UNIVERSITA' COMMERCIALE "LUIGI BOCCONI"
PhD SCHOOL

PhD program in Statistics
Cycle: XXXIV
Disciplinary Field: SECS-S/01

Modelling with discrete random probability measures

Advisor: Antonio LIJOI
Co-Advisor: Igor PRÜNSTER

PhD Thesis by
Francesco GAFFI
ID number: 3084759

Year 2023

Alla maestra Franca

Contents

Abstract

Random probability measures are a cornerstone of Bayesian nonparametrics. By virtue of de Finetti's representation theorem, their law acts as the prior distribution for exchangeable observations. Mostly used Bayesian nonparametric procedures, in this framework, rely on laws selecting almost surely discrete probability measures, such as the celebrated Dirichlet process and its several extensions.

The first part of this thesis is dedicated to problems related to random probability measures arising in the exchangeable regime. We explore properties of functionals of noteworthy discrete random probability measures in order to provide prior elicitation. In particular we retrieve explicit expressions for base measures inducing a broad class of distributions on the random mean of a Dirichlet process, a normalized stable process and a Pitman–Yor process. We furthermore provide an application to widely employed mixture models. These results have led us to further theoretical investigations regarding the connection between Dirichlet random means and continual Young diagrams.

The second part of the thesis is instead devoted to the partially exchangeable regime, a generalization of exchangeability which encompasses a more complex dependence structure among observations naturally divided in groups. We rely on hierarchical discrete random probability measures to enforce such distributional invariance in a model for clustering of nodes in multilayer networks. The induced distribution on the space of sequences of consistent partitions, determined by partially exchangeable partition probability functions, allows for theoretically validated prediction regarding new nodes incoming into the network.

Chapter 1

Preliminaries

1.1 Introduction

Random probability measures are the basic building block of popular and well-established Bayesian nonparametric procedures. This is apparent in the standard framework of exchangeable observations, which stands as the natural counterpart of the independence and identity in distribution assumption in frequentist statistics and entails invariance of the finite dimensional distributions with respect to permutation of the indices. This fundamental form of probabilistic symmetry encompasses the idea that the order in which we observe data realizations is not important for performing inference. By virtue of the celebrated representation theorem in de Finetti (1937), a sequence $(X_n)_{n\geq 1}$ of random elements valued in some space $\mathbb{X}$, is exchangeable if and only if there exists a random probability measure $\tilde{P}$ on $\mathbb{X}$, such that

$$X_1, \dots, X_n \mid \tilde{P} \overset{\text{iid}}{\sim} \tilde{P} \tag{1.1}$$

for any $n \geq 1$. Hence, the law of $\tilde{P}$ plays the role of prior distribution and reflects pre-experimental information on the data generating distribution.

In this thesis we focus on discrete random probabilities

$$\tilde{P} = \sum_{i\in\mathcal{I}} \omega_i \, \delta_{Z_i}, \tag{1.2}$$

on some space $\mathbb{X}$, where $\mathcal{I}$ is countable, the sequences $(\omega_i)_{i\in\mathcal{I}}$ and $(Z_i)_{i\in\mathcal{I}}$ are independent, with $\sum_{i\in\mathcal{I}} \omega_i = 1$, almost surely. Moreover the Z_i's are independent and identically distributed from some probability measure P_0 on $\mathbb{X}$ that takes on the name of *parameter measure* or *baseline measure*. It is further easily seen that $\mathbb{E}[\,\tilde{P}\,] = P_0$. Such random probability measures have been extensively employed in Bayesian statistical modeling, as accounted, *e.g.*, in Müller et al. (2015), Phadia

(2016) and Ghosal and van der Vaart (2017), as well as thoroughly studied in probabilistic contexts, because of their intrinsic connection with random partitions theory and in general with combinatorial stochastic processes. See Kingman (1975, 1978, 1982) and Pitman (1995, 1996, 2006).

The well-known Dirichlet process, introduced in Ferguson (1973), is the most celebrated example of such nonparametric priors. Moreover, it has spurred several well-established extensions: from the two-parameter version in Pitman and Yor (1997b), to normalized completely random measures in Regazzini et al. (2003), to Gibbs-type priors in Gnedin and Pitman (2005) and De Blasi et al. (2015). A noteworthy use of such discrete random probability measures is in mixture models for density estimation. See Lo (1984) for the Dirichlet process and, *e.g.*, Ishwaran and James (2001), Lijoi et al. (2007) and Barrios et al. (2013) for its extensions. Other important early generalizations are represented by tail-free processes and neutral-to-the-right processes, in Fabius (1973), Ferguson (1974) and Doksum (1974).

Theoretical investigations on random probability measures have been fueled by inferential problems in Bayesian statistics, leading to remarkable advances and the proposal of new modeling tools. An inspiring and appealing area of research is the one focusing on linear functionals of random probability measures. After the seminal contribution by Cifarelli and Regazzini (1990), the developments in the field have been impressive and several connections with seemingly unrelated areas of mathematics have emerged. See Lijoi and Prünster (2009) for a review.

The first part of this thesis, namely Chapters 2 and 3, focuses on means of random probability measures. Chapter 2 is mainly devoted to providing prior elicitation for commonly used nonparametric priors by enforcing a desired distribution on their linear functionals. More specifically, following an inverse path with respect to the existing literature on the topic, we establish explicit expressions for the base measure of a discrete random probability measure inducing a broad class of distributions on the mean, when such random measure is either a Dirichlet process, a normalized stable process, or a Pitman–Yor process. We further show that these results can be extended to popular nonparametric mixture models that are customarily used for Bayesian density estimation. This investigation naturally leads to theoretical questions about the structure of the space of such mean distributions, which find parallels in different fields of mathematics, from combinatorics to statistical physics. In Chapter 3, we explore in particular the connection with transition measures induced by hook walks on continual Young diagrams.

On the other hand, the great availability of data featuring complex dependence structures has driven the investigation to explore the statistical implications of probabilistic symmetries more general than exchangeability. Indeed, in the last two decades a lively area of research has been contributing to the development of theory and models encompassing more flexible forms of distributional invariance. Among these, one of the most popular is partial exchangeability, that was introduced by de Finetti (1938).

In the second part of the thesis, namely Chapter 4, we propose a model for a connectivity-driven clustering of nodes in multilayer node-colored networks, based on hierarchical priors on the space of partitions, inducing partial exchangeability on a latent array of characteristics of the nodes. The proposal has a probabilistically coherent structure. Indeed, we are able to evaluate partially exchangeable partition probability functions, which entail distributions on the space of consistent sequences of partitions of a growing network. This, differently from previous proposals in the network analysis literature, allows for a theoretically validated prediction. Moreover, the analytical tractability of these priors grants explicit evaluation of predictive clustering and co-clustering probabilities, and in turn permits elicitation of hyperparameters choice. We devise Markov chain Monte Carlo algorithms to perform posterior inference and prediction both on the allocations and on the connections of new nodes entering the network. Further work addressing the construction of models enforcing forms of distributional invariance other than partial exchangeability, that better suit different types of multilayer networks, is briefly outlined.

In the rest of this Chapter we provide a brief overview of concepts and tools which are fundamental for the derivation of the main results in the thesis. We start from exchangeability, as it represents a cornerstone of the Bayesian inferential scheme and exemplifies the tight connection with the theory of random probability measures. An effective strategy for defining discrete random probability measures is through transformation of completely random measures. Such constructions, combined with powerful analytical tools that we are going to present, have been fundamental both for the study of random means in Chapters 2 and 3, and for the formulation of the model for clustering in multilayer networks in Chapter 4. As far as the latter is concerned, partial exchangeability plays a key role. It is worth noting that for arrays fulfilling this distributional invariance a representation theorem holds true and it is based on the joint law of vectors of random probability measures. A particular way of constructing partially exchangeable arrays of observables is given by hierarchical random probability measures. Such hierarchies, together with the almost sure discreteness of the involved random probability measures, induce distributions on the space of partitions to be used as priors in structured clustering problems, included the one we confront in Chapter 4. Hierarchical compositions of discrete random probability measures obtained via transformations of completely random measures are, in these regards, powerful and analytically tractable tools, since they inherit useful properties and induce, in turn, handy distributions on the space of sequences of consistent partitions. The probability mass function determining such a distribution is called partially exchangeable partition probability function and is a brilliant example of an elegant mathematical object that has not been thoroughly investigated in the probabilistic literature yet.

1.2 Exchangeable sequences

Let $\mathbb{X}$ be a complete and separable metric space equipped with the Borel σ-algebra $\mathscr{X}$ and denote by $\mathscr{P}$ the space of probability distributions defined on $(\mathbb{X}, \mathscr{X})$ and endowed with the topology of weak convergence. We denote with $\sigma(\mathscr{P})$ the Borel σ-algebra of subsets of $\mathscr{P}$. We will indicate with $[n]$ the set of the first n natural numbers $\{1, \ldots, n\}$, for any $n \in \mathbb{N}$. For a sequence of $\mathbb{X}$-valued random elements $\boldsymbol{X} = (X_n)_{n\geq 1}$, defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ we have the following.

Definition 1.1. *$\boldsymbol{X}$ is exchangeable if for any $n \geq 1$*

$$(X_1, \ldots, X_n) \stackrel{d}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)}) \tag{1.3}$$

for any $\pi \in S_n$, the symmetric group of $[n]$.

By virtue of the representation result in de Finetti (1937), we have the following.

Theorem 1.1 (de Finetti)**.** *$\boldsymbol{X}$ is exchangeable if and only if there exists a probability measure Q on the space of probability distributions $(\mathscr{P}, \sigma(\mathscr{P}))$ such that*

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \int_{\mathscr{P}} \prod_{i=1}^{n} P(A_i)\, Q(\mathrm{d}P) \tag{1.4}$$

for any $A_1, \ldots, A_n$ in $\mathscr{X}$ and $n \geq 1$.

The probability measure Q directing the exchangeable sequence $(X_n)_{n\geq 1}$ is also termed *de Finetti measure* and takes on the interpretation of prior distribution in the Bayesian framework. The representation theorem can be equivalently stated by saying that, given an exchangeable sequence $(X_n)_{n\geq 1}$, there exists a random probability measure $\tilde{P}$, defined on $(\mathbb{X}, \mathscr{X})$ and taking values in $(\mathscr{P}, \sigma(\mathscr{P}))$, such that, for any $n \in \mathbb{N}$

$$\begin{aligned} X_1, \ldots, X_n \mid \tilde{P} &\stackrel{\text{iid}}{\sim} \tilde{P} \\ \tilde{P} &\sim Q \end{aligned} \tag{1.5}$$

Discrete nonparametric priors Q, namely priors Q selecting discrete distributions with probability 1, represent a fundamental tool in Bayesian nonparametrics. The most popular example is the Dirichlet process (DP), introduced in Ferguson (1973). However, as shown in Lijoi and Prünster (2010) most classes of discrete nonparametric priors, including the DP, can be seen as suitable transformations of completely random measures, presented in Kingman (1967). In order to introduce the other specific processes we will deal with, it is therefore worth briefly describing this unifying concept.

Let $\mathscr{M}$ be the set of boundedly finite measures on $\mathbb{X}$ equipped with the corresponding Borel σ-algebra on $\sigma(\mathscr{M})$. For details on the definition of this σ-algebra, see Daley and Vere-Jones (2007). A completely random measure (CRM) $\tilde{\mu}$ on $(\mathbb{X}, \mathscr{X})$ is a measurable function on $(\Omega, \mathscr{F}, \mathbb{P})$ taking values in $\mathscr{M}$ such that for any $k \geq 2$ and pairwise disjoint sets $A_1, \ldots, A_k$ in $\mathscr{X}$ the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_k)$ are independent. CRMs have been introduced in Kingman (1967). Detailed treatments can be found in Kingman (1993) and Daley and Vere-Jones (2007). Any CRM $\tilde{\mu}$ fulfills the following representation. For $A \in \mathscr{X}$

$$\tilde{\mu}(A) = \sum_{k \geq 1} U_k \delta_{x_k} + \beta(A) + \int_0^\infty s\, \tilde{N}(\mathrm{d}s,\, A) \tag{1.6}$$

where $(x_k)_{k\geq 1}$ is a (countable) sequence in $\mathbb{X}$, $(U_k)_{k\geq}$ is a sequence of independent non-negative random variables, β is a fixed non-atomic boundedly finite measure on $(\mathbb{X}, \mathscr{X})$, and $\tilde{N}$ is a Poisson process on $\mathbb{R}^+ \times \mathbb{X}$ independent of $(U_k)_{k\geq}$ and whose parameter measure ν satisfies

$$\int_{\mathbb{R}^+} \int_B \min\{s, 1\}\, \nu(\mathrm{d}s, \mathrm{d}x) < \infty \tag{1.7}$$

for any bounded B in $\mathscr{X}$. In words, we can think of a CRM as a superposition of a random measure with fixed atoms, a deterministic non-atomic drift and a part characterized by random jumps and random locations, whose intensities, distribution and mutual dependence are governed by a Poisson process.

We will focus on CRMs $\tilde{\mu}$ with no fixed atoms and no drift. For our purposes it is important to remind that such CRMs are almost surely discrete. Their Laplace functional admits the following *Lévy–Khintchine* representation.

$$\mathbb{E}\left[\mathrm{e}^{-\int_{\mathbb{X}} f(x)\, \tilde{\mu}(\mathrm{d}x)}\right] = \exp\left\{-\int_{\mathbb{R}^+ \times \mathbb{X}} \left[1 - \mathrm{e}^{-sf(x)}\right]\, \nu(\mathrm{d}s, \mathrm{d}x)\right\} \tag{1.8}$$

where $f : \mathbb{X} \to \mathbb{R}$ is a measurable function such that $\int |f|\, \mathrm{d}\tilde{\mu} < \infty$ almost surely. The measure ν is known as the *Lévy intensity* of $\tilde{\mu}$ and regulates the intensity of the jumps of a CRM and their locations. By virtue of (1.8), it characterizes the CRM $\tilde{\mu}$.

Let us introduce some noteworthy CRMs which will be employed in the rest of the thesis. Let α be a measure on $(\mathbb{X}, \mathscr{X})$, $\sigma \in (0, 1)$ and consider a CRM $\tilde{\mu}_\sigma$ with Lévy intensity defined by

$$\nu(\mathrm{d}s, \mathrm{d}x) = \frac{\sigma}{\Gamma(1-\sigma)} s^{-1-\sigma}\, \mathrm{d}s\, \alpha(\mathrm{d}x) \tag{1.9}$$

Then $\tilde{\mu}_\sigma$ is a *σ-stable CRM* with parameter measure α on $\mathbb{X}$. Moreover, for any measurable function $f : \mathbb{X} \to \mathbb{R}$ such that $\int |f|^\sigma\, \mathrm{d}\alpha < \infty$, the Laplace functional is of the form

$$\mathbb{E}\left[\mathrm{e}^{-\int f\,\mathrm{d}\tilde{\mu}_\sigma}\right] = \mathrm{e}^{-\int f^\sigma\,\mathrm{d}\alpha} \tag{1.10}$$

Hence, for any $B \in \mathscr{X}$ with $\alpha(B)$, the Laplace transform of $\tilde{\mu}_\sigma(B)$ is that of a positive stable random variable, namely

$$\mathbb{E}\left[\mathrm{e}^{-\lambda\tilde{\mu}_\sigma(B)}\right] = \mathrm{e}^{-\lambda^\sigma\,\alpha(B)} \tag{1.11}$$

If instead we consider the following Lévy intensity

$$\nu(\mathrm{d}s,\mathrm{d}x) = e^{-s}s^{-1}\,\mathrm{d}s\,\alpha(\mathrm{d}x) \tag{1.12}$$

we shall obtain a *gamma CRM* $\tilde{\mu}$. It is characterized by its Laplace functional

$$\mathbb{E}\left[\mathrm{e}^{-\int f\,\mathrm{d}\tilde{\mu}}\right] = \mathrm{e}^{-\int \log(1+f)\,\mathrm{d}\alpha} \tag{1.13}$$

for any measurable real-valued function f such that $\int \log(1+|f|)\,\mathrm{d}\alpha < \infty$. Therefore, for any $B \in \mathscr{X}$ such that $\alpha(B) < \infty$

$$\mathbb{E}\left[\mathrm{e}^{-\lambda\tilde{\mu}(B)}\right] = (1+\lambda)^{-\alpha(B)} \tag{1.14}$$

that implies a gamma distribution with scale parameter equal to 1 and shape parameter equal to $\alpha(B)$ for the random variable $\tilde{\mu}(B)$.

Lastly, an allied CRM is given by the *generalized gamma CRM*, introduced in Brix (1999). We will denote it with $\tilde{\mu}_{\sigma,\tau}$. It is characterized by a Lévy intensity of the form

$$\nu(\mathrm{d}s,\mathrm{d}x) = \frac{1}{\Gamma(1-\sigma)}\,s^{-1-\sigma}\,\mathrm{e}^{-\tau s}\mathrm{d}s\,\alpha(\mathrm{d}x) \tag{1.15}$$

where $\sigma \in (0,1)$ and $\tau > 0$. Note that if $\tau = 0$ then it coincides with the σ-stable CRM $\tilde{\mu}_\sigma$, whereas if $\sigma \to 0$ one obtains the gamma CRM. The Laplace functional is of the form

$$\mathbb{E}\left[\mathrm{e}^{-\int f\,\mathrm{d}\tilde{\mu}_{\sigma,\tau}}\right] = \mathrm{e}^{-\frac{1}{\sigma}\int(\tau+f)^\sigma\,\mathrm{d}\alpha - \tau^\sigma\alpha(\mathbb{X})} \tag{1.16}$$

for any measurable real-valued function f such that $\int(\tau+|f|)^\sigma\,\mathrm{d}\alpha < \infty$. If we choose, fo example, $\sigma = 1/2$, then for $B \in \mathbb{X}$ such that $\alpha(B) < \infty$ we have

$$\mathbb{E}\left[\mathrm{e}^{-\lambda\tilde{\mu}_{\sigma,\tau}}(B)\right] = \mathrm{e}^{-\alpha(B)\sqrt{\tau+\lambda}-\sqrt{\tau}} \tag{1.17}$$

that is $\tilde{\mu}_{\sigma,\tau}(B)$ has an inverse Gaussian distribution with parameters τ and $\alpha(B)$.

The first transformation of CRMs we are going to consider for obtaining a random probability measure is normalization. If we impose the condition $0 < \tilde{\mu}(\mathbb{X}) < \infty$ almost surely, which is implied by $\nu(\mathbb{R}^+ \times \mathbb{X}) = \infty$ and

$$\int_{\mathbb{R}^+\times\mathbb{X}} \left[1-\mathrm{e}^{-\lambda s}\right]\,\nu(\mathrm{d}s,\mathrm{d}x) < \infty \tag{1.18}$$

for any $\lambda > 0$, as proved in Regazzini et al. (2003), then it is possible to define

$$\tilde{P}(A) := \frac{\tilde{\mu}(A)}{\tilde{\mu}(\mathbb{X})} \tag{1.19}$$

for any $A \in \mathscr{X}$. Random probability measures as in (1.19) form the class of *normalized random measures with independent increments* (NRMIs), introduced in Regazzini et al. (2003), and they are characterized by ν. Moreover, in the case of $\tilde{\mu}$ being a *homogeneous CRM*, that is such that its Lévy intensity factorizes as $\nu(\mathrm{d}s, \mathrm{d}x) = \rho(s)\,\mathrm{d}s\,\alpha(\mathrm{d}x)$ for some measurable positive function ρ on $\mathbb{R}^+$ and some measure α on $\mathbb{X}$, we may represent $\tilde{\mu}$ through ρ and α. Intensities in (1.9) and (1.12) are in this class. Notice that (1.18) implies the finiteness of α. Setting $c := \alpha(\mathbb{X})$, we have

$$\nu(\mathrm{d}s, \mathrm{d}x) = \rho(s)\,\mathrm{d}s\,c\,G(\mathrm{d}x) \tag{1.20}$$

where $G(\cdot) := \frac{\alpha(\cdot)}{\alpha(\mathbb{X})}$ is now a probability measure on $\mathbb{X}$. We may write then

$$\tilde{P} \sim \mathrm{NRMI}(\rho,\, c,\, G) \tag{1.21}$$

for some measurable positive ρ, some $c > 0$ and some probability measure G. Notice that $\mathbb{E}\left[\tilde{P}(A)\right] = G(A)$ for any $A \in \mathscr{X}$, which means that, if $X \mid \tilde{P} \sim \tilde{P}$ then $\mathbb{P}\,(X \in A) = G(A)$ for any $A \in \mathscr{X}$. For particular choices of ρ one gets noteworthy nonparametric priors. Taking $\rho(s) = e^{-s}s^{-1}$ as in (1.12), we are normalizing a gamma CRM: the correspondent $\tilde{P}$ in (1.19) is a DP, which we will denote as $\tilde{\mathscr{D}}_\alpha$. In this case we will indicate the total mass c, also called *concentration parameter*, with $\theta > 0$. Hence, we shall also write $\tilde{P} \sim \mathrm{DP}(\theta, P_0)$, where $\theta = \alpha(\mathbb{X})$ and $P_0 = \mathbb{E}[\tilde{P}]$.

If instead we choose ρ as in (1.9), $\tilde{P}$ in (1.19) is called *normalized stable process* (NSP), introduced in Kingman (1975). With the intensity in (1.15), we obtain the *normalized generalized gamma*.

The popular Pitman–Yor process (PYP) is another noteworthy example of prior that is related to CRMs, more precisely the σ-stable CRM, through a different type of transformation. See Pitman and Yor (1997b). Denote by $\mathbb{P}_\sigma$ the law of a σ-stable CRM, $\tilde{\mu}_\sigma$. Now one can define a random measure $\tilde{\mu}_{\sigma,\theta}$ with distribution $\mathbb{P}_{\sigma,\theta}$ absolutely continuous with respect to $\mathbb{P}_\sigma$ and such that

$$\frac{\mathrm{d}\mathbb{P}_{\sigma,\theta}}{\mathrm{d}\mathbb{P}_\sigma}(\tilde{\mu}) = \frac{[\tilde{\mu}(\mathbb{X})]^{-\theta}}{\mathbb{E}[\tilde{\mu}_\sigma(\mathbb{X})^{-\theta}]} \tag{1.22}$$

for some $\theta > -\sigma$. Note that $\tilde{\mu}_{\sigma,\theta}$ is not a CRM, but we can still normalize it, defining for any $A \in \mathscr{X}$

$$\tilde{P}_{\sigma,\theta}(A) = \frac{\tilde{\mu}_{\sigma,\theta}(A)}{\tilde{\mu}_{\sigma,\theta}(\mathbb{X})} \tag{1.23}$$

which is distributed as a PYP. We will write $\tilde{P}_{\sigma,\theta} \sim \text{PYP}(\sigma,\, \theta,\, G)$. It is indeed easy to see that $\mathbb{E}\left[\tilde{P}_{\sigma,\theta}(A)\right] = G(A)$, where G is the parameter measure of $\tilde{\mu}_\sigma$ in (1.20).

Interestingly, as shown in Pitman (2003), also the normalized generalized gamma can be obtained via an argument similar to the above and still from the σ-stable CRM. In fact via an exponential tilting one can define a random measure $\tilde{\mu}_{\sigma,\tau}$ via

$$\frac{\mathrm{d}\mathbb{P}_{\sigma,\tau}}{\mathrm{d}\mathbb{P}_\sigma}(\tilde{\mu}) = \mathrm{e}^{\tau - \tau^{1/\sigma}\tilde{\mu}(\mathbb{X})}, \tag{1.24}$$

which coincides in distribution with the generalized gamma CRM, characterized by (1.15). Therefore, in contrast with the previous case, one still has a CRM and, obviously, by normalization one obtains a normalized generalized gamma. Therefore both the Pitman–Yor process and the normalized generalized gamma are suitable transformations of the σ-stable CRM $\tilde{\mu}_\sigma$.

Being obtained via normalization or power tilting of CRMs, both NRMIs and PYPs naturally inherit almost sure discreteness.

It is worth to mention that another popular way to introduce discrete random probability measures, and hence nonparametric priors, is the *stick-breaking* construction. It focuses on modeling the weights ω_i of the series representation in (1.2) and is based on the idea that, since these weights have to sum up to 1, they can be seen as the lengths of pieces obtained by sequentially breaking a stick of length 1. In symbols

$$\begin{aligned} \omega_i &= V_i \prod_{k=1}^{i-1}(1 - V_k) \quad i \geq 2 \\ \omega_1 &= V_1 \end{aligned} \tag{1.25}$$

for a sequence of random variables $(V_k)_{k\geq 1}$ valued in $[0, 1]$. Hence, each ω_i is the V_i portion of the remaining part after the $(i-1)$-th break. By choosing a law for the sequence $(V_k)_{k\geq 1}$, one obtains a distribution on discrete probability measures. With

$$V_1, \ldots, V_k, \ldots \overset{\text{iid}}{\sim} \text{beta}(1, \theta) \tag{1.26}$$

a Dirichlet process with concentration parameter θ is recovered. This construction has been introduced in Sethuraman (1994) for the DP. The stick-breaking representation of the PYP is given in Perman et al. (1992): for a PYP(σ,θ), we choose

$$V_i \overset{\text{ind}}{\sim} \text{beta}(1 - \sigma, \theta + i\sigma) \tag{1.27}$$

This construction has become widely popular mostly for computational reasons. Nonetheless, when it comes to studying theoretical properties of such discrete random probability measures, they are not very convenient. This explains why we will not refer to stick-breaking representations henceforth.

1.3 Partially exchangeable arrays

Introduced in de Finetti (1938), partial exchangeability is often a natural hypothesis of dependence for random elements grouped in a finite number of blocks. Given an array $(X_{ji})_{i=1,\dots,N_j}^{j=1,\dots,d}$ of $\mathbb{X}$-valued random elements, where the d rows represent the groups and, for any $j \in [d]$, N_j is the number of individuals in each group, while $N := \sum_{j=1}^{d} N_j$ is the total number of individuals, we can state the following.

Definition 1.2. $(X_{ji})_{i=1,\dots,N_j}^{j=1,\dots,d}$ *is partially exchangeable if*

$$(X_{11}, \dots, X_{1N_1}, \dots, X_{d1}, \dots, X_{dN_d}) \stackrel{d}{=} \\ \stackrel{d}{=} (X_{1\pi_1(1)}, \dots, X_{1\pi_1(N_1)}, \dots, X_{d\pi_d(1)} \dots, X_{d\pi_d(N_d)}) \quad (1.28)$$

for $\pi_j \in S_{N_j}$, *the symmetric group of* $[N_j]$, *for any* $j \in [d]$.

In words, the joint distribution of the array's entries has to be invariant with respect to *intra*-group permutations, but not necessarily with respect to *inter*-groups ones. An infinite array $\boldsymbol{X} = \{(X_{ji})_{i\geq 1} \mid j \in [d]\}$, is partially exchangeable if it fulfills (1.28) for any $(N_1, \dots, N_d) \in \mathbb{N}^d$. For such partially exchangeable infinite arrays, it holds the following representation.

Theorem 1.2 (de Finetti)**.** *The infinite random array* $\boldsymbol{X}$ *is partially exchangeable if and only if there exists a measure* Q *on the space of* d*-dimensional vectors of probability measures* $\mathscr{P}^d$ *endowed with the product* σ*-algebra* $\bigotimes_{j=1}^{d} \sigma(\mathscr{P})$ *such that*

$$\mathbb{P}\left(X_{11} \in A_{11}, \dots, X_{1N_1} \in A_{1N_1}, \dots, X_{d1} \in A_{d1}, \dots, X_{dN_d} \in A_{dN_d}\right) = \\ = \int_{\mathscr{P}^d} \prod_{j=1}^{d} \prod_{i=1}^{N_j} P_j(A_{ji})\, Q(\mathrm{d}P_1, \dots, \mathrm{d}P_d) \quad (1.29)$$

for any $A_{11}, \dots, A_{dN_d} \in \mathscr{X}$ *and any* $(N_1, \dots, N_d) \in \mathbb{N}^d$.

Again the probability measure Q is called *de Finetti measure*. An equivalent statement of Theorem 1.2 is the following. Given a partially exchangeable infinite array $\boldsymbol{X} = \{(X_{ji})_{i\geq 1} \mid j \in [d]\}$, there exists a vector of random probability measures $(\tilde{P}_1, \dots, \tilde{P}_d)$, each defined on $(\mathbb{X}, \mathscr{X})$, and taking values in $\left(\mathscr{P}^d, \bigotimes_{j=1}^{d} \sigma(\mathscr{P})\right)$, such that, for any $(N_1, \dots, N_d) \in \mathbb{N}^d$

$$\begin{aligned} X_{j1}, \dots, X_{jN_j} \mid (\tilde{P}_1, \dots, \tilde{P}_d) &\stackrel{\text{iid}}{\sim} \tilde{P}_j \qquad \forall j \in [d] \\ (\tilde{P}_1, \dots, \tilde{P}_d) &\sim Q \end{aligned} \quad (1.30)$$

Notice that partial exchangeability is indeed a generalization of exchangeability, since for $\tilde{P}_j = \tilde{P}$ almost surely for any $j \in [d]$ and some random probability measure $\tilde{P}$, then (1.30) boils down to (1.5). On the other hand if $(\tilde{P}_j)_{j=1,\dots,d}$ are independent, that is Q is a product law on $\mathscr{P}^d$, then the rows of the array $\boldsymbol{X}$ are independent. These particular cases represent the extremes in the range of borrowing of information among observables coming from different sub-populations. For an approach to measuring the distance from full exchangeability of a couple of random probability measures, see Catalano et al. (2021).

Defining a partially exchangeable model for an array, then, amounts to choosing a de Finetti measure Q for a vector of random probability measures. Several constructions have been proposed in the literature, starting from the pioneering contributions of MacEachern (1999, 2000). Among these we mention additive structures as in Müller et al. (2004) and Lijoi et al. (2014), nested structures as in Rodríguez et al. (2008), hierarchical structures as in Teh et al. (2006) and Camerlenghi et al. (2019), or even a combination of the last two as in Beraha et al. (2021) and Lijoi et al. (2022).

Since in Chapter 4 we confront a clustering task for nodes of multilayer networks, which we shall see as statistical units belonging to different sub-populations, and we want to put positive prior mass on partitions including individuals from different groups, we will investigate hierarchical compositions of discrete random probability measures. In Section 1.2 we recalled the construction of NRMIs and PYP. Here, as anticipated, we will give definitions for hierarchical compositions of such nonparametric priors, called *hierarchical normalized random measures with independent increments* (H-NRMIs) and *hierarchical Pitman–Yor process* (H-PYP). These prominent examples of partially exchangeable models are presented and thoroughly studied in Camerlenghi et al. (2019). We can define a vector of random probability measures being H-NRMIs or having H-PYP distribution as follows.

Definition 1.3. *$(\tilde{P}_1, \dots, \tilde{P}_d)$ is a vector of H-NRMIs on $(\mathbb{X}, \mathscr{X})$ with parameters $(\rho, \rho_0, c, c_0, P_0)$ if*

$$\begin{aligned} \tilde{P}_1, \dots, \tilde{P}_d \mid \tilde{P}_0 &\overset{iid}{\sim} \mathrm{NRMI}\left(\rho, c, \tilde{P}_0\right) \\ \tilde{P}_0 &\sim \mathrm{NRMI}\left(\rho_0, c_0, P_0\right) \end{aligned} \tag{1.31}$$

where $\tilde{\nu}(\mathrm{d}s, \mathrm{d}x) = \rho(s)\, \mathrm{d}s\, c\, \tilde{P}_0(\mathrm{d}x)$ and $\nu_0(\mathrm{d}s, \mathrm{d}x) = \rho_0(s)\, \mathrm{d}s\, c_0\, P_0(\mathrm{d}x)$ are the Lévy intensities of $\tilde{P}_j \mid \tilde{P}_0$ for any $j \in [d]$ and $\tilde{P}_0$, respectively, with ρ and ρ_0 measurable positive functions, $c, c_0 > 0$ and P_0 a diffuse probability measure on $(\mathbb{X}, \mathscr{X})$.

Definition 1.4. *The vector $(\tilde{P}_1, \dots, \tilde{P}_d)$ has H-PYP distribution with parameters*

$(\sigma,\, \sigma_0,\, \theta,\, \theta_0,\, P_0)$ *if*

$$\begin{aligned} \tilde{P}_1, \dots, \tilde{P}_d \mid \tilde{P}_0 &\overset{iid}{\sim} \mathrm{PYP}\left(\sigma,\, \theta,\, \tilde{P}_0\right) \\ \tilde{P}_0 &\sim \mathrm{PYP}\left(\sigma_0,\, \theta_0,\, P_0\right) \end{aligned} \tag{1.32}$$

with $\sigma,\, \sigma_0 \in (0,1)$, $\theta > -\sigma$, $\theta_0 > -\sigma_0$ *and* P_0 *a diffuse probability measure on* $(\mathbb{X},\, \mathscr{X})$.

Example 1.1. If on both the levels of hierarchy in (1.31) we put a DP we obtain the hierarchical Dirichlet process (H-DP), introduced in Teh et al. (2006). We write

$$(\tilde{P}_1, \dots, \tilde{P}_d) \sim \text{H-DP}(\theta,\, \theta_0,\, P_0) \tag{1.33}$$

If instead we choose two levels of NSP, we get the hierarchical normalized stable process (H-NSP), discussed in Camerlenghi et al. (2019). We denote it as

$$(\tilde{P}_1, \dots, \tilde{P}_d) \sim \text{H-NSP}(\sigma,\, \sigma_0,\, P_0) \tag{1.34}$$

Being (1.31) and (1.32) particular ways of setting the de Finetti measure Q in (1.30), we have that an array $\boldsymbol{X}$ as in (1.30) whose j-th row $(X_{j1}, \dots, X_{jN_j})$ is a conditionally iid drawn from $\tilde{P}_j$ as in (1.31) or (1.32) for any $j \in [d]$, is partially exchangeable according to its rows. Moreover, because of the almost sure discreteness of each $\tilde{P}_j$ and $\tilde{P}_0$, the probability of a tie in the array $\boldsymbol{X}$ is positive both within- and across-rows, that is

$$\mathbb{P}\left(X_{ij} = X_{i'j'}\right) > 0 \tag{1.35}$$

for any $i \in [N_j]$, $i' \in [N_{j'}]$ and $j, j' \in [d]$. This fact will be crucial to obtain clusters sharing elements in different sub-populations, defined by the ties among the entries of the partially exchangeable array. In Chapter 4 we will recall further properties of H-NRMIs and H-PYP, above all the induced partially exchangeable partition probability functions driving the distribution on the space of partitions.

Chapter 2

Random probability measures with fixed mean distributions

Abstract

Linear functionals, or means, of discrete random probability measures play a key role in several areas of mathematics, including statistics, combinatorics, special functions, excursions of stochastic processes and financial mathematics, among others. Our interest is motivated by its relevance in Bayesian nonparametric inference, where the law of a random probability measure acts as a prior distribution. The literature on the topic has aimed at determining the distribution of such linear functionals when the prior is a Dirichlet process, a Pitman–Yor process and a normalized random measure with independent increments. This work addresses the inverse problem and focuses on the identification of the baseline measure of a discrete random probability measure yielding a specific mean distribution. This is extremely useful in Bayesian inference as it is often the case that a statistician has pre-experimental information about a finite-dimensional projection of the data generating distribution, such as the mean, that may be of better use in prior elicitation than an infinite-dimensional parameter. This research direction has been pursued, with different motivations, in the combinatorial literature and the only available result concerns the Dirichlet process with unit concentration parameter. Here we address the problem in greater generality to cover the Dirichlet process with concentration parameter not necessarily equal to 1, the normalized stable process and the Pitman–Yor process. Finally, we deal with an extension of our findings to a popular class of mixture models used for density estimation and clustering.

2.1 Introduction

As discussed in Section 1.1, discrete random probability measures, on top of being probabilistic objects of undoubted interest, are crucial in Bayesian nonparametrics. In this framework, and referring to a discrete random probability measure $\tilde{P}$ as in

(1.2), linear functionals of the type

$$M_h(\tilde{P}) := \int_{\mathbb{X}} h(x)\tilde{P}(\mathrm{d}x) = \sum_{i\in\mathcal{I}} \omega_i\, h(Z_i), \tag{2.1}$$

where $h : \mathbb{X} \to \mathbb{R}$ is some measurable function such that $\int |h|\mathrm{d}\tilde{P} < +\infty$ almost surely, inherit a pivotal position in nonparametric modeling, as finite dimensional projections of $\tilde{P}$. The investigation of distributional properties of linear functional of random probability measures has been pioneered by Cifarelli and Regazzini (1990). Since then, it has become a lively area of research in statistics that has generated many important results, from different mathematical perspectives, as reported in Diaconis and Kemperman (1996). On the other hand, it is worth stressing that the subject has attracted considerable interest in several other, and seemingly unrelated, fields of mathematics ranging from combinatorics to statistical physics. For example, Kerov (1993) and Kerov (1998) frame the investigation of the distribution of $M_h(\tilde{P})$ in terms of Markov and Hausdorff moment problems, while introducing the connection with transition measures induced by continual Young diagrams. These works have further spurred an extensive literature and means of discrete random probability measures have been treated in Tsilevich (1999), Kerov and Tsilevich (2004) and Vershik et al. (2004). Additionally, important connections with the theory of multivariate hypergeometric functions have been pointed out in Lijoi and Regazzini (2004) and these are used to determine new closed form expressions for the distribution of $M_h(\tilde{P})$, when $\tilde{P}$ is a Dirichlet process, and a representation of its characteristic function. Furthermore, general assumptions for an explicit expression of the mean density and the mean cumulative distribution function to be derived, as well as results on symmetry of the mean distribution and on vectors of Dirichlet random means can be found in Regazzini et al. (2002).

Means of Pitman–Yor processes are relevant in connection with that study of the excursions of Bessel processes, as thoroughly discussed in Perman et al. (1992) and Pitman and Yor (1997a). Such results can be traced back to the seminal works of Lévy (1939), where the Brownian case is covered, and of Lamperti (1958), where the arcsine laws are treated. Finally, we recall that means as in (2.1) appear also in the statistical physics literature, in relation to zero-range process models (see Evans and Hanney (2005) for a review), as can be found in Pulkkinen (2007). An extensive review on random means and on their uses in statistics, and beyond, can be found in Lijoi and Prünster (2009), whereas Lijoi and Prünster (2011) provides a historical perspective on the development of the subject in Bayesian statistics.

The existing literature, which we have referred to so far, addresses the problem of determining the probability distribution of $M_h(\tilde{P})$, once one has completely specified $\tilde{P}$ as in (1.2). See Cifarelli and Regazzini (1990) if $\tilde{P}$ is a Dirichlet process, James et al. (2008) if $\tilde{P}$ is a Pitman-Yor process and Regazzini et al. (2003) if $\tilde{P}$

is a normalized random measure with independent increments.

The work presented in this Chapter pursues a different, and in a sense opposite, task. Indeed, we aim at determining which $\tilde{P}$, within a specific class of discrete random probability measures, yields a specific distribution for the mean $M_h(\tilde{P})$. This amounts to identifying the parameter measure $\mathbb{E}[\tilde{P}] = P_0$, if there exists any, inducing the pre-specified law of the mean of $\tilde{P}$. The motivation for tackling this problem, from a Bayesian nonparametric modeling standpoint, is pragmatic. Indeed, as shown *e.g.* in Kessler et al. (2015), in many applications one might have enough *a priori* information for eliciting the distribution of an interpretable (and finite-dimensional) parameter of a nonparametric prior, such as its mean. The same arguments apply also to the case where the data generating distribution is absolutely continuous with respect to the Lebesgue measure and the prior is induced, *e.g.*, by the popular Dirichlet process mixture model for density estimation. Indeed, since the prior coincides with the law of the random density, $\tilde{f}(y) = \int k(y;x)\,\tilde{P}(\mathrm{d}x)$ for some fixed transition kernel density $k(\,\cdot\,;\,\cdot\,)$, the mean

$$\int_{\mathbb{R}} h(y)\tilde{f}(y)\,\mathrm{d}y = M_{\bar{h}}(\tilde{P}),$$

where $x \mapsto \bar{h}(x) = \int_{\mathbb{R}} h(y)\,k(y;x)\,\mathrm{d}y$, is still a linear functional of the Dirichlet process $\tilde{P}$. If we assume that available pre-experimental information allows for the elicitation of the law of $M_{\bar{h}}(\tilde{P})$ and $\tilde{P}$ is identified up to its parameter measure P_0, the latter can be specified so to enforce such prior knowledge on the mean. Hence, one notably achieves the desired distribution of $M_{\bar{h}}(\tilde{P})$, while still relying on a Dirichlet process mixture model that is very convenient for Bayesian computations and inference. This extends also to the cases where $\tilde{P}$ is a Pitman–Yor process or a normalized stable process.

Our interest in the problem has been further spurred by some intriguing work in combinatorics. Indeed, Romik (2004) investigates transition measures induced by hook walk on continual Young diagrams and refers to the connection with random means introduced in Kerov (1993). The problem addressed in these papers is basically the same we are considering here, though confined to the Dirichlet process. Indeed, as described in Chapter 3, for a specific instance of the Dirichlet process the baseline measure yielding a given distribution of the mean can be deduced from results in Romik (2004). Here we address the problem without resorting to this connection, which nonetheless has been inspiring in the interpretation of our results.

Taking the described inverse path and defining random probability measures with fixed mean distribution, leads to face additional difficulties. If on one side we rely on integral identities, known as Markov-Krein correspondences or Cifarelli-Regazzini identities, as well as on generalized Cauchy-Stieltjes transform inversion formulas, on the other, these classical tools cannot be directly applied in our case and new proof strategies need to be worked out to tackle the problem. Moreover, we need to assess existence and regularity for singular integrals, in order to identify hypotheses that allow a broad class of mean densities to be included. This allows us to determine closed form expressions for the parameter measure of a Dirichlet process, of a normalized stable process and of a Pitman-Yor process inducing a broad class of mean distributions. We further show that these results can be extended to popular nonparametric mixture models that are customarily used for Bayesian density estimation. Overall we believe that our findings represent an important step forward in nonparametric prior elicitation. Interestingly, our study unravels some relevant features of the underlying discrete random probability measures and of the corresponding space of mean distributions. For example, we show a surprising (at least to us) fact according to which not every absolutely continuous law with compact support is the mean distribution of a certain discrete random probability measure with given parameters.

For the sake of clarity of exposition, we will be assuming that $\tilde{P}$ or, equivalently, P_0 have compact support. This is not a restrictive assumption both from a theoretical and an applied perspective. On the one hand, the extension to the unbounded support case requires some strengthening of the hypotheses, without affecting the type of results we get though not providing any further insights to our contribution. On the other hand, nonparametric priors with unbounded support are most often approximated through random probability measures having compact, or even finite, support. This is typically the case when computational algorithms are used for the actual implementation of Bayesian nonparametric models to applied problems. It is also an effective strategy for determining the distribution of the mean, as in Guglielmi (1998) and Lijoi and Regazzini (2004), or for determining the posterior distribution of $\tilde{P}$, as in Regazzini and Sazonov (2000) and James et al. (2009). Hence, we are ultimately considering a general framework and pursuing a research direction that has been unexplored so far in the literature on random means and their applications. Our results also encompass means of random probability measures that are either discrete or absolutely continuous with respect to the Lebesgue measure, thus showing that the techniques we introduce are effective in addressing a wide variety of problems.

The structure of the Chapter is the following. Relying on general aspects concerning exchangeable sequences, completely random measures and normalized random measures with independent increments reviewed in Section 1.2, we briefly recall in Section 2.2 some fundamental results achieved in the study of distributional properties of random means and illustrate tools that will be crucial for achieving our goals. Section 2.3, thus, illustrates our main results. More specifically, we establish explicit expressions for the base measure P_0 inducing a broad class of distributions on the mean, when $\tilde{P}$ is either a Dirichlet process with concentration parameter $\theta < 1$, or a normalized stable process or a Pitman–Yor process. Moreover, we give examples of usage of our formulas for recovering noteworthy examples of mean distributions and discuss cases for which we cannot apply our main result, though we are still able to identify the baseline measure: this is very helpful for gaining some insight about the admissible sets of random mean distributions. Lastly, in Section 2.4 we deal with nonparametric mixtures widely used for density estimation.

2.2 Random means

The main object of interest of the work presented in this Chapter are linear functionals of random probability measures, that is

$$M_f(\tilde{P}) := \int_{\mathbb{X}} f(x)\tilde{P}(\mathrm{d}x),$$

where $f : \mathbb{X} \to \mathbb{R}$ is some measurable function and $\tilde{P}$ a random probability measure.

Key steps for the determination of the distribution of means $\int f \,\mathrm{d}\tilde{P}$ typically consist in representing them via suitable integral transforms and then in applying appropriate inversion formulae. We concisely describe the two most successful approaches to date, which we will partially exploit also in our results.

A first very convenient tool is the *generalized Cauchy–Stieltjes transform*: for a generic function $g : \mathbb{R}^+ \to \mathbb{R}$, it is defined as

$$\mathscr{S}_\lambda[z; g] := \int_{\mathbb{R}^+} \frac{g(x)}{(z + x)^\lambda} \,\mathrm{d}x \tag{2.2}$$

for any $\lambda > 0$ and for and $z \in \mathbb{C}$ such that $|\arg(z)| < \pi$. Inversion formulae for (2.2) are available and can be found, *e.g.*, in Sumner (1949) and Schwarz (2005). Under suitable conditions, for example $|z^\beta \mathscr{S}_\lambda[z; g]|$ is bounded at infinity for some $\beta > 0$, from Schwarz (2005) one has

$$g(x) = -\frac{x^\lambda}{2\pi\mathrm{i}} \int_{\mathscr{W}} (1 + w)^{\lambda-1} \mathscr{S}'_\lambda[xw; g] \,\mathrm{d}w \tag{2.3}$$

where $\mathscr{W}$ is a contour in the complex plane starting and ending at the point $w = -1$ and enclosing the origin in a counterclockwise sense, while $\mathscr{S}'_\lambda[xw; g] = \frac{\mathrm{d}}{\mathrm{d}z} \mathscr{S}_\lambda[z; g] \,|_{z=xw}$. If $\lambda > 1$, then one can integrate (2.3) by parts obtaining

$$g(x) = \frac{\lambda - 1}{2\pi\mathrm{i}} \, x^{\lambda-1} \int_{\mathscr{W}} (1 + w)^{\lambda-2} \mathscr{S}_\lambda[xw; g] \,\mathrm{d}w \tag{2.4}$$

For the case $\lambda = 1$, (2.3) reduces to the inversion formula originally derived by Widder, which can be found in Widder (2015). Dealing with a Dirichlet process-distributed random probability measure $\tilde{\mathscr{D}}_\alpha$, a closed form expression for the distribution of the mean of $\tilde{\mathscr{D}}_\alpha$ has been derived in Cifarelli and Regazzini (1990) by resorting to the inversion formula in Sumner (1949). A similar strategy has been pursued in James et al. (2008) for the evaluation of the (*a priori*) probability distribution of the mean functional of a Pitman–Yor process $\tilde{P}_{\sigma,\theta}$.

A second fruitful approach makes use of an inversion formula of the characteristic function due to Gurland (1948). If F is a cumulative distribution function on $\mathbb{R}$ and ϕ the corresponding characteristic function, then

$$F(y) - F(y-) = 1 - \frac{2}{\pi} \lim_{\varepsilon\downarrow 0,\, T\uparrow\infty} \int_\varepsilon^T \frac{1}{t} \operatorname{Im}\left[\mathrm{e}^{-\mathrm{i}yt}\,\phi(t)\right]\,\mathrm{d}t, \tag{2.5}$$

where $F(x-)$ is the left limit of F at y. Such an inversion formula is very useful when one aims at determining the distribution of ratios of random variables and, thus, suits perfectly to the case of NRMIs. To see, this let f be such that $\int |f|\,\mathrm{d}\tilde{\mu} < \infty$, almost surely and denote the cumulative distribution function of $\int f\,\mathrm{d}\tilde{P}$ by $y \mapsto Q(y) = \mathbb{P}\left(\int f\,\mathrm{d}\tilde{P} \le y\right)$, where $\tilde{P}$ is a NRMI. Following Regazzini et al. (2003), a crucial step consists in noting that

$$Q(y) := \mathbb{P}\left(\int_{\mathbb{X}} f(x)\tilde{P}(\mathrm{d}x) \le y\right) = \mathbb{P}\left(\int_{\mathbb{X}} [f(x) - y]\tilde{\mu}(\mathrm{d}x) \le 0\right) \tag{2.6}$$

which reduces the problem of studying a mean of a NRMI to the problem of studying a linear functional of a CRM. And, importantly, the characteristic functions of linear functionals of CRMs, analogously to the Laplace functional transform recalled in (1.8), have Lévy–Khintchine representation in terms of the underlying Lévy intensity measure. Therefore, from (2.5) one obtains

$$\begin{aligned}
&\frac{1}{2}\left\{Q(y) + Q(y-)\right\} \\
&= \frac{1}{2} - \frac{1}{\pi} \lim_{\epsilon\downarrow 0, T\uparrow+\infty} \int_\epsilon^T \frac{1}{t} \operatorname{Im} \exp\left\{-\int_{\mathbb{X}\times\mathbb{R}^+} [1 - \mathrm{e}^{\mathrm{i}tv\,(f(x)-y)}]\nu(\mathrm{d}v, \mathrm{d}x)\right\} \mathrm{d}t
\end{aligned}$$

where $\operatorname{Im} z$ stands for the imaginary part of $z \in \mathbb{C}$. The details for obtaining expressions for prior and posterior distributions of means of NRMI can be found in Regazzini et al. (2003) and James et al. (2010).

2.3 Fixing the distribution of the mean

In this Section we will focus on our principal task: given a random probability measure $\tilde{P}$ on $[0, 1]$, to determine the parameter measure inducing a desired distribution on the random mean

$$M(\tilde{P}) := \int_0^1 x\, \tilde{P}(\mathrm{d}x) \tag{2.7}$$

provided that such a measure is unique. Specifically, we shall give an explicit expression for the cumulative distribution function (cdf) of the base measure of $\tilde{P}$ enforcing a broad class of distributions on the random mean, when $\tilde{P}$ is Dirichlet, normalized stable and Pitman–Yor distributed. In the following we will see how considering (2.7) instead of a generic linear functional as in (2.1), is not restrictive. Moreover, the given statements can be easily extended to cover the case where the support of $\tilde{P}$ is any bounded interval of $\mathbb{R}$.

The problem of determining the base measure $\alpha := \theta P_0$, where $\theta > 0$ and P_0 is a probability measure on $[0, 1]$, of a Dirichlet process $\tilde{\mathscr{D}}_\alpha$ that leads to a specific probability distribution for the mean functional $M(\tilde{\mathscr{D}}_\alpha) = \int x\, \tilde{\mathscr{D}}_\alpha(\mathrm{d}x)$ is considered in Cifarelli and Regazzini (1993). Let $\mathbb{F}$ be the set of finite and non–null measures on $([0, 1], \mathscr{B}([0, 1]))$ and $\mathbb{F}_\theta := \{\alpha \in \mathbb{F} : \alpha([0, 1]) = \theta\}$. Moreover,

$$\mathbb{M}_\theta := \{\mathbb{P} \circ (M(\tilde{\mathscr{D}}_\alpha))^{-1} : \alpha \in \mathbb{F}_\theta\} \tag{2.8}$$

is the set of all probability distributions of the random Dirichlet mean $M(\tilde{\mathscr{D}}_\alpha)$ as α varies in $\mathbb{F}_\theta$. According to Theorem 2 in Lijoi and Regazzini (2004) any measure α in $\mathbb{F}_\theta$ is determined by the corresponding distribution Q_α in $\mathbb{M}_\theta$. This implies that, for random Dirichlet means, the total mass θ and Q_α in $\mathbb{M}_\theta$ uniquely identify the base measure $\alpha \in \mathbb{F}_\theta$. Furthermore, as a consequence of Theorem 10 in Lijoi and Regazzini (2004), Q_α is absolutely continuous with respect to the Lebesgue measure on $[0, 1]$ and we shall denote by q_α its density function. The correspondence between q_α and α is expressed by the Cifarelli–Regazzini (CR) identity

$$\mathscr{S}_\theta[z; q_\alpha] = \exp\left\{ -\theta \int_0^1 \log(z + x)\, P_0(\mathrm{d}x) \right\} \qquad z \in \mathbb{C} \setminus [-1, 0] \tag{2.9}$$

where $\mathscr{S}_\theta$ denotes the generalized Cauchy–Stieltjes transform of order θ as defined in (2.2). As mentioned, the explicit determination of q_α for a given P_0 is achieved in Lijoi and Regazzini (2004) for any $\theta > 0$.

When $\theta = 1$, an explicit solution to the inverse problem, that is the determination of $\alpha = P_0$ inducing a suitably smooth q_α, can be extrapolated from Romik (2004). In this work, *continual Young diagrams* and the *transition measure* they induce on a compact interval via *hook walks* are considered. See also Kerov (1993) for definitions, early results, and links to the Markov moment problem. If the Young diagram is convex, then it can be seen as a primitive function of a cumulative distribution function, which then corresponds to a probability distribution on the compact interval. In this case, the correspondence between the diagram and the induced transition measure is the same as the one between the base measure of a Dirichlet process with concentration parameter $\theta = 1$ and its mean distribution. Since in Romik (2004) an explicit expression of the derivative of the diagram as a function of the transition density is given, it is possible to leverage such result and obtain

$$P_0([0,x]) = \frac{1}{\pi} \operatorname{arccot}\left(\frac{1}{\pi\, q(x)} \, \mathrm{PV} \int_0^1 \frac{q(t)}{t-x} \, \mathrm{d}t \right) \tag{2.10}$$

for P_0 being the base measure of a Dirichlet process with $\theta = 1$, q being the density of the mean distribution and where $\mathrm{PV} \int$ stands for the Cauchy principal value integral (see Estrada and Kanwal (2012) for a full account on these analytical tools). The identity (2.10) has been proven to hold true for q piecewice C^1 with bounded derivative. This means that however we choose a mean density q with such regularity, we can explicitly identify the parameter measure P_0 such that if $\tilde{P} \sim \mathrm{DP}(1, P_0)$ then $M(\tilde{P}) \sim q$. Hook walks on continual Young diagrams, transition measures and this surprising connection with the Dirichlet process are treated in Chapter 3.

We provide results yielding the base measure that corresponds to a specific choice of the distribution of the mean for a Dirichlet process (DP) with concentration parameter $\theta < 1$ and for a normalized stable process (NSP). The interest for the latter case does not have only a theoretical motivation. In terms of applications, especially in the context of popular hierarchical mixture models, the NSP exhibits an appealing properties for density estimation and for inference on the clustering structure featured by the data. For instance, in Barrios et al. (2013) the NSP is suggested as a default prior because of its consistently good performance regardless of misspecifications of the prior parameters. Moreover, it also represents the basic building block for defining various alternative and widely used nonparametric priors for Bayesian inference. In view of this, indeed, we shall be able to deduce a similar representation for α that corresponds to a specified distribution of the mean of a Pitman–Yor process (PYP).

2.3.1 Base measure of a Dirichlet process

As previously reminded, the mean density of a DP given its base measure is determined in Lijoi and Regazzini (2004) for every $\theta > 0$, while an explicit expression of the base (probability) measure given a C^1 mean density can be derived from Romik (2004) in the case of concentration parameter $\theta = 1$. We solve the latter problem for every $\theta \in (0, 1)$.

For a density function f such that

$$\int_0^1 \frac{f(x)}{|x-t|^\theta} \, \mathrm{d}x < \infty \qquad \forall t \in [0,1]$$

with $\theta \in (0,1)$, we define

$$\mathscr{I}_\theta[\,f;\,t\,] := \frac{\displaystyle\int_t^1 \frac{f(x)}{|x-t|^\theta}\,\mathrm{d}x}{\displaystyle\int_0^t \frac{f(x)}{|x-t|^\theta}\,\mathrm{d}x} \qquad t \in (0,1] \tag{2.11}$$

Notice that $\lim_{t\to 0} \mathscr{I}_\theta[\,f;\,t\,] = \infty$ and $\mathscr{I}_\theta[\,f;\,1\,] = 0$. Moreover $\mathscr{I}_\theta[f;\,\cdot\,]$ is a monotonically decreasing function, hence we can consider its right continuous version, suitably modifying it in its at most countable jump discontinuity points. With a slight abuse of notation, we shall denote this version with $\mathscr{I}_\theta[\,\cdot\,;\,t\,]$ as well. Then we state the following.

Theorem 2.1. *Let $\theta \in (0,1)$ and q_α be the density of $M(\tilde{\mathscr{D}}_\alpha)$ with $\mathrm{supp}(q_\alpha) = [0,1]$. If*

$$\int_0^1 \frac{q_\alpha(x)}{|x-t|^\theta}\,\mathrm{d}x < \infty \qquad \forall t \in [0,1] \tag{2.12}$$

then the cdf of the base measure P_0 is given by

$$F_0(t) = \\ = \left\{ \frac{1}{\theta\pi} \arctan\left(\frac{\sin(\theta\pi)}{\cos(\theta\pi) + \mathscr{I}_\theta[\,q_\alpha;\,t\,]} \right) + \frac{1}{\theta} \mathbb{1}_{(t_*,\infty)}(t) \right\} \mathbb{1}_{(0,1)}(t) + \mathbb{1}_{[1,\infty)}(t) \tag{2.13}$$

where

$$t_* = \inf \left\{ t \in [0,1] \,\middle|\, \mathscr{I}_\theta[\,q_\alpha;\,t\,] \le -\cos(\theta\pi) \right\} \tag{2.14}$$

Remark 2.1. It is easy to verify that F_0 in (2.13) is indeed a cdf. Clearly $F_0(0) = 0$ and $F_0(1) = 1$. Moreover F_0 is increasing since $\mathscr{I}_\theta[\,q_\alpha;\,\cdot\,]$ is decreasing. Finally, being $\mathscr{I}_\theta[\,q_\alpha;\,\cdot\,]$ right continuous and since

$$F_0(t_*^-) = \frac{1}{2\theta} = F_0(t_*^+)$$

when the set in (2.14) is non-empty, F_0 is right continuous, too. Notice that $t_* = \infty$ for $\theta < \frac{1}{2}$.

Remark 2.2. For every $\theta < 1$, we have $F_0(0^+) = 0$ and $F_0(1^-) = 1$, that is the parameter measure cannot have positive mass on 0 or 1. The reason is that Dirichlet mean densities corresponding to such parameter measures are ruled out by the integrability hypothesis (2.12). Consider, for instance, $\alpha(\,\cdot\,) = \theta_0\delta_{\{0\}}(\,\cdot\,) + \theta_1\delta_{\{1\}}(\,\cdot\,)$. In this case $M(\tilde{\mathscr{D}}_\alpha) \overset{d}{=} \tilde{\mathscr{D}}_\alpha(\{1\})$, hence $M(\tilde{\mathscr{D}}_\alpha) \sim \mathrm{beta}(\theta_1, \theta_0)$, and its density does not fulfill (2.12) for $t \in \{0,1\}$, since $\theta + 1 - \theta_i > 1$ for $i = 0,\,1$. Notice however that for any α with total mass $\theta < 1$ and point masses in $\{0\}$ and $\{1\}$, such point masses can be treated separately in the previous straightforward way. Therefore, since $\theta < 1$, the hypothesis (2.12) is not restrictive. Moreover we do not impose on the mean density any local regularity properties, which are needed for (2.10) to be true in case of $\theta = 1$.

Proof of Theorem 2.1. We rewrite the left hand side of the CR identity

$$\exp\left\{-\theta\int_0^1 \log(z+x)\, P_0(\mathrm{d}x)\right\} = \int_0^1 \frac{q_\alpha(x)}{(z+x)^\theta}\,\mathrm{d}x \qquad z \in \mathbb{C}\setminus[-1,0] \tag{2.15}$$

leveraging on the fact that

$$\int_0^x \frac{1}{t+z}\,\mathrm{d}t = \log(z+x) - \log(z) \qquad \text{for} \quad \mathrm{Im}\, z \neq 0 \tag{2.16}$$

and obtaining

$$\begin{aligned}\exp\left\{-\theta\int_0^1 \log(z+x)\, P_0(\mathrm{d}x)\right\} &= \\ &= \exp\left\{-\theta\log(z) - \theta\int_0^1\int_t^1 P_0(\mathrm{d}x)\frac{\mathrm{d}t}{t+z}\right\} = \\ &= \frac{1}{(1+z)^\theta}\exp\left\{\theta\int_0^1 \frac{F_0(t)}{t+z}\,\mathrm{d}t\right\}\end{aligned} \tag{2.17}$$

where F_0 is the cdf of P_0. Hence (2.15) becomes

$$\exp\{\theta\mathscr{S}_1[F_0;z]\} = (1+z)^\theta\int_0^1 \frac{q_\alpha(x)}{(z+x)^\theta}\,\mathrm{d}x$$

with $\mathscr{S}_1$ according to the definition in (2.2). Applying the principal value of the complex logarithm to both sides

$$\mathscr{S}_1[F_0;z] = \frac{1}{\theta}\log\left\{(1+z)^\theta\mathscr{S}_\theta[q_\alpha;z]\right\} + \frac{2k(z)\pi\mathrm{i}}{\theta}$$

where

$$k(z) := -\theta\,\mathrm{Im}\,\mathscr{S}_1[F_0;z] \setminus \pi$$

for $\setminus$ denoting the integer division. Then, by the Cauchy–Stieltjes transform inversion formula in Widder (2015), we have for $t \in [0,1]$

$$F_0(t) \quad = \quad \lim_{\varepsilon\downarrow 0}\left\{-\frac{1}{\theta\pi}\,\mathrm{Im}\left(\log\left\{(1-t+\mathrm{i}\varepsilon)^\theta\mathscr{S}_\theta[q_\alpha;-t+\mathrm{i}\varepsilon]\right\}\right) + \frac{2k(-t+\mathrm{i}\varepsilon)}{\theta}\right\} =$$

$$= \lim_{\varepsilon\downarrow 0}\left\{ -\frac{1}{\theta\pi}\mathrm{Arg}\Big((1-t+\mathrm{i}\varepsilon)^{\theta}\mathscr{S}_{\theta}[q_\alpha;-t+\mathrm{i}\varepsilon]\Big) + \frac{2k(-t+\mathrm{i}\varepsilon)}{\theta}\right\}$$

where $\mathrm{Arg}(w)$ denotes the principal argument of $w \in \mathbb{C}$. If we write

$$\Im_{\varepsilon}[\,q_\alpha;t\,] := \mathrm{Im}\left((1-t+\mathrm{i}\varepsilon)^{\theta}\mathscr{S}_{\theta}[q_\alpha;-t+\mathrm{i}\varepsilon]\right) \tag{2.18}$$

and

$$\Re_{\varepsilon}[\,q_\alpha;t\,] := \mathrm{Re}\left((1-t+\mathrm{i}\varepsilon)^{\theta}\mathscr{S}_{\theta}[q_\alpha;-t+\mathrm{i}\varepsilon]\right) \tag{2.19}$$

then

$$\mathrm{Arg}\Big((1-t+\mathrm{i}\varepsilon)^{\theta}\mathscr{S}_{\theta}[q_\alpha;-t+\mathrm{i}\varepsilon]\Big) = \\ = \arctan\left(\frac{\Im_{\varepsilon}[\,q_\alpha;t\,]}{\Re_{\varepsilon}[\,q_\alpha;t\,]}\right) + \pi\mathbb{1}_{\{\Re_{\varepsilon}[\,q_\alpha;t\,]<0\}}\mathrm{sign}\Big(\Im_{\varepsilon}[\,q_\alpha;t\,]\Big) \tag{2.20}$$

Since

$$\lim_{\varepsilon\downarrow 0}\mathrm{Im}\left((1-t+\mathrm{i}\varepsilon)^{\theta}\right) = 0 \qquad \text{and} \qquad \lim_{\varepsilon\downarrow 0}\mathrm{Re}\left((1-t+\mathrm{i}\varepsilon)^{\theta}\right) = (1-t)^{\theta}$$

we shall neglect a summand and obtain

$$\lim_{\varepsilon\downarrow 0}\Im_{\varepsilon}[\,q_\alpha;t\,] = \\ = -(1-t)^{\theta}\lim_{\varepsilon\downarrow 0}\int_0^1 \frac{\sin\big(\theta\arctan\left(\varepsilon/(x-t)\right)+\theta\pi\mathbb{1}_{(0,t)}(x)\big)}{\left((x-t)^2+\varepsilon^2\right)^{\theta/2}}q_\alpha(x)\,\mathrm{d}x = \\ = -(1-t)^{\theta}\sin(\theta\pi)\int_0^t \frac{q_\alpha(x)}{|x-t|^{\theta}}\,\mathrm{d}x \tag{2.21}$$

and

$$\lim_{\varepsilon\downarrow 0}\Re_{\varepsilon}[\,q_\alpha;t\,] = \\ = (1-t)^{\theta}\lim_{\varepsilon\downarrow 0}\int_0^1 \frac{\cos\big(\theta\arctan\left(\varepsilon/(x-t)\right)+\theta\pi\mathbb{1}_{(0,t)}(x)\big)}{\left((x-t)^2+\varepsilon^2\right)^{\theta/2}}q_\alpha(x)\,\mathrm{d}x = \\ = (1-t)^{\theta}\left\{\cos(\theta\pi)\int_0^t \frac{q_\alpha(x)}{|x-t|^{\theta}}\,\mathrm{d}x + \int_t^1 \frac{q_\alpha(x)}{|x-t|^{\theta}}\,\mathrm{d}x\right\} \tag{2.22}$$

applying Lebesgue's dominated convergence theorem, which holds because of (2.12). To deal with the indicator and sign functions in (2.20), it suffices to check their discontinuity points. We have

$$-(1-t)^{\theta}\sin(\theta\pi)\int_0^t \frac{q_\alpha(x)}{|x-t|^{\theta}}\,\mathrm{d}x \leq 0$$

with equal holding for $t \in \{0,1\}$, while

$$(1-t)^{\theta}\left\{\cos(\theta\pi)\int_0^t \frac{q_\alpha(x)}{|x-t|^{\theta}}\,\mathrm{d}x + \int_t^1 \frac{q_\alpha(x)}{|x-t|^{\theta}}\,\mathrm{d}x\right\} \lesseqgtr 0 \iff \mathscr{I}_\theta[q_\alpha;t] \gtreqless -\cos(\theta\pi)$$

for $t \neq 0$, with equal holding at most in one point, by monotonicity. Finally we just need to prove that $k(-t+\mathrm{i}\varepsilon) \to 0$ for $\varepsilon \downarrow 0$. Since $\varepsilon > 0$, by Cauchy–Stieltjes transfrom properties, reported for instance in Karp and Prilepkina (2012), $\operatorname{Im}\mathscr{S}_1[F_0;-t+\mathrm{i}\varepsilon] \leq 0$, hence it suffices to notice that

$$\begin{aligned}-\operatorname{Im}\mathscr{S}_1[F_0;-t+\mathrm{i}\varepsilon] \leq \int_0^1 \frac{\varepsilon}{(s-t)^2-\varepsilon^2}\,\mathrm{d}s = \\ = \arctan\left(\frac{1-t}{\varepsilon}\right) - \arctan\left(-\frac{t}{\varepsilon}\right) \xrightarrow{\varepsilon} \pi \end{aligned} \tag{2.23}$$

□

As an application of the previous result we consider a few interesting cases, namely where one wants to determine the base measure of a DP such that the corresponding mean functional $M(\tilde{\mathscr{D}}_\alpha)$ has a uniform and a triangular distribution on $[0,1]$.

Example 2.3 (Uniform case)**.** Let $q_\alpha(x) = \mathbb{1}_{[0,1]}(x)$. Since

$$\mathscr{I}_\theta[\,\mathbb{1}_{[0,1]};\,t\,] = \left(\frac{1-t}{t}\right)^{1-\theta} \tag{2.24}$$

the cdf of P_0 is

$$F_0(t) = \frac{1}{\theta\pi}\arctan\left(\frac{\sin(\theta\pi)}{\cos(\theta\pi)+\left(\frac{1-t}{t}\right)^{1-\theta}}\right) + \frac{1}{\theta}\mathbb{1}_{(t_*,1)}(t)\mathbb{1}_{(\frac{1}{2},1)}(\theta) \tag{2.25}$$

for $t \in (0,1)$, where for $\theta \in \left(\frac{1}{2},1\right)$

$$t_* := \frac{1}{1+(-\cos(\theta\pi))^{\theta}} \tag{2.26}$$

In particular, for $\theta = \frac{1}{2}$

$$F_0(t) = \frac{2}{\pi}\arctan\sqrt{\frac{t}{1-t}} \qquad t \in (0,1) \tag{2.27}$$

Example 2.4 (Triangular case)**.** Let $q_\alpha(x) = 4x\mathbb{1}_{[0,\frac{1}{2})}(x) + 4(1-x)\mathbb{1}_{[\frac{1}{2},1]}(x)$. We have

$$\mathscr{I}_\theta[\,q_\alpha;\,t\,] = \left\{\frac{(1-t)^{2-\theta} - 2\left(\frac{1}{2}-t\right)^{2-\theta}}{t^{2-\theta}}\right\}\mathbb{1}_{(0,\frac{1}{2}]}(t) + \\ + \left\{\frac{(1-t)^{2-\theta}}{t^{2-\theta} - 2\left(t-\frac{1}{2}\right)^{2-\theta}}\right\}\mathbb{1}_{(\frac{1}{2},1]}(t) \quad (2.28)$$

Since $\mathscr{I}_\theta[\,q_\alpha;\,t\,]$ is decreasing and $\mathscr{I}_\theta\left[\,q_\alpha;\,\frac{1}{2}\,\right] = 1$, we can write the cdf of P_0, for $t \in (0,1)$, as follows.

$$F_0(t) = \left\{\frac{1}{\theta\pi}\arctan\left(\frac{t^{2-\theta}\sin(\theta\pi)}{t^{2-\theta}\cos(\theta\pi) + (1-t)^{2-\theta} - 2\left(\frac{1}{2}-t\right)^{2-\theta}}\right)\right\}\mathbb{1}_{(0,\frac{1}{2}]}(t) + \\ + \left\{\frac{1}{\theta\pi}\arctan\left(\frac{\left(t^{2-\theta} - 2(t-\frac{1}{2})^{2-\theta}\right)\sin(\theta\pi)}{\left(t^{2-\theta} - 2(t-\frac{1}{2})^{2-\theta}\right)\cos(\theta\pi) + (1-t)^{2-\theta}}\right)\right\}\mathbb{1}_{(\frac{1}{2},1)}(t) + \\ + \frac{1}{\theta}\mathbb{1}_{(t_*,1)}(t)\mathbb{1}_{(\frac{1}{2},1)}(\theta) \quad (2.29)$$

where t_* is such that

$$\frac{(1-t_*)^{2-\theta}}{t_*^{2-\theta} - 2\left(t_*-\frac{1}{2}\right)^{2-\theta}} = -\cos(\theta\pi) \quad (2.30)$$

Now, we deal with some particular cases, not covered by Theorem 2.1, but where still it is possible, because of the constructive nature of the proof, to retrieve consistent results. For $\theta > 1$, the integrability condition (2.12) rules out every probability density. Notice however that such hypothesis is just needed to perform a limit/integral switch in (2.21) and (2.22), hence whenever one selects a particular mean density such that the generalized Cauchy–Stieltjes transforms in (2.18) and (2.19) have explicit expressions, one can obtain analogous results even in this case. Yet the set $\mathbb{M}_\theta$ varies with θ and we have no guarantees that any absolutely continuous distribution can be a DP(θ) mean distribution for any θ; therefore with this procedure, one may obtain a function which is not a cdf. The fact that this is actually happening, as we will present in an instance, is in line with the $\theta = 1$ case: in Romik (2004) a homeomorphism is built between diagrams and probability distributions on $[0,1]$, but convex diagrams form a proper subset of all diagrams, hence some transition measures are bound to correspond with non-convex diagrams, which do not represent base probability measures. Such eventualities are treated in the following Proposition and Example.

Proposition 2.2. *Let $\theta \in (1,2)$ and $M(\tilde{\mathscr{D}}_\alpha) \sim q_\alpha$ with $q_\alpha(x) = \mathbb{1}_{[0,1]}(x)$. Then the cdf of the base measure P_0 is*

$$F_0(t) = \\ = \left\{ \frac{1}{\theta\pi} \arctan\left(\frac{\sin(\theta\pi)}{\cos(\theta\pi) + \left(\frac{t}{1-t}\right)^{\theta-1}} \right) + \frac{1}{\theta}\mathbb{1}_{(t_*,1)}(t) \right\} \mathbb{1}_{(0,1)}(t) + \mathbb{1}_{(1,\infty)}(t) \quad (2.31)$$

where

$$t_* = \frac{(-\cos(\theta\pi))^\theta}{1+(-\cos(\theta\pi))^\theta}\mathbb{1}_{(1,\frac{3}{2})}(\theta) \quad (2.32)$$

Remark 2.5. Notice that, unlike in $\theta \in (0,1)$ case, F_0 in (2.31) defines a distribution with $\frac{\theta-1}{\theta}$ masses in 0 and 1, while the rest of the mass is (symmetrically) diffuse in $(0,1)$.

Proof of Proposition 2.2. Since

$$\mathscr{S}_\theta[\mathbb{1}_{[0,1]}, -t+\mathrm{i}\varepsilon] = \frac{1}{\theta-1}\left\{ \frac{1}{(-t+\mathrm{i}\varepsilon)^{\theta-1}} - \frac{1}{(1-t+\mathrm{i}\varepsilon)^{\theta-1}} \right\} \quad (2.33)$$

we have

$$\mathrm{Im}\left(\mathscr{S}_\theta[\mathbb{1}_{[0,1]}, -t+\mathrm{i}\varepsilon]\right) = \\ = \frac{\left(t^2+\varepsilon^2\right)^{\frac{1-\theta}{2}}}{\theta-1} \sin\left((\theta-1)\arctan\left(\frac{\varepsilon}{t}\right) - (\theta-1)\pi\right) + \\ + \frac{\left((1-t)^2+\varepsilon^2\right)^{\frac{1-\theta}{2}}}{\theta-1} \sin\left((\theta-1)\arctan\left(\frac{\varepsilon}{1-t}\right)\right) \quad (2.34)$$

and

$$\mathrm{Re}\left(\mathscr{S}_\theta[\mathbb{1}_{[0,1]}, -t+\mathrm{i}\varepsilon]\right) = \\ = \frac{\left(t^2+\varepsilon^2\right)^{\frac{1-\theta}{2}}}{\theta-1} \cos\left((\theta-1)\arctan\left(\frac{\varepsilon}{t}\right) + (\theta-1)\pi\right) - \\ - \frac{\left((1-t)^2+\varepsilon^2\right)^{\frac{1-\theta}{2}}}{\theta-1} \cos\left((\theta-1)\arctan\left(\frac{\varepsilon}{1-t}\right)\right) \quad (2.35)$$

Therefore, using the notation established in (2.21) and (2.22), we obtain

$$\lim_{\varepsilon\downarrow 0} \Im_\varepsilon[\mathbb{1}_{[0,1]}; -t+\mathrm{i}\varepsilon] = \frac{(1-t)^\theta}{\theta-1} t^{\theta-1}\sin(\theta\pi) \quad (2.36)$$

and

$$\lim_{\varepsilon \downarrow 0} \Re_\varepsilon[\mathbb{1}_{[0,1]}; -t + \mathrm{i}\varepsilon] = \frac{(1-t)^\theta}{1-\theta} t^{\theta-1} \left\{ \cos(\theta\pi) + \left(\frac{t}{1-t}\right)^{\theta-1} \right\} \tag{2.37}$$

Hence, proceeding as in the proof of Theorem 2.1, we can conclude. □

Example 2.6. Let $\theta \in (2,3)$ and $M(\tilde{\mathscr{D}}_\alpha) \sim q_\alpha$ with $q_\alpha(x) = \mathbb{1}_{[0,1]}(x)$.Then it is possible to reason as in Proposition 2.2 and one obtains again

$$\frac{1}{\theta\pi} \arctan\left(\frac{\sin(\theta\pi)}{\cos(\theta\pi) + \left(\frac{t}{1-t}\right)^{\theta-1}} \right) \tag{2.38}$$

disregarding additive constants. However, for $\theta \in (2,3)$ this is a decreasing function, hence it cannot be the cdf of a probability measure. This implies that the uniform distribution cannot be the mean distribution of a DP with concentration parameter $\theta \in (2,3)$.

2.3.2 Base measure of a normalized stable process

The definition of the NSP as special case of NRMI has been recalled in Section 1.2. As for the random Dirichlet mean case, a preliminary step in our process of determining the base measure P_0 yielding a specific probability distribution for $M(\tilde{P}_\sigma)$ consists in establishing whether a correspondence similar to the one highlighted between $\mathbb{F}_\theta$ and $\mathbb{M}_\theta$ for the Dirichlet case holds true in this setting as well. In this respect, one can prove the following.

Theorem 2.3. *The probability distribution of the mean $M(\tilde{P}_\sigma)$ of a NSP $\tilde{P}_\sigma$ is determined by σ and $\mathbb{E}[\tilde{P}_\sigma] = P_0$.*

Proof. The result follows after evaluating the moments of any order of $M(\tilde{P}_\sigma)$ and noting that they do not depend on c in (1.21). To this end, set

$$Z(n,k) := \{\boldsymbol{m} = (m_1, \ldots, m_n) \in \mathbb{Z}_+^n : \sum_i i m_i = n, \sum_i m_i = k\} \tag{2.39}$$

where $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$ and $r_j = \int_0^1 x^j \, P_0(\mathrm{d}x)$. One can, now, rely on Theorem 3.3 in Lijoi and Prünster (2009) and obtain

$$\mathbb{E}\left[M^n(\tilde{P}_\sigma)\right] = \\ = \frac{1}{\Gamma(n)} \sum_{k=1}^n \sum_{\boldsymbol{m} \in Z(n,k)} \frac{n!}{\prod_{j=1}^n (j!)^{m_j} \, m_j!} \prod_{j=1}^n (r_j \, (1-\sigma)_{j-1})^{m_j} \times$$

$$\times\ \sigma^k\, c^k \int_0^\infty u^{k\sigma-1}\, \mathrm{e}^{-cu^\sigma}\, \mathrm{d}u =$$
$$= \sum_{k=1}^{n} \frac{\sigma^{k-1}\,\Gamma(k)}{\Gamma(n)} \sum_{\boldsymbol{m}\in Z(n,k)} \frac{n!}{\prod_{j=1}^{n}(j!)^{m_j}\, m_j!} \prod_{j=1}^{n} (r_j\,(1-\sigma)_{j-1})^{m_j} \tag{2.40}$$

which depends on the base measure $\alpha = c\,P_0$ only through the moments r_j of P_0. ☐

This result implies that, unlike in the DP case, any α in (1.9) such that $\alpha = c\,P_0$ leads to the same probability distribution for $M(\tilde{P}_\sigma)$, regardless of the value of the total mass c. For this reason we shall henceforth set $c = 1$ and focus on the determination of P_0.

Before displaying the main result which links the distribution of the mean to its base measure, it is important to point out that the key idea of the proof makes use of an analogue of the CR identity, for the mean of a NSP. If q_σ is the density function of the random mean $M(\tilde{P}_\sigma)$, where $\tilde{P}_\sigma$ is obtained by normalizing a σ-stable CRM with Lévy intensity as in (1.9) with $\alpha = P_0$, then, as shown in Tsilevich (1999), one has

$$\exp\left\{\int \log(z+x)^\sigma\, q_\sigma(x)\,\mathrm{d}x\right\} = \int (z+x)^\sigma\, P_0(\mathrm{d}x). \tag{2.41}$$

Here, we shall consider again q_σ and P_0 supported on $[0,1]$.

Theorem 2.4. *Let the density q_σ of $M(\tilde{P}_\sigma)$ be piecewise Hölder continuous and such that*

$$\int_0^1 |\log|x-t||\ q_\sigma(x)\,\mathrm{d}x < \infty \tag{2.42}$$

Lebesgue-almost everywhere. Then the base measure P_0 has cdf given by

$$\begin{aligned} F_0(y) = \frac{1}{\pi}\int_0^y\ & (y-t)^{-\sigma}\, \mathrm{e}^{\sigma\int_0^1 \log|x-t|\, q_\sigma(x)\,\mathrm{d}x} \\ & \left\{\pi\, q_\sigma(t)\, \cos(\sigma\pi\, Q_\sigma(t)) + \sin(\sigma\pi Q_\sigma(t))\,\mathrm{PV}\int_0^1 \frac{q_\sigma(x)}{t-x}\,\mathrm{d}x\right\}\,\mathrm{d}t \end{aligned} \tag{2.43}$$

for any $y \in (0,1)$, where Q_σ is the cdf of q_σ.

Proof. First note that $(z+x)^\sigma = z^\sigma + \sigma\int_0^x (z+s)^{\sigma-1}\,\mathrm{d}s$ for any x in $[0,1]$ and $\mathrm{Im}(z) \neq 0$. This implies that

$$\int_0^1 (z+x)^\sigma\, P_0(\mathrm{d}x) =$$

$$= z^{\sigma} + \sigma \int_0^1 (z+s)^{\sigma-1} \int_s^1 P_0(\mathrm{d}x)\,\mathrm{d}s = z^{\sigma} + \sigma \int_0^1 \frac{1-F_0(s)}{(z+s)^{1-\sigma}}\,\mathrm{d}s \quad (2.44)$$

from which

$$\int_0^1 \frac{F_0(s)}{(z+s)^{1-\sigma}}\,\mathrm{d}s = \frac{(z+1)^{\sigma}}{\sigma} - \frac{1}{\sigma}\int_0^1 (z+x)^{\sigma}\,P_0(\mathrm{d}x) \quad (2.45)$$

By virtue of the identity in (2.41), one can rewrite the right hand side of the equation above and obtains

$$\mathscr{S}_{1-\sigma}[z;F_0] = \frac{(z+1)^{\sigma}}{\sigma} - \frac{1}{\sigma}\exp\left\{\sigma\int_0^1 \log(z+x)\,q_{\sigma}(x)\,\mathrm{d}x\right\}. \quad (2.46)$$

At this point, one just needs to apply an inversion formula for the generalized Cauchy–Stieltjes transform $\mathscr{S}_{1-\sigma}$ and obtain F_0. This is the strategy we will undertake. We can apply the following alternative version of the inversion formula in (2.3), displayed in Schwarz (2005):

$$F_0(y) = \int_0^y (y-t)^{-\sigma}\Delta'(t)\,\mathrm{d}t. \quad (2.47)$$

where

$$\Delta(t) := \frac{1}{2\pi\mathrm{i}}\lim_{\varepsilon\downarrow 0}\left\{\mathscr{S}_{1-\sigma}[-t-\mathrm{i}\varepsilon;F_0] - \mathscr{S}_{1-\sigma}[-t+\mathrm{i}\varepsilon;F_0]\right\}. \quad (2.48)$$

Such alternative version holds whenever the involved integral does exist, and this will be clear *a posteriori*. Since the generalized Cauchy–Stieltjes transform is a holomorphic function on $\mathbb{C}\setminus\mathbb{R}^-$, we have

$$\Delta(t) = \frac{1}{\pi}\lim_{\varepsilon\downarrow 0}\mathrm{Im}\left(\mathscr{S}_{1-\sigma}[\,-t-\mathrm{i}\varepsilon;\,F_0\,]\right) = \frac{-1}{\pi\sigma}\lim_{\varepsilon\downarrow 0}\ \mathrm{Im}\left(\mathscr{L}_{\sigma}(-t-\mathrm{i}\varepsilon)\right) \quad (2.49)$$

because $\lim_{\varepsilon\downarrow 0}\mathrm{Im}(1-t-\mathrm{i}\varepsilon)^{\sigma} = 0$, where

$$\mathscr{L}_{\sigma}(z) := \exp\left\{\sigma\int_0^1 \log(x+z)\,q_{\sigma}(x)\,\mathrm{d}x\right\} \qquad z\in\mathbb{C}. \quad (2.50)$$

Now, since

$$\log(x-t-\mathrm{i}\varepsilon) = \frac{1}{2}\log((x-t)^2+\varepsilon^2) + \mathrm{i}\left\{\arctan\left(\frac{-\varepsilon}{x-t}\right) - \mathbb{1}_{\{(0,t)\}}(x)\pi\right\} \quad (2.51)$$

we have

$$\mathrm{Im}\left(\mathscr{L}_{\sigma}(-t-\mathrm{i}\varepsilon)\right) = -\exp\left\{\frac{\sigma}{2}\int_0^1 \log((x-t)^2+\varepsilon^2)q_{\sigma}(x)\,\mathrm{d}x\right\}\times$$

$$\times \sin\left(\sigma \int_0^1 \arctan\left(\frac{\varepsilon}{x-t}\right) q_\sigma(x)\, \mathrm{d}x + \sigma\pi Q_\sigma(t)\right) \tag{2.52}$$

Hence, by monotone and Lebesgue's dominated convergence theorems (the latter of which applies because of (2.42)) we obtain

$$\Delta(t) = \frac{1}{\sigma\pi}\, \mathrm{e}^{\sigma \int_0^1 \log|x-t| q_\sigma(x)\, \mathrm{d}x} \, \sin\left(\sigma\pi Q_\sigma(t)\right). \tag{2.53}$$

Finally, as can be found in Estrada and Kanwal (2012), we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\int_0^1 \log|x-t| q_\sigma(x)\, \mathrm{d}x\right) = \mathrm{PV}\int_0^1 \frac{q_\sigma(x)}{t-x}\, \mathrm{d}x \tag{2.54}$$

whenever the Cauchy principal value integral in the right hand side exists. It is easy to show that if q_α is Hölder continuous in the singularity point t, then the principal value in (2.54) exists and it is finite. See *e.g.* Estrada and Kanwal (2012). For arguments which weaken this condition, involving even and odd part of the density function, see Martin and Rizzo (1996). Hence, since q_σ is piecewice Hölder continuous, (2.54) holds for Lebesgue-almost every $t \in [0, 1]$. Therefore, differentiating (2.53) and substituting in (2.47), we get the expression in (2.43). □

Also here, we consider the cases where one wants to determine the base measure of a NSP such that the corresponding mean functional $M(\tilde{P}_\sigma)$ has a uniform and a triangular distribution on $[0, 1]$.

Example 2.7 (Uniform case)**.** Let $q_\sigma(x) = \mathbb{1}_{[0,1]}(x)$. Since

$$\mathrm{PV}\int_0^1 \frac{\mathrm{d}x}{t-x} = \lim_{\varepsilon\downarrow 0}\left(\int_0^{t-\varepsilon} + \int_{t+\varepsilon}^1\right)\frac{\mathrm{d}x}{t-x} =$$
$$= \lim_{\varepsilon\downarrow 0}\{\log t - \log\varepsilon - \log(1-t) + \log\varepsilon\} = \log\frac{t}{1-t} \tag{2.55}$$

by Theorem 2.4 the cdf of P_0 is

$$F_0(y) = \mathbb{1}_{[1,\infty)}(y) + \frac{1}{\mathrm{e}^\sigma \pi}\int_0^y \left(\frac{1-t}{y-t}\right)^\sigma \left(\frac{t}{1-t}\right)^{\sigma t} \times$$
$$\times \left\{\pi \cos(\sigma\pi t) - \sin(\sigma\pi t)\log\frac{t}{1-t}\right\} \mathrm{d}t\, \mathbb{1}_{[0,1)}(y) \tag{2.56}$$

Example 2.8 (Triangular case)**.** Let $q_\sigma(x) = 4x\mathbb{1}_{[0,\frac{1}{2})}(x) + 4(1-x)\mathbb{1}_{[\frac{1}{2},1]}(x)$. If $t \in \left[0, \frac{1}{2}\right)$ then

$$\begin{aligned}\mathrm{PV}\int_0^1 \frac{q_\sigma(x)}{t-x}\,\mathrm{d}x &= 4\lim_{\varepsilon\downarrow 0}\left\{\int_0^{t-\varepsilon}+\int_{t+\varepsilon}^{\frac{1}{2}}\right\}\frac{x}{t-x}\,\mathrm{d}x + 4\int_{\frac{1}{2}}^1 \frac{1-x}{t-x}\,\mathrm{d}x = \\ &= 4\lim_{\varepsilon\downarrow 0}\left\{-t+\varepsilon-t\log\varepsilon+t\log t-\frac{1}{2}\log\left(\frac{1}{2}-t\right)+t\varepsilon+t\log\varepsilon\right\}+ \\ &\quad + 2-4(1-t)\log(1-t)+4(1-t)\log\left(\frac{1}{2}-t\right)= \\ &= 2t\log t-2(1-t)\log(1-t)+2(1-2t)\log\left|\frac{1}{2}-t\right| \qquad (2.57)\end{aligned}$$

A similar expression holds true when $t\in\left[\frac{1}{2},1\right]$. Moreover, for any $t\in(0,1)$ one has

$$\begin{aligned}\int_0^1 &\log|t-x|\,q_\sigma(x)\,\mathrm{d}x = \\ &= 2t^2\log t+2(1-t)^2\log(1-t)-4\left(t-\frac{1}{2}\right)^2\log\left|t-\frac{1}{2}\right| \qquad (2.58)\end{aligned}$$

Hence by Theorem 2.4

$$\begin{aligned}F_0(y) &= \frac{1}{\pi}\int_0^y (y-t)^{-\sigma}\frac{(1-t)^{2\sigma(1-t)^2}\,t^{2\sigma t^2}}{\left|\frac{1}{2}-t\right|^{4\sigma(t-\frac{1}{2})^2}}\times \\ &\times\left\{\pi\,q_\sigma(t)\,\cos(\sigma\pi Q_\sigma(t))-\sin(\sigma\pi Q_\sigma(t))\,\log\frac{t^{2t}\,\left|t-\frac{1}{2}\right|^{2(1-2t)}}{(1-t)^{2(1-t)}}\right\}\mathrm{d}t \qquad (2.59)\end{aligned}$$

for $y\in(0,1)$, where

$$Q_\sigma(t)=\mathbb{1}_{[1,\infty)}(t)+2t^2\,\mathbb{1}_{[0,\frac{1}{2})}(t)+\left\{-2t^2+4t-1\right\}\mathbb{1}_{[\frac{1}{2},1)}(t) \qquad (2.60)$$

2.3.3 Base measure of a Pitman–Yor process

The general result in Theorem 2.4, besides being relevant for studying the mean functional of a NSP, forms also the basis for the derivation of an analogous result for the PYP mean functional. Indeed, Theorem 2.1 in James et al. (2008), establishes an important representation of PYP means in terms of DP and NSP means. Specifically, let $\tilde{P}_\sigma$ be a NSP with base measure P_0, denote by q_σ the density of its mean $M(\tilde{P}_\sigma)$ and consider a DP whose base measure is defined by $\alpha(B)=\theta\int_B q_\sigma(x)\,\mathrm{d}x$ for any Borel set B. Then one has

$$\int x\,\tilde{P}_{\sigma,\theta}(\mathrm{d}x)\overset{d}{=}\int x\,\tilde{\mathscr{D}}_\alpha(\mathrm{d}x) \qquad (2.61)$$

where $\mathbb{E}[\tilde{P}_{\sigma,\theta}] = P_0$. In words, this representation states that a PYP$(\sigma,\, \theta)$ mean has the same distribution as a DP(θQ) mean, where the probability measure Q is given by a NSP mean.
In the following Proposition a real version of the original CR identity is stated. Such result is crucial in order to obtain sufficient conditions on the density of a PYP mean which allow to recover an expression of the cdf of the base measure, combining results on DP and NSP means via the distributional identity (2.61).

Proposition 2.5. *Let q_α be the density of the mean $M(\tilde{\mathscr{D}}_\alpha)$ where $\tilde{\mathscr{D}}_\alpha \sim \mathrm{DP}(1,\, P_0)$. If q_α is piecewise Hölder continuous, then*

$$\cos\left(\pi P_0\left([0,\, t\,)\right) + \frac{\pi}{2} P_0\left(\{t\}\right)\right) \mathrm{e}^{-\int_0^1 \log|x-t|\, P_0(\mathrm{d}x)} = \mathrm{PV} \int_0^1 \frac{q_\alpha(x)}{x-t}\, \mathrm{d}x \tag{2.62}$$

for Lebesgue-almost every $t \in [0,1]$.

Proof. Let us consider the CR identity for concentration parameter equal to 1, which holds in this case

$$\exp\left\{-\int_0^1 \log(z+x)\, P_0(\mathrm{d}x)\right\} = \int_0^1 \frac{q_\alpha(x)}{z+x}\, \mathrm{d}x \qquad z \in \mathbb{C} \setminus [-1,0] \tag{2.63}$$

Substituting $z = -t + \mathrm{i}\varepsilon$, with $t \in [0,1]$ and $\varepsilon > 0$, and taking the real part we have on the left hand side

$$\cos\left(-\int_0^1 \left\{\arctan\left(\frac{\varepsilon}{x-t}\right) + \pi \mathbb{1}_{[0,t)}(x) + \frac{\pi}{2} \mathbb{1}_{\{t\}}(x)\right\} P_0(\mathrm{d}x)\right) \times \\ \times\ \exp\left\{-\int_0^1 \log\sqrt{(x-t)^2+\varepsilon^2}\, P_0(\mathrm{d}x)\right\} \tag{2.64}$$

and on the right hand side

$$\int_0^1 \frac{x-t}{(x-t)^2+\varepsilon^2}\, q_\alpha(x)\, \mathrm{d}x. \tag{2.65}$$

Since q_α is piecewice Hölder continuous, we have

$$\lim_{\varepsilon \downarrow 0} \int_0^1 \frac{x-t}{(x-t)^2+\varepsilon^2}\, q_\alpha(x)\, \mathrm{d}x = \mathrm{PV} \int_0^1 \frac{q_\alpha(x)}{x-t}\, \mathrm{d}x \tag{2.66}$$

and the limit is finite for Lebesgue-almost every $t \in [0,1]$. Hence, taking the limit for $\varepsilon \downarrow 0$ also in the left hand side, by virtue of monotone convergence theorem, we obtain (2.62). Notice that, *a fortiori*,

$$\int_0^1 |\log|x-t||\, P_0(\mathrm{d}x) < \infty \tag{2.67}$$

for Lebesgue-almost every $t \in [0,1]$.

□

Now we can state the general result for PYP means.

Theorem 2.6. *Let the density $q_{\sigma,1}$ of the mean $M(\tilde{P}_{\sigma,1})$ of a PYP with parameters $(\sigma, 1)$ be piecewise C^1 with piecewise Hölder continuous derivative. Then the base measure P_0 of $\tilde{P}_{\sigma,1}$ has cdf given by*

$$F_0(y) = \frac{1}{\pi} \int_0^y (y-t)^{-\sigma} \, e^{\sigma \int_0^1 \log|x-t| \, q_\sigma(x) \, \mathrm{d}x} \times \\ \times \left\{ \pi \, q_\sigma(t) \, \cos(\sigma\pi \, Q_\sigma(t)) + \sin(\sigma\pi Q_\sigma(t)) \, \mathrm{PV} \int_0^1 \frac{q_\sigma(x)}{t-x} \, \mathrm{d}x \right\} \mathrm{d}t \quad (2.68)$$

with q_σ having cdf given by

$$Q_\sigma(t) = \frac{1}{\pi} \operatorname{arccot} \left(\frac{1}{\pi \, q_{\sigma,1}(t)} \, \mathrm{PV} \int_0^1 \frac{q_{\sigma,1}(x)}{x-t} \, \mathrm{d}x \right) \quad (2.69)$$

Proof. The result follows immediately by resorting to the distributional identity in (2.61), which allows to apply iteratively the representation in (2.10) and Theorem 2.4 leading to the desired result. We only need to check that the conditions on $q_{\sigma,1}$ are sufficient to apply the DP and the NSP results. Firstly, since $q_{\sigma,1}$ is piecewice C^1 with bounded derivative, by (2.10), Q_σ as defined in (2.69) is the cdf of the base measure of a DP whose mean has density $q_{\sigma,1}$.
Now, since $q_{\sigma,1}$ is piecewise Hölder continuous, we can apply Proposition 2.5 and obtain

$$\cos\left(\pi Q_\sigma(t)\right) \exp\left\{ - \int_0^1 \log|x-t| \, q_\sigma(x) \, \mathrm{d}x \right\} = \mathrm{PV} \int_0^1 \frac{q_{\sigma,1}(x)}{x-t} \, \mathrm{d}x \quad (2.70)$$

for Lebesgue-almost every $t \in [0,1]$. Therefore we immediately recover the integrability condition (2.42) on q_σ. Hence in order to apply Theorem 2.4, we only need the Hölder continuity of q_σ, which is used in the proof to establish the derivative in (2.54). But, as recalled in Martin and Rizzo (1996), the derivative of the singular integral

$$\mathrm{PV} \int_0^1 \frac{f(x)}{x-t} \, \mathrm{d}x \quad (2.71)$$

exists and it is equal to the hypersingular integral

$$\mathrm{H} \int_0^1 \frac{f(x)}{(x-t)^2} \, \mathrm{d}x \quad (2.72)$$

named *Hadamard finite part integral*, whenever the density f is Hölder continuous with Hölder continuous derivative. Therefore, since this is the case for $q_{\sigma,1}$, both sides of (2.70) are differentiable, (2.54) holds and we can apply Theorem 2.4. □

The following Corollary of Proposition 2.5 highlights the connection between a PYP mean density and the mean density of the NSP obtained by normalizing the σ-stable CRM underling the PYP, according to the construction showed in (1.22).

Corollary 2.7. *Let $q_{\sigma,1}$ be the density of the mean $M(\tilde{P}_{\sigma,1})$, for $\tilde{P}_{\sigma,1} \sim \mathrm{PYP}(\sigma, 1)$. If $q_{\sigma,1}$ is piecewise Hölder continuous, then*

$$\frac{\cos\left(\pi Q_\sigma(t)\right)}{1-t} \exp\left\{ \mathrm{PV}\int_0^1 \frac{Q_\sigma(x)}{x-t}\,\mathrm{d}x \right\} = \mathrm{PV}\int_0^1 \frac{q_{\sigma,1}(x)}{x-t}\,\mathrm{d}x \tag{2.73}$$

for Lebesgue-almost every $t \in (0,1)$, where Q_σ is the mean distribution function of the NSP $\tilde{P}_\sigma$ in (1.22).

Proof. It suffices to apply Proposition 2.5 in view of representation (2.61), as in the proof of Theorem 2.6, and consider that

$$\begin{aligned}\int_0^1 \log|x-t| q_\sigma(x)\,\mathrm{d}x = \\ = \int_0^1 \frac{Q_\sigma(t)-Q_\sigma(x)}{x-t}\,\mathrm{d}x + Q_\sigma(t)\log t + (1-Q_\sigma(t))\log(1-t)\end{aligned} \tag{2.74}$$

and

$$\begin{aligned}\mathrm{PV}\int_0^1 \frac{Q_\sigma(x)}{x-t}\,\mathrm{d}x = Q_\sigma(t)\,\mathrm{PV}\int_0^1 \frac{1}{x-t}\,\mathrm{d}x + \int_0^1 \frac{Q_\sigma(x)-Q_\sigma(t)}{x-y}\,\mathrm{d}x = \\ = Q_\sigma(t)\log\left(\frac{1-t}{t}\right) - \int_0^1 \frac{Q_\sigma(t)-Q_\sigma(x)}{x-y}\,\mathrm{d}x\end{aligned} \tag{2.75}$$

Notice that, by Proposition 2.5, the Hölder continuity of $q_{\sigma,1}$ entails that q_σ integrates logarithmic singularities, which in turns implies the existence (and finiteness) of the principal value in (2.75). □

We shall close this Section giving an instance of application of Theorem 2.6, by treating the uniform case. Namely, suppose we wish to determine the parameter measure that makes the distribution of $M(\tilde{P}_{\sigma,1})$ uniform on $[0,1]$. One, then, sets $q_{\sigma,1}(x) = \mathbb{1}_{(0,1)}(x)$ and from (2.69) and

$$Q_\sigma(x) = \mathbb{1}_{[1,\infty)}(x) + \frac{1}{\pi}\,\mathrm{arccot}\left(\frac{1}{\pi}\,\log\frac{1-x}{x}\right)\mathbb{1}_{(0,1)}(x) \tag{2.76}$$

We shall denote as q_σ the density function corresponding to Q_σ. Finally, set

$$\xi(t) = \frac{1}{t(1-t)}\int_0^1 (1-x-t)\,q_\sigma(x)\,\mathrm{d}x +$$

$$+\frac{2}{t(1-t)}\int_0^1 \frac{(\log|t-x|)\left(\log\frac{1-x}{x}\right)}{\pi^2+\log^2\frac{1-x}{x}}\,q_\sigma(x)\,\mathrm{d}x \quad (2.77)$$

for any $t \in (0,1)$. By virtue of Theorem 2.6 one can state the following

Proposition 2.8. *The distribution of the mean $M(\tilde{P}_{\sigma,1})$, where $\tilde{P}_{\sigma,1} \sim \mathrm{PYP}(\sigma,1)$ is uniform on $(0,1)$ if and only if its base measure P_0 has cumulative distribution function as follows*

$$F_0(y) = \frac{1}{\pi}\mathrm{e}^{\sigma\int_0^1 \log|y-x|\,q_\sigma(x)\,\mathrm{d}x}\left\{\pi\,q_\sigma(t)\,\cos(\sigma\pi\,Q_\sigma(t)) + \xi(y)\,\sin(\sigma\pi Q_\sigma(t))\right\} \quad (2.78)$$

for any $y \in (0,1)$, where Q_σ and ξ are as in (2.76) and (2.77), respectively.

Proof. The density function corresponding to (2.76) is

$$q_\sigma(x) = \frac{1}{x(1-x)}\,\frac{1}{\left\{\pi^2+\log^2\frac{1-x}{x}\right\}}\,\mathbb{1}_{(0,1)}(x) \quad (2.79)$$

As for the evaluation of the principal value integral appearing in (2.68), note that

$$\begin{aligned}
&\left(\int_0^{t-\varepsilon}+\int_{t+\varepsilon}^1\right)\frac{1}{x(t-x)}\,\frac{1}{\left\{\pi^2+\log^2\frac{1-x}{x}\right\}}\,\mathrm{d}x = \\
&\qquad = \frac{1}{t}\left(\int_0^{t-\varepsilon}+\int_{t+\varepsilon}^1\right)\frac{1}{x}\,\frac{1}{\left\{\pi^2+\log^2\frac{1-x}{x}\right\}}\,\mathrm{d}x+ \\
&\qquad + \frac{1}{t}\left(\int_0^{t-\varepsilon}+\int_{t+\varepsilon}^1\right)\frac{1}{t-x}\,\frac{1}{\left\{\pi^2+\log^2\frac{1-x}{x}\right\}}\,\mathrm{d}x =: I_{1,\varepsilon}+I_{2,\varepsilon} \quad (2.80)
\end{aligned}$$

To shorten notation below, set $\zeta_{1,\varepsilon} := (\log\varepsilon)/\{\pi^2+\log^2[(1-t+\varepsilon)/(t-\varepsilon)]\}$ and $\zeta_{2,\varepsilon} := (\log\varepsilon)/\{\pi^2+\log^2[(1-t-\varepsilon)/(t+\varepsilon)]\}$. A simple change of variable, now, leads to

$$\begin{aligned}
I_{2,\varepsilon} &= \frac{1}{t}\int_\varepsilon^t \frac{1}{y}\,\frac{1}{\left\{\pi^2+\log^2\frac{1-t+y}{t-y}\right\}}\,\mathrm{d}y - \frac{1}{t}\int_\varepsilon^{1-t}\frac{1}{y}\,\frac{1}{\left\{\pi^2+\log^2\frac{1-t-y}{t+y}\right\}}\,\mathrm{d}y \\
&= \frac{1}{t}\Bigg\{-\zeta_{1,\varepsilon}+2\int_\varepsilon^t \frac{(\log y)\left(\log\frac{1-t+y}{t-y}\right)}{(t-y)(1-t+y)\left\{\pi^2+\log^2\frac{1-t+y}{t-y}\right\}^2}\,\mathrm{d}y \\
&\qquad + \zeta_{2,\varepsilon}+2\int_\varepsilon^{1-t}\frac{(\log y)\left(\log\frac{1-t-y}{t+y}\right)}{(t+y)(1-t-y)\left\{\pi^2+\log^2\frac{1-t-y}{t+y}\right\}^2}\,\mathrm{d}y\Bigg\} \quad (2.81)
\end{aligned}$$

Note also that

$$\left(\int_0^{t-\varepsilon} + \int_{t+\varepsilon}^1\right) \frac{1}{(1-x)(t-x)} \frac{1}{\left\{\pi^2 + \log^2 \frac{1-x}{x}\right\}} \,\mathrm{d}x = \\ = -\frac{1}{1-t}\left(\int_0^{t-\varepsilon} + \int_{t+\varepsilon}^1\right) \frac{1}{1-x} \frac{1}{\left\{\pi^2 + \log^2 \frac{1-x}{x}\right\}} \,\mathrm{d}x + \\ + \frac{1}{1-t}\left(\int_0^{t-\varepsilon} + \int_{t+\varepsilon}^1\right) \frac{1}{t-x} \frac{1}{\left\{\pi^2 + \log^2 \frac{1-x}{x}\right\}} \,\mathrm{d}x =: J_{1,\varepsilon} + J_{2,\varepsilon} \quad (2.82)$$

Moreover, $J_{2,\varepsilon} = tI_{2,\varepsilon}/(1-t)$, for $i = 1, 2$. □

2.4 Application to mixture models

Despite the results in Theorems 2.1, 2.4 and 2.6 have been stated for a simple mean of the form $M(\tilde{P}) = \int x\, \tilde{P}(\mathrm{d}x)$, they can be easily extended to include the case where one is willing to study a linear functional of the type

$$\int f \,\mathrm{d}\tilde{P} \quad (2.83)$$

for any measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $\int |f| \,\mathrm{d}\tilde{P} < \infty$, almost surely. This follows from the fact that

$$\int f \,\mathrm{d}\tilde{P} \stackrel{d}{=} \int x\, \tilde{P}_f(\mathrm{d}x) \quad (2.84)$$

where $\tilde{P}_f = \tilde{P} \circ f^{-1}$. If $\tilde{P}$ is obtained, as in (1.19), by normalizing a CRM $\tilde{\mu}$ with Lévy intensity $\nu(\mathrm{d}s, \mathrm{d}x)$, this is equivalent to saying that $\tilde{P}_f$ is obtained by normalizing a CRM $\tilde{\mu}_f$ whose Lévy intensity ν_f is such that

$$\int_B \int_A \nu_f(\mathrm{d}s,\, \mathrm{d}x) = \int_{f^{-1}(B)} \int_A \nu(\mathrm{d}s,\, \mathrm{d}x) \quad (2.85)$$

for any $A \in \mathscr{B}(\mathbb{R}^+)$ and $B \in \mathscr{B}(\mathbb{R})$. Hence, in (2.43) one just needs to interpret F_0 as the cumulative distribution function of $P_0 \circ f^{-1}$. On the other hand, (2.61) can be rewritten as

$$\int f(x)\, \tilde{P}_{\sigma,\theta}(\mathrm{d}x) \stackrel{d}{=} \int x\, \tilde{\mathscr{D}}_{q_\sigma}(\mathrm{d}x) \quad (2.86)$$

where now q_σ is the density function of $\int f \,\mathrm{d}\tilde{P}_\sigma$ and $\mathbb{E}[\tilde{P}_{\sigma,\theta}] = \mathbb{E}[\tilde{P}_\sigma] = P_0$. And an obvious extension of Theorem 2.6 follows.
In view of this, the results immediately carry over to consider some special cases of great relevance in statistical practice. An example is represented by mixture

models, in which one is interested in assigning a prescribed distribution to the mean of the mixture model itself, such as in, *e.g.*, Kessler et al. (2015). In particular, letting $\mathbb{Y}$ be a Polish space equipped with the Borel σ-algebra $\mathscr{Y}$, one defines a random mixture density (absolutely continuous with respect to some σ-finite measure ν on $\mathbb{Y}$) as

$$\tilde{f}(y) = \int_{\mathbb{X}} k(y;x)\,\tilde{P}(\mathrm{d}x) \tag{2.87}$$

where $\{k(\,\cdot\,;x) : \; x \in \mathbb{X}\}$ is a collection of density functions on $\mathbb{Y}$ indexed by a parameter taking values in $\mathbb{X}$. When $\tilde{P}$ is a DP one obtains the popular Dirichlet process mixture introduced by Lo (1984). Mixtures based on NSPs or PYPs represent valid alternatives with appealing features especially in terms of clustering and robustness. See, e.g., Ishwaran and James (2001), Lijoi et al. (2007) and Barrios et al. (2013).
Moreover, if one is interested in a mean of the mixture (2.87), then the problem of studying a linear functional of the mixture can be easily reduced to studying a (different) linear functional of the underlying $\tilde{P}$ by noting that

$$\int_{\mathbb{Y}} g(y)\,\tilde{f}(y)\nu(\mathrm{d}y) = \int_{\mathbb{X}} h(x)\,\tilde{P}(\mathrm{d}x) \tag{2.88}$$

where $h(x) = \int_{\mathbb{Y}} g(y)k(y,x)\nu(\mathrm{d}y)$. This strategy was applied in Nieto-Barajas et al. (2004) and James et al. (2010) for deriving the distribution of means DP and NRMI mixtures. From (2.88), it follows immediately that Theorems 2.1, 2.4 and 2.6 hold also in the case of mixture models by suitably adapting the specification of f yielding an important tool for prior specification in mixture models. Furthermore, notice that the mixture procedure allows the extension of our results on discrete random probability measures to absolutely continuous ones.

Theorem 2.9. *Suppose $g : \mathbb{Y} \to \mathbb{R}^+$ is a measurable function such that the probability distribution of the mean (2.88), where $\tilde{f}$ is defined as in (2.87) with $\tilde{P} = \tilde{\mathscr{D}}_\alpha$ and $\alpha = P_0$, has density q pointwise C^1 with bounded derivative and* $\mathrm{supp}(q) = [0,1]$. *Then*

$$P_0 \circ h^{-1}([0,x]) = \frac{1}{\pi}\,\mathrm{arccot}\left(\frac{1}{\pi\, q_\alpha(x)}\,\mathrm{PV}\!\int_0^1 \frac{q_\alpha(t)}{t-x}\,\mathrm{d}t\right) \tag{2.89}$$

for any $x \in (0,1)$, where $h(x) = \int_{\mathbb{Y}} g(y)\,k(y,x)\,\nu(\mathrm{d}x)$.
If instead $\alpha = \theta P_0$ for some $\theta \in (0,1)$ and the density q also satisfies condition (2.12), then

$$P_0 \circ h^{-1}([0,x]) = \frac{1}{\theta\pi}\arctan\left(\frac{\sin(\theta\pi)}{\cos(\theta\pi) + \mathscr{I}_\theta[\,q;\,x\,]}\right) + \frac{1}{\theta}\mathbb{1}_{(x_*,\infty)}(x) \tag{2.90}$$

for any $x \in (0,1)$, *where* $\mathscr{I}_\theta$ *is defined in* (2.11) *and*

$$x_* = \inf \left\{ x \in [0,1] \,\middle|\, \mathscr{I}_\theta[\, q_\alpha;\, x\,] \le -\cos(\theta\pi) \right\}$$

On the other hand, if $\tilde{P} = \tilde{P}_\sigma$ *in* (2.87) *with* $P_0 = \mathbb{E}\left[\tilde{P}_\sigma\right]$ *and the density* q *also satisfies condition* (2.42) *Lebesgue-almost everywhere, then*

$$\begin{aligned} P_0 \circ h^{-1}([0,x]) = \frac{1}{\pi} \int_0^x & (x-t)^{-\sigma}\, \mathrm{e}^{\sigma \int_0^1 \log|s-t|\, q(s)\, \mathrm{d}s} \times \\ & \times \left\{ \pi\, q(t)\, \cos(\sigma\pi\, Q(t)) + \sin(\sigma\pi Q(t))\, \mathrm{PV} \int_0^1 \frac{q(s)}{t-s}\, \mathrm{d}s \right\} \mathrm{d}t \end{aligned} \quad (2.91)$$

for any $x \in (0,1)$, *where* Q *is the distribution function of the mean* (2.88).

The expressions in (2.89), (2.90) and (2.91) can be used to determine the parameter measure yielding a specified probability distribution for a mean of a mixture model governed either by a DP or by a NSP. Here we describe an example involving the σ-stable case.

Example 2.9. Suppose $\tilde{f}$ is defined as in (2.87) with $\tilde{P} = \tilde{P}_\sigma$. Moreover, set $k(y,x) = x\, \mathrm{e}^{-xy} \mathbb{1}_{(0,\infty)}(y)$ and $f(y) = y$. In this setting we, then, aim at determining the parameter measure P_0 that induces a specified distribution for the mean of a σ-stable mixture of exponential densities. Hence $g(x) = x^{-1}$ and if we set $q_\sigma(x) = \mathbb{1}_{[0,1]}(x)$ it can be easily seen that

$$\begin{aligned} P_0 \circ g^{-1}((0,x]) = \mathbb{1}_{[1,\infty)}(x) + \frac{\mathrm{e}^\sigma}{\pi} \int_0^x & (x-t)^{-\sigma}\, (1-t)^{\sigma(1-t)}\, t^{\sigma t} \times \\ & \times \left\{ \pi \cos(\sigma\pi t) - \sin(\sigma\pi t)\, \log \frac{1-t}{t} \right\} \mathrm{d}t\, \mathbb{1}_{(0,1)}(x) \end{aligned} \quad (2.92)$$

From this one finds out that $\mathrm{supp}(P_0) = [1,\infty)$ and

$$\begin{aligned} P_0((0,x]) = \mathbb{1}_{[1,\infty)}(x) \Bigg\{ 1 - \frac{\mathrm{e}^\sigma x^\sigma}{\pi} \int_0^{1/x} & (1-xt)^{-\sigma}\, (1-t)^{\sigma(1-t)}\, t^{\sigma t} \times \\ & \times \left[\pi \cos(\sigma\pi t) - \sin(\sigma\pi t)\, \log \frac{1-t}{t} \right] \mathrm{d}t \Bigg\} \end{aligned} \quad (2.93)$$

In a similar fashion one can proceed if $f(y) = \mathbb{1}_{[T,\infty)}(y)$ for some $T > 0$, with the same exponential kernel as above. In this case $g(x) = \exp\{-xT\}$ and the distribution of the mean of the mixture is the distribution of an average survival probability $\int \exp\{-xT\}\, \tilde{P}_\sigma(\mathrm{d}x)$ at T. If one again wishes to specify P_0 in such a way that the mean of the mixture has uniform distribution, then

$$P_0((0,x]) = \mathbb{1}_{[0,\infty)}(x) \left\{ 1 - \frac{\mathrm{e}^\sigma}{\pi} \int_0^{\mathrm{e}^{-xT}} \left(\mathrm{e}^{-xT} - t\right)^{-\sigma} (1-t)^{\sigma(1-t)}\, t^{\sigma t} \right.$$
$$\left. \times \left[\pi \cos(\sigma \pi t) - \sin(\sigma \pi t)\, \log \frac{1-t}{t} \right] \mathrm{d}t \right\} \quad (2.94)$$

2.5 Conclusions

The determination of the parameter measure of a random probability measure yielding to a fixed random mean distribution is a challenging problem linked to many fields in probability, statistics and mathematics in general. In Bayesian nonparmetric modeling, solving this problem yields important prior elicitation, allowing the direct enforcement of prior information on an interpretable finite dimensional feature of the random probability measure. In this Chapter we provided explicit expressions for the distribution functions of the parameter measure in Dirichlet process with concentration parameter less than 1, normalized stable random measure and Pitman–Yor process cases, for a fairly broad class of fixed mean distributions. In doing so, we devised procedures and techniques which can be applied even to cases not covered by our general results, giving insights on the structure of sets of random means of discrete nonparametric priors. As an extra motivation, results apply also to mixture models with, *e.g.*, fixed population mean or average survival probability, opening the way to a wide range of applications.

Chapter 3

Dirichlet random means and continual Young diagrams

Abstract

In this work we present the connection between hook walks on continual Young diagrams and the Dirichlet process. In particular, this link is in the fact that the base measure of a Dirichlet process with concentration parameter equal to 1 and its mean distribution are each other uniquely determined by the same integral identity, that is the Cifarelli–Regazzini identity. Such equation also defines a bijection between a continual Young diagram and the so-called transition measure induced by it via a hook walk. We present the involved combinatorial objects, which are studied in many fields of mathematics, from probability to representation theory, we make explicit the connection with the Dirichlet process and give some implications of it. Finally, we discuss strategies for extending such a connection to general concentration parameters, and give a characterization in terms of operatorial inequlities, of the space of differentiable random Dirichlet means.

3.1 Introduction

As mentioned in Chapter 2, one of the interesting features of the area of research exploring properties of linear functionals of random probability measures is the pervasive character of the connections it provides with several and seemingly unrelated fields of mathematics. Our investigation for retrieving the parameter measure of certain classes of discrete random probability measures, still following the line of research opened by Cifarelli and Regazzini (1990), has been for sure inspired by one of these connection. Namely, in Kerov (1993), one can find the description of a Markov process named hook walk, defined on a continual Young diagram. The distribution of arrival point of such time-infinite random walk is called transition measure and it is linked in a one-to-one correspondence with the diagram itself. In Kerov (1993) the topic is framed in terms of Hausdorff and Markov moment problems. It is noted in Romik (2004) that the integral identity which entails the previously mentioned bijection is indeed, once extended by analytical continuation, a special case of the Cifarelli-Regazzini identity (CR identity), presented in Chapter 2, and linking the base measure of a Dirichlet process to its random mean density. In this Chapter we explore the connection that this parallel creates between hook walks on continual Young diagrams and the Dirichlet process. In the next Section we shall give an account of continual Young diagrams and their approximation via rectangular diagram, as well as present the hook walk algorithm and an alternative and equivalent way to sample a trajectory of this process and hence to sample (approximately) from the transition measure. In Section 3.3 we describe in details the connection with the Dirichlet process with concentration parameter equal to 1, explicitly constructing the cumulative distribution function of the base measure from the profile of the diagram. We further show the equivalence of the integral identities. Finally, open research direction spurred by this inspiring connection are presented in the last Section: the generalization to any concentration parameter, and ultimately the characterization of the space of random Dirichlet means.

3.2 Hook walks on continual Young diagrams

Continual Young diagrams are a limiting case of Young tableux, combinatorial objects used to visualize the representations of the symmetric group.
In details, given an interval $[a, b] \subset \mathbb{R}$, a *continual Young diagram* on $[a, b]$ is a 1-Lipschitz function ω defined on $\mathbb{R}$ such that

$$z := a + \omega(a) = b - \omega(b)$$

and

$$\forall x \notin [a, b] \qquad \omega(x) = |x - z|$$

z is called the *center* of the diagram ω. That is, a diagram is a Lipschitz function on (a, b) which hinges to the function $|x - z|$ at the extremes of the interval. An example in Figure 3.1.

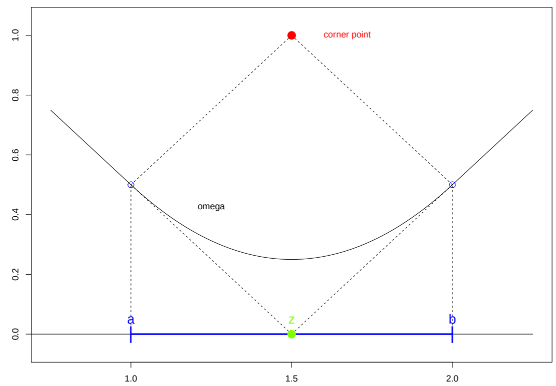
The *domain* of ω is

$$D_\omega := \{(x, y) \mid a \leq x \leq b, \, \omega(x) \leq y \leq \min(\omega(a) + x - a, \, \omega(b) + b - x)\}$$

that is the area delimited by the graph of the diagram, denoted $\mathcal{G}(\omega)$, and the two lines of slope 1 and -1 passing, respectively, through $(a, \omega(a))$ and $(b, \omega(b))$. The intersection of such lines is$(b - \omega(a), \, \omega(a) + \omega(b))$ and is called *corner point*. For a point $(x, y) \in D_\omega$ the *hook* of (x, y) is the set

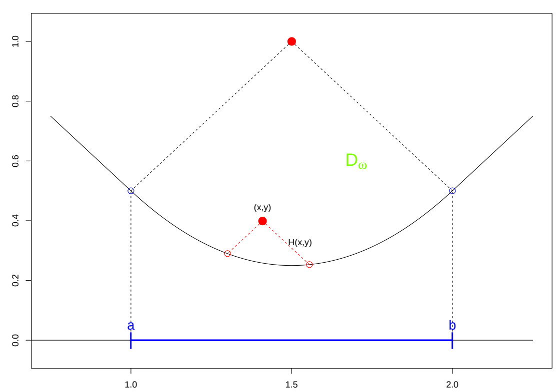
$$H(x, y) := \big\{(x', y') \in D_\omega \mid (x' \leq x \wedge y - y' = x - x') \vee (x' > x \wedge y - y' = x' - x)\big\}$$

Figure 3.1: A continual Young diagram



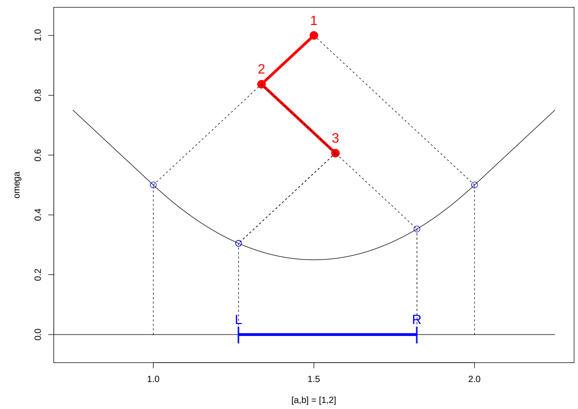
In words, the hook of a point (x, y) in the domain is obtained by the union of the segments of the lines of slope 1 and -1 passing through the point, included between the point itself and the intersections with the diagram. An example in Figure 3.2.

Figure 3.2: The hook of a point in the domain of a diagram



A *hook walk* on ω is a random sequence $\{(X_n, Y_n)\}_{n\geq 1}$ on D_ω, starting from the corner point, that at each step from a given point changes to a point uniformly, by arc length, chosen on the hook of the previous one. An example in Figure 3.3. That is, if $(X_n, Y_n) = (x, y)$, then $(X_{n+1}, Y_{n+1}) \in H(x, y)$, $\Xi(T) = (X_{n+1}, Y_{n+1})$ and $T \sim U[0, 1]$, where $\Xi : [0, 1] \longrightarrow H(x, y)$ is a parametric (one-to-one) rep-

Figure 3.3: First steps of a hook walk



resentation of $H(x,y)$. Notice that, in practice, to get the $n+1$ step it suffices to extract uniformly a value X^* from $[L,R]$, where L and R are the projections on the x-axis of the intersection points between $H(x,y)$ and ω, then to take $(X_{n+1},Y_{n+1}) = (X^*,Y^*) \in H(x,y)$, that is the only point in $H(x,y)$ having X^* as first component.

It is clear that the hook walk converges almost surely to a point in $\mathcal{G}(\omega)$, even if in a time that is almost surely infinite. Indeed, at every step n,

$$\mathbb{P}\left((X_{n+1},Y_{n+1}) \in \mathcal{G}(\omega) \mid (X_n,Y_n) = (x,y)\right) = 0.$$

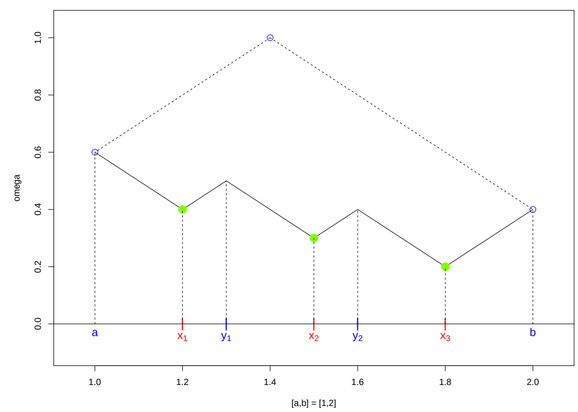
If $X_\infty := \lim_n X_n$ is the first component of the limiting point of the random walk, the distribution of X_∞ is called *transition measure* and we denote it with μ. The diagram ω and the transition measure μ are linked, and this connection is the object of the work in Romik (2004). Here, an expression of the density g of the transition measure μ is given for a class of smooth diagrams. Denoting with $\mathcal{D}[a,b]$ the set of diagrams on [a,b], define:

$$\mathcal{S}[a,b] := \left\{\omega \in \mathcal{D}[a,b]\,\text{piecewice}\ C^2,\ \omega''\,\text{bounded},\ c_1 < \omega' < c_2\,\text{with}\,c_1,c_2 \in (-1,1)\right\}$$

If $\omega \in \mathcal{S}[a,b]$ then

$$g(x) = \frac{1}{\pi}\cos\left(\pi\frac{\omega'(x)}{2}\right)(x-a)^{-\frac{1+\omega'(x)}{2}}(b-x)^{-\frac{1-\omega'(x)}{2}} \times \\ \times \exp\left\{\frac{1}{2}\int_a^b \frac{\omega'(u)-\omega'(x)}{u-x}\,\mathrm{du}\right\} \tag{3.1}$$

Figure 3.4: A rectangular diagram



This result is obtained by means of density arguments, that is a simpler class of diagrams is defined and proved to be dense in the general class. If ω is piecewise-linear with slope 1 or -1 in $[a, b]$, we call it *rectangular diagram*; an example in Figure 3.4. We denote with $\mathcal{D}_0[a, b]$ the set of rectangular diagrams. It is easy to see that $\omega \in \mathcal{D}_0[a, b]$ is uniquely determined by the *interlacing sequence* of its minimum and maximum points:

$$x_1 < y_1 < x_2 < \cdots < y_{n-1} < x_n.$$

The reason why this class results to be simpler to handle is that the transition measure relative to a rectangular diagram is a discrete measure with atoms in the minimum points:

$$\mu(\cdot) = \sum_{k=1}^{n} \delta_{x_k}(\cdot)\mu_k$$

and, plus, a characterization for the masses μ_k is available thanks to the work in Kerov (1993).

In this article, it is made use of a procedure of *random splitting*, directed by a rectangular diagram, of the interval $[a, b]$. This procedure is proved to have, in law, the same limiting point of a hook walk on the diagram. The random splitting works as follows. It is defined a *shrinkage process* as a Markov chain with state space given by the set of intervals $[x_i, x_j]$ with $1 \leq i \leq j \leq n$ and with μ as limiting distribution. From state $[x_i, x_j]$, one chooses uniformly $\gamma \in [x_i, x_j]$, then:

$$\begin{cases} \text{if } \gamma \in [x_k, y_k] \text{ for some } k \implies [x_i, x_k] \\ \text{if } \gamma \in [y_{k-1}, x_k] \text{ for some } k \implies [x_k, x_j] \end{cases}$$

For the limiting distribution μ (that is clearly discrete) of such chain it is proved that:

$$\sum_{k=1}^{n} \frac{\mu_k}{x - x_k} = \frac{\prod_{i=1}^{n-1}(x - y_i)}{\prod_{i=1}^{n}(x - x_i)} \tag{3.2}$$

$x \notin [a, b]$, where $\{\mu_k\}_1^n$ are the masses in the atoms $\{x_k\}_1^n$.

We notice here that the hook walk gives rise to another random splitting procedure of the interval $[a, b]$, perhaps more obvious and simpler to implement in case of simulations. We already noticed that the uniform sample from the hook of a point can be substituted by a uniform sample in $[L, R]$ and the selection on the hook of the only point having that x-coordinate. Hence, one just needs to find the intersection between the new branch of the hook and $\mathcal{G}(\omega)$ and this can be easily done in practice using well-known numerical methods. The projection on the x-axis of such intersection will determine the new extreme of the interval, from which the one can start over. This is the procedure adopted to obtain a simulator of a hook walk on a generic diagram, which produces in turn a sampler from the transition measure. In Figure 3.5 a recovery of the density of the transition measure for a diagram having a quadratic profile.

Coming back to transition measures, if we define the *charge* of the diagram ω as

$$\sigma(x) := \frac{\omega(x) - |x|}{2},$$

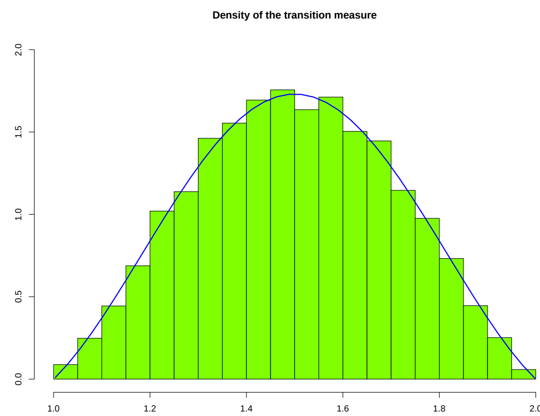
the equation (3.2) can be rewritten as

$$\int_{\mathbb{R}} \frac{\mu(\mathrm{d}t)}{x - t} = \frac{1}{x} \exp\left\{ \int_{\mathbb{R}} \frac{\mathrm{d}\sigma(t)}{t - x} \right\} \qquad \forall x \notin [a, b] \tag{3.3}$$

where the integral in the right hand side is intended in the Stieltjes sense. Since, as anticipated, $\mathcal{D}_0[a, b]$ is dense in $\mathcal{D}[a, b]$ with respect to the uniform norm, (3.3) can be proved to be true *for every* diagram. Now, this identity is closely related to Cifarelli-Regazzini identity, as it will be explained, and it used to prove the existence of a *homeomorphism* between $\mathcal{D}[a, b]$ with the topology induced by the uniform norm, and $\mathcal{M}[a, b]$, the set of probability measures on $[a, b]$ endowed with the weak topology. Hence, a diagram and its transition measure are not only connected, but in a one-to-one correspondence. This is the argument to prove both the expression of the density of the transition measure (3.1) and an inversion formula which allows to write the derivative of the diagram. If the transition measure is absolutely continuous, with density g piecewise C^1, having bounded derivative and it is bounded away from 0, then

$$\omega'(x) = -1 + \frac{2}{\pi} \operatorname{arccot}\left(\frac{1}{\pi} \log\left(\frac{b - x}{x - a} \right) + \frac{1}{\pi g(x)} \int_a^b \frac{g(u) - g(x)}{u - x} \mathrm{d}u \right) \tag{3.4}$$

Figure 3.5: In blue the analytic expression. The histogram is obtained by simulation



Both the formulas are proved by approximation arguments. In particular, in order to prove (3.1) a smooth diagram (that is in $\mathcal{S}[a, b]$) is approximated by a piecewise-linear, that is in turn approximated by an interpolating rectangular diagram, whose discrete transition measure density and its limit are retrieved by the use of (3.2). The inverse formula is obtained working on a step function density, whose corresponding measure is a mixture of discrete uniform measures, which have a related rectangular diagram; exploiting the homeomorphism, then, it suffices to calculate the uniform limit of such diagram.

3.3 Link with the Dirichlet process

The connection between continual Young diagrams and the Dirichlet process comes when one considers a convex diagram ω. In this case, a proper translation of the derivative of the charge of ω, is a càdlàg, non-decreasing function with values in $[0, 1]$, that is a cumulative distribution function of some probability measure on $[a, b]$.

Let us focus on the case $a = 0$, $b = 1$. If ω is a convex function, then it admits right and left derivatives everywhere and the non-differentiability points correspond to jump discontinuities of such derivatives. We denote with ω'_+ the right derivative of ω, which is a non-decreasing right continuous and bounded function. Now we have

$$\sigma_+'(x) = \begin{cases} \dfrac{\omega_+'(x)+1}{2} & x<0 \\[2ex] \dfrac{\omega_+'(x)-1}{2} & x \geq 0 \end{cases}$$

which means

$$\sigma_+'(x) = \left\{ \frac{\omega_+'(x)-1}{2} \right\} \mathbb{1}_{[0,\infty)}(x)$$

Hence it is clear that if we define

$$F_0(x) := \sigma_+'(x) + \mathbb{1}_{[0,\infty)}(x) \tag{3.5}$$

F_0 is the cumulative distribution function of a probability measure, that we denote

$$P_0((-\infty, x]) := F_0(x).$$

Notice that, since a diagram is constrained at the hinges in 0 and 1, if two diagrams differ by a constant c then $c = 0$. Indeed let $\omega_c = \omega + c$ with ω, ω_c diagrams; then, since $\omega_c(0) = 1 - \omega_c(1)$ and $\omega(0) = 1 - \omega(1)$, we have

$$\omega(0) + c = 1 - \omega(1) - c \quad \Longrightarrow \quad c = 0.$$

This implies that any F_0 obtained as in (3.5) comes from a unique diagram. Let us consider $x \notin [0,1]$, then

$$\begin{aligned}
\frac{1}{x}\exp\left\{\int_{\mathbb{R}} \frac{\mathrm{d}\sigma(t)}{t-x}\right\} &= \frac{1}{x}\exp\left\{\int_{\mathbb{R}} \frac{\sigma_+'(t)\mathrm{d}t}{t-x}\right\} = \\
&= \frac{1}{x}\exp\left\{\int_{-\infty}^{0} \frac{F_0(t)}{t-x}\,\mathrm{d}t - \int_0^\infty \frac{1-F_0(t)}{t-x}\,\mathrm{d}t\right\} = \\
&= \frac{1}{x}\exp\left\{\int_{-\infty}^{0}\int_{-\infty}^{t} P_0(\mathrm{d}s)\frac{\mathrm{d}t}{t-x} - \int_0^\infty\int_t^\infty P_0(\mathrm{d}s)\frac{\mathrm{d}t}{t-x}\right\} = \\
&= \frac{1}{x}\exp\left\{\int_{-\infty}^{0}\int_{s}^{0} \frac{\mathrm{d}t}{t-x}\,P_0(\mathrm{d}s) - \int_0^\infty\int_0^s \frac{\mathrm{d}t}{t-x}\,P_0(\mathrm{d}s)\right\} = \\
&= \frac{1}{x}\exp\left\{-\int_{\mathbb{R}}\int_0^s \frac{\mathrm{d}t}{t-x}\,P_0(\mathrm{d}s)\right\} = \\
&= \exp\left\{-\int_{\mathbb{R}} \log|x-s|\,P_0(\mathrm{d}s)\right\}
\end{aligned}$$

Hence identity (3.3) can be rewritten as

$$\exp\left\{-\int_{\mathbb{R}} \log|x-s|\,P_0(\mathrm{d}s)\right\} = \int_{\mathbb{R}} \frac{\mu(\mathrm{d}s)}{x-s} \tag{3.6}$$

for any $x \notin [0,1]$. Finally, (3.6) can be extended by analytic continuation, obtaining

$$\exp\left\{\int_{\mathbb{R}} \log(z+s)\, P_0(\mathrm{d}s)\right\} = \int_{\mathbb{R}} \frac{\mu(ds)}{z+s} \tag{3.7}$$

for any $z \in \mathbb{C} \setminus [-1,0]$. It is apparent that we recovered CR identity for concentration parameter $\theta = 1$, displayed in (2.63).

Hence a convex diagram ω on $[0,1]$ defines a probability measure P_0 on the interval that is related to the transition measure μ in the same way in which the parameter (probability) measure of a Dirichlet process $\mathscr{D}_\alpha$ on $[0,1]$ with $\alpha(\,\cdot\,) = P_0(\,\cdot\,)$ is related to the distribution of the random mean $M(\mathscr{D}_\alpha)$. Therefore, coming back to the simulation of a hook walk on a continual Young diagram, we have the immediate possibility of sampling from the random mean of a Dirichlet with concentration parameter 1. In particular, what we actually see in Figure 3.5 is a recovery of the density of a Dirichlet mean with base measure a uniform distribution on $[1,2]$.

This framework, then, is both more general and more particular. An integral identity is proved also for non-convex diagrams, where the connection between a diagram and a probability measure is not available anymore. On the other hand, the result is limited to probability base measures ($\theta = 1$) on a closed interval. However, in this environment, it seems to be more natural to solve the inverse problem: leveraging on the homeomorphism and on the density of sets of simpler diagrams (and atomic transition measures), the asymmetry in difficulty one experiences working analytically on inversion of Cauchy–Stieltjes transforms is flattened. Considering that hypotheses on the transition density imply the existence of the its Cauchy singular integral, we can rewrite (3.4) as

$$P_0((0,x]) = \frac{1}{\pi}\operatorname{arccot}\left(\frac{1}{\pi q(x)}\right) \mathrm{PV}\int_0^1 \frac{q(s)}{s-x}\,\mathrm{d}s \tag{3.8}$$

where q is the transition density, which the formula we reported in (2.10).

The possibility of going more smoothly back and forth between the two measures, if seen in this external and seemingly far mathematical environment, has been the inspiration for trying to find a connection between our results in Chapter 2 and the framework accounted in this Chapter. Open research directions in this matter are presented in the next Section.

3.4 Further work

The homeomorphism between the set of all diagrams on $[0,1]$ and all probability distributions on the same interval is fascinating. However, our argument to connect the Dirichlet process to such homeomorphism creates another link between diagrams and probability measures. Indeed we take a translation of the derivative of a convex diagram and we notice that it is a cumulative distribution function. In the Dirichlet world, this is the base measure, while the transition measure becomes the random mean distribution. We already noticed that the derivative identifies a diagram, since two diagrams cannot differ by a constant. Moreover it seems that our procedure can be performed the other way around to obtain a convex diagram from any cumulative distribution function. Since convex diagrams are a proper subset of all diagrams and the diagrams are in (homeomorphic) bijection with probability measures, this leads to a contradiction.

Once the previous point is solved, one would like to use the structure that the homeomorphism brings, for instance to evaluate how distances between base measures reflect on distances between mean distributions. This could even become a way of measuring consistency *a posteriori*. A possible obstacle in this matter is again the way we use diagrams to define base measures: the homeomorphism is built for the space of diagrams endowed with the topology of the uniform convergence, while we consider derivatives of diagrams, and as known, uniform convergence does not pass in general to derivatives. A possible solution is to rebuild the homeomorphism using the Lipschitz seminorm, which in the case of diagrams, is actually a norm.

Another natural question is the following. Is it possible to extend the diagrams-Dirichlet connection to general concentration parameters, and further on to more general processes? For what concerns the first question, it is easy to device a generalization of diagrams which correspond to general finite measures: it suffices to consider θ-Lipschitz functions. However, the hook walk, or equivalently the random splitting procedure, seems to induce on the interval the same transition measure induced by a 1-Lipschitz suitable deformation of the original θ-diagram. This is clear with rectangular diagrams, where a θ-rectangular diagram induces the same transition measure than the 1-rectangular diagram defined by the same interlacing sequence.

In general, the set of random means of discrete nonparametric priors is still fairly mysterious. We know from the homoemorphism, as we recalled in Chapter 2, that not every absolutely continuous distribution can be the mean distribution of a Dirichlet mean with concentration parameter θ equal to 1. This eventuality seems to be confirmed also for $\theta < 1$ by our results in Chapter 2 and in particular in Example 2.6. However, a proper characterization of the set is still missing. Resorting on the expression in (3.8), we can state that a differentiable density q can be a mean density of a Dirichlet(1) if and only if

$$q'(x)\,\mathrm{PV}\int_0^1 \frac{q(y)}{y-x}\,\mathrm{d}y \geq q(x)\,\mathrm{PV}\int_0^1 \frac{q'(y)}{y-x}\,\mathrm{d}y \tag{3.9}$$

or equivalently

$$q'(x)\,\mathrm{PV}\int_0^1 \frac{q(y)}{y-x}\,\mathrm{d}y \geq q(x)\,\mathrm{H}\int_0^1 \frac{q(y)}{(y-x)^2}\,\mathrm{d}y \tag{3.10}$$

Where the singular integral on the right hand side is the Hadamard finite part integral, introduced in 2.

Therefore the elements of the space of random Dirichlet means are the solutions an operatorial inequality where differential operators, singular integral operators and their compositions are involved.

Chapter 4

Partially exchangeable multilayer stochastic block models

Abstract

There is an increasing availability of complex network data encoding connectivity information among a set of nodes, often belonging to different layers. A challenging task is represented by inferring grouping structures among nodes based on common connectivity patterns, while considering the layer division. Although it could be reasonable in some cases to expect such connectivity blocks to coincide with layers, this assumption is in general too strong and fails to learn sub-blocks within each layer as well as across-layer clusters. To incorporate these mixed architectures while accounting for layer information in a principled manner, we propose a new generation of partially exchangeable multilayer stochastic block models relying on a hierarchical random partition prior for the node allocations driven by the urn scheme of a hierarchical normalized completely random measure or a hierarchical Pitman–Yor process. The partial exchangeability assumption among nodes according to layer partitions allows to infer both within- and across-layer blocks, while preserving probabilistic coherence, principled uncertainty quantification and formal inclusion of prior information from layer membership. The mathematical tractability of such priors further allows to analytically derive and compare predictive within- and across-layer co-clustering probabilities, thereby providing conditions on hyperparameters to enforce interpretable features on the grouping structures. Moreover, the predictive structure of the model naturally entails a way to infer both connections and allocation of new nodes incoming into the network. The applied potentials of this new class of Bayesian nonparametric models are illustrated in criminal network studies.

4.1 Introduction

Network theory is widely employed to describe and analyze complex systems in social, biological, physical and engineering sciences. Consequently, methodological investigations regarding the analysis of network data are becoming more and more spread in statistical literature. As the systems that need to be modeled become more structured, as the urge for models able to encode different information at several levels increases. Multilayer networks are a prominent example of such elaborated systems: on top of the usual network structure, given by nodes linked by (directed, undirected or weighted) edges, they include layerizations of the network itself, according to characteristics (also multivariate ones) of the nodes, of the edges or both, accounting for, *e.g.*, *intra*- and *inter*-layer connections, several directions of layerization, copies of the same node existing in different layers, and so on. For insightful reviews on multilayer networks, see Boccaletti et al. (2014) and Kivelä et al. (2014). Modelling such networks naturally represents a challenging statistical task: in order to take into account the entirety of such structures but still being able to perform reliable inference on features of the network as well as validated prediction about an unobserved part of it, we want to stick to a model-based and fully probabilistic approach.

One of the most significant tasks in the analysis of networks is the detection of grouping structures based on the connection activity of the nodes. Early approaches have often relied on community detection algorithms, as in Girvan and Newman (2002), Newman and Girvan (2004) and Newman (2006). These methods are focused on recognizing groups with dense connectivity inside and sparse communication outside, and hence tend to oversimplify the grouping structure. On the other hand, also spectral clustering techniques, which show better flexibility properties, have been employed. See *e.g.* Von Luxburg (2004). However, such algorithms lack of an integrated statistical procedure allowing for joint modeling of different aspects of the clustering, like the number of groups, as well as uncertainty quantification for inference and prediction. More on the side of a model-based approach lie stochastic block models (SBMs). See Holland et al. (1983) and Nowicki and Snijders (2001). In this class of models, the partition of the nodes is a latent feature of the network: for each couple of nodes, the probability of creating an edge is given conditionally on the allocations of the couple. In a Bayesian framework, in order to perform inference on such allocations, a prior is placed on the space of partitions and a posterior distribution is retrieved by updating the prior with the observed connections. This entails the inference of block structures which tend to cluster together nodes sharing among them the same patterns of communication with the members of the other blocks, hence allowing for the recognition of more complex and connectivity-driven grouping structures. The favorable ratio between simplicity of the structure and flexibility of the inferred clustering, together with the availability of easy and efficient inference techniques, have brought to further developments and generalizations of such models, encompassing also the use of more and more flexible priors on the random partition, from the classical Dirichlet-multinomial to nonparametric solutions as the Dirichlet process or mixture-of-finite-mixtures, as in See Schmidt and Morup (2013), Kemp et al. (2006)

and Geng et al. (2019). For a general framework under the unifying concept of Gibbs-type priors, a careful comparison and an insightful application to criminal networks, see Legramanti et al. (2022).

In this work we focus on the inference of grouping structures reflecting connectivity patterns among nodes of a special kind of multilayer networks called *node-colored* networks. They are multilayer graphs with one direction of layerization, based on a characteristic of the nodes (*e.g.* affiliation to a certain university in a researcher network), and such that each node has no copies in different layers. We propose a stochastic block model for the *supra*-graph, which clusters together nodes with similar connectivity behavior, also belonging to different layers, while unloading the information of the division in layers on the latent level of the random partition. This is achieved by choosing hierarchical nonparametric priors on the space of partitions which structurally enforce *homophily* among nodes in the same layer and *heterogeneity* between nodes in different layers. Thinking of each node as a statistical unit and of their division in layers as the division in sub-populations of individuals, we induce partial exchangeability on a latent characteristic of the nodes, this being the most natural assumption for data coming from different sub-populations and encompassing exactly the similarity-within/dissimilarity-across scheme just described. Hence we suppose a hierarchical random partition prior on the nodes' allocations which implies such condition. In particular, we employ random partition distributions driven by hierarchical discrete random probability measures, such as the hierarchical Dirichlet process (H-DP), introduced in Teh et al. (2006) and, its generalizations given by the hierarchical normalized random measures with independent increments (H-NRMI) and the hierarchical Pitman–Yor process (H-PYP), whose constructions and properties are thoroughly studied in Camerlenghi et al. (2019). The result is a fully-probabilistic model including the layer information, allowing a flexible *inter*- and *intra*-layer clustering, ensuring principled uncertainty quantification and a theoretically-validated inference. Analytical expressions of characteristic posterior co-clustering probabilities can be derived.

We devise a Gibbs sampler algorithm to perform inference on the posterior distribution of the node allocation, by leveraging both the collapsed structure typical of SBMs and the augmented urn scheme of a random partition induced by H-NRMI and H-PYP.

Moreover, the structure of the model allows the construction of a prediction algorithm both for the allocation and for the connections of a new node incoming into the network, on whichever of the existing layers. The posterior sample of the old nodes' latent allocations, inferred from their connections and layer memberships, can be indeed used in a Monte Carlo procedure, together with the new layer membership, to estimate a probability distribution on the joint space of possible allocations and connections of the new node. Therefore, we provide a tool that, once the inference on the network is performed, employs its output to produce principled prediction, naturally provided of uncertainty quantification, on a new node's allocation and connections, without the need of re-run the posterior sampler on the augmented network. To overcome the exponentially-growing cardinality of the product space, we device a sampler for the joint predictive distribution too. The predictive construction of the model, based on exchangeable partition probability functions (EPPFs), provides the theoretical guarantees for the coherence between the updated model and the full model. Namely, our prior is placed on the space of partitions of a growing number of objects and automatically entails Kolmogorov consistency of such sequence of random partitions. It is natural and probabilistically coherent then to rely on the posterior samples of old nodes' allocations to predict features of nodes joining the network.

The Chapter is structured as follows. In Section 4.2 we illustrate the aim of such new generation of models, while giving some preliminaries on multilayer networks and the partially exchangeable regime. Section 4.3 is devoted to the illustration of the model and the hierarchical structure of the prior, its specification via conditions on the hyperparameters to enforce desired clustering patterns in the model. Moreover we present explicit analytic expressions for posterior and predictive co-clustering probabilities. Section 4.4 presents the posterior inference techniques adopted, the posterior sampler and the prediction scheme for new nodes. Results of applications to simulation scenarios and real data are reported in Section 4.5.

4.2 Motivation and preliminaries

We pursue the construction of probabilistic models for clustering nodes in networks taking a multilayer structure in account, in order to be able to tackle in a principled way inferential problems arising in several applications. In particular, we focus on multilayer networks such that layers realize a division of the nodes according to some characteristics, expressed for instance by a categorical variable. We do so inspired by different applicative scenarios.

Example 4.1. Bill co-sponsorship networks are a motivating example. In such networks the nodes correspond to representatives in a parliament, each one afferent to a political party. An edge exists between two nodes whenever the candidates both voted a bill, or a certain number of bills. Note that, clearly, edges between nodes in different parties are possible. If one wants to infer grouping structures characterized by similar behaviors in co-sponsoring bills, then, on one hand it would be too restrictive to let these blocks of similar co-sponsoring choices to coincide with the parties, on the other hand it would be rather simplistic to completely disregard the information enclosed in the party memberships and treat data as if we were in the *planar* case. Moreover, in this specific application, direct comparisons between the party division and a connectivity-driven inferred clustering should be a trait of particular interest.

Example 4.2. Another important application field is the investigation of criminal networks. Here, nodes are registered criminals and connections may represent different forms of communication that can be assessed by law enforcement, such as phone calls, messages or co-participation in meetings. Moreover it is quite common, above all in organized crime, that criminals can be divided according to affiliation to subgroups. Nonetheless, such divisions might not reflect the operational connectivity patterns of the whole organization, disregarding collaborations, internal feuds or consequences of disruptive events (like murders or arrests). Also in this case then, it arises the need for models enforcing the subgroup division information, but still inferring grouping structures according to connectivity patterns. Moreover the ability of dealing with new nodes entering the network appears as a strongly desirable feature for models to be applied in this context, as it is natural to add new registered criminals while the investigation is still on-going. For this reason, we propose models whose construction is prediction-based, in order to naturally perform a theoretically-validated inference on the clustering of new nodes and also an as well principled prediction on their connections.

From both Example 4.1 and 4.2, it is clear that a SBM for the connections fits our inferential interests, since it enforces grouping based on connectivity patterns on the latent clustering it entails. Indeed, these models assume that, given a latent allocation and a matrix of group-wise connection probabilities, the existence of edges between nodes is the result of independent Bernoulli trials, whose success probabilities are, as said, group-specific. These probabilities are, in turn, given the allocations, independent and beta-distributed, while a prior is placed on the latent allocations. Intuitively, the posterior distribution of the latent clustering is driven towards an allocation reflecting similar connection behaviours, given by the shared connection probabilities: a group is such whenever its members connect to other groups' members in the same way.

On the other hand, concerning the introduction of the layer division information, the parallel between such grouping of the nodes and the division in subpopulations typical of a partially exchangeable sampling scheme, recalled in Section 1.3, is natural. From such parallel comes the idea of an allocation mechanism *a priori* driven by a scheme enforcing such distributional invariance. Nonetheless, as it will be clear in next Sections, to suppose partial exchangeability for labels attached to nodes is completely appropriate just in the special case of multilayer networks we consider in this work, that is the node-colored ones. Finally, since we want the inferred clusters to be across-layers with non-zero probability, we will employ hierarchical discrete random measures, that is the de Finetti measure of the random probability measures' vector inducing partial exchangeability in (1.30) will be defined in a hierarchical way and it will have support on discrete distributions.

In the rest of this Section we will present the background for the construction of the novel generation of models for clustering in networks we propose. In particular, we will frame Example 4.1 and 4.2 in the general environment of multilayer networks, contextualize the idea of enforcing partial exchangeability and recall the construction of random partitions driven by hierarchical discrete random measures, such as H-NRMIs and the H-PYP, which we introduced in Section 1.3. Analytical tractability of such hierarchical processes, expressed in terms of partially exchangeable partition probability functions (pEPPFs), will be key in retrieving explicit expressions for posterior and predictive clustering and co-clustering probability, employed in Sections 4.3 and 4.4 for prior elicitation, posterior inference and prediction.

4.2.1 Multilayer networks

In order to model complex connectivity architectures, notions of *multilayer network* has been introduced in the literature of several fields. The objective is to incorporate additional information in the usual graph representation of a network. For example one may want to divide the nodes according to the levels of one or more factors, or to include different kind of connections, or both. An easy and comprehensive way to do so is to consider the nodes as laying in different layers, even allowing the same node to exist in different layers. Following Kivelä et al. (2014), we give a fairly general definition of a multilayer network, which encompasses a broad variety of special cases. For the sake of easing the notation, nodes and elementary layers will be directly identified with natural numbers. We recall that the notation $[n] := \{1, \dots, n\}$ indicates the set of the first n natural numbers, for any $n \in \mathbb{N}$. Hence $V := [N]$ is a set of nodes, for $N \in \mathbb{N}$. Then we call $\delta \in \mathbb{N}$ the number of *aspects* of the network, that is the number of directions of layerization. Namely, for any $l \in [\delta]$, we define a set of *elementary layers* $L_l = [d_l]$ for $d_l \in \mathbb{N}$. A *layer* is an element of the product space $L_1 \times \cdots \times L_\delta$, *i.e.* a vector of coordinates identifying a location in the multilayer network. Then we define a set $V_M \subset V \times L_1 \times \cdots \times L_\delta$ of *existing nodes*. In symbols, the node i exists in layer $(j_1, \dots, j_\delta)$ whenever $(i, j_1, \dots, j_\delta) \in V_M$, for $j_l \in L_l$ and $l \in [\delta]$. Each existing node $(i, j_1, \dots, j_\delta)$ is a *copy* of node $i \in V$. Finally we consider a set $E_M \subset V_M \times V_M$ of edges, *i.e.* of couples $((i, j_1, \dots, j_\delta), (i', j'_1, \dots, j'_\delta))$ such that there is an arc between the existing nodes $(i, j_1, \dots, j_\delta)$ and $(i', j'_1, \dots, j'_\delta)$. Now we can state the following.

Definition 4.1. *A δ-multilayer network is a quadruplet $M := (V_M, E_M, V, L)$ where $L = L_1 \times \cdots \times L_\delta$ is the product space of layers, V is the set of nodes, V_M is the set of coordinates of existing nodes and E_M is the set of edges.*

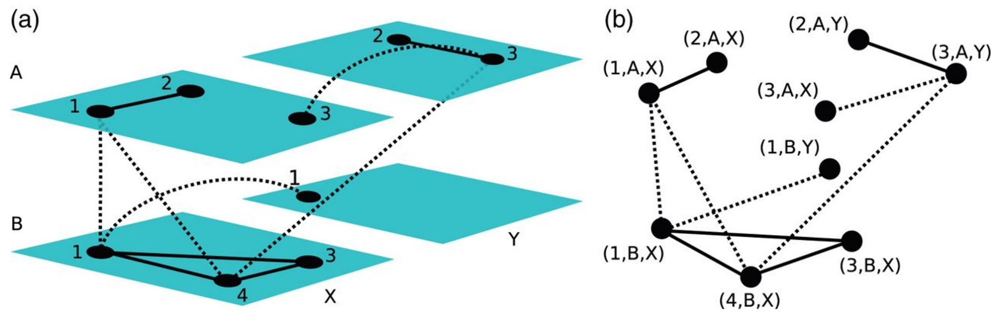


Figure 4.1: A multilayer network. (a) nodes $V = \{1, 2, 3, 4\}$, 2 aspects, elementary layers $\boldsymbol{L} = \{L_1 = \{A, B\}, L_2 = \{X, Y\}\}$; (b) the *supra*-graph of the existing nodes Kivelä et al. (2014).

Notice that we are allowing the existence of edges between nodes in the same layer, as well as across-layers, and even between copies of the same node existing in different layers. A multilayer network M can be flattened in the graph with labelled nodes given by the couple (V_M, E_M), which is called the *supra*-graph of M. See Figure 4.1.

We recall that given a graph with N nodes, its adjacency matrix is a $N \times N$ matrix such that entry (r, s) is 1 whenever there is an edge between node r and node s, and 0 otherwise. As a consequence, a δ-multilayer network can be represented by a *supra*-adjacency matrix $\boldsymbol{Y}$, that is the adjacency matrix of the *supra*-graph. Once chosen an order in the product space of the layers L, *e.g.* the lexicographic order, based on the natural order of each entry, then it is easy to describe the matrix $\boldsymbol{Y}$ as a block matrix whose diagonal blocks are the adjacency matrices of each layer, and the off-diagonal blocks are the matrices of across-layers connections.

As mentioned, a common configuration which can be represented in a multilayer network framework is the one of networks with connections classified through the levels of a multivariate categorical variable. Now, considering a δ-multilayer network, as in Definition 4.1, given by a block-diagonal *supra*-adjacency matrix, namely excluding connections across-layers, one obtains such a network as a special case. If for simplicity we consider the case in which each node exists in each layer, the network can be seen as a superposition of different sets of edges on the same collection of nodes, hence the layerization is simply given by considering a copy of the nodes for each set of edges. When $\delta = 1$, these are called *edge-colored* networks. Examples are given by transport networks allowing different means of transportation, or by social networks where connections are divided per social media. See Figure 4.2.

Now, it is clear how an edge-colored network is not fitting the hypothesis of

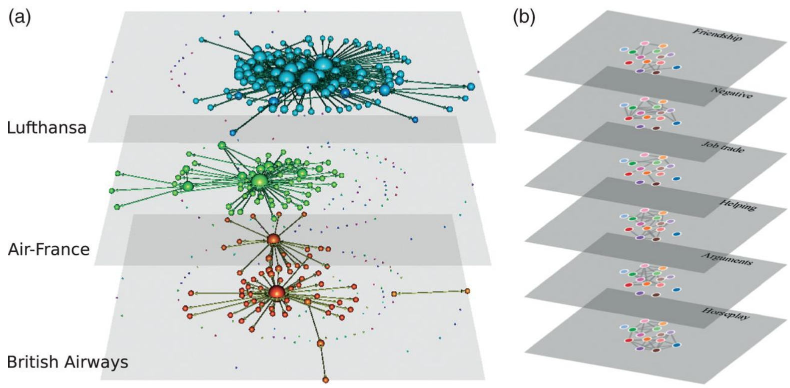


Figure 4.2: Edge-colored network: (a) air-transportation network; (b) bank wiring network Kivelä et al. (2014).

partial exchangeability for any array of random variables attached to its nodes, according to the layer division. In fact, in this case, each node is linked to its copies in other layers, usually by the fact that they represent the same statistical unit. But partial exchangeability is not tight enough as a distributional constraint to enforce such connection and one ends up treating several copies of the same statistical unit as different units.

Another broadly used network architecture is the one where nodes are attached with a multivariate categorical variable, reflecting some set of individual features for each node. To work out this case as a specialization of the δ-multilayer network, it suffices to impose the following disjointness condition.

$$\forall\, i \in V \quad \exists!\,(j_1, \ldots, j_\delta) \in L \quad \text{s.t.} \quad (i,\, j_1, \ldots, j_\delta) \in V_M \tag{4.1}$$

Namely, since in this case the existence of a node in a particular layer represents its individual features, a node cannot have copies in different layers. Here, the division in layers can be identified with a partition of V, where nodes in the same subset share the same realization of the multivariate categorical variable. If $\delta = 1$, that is such categorical variable in univariate, these are called *node-colored* networks. See Figure 4.3. Examples 4.1 and 4.2 fall in this special case: layers are, respectively, political parties and criminal sub-groups of affiliation. Notice that, since layers represent a division of the network which is solely based on nodes' characteristics, and each node represents a different individual across layers, indeed this special case entails a reasonable scenario for assuming a partially exchangeable labelling regime.

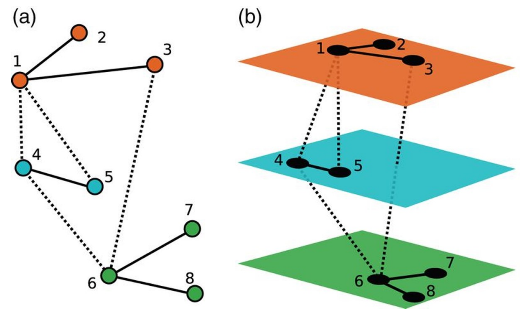


Figure 4.3: A multilayer network. (a) nodes $V = \{1, 2, 3, 4\}$, 2 aspects, elementary layers $\boldsymbol{L} = \{L_1 = \{A, B\}, L_2 = \{X, Y\}\}$; (b) the *supra*-graph of the existing nodes Kivelä et al. (2014).

4.2.2 Random partitions induced by H-NRMI and H-PYP

As mentioned before, we are willing to employ, for the network's connections, a SBM, which is defined conditionally on latent allocations, whose posterior distribution will be the object of our inference. Hence, we need to place a prior on the latent partition of the nodes and, since we consider a node-colored network, we want such prior to incorporate the layer information. Incidentally, as explained previously in this Section, the division in layers of the nodes, in this particular case, has a natural parallel with the division in blocks of statistical units proper of partial exchangeability, which we recalled in Section 1.3.

Hence the idea is to attach to each node of a node-colored network with d layers and N_j nodes in layer $j \in [d]$, an entry of a partially exchangeable random array $\boldsymbol{X}$ as in Definition 1.2 and to deduce from a realization of $\boldsymbol{X}$, the realization of a random partition for the nodes. Namely, we are modeling a latent characteristic of the nodes via $\boldsymbol{X}$, inducing a form of *distributional homophily* between nodes in the same layer via partial exchangeability, and deducing a partition from this characteristic. A natural way to deduce a random partition from $\boldsymbol{X}$ is to choose the *directing measure* Q of Theorem 1.2 to be putting all its mass on vectors of discrete probability measures and interpret then the ties among entries of a realization of $\boldsymbol{X}$ as a co-clustering relationship. In symbols, we define the random array $\boldsymbol{Z} = (Z_{ji})_{i=1,\dots,N_j}^{j=1,\dots,d}$ of allocations as

$$Z_{ji} = h \iff X_{ji} = X_h^* \tag{4.2}$$

for $h \in [H]$, where $\boldsymbol{X^*} = (X_1^*, \dots, X_H^*)$ is the vector of the unique values of $\boldsymbol{X}$ in order of occurrence and H is the number of clusters, which can be random. Indeed, we are inducing a distribution on the space of random partitions of $[N]$.

Now, for the motivation brought in Example 4.1 and 4.2, we want the support of our prior to include across-layers partitions. Since the labeling of the nodes of each layer j is driven by a distinct random probability measure $\tilde{P}_j$, as can be seen in (1.30), we need to consider a hierarchical structure in assigning the directing measure Q in order to allow ties both across-column and across-rows in the array $\boldsymbol{X}$. In particular, we employ H-NRMIs and the H-PYP, which, as stated in (1.35), fulfill this property.

As mentioned, we recalled the basics of the construction and definition of H-NRMIs and H-PYP in Section 1.3. In this Chapter, however, we will need a deeper recollection of their properties, also because we are going to explore and asses new features of such hierarchical priors, in order to specify properties of our model. First of all, we take advantage of the fact that we employ such nonparametric priors just for clustering: as stated earlier, in fact, for our purposes, the realization of the array $\boldsymbol{X}$ lay in latent level of the model which drives the prior on the allocations array $\boldsymbol{Z}$. This implies that we can actually work directly on the distribution induced on the space of partitions. From Camerlenghi et al. (2019) we know that the distribution of a random partition driven by a H-NRMI or a H-PYP is characterized by its partially exchageable partition probability function (pEPPF) Π, which is derived as a hierarchical interaction of exchangeable partition probability functions (EPPFs). See Pitman (1996) for a full account on these objects. Indeed, for any array of positive integers $(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_d)$ with $\boldsymbol{n}_j = (n_{j1}, \ldots, n_{jH})$ such that $\sum_{h=1}^{H} n_{jh} = N_j$ for some $H > 0$, representing an allocation of N objects, divided in d groups each of N_j objects, in H clusters, we have that

$$\Pi_H^{(N)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_d) = \sum_{\boldsymbol{\ell}} \sum_{\boldsymbol{q}} \Phi_{H,0}^{(|\boldsymbol{\ell}|)}(\ell_{\cdot 1}, \ldots, \ell_{\cdot H}) \prod_{j=1}^{d} \prod_{h=1}^{H} \frac{1}{\ell_{jh}!} \binom{n_{jh}}{q_{jh1}, \ldots, q_{jh\ell_{jh}}} \Phi_{\ell_{j\cdot}, j}^{(N_j)}(\boldsymbol{q}_{j1}, \ldots, \boldsymbol{q}_{jH}) \tag{4.3}$$

determines the probability of such allocation according to the law induced by a H-NRMI, where $\boldsymbol{\ell} = (\ell_{jh})_{j=1,\ldots,d}^{h=1\ldots H}$ is a matrix with $\ell_{jh} \in [n_{jh}]$, $\ell_{j\cdot} = \sum_{h=1}^{H} \ell_{jh}$, $\ell_{\cdot h} = \sum_{j=1}^{d} \ell_{jh}$, $|\boldsymbol{\ell}| = \sum_{j=1}^{d} \ell_{j\cdot}$, while $\boldsymbol{q}_{jh} = (q_{jh1}, \ldots, q_{jh\ell_{jh}})$ is a vector of integers such that $\sum_{t=1}^{\ell_{jh}} q_{jht} = n_{jh}$. The functions $\Phi_{\ell_{j\cdot}, j}^{(N_j)}$ and $\Phi_{H,0}^{(|\boldsymbol{\ell}|)}$ are the EPPFs driven by NRMIs with parameters c, ρ and c_0, ρ_0 respectively. Indeed they give probabilities of partitions of N_j elements in $\ell_{j\cdot}$ clusters and $|\ell|$ elements in H clusters, respectively. Their expressions can be found in James et al. (2009). Similarly, for a partition driven by a H-PYP with parameters $(\sigma, \sigma_0, \theta, \theta_0)$ we have

$$\Pi_H^{(N)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_d) = \sum_{\boldsymbol{\ell}} \frac{\prod_{k=1}^{H-1}(\theta_0 + k\sigma_0)}{(\theta_0 + 1)_{|\boldsymbol{\ell}|-1}} \prod_{h=1}^{H} (1 - \sigma_0)_{\ell_{\cdot h} - 1}$$

$$\prod_{j=1}^{d} \frac{\prod_{t=1}^{\ell_{j\cdot}-1}(\theta+t\sigma)}{(\theta+1)_{N_j-1}} \prod_{h=1}^{H} \frac{\mathscr{C}(n_{jh},\ell_{jh},\sigma)}{\sigma^{\ell_{jh}}} \tag{4.4}$$

where $(a)_n = \Gamma(a+n)/\Gamma(a)$ denotes the ascending factorial and $\mathscr{C}(n,k,\sigma)$ the generalized factorial coefficients.

To sum up, given an array $\boldsymbol{X}$ such that

$$\begin{aligned} X_{j1},\ldots,X_{jN_j} \mid (\tilde{P}_1,\ldots,\tilde{P}_d) &\overset{\text{iid}}{\sim} \tilde{P}_j \qquad \forall j \in [d] \\ (\tilde{P}_1,\ldots,\tilde{P}_d) &\sim \text{H-NRMI}(\rho,\, \rho_0,\, c,\, c_0,\, P_0) \end{aligned} \tag{4.5}$$

and being $\boldsymbol{Z}$ the random allocation array obtained as in (4.2), we write

$$\boldsymbol{Z} \sim \text{pEPPF}(\rho,\, \rho_0,\, c,\, c_0)$$

Similarly if

$$\begin{aligned} X_{j1},\ldots,X_{jN_j} \mid (\tilde{P}_1,\ldots,\tilde{P}_d) &\overset{\text{iid}}{\sim} \tilde{P}_j \qquad \forall j \in [d] \\ (\tilde{P}_1,\ldots,\tilde{P}_d) &\sim \text{H-PYP}(\sigma,\, \sigma_0,\, \theta,\, \theta_0,\, P_0) \end{aligned} \tag{4.6}$$

then we write

$$\boldsymbol{Z} \sim \text{pEPPF}(\sigma,\, \sigma_0,\, \theta,\, \theta_0)$$

Example 4.3. If we choose in particular a H-DP distribution with parameters $\theta,\, \theta_0 > 0$for the random probability measures vector in (4.5), we denote the induced distribution on the partition as

$$\boldsymbol{Z} \sim \text{pEPPF}(\theta,\, \theta_0)$$

Similarly, if choose H-NSP in (4.5), we write

$$\boldsymbol{Z} \sim \text{pEPPF}(\sigma,\, \sigma_0)$$

A convenient and intuitive framework to treat the partitions induced by these kind of hierarchical models is for sure given by the *restaurant franchise metaphor*, introduced in Teh et al. (2006) for the H-DP. Such useful metaphor holds also for H-NRMI and H-PYP. Paralleling directly with the node-colored network environment, we can think of each node in its layer as a customer in a restaurant: customers are seated at tables restaurant-wise, and each table is served with a dish taken from a common across-restaurant menu. We consider as clustered together all the customers eating the same dish. Nonetheless, the division in tables generates a nested and *inter*-restaurant further partition. In this metaphor each $\tilde{P}_j$ in (1.31) or (1.32) is directing the sitting of the customers at the tables in each restaurant, while $\tilde{P}_0$ directs the serving of the dishes to the tables from the same menu. The almost sure discreteness of $\tilde{P}_0$ results in the fact that the same dish can be eaten at different tables both within- and across-restaurants. In the metaphor, the matrix $\boldsymbol{\ell}$ and the arrays $\boldsymbol{q}$ and $\boldsymbol{n}$ employed in (4.3) and (4.4) find intuitive interpretations: ℓ_{jh} is the number of tables in restaurant j eating dish h, q_{jht} is the number of customers in restaurant j, eating dish h at table t. As a consequence $\ell_{j\cdot}$ is the number of tables in restaurant j, $\ell_{\cdot h}$ is the number of tables eating dish h and n_{jh} is the number of customers in restaurant j eating dish h. The metaphor is useful to exemplify the role of the EPPFs $\Phi^{(N_j)}_{\ell_{j\cdot},j}$ and $\Phi^{(|\boldsymbol{\ell}|)}_{H,0}$ in the sampling of the two nested partitions: the first ones give, for any $j \in [d]$, the distribution of the division in tables for any restaurant, and the second drives the dish labeling of the tables across-restaurants.

Leveraging the nested clustering, by making explicit the table labelling, we obtain the following sequential way of sampling an array $\boldsymbol{X}$ from a H-NRMI or a H-PYP, which is also well demonstrating the hierarchical structure of these non-parametric priors.

1. For each restaurant $j \in [d]$, sample the table labels

$$\begin{aligned} T_{j1}, \ldots, T_{jN_j} \mid \tilde{q}_j &\overset{\text{iid}}{\sim} \tilde{q}_j \\ \tilde{q}_j &\sim \text{NRMI}(\rho, c, G) \qquad \text{or} \\ \tilde{q}_j &\sim \text{PYP}(\sigma, \theta, G) \end{aligned} \tag{4.7}$$

   for some diffuse G.

2. For each restaurant $j \in [d]$, allocate the nodes in tables considering the partition induced by ties in $\boldsymbol{T}$, defining the sub-allocation array $\boldsymbol{W}$:

$$W_{ji} = t \iff T_{ji} = T^*_{jt} \tag{4.8}$$

   for $t \in [\ell_{j\cdot}]$, where $\boldsymbol{T}^*_j = (T^*_{j1}, \ldots, T^*_{j\ell_{j\cdot}})$ is the vector of the unique values of $\boldsymbol{T}_j$ in order of occurrence.

3. Sample the dish labels

$$(D_{jt})^{j=1,\ldots,d}_{t=1,\ldots,\ell_{j\cdot}} \mid \tilde{P}_0 \overset{\text{iid}}{\sim} \tilde{P}_0 \tag{4.9}$$

   for each table, with $\tilde{P}_0$ as in (1.31) or (1.32).

4. Define $X_{ji} := D_{jt}$ whenever $W_{ji} = t$.

Notice that, since we are just considering ties for clustering, the probability measure G in (4.7), as well as P_0 in (1.31) and (1.32), can be arbitrary, as long as they are diffuse. The sub-allocation array $\boldsymbol{W}$ and in general the nested partition it represents, namely the table labeling, is crucial for recovering the full conditionals we will need in order to perform posterior inference and prediction.

4.3 Hierarchical random partitions for multilayer networks

Let us consider a 1-multilayer network M, as in Definition 4.1, represented by an undirected *supra*-graph (V_M, E_M) with no self-loops allowed (that is its *supra*-adjacency matrix is symmetric with null diagonal) and satisfying the disjointness condition (4.1) with d layers, where layer j includes N_j nodes for any $j \in [d]$ and $N = \sum_{j=1}^{d} N_j$. For tackling in a principled manner the inferential task of clustering the nodes of such kind of networks, we propose a class of partially exchangeable stochastic block models (PEx-SBMs), a novel generation of probabilistic models taking into account both the connection patterns in the network and the layer division of the nodes. In particular we combine stochastic block models (SBMs) with hierarchical random partition priors, whose use we reviewed and motivated in Section 4.2. In this Section we determine the model and we study some of its theoretical characteristics, giving explicit expressions for predictive clustering and co-clustering probabilities proper of random partitions induced by H-NRMIs and H-PYPs, enlightening unexplored features of such important nonparametric priors. These results, applied to specific cases, bring prior elicitation tools for the determination of hyperparameters enforcing a desired behaviour in the predictive mechanism.

4.3.1 Model structure

Let $\boldsymbol{Y}$ be the $N \times N$ symmetric *supra*-adjacency matrix of M. If $\boldsymbol{Z}$ is the array of node allocations, $[H]$ the set of labels of occupied clusters in a realization of $\boldsymbol{Z}$ and $\boldsymbol{\Xi} = (\xi_{hk})$ is the $H \times H$ matrix of connection probabilities, each entry for each couple of clusters, then we can characterize the models as

$$
\begin{aligned}
p\left(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\Xi}\right) &= \prod_{h=1}^{H} \prod_{k=1}^{h} \xi_{hk}^{m_{hk}} (1-\xi_{hk})^{\overline{m}_{hk}} \\
p\left(\boldsymbol{\Xi} \mid \boldsymbol{Z}\right) &= \prod_{h=1}^{H} \prod_{k=1}^{h} \frac{\xi_{hk}^{a-1}(1-\xi_{hk})^{b-1}}{B(a,b)} \\
\boldsymbol{Z} &\sim \mathrm{pEPPF}(\rho,\, \rho_0,\, c,\, c_0) \qquad \text{or} \\
\boldsymbol{Z} &\sim \mathrm{pEPPF}(\sigma,\, \sigma_0,\, \theta,\, \theta_0)
\end{aligned}
\tag{4.10}
$$

for some parameters $a, b > 0$, where m_{hk} is the number of edges between a node in cluster h and one in cluster k while $\overline{m}_{hk}$ is the number of non-edges and $B\left(\cdot\,,\cdot\right)$ denotes the beta function. Notice that, since $\boldsymbol{Y}$ is symmetric (that is, the *supra*-graph $(V_M,\, E_M)$ is undirected), we are modeling just its lower-triangular part. The extension to directed *supra*-graphs is straightforward. As noticed in Schmidt and Morup (2013) and Legramanti et al. (2022), through conjugacy, the probabilities $\boldsymbol{\Xi}$ in (4.10) can be integrated out to obtain the likelihood

$$
p\left(\boldsymbol{Y} \mid \boldsymbol{Z}\right) = \prod_{h=1}^{H} \prod_{k=1}^{h} \frac{B(a+m_{hk}, b+\overline{m}_{hk})}{B(a,b)} \tag{4.11}
$$

In words, the first two lines of (4.10) define a SBM likelihood, as in Holland et al. (1983) and Nowicki and Snijders (2001), with conditionally independent Bernoulli trials for the connections and conditionally independent beta random variables for the group-specific connection probabilities. However, in our case, the model is placed on a *supra*-adjacency matrix, since our network is layered. This implies that, at the likelihood level, we are modeling connections within- and across-layers without any distinction. The layer information is indeed completely unloaded on the prior we put on the random partition, that is the distribution induced by a H-NRMI or a H-PYP. As described in Section 4.2, the division in layers is in fact mirrored by the division in sub-population of a partially exchangeable latent random array attached to the set of nodes, representing an unobserved characteristic of each individual, whose realization is driving our prior distribution on the random partition. In this way, we put prior support also on across-layers clusters and allow a proper update through the observation of connections both within- and across-layers. A somewhat similar goal is achieved by the extended stochastic block model (ESBM) in Legramanti et al. (2022), where the SBM is completed with Gibbs-type priors on the random partition. A *supervised* version of the model is presented, meaning that categorical covariates, conditionally independent on the connections given the allocations, are enforced in the prior through a cohesion function, with a product partition model technique. Such covariates can be used to convey the layer division information. In our case, the layer labeling is built-in the prior structure, resulting in a fully-probabilistic model. Moreover, the use of random partition models based on EPPFs gives for free theoretical guarantees on their Kolmogorov consistency: as can be found in Pitman (1996) for example, an EPPF, as well as a pEPPF, defines a distribution on a consistent *sequence* of nested random partitions of $[N]$, for $N \to \infty$. This implies that the *a priori* structure of the model is theoretically built to perform predictive tasks with a sequential mechanism: in practice, such tasks can be accomplished thanks to the joint urn scheme characterizing H-NRMI and H-PYP. On the other hand, the product partition model that the supervised ESBM is based on, breaks the sequential structure of the urn scheme and makes the model lose such consistency for N growing. These lacks become even more sensible whenever one wants to perform m-step-ahead prediction, that is prediction for m new nodes joining the network.

4.3.2 Clustering and co-clustering probabilities

In this Section we give some properties of the predictive probabilities induced by H-NRMI and H-PYP which have not been explicitly underlined in the literature, or have not been studied at all. We specialize the H-NRMI results for the noteworthy cases of H-DP and H-NSP.

In order to do so, here and in the rest of the Chapter, with an abuse of notation, we will denote with $\boldsymbol{Z}$ and $\boldsymbol{W}$ the *vectors* of allocations $(Z_1, \dots, Z_N)$ and $(W_1, \dots, W_N)$ obtained from the original arrays defined in (4.2) and (4.8) by simple juxtaposition of their rows. The apex $^{-r}$ will in general denote the fact that the quantity has been calculated without considering the r-th node of the network, hence disregarding its connections and allocation. In particular $\boldsymbol{X}^{-r}$ is obtained from $\boldsymbol{X}$ by marginalizing out entry (j_r, i_r), where $j_r := r \operatorname{div} d$ and $i_r := r \bmod d$, with div and mod denoting the integer division and its remainder, respectively. Instead $\boldsymbol{Z}^{-r}$ is the $(N-1)$-dimensional vector of allocations of all but the r-th nodes, with labels in order of occurrence in $\boldsymbol{X}^{-r}$. Notice that this has a different distribution than the vector obtained marginalizing out the r-th component from $\boldsymbol{Z}$, but this difference is just up to labels: indeed the correspondent distribution on the random partition is consistent, as explained in Section 4.3.1. The same holds for $\boldsymbol{W}^{-r}$. Recalling the summary statistics $\boldsymbol{\ell}$ and $\boldsymbol{q}$ defined in Section 4.2.2 and keeping in mind their interpretation in the restaurant franchise metaphor, we can state the following.

Proposition 4.1. *Let $\boldsymbol{Z}$ be a random allocation vector such that*

$$\boldsymbol{Z} \sim pEPPF(\rho,\, \rho_0,\, c,\, c_0) \tag{4.12}$$

for positive measurable functions ρ, ρ_0 and positive constants c, c_0. If $j := r \operatorname{div} d$, *then for any* $k \in [H^{-r}+1]$

$$\mathbb{P}\left(Z_r = k \,\middle|\, \boldsymbol{Z}^{-r}, \boldsymbol{W}^{-r}\right) = \frac{\sum_{t=1}^{\ell_{jk}^{-r}} \Phi^{(N_j)}_{\ell_{j\cdot}^{-r},j}\left(q_{j1}^{-r},\ldots,q_{jk}^{-r}+\boldsymbol{e}_t,\ldots,q_{jH^{-r}}^{-r}\right)}{\Phi^{(N_j-1)}_{\ell_{j\cdot}^{-r},j}\left(q_{j1}^{-r},\ldots,q_{jH^{-r}}^{-r}\right)} + \\ + \frac{\Phi^{(N_j)}_{\ell_{j\cdot}^{-r}+1,j}\left(q_{j1}^{-r},\ldots,(q_{jk}^{-r},1),\ldots,q_{jH_k^{-r}}^{-r}\right)}{\Phi^{(N_j-1)}_{\ell_{j\cdot}^{-r},j}\left(q_{j1}^{-r},\ldots,q_{jH^{-r}}^{-r}\right)} \times \\ \times \frac{\Phi^{(|\boldsymbol{\ell}^{-r}|+1)}_{H_k^{-r},0}\left(\ell_{\cdot 1}^{-r},\ldots,\ell_{\cdot k}^{-r}+1,\ldots,\ell_{\cdot H_k^{-r}}^{-r}\right)}{\Phi^{(|\boldsymbol{\ell}^{-r}|)}_{H^{-r},0}\left(\ell_{\cdot 1}^{-r},\ldots,\ell_{\cdot H^{-r}}^{-r}\right)} \tag{4.13}$$

where $\boldsymbol{e}_t$ *is the* t*-th vector of the* ℓ_{jk}^{-r}*-dimensional canonical basis and* $H_k^{-r} := k \wedge H^{-r}$

Remark 4.4. Here an interpretation of the previous result, speaking into the metaphor: the first summand is the probability of customer r sitting at an old table where dish k is eaten, while the second is the probability of sitting at a new table (first factor) and being served with dish k (second factor). Notice that, if dish k is not eaten yet in restaurant j, then $\ell_{jk}^{-r} = 0$ and the sum in the first summand disappears: k can be eaten by r just sitting at a new table. Notice also that, since labels in $\boldsymbol{Z}^{-r}$ are in order of occurrence in $\boldsymbol{X}^{-r}$, $k = H^{-r}+1$ represents the case of a dish new to the all franchise.

Proof of Proposition 4.1. Firstly, for some fixed and coherent vector $(\boldsymbol{z}^{-r}, \boldsymbol{w}^{-r})$ of allocations and sub-allocations of all but the r-th node, we have

$$\mathbb{P}\left(Z_r = k, W_r = t \,\middle|\, \boldsymbol{Z}^{-r} = \boldsymbol{z}^{-r}, \boldsymbol{W}^{-r} = \boldsymbol{w}^{-r}\right) = \\ = \frac{\mathbb{P}\left(Z_r = k,\, W_r = t,\, \boldsymbol{Z}^{-r} = \boldsymbol{z}^{-r}, \boldsymbol{W}^{-r} = \boldsymbol{w}^{-r}\right)}{\mathbb{P}\left(\boldsymbol{Z}^{-r} = \boldsymbol{z}^{-r}, \boldsymbol{W}^{-r} = \boldsymbol{w}^{-r}\right)} \tag{4.14}$$

Now, denoting with $\boldsymbol{\ell}^{-r}$ and $\boldsymbol{q}^{-r}$ the summaries defined in Section 4.2.2 relative to the allocations $(\boldsymbol{z}^{-r}, \boldsymbol{w}^{-r})$, the probability at the denominator in (4.14) can be retrieved from the pEPPF in (4.3). It suffices to notice that, having fixed a particular configuration for allocations and sub-allocation, instead of just an array of cluster frequencies, we do not need to sum over all the configurations nor multiply by the product of multinomial factors, which account for all the equiprobable configurations represented by the same summaries. Hence

$$\mathbb{P}\left(\boldsymbol{Z}^{-r}=\boldsymbol{z}^{-r},\boldsymbol{W}^{-r}=\boldsymbol{w}^{-r}\right)= \\ =\Phi^{(|\boldsymbol{\ell}^{-r}|)}_{H^{-r},0}(\ell^{-r}_{\cdot 1},\dots,\ell^{-r}_{\cdot H^{-r}})\times \\ \times\Phi^{(N_j-1)}_{\ell^{-r}_{j\cdot},j}(\boldsymbol{q}^{-r}_{j1},\dots,\boldsymbol{q}^{-r}_{jH^{-r}})\prod_{\substack{j'=1\\ j'\neq j}}^{d}\Phi^{(N_{j'})}_{\ell^{-r}_{j'\cdot},j'}(\boldsymbol{q}^{-r}_{j'1},\dots,\boldsymbol{q}^{-r}_{j'H^{-r}}) \tag{4.15}$$

Let us now express the numerator in (4.14) as a function of $\boldsymbol{\ell}^{-r}$, $\boldsymbol{q}^{-r}$, k and t. For any $t=w^{-r}_s$ for some $s\in[r-1]$ (that is any old table) such that $z^{-r}_s=k$ (that is eating dish k), we need to increase the frequency q^{-r}_{jkt}, while the frequencies in $\boldsymbol{\ell}^{-r}$ remain unchanged. Hence we have

$$\mathbb{P}\left(Z_r=k,W_r=t,\,\boldsymbol{Z}^{-r}=\boldsymbol{z}^{-r},\boldsymbol{W}^{-r}=\boldsymbol{w}^{-r}\right)= \\ =\Phi^{(|\boldsymbol{\ell}^{-r}|)}_{H^{-r},0}(\ell^{-r}_{\cdot 1},\dots,\ell^{-r}_{\cdot H^{-r}})\prod_{\substack{j'=1\\ j'\neq j}}^{d}\Phi^{(N_{j'}-1)}_{\ell^{-r}_{j'\cdot},j'}(\boldsymbol{q}^{-r}_{j'1},\dots,\boldsymbol{q}^{-r}_{j'H^{-r}}) \\ \Phi^{(N_j)}_{\ell^{-r}_{j\cdot},j}\left(\boldsymbol{q}^{-r}_{j1},\dots,\boldsymbol{q}^{-r}_{jk}+\boldsymbol{e}_t,\dots,\boldsymbol{q}^{-r}_{jH^{-r}}\right) \tag{4.16}$$

for $\boldsymbol{e}_t$ being a ℓ^{-r}_{jk}-dimensional vector of 0s, with a 1 in the t-th entry. Instead, for any $k\in[H^{-r}+1]$ and $t=\max\{w^{-r}_s\mid z^{-r}_s=k\}\wedge 1$, we have

$$\mathbb{P}\left(Z_r=k,\,W_r=t,\,\boldsymbol{Z}^{-r}=\boldsymbol{z}^{-r},\boldsymbol{W}^{-r}=\boldsymbol{w}^{-r}\right)= \\ =\Phi^{(|\boldsymbol{\ell}^{-r}|+1)}_{H^{-r}_k,0}\left(\ell^{-r}_{\cdot 1},\dots,\ell^{-r}_{\cdot k}+1,\dots,\ell^{-r}_{\cdot H^{-r}_k}\right)\prod_{\substack{j'=1\\ j'\neq j}}^{d}\Phi^{(N_{j'}-1)}_{\ell^{-r}_{j'\cdot},j'}(\boldsymbol{q}^{-r}_{j'1},\dots,\boldsymbol{q}^{-r}_{j'H^{-r}}) \\ \Phi^{(N_j)}_{\ell^{-r}_{j\cdot}+1,j}\left(\boldsymbol{q}^{-r}_{j1},\dots,(\boldsymbol{q}^{-r}_{jk},1),\dots,\boldsymbol{q}^{-r}_{jH^{-r}_k}\right) \tag{4.17}$$

In (4.17) we are considering, in the metaphor, the event of r sitting at a new table, eating either an old dish ($k\in[H^{-r}]$) or a new one ($k=H^{-r}+1$), hence we are adding a new entry to $\boldsymbol{q}^{-r}_{jk}$ and increasing ℓ^{-r}, for the creating of the new table. Notice that if $k=H^{-r}+1$ then $(\boldsymbol{q}^{-r}_{jk},1)=1$.

Now, taking the ratio in (4.14) and summing over $t\in[\ell^{-r}_{jk}+1]$, we obtain (4.13). □

Similarly we get the following for the H-PYP case.

Proposition 4.2. *Let $\boldsymbol{Z}$ be a random allocation vector such that*

$$\boldsymbol{Z}\sim pEPPF(\sigma,\,\sigma_0,\,\theta,\,\theta_0) \tag{4.18}$$

for parameters σ, $\sigma_0 \in (0,1)$ and $\theta > \sigma$, $\theta_0 > \sigma_0$. If $j := r \operatorname{div} d$, then for any $k \in [H^{-r}+1]$

$$\mathbb{P}\left(Z_r = k \,\middle|\, \boldsymbol{Z}^{-r}, \boldsymbol{W}^{-r}\right) = \mathbb{1}_{\{\ell_{\cdot k}^{-r} \neq 0\}} \left\{ \frac{\ell_{\cdot k}^{-r} - \sigma_0}{\theta_0 + |\boldsymbol{\ell}^{-r}|} \frac{\theta + \ell_{j\cdot}^{-r}\sigma}{\theta + N_j - 1} + \frac{n_{jk}^{-r} - \ell_{jk}^{-r}\sigma}{\theta + N_j - 1} \right\} + \\ + \mathbb{1}_{\{\ell_{\cdot k}^{-r} = 0\}} \frac{\theta_0 + H^{-r}\sigma_0}{\theta_0 + |\boldsymbol{\ell}^{-r}|} \frac{\theta + \ell_{j\cdot}^{-r}\sigma}{\theta + N_j - 1} \tag{4.19}$$

Remark 4.5. The expression of the probability in (4.19) makes explicit the division in the two possible scenarios. Speaking into the metaphor, if k is a new dish for the franchise we have $\ell_{\cdot k}^{-r} = 0$, hence the probability of customer r of eating it is the probability of eating a new dish at a new table. While if $\ell_{\cdot k}^{-r} \neq 0$, k is an old dish for the franchise, therefore we have the probability of serving dish k to a new table plus the probability of sitting at an old table already eating k, which is 0 if the dish is new for the restaurant ($n_{jk}^{-r} = 0$).

An immediate result we deduce from the proofs of Propositions 4.1 and 4.2 is the following.

Corollary 4.3. *Let $\boldsymbol{Z}$ be a random allocation vector as in (4.12) or (4.18) and the sub-allocation vector defined as in (4.8). Then $(\boldsymbol{\ell}, \boldsymbol{q})$ is a predictive sufficient statistics for $(\boldsymbol{Z}, \boldsymbol{W})$, that is*

$$\mathbb{P}\left(Z_r = k,\, W_r = t \,\middle|\, \boldsymbol{Z}^{-r}, \boldsymbol{W}^{-r}\right) = \mathbb{P}\left(Z_r = k,\, W_r = t \,\middle|\, \boldsymbol{\ell}^{-r}, \boldsymbol{q}^{-r}\right) \tag{4.20}$$

Example 4.6. By leveraging the expressions of the EPPF induced by a DP and a NSP, we can specialize the result in Proposition 4.1 to the H-DP and H-NSP cases. If

$$\boldsymbol{Z} \sim \mathrm{pEPPF}(\theta, \theta_0) \tag{4.21}$$

for θ, $\theta_0 > 0$, then

$$\mathbb{P}\left(Z_r = k \,\middle|\, \boldsymbol{Z}^{-r}, \boldsymbol{W}^{-r}\right) = \mathbb{1}_{\{\ell_{\cdot k}^{-r} = 0\}} \frac{\theta_0}{\theta_0 + |\boldsymbol{\ell}^{-r}|} \frac{\theta}{\theta + N_j - 1} + \\ + \mathbb{1}_{\{\ell_{\cdot k}^{-r} \neq 0\}} \left\{ \frac{\ell_{\cdot k}^{-r}}{\theta_0 + |\boldsymbol{\ell}^{-r}|} \frac{\theta}{\theta + N_j - 1} + \frac{n_{jk}^{-r}}{\theta + N_j - 1} \right\} \tag{4.22}$$

If

$$\boldsymbol{Z} \sim \mathrm{pEPPF}(\sigma, \sigma_0) \tag{4.23}$$

for σ, $\sigma_0 \in (0,1)$, then

$$\mathbb{P}\left(Z_r = k \,\middle|\, \boldsymbol{Z}^{-r}, \boldsymbol{W}^{-r}\right) = \mathbb{1}_{\{\ell_{\cdot k}^{-r} = 0\}} \frac{H^{-r}\sigma_0}{|\boldsymbol{\ell}^{-r}|} \frac{\ell_{j\cdot}^{-r}\sigma}{N_j - 1} + \\ + \mathbb{1}_{\{\ell_{\cdot k}^{-r} \neq 0\}} \left\{ \frac{\ell_{\cdot k}^{-r} - \sigma_0}{|\boldsymbol{\ell}^{-r}|} \frac{\ell_{j\cdot}^{-r}\sigma}{N_j - 1} + \frac{n_{jk}^{-r} - \ell_{jk}^{-r}\sigma}{N_j - 1} \right\} \tag{4.24}$$

Notice that, as expected due to the relations linking these nonparametric priors, putting $\sigma = \sigma_0 = 0$ and $\theta = \theta_0 = 0$ in (4.19) gives (4.22) and (4.24) respectively.

Now we will give expressions for predictive co-clustering probabilities for the general case of H-NRMI and then specialized to H-DP.

Theorem 4.4. *Let $\boldsymbol{Z}$ be a random allocation vector such that*

$$\boldsymbol{Z} \sim pEPPF(\rho,\, \rho_0,\, c,\, c_0) \tag{4.25}$$

then, if $j = r \operatorname{div} d = s \operatorname{div} d$

$$\begin{aligned}
&\mathbb{P}\left(\{Z_r = Z_s\} \,\middle|\, \boldsymbol{\ell}^{-rs}, \boldsymbol{q}^{-rs}\right) = \\
&= \sum_{k=1}^{H^{-rs}+1} \left\{ \frac{\sum_{t=1}^{\ell_{jk}^{-rs}} \Phi^{(N_j)}_{\ell_{j\cdot}^{-rs},\, j}\left(\boldsymbol{q}_{j1}^{-rs}, \dots, \boldsymbol{q}_{jk}^{-rs} + 2\boldsymbol{e}_t, \dots, \boldsymbol{q}_{jH^{-rs}}^{-rs}\right)}{\Phi^{(N_j-2)}_{\ell_{j\cdot}^{-rs},\, j}\left(\boldsymbol{q}_{j1}^{-rs}, \dots, \boldsymbol{q}_{jH^{-rs}}^{-rs}\right)} + \right. \\
&\quad + \frac{\sum_{A \in C_{\ell_{jk}^{-rs},\, 2}} \Phi^{(N_j)}_{\ell_{j\cdot}^{-rs},\, j}\left(\boldsymbol{q}_{j1}^{-rs}, \dots, \boldsymbol{q}_{jk}^{-rs} + \boldsymbol{e}_A, \dots, \boldsymbol{q}_{jH^{-rs}}^{-rs}\right)}{\Phi^{(N_j-2)}_{\ell_{j\cdot}^{-rs},\, j}\left(\boldsymbol{q}_{j1}^{-rs}, \dots, \boldsymbol{q}_{jH^{-rs}}^{-rs}\right)} + \\
&\quad + \frac{\Phi^{(|\boldsymbol{\ell}^{-rs}|+1)}_{H_k^{-rs},0}\left(\ell_{\cdot 1}^{-rs}, \dots, \ell_{\cdot k}^{-rs} + 1, \dots, \ell_{\cdot H_k^{-rs}}^{-rs}\right)}{\Phi^{(|\boldsymbol{\ell}^{-rs}|)}_{H^{-rs},0}\left(\ell_{\cdot 1}^{-rs}, \dots, \ell_{\cdot H^{-rs}}^{-rs}\right)} \times \\
&\quad \times \left\{ \frac{\Phi^{(N_j)}_{\ell_{j\cdot}^{-rs}+1,\, j}\left(\boldsymbol{q}_{j1}^{-rs}, \dots, (\boldsymbol{q}_{jk}^{-rs}, 2), \dots, \boldsymbol{q}_{jH_k^{-rs}}^{-rs}\right)}{\Phi^{(N_j-2)}_{\ell_{j\cdot}^{-rs},\, j}\left(\boldsymbol{q}_{j1}^{-rs}, \dots, \boldsymbol{q}_{jH^{-rs}}^{-rs}\right)} + \right. \\
&\quad \left. + \frac{\sum_{t=1}^{\ell_{jk}^{-rs}} \Phi^{(N_j)}_{\ell_{j\cdot}^{-rs}+1, j}\left(\boldsymbol{q}_{j1}^{-rs}, \dots, (\boldsymbol{q}_{jk}^{-rs} + \boldsymbol{e}_t,\, 1), \dots, \boldsymbol{q}_{jH_k^{-rs}}^{-rs}\right)}{\Phi^{(N_j-2)}_{\ell_{j\cdot}^{-rs},\, j}\left(\boldsymbol{q}_{j1}^{-rs}, \dots, \boldsymbol{q}_{jH^{-rs}}^{-rs}\right)} \right\} + \\
&\quad + \frac{\Phi^{(|\boldsymbol{\ell}^{-rs}|+2)}_{H_k^{-rs},0}\left(\ell_{\cdot 1}^{-rs}, \dots, \ell_{\cdot k}^{-rs} + 2, \dots, \ell_{\cdot H_k^{-rs}}^{-rs}\right)}{\Phi^{(|\boldsymbol{\ell}^{-rs}|)}_{H^{-rs},0}\left(\ell_{\cdot 1}^{-rs}, \dots, \ell_{\cdot H^{-rs}}^{-rs}\right)} \times \\
&\quad \left. \times \frac{\Phi^{(N_j)}_{\ell_{j\cdot}^{-rs}+2,\, j}\left(q_{j1}^{-rs}, \dots, (q_{jk}^{-rs}, 1, 1), \dots, q_{jH_k^{-rs}}^{-rs}\right)}{\Phi^{(N_j-2)}_{\ell_{j\cdot}^{-rs},\, j}\left(q_{j1}^{-rs}, \dots, q_{jH^{-rs}}^{-rs}\right)} \right\}
\end{aligned} \tag{4.26}$$

while if $r \operatorname{div} d = j_r$ *and* $s \operatorname{div} d = j_s$*, with* $j_r \neq j_s$

$$
\begin{aligned}
&\mathbb{P}\left(\{Z_r = Z_s\} \,\middle|\, \boldsymbol{\ell}^{-rs}, \boldsymbol{q}^{-rs}\right) = \\
&= \sum_{k=1}^{H^{-rs}+1} \left\{ \frac{\sum_{t=1}^{\ell_{j_r k}^{-rs}} \Phi^{(N_{j_r})}_{\ell_{j_r \cdot}, j_r} \left(\boldsymbol{q}_{j_r 1}^{-rs}, \dots, \boldsymbol{q}_{j_r k}^{-rs} + \boldsymbol{e}_t, \dots, \boldsymbol{q}_{j_r H^{-rs}}^{-rs}\right)}{\Phi^{(N_{j_r}-1)}_{\ell_{j_r \cdot}, j_r} \left(\boldsymbol{q}_{j_r 1}^{-rs}, \dots, \boldsymbol{q}_{j_r H^{-rs}}^{-rs}\right)} \right. \\
&\quad \frac{\sum_{t=1}^{\ell_{j_s k}^{-rs}} \Phi^{(N_{j_s})}_{\ell_{j_s \cdot}^{-rs}, j_s} \left(\boldsymbol{q}_{j_s 1}^{-rs}, \dots, \boldsymbol{q}_{j_s k}^{-rs} + \boldsymbol{e}_t, \dots, \boldsymbol{q}_{j_s H^{-rs}}^{-rs}\right)}{\Phi^{(N_{j_s}-1)}_{\ell_{j_s \cdot}^{-rs}, j_s} \left(\boldsymbol{q}_{j_s 1}^{-rs}, \dots, \boldsymbol{q}_{j_s H^{-rs}}^{-rs}\right)} + \\
&\quad + \frac{\Phi^{(|\boldsymbol{\ell}^{-rs}|+1)}_{H_k^{-rs}, 0} \left(\ell_{\cdot 1}^{-rs}, \dots, \ell_{\cdot k}^{-rs} + 1, \dots, \ell_{\cdot H_k^{-rs}}^{-rs}\right)}{\Phi^{(|\boldsymbol{\ell}^{-rs}|)}_{H^{-rs}, 0} \left(\ell_{\cdot 1}^{-rs}, \dots, \ell_{\cdot H^{-rs}}^{-rs}\right)} \times \\
&\quad \times \left\{ \frac{\Phi^{(N_{j_s})}_{\ell_{j_s \cdot}^{-rs}+1, j_s} \left(\boldsymbol{q}_{j_s 1}^{-rs}, \dots, (\boldsymbol{q}_{j_s k}^{-rs}, 1), \dots, \boldsymbol{q}_{j_s H_k^{-rs}}^{-rs}\right)}{\Phi^{(N_{j_s}-1)}_{\ell_{j_s \cdot}^{-rs}, j_s} \left(\boldsymbol{q}_{j_s 1}^{-rs}, \dots, \boldsymbol{q}_{j_s H^{-rs}}^{-rs}\right)} \times \right. \\
&\quad \times \frac{\sum_{t=1}^{\ell_{j_r k}^{-rs}} \Phi^{(N_{j_r})}_{\ell_{j_r \cdot}^{-rs}, j_r} \left(\boldsymbol{q}_{j_r 1}^{-rs}, \dots, \boldsymbol{q}_{j_r k}^{-rs} + \boldsymbol{e}_t, \dots, \boldsymbol{q}_{j_r H^{-rs}}^{-rs}\right)}{\Phi^{(N_{j_r}-1)}_{\ell_{j_r \cdot}^{-rs}, j_r} \left(\boldsymbol{q}_{j_r 1}^{-rs}, \dots, \boldsymbol{q}_{j_r H^{-rs}}^{-rs}\right)} + \\
&\quad + \frac{\Phi^{(N_{j_r})}_{\ell_{j_r \cdot}^{-rs}+1, j_r} \left(\boldsymbol{q}_{j_r 1}^{-rs}, \dots, (\boldsymbol{q}_{j_r k}^{-rs}, 1), \dots, \boldsymbol{q}_{j_r H_k^{-rs}}^{-rs}\right)}{\Phi^{(N_{j_r}-1)}_{\ell_{j_r \cdot}^{-rs}, j_r} \left(\boldsymbol{q}_{j_r 1}^{-rs}, \dots, \boldsymbol{q}_{j_r H^{-rs}}^{-rs}\right)} \times \\
&\quad \left. \times \frac{\sum_{t=1}^{\ell_{j_s k}^{-rs}} \Phi^{(N_{j_s})}_{\ell_{j_s \cdot}^{-rs}, j_s} \left(\boldsymbol{q}_{j_s 1}^{-rs}, \dots, \boldsymbol{q}_{j_s k}^{-rs} + \boldsymbol{e}_t, \dots, \boldsymbol{q}_{j_s H^{-rs}}^{-rs}\right)}{\Phi^{(N_{j_s}-1)}_{\ell_{j_s \cdot}^{-rs}, j_s} \left(\boldsymbol{q}_{j_s 1}^{-rs}, \dots, \boldsymbol{q}_{j_s H^{-rs}}^{-rs}\right)} \right\} + \\
&\quad + \frac{\Phi^{(|\boldsymbol{\ell}^{-rs}|+2)}_{H_k^{-rs}, 0} \left(\ell_{\cdot 1}^{-rs}, \dots, \ell_{\cdot k}^{-rs} + 2, \dots, \ell_{\cdot H_k^{-rs}}^{-rs}\right)}{\Phi^{(|\boldsymbol{\ell}^{-rs}|)}_{H^{-rs}, 0} \left(\ell_{\cdot 1}^{-rs}, \dots, \ell_{\cdot H^{-rs}}^{-rs}\right)} \\
&\quad \frac{\Phi^{(N_{j_r})}_{\ell_{j_r \cdot}^{-rs}+1, j_r} \left(\boldsymbol{q}_{j_r 1}^{-rs}, \dots, (\boldsymbol{q}_{j_r k}^{-rs},\ 1), \dots, \boldsymbol{q}_{j_r H_k^{-rs}}^{-rs}\right)}{\Phi^{(N_{j_r}-1)}_{\ell_{j_r \cdot}^{-rs}, j_r} \left(\boldsymbol{q}_{j_r 1}^{-rs}, \dots, \boldsymbol{q}_{j_r H^{-rs}}^{-rs}\right)} \times
\end{aligned}
$$

$$\times \left. \frac{\Phi^{(N_{j_s})}_{\ell^{-rs}_{j_s \cdot}+1, j_s}\left(\boldsymbol{q}^{-rs}_{j_s 1}, \dots, (\boldsymbol{q}^{-rs}_{j_s k}, 1), \dots, \boldsymbol{q}^{-rs}_{j_s H^{-rs}_k}\right)}{\Phi^{(N_{j_s}-1)}_{\ell^{-rs}_{j_s \cdot}, j_s}\left(\boldsymbol{q}^{-rs}_{j_s 1}, \dots, \boldsymbol{q}^{-rs}_{j_s H^{-rs}}\right)} \right\} \quad (4.27)$$

with $C_{h,2}$ the set of 2*-combinations of* $[h]$ *and for* $A \in C_{h,2}$*,* $\boldsymbol{e}_A$ *is a* h*-dimensional vector with* $\boldsymbol{e}_i = 1$ *if* $i \in A$ *and* 0 *otherwise, and* $H^{-rs}_k = H^{-rs} \wedge k$*.*

Remark 4.7. We can interpret in the restaurant metaphor the components of (4.26) and (4.27) as follows. In the first case we have, in order, the probability of customers r and s sitting at the same old table, plus the probability of sitting at two different old tables but eating the same dish, plus the probability of creating a new table and either both sit there, either one sits at the new and the other at an old table eating the same dish, plus, finally, the probability of creating two new tables served with the same dish. In the second case, being in different restaurants, we have the probability of sitting at two old tables eating the same dish, plus the probability of creating a new table in one of the two restaurants, while the other customer sits in an old table eating the same dish, plus the probability of creating new tables at both restaurants, both served with the same dish.

Proof of Proposition 4.4. As in proof of Proposition 4.1, we have

$$\mathbb{P}\left(\{Z_r = Z_s\}, W_r = t, W_s = t' \,\middle|\, \boldsymbol{\ell}^{-rs}, \boldsymbol{q}^{-rs}\right) = \\ = \frac{p(\{Z_r = Z_s\}, W_r = t, W_s = t', \boldsymbol{\ell}^{-rs}, \boldsymbol{q}^{-rs})}{p(\boldsymbol{\ell}^{-rs}, \boldsymbol{q}^{-rs})} \quad (4.28)$$

and

$$p(\{Z_r = Z_s\}, W_r = t, W_s = t', \boldsymbol{\ell}^{-rs}, \boldsymbol{q}^{-rs}) = \\ = \sum_{k=1}^{H^{-rs}+1} p(Z_r = k, Z_s = k, W_r = t, W_s = t', \boldsymbol{\ell}^{-rs}, \boldsymbol{q}^{-rs}) \quad (4.29)$$

Now, if $j = j_r = j_s$, the probability in the sum, for any k, t, t' will comprehend the common factor

$$\prod_{\substack{j'=1 \\ j' \neq j}}^{d} \Phi^{(N_{j'})}_{\ell^{-rs}_{j' \cdot}, j'}(\boldsymbol{q}^{-rs}_{j' 1}, \dots, \boldsymbol{q}^{-rs}_{j' H^{-rs}}) \quad (4.30)$$

given by the allocations in all the other layers. The form of the remaining factor depends on t and t': if they are old tables we have

$$\Phi^{(|\boldsymbol{\ell}^{-rs}|)}_{H^{-rs}, 0}(\ell^{-rs}_{\cdot 1}, \dots, \ell^{-rs}_{\cdot H^{-rs}}) \Phi^{(N_j)}_{\ell^{-rs}_{j \cdot}, j}\left(\boldsymbol{q}^{-rs}_{j1}, \dots, \boldsymbol{q}^{-rs}_{jk} + \boldsymbol{e}_t + \boldsymbol{e}_{t'}, \dots, \boldsymbol{q}^{-rs}_{jH^{-rs}}\right) \quad (4.31)$$

if t is old and t' is new we have

$$\Phi^{(|\boldsymbol{\ell}^{-rs}|)}_{H^{-rs},0}(\ell^{-rs}_{\cdot 1},\dots,\ell^{-rs}_{\cdot k}+1,\dots,\ell^{-rs}_{\cdot H^{-rs}})\times \\ \times \Phi^{(N_j)}_{\ell^{-rs}_{j\cdot},j}\left(\boldsymbol{q}^{-rs}_{j1},\dots,(\boldsymbol{q}^{-rs}_{jk}+\boldsymbol{e}_t,1),\dots,\boldsymbol{q}^{-rs}_{jH^{-rs}}\right) \tag{4.32}$$

if $t=t'$ is new

$$\Phi^{(|\boldsymbol{\ell}^{-rs}|)}_{H^{-rs},0}(\ell^{-rs}_{\cdot 1},\dots,\ell^{-rs}_{\cdot k}+1,\dots,\ell^{-rs}_{\cdot H^{-rs}})\times \\ \times \Phi^{(N_j)}_{\ell^{-rs}_{j\cdot},j}\left(\boldsymbol{q}^{-rs}_{j1},\dots,(\boldsymbol{q}^{-rs}_{jk},2),\dots,\boldsymbol{q}^{-rs}_{jH^{-rs}}\right) \tag{4.33}$$

while if they are both new and different

$$\Phi^{(|\boldsymbol{\ell}^{-rs}|)}_{H^{-rs},0}(\ell^{-rs}_{\cdot 1},\dots,\ell^{-rs}_{\cdot k}+2,\dots,\ell^{-rs}_{\cdot H^{-rs}})\times \\ \times \Phi^{(N_j)}_{\ell^{-rs}_{j\cdot},j}\left(\boldsymbol{q}^{-rs}_{j1},\dots,(\boldsymbol{q}^{-rs}_{jk},1,1),\dots,\boldsymbol{q}^{-rs}_{jH^{-rs}}\right) \tag{4.34}$$

It is easy to see that the denominator in (4.28) is

$$\Phi^{(|\boldsymbol{\ell}^{-rs}|)}_{H^{-rs},0}(\ell^{-rs}_{\cdot 1},\dots,\ell^{-rs}_{\cdot H^{-rs}})\times \\ \times \Phi^{(N_j-2)}_{\ell^{-rs}_{j\cdot},j}\left(\boldsymbol{q}^{-rs}_{j1},\dots,\boldsymbol{q}^{-rs}_{jH^{-rs}}\right)\prod_{\substack{j'=1\\ j'\neq j}}^{d}\Phi^{(N_{j'})}_{\ell^{-rs}_{j'\cdot},j'}(\boldsymbol{q}^{-rs}_{j'1},\dots,\boldsymbol{q}^{-rs}_{j'H^{-rs}}) \tag{4.35}$$

Taking the ratio and summing over all the possible choices of t and t', we obtain (4.26).

The proof of (4.27) follows the same line of reasoning. It suffices to notice that, being $j_r \neq j_s$, it is not possible for r ad s to sit at the same table, neither new nor old. $\square$

If we substitute the expression of the EPPF induced by a DP at the two levels of hierarchy in (4.26) and (4.27), we obtain the following.

Corollary 4.5. *Let $\boldsymbol{Z}$ be a random allocation vector such that*

$$\boldsymbol{Z} \sim pEPPF(\theta,\,\theta_0)$$

then, if both nodes r and s are in layer j

$$\mathbb{P}\left(\{Z_r = Z_s\}\,\middle|\,\boldsymbol{\ell}^{-rs},\boldsymbol{q}^{-rs}\right) =$$

$$\frac{1}{(\theta+N_j-2)(\theta+N_j-1)}\left\{\frac{\theta_0\theta}{\theta_0+|\boldsymbol{\ell}^{-rs}|}+\frac{\theta_0\theta^2}{(\theta_0+|\boldsymbol{\ell}^{-rs}|)(\theta_0+|\boldsymbol{\ell}^{-rs}|+1)}+\right.$$
$$+\sum_{\substack{h=0\\ \ell_{\cdot h}\neq 0}}^{H^{-rs}} n_{jh}^{-rs}(n_{jh}^{-rs}+1)+\frac{|\boldsymbol{\ell}^{-rs}|N_j\theta}{\theta_0+|\boldsymbol{\ell}^{-rs}|}+$$
$$\left.+\sum_{\substack{h=0\\ \ell_{\cdot h}\neq 0}}^{H^{-rs}} \frac{\ell_{\cdot h}^{-rs}(\ell_{\cdot h}^{-rs}+1)\theta^2}{(\theta_0+|\boldsymbol{\ell}^{-rs}|)(\theta_0+|\boldsymbol{\ell}^{-rs}|+1)}\right\} \tag{4.36}$$

while if node r is in layer j_r and node s is in layer j_s, with $j_r \neq j_s$

$$\mathbb{P}\left(\{Z_r=Z_s\}\,\middle|\,\boldsymbol{\ell}^{-rs},\boldsymbol{q}^{-rs}\right)=$$
$$\frac{1}{(\theta+N_{j_r}-1)(\theta+N_{j_s}-1)}\left\{\frac{\theta_0\theta^2}{(\theta_0+|\boldsymbol{\ell}^{-rs}|)(\theta_0+|\boldsymbol{\ell}^{-rs}|+1)}+\right.$$
$$+\sum_{\substack{h=0\\ \ell_{\cdot k}\neq 0}}^{H^{-rs}} n_{j_r h}^{-rs}n_{j_s h}^{-rs}+\frac{|\boldsymbol{\ell}^{-rs}|\{N_{j_r}+N_{j_s}-2\}\theta}{\theta_0+|\boldsymbol{\ell}^{-rs}|}+$$
$$\left.+\sum_{\substack{h=0\\ \ell_{\cdot h}\neq 0}}^{H^{-rs}} \frac{\ell_{\cdot h}^{-rs}(\ell_{\cdot h}^{-rs}+1)\theta^2}{(\theta_0+|\boldsymbol{\ell}^{-rs}|)(\theta_0+|\boldsymbol{\ell}^{-rs}|+1)}\right\} \tag{4.37}$$

4.3.3 Prior elicitation

When we employ nonparametric priors, understanding how to induce elicited prior information is often a challenging task, even more if the prior is placed on the space of distributions of random partitions, as in the case of our model. However, leveraging results obtained in Section 4.3.2 about predictive probabilities of H-NRMI priors, it is possible to give some insight on the kind of behaviour we enforce on the predictive mechanism via the hyperparameters of our prior. In particular, in the case of a H-DP prior, we give here conditions on the hyperparameters so that the allocation of a node in a cluster already present in its layer has always higher predictive probability than the allocation to a cluster new to the layer, both considering or not the creation of a (totally) new cluster. Namely, we give conditions on the parameters to enforce a specific predictive clustering behaviour, given the sub-cluster frequencies. In general, as can be argued from (4.22), the probability of allocating a node r in a cluster with representatives in its layer j or

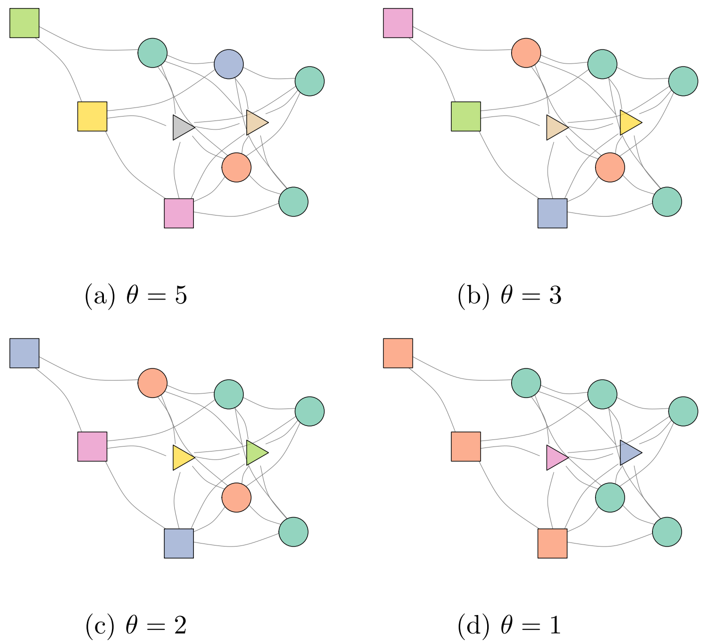


Figure 4.4: Posterior point estimate in a small network. *The prior on the allocations is a H-DP with $\theta_0 = 5$ while θ varies as in the captions.*

in a 'foreign' cluster depends on the proportion

$$p_r = \frac{1}{|\boldsymbol{\ell}^{-r}|} \sum_{k \in D_j^{-r}} \ell_{\cdot k}^{-r} \tag{4.38}$$

where D_j^{-r} is the set of unique cluster labels of nodes in layer j, which may be seen, in the metaphor, as an indicator of the popularity of the dishes eaten at restaurant j. Intuitively we have more ways of clustering within-layer, since we can either sit at an old table or create a new table eating an old dish for the restaurant, while to cluster strictly across-layers we have only the latter option. But, still, if 'foreign' dishes are very popular, meaning that lots of tables are served with them, and hence p_r in (4.38) is small enough, then a clustering strictly across-layers becomes more probable. Nonetheless, in the case of H-DP, it is possible to determine conditions on the parameters so to eliminate the dependence on the proportion p_r, as in the following.

Proposition 4.6. *Let $\boldsymbol{Z}$ be a random allocation vector such that*

$$\boldsymbol{Z} \sim pEPPF(\theta,\, \theta_0)$$

then

$$\theta \leq (N_j - 1)\left(\frac{\theta_0}{|\boldsymbol{\ell}^{-r}|} + 1\right) \implies$$
$$\implies \mathbb{P}\left(Z_r \in D_j^{-r} \,\middle|\, \boldsymbol{\ell}^{-r}, \boldsymbol{q}^{-r}\right) \geq \mathbb{P}\left(Z_r \in [H^{-r}] \setminus D_j^{-r} \,\middle|\, \boldsymbol{\ell}^{-r}, \boldsymbol{q}^{-r}\right) \quad (4.39)$$

Moreover, we have that

$$\theta \;\leq\; N_j - 1 \implies \mathbb{P}\left(Z_r \in D_j^{-r} \,\middle|\, \boldsymbol{\ell}^{-r}, \boldsymbol{q}^{-r}\right) \;\geq\; \mathbb{P}\left(Z_r \notin D_j^{-r} \,\middle|\, \boldsymbol{\ell}^{-r}, \boldsymbol{q}^{-r}\right) \quad (4.40)$$

Remark 4.8. With both conditions in Proposition 4.6 we are enforcing the clustering to be more adherent to the layer division, by making *intra*-layer clusters more probable than *inter*-layer ones. Notice that conditions in (4.40) imply the ones in (4.39), since $\{Z_r \notin D_j^{-r}\}$ includes Z_r being a new cluster label. Moreover notice that for large enough networks, (or in large enough layers) the second conditions are always attained. Notice, finally, that in general, taking $\theta << \theta_0$ means to favor a situation in with new tables are not so often created, but when they are, it is more probable that they are served with dishes new to the all franchise.

Proof of Proposition 4.6. Summing over $k \in D_j^{-r}$ and $k \in [H^{-r} \setminus D_j^{-r}]$ the predictive probabilities in (4.48), we have that the inequality at the right hand side of (4.39) is satisfied if and only if

$$\frac{|\boldsymbol{\ell}^{-r}|p_r}{\theta_0 + |\boldsymbol{\ell}^{-r}|}\frac{\theta}{\theta + N_j - 1} + \frac{N_j - 1}{\theta + N_j - 1} \geq \frac{|\boldsymbol{\ell}^{-r}|(1-p_r)}{\theta_0 + |\boldsymbol{\ell}^{-r}|}\frac{\theta}{\theta + N_j - 1} \quad (4.41)$$

where p_r is defined in (4.38). This is equivalent to

$$p_r \geq \frac{1}{2}\left\{1 - \frac{|\boldsymbol{\ell}^{-r}| + \theta_0}{|\boldsymbol{\ell}^{-r}|}\frac{N_j - 1}{\theta}\right\} \quad (4.42)$$

Hence whenever the right hand side of (4.42) is negative, regardless of the value of the proportion p_r of tables eating dishes eaten in restaurant j, the inequality is always satisfied. This is implied by the left hand side of (4.39).

Summing to the right hand side of (4.41) the probability of r eating a new dish, that is

$$\frac{\theta_0}{\theta_0 + |\boldsymbol{\ell}^{-r}|}\frac{\theta}{\theta + N_j - 1} \quad (4.43)$$

we obtain the inequality in the right hand side of (4.40). With the previous strategy we retrieve the sufficient condition. □

An illustration of such property, showing the behaviour of the prior on small networks with unbalanced layer division is given in Figure 4.4. Here, the shape of the nodes represent the layer and the color the inferred partition. It is clear how, decreasing the parameter θ under the threshold of the numerosity of each layer, the inferred clustering becomes more and more adherent to the division in layers, starting from larger layers.

4.4 Posterior inference and prediction

Our goal is to retrieve the posterior distribution of the allocation $\boldsymbol{Z}$ given the connections $\boldsymbol{Y}$, as well as deal with new nodes incoming into the network. We rely on MCMC algorithms to sample from such distribution and perform prediction, as well as on Monte Carlo approximations to evaluate functionals of the posterior. In particular, we devise a Gibbs sampler for posterior inference, combining the variable augmentation given by the labels in (4.7), the collapsed structure typical of Bayesian SBMs, also leveraged in Legramanti et al. (2022), and the joint predictive probabilities of the nested clustering directed by a H-NRMI or a H-PYP. As far as prediction is concerned, we provide an algorithm that can be adapted to the case of prediction of the allocation of a new node only, once one knows its connections, but in general performs joint prediction on both allocation and connections of the new node.

4.4.1 Gibbs sampler

The objective distribution is $p(\boldsymbol{Z} \mid \boldsymbol{Y})$. As mentioned, we leverage the (further) latent table labeling, to obtain closed-form expressions for the conditional probabilities needed to built a Gibbs sampler. Hence we target the distribution $p(\boldsymbol{Z}, \boldsymbol{W} \mid \boldsymbol{Y})$. We determine the blocked-conditionals we need as follows. Firstly

$$\begin{aligned}\mathbb{P}\left(Z_r = h, W_r = t \,\middle|\, \boldsymbol{Y}, \boldsymbol{Z}^{-r}, \boldsymbol{W}^{-r}\right) = \\ = \mathbb{P}\left(Z_r = h, W_r = t \,\middle|\, \boldsymbol{\ell}^{-r}, \boldsymbol{q}^{-r}\right) \frac{p(\boldsymbol{Y} \mid Z_r = h, \boldsymbol{Z}^{-r})}{p(\boldsymbol{Y} \mid \boldsymbol{Z}^{-r}, \boldsymbol{W}^{-r})} \qquad (4.44)\end{aligned}$$

for any $h \in [H^{-r}+1]$ and $t \in [\ell_{j\cdot}+1]$, with $j = r \text{ div } d$, where we simply employed Bayes' theorem, Corollary 4.3 and that, by construction $\boldsymbol{Y} \mid \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{W}$. Indeed the SBM is defined conditionally on the clustering in $\boldsymbol{Z}$ and the sub-clustering $\boldsymbol{W}$ does not add any information. Now, since

$$p\left(\boldsymbol{Y} \mid \boldsymbol{Z}^{-r}, \boldsymbol{W}^{-r}\right) = p\left(\boldsymbol{Y}^{-r} \mid \boldsymbol{Z}^{-r}\right) p\left(\boldsymbol{y}_r \mid \boldsymbol{Z}^{-r}, \boldsymbol{W}^{-r}\right) \qquad (4.45)$$

where $\boldsymbol{y}_r$ denotes the vector of connections of the r-th node, the ratio in (4.44) is such that

$$\frac{p\left(\boldsymbol{Y} \mid Z_r = h,\, \boldsymbol{Z}^{-r}\right)}{p\left(\boldsymbol{Y} \mid \boldsymbol{Z}^{-r},\, \boldsymbol{W}^{-r}\right)} \propto \frac{p\left(\boldsymbol{Y} \mid Z_r = h,\, \boldsymbol{Z}^{-r}\right)}{p\left(\boldsymbol{Y}^{-r} \mid \boldsymbol{Z}^{-r}\right)} \tag{4.46}$$

Notice that, in general, $p\left(\boldsymbol{y}_r \mid \boldsymbol{Z}^{-r},\, \boldsymbol{W}^{-r}\right) \neq p\left(\boldsymbol{y}_r \mid \boldsymbol{Z}^{-r}\right)$. Intuitively, the conditional distribution of the connection probabilities we need to integrate out to get $p\left(\boldsymbol{y}_r \mid \boldsymbol{Z}^{-r},\, \boldsymbol{W}^{-r}\right)$ depends on whether Z_r is a new allocation or not, and the probability of these events depends in turn on the sub-allocations in $\boldsymbol{W}^{-r}$. Anyways, the right hand side of (4.46) simplifies, as suggested in Schmidt and Morup (2013) and leveraged in Legramanti et al. (2022), employing the expression in (4.11), and allowing us to collapse the sampling of the cluster-specific probabilities $\boldsymbol{\Xi}$. Hence we get

$$\begin{aligned}
&\mathbb{P}\left(Z_r = h, W_r = t \,\middle|\, \boldsymbol{Y},\, \boldsymbol{Z}^{-r}, \boldsymbol{W}^{-r}\right) \propto \\
&\qquad \propto \mathbb{P}\left(Z_r = h, W_r = t \,\middle|\, \boldsymbol{\ell}^{-r}, \boldsymbol{q}^{-r}\right) \prod_{k=1}^{H^{-r}} \frac{B(a + m_{hk}^{-r} + v_{rk}, b + \overline{m}_{hk}^{-r} + \overline{v}_{rk})}{B(a + m_{hk}^{-r}, b + \overline{m}_{hk}^{-r})}
\end{aligned} \tag{4.47}$$

where v_{rk} and $\overline{v}_{rk}$ denote the number of edges and non-edges between node r and nodes in cluster k. The quantity we need to evaluate at each step of the sampler, and for any couple (h, t), factorizes then in a 'prior term', given by the joint predictive probabilities of allocation, and a 'likelihood term', dependent on observed edges. The latter can be simply computed by obtaining, from the adjacency matrix $\boldsymbol{Y}$, the matrices $\boldsymbol{m}^{-r} = (m_{hk}^{-r})$, $\overline{\boldsymbol{m}}^{-r} = (\overline{m}_{hk}^{-r})$ of the number of edges and non-edges between clusters (given by $\boldsymbol{Z}^{-r}$), disregarding node r and the vectors $\boldsymbol{v}_r = (v_{rk})$, $\overline{\boldsymbol{v}}_r = (\overline{v}_{rk})$ for any $r \in [N]$. For what concern the 'prior term', one may notice that the $(H^{-r} + 1) \times (\ell_{j\cdot}^{-r} + 1)$ matrix of probabilities we need to evaluate for each node r is actually very sparse. Indeed, speaking in the metaphor, the probability of eating dish h at table t is non-zero just if t is an old table already served with h, or t is a new table. Hence every column of the matrix but the last one has just one non-zero element. General expressions for such joint predictive probabilities can be retrieved as in the proofs of Propositions 4.1 and 4.2. In the following we give expressions for the H-DP, H-NSP and H-PYP cases, as well as a scheme of the sampler in Algorithm 1.

Example 4.9. If $\boldsymbol{Z} \sim \text{pEPPF}(\theta,\, \theta_0)$, for $\theta,\, \theta_0 > 0$ and $j = r \text{ div } d$, we have

$$\begin{aligned}
&\mathbb{P}\left(Z_r = h, W_r = t \,\middle|\, \boldsymbol{\ell}^{-r}, \boldsymbol{q}^{-r}\right) = \\
&\qquad = \mathbb{1}_{\{\ell_{\cdot h}^{-r} = 0\}} \left\{ \frac{\theta_0}{\theta_0 + |\boldsymbol{\ell}^{-r}|} \frac{\theta}{\theta + N_j - 1} \right\} + \\
&\qquad\quad + \mathbb{1}_{\{\ell_{\cdot h}^{-r} \neq 0\}} \left\{ \frac{\ell_{\cdot h}^{-r}}{\theta_0 + |\boldsymbol{\ell}^{-r}|} \frac{\theta}{\theta + N_j - 1} + \frac{q_{jht}^{-r}}{\theta + N_j - 1} \right\}
\end{aligned} \tag{4.48}$$

for any $h \in [H^{-r}+1]$ and $t \in [\ell_{j\cdot}+1]$.

Remark 4.10. In the H-DP case, the result in Corollary 4.3 can be pushed forward saying that the vector of the column-wise sums $(\ell_{\cdot 1}, \ldots, \ell_{\cdot H})$ of the matrix $\boldsymbol{\ell}$ and the array $\boldsymbol{q}$ are predictive sufficient for the couple $(\boldsymbol{Z}, \boldsymbol{W})$.

Example 4.11. If $\boldsymbol{Z} \sim$ pEPPF(σ, σ_0), for σ, $\sigma_0 \in (0,1)$ and $j = r$ div d, we have

$$\begin{aligned}\mathbb{P}\left(Z_r = h, W_r = t \,\middle|\, \boldsymbol{\ell}^{-r}, \boldsymbol{q}^{-r}\right) &= \\ &= \mathbb{1}_{\{\ell_{\cdot h}^{-r}=0\}} \frac{H^{-r}\sigma_0}{|\boldsymbol{\ell}^{-r}|} \frac{\ell_{j\cdot}^{-r}\sigma}{N_j - 1} + \\ &\quad + \mathbb{1}_{\{\ell_{\cdot h}^{-r}\neq 0\}} \left\{ \frac{\ell_{\cdot h}^{-r} - \sigma_0}{|\boldsymbol{\ell}^{-r}|} \frac{\ell_{j\cdot}^{-r}\sigma}{N_j - 1} + \frac{q_{jht}^{-r} - \sigma}{N_j - 1} \right\}\end{aligned} \tag{4.49}$$

Example 4.12. If $\boldsymbol{Z} \sim$ pEPPF(θ, θ_0, σ, σ_0), for $\theta > \sigma$, $\theta_0 > \sigma_0$ and σ, $\sigma_0 \in (0,1)$ and $j = r$ div d, we have

$$\begin{aligned}\mathbb{P}\left(Z_r = h, W_r = t \,\middle|\, \boldsymbol{\ell}^{-r}, \boldsymbol{q}^{-r}\right) &= \\ &= \mathbb{1}_{\{\ell_{\cdot h}^{-r}=0\}} \frac{\theta_0 + H^{-r}\sigma_0}{\theta_0 + |\boldsymbol{\ell}^{-r}|} \frac{\theta + \ell_{j\cdot}^{-r}\sigma}{\theta + N_j - 1} + \\ &\quad + \mathbb{1}_{\{\ell_{\cdot h}^{-r}\neq 0\}} \left\{ \frac{\ell_{\cdot h}^{-r} - \sigma_0}{\theta_0 + |\boldsymbol{\ell}^{-r}|} \frac{\theta + \ell_{j\cdot}^{-r}\sigma}{\theta + N_j - 1} + \frac{q_{jht}^{-r} - \sigma}{\theta + N_j - 1} \right\}\end{aligned} \tag{4.50}$$

Remark 4.13. Notice that predictive probabilities in Examples 4.9, 4.11 and 4.12, if $j = r$ div d, depend on $\boldsymbol{q}$ just through the matrix $\boldsymbol{q}_j$ of numbers of customers eating each dish at each table in restaurant j, respectively. Nonetheless, even using those priors, in Algorithm 1, removing a node, one has to update the whole array $\boldsymbol{q}$, since if one removes a singleton and relabels dishes and tables, all the indices needs to be shifted.

4.4.2 Joint prediction

As mentioned in previous Sections, PEx-SBMs inherit a consistency property from prior distributions based on pEPPFs which guarantee a validated prediction for new nodes incoming into the network. In particular, it is natural to be interested in the joint conditional law of the allocation Z_{N+1} and the connections $\boldsymbol{y}_{N+1} = (y_{N+1}^{(i)})_{i=1}^{N+1}$ of a new node, given the connections observed between the N old nodes in $\boldsymbol{Y}$, as well as, given the consistency, obtain such law integrating out the posterior allocations of the old nodes from the joint posterior distribution of all allocations and new connections. In symbols,

Algorithm 1 Gibbs sampler for PEx-SBM

Initialize coherently $\boldsymbol{Z}$ and $\boldsymbol{W}$ ▷ Ex: N different tables and N different dishes
for $s \in [n_{\text{iter}}]$ **do**
  for $r \in [N]$ **do**
    $j \leftarrow r \text{ div } d$
    remove node r:
      reorder labels to get $(\boldsymbol{Z}^{-r},\, \boldsymbol{W}^{-r})_{s-1}$
      get $\boldsymbol{m}^{-r}$, $\overline{\boldsymbol{m}}^{-r}$ and $\boldsymbol{v}_r$, $\overline{\boldsymbol{v}}_r$
      get $\boldsymbol{\ell}^{-r}$ and $\boldsymbol{q}^{-r}$
    for $(h, t) \in [H^{-r} + 1] \times [\ell_{j\cdot}^{(-r)} + 1]$ **do**
      compute $\dfrac{p\left(\boldsymbol{Y} \mid Z_r = h,\, \boldsymbol{Z}^{-r}\right)}{p\left(\boldsymbol{Y}^{-r} \mid \boldsymbol{Z}^{-r}\right)}$
      compute $\mathbb{P}\left(Z_r = h, W_r = t \,\middle|\, \boldsymbol{\ell}^{-r}, \boldsymbol{q}^{-r}\right)$
    end for
    sample jointly $(Z_r,\, W_r)$
    obtain $\boldsymbol{m}$, $\boldsymbol{\ell}$, $\boldsymbol{q}$
  end for
  reorder labels
  record $(\boldsymbol{Z}, \boldsymbol{W})_s$
end for

$$\mathbb{P}\left(\boldsymbol{y}_{N+1} = c,\, Z_{N+1} = h,\, W_{N+1} = t\,|\,\boldsymbol{Y}\,\right) = \\ = \int \mathbb{P}\left(\boldsymbol{y}_{N+1} = c,\, Z_{N+1} = h,\, W_{N+1} = t\,|\,\boldsymbol{Y},\, \boldsymbol{Z},\, \boldsymbol{W}\,\right) \mathscr{L}_{\boldsymbol{Y}}(\mathrm{d}\boldsymbol{Z},\, \mathrm{d}\boldsymbol{W}) \tag{4.51}$$

where $c \in \{0,1\}^{N+1}$ and $\mathscr{L}_{\boldsymbol{Y}}$ is the conditional law of $(\boldsymbol{Z},\, \boldsymbol{W})$ given $\boldsymbol{Y}$, for any $(h,\, t) \in [H+1] \times [\ell_{d\cdot} + 1]$. Notice that the new node is assumed to be in the d-th layer just for a matter of notation and the procedure can be straightforwardly applied to new nodes lying in any of the d layers.

Let us suppose of having performed inference on the allocations of the old nodes and let us consider the sequence $\{(\boldsymbol{z}, \boldsymbol{w})_s\}_{s=1}^{n}$ of outputs of the last n iterations of the Gibbs sampler described in Algorithm 1. Then we can approximate the integral in (4.51) as follows.

$$\mathbb{P}\left(\boldsymbol{y}_{N+1} = c,\, Z_{N+1} = h,\, W_{N+1} = t\,|\,\boldsymbol{Y}\,\right) \simeq \\ \simeq \frac{1}{n}\sum_{s=1}^{n} \mathbb{P}\left(\boldsymbol{y}_{N+1} = c,\, Z_{N+1} = h,\, W_{N+1} = t\,|\,\boldsymbol{Y}, (\boldsymbol{z}, \boldsymbol{w})_s\,\right) \tag{4.52}$$

As in (4.47), we have

$$\mathbb{P}\left(\boldsymbol{y}_{N+1} = c,\, Z_{N+1} = h,\, W_{N+1} = t\,|\,\boldsymbol{Y}, (\boldsymbol{z}, \boldsymbol{w})_s\,\right) = \\ = \mathbb{P}\left(Z_{N+1} = h,\, W_{N+1} = t\,|(\boldsymbol{z}, \boldsymbol{w})_s\,\right) \times \\ \times \prod_{k=1}^{H^{(s)}} \frac{B\left(a + m_{hk}^{(s)} + \sum_{i \in \mathcal{I}_k} c_i, b + \overline{m}_{hk}^{(s)} + \sum_{i \in \mathcal{I}_k} (1 - c_i)\right)}{B(a + m_{hk}^{(s)}, b + \overline{m}_{hk}^{(s)})} \tag{4.53}$$

for

$$\mathcal{I}_k := \{i \in [N] \,:\, z_i = k\} \tag{4.54}$$

where $H^{(s)}$, $\boldsymbol{m}^{(s)}$, $\overline{\boldsymbol{m}}^{(s)}$ denote the number of occupied clusters, the number of edges and non-edges between clusters referring to the allocations given by the s-th MCMC sample. Hence, by combining (4.52) and (4.53), we can fully determine the probability distribution on the product space

$$\left[\max_{s=1,\dots,n} H^{(s)} + 1\right] \times \left[\max_{s=1,\dots,n} \ell_{d\cdot}^{(s)} + 1\right] \times \{0,\, 1\}^{N+1} \tag{4.55}$$

of joint allocations and connections of the new node. However, since the cardinality of this space grows exponentially with the number of nodes, it is clear that this route is infeasible for large networks. Therefore, we devise a second Gibbs sampler to draw from such joint predictive distribution. To achieve this goal, we need the blocked conditional distribution

$$\begin{aligned}\mathbb{P}\left(Z_{N+1}=h,\, W_{N+1}=t\,|\,\boldsymbol{Y},\boldsymbol{y}_{N+1}=c\right) &= \\ &= \frac{\mathbb{P}\left(\boldsymbol{y}_{N+1}=c,\, Z_{N+1}=h,\, W_{N+1}=t\,|\,\boldsymbol{Y}\right)}{\mathbb{P}\left(\boldsymbol{y}_{N+1}=c\,|\,\boldsymbol{Y}\right)} \propto (4.52)\end{aligned} \tag{4.56}$$

and the full conditionals for the connections vector, given by

$$\begin{aligned}\mathbb{P}\left(y_{N+1}^{(i)}=1\,\middle|\,\boldsymbol{Y},\, \boldsymbol{y}_{N+1}^{(-i)}=c^{(-i)},\, Z_{N+1}=h,\, W_{N+1}=t\right) &= \\ &= \int \mathbb{P}\left(y_{N+1}^{(i)}=1\,\middle|\,\boldsymbol{Y},\, \boldsymbol{y}_{N+1}^{(-i)}=c^{(-i)},\, Z_{N+1}=h,\, \boldsymbol{Z}\right) \mathscr{L}_{\boldsymbol{Y}}(\mathrm{d}\boldsymbol{Z})\end{aligned} \tag{4.57}$$

again because of connections being independent of the sub-allocations $\boldsymbol{W}$ given (all) the allocations $\boldsymbol{Z}$. Here, with an slight abuse of notation, we are denoting with $\mathscr{L}_{\boldsymbol{Y}}$ the posterior distribution of $\boldsymbol{Z}$. Now,

$$\begin{aligned}\mathbb{P}\left(y_{N+1}^{(i)}=1\,\middle|\,\boldsymbol{Y},\, \boldsymbol{y}_{N+1}^{(-i)}=c^{(-i)},\, Z_{N+1}=h,\, \boldsymbol{Z}\right) &= \\ &= \mathbb{E}\left[\xi_{hZ_i}\,\middle|\,\boldsymbol{Y},\, \boldsymbol{y}_{N+1}^{(-i)}=c^{(-i)},\, Z_{N+1}=h,\, \boldsymbol{Z}\right]\end{aligned} \tag{4.58}$$

and it is easy to see that, by conjugacy, posterior distributions of the connection probabilities are still beta laws. Finally, using Monte Carlo approximation for the integral in (4.57)

$$\begin{aligned}\mathbb{P}\left(y_{N+1}^{(i)}=1\,\middle|\,\boldsymbol{Y},\, \boldsymbol{y}_{N+1}^{(-i)}=c^{(-i)},\, Z_{N+1}=h\right) &\simeq \\ &\simeq \frac{1}{n}\sum_{s=1}^{n} \frac{a+m_{h,z_i^{(s)}}^{(s)}+\sum_{j\in\mathcal{I}_{z_i^{(s)}}\setminus\{i\}} c_j}{a+m_{h,z_i^{(s)}}^{(s)}+\sum_{j\in\mathcal{I}_{z_i^{(s)}}\setminus\{i\}} c_j+b+\overline{m}_{h,z_i^{(s)}}^{(s)}+\sum_{j\in\mathcal{I}_{z_i^{(s)}}\setminus\{i\}}(1-c_j)}\end{aligned} \tag{4.59}$$

where $\mathcal{I}_{z_i^{(s)}}$ follows the definition in (4.54) and we used the expression of the beta expectation. In words, the sums in (4.59) are counting ones and zeros in vector c corresponding to indices of nodes assigned to the same cluster as node i, disregarding the i-th entry itself. A scheme of the sampler is given in Algorithm 2, where n_{MC} and n_{GIBBS} denote respectively the number of posterior sample used for the Monte Carlo averages and the number of iteration of the Gibbs sampler with the joint predictive distribution in (4.51) as stationary one.

Notice that if we are in the situation in which we know the connections established by the new node joining the network and we just want to infer its allocation, then we don't need sampling. Indeed, this task boils down to evaluate

Algorithm 2 Joint prediction with PEx-SBM

Input: $\{(\boldsymbol{z}, \boldsymbol{w})_s\}_{s=1,\ldots,n_{\mathrm{MC}}}$
$j \leftarrow v$ div d
for $s \in [n_{\mathrm{MC}}]$ **do**
  get and store $\boldsymbol{m}^{(s)}$, $\overline{\boldsymbol{m}}^{(s)}$
  get $\boldsymbol{\ell}^{(s)}$, $\boldsymbol{q}^{(s)}$
  for $(h, t) \in [H^{(s)} + 1] \times [\ell_{d\cdot}^{(s)} + 1]$ **do**
    compute $\mathbb{P}\left(Z_{N+1} = h,\, W_{N+1} = t \,|(\boldsymbol{z}, \boldsymbol{w})_s\right)$
  end for
end for
Initialize $(Z_{N+1},\, W_{N+1})$ and $\boldsymbol{y}_{N+1}$
for $v \in [n_{\mathrm{GIBBS}}]$ **do**
  for $s \in [n_{\mathrm{MC}}]$ **do**
    compute $\sum_{i \in \mathcal{I}_k} y_{N+1}^{(i)}$ and $\sum_{i \in \mathcal{I}_k} (1 - y_{N+1}^{(i)})$ for any $k \in [H^{(s)}]$
    compute likelihood in (4.53)
  end for
  compute MC average in (4.52)
  sample $(Z_{N+1},\, W_{N+1})$
  for $i \in [N]$ **do**
    for $s \in [n_{\mathrm{MC}}]$ **do**
      compute $\sum_{j \in \mathcal{I}_{z_i^{(s)}} \setminus \{i\}} y_{N+1}^{(j)}$ and $\sum_{j \in \mathcal{I}_{z_i^{(s)}} \setminus \{i\}} (1 - y_{N+1}^{(j)})$
      if Z_{N+1} is new **then**
        adjust $\boldsymbol{m}^{(s)}$ and $\overline{\boldsymbol{m}}^{(s)}$
      end if
      compute expectation in (4.58)
    end for
    compute MC average in (4.59)
    sample $y_{N+1}^{(i)}$
  end for
  record $(Z_{N+1},\, y_{N+1})_v$
end for

$$\mathbb{P}\left(Z_{N+1}=h,\, W_{N+1}=t \,|(\boldsymbol{z},\boldsymbol{w})_s\right) \times \\ \times \prod_{k=1}^{H^{(s)}} \frac{B\left(a+m_{hk}^{(s)}+\sum_{i\in\mathcal{I}_k} c_i, b+\overline{m}_{hk}^{(s)}+\sum_{i\in\mathcal{I}_k}(1-c_i)\right)}{B(a+m_{hk}^{(s)}, b+\overline{m}_{hk}^{(s)})} \tag{4.60}$$

for any $(h,t) \in \left[H^{(s)}+1\right] \times \left[\ell_{d\cdot}^{(s)}+1\right]$ and any sample s, take the average as in (4.52) and normalize it. In this way, we obtain a vector of allocation probabilities for the new node as a functional of the posterior distribution of the allocation of all the other nodes. A point estimate for such a situation is proposed in Legramanti et al. (2022) by means of a plug-in estimator.

4.5 Application

In Examples 4.1 and 4.2 we presented some real-world cases where a model for clustering nodes of a network divided in different layers, taking into account both such division and the connectivity patterns between the nodes. In this Section we apply PEx-SBM both to simulated data and to a real criminal dataset, named *Infinito* network, which represent the principal application also in Legramanti et al. (2022). Summaries of the posterior distribution from the MCMC samples we obtain as described in Section 4.4, such as point estimates and credible regions, are retrieved using *variation of information* (VI) based algorithms presented in Wade and Ghahramani (2018). Briefly, the VI metric, introduced in Meilă (2007), measures the discrepancy between partitions evaluating their entropies, relatively to their shared information. In optimization terms, a point estimate $\hat{\boldsymbol{z}}$ is obtained minimizing the posterior expected VI distance from a posterior drawn, that is

$$\hat{\boldsymbol{z}} = \arg\min_{z} \mathbb{E}\left[\,\mathrm{VI}(\boldsymbol{Z},\boldsymbol{z})\,|\boldsymbol{Y}\,\right] \tag{4.61}$$

with $\boldsymbol{z}$ varying in the set of all partitions of $[N]$.

4.5.1 Simulation scenarios

In order to establish the clustering performances of our model, we apply it to synthetic networks, where a 'true' partition driving the connections exists. In particular, inspired by real-data application we present in Section 4.5.2, we built a criminal-like scenario where the true clusters are *intra*-layer, *inter*-layer and also exactly coincident with layers. We also cared of considering both layers and clusters of various numerosities. A realization of the adjacency matrix of such network is plotted in the left part of Figure 4.5. The network is composed as follows. We have $d = 4$ layers with $N = N_1 + N_2 + N_3 + N_4 = 30 + 30 + 15 + 5$, the first 3 of which share the same structure: $2/3$ of the nodes of each layer are in a *intra*-layer cluster,

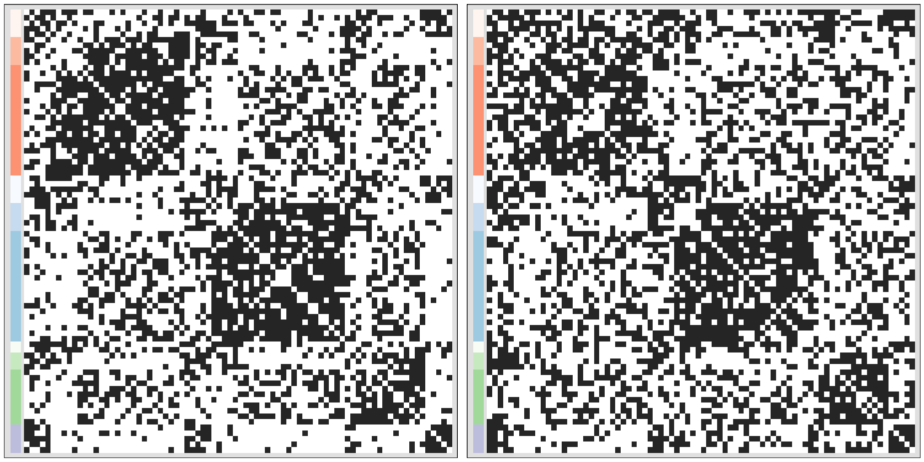


Figure 4.5: Criminal-like simulation scenarios: plain and noisy

1/6 is in an another *intra*-layer one, while the remaining 1/6 is in a common *inter*-layer cluster; finally, layer 4 coincide with a cluster. Hence the true number of clusters is $H = 8$. In Figure 4.5, shades of the same color denotes the division in layer (red, blue, green, purple), while the particular shade is a cluster: the lighter, almost white shade of red green and blue indicates the *inter*-layer cluster. In the criminal metaphor, layers are subgroups of a criminal organization, *e.g.* with a different geographic influence area, while clusters are actual collaborative groups in criminal activities, the former being known, while the latter being the object of our inference. We designed the connection probabilities considering the large *intra*-layer cluster in each layer as *operatives*, the small one as *local supervisors*, the *inter*-layer one as *shared supervisors* and the layer-coincident one as *bosses*. A heatmap of the matrix of connection probabilities is plotted in the left half of Figure 4.6, where clusters are ordered as follows.

Cluster	
1	red operatives
2	blue operatives
3	green operatives
4	red supervisors
5	blue supervisors
6	green supervisors
7	shared supervisors
8	bosses

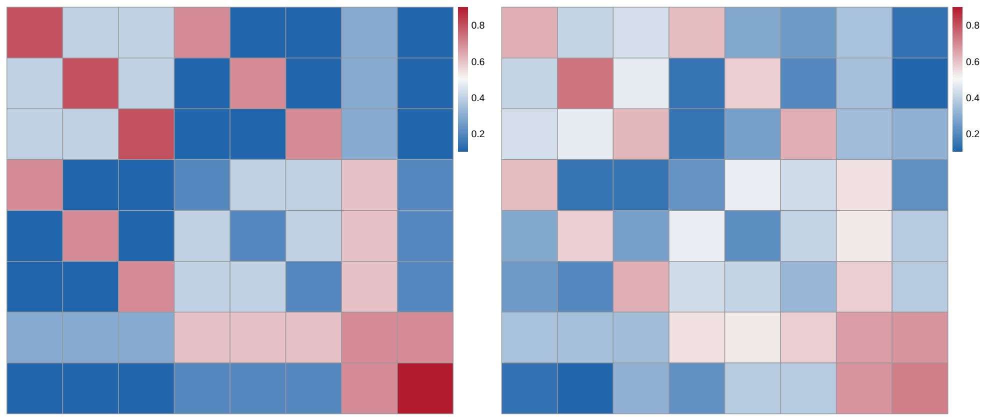


Figure 4.6: Connection probabilities for criminal-like simulation scenario: plain and noisy

This choice favours a hierarchical structure for the communication chain, which is a reasonable assumption for a criminal network: *e.g.*, operatives are probably in contact with their fellows and their local supervisors, not so much with operatives in other layers or shared supervisors, at all with the bosses, who are very connected among them and communicate directly just with shared supervisors. Given the connection probabilities and the described partition, we generated edges via independent Bernoulli trials, as in a SBM. We applied PEx-SBM posterior sampling algorithm to this network, choosing a H-DP directed prior on the partition. A point estimate of the posterior partition distribution, obtained via the VI-based algorithms in Wade and Ghahramani (2018) from posterior samples, is plotted in Figure 4.7. As in Section 4.3.3, the shape of the nodes indicates their layer, while the color represents the inferred clustering, matching cluster colors in Figure 4.5. Such point estimate coincides with the true partition.

We also considered noisy versions of such synthetic networks: we perturbed connection probabilities towards $1/2$, by drawing for each entry from a beta distribution re-scaled between the actual value and $1/2$, so that the beta parameters control the amount of noise. Such perturbation results in a general *whitening effect*, appreciable in Figure 4.6. Also, in a realization of the adjacency matrix plotted in the right half of Figure 4.5 we see that the connection patterns are significantly less clear, above all for small clusters. A point estimate of the posterior clustering for a realization of such noisy network is given in Figure 4.8: again this summary coincides with the true partition. The effect of the perturbation is also visible in the representation of the network itself in Figure 4.8, which employs the algorithm described in Fruchterman and Reingold (1991), based on equilibrium states of a system in which nodes are electrically charged and edges act like springs. For example,

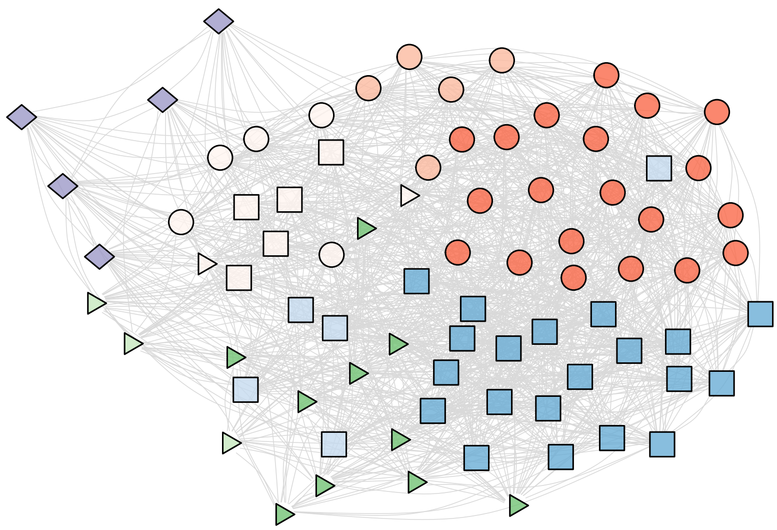


Figure 4.7: Criminal-like simulation scenario: point estimate

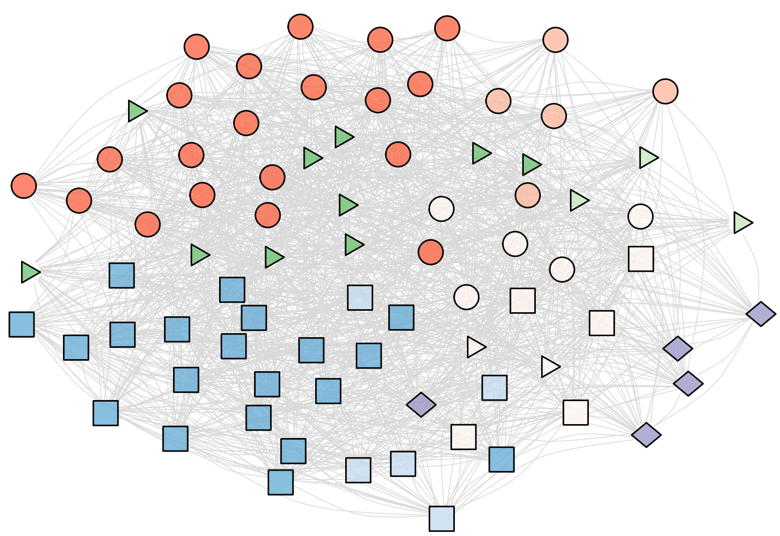


Figure 4.8: Noisy criminal-like simulation scenario: point estimate

	$\mathbb{E}\,[\,\mathrm{VI}(\boldsymbol{Z},\boldsymbol{Z}_0)\,\vert\,\boldsymbol{Y}\,]$	$\mathrm{VI}(\boldsymbol{Z},\boldsymbol{Z}_b)$
HDP-SBM	**0.026198**	**0.130944**
ESBM	0.067406	0.230043

Table 4.1: Criminal-like simulation scenario: performance comparison

as an effect of the noise, one of the bosses, which are the more sparsely connected nodes, creates enough edges to reach the middle of the network. Nonetheless, the point estimate places it in the right cluster.

Moreover, we compared the performances of PEx-SBM with the supervised version of the ESBM described in Legramanti et al. (2022). As mentioned in Section 4.3.1, in this model it is possible to enforce the layer division information via a product partition prior with covariates on the random partition, as in Müller et al. (2011). In particular we employed a DP prior with a Dirichlet-multinomial scheme for enforcing categorical covariates representing the division in layers. In Tables 4.1 and 4.2 we compare results, for both the plain and noisy case, in terms of posterior expected VI distance from the true clustering and VI radius of 95% credible balls, that is VI distance between the point estimate and a partition on the boundary of the credible ball. See Wade and Ghahramani (2018) for details on such validation quantities. Furthermore, we performed leave-one-out prediction on the criminal-like simulation scenario for the allocation of every node in the network, using the strategy described in conclusion of Section 4.4.2. We report some results on uncertainty quantification of the prediction. For node 11, that is a red operative, *i.e.* in cluster 1, the posterior allocation probabilities are the following

$$\hat{p}_{11} = (\mathbf{0.9213}, 0.0008, 0, 0.0017, 0.0762, 0, 0, 0, 0) \tag{4.62}$$

where the last entry is the probability that node 11 creates a new cluster, not yet seen in all the others. For node 77, that is a boss, *i.e.* in cluster 8, we have

$$\hat{p}_{77} = (0.0012, 0.0152, 0.0017, 0.0131, 0.1189, 0.0021, 0, \mathbf{0.7100}, 0.1377) \tag{4.63}$$

In general we compared the true clustering of the full network with the estimate obtained as follows. We consider the VI point estimate obtained with HDP-

	$\mathbb{E}\,[\,\mathrm{VI}(\boldsymbol{Z},\boldsymbol{Z}_0)\,\vert\,\boldsymbol{Y}\,]$	$\mathrm{VI}(\boldsymbol{Z},\boldsymbol{Z}_b)$
HDP-SBM	**0.368974**	**0.705689**
ESBM	0.523805	0.781609

Table 4.2: Noisy criminal-like simulation scenario: performance comparison

SBM of the clustering for the leave-one-out network and we complete it with the most probable predictive allocation for the left-out node. Doing this for nodes $1, 6, 11, 31, 36, 41, 61, 63, 66, 76$, which are representative of all the others (there is one node for any available choice of levels for the couple layer-cluster), we obtain perfect match between the estimate and the truth in any case.

4.5.2 Infinito network

The real-world data we analyze for illustrating the clustering performances of our model are derived by an Italian law enforcement operation, named *Operazione Infinito*, aiming to disrupt the Lombardy branch of *'Ndrangheta*, a criminal organization with roots in Calabria, but known to be operating all over Italy, with worldwide connections. Data, together with the pre-trial detention order triggered by the *Giudice per le indagini preliminari di Milano* from which they are extrapolated, are available at https://sites.google.com/site/ucinetsoftware/datasets/covert-networks. For a thorough account on the raw data see Legramanti et al. (2022). We used the clean data available at https://github.com/danieledurante/ESBM/tree/master/Application, which are structured as follows. We have 84 registered criminals, each with an affiliation to a geographical sub-group of the organization, called *locale*. The division in *locali* will be the division in layers in our model and there are 5 of them. An undirected connection between two criminals exists whenever they co-participated in a meeting, up to law enforcement knowledge. As an additive information, which we are not enforcing into the model, criminals are also classified as *bosses* or *affiliates* of their own *locale*, again according to law enforcement judgement. We applied PEx-SBM to this node-colored network, with a H-DP driven prior on the partition. For the shown results, hyperparameters are set to $\theta = \theta_0 = 5$. The resulting point estimate for the posterior clustering is showed in Figure 4.9. The estimate is again obtained by means of VI algorithms acting on the posterior samples given by Algorithm 1. In the representation, again the shape of the nodes is giving the layer they belong to, that is the *locale*, while the color indicates the cluster according to the posterior inference. Moreover, large nodes are those classified as bosses by law enforcement.

We report here some remarks on the posterior clustering. There are several characteristics of the Infinito network which are remarkably grasped by PEx-SBM inference. Indeed, with the full reports of the investigations in hand, one can observe that a number of dynamics of the network are well depicted by the point estimate showed in Figure 4.9.

At first, we notice that, even if such information is not included in the model, the PEx-SBM clustering manages to disentangle core-periphery structures given by boss-affiliate dynamics. For 3 of the *locali* (circles, triangles and rectangles) affiliates and bosses are clustered separately, suggesting that in those *locali* the

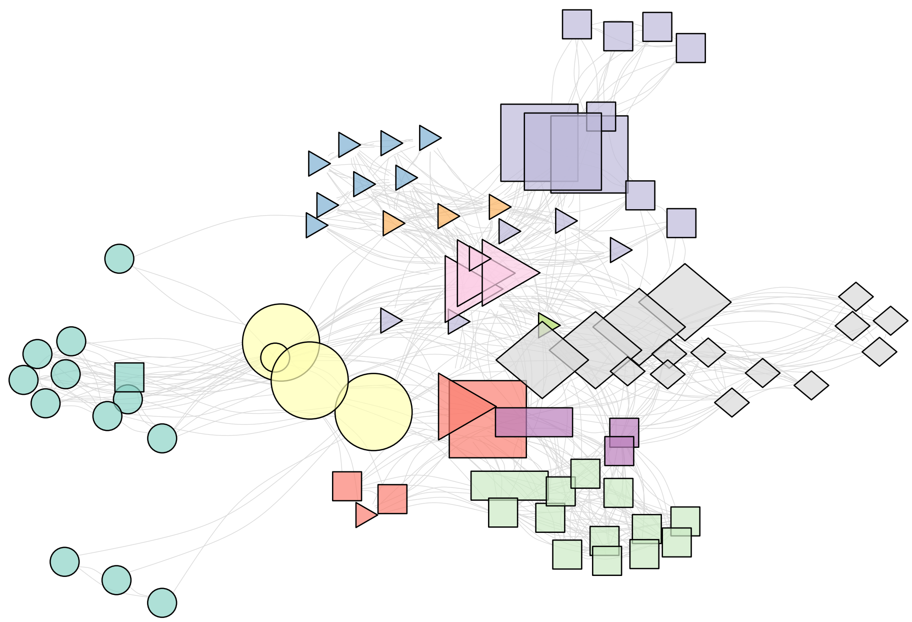


Figure 4.9: Infinito network: point estimate

connectivity patterns of these two kinds of nodes are different, as may be expected. Also, often some affiliates are allocated in a bosses cluster of his *locale*: this suggests that, since the communicative behaviour of such nodes is so similar to the bosses' one, they may have been misclassified by law enforcement or in any case they should be considered very close to the high part of the hierarchy of their *locale*.

Inter-layers clusters are inferred, as desired. They should reflect collaborative schemes between criminals in different *locali*. In particular the cluster of circle affiliates includes also a square affiliate: such allocation encompasses the fact that, as can be found in the reports, the square affiliate was attempting to create a new *locale*, and looking for affiliates in a different influence area from his original one.

The reports often refer to the murder of one of the triangles *locale* bosses. The consequences in the connectivity behaviour of such disruptive event can be read in our posterior clustering. The triangles one is the most divided layer. Some of its affiliates are allocated in the cluster of the squares *locale*. Moreover, even if a bosses cluster is present (pink), there is also a small cluster of affiliates (orange) which seems to form a core-periphery structure with other affiliates (blue). Nodes in orange cluster happen to be more closely involved in the murder, according to reports.

Finally, the posterior clustering seems to be able to spot special nodes of the network. The only inferred singleton (green triangle) should have a unique pattern of connections and in turn a peculiar role in the criminal organization. Indeed, this is the case: the green triangle is the oldest node in the network, and, as can be found in the reports, their role is considered to be of coordination and communication with Calabria-based branches of the same macro-organization.

The application of the supervised version of ESBM to the Infinito network, with *locale* affiliation as covariate, results in a similar clustering point estimate, but in wider credible balls, as showed in Table 4.3.

	$\text{VI}(\boldsymbol{Z}, \boldsymbol{Z}_b)$
HDP-SBM	**0.457569**
ESBM	0.4903655

Table 4.3: Infinito network: performance comparison

4.6 Conclusions and further work

In this work we propose a new generation of models for clustering nodes in multilayer networks. In particular, looking at nodes as statistical units, we parallel their division in layers in node-colored networks and the division in sub-populations of individuals in partially exchangeable schemes. This analogy brings us to structure a model which enforces such distributional invariance on latent characteristic of the nodes, directing a partition conditionally on which we define a connectivity model. Having the goal of an across-layers clustering, we do so by employing hierarchical discrete random probability measures. As we underlined in previous Sections, partial exchangeability is a good fit for the specific kind of multilayer networks we focus on in this work, but not for others. We described edge-colored networks in Section 4.2 and we argued how the strong link existing between nodes in different layers is not encompassed by partial exchangeability. However, other invariance structures may be appropriate for other types of multilayer networks. In general, we believe this work to be the first step into a way of designing models for complex networks, by enclosing their structures in invariance hypotheses. In particular an extension in this direction is represented by employing *separate exchangeability* for edge-colored networks. This hypothesis allows the distribution of a matrix of observables to be invariant with respect to rigid permutations of rows and columns. In symbols, let $\boldsymbol{X} = (X_{ji})_{j=1,\dots,d}^{i=1,\dots,N}$ then

$$
\begin{aligned}
&(X_{11}, \dots, X_{1N}, \dots, X_{d1}, \dots, X_{dN}) \stackrel{d}{=} \\
&\qquad \stackrel{d}{=} (X_{\pi_1(1)\pi_2(1)}, \dots, X_{\pi_1(1)\pi_2(N)}, \dots, X_{\pi_1(d)\pi_2(1)}, \dots, X_{\pi_1(d)\pi_2(N)}) \quad (4.64)
\end{aligned}
$$

for any $\pi_1 \in S_d$ and $\pi_2 \in S_N$, the symmetric groups of $[d]$ and $[N]$ respectively. See Lin et al. (2021) for a recent review, a comparison with partial exchangeability and insightful modeling instances. It is clear that this kind of invariance is suitable for edge-colored networks: if we think of each entry of $\boldsymbol{X}$ as attached to a node, with different rows referring to different layers, while a column contains the copies of the same node across layers, if we just allow for rigid permutations we are keeping track of the link between copies, as well as assuming exchangeability for the colors of the edges.

Another fundamental work direction is the implementation of algorithms for the m-step-ahead prediction with PEx-SBM, that is the joint modeling of allocations and connections of m unobserved new nodes joining the network, given the connections of the old nodes. In this task, we expect the Kolmogorov consistency of PEx-SBM to be a real game-changer in terms of predictive power, as we can fully leverage the intrinsic predictive structure of the model and the available urn scheme of the hierarchical prior.

Acknowledgments

I thank my advisors Antonio Lijoi and Igor Prünster, for directing my research work and my early steps in the academic world: together, they convey the perfectly balanced mixture of guidance and fostering of independence. I thank my Master supervisor Domenico Marinucci because, among other merits, he predicted that the Ph.D. would have been one of the best period of my life. Hence, I cannot thank enough the friends who made that prediction accurate: Filippo for being unreasonably kind and inspirational; Tommaso and Marta for their priceless guidance; Gregor for stepping in at the right moment; Niccolò because we found the right equilibrium; Michał for his open-hearted affection; Giovanni, Beatrice, and Laura mainly for the dancing. The Statistics group at our Department evolved in an exciting environment where research thrives: it has been a privilege to be a part of it. I thank my lifelong friends Paolo, Luca, Fabrizio, Cristiano, and Emanuele, for making friendship effortless. I thank my parents and brother for still being overly proud of me. Clarissa is beyond any acknowledgment: everything I have, we made it together.

4. Acknowledgments

Bibliography

Barrios, E., Lijoi, A., Nieto-Barajas, L. E., and Prünster, I. Modeling with normalized random measure mixture models. *Stat. Sci.*, 28(3):313 – 334, 2013.

Beraha, M., Guglielmi, A., and Quintana, F. A. The semi-hierarchical Dirichlet process and its application to clustering homogeneous distributions. *Bayesian Analysis*, 16(4):1187 – 1219, 2021.

Boccaletti, S., Bianconi, G., Criado, R., del Genio, C., Gómez-Gardeñes, J., Romance, M., Sendiña-Nadal, I., Wang, Z., and Zanin, M. The structure and dynamics of multilayer networks. *Phys. Rep.*, 544(1):1–122, 2014.

Brix, A. Generalized gamma measures and shot-noise Cox processes. *Adv. Appl. Probab.*, 31(4):929–953, 1999.

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. Distribution theory for hierarchical processes. *Ann. Stat.*, 47(1):67 – 92, 2019.

Catalano, M., Lijoi, A., and Prünster, I. Measuring dependence in the Wasserstein distance for Bayesian nonparametric models. *Ann. Stat.*, 49(5):2916–2947, 2021.

Cifarelli, D. M. and Regazzini, E. Distribution functions of means of a Dirichlet process. *Ann. Stat.*, pages 429–442, 1990.

Cifarelli, D. M. and Regazzini, E. Some remarks on the distribution functions of means of a Dirichlet process. *Technical Report of IMATI–CNR, Milano*, 93.4, 1993.

Daley, D. and Vere-Jones, D. *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Probability and Its Applications. Springer New York, 2007.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. Are gibbs-type priors the most natural generalization of the dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):212–229, 2015.

de Finetti, B. La prévision : ses lois logiques, ses sources subjectives. *Ann. Henri Poincaré*, (1):1–68, 1937.

de Finetti, B. Sur la condition d'équivalence partielle. *Actual. Sci. Ind.*, (739): 5–18, 1938.

Diaconis, P. and Kemperman, J. Some new tools for Dirichlet priors. In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics*, volume 5, pages 97–106. Oxford University Press, 1996.

Doksum, K. Tailfree and neutral random probabilities and their posterior distributions. *Ann. Prob.*, 2(2):183–201, 1974.

Estrada, R. and Kanwal, R. P. *Singular integral equations.* Birkhäuser Boston, 2012.

Evans, M. R. and Hanney, T. Nonequilibrium statistical mechanics of the zero-range process and related models. *J. Phys. A Math. Theor.*, 38(19):R195–R240, apr 2005.

Fabius, J. Neutrality and Dirichlet distributions. In *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (Tech. Univ. Prague, Prague, 1971; dedicated to the memory of Antonín Špaček)*, pages 175–181. 1973.

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

Ferguson, T. S. Prior distributions on spaces of probability measures. *Ann. Stat.*, 2:615–629, 1974.

Fruchterman, T. M. J. and Reingold, E. M. Graph drawing by force-directed placement. *Softw. - Pract. Exp.*, 21(11):1129–1164, 1991.

Geng, J., Bhattacharya, A., and Pati, D. Probabilistic community detection with unknown number of communities. *J. Am. Stat. Assoc.*, 114(526):893–905, 2019.

Ghosal, S. and van der Vaart, A. *Fundamentals of nonparametric Bayesian inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, 2017.

Girvan, M. and Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.*, 99(12):7821–7826, 2002.

Gnedin, A. and Pitman, J. Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci.*, 138, 01 2005. doi: 10.1007/s10958-006-0335-z.

Guglielmi, A. A simple procedure calculating the generalized Stieltjes transform of the mean of a Dirichlet process. *Statist. Probab. Lett.*, 38(4):299–303, 1998.

Gurland, J. Inversion formulae for the distribution of ratios. *Ann. Math. Stat.*, 19 (2):228 – 237, 1948.

Holland, P. W., Laskey, K. B., and Leinhardt, S. Stochastic blockmodels: First steps. *Soc. Networks*, 5(2):109–137, 1983.

Ishwaran, H. and James, L. F. Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc.*, 96(453):161–173, 2001.

James, L. F., Lijoi, A., and Prünster, I. Distributions of linear functionals of two parameter Poisson–Dirichlet random measures. *Ann. Appl. Probab.*, 18(2): 521–551, 2008.

James, L. F., Lijoi, A., and Prünster, I. Posterior analysis for normalized random measures with independent increments. *Scand. J. Stat.*, 36(1):76–97, 2009.

James, L. F., Lijoi, A., and Prünster, I. On the posterior distribution of classes of random means. *Bernoulli*, 16(1):155–180, 2010.

Karp, D. and Prilepkina, E. Generalized Stieltjes functions and their exact order. *J. Class. Anal.*, 1(1):53–74, 2012.

Kemp, C., Tenenbaum, J., Griffiths, T., Yamada, T., and Ueda, N. Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 381–388, 2006.

Kerov, S. V. Transition probabilities for continual Young diagrams and Markov moment problem. *Funct. Anal. its Appl.*, 27(2):104–117, 1993.

Kerov, S. V. Interlacing measures. In *Kirillov's Seminar on Representation Theory*, volume 181 of *AMS Translation Series 2*, pages 35–83. American Mathematical Society, Providence, 1998.

Kerov, S. V. and Tsilevich, N. The Markov–Krein correspondence in several dimensions. *J. Math. Sci.*, 121(3):2345–2359, 2004.

Kessler, D., Hoff, P., and Dunson, D. Marginally specified priors for nonparametric Bayesian estimation. *J. R. Stat. Soc. Series B Stat. Methodol.*, 77:35–58, 2015.

Kingman, J. F. C. Completely random measures. *Pac. J. Math.*, 21(1):59 – 78, 1967.

Kingman, J. F. C. Random discrete distributions. *J. R. Stat. Soc. Series B Stat. Methodol.*, 37(1):1–22, 1975.

Kingman, J. F. C. The representation of partition structures. *J. Lond. Math. Soc.*, s2-18(2):374–380, 1978.

Kingman, J. F. C. The coalescent. *Stoch. Process. Their Appl.*, 13(3):235–248, 1982.

Kingman, J. F. C. *Poisson processes*. Oxford Studies in Probability. Oxford University Press, 1993. ISBN 0198536933.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. Multilayer networks. *J. Complex Netw.*, 2(3):203–271, 2014.

Lamperti, J. W. An occupation time theorem for a class of stochastic processes. *Trans. Am. Math. Soc.*, 88:380–387, 1958.

Legramanti, S., Rigon, T., Durante, D., and Dunson, D. B. Extended stochastic block models with application to criminal networks. *Ann. Appl. Stat*, In press, 2022.

Lévy, P. Sur certains processus stochastiques homogènes. *Compos. Math.*, 7:283–339, 1939.

Lijoi, A. and Prünster, I. Distributional properties of means of random probability measures. *Stat. Surv.*, 3:47–95, 2009.

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

Lijoi, A. and Prünster, I. A conversation with Eugenio Regazzini. *Stat. Sci.*, 26 (4):647 – 672, 2011.

Lijoi, A. and Regazzini, E. Means of a Dirichlet process and multiple hypergeometric functions. *Ann. Probab.*, 32(2):1469–1495, 2004.

Lijoi, A., Mena, R. H., and Prünster, I. Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Stat. Soc. Series B Stat. Methodol.*, 69 (4):715–740, 2007.

Lijoi, A., Nipoti, B., and Prünster, I. Bayesian inference with dependent normalized completely random measures. *Bernoulli*, 20(3):1260–1291, 2014.

Lijoi, A., Prünster, I., and Rebaudo, G. Flexible clustering via hidden hierarchical Dirichlet priors. *Scand. J. Stat.*, Forthcoming, 2022.

Lin, Q., Rebaudo, G., and Müller, P. Separate exchangeability as modeling principle in Bayesian nonparametrics. *https://arxiv.org/abs/2112.07755*, 2021.

Lo, A. Y. On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Stat.*, 12(1):351 – 357, 1984.

MacEachern, S. Dependent nonparametric processes. In *ASA 1999 Proceedings of the Section on Bayesian Statistics*. 1999.

MacEachern, S. Dependent Dirichlet processes. Technical report, Ohio State University, 2000.

Martin, P. and Rizzo, F. Hypersingular integrals: how smooth must the density be? *Int. J. Numer. Methods Eng.*, 39:687–704, 1996.

Meilă, M. Comparing clusterings—an information based distance. *J. Multivar. Anal.*, 98(5):873–895, 2007.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. *Bayesian nonparametric data analysis*. Springer Series in Statistics. Springer, Cham, 2015.

Müller, P., Quintana, F., and Rosner, G. A method for combining inference across related nonparametric Bayesian models. *J. R. Stat. Soc. Series B Stat. Methodol.*, 66(3):735–749, 2004.

Müller, P., Quintana, F., and Rosner, G. A product partition model with regression on covariates. *J. Comput. Graph. Stat.*, 20:260–278, 03 2011. doi: 10.1198/jcgs.2011.09066.

Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.*, 103(23):8577–8582, 2006.

Newman, M. E. J. and Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, Feb 2004.

Nieto-Barajas, L., Prünster, I., and Walker, S. Normalized random measures driven by increasing additive processes. *Ann. Stat.*, 32:2343–2360, 2004.

Nowicki, K. and Snijders, T. A. B. Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Assoc.*, 96(455):1077–1087, 2001.

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

Phadia, E. *Prior Processes and Their Applications: Nonparametric Bayesian Estimation*. Springer Series in Statistics. Springer International Publishing, 2016.

Pitman, J. Exchangeable and partially exchangeable random partitions. *Probab. Theory Relat. Fields*, 102(2):145–158, 1995.

Pitman, J. Some developments of the Blackwell–MacQueen urn scheme. In Ferguson, T., Shapley, J., and MacQueen, J., editors, *Lecture Notes - Monograph Series*, Statistics, probability and game theory: papers in honor of David Blackwell, pages 245–267. Institute of Mathematical Statistics, 1996.

Pitman, J. Poisson–Kingman partitions. In *Statistics and science: a Festschrift for Terry Speed*, pages 1–34. Institute of Mathematical Statistics, 2003.

Pitman, J. *Combinatorial stochastic processes: Ecole d'eté de probabilités de Saint-Flour XXXII-2002 (J. Picard ed.)*. Number 1875 in Lecture Notes in Mathematics. Springer, 2006.

Pitman, J. and Yor, M. On the relative lengths of excursions derived from a stable subordinator. In Azéma, J., Yor, M., and Émery, M., editors, *Séminaire de probabilités XXXI*, pages 287–305. Springer, Berlin, 1997a.

Pitman, J. and Yor, M. The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, pages 855–900, 1997b.

Pulkkinen, O. Boundary driven zero-range processes in random media. *J. Stat. Phys.*, 128(6):1289–1305, 2007.

Regazzini, E. and Sazonov, V. V. Approximation of distributions of random probabilities by mixtures of Dirichlet distributions with applications to nonparametric Bayesian statistical inferences. *Teor. Veroyatnost. i Primenen.*, 45(1):103–124, 2000.

Regazzini, E., Guglielmi, A., and Nunno, G. D. Theory and numerical analysis for exact distributions of functionals of a Dirichlet process. *Ann. Stat.*, 30(5):1376 – 1411, 2002.

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

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. The nested Dirichlet process. *J. Am. Stat. Assoc.*, 103(483):1131–1154, 2008.

Romik, D. Explicit formulas for hook walks on continual Young diagrams. *Adv. Appl. Math.*, 32(4):625–654, 2004.

Schmidt, M. N. and Morup, M. Nonparametric Bayesian modeling of complex networks: an introduction. *IEEE Signal Process. Mag.*, 30(3):110–128, 2013.

Schwarz, J. H. The generalized Stieltjes transform and its inverse. *J. Math. Phys.*, 46, 2005.

Sethuraman, J. A constructive definition of Dirichlet priors. *Statist. Sinica*, 4(2): 639–650, 1994.

Sumner, D. B. An inversion formula for the generalized Stieltjes transform. *Bull. Am. Math. Soc.*, 55:174–183, 1949.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.*, 101(476):1566–1581, 2006.

Tsilevich, N. Distribution of the mean value for certain random measures. *J. Math. Sci.*, 96(5):3616–3623, 1999.

Vershik, A. M., Yor, M., and Tsilevich, N. On the Markov–Krein identity and quasi-invariance of the gamma process. *J. Math. Sci.*, 121(3):2303–2310, 2004.

Von Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.*, 17:395–416, 01 2004. doi: 10.1007/s11222-007-9033-z.

Wade, S. and Ghahramani, Z. Bayesian cluster analysis: Point estimation and credible balls. *Bayesian Anal.*, 13(2):559 – 626, 2018.

Widder, D. V. *Laplace transform.* Princeton University Press, 2015.