $\tilde{\pi}_0$ with EPPF Φ_0. With these weights, Proposition 3.1 allows us to generate samples from a HNRMI by the means of Algorithm 3.1, where $\mathrm{x}^*_{i,1}, \ldots, \mathrm{x}^*_{i,\bar{\ell}_{i\bullet}}$ are the labels of the $\bar{\ell}_{i\bullet}$ tables at restaurant i for $i = 1, \ldots, m$. A graphical depiction is shown in Figure 3.10.

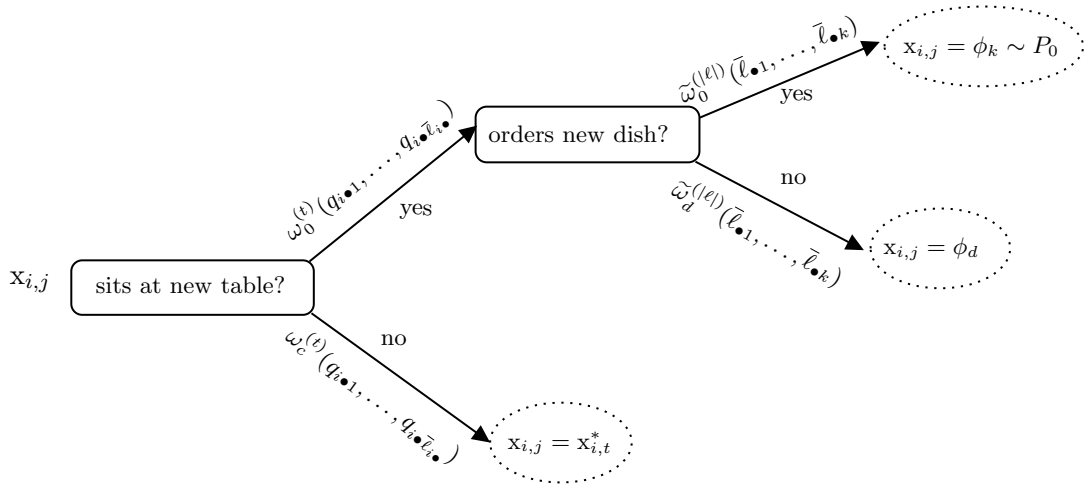


Figure 3.10: Diagram of the sampling scheme described in Algorithm 3.1.

Algorithm 3.1: Sample $\left(\mathbf{x}_i^{(n_i)}\right)_{i=1}^{m}$ using the EPPFs.

for $i = 1, 2, \ldots, m$ **do**

  if $i = 1$ **then**

    Sample $\phi_1 \sim P_0$ and assign $\mathrm{x}_{1,1} = \mathrm{x}^*_{1,1} = \phi_1$

    Set $k = 1$, $\bar{\ell}_{1\bullet} = 1$, $\bar{\ell}_{\bullet 1} = 1$, $q_{1\bullet 1} = 1$

    else

      Take $\mathrm{x}_{i,1} = \mathrm{x}^*_{i,1}$, where $\mathrm{x}^*_{i,1}$ is sampled from G_{it}

      Set $\bar{\ell}_{i\bullet} = 1$ and $q_{i\bullet 1} = 1$

  for $j = 1, \ldots, n_i$ **do**

    Sample $\mathrm{x}_{i,j} \,|\, \mathrm{x}_{i,1}, \ldots, \mathrm{x}_{i,j-1}$ from $G^*_{it}(\,\cdot\,) + \omega_0^{(t)}\left(q_{i\bullet 1}, \ldots q_{i\bullet\bar{\ell}_{i\bullet}}\right) G_{it}(\,\cdot\,)$, where

$$G^*_{it}(\,\cdot\,) = \sum_{t=1}^{\bar{\ell}_{i\bullet}} \omega_t^{(t)}\left(q_{i\bullet 1}, \ldots q_{i\bullet\bar{\ell}_{i\bullet}}\right) \delta_{\mathrm{x}^*_{i,t}}(\cdot)$$

$$G_{it}(\,\cdot\,) = \tilde{G}_{it}(\,\cdot\,) + \widetilde{\omega}_0^{(|\boldsymbol{\ell}|)}\left(\bar{\ell}_{\bullet 1}, \ldots, \bar{\ell}_{\bullet k}\right) P_0(\,\cdot\,)$$

$$\tilde{G}_{it}(\,\cdot\,) = \sum_{d=1}^{k} \widetilde{\omega}_d^{(|\boldsymbol{\ell}|)}\left(\bar{\ell}_{\bullet 1}, \ldots, \bar{\ell}_{\bullet k}\right) \delta_{\phi_d}(\cdot)$$

    if $\mathrm{x}_{i,j}$ *is sampled from* G^*_{it} **then**

      `// Sits at old table`

      Set $\mathrm{x}_{i,j} = \mathrm{x}^*_{i,t}$ for the chosen t and set $\mathrm{t}_{i,j} = t$

      Increment customer-table count $q_{i\bullet t} = q_{i\bullet t} + 1$

    else if $\mathrm{x}_{i,j}$ *is sampled from* G_{it} **then**

      `// Sits at new table`

      Set $\mathrm{x}_{i,j} = \mathrm{x}^*_{i,\bar{\ell}_{i\bullet}}$ and $\mathrm{t}_{i,j} = \bar{\ell}_{i\bullet}$

      Increment table counts $\bar{\ell}_{i\bullet} = \bar{\ell}_{i\bullet} + 1$ and $|\boldsymbol{\ell}| = |\boldsymbol{\ell}| + 1$

      if $\mathrm{x}_{i,j}$ *is sampled from* $\tilde{G}_{it}$ **then**

        `// Orders old dish`

        Set $\mathrm{x}_{i,\bar{\ell}_{i\bullet}} = \phi_d$ for the chosen d and $\mathrm{d}_{i,\bar{\ell}_{i\bullet}} = d$

        Increment customer-mish count $\bar{\ell}_{\bullet d} = \bar{\ell}_{\bullet d} + 1$

      else if $\mathrm{x}_{i,j}$ *is sampled from* P_0 **then**

        `// Orders new dish`

        Set $\phi_k = \mathrm{x}_{i,j}$, and $\mathrm{d}_{i,\bar{\ell}_{i\bullet}} = k$

        Increment the global dish count $k = k + 1$

Remark. The reason why we choose to derive Algorithm 3.1 in terms of the EPPFs $\widetilde{\Phi}_i$ instead of using the partition probability functions Φ_i as defined in Theorem 3.3 is only for the sake of simplicity. One could express the exact same posterior characterization using solely Φ_i as follows:

$$\begin{aligned}
\mathbb{P}[\mathrm{x}_{i,j} = \text{new dish}, \mathrm{t}_{i,j} = \text{new table} \,|\, ...] &= \mathbb{P}[\mathrm{x}_{i,j} = \phi_{k+1}, \mathrm{t}_{i,j} = t^{\text{new}} \,|\, ...] \\
&= \widetilde{\omega}_0^{(|\boldsymbol{\ell}|)} \frac{\Phi^{(n_i+1)}_{\bar{\ell}_{i\bullet}+1,i}\left(\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,k}, 1\right)}{\Phi^{(n_i)}_{\bar{\ell}_{i\bullet},i}\left(\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,k}\right)} \\
\mathbb{P}[\mathrm{x}_{i,j} = \text{old dish}, \mathrm{t}_{i,j} = \text{new table} \,|\, ...] &= \mathbb{P}[\mathrm{x}_{i,j} = \phi_{d}, \mathrm{t}_{i,j} = t^{\text{new}} \,|\, ...] \\
&= \widetilde{\omega}_d^{(|\boldsymbol{\ell}|)} \frac{\Phi^{(n_i+1)}_{\bar{\ell}_{i\bullet}+1,i}\left(\mathbf{q}_{i,1}, \ldots, (\mathbf{q}_{i,d}, 1), \ldots \mathbf{q}_{i,k}\right)}{\Phi^{(n_i)}_{\bar{\ell}_{i\bullet},i}\left(\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,k}\right)} \\
\mathbb{P}[\mathrm{x}_{i,j} = \text{old dish}, \mathrm{t}_{i,j} = \text{old table} \,|\, ...] &= \mathbb{P}[\mathrm{x}_{i,j} = \phi_{\mathrm{d}_{i,t^{\text{old}}}}, \mathrm{t}_{i,j} = t^{\text{old}} \,|\, ...] \\
&= \frac{\Phi^{(n_i)}_{\bar{\ell}_{i\bullet},i}\left(\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,\mathrm{d}_{i,t^{\text{old}}}} + \mathbf{1}_{t^{\text{old}}}, \ldots, \mathbf{q}_{i,k}\right)}{\Phi^{(n_i)}_{\bar{\ell}_{i\bullet},i}\left(\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,k}\right)},
\end{aligned}$$

where $\mathbf{1}_{t^{\text{old}}}$ is a vector of zeros of length $\bar{\ell}_{i,\mathrm{d}_{i,t^{\text{old}}}}$, with a one at position t^{old}. This scheme is very similar to that of Algorithm 3.1, the main difference being in here that the dishes are coupled within the frequency vectors $\mathbf{q}_{i,j}$, as these record the customer-table counts but arranged by dish. Using $\widetilde{\Phi}_i$ instead of Φ_i allows us to de-couple these frequencies so that we can use the actual customer-table counts $q_{i\bullet t}$, separately from the dish counts.

3.3 Examples

Now we will present some examples of hierarchical processes where $\tilde{p}_i$ and $\tilde{p}_0$ are of the same kind of process, although this is not necessary and one can construct any mixed case.

3.3.1 Hierarchies of Dirichlet processes

If $\rho(v) = \rho_0(v) = v^{-1}e^{-v}$, then $\tilde{p}_0$ is a Dirichlet process and $\tilde{p}_i$'s are, conditional on $\tilde{p}_0$, independent and identically distributed Dirichlet processes, so that

$$\begin{aligned}
\tilde{p}_i \,|\, \tilde{p}_0 &\overset{\text{i.i.d.}}{\sim} \mathfrak{D}(\theta \tilde{p}_0) \quad \text{for } i = 1, \ldots, m \\
\tilde{p}_0 &\sim \mathfrak{D}(\theta_0 P_0).
\end{aligned}$$

Therefore, $(\tilde{p}_1 \ldots, \tilde{p}_m)$ is a *vector of hierarchical Dirichlet processes* (HDP for short) as in Teh et al. (2006). We previously proved that $\tau_{m,0}(u) = \frac{\Gamma(m)}{(1+u)^m} = \tau_m(u)$, so that a straightforward application of Theorem 3.2 yields that for $A \in \mathcal{X}$ and $i \neq j \in \{1, \ldots, m\}$

$$\operatorname{corr}\left(\tilde{p}_i(A), \tilde{p}_j(A)\right) = \frac{1+\theta}{1+\theta+\theta_0}.$$

The correlation is increasing as a function of θ and decreasing in θ_0. We can distinguish two limiting cases: as $\theta_0 \uparrow \infty$, the distribution of $\tilde{p}_0$ degenerates on its base measure P_0 and the $\tilde{p}_i$'s are independent, consequently $\mathrm{corr}(\tilde{p}_i(A), \tilde{p}_j(A))$ converges to 0. On the other hand, if $\theta \uparrow \infty$, the distribution of each $\tilde{p}_i$, conditional on $\tilde{p}_0$, degenerates on $\tilde{p}_0$ and hence the correlation coefficient between any pair $\tilde{p}_i(A)$ and $\tilde{p}_j(A)$ converges to 1.

To compute the pEPPF, note that $\Phi^{(n_i)}_{\bar{\ell}_{i\bullet},i}(\mathbf{q}_{i,1},\ldots,\mathbf{q}_{i,k})$ can be determined easily as

$$\begin{aligned}
\Phi^{(n_i)}_{\bar{\ell}_{i\bullet},i}(\mathbf{q}_{i,1},\ldots,\mathbf{q}_{i,k}) &= \frac{\theta^{\bar{\ell}_{i\bullet}}}{\Gamma(n_i)} \int_{\mathbb{R}_+} u^{n_i-1}(1+u)^{-\theta} \prod_{j=1}^{k}\prod_{t=1}^{\ell_{i,j}} \frac{\Gamma(q_{i,j,t})}{(1+u)^{q_{i,j,t}}}\mathrm{d}u \\
&= \frac{\theta^{\bar{\ell}_{i\bullet}}}{\Gamma(n_i)} \prod_{j=1}^{k}\prod_{t=1}^{\ell_{i,j}} \Gamma(q_{i,j,t}) \int_{\mathbb{R}_+} \frac{u^{n_i-1}}{(1+u)^{n_i+\theta}}\mathrm{d}u \\
&= \frac{\theta^{\bar{\ell}_{i\bullet}}}{(\theta)_{n_i}} \prod_{j=1}^{k}\prod_{t=1}^{\ell_{i,j}} \Gamma(q_{i,j,t}).
\end{aligned}$$

As we previously proved that $\Phi^{(|\boldsymbol{\ell}|)}_{k,0}(\bar{\ell}_{\bullet 1},\ldots,\bar{\ell}_{\bullet k}) = \frac{\theta_0^k}{(\theta_0)_{|\boldsymbol{\ell}|}} \prod_{j=1}^{k}(\bar{\ell}_{\bullet j}-1)!$, a straightforward application of Theorem 3.3 leads to the following pEPPF

$$\begin{aligned}
\boldsymbol{\Pi}^{(N)}_k(\mathbf{n}_1,\ldots,\mathbf{n}_m) &= \sum_{\boldsymbol{\ell}}\sum_{\mathbf{q}} \frac{\theta_0^k}{(\theta_0)_{|\boldsymbol{\ell}|}} \prod_{j=1}^{k}(\bar{\ell}_{\bullet j}-1)! \prod_{i=1}^{m}\frac{1}{\ell_{i,j}!}\binom{n_{i,j}}{q_{i,j,1}\cdots,q_{i,j,\ell_{i,j}}} \frac{\theta^{\bar{\ell}_{i\bullet}}}{(\theta)_{n_i}} \\
&\quad \times \prod_{t=1}^{\ell_{i,j}} \Gamma(q_{i,j,t}) \\
&= \frac{\theta_0^k}{\prod_{i=1}^{m}(\theta)_{n_i}} \sum_{\boldsymbol{\ell}} \frac{\theta^{|\boldsymbol{\ell}|}}{(\theta_0)_{|\boldsymbol{\ell}|}} \prod_{j=1}^{k}(\bar{\ell}_{\bullet j}-1)! \prod_{i=1}^{m}\sum_{\mathbf{q}} \frac{n_{i,j}!}{\ell_{i,j}!}\frac{1}{q_{i,j,1}\cdots q_{i,j,\ell_{i,j}}} \\
&= \frac{\theta_0^k}{\prod_{i=1}^{m}(\theta)_{n_i}} \sum_{\boldsymbol{\ell}} \frac{\theta^{|\boldsymbol{\ell}|}}{(\theta_0)_{|\boldsymbol{\ell}|}} \prod_{j=1}^{k}(\bar{\ell}_{\bullet j}-1)! \prod_{i=1}^{m} |\mathfrak{s}_{n_{i,j},\ell_{i,j}}|,
\end{aligned}$$

where we have used that the unsigned Stirling numbers of the first kind can be written as $|\mathfrak{s}_{n,k}| = \frac{n!}{k!}\sum_{(r_1,\ldots,r_k)\in\mathcal{C}^k_{[n]}} \frac{1}{r_1\cdots r_k}$, see Charalambides (2002).

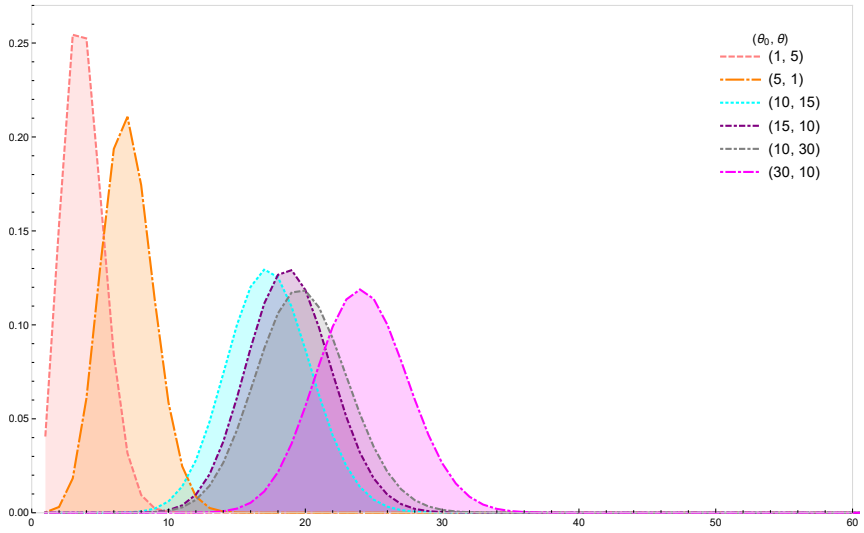
To analyze the distribution of the number of distinct values, K_N, note that on one hand we have that

$$\begin{aligned}
\mathbb{P}[\mathrm{K}_{0,t}=k] &= \frac{\theta_0^k}{(\theta_0)_t}\frac{1}{k!} \sum_{(r_1,\ldots,r_k)\in\mathcal{C}^k_{[N]}} \binom{t}{r_1\cdots r_k} \prod_{j=1}^{k}(r_j-1)! \\
&= \frac{\theta_0^k}{(\theta_0)_t}\frac{t!}{k!} \sum_{(r_1,\ldots,r_k)\in\mathcal{C}^k_{[N]}} \frac{\prod_{j=1}^{k}(r_j-1)!}{\prod_{j=1}^{k} r_j!} = \frac{\theta_0^k}{(\theta_0)_t}|\mathfrak{s}_{t,k}|,
\end{aligned}$$

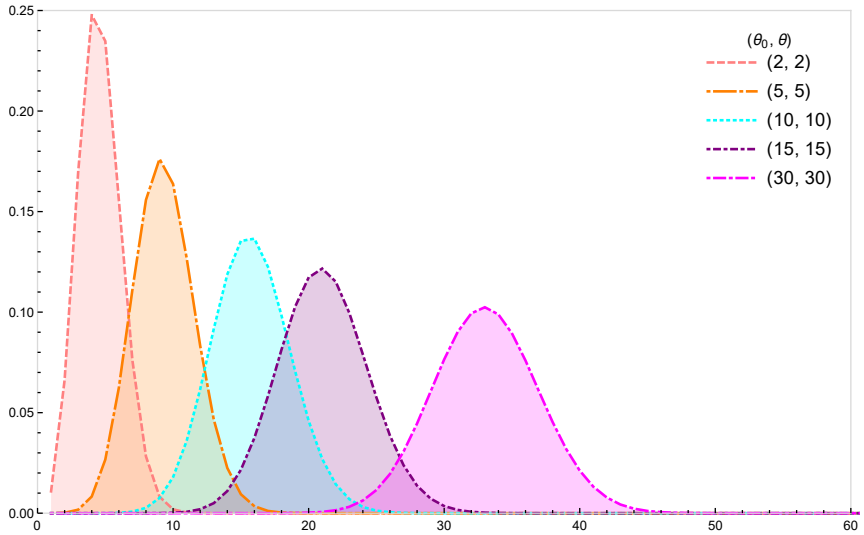
for any $k \in \{1,\ldots,t\}$, and analogously for $i = 1,\ldots,m$

$$\mathbb{P}[\mathrm{K}_{i,n_i}=\zeta] = \frac{\theta^{\zeta}}{(\theta)_{n_i}}|\mathfrak{s}_{n_i,\zeta}| \quad \text{for } \zeta \in \{1,\ldots,n_i\}.$$

To illustrate the effect that the parameters θ_0 and θ have on the distribution of K_N, Figure 3.11 shows plots of the distribution of K_N for two partially exchangeable sequences governed by a vector a hierarchical Dirichlet processes, where $n_1 = 50 = n_2$. From here we see that when both parameters grow, K_N favors larger values and this happens only when both (θ, θ_0) grow simultaneously, as whenever only one of is large, the effect on the location of the mode is diminished by the smaller parameter. Regardless of the placement of the mode, the a priori distribution of K_N is leptokurtic.



(a) Distribution of K_{100} *for the HDP with* $\theta_0 \neq \theta$.



(b) Distribution of K_{100} *for the HDP with* $\theta_0 = \theta$.

Figure 3.11: Distribution of K_{100} *for the HDP for different choices of* θ_0 *and* θ.

To complete the description of the hierarchical Dirichlet Process, note that

$$f_0(u) \propto \frac{u^{|\boldsymbol{\ell}|-1}}{(1+u)^{\theta_0+|\boldsymbol{\ell}|}}$$

so that $\frac{U_0}{U_0+1} \sim \mathrm{Be}(|\boldsymbol{\ell}|, \theta_0)$. The posterior distribution of $\tilde{\mu}_0$ is composed of the two summands:

- η_0^* is a Gamma CRM with intensity $\nu_0(\mathrm{d}v, \mathrm{d}x) = e^{-(1+U_0)v} v^{-1} \mathrm{d}v \theta_0 P_0(\mathrm{d}x)$.
- The jumps $(I_j)_{j=1}^k$ have density $f_j(v \,|\, \mathbf{x}, \mathbf{t}) \propto v^{\bar{\ell}_{\bullet j}-1} e^{-(1+U_0)v}$, meaning that $I_j \overset{\text{ind}}{\sim} \mathrm{Ga}(\bar{\ell}_{\bullet j},\, 1+U_0)$ for $j = 1, \ldots, k$.

Therefore, $\tilde{p}_0^* = \frac{\tilde{\mu}_0^*}{\tilde{\mu}_0^*(\mathbb{X})}$ satisfies

$$\tilde{p}_0^* \sim \mathfrak{D}\left(\theta_0 P_0 + \sum_{j=1}^k \bar{\ell}_{\bullet j} \delta_{X_j}^*\right)$$

as the normalizing constants of η_0^* and the jumps I_j do not depend on the scale U_0. Now concerning the vector $(\tilde{p}_1, \ldots, \tilde{p}_m)$, one has that

$$f_i(u \,|\, \mathbf{x}, \mathbf{t}) \propto \frac{u^{n_i - 1}}{(1+u)^{\theta + n_i}}$$

Hence $\frac{U_i}{U_i+1} \sim \mathrm{Be}(\theta,\, n_i)$ and, conditional on $\tilde{p}_0^*$ and $\mathbf{x}, \mathbf{t}, \mathbf{U}$, the CRMs $\tilde{\mu}_1, \ldots, \tilde{\mu}_m$ are independent and distributed as $\tilde{\mu}_i^* + \sum_{j=1}^k \sum_{t=1}^{\ell_{i,j}} J_{i,j,t} \delta_{\mathrm{x}_{i,j}}^*$, where

- $\tilde{\mu}_i^*$ is a Gamma CRM with intensity $e^{-(1+U_i)v} v^{-1} \mathrm{d}v \theta \tilde{p}_0^*(\mathrm{d}x)$.
- The jumps $J_{i,j,t}$ have density $f_{i,j,t}(v) \propto e^{-(1+U_i)v} v^{q_{i,j,t}-1}$, meaning that $J_{i,j,t} \overset{\text{ind}}{\sim} \mathrm{Ga}(q_{i,j,t},\, 1+U_i)$. This means that $\sum_{t=1}^{\ell_{i,j}} J_{i,j,t} \sim \mathrm{Ga}(n_{i,j},\, 1+U_i)$ whenever $n_{i,j} \geq 1$ and $G_{i,j} = 0$ a.s. if $n_{i,j} = 0$.

Again one has that

$$\tilde{p}_i \,|\, \mathbf{x}, \mathbf{t}, \tilde{p}_0^* \sim \mathfrak{D}\left(\theta \tilde{p}_0^* + \sum_{j=1}^k n_{i,j} \delta_{\mathrm{x}_{i,j}^*}\right)$$

for $i = 1, \ldots m$.

The sampling scheme described in Algorithm 3.1 specifies as follows.

3.3.2 Hierarchies of normalized stable processes

The *hierarchical stable NRMI* arises by setting

$$\rho(v) = \frac{\sigma v^{-1-\sigma}}{\Gamma(1-\sigma)} \quad \text{and } \rho_0(v) = \frac{\sigma_0 v^{-1-\sigma_0}}{\Gamma(1-\sigma_0)}$$

for some σ and σ_0 in $(0, 1)$. This means that $\tilde{p}_0$ is a σ_0-stable NRMI and, conditional on $\tilde{p}_0$, the $\tilde{p}_i$'s are independent and identically distributed as a σ-stable NRMI. We will refer to $(\tilde{p}_1, \ldots, \tilde{p}_m)$ as a *vector of hierarchical stable NRMIs* (HSP for short).

As we proved in the previous chapter, $\tau_{m,0}(u) = \frac{\sigma_0 \Gamma(m-\sigma_0)}{\Gamma(1-\sigma_0) u^{m-\sigma_0}}$ and $\tau_m(u) = \frac{\sigma \Gamma(m-\sigma)}{\Gamma(1-\sigma) u^{m-\sigma}}$. A plain application of Theorem 3.2 leads to

$$\mathrm{corr}(\tilde{p}_i(A), \tilde{p}_j(A)) = \frac{1-\sigma_0}{1-\sigma\sigma_0},$$

Algorithm 3.2: Sample $\left(\mathbf{x}_i^{(n_i)}\right)_{i=1}^m$ from the HDP.

for $i = 1, 2, \ldots, m$ **do**
  if $i = 1$ **then**
    Sample $\phi_1 \sim P_0$ and assign $\mathrm{x}_{1,1} = \mathrm{x}^*_{1,1} = \phi_1$
    Set $k = 1$, $\bar{\ell}_{1\bullet} = 1$, $\bar{\ell}_{\bullet 1} = 1$, $q_{1\bullet 1} = 1$
    else
      Take $\mathrm{x}_{i,1} = \mathrm{x}^*_{i,1}$, where $\mathrm{x}^*_{i,1}$ is sampled from G_{it}
      Set $\bar{\ell}_{i\bullet} = 1$ and $q_{i\bullet 1} = 1$
  for $j = 1, \ldots, n_i$ **do**
    Sample $\mathrm{x}_{i,j} \,|\, \mathrm{x}_{i,1}, \ldots, \mathrm{x}_{i,j-1}$ according to

$$\mathrm{x}_{i,j} \,|\, \mathrm{x}_{i,1}, \ldots, \mathrm{x}_{i,j-1} \sim \sum_{t=1}^{\bar{\ell}_{i\bullet}} \frac{q_{i\bullet t}}{\theta + t - 1} \delta_{\mathrm{x}^*_{i,t}}(\cdot) + \frac{\theta}{\theta + t - 1} G_{it}(\cdot)$$

$$G_{it}(\cdot) = \sum_{d=1}^{k} \frac{\bar{\ell}_{\bullet d}}{\theta_0 + |\boldsymbol{\ell}|} \delta_{\phi_d}(\cdot) + \frac{\theta_0}{\theta_0 + |\boldsymbol{\ell}|} P_0(\cdot)$$

for any measurable A and $i \neq j$. The correlation is increasing in σ and decreasing in σ_0, and the two limiting cases arise as follows: if $\sigma \uparrow 1$ then $\mathrm{corr}(\tilde{p}_i(A), \tilde{p}_j(A)) \uparrow 1$, whereas when $\sigma_0 \uparrow 1$ implies $\mathrm{corr}(\tilde{p}_i(A), \tilde{p}_j(A)) \downarrow 0$.

Recalling that $\Psi_0(u) = u^{\sigma_0}$ and $\Psi(u) = u^{\sigma}$, to determine the pEPPF note that

$$\begin{aligned}
\Phi^{(n_i)}_{\bar{\ell}_{i\bullet},i}(\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,k}) &= \frac{\theta^{\bar{\ell}_{i\bullet}}}{\Gamma(n_i)} \int_{\mathbb{R}_+} u^{n_i - 1} e^{-\theta u^{\sigma}} \prod_{j=1}^{k} \prod_{t=1}^{\ell_{i,j}} \frac{\sigma \Gamma(q_{i,j,t} - \sigma)}{\Gamma(1-\sigma) u^{q_{i,j,t} - \sigma}} \mathrm{d}u \\
&= \frac{\theta^{\bar{\ell}_{i\bullet}}}{\Gamma(n_i)} \sigma^{\bar{\ell}_{i\bullet}} \prod_{j=1}^{k} \prod_{t=1}^{\ell_{i,j}} (1-\sigma)_{q_{i,j,t}-1} \int_{\mathbb{R}_+} u^{\sigma \bar{\ell}_{i\bullet} - 1} e^{-\theta u^{\sigma}} \mathrm{d}u \\
&= \frac{\Gamma(\bar{\ell}_{i\bullet})}{\Gamma(n_i)} \sigma^{\bar{\ell}_{i\bullet} - 1} \prod_{j=1}^{k} \prod_{t=1}^{\ell_{i,j}} (1-\sigma)_{q_{i,j,t}-1}.
\end{aligned}$$

Since $\Phi^{(|\boldsymbol{\ell}|)}_{k,0}(\bar{\ell}_{\bullet 1}, \ldots, \bar{\ell}_{\bullet k}) = \frac{\sigma_0^{k-1}\Gamma(k)}{\Gamma(|\boldsymbol{\ell}|)} \prod_{j=1}^{k} (1-\sigma_0)_{\bar{\ell}_{\bullet j} - 1}$, the pEPPF equals

$$\begin{aligned}
\boldsymbol{\Pi}^{(N)}_k(\mathbf{n}_1, \ldots, \mathbf{n}_m) = \frac{\sigma_0^{k-1}\Gamma(k)}{\prod_{i=1}^{m} \Gamma(n_i)} \sum_{\boldsymbol{\ell}} \frac{\sigma^{|\boldsymbol{\ell}| - m} \prod_{i=1}^{m} \Gamma(\bar{\ell}_{i\bullet})}{\Gamma(|\boldsymbol{\ell}|)} \prod_{j=1}^{k} (1-\sigma_0)_{\bar{\ell}_{\bullet j} - 1} \\
\times \prod_{j=1}^{k} \prod_{i=1}^{m} \frac{1}{\ell_{i,j}!} \binom{n_{i,j}}{q_{i,j,1} \cdots, q_{i,j,\ell_{i,j}}} \sum_{\mathbf{q}} \prod_{t=1}^{\ell_{i,j}} (1-\sigma)_{q_{i,j,t}-1}
\end{aligned}$$

Therefore

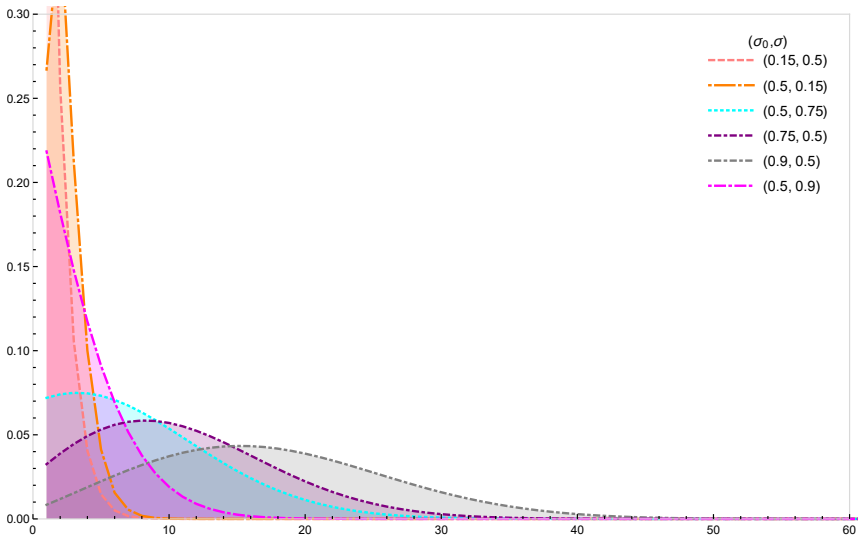
$$\boldsymbol{\Pi}^{(N)}_k(\mathbf{n}_1, \ldots, \mathbf{n}_m) = \frac{\sigma_0^{k-1}\Gamma(k)}{\prod_{i=1}^{m} \Gamma(n_i)} \sum_{\boldsymbol{\ell}} \frac{\sigma^{|\boldsymbol{\ell}| - m} \prod_{i=1}^{m} \Gamma(\bar{\ell}_{i\bullet})}{\Gamma(|\boldsymbol{\ell}|)} \prod_{j=1}^{k} (1-\sigma_0)_{\bar{\ell}_{\bullet j} - 1} \prod_{i=1}^{m} \prod_{j=1}^{k} S^{\sigma}_{n_{i,j}, \ell_{i,j}},$$

where $S^{\sigma}_{n,k}$ denotes the generalized Stirling number. Unsurprisingly, due to the properties of the stable CRM, neither the correlation coefficient nor the pEPPF depend on the total masses θ_0 and θ.

To study the distribution of K_N, note that for $k \in \{1, \dots, t\}$

$$\begin{aligned}
\mathbb{P}[\mathrm{K}_{0,t} = k] &= \frac{1}{k!} \sum_{(r_1,\dots,r_k)\in\mathcal{C}^k_{[t]}} \binom{t}{r_1 \cdots r_k} \frac{\sigma_0^{k-1}\Gamma(k)}{\Gamma(t)} \prod_{j=1}^{k} (1-\sigma_0)_{r_j-1} \\
&= \frac{\sigma_0^{k-1}\Gamma(k)}{\Gamma(t)} \sum_{(r_1,\dots,r_k)\in\mathcal{C}^k_{[t]}} \binom{t}{r_1 \cdots r_k} \frac{1}{k!} \prod_{j=1}^{k} (1-\sigma_0)_{r_j-1} \\
&= \frac{\sigma_0^{k-1}\Gamma(k)}{\Gamma(t)} S^{\sigma_0}_{t,k},
\end{aligned}$$

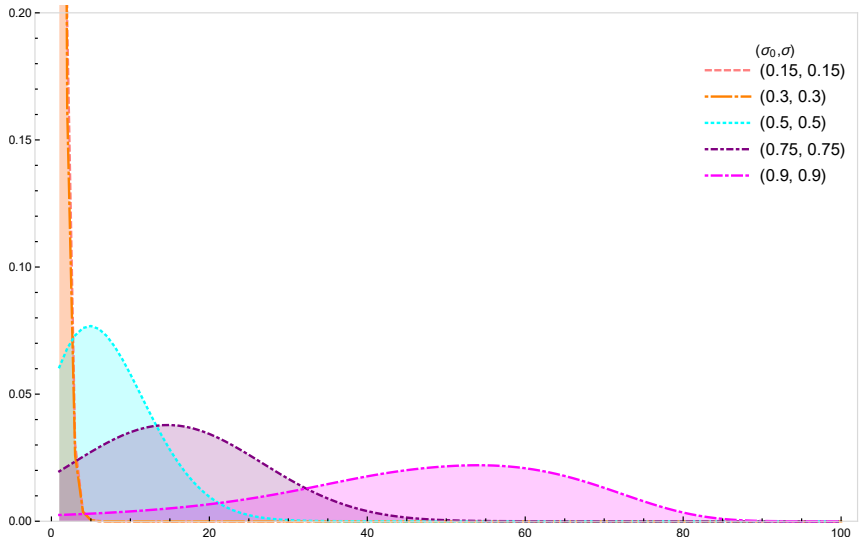
and similarly for $\zeta \in \{1, \dots, n_i\}$, one has that $\mathbb{P}[\mathrm{K}_{i,n_i} = \zeta] = \frac{\sigma^{k-1}\Gamma(\zeta)}{\Gamma(n_i)} S^{\sigma}_{n_i,\zeta}$. Figure 3.12 shows the distribution of K_N for two partially exchangeable sequences, with $n_1 = 50 = n_2$ and different choices of σ_0, σ. It is easy to see from Figure 3.12a that the parameters control the kurtosis: smaller values of either σ or σ_0 correspond to leptokurtic distributions whereas as either one of them grows, the curves become more platykurtic. The same behavior is observed in Figure 3.12b, where $\sigma_0 = \sigma$: larger parameters produce flatter curves as opposed to small parameters.



(a) Distribution of K_{100} *for the HSP with* $\sigma_0 \neq \sigma$.

For the posterior characterization, one has that

$$\begin{aligned}
f_0(u \,|\, \mathbf{x}, \mathbf{t}) &\propto u^{|\boldsymbol{\ell}|-1} e^{-\theta_0 u^{\sigma_0}} \prod_{j=1}^{k} \sigma_0 (1-\sigma_0)_{\bar{\ell}_{\bullet j}-1} u^{\sigma_0 - \bar{\ell}_{\bullet j}} \\
&= u^{\sigma_0 k - 1} \sigma_0^k \prod_{j=1}^{k} (1-\sigma_0)_{\bar{\ell}_{\bullet j}-1},
\end{aligned}$$



(b) Distribution of K_{100} *for the HSP with* $\sigma_0 = \sigma$.

Figure 3.12: Distribution of K_{100} *for the HSP for different choices of* σ_0 *and* σ.

so that $U_0 \sim \mathrm{Ga}(k, \theta_0)$. Hence the distribution of U_0 depends on the observations only through the number of distinct values k. Moreover

- η_0^* is a generalized Gamma CRM as in Example 1.5, whose intensity is

$$\frac{\sigma_0}{\Gamma(1-\sigma_0)} \frac{u^{-U_0 v}}{v^{\sigma_0+1}} \mathrm{d}v \theta_0 P_0(\mathrm{d}x)$$

- The jump random variables are independently distributed with density $f_j(v \,|\, \mathbf{x}, \mathbf{t}) \propto v^{\bar{\ell}_{\bullet j} - \sigma_0 - 1} e^{-vU_0}$, that is $I_j \overset{\text{ind}}{\sim} \mathrm{Ga}(\bar{\ell}_{\bullet j} - \sigma_0, U_0)$.

Thus $\tilde{p}_0^* = \frac{\eta_0^* + \sum_{j=1}^k I_j \delta_{\mathrm{x}_j^*}}{\eta_0^*(\mathbb{X}) + \sum_{j=1}^k I_j}$ and, conditional on $\tilde{p}_0^*$ and $\mathbf{x}, \mathbf{t}, \mathbf{U}$, the completely random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_m$ are independent. Each $\tilde{\mu}_i$ distributes as $\tilde{\mu}_i^* + \sum_{j=1}^{k_i} \sum_{t=1}^{\ell_{i,j}} J_{i,j,t} \delta_{\mathrm{x}_{i,j}^*}$, where

- $\tilde{\mu}_i^*$ is a generalized Gamma CRM whose intensity is

$$\frac{\sigma}{\Gamma(1-\sigma)} \frac{e^{-U_i v}}{v^{\sigma+1}} \mathrm{d}v \theta \tilde{p}_0^*(\mathrm{d}x).$$

- The jump random variables $J_{i,j,t}$ are independent and each with density $f_{i,j,t}(v) \propto e^{-U_i v} v^{q_{i,j,t} - \sigma - 1}$, i.e. $J_{i,j,t} \overset{\text{ind}}{\sim} \mathrm{Ga}(q_{i,j,t} - \sigma, U_i)$ for $t = 1, \ldots, \ell_{i,j}$. This implies that $\sum_{t=1}^{\ell_{i,j}} J_{i,j,t} \sim \mathrm{Ga}(n_{i,j} - \sigma \ell_{i,j}, U_i)$ if $n_{i,j} \geq 1$ and $\sum_{t=1}^{\ell_{i,j}} J_{i,j,t} = 0$ a.s. if $n_{i,j} = 0$.

A sampling scheme is presented in Algorithm 3.3.

Algorithm 3.3: Sample $\left(\mathbf{x}_i^{(n_i)}\right)_{i=1}^m$ from the HSP.

for *iteration* $i = 1, 2, \ldots, m$ **do**
  if $i = 1$ **then**
    Sample $\phi_1 \sim P_0$ and assign $\mathrm{x}_{1,1} = \mathrm{x}^*_{1,1} = \phi_1$
    Set $k = 1$, $\bar{\ell}_{1\bullet} = 1$, $\bar{\ell}_{\bullet 1} = 1$, $q_{1\bullet 1} = 1$
    else
      Take $\mathrm{x}_{i,1} = \mathrm{x}^*_{i,1}$, where $\mathrm{x}^*_{i,1}$ is sampled from G_{it}
      Set $\bar{\ell}_{i\bullet} = 1$ and $q_{i\bullet 1} = 1$
  for $j = 1, \ldots, n_i$ **do**
    Sample $\mathrm{x}_{i,j} \,|\, \mathrm{x}_{i,1}, \ldots, \mathrm{x}_{i,j-1}$ according to

$$\mathrm{x}_{i,j} \,|\, \mathrm{x}_{i,1}, \ldots, \mathrm{x}_{i,j-1} \sim \sum_{t=1}^{\bar{\ell}_{i\bullet}} \frac{q_{i\bullet t} - \sigma}{t-1} \delta_{\mathrm{x}^*_{i,t}}(\cdot) + \frac{\bar{\ell}_{i\bullet}\sigma}{t-1} G_{it}(\cdot)$$

$$G_{it}(\cdot) = \sum_{d=1}^{k} \frac{\bar{\ell}_{\bullet d} - \sigma_0}{|\boldsymbol{\ell}| + \theta_0} \delta_{\phi_d}(\cdot) + \frac{k\sigma_0}{|\boldsymbol{\ell}| + \theta_0} P_0(\cdot)$$

3.4 Hierarchies of Pitman-Yor processes

Recall that the Pitman-Yor process is obtained by normalizing the random measure $\tilde{\mu}_{\sigma,\theta}$, whose distribution of $P_{\sigma,\theta}$ satisfies the relationship

$$\frac{\mathrm{d}P_{\sigma,\theta}(m)}{\mathrm{d}P_\sigma} = \frac{\Gamma(\theta+1)}{\Gamma\left(\frac{\theta}{\sigma}+1\right)} m(\mathbb{X})^{-\theta} \quad \text{for } m \in \mathrm{M}_{\mathbb{X}}, \tag{3.8}$$

where P_σ is the distribution of a positive σ-stable random variable.

Suppose that $\mathrm{x}_{i,j}$ are partially exchangeable as in (3.2), with Q_m characterized by

$$\begin{aligned} \tilde{p}_i \,|\, \tilde{p}_0 &\overset{\text{i.i.d.}}{\sim} \mathcal{PY}(\sigma, \theta, \tilde{p}_0) \quad i = 1, \ldots, m \\ \tilde{p}_0 &\sim \mathcal{PY}(\sigma_0, \theta_0, P_0), \end{aligned} \tag{3.9}$$

where $\sigma, \sigma_0 \in (0,1)$, $\theta > -\sigma$, $\theta_0 > -\sigma_0$ and P_0 nonatomic. Each $\tilde{p}_i$ is then the normalization of a measure $\tilde{\mu}_i$ that is not completely random, and whose law is absolutely continuous with respect to the law of a σ-stable CRM. The above results presented for hierarchical NRMIs can be extended to hierarchies of Pitman-Yor processes by taking into account the change of measure (3.8) and working directly with the CRMs. We will refer to $(\tilde{p}_1, \ldots, \tilde{p}_m)$ as a *vector of hierarchical Pitman-Yor processes* (HPYP).

Theorem 3.7. *Suppose that* $\tilde{p}_i \,|\, \tilde{p}_0 \overset{i.i.d.}{\sim} \mathcal{PY}(\sigma, \theta, \tilde{p}_0)$, *for* $i = 1, \ldots, m$, *and that* $\tilde{p}_0 \sim \mathcal{PY}(\sigma_0, \theta_0, P_0)$ *with* P_0 *nonatomic,* $\sigma, \sigma_0 \in (0,1)$ *and* $\theta > -\sigma$, $\theta_0 > -\sigma_0$. *Then, for any* $A \in \mathcal{X}$ *and* $i \neq j \in \{1, \ldots, m\}$,

$$corr\left(\tilde{p}_i(A), \tilde{p}_j(A)\right) = \left\{1 + \frac{1-\sigma}{1-\sigma_0} \frac{\theta_0 + \sigma_0}{\theta + 1}\right\}^{-1}.$$

Theorem 3.8. *Let* $\{(\mathrm{x}_{i,j})_{j\geq 1} : i = 1\ldots, m\}$ *be partially exchangeable as in* (3.9)*, and suppose that we have sampled* $\left(\mathbf{x}_i^{(n_i)}\right)_{i=1}^m$*. Then*

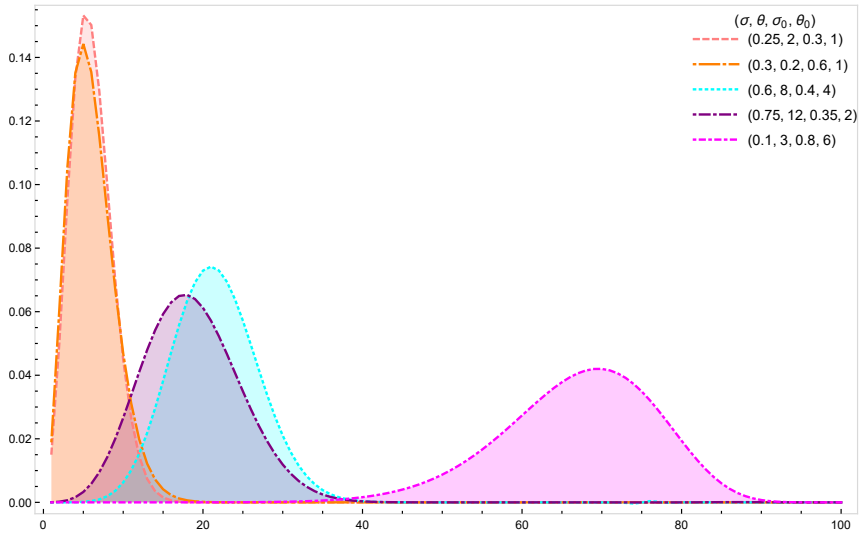
$$\begin{aligned}\boldsymbol{\Pi}_k^{(N)}(\mathbf{n}_1,\ldots,\mathbf{n}_m) = \sum_{\boldsymbol{\ell}} \frac{\prod_{r=1}^{k-1}(\theta_0 + r\sigma_0)}{(\theta_0+1)_{|\boldsymbol{\ell}|-1}} \prod_{j=1}^{k}(1-\sigma_0)_{\bar{\ell}_{\bullet j}-1} \\ \times \prod_{i=1}^{m} \frac{\prod_{r=1}^{\bar{\ell}_{i\bullet}-1}(\theta+r\sigma)}{(\theta+1)_{n_i-1}} \prod_{j=1}^{k} S^{\sigma}_{n_{i,j},\ell_{i,j}}.\end{aligned} \tag{3.10}$$

A closed expression for the distribution of the number of blocks holds.

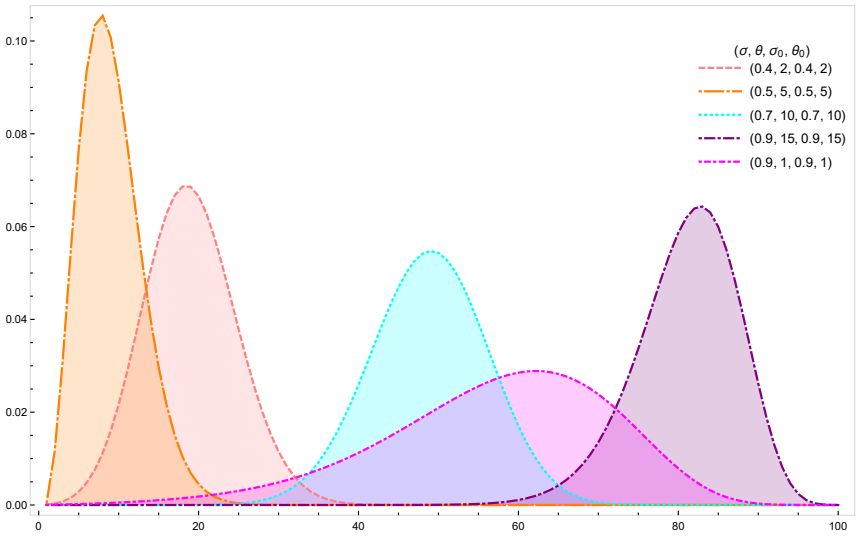
Theorem 3.9. *Suppose that* K_N *is the number of distinct values in the* m *partially exchangeable samples* $\mathbf{x}$*, governed by a vector of hierarchical Pitman-Yor processes. Then*

$$\begin{aligned}\mathbb{P}[\mathrm{K}_N = k] = \sum_{t=k}^{N} \frac{\prod_{r=1}^{k-1}(\theta_0 + r\sigma_0)}{(\theta_0+1)_{t-1}} S^{\sigma_0}_{t,k} \\ \times \sum_{(\zeta_1,\ldots,\zeta_m)\in\mathcal{C}^m_{[t]}} \prod_{i=1}^{m} \frac{\prod_{r=1}^{\zeta_i-1}(\theta+r\sigma)}{(\theta+1)_{n_i-1}} S^{\sigma}_{n_i,\zeta_i}.\end{aligned}$$

Figure 3.13 shows plots of the distribution of K_N for two partially exchangeable sequences with $n_1 = 100 = n_2$. As expected, the hierarchical Pitman-Yor process allows more flexibility than the hierarchical Dirichlet and stable processes, even in the homogeneous case where $(\sigma_0, \theta_0) = (\sigma, \theta)$, as the distribution of K_N can take a variety of forms depending on the choice of the parameters. Changes in the parameters σ and σ_0 affect mainly the dispersion.



(a) Distribution of K_{100} *for the HPYP with* $\sigma_0 \neq \sigma$ *and* $\theta \neq \theta_0$*.*



(b) Distribution of K_{100} *for the HPYP with* $\sigma_0 = \sigma$ *and* $\theta_0 = \theta$.

Figure 3.13: Distribution of K_{100} *for the HPYP for different choices of* (σ_0, θ_0) *and* (σ, θ).

To characterize the posterior distribution, let U_0 be a positive random variable with density with respect to the Lebesgue measure such that, conditionally on $\mathbf{x}$ and on the latent tables $\mathbf{t}$, is given by

$$f_0(u \,|\, \mathbf{x}, \mathbf{t}) = \frac{\sigma_0}{\Gamma\left(k + \frac{\theta_0}{\sigma_0}\right)} u^{k\sigma_0 + \theta_0 - 1} e^{-u\sigma_0}.$$

Theorem 3.10. *Assume that the data* $\mathbf{x}$ *are partially exchangeable and modeled as in* (3.9). *Then*

$$\tilde{\mu}_0 \,|\, \mathbf{x}, \mathbf{t}, U_0 \sim \eta_0^* + \sum_{j=1}^{k} I_j \delta_{\mathrm{x}_j^*}, \tag{3.11}$$

where η_0^* *and* $\sum_{j=1}^{k} I_j \delta_{\mathrm{x}_j^*}$ *are independent and*

- η_0^* *is a generalized Gamma CRM with intensity*

$$e^{-vU_0} \frac{1}{v^{1+\sigma_0}} \frac{\sigma_0}{\Gamma(1-\sigma_0)} \mathrm{d}v P_0(\mathrm{d}x).$$

- *The jump random variables* $(I_j)_{j=1}^{k}$ *are independent, nonnegative and* $I_j \sim \mathrm{Ga}(\ell^-_{\bullet j} - \sigma_0, U_0)$.

To characterize the posterior distribution of $(\tilde{\mu}_1, \ldots, \tilde{\mu}_m)$, let $\mathbf{U} = (U_1, \ldots, U_m)$, where the components are independent and admit, conditional on $\mathbf{x}, \mathbf{t}$, a density with respect to the Lebesgue measure given by

$$f_i(u \,|\, \mathbf{x}, \mathbf{t}) = \frac{\sigma}{\Gamma\left(k_i + \frac{\theta}{\sigma}\right)} u^{\sigma k_i + \theta - 1} e^{-u^\sigma} \quad i = 1, \ldots, m.$$

Theorem 3.11. *Assume that we have sampled* $\mathbf{x}$ *from a partially exchangeable sequence modeled as in* (3.9)*. The posterior distribution of* $(\tilde{\mu}_1, \ldots, \tilde{\mu}_m)$*, given the observations, the latent tables and* $\tilde{p}_0$ *coincides with*

$$(\tilde{\mu}_1, \ldots, \tilde{\mu}_m) \,|\, (\mathbf{x}, \mathbf{t}, \mathbf{U}, \tilde{\mu}_0) \sim (\tilde{\mu}_1^*, \ldots, \tilde{\mu}_m^*) + \left(\sum_{j=1}^{k} \sum_{t=1}^{\ell_{1,j}} J_{1,j,t} \delta_{\mathrm{x}_j^*}, \ldots, \sum_{j=1}^{k} \sum_{t=1}^{\ell_{d,j}} J_{d,j,t} \delta_{\mathrm{x}_j^*} \right),$$

where the two summands on the right hand side are independent, $\sum_{t=1}^{\ell_{i,j}} J_{i,j,t} \equiv 0$ *if* $n_{i,j} = 0$ *and*

- $(\tilde{\mu}_1^*, \ldots, \tilde{\mu}_m^*)$ *is a vector of hierarchical CRMs such that, conditional on* $\tilde{\mu}_0^* = \eta_0^* + \sum_{j=1}^{k} I_j \delta_{X_j^*}$ *as in* (3.11)*, each* $\tilde{\mu}_i^*$ *is generalized Gamma CRM with intensity*

$$e^{-vU_i} \frac{1}{v^{1+\sigma}} \frac{\sigma}{\Gamma(1-\sigma)} \mathrm{d}v \tilde{p}_0(\mathrm{d}x).$$

- *The jump random variables* $\{(J_{i,j,t})\}$ *are independent and nonnegative, where each* $J_{i,j,t} \sim \mathrm{Ga}(q_{i,j,t} - \sigma, U_i)$ *when* $n_{i,j} \geq 1$*, whereas* $J_{i,j,t} = 0$ *a.s. if* $n_{i,j} = 0$*. In particular* $\sum_{t=1}^{\ell_{i,j}} J_{i,j,t} \sim \mathrm{Ga}(n_{i,j} - \ell_{i,j}\sigma, U_i)$*.*

Proofs of all of the above theorems are attached in Appendix B. Finally, the sampling scheme of Algorithm 3.1 specializes as follows.

Algorithm 3.4: Sample $\left(\mathbf{x}_i^{(n_i)}\right)_{i=1}^{m}$ from the HPYP.

for $i = 1, 2, \ldots, m$ **do**
  if $i = 1$ **then**
    Sample $\phi_1 \sim P_0$ and assign $\mathrm{x}_{1,1} = \mathrm{x}_{1,1}^* = \phi_1$
    Set $k = 1$, $\bar{\ell}_{1\bullet} = 1$, $\bar{\ell}_{\bullet 1} = 1$, $q_{1\bullet 1} = 1$
    else
      Take $\mathrm{x}_{i,1} = \mathrm{x}_{i,1}^*$, where $\mathrm{x}_{i,1}^*$ is sampled from G_{it}
      Set $\bar{\ell}_{i\bullet} = 1$ and $q_{i\bullet 1} = 1$
  for $j = 1, \ldots, n_i$ **do**
    Sample $\mathrm{x}_{i,j} \,|\, \mathrm{x}_{i,1}, \ldots, \mathrm{x}_{i,j-1}$ according to

$$\mathrm{x}_{i,j} \,|\, \mathrm{x}_{i,1}, \ldots, \mathrm{x}_{i,j-1} \sim \sum_{t=1}^{\bar{\ell}_{i\bullet}} \frac{q_{i\bullet t} - \sigma}{\theta + t - 1} \delta_{X_{i,c}^*}(\cdot) + \frac{\theta + \bar{\ell}_{i\bullet}\sigma}{\theta + t - 1} G_{it}(\,\cdot\,)$$

$$G_{it}(\,\cdot\,) = \sum_{d=1}^{k} \frac{\bar{\ell}_{\bullet d} - \sigma_0}{|\boldsymbol{\ell}| + \theta_0} \delta_{\phi_d}(\cdot) + \frac{\theta_0 + k\sigma_0}{|\boldsymbol{\ell}| + \theta_0} P_0(\,\cdot\,)$$

Algorithms

The purpose of this chapter is to illustrate the usefulness of Bayesian nonparametric methods based on hierarchical processes. As mentioned before, hierarchical processes are useful in problems in which there are multiple groups of data somehow related; here this relationship will take the form of mixture components that are shared across m sets of data. The borrowing of strength phenomenon should then allow us to model these densities jointly rather than as m separate mixtures. We rely mostly on the results displayed in Section 3.2.4 and, by means of Markov chain Monte Carlo methods, we describe an algorithm for hierarchical mixtures models using some of the prior processes described in Chapter 3. The first thing we do in Section 4.1 is a brief review on mixture and infinite mixture models. In here we describe some MCMC methods for adjusting nonparametric mixture models and define the log-pseudo marginal likelihood that will be used as a measure of the goodness of fit. In Section 4.2 we generalize this setting to adjust nonparametric mixtures to partially exchangeable data. Finally, in Section 4.3 we will perform two small experiments to compare the performance of several hierarchical prior processes. In order to avoid the proliferation of the symbols, from now on we shall use the same letters to denote both the random variables and their realizations.

4.1 Infinite mixtures

Suppose that we are interested in modelling data $\{\mathrm{y}_1, \ldots, \mathrm{y}_n\}$ that presents no repetitions. The a.s. discrete nature of NRMIs makes them unsuitable to model continuous distributions directly, however, it is possible to define nonparametric priors whose realizations yield a.s. probability distributions that admit a density with respect to some reference measure ν acting over the space $\mathbb{X}$ on which we are interested. Namely, for density estimation purposes, it is custom in Bayesian nonparametric statistics the use of random mixture modelling, oftentimes called *infinite mixtures*.

First let us recall that if $\{\mathrm{y}_1, \ldots, \mathrm{y}_n\}$ is modeled as independent draws from a *mixture* distribution function with a fixed number of components $M > 1$, then the distribution of each y_i is a convex combination of components of the form

$$\mathrm{y}_i \sim \sum_{j=1}^{M} \eta_j \mathcal{K}(\cdot \,|\, \mathrm{x}_j), \tag{4.1}$$

where $\mathcal{K}(\cdot,\cdot) : \Theta \times \mathcal{X} \to [0,1]$ is a diffuse probability kernel and we have adopted the notation $\mathcal{K}(\mathrm{x},\cdot) = \mathcal{K}(\cdot \,|\, \mathrm{x})$. The constants $(\eta_1,\ldots,\eta_M)$ are the mixture proportions, which are constraint to satisfy

$$\eta_j \geq 0 \quad j = 1,\ldots,M \quad \text{and} \sum_{j=1}^{M} \eta_j = 1.$$

Usually one considers $\{\mathcal{K}(\cdot \,|\, \mathrm{x}) : \mathrm{x} \in \Theta\}$ to be a parametric family, with Θ being its parameter space. Note that model (4.1) can be alternatively expressed in an integral form

$$\mathrm{y}_i \sim \int \mathcal{K}(\cdot \,|\, \mathrm{x}) \tilde{\mu}(\mathrm{dx}), \tag{4.2}$$

where $\tilde{\mu}$ is a discrete measure that places probability mass η_j on the atom x_j for $j = 1,\ldots,M$. In this setting, the measure $\tilde{\mu}$ is referred to as the *mixing measure* or the *mixing distribution*.

An alternative starting point for setting up a mixture model is to introduce a collection of *latent allocation random variables* $(\mathrm{z}_i)_{i=1}^n$ that record the information about which component of the mixture was y_i sampled from. This means that $\mathrm{z}_i = j$ if y_i is drawn from $\mathcal{K}(\cdot \,|\, \mathrm{x}_j)$. Suppose the population from which we are sampling consists of M different groups, each present in the population in proportion η_j, $j = 1,\ldots,M$. Whenever we are sampling from group j, observations are assumed drawn from $\mathcal{K}(\cdot \,|\, \mathrm{x}_j)$, hence we can imagine that a single observation y_i arises in two steps: first, the mixture component z_i is drawn according to $\mathbb{P}[\mathrm{z}_i = j] = \eta_j$ and secondly, given z_i, y_i is drawn from $\mathcal{K}(\cdot \,|\, \mathrm{x}_{\mathrm{z}_i})$. Thus the data-generating mechanism can be expressed as

$$\mathrm{z}_i \,|\, \boldsymbol{\eta} \sim \text{Categorical}(\boldsymbol{\eta}), \quad \mathrm{y}_i \,|\, \mathrm{z}_i \overset{\text{ind}}{\sim} \mathcal{K}(\cdot \,|\, \mathrm{x}_{\mathrm{z}_i}), \tag{4.3}$$

where $\boldsymbol{\eta} = (\eta_1,\ldots,\eta_M)$. Model (4.1) is obtained by marginalizing out the latent allocation variables in (4.3). In a Bayesian setting, the formulation of a mixture model is completed with a prior distribution on the mixing proportions $\mathrm{p}(\boldsymbol{\eta})$ and on the unknown parameters $(\mathrm{x}_1,\ldots,\mathrm{x}_M)$. Written in a hierarchical way, this entails

$$\begin{aligned}
\mathrm{y}_i \,|\, \mathrm{z}_i &\overset{\text{ind}}{\sim} \mathcal{K}(\cdot \,|\, \mathrm{x}_{\mathrm{z}_i}) \\
\mathrm{z}_1,\ldots,\mathrm{z}_n \,|\, \boldsymbol{\eta} &\overset{\text{i.i.d.}}{\sim} \text{Categorical}(\boldsymbol{\eta}) \\
(\mathrm{x}_j)_{j=1}^M &\overset{\text{i.i.d.}}{\sim} P_0 \\
\eta_1,\ldots,\eta_M \,|\, M &\sim \pi_M,
\end{aligned}$$

where P_0 is a diffuse probability measure and π_M is a probability measure. In principle, one could additionally assign a prior over the number of components M.

Now let us assume that $\tilde{\mu} = \sum_{j=1}^{\infty} \eta_j \delta_{\mathrm{x}_j}$ is a discrete random probability measure, where $(\eta_j)_{j\geq 1}$ is a sequence of nonnegative random weights such that $\sum_{j=1}^{\infty} \eta_j = 1$ a.s. and $(\mathrm{x}_j)_{j\geq 1}$ is a sequence of $\mathbb{X}$-valued random locations independent of $(\eta_j)_{j\geq 1}$. In this context, the mixture model at (4.2) becomes

$$\mathrm{y}_i \,|\, \tilde{\mu} \sim \int \mathcal{K}(\mathrm{y}_i \,|\, \mathrm{x}) d\tilde{\mu}(\mathrm{x}) = \sum_{j=1}^{\infty} \eta_j \mathcal{K}(\mathrm{y}_i \,|\, \mathrm{x}_j).$$

The above expression resembles model (4.1), the main difference being that the number of mixture components M is set to infinity. Such models are termed *infinite mixtures* and are often called *nonparametric mixtures*. Accordingly, in a Bayesian nonparametric setting it becomes natural to consider a nonparametric prior for the unknown mixing measure $\tilde{\mu}$, so that the data generating mechanism of a sample $\{\mathrm{y}_1, \ldots, \mathrm{y}_n\}$ is modeled as

$$\mathrm{y}_i \,|\, \mathrm{x}_i \overset{\text{ind}}{\sim} \mathcal{K}(\cdot \,|\, \mathrm{x}_i), \quad \mathrm{x}_i \,|\, \tilde{\mu} \overset{\text{i.i.d.}}{\sim} \tilde{\mu}, \quad \tilde{\mu} \sim \mathrm{Q}, \tag{4.4}$$

where Q is the law of the random probability measure $\tilde{\mu}$. This means that now the probabilities of the categorical distribution on $\boldsymbol{\eta}$ in (4.3) are replaced with $(\eta_j)_{j\geq 1}$, which is a sequence of random probabilities and the parameters $(\mathrm{x}_j)_{j\geq 1}$ are i.i.d., distributed according to the base measure of $\tilde{\mu}$. Alternatively, the allocation variables z_i are no longer drawn from a categorical distribution, but from a random probability measure. Intuitively this would translate onto assigning a prior distribution over M as an indicator function, centered at $\{\infty\}$. It is important to make the clarification that assuming that the number of components present in the population model is unlimited in no way implies that infinitely many components are occupied by a sample. Rather only a finite (but varying) number of components will be used, as each data item is associated with exactly one component but each component can be associated with multiple data items. In general there will be empty components x_r for which $\mathrm{z}_i \neq r$ for all $i = 1, \ldots, n$. Adopting these kinds of models avoids difficulties related with choosing the number of components M, as inference in infinite mixture models automatically recovers both the unknown number of components to use and the parameters of the components. This provides flexibility in the sense that we can freely introduce new mixture components as data arrives.

Example 4.1. Consider P_0 a diffuse probability measure and θ the precision parameter of a Dirichlet process. A *Dirichlet process mixture* with mixing kernel $\mathcal{K}$, introduced by Lo (1984), is defined as

$$\mathrm{y}_i \,|\, \mathrm{x}_i \sim \mathcal{K}(\cdot \,|\, \mathrm{x}_i), \quad \mathrm{x}_i \,|\, \tilde{\mu} \sim \tilde{\mu}, \quad \tilde{\mu} \sim \mathfrak{D}(\theta, P_0).$$

For a throughout account on methods and implementation on mixture models, refer to Frühwirth-Schnatter et al. (2018).

4.1.1 Gibbs samplers

A key question arising on infinite mixtures is how to conduct the posterior computation

$$\mathrm{p}(\mathbf{x} \,|\, \mathbf{y}) \propto \mathrm{p}(\mathbf{x})\mathrm{p}(\mathbf{y} \,|\, \mathbf{x}),$$

as this initially seems problematic in that the mixing measure is characterized by infinitely many parameters. However, one can recur to a Markov Chain Monte Carlo method (MCMC), such as the Gibbs sampler, to draw samples from $\mathrm{p}(\mathbf{x} \,|\, \mathbf{y})$. For a complete review on MCMC methods for nonparametric mixtures, refer to Favaro and Teh (2013).

Definition 4.1. Let $\mathbf{x} \in \mathbb{R}^n$, with $n > 1$, with joint density function $\mathrm{p}(\mathbf{x})$. If we set $\mathbf{x}^{-i} = \mathbf{x} \setminus \{\mathrm{x}_i\}$, then the *full conditional density* of x_i given $\mathbf{x}^{-i}$ equals

$$\mathrm{p}(\mathrm{x}_i \,|\, \mathbf{x}^{-i}) = \frac{\mathrm{p}(\mathrm{x}_i, \mathbf{x}^{-i})}{\mathrm{p}(\mathbf{x}^{-i})} = \frac{\mathrm{p}(\mathbf{x})}{\int \mathrm{p}(\mathbf{x})\mathrm{d}\mathrm{x}_i}.$$

If these full conditionals are easy to sample from, one can define a Markov Chain whose transition kernel is based on $\{\mathrm{p}(\mathrm{x}_i \,|\, \mathbf{x}^{-i})\}_{i=1}^n$, that has p as the stationary distribution and equals

$$K(\mathbf{x}, \mathbf{w}) = \mathrm{p}(\mathrm{w}_1 \,|\, \mathrm{x}_1, \ldots, \mathrm{x}_n)\mathrm{p}(\mathrm{w}_2 \,|\, \mathrm{w}_1, \mathrm{x}_3, \ldots, \mathrm{x}_n) \cdots \mathrm{p}(\mathrm{w}_n \,|\, \mathrm{w}_1, \ldots, \mathrm{x}_{n-1}) \tag{4.5}$$

for $\mathbf{w}, \mathbf{x} \in \mathbb{R}^n$. The steps of the Gibbs sampler can be summarized follows.

- Initialize $(\mathrm{x}_1^{(0)}, \ldots, \mathrm{x}_n^{(0)})$.
- For, $t \geq 1$ sample $(\mathrm{x}_1^{(t)}, \ldots, \mathrm{x}_n^{(t)})$ from $\mathbf{y}^{(t-1)}$ according to the transition kernel (4.5)

$$\begin{aligned}
\mathrm{x}_i^{(t)} &\sim \mathrm{p}(\mathrm{x}_i \,|\, \mathrm{x}_2^{(t-1)}, \mathrm{x}_3^{(t-1)}, \ldots, \mathrm{x}_n^{(t-1)}) \\
\mathrm{x}_2^{(t)} &\sim \mathrm{p}(\mathrm{x}_2 \,|\, \mathrm{x}_i^{(t)}, \mathrm{x}_3^{(t-1)}, \ldots, \mathrm{x}_n^{(t-1)}) \\
&\vdots \\
\mathrm{x}_{n-1}^{(t)} &\sim \mathrm{p}(\mathrm{x}_{n-1} \,|\, \mathrm{x}_i^{(t)}, \mathrm{x}_2^{(t)}, \ldots, \mathrm{x}_n^{(t-1)}) \\
\mathrm{x}_n^{(t)} &\sim \mathrm{p}(\mathrm{x}_n \,|\, \mathrm{x}_i^{(t)}, \mathrm{x}_2^{(t)}, \ldots, \mathrm{x}_{n-1}^{(t)}).
\end{aligned}$$

Going back to our case of interest, suppose that the EPPF of the partition induced by the prior process is known. The exchangeability of $(\mathrm{x}_i)_{i \geq 1}$ and the fact that the labels of the components are completely arbitrary allow us to treat each x_i as the last one being sampled. Using the predictive distribution whose weights are given by $\omega_0^{(n)}$ and $\omega_j^{(n)}$ as in (2.4) and (2.5), one has that

$$\mathrm{x}_i \,|\, \mathbf{x}^{-i} \sim \omega_0^{(n)}(\mathrm{n}_1^{-i}, \ldots, \mathrm{n}_k^{-i})P_0 + \sum_{j=1}^{k^{-i}} \omega_j^{(n)}(\mathrm{n}_1^{-i}, \ldots, \mathrm{n}_k^{-i})\delta_{\mathrm{x}_j^*},$$

where $\{\mathrm{x}_j^*\}_{j=1}^{k^{-i}}$ are the k^{-i} unique values displayed on $\mathbf{x}^{-i}$ and n_j^{-i} its corresponding frequencies. This prediction rule allows us to compute the full conditional needed in the Gibbs sampler, by combining the previous expression with the sampling distribution

$$\begin{aligned}\mathrm{p}(\mathrm{x}_i \,|\, \mathbf{y}, \mathbf{x}^{-i}) &\propto \mathrm{p}(\mathbf{y} \,|\, \mathbf{x})\mathrm{p}(\mathrm{x}_i \,|\, \mathbf{x}^{-i}) \\ &\propto \omega_0^{(n)} P_0(\mathrm{x}_i)\mathcal{K}(\mathrm{y}_i \,|\, \mathrm{x}_i) + \sum_{j=1}^{k^{-i}} \omega_j^{(n)} \mathcal{K}(\mathrm{y}_i \,|\, \mathrm{x}_j^*)\delta_{\mathrm{x}_j^*}(\mathrm{x}_i) \\ &\propto q_0 \frac{P_0(\mathrm{y}_i)\mathcal{K}(\mathrm{y}_i \,|\, \mathrm{x}_i)}{\int \mathcal{K}(\mathrm{y}_i \,|\, x) P_0(\mathrm{d}x)} + \sum_{j=1}^{k^{-i}} q_j \delta_{\mathrm{x}_j^*}(\mathrm{x}_i), \end{aligned} \tag{4.6}$$

where

$$q_0 = \omega_0^{(n)} \int \mathcal{K}(\mathrm{y}_i \,|\, x) P_0(\mathrm{z}x), \quad q_j = \omega_j^{(n)} \mathcal{K}(\mathrm{y}_i \,|\, \mathrm{x}_j^*)$$

This provides an easy algorithm for the first T iterations of the Gibbs sampler.

Algorithm 4.1: T iterations of a Gibbs sampler by direct assignment.

Initialize $(\mathrm{x}_1^{(0)}, \ldots, \mathrm{x}_n^{(0)})$.
for *iteration* $t = 1, \ldots, T$ **do**
  for $i = 1, \ldots, n$ **do**
    Set $\mathrm{x}_i^{(t)} = \mathrm{x}_i$, where x_i is sampled from (4.6)

The simplest situation occurs when the base distribution P_0 is conjugate to the mixture kernel $\mathcal{K}$, as in this case the integral $\int \mathcal{K}(\mathrm{y}_i \,|\, x) P_0(\mathrm{d}x)$ can be calculated analytically.

Example 4.2. Recall that for a Dirichlet process with precision parameter θ and base measure P_0, the weights of the predictive distribution coincide with $\omega_0^{(n)} = \frac{\theta}{\theta+n-1}$ and $\omega_j^{(n)} = \frac{n_j}{\theta+n-1}$. Escobar (1994) proposed the first posterior Gibbs sampler for the Dirichlet Process mixture, based on transition probabilities that update x_i by draws from the complete conditional posterior $\mathrm{p}(\mathrm{x}_i \,|\, \mathbf{y}, \mathbf{x}^{-i})$ as in (4.6), where

$$q_j = \frac{n_j}{\theta + n - 1} \mathcal{K}(\mathrm{y}_i \,|\, \mathrm{x}_j^*), \quad q_0 = \frac{\theta}{\theta + n - 1} \int \mathcal{K}(\mathrm{y}_i \,|\, x) P_0(\mathrm{d}x).$$

It is well known that the Gibbs sampler described in Algorithm 4.1 tends to mix slowly: when the values of q_j exceed by far q_0, we may need many iterations before a new value is generated. In order to avoid this problem, Maceachern (1994) proposed a variation that can speed up the algorithm by re sampling the distinct values $(\mathrm{x}_i^*)_{i=1}^k$. If we let $\mathbf{z} = (\mathrm{z}_i)_{i=1}^n$, then the parameters of the mixture are generated from $\mathrm{p}(\mathrm{x}_j^* \,|\, \mathbf{z}, \mathbf{y})$, and using the fact that $\mathrm{p}(\mathrm{x}_j^*) = P_0(\mathrm{x}_j^*)$, we obtain

$$\mathrm{p}(\mathrm{x}_j^* \,|\, \mathbf{z}, \mathbf{y}) \propto P_0(\mathrm{x}_j^*) \prod_{i \in \Pi_j} \mathcal{K}(\mathrm{y}_i \,|\, \mathrm{x}_j^*) \quad \text{for } j = 1, \ldots, k, \tag{4.7}$$

where $\Pi_j = \{i : \mathrm{x}_i = \mathrm{x}_j^*\}$ is the j-th block of the partition induced by $\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ or, in terms of the mixture model, it is the set of the indexes of all observations assigned to component j, so that $\Pi_j = \{i : \mathrm{z}_i = j\}$. This transition essentially entails that the posterior on x_j^* is just that of a parametric model with prior distribution P_0 and likelihood $\mathrm{p}(\mathrm{y}_i \,|\, \mathrm{x}_j^*)$, restricted to data within the same component j. This would mean that we have to add an additional step to Algorithm 4.1, so that at each iteration t we additionally sample x_j^* from (4.7).

Regardless of the sampling scheme, given the output $\left(\mathrm{x}_1^{(t)}, \ldots, \mathrm{x}_n^{(t)}\right)_{t=1}^{T}$, we can approximate the random density induced by the integral mapping in (4.2) by the means of

$$f(y) \approx \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n} \sum_{j=1}^{\mathrm{K}_n^{(t)}} n_j^{(t)} \mathcal{K}(y \,|\, \mathrm{x}_j^{*(t)}),$$

where $\mathrm{K}_n^{(t)}$ is the number of distinct values among $\{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ at iteration t. Furthermore, the posterior distribution of the number of distinct values K_n can be approximated as

$$\mathbb{P}[\mathrm{K}_n = k \,|\, \mathbf{y}] \approx \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{\{\mathrm{K}_n^{(t)} = k\}}.$$

If we were to disregard the first T_0 iterations as a burn-in period, the starting point of both sums is shifted to start at $t = T - T_0$ and we substitute $\frac{1}{T}$ with $\frac{1}{T - T_0}$.

4.1.2 Goodness of fit

Now suppose that there are J proposed mixture models $M_1, \ldots, M_J$ that could have generated the data $\mathbf{y}$. The *log-pseudo marginal likelihood*, introduced in Gelman and Rubin (1992), is a criterion that uses the *conditional predictive densities* under model j, defined by (4.8), to compare alternative models in terms of their predictive abilities.

$$\mathrm{CPO}_{ij} = \mathcal{K}(\mathrm{y}_i \,|\, \mathbf{y}^{-i}, \mathrm{x}_j). \tag{4.8}$$

Here $\mathcal{K}(\mathrm{y}_i \,|\, \mathbf{y}^{-i}, \mathrm{x}_j)$ indicates that we are considering the estimated parameters from model j. Gelfand and Dey (1994) provide an easy method for approximating the conditional predictive ordinates as

$$\mathrm{CPO}_{ij} = \left(\frac{1}{T} \sum_{t=1}^{T} \frac{1}{\mathcal{K}(\mathrm{y}_i \,|\, \mathrm{x}_j^{(t)}, M_j)} \right)^{-1},$$

where T stands for the total number of iterations in the MCMC simulation and $\mathrm{x}_j^{(t)}$ is the estimated model parameters on iteration t, belonging to model M_j. The product of CPOs across all observations gives the pseudo marginal likelihood for model j, denoted as $\mathcal{L}_j$

$$\mathcal{L}_j = \prod_{i=1}^{n} \mathrm{CPO}_{ij}.$$

Alternatively one can compute the logarithm of the marginal likelihood for model j, abbreviated as LPML

$$\text{LPML}_j = \sum_{i=1}^{n} \log(\text{CPO}_{ij}). \tag{4.9}$$

The model with the highest $\mathcal{L}_j$ indicates a better fit to the data, so M_{j^*} is selected as the most appropriate of the models being considered if the index j^* maximizes (4.9).

4.2 Hierarchical processes mixtures

All mixture models described the previous sections are applied in settings where observations are assumed to be homogeneous or exchangeable and, as we discussed in the beginning of Chapter 3, there are many situations on which this assumption is not the most appropriate. There has been efforts in Bayesian nonparametric statistics to develop extensions of infinite mixtures that can handle this heterogeneous setting: in Bassetti et al. (2020), a throughout analysis of several hierarchical species sampling models is performed, whilst Argiento et al. (2020) explore mixtures with hierarchies of normalized generalized gamma processes. In other contexts such as topic modelling, both the hierarchical Dirichlet and Pitman-Yor process have been used by Teh (2006) successfully as the prior for the topic distributions for documents, acting as a nonparametric extension of the LDA algorithm.

4.2.1 Chinese restaurant franchise sampler

In order to derive the full conditionals needed in the Gibbs sampler, we will make use of the sampling scheme exposed in Algorithm 3.1. Suppose that we encounter m samples of a partially exchangeable array $\{\mathrm{y}_{i,j} : j = 1, \ldots, n_i \text{ and } i = 1, \ldots, m\}$ and that, given an unobserved parameter $\mathrm{x}_{i,j}$, the data generating mechanism is

$$\begin{aligned} \mathrm{y}_{i,j} \,|\, \mathrm{x}_{i,j} &\overset{\text{ind}}{\sim} \mathcal{K}(\cdot \,|\, \mathrm{x}_{i,j}) \quad j = 1, \ldots, n_i,\ i = 1, \ldots, m \\ \mathrm{x}_{i,j} \,|\, \tilde{p}_0 &\sim \text{NRMI}(\rho, \theta, \tilde{p}_0) \\ \tilde{p}_0 &\sim \text{NRMI}(\rho_0, \theta_0, P_0) \end{aligned} \tag{4.10}$$

where $\mathcal{K}$ is a suitable kernel density. Let

$$\begin{aligned} \mathbf{t}_i &= [\mathrm{t}_{i,j} : j = 1, \ldots, n_i] \\ \mathbf{t} &= [\mathbf{t}_i : i = 1, \ldots, m] \\ \mathbf{d} &= \left[\mathrm{d}_{i,t} : i = 1, \ldots, m \text{ and } t = 1, \ldots, \bar{\ell}_{i\bullet}\right] \\ \boldsymbol{\phi} &= \left[\phi_d : d = \mathrm{d}_{i,t} \text{ for some } i \in \{1, \ldots, m\} \text{ and } t \in \{1, \ldots, \bar{\ell}_{i\bullet}\right] \end{aligned}$$

Let us assume that the base measure of the top level of the hierarchy, P_0, admits a density h (with respect to the Lebesgue measure or any reference measure) and furthermore we will assume that P_0 and $\mathcal{K}$ constitute a conjugate pair, so that the atoms can be integrated out analytically. In this scenario, instead of sampling directly over the parameters $\mathrm{x}_{i,j}$, one

can sample the allocation variables $\mathbf{t}$ and $\mathbf{d}$ and then sample the corresponding posterior parameters ϕ. The model thus becomes

$$\begin{aligned} \mathrm{y}_{i,j} \,|\, \boldsymbol{\phi}, \mathbf{t}, \mathbf{d} &\overset{\text{ind}}{\sim} \mathcal{K}\left(\,\cdot \,|\, \phi_{\mathbf{d}_{i,t_{ij}}}\right) \\ \boldsymbol{\phi} \,|\, \mathbf{t}, \mathbf{d} &\overset{\text{i.i.d.}}{\sim} h(\,\cdot\,) \\ [\mathbf{t}, \mathbf{d}] &\sim \text{HNRMI} \end{aligned} \tag{4.11}$$

Here $[\mathbf{t}, \mathbf{d}] \sim$ HNRMI means that the distribution of the labels $\mathbf{t}, \mathbf{d}$ has been obtained as in Proposition 3.1. Recalling that $\mathrm{d}^*_{i,j} = \mathrm{d}_{i,\mathrm{t}_{i,j}}$ and defining

$$\mathbf{d}^* := [\mathrm{d}^*_{i,j} : i \in \{1, \dots, m\}, j \in \{1, \dots, n_i\}]$$

it is easy to see that $\mathbf{d}^*$ is a function of $\mathbf{d}$ and $\mathbf{t}$, while $\mathbf{d}$ is a function of $\mathbf{d}^*$ and $\mathbf{t}$. Hence both $[\mathbf{t}, \mathbf{d}]$ and $[\mathbf{t}, \mathbf{d}^*]$ contain the same information about the sample.

Now we describe a Gibbs sampler in which the table and dish assignment variables $\mathbf{t}$ and $\mathbf{d}$ are sequentially sampled, conditioned on the state of all other variables. Let $\mathbf{y}_i^{(n_i)} = (\mathrm{y}_{i,j})_{j=1}^{n_i}$ and $\mathbf{y} = \left(\mathbf{y}_i^{(n_i)}\right)_{i=1}^{m}$. It is clear from Algorithm 3.1 that we have three types of output whenever we are sampling the full conditional of $[\mathrm{t}_{i,j}, \mathrm{d}^*_{i,j}]$: either we locate $\mathrm{y}_{i,j}$ in an old cluster in group i and it gets assigned the component associated with that cluster, or at a new one. If the latter occurs, then two disjoint events are possible: either this new clusters gets assigned a new mixture component or a new one.

Refer to $\omega_0^{(n)}$ and $\omega_j^{(n)}$ as the weights of the predictive distribution of the random partition with EPPF $\widetilde{\Phi}_i$ and analogously as $\tilde{\omega}_0^{(n)}$ and $\tilde{\omega}_j^{(n)}$ the weights of the predictive distribution of the random partition with EPPF Φ_0. The conjugacy assumption allows us to integrate out analytically the mixture components $\boldsymbol{\phi}$ at each step, and the full conditional of $[\mathrm{t}_{i,j}, \mathrm{d}^*_{i,j}]$ is given by

$$\begin{aligned} \mathrm{p}(\mathrm{t}_{i,j} = t^{\text{old}}, \mathrm{d}^*_{i,j} = \mathrm{d}^*_{i,t^{\text{old}}} \,|\, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij}) &\propto \omega_{t^{\text{old}}}^{(n_i-1)}\left(\mathbf{t_i}^{-ij}\right) f_{d^*_{i,t^{\text{old}}}}(\mathrm{y}_{i,j}) \\ \mathrm{p}(\mathrm{t}_{i,j} = t^{\text{new}}, \mathrm{d}^*_{i,j} = d^{\text{old}} \,|\, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij}) &\propto \omega_0^{(n_i-1)}\left(\mathbf{t_i}^{-ij}\right) \tilde{\omega}_{d^{\text{old}}}^{|\boldsymbol{\ell}|^{-ij}}\left(\mathbf{d}^{-ij}\right) f_{d^{\text{old}}}(\mathrm{y}_{i,j}) \\ \mathrm{p}(\mathrm{t}_{i,j} = t^{\text{new}}, \mathrm{d}^*_{i,j} = d^{\text{new}} \,|\, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij}) &\propto \omega_0^{(n_i-1)}\left(\mathbf{t_i}^{-ij}\right) \tilde{\omega}_0^{|\boldsymbol{\ell}|^{-ij}}\left(\mathbf{d}^{-ij}\right) f_{d^{\text{new}}}(\mathrm{y}_{i,j}), \end{aligned} \tag{4.12}$$

where

$$\begin{aligned} \omega_{t^{\text{old}}}^{(n_i-1)}\left(\mathbf{t_i}^{-ij}\right) &= \omega_{t^{\text{old}}}^{(n_i-1)}\left(q_{i\bullet 1}^{-ij}, \dots, q_{i\bullet \bar{\ell}_{i\bullet}^{-ij}}^{-ij}\right) \\ \omega_0^{(n_i-1)}\left(\mathbf{t_i}^{-ij}\right) &= \omega_0^{(n_i-1)}\left(q_{i\bullet 1}^{-ij}, \dots, q_{i\bullet \bar{\ell}_{i\bullet}^{-ij}}^{-ij}\right) \end{aligned}$$

and for an arbitrary set of indexes $\mathcal{S}$

$$f_d(\{\mathrm{y}_{i,t}\}_{i,t\in\mathcal{S}}) = \frac{\int \prod_{i't' \in \mathcal{S}_d \cup \mathcal{S}} \mathcal{K}(\mathrm{y}_{i',t'} \,|\phi) h(\phi) \mathrm{d}\phi}{\int \prod_{i't' \in \mathcal{S}_d \setminus \mathcal{S}} \mathcal{K}(\mathrm{y}_{i',t'} \,|\, \phi) h(\phi) \mathrm{d}\phi},$$

where $\mathcal{S}_d$ denotes the d-th component of the mixture. The Gibbs update for the component served at table c is derived similarly, as again we have two possibilities: either the table gets assigned an old mixture component or a new one.

$$\begin{aligned} \mathrm{p}(\mathrm{d}_{i,c} = d^{\text{new}} \,|\, \mathbf{t}, \mathbf{d}^{-ic}) &= \widetilde{\omega}_0^{|\boldsymbol{\ell}|^{-ic}}\left(\mathbf{d}^{-ic}\right) \\ \mathrm{p}(\mathrm{d}_{i,ct} = d^{\text{old}} \,|\, \mathbf{t}, \mathbf{d}^{-ic}) &= \widetilde{\omega}_{d^{\text{old}}}^{|\boldsymbol{\ell}|^{-ic}}\left(\mathbf{d}^{-ic}\right) \text{ for } d^{\text{old}} \in \mathcal{D}^{-ic}, \end{aligned} \tag{4.13}$$

where

$$\begin{aligned} \widetilde{\omega}_{d^{\text{old}}}^{|\boldsymbol{\ell}|^{-ic}}\left(\mathbf{d}^{-ic}\right) &= \widetilde{\omega}_{d^{\text{old}}}^{|\boldsymbol{\ell}|^{-ic}}(\bar{\ell}_{\bullet 1}^{-ic}, \ldots, \bar{\ell}_{\bullet|\mathcal{D}^{-ic}|}^{-ic}) \\ \widetilde{\omega}_0^{|\boldsymbol{\ell}|^{-ic}}\left(\mathbf{d}^{-ic}\right) &= \widetilde{\omega}_0^{|\boldsymbol{\ell}|^{-ic}}(\bar{\ell}_{\bullet 1}^{-ic}, \ldots, \bar{\ell}_{\bullet|\mathcal{D}^{-ic}|}^{-ic}). \end{aligned}$$

Finally, one can sample the values of $\boldsymbol{\phi}$ given $\mathbf{y}, \mathbf{t}, \mathbf{d}$ by

$$\mathrm{p}(\boldsymbol{\phi} \,|\, \mathbf{y}, \mathbf{t}, \mathbf{d}) \propto \prod_{d \in \mathcal{D}} h(\phi_d) \prod_{(i,j):\mathrm{d}_{i,j}^* = d} \mathcal{K}(\mathrm{y}_{i,j} \,|\, \phi_d). \tag{4.14}$$

Full details can be consulted on Appendix C. When sampling $\mathbf{t}$ one needs to sample jointly $[\mathbf{t}, \mathbf{d}^*]$ and since $\mathbf{d}$ is a function of $[\mathbf{t}, \mathbf{d}^*]$, one implicitly obtains a sample for $\mathbf{d}$. However, re-sampling $\mathbf{d}$ given $\mathbf{t}$ in a second step improves the mixing of the Markov Chain. The Gibbs sampler is described in Algorithm 4.2.

Algorithm 4.2: T iterations of the Chinese Restaurant Franchise Gibbs sampler.

Data: Observations $\left(\mathbf{y}_i^{(n_i)}\right)_{i=1}^m$ assumed as in model (4.10).
Initialize $\mathbf{t}^{(0)}$ and $\mathbf{d}^{*(0)}$.
for *iteration* $t = 1, \ldots, T$ **do**
  for $i = 1, \ldots, m$ **do**
    for $j = 1, \ldots, n_i$ **do**
      Sample $\left[\mathrm{t}_{i,j}^{(t)}, \mathrm{d}_{i,j}^{*\,(t)}\right]$ from (4.12)
    for $c = 1, \ldots \bar{\ell}_{i\bullet}$ **do**
      Sample $\mathrm{d}_{i,c}^{(t)}$ from (4.13)
    for $d = 1, \ldots, \mathrm{K}_N^{(t)}$ **do**
      Sample ϕ_d according to (4.14).

The marginal density of each group will be approximated by the means of

$$\hat{f}_i(\mathrm{y}_i) \approx \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n_i} \sum_{j=1}^{\bar{\ell}_{i\bullet}^{(t)}} q_{i\bullet j}^{(t)} \mathcal{K}\left(\mathrm{y}_i \,\middle|\, \phi_{k_{i,\mathrm{t}_{i,j}}{}^{(t)}}\right),$$

where $\bar{\ell}_{i\bullet}^{(t)}$ and $q_{i\bullet j}^{(t)}$ are the number of groups and the frequency of each group on iteration t respectively. The posterior distribution of the number of groups on each group i, K_{i,n_i}, can be approximated by

$$\mathbb{P}[\mathrm{K}_{i,n_i} = k \,|\, \mathbf{y}] \approx \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\left\{\bar{\ell}_{i\bullet}^{(t)}=k\right\}}$$

for $i = 1, \ldots, m$, and the distribution of the global number of components can be approximated by the means of

$$\mathbb{P}[\mathrm{K}_N = k \,|\, \mathbf{y}] \approx \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\left\{\mathrm{K}_N^{(t)}=k\right\}},$$

where $\mathrm{K}_N^{(t)}$ is the number of unique global labels at iteration t.

4.3 Simulation study

We will assume a Gaussian kernel with random mean and variance and, to attain conjugacy, P_0 will be a Normal Inverse Gamma distribution, so that $\phi_{i,j} = \left(m_{i,j}, \sigma_{i,j}^2\right)$ and $P_0(\mathrm{d}m, \mathrm{d}\tau^2) = \mathrm{N}\left(\mathrm{d}m \,\middle|\, m_0, \frac{\sigma^2}{\tau_0}\right) \mathrm{InvGa}\left(\mathrm{d}\sigma^2 \,|\, a, b\right)$. To compare and asses each model's adequacy, we will use the LPML as a measure of the goodness of fit. We will compare between the three hierarchical processes we studied in Chapter 3 and also mixed cases, e.g. $\tilde{p}_i \,|\, \tilde{p}_0 \sim \mathfrak{D}(\theta\tilde{p}_0)$, $\tilde{p}_0 \sim \mathcal{PY}(\sigma, \theta, G)$.

4.3.1 One shared component

For the first experiment, we simulated 200 data points from the mixtures

$$\begin{aligned}
\mathrm{y}_{1,j} &\sim 0.6\mathrm{N}(-4, 1) + 0.2\mathrm{N}(0, 0.5) + 0.2\mathrm{N}(3.5, 1.25) \quad j = 1, \ldots, 100 \\
\mathrm{y}_{2,j} &\sim 0.7\mathrm{N}(0, 0.5) + 0.3\mathrm{N}(-3.5, 1) \quad j = 1, \ldots, 100
\end{aligned}$$

The common mixture component is $\mathrm{N}(0, 0.5)$ but with significantly different weights. The parameters of the prior processes were chosen in such a way that, marginally, $\mathbb{E}[K_{i,100}] = 5$ for $i = 1, 2$ and, globally, $\mathbb{E}[K_{200}] = 6$. The hyperparameters of P_0, (m_0, τ_0, a, b), were chosen as the ones minimizing the LPML and correspond to assuming that the prior moments of P_0 satisfy $\mathbb{E}[m] = \bar{\mathbf{y}}$, $\mathbb{E}[\sigma^2] = \frac{\mathrm{var}(\mathbf{y})}{2}$, where $\bar{\mathbf{y}}$ and $\mathrm{var}(\mathbf{y})$ are the overall mean and variance of the data, for all seven HNRMIs and $\mathrm{var}[m] = 7 = \mathrm{var}[\sigma^2]$ for the hierarchical Pitman-Yor Dirichlet Process (HPYDP), the hierarchical Pitman-Yor Stable Process (HPYSP) and the hierarchical Dirichlet Pitman-Yor Process (HDPYP). As for the hierarchical Dirichlet Process (HDP) and the hierarchical Stable Pitman-Yor Process (HSPYP), the hyperparameters are such that $\mathrm{var}[m_0] = 7$ and $\mathrm{var}[\sigma^2] = 5$. Finally, the setting $\mathrm{var}[m] = 5$ and $\mathrm{var}[\sigma^2] = 7$ correspond to the hyperparameters for the hierarchical Stable Process (HSP) and the hierarchical Pitman-Yor Process. Details can be consulted at Appendix C, where a thorough sensitivity analysis is attached.

The estimated densities for $(y_{1,j})_{j=1}^{100}$ and $(y_{2,j})_{j=1}^{100}$ are shown in Figures 4.1 and 4.2, taking into account 4000 iterations of the Gibbs sampler after a burn-in period of 5000 iterations. From here we see that all prior processes do a good job at estimating the density and furthermore, all of them recover the three modes present in $\mathbf{y}_1^{(100)}$ and the two modes at $\mathbf{y}_2^{(100)}$. The fit is fairly similar at most points from model to model. In spite of that, if we look closely to the left of the first histogram and at the center of the second histogram, we see that the density estimated by the means of the HPYP differs slightly from the others. On the other hand, on the right side of Figure 4.2 we see that the HSPYP and the HDP differ in the size of this second mode, placing less mass than the other models.

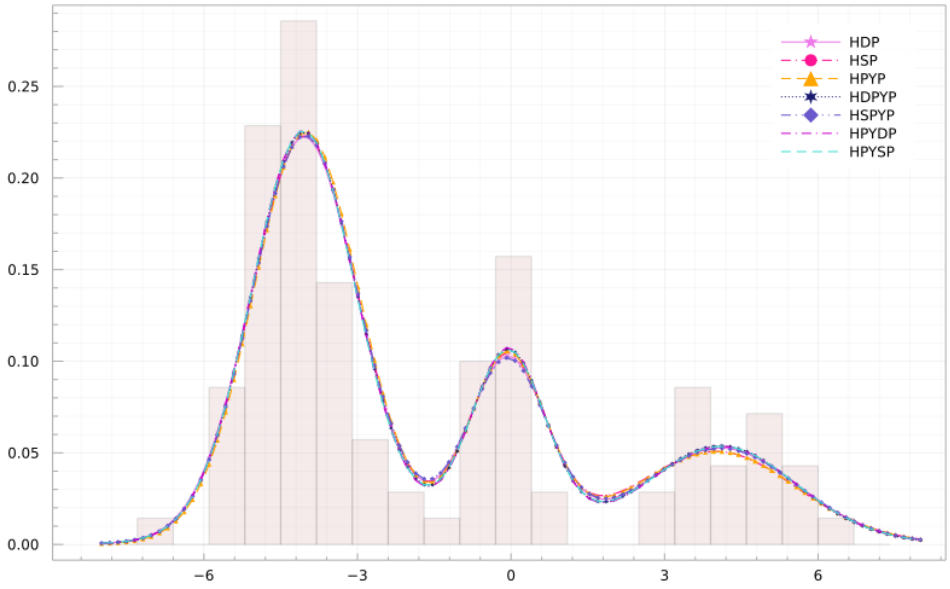


Figure 4.1: Estimated density for $(y_{1,j})_{j=1}^{100}$.

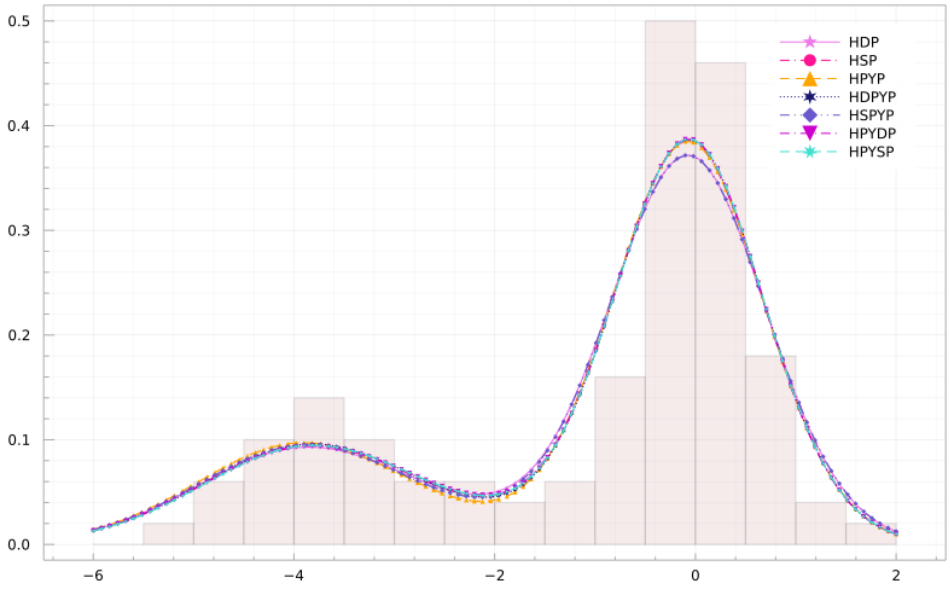


Figure 4.2: Estimated density for $(y_{2,j})_{j=1}^{100}$.

Figure 4.3 shows the posterior distribution of the global number of clusters, K_{200}, and Table 4.1 exhibits the modes and its corresponding masses. We see that all processes give high probability to numbers close to four, which is the true number of global components from which the data was sampled, however only the HPYP place the mode exactly at four. The rest of the models tend to give higher probability to larger values of K_{200}, meaning that these models use more components to estimate the densities from Figures 4.1 and 4.2. Particularly, the inclusion of a stable process at any level of the hierarchy leads to more platykurtic distributions as opposed to the inclusion of a Dirichlet process, which produces more leptokurtic distributions.

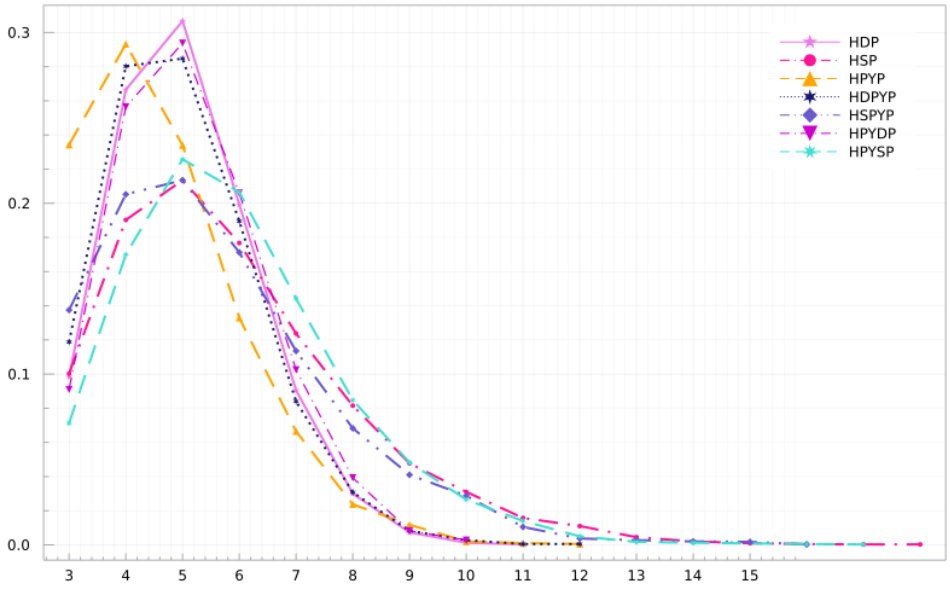


Figure 4.3: Posterior distribution of K_{200}.

	HDP	**HSP**	**HPYP**	**HDPYP**	**HSPYP**	**HPYDP**	**HPYSP**
Mode	5	5	4	5	5	5	5
Mass	0.3067	0.2135	0.2932	0.2847	0.2135	0.2937	0.22575

Table 4.1: Modes for the posterior distribution of K_{200} *for different prior processes.*

Figures 4.4 and 4.5 show the posterior distribution of $K_{1,100}$ and $K_{2,100}$. Most of the prior processes place high probability to number significantly close to the true number of components, that is three and two for group one and group two respectively. The behavior of the HPYP resembles more to the one of the HSP process rather than the HDP, as in both $K_{1,100}$ and $K_{2,100}$ it tends to use more components than any other model. In general the same behavior as in the global posterior distribution of K_{200} is observed: models that include the Dirichlet process place higher masses at smaller values while models that involve a stable process produce posterior distributions of K_{i,n_i} with a larger support. Clearly the distribution of the global number of components involves smaller values than the sum of the marginal number of components, indicating that all the hierarchical prior processes considered allow for various degrees of information pooling across the two different groups.

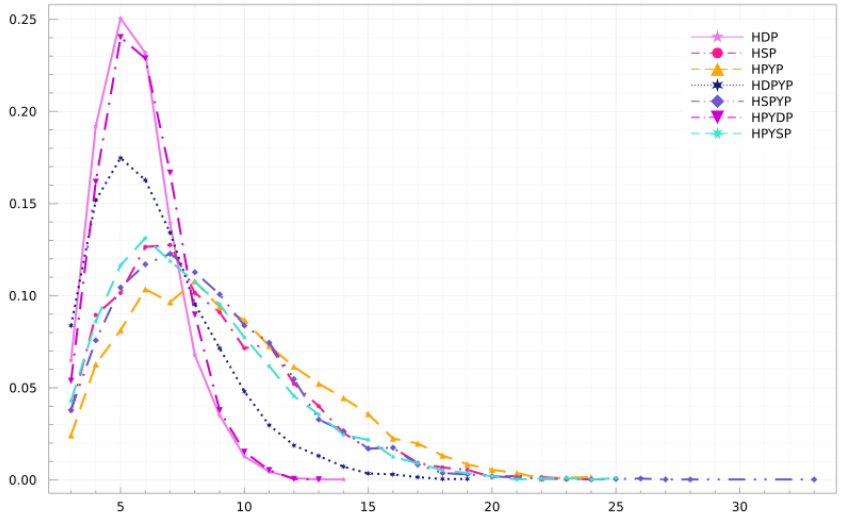


Figure 4.4: Posterior distribution of $\mathrm{K}_{1,100}$.

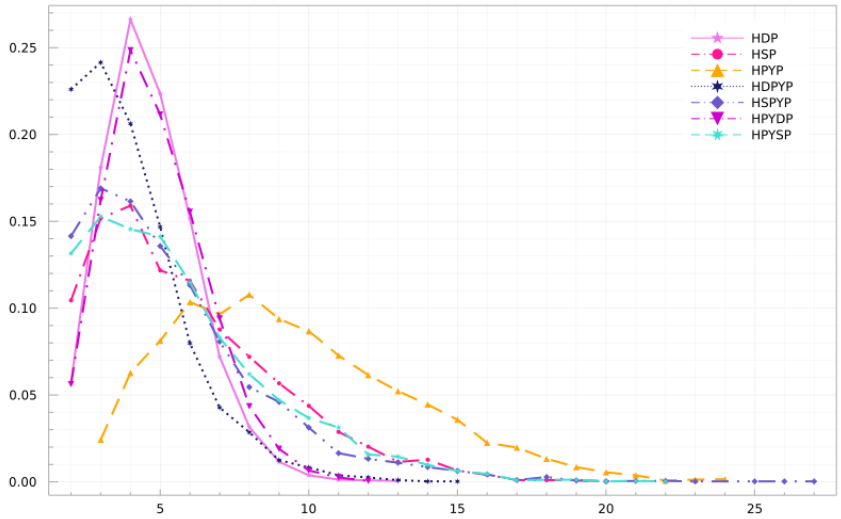


Figure 4.5: Posterior distribution of $\mathrm{K}_{2,100}$

Finally, Table 4.2 shows the LPML for each model.

	HDP	**HSP**	**HPYP**	**HDPYP**	**HSPYP**	**HPYDP**	**HPYSP**
Global LPML	-1.062	-1.062	**-1.057**	-1.061	-1.06	-1.069	-1.067

Table 4.2: LPML (10^3) *for different prior processes.*

In general the effect of the random probability measure or its parameters cannot be distinguished speaking on the estimated densities. This can be due to the initial selection of the parameters, as all of them were chosen under the same criteria and fairly close to the true number of components, and due to the fact that Gaussian mixtures are very flexible models. In terms of the number of clusters, the inclusion of the Dirichlet process at

some point of the hierarchy produces models that tend to use fewer components, whereas the inclusion of a stable process produces greater dispersion on the number of clusters both globally and marginally. The LPML indicates a similar goodness of fit a across all hierarchical processes, although the HPYP performs slightly better.

4.3.2 Two shared components and large heterogeneity

Now we will perform an experiment on which large heterogeneity exist among the data. We will generate six samples from the mixtures

$$\begin{aligned} \mathrm{y}_{i,j} &\sim 0.4\mathrm{N}(-3,1) + 0.3\mathrm{N}(-5-i,1) + 0.3\mathrm{N}(-2+i,1) \text{ for } i = 1,2 \\ \mathrm{y}_{i,j} &\sim 0.4\mathrm{N}(-3,1) + 0.3\mathrm{N}(-2-i,1) + \mathrm{N}(-4-i,1) \text{ for } i = 3,4,5,6 \end{aligned}$$

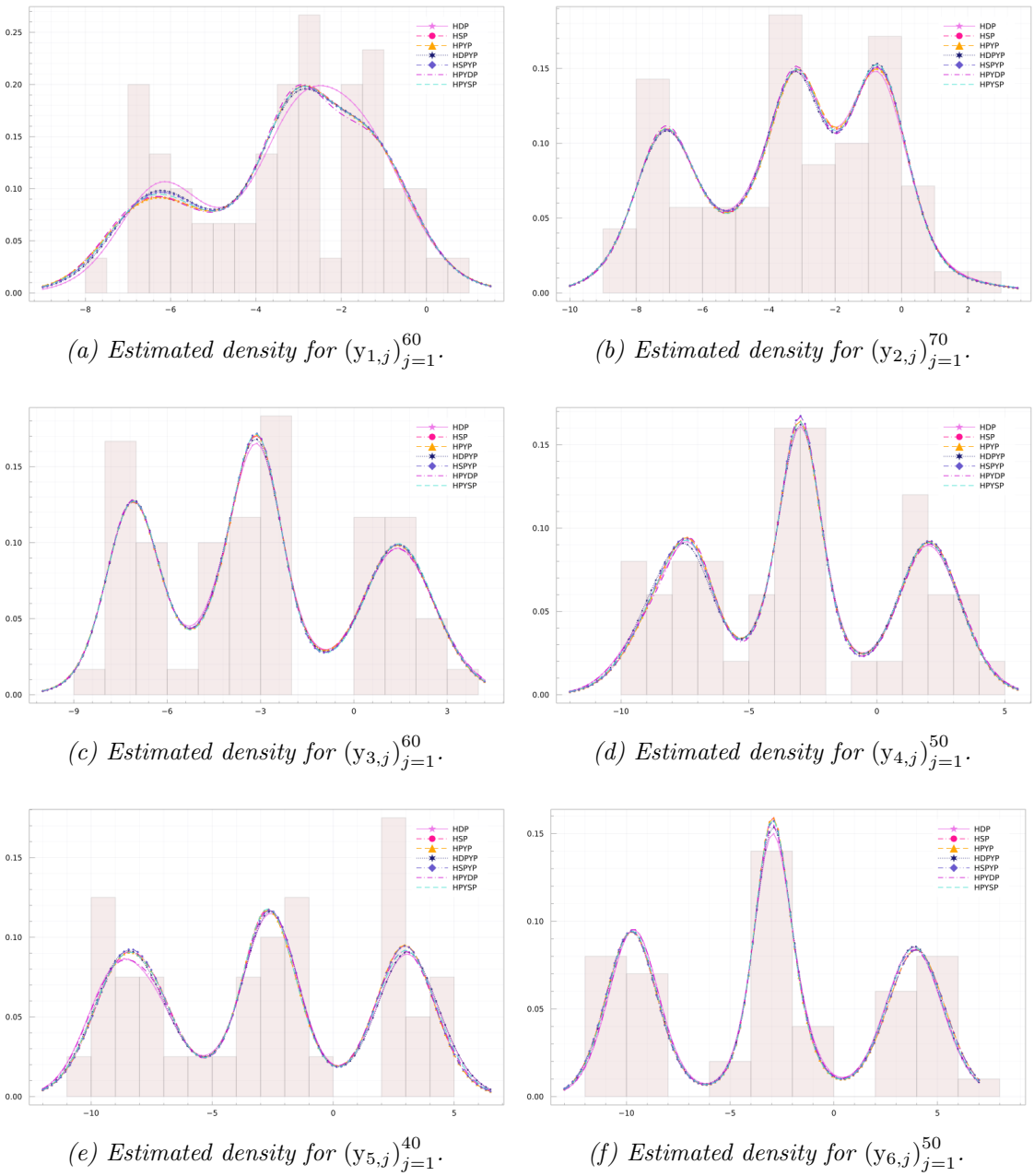
with $j \in \{60, 70, 60, 50, 40, 50\}$ respectively. In this setting, there is a larger number of group specific components and the common component to all 6 samples $\mathrm{N}(-3,1)$ has a relatively smaller weight. $\mathbf{y}_2^{(n_2)}$ and $\mathbf{y}_5^{(n_5)}$ additionally share $\mathrm{N}(-7,1)$. The hyperparameters of P_0 are going to remain fixed at the values that correspond to solving the system

$$\begin{aligned} \mathbb{E}[m_0] = \bar{\mathbf{y}}, \quad \mathbb{E}[\sigma^2] = \frac{\mathrm{var}(\mathbf{y})}{6} \\ \mathrm{var}[m_0] = 5, \quad \mathrm{var}[\sigma^2] = 5 \end{aligned}$$

which are $(m_0, \tau_0, a, b) = (-3.114,\ 0.272,\ 2.37,\ 1.865)$. The underlying prior process parameters are selected such that $\mathbb{E}[\mathrm{K}_{330}] = 50$, far from the true value to emphasize the reinforcement. For the HDP, the HSP and the HPYP the parameters are chosen such that $\mathbb{E}[\mathrm{K}_{i,40}] = 15$. For the other cases $\mathbb{E}[\mathrm{K}_{i,40}] = 14$, just to avoid having equal parameters.

The estimated densities are shown in Figure 4.6, taking into account 4000 iterations of the Gibbs sampler after a burn-in period of 6000. Again we see that all models provide an overall fairly good estimation of the densities, and the fit is quite similar across distinct models. Most of the hierarchical processes are able to capture the three modes present in each sample $\mathbf{y}_i^{(n_i)}$ for $i = 2, \ldots, 6$. For $\mathbf{y}_1^{(n_1)}$, from Figure 4.6a we see that none of the models was able to recover well the subtle changes in the histogram, as the two modes at the left are barely distinguished from each other. This could be due to the fact that the components $\mathrm{N}(-3,1)$ and $\mathrm{N}(-1,1)$ have means that are quite close to each other and hence $\mathrm{N}(-3,1)$ could produce samples that can easily be confused as if they come from $\mathrm{N}(-1,1)$ and vice-versa. The HDP is the one who struggles the most and completely fails to differentiate the second and third modes, while the HPYDP differs slightly in the height of the gap between the second and third mode.

In Figures 4.6d, 4.6e and 4.6f we see that the main differences between all fits occur at the height of the modes, and the HDP is the one that differs the most from the other processes.



(a) Estimated density for $(\mathrm{y}_{1,j})_{j=1}^{60}$.

(b) Estimated density for $(\mathrm{y}_{2,j})_{j=1}^{70}$.

(c) Estimated density for $(\mathrm{y}_{3,j})_{j=1}^{60}$.

(d) Estimated density for $(\mathrm{y}_{4,j})_{j=1}^{50}$.

(e) Estimated density for $(\mathrm{y}_{5,j})_{j=1}^{40}$.

(f) Estimated density for $(\mathrm{y}_{6,j})_{j=1}^{50}$.

Figure 4.6

Figure 4.7 shows the posterior distribution of the global number of clusters K_{330}. We observe a behavior similar to that of the previous experiment: the HDP, the HPYDP and the HDPYP have more leptokurtic posterior distributions for K_{330}, and they tend to use fewer components since the mode is placed on lower values. These three models put the majority of their mass around 25, just as the HPYDP and the HPYSP, although they have more dispersion. The HSP and the HPYP are the ones that have more platykurtic distributions and a greater support, as they assign positive probability to numbers above 50. The true number of components is 12, so the placement of the mode is quite far from the true value in most cases. Table 4.3 exhibits the modes and its corresponding mass.

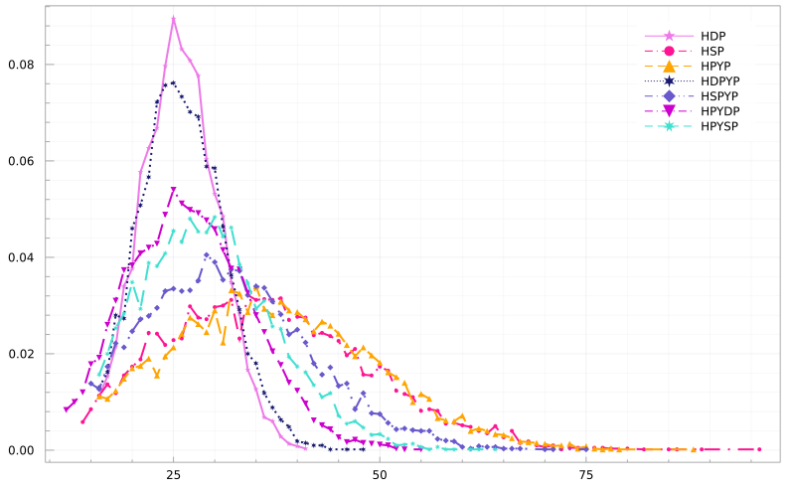
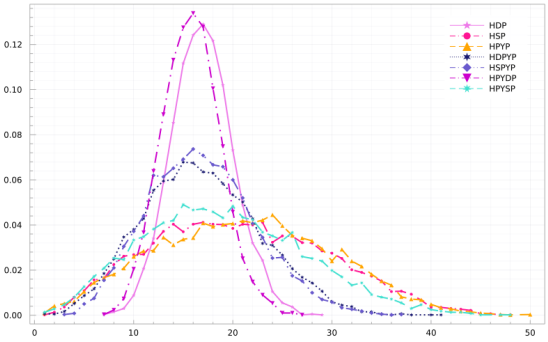


Figure 4.7: Posterior distribution of K_{330}.

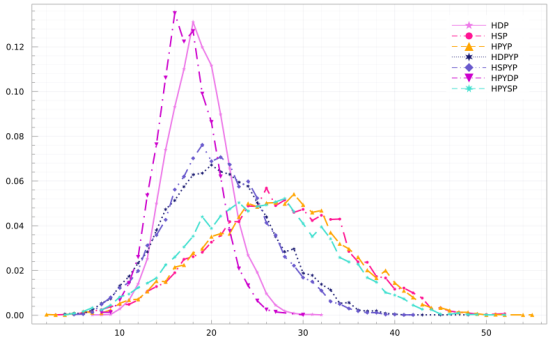
	HDP	**HSP**	**HPYP**	**HPYDP**	**HPYSP**	**HSPYP**	**HDPYP**
Mode	25	36	35	25	35	30	25
Mass	0.090	0.0306	0.034	0.054	0.048	0.0405	0.076

Table 4.3: Modes for the posterior distribution of K_{330} *for different prior processes.*

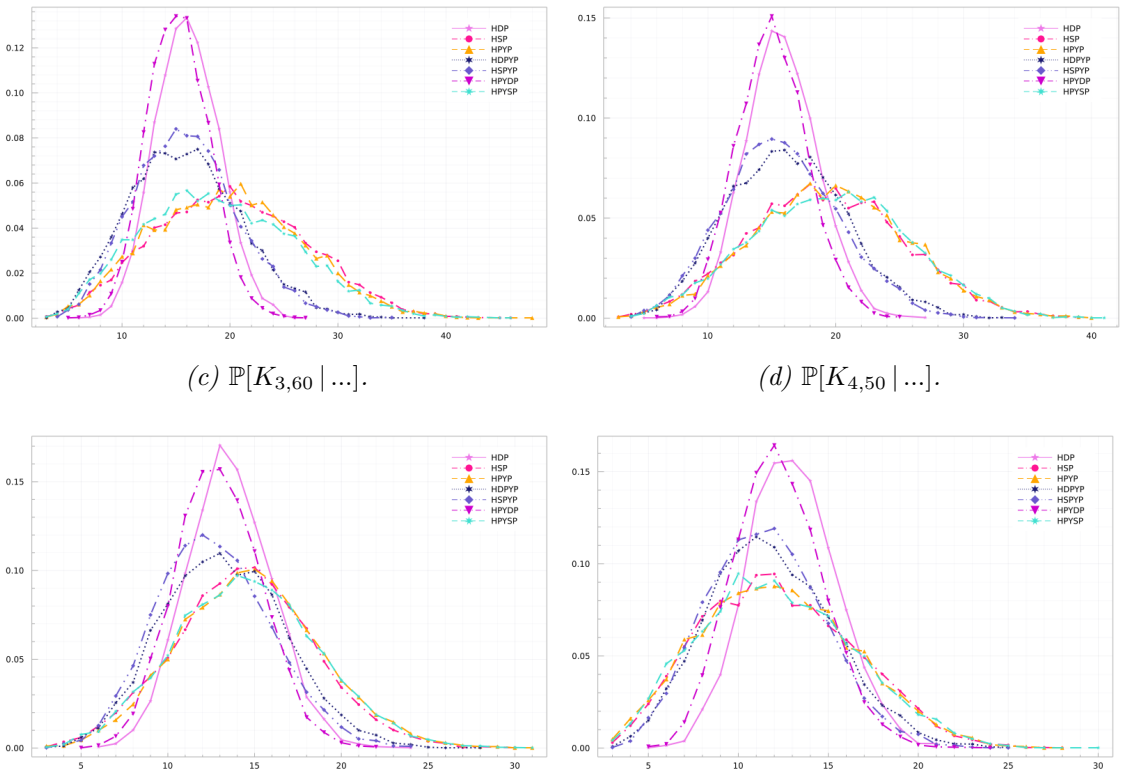
Figure 4.8 shows the posterior distributions of the marginal number of groups, where we see something consistent with the first experiment: the posteriors of K_{i,n_i} for the HDP and the HPYDP tend to have a smaller support and a leptokurtic shape. It is interesting to note that in Figure 4.8f, the posterior distribution of $\mathrm{K}_{6,50}$ is not significantly different between all prior processes, this could be due to the fact that this data set is the one on which the modes are more separated from each other, whilst in Figure 4.8a we see that the posterior of $\mathrm{K}_{1,60}$ is the one on which there exist more discrepancy between all prior processes. Most processes place the mode between 10 and 15, which is quite far from the true value (three) but close to the value k that we chose as the expected value a priori of the number of cluster in each group.



(a) $\mathbb{P}[K_{1,60} \,|\, ...]$.



(b) $\mathbb{P}[K_{2,70} \,|\, ...]$.



(c) $\mathbb{P}[K_{3,60} \mid ...]$. *(d)* $\mathbb{P}[K_{4,50} \mid ...]$.

(e) $\mathbb{P}[K_{5,40} \mid ...]$. *(f)* $\mathbb{P}[K_{6,50} \mid ...]$.

Figure 4.8: Posterior distribution of K_{i,n_i}.

Table 4.4 shows the LPML estimates, with the highest score highlighted in bold. From here we see that again the HPYP gives a better fit than the other processes, but in general the inclusion of a Pitman-Yor process at some point of the hierarchy improves the goodness of fit, as the HSP and the HDP are the ones with the worst score.

	HDP	**HSP**	**HPYP**	**HPYDP**	**HPYSP**	**HSPYP**	**HDPYP**
Global LPML	-6.379	-6.339	**-6.194**	-6.261	-6.251	-6.262	-6.301

Table 4.4: LPML (10^2) *for different prior processes.*

It is clear that the a priori choice of the parameters such that $\mathbb{E}[\mathrm{K}_N] = k$, for a k close to the true value of components, produces better results, as it was the case in the previous experiment. In the two experiments, except for the HDP, the HPYDP and the HDPYP, the prior processes produce platykurtic posterior distributions of K_N and with much larger support. It would seem that in the presence of larger heterogeneity, a Pitman-Yor process at some point of the hierarchy improves the fit in terms of the LPML, although in general the estimated densities are practically the same, regardless of the prior process choice.

Concluding remarks

Extensions of Bayesian nonparametric models to partially exchangeable settings is analytically challenging, although not impossible. Having a closed form result for the pEPPF allows us to study important quantities such as the distribution of the number of groups, K_N, which has an intuitive interpretation. This fact combined with having a closed expression for the correlation coefficient allows us to develop guidelines on the choice of the processes' parameters, as in the small experiments presented in Chapter 4.

Additionally, CRM–based dependent priors look promising as, conditionally on a suitable latent random variable or vector of random variables, they typically display distributional properties reminiscent of those available in the exchangeable case. These completely explicit posterior representations allows us to device marginal and/or conditional samplers. Moreover, most of the properties and posterior distributions concerning hierarchical NRMIs are quite general, as they can easily be adapted to random probability measures constructed from transformations of CRMs, as it was the case with the hierarchical Pitman-Yor process. One only needs to take into account the proper adaptations depending on the specific transformations of the CRMs.

A possible area of research is to study the dependence between these random probability measures, i.e. to develop some kind of metric that allow us to understand how much the model is far from an exchangeable situation.

Appendix A

Proof of 1.1. For simplicity we will limit ourselves to the case where $\mathbb{X} = \mathbb{R}$, and the proof will be based upon the theory of martingales and the work of Kingman (1978).

A random variable is *n-symmetric* if it is a function of an infinite sequence $\mathbf{x} = \{\mathrm{x}_i\}_{i=1}^{\infty}$ and it is invariant under permutations of the first n entries of $\mathbf{x}$, for example $\mathrm{x}_1\mathrm{x}_2 + \mathrm{x}_3$ is 2-symmetric but not 3-symmetric. Let $\mathcal{F}_n$ be the smallest σ-algebra with respect to which all n-symmetric functions are measurable. Note that $\mathcal{F}_{n+1} \subseteq \mathcal{F}_n$, since a $n+1$-symmetric function f can be written as $f(\mathrm{x}_1, \ldots, \mathrm{x}_{n+1}) = g(h(\mathrm{x}_1, \ldots, \mathrm{x}_n), \mathrm{x}_{n+1})$ with h being n-symmetric.
Let $g : \mathbb{R} \to \mathbb{R}$ be a bounded measurable function and γ a bounded n-symmetric function. Exchangeability of $\mathbf{x}$ implies that for $1 \leq k \leq n$

$$\begin{aligned}\mathbb{E}\left[g(\mathrm{x}_k)\gamma(\mathbf{x})\right] &= \mathbb{E}\left[g(\mathrm{x}_1)\gamma(\mathrm{x}_j, \mathrm{x}_2, \ldots, \mathrm{x}_{k-1}, \mathrm{x}_1, \mathrm{x}_{k+1}, \ldots)\right] \\ &= \mathbb{E}\left[g(\mathrm{x}_1)\gamma(\mathbf{x})\right],\end{aligned}$$

so that

$$\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} g(\mathrm{x}_j)\gamma(\mathbf{x})\right] = \mathbb{E}\left[g(\mathrm{x}_1)\gamma(\mathbf{x})\right].$$

Noting that $\gamma(\mathbf{x}) = \mathbb{1}_A(\mathbf{x})$ is n-symmetric and bounded for any $A \in \mathcal{F}_n$, and that $\frac{1}{n}\sum_{j=1}^{n} g(\mathrm{x}_j)$ is also n-symmetric, last equation can be rewritten as

$$\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} g(\mathrm{x}_j)\mathbb{1}_A\right] = \mathbb{E}\left[g(\mathrm{x}_1)\mathbb{1}_A\right],$$

meaning that $\frac{1}{n}\sum_{j=1}^{n} g(\mathrm{x}_j)$ is a version of the conditional expectation of $g(\mathrm{x}_1)$ with respect to $\mathcal{F}_n$, i.e.

$$\frac{1}{n}\sum_{j=1}^{n} g(\mathrm{x}_j) = \mathbb{E}\left[g(\mathrm{x}_1) \,|\, \mathcal{F}_n\right]. \tag{A.1}$$

Thus $\left\{\frac{1}{n}\sum_{j=1}^{n} g(\mathrm{x}_j)\right\}_{n=1}^{\infty}$ defines a closed and backwards martingale w.r.t. $\{\mathcal{F}_n\}_{n=1}^{\infty}$, so by the backwards martingale convergence theorem, as $n \to \infty$

$$\frac{1}{n}\sum_{j=1}^{n} g(\mathrm{x}_j) \overset{\text{a.s.}}{\to} \mathbb{E}\left[g(\mathrm{x}_1) \,|\, \mathcal{F}_\infty\right], \tag{A.2}$$

where $\mathcal{F}_\infty = \bigcap_{n=1}^{\infty} \mathcal{F}_n$. For the particular choice of $g(x) = \delta_x(A)$ for $A \in \mathcal{F}_n$, (A.2) becomes

$$P_n(A) = \frac{1}{n}\sum_{j=1}^{n} \delta_{\mathrm{x}_j}(A) \overset{\text{a.s.}}{\to} \tilde{p}(A), \tag{A.3}$$

where

$$\tilde{p}(A) = \mathbb{P}[\mathrm{x}_1 \in A \,|\, \mathcal{F}_\infty].$$

Note that for $A \in \mathcal{B}(\mathbb{R})$ fixed, $\tilde{p}(A)$ is a $\mathcal{F}_\infty$-measurable random variable, and that this proofs the second part of the representation theorem. The generalization of the argument leading to (A.1) is as follows. Let us define for $n \geq 1$ and $1 \leq k < n$ the sets of indexes

$$\begin{aligned} I_{n,k} &:= \{(j_1, \ldots, j_k) \,:\, j_i \in \{1, \ldots, n\},\, i = 1, \ldots, k\} \\ J_{n,k} &:= \{(j_1, \ldots, j_k) \,:\, j_i \in I_{n,k}\,, j_i \neq j_l,\, i \neq l\}. \end{aligned}$$

Notice that $|\, J_{n,k} \,| = (n)_k$. For $g : \mathbb{R}^k \to \mathbb{R}$ a bounded and measurable function, let

$$A_n(g) := \frac{1}{(n)_k} \sum_{(j_1,\ldots,j_k) \in J_{n,k}} g(\mathrm{x}_{j_1}, \ldots, \mathrm{x}_{j_k}) = \frac{1}{(n)_k} \sum_{j_1=1}^{n} \cdots \sum_{j_k=1}^{n} g(\mathrm{x}_{j_1}, \ldots, \mathrm{x}_{j_k}).$$

Due to exchangeability, one has that

$$\mathbb{E}\left[g(\mathrm{x}_1, \ldots, \mathrm{x}_k) \,|\, \mathcal{F}_n\right] = \mathbb{E}\left[g(\mathrm{x}_{j_1}, \ldots, \mathrm{x}_{j_k}) \,|\, \mathcal{F}_n\right].$$

As $A_n(g)$ is a n-symmetric function and hence $\mathcal{F}_n$-measurable

$$\begin{aligned} A_n(g) &= \mathbb{E}\left[A_n(g) \,|\, \mathcal{F}_n\right] \\ &= \frac{1}{(n)_k} \sum_{j \in J_{n,k}} \mathbb{E}\left[g(\mathrm{x}_{j_1}, \ldots, \mathrm{x}_{j_k}) \,|\, \mathcal{F}_n\right] \\ &= \mathbb{E}\left[g(\mathrm{x}_1, \ldots, \mathrm{x}_k) \,|\, \mathcal{F}_n\right]. \end{aligned}$$

By the backwards martingale convergence theorem, as $n \to \infty$,

$$\begin{aligned}\mathbb{E}\left[g(\mathrm{x}_1,\ldots,\mathrm{x}_k)\,|\,\mathcal{F}_\infty\right] &= \lim_{n\to\infty} A_n(g) \\ &= \lim_{n\to\infty} \frac{1}{(n)_k} \sum_{j_1=1}^{n} \cdots \sum_{j_k=1}^{n} g(\mathrm{x}_{j_1},\ldots,\mathrm{x}_{j_k}) \\ &= \lim_{n\to\infty} \frac{1}{n^k} \sum_{j_1=1}^{n} \cdots \sum_{j_k=1}^{n} g(\mathrm{x}_{j_1},\ldots,\mathrm{x}_{j_k}) \qquad \text{(A.4)}\end{aligned}$$

almost surely. The last equality follows from the fact that, for a fixed k, there are $|\,I_{n,k}\,| - |J_{n,k}\,| = n^k - (n)_k$ combinations of k indexes out of $\{1,\ldots,n\}$ that have coincidences and

$$\lim_{n\to\infty} \frac{n^k - (n)_k}{n^k} = 0,$$

meaning that, as n grows larger, the proportion of vectors $(j_1,\ldots,j_k)$ with coincidences vanishes. That is, there is no difference between considering index vectors within $J_{n,k}$ or the whole set $I_{n,k}$.
In particular, if $g_i : \mathbb{R} \to \mathbb{R}$ are bounded and measurable functions and

$$g(\mathrm{x}_1,\ldots,\mathrm{x}_k) = \prod_{i=1}^{k} g_i(\mathrm{x}_i),$$

then g is bounded and (A.4) implies that

$$\frac{1}{n^k} \sum_{j_1=1}^{n} \cdots \sum_{j_k=1}^{n} g(\mathrm{x}_{j_1},\ldots,\mathrm{x}_{j_k}) \overset{\text{a.s}}{\to} \mathbb{E}\left[g(\mathrm{x}_1,\ldots,\mathrm{x}_k)\,|\,\mathcal{F}_\infty\right] = \mathbb{E}\left[\prod_{i=1}^{k} g_i(\mathrm{x}_j)\,\middle|\,\mathcal{F}_\infty\right].$$

Noting that $\frac{1}{n^k}\sum_{j_1=1}^{n}\cdots\sum_{j_k=1}^{n} g(\mathrm{x}_{j_1},\ldots,\mathrm{x}_{j_k}) = \prod_{i=1}^{k}\left(\frac{1}{n}\sum_{j_i=1}^{n} g_i(\mathrm{x}_{j_i})\right)$, A.1 leads to

$$\frac{1}{n^k} \sum_{j_1=1}^{n} \cdots \sum_{j_k=1}^{n} g(\mathrm{x}_{j_1},\ldots,\mathrm{x}_{j_k}) \overset{\text{a.s}}{\to} \prod_{i=1}^{k} \mathbb{E}\left[g_i(\mathrm{x}_i)\,|\,\mathcal{F}_\infty\right]$$

and thus, almost surely

$$\mathbb{E}\left[\prod_{i=1}^{k} g_i(\mathrm{x}_i)\,\middle|\,\mathcal{F}_\infty\right] = \prod_{i=1}^{k} \mathbb{E}\left[g_i(\mathrm{x}_i)\,|\,\mathcal{F}_\infty\right]. \qquad \text{(A.5)}$$

This means that the sequence of random variables $\{g(\mathrm{x}_i)\}_{i=1}^k$ are i.i.d. given $\mathcal{F}_\infty$. For the particular choice of $g_i(x) = \delta_x(A_i)$ for $A_i \in \mathcal{F}_n$, (A.5) reads as

$$\begin{aligned}\mathbb{P}\left[\mathrm{x}_1 \in A_1, \ldots, \mathrm{x}_k \in A_k \,|\, \mathcal{F}_\infty\right] &= \mathbb{E}\left[\prod_{i=1}^k \delta_{\mathrm{x}_i}(A_i) \,\middle|\, \mathcal{F}_\infty\right] \\ &= \prod_{i=1}^k \mathbb{P}\left[\mathrm{x}_i \in A_i \,|\, \mathcal{F}_\infty\right] \\ &= \prod_{i=1}^k \tilde{p}(A_i). \end{aligned} \tag{A.6}$$

Using the tower property of conditional expectation and (A.6)

$$\begin{aligned}\mathbb{P}\left[\mathrm{x}_1 \in A_1, \ldots, \mathrm{x}_k \in A_k \,|\, \tilde{p}\right] &= \mathbb{E}\left[\mathbb{P}\left[\mathrm{x}_1 \in A_1, \ldots, \mathrm{x}_k \in A_k \,|\, \mathcal{F}_\infty\right] \,|\, \tilde{p}\right] \\ &= \mathbb{E}\left[\prod_{i=1}^k \tilde{p}(A_i) \,\middle|\, \tilde{p}\right] \\ &= \prod_{i=1}^k \tilde{p}(A_i). \end{aligned} \tag{A.7}$$

Let Q be the distribution of $\tilde{p}$ over $\mathrm{P}_\mathbb{R}$. Equation (A.7) implies that

$$\mathbb{P}\left[\mathrm{x}_1 \in A_1, \ldots, \mathrm{x}_k \in A_k\right] = \int_{\mathrm{P}_\mathbb{R}} \prod_{i=1}^k \tilde{p}(A_i) Q(\mathrm{d}\tilde{p}), \tag{A.8}$$

which finishes the proof. Finally, note that if the integral representation of (A.8) holds, then clearly the sequence is exchangeable as the product function is invariant under permutations. ■

Proof of 1.5. The idea is, as usual, to prove the statement for non-negative simple functions and then to extend it to arbitrary measurable functions by integration theory. Let $\{A_i\}_{i=1}^n$ be a measurable partition of $\mathbb{X}$ and $\{a_i\}_{i=1}^n \subseteq \mathbb{R}^+$, and consider the simple function

$$f(x) = \sum_{i=1}^n a_i \mathbb{1}_{A_i}(x).$$

By the definition of a Poisson process, the random variables $\{\mathrm{N}(A_i)\}_{i=1}^n$ are independent and $\mathrm{N}(A_i) \sim \text{Poisson}(\mu(A_i))$ for $i = 1, \ldots, n$. The moment-generating function of each $\mathrm{N}(A_i)$ is given by

$$\mathbb{E}\left[e^{t\mathrm{N}(A_i)}\right] = \exp\left[(e^t - 1)\mu(A_i)\right],$$

and thus

$$\begin{aligned}
\mathbb{E}\left[e^{t\Sigma}\right] &= \mathbb{E}\left[\exp\left(t\sum_{\upsilon\in\Upsilon}\sum_{i=1}^{n} a_i \mathbb{1}_{A_i}(\upsilon)\right)\right] \\
&= \mathbb{E}\left[\exp\left(\sum_{i=1}^{n} t a_i \mathrm{N}(A_i)\right)\right] \\
&= \prod_{i=1}^{n} \mathbb{E}\left[e^{t a_i \mathrm{N}(A_i)}\right] \\
&= \prod_{i=1}^{n} \exp\left[(e^{t a_i} - 1)\mu(A_i)\right] \\
&= \exp\left[\sum_{i=1}^{n} \int_{\mathbb{X}} \left(e^{t a_i} - 1\right) \mathbb{1}_{A_i}(\upsilon)\mu(\mathrm{d}x)\right] \\
&= \exp\left[\int_{\mathbb{X}} \left(e^{t\sum_{i=1}^{n} a_i \mathbb{1}_{A_i}(\upsilon)} - 1\right)\mu(\mathrm{d}\upsilon)\right] \\
&= \exp\left[\int_{\mathbb{X}} \left(e^{t f(\upsilon)} - 1\right)\mu(\mathrm{d}\upsilon)\right].
\end{aligned}$$

Now assume that f is a non-negative measurable function. Then f can be written as the limit of an increasing sequence of non-negative simple functions $\{f_n\}_{n=1}^{\infty}$ and by the Monotone Convergence Theorem

$$\begin{aligned}
\mathbb{E}\left[e^{t\Sigma}\right] &= \lim_{n\to\infty} \mathbb{E}\left[e^{t\sum_{\upsilon\in\Upsilon} f_n(\upsilon)}\right] \\
&= \lim_{n\to\infty} \exp\left[-\int_{\mathbb{X}} \left(1 - e^{t f_n(\upsilon)}\right)\mu(\mathrm{d}\upsilon)\right] \\
&= \exp\left[-\lim_{n\to\infty}\int_{\mathbb{X}} \left(1 - e^{t f_n(\upsilon)}\right)\mu(\mathrm{d}\upsilon)\right] \\
&= \exp\left[-\int_{\mathbb{X}} \left(1 - e^{t f(\upsilon)}\right)\mu(\mathrm{d}\upsilon)\right].
\end{aligned}$$

If $\int_{\mathbb{X}} \min(|f(x)|, 1)\mu(\mathrm{d}x) < \infty$ holds, then the right hand side of last equality converges and Σ is finite a.s. If not, then $\mathbb{E}[e^{t\Sigma}] = 0$ for $t < 0$ and thus $\Sigma = \infty$ a.s.

Lastly, if f is any measurable function, $f = f^+ - f^-$ where f^+, f^- are non-negative measurable functions. Σ is absolutely convergent if and only if

$$\Sigma_+ = \sum_{\upsilon\in\Upsilon_+} f^+(\upsilon) \quad \text{and} \quad \Sigma_- = \sum_{\upsilon\in\Upsilon_-} f^-(\upsilon)$$

are convergent, where

$$\Upsilon_+ = \{\upsilon \in \Upsilon : f(\upsilon) > 0\} \quad \text{and} \quad \Upsilon_- = \{\upsilon \in \Upsilon : f(\upsilon) < 0\}.$$

By the restriction theorem, Υ_+ and Υ_- are independent Poisson processes, as they are the restriction of Υ to disjoint subsets and hence

$$\begin{aligned}
\mathbb{E}\left[e^{t\Sigma}\right] &= \mathbb{E}\left[e^{t\Sigma_+ - t\Sigma_-}\right] \\
&= \mathbb{E}\left[e^{t\Sigma_+}\right]\mathbb{E}\left[e^{-t\Sigma_-}\right] \\
&= \exp\left[-\int_{\mathbb{X}}\left(1-e^{tf^+(\upsilon)}\right)\mu(\mathrm{d}\upsilon)\right]\exp\left[-\int_{\mathbb{X}}\left(1-e^{-tf^-(\upsilon)}\right)\mu(\mathrm{d}\upsilon)\right] \\
&= \exp\left[-\int\limits_{\{\upsilon:f(\upsilon)>0\}}\left(1-e^{tf^+(\upsilon)}\right)\mu(\mathrm{d}\upsilon) - \int\limits_{\{\upsilon:f(\upsilon)<0\}}\left(1-e^{tf^-(\upsilon)}\right)\mu(\mathrm{d}\upsilon)\right] \\
&= \exp\left[-\int_{\mathbb{X}}\left(1-e^{tf(\upsilon)}\right)\mu(\mathrm{d}\upsilon)\right].
\end{aligned}$$

This finishes the proof. ■

Proof of 1.6. For any measurable function f on $\mathbb{X}\times T$, define

$$\Sigma^* = \sum_{\upsilon\in\Upsilon} f(\upsilon, m_\upsilon)$$

and

$$f_*(\upsilon) = -\log\left(\int_T e^{-f(\upsilon,m)}p(\upsilon,\mathrm{d}m)\right).$$

Given Υ, Σ^* is a sum of independent random variables, hence using conditional expectation

$$\begin{aligned}
\mathbb{E}\left[e^{-\Sigma^*}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{-\Sigma^*}\mid\Upsilon\right]\right] \\
&= \mathbb{E}\left[\prod_{\upsilon\in\Upsilon}\mathbb{E}\left[e^{-f(\upsilon,m_\upsilon)}\mid\Upsilon\right]\right] \\
&= \mathbb{E}\left[\prod_{\upsilon\in\Upsilon}\int_T e^{-f(\upsilon,m)}p(\upsilon,\mathrm{d}m)\right] \\
&= \mathbb{E}\left[\prod_{\upsilon\in\Upsilon}\int_T e^{-f(\upsilon,m)}p(\upsilon,\mathrm{d}m)\right] \\
&= \mathbb{E}\left[e^{-\sum_{\upsilon\in\Upsilon}f_*(\upsilon)}\right].
\end{aligned}$$

By replacing f with f_* on the characteristic functional of Υ given on (1.1) we obtain

$$\begin{aligned}
\mathbb{E}\left[e^{-\Sigma^*}\right] &= \exp\left(-\int_{\mathbb{X}}\left(1-e^{-f_*(x)}\right)\mu(\mathrm{d}x)\right) \\
&= \exp\left(-\int_{\mathbb{X}}\left(\int_T p(\upsilon,\mathrm{d}m) - \int_T e^{-f(x,m)}p(\upsilon,\mathrm{d}m)\right)\mu(\mathrm{d}\upsilon)\right) \\
&= \exp\left(-\int_{\mathbb{X}}\int_T\left(1-e^{-f(\upsilon,m)}\right)p(\upsilon,\mathrm{d}m)\mu(\mathrm{d}\upsilon)\right) \\
&= \exp\left(-\int_{\mathbb{X}}\int_T\left(1-e^{-f(x,m)}\right)\mu^*(\mathrm{d}\upsilon,\mathrm{d}m)\right),
\end{aligned}$$

showing that Υ^* follows a Poisson process with mean measure μ^*. ■

Proof of 1.7. Before proving the main result, note that for any measurable collection of sets $(B_i)_{i=1}^n$ (not necessarily disjoint), using the independence property of $\tilde{\mu}$, we can write each $\tilde{\mu}(B_j)$ as a sum of elements of the form $\tilde{\mu}(C_1 \cap \cdots \cap C_n)$, where the sets C_i are either B_i or $\mathbb{X} \setminus B_i$ for $i, j = 1, \ldots, n$. Thus, we can recover the joint distribution $(\tilde{\mu}(B_1), \ldots, \tilde{\mu}(B_n))$ once we know the distribution of $\tilde{\mu}(A)$ for every $A \in \mathcal{X}$. A way to characterize this distribution for each A is given in the following proposition, due to Kingman (1967).

Proposition A.1. Let $\tilde{\mu}$ be a completely random measure on $(\mathbb{X}, \mathcal{X})$. For each measurable A and $t > 0$ define the *cumulant* of $\tilde{\mu}(A)$ as the function

$$\lambda_t(A) = -\log\left(\mathbb{E}\left[e^{-t\tilde{\mu}(A)}\right]\right).$$

Then

1. λ_t and $\tilde{\mu}$ are mutually absolutely continuous, i.e. $\lambda_t(A) = 0 \iff \tilde{\mu}(A) = 0$.
2. λ_t and $\tilde{\mu}$ are infinite or finite together.

Proof of A.1. Note that $\lambda_t(\emptyset) = -\log(\mathbb{E}[1]) = 0$. For a disjoint sequence of measurable sets $(A_i)_{i \geq 1}$, the independence property of $\tilde{\mu}$ implies that $\left(e^{-t\tilde{\mu}A_i)}\right)_{i \geq 1}$ are independent random variables and hence

$$\begin{aligned}
\lambda_t\left(\bigcup_{i=1}^{\infty} A_i\right) &= -\log\left(\mathbb{E}\left[e^{-t\tilde{\mu}\left(\bigcup_{i=1}^{\infty} A_i\right)}\right]\right) \\
&= -\log\left(\mathbb{E}\left[e^{-t\sum_{i=1}^{\infty}\tilde{\mu}(A_i)}\right]\right) \\
&= -\log\left(\mathbb{E}\left[\prod_{i=1}^{\infty} e^{-t\tilde{\mu}(A_i)}\right]\right) \\
&= -\log\left(\prod_{i=1}^{\infty}\mathbb{E}\left[e^{-t\tilde{\mu}(A_i)}\right]\right) \\
&= \sum_{i=1}^{\infty}\lambda_t(A_i),
\end{aligned}$$

so that $\lambda_t(\,\cdot\,)$ is a measure on $\mathbb{X}$ and further on, by construction, $0 \leq \lambda_t(\,\cdot\,) \leq \infty$.

To prove the first statement, note that if $\tilde{\mu}(A) = 0$ then $\lambda_t(A) = -\log(1) = 0$. On the other hand, $\lambda_t(A) = 0$ implies that $\mathbb{E}\left[e^{-t\tilde{\mu}(A)}\right] = 1$ and therefore $\tilde{\mu}(A) = 0$, since $t > 0$. Moving on to the second statement, note that

$$\lambda_t(A) = \infty \text{ a.s.} \iff e^{-t\tilde{\mu}(A)} = 0 \text{ a.s.} \iff \tilde{\mu}(A) = \infty \text{ a.s.}$$

■

As a first step, assume that λ_t is σ-finite for some t, which means that there exists a countable partition $\{S_i\}_{i=1}^{\infty} \subseteq \mathcal{X}$ such that $\lambda_t(S_i) < \infty$ for all i. By the first part of Proposition A.1, $\tilde{\mu}$ is Σ-finite. An immediate consequence of the second part of Proposition A.1 is that both λ_t and $\tilde{\mu}$ share the same atoms, namely if $\mathcal{A} = \{x \in \mathbb{X} : \lambda_t(\{x\}) > 0\}$ is the set of atoms of λ_t, then

$$x \in \mathcal{B} \iff \mathbb{P}\left[\tilde{\mu}(\{x\}) > 0\right] > 0.$$

λ_t being σ-finite leads to $\mathcal{A}$ being at most countable, hence the CRM $\tilde{\mu}$ has at most countable many fixed atoms. For any measurable set A, let

$$\tilde{\mu}_f(A) = \tilde{\mu}(A \cap \mathcal{A}) \quad \text{and} \quad \tilde{\mu}_1(A) = \tilde{\mu}(A \cap (\mathbb{X} \setminus \mathcal{A})).$$

Both $\tilde{\mu}_f$ and $\tilde{\mu}_1$ are completely random measures, as for pairwise disjoint measurable sets $\{A_i\}_{i=1}^d$, the sets $\{A_i \cap \mathcal{A}\}_{i=1}^d$ and $\{A_i \cap (\mathbb{X} \setminus \mathcal{A})\}_{i=1}^d$ are disjoint. Additionally, $\tilde{\mu}_f$ and $\tilde{\mu}_1$ are independent. We can now state the first decomposition for Σ-finite completely random measures as

$$\tilde{\mu} = \tilde{\mu}_f + \tilde{\mu}_1.$$

Setting $\varrho(x) := \tilde{\mu}(\{x\})$, the random variables $\{\varrho(x)\}_{x \in \mathcal{A}}$ are independent and we can express $\tilde{\mu}_f$ in terms of the independent masses $\varrho(x)$ as

$$\tilde{\mu}_f(\,\cdot\,) = \sum_{x \in \mathcal{A}} \varrho(x) \delta_x(\,\cdot\,).$$

The measure $\tilde{\mu}_1$ has no fixed atoms since for $y \in \mathbb{X}$, $\tilde{\mu}_1(\{y\})$ is either $\tilde{\mu}(\emptyset) = 0$ a.s. if $y \in \mathcal{A}$ or $\tilde{\mu}(\{y\})$ if $y \notin \mathcal{A}$, but in this case $\tilde{\mu}(\{y\}) = 0$ a.s. as well. Attention will be put on the non-fixed component $\tilde{\mu}_1$ of $\tilde{\mu}$, as the fixed atoms of λ_t that define $\tilde{\mu}_f$ can be removed, meaning that from now on $\tilde{\mu} = \tilde{\mu}_1$ or equivalently $\lambda_t(\,\cdot\,)$ is non-atomic. Before moving on, for completeness, we will state a definition and a result that will be used in what follows.

Definition A.1. A distribution μ is *infinitely divisible* if for each $n \in \mathbb{N}$ we can write μ as the n-fold self-convolution $\mu^{(n)} * \cdots * \mu^{(n)}$ of some distribution $\mu^{(n)}$.

The celebrated Lévy-Khintchine Theorem, whose proof can be found at Sato (1999), allows us to characterize infinitely divisible random variables by their characteristic function.

Theorem A.1. *(Lévy-Khintchine) A distribution μ on $\mathbb{R}$ is infinitely divisible if for any $u \in \mathbb{R}$ its characteristic function $\varphi(u)$ can be represented in the form*

$$\varphi(u) = exp\left[iau - \frac{\sigma^2 u^2}{2} - \int_{\mathbb{R}} \left(1 - e^{-iux} + iux \mathbb{1}_{\{|x|<1\}}\right) \nu(\mathrm{d}x)\right], \tag{A.9}$$

where $a \in \mathbb{R}$, $\sigma \geq 0$ and ν is a measure on $\mathbb{R} \setminus \{0\}$ such that $\int_{\mathbb{R}} \left(1 \wedge x^2\right) \nu(\mathrm{d}x) < \infty$. Conversely, if ν is a measure satisfying the last condition, $a \in \mathbb{R}$ and $\sigma \geq 0$, there exists an infinitely divisible distribution whose characteristic function is given by (A.9)*. We call a the drift, σ^2 the Gaussian variance, ν the Lévy measure and (a, σ^2, ν) the generating triplet.*

Now let $A \in \mathcal{X}$ be such that $\lambda_1(A) = a < \infty$. λ_1 being non-atomic and a special case of Lyapunov's Theorem [1] imply that for any $n \in \mathbb{N}$ there is a measurable partition $\{A_{n,j}\}_{j=1}^n$ of A such that $\lambda_1(A_{n,j}) = \frac{a}{n}$ and hence

$$\mathbb{E}\left[e^{-\tilde{\mu}(A_{n,j})}\right] = e^{-t\lambda_1(A_{n,j})} = e^{-\frac{a}{n}}.$$

This fact together with Markov's inequality lead to

$$\mathbb{P}\left[\tilde{\mu}(A_{n,j}) \geq c\right] \leq \frac{1 - e^{-\frac{a}{n}}}{1 - e^{-c}}$$

for $c > 0$, meaning that the array of random variables $\{\tilde{\mu}(A_{n,j})\}_{j=1}^n$ is uniformly asymptotically negligible. As we can express, for every $n \in \mathbb{N}$, $\tilde{\mu}(A)$ as a sum of independent random variables of the form

$$\tilde{\mu}(A) = \sum_{j=1}^{n} \tilde{\mu}(A_{n,j}),$$

the conclusion is that $\tilde{\mu}(A)$ is a infinitely divisible random variable. The Lévy-Khintchine formula for non-negative random variables allows us to represent the characteristic function of $\tilde{\mu}(A)$ as

$$\mathbb{E}\left[e^{-t\tilde{\mu}(A)}\right] = \exp\left[-\beta(A)t + \int_{\mathbb{R}_+} (1 - e^{tx})\nu(A, \mathrm{d}x)\right] \tag{A.10}$$

for $t > 0$, where $\nu(A, \cdot)$ is a Lévy measure on $\mathbb{R}_+$. In other words, we must have

$$\lambda_t(A) = \beta(A)t + \int_{\mathbb{R}_+} (1 - e^{tx})\nu(A, \mathrm{d}x) \tag{A.11}$$

and this relationship determines $\beta(A)$ and $\nu(A, \cdot)$ uniquely. From the σ-additivity of $\lambda_t(\cdot)$ it follows that for a disjoint sequence of measurable sets $\{A_i\}_{i \geq 1}$

$$\beta\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \beta\left(A_i\right) \quad \text{and} \quad \nu\left(\bigcup_{i=1}^{\infty} A_i, \cdot\right) = \sum_{i=1}^{\infty} \nu\left(A_i, \cdot\right).$$

[1] Let $\lambda_t : E \to \mathbb{R}$ $t = 1, \ldots, k$ be finite, σ-finite non-atomic measures on a measurable space $(E, \mathcal{E})$. Given any positive integer n there exists a measurable partition $\{A_{n,j}\}_{j=1}^n$ of E such that $\lambda_j(F_j) = \lambda_j(E)/n$, $j = 1, \ldots, k$.

This means that the distribution of a Σ-finite completely random measure with no fixed atoms can be characterized by $\lambda_t(A)$ given in (A.11) and (A.10), where β is a measure on $\mathbb{X}$ and ν is such that for $A \subseteq \mathbb{X}$, $\nu(A, \cdot)$ is a measure on $\mathbb{R}_+$ and for $B \subseteq \mathbb{R}_+$, $\nu(\cdot, B)$ is a measure on $\mathbb{X}$. Both β and ν are non-atomic as a consequence of λ_t being non-atomic.

The next step is to examine how to construct a completely random measure from β and ν, for which the Poisson process will be useful. For now, emphasis will be placed on analyzing the function ν. Define the measure μ over $\mathbb{X}$ as

$$\mu(A) = \nu(A, \mathbb{R}_+).$$

Additionally, suppose that μ is σ-finite. In this case, for $s > 0$ the measure

$$\mu_s(A) = \nu(A, (0, s])$$

is absolutely continuous with respect to μ and so has a Radon-Nikodym derivative $F(x, s)$, uniquely defined up to a μ-null set for each s, that satisfies

$$\mu_s(A) = \int_A F(x, s)\mu(\mathrm{d}x).$$

We will now see that $F(x, s)$ is the distribution function for some random variable. For each $s \in \mathbb{Q}_+$ let $F(x, s)$ be a version of the Radon-Nikodym derivative. If $s_1 < s_2$ are rational, then $\mu_{s_1} \leq \mu_{s_2}$, so that $F(x, s_1) \leq F(x, s_2)$ except on a μ-null set. Since $\mathbb{Q}_+$ is countable, $F(x, \cdot)$ is a non decreasing function of $s \in \mathbb{Q}_+$, and for any $s \in \mathbb{Q}_+$ and measurable A,

$$\mu_s(A) = \lim_{n\to\infty} \mu_{s+\frac{1}{n}}(A) = \lim_{n\to\infty} \int_A F\left(x, s + \frac{1}{n}\right) \mu(\mathrm{d}x) = \int_A F\left(x, s^+\right) \mu(\mathrm{d}x)$$

so that, for almost all x, F is right continuous. Lastly note that

$$\mu(A) = \lim_{n\to\infty} \mu_n(A) = \int_A \lim_{n\to\infty} F(x, n)\mu(\mathrm{d}x)$$

implies that $\lim_{n\to\infty} F(x, n) = 1$. We can extend this argument beyond $\mathbb{Q}_+$ to $\mathbb{R}_+$ by right continuity and therefore, for each fixed $x \in \mathbb{X}$, $F(x, s)$ is the distribution function of some random variable that takes values on $\mathbb{R}_+$.

Now consider the measure μ^* defined over the product space $\mathbb{X}^* = \mathbb{X} \times \mathbb{R}_+$ given by

$$\mu^*(A^*) = \iint_{A^*} \mathrm{d}F(x, z)\mu(\mathrm{d}x) \quad \forall A^* \subseteq \mathbb{X}^*.$$

If $A^* = A \times B$ with $A \in \mathcal{X}$ and $B \in \mathcal{B}(\mathbb{R}_+)$, μ^* becomes

$$\mu^*(A \times B) = \int_A \int_B \mathrm{d}F(x, z)\mu(\mathrm{d}x) = \int_B \mathrm{d}\nu(A, (0, s]) = \nu(A, B).$$

Although we cannot ensure that μ is σ-finite, we do know that ν must satisfy

$$\int \left(1 - e^{-z}\right) \nu(S_j, \mathrm{d}z) < \infty$$

over a partition $\{S_j\}_{j \geq 1}$, so that $\nu(S_j, (\epsilon, \infty]) < \infty$ for any $\epsilon > 0$. As a result we can define for $k \in \mathbb{N}$ the σ-finite measure

$$\mu^{(k)}(A) := \nu\left(A, \left[\frac{1}{k+1}, \frac{1}{k}\right)\right).$$

We can apply to each $\mu^{(k)}$ the previous argument that was developed for μ under the assumption of σ-finiteness. As we can express μ as $\mu(\,\cdot\,) = \sum_{k=1}^{\infty} \mu^{(k)}(\,\cdot\,)$, this proves that there is a measure μ^* on $\mathbb{X}^*$ such that

$$\mu^*(A \times B) = \nu(A, B) \quad \text{ for } A \in \mathcal{X}, B \in \mathcal{B}(\mathbb{R}_+).$$

Let Υ^* be a Poisson process on $\mathbb{X}^*$ with mean measure μ^* and set

$$\Psi(A) = \sum_{i=1}^{\infty} z_i \delta_A(x_i) \quad \text{ for } (x_i, z_i) \in \Upsilon^*.$$

Ψ is a purely atomic measure on $\mathbb{X}$ whose atoms correspond to the points of Υ^* and each atom x_i has mass z_i, distributed as $F(x_i, \,\cdot\,)$. From the definition of a Poisson process, for disjoint sets $(A_i)_{i=1}^d$, the random variables $(\Psi(A_i))_{i=1}^d$ are independent and Poisson distributed and therefore Ψ is a completely random measure. The distribution of Ψ can be computed by Campbell's Theorem for $t > 0$ and $A \in \mathcal{X}$ as

$$\mathbb{E}\left[e^{-t\Psi(A)}\right] = \exp\left[-\int_A \int_{\mathbb{R}_+} \left(1 - e^{-tx}\right) \mu^*(\mathrm{d}s, \mathrm{d}x)\right] = \exp\left[-\int_{\mathbb{R}_+} \left(1 - e^{-tx}\right) \nu(A, \mathrm{d}x)\right].$$

The right-hand side of last equality was encountered on the expression $\mathbb{E}\left[e^{-t\tilde{\mu}(A)}\right]$, given by (A.10). Therefore, $\tilde{\mu}(A)$ has the same distribution as $\beta(A) + \Psi(A)$. By adding back the fixed atoms the result follows. ■

Proof of 2.2. The EPPF is given by

$$\begin{aligned}\Phi_k^{(n)}(\mathrm{n}_1, \ldots, \mathrm{n}_k) &= \mathbb{E}\left[\int_{\mathbb{X}^k} \prod_{j=1}^{k} \tilde{p}^{\mathrm{n}_j}(\mathrm{d}x_j)\right] = \int_{\mathbb{X}^k} \mathbb{E}\left[\prod_{j=1}^{k} \tilde{p}^{\mathrm{n}_j}(\mathrm{d}x_j)\right] \\ &= \int_{\mathbb{X}^k} \mathbb{E}\left[\frac{\prod_{j=1}^{k} \tilde{\mu}^{\mathrm{n}_j}(\mathrm{d}x_j)}{\tilde{\mu}^n(\mathbb{X})}\right],\end{aligned}$$

where we've applied Fubini's Theorem and the definition of $\tilde{p}$. By the Gamma identity $\frac{1}{\beta^\alpha} = \frac{1}{\Gamma(\alpha)} \int_{\mathbb{R}_+} u^{\alpha-1} e^{-u\beta} \mathrm{d}u$, the integrand can be written as

$$\mathbb{E}\left[\frac{\prod_{j=1}^k \tilde{\mu}^{\mathrm{n}_j}(\mathrm{d}x_j)}{\tilde{\mu}^n(\mathbb{X})}\right] = \frac{1}{\Gamma(n)} \int_{\mathbb{R}_+} \mathbb{E}\left[e^{-u\tilde{\mu}(\mathbb{X})} \prod_{j=1}^k \tilde{\mu}^{\mathrm{n}_j}(\mathrm{d}x_j)\right] u^{n-1} \mathrm{d}u.$$

Let $\mathbb{X}^* := \mathbb{X} \setminus \{\mathrm{d}x_j\}_{j=1}^k$, so that $\mathbb{X}$ equals the disjoint union $\mathbb{X}^* \cup \{\mathrm{d}x_j\}_{j=1}^k$. Exploiting the independence property of $\tilde{\mu}$ and the definition of the Laplace transform we get

$$\begin{aligned}
\mathbb{E}\left[\frac{\prod_{j=1}^k \tilde{\mu}^{\mathrm{n}_j}(\mathrm{d}x_j)}{\tilde{\mu}^n(\mathbb{X})}\right] &= \frac{1}{\Gamma(n)} \int_{\mathbb{R}_+} \mathbb{E}\left[e^{-u\tilde{\mu}(\mathbb{X})} \prod_{j=1}^k \tilde{\mu}^{\mathrm{n}_j}(\mathrm{d}x_j)\right] u^{n-1} \mathrm{d}u \\
&= \frac{1}{\Gamma(n)} \int_{\mathbb{R}_+} \mathbb{E}\left[e^{-u\tilde{\mu}(\mathbb{X}^*)} \prod_{j=1}^k e^{-u\tilde{\mu}(\mathbb{X}^{*C})} \tilde{\mu}^{\mathrm{n}_j}(\mathrm{d}x_j)\right] u^{n-1} \mathrm{d}u \\
&= \frac{1}{\Gamma(n)} \int_{\mathbb{R}_+} \mathbb{E}\left[e^{-u\tilde{\mu}(\mathbb{X}^*)}\right] \prod_{j=1}^k \mathbb{E}\left[e^{-u\tilde{\mu}(\mathrm{d}x_j)} \tilde{\mu}^{\mathrm{n}_j}(\mathrm{d}x_j)\right] u^{n-1} \mathrm{d}u \\
&= \frac{1}{\Gamma(n)} \int_{\mathbb{R}_+} e^{-P_0(\mathbb{X}^*)\psi(u)} \prod_{j=1}^k \mathbb{E}\left[e^{-u\tilde{\mu}(\mathrm{d}x_j)} \tilde{\mu}^{\mathrm{n}_j}(\mathrm{d}x_j)\right] u^{n-1} \mathrm{d}u. \qquad \text{(A.12)}
\end{aligned}$$

But, for $j = 1, \ldots, k$ we have that

$$\begin{aligned}
\mathbb{E}\left[e^{-u\tilde{\mu}(\mathrm{d}x_j)} \tilde{\mu}^{\mathrm{n}_j}(\mathrm{d}x_j)\right] &= (-1)^{\mathrm{n}_j} \frac{\mathrm{d}^{\mathrm{n}_j}\left(\mathbb{E}[e^{-u\tilde{\mu}(\mathrm{d}x_j)}]\right)}{\mathrm{d}u^{\mathrm{n}_j}} \\
&= (-1)^{\mathrm{n}_j} \frac{\mathrm{d}^{\mathrm{n}_j}\left(e^{-P_0(\mathrm{d}x_j)\theta\Psi(u)}\right)}{\mathrm{d}u^{\mathrm{n}_j}} \\
&= (-1)^{\mathrm{n}_j} P_0(\mathrm{d}x_j) e^{-P_0(\mathrm{d}x_j)\theta\Psi(u)} \frac{\mathrm{d}^{\mathrm{n}_j}}{\mathrm{d}u^{\mathrm{n}_j}} \left(-\theta\Psi(u) + o\left(P_0(\mathrm{d}x_j)\right)\right).
\end{aligned}$$

The last equality comes from applying the Faa di Bruno's formula for the n_j-th derivative, with $f(u) = e^u$ and $g(u) = -P_0(\mathrm{d}x_j)\psi(u)$. Computing the n_j-th derivative of $-\psi(u)$ yields

$$\begin{aligned}
\frac{\mathrm{d}^{\mathrm{n}_j}\left(-\theta\Psi(u)\right)}{\mathrm{d}u^{\mathrm{n}_j}} &= -\frac{\mathrm{d}^{\mathrm{n}_j}}{\mathrm{d}u^{\mathrm{n}_j}} \left(\theta \int_{\mathbb{R}_+} \left(1 - e^{-uv}\right) \rho(\mathrm{d}v)\right) \\
&= -\theta \int_{\mathbb{R}_+} \frac{\mathrm{d}^{\mathrm{n}_j}\left(1 - e^{-uv}\right)}{\mathrm{d}u^{\mathrm{n}_j}} \rho(\mathrm{d}v) \\
&= \theta\,(-1)^{\mathrm{n}_j}\, \tau_{\mathrm{n}_j}(u).
\end{aligned}$$

Therefore

$$\mathbb{E}\left[e^{-u\tilde{\mu}(\mathrm{d}x_j)} \tilde{\mu}^{\mathrm{n}_j}(\mathrm{d}x_j)\right] = \theta P_0(\mathrm{d}x_j) e^{-P_0(\mathrm{d}x_j)\theta\Psi(u)} \tau_{\mathrm{n}_j}(u) + o\left(P_0(\mathrm{d}x_j)\right).$$

Substituting back these expressions at (A.12)

$$\mathbb{E}\left[e^{-u\tilde{\mu}(\mathrm{d}x_j)}\tilde{\mu}^{\mathrm{n}_j}(\mathrm{d}x_j)\right] = \frac{\theta^k}{\Gamma(n)}\int_{\mathbb{R}_+} e^{-\theta\Psi(u)\{P_0(A)+P_0(A^C)\}}\prod_{j=1}^{k}\tau_{\mathrm{n}_j}(u)u^{n-1}\mathrm{d}u\prod_{j=1}^{k}P_0(\mathrm{d}x_j) + o\left(\prod_{j=1}^{k}P_0(\mathrm{d}x_j)\right). \tag{A.13}$$

By integrating over $\mathbb{X}^k$ we obtain

$$\Phi_k^{(n)}(\mathrm{n}_1,\ldots,\mathrm{n}_k) = \int_{\mathbb{X}^k}\mathbb{E}\left[\frac{\prod_{j=1}^{k}\tilde{\mu}^{\mathrm{n}_j}(\mathrm{d}x_j)}{\tilde{\mu}^n(\mathbb{X})}\right] = \frac{\theta^k}{\Gamma(n)}\int_{\mathbb{R}_+} e^{-\theta\Psi(u)}\prod_{j=1}^{k}\tau_{\mathrm{n}_j}(u)u^{n-1}\mathrm{d}u.$$

This finishes the proof. ■

Appendix B

Proof of Theorem 3.2. For any $i \neq j$, by the tower property of conditional expectation, one has that

$$\begin{aligned}\operatorname{cov}\left(\tilde{p}_i(A), \tilde{p}_j(A)\right) &= \mathbb{E}\mathbb{E}\left[\tilde{p}_i(A)\tilde{p}_j(A) \,|\, \tilde{p}_0\right] - \left(\mathbb{E}\mathbb{E}\left[\tilde{p}_i(A) \,|\, \tilde{p}_0\right]\right)\left(\mathbb{E}\mathbb{E}\left[\tilde{p}_j(A) \,|\, \tilde{p}_0\right]\right) \\ &= \mathbb{E}\left[\mathbb{E}\left[\tilde{p}_i(A) \,|\, \tilde{p}_0\right]\mathbb{E}\left[\tilde{p}_j(A) \,|\, \tilde{p}_0\right]\right] - \left(\mathbb{E}[\tilde{p}_0(A)]\right)^2.\end{aligned}$$

To simplify notation, consider

$$\mathcal{J}_m = \int_{\mathbb{R}_+} u e^{-\theta\Psi(u)}\tau_m(u)\mathrm{d}u \quad \mathcal{J}_{m,0} = \int_{\mathbb{R}_+} u e^{-\theta_0\Psi_0(u)}\tau_{m,0}(u)\mathrm{d}u$$

for $m \geq 1$. Proposition 2.1 states that, for any $A \in \mathcal{X}$,

$$\begin{aligned}\operatorname{var}\left(\tilde{p}_0(A)\right) &= \theta_0 P_0(A)(1 - P_0(A))\mathcal{J}_{2,0} \\ \operatorname{var}\left(\tilde{p}_i(A) \,|\, \tilde{p}_0\right) &= \theta\tilde{p}_0(A)(1 - \tilde{p}_0(A))\mathcal{J}_2.\end{aligned}$$

Hence, for any $i \in \{1, \ldots, m\}$,

$$\begin{aligned}\operatorname{var}(\tilde{p}_i(A)) &= \mathbb{E}\left[\operatorname{var}(\tilde{p}_i(A) \,|\, \tilde{p}_0)\right] + \operatorname{var}(\tilde{p}_0(A)) \\ &= \theta\mathcal{J}_2\mathbb{E}\left[\tilde{p}_0(A)(1 - \tilde{p}_0(A))\right] + \operatorname{var}(\tilde{p}_0(A)) \\ &= \theta_0 P_0(A)(1 - P_0(A))\left\{\theta\theta_0\mathcal{J}_2\mathcal{J}_{1,0} + \mathcal{J}_{2,0}\right\}.\end{aligned}$$

Therefore

$$\begin{aligned}\operatorname{cov}\left(\tilde{p}_i(A), \tilde{p}_j(A)\right) &= \mathbb{E}\left[\mathbb{E}\left[\tilde{p}_i(A) \,|\, \tilde{p}_0\right]\mathbb{E}\left[\tilde{p}_j(A) \,|\, \tilde{p}_0\right]\right] - \left(\mathbb{E}[\tilde{p}_0(A)]\right)^2 \\ &= \mathbb{E}[\tilde{p}_0^2(A)] - \left(\mathbb{E}[\tilde{p}_0(A)]\right)^2 = \operatorname{var}(\tilde{p}_0(A)) \\ &= \theta_0 P_0(A)(1 - P_0(A))\mathcal{J}_{2,0}.\end{aligned}$$

It is now easy to see that

$$\begin{aligned}\operatorname{corr}\left(\tilde{p}_i(A), \tilde{p}_j(A)\right) &= \frac{\operatorname{cov}\left(\tilde{p}_i(A), \tilde{p}_j(A)\right)}{\sqrt{\operatorname{var}(\tilde{p}_i(A))\operatorname{var}(\tilde{p}_j(A))}} \\ &= 1 + \frac{\mathcal{J}_{2,0}}{\theta\theta_0\mathcal{J}_2\mathcal{J}_{1,0}},\end{aligned}$$

and the result follows. ■

Proof of Theorem 3.3. First we need a technical lemma regarding the n-th derivative of $e^{-m\Psi(u)}$.

Lemma B.1. If $\tau_m(u) = \int_0^\infty v^m e^{-uv}\rho(\mathrm{d}v)$ for $m \geq 1$ and

$$\varrho_{n,i}(u) = \sum_{(q_1,\ldots,q_i)\in\mathcal{C}^i_{[n]}} \frac{1}{i!}\binom{n}{q_1,\cdots,q_i}\prod_{t=1}^{i}\tau_{q_t}(u),$$

then the following relationship holds

$$(-1)^n\frac{\mathrm{d}^n e^{-m\Psi(u)}}{\mathrm{d}u^n} = e^{-m\Psi(u)}\sum_{i=1}^{n} m^i\varrho_{n,i}(u).$$

Proof of B.1. According to Faà di Bruno's formula

$$\begin{aligned}\frac{\mathrm{d}^n e^{-m\Psi(u)}}{\mathrm{d}u^n} &= \sum_{\pi\in\mathcal{P}_{[n]}}\frac{\mathrm{d}^{|\pi|}}{\mathrm{d}u^{|\pi|}}\left(e^{-mx}\right)\big|_{x=\Psi(u)}\prod_{A\in\pi}\frac{\mathrm{d}^{|A|}\Psi(u)}{\mathrm{d}u^{|A|}}\\ &= \sum_{i=1}^{n}(-m)^i e^{-m\Psi(u)}\sum_{\pi\in\mathcal{P}^i_{[n]}}\prod_{A\in\pi}\frac{\mathrm{d}^{|A|}\Psi(u)}{\mathrm{d}u^{|A|}}.\end{aligned}$$

As to each unordered partition $\pi \in \mathcal{P}^i_{[n]}$ there are $i!$ ordered partitions, obtained by permuting the elements of π, last equality can be re written as

$$\frac{\mathrm{d}^n e^{-m\Psi(u)}}{\mathrm{d}u^n} = \sum_{i=1}^{n}(-m)^i e^{-m\Psi(u)}\frac{1}{i!}\sum_{(\circ)}\prod_{A\in\pi}\frac{\mathrm{d}^{|A|}\Psi(u)}{\mathrm{d}u^{|A|}},$$

where $(\circ)$ denotes that we are summing over ordered partitions of $[n]$ into i blocks.

The derivative of $\Psi(u)$ depends only on the cardinality of each of the elements of the partition π, and the number of partitions $\pi \in \mathcal{P}^i_{[n]}$ that have an associated composition $(q_1,\ldots,q_i) \in \mathcal{C}^i_{[n]}$ equals to $\frac{n!}{q_1!q_2!\cdots q_i!}$, that is, the multinomial coefficient $\binom{n}{q_1,\cdots,q_i}$. In the proof of Proposition 2.2 we proved that $(-1)^m\tau_m(u) = \frac{\mathrm{d}^m\Psi(u)}{\mathrm{d}u^m}$, so that

$$\sum_{(\circ)}\prod_{A\in\pi}\frac{\mathrm{d}^{|A|}\Psi(u)}{\mathrm{d}u^{|A|}} = \sum_{(q_1,\ldots,q_i)\in\mathcal{C}^i_{[n]}}\frac{1}{i!}\binom{n}{q_1,\cdots,q_i}\frac{\mathrm{d}^{q_1}\Psi(u)}{\mathrm{d}u^{q_1}}\cdots\frac{\mathrm{d}^{q_i}\Psi(u)}{\mathrm{d}u^{q_i}} = \vartheta_{n,i}(u).$$

Finally

$$\frac{\mathrm{d}^n e^{-m\Psi(u)}}{\mathrm{d}u^n} = e^{-m\Psi(u)}\sum_{i=1}^{n}(-m)^i\vartheta_{n,i}(u).$$

■

In view of this, for $x_1 \neq \cdots \neq x_k \in \mathbb{X}$, define

$$M_{\mathbf{n}_1,\ldots,\mathbf{n}_m}(\mathrm{d}x_1,\ldots,\mathrm{d}x_k) = \mathbb{E}\left[\prod_{j=1}^{k}\prod_{i=1}^{m}\tilde{p}_i^{n_{i,j}}(\mathrm{d}x_j)\right],$$

so that an application of Fubini's theorem allows us to write the pEPPF as

$$\begin{aligned}\boldsymbol{\Pi}_k^{(N)}(\mathbf{n}_1,\ldots,\mathbf{n}_m) &= \mathbb{E}\left[\int_{\mathbb{X}^k}\prod_{j=1}^{k}\prod_{i=1}^{m}\tilde{p}_i^{n_{i,j}}(\mathrm{d}x_j)\right]\\ &= \int_{\mathbb{X}^k} M_{\mathbf{n}_1,\ldots,\mathbf{n}_m}(\mathrm{d}x_1,\ldots,\mathrm{d}x_k).\end{aligned}$$

Let $\epsilon > 0$ be such that the balls with center at x_j and radius ϵ,

$$A_{j,\epsilon} = \{y \in \mathbb{X} : d_{\mathbb{X}}(y, x_j) < \epsilon\}, \tag{B.1}$$

where $d_{\mathbb{X}}$ is the distance in $\mathbb{X}$, are pairwise disjoint. Using the tower property of conditional expectation and the fact that $\tilde{p}_i$ are conditionally independent given $\tilde{p}_0$, we can write

$$\begin{aligned}M_{\mathbf{n}_1,\ldots,\mathbf{n}_m}(A_{1,\epsilon}\times\cdots\times A_{k,\epsilon}) &= \mathbb{E}\left[\prod_{j=1}^{k}\prod_{i=1}^{m}\tilde{p}_i^{n_{i,j}}(A_{j,\epsilon})\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[\prod_{j=1}^{k}\prod_{i=1}^{m}\tilde{p}_i^{n_{i,j}}(A_{j,\epsilon})\,\middle|\,\tilde{p}_0\right]\right]\\ &= \mathbb{E}\left[\prod_{i=1}^{m}\mathbb{E}\left[\prod_{j=1}^{k}\tilde{p}_i^{n_{i,j}}(A_{j,\epsilon})\,\middle|\,\tilde{p}_0\right]\right]\\ &= \mathbb{E}\left[\prod_{i=1}^{m}\mathbb{E}\left[\frac{\prod_{j=1}^{k}\tilde{\mu}_i^{n_{i,j}}(A_{j,\epsilon})}{\tilde{\mu}_i^{n_i}(\mathbb{X})}\,\middle|\,\tilde{p}_0\right]\right].\end{aligned} \tag{B.2}$$

Let $\mathbb{X}_\epsilon^* = \mathbb{X}\setminus\{A_{j,\epsilon}\}_{j=1}^k$, so that $\mathbb{X} = \mathbb{X}_\epsilon^* \cup \{A_{j,\epsilon}\}_{j=1}^k$ is a disjoint union. By the Gamma identity and the independence property of each CRM $\tilde{\mu}_i$

$$\begin{aligned}\mathbb{E}\left[\frac{\prod_{j=1}^{k}\tilde{\mu}_i^{n_{i,j}}(A_{j,\epsilon})}{\tilde{\mu}_i^{n_i}(\mathbb{X})}\,\middle|\,\tilde{p}_0\right] &= \frac{1}{\Gamma(n_i)}\int_{\mathbb{R}_+}\mathbb{E}\left[e^{-u\tilde{\mu}_i(\mathbb{X})}\prod_{j=1}^{k}\tilde{\mu}_i^{n_{i,j}}(A_{j,\epsilon})\,\middle|\,\tilde{p}_0\right]u^{n_i-1}\mathrm{d}u\\ &= \frac{1}{\Gamma(n_i)}\int_{\mathbb{R}_+}\mathbb{E}\left[e^{-u\tilde{\mu}_i(\mathbb{X}_\epsilon^*)}\prod_{j=1}^{k}\tilde{\mu}_i^{n_{i,j}}(A_{j,\epsilon})e^{-u\tilde{\mu}_i(A_{j,\epsilon})}\,\middle|\,\tilde{p}_0\right]u^{n_i-1}\mathrm{d}u\\ &= \frac{1}{\Gamma(n_i)}\int_{\mathbb{R}_+}\mathbb{E}\left[e^{-u\tilde{\mu}_i(\mathbb{X}_\epsilon^*)}\,|\,\tilde{p}_0\right]\\ &\quad\times\prod_{j=1}^{k}\mathbb{E}\left[\tilde{\mu}_i^{n_{i,j}}(A_{j,\epsilon})e^{-u\tilde{\mu}_i(A_{j,\epsilon})}\,\middle|\,\tilde{p}_0\right]u^{n_i-1}\mathrm{d}u.\end{aligned}$$

The expressions inside the integral correspond to the Laplace transform of $\tilde{\mu}_i(\mathbb{X}_\epsilon^*)$ and the $n_{i,j}$-th derivative of $\tilde{\mu}_i(A_{j,\epsilon})$ respectively. By the virtue of Lemma B.1, $M_{\mathbf{n}_1,\dots,\mathbf{n}_m}(A_{1,\epsilon} \times \cdots \times A_{k,\epsilon})$ can be expressed as

$$\begin{aligned}
M_{\mathbf{n}_1,\dots,\mathbf{n}_m} &= \mathbb{E}\prod_{i=1}^{m}\frac{1}{\Gamma(n_i)}\int_{\mathbb{R}_+} e^{-\theta\Psi(u)\tilde{p}_0(\mathbb{X}_\epsilon^*)}\prod_{j=1}^{k}\left[(-1)^{n_{i,j}}\frac{\mathrm{d}^{n_{i,j}}e^{-\theta\Psi(u)\tilde{p}_0(A_{j,\epsilon})}}{\mathrm{d}u^{n_{i,j}}}\right]u^{n_i-1}\mathrm{d}u \\
&= \mathbb{E}\left[\prod_{i=1}^{m}\frac{1}{\Gamma(n_i)}\int_{\mathbb{R}_+} e^{-\theta\Psi(u)}\prod_{j=1}^{k}\sum_{\ell_{i,j}=1}^{n_{i,j}}\theta^{\ell_{i,j}}\tilde{p}_0(A_{j,\epsilon})^{\ell_{i,j}}\varrho_{n_{i,j},\ell_{i,j}}(u)u^{n_i-1}\mathrm{d}u\right] \\
&= \sum_{\boldsymbol{\ell}}\mathbb{E}\left[\prod_{j=1}^{k}\tilde{p}_0^{\bar{\ell}_{\bullet j}}(A_{j,\epsilon})\right]\prod_{i=1}^{m}\left(\frac{\theta^{\bar{\ell}_{i\bullet}}}{\Gamma(n_i)}\int_{\mathbb{R}_+}u^{n_i-1}e^{-\theta\Psi(u)}\prod_{j=1}^{k}\varrho_{n_{i,j},\ell_{i,j}}(u)\mathrm{d}u\right).
\end{aligned}$$

As $\epsilon \downarrow 0$, using equation (A.13) from the proof of Proposition 2.2 we can neglect the superior order terms due to P_0 being non-atomic, so that

$$\begin{aligned}
\mathbb{E}\left[\prod_{j=1}^{k}\tilde{p}_0^{\bar{\ell}_{\bullet j}}(\mathrm{d}x_j)\right] &= \mathbb{E}\left[\frac{\prod_{j=1}^{k}\tilde{\mu}_0^{\bar{\ell}_{\bullet j}}(\mathrm{d}x_j)}{\tilde{\mu}_0^{|\boldsymbol{\ell}|}(\mathbb{X})}\right] \\
&= \left(\prod_{j=1}^{k}P_0(\mathrm{d}x_j)\right)\int_{\mathbb{R}_+}u^{|\boldsymbol{\ell}|-1}e^{-\theta_0\Psi_0(u)}\prod_{j=1}^{k}\tau_{\bar{\ell}_{\bullet j},0}(u)\mathrm{d}u \\
&= \left(\prod_{j=1}^{k}P_0(\mathrm{d}x_j)\right)\Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\bar{\ell}_{\bullet 1},\dots,\bar{\ell}_{\bullet k}).
\end{aligned}$$

This means that

$$\begin{aligned}
M_{\mathbf{n}_1,\dots,\mathbf{n}_m}(\mathrm{d}x_1,\dots,\mathrm{d}x_k) = \left(\prod_{j=1}^{k}P_0(\mathrm{d}x_j)\right)\sum_{\boldsymbol{\ell}}\Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\bar{\ell}_{\bullet 1},\dots,\bar{\ell}_{\bullet k}) \\
\times\prod_{i=1}^{m}\underbrace{\left(\frac{\theta^{\bar{\ell}_{i\bullet}}}{\Gamma(n_i)}\int_{\mathbb{R}_+}u^{n_i-1}e^{-\theta\Psi(u)}\prod_{j=1}^{k}\varrho_{n_{i,j},\ell_{i,j}}(u)\mathrm{d}u\right)}_{=\mathcal{J}_{n_i}}.
\end{aligned}$$

By the definition of $\varrho_{n_{i,j},\ell_{i,j}}(u)$, one has that

$$\begin{aligned}
\mathcal{J}_{n_i} &= \frac{\theta^{\bar{\ell}_{i\bullet}}}{\Gamma(n_i)}\int_{\mathbb{R}_+}u^{n_i-1}e^{-\theta\Psi(u)}\prod_{j=1}^{k}\sum_{\mathbf{q}}\frac{1}{\ell_{i,j}!}\binom{n_{i,j}}{q_{i,j,1},\cdots,q_{i,j,\ell_{i,j}}}\prod_{t=1}^{\ell_{i,j}}\tau_{q_{i,j,t}}(u)\mathrm{d}u \\
&= \sum_{\mathbf{q}}\prod_{j=1}^{k}\frac{1}{\ell_{i,j}!}\binom{n_{i,j}}{q_{i,j,1},\cdots,q_{i,j,\ell_{i,j}}}\frac{\theta^{\bar{\ell}_{i\bullet}}}{\Gamma(n_i)}\int_{\mathbb{R}_+}u^{n_i-1}e^{-\theta\Psi(u)}\prod_{j=1}^{k}\prod_{t=1}^{\ell_{i,j}}\tau_{q_{i,j,t}}(u)\mathrm{d}u \\
&= \sum_{\mathbf{q}}\prod_{j=1}^{k}\frac{1}{\ell_{i,j}!}\binom{n_{i,j}}{q_{i,j,1},\cdots,q_{i,j,\ell_{i,j}}}\Phi_{\bar{\ell}_{i\bullet},i}^{(n_i)}(\mathbf{q}_{i,1},\dots,\mathbf{q}_{i,k}). \qquad \text{(B.3)}
\end{aligned}$$

This reads as

$$M_{\mathbf{n}_1,\ldots,\mathbf{n}_m}(\mathrm{d}x_1,\ldots,\mathrm{d}x_k) = \left(\prod_{j=1}^{k} P_0(\mathrm{d}x_j)\right) \sum_{\boldsymbol{\ell}} \sum_{\mathbf{q}} \Phi_{k,\,0}^{(|\boldsymbol{\ell}|)}(\bar{\ell}_{\bullet 1},\ldots,\bar{\ell}_{\bullet k}) \times \prod_{i=1}^{m}\prod_{j=1}^{k} \frac{1}{\ell_{i,j}!}\binom{n_{i,j}}{q_{i,j,1},\cdots,q_{i,j,\ell_{i,j}}} \Phi_{\bar{\ell}_{i\bullet},i}^{(n_i)}(\mathbf{q}_{i,1},\ldots,\mathbf{q}_{i,k}).$$

Integrating over $\mathbb{X}^k$ yields the desired result. ■

Proof of Theorem 3.4. Introduce the quantities $\bar{n}_j$, where

$$\bar{n}_0 \equiv 0 \quad \text{and} \quad \bar{n}_i = \sum_{j=1}^{i} n_j.$$

Suppose that $\bar{\pi}_1,\ldots,\bar{\pi}_m$ denote independent random partitions of $\mathbb{N}$ such that the restriction of $\bar{\pi}_i$ to $\{\bar{n}_{i-1}+1,\ldots,\bar{n}_i\}$, which will be written as $\bar{\pi}_{i,n_i}$, has a probability distribution $\Phi_{\zeta_i,i}^{n_i}$ as defined in equation (3.7). Additionally, $\tilde{\pi}_0$ is a random partition of $\mathbb{N}$ such that, conditional on $(\tilde{\pi}_{1,n_1},\ldots,\tilde{\pi}_{m,n_m})$, its restriction $\tilde{\pi}_{0,h}$ to $[h]$ has probability distribution $\Phi_{k,0}^{(h)}$ in (3.6), where $h = \sum_{i=1}^{m} |\tilde{\pi}_{i,n_i}|$, that is, the sum of the number of blocks in each partition $\tilde{\pi}_{i,n_i}$. Considering the restriction of $\tilde{\pi}_i$ to $\{\bar{n}_{i-1}+1,\ldots,\bar{n}_i\}$ instead of $[n_i]$ allows us to scroll the indexes of the elements of the blocks, by taking into account the number of observations placed in the previous groups, i.e. the number of costumers that have sat down in the previous $i-1$ restaurants.

Expressing Theorem 3.3 in terms of partitions instead of compositions, one has that

$$\Pi_k^{(N)}(\mathbf{n}_1,\ldots,\mathbf{n}_m) = \sum_{h=k}^{N}\prod_{i=1}^{m} \mathbb{P}\left[\tilde{\pi}_{i,n_i} = \{B_{i,1},\ldots,B_{i,\zeta_i}\}\right] \times \sum \mathbb{P}\left[\tilde{\pi}_{0,h} = \{C_1,\ldots,C_k\} \,\middle|\, \bigcap_{i=1}^{m}\{\tilde{\pi}_{i,n_i} = \{B_{i,1},\ldots,B_{i,\zeta_i}\}\}\right]$$

where, for any $h \in \{k,\ldots,N\}$, the sums are taken over all partitions such that $\sum_{i=1}^{m}\zeta_i = h$, and

$$\sum_{\{t:\bar{\zeta}_{i-1}+1\in C_j\}\cap\{1,\ldots,\zeta_i\}} |B_{i,t}| = n_{i,j},$$

where $\bar{\zeta}_0 \equiv 0$ and $\bar{\zeta}_i = \sum_{r=1}^{i}\zeta_r$, for each $i = 1,\ldots,m$. If $\{t : \bar{\zeta}_{i-1}+1 \in C_j\}\cap\{1,\ldots,\zeta_i\} = \emptyset$, then $n_{i,j} = 0$, meaning that in the Chinese restaurant franchise metaphor, the j-th dish is not served at restaurant i. According to this, the number of distinct table labels at each restaurant i is given by $\mathrm{K}_{i,n_i} = |\tilde{\pi}_{i,n_i}|$, whereas on the root level of the hierarchy $\mathrm{K}_{0,h} = |\tilde{\pi}_{0,h}|$. This implies that

$$\mathbb{P}[\mathrm{K}_N = k] = \mathbb{P}\left[\bigcup_{t=k}^{N}\bigcup_{(t_1,\ldots,t_d)\in\mathcal{C}_t^m} \{\mathrm{K}_{1,n_1} = t_1,\ldots \mathrm{K}_{m,n_m} = t_m\} \cap \{\mathrm{K}_{0,t} = k\}\right].$$

The conclusion follows from the fact that the random variables $\{\mathrm{K}_{i,n_i}\}_{i=1}^m$ are independent, and that

$$\mathbb{P}\left[\mathrm{K}_{0,t}=k\,\Big|\,\bigcap_{i=1}^m\{\mathrm{K}_{i,n_i}=t_i\}\right]=\mathbb{P}[\mathrm{K}_{0,t}=k]\mathbb{1}_{\mathcal{C}^m_{[t]}}(t_1,\ldots,t_m).$$

■

Proof of Theorem 3.5. The goal is to determine the posterior Laplace functional of $\tilde{\mu}_0$ for any measurable function $f:\mathbb{X}\to\mathbb{R}_+$ via a limiting argument. Consider the sets $A_{j,\epsilon}$ as in the proof of Theorem 3.3. The posterior Laplace functional equals

$$\mathbb{E}\left[e^{-\tilde{\mu}_0(f)}\,\Big|\,\mathbf{x}\right]=\lim_{\epsilon\downarrow 0}\frac{\mathbb{E}\left[e^{-\tilde{\mu}_0(f)}\prod_{j=1}^k\prod_{i=1}^m\tilde{p}_i^{n_{i,j}}(A_{j,\epsilon})\right]}{\mathbb{E}\left[\prod_{j=1}^k\prod_{i=1}^m\tilde{p}_i^{n_{i,j}}(A_{j,\epsilon})\right]}. \tag{B.4}$$

The denominator is $M_{\mathbf{n}_1,\ldots,\mathbf{n}_m}(A_{1,\epsilon}\times\cdots\times A_{k,\epsilon})$, which we proved that, as $\epsilon\downarrow 0$, equals to

$$\left(\prod_{j=1}^k P_0(A_{j,\epsilon})\right)\sum_{\boldsymbol{\ell},\mathbf{q}}\Phi^{(|\boldsymbol{\ell}|)}_{K,0}(\bar{\ell}_{\bullet 1},\ldots,\bar{\ell}_{\bullet K})\times\prod_{i=1}^m\prod_{j=1}^k\frac{1}{\ell_{i,j}!}\binom{n_{i,j}}{q_{i,j,1},\cdots,q_{i,j,\ell_{i,j}}}\Phi^{(n_i)}_{\bar{\ell}_{i\bullet},i}(\mathbf{q}_{i,1},\ldots,\mathbf{q}_{i,k})$$

plus $\lambda_{k,\epsilon}$, where $\lambda_{k,\epsilon}=o\left(\prod_{j=1}^k P_0(A_{j,\epsilon})\right)$.

The numerator of (B.4) can be determined by following along the same steps, that is conditioning with respect to $\tilde{\mu}_0$, setting $\mathbb{X}^*_\epsilon=\mathbb{X}\setminus\{A_{j,\epsilon}\}_{j=1}^k$ and recalling the expression for $\mathcal{J}_{n_i}$ that we found in (B.3)

$$\begin{aligned}
\mathbb{E}\left[e^{-\tilde{\mu}_0(f)}\prod_{j=1}^k\prod_{i=1}^m\tilde{p}_i^{n_{i,j}}(A_{j,\epsilon})\right]&=\mathbb{E}\left[e^{-\tilde{\mu}_0(f)}\prod_{i=1}^m\mathbb{E}\left[\frac{\prod_{j=1}^k\tilde{\mu}_i^{n_{i,j}}(A_{j,\epsilon})}{\tilde{\mu}_i^{n_i}(\mathbb{X})}\,\Big|\,\tilde{\mu}_0\right]\right]\\
&=\sum_{\boldsymbol{\ell}}\prod_{i=1}^m\underbrace{\left(\frac{\theta^{\bar{\ell}_{i\bullet}}}{\Gamma(n_i)}\int_{\mathbb{R}_+}u^{n_i-1}e^{-\theta\Psi(u)}\prod_{j=1}^k\varrho_{n_{i,j},\ell_{i,j}}(u)\mathrm{d}u\right)}_{\mathcal{J}_{n_i}}\\
&\quad\times\mathbb{E}\left[e^{-\tilde{\mu}_0(f)}\prod_{j=1}^k\tilde{p}_0^{\bar{\ell}_{\bullet j}}(A_{j,\epsilon})\right]\\
&=\sum_{\boldsymbol{\ell},\mathbf{q}}\prod_{i=1}^m\prod_{j=1}^k\frac{1}{\ell_{i,j}!}\binom{n_{i,j}}{q_{i,j,1},\cdots,q_{i,j,\ell_{i,j}}}\Phi^{(n_i)}_{\bar{\ell}_{i\bullet},i}(\mathbf{q}_{i,1},\ldots,\mathbf{q}_{i,k})\\
&\quad\times\mathbb{E}\left[e^{-\tilde{\mu}_0(f)}\prod_{j=1}^k\tilde{p}_0^{\bar{\ell}_{\bullet j}}(A_{j,\epsilon})\right].
\end{aligned}\tag{B.5}$$

Now note that, as $\epsilon\downarrow 0$,

$$\mathbb{E}\left[e^{-\tilde{\mu}_0(f)}\prod_{j=1}^{k}\tilde{p}_0^{\bar{\ell}_{\bullet j}}(A_{j,\epsilon})\right] = \mathbb{E}\left[e^{-\tilde{\mu}_0(f)}\frac{\prod_{j=1}^{k}\tilde{\mu}_0^{\bar{\ell}_{\bullet j}}(A_{j,\epsilon})}{\tilde{\mu}_0^{|\boldsymbol{\ell}|}(\mathbb{X})}\right]$$

$$= \frac{1}{\Gamma(|\boldsymbol{\ell}|)}\int_{\mathbb{R}_+}\mathbb{E}\left[e^{-\int_{\mathbb{X}}(f(x)+u)\tilde{\mu}_0(\mathrm{d}x)}\prod_{j=1}^{k}\tilde{\mu}_0^{\bar{\ell}_{\bullet j}}(A_{j,\epsilon})\right]u^{|\boldsymbol{\ell}|-1}\mathrm{d}u$$

$$= \frac{1}{\Gamma(|\boldsymbol{\ell}|)}\int_{\mathbb{R}_+}\mathbb{E}u^{|\boldsymbol{\ell}|-1}e^{-\int_{\mathbb{X}_\epsilon^*}(f(x)+u)\tilde{\mu}_0(\mathrm{d}x)}$$

$$\times\prod_{j=1}^{k}\tilde{\mu}_0^{\bar{\ell}_{\bullet j}}(A_{j,\epsilon})e^{-\int_{A_{j,\epsilon}}(f(x)+u)\tilde{\mu}_0(\mathrm{d}x)}\mathrm{d}u$$

$$= \frac{1}{\Gamma(|\boldsymbol{\ell}|)}\int_{\mathbb{R}_+}u^{|\boldsymbol{\ell}|-1}\mathbb{E}\left[e^{-\tilde{\mu}_0\left((f+u)\mathbb{1}_{X_\epsilon^*}\right)}\right]$$

$$\times\prod_{j=1}^{k}\mathbb{E}\left[\tilde{\mu}_0^{\bar{\ell}_{\bullet j}}(A_{j,\epsilon})e^{-\tilde{\mu}_0\left((f+u)\mathbb{1}_{A_{j,\epsilon}}\right)}\right]\mathrm{d}u$$

$$= \frac{\theta_0^k\prod_{j=1}^{k}P_0(A_{j,\epsilon})}{\Gamma(|\boldsymbol{\ell}|)}\int_{\mathbb{R}_+}u^{|\boldsymbol{\ell}|-1}e^{-\theta_0\Psi_0(f+u)}$$

$$\times\prod_{j=1}^{k}\tau_{\bar{\ell}_{\bullet j}}\left(u+f(\mathrm{x}_\mathrm{j}^*)\right)\mathrm{d}u+\lambda_{k,\epsilon}. \tag{B.6}$$

Combining (B.5) and (B.6) one has that

$$\mathbb{E}\left[e^{-\tilde{\mu}_0(f)}\prod_{j=1}^{k}\prod_{i=1}^{m}\tilde{p}_i^{n_{i,j}}(A_{j,\epsilon})\right] = \prod_{j=1}^{k}P_0(A_{j,\epsilon})$$

$$\times\sum_{\boldsymbol{\ell},\mathbf{q}}\prod_{i=1}^{m}\prod_{j=1}^{k}\frac{1}{\ell_{i,j}!}\binom{n_{i,j}}{q_{i,j,1},\cdots,q_{i,j,\ell_{i,j}}}\Phi_{\bar{\ell}_{i\bullet},i}^{(n_i)}(\mathbf{q}_{i,1},\ldots,\mathbf{q}_{i,k})$$

$$\times\frac{\theta_0^k}{\Gamma(|\boldsymbol{\ell}|)}\int_{\mathbb{R}_+}u^{|\boldsymbol{\ell}|-1}e^{-\theta_0\Psi_0(f+u)}\prod_{j=1}^{k}\tau_{\bar{\ell}_{\bullet j}}\left(u+f(\mathrm{x}_\mathrm{j}^*)\right)\mathrm{d}u+\lambda_{k,\epsilon}.$$

Hence, by making $\epsilon\downarrow 0$ and further conditioning over $\mathbf{t}$ we can conclude that

$$\mathbb{E}\left[e^{-\tilde{\mu}_0(f)}\,\middle|\,\mathbf{x},\mathbf{t}\right]\to\frac{\int_{\mathbb{R}_+}u^{|\boldsymbol{\ell}|-1}e^{-\theta_0\Psi_0(f+u)}\prod_{j=1}^{k}\tau_{\bar{\ell}_{\bullet j}}\left(u+f(\mathrm{x}_\mathrm{j}^*)\right)\mathrm{d}u}{\Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\bar{\ell}_{\bullet 1},\ldots,\bar{\ell}_{\bullet k})}.$$

Finally, note that

$$\Phi_{K,0}^{(|\boldsymbol{\ell}|)}(\bar{\ell}_{\bullet 1},\ldots,\bar{\ell}_{\bullet k}) = \int_{\mathbb{R}_+}u^{|\boldsymbol{\ell}|-1}e^{-\theta_0\Psi_0(u)}\prod_{j=1}^{k}\tau_{\bar{\ell}_{\bullet j},0}(u)\mathrm{d}u = \int_{\mathbb{R}_+}f_0(u\,|\,\mathbf{x},\mathbf{t})\mathrm{d}u.$$

Hence the denominator of the posterior Laplace functional is the normalizing constant of the density $f_0(u \,|\, \mathbf{x}, \mathbf{t})$. The results follows from here by taking into account the definition of $\tau_{\bar{\ell}_{\bullet j}}(u + f(\mathrm{x}_{\mathrm{j}}^*))$. ■

Proof of Theorem 3.6. By the tower property of conditional expectation, one has that

$$\mathbb{E}\left[e^{-\sum_{i=1}^m \tilde{\mu}_i(f_i)} \,\middle|\, \mathbf{x}, \mathbf{t}\right] = \mathbb{E}\left[\mathbb{E}\left[e^{-\sum_{i=1}^m \tilde{\mu}_i(f_i)} \,\middle|\, \mathbf{x}, \mathbf{t}, \tilde{\mu}_0\right] \,\middle|\, \mathbf{x}, \mathbf{t}\right],$$

for any collection of measurable functions $f_1, \ldots, f_m$, where $f_i : \mathbb{X} \to \mathbb{R}_+$, for $i = 1, \ldots, m$. Hence the proof consists on calculating the posterior Laplace functional

$$\mathbb{E}\left[e^{-\sum_{i=1}^m \tilde{\mu}_i(f_i)} \,\middle|\, \mathbf{x}, \mathbf{t}, \tilde{\mu}_0\right] = \lim_{\epsilon \downarrow 0} \frac{\mathbb{E}\left[e^{-\sum_{i=1}^m \tilde{\mu}_i(f_i)} \prod_{i=1}^m \prod_{j=1}^k \tilde{p}_i^{n_{i,j}}(A_{j,\epsilon}) \,\middle|\, \mathbf{t}, \tilde{\mu}_0\right]}{\mathbb{E}\left[\prod_{i=1}^m \prod_{j=1}^k \tilde{p}_i^{n_{i,j}}(A_{j,\epsilon}) \,\middle|\, \mathbf{t}, \tilde{\mu}_0\right]}, \tag{B.7}$$

where the sets $A_{j,\epsilon}$ are as in (B.1). By following along the same steps as in the proof of Theorem 3.3, the denominator $\mathbb{E}\left[\prod_{i=1}^m \prod_{j=1}^k \tilde{p}_i^{n_{i,j}}(A_{j,\epsilon}) \,\middle|\, \mathbf{t}, \tilde{\mu}_0\right]$ equals to

$$\left(\prod_{j=1}^k \prod_{i=1}^m \tilde{p}_0^{\bar{\ell}_{i,j}}(A_{j,\epsilon})\right) \prod_{i=1}^m \frac{\theta^{\bar{\ell}_{i\bullet}}}{\Gamma(n_i)} \int_{\mathbb{R}_+} u^{n_i - 1} e^{-\theta\Psi(u)} \prod_{j=1}^k \prod_{t=1}^{\ell_{i,j}} \tau_{q_{i,j,t}}(u) \mathrm{d}u.$$

As far as the numerator is concerned, by following a similar reasoning as the one that led to (B.6), one has that

$$\begin{aligned}
\mathbb{E}\left[e^{-\sum_{i=1}^m \tilde{\mu}_i(f_i)} \prod_{i=1}^m \prod_{j=1}^k \tilde{p}_i^{n_{i,j}}(A_{j,\epsilon}) \,\middle|\, \mathbf{t}, \tilde{\mu}_0\right] &= \mathbb{E}\left[\prod_{i=1}^m \prod_{j=1}^k e^{-\tilde{\mu}_i(f_i)} \tilde{p}_i^{n_{i,j}}(A_{j,\epsilon}) \,\middle|\, \mathbf{t}, \tilde{\mu}_0\right] \\
&= \prod_{i=1}^m \mathbb{E}\left[e^{-\tilde{\mu}_i(f_i)} \prod_{j=1}^k \tilde{p}_i^{n_{i,j}}(A_{j,\epsilon}) \,\middle|\, \mathbf{t}, \tilde{\mu}_0\right] \\
&= \left(\prod_{j=1}^k \prod_{i=1}^m \tilde{p}_0^{\ell_{i,j}}(A_{j,\epsilon})\right) \\
&\times \prod_{i=1}^m \frac{\theta^{\bar{\ell}_{i\bullet}}}{\Gamma(n_i)} \times \int_{\mathbb{R}_+} u^{n_i - 1} e^{-\theta\tilde{\Psi}(u + f_i)} \\
&\times \prod_{j=1}^k \prod_{i=1}^m \tau_{q_{i,j,t}}(u + f_i(\mathrm{x}_j^*)) \mathrm{d}u,
\end{aligned}$$

where

$$\tilde{\Psi}(f) = \int_{\mathbb{X}} \int_{\mathbb{R}_+} \left(1 - e^{-vf(x)}\right) \rho(v) \mathrm{d}v \tilde{p}_0(\mathrm{d}x).$$

Thus, as $\epsilon \downarrow 0$

$$\mathbb{E}\left[e^{-\sum_{i=1}^m \tilde{\mu}_i(f_i)} \,\middle|\, \mathbf{x}, \mathbf{t}, \tilde{\mu}_0\right] \to \prod_{i=1}^m \frac{\int_{\mathbb{R}_+} u^{n_i - 1} e^{-\theta\tilde{\Psi}(u + f_i)} \prod_{j=1}^k \prod_{i=1}^m \tau_{q_{i,j,t}}(u + f_i(\mathrm{x}_j^*)) \mathrm{d}u}{\int_{\mathbb{R}_+} u^{n_i - 1} e^{-\theta\Psi(u)} \prod_{j=1}^k \prod_{t=1}^{\ell_{i,j}} \tau_{q_{i,j,t}}(u) \mathrm{d}u}.$$

If we further condition over $\mathbf{U}$, this entails

$$\mathbb{E}\left[e^{-\sum_{i=1}^{m}\tilde{\mu}_i(f_i)}\,\middle|\,\mathbf{x},\mathbf{t},\mathbf{U},\tilde{\mu}_0\right]=\prod_{i=1}^{m}\prod_{j=1}^{k}\prod_{t=1}^{\ell_{i,j}}\frac{\tau_{q_{i,j,t}}(U_i+f_i(\mathrm{x}_j^*))}{\tau_{q_{i,j,t}}(U_i)}$$
$$\times\prod_{i=1}^{m}\exp\left\{-\theta\int_{\mathbb{X}}\int_{\mathbb{R}_+}\left(1-e^{-vf_i(x)}\right)e^{-vU_i}\rho(v)\mathrm{d}v\tilde{p}_0(\mathrm{d}x)\right\}.$$

■

Proof of Theorem 3.7. The same steps of the proof of Theorem 3.2 allows us to conclude that for $i\neq j$ and $A\in\mathcal{X}$

$$\mathrm{cov}\left(\tilde{p}_i(A),\tilde{p}_j(A)\right)=\mathrm{var}(\tilde{p}_0(A)).$$

By taking into account the polynomial tilting that defines the Pitman-Yor process, one has that

$$\begin{aligned}\mathrm{var}(\tilde{p}_0(A))&=P_0(A)(1-P_0(A))\frac{\sigma_0^2(1-\sigma_0)}{\theta_0(\theta_0+1)\Gamma\left(\frac{\theta_0}{\sigma_0}\right)}\int_{\mathbb{R}_+}u^{\theta_0+\sigma_0-1}e^{u^{\sigma_0}}\mathrm{d}u\\&=\frac{1-\sigma_0}{\theta_0+1}P_0(A)(1-P_0(A)).\end{aligned}$$

Furthermore, the equality $\mathrm{var}(\tilde{p}_i(A))=\mathbb{E}\left[\mathrm{var}(\tilde{p}_i(A)\,|\,\tilde{p}_0)\right]+\mathrm{var}(\tilde{p}_0(A))$ holds true and therefore

$$\begin{aligned}\mathrm{var}(\tilde{p}_i(A))&=\frac{1-\sigma}{\theta+1}\mathbb{E}\left[\tilde{p}_0(A)(1-\tilde{p}_0(A))\right]+\frac{1-\sigma_0}{\theta_0+1}P_0(A)(1-P_0(A))\\&=\frac{P_0(A)(1-P_0(A))}{\theta_0+1}\left\{(1-\sigma_0)+(\theta_0+\sigma_0)\frac{1-\sigma}{\theta+1}\right\}.\end{aligned}$$

This entails

$$\begin{aligned}\mathrm{corr}\left(\tilde{p}_i(A),\tilde{p}_j(A)\right)&=\frac{\mathrm{cov}\left(\tilde{p}_i(A),\tilde{p}_j(A)\right)}{\sqrt{\mathrm{var}(\tilde{p}_i(A))\mathrm{var}(\tilde{p}_j(A))}}\\&=\frac{1-\sigma_0}{(1-\sigma_0)+(\theta_0+\sigma_0)\frac{1-\sigma}{\theta+1}}\\&=1+\frac{1-\sigma_0}{1-\sigma}\frac{\theta+1}{\theta_0+\sigma_0}.\end{aligned}$$

■

Proof of Theorem 3.8. Following the same notation as in the proof of Theorem 3.3, one has that

$$\mathbf{\Pi}_k^{(N)}(\mathbf{n}_1,\ldots,\mathbf{n}_m)=\int_{\mathbb{X}^k}M_{\mathbf{n}_1,\ldots,\mathbf{n}_m}(\mathrm{d}x_1,\ldots,\mathrm{d}x_k).$$

Consider the sets $A_{j,\epsilon}$ as in equation (B.1). The change of measure (3.8) yields

$$\begin{aligned}\mathbb{E}\left[\prod_{j=1}^{k}\tilde{p}_i^{n_{i,j}}(A_{j,\epsilon})\right] &= \mathbb{E}\left[\mathbb{E}\left[\prod_{j=1}^{k}\tilde{p}_i^{n_{i,j}}(A_{j,\epsilon})\,\Big|\,\tilde{p}_0\right]\right]\\ &= \frac{\Gamma(\theta+1)}{\Gamma\left(\frac{\theta}{\sigma}+1\right)}\frac{1}{\Gamma(\theta+n_i)}\int_{\mathbb{R}_+}u^{\theta+n_i-1}\mathbb{E}\mathbb{E}\left[e^{-u\tilde{\mu}_i(\mathbb{X})}\prod_{j=1}^{k}\tilde{\mu}_i^{n_{i,j}}(A_{j,\epsilon})\,\Big|\,\tilde{p}_0\right]\mathrm{d}u\end{aligned}$$

where, conditional on $\tilde{p}_0$, each $\tilde{\mu}_i$ is a σ-stable CRM with $\mathbb{E}[\tilde{p}_i\,|\,\tilde{p}_0]=\tilde{p}_0$. Recalling that for a σ-stable CRM, $\Psi(u)=u^\sigma$, from here it follows that

$$\begin{aligned}M_{\mathbf{n}_1,\ldots,\mathbf{n}_m}(A_{1,\epsilon}\times\cdots\times A_{k,\epsilon}) &= \mathbb{E}\left[\prod_{i=1}^{m}\mathbb{E}\left[\prod_{j=1}^{k}\tilde{p}_i^{n_{i,j}}(A_{j,\epsilon})\,\Big|\,\tilde{p}_0\right]\right]\\ &= \frac{\Gamma^m(\theta+1)}{\Gamma^m\left(\frac{\theta}{\sigma}+1\right)}\mathbb{E}\prod_{i=1}^{m}\frac{1}{\Gamma(\theta+n_i)}\int_{\mathbb{R}_+}u^{\theta+n_i-1}e^{-u^\sigma}\\ &\times\prod_{j=1}^{k}\sum_{\ell_{i,j}=1}^{n_{i,j}}\tilde{p}_0^{\ell_{i,j}}(A_{j,\epsilon})\varrho_{n_{i,j},\ell_{i,j}}(u)\mathrm{d}u\\ &= \frac{\Gamma^m(\theta+1)}{\Gamma^m\left(\frac{\theta}{\sigma}+1\right)}\sum_{\boldsymbol{\ell}}\left[\mathbb{E}\prod_{j=1}^{k}\tilde{p}_0^{\bar{\ell}_{\bullet j}}(A_{j,\epsilon})\right]\\ &\times\prod_{i=1}^{m}\left(\frac{1}{\Gamma(n_i)}\int_{\mathbb{R}_+}u^{n_i-1}e^{-u^\sigma}\prod_{j=1}^{k}\varrho_{n_{i,j},\ell_{i,j}}(u)\mathrm{d}u\right),\end{aligned}$$

where

$$\varrho_{n,k}(u)=\frac{\sigma^k}{u^{n-k\sigma}}\sum_{(q_1,\ldots,q_k)\in\mathcal{C}^k_{[n]}}\frac{1}{k!}\binom{n}{q_1,\cdots,q_k}\prod_{t=1}^{k}(1-\sigma)_{q_t-1}=\frac{\sigma^k}{u^{n-k\sigma}}S^\sigma_{n,k}. \tag{B.8}$$

As P_0 is assumed to be non-atomic,

$$\mathbb{E}\prod_{j=1}^{k}\tilde{p}_0^{\bar{\ell}_{\bullet j}}(A_{j,\epsilon})=\prod_{j=1}^{k}P_0(A_{j,\epsilon})\frac{\prod_{i=1}^{k-1}(\theta_0+i\sigma_0)}{(\theta_0+1)_{|\boldsymbol{\ell}|-1}}\prod_{j=1}^{k}(1-\sigma_0)_{\bar{\ell}_{\bullet j}-1}+\lambda_{k,\epsilon}.$$

Hence, as $\epsilon \downarrow 0$

$$
\begin{aligned}
M_{\mathbf{n}_1,\ldots,\mathbf{n}_m}(A_{1,\epsilon}\times\cdots\times A_{k,\epsilon}) \to{}& \frac{\Gamma^m(\theta+1)}{\Gamma^m\left(\frac{\theta}{\sigma}+1\right)}\sum_{\boldsymbol{\ell}}\prod_{j=1}^{k}P_0(A_{j,\epsilon})\frac{\prod_{i=1}^{k-1}(\theta_0+i\sigma_0)}{(\theta_0+1)_{|\boldsymbol{\ell}|-1}}\\
&\times\prod_{j=1}^{k}(1-\sigma_0)_{\bar{\ell}_{\bullet j}-1}\prod_{i=1}^{m}\left(\sigma^{\bar{\ell_{i\bullet}}-1}\frac{\Gamma\left(\bar{\ell_{i\bullet}}+\frac{\theta}{\sigma}\right)}{\Gamma(n_i+\theta)}\prod_{j=1}^{k}S^{\sigma}_{n_{i,j},\ell_{i,j}}\right)\\
={}&\sum_{\boldsymbol{\ell}}\prod_{j=1}^{k}P_0(A_{j,\epsilon})\frac{\prod_{i=1}^{k-1}(\theta_0+i\sigma_0)}{(\theta_0+1)_{|\boldsymbol{\ell}|-1}}\prod_{j=1}^{k}(1-\sigma_0)_{\bar{\ell}_{\bullet j}-1}\\
&\times\prod_{i=1}^{m}\left(\frac{\Gamma(\theta+1)}{\Gamma(n_i+\theta)}\frac{\sigma^{\bar{\ell_{i\bullet}}-1}\Gamma\left(\frac{\theta}{\sigma}+1+\bar{\ell_{i\bullet}}-1\right)}{\Gamma\left(\frac{\theta}{\sigma}+1\right)}\prod_{j=1}^{k}S^{\sigma}_{n_{i,j},\ell_{i,j}}\right)\\
={}&\sum_{\boldsymbol{\ell}}\prod_{j=1}^{k}P_0(A_{j,\epsilon})\frac{\prod_{i=1}^{k-1}(\theta_0+i\sigma_0)}{(\theta_0+1)_{|\boldsymbol{\ell}|-1}}\prod_{j=1}^{k}(1-\sigma_0)_{\bar{\ell}_{\bullet j}-1}\\
&\times\prod_{i=1}^{m}\left(\frac{\prod_{r=1}^{\bar{\ell}_{i\bullet}}(\theta+r\sigma)}{(\theta+1)_{n_i-1}}\prod_{j=1}^{k}S^{\sigma}_{n_{i,j},\ell_{i,j}}\right).
\end{aligned}
$$

The result follows by integrating over $\mathbb{X}^k$. ■

Proof of Theorem 3.9. Following along the exact same steps as in 3.4 allows us to conclude that

$$
\mathbb{P}[\mathrm{K}_N=k]=\mathbb{P}\left[\bigcup_{t=k}^{N}\bigcup_{(t_1,\ldots,t_d)\in\mathcal{C}^m_t}\{\mathrm{K}_{1,n_1}=t_1,\ldots\mathrm{K}_{m,n_m}=t_d\}\cap\{\mathrm{K}_{0,t}=k\}\right],
$$

with the only difference being that the random variables K_{i,n_i} are the number of distinct values corresponding to independent random partitions generated from $\mathrm{PY}(\sigma,\theta,G)$ for some diffuse probability measure G, while $\mathrm{K}_{0,t}$ is the number of blocks of a random partition induced by $\mathrm{PY}(\sigma,\theta,P_0)$. Hence the result easily follows from the fact that

$$
\mathbb{P}[\mathrm{K}_N=k]=\sum_{t=k}^{N}\mathbb{P}\left[\mathrm{K}_{0,t}=k\right]\sum_{(\zeta_1,\ldots,\zeta_d)\in\mathcal{C}^m_{[t]}}\prod_{i=1}^{m}\mathbb{P}\left[\mathrm{K}_{i,n_i}=\zeta_i\right]
$$

and that the probability of the number of blocks being equal to k on a random partition generated by a Pitman-Yor process is given by $\frac{\prod_{i=1}^{k-1}(\theta+i\sigma)}{(\theta+1)_{n-1}}S^{\sigma}_{n,k}$. ■

Proof of Theorem 3.10. The proof consists on computing the posterior Laplace functional given in (B.4) for any measurable function $f:\mathbb{X}\to\mathbb{R}_+$.

Let ς_i and ς_0 denote stable CRMs with parameters σ and σ_0 in $(0,1)$ respectively, where the base measure of ς_0 is a nonatomic measure P_0 and ς_i, conditional of ς_0, are independent and identically distributed CRMs with base measure $\frac{\varsigma_0}{\varsigma_0(\mathbb{X})}$. To compute the numerator, first note that

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{m}\frac{\prod_{j=1}^{k}\tilde{\mu}_i^{n_{i,j}}(A_{j,\epsilon})}{\tilde{\mu}_i^{n_i}(\mathbb{X})}\,\Big|\,\tilde{\mu}_0\right] &= \frac{\Gamma(\theta+1)}{\Gamma\left(\frac{\theta}{\sigma}+1\right)}\mathbb{E}\left[\prod_{j=1}^{k}\frac{\varsigma^{n_{i,j}}(A_{j,\epsilon})}{\varsigma_i^{n_i}(\mathbb{X})}\,\Big|\,\tilde{\mu}_0\right]\\
&= \frac{\Gamma\left(\frac{\theta}{\sigma}+1\right)^{-1}}{(\theta+1)_{n_i-1}}\int_{\mathbb{R}_+}u^{\theta+n_i-1}\mathbb{E}\left[e^{-u\varsigma_i(\mathbb{X})}\prod_{j=1}^{k}\varsigma_i^{n_{i,j}}(A_{j,\epsilon})\,\Big|\,\tilde{\mu}_0\right]\mathrm{d}u\\
&= \frac{\Gamma\left(\frac{\theta}{\sigma}+1\right)^{-1}}{(\theta+1)_{n_i-1}}\int_{\mathbb{R}_+}u^{\theta+n_i-1}e^{-u^\sigma}\prod_{j=1}^{k}\tilde{p}_0^{\ell_{i,j}}(A_{j,\epsilon})\varrho_{n_{i,j},\ell_{i,j}}(u)\mathrm{d}u.
\end{aligned}$$

Therefore one has that the numerator equals

$$\begin{aligned}
\mathbb{E}\left[e^{-\tilde{\mu}_0(f)}\prod_{j=1}^{k}\prod_{i=1}^{m}\tilde{p}_i^{n_{i,j}}(A_{j,\epsilon})\right] &= \mathbb{E}\left[e^{-\tilde{\mu}_0(f)}\mathbb{E}\left[\prod_{i=1}^{m}\frac{\prod_{j=1}^{k}\tilde{\mu}_i^{n_{i,j}}(A_{j,\epsilon})}{\tilde{\mu}_i^{n_i}(\mathbb{X})}\,\Big|\,\tilde{\mu}_0\right]\right]\\
&= \frac{1}{(\theta+1)_{n_i-1}}\frac{1}{\Gamma\left(\frac{\theta}{\sigma}+1\right)}\\
&\times\int_{\mathbb{R}_+}u^{\theta+n_i-1}e^{-u^\sigma}\prod_{j=1}^{k}\sum_{\ell_{i,j}=1}^{n_{i,j}}\tilde{p}_0^{\ell_{i,j}}(A_{j,\epsilon})\varrho_{n_{i,j},\ell_{i,j}}(u)\mathrm{d}u\\
&\times\mathbb{E}\left[e^{-\tilde{\mu}_0(f)}\prod_{j=1}^{k}\tilde{p}_0^{\bar{\ell}_{\bullet j}}(A_{j,\epsilon})\right],
\end{aligned}$$

where $\varrho_{n,k}$ is as in (B.8). Now it remains only to compute

$$\begin{aligned}
\mathbb{E}\left[e^{-\tilde{\mu}_0(f)}\prod_{j=1}^{k}\tilde{p}_0^{\bar{\ell}_{\bullet j}}(A_{j,\epsilon})\right] &= \frac{\Gamma(\theta_0+1)}{\Gamma\left(\frac{\theta_0}{\sigma_0}+1\right)}\mathbb{E}\left[\varsigma_0(\mathbb{X})^{-\theta_0}e^{-\varsigma_0(f)}\frac{\prod_{j=1}^{k}\varsigma_0(A_{j,\epsilon})^{\bar{\ell}_{\bullet j}}}{\varsigma_0(\mathbb{X})^{|\boldsymbol{\ell}|}}\right]\\
&= \frac{\sigma_0^k\Gamma\left(\frac{\theta_0}{\sigma_0}+1\right)^{-1}}{(\theta_0+1)_{|\boldsymbol{\ell}|-1}}\left(\prod_{j=1}^{k}P_0(A_{j,\epsilon})(1-\sigma_0)_{\bar{\ell}_{\bullet j}-1}\right)\\
&\times\int_{\mathbb{R}_+}z^{\theta_0+|\boldsymbol{\ell}|-1}e^{-\int_{\mathbb{R}_+}(z+f(x))^{\sigma_0}P_0(\mathrm{d}x)}\prod_{j=1}^{k}(z+f(\mathrm{x}_j^*))^{-\bar{\ell}_{\bullet j}-\sigma_0}\mathrm{d}z\\
&+\lambda_{k,\epsilon},
\end{aligned}$$

where $\lambda_{k,\epsilon}=o\left(\prod_{j=1}^{k}P_0(A_{j,\epsilon})\right)$.

Conditioning over $\mathbf{t}$ we get

$$\mathbb{E}\left[e^{-\tilde{\mu}_0(f)}\prod_{j=1}^{k}\prod_{i=1}^{m}\tilde{p}_i^{n_{i,j}}(A_{j,\epsilon})\,\middle|\,\mathbf{t}\right] = \frac{\sigma_0^k\Gamma\left(\frac{\theta_0}{\sigma_0}+1\right)^{-1}}{(\theta_0+1)_{|\boldsymbol{\ell}|-1}}\prod_{i=1}^{m}\frac{\prod_{r=1}^{\bar{\ell}_{i\bullet}-1}(\theta+r\sigma)}{(\theta+1)_{n_i-1}}$$
$$\times\prod_{j=1}^{k}P_0(A_{j,\epsilon})(1-\sigma_0)_{\bar{\ell}_{\bullet j}}\prod_{i=1}^{m}S^{\sigma}_{n_{i,j},\ell_{i,j}}$$
$$\times\int_{\mathbb{R}_+}z^{\theta_0+k\sigma_0-1}e^{-z^{\sigma_0}}e^{-\int_{\mathbb{R}_+}(z+f(x))^{\sigma_0}P_0(\mathrm{d}x)}$$
$$\times\left(\prod_{j=1}^{k}\mathbb{E}\left[e^{-I_jf(\mathrm{x}_j^*)}\right]\right)\mathrm{d}z$$

on the numerator. Noting that the denominator is $M_{\mathbf{n}_1,\ldots,\mathbf{n}_m}(A_{1,\epsilon}\times\cdots\times A_{k,\epsilon})$, which has been identified in the proof of Theorem 3.8, leads to the result. ■

Proof of Theorem 3.11. Again we aim at determining the posterior Laplace functional given in (B.7) for any collection of measurable functions $f_1,\ldots,f_m$ on $\mathbb{X}$. The numerator $\mathbb{E}\left[e^{-\sum_{i=1}^{m}\tilde{\mu}_i(f_i)}\prod_{i=1}^{m}\prod_{j=1}^{k}\tilde{p}_i^{n_{i,j}}(A_{j,\epsilon})\,\middle|\,\mathbf{t},\tilde{\mu}_0\right]$ coincides with

$$\prod_{i=1}^{m}\mathbb{E}\left[e^{-\tilde{\mu}_i(f_i)}\prod_{j=1}^{k}\tilde{p}_i^{n_{i,j}}(A_{j,\epsilon})\,\middle|\,\mathbf{t},\tilde{\mu}_0\right] = \frac{1}{(\theta+1)_{n_i-1}}\frac{1}{\Gamma\left(\frac{\theta}{\sigma}+1\right)}$$
$$\times\int_{\mathbb{R}_+}v^{\theta+n_i-1}\mathbb{E}\left[e^{-\varsigma_i(v+f_i)}\prod_{j=1}^{k}\varsigma_i^{n_{i,j}}(A_{j,\epsilon})\,\middle|\,\mathbf{t},\tilde{\mu}_0\right]\mathrm{d}v$$
$$=\frac{1}{(\theta+1)_{n_i-1}}\frac{1}{\Gamma\left(\frac{\theta}{\sigma}+1\right)}\prod_{j=1}^{k}\tilde{p}_0^{\ell_{i,j}}(A_{j,\epsilon})\sigma^{\ell_{i,j}}S^{\sigma}_{n_{i,j},\ell_{i,j}}$$
$$\times\int_{\mathbb{R}_+}v^{\theta+n_i-1}e^{-\int_{\mathbb{R}_+}(v+f(x))^{\sigma}\tilde{p}_0(\mathrm{d}x)}$$
$$\times\prod_{j=1}^{k}(v+f_i(\mathrm{x}_j^*))^{-(n_{i,j}-\ell_{i,j}\sigma)}\mathrm{d}v$$
$$=\frac{1}{(\theta+1)_{n_i-1}}\frac{1}{\Gamma\left(\frac{\theta}{\sigma}+1\right)}\prod_{j=1}^{k}\tilde{p}_0^{\ell_{i,j}}(A_{j,\epsilon})\sigma^{\ell_{i,j}}S^{\sigma}_{n_{i,j},\ell_{i,j}}$$
$$\times\int_{\mathbb{R}_+}v^{\theta+\bar{\ell}_{i\bullet}\sigma-1}e^{-\int_{\mathbb{R}_+}(v+f(x))^{\sigma}-v^{\sigma}\tilde{p}_0(\mathrm{d}x)}$$
$$\times\prod_{j=1}^{k}\left(1+\frac{f_i(\mathrm{x}_j^*)}{v}\right)^{-(n_{i,j}-\ell_{i,j}\sigma)}\mathrm{d}v$$
$$=\frac{\prod_{r=1}^{\bar{\ell}_{i\bullet}-1}(\theta+r\sigma)}{(\theta+1)_{n_i-1}}\prod_{j=1}^{k}\tilde{p}_0^{\ell_{i,j}}(A_{j,\epsilon})S^{\sigma}_{n_{i,j},\ell_{i,j}}$$
$$\times\int_{\mathbb{R}_+}h_i(v)\mathbb{E}_v\left[e^{-\tilde{\mu}_i^*(f_i)}\right]\prod_{j=1}^{k}\mathbb{E}_v\left[e^{-H_{i,j}f_i(\mathrm{x}_j^*)}\right]\mathrm{d}v,$$

where h_i is the density of a random variable $U_i \sim \text{Ga}\left(\bar{\ell}_{i\bullet} + \frac{\theta}{\sigma}, 1\right)$, and $\tilde{\mu}_i^*$ is, conditionally on U_i and $\tilde{\mu}_0$, a generalized Gamma CRM with parameters (U_i, σ) and base measure $\tilde{p}_0 = \frac{\tilde{\mu}_0}{\tilde{\mu}_0(\mathbb{X})}$, for $i = 1, \ldots, m$. The random variables $H_{i,j}$ are independent Gamma random variables, each with parameters $n_{i,j} - \ell_{i,j}\sigma, U_i$. Concerning the denominator, one has that

$$\mathbb{E}\left[\prod_{j=1}^{k}\prod_{i=1}^{m}\tilde{p}_i^{n_{i,j}}(A_{j,\epsilon}) \,\Big|\, \mathbf{t}, \tilde{\mu}_0\right] = \frac{\prod_{r=1}^{\bar{\ell}_{i\bullet}}(\theta + r\sigma)}{(\theta+1)_{n_i-1}}\prod_{j=1}^{k}\tilde{p}_0^{\ell_{i,j}}(A_{j,\epsilon})S^{\sigma}_{n_{i,j},\ell_{i,j}},$$

and thus the proof is completed. ■

Appendix C

Chinese restaurant franchise sampler

To device a sampler for the mixture model (4.11), let

$$\mathcal{T}_i = \{t : \mathrm{t}_{i,j} \text{ for some } j = 1, \dots, n_i\}$$
$$\mathcal{D} = \{d : d = \mathrm{d}_{i,t} \text{ for some } i \in \{1, \dots, m\} \text{ and } t \in \{1, \dots, \bar{\ell}_{i\bullet}\}\},$$

that is, $\mathcal{D}$ is the set of indexes of the served dishes across the d groups, i.e. the *active dishes* displayed in the sample and $\mathcal{T}_i$ are the tables in restaurant i. Let the subscript $^{-ij}$ denote either the counts or the sets in which the observation $\mathrm{y}_{i,j}$ (customer j in restaurant i) is removed and, analogously, with the subscript $^{-ic}$ we indicate the counts and sets in which all the customers in table c of restaurant i are removed. That means that

$$\mathbf{y}^{-ij} = \mathbf{y} \setminus \{\mathrm{y}_{i,j}\} \quad \mathbf{t}^{-ij} = \mathbf{t} \setminus \{\mathrm{t}_{i,j}\} \quad \mathcal{T}_i^{-ij} = \mathcal{T}_i \setminus \{\mathrm{t}_{i,j}\}$$
$$\mathbf{y}^{-ic} = \mathbf{y} \setminus \{\mathrm{y}_{i,j} : \mathrm{t}_{i,j} = c\} \quad \mathbf{d}^{-ic} = \mathbf{d} \setminus \{\mathrm{d}_{i,c}\}$$
$$\bar{\ell}_{\bullet p}^{-ij} = \sum_{(i',p)\neq(i,j)} \ell_{i',p} \quad \bar{\ell}_{r\bullet}^{-ij} = \sum_{(r,j')\neq(i,j)} \ell_{r,j'} \quad |\boldsymbol{\ell}|^{-ij} = \sum_{(i',j')\neq(i,j)} \ell_{i',j}$$
$$q_{i\bullet c}^{-ij} = \sum_p q_{i,p,c}$$

To determine the full conditionals, we start with

$$\mathrm{p}(\mathbf{y}, \boldsymbol{\phi}, \mathbf{t}, \mathbf{d}) = \mathrm{p}(\mathbf{y} \,|\, \boldsymbol{\phi}, \mathbf{t}, \mathbf{d})\mathrm{p}(\boldsymbol{\phi} \,|\, \mathcal{D})\mathrm{p}(\mathbf{t}, \mathbf{d}),$$

where

$$\mathrm{p}(\mathbf{y} \,|\, \boldsymbol{\phi}, \mathbf{t}, \mathbf{d}) = \prod_{i=1}^{m}\prod_{j=1}^{n_i} \mathcal{K}(\mathrm{y}_{i,j} \,|\, \phi_{d_{i,\mathrm{t}_{i,j}}}), \quad \mathrm{p}(\boldsymbol{\phi} \,|\, \mathcal{D}) = \prod_{d\in\mathcal{D}} h(\phi_d).$$

Now note that the marginal distribution of $[\mathbf{y}, \mathbf{t}, \mathbf{d}]$ factorizes as follows

$$\mathrm{p}(\mathbf{y}, \mathbf{t}, \mathbf{d}) = \mathrm{p}(\mathbf{y} \,|\, \mathbf{t}, \mathbf{d})\mathrm{p}(\mathbf{t}, \mathbf{d}) = \mathrm{p}(\mathbf{t}, \mathbf{d}) \prod_{d\in\mathcal{D}} \int \prod_{(i,j):d_{i,\mathrm{t}_{i,j}}=d} \mathcal{K}(\mathrm{y}_{i,j} \,|\, \phi) h(\phi) \mathrm{d}\phi, \tag{C.1}$$

where the product runs over every observation $\mathrm{y}_{i,j}$ that shares the same mixture component ϕ_d (or equivalently every customer that sits on a table on which dish d is being served). But recalling that $\mathrm{d}_{i,\mathrm{t}_{i,j}} = \mathrm{d}^*_{i,j}$, we can write

$$\mathrm{p}(\mathbf{y}, \boldsymbol{\phi}, \mathbf{t}, \mathbf{d}^*) = \mathrm{p}(\mathbf{y} \,|\, \boldsymbol{\phi}, \mathbf{d}^*)\mathrm{p}(\boldsymbol{\phi} \,|\, \mathcal{D})\mathrm{p}(\mathbf{t}, \mathbf{d}^*), \tag{C.2}$$

where

$$\mathrm{p}(\mathbf{y} \,|\, \boldsymbol{\phi}, \mathbf{t}, \mathbf{d}^*) = \prod_{i=1}^{m}\prod_{j=1}^{n_i} \mathcal{K}(\mathrm{y}_{i,j} \,|\, \phi_{\mathrm{d}^*_{i,j}}) \tag{C.3}$$

and $\mathcal{D} = \{\mathrm{d}^*_{i,j} : i \in \mathcal{I}, j = 1, \ldots, n_i\}$. Now from (C.1) it is clear that, given $[\mathbf{d}^*, \boldsymbol{\phi}]$, $\mathbf{y}$ and $\mathbf{t}$ are conditionally independent and since $\mathbf{d}^* \overset{d}{=} \mathbf{d}$, we obtain

$$\begin{aligned}\mathrm{p}(\mathbf{y}, \mathbf{t}, \mathbf{d}^*) &= \mathrm{p}(\mathbf{t}, \mathbf{d}^*)\mathrm{p}(\mathbf{y} \,|\, \mathbf{d}^*) \\ &= \mathrm{p}(\mathbf{t}, \mathbf{d}^*) \prod_{d \in \mathcal{D}} \int \prod_{(i,j):\mathrm{d}^*_{i,j}=d} \mathcal{K}(\mathrm{y}_{i,j} \,|\, \phi) h(\phi) \mathrm{d}\phi.\end{aligned} \tag{C.4}$$

where again the indexes of the product indicate that we are iterating over observations that share the same mixture component. Finally, let $\mathcal{S}_d$ denote the set of indexes of the observations assigned to the d-th component of the mixture, that is

$$\mathcal{S}_d = \{(i,t) : \mathrm{d}^*_{i,t} = d\}.$$

For an arbitrary index $\mathcal{S}$, the marginal conditional density of $\{\mathrm{y}_{i,t} : i, t \in \mathcal{S}\}$ given all the observations assigned to component d is

$$\mathrm{p}(\{\mathrm{y}_{i,t}\}_{i,t\in\mathcal{S}} \,|\, \{\mathrm{y}_{i',t'} : (i',t') \in \mathcal{S}_d \setminus \mathcal{S}\}, \mathbf{t}, \mathbf{d}) = \frac{\int \prod_{i't' \in \mathcal{S}_d \cup \mathcal{S}} \mathcal{K}(\mathrm{y}_{i',t'} \,|\phi) h(\phi) \mathrm{d}\phi}{\int \prod_{i't' \in \mathcal{S}_d \setminus \mathcal{S}} \mathcal{K}(\mathrm{y}_{i',t'} \,|\, \phi) h(\phi) \mathrm{d}\phi}. \tag{C.5}$$

We will refer to this density as $f_d(\{\mathrm{y}_{i,t}\}_{i,t\in\mathcal{S}})$.

Full conditionals for $(\mathrm{t}_{i,j}, \mathrm{d}^*_{i,j})$.

As described in Algorithm 3.1, the outcomes of the sampling are of three types. Observation $\mathrm{y}_{i,j}$ can be either located at an old cluster in group i and therefore it gets assigned the mixture component corresponding to that cluster or it can be allocated at a new cluster. If the latter occurs, then two disjoint events are possible: either this new cluster gets assigned an old mixture component or a completely new one. This corresponds to the customer either seating at an old table and sharing the same dish of that table, sitting at a new table and either ordering an old dish or a new dish. In terms of the index random variables, first scenario reads as $\mathrm{t}_{i,j} = t^{\mathrm{old}}$, where t^{old} is a table already present in $\mathcal{T}_i^{-ij}$. In this case, $\mathrm{d}^*_{i,j} = d^*_{i,t^{\mathrm{old}}} \in \mathcal{D}^{-ij}$. In the second and third scenario, $\mathrm{t}_{i,j} = t^{\mathrm{new}}$ and either $\mathrm{d}^*_{i,j} = d$ for some $d \in \mathcal{D}^{-ij}$ or $\mathrm{d}^*_{i,j} = d^{\mathrm{new}}$.

From (C.2), (C.3) and (C.4) one has that

$$\mathrm{p}(\mathbf{y}^{-ij}, \mathbf{t}, \mathbf{d}^*) = \mathrm{p}(\mathbf{t}, \mathbf{d}^*)\mathrm{p}(\mathbf{y}^{-ij} \,|\, \mathbf{t}, \mathbf{d}^*) = \mathrm{p}(\mathbf{t}, \mathbf{d}^*)\mathrm{p}(\mathbf{y}^{-ij} \,|\, \mathbf{d}^{*-ij}),$$

which implies that $\mathrm{p}(\mathbf{y}^{-ij} \,|\, \mathbf{t}, \mathbf{d}^*) = \mathrm{p}(\mathbf{y}^{-ij} \,|\, \mathbf{d}^{*-ij})$. Using this fact, we obtain

$$\begin{aligned}\mathrm{p}(\mathbf{y}, \mathrm{t}_{i,j}, \mathrm{d}^*_{i,j}, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij}) &= \mathrm{p}(\mathbf{t}^{-ij}, \mathbf{d}^{*-ij})\mathrm{p}(\mathrm{t}_{i,j}, \mathrm{d}^*_{i,j} \,|\, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij})\mathrm{p}(\mathbf{y}^{-ij} \,|\, \mathbf{t}, \mathbf{d}^*)\mathrm{p}(\mathrm{y}_{i,j} \,|\, \mathbf{y}^{-ij}, \mathbf{t}, \mathbf{d}^*) \\ &= \mathrm{p}(\mathbf{t}^{-ij}, \mathbf{d}^{*-ij})\mathrm{p}(\mathrm{t}_{i,j}, \mathrm{d}^*_{i,j} \,|\, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij})\mathrm{p}(\mathbf{y}^{-ij} \,|\, \mathbf{d}^{*-ij})\mathrm{p}(\mathrm{y}_{i,j} \,|\, \mathbf{y}^{-ij}, \mathbf{t}, \mathbf{d}^*).\end{aligned}$$

Hence

$$\begin{aligned}\mathrm{p}(\mathrm{t}_{i,j}, \mathrm{d}^*_{i,j} \,|\, \mathbf{y}, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij}) &\propto \mathrm{p}(\mathbf{y}, \mathrm{t}_{i,j}, \mathrm{d}^*_{i,j}, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij}) \\ &= \mathrm{p}(\mathrm{t}_{i,j}, \mathrm{d}^*_{i,j} \,|\, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij})\mathrm{p}(\mathrm{y}_{i,j} \,|\, \mathbf{y}^{-ij}, \mathbf{d}^{*-ij}, \mathrm{d}^*_{i,j})\end{aligned}$$

The three scenarios described in the beginning of this section are encoded within

$$\mathrm{p}(\mathrm{t}_{i,j}, \mathrm{d}^*_{i,j} \,|\, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij}).$$

Note that the labels inherit the partial exchangeability assumption, hence labels within the same group i can be treated as exchangeable and accordingly, can be re ordered so that $\mathrm{t}_{i,j}$ and $\mathrm{d}^*_{i,j}$ are sampled at last. Let us denote as $\omega_0^{(n)}$ and $\omega_t^{(n)}$ the weights of the predictive distributions of the random partition with EPPF $\widetilde{\Phi}_i$, for $i = 1, \ldots, m$ and let $\tilde{\omega}_0^{(n)}$ and $\tilde{\omega}_j^{(n)}$ denote the weights of the predictive distribution of the random partitions with EPPF Φ_0 defined analogously by using Φ_0 in place of $\widetilde{\Phi}_i$. Using the exchangeability of the labels, we obtain

$$\begin{aligned}\mathrm{p}(\mathrm{t}_{i,j} = t^{\mathrm{old}}, \mathrm{d}^*_{i,j} = d^*_{i,t^{\mathrm{old}}} \,|\, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij}) &= \omega_{t^{\mathrm{old}}}^{(n_i-1)}\left(\mathbf{t_i}^{-ij}\right) \\ \mathrm{p}(\mathrm{t}_{i,j} = t^{\mathrm{new}}, \mathrm{d}^*_{i,j} = d^{\mathrm{old}} \,|\, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij}) &= \omega_0^{(n_i-1)}\left(\mathbf{t_i}^{-ij}\right)\widetilde{\omega}_{d^{\mathrm{old}}}^{|\boldsymbol{\ell}|^{-ij}}\left(\mathbf{d}^{-ij}\right) \\ \mathrm{p}(\mathrm{t}_{i,j} = t^{\mathrm{new}}, \mathrm{d}^*_{i,j} = d^{\mathrm{new}} \,|\, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij}) &= \omega_0^{(n_i-1)}\left(\mathbf{t_i}^{-ij}\right)\widetilde{\omega}_0^{|\boldsymbol{\ell}|^{-ij}}\left(\mathbf{d}^{-ij}\right),\end{aligned} \tag{C.6}$$

where

$$\begin{aligned}\omega_{t^{\mathrm{old}}}^{(n_i-1)}\left(\mathbf{t_i}^{-ij}\right) &= \omega_{t^{\mathrm{old}}}^{(n_i-1)}\left(q_{i\bullet 1}^{-ij}, \ldots, q_{i\bullet \bar{\ell}_{i\bullet}^{-ij}}^{-ij}\right) \\ \omega_0^{(n_i-1)}\left(\mathbf{t_i}^{-ij}\right) &= \omega_0^{(n_i-1)}\left(q_{i\bullet 1}^{-ij}, \ldots, q_{i\bullet \bar{\ell}_{i\bullet}^{-ij}}^{-ij}\right).\end{aligned}$$

These are the weights of the predictive distribution associated with the EPPF $\widetilde{\Phi}_i$, evaluated at the block sizes without taking into account observation $\mathrm{y}_{i,j}$, i.e. each $q_{i\bullet c}^{-ij}$ represents the number of observations sitting at table c without counting y_{ij}. Particularly, the last entry $q_{i\bullet \bar{\ell}_{i\bullet}^{-ij}}^{-ij}$ represents the number of people sitting at the last table, without taking into

account the table in which $\mathrm{y}_{i,j}$ is sitting. Analogously,

$$\widetilde{\omega}_{d^{\text{old}}}^{|\boldsymbol{\ell}|^{-ij}}\left(\mathbf{d}^{-ij}\right) = \widetilde{\omega}_{d^{\text{old}}}^{|\boldsymbol{\ell}|^{-ij}}(\bar{\ell}_{\bullet 1}^{-ij}, \ldots, \bar{\ell}_{\bullet|\mathcal{D}^{-ij}|}^{-ij})$$
$$\widetilde{\omega}_{0}^{|\boldsymbol{\ell}|^{-ij}}\left(\mathbf{d}^{-ij}\right) = \widetilde{\omega}_{0}^{|\boldsymbol{\ell}|^{-ij}}(\bar{\ell}_{\bullet 1}^{-ij}, \ldots, \bar{\ell}_{\bullet|\mathcal{D}^{-ij}|}^{-ij}),$$

where the weights of the predictive distribution associated with the EPPF Φ_0 are evaluated at the block sizes $\bar{\ell}_{\bullet m}^{-ij}$, which represent the number of tables that serve dish m without taking into account the table on which $\mathrm{y}_{i,j}$ is sitting. On the other hand, by the virtue of equation (C.5) with $\mathcal{S} = \{(i,j)\}$

$$\mathrm{p}(\mathrm{y}_{i,j} \,|\, \mathbf{y}^{-ij}, \mathbf{d}^{*-ij}, \mathrm{d}_{i,j}^{*}) = f_{\mathrm{d}_{i,j}^{*}}(\mathrm{y}_{i,j}). \tag{C.7}$$

Putting together (C.6) and (C.7) in (C.6) yields

$$\begin{aligned}
\mathrm{p}(\mathrm{t}_{i,j} = t^{\text{old}}, \mathrm{d}_{i,j}^{*} = d_{i,t^{\text{old}}}^{*} \,|\, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij}) &\propto \omega_{t^{\text{old}}}^{(n_i-1)}\left(\mathbf{t_i}^{-ij}\right) f_{d_{i,t^{\text{old}}}^{*}}(\mathrm{y}_{i,j}) \\
\mathrm{p}(\mathrm{t}_{i,j} = t^{\text{new}}, \mathrm{d}_{i,j}^{*} = d^{\text{old}} \,|\, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij}) &\propto \omega_{0}^{(n_i-1)}\left(\mathbf{t_i}^{-ij}\right) \widetilde{\omega}_{d^{\text{old}}}^{|\boldsymbol{\ell}|^{-ij}}\left(\mathbf{d}^{-ij}\right) f_{d^{\text{old}}}(\mathrm{y}_{i,j}) \\
\mathrm{p}(\mathrm{t}_{i,j} = t^{\text{new}}, \mathrm{d}_{i,j}^{*} = d^{\text{new}} \,|\, \mathbf{t}^{-ij}, \mathbf{d}^{*-ij}) &\propto \omega_{0}^{(n_i-1)}\left(\mathbf{t_i}^{-ij}\right) \widetilde{\omega}_{0}^{|\boldsymbol{\ell}|^{-ij}}\left(\mathbf{d}^{-ij}\right) f_{d^{\text{new}}}(\mathrm{y}_{i,j}).
\end{aligned} \tag{C.8}$$

Full conditionals for $\mathrm{d}_{i,t}$.

To sample the dish labels of the $\bar{\ell}_{i\bullet}$ tables in restaurant i, there are two scenarios: either we re-sample an old dish label or sample a new one from P_0. Analogously to the definition of $\mathcal{S}_d$, the set of indexes of observations seating at table c in restaurant equals

$$\mathcal{S}_{ic} = \{(i,j) : \mathrm{t}_{i,j} = c\} \text{ for } i = 1, \ldots, m$$

that is, the only index that changes is j and i remains fixed. From (C.1) we get

$$\mathrm{p}(\mathrm{d}_{i,t} = d \,|\, \mathbf{y}, \mathbf{t}, \mathbf{d}^{-ic}) \propto \mathrm{p}(\mathrm{d}_{i,t} = d \,|\, \mathbf{t}, \mathbf{d}^{-ic}) \mathrm{p}(\{\mathrm{y}_{i,j} : (i,j) \in \mathcal{S}_{ic}\} \,|\, \{\mathrm{y}_{i',j'} : (i',j') \in \mathcal{S}_d \setminus \mathcal{S}_{i,c}\}, \mathbf{t}, \mathbf{d}^{-ic}, d)$$

The weights of the predictive distribution associated with the EPPF Φ_0, leads to

$$\begin{aligned}
\mathrm{p}(\mathrm{d}_{i,t} = d^{\text{new}} \,|\, \mathbf{t}, \mathbf{d}^{-ic}) &= \widetilde{\omega}_{0}^{|\boldsymbol{\ell}|^{-ic}}\left(\mathbf{d}^{-ic}\right) \\
\mathrm{p}(\mathrm{d}_{i,t} = d^{\text{old}} \,|\, \mathbf{t}, \mathbf{d}^{-ic}) &= \widetilde{\omega}_{d^{\text{old}}}^{|\boldsymbol{\ell}|^{-ic}}\left(\mathbf{d}^{-ic}\right) \text{ for } d^{\text{old}} \in \mathcal{D}^{-ic},
\end{aligned} \tag{C.9}$$

where

$$\widetilde{\omega}_{d^{\text{old}}}^{|\boldsymbol{\ell}|^{-ic}}\left(\mathbf{d}^{-ic}\right) = \widetilde{\omega}_{d^{\text{old}}}^{|\boldsymbol{\ell}|^{-ic}}\left(\bar{\ell}_{\bullet 1}^{-ic}, \ldots, \bar{\ell}_{\bullet|\mathcal{D}^{-ic}|}^{-ic}\right)$$
$$\widetilde{\omega}_{0}^{|\boldsymbol{\ell}|^{-ic}}\left(\mathbf{d}^{-ic}\right) = \widetilde{\omega}_{0}^{|\boldsymbol{\ell}|^{-ic}}\left(\bar{\ell}_{\bullet 1}^{-ic}, \ldots, \bar{\ell}_{\bullet|\mathcal{D}^{-ic}|}^{-ic}\right),$$

that is, the weights are evaluated in on the block sizes $\ell_{\bullet d}^{-ic}$ that represent of the number of tables that serve dish d without taking into account table c. Now note that

$$\mathrm{p}(\{\mathrm{y}_{i,j} : (i,j) \in \mathcal{S}_{ic}\} \,|\, \{\mathrm{y}_{i',j'} : (i',j') \in \mathcal{S}_d \setminus \mathcal{S}_{i,c}\}, \mathbf{t}, \mathbf{d}^{-ic}, d) = f_d(\{\mathrm{y}_{i,j} : (i,j) \in S_{ic}\}). \tag{C.10}$$

By putting together (C.9) and (C.10) we obtain

$$\begin{aligned}
\mathrm{p}(\mathrm{d}_{i,t} = d^{\mathrm{new}} \,|\, \mathbf{y}, \mathbf{t}, \mathbf{d}^{-ic}) &\propto \widetilde{\omega}_0^{|\boldsymbol{\ell}|^{-ic}} \left(\mathbf{d}^{-ic}\right) f_{d^{\mathrm{new}}}(\{\mathrm{y}_{i,j} : (i,j) \in S_{ic}\}) \\
\mathrm{p}(\mathrm{d}_{i,t} = d^{\mathrm{old}} \,|\, \mathbf{y}, \mathbf{t}, \mathbf{d}^{-ic}) &\propto \widetilde{\omega}_{d^{\mathrm{old}}}^{|\boldsymbol{\ell}|^{-ic}} \left(\mathbf{d}^{-ic}\right) f_{d^{\mathrm{old}}}(\{\mathrm{y}_{i,j} : (i,j) \in S_{ic}\}).
\end{aligned}$$

Sensitivity analysis for the first experiment

The parameters of the prior processes were chosen is such a way that, marginally, $\mathbb{E}[\mathrm{K}_{i,n_i}] = 5$ for $i = 1, 2$ and $\mathbb{E}[\mathrm{K}_N] = 6$, where $n_1 = 100 = n_2$ and $N = 200$. The distribution of $\mathrm{K}_{i,100}$ and K_{200} were computed using (3.7) and (3.5) respectively.

Now we will perform a sensitivity analysis of the effect of the hyperparameters of P_0 by imposing values of the a priori moments of m and σ^2 in a data-driven way. Consider the following systems of equations to solve for (τ_0, a, b)

$$\begin{aligned}
\mathbb{E}\left[m\right] =& m_0 = \bar{\mathbf{y}} \\
\mathbb{E}[\sigma^2] =& \frac{b}{a-1} = \frac{\mathrm{var}(\mathbf{y})}{2} \\
\mathrm{v}[m] =& \frac{b}{(a-1)\tau_0} = \frac{\mathbb{E}[\sigma^2]}{\tau_0} \\
\mathrm{v}[\sigma^2] =& \frac{b^2}{(a-1)^2(a-2)} = \left(\mathbb{E}[\sigma^2]\right)^2 \frac{b}{(a-2)}
\end{aligned}$$

where the variances of m and σ^2 take values on $\{0.1, 1, 5, 7\}$, $\bar{\mathbf{y}}$ and $\mathrm{var}(\mathbf{y})$ are the overall mean and variance of the data. The hyperparameters of P_0 will be chosen as the ones that minimize the LPML, while also taking into account the posterior moments of K_N. The results are displayed in the following pages, indicating the best LPML score in bold.

	$\mathbb{E}[\mathrm{K}_N \mid \ldots]$					$\mathrm{var}[\mathrm{K}_N \mid \ldots]$			
	$\mathrm{var}(\sigma^2) = 0.1$	$\mathrm{var}(\sigma^2) = 1$	$\mathrm{var}(\sigma^2) = 5$	$\mathrm{var}(\sigma^2) = 7$		$\mathrm{var}(\sigma^2) = 0.1$	$\mathrm{var}(\sigma^2) = 1$	$\mathrm{var}(\sigma^2) = 5$	$\mathrm{var}(\sigma^2) = 7$
$\mathrm{var}(m_0) = 0.1$	5.472	5.127	5.19	5.22	$\mathrm{var}(m_0) = 0.1$	2.18	2.28	2.25	2.15
$\mathrm{var}(m_0) = 1$	6.272	5.18	4.765	4.739	$\mathrm{var}(m_0) = 1$	2.53	1.95	1.5	1.55
$\mathrm{var}(m_0) = 5$	6.303	4.9	5.097	5.125	$\mathrm{var}(m_0) = 5$	2.31	1.8	1.72	1.67
$\mathrm{var}(m_0) = 7$	6.262	4.761	5.032	5.132	$\mathrm{var}(m_0) = 7$	2.25	1.65	1.58	1.66

Table C.1: Posterior moments of K_N *for the hierarchical Dirichlet Process.*

LPML				
	$\text{var}(\sigma^2) = 0.1$	$\text{var}(\sigma^2) = 1$	$\text{var}(\sigma^2) = 5$	$\text{var}(\sigma^2) = 7$
$\text{var}(m_0) = 0.1$	-1.213	-1.206	-1.205	-1.205
$\text{var}(m_0) = 1$	-1.177	-1.152	-1.115	-1.1
$\text{var}(m_0) = 5$	-1.179	-1.121	-1.063	-1.063
$\text{var}(m_0) = 7$	-1.171	-1.119	**-1.061**	-1.063

Table C.2: LPML (10^3) for the hierarchical Dirichlet Process.

$\mathbb{E}[\text{K}_N \mid \ldots]$					$\text{var}[\text{K}_N \mid \ldots]$				
	$\text{var}(\sigma^2) = 0.1$	$\text{var}(\sigma^2) = 1$	$\text{var}(\sigma^2) = 5$	$\text{var}(\sigma^2) = 7$		$\text{var}(\sigma^2) = 0.1$	$\text{var}(\sigma^2) = 1$	$\text{var}(\sigma^2) = 5$	$\text{var}(\sigma^2) = 7$
$\text{var}(m_0) = 0.1$	4.927	4.896	4.912	4.595	$\text{var}(m_0) = 0.1$	6.516	7.843	6.458	5.711
$\text{var}(m_0) = 1$	7.523	5.871	5.232	5.195	$\text{var}(m_0) = 1$	9.981	5.417	3.875	3.536
$\text{var}(m_0) = 5$	7.589	5.37	5.716	5.848	$\text{var}(m_0) = 5$	8.268	4.46	4.033	4.338
$\text{var}(m_0) = 7$	7.418	5.244	5.548	5.988	$\text{var}(m_0) = 7$	8.261	3.945	3.761	4.492

Table C.3: Posterior moments of K_N for the hierarchical Stable process.

LPML				
	$\text{var}(\sigma^2) = 0.1$	$\text{var}(\sigma^2) = 1$	$\text{var}(\sigma^2) = 5$	$\text{var}(\sigma^2) = 7$
$\text{var}(m_0) = 0.1$	-1.219	-1.205	-1.205	-1.205
$\text{var}(m_0) = 1$	-1.177	-1.153	-1.113	-1.112
$\text{var}(m_0) = 5$	-1.169	-1.12	-1.074	**-1.061**
$\text{var}(m_0) = 7$	-1.168	-1.123	-1.071	-1.065

Table C.4: LPML (10^3) for the hierarchical Stable process.

$\mathbb{E}[\text{K}_N \mid \ldots]$					$\text{var}[\text{K}_N \mid \ldots]$				
	$\text{var}(\sigma^2) = 0.1$	$\text{var}(\sigma^2) = 1$	$\text{var}(\sigma^2) = 5$	$\text{var}(\sigma^2) = 7$		$\text{var}(\sigma^2) = 0.1$	$\text{var}(\sigma^2) = 1$	$\text{var}(\sigma^2) = 5$	$\text{var}(\sigma^2) = 7$
$\text{var}(m_0) = 0.1$	3.507	3.496	3.478	3.563	$\text{var}(m_0) = 0.1$	2.267	2.356	2.328	2.661
$\text{var}(m_0) = 1$	3.244	2.213	1.394	1.539	$\text{var}(m_0) = 1$	3.244	2.213	1.394	1.539
$\text{var}(m_0) = 5$	5.466	4.341	4.484	4.598	$\text{var}(m_0) = 5$	3.67	1.857	1.862	2.07
$\text{var}(m_0) = 7$	5.49	4.231	4.508	4.766	$\text{var}(m_0) = 7$	3.579	1.591	1.841	2.241

Table C.5: Posterior moments of K_N for the hierarchical Pitman-Yor process.

LPML				
	$\text{var}(\sigma^2) = 0.1$	$\text{var}(\sigma^2) = 1$	$\text{var}(\sigma^2) = 5$	$\text{var}(\sigma^2) = 7$
$\text{var}(m_0) = 0.1$	-1.22	-1.205	-1.205	-1.205
$\text{var}(m_0) = 1$	-1.177	-1.148	-1.109	-1.101
$\text{var}(m_0) = 5$	-1.168	-1.12	-1.074	**-1.058**
$\text{var}(m_0) = 7$	-1.165	-1.125	-1.065	-1.06

Table C.6: LPML (10^3) for the hierarchical Pitman-Yor process.

	$\mathbb{E}[\mathrm{K}_N \mid \ldots]$					$\mathrm{var}[\mathrm{K}_N \mid \ldots]$			
	$\mathrm{var}(\sigma^2) = 0.1$	$\mathrm{var}(\sigma^2) = 1$	$\mathrm{var}(\sigma^2) = 5$	$\mathrm{var}(\sigma^2) = 7$		$\mathrm{var}(\sigma^2) = 0.1$	$\mathrm{var}(\sigma^2) = 1$	$\mathrm{var}(\sigma^2) = 5$	$\mathrm{var}(\sigma^2) = 7$
$\mathrm{var}(m_0) = 0.1$	4.687	4.869	4.388	4.394	$\mathrm{var}(m_0) = 0.1$	3.967	4.47	3.698	3.461
$\mathrm{var}(m_0) = 1$	6.44	5.967	5.472	5.601	$\mathrm{var}(m_0) = 1$	4.076	3.442	2.736	2.851
$\mathrm{var}(m_0) = 5$	6.808	5.997	6.085	6.199	$\mathrm{var}(m_0) = 5$	3.993	3.386	3.04	3.177
$\mathrm{var}(m_0) = 7$	6.681	5.884	5.906	6.218	$\mathrm{var}(m_0) = 7$	4.163	3.288	2.79	3.188

Table C.7: Posterior moments of K_N *for the hierarchical Dirichlet-Pitman-Yor process.*

LPML				
	$\mathrm{var}(\sigma^2) = 0.1$	$\mathrm{var}(\sigma^2) = 1$	$\mathrm{var}(\sigma^2) = 5$	$\mathrm{var}(\sigma^2) = 7$
$\mathrm{var}(m_0) = 0.1$	-1.216	-1.205	-1.204	-1.205
$\mathrm{var}(m_0) = 1$	-1.175	-1.154	-1.112	-1.114
$\mathrm{var}(m_0) = 5$	-1.169	-1.127	-1.068	-1.066
$\mathrm{var}(m_0) = 7$	-1.177	-1.119	-1.071	**-1.065**

Table C.8: LPML (10^3) *for the hierarchical Dirichlet-Pitman-Yor process.*

	$\mathbb{E}[\mathrm{K}_N \mid \ldots]$					$\mathrm{var}[\mathrm{K}_N \mid \ldots]$			
	$\mathrm{var}(\sigma^2) = 0.1$	$\mathrm{var}(\sigma^2) = 1$	$\mathrm{var}(\sigma^2) = 5$	$\mathrm{var}(\sigma^2) = 7$		$\mathrm{var}(\sigma^2) = 0.1$	$\mathrm{var}(\sigma^2) = 1$	$\mathrm{var}(\sigma^2) = 5$	$\mathrm{var}(\sigma^2) = 7$
$\mathrm{var}(m_0) = 0.1$	4.687	4.869	4.388	4.394	$\mathrm{var}(m_0) = 0.1$	6.436	7.951	5.598	5.859
$\mathrm{var}(m_0) = 1$	6.44	5.967	5.472	5.601	$\mathrm{var}(m_0) = 1$	8.761	5.426	4.014	3.873
$\mathrm{var}(m_0) = 5$	6.808	5.997	6.085	6.199	$\mathrm{var}(m_0) = 5$	9.644	5.734	4.684	4.894
$\mathrm{var}(m_0) = 7$	6.681	5.884	5.906	6.218	$\mathrm{var}(m_0) = 7$	8.834	5.183	4.369	4.806

Table C.9: Posterior moments of K_N *for the hierarchical Stable-Pitman-Yor process.*

LPML				
	$\mathrm{var}(\sigma^2) = 0.1$	$\mathrm{var}(\sigma^2) = 1$	$\mathrm{var}(\sigma^2) = 5$	$\mathrm{var}(\sigma^2) = 7$
$\mathrm{var}(m_0) = 0.1$	-1.22	-1.205	-1.205	-1.205
$\mathrm{var}(m_0) = 1$	-1.175	-1.151	-1.113	-1.109
$\mathrm{var}(m_0) = 5$	-1.178	-1.133	-1.063	-1.066
$\mathrm{var}(m_0) = 7$	-1.17	-1.12	**-1.06**	-1.061

Table C.10: LPML (10^3) *for the hierarchical Stable-Pitman-Yor process.*

	$\mathbb{E}[\mathrm{K}_N \mid \ldots]$					$\mathrm{var}[\mathrm{K}_N \mid \ldots]$			
	$v(\sigma^2) = 0.1$	$\mathrm{var}(\sigma^2) = 1$	$\mathrm{var}(\sigma^2) = 5$	$\mathrm{var}(\sigma^2) = 7$		$\mathrm{var}(\sigma^2) = 0.1$	$\mathrm{var}(\sigma^2) = 1$	$\mathrm{var}(\sigma^2) = 5$	$\mathrm{var}(\sigma^2) = 7$
$\mathrm{var}(m_0) = 0.1$	5.374	5.195	5.241	5.235	$\mathrm{var}(m_0) = 0.1$	2.437	2.516	2.513	2.548
$\mathrm{var}(m_0) = 1$	6.405	5.217	4.628	4.639	$\mathrm{var}(m_0) = 1$	3.026	2.225	1.54	1.695
$\mathrm{var}(m_0) = 5$	6.48	4.715	4.973	5.065	$\mathrm{var}(m_0) = 5$	2.689	1.797	1.711	1.76
$\mathrm{var}(m_0) = 7$	6.304	4.696	5.004	5.139	$\mathrm{var}(m_0) = 7$	2.559	1.76	1.833	1.775

Table C.11: Posterior moments of K_N *for the hierarchical Pitman-Yor-Dirichlet process.*

LPML				
	$\mathrm{var}(\sigma^2) = 0.1$	$\mathrm{var}(\sigma^2) = 1$	$\mathrm{var}(\sigma^2) = 5$	$\mathrm{var}(\sigma^2) = 7$
$\mathrm{var}(m_0) = 0.1$	-1.216	-1.206	-1.204	-1.205
$\mathrm{var}(m_0) = 1$	-1.178	-1.153	-1.119	-1.101
$\mathrm{var}(m_0) = 5$	-1.174	-1.124	-1.075	-1.076
$\mathrm{var}(m_0) = 7$	-1.167	-1.118	-1.064	**-1.06**

Table C.12: LPML (10^3) *for the hierarchical Pitman-Yor-Dirichlet process.*

	$\mathbb{E}[\mathrm{K}_N \mid \ldots]$					$\mathrm{var}[\mathrm{K}_N \mid \ldots]$			
	$v(\sigma^2) = 0.1$	$\mathrm{var}(\sigma^2) = 1$	$\mathrm{var}(\sigma^2) = 5$	$\mathrm{var}(\sigma^2) = 7$		$\mathrm{var}(\sigma^2) = 0.1$	$\mathrm{var}(\sigma^2) = 1$	$\mathrm{var}(\sigma^2) = 5$	$\mathrm{var}(\sigma^2) = 7$
$\mathrm{var}(m_0) = 0.1$	5.267	4.768	4.392	4.2	$\mathrm{var}(m_0) = 0.1$	5.381	5.174	4.204	3.899
$\mathrm{var}(m_0) = 1$	6.844	5.91	5.258	5.421	$\mathrm{var}(m_0) = 1$	6.594	4.395	3.078	3.242
$\mathrm{var}(m_0) = 5$	7.25	5.696	5.869	5.933	$\mathrm{var}(m_0) = 5$	6.503	4.177	3.874	3.959
$\mathrm{var}(m_0) = 7$	7.332	5.485	5.67	5.992	$\mathrm{var}(m_0) = 7$	6.474	3.675	3.288	3.635

Table C.13: Posterior moments of K_N *for the hierarchical Pitman-Yor-Stable process.*

LPML				
	$\mathrm{var}(\sigma^2) = 0.1$	$\mathrm{var}(\sigma^2) = 1$	$\mathrm{var}(\sigma^2) = 5$	$\mathrm{var}(\sigma^2) = 7$
$\mathrm{var}(m_0) = 0.1$	-1.213	-1.205	-1.205	-1.205
$\mathrm{var}(m_0) = 1$	-1.178	-1.155	-1.115	-1.111
$\mathrm{var}(m_0) = 5$	-1.167	-1.125	-1.072	-1.085
$\mathrm{var}(m_0) = 7$	-1.172	-1.127	-1.064	**-1.058**

Table C.14: LPML (10^3) *for the hierarchical Pitman-Yor-Stable process.*

It is clear from the LPML estimates that higher a priori variances for m_0 and σ^2 yield a better fit. In particular, it can be seen how an increased variability in the a priori number of clusters (i.e. higher values of the hyperparameter) is associated with a better fit in terms of LPML values. In terms of the posterior moments of K_{200}, smaller variances a priori produce greater dispersion and also a greater posterior mean in most cases. Table C.15 shows the final parameters and hyperparameters values used in the simulation for the first experiment.

	m_0	τ_0	a	b	θ	θ_0	σ	σ_0
HDP	-1.3167	0.4776	4.2355	10.8177	0.9475	5.5837		
HSP	-1.3167	0.6686	3.5968	8.6821			0.3253	0.7406
HPYP	-1.3167							
HDPY	-1.3167	0.4776	3.5968	8.6821	0.0024	6.4431	0.3243	
HSPYP	-1.3167	0.4776	4.2355	10.8177	0.0024		0.3243	0.7488
HPYDP	-1.3167	0.4776	3.5968	8.6821	0.9475	3.3258	0.28869	
HPYSP	-1.3167	0.4776	3.5968	8.6821		1.7147	0.3253	0.5281

Table C.15: Final parameters used in the simulations.

Bibliography

Aldous, D. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII*, volume 1117. Springer.

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598.

Argiento, R., Cremaschi, A., and Vannucci, M. (2020). Hierarchical Normalized Completely Random Measures to Cluster Grouped Data. *Journal of the American Statistical Association*, 115(529):318–333.

Bassetti, F., Casarin, R., and Rossini, L. (2020). Hierarchical species sampling models. *Bayesian Analysis*, 15(3):809–838.

Bertoin, J. (1996). *Lévy Processes*. Cambridge University Press.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, 1(2):353–355.

Brix, A. (1999). Generalized gamma measures and shot-noise cox processes. *Advances in Applied Probability*, 31:929–953.

Camerlenghi, F. (2015). *Hierarchical and nested random probability measures with statistical applications*. PhD thesis, University of Pavia.

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., and Rodríguez, A. (2019a). Latent Nested Nonparametric Priors (with Discussion). *Bayesian Analysis*, 14(4):1303 – 1356.

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019b). Distribution theory for hierarchical processes. *The Annals of Statistics*, 47(1):67–92.

Charalambides, C. (2002). *Enumerative Combinatorics*. Chapman and Hall.

Charalambides, C. and Singh, J. (1988). A review of the Stirling numbers, their generalizations and statistical applications. *Communication in Statistics- Theory and Methods*, 17:2533–2595.

Cifarelli, D. and Regazzini, E. (1978). Nonparametric statistical problems under partial exchangeability: The role of associative means. Technical report, University of Turin.

de Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Accademia Nazionale dei Lincei, Ser. 6. Memorie, Classe di Scienze Fisiche, Matematiche e Naturali*, 4:251–299. Translated by Alvarez-Melis, D. and Broderick, T. (2015).

de Finetti, B. (1937). Sur la condition d'equivalence partielle. *Actualités Scientifiques et Industrielles*, 739. Translated by Benacerraf, P. and Jeffrey, R. (1980).

Diaconis, P. and Freedman, D. (1978). de Finetti's generalizations of exchangeability. Technical report, Standford University.

Escobar, M. D. (1994). Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277.

Favaro, S. and Teh, Y. W. (2013). MCMC for Normalized Random Measure Mixture Models. *Statistical Science*, 28(3):335 – 359.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 2(1):209–230.

Ferguson, T. S. and Klass, M. J. (1972). A Representation of Independent Increment Processes without Gaussian Components. *The Annals of Mathematical Statistics*, 43(5):1634 – 1643.

Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P. (2018). *Handbook of Mixture Analysis*. CRC Press.

Gelfand, A. and Dey, D. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56:501–514.

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457 – 472.

Gil-Leyva, M. F. (2016). Random partitions models. Master's thesis, Universidad Nacional Autónoma de México.

Gil-Leyva, M. F. (2021). *Stick-breaking processes and related random probability measures*. PhD thesis, Universidad Nacional Autónoma de México.

Goldstein, M. (2013). Observables and models: exchangeability and the inductive argument. In *Bayesian theory and applications*, pages 3–18. Oxford University Press.

Hewitt, E. and Savage, L. J. (1955). Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, 80:470–501.

James, L., Lijoi, A., and Prünster, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scandinavian Journal of Statistics*, 33(1):105–120.

James, L., Lijoi, A., and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1):76–97.

Kallenberg, O. (2017). *Random Measures, Theory and Applications*. Springer.

Kingman, J. (1967). Completely Random Measures. *Pacific Journal of Mathematics*, 21(1):59–78.

Kingman., J. (1975). Random discrete distributions. *Journal of the Royal Statistical Society*, 37(4):1–22.

Kingman, J. (1978). Uses of exchangeability. *The Annals of Probability*, 6(2):183–197.

Kingman, J. (1993). *Poisson processes*. Oxford University Press.

Korwar, R. M. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *The Annals of Probability*, 1(4):705–711.

Lo, A. Y. (1984). On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics*, 12(1):351 – 357.

Maceachern, S. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-simulation and Computation*, 23:727–741.

MacEachern, S. (1999). Dependent nonparametric processes. *Proceedings of the Bayesian Statistical Science Section*, pages 50–55.

Müller, P., Quintana, F., and Rosner, G. (2004). A method for combining inference across related nonparametric bayesian models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66(3):735–749.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and related Fields*, 31(102):145–158.

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

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

Pitman, J. (2006). *Combinatorial stochastic processes*. Springer. Lecture notes from Ecole d'Eté de Probabilités de Saint-Flour XXXII - 2002.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900.

Regazzini, E., Lijoi, A., and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2):560–585.

Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press.

Schervish, M. (1996). *The Theory of Statistics*. Springer.

Teh, Y. (2006). A hierarchical bayesian language model based on pitman-yor processes.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, pages 1566–1581.