Università Commerciale "Luigi Bocconi"
PhD School

PhD program in: Statistics and Computer Science
Cycle: XXXVI
Disciplinary Field: SECS-S/01

Contributions to dependent processes in Bayesian nonparametrics

Advisor: Igor Prünster
Co-Advisor: Antonio Lijoi

PhD Thesis by
Claudio Del Sole
ID number: 3147519

Year 2025

Abstract

Random measures represent the fundamental building blocks for defining flexible priors in Bayesian nonparametric models. Over the last three decades, there has been a widespread diffusion of proposals aimed at introducing dependence among different random measures, in order to properly account for various forms of heterogeneity in the observations while preserving the possibility to borrow information across them.

The first part of the thesis focuses on vectors of dependent random measures defined through hierarchical structures, arguably the most natural and popular strategy to specify nonparametric priors for partially exchangeable data. These prior processes induce a random nested partition structure on the observations, which is usually taken into account by introducing additional latent variables. In this work, we propose a unified approach to hierarchies of random measures, in which such latent variables are directly inserted into the generative model for the data; within this framework, we identify a common structure shared by different hierarchical constructions proposed in the literature, and highlight a key identity which plays a prominent role in the derivation of quantities of interest. Furthermore, we consider hierarchical completely random measures as mixing measures to define dependent mixture hazard rates, which are in turn employed in the context of survival analysis to model competing risks data. We derive analytical results for Bayesian inference and prediction, as well as efficient posterior sampling algorithms; the effectiveness of our proposal is tested on simulated and clinical datasets.

The second part of the thesis contributes to the Bayesian nonparametric regression framework, by introducing a collection of dependent random probability measures indexed by covariates, which enter the model specification through a multiplicative kernel structure. This construction induces a predictor-dependent random partition model, characterized by great flexibility and inherent consistency for new observations, while retaining some analytical tractability. A noteworthy example arises when the distribution of such random measures is a transformation of the distribution of a stable process; moreover, the structure of the posterior distribution implied by such specification suggests the introduction of a novel nonhomogeneous process, which extends the two-parameter Poisson-Dirichlet process and acts as quasi-conjugate prior for our proposal. The last chapter further develops this novel nonparametric process in the exchangeable setting, and characterizes its prior-posterior updating mechanism, as well as its predictive structure.

Contents

Abstract **i**

1 Bayesian nonparametric models **1**

1.1 Completely random measures 2

1.2 Nonparametric priors based on CRMs 4

1.3 Dirichlet and Pitman-Yor processes as normalized random measures 6

2 A Unified Approach to Hierarchical Random Measures **9**

2.1 Introduction 9

2.2 Hierarchical random measures 12

2.3 Random partitions and latent variables 13

2.4 Key identity for random measures 16

2.5 Bayesian nonparametric hierarchical models 18

2.6 Posterior characterizations 22

2.7 Generalized gamma CRMs as natural conjugate priors 25

2.8 Remarks on the dependence structure of hierarchical CRMs 26

3 Hierarchically dependent mixture hazards for modelling competing risks **31**

3.1 Competing risks in survival analysis 32

3.2 Modelling mixture hazard rates via random measures 35

3.3 Latent partition structure 37

3.4 Prediction curves 40

3.5 Posterior characterization and estimates 42

3.6 Gibbs sampling schemes 46

3.7 Numerical illustration and simulation study 54

3.8 Applications to clinical datasets 60

4 Kernel-weighted random measures and covariate-dependent partitions **65**

4.1 Bayesian nonparametric regression framework 66

4.2 Kernel-weighted normalized random measures 69

4.3 Marginal distribution and partition probability function 72

4.4 Latent variables and posterior characterization 77

4.5 Predictive distribution 80

4.6 Future developments 85

5 Nonhomogeneous Pitman-Yor process **88**

5.1 Construction of nonhomogeneous processes 88

5.2 Marginal distribution and partition function 91

5.3 Latent variable and posterior distribution 95

5.4 Predictive distribution 97

5.5 Future developments 98

A Proofs **102**

A.1 Proofs of Chapter 2 102

A.2 Proofs of Chapter 3 104

A.3 Proofs of Chapter 4 120

A.4 Proofs of Chapter 5 135

Bibliography **144**

Chapter 1

Bayesian nonparametric models

In a Bayesian framework, observations from a homogeneous population are usually modeled as an (infinitely extendable) exchangeable sequence, meaning that their distribution does not depend on their order of appearance (de Finetti, 1937). Formally, an infinite sequence of observations $(X_n)_{n\geq 1}$ taking values in a complete and separable metric space $\mathbb{X}$ is *exchangeable* if and only if, for each $n \geq 1$ and each finite permutation π of the indices $1, \ldots, n$ the probability distribution of the random vector $(X_1, \ldots, X_n)$ coincides with the probability distribution of the permuted random vector $(X_{\pi(1)}, \ldots, X_{\pi(n)})$. The paramount role played by exchangeability within the Bayesian approach is motivated by the celebrated de Finetti's representation theorem (de Finetti, 1937), which states that an exchangeable sequence can be equivalently represented as a conditionally independent and identically distributed (i.i.d.) sequence, given a random probability measure $\tilde{p}$, that is

$$\begin{aligned} X_1, \ldots, X_n \mid \tilde{p} &\overset{\text{i.i.d.}}{\sim} \tilde{p}, \qquad n \geq 1, \\ \tilde{p} &\sim \mathcal{Q}. \end{aligned} \tag{1.1}$$

The random probability measure $\tilde{p}$ takes values in the space $\mathcal{P}_{\mathbb{X}}$ of probability measures on $\mathbb{X}$, and is sometimes referred to as the *directing random measure* (Aldous, 1985), while its probability distribution $\mathcal{Q}$ is known as the *de Finetti measure* and acts as the prior distribution for Bayesian inference; if the support of $\mathcal{Q}$ is an infinite-dimensional subset of $\mathcal{P}_{\mathbb{X}}$, the resulting inferential problem is typically termed *nonparametric*, which is the setting considered in this work.

In light of de Finetti's result, the Bayesian modeling choice reduces to the identification of a suitable distribution $\mathcal{Q}$ for the random probability measure $\tilde{p}$, that is, the selection of a suitable prior. A probability measure may be characterized in many different ways: for example, through its probability mass or density function, cumulative distribution function or survival function, cumulative hazard, and hazard rate function. Each representation highlights different features of the distribution, and may be preferred to the others depending on the main target of the analysis: for this reason, different Bayesian nonparametric models have focused on each of these representations. In particular, the renowned Dirichlet process (Ferguson, 1973,

1974), which represents the cornerstone of Bayesian nonparametric modeling, may successfully characterize both random probability mass functions and random density functions, via mixtures with countinuous probability kernels (Lo, 1984). In the setting of survival analysis, random cumulative distribution functions can be modeled as neutral-to-the right processes (Doksum, 1974; Walker and Muliere, 1997), while models for random cumulative hazards and random hazard rate functions have been proposed in Hjort (1990) and Dykstra and Laud (1981). The seminal contributions mentioned above share one common feature: the random probability measure $\tilde{p}$ is modeled as a suitable transformation of a completely random measure. Indeed, completely random measures (Kingman, 1967) are a remarkable class of discrete random measures that lends itself to modeling both discrete functions, thanks to their almost-sure discrete nature, and continuous ones, typically through kernel smoothing; moreover, completely random measures feature an infinite number of random atoms and random weights, which guarantees full-modeling flexibility of many quantities of interest. A comprehensive review highlighting on the role of completely random measures as unifying concept in Bayesian nonparametrics is proposed in Lijoi and Prünster (2010), on which the rest of this chapter is widely based.

1.1 Completely random measures

Let $(\mathbb{X}, \mathcal{X})$ be a complete and separable metric space, and denote by $\mathbb{M}$ the space of boundedly finite measures on $(\mathbb{X}, \mathcal{X})$ equipped with the corresponding Borel σ-algebra $\mathcal{M}$, that is, the smallest σ-algebra that makes the projections $A \mapsto \mu(A)$ measurable for every measure μ and every bounded set A; see Daley and Vere-Jones (2007) for details. A random element $\tilde{\mu}$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in $(\mathbb{M}, \mathcal{M})$ is termed *random measure*.

Definition 1.1. (Kingman, 1967) A completely random measure (CRM) $\tilde{\mu}$ is a random measure on $(\mathbb{X}, \mathcal{X})$ such that, for any collection of $n \geq 1$ bounded and pairwise disjoint sets $A_1, \ldots, A_n \in \mathcal{X}$, the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_n)$ are mutually independent.

Completely random measures can be regarded as the natural extension to general Polish spaces of stochastic processes on the real line with non-negative and independent increments, as highlighted in (Kingman, 1967). In this work, we consider CRMs without deterministic components and without fixed points of discontinuity: a CRM belonging to this class has almost-surely discrete realizations, and can be represented as

$$\tilde{\mu}(dx) = \sum_{h \geq 1} S_h \, \delta_{X_h^*}(dx), \tag{1.2}$$

where $(S_h)_{h \geq 1}$ is a sequence of positive random jumps and $(X_h^*)_{h \geq 1}$ is a sequence of random locations (points of discontinuity) in $\mathbb{X}$. In fact, the almost-sure discreteness of their realizations

stems from their characterization as linear functionals of Poisson random measures,

$$\tilde{\mu}(dx) \stackrel{\mathrm{d}}{=} \int_{\mathbb{R}^+} s\, \tilde{N}(ds, dx), \tag{1.3}$$

where $\tilde{N}$ is a Poisson random measure on the product space $\mathbb{R}^+ \times \mathbb{X}$. The distribution of CRMs is characterized by their Laplace functional transform at any non-negative measurable function $f\colon \mathbb{X} \mapsto \mathbb{R}^+$, namely

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} f(x)\, \tilde{\mu}(dx)\right\}\right] = \exp\left\{-\int_{\mathbb{R}^+\times\mathbb{X}} (1 - e^{-s\, f(x)})\, \nu(ds, dx)\right\}, \tag{1.4}$$

where ν is the Lévy intensity measure uniquely identifying $\tilde{\mu}$; this characterization motivates the notation $\tilde{\mu} \sim \mathrm{CRM}(\nu)$. Note that ν is the mean intensity measure of the Poisson random measure $\tilde{N}$ in (1.3), and must satisfy the condition

$$\int_{\mathbb{R}^+\times\mathbb{X}} \min(s, 1)\, \nu(ds, dx) < \infty.$$

The Lévy intensity measure ν can be always disintegrated as

$$\nu(ds, dx) = \rho(ds|x)\, \alpha(dx), \tag{1.5}$$

where $\rho\colon \mathcal{B}(\mathbb{R}^+) \times \mathbb{X} \mapsto \mathbb{R}^+$ is a transition kernel characterizing the random jumps and α is a σ-finite measure on $(\mathbb{X}, \mathcal{X})$ characterizing the random locations in (1.2). In view of the discussion contained in Chapter 2, it is relevant to point out that, when α is a diffuse measure, such random locations are almost-surely distinct. Lévy intensities, and the corresponding CRMs, are termed *homogeneous* if $\rho(\cdot\,|x) = \rho(\cdot)$ is a measure not depending on $x \in \mathbb{X}$, and *nonhomogeneous* otherwise. An exhaustive account on CRMs can be found in Kingman (1993).

In the following, homogeneous CRMs are often considered for their simplicity and tractability; intuitively, this is equivalent to the independence between atoms and jumps of the corresponding random measure. Moreover, the measure α appearing in (1.5) is assumed to be finite, which ensures that the random measure is finite almost surely. This implies that the total mass of α can be included into the transition kernel ρ, and the Lévy intensity measure is uniquely disintegrated as

$$\nu(ds, dx) = \rho(ds)\, P_0(dx),$$

with P_0 a diffuse probability measure on $(\mathbb{X}, \mathcal{X})$, termed *base probability measure*; in the following, it will be sometimes useful to resort to this representation rather than on (1.5). The *infinite activity* property of the Lévy intensity measure is also assumed, namely $\rho(\mathbb{R}^+) = \infty$, which implies that the corresponding random measure has an infinite number of jumps on any bounded set, and thus ensures it is non-zero almost surely. Further details can be found in Regazzini et al. (2003).

The fundamental example of homogeneous CRM is represented by the *gamma* process, which corresponds to P_0 being a diffuse probability measure and

$$\rho(ds) = \theta\, s^{-1}\, e^{-\beta\, s}\, ds, \tag{1.6}$$

where $\theta > 0$ and $\beta > 0$ are positive parameters; from the characterization in (1.4), it follows that $\tilde{\mu}(A)$ has gamma distribution, for every A such that $\alpha(A) > 0$. The *generalized gamma* process (Brix, 1999) corresponds to the specification

$$\rho(ds) = \frac{\theta}{\Gamma(1-\sigma)}\, s^{-\sigma-1}\, e^{-\beta s}\, ds, \tag{1.7}$$

where $\sigma \in [0, 1)$, and includes, as notable special cases, the above-mentioned *gamma* process, obtained setting $\sigma = 0$, and the σ–*stable* process, characterized by the choice $\beta = 0$.

On the other hand, nonhomogeneous CRMs appearing in the literature are typically the result of a change of measure with respect to homogeneous CRMs; indeed, a nonhomogeneous completely random measure $\tilde{\mu}_g$ can be defined, for every set $A \in \mathcal{X}$, as

$$\tilde{\mu}_g(A) := \int_A g(x)\, \tilde{\mu}(dx), \tag{1.8}$$

where $\tilde{\mu}$ is a homogeneous CRM and $h\colon \mathbb{X} \mapsto \mathbb{R}^+$ a measurable non-negative function, playing the role of Radon-Nikodym derivative. Intuitively, the jumps of the homogeneous random measure $\tilde{\mu}$ are rescaled according to the value of function $g(\cdot)$ at the corresponding locations, thus introducing dependence between jumps and atoms. This construction has been considered by Dykstra and Laud (1981) to define the *extended gamma* process, characterized by

$$\rho(ds|x) = \theta\, s^{-1}\, e^{-\beta(x)\, s}\, ds, \tag{1.9}$$

where $\theta > 0$ and $\beta\colon \mathbb{X} \mapsto \mathbb{R}^+$ is a measurable non-negative function; the same specification is investigated by Lo (1982), where is termed *weighted gamma* process.

1.2 Nonparametric priors based on CRMs

In Bayesian nonparametrics, (completely) random measures represent a very natural and convenient choice of discrete random measures, and can be effectively exploited as the basic building block for the construction of nonparametric priors, as remarked in the introduction to this chapter. The present work considers random probability measures and random hazard rates obtained from two suitable transformations of random measures: normalizations and kernel mixtures.

Normalized random measures. An almost surely discrete random probability measure $\tilde{p}$ can be defined via normalization of a random measure $\tilde{\mu}$ as

$$\tilde{p}(dx) := \frac{\tilde{\mu}(dx)}{\tilde{\mu}(\mathbb{X})}, \tag{1.10}$$

provided that $0 < \tilde{\mu}(\mathbb{X}) < \infty$ almost surely; in case $\tilde{\mu}$ is a homogeneous CRM, these conditions are guaranteed by the finiteness of α and the infinite activity property, as proved in Regazzini et al. (2003), where this normalization procedure has first been introduced. A posterior characterization of normalized CRMs has been derived in James et al. (2009) for the exchangeable setting. Popular special instances include the Dirichlet process (Ferguson, 1973), which arises from normalization of gamma CRMs (Ferguson, 1973, 1974), the normalized σ–stable process (Kingman, 1975), the normalized inverse Gaussian process (Lijoi et al., 2005) and the normalized generalized gamma process (Lijoi et al., 2007); their use in mixture models is reviewed in Barrios et al. (2013). Henceforth, whenever $\tilde{\mu} \sim \text{CRM}(\nu)$, its normalization $\tilde{p}$ in (1.10) is denoted by $\tilde{p} \sim \text{nCRM}(\nu)$.

Random mixture hazard rates. In survival analysis, the time to failure is usually modeled by a random variable T, taking values on $\mathbb{R}^+$, whose probability distribution is here assumed to be absolutely continuous with respect to the Lebesgue measure. The hazard rate function of T represents the instantaneous risk of failure, and is defined as

$$h(t) := \frac{f(t)}{S(t)}, \qquad t \in \mathbb{R},$$

where f is the density function of T and S is its survival function. A random hazard rate function can be effectively modeled as a mixture of a suitable smoothing kernel over a random measure $\tilde{\mu}$ as

$$\tilde{h}(t) := \int_{\mathbb{X}} k(t;x)\, \tilde{\mu}(dx), \tag{1.11}$$

where $k\colon \mathbb{R}^+ \times \mathbb{X} \mapsto \mathbb{R}^+$ is a deterministic non-negative kernel. Such appealing mixture structure was introduced in the pioneering papers by Dykstra and Laud (1981) and Lo and Weng (1989), where the authors restrict to a gamma process on $\mathbb{R}^+$ as mixing random measure. Further developments considering general kernel choices and arbitrary CRMs are described in James (2005), where the posterior characterization of the mixing random measure is explicitly derived; see also Ishwaran and James (2004).

Given the mixture representation in (1.11), the random survival function is expressed in terms of the random hazard rate as

$$\tilde{S}(t) = \exp\left\{ -\int_0^t \int_{\mathbb{X}} k(s;x)\, \tilde{\mu}(dx)\, ds \right\}, \tag{1.12}$$

which is a proper survival function whenever $\lim_{t\to\infty} \tilde{S}(t) = 0$ almost surely. In case $\tilde{\mu}$ is a

homogeneous CRM, this condition is guaranteed by infinite activity and requiring

$$\int_{\mathbb{R}^+} k(s;x)\,ds = \infty, \qquad P_0 - \text{a.s.} \tag{1.13}$$

A very popular kernel specification was proposed in Dykstra and Laud (1981), where the authors consider $\mathbb{X} = \mathbb{R}^+$ and define $k(t;x) = \gamma(x)\,\mathbb{1}(t \geq x)$, for γ positive and right-continuous function. Such kernel satisfies the condition in (1.13) and represents the reference choice for modelling increasing hazard rates. For simplicity, this work considers $\gamma(x) = \gamma$ to be a constant function, though results presented in the following can be easily extended to non-constant specifications of γ function. Alternative kernel choices which are not bound to the modelling of increasing hazard rates are the rectangular kernel $k(t;x) = \gamma(x)\,\mathbb{1}(0 \leq t - x \leq \tau)$, with bandwidth $\tau > 0$, and the exponential (or Ornstein-Uhlenbeck) kernel $k(t;x) = \sqrt{2\kappa}\,\exp(-\kappa(t-x))\,\mathbb{1}(t \geq x)$, with rate parameter $\kappa > 0$ (Ishwaran and James, 2004; De Blasi et al., 2009); however, these kernel specifications do not satisfy the condition in (1.13). Further methodological and computational investigations can be found, e.g., in Ishwaran and James (2004); Nieto-Barajas and Walker (2004); Catalano et al. (2020).

1.3 Dirichlet and Pitman-Yor processes as normalized random measures

The Dirichlet process (DP) stands out as the fundamental and most celebrated nonparametric prior since its introduction in the seminal paper by Ferguson (1973); its relevance and influence within Bayesian nonparametrics is pointed out by Ghosal and van der Vaart (2017):

> ... The importance of the Dirichlet process in Bayesian nonparametrics is comparable to that of the normal distribution in probability and general statistics. ...

A number of equivalent constructions of the Dirichlet process have been proposed in the literature, each highlighting different properties and paving the way to its many possible extensions. Ferguson (1973) introduced the Dirichlet process through its finite-dimensional distributions, that is, as the random probability measure on a metric space whose evaluations on a finite and measurable partition of such space have Dirichlet distribution; the predictive characterization and connection with Polya urn schemes was proposed by Blackwell and MacQueen (1973), while Sethuraman (1994) popularized the stick-breaking construction within the statistical framework. This section focuses on the construction of the Dirichlet process as a normalization of a gamma process, which was already proposed as an alternative definition by Ferguson (1973); indeed, the intuition behind this characterization is, in Ferguson's words,

> ... since the Dirichlet distribution is definable ... as the joint distribution of a set of independent gamma variables divided by their sum, so also should the Dirichlet

> process be definable as a gamma process with independent "increments" divided by the sum. ...

Specifically, consider a gamma process, characterized by the Lévy intensity with diffuse base probability measure P_0 and jump component specified in (1.6),

$$\rho(ds) = \theta\, s^{-1}\, e^{-\beta\, s}\, ds, \tag{1.6}$$

where θ and β are positive parameters; the random probability measures defined by the normalization of such gamma process, as in (1.10), is distributed according to a Dirichlet process with base probability measure P_0 and *concentration* parameter $\theta > 0$. Note that the additional parameter $\beta > 0$ disappears throughout the normalization, and $\beta = 1$ is usually assumed in this context. From the Bayesian perspective, the prominent role of the Dirichlet process within the class of normalized random measures is highlighted in James et al. (2006), where it is characterized as the only conjugate prior among homogeneous normalized CRMs. Another interesting construction of the Dirichlet process based on completely random measures stems from its characterization as neutral-to-the right process (Doksum, 1974), and was first proposed in Ferguson (1974); see also Walker and Muliere (1997) for an extension of such approach. A complete account of the fundamental properties of the Dirichlet process and their implications can be found, e.g., in Ghosal and van der Vaart (2017).

Over the last five decades, possible extensions of the Dirichlet process in various directions have been explored in the literature, with the aim of enhancing its flexibility or accommodating different types of data; among them, those based on suitable transformations of completely random measures stand out for their analytical tractability (Lijoi and Prünster, 2010). A special role in this framework is played by the two-parameter Poisson-Dirichlet process, first introduced in Perman et al. (1992) and often referred to as *Pitman-Yor* process after the foundational paper by Pitman and Yor (1997), where many of its properties are derived. The characterization of the Pitman-Yor process (PY) as a normalized random measure is clarified in Pitman (2003), where it is presented as a special case within the class of Poisson-Kingman distributions for sequences of ranked random probability masses; indeed, the Pitman-Yor process is not a normalized CRM, but arises as the normalization of a random measure defined through a change of measure with respect to a stable CRM. Specifically, consider a σ-stable process $\tilde{\mu}_\sigma$, characterized by the Lévy intensity with diffuse base probability measure P_0 and jump component,

$$\rho(ds) = \frac{\sigma}{\Gamma(1-\sigma)}\, s^{-1-\sigma}\, ds,$$

where $\sigma \in [0, 1)$, and define the probability distribution of a random measure $\tilde{\mu}_{\sigma,\theta}$ as absolutely continuous with respect to the distribution of $\tilde{\mu}_\sigma$, having Radon-Nikodym derivative

$$\frac{d\mathcal{L}(\tilde{\mu}_{\sigma,\theta})}{d\mathcal{L}(\tilde{\mu}_\sigma)}(m) = \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)}\, m(\mathbb{X})^{-\theta}, \tag{1.14}$$

where $\theta > -\sigma$; the random probability measure defined by the normalization of $\tilde{\mu}_{\sigma,\theta}$, as in (1.10), is distributed according to a Pitman-Yor process with base probability measure P_0, *concentration* parameter θ and *discount* parameter σ. An alternative construction of the random measure $\tilde{\mu}_{\sigma,\theta}$, which is arguably simpler but less convenient for the purposes of this work, is obtained by considering a generalized gamma process, as specified in (1.7), and placing a gamma hyperprior on its concentration parameter (Pitman and Yor, 1997). Note that the Pitman-Yor process also admits a convenient representation as stick-breaking prior, as already remarked in Pitman (1995). Moreover, it stands out as a notable example among Gibbs-type priors, introduced by Gnedin and Pitman (2006) and further studied from the statistical perspective in Lijoi et al. (2007, 2008); in particular, Lijoi et al. (2008) characterize the Pitman-Yor process and the only quasi-conjugate Gibbs-type prior.

Chapter 2

A Unified Approach to Hierarchical Random Measures

Hierarchical models enjoy great popularity due to their ability to handle heterogeneous groups of observations by leveraging on their underlying common structure. In a Bayesian nonparametric framework, the hierarchy is introduced at the level of group-specific random measures, and then translated at the observations' level via suitable transformations. This chapter proposes a new strategy to derive closed-form expressions for the marginal and posterior distributions of each group. Indeed, a common structural core shared by the different hierarchical constructions proposed in the Bayesian nonparametric literature is unraveled by directly inserting a suitable set of latent variables into the generative model for the data. Specifically, this chapter identifies a key identity that underlies such hierarchical models and highlights its role in the derivation of quantities of interest.

2.1 Introduction

Recent developments in Bayesian nonparametrics are focused on flexible ways to account for different forms of heterogeneity across observations; see Cifarelli and Regazzini (1978) and MacEachern (1999, 2000) for pioneering works, and Quintana et al. (2022) for a recent review. A particularly relevant framework is that of partial exchangeability (de Finetti, 1938), which allows to model multiple groups of observations sharing similar features, with homogeneity (exchangeability) holding within each group, while preserving some heterogeneity across different groups. Typical instances include patients affected by the same disease and treated in different hospitals, or children of the same age raised in different countries. A generalization of de Finetti's representation theorem for multiple groups of observations characterizes partially exchangeable sequences as conditionally independent sequences, given group-specific random probability

measures; since groups are assumed to share similar features, it is important to incorporate borrowing of information across different groups. This objective is naturally met by the Bayesian paradigm: if dependence among the group-specific random probability measures is introduced a priori, the posterior distribution for each group should also make use of the information contained in the other groups. This typically induces a shrinkage effect that makes the posterior estimates more reliable, and disappears as the number of observations diverges.

The previous discussion raises the fundamental question of how to introduce dependence among random (probability) measures, which has been addressed from several different perspectives in the vast literature on the topic. De Iorio et al. (2004) develop the dependent Dirichlet process framework of MacEachern (1999, 2000) by imposing an ANOVA-type structure on the atoms, while Dunson and Park (2008) and Rodriguez and Dunson (2011) model predictor-dependent weights via kernel and probit transformations. As for the partially exchangeable setting, proposals inducing dependence across different groups of observations include additive (Müller et al., 2004; Griffin et al., 2013; Lijoi et al., 2014), nested (Rodriguez et al., 2008; Camerlenghi et al., 2019; Lijoi et al., 2023) and hierarchical structures (Teh et al., 2006; Camerlenghi et al., 2019, 2021). Further interesting constructions of dependent completely random measures, based on multivariate Lévy intensities, can be found in Epifani and Lijoi (2010); Riva-Palacio and Leisen (2018, 2021), which exploit Lévy copulas, and in Griffin and Leisen (2017), where the authors introduced compound random measures (CoRMs), a general class of dependent random measures including both superpositions of CRMs (Lijoi and Nipoti, 2014; Lijoi et al., 2014) and thinned CRMs (Chen et al., 2013; Lau and Cripps, 2022). See also the review by Quintana et al. (2022) and references therein. Among these constructions, hierarchical forms of dependence are arguably the most natural ones within the Bayesian framework: as dependence among homogeneous (exchangeable) observations is typically introduced through conditional independence, it is conceptually straightforward to introduce dependence among the random (probability) measures through conditional independence as well. Thanks to de Finetti's representation theorem, this approach leads to an (infinitely extendable) exchangeable sequence of random measures. A compelling strategy to define conditionally independent completely random measures consists in assuming a random base measure, which is modeled either through a normalized completely random measure or directly through a completely random measure. The first approach has been mainly used to model dependent random discrete probability measures (Teh et al., 2006; Camerlenghi et al., 2019; Argiento et al., 2020), whereas the second has been used to provide the main ingredients to model dependent random hazard functions (Camerlenghi et al., 2021), though it is probably interesting to remark that, in principle, they could both be used to model both quantities.

These two classes of hierarchical models entail different dependence assumptions on the random measures; however, they also present significant conceptual and mathematical similarities. This chapter investigate such similarities and proposes a unifying framework that sheds light on their

common structure and on intriguing analogies in their posterior and predictive representations. Indeed, even if dealing with hierarchical models may appear more challenging than treating simpler exchangeable ones, they both rely on the same identity, which can be applied recursively to reduce the analysis of the multi-group framework to the easier single-group scenario. Specifically, consider a completely random measure $\tilde{\mu}$ with a diffuse base probability measure, i.e. whose atoms are distinct almost surely. For any non-negative measurable function f, and mutually disjoint balls $B_\varepsilon(x_j^*) = \{\, x : d(x, x_j^*) < \varepsilon \,\}$, the limiting behavior of

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} f(x)\,\tilde{\mu}(dx)\right\}\prod_{j=1}^{k}\tilde{\mu}(B_\varepsilon(x_j^*))^{n_j}\right], \tag{2.1}$$

as $\varepsilon \to 0$, can be explicitly characterized as recalled in (2.10) of Section 2.4. Thanks to this identity, one can derive the distribution of the random partition for exchangeable observations according to their ties, which in turn determines both the predictive and posterior distributions, as shown in James et al. (2006, 2009).

At first sight, it seems difficult to extend this strategy to the multi-group hierarchical framework: indeed, the base probability measure of each exchangeable completely random measure is modeled as an almost-surely discrete random measure itself (Section 2.2). This implies that the atoms of such measures display ties with positive probability, and thus do not fit into the above framework. Camerlenghi et al. (2019, 2021) work around this issue by exploiting the celebrated Faà di Bruno's formula, expressing higher order derivatives of compositions of functions: eventually, the inherent combinatorial structure induced by this formula turns out to be effectively represented by introducing suitable latent variables. Leveraging on this observation, this chapter proposed an alternative approach that bypasses the Faà di Bruno formula by directly inserting the latent variables into the data generative model. From an analytical point of view, this strategy substantially mitigates the combinatorial burden connected to the Faà di Bruno's formula, at the negligible cost of an augmented Lévy intensity measure. In addition, the description of the induced random partition structure becomes more transparent, as it can be considered a mere consequence of the ties within sequences of latent (unobserved) variables.

The intuition behind such proposal is the following: by adding a diffuse independent mark to each atom of the exchangeable random measure, one derives a completely random measure on the joint space of the atoms and the marks with the compelling property of not displaying ties almost surely. Then, one may use (2.1) on the augmented random measure and possibly remove the auxiliary latent marks through marginalization. Remarkably, this strategy leads to posterior and predictive representations for both classes of hierarchical models, thanks to a fundamental identity that extends (2.1) to hierarchical random measures. In Basu and Tiwari (1982), which stands out as an insightful contribution to the foundations of the Dirichlet process, the authors state a crucial desirable property for nonparametric models:

> ... If a prior is selected from this class, then the posterior distribution given a sample of observations from P is manageable analytically, and it belongs to the class, i.e. the class is closed under 'Bayesian operation'. ...

The present contribution shows that the two considered classes of hierarchical nonparametric priors meet such *desideratum*, by developing the necessary analytical tools and suitably extending the notion of closure to fit the more complex partially exchangeable framework.

The chapter is structured as follows. Section 2.2 introduces dependence between completely random measure through two related hierarchical constructions. Section 2.3 provides some intuition on the induced partition structure and introduces a convenient set of latent variables, representing the marks of the atoms in the previous discussion; its formal treatment can be found in Section 2.4, together with the key identity mentioned above for closed-form computations involving CRMs, which is then exploited to derive marginal distributions. The posterior characterization of CRMs is described in Section 2.6 for both classes of hierarchical models; in light of these results, the generalized gamma completely random measure arises as the natural conjugate prior, as discussed in Section 2.7. Finally, Section 2.8 provides further insights on the dependence structure between hierarchical random measures, together with some intuition on how to enhance its flexibility.

2.2 Hierarchical random measures

Hierarchical structures represent a natural way to construct vectors of dependent random measures (Section 2.1). Indeed, dependence in a vector of random measures $\tilde{\boldsymbol{\mu}}$ can be induced through the following hierarchical scheme:

$$\begin{aligned} \tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_D) \;\mid\; \tilde{\mu}_0 &\overset{\text{i.i.d.}}{\sim} \tilde{\mathcal{G}}, \\ \tilde{\mu}_0 &\sim \mathcal{G}_0, \end{aligned} \tag{2.2}$$

where $\tilde{\mathcal{G}}$ is the (random) conditional distribution of each $\tilde{\mu}_d$, for $d = 1, \ldots, D$, given the random measure $\tilde{\mu}_0$, which represents the root of the hierarchy and is distributed according to $\mathcal{G}_0$. Completely random measures are particularly well-suited to this hierarchical structure, as the random measure $\tilde{\mu}_0$ at the root of the hierarchy can be easily incorporated into the Lévy intensity measure characterizing the distribution $\tilde{\mathcal{G}}$ of the vector at the lower level of the hierarchy. This chapter considers two different hierarchical constructions that are commonly used to model dependent random probabilities and dependent random hazard rates, respectively.

A vector of dependent discrete random probability measures can be defined through the hierarchical structure

$$\begin{aligned} \tilde{\boldsymbol{p}} = (\tilde{p}_1, \ldots, \tilde{p}_D) \;\mid\; \tilde{p}_0 &\overset{\text{i.i.d.}}{\sim} \text{nCRM}(\tilde{\nu}_{\text{norm}}), \\ \tilde{p}_0 &\sim \text{nCRM}(\nu_0), \end{aligned} \tag{2.3}$$

where $\tilde{p}_1, \ldots, \tilde{p}_D$ are conditionally independent normalized CRMs with random Lévy intensity measure

$$\tilde{\nu}_{\text{norm}}(ds, dx) = \rho(ds)\, \tilde{p}_0(dx),$$

and $\tilde{p}_0$ is the normalized CRM at the root of the hierarchy, with Lévy intensity measure $\nu_0(ds, dx) = \rho_0(ds)\, P_0(dx)$. This hierarchical nonparametric construction was first introduced in Teh et al. (2006) for the special case of hierarchical Dirichlet processes, and extensively studied in Camerlenghi et al. (2019) for the more general class of processes considered here. Further investigations beyond the Dirichlet setup can be found, e.g., in Teh and Jordan (2010); Camerlenghi et al. (2017, 2018); Catalano et al. (2024).

In the case of random mixture hazard models, normalization of random measures is not needed to define the nonparametric prior, since it results from the combination of dependent non-normalized random measures with suitable kernels. The hierarchical structure defining the vector of underlying dependent random measures is therefore

$$\begin{aligned} \tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_D) \;\mid\; \tilde{\mu}_0 &\overset{\text{i.i.d.}}{\sim} \text{CRM}(\tilde{\nu}), \\ \tilde{\mu}_0 &\sim \text{CRM}(\nu_0), \end{aligned} \tag{2.4}$$

where $\tilde{\mu}_1, \ldots, \tilde{\mu}_D$ are conditionally independent CRMs with random Lévy intensity measure

$$\tilde{\nu}(ds, dx) = \rho(ds)\, \tilde{\mu}_0(dx),$$

and $\tilde{\mu}_0$ is the CRM at the root of the hierarchy, with Lévy intensity measure $\nu_0(ds, dx) = \rho_0(ds)\, P_0(dx)$. This class of mixture hazard rates, based on a hierarchical dependence structure of the underlying random measures, was introduced in Camerlenghi et al. (2021); an alternative approach to define dependent mixture hazards can be found in Lijoi et al. (2014).

2.3 Random partitions and latent variables

Consider an array of partially exchangeable sequences with de Finetti measure featuring the hierarchically dependent specification discussed in the previous section. The almost-sure discreteness of (normalized) CRMs naturally induces a random nested partition structure, with groups including elements both within and across such partially exchangeable sequences. This partition structure can be regarded as a two-level extension of the random partition induced by discrete random probability measures in the exchangeable setting, conveniently described by the well-known *Chinese restaurant process* metaphor (Aldous, 1985; Pitman, 1996); indeed, it was first characterized for hierarchical Dirichlet processes by Teh et al. (2006), where the authors introduce the *Chinese restaurant franchise* metaphor, and later generalized for hierarchical (normalized) CRMs in Camerlenghi et al. (2019), Argiento et al. (2020) and Camerlenghi et al. (2021). The characterization of such partition structure is proposed in this section through the hierarchical

prior specification in (2.3), based on normalized random measures; however, it characterizes every Bayesian nonparametric model built upon hierarchical CRMs priors, being a property of the hierarchical structure itself rather than of the specific model.

Consider the array of D partially exchangeable samples

$$\begin{aligned} \mathbf{X}_d = (X_{d1}, \dots, X_{dN_d}) \mid \tilde{p}_d &\overset{\text{i.i.d.}}{\sim} \tilde{p}_d, \qquad d = 1, \dots, D, \\ \tilde{\boldsymbol{p}} = (\tilde{p}_1, \dots, \tilde{p}_D) &\sim \mathcal{Q}_D, \end{aligned} \tag{2.5}$$

for integers $N_1, \dots, N_D \geq 1$, where $\mathcal{Q}_D$ is the hierarchical prior in (2.3). The almost-sure discreteness and dependence of the random probability measures in $\tilde{\boldsymbol{p}}$ imply that tied values occur with positive probability both within and across samples, that is $\mathbb{P}(X_{\ell i} = X_{\kappa j}) > 0$ for any ℓ and κ. Therefore, a random partition of the observations is naturally induced, whereby two elements are in the same partition group if and only if they have the same value. Denote by $X_1^*, \dots, X_k^*$ the k distinct values assumed in the D partially exchangeable samples, with respective multiplicities $n_1, \dots, n_k$. As a consequence, elements belonging to the same partition group may or may not belong to the same sample, which means that there is no structural relationship between the random partition induced by tied values and the natural one determined by the D samples. Let n_{dj} be the number of elements in sample d belonging to group j,

$$n_{dj} = \sum_{i=1}^{N_d} (X_{di} = X_j^*), \qquad d = 1, \dots, D, \ j = 1, \dots, k;$$

given the two-fold nature of the partition structure, observed values are not enough to fully characterize its complexity. It is therefore convenient to introduce corresponding sequences of latent variables

$$\mathbf{Z}_d = (Z_{d1}, \dots, Z_{dN_d}), \qquad d = 1, \dots, D,$$

taking values on a complete and separable metric space $(\mathbb{T}, \mathcal{T})$, which themselves admit ties with positive probability. These latent variables are formally introduced in the next section for both normalized random measures and random mixture hazards models, and allow to describe a finer partition structure, featuring ties within each sample (but not across samples); this greatly simplifies the learning scheme. In particular, for each sample d and each group j, consider the n_{dj} elements in $\mathbf{X}_d$ for which $X_{di} = X_j^*$ holds. The corresponding elements in $\mathbf{Z}_d$ may themselves show ties: denote by $Z_{dj1}^*, \dots, Z_{djr_{dj}}^*$ their r_{dj} distinct values, with multiplicities $q_{dj1} + \cdots + q_{djr_{dj}} = n_{dj}$. Notice that, whenever $Z_{di} = Z_{di'}$ then $X_{di} = X_{di'}$, i.e. a tie among values in $\mathbf{Z}_d$ implies a tie among the corresponding elements in $\mathbf{X}_d$, while the converse is not necessarily true. Moreover, let r_j be the partial sum of elements in $(r_{dj})_{dj}$ with respect to d, and let r be their total sum.

As briefly mentioned above, an intuitive description of the partition structure introduced in this section is provided by the well-known *Chinese restaurant franchise* metaphor, first presented

in Teh et al. (2006) for the hierarchical Dirichlet process. According to this scheme, a franchise consists of D restaurants sharing the same menu, which includes an infinite number of dishes; each restaurant has infinitely many tables and the customers seated at the same table eat the same dish. Customers arriving at each restaurant may choose to either sit at a table with other customers, thus eating the dish already served at that table, or sit at an empty table, either eating a dish already served at other tables in the franchise or eating a new dish from the menu. Notice that, in contrast to the simple Chinese restaurant metaphor, the same dish can be served at different tables within the same restaurant and across different restaurants. Embedding the partition structure described above in the metaphor, element X_{di} represents the dish served in restaurant d to customer i, and the distinct values $X_1^*, \ldots, X_k^*$ represent the k distinct dishes served in the franchise, with n_{dj} being the number of customers eating dish j at restaurant d. Likewise, the latent variable Z_{di} represents the table in restaurant d at which customer i is seated, with r_{dj} being the number of tables in restaurant d at which dish j is served, and q_{djh} being the number of customers in restaurant d eating dish j at table h.

The sequences of latent variables introduced above can be explicitly included in the hierarchical prior specifications by extending the random measures at the lower level of the hierarchies to a larger space (as further clarified in the next section). Specifically, given the root of the hierarchy, the conditionally independent (normalized) CRMs appearing in (2.3) and (2.4), namely $\tilde{p}_1, \ldots, \tilde{p}_D$ and $\tilde{\mu}_1, \ldots, \tilde{\mu}_D$, can be extended as random measures on $\mathbb{T} \times \mathbb{X}$, and characterized, respectively, by the augmented random Lévy intensities

$$\tilde{\nu}^e_{\text{norm}}(ds, dz, dx) = \rho(ds)\, H(dz)\, \tilde{p}_0(dx), \qquad \tilde{\nu}^e(ds, dz, dx) = \rho(ds)\, H(dz)\, \tilde{\mu}_0(dx),$$

where H is an arbitrary diffuse probability measure on $(\mathbb{T}, \mathcal{T})$. As a consequence, each extended random measure at the lower level of the hierarchy can be represented, respectively, as

$$\tilde{p}^e_d(dz, dx) = \sum_{h \geq 1} S_{dh}\, \delta_{(Z^*_{dh}, X^*_{dh})}(dz, dx), \qquad \tilde{\mu}^e_d(dz, dx) = \sum_{h \geq 1} S_{dh}\, \delta_{(Z^*_{dh}, X^*_{dh})}(dz, dx), \tag{2.6}$$

where $(S_{dh})_{h \geq 1}$ and $(Z^*_{dh})_{h \geq 1}$ are independent sequences of random (probability) jumps and marks, while $(X^*_{dh})_{h \geq 1}$ is a sequence of independent locations sampled proportionally to, respectively, $\tilde{p}_0$ and $\tilde{\mu}_0$. Note that the random measures on $(\mathbb{X}, \mathcal{X})$ introduced in the original definition are easily recovered via marginalization of the marks.

An interesting feature of this analytical device is the extension of the (random) *discrete* components, defined on $\mathbb{X}$, of the Lévy intensity measures characterizing CRMs at the lower level of the hierarchies, to *diffuse* components, defined on $\mathbb{T} \times \mathbb{X}$, as a consequence of H being a diffuse measure. Such property turns out to be fundamental for the recursive application of the result in (2.10) in presence of hierarchical schemes, as discussed in the following section.

2.4 Key identity for random measures

The availability of a framework to describe the random partition structure induced by discrete hierarchical priors is a prerequisite for the determination of marginal and posterior distributions. In this respect, the possibility to marginalize quantities of interest with respect to the prior distribution represents the cornerstone of computations with CRMs, and relies on a specific core structure, which leads to closed-form and tractable expressions.

Let $\tilde{\mu}$ be a homogeneous CRM with Lévy intensity measure $\nu(ds, dx) = \rho(ds)\alpha(dx)$, where α is a finite diffuse measure on $\mathbb{X}$; both its Laplace exponent and cumulants, defined respectively by

$$\psi(u) = \int_{\mathbb{R}^+} \left(1 - e^{-us}\right) \rho(ds), \qquad \tau(m; u) = \int_{\mathbb{R}^+} s^m e^{-us} \rho(ds), \tag{2.7}$$

play a crucial role in the investigation of the distributional properties of $\tilde{\mu}$. The core quantity representing the building block for every computation with CRMs is

$$\exp\left\{-\int_{\mathbb{X}} f(x)\tilde{\mu}(dx)\right\} \prod_{j=1}^{k} \tilde{\mu}(B_\varepsilon(x_j^*))^{n_j} \tag{2.8}$$

where $B_\varepsilon(x_j^*) = \{\, x \in \mathbb{X} : d(x, x_j^*) < \varepsilon \,\}$, for $j = 1, \ldots, k$, are ε-balls centered at distinct values $x_1^*, \ldots, x_k^* \in \mathbb{X}$, with multiplicities $n_1, \ldots, n_k \geq 1$ such that $\sum_{j=1}^k n_j = n$, and $f : \mathbb{X} \mapsto \mathbb{R}^+$ is any non-negative measurable function. Computing the expectation of the quantity in (2.8), one can prove that

$$\begin{aligned} \lim_{\varepsilon \to 0} \frac{\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} f(x)\tilde{\mu}(dx)\right\} \prod_{j=1}^{k} \tilde{\mu}(B_\varepsilon(x_j^*))^{n_j}\right]}{\prod_{j=1}^{k} \alpha(B_\varepsilon(x_j^*))} \\ = \exp\left\{-\int_{\mathbb{X}} \psi(f(x))\, \alpha(dx)\right\} \prod_{j=1}^{k} \tau(n_j; f(x_j^*)), \end{aligned} \tag{2.9}$$

which can be informally rewritten as

$$\begin{aligned} \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} f(x)\, \tilde{\mu}(dx)\right\} \prod_{j=1}^{k} \tilde{\mu}(dx_j^*)^{n_j}\right] \\ = \exp\left\{-\int_{\mathbb{X}} \psi(f(x))\, \alpha(dx)\right\} \prod_{j=1}^{k} \tau(n_j; f(x_j^*))\, \alpha(dx_j^*). \end{aligned} \tag{2.10}$$

A simplified proof of the identity (2.9), based on the characterization of completely random

measures through their Laplace function transform, is proposed in Appendix A.1. Formally, the left-hand side of (2.10) can be interpreted as the (first) moment measure of an exponentially tilted n-fold product random measure of the CRM $\tilde{\mu}$, restricted to a specific subset of $\mathbb{X}^n$: indeed, $\tilde{\mu}(\cdot)^{n_j}$ may be regarded as the restriction of the n_j-fold product measure of $\tilde{\mu}$ to the diagonal $\Delta_{n_j} = \{ (x_1, \dots, x_{n_j}) \in \mathbb{X}^{n_j} : x_1 = \cdots = x_{n_j} \}$, through the correspondence

$$\tilde{\mu}(A)^{n_j} = \tilde{\mu}^{n_j}\big(\{ (\underbrace{x, \dots, x}_{n_j}) : x \in A \} \big) = \tilde{\mu}^{n_j}\big((\underbrace{A \times \cdots \times A}_{n_j}) \cap \Delta_{n_j} \big),$$

which is non-zero due to the almost-sure discreteness of $\tilde{\mu}$. Similarly, $\prod_{j=1}^{k} \tilde{\mu}(dx_j^*)^{n_j}$ may be identified with the restriction of the n-fold product measure of $\tilde{\mu}$ to a particular subspace of $\mathbb{X}^n$. Specifically, for any $1 \leq k \leq n$, consider the linear subspaces of $\mathbb{X}^n$ of dimension k for which the n coordinates can be partitioned into k groups with multiplicities $n_1, \dots, n_k$, where coordinates in the same group take on the same value. The moment measure introduced above, restricted to each of such k-dimensional subspaces, is shown in (2.10) to be absolutely continuous with respect to the product measure α^k, i.e. the k-fold product of α with itself, with Radon-Nikodym derivative

$$\exp\left\{ -\int_{\mathbb{X}} \psi(f(x))\, \alpha(dx) \right\} \prod_{j=1}^{k} \tau(n_j; f(x_j^*)),$$

where $x_1^*, \dots, x_k^* \in \mathbb{X}$ are the k distinct values assumed by coordinates belonging to the same group. Note that this density only depends on the groups' multiplicities $n_1, \dots, n_k$ and distinct values $x_1^*, \dots, x_k^*$, while the coordinates' ordering identifies the specific k-dimensional subspace where the measure in (2.10) is concentrated. Moreover, such measure can be decomposed into the product of a constant exponential term and k independent measures on $\mathbb{X}$, each absolutely continuous with respect to the diffuse measure α.

This result is particularly suited to hierarchical structures of random measures: indeed, if the random measure $\tilde{\mu}$ is defined through a hierarchical scheme and α is itself a random measure, the same result can be applied recursively, since the structure in (2.8) reappears for measure α in the right-hand side of (2.10). Note that the equality above holds true only for diffuse choices of the finite measure α, making it potentially useless if α is a CRM (thus having almost surely discrete realizations). However, this issue is successfully addressed by replacing the random measure $\tilde{\mu}$ with its extended counterpart, as formally described hereunder. Let $\tilde{\mu}^e$ be a homogeneous CRM with random Lévy intensity measure

$$\tilde{\nu}^e(ds, dz, dx) = \rho(ds)\, H(dz)\, \tilde{\mu}_0(dx),$$

where H is a diffuse probability measure on $\mathbb{T}$ and $\tilde{\mu}_0$ is itself a homogeneous CRM with Lévy intensity measure $\nu_0(ds, dx) = \rho_0(ds)\, P_0(dx)$, for P_0 a diffuse probability measure on $\mathbb{X}$. The

core quantity of interest, playing the role of (2.8) for hierarchical schemes, is

$$\exp\left\{-\int_{\mathbb{T}\times\mathbb{X}} f(x)\,\tilde{\mu}^e(dz,dx)\right\}\prod_{j=1}^{k}\prod_{h=1}^{r_j}\tilde{\mu}^e(dz^*_{jh},dx^*_j)^{q_{jh}}, \tag{2.11}$$

where $x^*_1,\ldots,x^*_k \in \mathbb{X}$ are distinct values, $z^*_{j1},\ldots,z^*_{jr_j} \in \mathbb{T}$ are the $r_j \geq 1$ distinct values corresponding to the same x^*_j with multiplicities $q_{j1},\ldots,q_{jr_j} \geq 1$ and such that $\sum_{j=1}^{k} r_j = r$, and $f\colon \mathbb{X} \mapsto \mathbb{R}^+$ is any non-negative measurable function. The expectation of such quantity with respect to the random measure $\tilde{\mu}$ is computed by recursive application of (2.10), first at the lower level and then at the root of the hierarchy, obtaining

$$\begin{aligned}\mathbb{E}\left[\exp\left\{-\int_{\mathbb{T}\times\mathbb{X}} f(x)\,\tilde{\mu}^e(dz,dx)\right\}\prod_{j=1}^{k}\prod_{h=1}^{r_j}\tilde{\mu}^e(dz^*_{jh},dx^*_j)^{q_{jh}}\right]&\\ = \exp\left\{-\int_{\mathbb{X}}\psi_0(\psi(f(x)))\,P_0(dx)\right\}\prod_{j=1}^{k}\tau_0(r_j;\psi(f(x^*_j)))\,P_0(dx^*_j)&\\ \times\prod_{j=1}^{k}\prod_{h=1}^{r_j}\tau(q_{jh};f(x^*_j))\,H(dz^*_{jh}),&\end{aligned} \tag{2.12}$$

where τ_0 and ψ_0 are defined as in (2.7) replacing ρ with ρ_0. Similarly to the non-hierarchical case, (2.12) defines a (first) moment measure of an exponentially tilted n-fold product measure of the random measure $\tilde{\mu}^e$ with itself, which is here a measure on the space $(\mathbb{T}\times\mathbb{X})^n = \mathbb{T}^n \times \mathbb{X}^n$. Specifically, consider the linear subspaces of $\mathbb{T}^n \times \mathbb{X}^n$ of dimension $r \times k$ for which the $2n$ coordinates can be grouped according to the partition structure encoded into elements $(q_{jh})_{jh}$. This moment measure, restricted to each of such subspaces, is shown in (2.12) to be absolutely continuous with respect to the product measure $H^r \times P_0^k$, with Radon-Nikodym derivative

$$\exp\left\{-\int_{\mathbb{X}}\psi_0(\psi(f(x)))\,P_0(dx)\right\}\prod_{j=1}^{k}\tau_0(r_j;\psi(f(x^*_j)))\left(\prod_{h=1}^{r_j}\tau(q_{jh};f(x^*_j))\right),$$

where $x^*_1,\ldots,x^*_k \in \mathbb{X}$ are the k distinct values assumed by $\mathbb{X}$-valued coordinates belonging to the same group. Again, this measure can be decomposed into the product of a constant exponential term, the r-fold product of H with itself, and k independent measures on $\mathbb{X}$, each absolutely continuous with respect to the diffuse probability measure P_0.

2.5 Bayesian nonparametric hierarchical models

This section is devoted to the analysis of the likelihood functions associated to normalized random measures and random mixture hazards, assuming partially exchangeable models with a hierarchical prior specification. In particular, the introduction of latent variables specific to each

model, together with suitable analytical manipulations, recovers the core structure introduced in (2.11), which in turn allows for the computation of marginal distributions explicitly via the recursive application of (2.10).

Normalized random measures. Consider the array of D partially exchangeable samples

$$\begin{aligned} \mathbf{X}_d = (X_{d1}, \ldots, X_{dN_d}) \mid \tilde{p}_d &\overset{\text{i.i.d.}}{\sim} \tilde{p}_d, \qquad d = 1, \ldots, D, \\ \tilde{\boldsymbol{p}} = (\tilde{p}_1, \ldots, \tilde{p}_D) \mid \tilde{p}_0 &\overset{\text{i.i.d.}}{\sim} \text{nCRM}(\tilde{\nu}_{\text{norm}}), \\ \tilde{p}_0 &\sim \text{nCRM}(\nu_0), \end{aligned} \tag{2.13}$$

for integers $N_1, \ldots, N_D \geq 1$ and hierarchical prior (2.3). Introducing the corresponding latent variables, which represent the tables in the restaurant franchise metaphor, the likelihood function associated to the augmented sample $(\mathbf{X}_d, \mathbf{Z}_d)$ is

$$\mathcal{L}(\tilde{\mu}_d^e; \mathbf{X}_d, \mathbf{Z}_d) = \prod_{i=1}^{N_d} \tilde{p}_d^e(dZ_{di}, dX_{di}) = \tilde{\mu}_d^e(\mathbb{T}, \mathbb{X})^{-N_d} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \tilde{\mu}_d^e(dZ_{djh}^*, dX_j^*)^{q_{djh}}.$$

By using a simple analytical manipulation based on the density of a gamma random variable, the likelihood can be rewritten as

$$\mathcal{L}(\tilde{\mu}_d^e; \mathbf{X}_d, \mathbf{Z}_d) = \frac{1}{\Gamma(N_d)} \int_{\mathbb{R}^+} u_d^{N_d - 1} \exp\left\{ - \int_{\mathbb{T} \times \mathbb{X}} u_d \, \tilde{\mu}_d^e(dz, dx) \right\} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \tilde{\mu}_d^e(dZ_{djh}^*, dX_j^*)^{q_{djh}} \, du_d,$$

where u_d is an additional latent variable, thanks to which the core structure in (2.11) is successfully recovered in the likelihood. Therefore, the result in (2.10) can be applied to marginalize the expression with respect to the lower level of the hierarchical prior, i.e. conditionally on $\tilde{p}_0$, obtaining

$$\mathbb{P}(\mathbf{X}_d, \mathbf{Z}_d \mid \tilde{p}_0) = \frac{1}{\Gamma(N_d)} \int_{\mathbb{R}^+} u_d^{N_d - 1} \, e^{-\psi(u_d)} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \tau(q_{djh}; u_d) \, du_d \cdot \prod_{j=1}^{k} \tilde{p}_0(dX_j^*)^{r_{dj}} \prod_{h=1}^{r_{dj}} H(dZ_{djh}^*).$$

Note that only the random partition induced by ties in the sequence $\mathbf{Z}_d$, which is encoded into groups multiplicities $(q_{djh})_{jh}$, is relevant in the expression above, while their specific values are sampled independently from the measure H. Since these values are not even observed in the model, they can be safely ignored, and $\Pi_{\mathbf{Z}_d}$ can be written instead of $\mathbf{Z}_d$ to indicate that results depend on the partition induced by latent variables, rather than on their specific values.

The same analytical manipulation can be performed recursively for the random probability measure $\tilde{p}_0$, considering the likelihood associated to the D partially exchangeable samples. The joint marginal distribution of the observations $\mathbf{X}$ and the latent variables $\mathbf{Z}$ can be effectively expressed in terms of the random partitions they induce, respectively denoted by $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Z}}$,

and of the vectors of their distinct values, denoted by $\mathbf{X}^*$ and $\mathbf{Z}^*$, as

$$\mathbb{P}(\Pi_{\mathbf{X}}, \mathbf{X}^*, \Pi_{\mathbf{Z}}, \mathbf{Z}^*) = \prod_{d=1}^{D} \left(\frac{1}{\Gamma(N_d)} \int_{\mathbb{R}^+} u_d^{N_d - 1} \, e^{-\psi(u_d)} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \tau(q_{djh}; u_d) \, du_d \cdot \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} H(dZ^*_{djh}) \right)$$
$$\times \frac{1}{\Gamma(r)} \int_{\mathbb{R}^+} u_0^{r-1} \, e^{-\psi_0(u_0)} \prod_{j=1}^{k} \tau_0(r_j; u_0) \, du_0 \cdot \prod_{j=1}^{k} P_0(dX^*_j). \quad (2.14)$$

In analogy with the comment above, the distinct values in $\mathbf{X}^*$ do not enter the partition function as well, and are sampled independently from the base probability measure P_0. Therefore, marginalizing out the contribution of the distinct values $(\mathbf{X}^*, \mathbf{Z}^*)$ in (2.14), one obtains the partially exchangeable partition probability function (pEPPF):

$$\mathbb{P}(\Pi_{\mathbf{X}}, \Pi_{\mathbf{Z}}) = \prod_{d=1}^{D} \frac{1}{\Gamma(N_d)} \int_{\mathbb{R}^+} u_d^{N_d - 1} \, e^{-\psi(u_d)} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \tau(q_{djh}; u_d) \, du_d$$
$$\times \frac{1}{\Gamma(r)} \int_{\mathbb{R}^+} u_0^{r-1} \, e^{-\psi_0(u_0)} \prod_{j=1}^{k} \tau_0(r_j; u_0) \, du_0. \quad (2.15)$$

This expression clearly highlights the way random partitions are composed at the two levels of the hierarchy. At the level of single samples (i.e. restaurant level), the partition function depends on the number of customers seated at each table, that is on the partition induced by each sequence $\mathbf{Z}_d$. On the other hand, at the root level (i.e. franchise level), the partition function depends on the number of tables eating each dish, which describes how the finer partition induced by latent variables $\mathbf{Z}$ is related to the coarser partition induced by observations $\mathbf{X}$.

According to the interpretation discussed in the previous section, the joint marginal distribution in (2.14) represents a moment measure on the joint space $\mathbb{T}^n \times \mathbb{X}^n$, restricted to the linear subspace of dimension $r \times k$ which reflects the partition structure induced by observations and latent variables and encoded into elements $(q_{djh})_{djh}$. Specifically, this restricted measure is absolutely continuous with respect to the product measure $H^r \times P_0^k$, and has constant Radon-Nikodym derivative expressed by the pEPPF in (2.15). The structure of the pEPPF represents the cornerstone of computational developments: indeed, full conditional distributions derived from it can be exploited to devise marginal Gibbs sampling schemes, as extensively discussed in Camerlenghi et al. (2019) and Argiento et al. (2020).

Random mixture hazards. Consider the array of D partially exchangeable samples

$$\begin{aligned} \mathbf{T}_d = (T_{d1}, \ldots, T_{dN_d}) \mid \tilde{p}_d &\overset{\text{i.i.d.}}{\sim} \tilde{p}_d, \qquad d = 1, \ldots, D, \\ \tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_D) \mid \tilde{\mu}_0 &\overset{\text{i.i.d.}}{\sim} \text{CRM}(\tilde{\nu}), \\ \tilde{\mu}_0 &\sim \text{CRM}(\nu_0), \end{aligned} \quad (2.16)$$

for integers $N_1, \ldots, N_D \geq 1$ and hierarchical prior (2.4). The random probability measure $\tilde{p}_d$ is recovered from the expression of the random hazard rate in (1.11) and is expressed in terms of the random measures $\tilde{\mu}_d$ as

$$\tilde{p}_d(dt) = \int_{\mathbb{X}} k(t;x)\,\tilde{\mu}_d(dx)\,\exp\left\{-\int_0^t\int_{\mathbb{X}} k(s;x)\,\tilde{\mu}_d(dx)\,ds\right\}dt.$$

The further level of complexity represented by the kernel mixture requires the introduction of additional latent variables, namely the ones representing the latent samples from the random measures, i.e. the dishes in the restaurant franchise metaphor, which instead are directly observed in the previously discussed case of normalized random measures. The augmented random probability measure takes the more tractable form

$$\tilde{p}_d(dt,dx) = k(t;x)\,\tilde{\mu}_d(dx)\,\exp\left\{-\int_0^t\int_{\mathbb{X}} k(s;x)\,\tilde{\mu}_d(dx)\,ds\right\}dt.$$

Introducing the latent variables representing the tables in the restaurant franchise metaphor, the likelihood function associated to the augmented sample $(\mathbf{T}_d, \mathbf{X}_d, \mathbf{Z}_d)$ is

$$\begin{aligned}\mathcal{L}(\tilde{\mu}_d^e;\mathbf{T}_d,\mathbf{X}_d,\mathbf{Z}_d) &= \prod_{i=1}^{N_d} k(T_{di};X_{di})\,\tilde{\mu}_d^e(dZ_{di},dX_{di})\,\exp\left\{-\int_0^{T_{di}}\int_{\mathbb{T}\times\mathbb{X}} k(s;x)\,\tilde{\mu}_d^e(dz,dx)\,ds\right\}dT_{di}\\ &= Q(\mathbf{T}_d,\mathbf{X}_d)\,\exp\left\{-\int_{\mathbb{T}\times\mathbb{X}} K_d(x)\,\tilde{\mu}_d^e(dz,dx)\right\}\prod_{j=1}^{k}\prod_{h=1}^{r_{dj}}\tilde{\mu}_d^e(dZ^*_{djh},dX^*_j)^{q_{djh}},\end{aligned}$$

where, in order to ease the notation, the following quantities have been defined:

$$Q(\mathbf{T}_d,\mathbf{X}_d) = \prod_{i=1}^{N_d} k(T_{di};X_{di})\,dT_{di}, \qquad K_d(x) = \sum_{i=1}^{N_d}\int_0^{T_{di}} k(s;x)\,ds.$$

Again, the structure in (2.11) is recovered in the likelihood, with the constant term u_d for normalized random measures replaced by the function K_d. Therefore, the result in (2.10) can be applied recursively at the lower and root levels of the hierarchical prior, so that the joint marginal distribution of both the observations $\mathbf{T}$ and the latent variables $\mathbf{X}$ and $\mathbf{Z}$ becomes

$$\begin{aligned}\mathbb{P}(\mathbf{T},\Pi_{\mathbf{X}},\mathbf{X}^*,\Pi_{\mathbf{Z}},\mathbf{Z}^*) &= Q(\mathbf{T},\mathbf{X})\,\exp\left\{-\int_{\mathbb{X}}\psi_0\left(\sum_{d=1}^{D}\psi(K_d(X^*_j))\right)P_0(dx)\right\}\\ &\times\prod_{d=1}^{D}\prod_{j=1}^{k}\prod_{h=1}^{r_{dj}}\tau(q_{djh};K_d(X^*_j))\,H(dZ^*_{djh})\cdot\prod_{j=1}^{k}\tau_0\left(r_j;\sum_{d=1}^{D}\psi(K_d(X^*_j))\right)P_0(dX^*_j).\end{aligned} \tag{2.17}$$

The similarities with the joint distribution derived in (2.14) are apparent, as the composition of random partitions at the two levels of the hierarchy follows the same structure. An important

difference is represented by the dependence of the partition function from the specific distinct values $\mathbf{X}^*$ through the functions $K_1, \ldots, K_D$, so that their contribution cannot be marginalized out, and a proper pEPPF cannot be defined. This dependence on the specific values is reflected by a non-constant Radon-Nikodym derivative, when (2.17) is regarded as a moment measure.

The lack of a proper pEPPF also represents a computational drawback in this context, as the Gibbs resampling step for the latent distinct values $\mathbf{X}^*$ involves both the kernel term $Q(\mathbf{T}, \mathbf{X})$ and the value of the partition function. On the other hand, analytical tractability greatly benefits from the absence of the integrals with respect to $u_1, \ldots, u_D$ and u_0 appearing in the expression (2.15) for normalized random measures, which are merely a byproduct of analytical manipulations and need to be treated as additional latent variables.

2.6 Posterior characterizations

Another essential result which leverages on the random partition structure is the posterior characterization of the random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_D$ and $\tilde{\mu}_0$. Specifically, posterior distributions of CRMs are recovered via the determination of their conditional Laplace functional transforms: in these expressions, the distributions of jumps at fixed locations and the Lévy intensities of CRMs without fixed jump points can be identified. A structural conjugacy property is shown to hold, that is, a posteriori, the vector of random measures $\tilde{\boldsymbol{\mu}}$ retains its hierarchical form, with random measures at the lower level of the hierarchy being conditionally independent given the random measure at the root.

In the following, posterior updates of hierarchical CRMs priors are explicitly described for the partially exchangeable models based on both normalized random measures and random mixture hazards. For convenience of presentation, the intensities of the jump components ρ and ρ_0 at both levels of the hierarchy are assumed to be absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^+$, and suitably written as $\rho(ds) = \rho(s)\, ds$ and $\rho_0(ds) = \rho_0(s)\, ds$, respectively. Moreover, the results in this section consider the original definitions of random measures at the lower level of the hierarchies, as introduced in (2.3) and (2.4): the adaptation to their extended versions is straightforward.

Normalized random measures. Consider the partially exchangeable model described in (2.13). The posterior distributions of the non-normalized random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_D$ and $\tilde{\mu}_0$ are characterized conditionally on the observations $\mathbf{X}$, the latent variables $\mathbf{Z}$, and the additional latent variables $U_1, \ldots, U_D$ and U_0. As already mentioned, such latent variables are a byproduct of analytical manipulations in the likelihood, and are needed to recover a (conditional) structural conjugacy. Let $U_1, \ldots, U_D$ and U_0 be conditionally independent positive random variables with

density functions

$$f_d(u \mid \Pi_{\mathbf{X}_d}, \Pi_{\mathbf{Z}_d}) \propto u^{N_d - 1}\, e^{-\psi(u)} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \tau(q_{djh}; u), \qquad d = 1, \ldots, D,$$

$$f_0(u \mid \Pi_{\mathbf{X}}, \Pi_{\mathbf{Z}}) \propto u^{r-1}\, e^{-\psi_0(u)} \prod_{j=1}^{k} \tau_0(r_j; u),$$

where we recall that $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Z}}$ denote the random partitions induced by ties in the sequences $\mathbf{X}$ and $\mathbf{Z}$, respectively, which are encoded into groups multiplicities r_j's (coarser partition) and q_{djh}'s (finer partition). At the lower level of the hierarchy, the posterior distribution of each random measure $\tilde{\mu}_d$, given the observations $\mathbf{X}_d$, the latent variables $\mathbf{Z}_d$ and U_d, and the root random measure $\tilde{\mu}_0$ is

$$\tilde{\mu}_d(dx) \;\mid\; \Pi_{\mathbf{X}_d}, \mathbf{X}_d^*, \Pi_{\mathbf{Z}_d}, U_d, \tilde{\mu}_0 \;\sim\; \tilde{\mu}_d^*(dx) + \sum_{j=1}^{k} \sum_{h=1}^{r_{dj}} J_{djh}\, \delta_{X_j^*}(dx),$$

where the random elements in the sum are mutually independent, $\tilde{\mu}_d^* \sim \mathrm{CRM}(\tilde{\nu}^*_{\mathrm{norm},d})$ with homogeneous Lévy intensity measure

$$\tilde{\nu}^*_{\mathrm{norm},d}(ds, dx) = e^{-U_d\, s}\, \tilde{\nu}_{\mathrm{norm}}(ds, dx) = e^{-U_d\, s}\, \rho(ds)\, \tilde{p}_0(dx),$$

and each J_{djh} is a non-negative random variable with density function

$$f_{djh}(s) \propto s^{q_{djh}}\, e^{-U_d\, s}\, \rho(s), \qquad j = 1, \ldots, k, \quad h = 1, \ldots, r_{dj}.$$

Therefore, a posteriori and conditionally on $\tilde{\mu}_0$, each random measure $\tilde{\mu}_d$ is still a CRM, resulting from the sum of random jumps at fixed points of discontinuity and a CRM without fixed points of discontinuity. The latter is characterized by the Lévy intensity measure of the prior with an exponential updating term, while the fixed points of discontinuity correspond to the distinct values of the observations. Moreover, the random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ preserve their conditional independence, given $\tilde{\mu}_0$.

Similarly, the posterior distribution of the random measure $\tilde{\mu}_0$ at the root of the hierarchy, given the observations $\mathbf{X}$, the latent variables $\mathbf{Z}$ and U_0, is

$$\tilde{\mu}_0(dx) \;\mid\; \Pi_{\mathbf{X}}, \mathbf{X}^*, \Pi_{\mathbf{Z}}, U_0 \;\sim\; \tilde{\mu}_0^*(dx) + \sum_{j=1}^{k} I_j\, \delta_{X_j^*}(dx),$$

where the random elements in the sum are mutually independent, $\tilde{\mu}_0^* \sim \mathrm{CRM}(\nu_0^*)$ with homogeneous Lévy intensity measure

$$\nu_0^*(ds, dx) = e^{-U_0\, s}\, \nu_0(ds, dx) = e^{-U_0\, s}\, \rho_0(ds)\, P_0(dx),$$

and each I_j is a non-negative random variable with density function

$$f_j(s) \propto s^{r_j}\, e^{-U_0\, s}\, \rho_0(s), \qquad j = 1, \ldots, k.$$

Again, the random measure $\tilde{\mu}_0$ is still a CRM a posteriori, given by the sum of random jumps at fixed points of discontinuity, corresponding to the distinct observed values, and an exponentially updated CRM without fixed points of discontinuity.

An interesting feature of this result is that the prior-posterior updating mechanism preserves the homogeneity of the random measures. Indeed, the density functions of additional latent variables $U_1, \ldots, U_D$ and U_0, the exponential update of the Lévy intensity and the distributions of random jumps at the fixed points of discontinuity depend only on the partition structure induced by the observations and the latent variables, encoded into $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Z}}$, while observed distinct values $\mathbf{X}^*$ only determine the fixed locations of discontinuity points. This property clearly parallels the factorization property of the marginal distribution into pEPPF and independent sampling of distinct values, and represents a fundamental computational advantage when direct sampling from the posterior distribution of hierarchical CRMs is involved.

Random mixture hazards. Consider the partially exchangeable model described in (2.16). In this case, the posterior distributions of random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_D$ and $\tilde{\mu}_0$ are characterized conditionally on the observations $\mathbf{T}$ and the latent variables $\mathbf{X}$ and $\mathbf{Z}$, featuring a proper structural conjugacy. Specifically, the posterior distribution of each random measure $\tilde{\mu}_d$, given the observations $\mathbf{T}_d$, the latent variables $\mathbf{X}_d$ and $\mathbf{Z}_d$, and the root measure $\tilde{\mu}_0$ is

$$\tilde{\mu}_d(dx) \;\mid\; \mathbf{T}_d, \Pi_{\mathbf{X}_d}, \mathbf{X}_d^*, \Pi_{\mathbf{Z}_d}, \tilde{\mu}_0 \;\sim\; \tilde{\mu}_d^*(dx) + \sum_{j=1}^{k} \sum_{h=1}^{r_{dj}} J_{djh}\, \delta_{X_j^*}(dx),$$

where the random elements in the sum are mutually independent, $\tilde{\mu}_d^* \sim \mathrm{CRM}(\tilde{\nu}_d^*)$ with non-homogeneous Lévy intensity measure

$$\tilde{\nu}_d^*(ds, dx) = e^{-K_d(x)\, s}\, \rho(ds)\, \tilde{\mu}_0(dx),$$

and each J_{djh} is a non-negative random variable with density function

$$f_{djh}(s) \propto s^{q_{djh}}\, e^{-K_d(X_j^*)\, s}\, \rho(s), \qquad j = 1, \ldots, k, \quad h = 1, \ldots, r_{dj}.$$

Similarly, the posterior distribution of the random measure $\tilde{\mu}_0$ at the root of the hierarchy, given the observations $\mathbf{T}$ and the latent variables $\mathbf{X}$ and $\mathbf{Z}$, is

$$\tilde{\mu}_0(dx) \;\mid\; \mathbf{T}, \Pi_{\mathbf{X}}, \mathbf{X}^*, \Pi_{\mathbf{Z}} \;\sim\; \tilde{\mu}_0^*(dx) + \sum_{j=1}^{k} I_j\, \delta_{X_j^*}(dx),$$

where the random elements in the sum are mutually independent, $\tilde{\mu}_0^* \sim \text{CRM}(\nu_0^*)$ with non-homogeneous Lévy intensity measure

$$\nu_0^*(ds, dx) = \exp\left\{-\sum_{d=1}^{D} \psi(K_d(x))\, s\right\} \rho_0(ds)\, P_0(dx),$$

and each I_j is a non-negative random variable with density function

$$f_j(s) \propto s^{r_j} \exp\left\{-\sum_{d=1}^{D} \psi(K_d(X_j^*))\, s\right\} \rho_0(s), \qquad j = 1, \ldots, k.$$

As already highlighted for marginal distributions, the structural analogies with the posterior characterization for normalized random measures are apparent, and similar considerations apply. In particular, here the role of the random variables $U_1, \ldots, U_D$ is played by the non-random functions $K_1, \ldots, K_D$, which summarize the contribution of the observations $\mathbf{T}$ to the posterior update. However, the analytical and computational advantage represented by the absence of additional latent variables is partially overturned by the non-homogeneity of the Lévy intensity measures characterizing, a posteriori, the continuous part of the hierarchical CRMs. The challenges represented by non-homogeneous CRMs in conditional sampling algorithms are discussed in Camerlenghi et al. (2021), where a novel general-purpose approach is proposed, and further developed in Section 3.6.

2.7 Generalized gamma CRMs as natural conjugate priors

The practical implementation of Bayesian procedures involving hierarchical CRMs priors requires the specification of their Lévy intensity measures. In particular, a fundamental role is played by measures ρ and ρ_0, which characterize the jump component and deeply affect the induced partition structure. Indeed, such measures directly impact the distributions of random jumps at fixed points of discontinuity in the posterior characterizations of hierarchical CRMs, and also enter the definition of key quantities in (2.7), which constitute the core structure of the marginal distributions (2.14) and (2.17): the availability of closed-form and tractable expressions represents a computational advantage for both marginal and conditional algorithms. On the contrary, the specification of P_0 has a far lower impact from both analytical and computational points of view.

A natural choice of ρ and ρ_0 for hierarchical constructions is represented by the hierarchical *generalized gamma* CRM (1.7), corresponding to the specifications

$$\rho(ds) = \frac{1}{\Gamma(1-\sigma)}\, s^{-\sigma-1}\, e^{-\beta s}\, ds, \qquad \rho_0(ds) = \frac{1}{\Gamma(1-\sigma_0)}\, s^{-\sigma_0-1}\, e^{-\beta_0 s}\, ds, \tag{2.18}$$

with parameters $\beta, \beta_0 \in \mathbb{R}^+$ and $\sigma, \sigma_0 \in [0, 1)$; notable special cases are obtained setting $\sigma = \sigma_0 = 0$, which corresponds to the *gamma* hierarchical CRM, and $\beta = \beta_0 = 0$, characterizing the σ*–stable* hierarchical CRM. The generalized gamma hierarchical CRM allows for the explicit computation of the integrals defining the Laplace exponent and its cumulants in (2.7), namely,

$$\psi(u) = \int_{\mathbb{R}^+} (1 - e^{-us})\, \rho(ds) = \frac{(\beta + u)^\sigma - \beta^\sigma}{\sigma} \overset{\sigma=0}{=} \log\left(1 + \frac{u}{\beta}\right),$$
$$\tau(m; u) = \int_{\mathbb{R}^+} s^m\, e^{-us}\, \rho(ds) = \frac{\Gamma(m - \sigma)}{\Gamma(1 - \sigma)}\, (\beta + u)^{-m+\sigma}.$$

These quantities can be directly substituted into the expressions of the marginal distributions, from which full conditional distributions and predictive urn schemes are easily derived.

Moreover, the generalized gamma choice acts as the (conditionally) conjugate prior with respect to the posterior characterization of hierarchical CRMs for the partially exchangeable models discussed in this work. For example, considering the model in (2.13) based on normalized random measures, the posterior distribution of each random measure $\tilde{\mu}_d$ at the lower level of the hierarchy consists of the sum of random jumps at fixed points of discontinuity having gamma distribution, namely,

$$J_{djh} \sim \text{Gamma}(q_{djh} - \sigma, \beta + U_d),$$

and a CRM without fixed points of discontinuity, which still has the Lévy intensity measure of a generalized gamma CRM, with the exponential term characterized by the parameter update $\beta \mapsto \beta + U_d$. The same structure is observed for the root measure $\tilde{\mu}_0$, and for the model in (2.16) based on random mixture hazards, with the usual roles swap of the latent variables $U_1, \ldots, U_D$ and the functions $K_1, \ldots, K_D$, which convert the real parameters β and β_0 into functional parameters and make the Lévy intensity non-homogeneous (see Chapter 3).

2.8 Remarks on the dependence structure of hierarchical CRMs

The previous section have described two different hierarchical models, namely (2.3) and (2.4), that provide an intuitive and effective way to introduce prior dependence among the components of a vector of random measures. The amount of such dependence regulates the borrowing of information across groups, that is, how much inference and prediction for each group are influenced by the observations in other groups. In an ideal setting where infinite observations for each group are available, leveraging on the information contained in the other groups of observations would be useless, if not potentially harmful; however, in real situations, when only few observations per group are available, or datasets are strongly unbalanced, the borrowing of information from other groups can lead to crucial improvements in the estimates and meaningful reduction of their uncertainty.

Considering a vector $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_D)$ of random measures, its dependence structure admits

two extremes situations: in case the random measures are equal almost surely, i.e., $\tilde{\mu}_1 = \cdots = \tilde{\mu}_D$ a.s., there is maximal dependence and, since observations are treated as belonging to the same group, full borrowing of information; on the other hand, in case the random measures are mutually independent, there is no borrowing of information, since the inference for each group is not affected by the observations in other groups. Such extremes highlight the role of the prior dependence introduced in the model (i.e. the prior elicitation), as it has major consequences on the learning mechanism. In this respect, it is sufficient to remark that, when perfect dependence is assumed a priori, the posterior estimates for each group are exactly coincident, disregarding possible differences across the groups; on the contrary, if independence of random measures is assumed a priori, the posterior estimates for each group do not take into account the observations in other groups, with potential loss of information. Therefore, the quantification of such amount of induced prior dependence represent a fundamental task.

The simpler and arguably most natural summary of the dependence structure between random measures $\tilde{\mu}_i$ and $\tilde{\mu}_j$ is represented by their pairwise covariance structure, namely the quantity $\mathrm{Cov}(\tilde{\mu}_i(A), \tilde{\mu}_j(A))$, for any set $A \in \mathcal{X}$, and by its normalized version, the pairwise correlation, namely $\mathrm{Cor}(\tilde{\mu}_i(A), \tilde{\mu}_j(A))$; more elaborated proposals going beyond this pairwise comparisons and based on the Wasserstein distance on the joint distribution of the vector of random measures are discussed Catalano et al. (2021, 2024). Note that, in case the pairwise correlation is 1 for any set $A \in \mathcal{X}$, it is sufficient to consider random measures having the same marginal distribution in order to prove that the random variables $\tilde{\mu}_1(A), \ldots, \tilde{\mu}_D(A)$ are equal almost-surely; on the other hand, even if a pairwise correlation equal to 0 does not imply, in general, that the random variables $\tilde{\mu}_1(A), \ldots, \tilde{\mu}_D(A)$ are mutually independent, this statement actually holds for many specific models. Both these fundamental properties hold for the hierarchical structures considered in this chapter, namely,

$$\tilde{\boldsymbol{\mu}}^{(1)} = \big(\tilde{\mu}_1^{(1)}, \ldots, \tilde{\mu}_D^{(1)}\big) \mid \tilde{\mu}_0 \overset{\text{i.i.d.}}{\sim} \mathrm{CRM}\big(\tilde{\nu}_{\text{norm}}\big), \tag{2.19}$$

$$\tilde{\boldsymbol{\mu}}^{(2)} = \big(\tilde{\mu}_1^{(2)}, \ldots, \tilde{\mu}_D^{(2)}\big) \mid \tilde{\mu}_0 \overset{\text{i.i.d.}}{\sim} \mathrm{CRM}\big(\tilde{\nu}\big), \tag{2.20}$$

where the conditionally independent CRMs have random Lévy intensity measures, respectively,

$$\tilde{\nu}_{\text{norm}}(ds, dx) = \rho(ds)\, \frac{\tilde{\mu}_0(dx)}{\tilde{\mu}_0(\mathbb{X})} = \rho(ds)\, \tilde{p}_0(dx), \qquad \tilde{\nu}(ds, dx) = \rho(ds)\, \tilde{\mu}_0(dx),$$

and $\tilde{\mu}_0$ is itself a CRM with Lévy intensity measure $\nu_0(ds, dx) = \rho_0(ds)\, P_0(dx)$. The previous sections have considered a normalization for the random measures in (2.19), and a hazard mixture transformation for the random measures in (2.20). This section rather focuses on the comparison between their dependence structures at the level of the random measures, irrespective of the particular choice of transformation: indeed, even though focusing on the transformed random measures would be of interest as well, a direct comparison between (2.19) and (2.20) can help in disentangling the effects of the hierarchical constructions from the effects of the transformations.

		$\tilde{\boldsymbol{\mu}}^{(1)}$ as in (2.19)	$\tilde{\boldsymbol{\mu}}^{(2)}$ as in (2.20)
mean:	$\mathbb{E}\big[\tilde{\mu}_i(A)\big]$	$\theta P_0(A)$	$\theta\,\theta_0\,P_0(A)$
variance:	$\mathrm{Var}\big(\tilde{\mu}_i(A)\big)$	$\dfrac{\theta^2\,P_0(A)(1-P_0(A))}{1+\theta_0}+\theta\,P_0(A)$	$\theta\,(\theta+1)\,\theta_0\,P_0(A)$
covariance:	$\mathrm{Cov}\big(\tilde{\mu}_i(A),\,\tilde{\mu}_j(A)\big)$	$\dfrac{\theta^2\,P_0(A)(1-P_0(A))}{1+\theta_0}$	$\theta^2\,\theta_0\,P_0(A)$
correlation:	$\mathrm{Cor}\big(\tilde{\mu}_i(A),\,\tilde{\mu}_j(A)\big)$	$\dfrac{\theta\,(1-P_0(A))}{\theta\,(1-P_0(A))+1+\theta_0}$	$\dfrac{\theta}{1+\theta}$

Table 2.1: Expressions for the mean, variance, covariance and correlation, as defined in (2.21), for the hierarchical structures of random measures in (2.19) and (2.20), in case of the hierarchical gamma specification in (2.22).

A decisive advantage of the pairwise covariance (and correlation) is represented by its plain evaluation for hierarchical models. In particular, considering the vector of random measures $\tilde{\boldsymbol{\mu}}^{(1)} = \big(\tilde{\mu}_1^{(1)},\ldots,\tilde{\mu}_D^{(1)}\big)$ defined in (2.19), for any fixed set $A \in \mathcal{X}$ and $i \neq j$,

$$\begin{aligned}
\mathbb{E}\big[\tilde{\mu}_i^{(1)}(A)\big] &= \tau(1;0)\,\mathbb{E}[\tilde{p}_0(A)],\\
\mathrm{Var}\big(\tilde{\mu}_i^{(1)}(A)\big) &= \tau(1;0)^2\,\mathrm{Var}(\tilde{p}_0(A)) + \tau(2;0)\,\mathbb{E}[\tilde{p}_0(A)],\\
\mathrm{Cov}\big(\tilde{\mu}_i^{(1)}(A),\,\tilde{\mu}_j^{(1)}(A)\big) &= \tau(1;0)^2\,\mathrm{Var}(\tilde{p}_0(A)),
\end{aligned} \tag{2.21}$$

where the cumulants τ's are defined as in (2.7); the corresponding expressions for the vector $\tilde{\boldsymbol{\mu}}^{(2)} = \big(\tilde{\mu}_1^{(2)},\ldots,\tilde{\mu}_D^{(2)}\big)$ in (2.20) are obtained by replacing the normalized random measure $\tilde{p}_0(A)$ with its non-normalized counterpart $\tilde{\mu}_0(A)$. Note that, at the root level of the hierarchy, the mean and the variance of both $\tilde{\mu}_0$ and its normalization $\tilde{p}_0$ can be expressed in terms of their Lévy intensity measure ν_0 (James et al., 2006). For illustration purposes, consider the hierarchical *gamma* process, where random measures at both levels of the hierarchies are gamma CRM, corresponding to

$$\rho(ds) = \theta\,s^{-1}\,e^{-s}\,ds, \qquad \rho_0(ds) = \theta_0\,s^{-1}\,e^{-s}\,ds; \tag{2.22}$$

with such specifications, the quantities appearing in (2.21) for both hierarchical structures in (2.19) and (2.20) can be computed explicitly, and are collected in Table 2.1. Specifically, the cumulants are specified by $\tau(m;0) = \theta\,\Gamma(m)$, while, at the root of the hierarchy,

$$\begin{aligned}
\mathbb{E}[\tilde{\mu}_0(A)] &= \mathrm{Var}(\tilde{\mu}_0(A)) = \theta_0\,P_0(A),\\
\mathbb{E}[\tilde{p}_0(A)] &= P_0(A), \qquad \mathrm{Var}(\tilde{p}_0(A)) = \frac{P_0(A)(1-P_0(A))}{1+\theta_0}.
\end{aligned} \tag{2.23}$$

In order to correctly interpret the information contained in (2.21) and specified in Table 2.1,

consider two different flexibility properties that are desirable for the dependence structure of a model that induces positive association between the random measures. The first kind of flexibility ensures that, for every value $\gamma \in [0,1]$, there exists a specification of the model parameters such that the random measures have correlation equal to (or converging to) γ. This property holds for hierarchical models in general, and can be easily checked for the hierarchical gamma process considered in (2.22): in both cases, the values of θ and possibly θ_0 can be chosen so that the correlations are equal to (or converge to) every fixed value $\gamma \in [0,1]$. The second, and stronger, kind of flexibility requires that, for every marginal distribution of the random measures and for every value $\gamma \in [0,1]$, there exists a specification of the model parameters such that the random measures have correlation equal to (or converging to) γ. This kind of flexibility ensures that the marginal distribution of the random measures can be modeled separately from their dependence structure, a feature which is often desirable in practice, as they encode different aspects of the model. For simplicity, it is usually sufficient to restrict to a weaker version of this second flexibility property, whereby only the first and second moments of the random measures are fixed, instead of their whole marginal distributions.

Interestingly, most hierarchical models proposed in the literature do not achieve this second type of flexibility. For example, consider the vector of random measures $\tilde{\boldsymbol{\mu}}^{(2)}$ defined in (2.20), with the hierarchical gamma specification described in (2.22). As revealed by the expression of the correlation, in order to recover perfectly correlated random measures, and thus perfect dependence, one needs $\theta \to +\infty$; however, in such case, the mean of the marginal distributions diverges. This observation suggests that a good practice for hierarchical gamma random measures, in case the random measure at the root of the hierarchy is not normalized, is to fix $\theta_0 = 1/\theta$, so that $\mathbb{E}\big(\tilde{\mu}_i^{(2)}(A)\big) = P_0(A)$ and thus the dependence structure does not affect the mean of the random measure. Nevertheless, with such choice of parameters, one obtains $\mathrm{Var}\big(\tilde{\mu}_i^{(2)}(A)\big) = (\theta+1)\, P_0(A)$, which in turn implies that the only situation in which perfectly correlated random measures are recovered entails a (marginally) infinite variance: in conclusion, the flexibility of second kind cannot be achieved for the hierarchical structure in (2.20), when the hierarchical gamma specification in (2.22) is considered. On the other hand, such issues do not arise for the vector of random measures $\tilde{\boldsymbol{\mu}}^{(1)}$ defined in (2.19), as clarified by comparing the expressions for the variances and covariances in Table 2.1 and (2.23): indeed, if $\tilde{\mu}_0$ is a gamma random measure, its mean and variance coincide, whereas the variance of the normalized random measure $\tilde{p}_0$ can be adjusted separately from its mean, leveraging on the parameter θ_0. This observation suggests to consider other specifications for the random measures $\tilde{\mu}_0$, for which a further hyper-parameter can be exploited to flexibly account for different values of the variance.

In conclusion, when resorting to hierarchical constructions for modeling the dependence between random measures, the elicitation of the dependence structure requires particular attention, as adjusting for such dependence may also affect the marginal distributions. For the same reason, the covariance does not represent a reliable measure of dependence: given that changing the

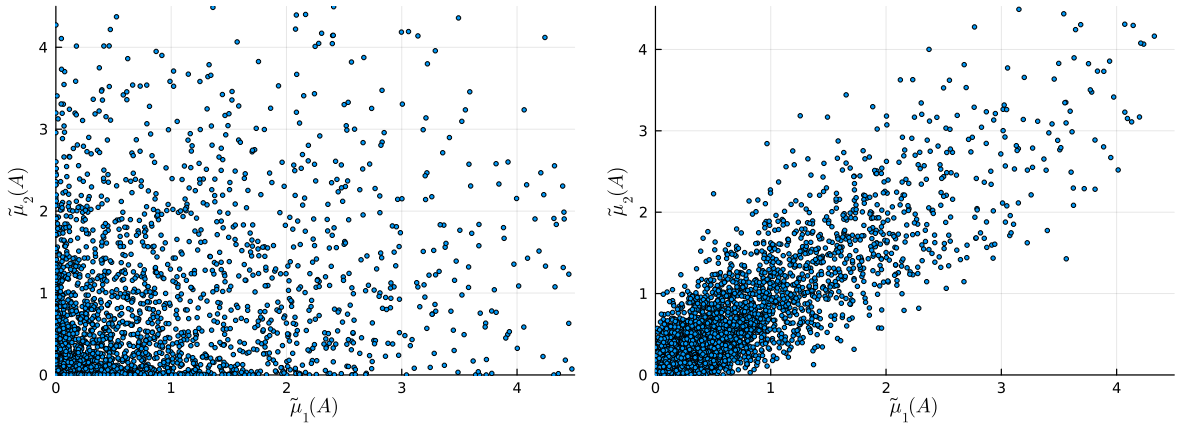

Figure 2.1: Samples from the bivariate vector of random variables $\tilde{\boldsymbol{\mu}}^{(2)}(A) = \big(\tilde{\mu}_1^{(2)}(A), \tilde{\mu}_2^{(2)}(A)\big)$, for parameters $\beta = 1$ (left) and $\beta = 9$ (right), where $A \in \mathcal{X}$ is such that $P_0(A) = 0.5$; the value of the covariance is the same for every $\beta > 0$ and equals $P_0(A) = 0.5$, while the correlation equals 0.5 for $\beta = 1$ (left) and 0.9 for $\beta = 9$ (right).

covariance also affects the variance, the normalization required to compute the correlation is not only a way to obtain values in $[0, 1]$, but additionally provides important information about the dependence structure. In order to effectively support this argument, consider the bivariate vector of hierarchical random measures $\tilde{\boldsymbol{\mu}}^{(2)} = \big(\tilde{\mu}_1^{(2)}, \tilde{\mu}_2^{(2)}\big)$ defined in (2.20), and impose gamma specifications for the Lévy intensity measures,

$$\rho(ds) = \beta\, s^{-1}\, e^{-\beta s}\, ds, \qquad \rho_0(ds) = s^{-1}\, e^{-s}\, ds,$$

with parameter $\beta > 0$: it easily follows from the expressions in (2.21) that, for each $A \in \mathcal{X}$, the covariance remains unchanged and equal to $P_0(A)$, for every value of β. On the other hand, the dependence structure of the vector of random measures appears substantially different for different values of β, as clearly shown in Figure 2.1 for the cases $\beta = 1$ and $\beta = 9$; this difference is correctly detected by the correlation, which coincides with $\beta/(1 + \beta)$ and thus equals 0.5 in the first case and 0.9 in the second case.

Chapter 3

Hierarchically dependent mixture hazards for modelling competing risks

A popular approach in Bayesian modelling of non-exchangeable data relies on the specification of hierarchical nonparametric priors, which induce dependence across groups of observations. In the analysis of grouped survival data, subject to a single disease or failure, hierarchies of completely random measures have been used as mixing measures to model multivariate dependent mixture hazard rates (Chapter 2). This chapter shows that such modelling approach can be recast to tackle a competing risks scenario, in which groups correspond to different diseases or causes of failures affecting each subject: in this setting, the multivariate construction acts at a latent level, as only the minimum time-to-event and the corresponding cause of death or failure are actually observed. The posterior hierarchy of random measures, as well as the posterior estimates of both survival function and cause-specific incidence functions are explicitly determined, conditionally on a suitable latent partition that admits a characterization in terms of a novel variant of the Chinese restaurant franchise process. Moreover, the derivation of *prediction curves*, namely the predictive probabilities for a future event to be due to each possible cause, as a function of the time at which the event occurs, represents a major contribution of this chapter. These results are pivotal for devising marginal and conditional sampling algorithms, which are tested on a synthetic dataset in order to assess their effectiveness. The performances of this proposal are also compared with those of its non-hierarchical counterpart, which models hazard rates independently for each disease. Finally, some applications to clinical datasets are discussed.

3.1 Competing risks in survival analysis

In the framework of survival analysis, several different types of events may be of interest to the researchers; for example, in clinical studies, a fatal outcome after treatment may be due to a number of different causes, which may be related or unrelated to the treatment. In situations where the occurrence of an event of interest, and thus the observation of the associated survival time, may be possibly precluded by the occurrence of one of the other events, the distinct event types are referred to as *competing events* or *competing risks* (Kalbfleisch and Prentice, 2002; Geskus, 2024). The primary domain of application for the competing risks framework are biomedical and clinical studies, in which subjects may die by different causes or diagnoses; another common setting concerns industrial life-testing, where the breakdown of a complex system may be due to the failure of one of its components.

In this chapter, the observed survival time is denoted by $T \in \mathbb{R}^+$ and assumed to be a random variable with absolutely continuous distribution (w.r.t. the Lebsegue measure), while $\Delta \in \{1, \dots, D\}$ is the categorical random variable identifying the observed event type. For each event type d, the cause-specific *hazard rate* is defined as

$$h_d(t) := \lim_{s \to 0} \frac{1}{s} \mathbb{P}(t \leq T < t + s,\, \Delta = d \mid T > t), \qquad t \geq 0,$$

and represents the instantaneous rate of occurrence of an event of type d, given survival up to that time point from *all* possible types of events (i.e., individuals experiencing a competing event are no longer at risk). The cause-specific hazard rates act as the main building blocks for several important quantities. For instance, the overall *survival function* depends on the sum of cumulative cause-specific hazard rates:

$$S(t) := \mathbb{P}(T > t) = \exp\left\{ -\sum_{d=1}^{D} \int_0^t h_d(s)\, ds \right\}, \qquad t \geq 0.$$

Similarly, the cause-specific *incidence functions*, or *subdensities*, representing the infinitesimal probability of occurrence of each event type, are defined in terms of the cause-specific hazard rates as

$$f_d(t) := \lim_{s \to 0} \frac{1}{s} \mathbb{P}(t \leq T < t + s,\, \Delta = d) = h_d(t) \exp\left\{ -\sum_{\ell=1}^{D} \int_0^t h_\ell(s)\, ds \right\}, \qquad t \geq 0;$$

the marginal probability of occurrence of each event type d, that is, the proportion π_d of subjects eventually experiencing an event of type d, is the limit of the corresponding *cumulative incidence function*, or *subdistribution*,

$$\pi_d = \mathbb{P}(\Delta = d) = \lim_{t \to \infty} \int_0^t f_d(s)\, ds.$$

The approach to the modeling of competing risks data introduced above, based on the specification of cause-specific hazard rates, is known as *multi-state approach*: indeed, these hazard rates can be regarded as transition rates in multi-state models (Andersen et al., 2002; Putter et al., 2007). Specifically, consider a Markov process having one transient state, termed *alive* and labeled as 0, and D absorbing states, termed *death from cause* d, for $d = 1, \dots, D$; within this setting, the cause-specific hazard rate h_d is the transition intensity from state 0 to state d. An alternative strategy relies on the specification of the subdistribution hazard rates; however, contrary to the hazard rate in standard survival analysis, they are not rates in the classical sense, since individuals experiencing competing events remain in the risks set, and their correct interpretation is not straightforward (Andersen and Keiding, 2012). Another common approach to competing risks data (Crowder, 2012) hypothesizes the existence of latent, or *potential*, survival times, one for each different event type, which are assumed to be almost-surely distinct; the observable random variables (T, Δ) are therefore the minimum of such survival times and the corresponding event type. This approach may appear rather intuitive, but has been extensively criticized for lack of plausibility and interpretability in biomedical applications, as one assumes the eventual occurrence of each type of event in each individual (Geskus, 2024). Moreover, from a mathematical perspective, the joint distribution of these latent survival times is not fully identifiable: more precisely, their marginal distributions cannot be identified, unless further assumptions on their dependence structure are imposed (Kalbfleisch and Prentice, 2002). In other words, even though a dependent parametric model is considered for the multivariate distribution of potential survival times, and model parameters are estimable, it is impossible to discern this model from a model with independent risks on the basis of observations (Cox, 1959; Tsiatis, 1975; Crowder, 1991). For this reason, a number of authors suggest focusing on the estimation of observable quantities only, and independence of the competing events is often assumed for mathematical convenience and interpretability. Nevertheless, this approach is particularly suited to the construction of simulated datasets (Beyersmann et al., 2009) (see Section 3.7).

The statistical analysis of competing risks data typically focuses on: (i) the estimation the cause-specific cumulative incidence functions, which allow to estimate the probabilities that each considered competing event occurs within some time interval; (ii) the investigation of the association of these quantities with different treatments or predictors of interest (Fine and Gray, 1999); (iii) the estimation of the overall survival function, or, equivalently, of the overall cumulative incidence function. Such statistical problems have received lengthily and widespread attention within the frequentist literature, as shown by the vast number of comprehensive reviews and textbooks; among them, Kalbfleisch and Prentice (2002), Lawless (2003), Crowder (2012) and Geskus (2015). On the other hand, Bayesian contributions on competing risks are less organically structured. In particular, Bayesian nonparametric models proposed in the literature are often based on gamma or beta process priors, following their successful adoption for modelling univariate survival data (Doksum, 1974; Dykstra and Laud, 1981; Lo and Weng, 1989; Hjort, 1990). Recently, Arfé et al. (2019) introduced a generalization of the beta-Stacy process (Walker

and Muliere, 1997) as prior distribution for cumulative incidence functions, whereas Lau and Cripps (2022) defined a mixture model for cause-specific hazard rates based on thinned completely random measures; the former contribution considers survival times on a discrete time scale, in contrast to the continuous survival times assumed in the latter. On the other hand, Xu et al. (2020) propose a model for estimating treatment effects with semi-competing risks which is based on the dependent Dirichlet process. An alternative nonparametric and flexible proposal, relying on Bayesian Additive Regression Trees is presented in Sparapani et al. (2020).

The standard frequentist inferential procedure based on the renowned Kaplan-Meier method, which allows for simple nonparametric estimation of the survival function in presence of right-censored observations, can be readily adapted to competing risks data (Kalbfleisch and Prentice, 2002). Specifically, the Nelson-Aalen estimator and the related Aalen-Johansen estimator are considered for exploratory analyses of cumulative cause-specific hazard rates and incidence functions, respectively (Allignol et al., 2011); however, the estimation of the standard error for the Aalen-Johansen estimator, and the consequent construction of confidence intervals, is not straightforward, and several estimators have been proposed (Geskus, 2024).

This chapter proposes a Bayesian nonparametric model for competing risks data with continuous survival times. The different competing events are characterized by conditionally independent and absolutely continuous distributions; in particular, their cause-specific hazard rates are specified as kernel mixtures with respect to almost-surely discrete random measures. An overview on this construction and its widespread adoption in Bayesian nonparametric literature is proposed in Section 1.2 and further discussed in Section 3.2. A hierarchical structure of completely random measures, recently introduced by Camerlenghi et al. (2021) and extensively discussed in Chapter 2, is considered as prior distribution on the mixing random measures. This prior choice allows for substantial flexibility in modelling the different hazard rates, without further restrictive assumptions on their relationships; moreover, it induces dependence among them through conditional independence, given a common random measure, and thus promotes borrowing of information whenever hazard rates share common features. Section 3.3 introduces additional sequences of latent variables, accounting for a nested partition structure similar to the one discussed in Section 2.3, which play a fundamental role in the characterization of the posterior distribution of the hierarchical random measures, as well as in the derivation of conditional posterior estimates of both the overall survival function and cause-specific incidence functions (Section 3.5). A major breakthrough in terms of prediction is achieved in Section 3.4 thanks to such principled and comprehensive framework, namely the introduction of a novel functional of interest, termed *prediction curve*, which corresponds to the conditional probability, given the past observations, that a future event is a certain type, as a function of the time at which such event occurs. Besides their foundational and methodological interest, prediction curves may potentially be useful in practice for decision making, given their straightforward interpretability. A marginal Gibbs sampling scheme for sequentially updating the latent partition structure

is devised in Section 3.6: unconditioned posterior estimates of quantities of interest are then computed by empirical marginalization of the conditional estimates obtained at each step of the MCMC algorithm. A procedure for sampling directly from the posterior distribution of hierarchical random measures is also presented as an alternative approach for computing posterior estimates and quantifying the uncertainty around them. Both methods are tested on a synthetic dataset in order to assess their effectiveness; moreover, the performances of the proposed model are compared with those of the corresponding Bayesian nonparametric model which considers independent hazard rates, by means of a simulation study (Section 3.7). Finally, an application to two clinical dataset are presented for illustrative purposes (Section 3.8).

3.2 Modelling mixture hazard rates via random measures

In a Bayesian nonparametric framework, random measures represent the building block to define prior laws over functions characterizing the multivariate distribution of cause-specific survival times. Specifically, when this distribution is assumed to be absolutely continuous, the cause-specific (random) hazard functions can be effectively modeled as mixtures of a deterministic smoothing kernel $k(\cdot;\cdot)$ over random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_D$ as

$$\tilde{h}_d(t) = \int_{\mathbb{X}} k(t;x)\, \tilde{\mu}_d(dx), \qquad d = 1, \ldots, D. \tag{3.1}$$

Such mixture structure was introduced, for modelling single-cause (exchangeable) survival data, in the seminal papers by Dykstra and Laud (1981) and Lo and Weng (1989) and is thoroughly discussed in Section 1.2. In the context of multi-sample (partially exchangeable) data, the random hazard specification in (3.1) is assumed in Lijoi et al. (2014) and Camerlenghi et al. (2021), in which dependence across different samples is introduced at the level of random measures; see Section 2.2 for further details. This chapter considers the hierarchical CRMs prior proposed in Camerlenghi et al. (2021) and detailed in (2.4).

The overall survival function is expressed in terms of the cause-specific hazard rates, and thus of the random measures, as

$$\tilde{S}(t) = \exp\left\{ -\sum_{d=1}^{D} \int_0^t \int_{\mathbb{X}} k(s;x)\, \tilde{\mu}_d(dx)\, ds \right\}; \tag{3.2}$$

as discussed in Camerlenghi et al. (2021), and pointed out in Section 1.2 for exchangeable survival data, such survival function is guaranteed to be proper, that is $\lim_{t\to\infty} \tilde{S}(t) = 0$ almost surely, whenever the hierarchical CRMs $\tilde{\mu}_1, \ldots, \tilde{\mu}_D$ are infinitely active and

$$\int_{\mathbb{R}^+} k(s;x)\, ds = \infty, \qquad P_0 - \text{a.s.} \tag{3.3}$$

The random mixture hazard rates (3.1) and the related random survival function (3.2) conveniently define nonparametric priors within a Bayesian statistical model for competing risks data. Specifically, as discussed in Section 3.1, consider a sequence of exchangeable observations, each consisting of a pair of values

$$(T_i,\, \Delta_i), \qquad i = 1, \dots, n,$$

where $T_i \in \mathbb{R}^+$ is the survival time of the observed (and thus firstly occurred) competing event, and $\Delta_i \in \{\, 1, \dots, D \,\}$ represents the corresponding event type. This work discusses the Bayesian nonparametric model

$$\begin{aligned} (T_1, \Delta_1), \dots, (T_n, \Delta_n) \mid \tilde{\boldsymbol{\mu}} &\overset{\text{i.i.d.}}{\sim} \tilde{p}, \\ \tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \dots, \tilde{\mu}_D) &\sim \mathcal{Q}, \end{aligned} \tag{3.4}$$

where $\mathcal{Q}$ is the prior distribution over the vector of random measures, and the directing random measure $\tilde{p}$ is defined from such random measures as

$$\tilde{p}(dt, \delta) = \int_{\mathbb{X}} k(t; x)\, \tilde{\mu}_\delta(dx)\, \exp\left\{ -\sum_{d=1}^{D} \int_0^t \int_{\mathbb{X}} k(s; x)\, \tilde{\mu}_d(dx)\, ds \right\} dt. \tag{3.5}$$

Note that the random measure $\tilde{p}$ is a probability measure on the product space $\mathbb{R}^+ \times \{\, 1, \dots, D \,\}$, and its first component is absolutely continuous with respect to the Lebesgue measure, almost-surely under $\mathcal{Q}$. As anticipated above, the prior distribution $\mathcal{Q}$ considered in this chapter is the hierarchical structure of CRMs proposed by Camerlenghi et al. (2019) and introduced in (2.4), namely

$$\begin{aligned} \tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \dots, \tilde{\mu}_D) \mid \tilde{\mu}_0 &\overset{\text{i.i.d.}}{\sim} \text{CRM}(\tilde{\nu}), \\ \tilde{\mu}_0 &\sim \text{CRM}(\nu_0), \end{aligned}$$

where $\tilde{\mu}_1, \dots, \tilde{\mu}_D$ are conditionally independent CRMs with random and homogeneous Lévy intensity measure $\tilde{\nu}(ds, dx) = \rho(ds)\, \tilde{\mu}_0(dx)$, and $\tilde{\mu}_0$ is the CRM at the root of the hierarchy, with homogeneous Lévy intensity measure $\nu_0(ds, dx) = \rho_0(ds)\, \theta P_0(dx)$; in these expressions, θ is a positive parameter and P_0 is a diffuse probability measure on $\mathbb{X}$. As a consequence of the random measures $\tilde{\mu}_1, \dots, \tilde{\mu}_D$ being identically distributed, the prior induced on the probabilities $(\pi_d)_d$ for a subject to eventually experience each competing event is uniform, that is, $\pi_d = \mathbb{P}(\Delta = d) = 1/D$ for each $d = 1, \dots, D$; different models in which random measures $\tilde{\mu}_1, \dots, \tilde{\mu}_D$ are conditionally independent but not identically distributed may be considered, so that the induced prior on $(\pi_d)_d$ is more informative.

The rest of the chapter is devoted to the posterior characterization of the hierarchically dependent random measures $\tilde{\mu}_1, \dots, \tilde{\mu}_D$, with the aim of evaluating Bayesian estimates of the overall survival function, cause-specific incidence functions, and prediction curves, as well as quantifying the uncertainty around those estimates.

3.3 Latent partition structure

Hierarchical nonparametric constructions are known to induce a random nested partition structure on the observations, which is possibly latent when random measures are smoothed via kernel mixtures; a thorough description of such latent partition structure is proposed in Section 2.3. In this section, the introduction of additional sequences of latent variables, accounting for the two-level partition structure, follows the approach proposed in Chapter 2. These latent variables play a crucial role in characterizing the posterior distribution of the hierarchical random measures, as discussed in Section 3.5.

Consider a sequence of exchangeable competing risks observations from the model in (3.4) and let $\boldsymbol{T} = (T_1, \dots, T_n)$ and $\boldsymbol{\Delta} = (\Delta_1, \dots, \Delta_n)$; for simplicity, it is here assumed that observed survival times are not censored, even though the accommodation of censored data is straightforward (see, e.g., James (2005); Lijoi and Nipoti (2014), and Section 3.8). The associated multiplicative intensity likelihood function (Kalbfleisch and Prentice, 2002) is

$$\mathcal{L}(\tilde{\mu}_1, \dots, \tilde{\mu}_D; \boldsymbol{T}, \boldsymbol{\Delta}) = \prod_{i=1}^{n} \int_{\mathbb{X}} k(T_i; x)\, \tilde{\mu}_{\Delta_i}(dx) \, \exp\left\{ - \sum_{d=1}^{D} \int_0^{T_i} \int_{\mathbb{X}} k(s; x)\, \tilde{\mu}_d(dx)\, ds \right\}, \tag{3.6}$$

which corresponds to the product of the directing random measure (3.5) evaluated at the observations. The same expression can be restated in terms of the extended random measures $\tilde{\mu}^e_1, \dots, \tilde{\mu}^e_D$ introduced in Section 2.3 as

$$\begin{aligned} &\mathcal{L}(\tilde{\mu}^e_1, \dots, \tilde{\mu}^e_D; \boldsymbol{T}, \boldsymbol{\Delta}) \\ &\qquad = \prod_{i=1}^{n} \int_{[0,1]\times\mathbb{X}} k(T_i; x)\, \tilde{\mu}^e_{\Delta_i}(dz, dx) \, \exp\left\{ - \sum_{d=1}^{D} \int_0^{T_i} \int_{[0,1]\times\mathbb{X}} k(s; x)\, \tilde{\mu}^e_d(dz, dx)\, ds \right\}. \end{aligned}$$

A common approach when dealing with mixture models suggests to remove the integration by introducing suitable sequences of latent variables, which essentially correspond to the latent locations sampled from the mixing random measures, leading to considerable simplifications in the following computations (James, 2005; Lijoi and Nipoti, 2014; Camerlenghi et al., 2021). Let $\boldsymbol{X} = (X_1, \dots, X_n)$ and $\boldsymbol{Z} = (Z_1, \dots, Z_n)$ be sequences of latent variables, where each $X_i \in \mathbb{X}$ and each $Z_i \in [0, 1]$; the resulting augmented likelihood is

$$\begin{aligned} &\mathcal{L}(\tilde{\mu}^e_1, \dots, \tilde{\mu}^e_D; \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}) \\ &\qquad = \prod_{i=1}^{n} k(T_i; X_i)\, \tilde{\mu}^e_{\Delta_i}(dZ_i, dX_i) \, \exp\left\{ - \sum_{d=1}^{D} \int_0^{T_i} \int_{[0,1]\times\mathbb{X}} k(s; x)\, \tilde{\mu}^e_d(dz, dx)\, ds \right\}. \end{aligned}$$

The almost-sure discreteness of random measures $\tilde{\mu}^e_1, \dots, \tilde{\mu}^e_D$ naturally induces the random nested partition structure mentioned above, which is encoded into the latent sequences $\boldsymbol{X}$ and $\boldsymbol{Z}$ through

tied values within each sequence. Specifically, the coarser level of the partition is induced by elements in $\boldsymbol{X}$, which assume k distinct values, $X_1^*, \ldots, X_k^*$, with multiplicities $n_1 + \cdots + n_k = n$, respectively. Resorting to the characterization in (2.6), these distinct values can be regarded as the latent locations from the random measure $\tilde{\mu}_0$ at the root of the hierarchy; since locations are shared among the random measures at the lower level of the hierarchy, the same location can be associated to observations from different random measures, that is, $X_i = X_\ell$ does not necessarily imply $\Delta_i = \Delta_\ell$. In other words, there is no structural relationship between the random latent partition induced by tied values in $\boldsymbol{X}$ and the observed partition encoded into $\boldsymbol{\Delta}$.

On the contrary, this relationship is accounted for at the finer level of the partition, induced by elements in $\boldsymbol{Z}$. More precisely, for each random measure d and each latent group j, consider the n_{dj} observations for which $\Delta_i = d$ and $X_i = X_j^*$; the corresponding elements in $\boldsymbol{Z}$ assume r_{dj} distinct values, $Z_{dj1}^*, \ldots, Z_{djr_{dj}}^*$, with multiplicities $q_{dj1} + \cdots + q_{djr_{dj}} = n_{dj}$, respectively. With reference to (2.6), these distinct values can be regarded as the latent marks from the random measure $\tilde{\mu}_d^e$ matched with the same location X_j^* from $\tilde{\mu}_0$. Moreover, for each latent group j, denote by r_j the partial sum of elements in $(r_{dj})_{dj}$ with respect to d, that is, the total number of different marks across random measures $\tilde{\mu}_1^e, \ldots, \tilde{\mu}_D^e$ associated with the same location. Notice that, whenever observations are associated with the same mark, i.e. $Z_i = Z_\ell$, they are also associated with the same location, i.e. $X_i = X_\ell$, while the converse is not necessarily true; stated differently, a tie between values in $\boldsymbol{Z}$ implies a tie between corresponding values in $\boldsymbol{X}$, that is, the two random partitions are nested.

In light of the partition structure described above, the augmented likelihood function can be expressed as

$$\begin{aligned}
&\mathcal{L}(\tilde{\mu}_1^e, \ldots, \tilde{\mu}_D^e; \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}) \\
&\qquad = Q(\boldsymbol{T}, \boldsymbol{X}) \, \exp\left\{ - \sum_{d=1}^{D} \int_{[0,1]\times\mathbb{X}} K_n(x) \, \tilde{\mu}_d^e(dz, dx) \right\} \prod_{d=1}^{D} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \tilde{\mu}_d^e(dZ_{djh}^*, dX_j^*)^{q_{djh}}, \quad (3.7)
\end{aligned}$$

where, in order to ease the notation, the following quantities have been defined:

$$Q(\boldsymbol{T}, \boldsymbol{X}) = \prod_{j=1}^{k} \prod_{\{i:\, X_i = X_j^*\}} k(T_i; X_j^*), \qquad K_n(x) = K_n(x; \boldsymbol{T}) = \sum_{i=1}^{n} \int_0^{T_i} k(s; x) \, ds. \quad (3.8)$$

The expression in (3.7) allows to conveniently isolate the contributions of the observed survival times and event types, and of the sequences of latent variables, to the joint likelihood function. Specifically, survival times $\boldsymbol{T}$ enter the likelihood through quantities $Q(\boldsymbol{T}, \boldsymbol{X})$ and $K_n(x; \boldsymbol{T})$, while event types $\boldsymbol{\Delta}$ enforce constraints on the possible configurations of the latent nested partition structure, which is encoded into multiplicities $(q_{djh})_{djh}$ at its finer $\boldsymbol{Z}$-level and $(r_{dj})_{dj}$ at its coarser $\boldsymbol{X}$-level.

The joint probability distribution of the exchangeable competing risks observations and latent variables introduced in this section is obtained by integrating the likelihood function in (3.7) with respect to the hierarchical prior distribution $\mathcal{Q}$; its derivation closely follows the unifying recursive approach discussed in Chapter 2. For this purpose, let ψ and τ be the Laplace exponent and cumulants of the random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_D$ at the lower level of the hierarchy, as defined in (2.7),

$$\psi(u) = \int_{\mathbb{R}^+} (1 - e^{-us})\, \rho(ds), \qquad \tau(m; u) = \int_{\mathbb{R}^+} s^m\, e^{-us}\, \rho(ds); \tag{3.9}$$

the corresponding quantities ψ_0 and τ_0 for the random measure $\tilde{\mu}_0$ at the root of the hierarchy are defined replacing ρ with ρ_0.

Proposition 3.1. *The joint marginal distribution of observations* $(\boldsymbol{T}, \boldsymbol{\Delta})$ *and latent variables* $(\boldsymbol{X}, \boldsymbol{Z})$ *is*

$$\begin{aligned}\mathbb{P}(\boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}) = Q(\boldsymbol{T}, \boldsymbol{X})\, \exp\left\{ - \int_{\mathbb{X}} \psi_0\big(D\, \psi(K_n(x))\big)\, \theta P_0(dx) \right\} \\ \times \prod_{j=1}^{k} \left(\prod_{d=1}^{D} \prod_{h=1}^{r_{dj}} \tau(q_{djh}; K_n(X_j^*))\, H(dZ_{djh}^*) \right) \tau_0\big(r_j; D\, \psi(K_n(X_j^*))\big)\, \theta P_0(dX_j^*).\end{aligned}$$

Note that the random partition induced by ties in the latent sequence of marks $\boldsymbol{Z}$, which is encoded into multiplicities $(q_{djh})_{djh}$ and denoted hereinafter by $\Pi_{\boldsymbol{Z}}$, is solely relevant in this expression, while their distinct values, denoted by $\boldsymbol{Z}^*$, are conditionally independent and sampled from the arbitrary probability measure H, and can be marginalized out. On the contrary, both the random partition induced by ties in the sequence of locations $\boldsymbol{X}$ and their distinct values, denoted by $\Pi_{\boldsymbol{X}}$ and $\boldsymbol{X}^*$, respectively, enter the marginal distribution; in particular, the distinct values $X_1^*, \ldots, X_k^*$ are tied to the observed survival times through the kernel product term $Q(\boldsymbol{T}, \boldsymbol{X})$. Therefore, upon marginalization with respect to the distinct values $\boldsymbol{Z}^*$, the result in Proposition 3.1 can be restated as

$$\begin{aligned}\mathbb{P}(\boldsymbol{T}, \boldsymbol{\Delta}, \Pi_{\boldsymbol{X}}, \boldsymbol{X}^*, \Pi_{\boldsymbol{Z}}) = Q(\boldsymbol{T}, \boldsymbol{X})\, \exp\left\{ - \int_{\mathbb{X}} \psi_0\big(D\, \psi(K_n(x))\big)\, \theta P_0(dx) \right\} \\ \times \prod_{j=1}^{k} \left(\prod_{d=1}^{D} \prod_{h=1}^{r_{dj}} \tau(q_{djh}; K_n(X_j^*)) \right) \tau_0\big(r_j; D\, \psi(K_n(X_j^*))\big)\, \theta P_0(dX_j^*).\end{aligned} \tag{3.10}$$

This expression represents the cornerstone of computational developments, discussed in Section 3.6: indeed, full conditional distributions obtained from (3.10) are exploited to devise a marginal Gibbs sampling scheme.

The specification of Proposition 3.1 in the special case of *generalized gamma* hierarchical CRM and Dykstra and Laud (1981) kernel choice concludes this section.

Example 3.2. Consider as hierarchical prior distribution $\mathcal{Q}$ the hierarchical *generalized gamma*

CRM introduced in (2.18); its Laplace exponent and cumulants (3.9) take the forms

$$\psi(u) = \frac{(\beta+u)^\sigma - \beta^\sigma}{\sigma} \overset{(\sigma=0)}{=} \log\left(1+\frac{u}{\beta}\right), \qquad \tau(m;u) = \frac{\Gamma(m-\sigma)}{\Gamma(1-\sigma)}(\beta+u)^{\sigma-m},$$

with ψ_0 and τ_0 defined replacing β and σ with β_0 and σ_0. Moreover, consider the Dykstra and Laud (1981) kernel with constant function γ, that is $k(t;x) = \gamma\,\mathbb{1}(t \geq x)$, so that the quantities in (3.8) become

$$Q(\boldsymbol{T},\boldsymbol{X}) = \gamma^n \prod_{j=1}^{k} \mathbb{1}\left(\min_{\{i:\, X_i = X_j^*\}} T_i \geq X_j^*\right), \qquad K_n(x) = \gamma \sum_{i=1}^{n} \max(T_i - x,\, 0).$$

The joint marginal distribution in (3.10) boils down to the closed-form expression

$$\begin{aligned}
\mathbb{P}(\boldsymbol{T},\boldsymbol{\Delta},\Pi_{\boldsymbol{X}},\boldsymbol{X}^*,\Pi_{\boldsymbol{Z}}) = {} & \prod_{j=1}^{k}\left(\prod_{d=1}^{D}\prod_{h=1}^{r_{dj}} \frac{\Gamma(q_{djh}-\sigma)}{\Gamma(1-\sigma)}\right)\frac{\Gamma(r_j-\sigma_0)}{\Gamma(1-\sigma_0)} \\
& \times \exp\left\{-\frac{\theta}{\sigma_0}\int_{\mathbb{X}}\left(\left(\beta_0 + \frac{D}{\sigma}\Big((\beta+K_n(x))^\sigma - \beta^\sigma\Big)\right)^{\sigma_0} - \beta_0^{\sigma_0}\right) P_0(dx)\right\} \\
& \times \prod_{j=1}^{k}\left(\beta + \frac{D}{\sigma}\Big((\beta+K_n(X_j^*))^\sigma - \beta^\sigma\Big)\right)^{\sigma_0-r_j} (\beta+K_n(X_j^*))^{r_j\sigma-n_j} \\
& \times \gamma^n\,\theta^k \prod_{j=1}^{k} \mathbb{1}\left(\min_{\{i:\, X_i = X_j^*\}} T_i \geq X_j^*\right) P_0(dX_j^*).
\end{aligned}$$

3.4 Prediction curves

An important by-product of Proposition 3.1 and (3.10) is the possibility to determine the predictive distribution for a future observation, namely the pair of future survival time T_{n+1} and corresponding event type Δ_{n+1}, given the previous observations and latent variables. Moreover, the predictive distribution for the event type Δ_{n+1}, given survival time T_{n+1} and previous observations and latent variables, may be also derived up to its normalizing constant. Such conditional distributions underlie the notion of *prediction curves*, which correspond to the probabilities for a future event to be of each type, as a function of the survival time, namely

$$t \;\mapsto\; \mathbb{P}(\Delta_{n+1} = d \mid T_{n+1} = t, \boldsymbol{T}, \boldsymbol{\Delta}, \Pi_{\boldsymbol{X}}, \boldsymbol{X}^*, \Pi_{\boldsymbol{Z}}), \qquad d = 1,\dots,D; \tag{3.11}$$

prediction curves represent the main contribution proposed in this chapter from both a methodological and practical point of view, as already discussed in Section 3.1.

Consider a conditional distribution of the random variables $(T_{n+1}, \Delta_{n+1}, X_{n+1})$, given the previous observations $(\boldsymbol{T},\boldsymbol{\Delta})$ and latent variables $(\Pi_{\boldsymbol{X}},\boldsymbol{X}^*,\Pi_{\boldsymbol{Z}})$, collectively denoted by $\mathcal{F}_n$; this distribution is absolutely continuous with respect to the Lebesgue measure in its first component,

while the third component is characterized by point masses at $X_1^*, \ldots, X_k^*$, and is absolutely continuous at any other point. Specifically, from (3.10), one obtains (see proof in Appendix A.2)

$$\begin{aligned}
&\mathbb{P}(T_{n+1} \in dt, \Delta_{n+1} = d, X_{n+1} = X_j^* \mid \mathcal{F}_n) \\
&\quad = k(t; X_j^*) \exp\left\{ - \int_{\mathbb{X}} (\psi_0(D\psi(K_{n+1}(y;t))) - \psi_0(D\psi(K_n(y)))) \, \theta \, P_0(dy) \right\} \\
&\qquad \times \prod_{\xi=1}^{k} \frac{\tau_0(r_\xi; D\psi(K_{n+1}(X_\xi^*;t)))}{\tau_0(r_\xi; D\psi(K_n(X_\xi^*)))} \prod_{\ell=1}^{D} \prod_{h=1}^{r_{\ell\xi}} \frac{\tau(q_{\ell\xi h}; K_{n+1}(X_\xi^*;t))}{\tau(q_{\ell\xi h}; K_n(X_\xi^*))} \\
&\qquad\quad \times \left\{ \sum_{h=1}^{r_{dj}} \frac{\tau(q_{djh}+1; K_{n+1}(X_j^*;t))}{\tau(q_{djh}; K_{n+1}(X_j^*;t))} \right. \\
&\qquad\qquad \left. + \tau(1; K_{n+1}(X_j^*;t)) \frac{\tau_0(r_j+1; D\psi(K_{n+1}(X_j^*;t)))}{\tau_0(r_j; D\psi(K_{n+1}(X_j^*;t)))} \right\} dt, \quad (3.12)
\end{aligned}$$

for each $j = 1, \ldots, k$, while for any $x \notin \{ X_1^*, \ldots, X_k^* \}$,

$$\begin{aligned}
&\mathbb{P}(T_{n+1} \in dt, \Delta_{n+1} = d, X_{n+1} \in dx \mid \mathcal{F}_n) \\
&\quad = k(t; x) \exp\left\{ - \int_{\mathbb{X}} (\psi_0(D\psi(K_{n+1}(y;t))) - \psi_0(D\psi(K_n(y)))) \, \theta \, P_0(dy) \right\} \\
&\qquad \times \prod_{j=1}^{k} \frac{\tau_0(r_j; D\psi(K_{n+1}(X_j^*;t)))}{\tau_0(r_j; D\psi(K_n(X_j^*)))} \prod_{\ell=1}^{D} \prod_{h=1}^{r_{\ell j}} \frac{\tau(q_{\ell jh}; K_{n+1}(X_j^*;t))}{\tau(q_{\ell jh}; K_n(X_j^*))} \\
&\qquad\quad \times \tau(1; K_{n+1}(x;t)) \, \tau_0(1; D\psi(K_{n+1}(x;t))) \, \theta \, P_0(dx) \, dt, \quad (3.13)
\end{aligned}$$

where $K_{n+1}(x;t) = K_n(x) + \int_0^t k(s,x)\, ds$ is the updated kernel term. Note that, in case the latent location X_{n+1} does not coincide with any of the distinct values $X_1^*, \ldots, X_k^*$, the expression above does not depend on d; therefore, the observations do not provide any information updating the uniform prior belief on the probabilities of experiencing each competing event. On the other hand, when $X_{n+1} = X_j^*$, the non-informative prior is updated according to the latent partition induced by the sequence $\boldsymbol{Z}$ and encoded into multiplicities (q_{djh}), for $h = 1, \ldots, r_{dj}$.

A predictive distribution for the future observation (T_{n+1}, Δ_{n+1}), given $\mathcal{F}_n$, is obtained from the expressions above by integrating out the latent location X_{n+1}. This distribution is the starting point to determine the predictive distribution of the future event type Δ_{n+1} conditionally on the time T_{n+1} at which such event occurs; the prediction curves defined in (3.11) are based on this latter distribution.

Proposition 3.3. *The predictive distribution of event type Δ_{n+1}, given survival time T_{n+1} and*

previous observations and latent variables, collected into $\mathcal{F}_n$*, is*

$$\mathbb{P}(\Delta_{n+1} = d \mid T_{n+1} = t, \mathcal{F}_n) \propto \sum_{j=1}^{k} k(t; X_j^*) \sum_{h=1}^{r_{dj}} \frac{\tau(q_{djh} + 1; K_{n+1}(X_j^*; t))}{\tau(q_{djh}; K_{n+1}(X_j^*; t))} + \sum_{j=1}^{k} k(t; X_j^*)\, \tau(1; K_{n+1}(X_j^*; t))\, \frac{\tau_0(r_j + 1; D\psi(K_{n+1}(X_j^*; t)))}{\tau_0(r_j; D\psi(K_{n+1}(X_j^*; t)))} + \int_{\mathbb{X}} k(t; x)\, \tau(1; K_{n+1}(x; t))\, \tau_0(1; D\psi(K_{n+1}(x; t)))\, P_0(dx).$$

The analytical form of the prediction curves, including both cause-specific and common terms, reflects the nested partition structure induced by hierarchical random measures; the interplay between these terms determines the way observations shape the prediction curves.

Example 3.4. The joint marginal distribution for the hierarchical *generalized gamma* CRM, introduced in (2.18), is described in Example 3.2; the corresponding prediction curves, for any $d = 1, \ldots, D$, are expressed as

$$\mathbb{P}(\Delta_{n+1} = d \mid T_{n+1} = t, \mathcal{F}_n) \propto \sum_{j=1}^{k} \mathbb{1}(t \geq X_j^*) \left\{ \frac{n_{dj} - r_{dj}\sigma}{\beta + K_{n+1}(X_j^*; t)} + \frac{(r_j - \sigma)\left(\beta + K_{n+1}(X_j^*; t)\right)^{\sigma-1}}{\beta_0 + \frac{D}{\sigma}\left((\beta + K_{n+1}(X_j^*; t))^{\sigma} - \beta^{\sigma}\right)} \right\} + \int_{(0,t]} \left(\beta + K_{n+1}(x; t)\right)^{\sigma-1} \left(\beta_0 + \frac{D}{\sigma}\left((\beta + K_{n+1}(x; t))^{\sigma} - \beta^{\sigma}\right)\right)^{\sigma_0 - 1} P_0(dx),$$

where the updated kernel term is $K_{n+1}(x; t) = K_n(x) + \gamma \max(t - x, 0)$. The expression above depends on the specific event type only though the multiplicities $(n_{dj})_j$ and $(r_{dj})_j$, respectively the number of subjects among those experiencing event type d and the number of distinct latent marks from random measure $\tilde{\mu}_d^e$ associated to each latent location.

3.5 Posterior characterization and estimates

The determination of posterior distributions represents a fundamental step in Bayesian analysis; in this section, the posterior distribution of the hierarchically dependent random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_D$ is explicitly characterized, conditionally on both the observations and the sequences of latent variables, introduced in Section 3.3. Specifically, random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_D$ and $\tilde{\mu}_0$ are still CRMs a posteriori, each consisting of a finite number of random jumps at fixed locations and a non-homogeneous CRM without fixed discontinuity points. Moreover, the hierarchical structure is preserved a posteriori, that is, random measures at the lower level of the hierarchy are conditionally independent given the random measure at the root. This structural conjugacy

property for hierarchical random measures parallels the findings in Camerlenghi et al. (2019) and Camerlenghi et al. (2021), extensively discussed in Section 2.6: indeed, the characterization of their posterior distributions through conditional Laplace transforms is completely analogous.

Proposition 3.5. *The posterior distribution of the random measure $\tilde{\mu}_0$ at the root of the hierarchy, given the observations and the latent variables, collected into $\mathcal{F}_n = (\boldsymbol{T}, \boldsymbol{\Delta}, \Pi_{\boldsymbol{X}}, \boldsymbol{X}^*, \Pi_{\boldsymbol{Z}})$, coincides with the distribution of the random measure*

$$\tilde{\mu}_0^*(dx) + \sum_{j=1}^{k} V_j \, \delta_{X_j^*}(dx),$$

where the random elements in the sum are mutually independent, $\tilde{\mu}_0^*$ is a CRM with non-homogeneous Lévy intensity measure

$$\nu_0^*(ds, dx) = \exp\big\{ - D\psi(K_n(x))\, s\big\}\, \rho_0(ds)\, \theta P_0(dx),$$

and each V_j is a non-negative random variable sampled proportionally to

$$s^{r_j} \, \exp\big\{ - D\psi(K_n(X_j^*))\, s\big\}\, \rho_0(ds), \qquad j = 1, \ldots, k.$$

At the lower level of the hierarchy, the posterior distribution of each random measure $\tilde{\mu}_d$, given observations, latent variables and the random measure $\tilde{\mu}_0$, coincides with the distribution of the random measure

$$\tilde{\mu}_d^*(dx) + \sum_{j=1}^{k} \sum_{h=1}^{r_{dj}} S_{djh} \, \delta_{X_j^*}(dx),$$

where the random elements in the sum are mutually independent, $\tilde{\mu}_d^*$ is a CRM with non-homogeneous Lévy intensity measure

$$\nu^*(ds, dx) = e^{-K_n(x)\, s}\, \rho(ds)\, \tilde{\mu}_0(dx),$$

and each S_{djh} is a non-negative random variable sampled proportionally to

$$s^{q_{djh}} \, e^{-K_n(X_j^*)\, s}\, \rho(ds), \qquad j = 1, \ldots, k, \quad h = 1, \ldots, r_{dj}.$$

This result sheds light on the way the nested partition structure enters the posterior distribution of hierarchical random measures. At the root of the hierarchy, the distribution of each random jump V_j depends on the multiplicity r_j, corresponding to the number of different marks across random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_D$ associated to the latent location X_j^*. On the other hand, at the lower level of the hierarchy, the distribution of each random jump S_{djh} from random measure $\tilde{\mu}_d$ depends on the number of observations q_{djh} associated to the latent mark Z_{djh}^*, which in turn is matched with the latent location X_j^*. In other words, the posterior distribution of each

random measure $\tilde{\mu}_d$ depends on the (finer) partition, induced by the latent sequence $\boldsymbol{Z}$, of the observations for which $\Delta_i = d$, i.e. on the association of observations with marks, while the posterior distribution of the root measure $\tilde{\mu}_0$ depends on the way the finer partition induced by $\boldsymbol{Z}$ is nested into the coarser partition induced by $\boldsymbol{X}$, i.e. on the association of marks with locations.

Furthermore, the components of the posterior random measures featuring random locations, namely random measures $\tilde{\mu}_1^*, \ldots, \tilde{\mu}_D^*$ and $\tilde{\mu}_0^*$, are characterized by non-homogeneous Lévy intensity measures, defined as an exponential tilting of the corresponding prior homogeneous Lévy intensities. Such prior-posterior updating mechanism only depends on the observed survival times $\boldsymbol{T}$ through the function $x \mapsto K_n(x; \boldsymbol{T})$ introduced in (3.8), and is unaffected by the latent partition structure. Note that, at both levels of the hierarchy, the distributions of the random jumps at each fixed latent location X_j^* also depend on the evaluation $K_n(X_j^*)$ of such function at that same location. The non-homogeneity of the posterior Lévy intensity measures is unavoidable for the class of random mixture hazards models (James, 2005; Lijoi and Nipoti, 2014; Camerlenghi et al., 2021; Lau and Cripps, 2022), and has crucial implications from the computational point of view, as detailed in Section 3.6.

Example 3.6. The hierarchical *generalized gamma* CRM, introduced in (2.18) and considered as prior distribution in Examples 3.2 and 3.4, allows for straightforward prior-posterior updates of its parameters, in view of Proposition 3.5. Specifically, the random measures $\tilde{\mu}_1^*, \ldots, \tilde{\mu}_D^*$ and $\tilde{\mu}_0^*$ are still (extended) generalized gamma CRMs, with the same values for σ and σ_0 and updated rate parameters (which are actually functions, due to non-homogeneity a posteriori)

$$\beta^{(\text{post})}(x) = \beta + K_n(x), \qquad \beta_0^{(\text{post})}(x) = \beta_0 + \frac{D}{\sigma}\Big((\beta + K_n(x))^\sigma - \beta^\sigma\Big).$$

Similarly, for each fixed latent location X_j^*, the random jumps $(S_{djh})_{dh}$ and V_j have Gamma distributions

$$S_{djh} \sim \text{Gamma}\big(q_{djh} - \sigma,\, \beta + K_n(X_j^*)\big), \qquad d = 1, \ldots, D, \quad h = 1, \ldots, r_{dj},$$
$$V_j \sim \text{Gamma}\left(r_j - \sigma_0,\, \beta_0 + \frac{D}{\sigma}\Big((\beta + K_n(X_j^*))^\sigma - \beta^\sigma\Big)\right).$$

Therefore, the hierarchical generalized gamma CRM can be regarded as the (conditionally) conjugate prior choice for the model (3.4) considered in this paper. Further considerations on the role of the generalized gamma CRMs as natural conjugate priors for Bayesian nonparametric models based on (hierarchical) random measures can be found in Section 2.7. Moreover, considering the Dykstra and Laud (1981) kernel choice, the prior-posterior updating mechanism is governed by the non-negative and non-increasing function

$$K_n(x) = \gamma \sum_{i=1}^{n} \max(T_i - x,\, 0);$$

such function reaches its maximum at $x = 0$, where $K_n(0) = \gamma \sum_{i=1}^n T_i$, and is null for values of x larger than the largest observed survival time, that is, $K_n(x) = 0$ whenever $x \geq \max_i T_i$. Hence, the deflating effect of the exponential tilting update in the posterior Lévy intensities is larger for smaller values of x, and becomes negligible for larger values of x. Additionally, each of the fixed latent locations $X_1^*, \ldots, X_k^*$ is necessarily smaller than the largest observed survival time, because of the kernel product term $Q(\boldsymbol{T}, \boldsymbol{X})$. As a result, the observations do not provide any information about the region beyond the largest observed survival time, where the prior and the posterior distributions of the random measures coincide. Similar considerations apply to the other kernel choices mentioned in Section 3.1.

The posterior distribution of hierarchical random measures, characterized in Proposition 3.5, is the starting point for Bayesian inference, as posterior estimates of many functionals of $\tilde{\mu}_1, \ldots, \tilde{\mu}_D$ can be determined resorting to it. In particular, within the competing risks setting discussed in this work, fundamental quantities to estimate are the overall survival function in (3.2) and the cause-specific incidence functions in (3.5); their posterior estimates, given both the observations and the latent variables, are obtained by marginalization with respect to the posterior random measures. For this purpose, it is convenient to explicitly define the non-homogeneous jump components of the posterior Lévy intensities as

$$\rho^*(ds \,|\, x) = e^{-K_n(x)\, s}\, \rho(ds), \qquad \rho_0^*(ds \,|\, x) = \exp\big\{ - D\psi(K_n(x))\, s \big\}\, \rho_0(ds).$$

Therefore, the Laplace exponent and cumulants of the posterior random measures $\tilde{\mu}_1^*, \ldots, \tilde{\mu}_D^*$ are

$$\psi^*(u \,|\, x) = \int_{\mathbb{R}^+} (1 - e^{-us})\, \rho^*(ds \,|\, x), \qquad \tau^*(m; u \,|\, x) = \int_{\mathbb{R}^+} s^m\, e^{-us}\, \rho^*(ds \,|\, x); \tag{3.14}$$

again, the corresponding quantities ψ_0^* and τ_0^* for the random measure $\tilde{\mu}_0^*$ are defined replacing ρ^* with ρ_0^*. In the following, the dependence of the above quantities on x is usually omitted, being clear from the context.

Proposition 3.7. *The posterior estimate of the overall survival function $\tilde{S}(t)$ with respect to a quadratic loss, given the observations and the latent variables, for $t > 0$, is*

$$\mathbb{E}\left[\tilde{S}(t) \mid \mathcal{F}_n\right] = \exp\left\{ - \int_{\mathbb{X}} \psi_0^*\big(D\psi^*(K_1(x;t))\big)\, \theta P_0(dx) \right\} \times \prod_{j=1}^{k} \left(\prod_{d=1}^{D} \prod_{h=1}^{r_{dj}} \frac{\tau^*(q_{djh}; K_1(X_j^*; t))}{\tau^*(q_{djh}; 0)} \right) \frac{\tau_0^*\big(r_j; D\psi^*(K_1(X_j^*; t))\big)}{\tau_0^*(r_j; 0)},$$

where $K_1(x;t) = \displaystyle\int_0^t k(s;x)\, ds$ is the integrated kernel up to time $t > 0$.

Proposition 3.8. *For each $\delta = 1, \ldots, D$, the posterior estimate of the cause-specific incidence function $\tilde{p}(dt, \delta)$ with respect to a quadratic loss, given the observations and the latent variables,*

for $t > 0$, *is*

$$\mathbb{E}\left[\tilde{p}(dt,\delta) \mid \mathcal{F}_n\right] = \mathbb{E}\left[\tilde{S}(t) \mid \mathcal{F}_n\right] \left\{ \int_{\mathbb{X}} k(t;x)\, \tau^*(1;K_t(x))\, \tau_0^*\big(1;D\psi^*(K_t(x))\big)\, \theta P_0(dx) + \sum_{j=1}^{k} k(t;X_j^*)\, \tau^*(1;K_t(X_j^*))\, \frac{\tau_0^*\big(r_j+1;D\psi^*(K_t(X_j^*))\big)}{\tau_0^*\big(r_j;D\psi^*(K_t(X_j^*))\big)} + \sum_{j=1}^{k} k(t;X_j^*) \sum_{h=1}^{r_{\delta j}} \frac{\tau^*(q_{\delta jh}+1;K_t(X_j^*))}{\tau^*(q_{\delta jh};K_t(X_j^*))} \right\} dt.$$

Remarkably, the posterior estimates of cause specific incidence functions coincide with the posterior estimate of the directing random measure, which in turn corresponds to the predictive distribution for the future observation (T_{n+1}, Δ_{n+1}), given the previous observations and latent variables; this distribution is extensively discussed in Section 3.4, where it is instead derived from the joint marginal distribution in Proposition 3.1. Although the two expression seemingly differ, they do actually coincide, as shown in Appendix A.2 by means of non-trivial analytical manipulations. As a consequence, the predictive distribution for the future event type given its time of occurrence, obtained in Proposition 3.3 and employed to define prediction curves, can be equivalently derived from Proposition 3.8; specifically, such predictive distribution is proportional to the posterior estimate of the directing random measure.

The closed-form expressions of Bayesian posterior estimates derived in Propositions 3.7 and 3.8 allow for pointwise evaluations of the overall survival function and cause-specific incidence functions, once the hierarchical prior $\mathcal{Q}$ and the kernel $k(\cdot;\cdot)$ are specified (e.g., hierarchical generalized gamma and Dykstra and Laud (1981) kernel, as in Examples 3.2 and 3.6). For their practical estimation, not depending on non-observable quantities, a further marginalization with respect to the posterior distribution of the latent variables, given the observations, is needed; however, such marginalization cannot be performed analytically, and is achieved resorting to a Gibbs sampling scheme, which is the detailed in the next section.

3.6 Gibbs sampling schemes

The latent partition structure described in Section 3.3 represents the fundamental building block for characterizing both marginal and posterior distributions throughout this work; in particular, the posterior distribution of hierarchical random measures, and consequently the posterior estimates of functionals of interest, are derived conditionally on such partition structure. This section is devoted to the description of a Gibbs sampling scheme for updating the sequences of latent variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ within an MCMC procedure; the stream of samples from their posterior distributions can then be exploited to compute unconditioned Bayesian estimates by marginalizing them out via Monte Carlo integration.

The sampling strategy introduced in the following is based on the analytical marginalization of the hierarchical prior distribution, hence belonging to the class of *marginal* methods. Such marginalization induces dependence among elements of the latent sequences: indeed, they are sequentially updated according to an urn scheme. In Bayesian nonparametrics, the adoption of self-reinforced urn schemes to describe predictive distributions dates back to the Blackwell and MacQueen (1973) characterization of the Dirichlet process through a Pólya urn model. However, their practical implementation within MCMC algorithms for posterior inference was popularized by Escobar and West (1995), where a marginal Gibbs sampling scheme is devised for Dirichlet process mixture models (Lo, 1984); this sampling strategy has been generalized to mixtures over stick-breaking priors (Ishwaran and James, 2001) and normalized CRMs (James et al., 2009). The first description of a predictive urn scheme for hierarchical priors, namely hierarchical Dirichlet processes, can be found in Teh et al. (2006), where it is interpreted in light of the *Chinese restaurant franchise* metaphor. The marginal Gibbs sampler presented in this section is an adaptation of the urn schemes developed in Camerlenghi et al. (2019, 2021).

Consider the i-th observation (T_i, Δ_i), meaning that subject i has experienced the competing event of type Δ_i at time T_i, and assume $\Delta_i = d$. The full conditional distribution of the corresponding latent variables X_i and Z_i, given the observations, namely the survival times $\boldsymbol{T}$ and event types $\boldsymbol{\Delta}$, and the other latent variables, conventionally denoted by $\boldsymbol{Z}^{-i}$ and $\boldsymbol{X}^{-i}$, is derived from the joint marginal distribution presented in Proposition 3.1 and restated in (3.10), whose product form greatly simplifies the predictive structure. Specifically, such full conditional distribution can be effectively described in terms of the reallocation of the i-th observation within the latent partition structure, distinguishing three alternative cases.

(1) Both X_i and Z_i display ties with other conditioning latent variables, say $X_i = X_j^*$ and $Z_i = Z_{djh}^*$, with probability

$$\mathbb{P}\left(X_i = X_j^*,\, Z_i = Z_{djh}^* \,|\, \cdots\right) \propto k(T_i; X_j^*)\, \frac{\tau(q_{djh}^{-i} + 1; K_n(X_j^*))}{\tau(q_{djh}^{-i}; K_n(X_j^*))},$$

where q_{djh}^{-i} denotes the number of latent marks in $\boldsymbol{Z}^{-i}$ taking value Z_{djh}^*; in other words, the i-th observation is associated to the latent mark Z_{djh}^* from random measure $\tilde{\mu}_d^e$, which in turn is matched with the latent location X_j^*.

(2) X_i displays a tie with values in $\boldsymbol{X}^{-i}$, say $X_i = X_j^*$, while Z_i assumes a new value, i.e. not included in $\boldsymbol{Z}^{-i}$, with probability

$$\mathbb{P}\left(X_i = X_j^*,\, Z_i = \text{`new'} \,|\, \cdots\right) \propto k(T_i; X_j^*)\, \tau(1; K_n(X_j^*))\, \frac{\tau_0\big(r_j^{-i} + 1; D\psi(K_n(X_j^*))\big)}{\tau_0\big(r_j^{-i}; D\psi(K_n(X_j^*))\big)},$$

where r_j^{-i} denotes the number of distinct marks in $\boldsymbol{Z}^{-i}$ associated with location X_j^*; in other words, the i-th observation is associated to a new latent mark from random measure

$\tilde{\mu}_d^e$, which is matched with the latent location X_j^*.

(3) Both X_i and Z_i assume new values, i.e. not included in $\boldsymbol{X}^{-i}$ and $\boldsymbol{Z}^{-i}$, respectively, with probability

$$\mathbb{P}\left(X_i = \text{'new'},\, Z_i = \text{'new'} \mid \cdots\right) \propto \int_{\mathbb{X}} k(T_i; x)\, \tau(1; K_n(x))\, \tau_0\big(1; D\psi(K_n(x))\big)\, \theta P_0(dx);$$

in other words, the i-th observation is associated to a new latent mark from random measure $\tilde{\mu}_d^e$, which is itself matched with a new latent location from the root random measure $\tilde{\mu}_0$.

As already discussed in Section 3.3, the new value possibly assigned to latent variable Z_i is arbitrary, and does not play any role within the Gibbs sampling scheme: the only relevant information encoded in the elements of the latent sequence of marks is the random partition structure induced by their ties. On the contrary, the new value possibly assigned to latent variable X_i in case (3) directly impacts the allocation probabilities in future steps of the Gibbs sampling scheme, and is sampled from a diffuse probability measure with density proportional to

$$k(T_i; x)\, \tau(1; K_n(x))\, \tau_0\big(1; D\psi(K_n(x))\big)\, P_0(dx).$$

Note that, for the actual implementation of the urn scheme introduced above within an MCMC algorithm, the role reserved here to the i-th observation is taken in turn by each of the n available observations, which are sequentially reallocated within the latent partition structure; indeed, groups multiplicities encoding such structure are recomputed at each step removing the latent variables corresponding to the observation to be reallocated.

Each iteration of the Gibbs sampling algorithm proposed in this section effectively updates the latent nested partition structure, which is sampled from the its full conditional distribution. A further step within the same Gibbs sampler consists in the independent resampling of latent locations $X_1^*, \ldots, X_k^*$, whose full conditional distributions are obtained from the same result. This additional *acceleration step* represents a common practice in marginal sampling algorithms based on urn schemes, and is known to enhance their mixing properties (Escobar and West, 1998; MacEachern, 1998; Ishwaran and James, 2001). More precisely, each latent location X_j^* is independently resampled, given the observations and latent partition structure, from the diffuse probability measure with density proportional to

$$\mathbb{P}\left(X_j^* \in dx \mid \cdots\right) \propto \left(\prod_{i:\, X_i = X_j^*} k(T_i; x)\right) \left(\prod_{d=1}^{D} \prod_{h=1}^{r_{dj}} \tau(q_{djh}; K_n(x))\right) \tau_0\big(r_j; D\psi(K_n(x))\big)\, P_0(dx).$$

Finally, hyperpriors are frequently imposed on the hyperparameters characterizing either the Lévy intensity measures ρ and ρ_0 or the kernel $k(\cdot;\cdot)$; in particular, the gamma distribution on θ is the conjugate prior with respect to the marginal distribution in (3.10). Therefore, if θ has prior distribution Gamma(a, b), its full conditional distribution, given the observed survival times $\boldsymbol{T}$

and the number k of distinct latent locations, is

$$\theta \mid \boldsymbol{T}, k \sim \text{Gamma}\left(a+k,\ b+\int_{\mathbb{X}} \psi_0\big(D\,\psi(K_n(x))\big)\,\theta P_0(dx)\right).$$

Example 3.9. Considering the hierarchical *generalized gamma* CRM and the Dykstra and Laud (1981) kernel, as in Examples 3.2 and 3.6, the marginal Gibbs sampling scheme for the reallocation of the i-th observation within the partition structure boils down to:

(1) for each latent location X_j^* such that $X_j^* \leq T_{n+1}$,

$$\mathbb{P}\left(X_i = X_j^*,\, Z_i = Z_{djh}^* \mid \cdots\right) \propto \frac{q_{djh}^{-i} - \sigma}{\beta + K_n(X_j^*)}, \qquad h = 1, \ldots, r_{dj}^{-i};$$

(2) for each latent location X_j^* such that $X_j^* \leq T_{n+1}$,

$$\mathbb{P}\left(X_i = X_j^*,\, Z_i = \text{'new'} \mid \cdots\right) \propto (\beta + K_n(X_j^*))^{\sigma-1}\ \frac{r_j^{-i} - \sigma_0}{\beta_0 + \frac{D}{\sigma}\big((\beta + K_n(X_j^*))^{\sigma} - \beta^{\sigma}\big)};$$

(3) both X_i and Z_i assume new values with probability

$$\mathbb{P}\left(X_i = \text{'new'},\, Z_i = \text{'new'} \mid \cdots\right) \propto \int_0^{T_i} (\beta + K_n(x))^{\sigma-1} \left(\beta_0 + \frac{D}{\sigma}\big((\beta + K_n(x))^{\sigma} - \beta^{\sigma}\big)\right)^{\sigma_0 - 1} \theta P_0(dx);$$

in this case, the new value assigned to X_i is sampled between 0 and T_i, proportionally to

$$(\beta + K_n(x))^{\sigma-1} \left(\beta_0 + \frac{D}{\sigma}\big((\beta + K_n(x))^{\sigma} - \beta^{\sigma}\big)\right)^{\sigma_0 - 1} P_0(dx).$$

Note that the quantities in case (1) and case (2) are easily computed in closed form, while the integral in case (3) can be evaluated via numerical approximations (e.g. quadrature formulas). Moreover, the new value possibly assigned to X_i can be sampled exploiting classical rejection sampling techniques. Finally, each latent location X_j^* is independently resampled between 0 and the minimum observed survival time associated with it, that is $\min_{\{i:\, X_i = X_j^*\}} T_i$, proportionally to

$$\mathbb{P}\left(X_j^* \in dx \mid \cdots\right) \propto (\beta + K_n(x))^{r_j\sigma - n_j} \left(\beta + \frac{D}{\sigma}\big((\beta + K_n(x))^{\sigma} - \beta^{\sigma}\big)\right)^{\sigma_0 - r_j} P_0(dx);$$

this task is routinely performed by means of a Metropolis-Hastings step.

The main reason for devising the marginal Gibbs sampling scheme detailed above is the

computation of Bayesian posterior estimates of functionals of interest, namely the overall survival function and the cause-specific incidence functions (or the cause-specific subdistributions), and the construction of the prediction curves. More precisely, at each iteration of the MCMC algorithm, the updated latent variables, which have been resampled according to their full conditional distributions, are plugged into the expressions presented in Propositions 3.7 and 3.8, in order to obtain a corresponding (conditional) posterior estimate of such functionals; the unconditioned posterior estimates are eventually computed by averaging over the sequence of estimates from the Markov chain.

Nevertheless, the marginal approach cannot be adopted for a reliable quantification of the uncertainty around posterior estimates. Indeed, the sequence of conditional estimates computed at each MCMC iteration is not a sample from the posterior distribution of the corresponding functionals, since the uncertainty associated to the posterior hierarchical random measures has been averaged out analytically in Propositions 3.7 and 3.8; see Lijoi and Nipoti (2014) for further details on this aspect. An alternative approach consists in sampling, at each iteration of the MCMC algorithm, a number of independent realizations from the posterior distribution of hierarchical random measures, given the observations and the updated latent variables, as characterized in Proposition 3.5; these samples can then be promptly exploited to compute an equal number of posterior estimates of functionals of interest, directly from their definitions (3.2) and (3.5). This *conditional* method entirely accounts for the posterior uncertainty, and estimates obtained through this approach are actual samples from the posterior distribution of the corresponding functionals: therefore, reliable credible bands can be effectively constructed based upon them.

The practical implementation of the conditional method requires sampling from the posterior distribution of hierarchical random measures; in particular, sampling both random jumps and random locations of the non-homogeneous random measures $\tilde{\mu}_1^*, \ldots, \tilde{\mu}_D^*$ and $\tilde{\mu}_0^*$ represents a challenging task, and involves the truncation of such infinite sequences. A popular sampling algorithm for CRMs is based on the Ferguson and Klass (1972) representation of independent increment processes, which guarantees the sequence of random jumps to be almost-surely decreasing. This compelling property ensures that, for each truncation level, the most relevant jumps are considered, and allows for a better control of the truncation error. In presence of non-homogeneous CRMs, an alternative and simpler sampling algorithm was suggested by Wolpert and Ickstadt (1998); however, their proposal does not induce an almost surely decreasing sequence of jumps, which represents a major drawback when a truncation is performed (Walker and Damien, 2000). An interesting result supporting the adoption of this techniques in Bayesian nonparametrics is discussed in Camerlenghi et al. (2021): when the non-homogeneity a posteriori is induced by an exponential tilting of a homogeneous prior, the sequence of jumps sampled from the posterior can be bounded above by a corresponding decreasing sequence obtained from the prior. Such theoretical guarantees on the truncation error underpin the conditional sampling

method proposed in the following.

Consider the non-homogeneous random measure $\tilde{\mu}_0^*$ at the root of the posterior hierarchy, introduced in Proposition 3.5 and characterized by the Lévy intensity measure

$$\nu_0^*(ds, dx) = \rho_0^*(ds \,|\, x)\, \theta P_0(dx) = \exp\big\{ - D\psi(K_n(x))\, s\big\}\, \rho_0(ds)\, \theta P_0(dx);$$

(1) the sequence of random locations $\big(X_h^{(0)}\big)_{h\geq 1}$ is sampled independently from base probability measure P_0;

(2) the sequence of random jumps $\big(V_h^{(0)}\big)_{h\geq 1}$ is sampled according to the Wolpert and Ickstadt (1998) algorithm, that is by solving sequentially with respect to $V_h^{(0)}$ the equations

$$N_h^{(0)} = \theta \int_{V_h^{(0)}}^{\infty} \exp\Big\{ - D\psi\big(K_n\big(X_h^{(0)}\big)\big)\, s\Big\}\, \rho_0(ds), \qquad h \geq 1,$$

where $\big(N_h^{(0)}\big)_{h\geq 1}$ is the sequence of jump times of a unit-rate Poisson process, namely $N_0^{(0)} = 0$ and, for $h \geq 1$, the increments $N_h^{(0)} - N_{h-1}^{(0)}$ are independent and exponentially distributed with unit mean.

As suggested in Camerlenghi et al. (2021), this procedure is stopped as soon as, for some H_0,

$$N_{H_0}^{(0)} > \theta \int_{\varepsilon}^{\infty} \rho_0(ds),$$

which guarantees each of the discarded jumps $V_{H_0+1}^{(0)}, V_{H_0+2}^{(0)}, \ldots$ to be smaller than some threshold $\varepsilon > 0$; in practical implementations, such threshold may either be a small and fixed value, or depend on the previously sampled jumps $V_1^{(0)}, \ldots, V_{H_0}^{(0)}$, e.g. through their sum. As a result, an approximate realization of the root random measure $\tilde{\mu}_0$ from its posterior distribution is given by

$$\sum_{j=1}^{k} V_j\, \delta_{X_j^*}(dx) + \sum_{h=1}^{H_0} V_h^{(0)}\, \delta_{X_h^{(0)}}(dx), \tag{3.15}$$

where the random jumps $V_1, \ldots, V_k$ at the fixed locations $X_1^*, \ldots, X_k^*$ are sampled as specified in Proposition 3.5. Note that, since P_0 is a diffuse probability measure, the random locations $X_1^{(0)}, \ldots, X_{H_0}^{(0)}$ assume distinct values almost surely, and differ from the fixed locations $X_1^*, \ldots, X_k^*$ as well.

Similarly, at the lower level of the posterior hierarchy, consider each non-homogeneous random measure $\tilde{\mu}_d^*$, characterized by the Lévy intensity measure

$$\nu^*(ds, dx) = \rho^*(ds \,|\, x)\, \tilde{\mu}_0(dx) = e^{-K_n(x)\, s}\, \rho(ds)\, \tilde{\mu}_0(dx);$$

(1) the sequence of random locations $\big(X_h^{(d)}\big)_{h\geq 1}$ is sampled independently and proportionally

to the discrete measure $\tilde{\mu}_0(dx)$, an approximation of which is obtained in (3.15);

(2) the sequence of random jumps $\left(S_h^{(d)}\right)_{h\geq 1}$ is sampled by solving sequentially with respect to $S_h^{(d)}$ the equations

$$N_h^{(d)} = \tilde{\mu}_0(\mathbb{X}) \int_{S_h^{(d)}}^{\infty} e^{-K_n(X_h^{(d)})\, s} \rho(ds), \qquad h \geq 1,$$

where $\tilde{\mu}_0(\mathbb{X})$ is the total mass of $\tilde{\mu}_0$ and $\left(N_h^{(d)}\right)_{h\geq 1}$ is the sequence of jump times of a unit-rate Poisson process.

Again, the procedure above is stopped as soon as, for some H_d,

$$N_{H_d}^{(d)} > \tilde{\mu}_0(\mathbb{X}) \int_{\varepsilon}^{\infty} \rho(ds),$$

which guarantees each of the discarded jumps $S_{H_d+1}^{(d)}, S_{H_d+2}^{(d)}, \ldots$ to be smaller than the threshold $\varepsilon > 0$. An approximate realization of the random measure $\tilde{\mu}_d$ from its posterior distribution is thus given by

$$\sum_{j=1}^{k} \sum_{h=1}^{r_{dj}} S_{djh}\, \delta_{X_j^*}(dx) + \sum_{h=1}^{H_d} S_h^{(d)}\, \delta_{X_h^{(d)}}(dx), \tag{3.16}$$

where the random jumps at fixed locations are sampled as specified in Proposition 3.5; in fact, the truncation of the infinite sequence of jumps in (3.16), and the conditioning on a truncated realization of the base measure $\tilde{\mu}_0$, entail two levels of approximation. Moreover, note that the random locations $X_1^{(d)}, \ldots, X_{H_d}^{(d)}$ display ties with positive probability, and take values among the finite collection of locations $X_1^*, \ldots, X_k^*$ and $X_1^{(0)}, \ldots, X_{H_0}^{(0)}$ characterizing the discrete and finitely supported measure approximating $\tilde{\mu}_0$ in (3.15). As a consequence, an alternative approach for sampling approximate realizations of each random measure $\tilde{\mu}_d$ from its posterior distribution consists in sampling its $k + H_0$ cumulative random jumps, associated to the locations inherited by the approximation in (3.15). Remarkably, the distributions of such cumulative jumps are infinitely divisible and characterized through their Laplace transforms, while their density functions are usually not available in closed form, with the notable exception of the hierarchical gamma CRM; random variables specified via their Laplace transforms or characteristic functions can be sampled exploiting the general algorithms described in Devroye (1981) and Ridout (2009). This exact sampling procedure avoids the approximation at the lower level of the hierarchy, and is explored for hierarchical normalized random measures in Lijoi et al. (2020).

Example 3.10. The conditional method for sampling from the posterior distribution of the hierarchical random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_D$, detailed above, is specialized here for the hierarchical *generalized gamma* CRM prior choice, previously considered in Examples 3.2, 3.6 and 3.9. A fundamental role in sampling the sequences of random jumps is played by the upper incomplete

gamma function, defined as

$$\Gamma(a,x) := \int_x^\infty s^{a-1}\, e^{-s}\, ds,$$

and evaluated for $x \geq 0$ and $a \in (-1, 0]$. At the root of the hierarchy, an approximate realization of $\tilde{\mu}_0$ from its posterior distribution is given by (3.15), where:

(1) the random jumps $V_1, \ldots, V_k$ at fixed locations $X_1^*, \ldots, X_k^*$ are sampled from the Gamma distribution

$$V_j \sim \text{Gamma}\left(r_j - \sigma_0,\, \beta_0 + \frac{D}{\sigma}\Big(\big(\beta + K_n(X_j^*)\big)^\sigma - \beta^\sigma\Big)\right), \qquad j = 1, \ldots, k;$$

(2) the sequence of random locations $\big(X_h^{(0)}\big)_{h\geq 1}$ is sampled independently from P_0;

(3) the sequence of random jumps $\big(V_h^{(0)}\big)_{h\geq 1}$ is sampled by solving sequentially with respect to $V_h^{(0)}$ the equations

$$\Gamma\left(-\sigma_0,\, V_h^{(0)}\left(\beta_0 + \frac{D}{\sigma}\Big(\big(\beta + K_n\big(X_h^{(0)}\big)\big)^\sigma - \beta^\sigma\Big)\right)\right) = \frac{N_h^{(0)}\,\Gamma(1-\sigma_0)}{\theta\left(\beta_0 + \frac{D}{\sigma}\Big(\big(\beta + K_n\big(X_h^{(0)}\big)\big)^\sigma - \beta^\sigma\Big)\right)^{\sigma_0}},$$

for $h \geq 1$, where $\big(N_h^{(0)}\big)_{h\geq 1}$ is the sequence of jump times of a unit-rate Poisson process. The truncation level H_0 is the first value of h for which

$$N_h^{(0)} > \frac{\Gamma(-\sigma_0,\, \beta_0\varepsilon)}{\Gamma(1-\sigma_0)}\,\beta_0^{\sigma_0}\theta \overset{(\beta_0=0)}{=} \frac{\theta\,\varepsilon^{-\sigma_0}}{\sigma_0\,\Gamma(1-\sigma_0)}.$$

Similarly, at the lower level of the hierarchy, an approximate realization of $\tilde{\mu}_d$ from its posterior distribution is given by (3.16), where:

(1) the random jumps $(S_{djh})_{jh}$ at fixed locations $X_1^*, \ldots, X_k^*$ can be sampled in groups from the Gamma distribution

$$\sum_{h=1}^{r_{dj}} S_{djh} \sim \text{Gamma}\big(n_{dj} - r_{dj}\sigma,\, \beta + K_n(X_j^*)\big), \qquad j = 1, \ldots, k;$$

(2) the sequence of random locations $\big(X_h^{(d)}\big)_{h\geq 1}$ is sampled independently and proportionally to the approximation of $\tilde{\mu}_0$ obtained above;

(3) the sequence of random jumps $\big(S_h^{(d)}\big)_{h\geq 1}$ is sampled by solving sequentially with respect to

$S_h^{(d)}$ the equation

$$\Gamma\left(-\sigma,\, S_h^{(d)}\left(\beta + K_n\big(X_h^{(d)}\big)\right)\right) = \frac{N_h^{(d)}\,\Gamma(1-\sigma)}{\tilde{\mu}_0(\mathbb{X})\left(\beta + K_n\big(X_h^{(d)}\big)\right)^{\sigma}}, \qquad h \geq 1,$$

where $\tilde{\mu}_0(\mathbb{X}) \approx \sum_{j=1}^{k} V_j + \sum_{h=1}^{H_0} V_h^{(0)}$ is the approximated total mass of $\tilde{\mu}_0$ and $\big(N_h^{(d)}\big)_{h\geq 1}$ is the sequence of jump times of a unit-rate Poisson process. Again, the truncation level H_d is the first value of h for which

$$N_h^{(d)} > \frac{\Gamma(-\sigma,\, \beta\varepsilon)}{\Gamma(1-\sigma)}\,\beta^{\sigma}\tilde{\mu}_0(\mathbb{X}) \overset{(\beta=0)}{=} \frac{\tilde{\mu}_0(\mathbb{X})\,\varepsilon^{-\sigma}}{\sigma\,\Gamma(1-\sigma)}.$$

For the actual implementation of this algorithm, the bottleneck is represented by the solution of the equations in steps (3). A simple but effective approach in practice consists in solving these equations with respect to the logarithm of the jump, exploiting Newton-Raphson method: indeed, the functions involved are strictly convex and their derivatives can be computed in closed form. An implementation of both marginal and conditional algorithms in Julia is available in the public Github repository `CompetingRisks.jl`.

3.7 Numerical illustration and simulation study

This section contains a numerical illustration of both the *marginal* and *conditional* strategies, introduced in Section 3.6, on a simulated dataset. The posterior estimates of the overall survival function and cause-specific incidence functions are compared with the functions controlling the data-generating mechanism, and with the corresponding frequentist estimates, in order to assess the effectiveness of the proposed approaches; prediction curves are also constructed and compared with their data-generating counterparts. Finally, a simulation study is performed with the twofold purpose of exploring the role of information borrowing among hazard rates, induced by the hierarchical structure, and evaluating its contribution to the accuracy of posterior estimates.

The data-generating model considered for the numerical illustration involves $D = 3$ competing sources of risk, which are assumed to be independent; their *potential* survival times (see Section 3.1) are sampled from the Weibull distributions

$$\begin{gathered} T_i^{(1)} \sim \text{Weibull}(\xi_1 = 1.2), \qquad T_i^{(2)} \sim \text{Weibull}(\xi_1 = 1.6), \\ T_i^{(3)} \sim \text{Weibull}(\xi_2 = 2.4), \end{gathered} \tag{3.17}$$

where ξ_1, ξ_2, ξ_3 are shape parameters and scale parameters are assumed to be unitary. The observed survival time T_i is the minimum of these latent survival times, $T_i = \min_d T_i^{(d)}$, while the corresponding event type $\Delta_i = d$ if and only if $T_i = T_i^{(d)}$. In the following, a simulated datasets consisting of $n = 300$ independent observations is considered. Note that, differently

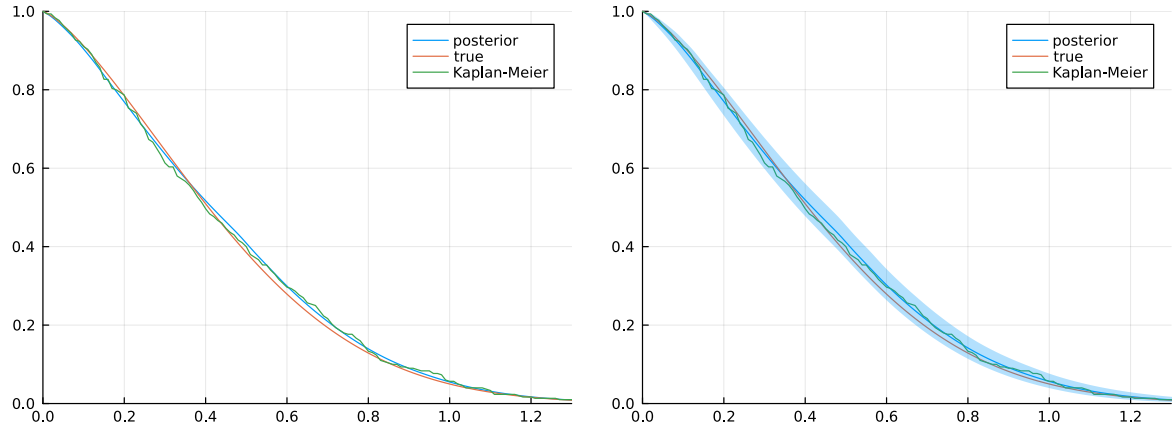

Figure 3.1: Posterior estimates of the overall survival function (blue), obtained via marginal (left) and conditional (right) algorithms, and Kaplan-Meier estimate (green), compared with the true survival function (orange), for a simulated datasets with $n = 300$ data points; a pointwise credible band at level 0.90 is also displayed for the conditional approach (right).

from the partially exchangeable settings (Lijoi and Nipoti, 2014; Camerlenghi et al., 2021), the number of subjects incurring in each competing event is a random variable itself. However, the parameter choice for the Weibull distributions in (3.17) implies a fairly balanced allocation of the observations to the event types: in particular, the number of subjects in the simulated dataset experiencing each competing event is 120, 99 and 81, respectively.

The marginal and conditional Gibbs sampling algorithms are implemented for the hierarchical *generalized gamma* CRM prior and Dykstra and Laud (1981) kernel, as specified in Examples 3.9 and 3.10. For illustration purposes, the hyperparameters characterizing the jump components of the hierarchical random measures are fixed at values $\beta = \beta_0 = 1.0$ and $\sigma = \sigma_0 = 0.25$, while the probability measure P_0 is assumed uniform on the interval $[0, 2]$, which entirely contains the simulated data. Moreover, exponential hyperpriors with large mean, corresponding to a non-informative specification, are imposed on θ and on the kernel parameter γ. The MCMC procedure is run for $75,000$ iterations, with a burn-in period of $25,000$ iterations; after thinning, $5,000$ samples from the latent partition structure are collected and exploited to compute posterior estimates within the marginal method. As for the conditional approach, 20 independent realizations from the posterior distribution of hierarchical random measures are sampled for each sample from the latent structure. The standard diagnostic tools suggest convergence of the Markov chains, which show a fairly good mixing; indeed, using the marginal algorithm, the average effective sample size for the posterior estimates of the survival function, evaluated on a grid of points, is about $3,200$.

Figure 3.1 compares the posterior estimates of the overall survival function, obtained via both the marginal and conditional strategies, with the true survival function, and with the frequentist Kaplan-Meier estimate. The true survival function, implied by the model in (3.17),

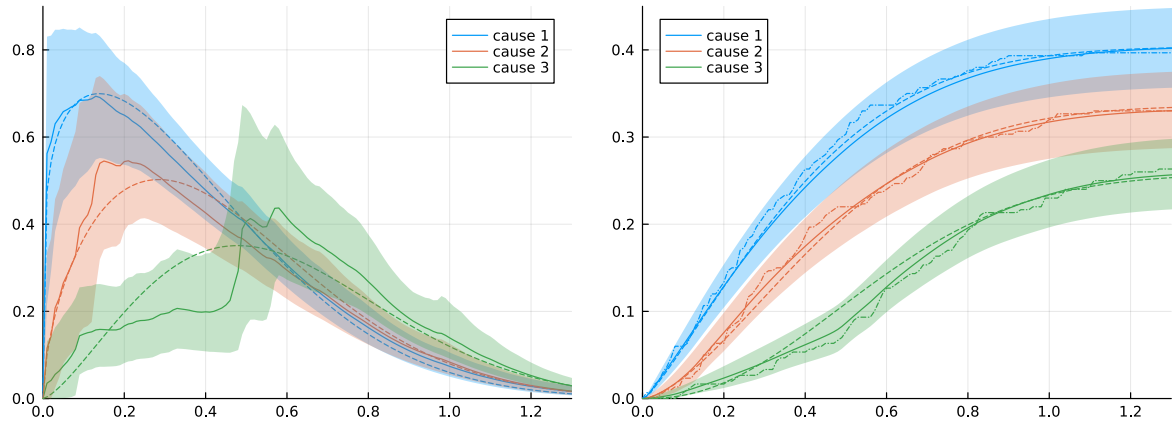

Figure 3.2: Posterior estimates of cause-specific incidence functions (left) and subdistributions (right), obtained via the conditional method, and corresponding pointwise credible bands at level 0.90, compared with the true functions (dashed), for a simulated dataset with $n = 300$ data points; Bayesian estimates of subdistributions are also compared with the frequentist Aalen-Johansen estimate (dash-dotted).

is successfully recovered by both methods, which provide similar estimates: their Kolmogorov distances from the true function (0.02404 and 0.02664, respectively) are substantially smaller than the Kolmogorov distances computed for the Kaplan-Meier estimate (0.03298). Furthermore, the pointwise credible band at level 0.90, constructed according to the conditional approach, entirely contains the true function; such credible band is up to twice as wide as the corresponding credible band constructed according to the marginal approach (not displayed), and represents a more conservative and reliable quantification of the uncertainty (see Section 3.6). The posterior estimates of the cause-specific incidence functions and cause-specific subdistributions, obtained via the conditional algorithm, are compared with the corresponding true functions in Figure 3.2. The mutual relationship among incidence functions, and their overall behaviour, are satisfactorily recovered: indeed, their pointwise credible bands just slightly overlap for times smaller than their crossing times, which are correctly detected. Moreover, posterior estimates of both subdistribution functions are consistent with the frequentist Aalen-Johanssen nonparametric estimates, and the corresponding pointwise credible bands entirely contain the true subdistribution functions. Finally, prediction curves obtained via the marginal sampling strategy are displayed in Figure 3.3; for comparison, the conditional probabilities, implied by the data-generating model, for an observation to be of a certain type as functions of the survival time are also reported. Remarkably, prediction curves feature just slight variations for larger values of t: indeed, observations do not inform prediction curves beyond the largest observed survival time, as already discussed in Example 3.6.

The possibility to share information between hazard rates related to different sources of risk represents the main reason for introducing dependence among such hazard rates through

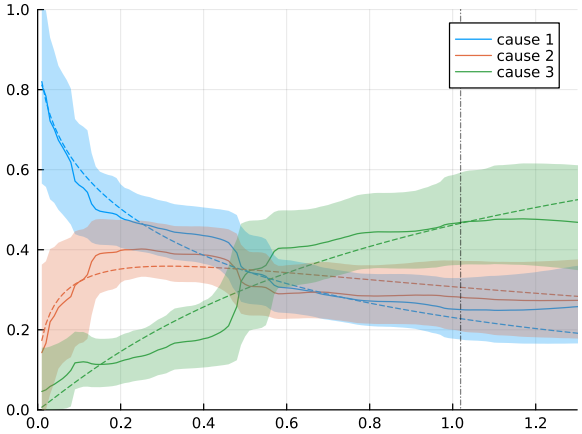

Figure 3.3: Prediction curves for each competing source of risk, obtained via the marginal method, and corresponding pointwise credible bands at level 0.90, compared with the true functions (dashed), for a simulated dataset with $n = 300$ data points; the vertical black dashed line marks the 0.95 empirical quantile of the observed survival times.

hierarchical priors. The simulation study presented in the following compares the performances of proposed model, featuring hierarchically dependent hazard rates, with those of the corresponding Bayesian nonparametric model which considers independent hazard rates for each source of risk. Specifically, the hierarchical prior distribution over the vector of random measures in (3.4) is replaced with the independent prior

$$\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \dots, \tilde{\mu}_D) \overset{\text{i.i.d.}}{\sim} \text{CRM}(\nu),$$

where $\tilde{\mu}_1, \dots, \tilde{\mu}_D$ are independent completely random measures with homogeneous Lévy intensity measure $\nu(ds, dx) = \rho(ds)\, \theta\, P_0(dx)$. Note that, according to this prior specification, the posterior estimate of the overall survival function coincides with the product of the *potential* survival functions for each event type, in the hypothesis of absence of competing risks, i.e., considering the occurrences of competing events as censored observations.

Consider a data-generating model involving $D = 3$ competing and independent sources of risk; for each event type, the corresponding potential survival times are sampled from a mixture of a cause-specific distribution and a common distribution, with equal weights. In particular, the common distribution is itself a mixture of (shifted) Weibull distributions, having density function

$$\bar{f}(t) = 0.5\ \text{Weibull}(t;\, \bar{\xi}_1 = 1.2) + 0.5\ \text{Weibull}(t - 1.0;\, \bar{\xi}_2 = 3.0), \qquad t \geq 0,$$

where $\bar{\xi}_1, \bar{\xi}_2$ are shape parameters, while the cause-specific distributions are specified as Weibull distributions with shape parameters $\xi_1 = 1.5$, $\xi_2 = 2.0$ and $\xi_3 = 2.5$, respectively. The cause-specific hazard rates and incidence functions resulting from such data-generating model are displayed in Figure 3.4: note that the hazard rate functions are essentially increasing, making

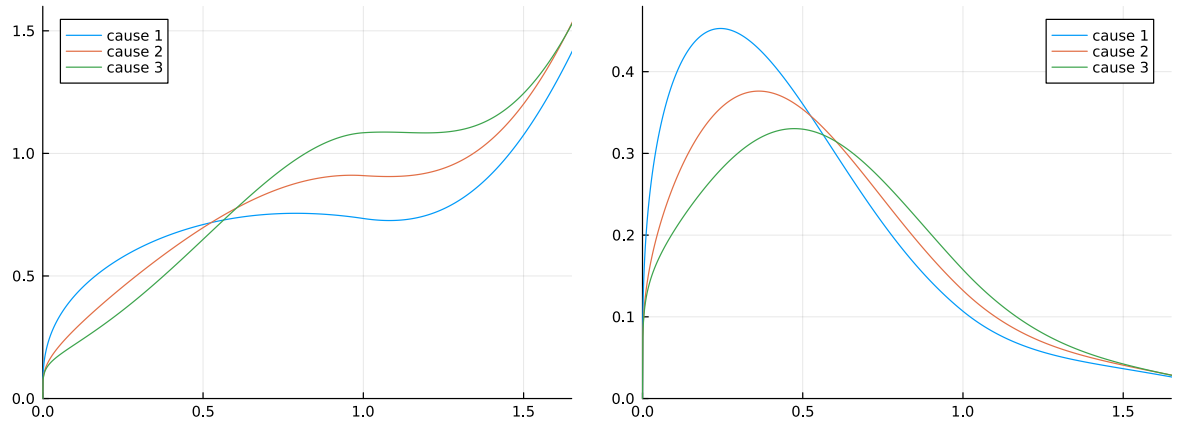

Figure 3.4: Cause-specific hazard rates (left) and incidence functions (right) for the $D = 3$ competing and independent sources of risks considered in the data-generating model for the simulation study; the hazard rate functions are basically increasing.

the Dykstra and Laud (1981) kernel an appropriate choice in this scenario. Results presented in the following are obtained averaging over $m = 100$ simulated datasets, each consisting of $n = 100$ observations, generated according to the potential survival times distributions specified above. Similarly to the previous illustration, the choice of the data-generating distributions entails a balanced allocation of the observations to the event types: the average number of subjects experiencing each competing event is 34.11, 33.21 and 32.68, respectively. Furthermore, the hyperparameters characterizing the distribution of the random measures, as well as the non-informative hyperpriors imposed on θ and γ, are specified according to the same values and distributions proposed in the illustration above.

Table 3.1 summarizes the outcomes of the simulation study, comparing the performances of the proposed hierarchical model with those of the alternative *independent* model, according to both the marginal and conditional sampling strategies; in particular, the average errors in estimating the cause-specific incidence function and subdistribution for each event type are displayed for both models and both sampling algorithms. In addition, the average estimation error for the frequentist Aalen-Johanssen estimator of the subdistribution functions is contained in the table; finally, estimation errors for the posterior estimates displayed in Figure 3.2 are also reported, in order to provide a visual reference. The comparison is based on total variation distances and Kolmogorov distances, which are rescaled according to the true probabilities of occurrence of each event type. Specifically, consider the measures of error

$$e_d^{\mathrm{TV}} = \frac{1}{2\pi_d} \int_{\mathbb{R}^+} |\hat{f}_d(t) - f_d(t)|\, dt, \qquad e_d^{\mathrm{K}} = \frac{1}{\pi_d} \sup_{t \in \mathbb{R}^+} |\hat{F}_d(t) - F(t)|, \qquad d = 1, \dots, D,$$

where $\hat{f}_d$ and f_d are the estimated and true incidence functions, respectively, $\hat{F}_d$ and F_d are the estimated and true subdistribution functions, and π_d is the true probability of occurrence, for

		incidence function			subdistribution		
		$d=1$	$d=2$	$d=3$	$d=1$	$d=2$	$d=3$
hierarchical	marginal	**.0928**	**.0862**	.1006	**.1270**	**.1217**	.1327
	conditional	.0956	.0871	.0994	.1312	.1237	**.1321**
independent	marginal	.1046	.0935	.0998	.1370	.1312	.1369
	conditional	.1067	.0946	**.0992**	.1402	.1335	.1373
frequentist					.1812	.1978	.2058
Figure 3.2		.0280	.0443	.1199	.0213	.0359	.0886

Table 3.1: Comparison of rescaled total variation distances between estimated and true incidence functions (first column) and rescaled Kolmogorov distances between estimated and true subdistribution functions (second column) for the hierarchical and independent model, using marginal and conditional sampling methods; distances are averaged over $m = 100$ simulated datasets.

each event type d. The aim of rescaling such standard distances is obtaining error measurements that are comparable for different event types: indeed, rescaled versions of the incidence and subdistribution functions are, respectively, proper density and distribution functions. The proposed hierarchical model is shown to quite consistently outperform the alternative independent model in estimating both incidence functions and subdistributions, using either the marginal or the conditional approach; in particular, the marginal method performs slightly better than the conditional method in most scenarios, with some exceptions appearing in the estimation of functionals of interest for the third source of risk, for which the differences between the hierarchical and independent models are also less pronounced. Interestingly, the average estimated number of latent locations shared by the dependent random measures in the hierarchical model is 15.21, while the alternative model employs an average of 32.16 latent locations among the independent random measures.

The quantification of the uncertainty around posterior estimates represents another inferential problem which benefits from information borrowing among hazard rates, as displayed in Table 3.2. The pointwise credible bands at level 0.90 around the posterior estimates of the survival function, incidence functions and subdistribution functions are constructed according to both marginal and conditional methods, and compared between the two alternative models; such comparison is carried out in terms of the average maximum width of the credible bands, and the proportion of times in the interval $[0, 2]$ the true functions are actually contained within the bands. Focusing on the estimation of subdistribution functions, pointwise credible bands constructed through the conditional algorithm are up to 2.75 times as wide as credible bands constructed through to the marginal algorithm. Accordingly, the coverage for the conditional method comes close to the 0.90 level, while the corresponding coverage for the marginal method can be as low as 0.40; these empirical findings support the theoretical evidence that the conditional approach provides a more

		survival function		incidence functions		subdistributions	
		bdwdth	coverg	bdwdth	coverg	bdwdth	coverg
hierarchical	marginal	.0925	.6198	.4181	.6840	.0520	.4109
	conditional	.1469	**.8860**	.4646	**.9012**	.1431	**.8886**
independent	marginal	.0832	.5162	.3914	.6072	.0513	.3904
	conditional	.1430	.7873	.4394	.8472	.1470	.8639

Table 3.2: Comparison of pointwise credible bands at level 0.90 around the posterior estimates of the survival function (first column), incidence functions (second column) and subdistribution functions (third column) for the hierarchical and independent model, using both marginal and conditional sampling methods, in terms of their average maximum width (*bdwdth*) and average proportion of times they include the corresponding true function values (*coverg*); figures are averaged over $m = 100$ simulated datasets and $D = 3$ sources of risks.

reliable quantification of the uncertainty. The differences in bandwidth and coverage between the different sampling approaches, obtained in estimating the survival function or incidence functions, are less pronounced. Moreover, pointwise credible bands for the independent model are consistently narrower, and their coverage definitely poorer, with respect to the hierarchical model; in particular, the values of coverage obtained via the conditional method approach the 0.90 level only adopting the proposed hierarchical model.

3.8 Applications to clinical datasets

This section contains an application of the proposed modelling approach to two clinical datasets that are publicly available and have been previously adopted to illustrate statistical techniques for competing risks data. The first dataset has been extracted from the bone marrow transplant registry of the European Blood and Marrow Transplant (EBMT) Group, and made available within the *crrSC* package of the statistical software R; these data have been used to compare different modelling approaches for the analysis of clustered competing risks data in Schmitt et al. (2023). The second dataset was collected by Drzewiecki et al. (1980) at the Odense University Hospital, Denmark, on patients affected by melanoma, and is available as part of the *timereg* package in R; Andersen et al. (1993) illustrate several survival analysis methods on such data, while Arfé et al. (2019) have recently analyzed them as an application of their modelling approach to competing risks data.

The analysis of real-world datasets requires the extension of the Bayesian nonparametric model thoroughly described in this paper, in order to accommodate both right-censored observations and categorical predictors. As for right-censored data, let the random variable Δ, representing the event type, assume the additional value 0 to denote a censored observation, that is, $\Delta_i = 0$ if

and only if the survival time T_i is right-censored. The distributional results discussed in Sections 3.3 and 3.5, as well as the sampling algorithms devised in Section 3.6, can be easily adapted to the presence of right-censored observations once the associated multiplicative intensity likelihood function considered in (3.6) is replaced by

$$\mathcal{L}(\tilde{\mu}_1, \ldots, \tilde{\mu}_D; \boldsymbol{T}, \boldsymbol{\Delta}) = \exp\left\{ -\sum_{d=1}^{D} \sum_{i=1}^{n} \int_0^{T_i} \int_{\mathbb{X}} k(s; x)\, \tilde{\mu}_d(dx)\, ds \right\} \prod_{i:\, \Delta_i \neq 0} \int_{\mathbb{X}} k(T_i; x)\, \tilde{\mu}_{\Delta_i}(dx);$$

in particular, additional random variables in the sequences $\boldsymbol{X}$ and $\boldsymbol{Z}$ are introduced exclusively for non-censored observations, which are therefore those accommodated within the latent partition structure. On the contrary, censored observations contribute to the augmented likelihood only through the quantity $K_n(x; \boldsymbol{T})$; see Lijoi et al. (2014) and the Supplementary Material in Camerlenghi et al. (2019) for similar derivations.

The inclusion of predictors is based on a Cox regression model (or *proportional hazard rates* model), which defines a semiparametric prior on the hazard functions; specifically, cause-specific hazard rates in (3.1) are extended to

$$\tilde{h}_d(t; \boldsymbol{c}) = \exp(\boldsymbol{\eta} \cdot \boldsymbol{c}) \int_{\mathbb{X}} k(t; x)\, \tilde{\mu}_d(dx), \qquad d = 1, \ldots, D,$$

where $\boldsymbol{\eta}$ is the vector of regression coefficients and $\boldsymbol{c}$ is the vector of categorical predictors. The resulting extension of distributional results and sampling algorithms requires minimal efforts, namely the redefinition of quantities in (3.8) as

$$Q(\boldsymbol{T}, \boldsymbol{X}, \boldsymbol{C}) = \prod_{j=1}^{k} \prod_{\{i:\, X_i = X_j^*\}} \exp(\eta \cdot \boldsymbol{C}_i)\, k(T_i; X_j^*),$$

$$K_n(x) = K_n(x; \boldsymbol{T}, \boldsymbol{C}) = \sum_{i=1}^{n} \exp(\eta \cdot \boldsymbol{C}_i) \int_0^{T_i} k(s; x)\, ds,$$

where $\boldsymbol{C} = (\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n)$ is the collection of the observed vectors of predictors. The applications proposed in this section consider a single binary predictor, that is, $\boldsymbol{C}_i = C_i \in \{0, 1\}$, and therefore involve a single regression coefficient $\boldsymbol{\eta} = \eta$, on which a non-informative centered Gaussian prior with large variance is specified. Furthermore, when dealing with survival data from clinical studies, the assumption of increasing hazard rates implied by the Dykstra and Laud (1981) kernel choice is often quite restrictive and hardly ever reasonable: therefore, the exponential (Ornstein-Uhlenbeck) kernel specification (see Section 1.2) is considered in the following applications, and a further non-informative hyperprior is imposed on its rate parameter.

The dataset from the bone marrow transplant registry of the EBMT Group includes data for 400 patients diagnosed with acute myeloid leukemia, who underwent a bone marrow transplantation in 153 different hospitals. The primary event of interest is the occurrence of either

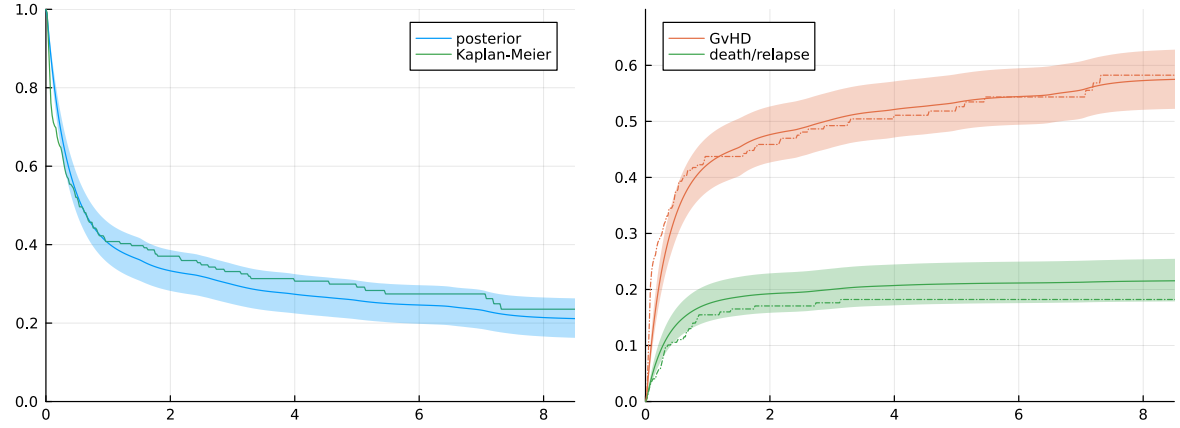

Figure 3.5: Leukemia dataset: posterior estimates of the overall survival function (left) and subdistributions (right) for the primary (GvHD) and competing (death/relapse) event types, compared with the corresponding frequentist estimates; curves are related to patients who experienced graft from bone marrow cells.

acute or chronic *Graft-versus-Host-Disease* (GvHD), while death or relapse without GvHD are competing events; survival times are expressed in years from the graft. Acute or chronic GvHD is observed in 194 patients (48.5%), while a competing event occurred for 74 patients (18.5%); the remaining observations are censored (33.0%). The median follow-up time is 0.50 years, and the maximum observed follow-up time for the primary event of interest is 7.30 years. The source of stem cells for transplantation represents the main categorical predictor considered in this study: graft from bone marrow cells was performed on 222 patients (55.5%), while for the others 178 patients (44.5%) the source of stem cells was peripheral blood. The posterior estimates of the overall survival function and subdistributions for the primary and competing sources of risk are displayed in Figure 3.5, for patients who experienced graft from bone marrow cells; pointwise credible bands at level 0.90 are obtained via the conditional approach, and the corresponding Kaplan-Meier and Aalen-Johansen frequentist estimators are also shown for comparison. As for the categorical predictor, the posterior estimate of the regression parameter η is -0.082, with credible interval at level 0.90 equal to $[-0.287, 0.121]$, suggesting that no significant difference between sources of stem cells for transplantation can be inferred from the data; these findings are consistent with the results obtained on a larger dataset from the same source by Zhou et al. (2012), where an estimate of -0.51 for this regression coefficient (p-value: 0.35), was obtained using frequentist techniques based on the Fine-Gray proportional hazards model (Fine and Gray, 1999). Finally, the posterior distribution of the exponential kernel rate parameter κ provides strong evidence against the increasing hazard rate assumption.

The clinical study from the Odense University Hospital involves 205 patients diagnosed with stage I melanoma, who underwent a surgical excision of the tumor during 1962-1977 period. The primary event of interest is death due to melanoma, which is observed in 57 patients (27.8%),

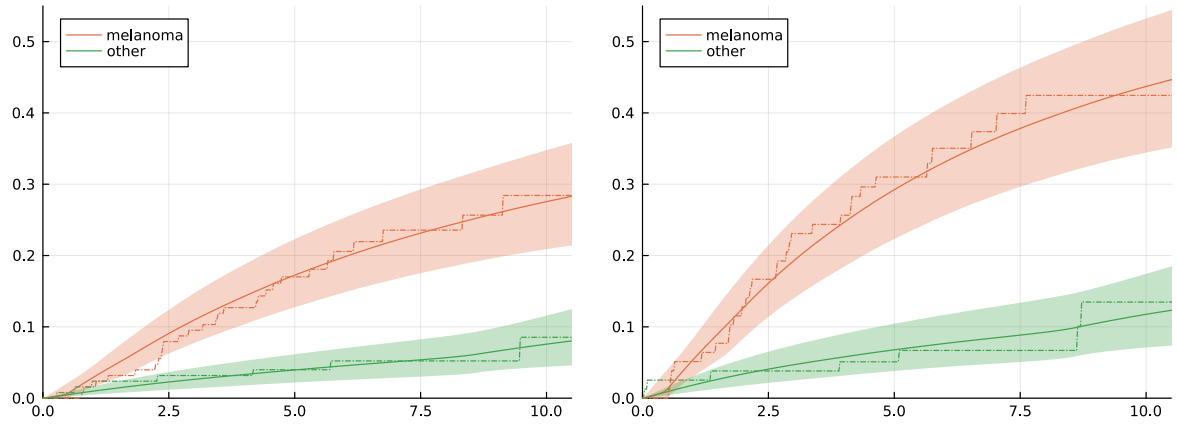

Figure 3.6: Melanoma dataset: posterior estimates of subdistribution functions for the primary (death due to melanoma) and competing (death from other causes) event types, for female (left) and male (right) patients, compared with the corresponding Aalen-Johansen estimators (dash-dotted); pointwise credible bands at level 0.90 are constructed via the conditional method.

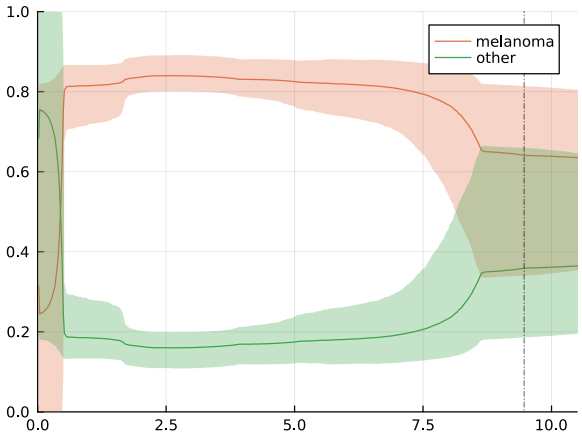

Figure 3.7: Melanoma dataset: prediction curves for the primary (death due to melanoma) and competing (death from other causes) event types, obtained via the marginal method, and corresponding pointwise credible bands at level 0.90; the vertical black dashed line marks the largest observed survival time.

while death from other causes is a competing event, occurred for 14 patients (6.8%); the remaining observations are censored (65.4%). The median survival time is 5.89 years, and the maximum observed survival time for the primary event of interest is 9.14 years. The categorical predictor considered in this application is the gender of the patient: overall, 126 participants in the study (61.5%) were women and 79 (38.5%) were men. Figure 3.6 displays the posterior estimates of subdistribution functions for the primary and competing event types, for both female and male patients, compared with the corresponding Aalen-Johansen estimators; again, pointwise credible bands at level 0.90 are obtained via the conditional approach. A significant difference is apparent between male and female patients: indeed, the posterior estimate of their hazard rate ratio $\exp(\eta)$ is 1.93, with credible interval at level 0.90 equal to $[1.26, 2.77]$, thus confirming such empirical evidence. Gender-related differences in the outcome for patients diagnosed with melanoma are well-established in clinical practice (Scoggins et al., 2006), with male patients showing an overall higher risk. Finally, prediction curves for the primary (death due to melanoma) and competing (death from other causes) event types are displayed in Figure 3.7; pointwise credible bands at level 0.90 are constructed via the marginal approach, and a vertical line marking the largest observed survival time is also displayed for reference.

Chapter 4

Kernel-weighted random measures and covariate-dependent partitions

In the Bayesian nonparametric framework, the inclusion of continuous covariates within a fully nonparametric regression model represents a challenging task, and existing models in the literature face a trade-off between flexibility in modelling the latent partition structure, its analytical tractability, and its consistency for new observations. This chapter introduces a novel class of covariate-dependent random probability measures, arising from the normalization of suitable random measures, which account for covariates through a kernel structure: specifically, the jumps of a common discrete random measure are rescaled via multiplication by a similarity kernel. A noteworthy example arises when the distribution of such random measure is a specific transformation of the distribution of a stable process: the resulting covariate-dependent random probability measure is termed *kernel-weighted Pitman-Yor process*. This construction induces a random partition model with dependence on covariates, which is characterized by great flexibility and inherent consistency for new observations, while retaining some analytical tractability, thanks to the introduction of suitable latent variables. The induced partition probability function and the posterior distribution of the common random measure are derived in closed form, conditionally on such latent variables, and the predictive scheme for both the new observations and corresponding latent variables is characterized. This proposal can be effectively exploited as the building block for the construction of nonparametric regression models, as well as a clustering or species sampling model which incorporates information available through different types of covariates.

4.1 Bayesian nonparametric regression framework

Bayesian nonparametric models often assume some kind of homogeneity among the observations, motivated by the exchangeability assumption in de Finetti's representation theorem. In presence of multiple groups of observations, homogeneity is usually assumed within each group and modeled through partial exchangeability; this setting can be regarded as a regression model with a categorical covariate, representing the group, and is extensively discussed in the literature (Section 2.1). On the other hand, including continuous covariates in a fully nonparametric fashion turns out to be a more challenging task, and a number of approaches has been proposed, typically building upon the different constructions of the Dirichlet process (Section 1.3).

The popularity of Bayesian nonparametric regression models has experienced a growing interest over the last decades, with particular focus on density regression problems, in which the entire distribution of a univariate response is allowed to change with a set of predictors[1]; in other words, the predictors are not assumed to affect some particular functional of the response distribution, as in parametric regression models, allowing for additional flexibility and showing improved performances, as for inference and prediction, with respect to classical approaches. The typical modelling choice consists in defining a collection of random density functions as

$$\tilde{f}(y \mid x) := \int \phi(y; x, \beta)\, \tilde{p}_x(d\beta), \tag{4.1}$$

where ϕ is the probability density associated to a parametric regression model and $\tilde{p}_x$ is a random probability measure on the space of the regression parameters, depending on the covariate $x \in \mathbb{X}$; within this framework, Bayesian nonparametric contributions are devoted to the specification of the prior distribution over the uncountable collection of covariate-dependent random probability measures,

$$\mathcal{P} := \{\, \tilde{p}_x \colon x \in \mathbb{X} \,\},$$

in a way that induces dependence among the elements of the collection. Chung and Dunson (2011) propose a list of desirable properties for such collection of random measures, which include: increasing pairwise dependence between random measures as the pairwise distance between the corresponding covariates decreases, simple and interpretable expressions for the marginal mean and variance, as well as for the pairwise correlation, and efficient algorithms for posterior inference. Note that the first requirement implies a topological structure for the space of covariates, which is usually not needed in presence of categorical covariates, that is, when $\mathcal{P}$ is a finite collection: this aspect is further highlighted in Section 4.2.

A class of popular and convenient proposals stems from the seminal works of MacEachern (1999, 2000), in which the dependent Dirichlet process was first introduced as a general framework. Such nonparametric prior builds upon the renowned stick-breaking construction of the Dirichlet

[1]The terms *predictor* and *covariate* are used as synonyms throughout this chapter.

process (Sethuraman, 1994), by allowing both the random locations and weights to change with the covariates; specifically, the measure-valued stochastic process $\mathcal{P}$ is defined as

$$\tilde{p}_x(d\beta) = \sum_{h\geq 1} \pi_h(x)\, \delta_{\beta_h^*(x)}(d\beta), \qquad \pi_h(x) = V_h(x) \prod_{\ell<h} (1 - V_h(x)), \tag{4.2}$$

where $(V_h)_{h\geq 1}$ and $(\beta_h^*)_{h\geq 1}$ are sequences of mutually independent stochastic processes, indexed by $x \in \mathbb{X}$ and taking values in the unit interval $[0, 1]$ and in the space of the regression parameters, respectively. The original definition of the dependent Dirichlet process assumes that, for each $h \geq 1$, the stochastic process V_h has marginal Beta distribution with parameters $(1, \theta)$ and β_h^* has marginal distribution P_0, so that the stochastic process $\mathcal{P}$ is marginally a Dirichlet process with base probability measure P_0 and concentration parameter θ. However, such assumption can be easily relaxed in different directions; for example, the marginal distributions may belong to the broader class of stick-breaking priors, introduced by Ishwaran and James (2001), encompassing both the Dirichlet process and Pitman-Yor process (Section 1.3).

The generality and flexibility of the dependent Dirichlet process construction has paved the way to a huge number of possible specifications and applications in the literature. The easiest and most convenient approach consists in considering common and predictor-independent random weights, introducing dependence only through random locations. This proposal has been successfully applied to the ANOVA and linear regression frameworks (De Iorio et al., 2004), and to the modeling of spatial processes (Gelfand et al., 2005) and dynamic phenomena (Rodriguez and ter Horst, 2008; Ascolani et al., 2021). The assumption on common random weights comes with simple and efficient algorithms for posterior inference, but results in limited flexibility in practice, as already noted in MacEachern (2000) and further discussed in Griffin and Steel (2006); in particular, the random partition of the observations induced by the prior process does not depend on the covariates, and the model can be regarded itself as a standard Dirichlet process, whose random locations are stochastic processes. This limitation is successfully overcome by more general formulations, which allow the weights to vary with the predictors; for example, Griffin and Steel (2006) propose an order-based dependent Dirichlet process, in which the ordering of the stick-breaking ratios depends on the covariates. More recent contributions loosen the assumption on the stochastic process $\mathcal{P}$ being marginally a Dirichlet process, in order to further enhance its flexibility and partially reduce the computational burden. Chung and Dunson (2009) and Rodriguez and Dunson (2011) define stick-breaking ratios as probit transformations of Gaussian random variables, so that the dependence on covariates can be dealt with in a more tractable framework; Rigon and Durante (2021) follow a similar approach, exploiting the logit link function to relate predictors with stick-breaking ratios. A further proposal which incorporates covariates into stick-breaking ratios via kernel functions is developed in Dunson and Park (2008), where the authors introduce kernel stick-breaking processes. The dependent Dirichlet process and its countless specifications and extensions are extensively reviewed in Quintana et al. (2022), while

Wade et al. (2023) propose a comprehensive survey on Bayesian dependent mixture models, with specific focus on their predictive structures.

A typical application of Bayesian nonparametric mixture models in the exchangeable setting is inference on the latent random partition structure (Müller and Mitra, 2013). Indeed, the almost-sure discreteness of the mixing random probability measures induces a random partition of the observations, which remains hidden behind the kernel mixing layer; observations belonging to the same group can be interpreted as a meaningful homogeneous subpopulation, whose distribution is usually characterized by a small number of parameters. In the context of Bayesian nonparametric regression models as (4.1), a latent random partition structure of the observations, depending on their predictors, is naturally induced whenever the elements in $\mathcal{P}$ are discrete random probability measures; in this setting, observations belonging to the same group are typically characterized by the same parametric regression model. The dependent Dirichlet process and related models discussed above stand out as prominent examples, even though they are rarely employed for inference on the induced random partitions, as noted in Quintana et al. (2022). Instead, a completely different and somehow orthogonal approach to the study of such latent random partition structures consists in directly defining their distributions, or equivalently their predictive urn schemes; the dependence on covariates is typically introduced by means of *similarity* functions, which measure the closeness of the predictors associated to observations belonging to the same group. A prominent example is represented by product partition models with regression on covariates, developed in Müller and Quintana (2010) and Müller et al. (2011), where product partition models (Hartigan, 1990; Barry and Hartigan, 1992) are extended to include covariates into the cohesion functions by multiplication with the similarity function; a similar construction is proposed in Park and Dunson (2010). The simple marginal and predictive structures, as well as the inherent computational advantages, represent the main appealing features of this class of models; indeed, deriving the distribution of the random partition, or the predictive scheme, induced by a collection $\mathcal{P}$ of dependent random probability measures can be difficult, and the resulting expressions are more involved. Conversely, the characterization of covariate-dependent random partition models in terms of a directing measure-valued stochastic process $\mathcal{P}$, inspired by de Finetti's representation theorem for exchangeable and partially exchangeable data, may represent an harder, or even impossible, task: in fact, most of such models are not inherently consistent with respect to the sample size (see Definition 4.3). The desirable consistency property is typically retrieved by specifying a statistical model for the covariates, that is, considering joint models for predictors and responses; the *induced conditional density approach* was first proposed by Müller et al. (1996) for modelling random mean functions, and lays behind the covariate-dependent product partition models discussed in Müller and Quintana (2010); Müller et al. (2011); Park and Dunson (2010). However, the assumption of endogenous predictors is often motivated by analytical or computational convenience, rather than suitable modeling choices; on the other hand, considering the predictors as fixed by design, or random but exogenous, is more appropriate when the focus is on the estimation of conditional densities (Quintana et al., 2022).

This chapter introduces a novel prior distribution for the collection $\mathcal{P}$ of covariate-dependent random probability measures, arising from the normalization of suitable random measures, in which dependence from the predictors is incorporated through multiplication by similarity kernels. An overview on kernel-weighted normalized random measures and their applications in the Bayesian nonparametric regression framework is presented in Section 4.2; remarkable special cases, as for analytical tractability, are the kernel-weighted normalized gamma process (or kernel-weighted Dirichlet process), the kernel-weighted stable process, and the kernel-weighted Pitman-Yor process, obtained from the former through a suitable change of measure and representing the main object of interest for this work. The marginal properties of such proposal are discussed in Section 4.3; in particular, the induced random partition model is carefully characterized, highlighting the role of the predictors into the partition structure. The posterior distribution of the stochastic process $\mathcal{P}$ is derived in Section 4.4, conditionally on a suitable sequence of latent variables. Remarkably, a quasi-conjugacy property is shown to hold, provided the change of measure which defines the prior process is regarded as a special case of a broader class of possible transformations which introduce non-homogeneity within normalized random measures; the properties and implications of such non-homogeneous prior processes choice are discussed in Chapter 5. Finally, Section 4.5 characterizes the predictive mechanism for both the new observations and corresponding latent variables, shedding light on the way predictors from previous observations, as well as the new predictors, affect the induced urn sampling scheme.

4.2 Kernel-weighted normalized random measures

In the Bayesian nonparametric regression literature, kernel weighting approaches have been successfully employed to define mixture models with covariate-dependent random mixing weights; a prominent example within the dependent Dirichlet process framework is represented by kernel stick-breaking processes (Dunson and Park, 2008), in which stick-breaking ratios are dampened by a suitable kernel, accounting for the similarity between the predictor and cluster-specific random locations. A similar approach underpins the construction of kernel-weighted normalized random measures.

Consider the complete and separable metric spaces $(\mathbb{X}, \mathcal{X})$ and $(\mathbb{B}, \mathcal{B})$, denoting, respectively, the space of covariates and the space of regression parameters, and let $\tilde{\mu}$ be a random measure on the product space $\mathbb{X} \times \mathbb{B}$. A kernel-weighted normalized random measure is a measure-valued stochastic process $\mathcal{P} := \{\, \tilde{p}_x \colon x \in \mathbb{X} \,\}$ such that, for each predictor $x \in \mathbb{X}$,

$$\tilde{p}_x(d\beta) := \frac{\displaystyle\int_{\mathbb{X}} k(x, \xi)\, \tilde{\mu}(d\xi, d\beta)}{\displaystyle\int_{\mathbb{X}\times\mathbb{B}} k(x, \xi)\, \tilde{\mu}(d\xi, dw)}, \tag{4.3}$$

is a random probability measure on the space of regression parameters, where $k \colon \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}^+$ is

a non-negative and bounded similarity kernel. The usual assumption in Bayesian nonparametrics is considering almost-surely discrete random measures, so that the random measure $\tilde{\mu}$, henceforth referred to as *common* random measure, can be represented as

$$\tilde{\mu}(d\xi, d\beta) = \sum_{h\geq 1} S_h \, \delta_{(\xi_h^*, \beta_h^*)}(d\xi, d\beta), \tag{4.4}$$

where $(S_h)_{h\geq 1}$ is a sequence of positive random jumps, while $(\xi_h^*)_{h\geq 1}$ and $(\beta_h^*)_{h\geq 1}$ are sequences of random locations in $\mathbb{X}$ and random parameters in $\mathbb{B}$, respectively; moreover, random locations and parameters are typically characterized by diffuse probability measures, which implies that the elements in the sequences $(\xi_h^*)_{h\geq 1}$ and $(\beta_h^*)_{h\geq 1}$ are almost-surely distinct. As a result, the random probability measure in (4.3) can be represented as the discrete measure

$$\tilde{p}_x(d\beta) := \sum_{h\geq 1} \pi_h(x) \delta_{\beta_h^*}(d\beta), \qquad \pi_h(x) := \frac{S_h \, k(x, \xi_h^*)}{\sum_{\ell\geq 1} S_\ell \, k(x, \xi_\ell^*)}, \tag{4.5}$$

that is, a normalized random measure with kernel-weighted jumps. In other words, for each predictor $x \in \mathbb{X}$, the random jumps of the common random measure $\tilde{\mu}$ are rescaled according to the similarities (measured by a similarity kernel) between the predictor and the corresponding random locations; these kernel-weighted jumps are then normalized and assigned to corresponding random parameters. In light of the interpretation of kernel evaluations as measures of similarities, each random location $\xi_h^* \in \mathbb{X}$ can be regarded as the typical or representative covariate associated to the regression model with parameters β_h^*; indeed, observations whose predictors have larger similarities with location ξ_h^* are more likely to be assigned to the regression model parameterized by β_h^*. On the other hand, the subset of predictors having larger similarities with location ξ_h^* represents the region of $\mathbb{X}$ in which the regression model parameterized by β_h^* is more appropriate; see Antoniano-Villalobos et al. (2014) for further comments on this interpretation.

The specification of the similarity kernel is a fundamental modeling aspect. A natural choice is represented by similarity kernels based on the distance between the predictor and the random location, whenever the space of covariates is assumed to be a metric space; however, in principle, every alternative measure of similarity or dissimilarity in more general spaces can be adopted to define a similarity kernel, as neither symmetry nor triangular inequality are actually needed. The most common specifications are the rectangular (or box) kernel and the square exponential (or Gaussian) kernel (Foti and Williamson, 2012), defined respectively as

$$k(x, \xi) = \mathbb{1}(d(x, \xi) \leq \eta), \qquad k(x, \xi) = \exp\left\{-\eta \, d(x, \xi)^2\right\},$$

where $\eta > 0$ is the kernel parameters and $d\colon \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}^+$ is a distance function. Note that one may assume, without loss of generality, that kernel functions take values in $[0, 1]$: indeed, any multiplicative constant would disappear with the normalization. Moreover, in presence of multiple types of covariates (e.g. continuous, ordinal, categorical, functional), different kernel

specifications may be combined.

Bayesian nonparametric regression models based on kernel-weighted normalized random measures have been already considered in the literature, particularly within the machine learning community. Rao and Teh (2009) introduce the spatial normalized gamma process and spatial normalized random measures, in which a subset of the locations and corresponding jumps are available, for each value of the predictor; their proposal can be regarded as a kernel-weighted normalized random measure with rectangular kernel and a completely random measure as common random measure. A similar approach is adopted by Griffin (2011), in which an exponential kernel is considered; such specification coincides with the normalization of a non-Gaussian Ornstein–Uhlenbeck processes, as defined by Barndorff-Nielsen and Shephard (2001). These constructions are extended to arbitrary kernel choices by Foti and Williamson (2012), where the authors devise a slice sampling algorithm for posterior inference; however, they restrict to a discrete space of covariates. Antoniano-Villalobos et al. (2014) also allow for general kernel specifications and consider stick-breaking priors (Ishwaran and James, 2001) as common random measures. A different approach is proposed by Dunson et al. (2007), where the common random measure is a finite collection of gamma random jumps located at the observed covariates; however, such sample-dependent prior specification lacks consistency and desirable marginalization properties, and is therefore not appealing from the Bayesian perspective (Dunson, 2010; Wade et al., 2023). Finally, Griffin and Leisen (2018) consider a Bayesian nonparametric regression model in which the squared distance function in the Gaussian kernel specification is replaced by a stochastic process. Their construction can be regarded as a more flexible version of a kernel-weighted normalized random measure: indeed, the random distance function $d(x, \xi)$, parameterized by the random variable ξ and interpreted as the distance from the representative covariate, is extended to a nonparametric stochastic process $r(x)$. On the other hand, such enhanced flexibility comes with less interpretability, as each mixture component cannot be interpreted as a *localized* parametric regression models, centered around its representative covariate.

A remarkable feature shared by the contributions considered above is their conditional approach: indeed, the analytical derivations discussed in these papers, as well as the posterior sampling algorithms proposed therein, are obtained conditionally on the random jumps and locations of the common random measure. Specifically, Foti and Williamson (2012) and Antoniano-Villalobos et al. (2014) design tailored conditional slice sampling algorithms, Griffin and Leisen (2018) suggest an hybrid marginal-conditional sampler, while Dunson et al. (2007) and Rao and Teh (2009) reframe their proposals as finite, sample-dependent mixtures of Dirichlet processes and devise suitable variations of the standard Polya urn scheme, conditionally on the allocations of the observations to mixture components. On the other hand, there is no systematic analysis of the marginal properties of the nonparametric regression models they introduce: for example, they lack any reference to the induced random partition structure, and its dependence on covariates. The construction proposed in this chapter aims at addressing this issue: indeed, it is characterized

by great flexibility and inherent consistency, but retains some analytical tractability after the marginalization of the common random measure (Section 4.3).

Consider a σ-stable completely random measure $\tilde{\mu}_\sigma$ on the product space $\mathbb{X} \times \mathbb{B}$, characterized by the Lévy intensity measure

$$\nu(ds, d\xi, d\beta) = \frac{\sigma}{\Gamma(1-\sigma)}\, s^{-1-\sigma} ds\, P_0(d\xi)\, Q_0(d\beta), \tag{4.6}$$

where $\sigma \in [0, 1]$, while P_0 and Q_0 are diffuse probability measures on $\mathbb{X}$ and $\mathbb{B}$, respectively. Along the lines of the construction of the Pitman-Yor process as a normalized random measure, presented in Section 1.3, define the distribution of the common random measure $\tilde{\mu}$ as absolutely continuous with respect to the distribution of $\tilde{\mu}_\sigma$, having Radon-Nikodym derivative

$$\frac{d\mathcal{L}(\tilde{\mu})}{d\mathcal{L}(\tilde{\mu}_\sigma)}(m) = \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)}\, m(\mathbb{X} \times \mathbb{B})^{-\theta}, \tag{4.7}$$

where $\theta > -\sigma$, as in (1.14). The kernel-weighted normalized random measure having $\tilde{\mu}$ as common random measure is termed *kernel-weighted Pitman-Yor process* with *concentration* parameter θ and *discount* parameter σ; the rest of the chapter is devoted to the characterization of its marginal, posterior and predictive properties within the Bayesian nonparametric regression framework.

4.3 Marginal distribution and partition probability function

Consider a sequence of observed responses $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and predictors $\boldsymbol{X} = (X_1, \ldots, X_n)$, where $Y_i \in \mathbb{R}$ is a continuous univariate response and $X_i \in \mathbb{X}$ is the corresponding covariate, which can potentially be multivariate and include elements of different types; the likelihood function associated to the nonparametric regression model in (4.1), with the kernel-weighted normalized random measure in (4.3) as directing random probability measure, is

$$\mathcal{L}(\tilde{\mu}; \boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^{n} \int_{\mathbb{B}} \phi(Y_i; X_i, \beta)\, \tilde{p}_{X_i}(d\beta) = \prod_{i=1}^{n} \frac{\displaystyle\int_{\mathbb{X}\times\mathbb{B}} \phi(Y_i; X_i, \beta)\, k(X_i, \xi)\, \tilde{\mu}(d\xi, d\beta)}{\displaystyle\int_{\mathbb{X}\times\mathbb{B}} k(X_i, \xi)\, \tilde{\mu}(d\xi, d\beta)}.$$

The integration at the numerator can be removed by introducing suitable sequences of latent variables, corresponding to the latent random locations and regression parameters from the common random measure, as typically proposed in presence of mixture models; refer to Section 2.5 and Section 3.3 for further applications of this approach. Specifically, let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ be the latent sequence of locations and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$ the latent sequence of regression parameters;

the resulting augmented likelihood is

$$\mathcal{L}(\tilde{\mu}; \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\xi}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\phi(Y_i; X_i, \beta_i)\, k(X_i, \xi_i)\, \tilde{\mu}(d\xi_i, d\beta_i)}{\int_{\mathbb{X}\times\mathbb{B}} k(X_i, \xi)\, \tilde{\mu}(d\xi, d\beta)}.$$

By using a simple analytical manipulation based on the density of an exponential random variable, which is commonly adopted for normalized random measures (James et al., 2009; Camerlenghi et al., 2019), the expression above can be rewritten as

$$\begin{aligned}&\mathcal{L}(\tilde{\mu}; \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\xi}, \boldsymbol{\beta}) \\ &\qquad = \prod_{i=1}^{n} \phi(Y_i; X_i, \beta_i)\, k(X_i, \xi_i)\, \tilde{\mu}(d\xi_i, d\beta_i) \int_{\mathbb{R}^+} \exp\left\{ - \int_{\mathbb{X}\times\mathbb{B}} u_i\, k(X_i, \xi)\, \tilde{\mu}(d\xi, d\beta) \right\} du_i,\end{aligned}$$

where $u_1, \ldots, u_n \geq 0$ are additional integration variables. As a consequence of the almost-sure discreteness of the common random measure $\tilde{\mu}$, as assumed in (4.4), each sequence of latent variables features ties among its values with positive probability; in particular, the random locations in $\boldsymbol{\xi}$ and random regression parameters in $\boldsymbol{\beta}$ both take k distinct values, denoted respectively by $\xi_1^*, \ldots, \xi_k^*$ and $\beta_1^*, \ldots, \beta_k^*$, with same multiplicities $n_1 + \cdots + n_k = n$. Indeed, the latent random partitions of the observations induced by the elements in $\boldsymbol{\xi}$ and $\boldsymbol{\beta}$ coincide almost-surely, since they share the same common random measure; therefore, the distinct values assumed within the two sequences can be ordered so that $\beta_i = \beta_j^*$ if and only if $\xi_i = \xi_j^*$, for each $i = 1, \ldots, n$. Therefore, the augmented likelihood function can be expressed by gathering together the observations belonging to the same group within the partition structure described above; specifically,

$$\begin{aligned}&\mathcal{L}(\tilde{\mu}; \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\xi}, \boldsymbol{\beta}) = R(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})\, Q(\boldsymbol{X}, \boldsymbol{\xi}) \\ &\qquad \times \int_{(\mathbb{R}^+)^n} \exp\left\{ - \int_{\mathbb{X}\times\mathbb{B}} \left(\sum_{i=1}^{n} u_i\, k(X_i, \xi) \right) \tilde{\mu}(d\xi, d\beta) \right\} \prod_{j=1}^{k} \tilde{\mu}(d\xi_j^*, d\beta_j^*)^{n_j}\, d\boldsymbol{u}, \quad (4.8)\end{aligned}$$

where $\boldsymbol{u} = (u_1, \ldots, u_n)$ and, in order to deal with a more compact notation, the following products of kernels have been defined:

$$R(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}) = \prod_{j=1}^{k} \prod_{\{i:\ \beta_i = \beta_j^*\}} \phi(Y_i; X_i, \beta_j^*), \qquad Q(\boldsymbol{X}, \boldsymbol{\xi}) = \prod_{j=1}^{k} \prod_{\{i:\ \xi_i = \xi_j^*\}} k(X_i, \xi_j^*). \quad (4.9)$$

The expression in (4.8) highlights the different contributions of the observed responses and their corresponding covariates to the likelihood function. Indeed, while the responses $\boldsymbol{Y}$ enter the likelihood only through the kernel terms in $R(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})$, the evaluations of the similarity kernel at the predictors $\boldsymbol{X}$ appear both in the multiplicative quantity $Q(\boldsymbol{X}, \boldsymbol{\xi})$ and into the exponential

term; this is a consequence of the normalizing constant for the random probability measures in $\mathcal{P}$ being covariate-dependent, and gives a clue on the crucial role the predictors play in the partition function (Proposition 4.2).

The joint marginal distribution of the observed responses and latent locations and regression parameters, given the predictors, is obtained by marginalization of the likelihood function in (4.8) with respect to the distribution of the common random measure. Note that the conditioning with respect to the predictors $\boldsymbol{X}$ is a slight abuse of notation, and should not be regarded as a formal conditioning statement (Müller et al., 2011): indeed, according to the induced conditional approach, there is no statistical model for the covariates. Henceforth, the dependence of probabilistic statements on the covariates is often deemed implicit, as it is for the other model parameters. The rest of this section focuses on the joint marginal distribution for the kernel-weighted Pitman-Yor process, having common random measure defined in (4.7) through a transformation of a stable completely random measure; the analytical derivations are based on the key identity for completely random measures (2.10), extensively discussed in Chapter 2. Specifically, let Δ^n be the n-dimensional unit simplex, that is, the set of $(n+1)$-dimensional vectors $\boldsymbol{v} = (v_0, \ldots, v_n)$ such that $v_i \geq 0$ for each $i = 0, \ldots, n$ and $v_0 + \cdots + v_n = 1$; moreover, denote by $(a)_m$ the ascending factorial, that is $(a)_m = \Gamma(a+m)/\Gamma(a)$.

Proposition 4.1. *The joint marginal distribution of responses $\boldsymbol{Y}$, latent locations $\boldsymbol{\xi}$ and latent regression parameters $\boldsymbol{\beta}$ for a kernel-weighted Pitman-Yor process is*

$$\mathbb{P}(\boldsymbol{Y}, \boldsymbol{\xi}, \boldsymbol{\beta} \mid \boldsymbol{X}) = R(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})\, Q(\boldsymbol{X}, \boldsymbol{\xi})\, \frac{\prod_{\ell=1}^{k-1}(\theta + \ell\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1}\, P_0(d\xi_j^*)\, Q_0(d\beta_j^*)$$
$$\times \int_{\Delta^n} H_n(\boldsymbol{v})^{-(\theta+n)} \prod_{j=1}^{k} \left(\frac{K_n(\xi_j^*, \boldsymbol{v})}{H_n(\boldsymbol{v})} \right)^{\sigma - n_j} \frac{\Gamma(n+\theta)}{\Gamma(\theta)}\, v_0^{\theta-1}\, d\boldsymbol{v},$$

where the following functions are defined, for each $\boldsymbol{v} \in \Delta^n$:

$$K_n(\xi, \boldsymbol{v}) = K(\xi, \boldsymbol{v}; \boldsymbol{X}) := v_0 + \sum_{i=1}^{n} v_i\, k(X_i, \xi), \qquad H_n(\boldsymbol{v}) := \left(\int_{\mathbb{X}} K_n(\xi, \boldsymbol{v})^{\sigma} P_0(d\xi) \right)^{1/\sigma}.$$

A remarkable property of the marginal distribution presented above is the possibility to factorize its expression into the kernel terms $R(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})$ and $Q(\boldsymbol{X}, \boldsymbol{\xi})$, the marginal distribution induced by the standard Pitman-Yor process on the product space $\mathbb{X} \times \mathbb{B}$, that is

$$\frac{\prod_{\ell=1}^{k-1}(\theta + \ell\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1}\, P_0(d\xi_j^*)\, Q_0(d\beta_j^*),$$

and a term depending on both the predictors $\boldsymbol{X}$ and the random locations $\boldsymbol{\xi}$, which is a mixture over the Dirichlet distribution on the n-dimensional unit simplex Δ^n with parameters $(\theta, 1, \ldots, 1)$,

having density function

$$f_{\theta,1,\ldots,1}(\boldsymbol{v}) = \frac{\Gamma(n+\theta)}{\Gamma(\theta)}\, v_0^{\theta-1};$$

such additional term represents the main difference with respect to the probability distribution of a standard nonparametric mixture model based on the Pitman-Yor process, which is retrieved by choosing a constant and unitary kernel specification. Moreover, the joint marginal distributions for the kernel-weighted Dirichlet process and kernel-weighted σ-stable process can be derived from the expression in Proposition 4.1 by taking the limits for $\sigma \to 0$ and $\theta \to 0$, respectively; in particular, the kernel-weighted Dirichlet process represents a noteworthy special case of the construction discussed in this chapter, and deserves further *ad hoc* investigations.

Proposition 4.1 is the starting point for the derivation of many interesting properties of the proposed formulation; among them, the characterization of the latent partition structure induced by ties in the sequences of locations and regression parameters is particularly relevant. For this purpose, denote by $\boldsymbol{\Pi}$ the random partition of the observations induced by ties in the latent sequence of locations (or regression parameters), encoded into multiplicities $(n_j)_j$, and let $\boldsymbol{A} = (A_1, \ldots, A_n)$ be the vector of allocation variables, that is, $A_i = j$ if and only if $\xi_i = \xi_j^*$, or equivalently $\beta_i = \beta_j^*$.

Proposition 4.2. *The probability distribution of the random partition $\boldsymbol{\Pi}$ induced by the kernel-weighted Pitman-Yor process is*

$$\mathbb{P}(\boldsymbol{\Pi} \mid \boldsymbol{X}) = \frac{\prod_{\ell=1}^{k-1}(\theta+h\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j-1} \int_{\Delta^n} H_n(\boldsymbol{v})^{-\theta} \prod_{j=1}^{k} S_j(\boldsymbol{v})\, f_{\theta,1,\ldots,1}(\boldsymbol{v})\, d\boldsymbol{v},$$

where, for each $\boldsymbol{v} \in \Delta^n$ and each $j = 1, \ldots, k$,

$$S_j(\boldsymbol{v}) = S_j(\boldsymbol{v}; \boldsymbol{X}) := \frac{\displaystyle\int_{\mathbb{X}} \prod_{\{i:\, A_i = j\}} \frac{k(X_i, \xi)}{K_n(\xi, \boldsymbol{v})}\, K_n(\xi, \boldsymbol{v})^{\sigma}\, P_0(d\xi)}{\displaystyle\int_{\mathbb{X}} K_n(\xi, \boldsymbol{v})^{\sigma}\, P_0(d\xi)}. \tag{4.10}$$

Remarkably, the expression for the probability distribution of $\boldsymbol{\Pi}$ can be rearranged as a mixture, over the Dirichlet distribution, of product partition functions, that is,

$$\mathbb{P}(\boldsymbol{\Pi} \mid \boldsymbol{X}) = \int_{\Delta^n} \frac{\prod_{\ell=1}^{k-1}(\theta+\ell\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j-1}\, S_j(\boldsymbol{v})\, H_n(\boldsymbol{v})^{-\theta}\, f_{\theta,1,\ldots,1}(\boldsymbol{v})\, d\boldsymbol{v}.$$

Specifically, the structure of such product partition functions, namely

$$\frac{\prod_{\ell=1}^{k-1}(\theta+\ell\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j-1}\, S_j(\boldsymbol{v}), \qquad \boldsymbol{v} \in \Delta^n,$$

resembles the characterization of PPMx models proposed in Müller and Quintana (2010) and Müller et al. (2011): indeed, for each group $j = 1, \dots, k$ of the partition, the quantity $(1-\sigma)_{n_j-1}$ plays the role of the *cohesion* function in the exchangeable setting (Hartigan, 1990), while the quantity $S_j(\boldsymbol{v})$ can be regarded as the *similarity* function of Müller and Quintana (2010). Moreover, the quantities in (4.10) can be rewritten as

$$S_j(\boldsymbol{v}) = S_j(\boldsymbol{v}; \boldsymbol{X}) := \int_{\mathbb{X}} \prod_{\{i:\ A_i=j\}} \frac{k(X_i, \xi)}{K_n(\xi, \boldsymbol{v})}\, g(\xi; \boldsymbol{v})\, d\xi, \qquad j = 1, \dots, k,$$

where $g(\xi; \boldsymbol{v}) \propto K_n(\xi, \boldsymbol{v})^{\sigma} P_0(d\xi)$ is a density function; this expression parallels the structure of the similarity function proposed for PPMx models, with the ratio $k(\cdot, \xi)/K_n(\xi, \boldsymbol{v})$ acting as the auxiliary probabilistic model for the covariates, and $g(\xi; \boldsymbol{v})$ representing the density function for the latent location ξ. However, differently from the PPMx construction, these quantities depend on the entire sequence of covariates $\boldsymbol{X}$ through the function $K_n(\xi, \boldsymbol{v})$, which is the convex combination of the similarities between the predictors and the latent location ξ, with proportions given by $\boldsymbol{v}$. Therefore, each ratio $k(X_i, \xi)/K_n(\xi, \boldsymbol{v})$ can be considered a *relative* measure of similarity, as it compares the absolute similarity between the covariate X_i and the latent location ξ with the average similarity $K_n(\xi, \boldsymbol{v})$ recorded by the covariates in $\boldsymbol{X}$; moreover, the distribution of the latent location ξ is a transformation of its prior distribution P_0, which takes into account where predictors in $\boldsymbol{X}$ are located.

The results discussed in this section may suggest that, in order to define a covariate-dependent product partition model which is inherently consistent with respect to the sample size, the similarity function should also depend on (some function of) the entire sequence of covariates, rather than just on the group-specific predictors. For the purpose of clarifying this fundamental and desirable property of Bayesian nonparametric regression models, a possible definition of consistency is presented.

Definition 4.3. Let Π_n be a partition of the first n observations, and denote by k_n the number of partition groups; a sequence $\mathcal{P}_1(\Pi_1), \mathcal{P}_2(\Pi_2), \dots$ of partition functions, that is, of probability distributions on the space of partitions, is *consistent* (with respect to the sample size) if, for every $n \geq 1$,

$$\mathcal{P}_n(\Pi_n) = \sum_{A_{n+1}=1}^{k_{n+1}} \mathcal{P}_{n+1}(\Pi_{n+1}), \qquad n \geq 1, \tag{4.11}$$

where A_{n+1} denotes the group, within the partition Π_{n+1}, to which observation $n+1$ is allocated.

In the context of partition models with dependence on predictors, the consistency condition defined above is required to hold for every possible value of the predictor X_{n+1} associated to the

$(n+1)$-th observation, that is

$$\mathcal{P}_n(\Pi_n \mid \boldsymbol{X}) = \sum_{A_{n+1}=1}^{k_{n+1}} \mathcal{P}_{n+1}(\Pi_{n+1} \mid X_{n+1}, \boldsymbol{X}), \qquad \forall\, X_{n+1} \in \mathbb{X},$$

where $\boldsymbol{X}$ denotes the predictors vectors associated to the previous n observations. Such requirement is satisfied by construction whenever the sequence of partition functions is obtained through the marginalization of a covariate-dependent random probability measure: therefore, the partition function described in Proposition 4.2 is inherently consistent. Remarkably, the verification of condition (4.11) starting from the explicit expression of the partition function is not straightforward. On the other hand, such consistency condition usually fails to hold for models specified through their predictive distributions, as already discussed in Section 4.1; in fact, consistency is typically retrieved by specifying a statistical model for the covariates, which amounts to relaxing the consistency condition in (4.11) to

$$\mathcal{P}_n(\Pi_n \mid \boldsymbol{X}) = \sum_{A_{n+1}=1}^{k_{n+1}} \int_{\mathbb{X}} \mathcal{P}_{n+1}(\Pi_{n+1} \mid x, \boldsymbol{X})\, q(x|\boldsymbol{X})\, dx,$$

where $q(x|\boldsymbol{X})$ denotes the predictive distribution for the auxiliary covariates model. The relaxed condition above is explicitly stated as a desirable coherence property in Müller et al. (2011), where it is shown to hold for the PPMx models; Park and Dunson (2010) also refer to such relaxed formulation. In conclusion, the relationships between covariate-dependent partition models enjoying convenient analytical structures (such as a product form) and consistency conditions with respect to the sample size deserve further investigations.

4.4 Latent variables and posterior characterization

The expression for joint marginal distribution in Proposition 4.1 involves an integral with respect to the integration variables $\boldsymbol{v} = (v_0, \ldots, v_n)$ on the n-dimensional unit simplex, which suggests the introduction of a further vector of auxiliary latent variables $\boldsymbol{V} = (V_0, \ldots, V_n)$; indeed, the result in Proposition 4.1 can be obtained by marginalizing, with respect to the additional latent variables $\boldsymbol{V}$, the augmented joint marginal distribution

$$\begin{aligned}\mathbb{P}(\boldsymbol{Y}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{V} \mid \boldsymbol{X}) = R(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})\, Q(\boldsymbol{X}, \boldsymbol{\xi})\, \frac{\prod_{\ell=1}^{k-1}(\theta + \ell\sigma)}{(\theta+1)_{n-1}}\, H_n(\boldsymbol{V})^{-(\theta+n)} \\ \times \prod_{j=1}^{k} (1-\sigma)_{n_j-1} \left(\frac{K_n(\xi_j^*, \boldsymbol{V})}{H_n(\boldsymbol{V})}\right)^{\sigma-n_j} P_0(d\xi_j^*)\, Q_0(d\beta_j^*)\, f_{\theta,1,\ldots,1}(\boldsymbol{V})\, d\boldsymbol{V}.\end{aligned}$$

As discussed in the previous section, the latent variables $\boldsymbol{V}$ can be interpreted as the proportions in the convex combination of similarities between predictors and each latent location, which

enters the marginal distribution through functions K_n and H_n; in other words, each variable V_i measures the relative influence of predictor X_i into the learning mechanism.

This augmented latent space is essential for the derivation of the posterior distribution of the common random measure $\tilde{\mu}$, which is characterized conditionally on the latent random locations, the corresponding regression parameters, and the additional latent variables introduced above. Specifically, let $\boldsymbol{V} = (V_0, \ldots, V_n)$ be a vector of latent variables on the n-dimensional unit simplex, having density function, given random locations $\boldsymbol{\xi}$, proportional to

$$f_{\boldsymbol{V}}(\boldsymbol{v} \mid \boldsymbol{\xi}) \propto H_n(\boldsymbol{v})^{-(\theta+n)} \prod_{j=1}^{k} \left(\frac{K_n(\xi_j^*, \boldsymbol{v})}{H_n(\boldsymbol{v})} \right)^{\sigma - n_j} f_{\theta,1,\ldots,1}(\boldsymbol{v}), \tag{4.12}$$

where $f_{\theta,1,\ldots,1}(\boldsymbol{v})$ is the density function of the Dirichlet distribution with parameters $(\theta, 1, \ldots, 1)$. Moreover, let T be an auxiliary random variable such that T^σ, given the number of groups k of the random partition induced by $\boldsymbol{\xi}$, has gamma distribution with shape parameter $k + \theta/\sigma$ and unit rate, that is, the random variable T has generalized gamma distribution with density function

$$f_T(t \mid \boldsymbol{\xi}) = \frac{\sigma}{\Gamma(k + \theta/\sigma)} \, t^{k\sigma + \theta - 1} e^{-t^\sigma}.$$

Proposition 4.4. *The posterior distribution of the common random measure $\tilde{\mu}$ for the kernel-weighted Pitman-Yor process, given the latent locations $\boldsymbol{\xi}$ and regression parameters $\boldsymbol{\beta}$, and the latent variables $\boldsymbol{V}$ and T, coincides with the distribution of the random measure*

$$\tilde{\mu}(d\xi, d\beta) \mid \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{V}, T \;\sim\; \tilde{\mu}^*(d\xi, d\beta) + \sum_{j=1}^{k} W_j \, \delta_{(\xi_j^*, \beta_j^*)}(d\xi, d\beta),$$

where the random elements in the sum are mutually independent, $\tilde{\mu}^*$ is a generalized gamma completely random measure, having non-homogeneous Lévy intensity

$$\nu^*(ds, d\xi, d\beta) = \frac{\sigma}{\Gamma(1-\sigma)} \, s^{-1-\sigma} \exp\left\{ -\frac{K_n(\xi, \boldsymbol{V})}{H_n(\boldsymbol{V})} \, T \, s \right\} ds \, P_0(d\xi) \, Q_0(d\beta),$$

and each W_j is a non-negative random variable having gamma distribution

$$W_j \;\sim\; \text{Gamma}\left(n_j - \sigma, \; \frac{K_n(\xi_j^*, \boldsymbol{V})}{H_n(\boldsymbol{V})} \, T \right).$$

The posterior characterization presented above highlights the role of the covariates within the learning mechanism. Indeed, the common random measure a posteriori is a nonhomogeneous random measure, that is, its random jumps depend on its (fixed or random) locations; such dependence is induced by function $K_n(\xi, \boldsymbol{v})$, which is a convex combination of predictors' similarities with the location ξ. Intuitively, the quantity $K_n(\xi, \boldsymbol{v})$ carries the information about the structure of the space of covariates induced by the predictors, suggesting that a careful choice

of the most suitable similarity function should be considered.

The auxiliary random variable T is introduced in order to derive a conditional posterior characterization of the common random measure $\tilde{\mu}$ as a completely random measure (Section 1.1), obtained as the sum of a generalized gamma completely random measure and a finite sequence of random jumps at fixed locations. However, the distribution of the continuous component $\tilde{\mu}^*$ of the posterior can be characterized unconditionally with respect to T, leading to a interesting quasi-conjugacy result; to this end, consider the σ-stable completely random measure $\tilde{\mu}_\sigma$ characterized by the Lévy intensity measure in (4.6).

Proposition 4.5. *The distribution of the random measure $\tilde{\mu}^*$ introduced in Proposition 4.4, given the locations $\boldsymbol{\xi}$, the regression parameters $\boldsymbol{\beta}$ and the latent variables $\boldsymbol{V}$, is absolutely continuous with respect to the distribution of $\tilde{\mu}_\sigma$, with Radon-Nikodym derivative*

$$\frac{d\mathcal{L}(\tilde{\mu}^*)}{d\mathcal{L}(\tilde{\mu}_\sigma)}(m) = \frac{\sigma\Gamma(k\sigma+\theta)}{\Gamma(k+\theta/\sigma)} \left(\int_{\mathbb{X}\times\mathbb{B}} \frac{K_n(\xi, \boldsymbol{V})}{H_n(\boldsymbol{V})}\, m(d\xi, d\beta) \right)^{-(k\sigma+\theta)}.$$

The concept of quasi-conjugacy is introduced by Lijoi et al. (2008) to describe a characterizing property of the Pitman-Yor process within the class of Gibbs-type priors: conditionally on the observed distinct values and multiplicities, the distribution of the process generating the new distinct values belongs to the same class of the prior distribution, with updated parameters. As for the kernel-weighted Pitman-Yor process considered in this chapter, the prior distribution of the common random measure $\tilde{\mu}$ is defined by the change of measure in (4.7),

$$\frac{d\mathcal{L}(\tilde{\mu})}{d\mathcal{L}(\tilde{\mu}_\sigma)}(m) = \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)} \left(\int_{\mathbb{X}\times\mathbb{B}} m(d\xi, d\beta) \right)^{-\theta},$$

while the distribution of the process generating the new distinct values, namely the random measure $\tilde{\mu}^*$, is characterized in Proposition 4.5, conditionally on the latent variables $\boldsymbol{V}$, as

$$\frac{d\mathcal{L}(\tilde{\mu}^*)}{d\mathcal{L}(\tilde{\mu}_\sigma)}(m) = \frac{\sigma\Gamma(k\sigma+\theta)}{\Gamma(k+\theta/\sigma)} \left(\int_{\mathbb{X}\times\mathbb{B}} \frac{K_n(\xi, \boldsymbol{V})}{H_n(\boldsymbol{V})}\, m(d\xi, d\beta) \right)^{-(k\sigma+\theta)},$$

where $\tilde{\mu}_\sigma$ is a σ-stable completely random measure. In light of the clear analogies between these expressions, consider the class of random measures on the product space $\mathbb{X}\times\mathbb{B}$ whose probability distribution is absolutely continuous with respect to the random measure $\tilde{\mu}_\sigma$, with Radon-Nikodym derivative

$$\frac{\sigma\Gamma(\eta)}{\Gamma(\eta/\sigma)} \left(\int_{\mathbb{X}\times\mathbb{B}} G(\xi)\, \tilde{\mu}_\sigma(d\xi, d\beta) \right)^{-\eta},$$

where $\eta > -\sigma$ and $G\colon \mathbb{X} \mapsto \mathbb{R}^+$ is a non-negative and bounded measurable function. The kernel-weighted Pitman-Yor process can be regarded as a *conditionally* quasi-conjugate prior distribution, since the random measures $\tilde{\mu}$ and $\tilde{\mu}^*$ both belong to the class described above; in

particular, the random measure $\tilde{\mu}$ is characterized by parameters $\eta = \theta$ and $G(\xi) = 1$, while the random measure $\tilde{\mu}^*$ is retrieved for $\eta = k\sigma + \theta$ and $G(\xi) \propto K_n(\xi, \boldsymbol{V})$. Random measures belonging to this novel class are extensively analyzed in Chapter 5, where they are normalized and considered as random probability measures within the exchangeable framework; the reader is referred there for further details.

4.5 Predictive distribution

Prediction is often deemed as the ultimate goal of the Bayesian paradigm, especially within the Bayesian nonparametric approach: this section is devoted to the characterization of the predictive structure and properties for the kernel-weighted Pitman-Yor process introduced in the previous sections. For this purpose, consider a sequence of m additional responses $\boldsymbol{Y}^+ = (Y_{n+1}, \dots, Y_{n+m})$ and corresponding predictors $\boldsymbol{X}^+ = (X_{n+1}, \dots, X_{n+m})$, where $Y_i \in \mathbb{R}$ is a continuous univariate response and $X_i \in \mathbb{X}$ a potentially multivariate predictor; moreover, let $\boldsymbol{\xi}^+ = (\xi_{n+1}, \dots, \xi_{n+m})$ be the latent sequence of locations and $\boldsymbol{\beta}^+ = (\beta_{n+1}, \dots, \beta_{n+m})$ the latent sequence of regression parameters. The joint predictive distribution of interest is

$$\mathbb{P}(\boldsymbol{Y}^+, \boldsymbol{\xi}^+, \boldsymbol{\beta}^+ \mid \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{V}),$$

where $\boldsymbol{V}$ is the vector of auxiliary latent variables introduced in Section 4.4. Note that, as already discussed, the dependence on the covariates is considered implicit, as there is no statistical model specified on them.

Along the same lines as Section 4.3, the almost-sure discreteness of the common random measure $\tilde{\mu}$ induces ties both within the sequences of additional latent locations and regression parameter and with values contained into the previous latent sequences. Specifically, the random locations in $\boldsymbol{\xi}^+$ and random regression parameters in $\boldsymbol{\beta}^+$ both take h additional distinct values, denoted respectively by $\xi^*_{k+1}, \dots, \xi^*_{k+h}$ and $\beta^*_{k+1}, \dots, \beta^*_{k+h}$, besides possibly taking some of the previous distinct values $\xi^*_1, \dots, \xi^*_k$ and $\beta^*_1, \dots, \beta^*_k$, respectively; the previous values are taken with multiplicities $m_1, \dots, m_k \geq 0$, while the new values are taken with multiplicities $m_{k+1}, \dots, m_{k+h} \geq 1$, such that $m_1 + \cdots + m_{k+h} = m$. Similarly, the random partitions of the additional observations induced by the elements in $\boldsymbol{\xi}^+$ and $\boldsymbol{\beta}^+$ coincide almost-surely, and therefore the additional distinct values can be ordered so that $\beta_i = \beta^*_j$ if and only if $\xi_i = \xi^*_j$, for each $i = n+1, \dots, n+m$. Finally, define the products of kernels

$$R(\boldsymbol{Y}^+, \boldsymbol{\beta}^+) = \prod_{j=1}^{k+h} \prod_{\{i>n:\ \beta_i = \beta^*_j\}} \phi(Y_i; X_i, \beta^*_j), \qquad Q(\boldsymbol{X}^+, \boldsymbol{\xi}^+) = \prod_{j=1}^{k+h} \prod_{\{i>n:\ \xi_i = \xi^*_j\}} k(X_i, \xi^*_j).$$

Proposition 4.6. *The joint predictive distribution for the additional responses $\boldsymbol{Y}^+$, and corresponding latent locations $\boldsymbol{\xi}^+$ and regression parameters $\boldsymbol{\beta}^+$, given the previous latent locations $\boldsymbol{\xi}$,*

regression parameters $\boldsymbol{\beta}$ *and latent variables* $\boldsymbol{V}$*, is*

$$\begin{aligned}\mathbb{P}(\boldsymbol{Y}^+, \boldsymbol{\xi}^+, \boldsymbol{\beta}^+ \mid \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{V}) = R(\boldsymbol{Y}^+, \boldsymbol{\beta}^+)\, Q(\boldsymbol{X}^+, \boldsymbol{\xi}^+)\, H_n(\boldsymbol{V})^{\theta+n} \prod_{j=1}^{k} \left(\frac{K_n(\xi_j^*, \boldsymbol{V})}{H_n(\boldsymbol{V})} \right)^{n_j - \sigma} \\ \times \frac{\prod_{\ell=0}^{h-1}(\theta + k\sigma + \ell\sigma)}{(\theta+n)_m} \prod_{j=1}^{k} (n_j - \sigma)_{m_j} \prod_{j=k+1}^{k+h} (1-\sigma)_{m_j - 1}\, P_0(d\xi_j^*)\, Q_0(d\beta_j^*) \\ \times \int_{\Delta^m} H_m^+(\boldsymbol{w}; \boldsymbol{V})^{-(\theta+n+m)} \prod_{j=1}^{k} \left(\frac{K_m^+(\xi_j^*, \boldsymbol{w}; \boldsymbol{V})}{H_m^+(\boldsymbol{w}; \boldsymbol{V})} \right)^{\sigma - n_j - m_j} \\ \times \prod_{j=k+1}^{k+h} \left(\frac{K_m^+(\xi_j^*, \boldsymbol{w}; \boldsymbol{V})}{H_m^+(\boldsymbol{w}; \boldsymbol{V})} \right)^{\sigma - m_j} f_{\theta+n, 1, \ldots, 1}(\boldsymbol{w})\, d\boldsymbol{w},\end{aligned}$$

where, for each $\boldsymbol{w} \in \Delta^m$*,*

$$K_m^+(\xi, \boldsymbol{w}; \boldsymbol{v}) = K_m^+(\xi, \boldsymbol{w}; \boldsymbol{v}, \boldsymbol{X}^+) := w_0\, K_n(\xi, \boldsymbol{v}) + \sum_{i=1}^{m} w_i\, k(X_{n+i}, \xi),$$

$$H_m^+(\boldsymbol{w}; \boldsymbol{v}) = \left(\int_{\mathbb{X}} K_m^+(\xi, \boldsymbol{w}; \boldsymbol{v})^{\sigma} P_0(d\xi) \right)^{1/\sigma}.$$

This result can be derived adopting two alternative approaches: (i) a *marginal* strategy, which computes the predictive distribution as the ratio between the joint marginal distribution of the entire sequence of $n+m$ observations and the joint marginal distribution of the first n observations; (ii) a *conditional* strategy, which marginalizes the augmented likelihood of the additional m observations with respect to the posterior distribution of the common random measure, given the previous n observations. As a consequence of the principled Bayesian standpoint considered in this chapter, both approaches lead to the same conclusion; proofs of Proposition 4.6 using both the marginal and conditional strategies can be found in Appendix A.3.

In analogy with the marginal distribution presented in Proposition 4.1, the expression above can be factorized into the kernel terms $R(\boldsymbol{Y}^+, \boldsymbol{\beta}^+)$ and $Q(\boldsymbol{X}^+, \boldsymbol{\xi}^+)$, the predictive distribution induced by the standard Pitman-Yor process on the product space $\mathbb{X} \times \mathbb{B}$, conditionally on hypothetical observations $(\boldsymbol{\xi}, \boldsymbol{\beta})$, that is

$$\frac{\prod_{\ell=0}^{h-1}(\theta + k\sigma + \ell\sigma)}{(\theta+n)_m} \prod_{j=1}^{k} (n_j - \sigma)_{m_j} \prod_{j=k+1}^{k+h} (1-\sigma)_{m_j - 1}\, P_0(d\xi_j^*)\, Q_0(d\beta_j^*),$$

and a term depending on both the predictors $\boldsymbol{X}^+$ and the random locations $\boldsymbol{\xi}^+$, as well as on the previous predictors $\boldsymbol{X}$ and locations $\boldsymbol{\xi}$, which is a mixture over the Dirichlet distribution on the m-dimensional unit simplex Δ^m with parameters $(\theta + n, 1, \ldots, 1)$. Note that the predictive distribution depends on the observations only through their latent locations $\boldsymbol{\xi}$, regression param-

eters $\boldsymbol{\beta}$ and the corresponding covariates $\boldsymbol{X}$, while the observed responses $\boldsymbol{Y}$ do not have a direct effect on the predictive mechanism; this is customary to mixture models, in which observations are conditionally independent given their allocation within the latent partition structure.

A relevant role in the Bayesian paradigm is played by the one-step-ahead predictive distribution, that is, the predictive distribution for the next observation, which is obtained restricting Proposition 4.6 to the case $m = 1$.

Corollary 4.7. *The joint predictive distribution for the additional response Y_{n+1}, and corresponding latent location ξ_{n+1} and regression parameter β_{n+1}, given the previous latent locations $\boldsymbol{\xi}$, regression parameters $\boldsymbol{\beta}$ and latent variables $\boldsymbol{V}$, is proportional to*

$$\begin{aligned}\mathbb{P}(Y_{n+1} \in dy, \xi_{n+1} \in d\xi, \beta_{n+1} \in d\beta \mid \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{V}) \;\propto\; \phi(y; X_{n+1}, \beta)\, dy\; k(X_{n+1}, \xi) \\ \times \int_0^1 \left(\frac{\theta + k\sigma}{\theta + n} \frac{H_1^+(w; \boldsymbol{V})^{1-\sigma}}{K_1^+(\xi, w; \boldsymbol{V})^{1-\sigma}} P_0(d\xi)\, Q_0(d\beta) + \sum_{j=1}^{k} \frac{n_j - \sigma}{\theta + n} \frac{H_1^+(w; \boldsymbol{V})}{K_1^+(\xi_j^*, w; \boldsymbol{V})} \delta_{(\xi_j^*, \beta_j^*)}(d\xi, d\beta) \right) \\ \times H_1^+(w; \boldsymbol{V})^{-(\theta+n+1)} \prod_{j=1}^{k} \left(\frac{K^+(\xi_j^*, w; \boldsymbol{V})}{H_1^+(w; \boldsymbol{V})} \right)^{\sigma - n_j} f_{1,\theta+n}(w)\, dw,\end{aligned}$$

where $f_{1,\theta+n}(w)$ is the density function of a Beta distribution with parameters 1 and $\theta + n$ and, for each $w \in [0, 1]$,

$$\begin{aligned} K_1^+(\xi, w; \boldsymbol{V}) &= w\, k(X_{n+1}, \xi) + (1 - w)\, K_n(\xi, \boldsymbol{V}), \\ H_1^+(w; \boldsymbol{V}) &= \left(\int_{\mathbb{X}} K_1^+(\xi, w; \boldsymbol{V})^{\sigma} P_0(d\xi) \right)^{1/\sigma}. \end{aligned}$$

As already noted for the marginal distribution, the integration with respect to the variable $w \in [0, 1]$ in the expression above suggests the introduction of a further latent variable W_{n+1}; this additional variable can be itself interpreted as the relative influence of the new predictor X_{n+1} into the learning mechanism, as highlighted by the expression of function K_1^+. The joint predictive distribution in Corollary 4.7 can be decomposed into a sequence of consecutive predictive distributions: this decomposition is actually the starting point for the construction of a marginal sampling algorithm (see Section 4.6)

The first and definitely more involved predictive distribution concerns the additional latent variable W_{n+1}, which has density function, given the previous locations $\boldsymbol{\xi}$ and latent variables $\boldsymbol{V}$,

proportional to

$$\left(\frac{\theta + k\sigma}{\theta + n} \int_{\mathbb{X}} \frac{k(X_{n+1}, \xi)}{K_1^+(\xi, w; \boldsymbol{v})} \left(\frac{K_1^+(\xi, w; \boldsymbol{v})}{H_1^+(w; \boldsymbol{v})} \right)^{\sigma} P_0(d\xi) + \sum_{j=1}^{k} \frac{n_j - \sigma}{\theta + n} \frac{k(X_{n+1}, \xi_j^*)}{K_1^+(\xi_j^*, w; \boldsymbol{v})} \right) \times H_1^+(w; \boldsymbol{v})^{-(\theta+n)} \prod_{j=1}^{k} \left(\frac{K_1^+(\xi_j^*, w; \boldsymbol{v})}{H_1^+(w; \boldsymbol{v})} \right)^{\sigma - n_j} f_{1,\theta+n}(w). \quad (4.13)$$

A convenient approximation of this intractable density function is obtained for large sample size, thanks to the following result; the proof presented in Appendix A.3 is an informal sketch, but the statement is expected to be formally true under mild assumptions on the sequence of predictors.

Lemma 4.8. *The predictive distribution of the latent variable W_{n+1}, given the previous locations $\boldsymbol{\xi}$ and latent variables $\boldsymbol{V}$, is such that, for $n \to \infty$,*

$$\left(\sum_{j=1}^{k} \frac{n_j}{n} \frac{k(X_{n+1}, \xi_j^*)}{K_n(\xi_j^*, \boldsymbol{V})} \right) n\, W_{n+1} \xrightarrow{d} \text{Exp}(1),$$

where convergence in distribution holds almost surely with respect to both the prior distribution $\mathcal{P}$ for the latent locations $\boldsymbol{\xi}$ and the conditional distribution of the latent variables $\boldsymbol{V}$.

The very first consequence of this Lemma is that the random variable W_{n+1} decreases to zero as $1/n$ for increasing sample size: this analytical result is actually consistent with the fundamental Bayesian assumption that the order in which observations are recorded should not impact the learning mechanism. Moreover, from the computational point of view, when the sample size n is large, an approximate sample of the latent variable W_{n+1} can be obtained dividing by n a sample from the exponential distribution with rate parameter

$$\sum_{j=1}^{k} \frac{n_j}{n} \frac{k(X_{n+1}, \xi_j^*)}{K_n(\xi_j^*, \boldsymbol{V})}.$$

This quantity is essentially the expectation, with respect to the empirical measure induced by latent locations $\boldsymbol{\xi}$, of the function $\xi \mapsto k(X_{n+1}, \xi)/K_n(\xi, \boldsymbol{V})$, which can be regarded as a relative measure of similarity for the new predictor X_{n+1}, as remarked in Section 4.3; intuitively, the less similar is the new covariate to the previous locations, compared to the other covariates, the larger the corresponding latent variable W_{n+1} is in expectation, and thus the larger the influence of the new observation in the learning mechanism.

Conditionally on the latent variable W_{n+1}, the predictive distribution for the new latent

location ξ_{n+1} can be framed as an urn sampling scheme:

$$\begin{aligned}&\mathbb{P}(\xi_{n+1} \in d\xi \mid \boldsymbol{\xi}, \boldsymbol{V}, W_{n+1} = w) \\ &\quad \propto \frac{\theta + k\sigma}{\theta + n} \frac{k(X_{n+1}, \xi)}{K_1^+(\xi, w; \boldsymbol{V})} \left(\frac{K_1^+(\xi, w; \boldsymbol{V})}{H_1^+(w; \boldsymbol{V})} \right)^{\sigma} P_0(d\xi) + \sum_{j=1}^{k} \frac{n_j - \sigma}{\theta + n} \frac{k(X_{n+1}, \xi_j^*)}{K_1^+(\xi_j^*, w; \boldsymbol{V})} \delta_{\xi_j^*}(d\xi). \quad (4.14)\end{aligned}$$

This expression is a generalization of the predictive urn scheme for the standard Pitman-Yor process,

$$\mathbb{P}(\xi_{n+1} \in d\xi \mid \boldsymbol{\xi}) = \frac{\theta + k\sigma}{\theta + n} P_0(d\xi) + \sum_{j=1}^{k} \frac{n_j - \sigma}{\theta + n} \delta_{\xi_j^*}(d\xi), \quad (4.15)$$

where the probabilities for the new location ξ_{n+1} to take either a new value or a value already taken by the previous locations $\boldsymbol{\xi}$ are rescaled according to a specific function of the new predictor. Specifically, such probabilities coincide with the allocation probabilities of a standard Pitman-Yor process on $\mathbb{X}$ with base probability measure $G(\xi) \propto K_1^+(\xi, w; \boldsymbol{V})^{\sigma} P_0(d\xi)$, rescaled by the relative similarity between the predictor X_{n+1} and location ξ, measured by the usual function $\xi \mapsto k(X_{n+1}, \xi)/K_1^+(\xi, w; \boldsymbol{V})$, which plays here the same role of a probability kernel in standard Bayesian nonparametric mixture models for exchangeable continuous observations.

A further interesting aspect of the predictive scheme discussed in this section is the updating mechanism for the function $K_n(\xi, \boldsymbol{v})$ measuring the weighted average similarity between location ξ and covariates in $\boldsymbol{X}$. The updated function after step $n+1$, incorporating the information from the additional predictor X_{n+1}, is

$$\begin{aligned}\xi \mapsto K_1^+(\xi, w; \boldsymbol{v}) &= w\, k(X_{n+1}, \xi) + (1 - w)\, K_n(\xi, \boldsymbol{v}) \\ &= w\, k(X_{n+1}, \xi) + (1 - w)\, v_0 + \sum_{i=1}^{n} (1 - w)\, v_i\, k(X_i, \xi);\end{aligned}$$

in other words, the updated function is a convex combination of the function at step n and the similarity for the new predictor, which is included with random weight W_{n+1}. Therefore, in the practical implementation of the algorithm, the vector of latent variables $\boldsymbol{V}$ at step $n+1$ is obtained from the same vector at step n and the freshly sampled latent variable W_{n+1} as

$$V_{n+1} \leftarrow W_{n+1}, \qquad V_i \leftarrow (1 - W_{n+1})\, V_i, \quad i = 0, \ldots, n.$$

The remaining predictive distributions are straightforward, once the new observation is assigned to a mixture component by the urn scheme in (4.14). Specifically, if the new observation is assigned to a new group, its regression parameters β_{n+1} are sampled from their base probability distribution Q_0; on the other hand, if $\xi_{n+1} = \xi_j^*$ for some $j = 1, \ldots, k$, that is, the new observation is assigned to group j, then $\beta_{n+1} = \beta_j^*$. In both cases, the predictive distribution for the new continuous response Y_{n+1} is characterized by the probability density $\phi(\cdot; X_{n+1}, \beta_{n+1})$.

4.6 Future developments

The marginal distribution derived in Proposition 4.1, the posterior characterization in Proposition 4.4 and the one-step-ahead predictive mechanism described in Section 4.5 provide the necessary analytical background for the development of a marginal Gibbs sampling scheme. The main computational challenge is represented by the posterior sampling of the vector of latent variables $\boldsymbol{V}$ from the density function in (4.12), as well as the sampling of the new latent variable W_{n+1} from the density in (4.13); in particular, the adoption of the approximation derived in Lemma 4.8 should be further investigated from the theoretical and algorithmic point of view, in order to carefully assess the induced approximation error.

Furthermore, a conditional slice sampling algorithm, inspired by the slice sampler proposed in Foti and Williamson (2012) for kernel-weighted normalized completely random measures, can be devised and implemented; remarkably, such sampling algorithm may benefit from the choice of a non-constant slicing threshold, which is expected to enhance the stability of the sampling scheme. In this regard, the evaluation of the Laplace transform term, which appears in the target distribution of most slice samplers, may pose some computational challenges. An interesting and convenient approach to such task is proposed by Griffin and Leisen (2018), where the authors suggest replacing the Laplace transform with an unbiased estimator; in particular, they resort to the Poisson estimator, which has been successfully adopted in the context of probabilistic inference for diffusion processes (Papaspiliopoulos, 2011).

As for the applications within the Bayesian nonparametric regression framework, the proposed kernel-weighted Pitman-Yor process can be compared with alternative models from the vast literature on the topic, extensively reviewed in Section 4.1; in particular, a thorough comparison in terms of modeling flexibility and predictive performances should be draw with respect to other kernel-based specifications, such as the kernel stick-breaking process (Dunson and Park, 2008). However, no remarkable differences are expected, as the constructions are based on similar approaches. Another interesting but less considered application is posterior inference on the induced random partition structure, as discussed in Section 4.1 and Quintana et al. (2022); indeed, the posterior characterization of the observations as samples from a mixture of *local* parametric regression models is itself appealing. On the other hand, relevant contributions may arise from the application of the kernel-weighted Pitman-Yor to species sampling problems in presence of spatial covariates; within this setting, the proposed formulation can be compared with existing partition models, such as the PPMx model (Müller and Quintana, 2010; Müller et al., 2011), which however suffer from inconsistency with respect to the sample size.

A simple exploratory illustration concludes this chapter, with the aim of providing a preliminary overview on the predictive properties of the proposed formulation. For this purpose, consider a sequence of n observations, and assume their latent partition structure is available, together with the corresponding predictors; in other words, for each observation $i = 1, \dots, n$,

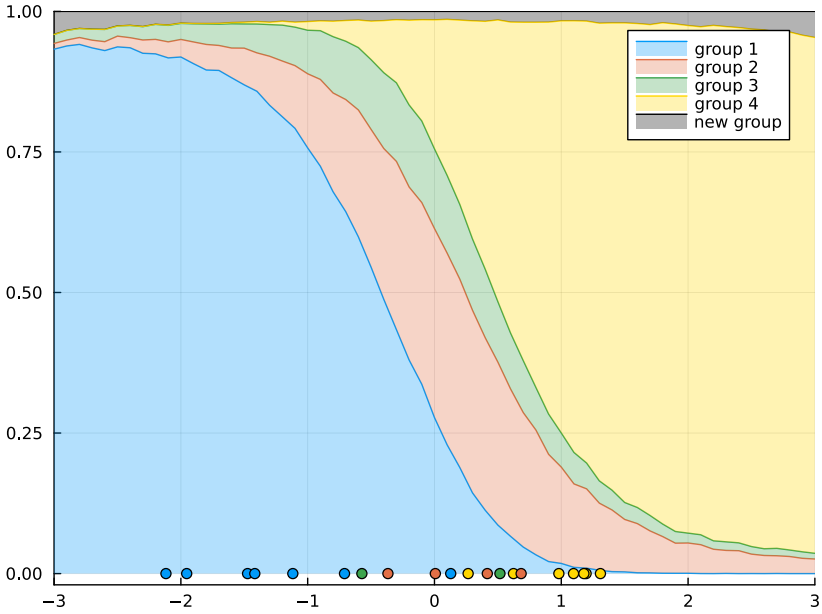

Figure 4.1: Predictive distributions for the allocation of an additional observation, as a function of its predictor, for a dataset of $n = 20$ observations partitioned in $k = 4$ groups; observed predictors and group labels (colors) are also reported on the horizontal axis.

the univariate continuous predictor $X_i \in \mathbb{R}$ and the allocation variable $A_i \in \{1, \ldots, k\}$, that is, the group label, are observed, while the distinct locations $\xi_1^*, \ldots, \xi_k^*$ and the auxiliary variables $\boldsymbol{V}$ are latent. The present illustration focuses on the predictive distribution for the allocation variable A_{n+1} of an additional observation, as a function of the corresponding predictor X_{n+1}, namely on the functions

$$x \mapsto \mathbb{P}(A_{n+1} \leq j \mid \boldsymbol{A}, \boldsymbol{X}, X_{n+1} = x), \qquad j = 1, \ldots, k+1,$$

where $A_{n+1} = k+1$ denotes the allocation of the additional observation to a new group. These probabilities can be computed as in (4.14), upon marginalization of the latent distinct locations and auxiliary variables, which is performed resorting to a marginal Gibbs sampling algorithm. For illustration purposes, consider a Gaussian kernel $k(x, \xi) = \exp\{-0.5\, d(x, \xi)^2\}$, and set the hyperparameters characterizing the commom random measure at values $\theta = 3.0$ and $\sigma = 0.5$, while P_0 is a Gaussian random variable with large variance, corresponding to a non-informative specification. Moreover, the number of observed groups is set to $k = 4$ and predictors are sampled from a mixture of Gaussian distributions. Results for a simulated dataset consisting of $n = 20$ observations are displayed in Figure 4.1, where the observed predictors, and corresponding group labels (colors) are also reported. The kernel-weighting approach has a considerable impact on the predictive probability distributions, which are highly influenced by the specific value of the predictor; for reference, the probabilities of allocation to the four observed groups for a standard

Pitman-Yor process are 0.295, 0.159, 0.068 and 0.295, respectively, while the probability of allocation to a new group is 0.183.

Chapter 5

Nonhomogeneous Pitman-Yor process

The kernel-weighted Pitman-Yor process introduced in Chapter 4 features a remarkable quasi-conjugacy property with respect to a novel class of nonhomogeneous random measures, the distributions of which are obtained as transformations of the distribution of a stable completely random measure. The random probability measures defined by normalizing such random measures are named nonhomogeneous Pitman-Yor processes, as they include the standard Pitman-Yor process as a special case. This chapter provides an extension of some well-established properties of the Pitman-Yor process as nonparametric prior for exchangeable observations; specifically, the prior-posterior updating mechanism of the proposed construction, as well as its predictive structure, are characterized, thanks to the introduction of an additional latent variable.

5.1 Construction of nonhomogeneous processes

Bayesian nonparametric mixture models for density estimation exploit discrete random probability measures as mixing measures; such nonparametric priors can be often defined via normalization of suitable random measures, as discussed in Section 1.2. Indeed, the renowned Dirichlet and Pitman-Yor processes may be both characterized as normalized homogeneous random measures (Section 1.3). More generally, random probability measures obtained by normalizing homogeneous random measures are considered in applications because of their analytical tractability; furthermore, the prior-posterior updating mechanism preserves their homogeneity (James et al., 2009). On the other hand, nonhomogeneous random measures are widely exploited in the Bayesian nonparametric modelling of survival data, while their normalization is discouraged by the intractability of most quantities of interest. This section introduces a novel nonhomogeneous random probability measure arising from the normalization of a nonhomogeneous random measure.

Consider a σ-stable completely random measure $\tilde{\mu}_\sigma$ on the complete and separable metric

space $\mathbb{X}$, characterized by the Lévy intensity measure

$$\nu(ds, dx) = \frac{\sigma}{\Gamma(1-\sigma)}\, s^{-1-\sigma}\, ds\, P_0(dx),$$

where $\sigma \in (0,1)$ and P_0 is a diffuse probability measure on $\mathbb{X}$. Along the lines of the discussion in Section 4.4, a random measure $\tilde{\mu}$ can be defined through a change of measure with respect to $\tilde{\mu}_\sigma$: specifically, the probability distribution of $\tilde{\mu}$ is absolutely continuous with respect to the random measure $\tilde{\mu}_\sigma$, with Radon-Nikodym derivative

$$\frac{d\mathcal{L}(\tilde{\mu})}{d\mathcal{L}(\tilde{\mu}_\sigma)}(m) = \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)} \left(\int_{\mathbb{X}} G(x)\, m(dx) \right)^{-\theta}, \tag{5.1}$$

where $\theta > 0$ and $G\colon \mathbb{X} \mapsto \mathbb{R}^+$ is a non-negative and bounded measurable function. Note that, in order for (5.1) to be a change of measure between probability distributions, the function $G(\cdot)$ should satisfy the condition

$$\mathbb{E}\left[\frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)} \left(\int_{\mathbb{X}} G(x)\tilde{\mu}_\sigma(dx) \right)^{-\theta} \right] = 1;$$

this amounts to the non-negative measure $Q_0(dx) = G(x)^\sigma P_0(dx)$ being a probability measure, or, equivalently,

$$\int_{\mathbb{X}} G(x)^\sigma P_0(dx) = 1. \tag{5.2}$$

The normalization of $\tilde{\mu}$ leads to a novel nonhomogeneous random probability measure.

Definition 5.1. The *nonhomogeneous Pitman-Yor process* is the random probability measure $\tilde{p}$ on $\mathbb{X}$ obtained from the normalization of the random measure $\tilde{\mu}$ defined in (5.1),

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

and denoted by $\tilde{p} \sim \mathrm{PY}(\sigma, \theta, P_0; G)$, where σ is the *discount* parameter, θ is the *concentration* parameter, P_0 is the base probability measure and $G(\cdot)$ is termed *nonhomogeneity* function.

The proposed terminology stems from the trivial observation that standard (homogeneous) Pitman-Yor process si retrieved by choosing the constant function $G(x) = 1$, as the expression for the change of measure in (5.1) boils down to (1.14). In fact, the nonhomogeneity of the process is introduced through the function G, as apparent from the expression of the change of measure in (5.1). More specifically, (5.1) can be regarded as a change of measure with respect to the distribution of the random measure

$$\tilde{\mu}_{\sigma,G}(A) := \int_A G(x)\, \tilde{\mu}_\sigma(dx), \qquad \forall\, A \in \mathcal{X}, \tag{5.3}$$

which is a nonhomogeneous completely random measure; indeed, its definition follows (1.8), and $\tilde{\mu}_{\sigma,G}$ can thus be termed *nonhomogeneous σ-stable* process. Remarkably, condition (5.2) on function G implies that the distribution of the total mass of $\tilde{\mu}_\sigma$ is preserved by the change of measure (5.3); indeed, its Laplace transform

$$\mathbb{E}\left[e^{-\lambda\tilde{\mu}_{\sigma,G}(\mathbb{X})}\right] = \mathbb{E}\left[\exp\left\{-\lambda\int_{\mathbb{X}} G(x)\,\tilde{\mu}_\sigma(dx)\right\}\right] = \exp\left\{-\lambda^\sigma\int_{\mathbb{X}} G(x)^\sigma\,P_0(dx)\right\} = e^{-\lambda^\sigma},$$

for $\lambda \geq 0$, coincides with the Laplace transform of a stable distribution with parameter σ. As a result, the construction of the random measure $\tilde{\mu}$ starting from a σ-stable process $\tilde{\mu}_\sigma$ may be split into a sequence of two steps: (i) define the nonhomogeneous σ-stable process $\tilde{\mu}_{\sigma,G}$ as in (5.3); (ii) apply the usual change of measure for the construction of the Pitman-Yor process, namely

$$\frac{d\mathcal{L}(\tilde{\mu})}{d\mathcal{L}(\tilde{\mu}_{\sigma,G})}(m) = \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)}\,m(\mathbb{X})^{-\theta}. \tag{1.14}$$

This sequential construction of the Pitman-Yor process suggests that a general procedure for the definition of nonhomogeneous nonparametric priors may consists in applying the same transformations, such as normalizations or changes of measure, considered for the homogeneous process to its nonhomogeneous version, obtained as in (1.8) and (5.3). Note that some conditions are typically required on the function introducing the nonhomogeneity: in particular, the distributional properties of the functionals of interest for the subsequent transformations should be preserved. A first example of such procedure may be the definition of a nonhomogeneous Dirichlet process, which is obtained via normalization of the extended gamma process (1.9). As a matter of fact, this process arises as a special case of the nonhomogeneous Pitman-Yor process in the limit for $\sigma \to 0$, just like the standard Dirichlet process is a special case of the standard Pitman-Yor process. The approach briefly outlined above may be of interest in Bayesian nonparametric methodological research, and deserves further investigations. Finally, a similar procedure may be applied to define a multivariate vector of dependent nonhomogeneous Pitman-Yor process, following the approach proposed in Zhu and Leisen (2015).

The rest of the chapter is devoted to the characterization of the main properties of this novel nonhomogeneous process when employed as nonparametric prior for a sequence of exchangeable observations; in particular, connections with well-known properties of the standard Pitman-Yor process are highlighted, as well as some similarities with the results described in Chapter 4 for the kernel-weighted Pitman-Yor process.

5.2 Marginal distribution and partition function

Consider a sequence of exchangeable observations $\boldsymbol{X} = (X_1, \ldots, X_n)$ taking values in the space $\mathbb{X}$, and assume the following Bayesian nonparametric model:

$$\begin{aligned} X_1, \ldots, X_n \mid \tilde{p} &\overset{\text{i.i.d.}}{\sim} \tilde{p}, \\ \tilde{p} &\sim \mathrm{PY}(\theta, \sigma, P_0; G). \end{aligned} \tag{5.4}$$

The almost-sure discreteness of the random measure $\tilde{p}$ implies that the sequence of observations displays ties with positive probability; specifically, denote by $X_1^*, \ldots, X_k^*$ the k distinct values taken by the observations $X_1, \ldots, X_n$, with multiplicities $n_1 + \cdots + n_k = n$. The likelihood function associated to the model in (5.4) is expressed as

$$\mathcal{L}(\tilde{\mu}; \boldsymbol{X}) = \prod_{i=1}^{n} \tilde{p}(dX_i) = \tilde{\mu}(\mathbb{X})^{-n} \prod_{i=1}^{n} \tilde{\mu}(dX_i) = \tilde{\mu}(\mathbb{X})^{-n} \prod_{j=1}^{k} \tilde{\mu}(dX_j^*)^{n_j}.$$

The marginal distribution of the observations is obtained by marginalizing the likelihood function above with respect to the distribution of the nonhomogeneous Pitman-Yor process; for this purpose, define the functions $G_0(x) = G(x) - 1$ and, for $v \in [0, 1]$,

$$H(v) := \left(\int_{\mathbb{X}} (1 + v\, G_0(x))^{\sigma} P_0(dx) \right)^{1/\sigma}. \tag{5.5}$$

Proposition 5.2. *The marginal distribution of the observations $\boldsymbol{X}$ from a nonhomogeneous Pitman-Yor process is*

$$\begin{aligned} \mathbb{P}(\boldsymbol{X}) = \frac{\prod_{\ell=1}^{k-1}(\theta + \ell\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1}\, P_0(dX_j^*) \\ \times \int_0^1 H(v)^{-(\theta+n)} \prod_{j=1}^{k} \left(\frac{1 + v\, G_0(X_j^*)}{H(v)} \right)^{\sigma - n_j} \frac{\Gamma(n+\theta)}{\Gamma(n)\,\Gamma(\theta)}\, v^{\theta-1}\, (1-v)^{n-1}\, dv. \end{aligned}$$

Similarly to the marginal distribution in Proposition 4.1, the expression above factorizes into the product of the marginal distribution induced by the homogeneous Pitman-Yor process,

$$\frac{\prod_{\ell=1}^{k-1}(\theta + \ell\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1}\, P_0(dX_j^*),$$

and a term involving evaluations of the nonhomogeneity function G (specifically, G_0) at the observed distinct values $X_1^*, \ldots, X_k^*$; more precisely, this latter term can be regarded as an expectation with respect to the Beta distribution with parameters θ and n, henceforth denoted

by $f_{\theta,n}(v)$, of the function in product form

$$v \in [0,1] \;\mapsto\; H(v)^{-(\theta+n)} \prod_{j=1}^{k} \left(\frac{1 + v\, G_0(X_j^*)}{H(v)} \right)^{\sigma - n_j}$$

Remarkably, the function above equals 1 for every $v \in [0,1]$ in case $G(x) = 1$, that is $G_0(x) = 0$, for every $x \in \mathbb{X}$, so that the marginal distribution for the homogeneous Pitman-Yor process is correctly retrieved.

As anticipated in Section 5.1, a nonhomogeneous version of the Dirichlet process may be defined by normalizing an extended gamma process. The following result establishes a connection between normalized extended gamma processes, as considered in James et al. (2009), and the proposed nonhomogeneous Pitman-Yor process: specifically, the marginal distribution for a normalized extended gamma process can be retrieved from the expression of the marginal distribution in Proposition 5.2 in the limit for $\sigma \to 0$.

Corollary 5.3. *In the limit for $\sigma \to 0$, the marginal distribution of the observations $\boldsymbol{X}$ is*

$$\begin{aligned}\mathbb{P}(\boldsymbol{X}) = \frac{\Gamma(\theta)}{\Gamma(n+\theta)} \prod_{j=1}^{k} \Gamma(n_j)\, \theta P_0(dX_j^*) \\ \times \int_0^1 \exp\left\{ - \int_{\mathbb{X}} \log(1 + v\, G_0(x))\, \theta P_0(dx) \right\} \prod_{j=1}^{k} (1 + v\, G_0(X_j^*))^{-n_j}\, f_{\theta,n}(v)\, dv,\end{aligned}$$

which coincides with the marginal distribution induced by a normalized extended gamma CRM, characterized by the Lévy intensity

$$\nu(ds, dx) = \theta\, s^{-1} e^{-G(x)\, s}\, ds\, P_0(dx).$$

As expected, the expression above factorizes into the product of the marginal distribution induced by the Dirichlet process, and a term involving evaluations of the function G_0 at the observations, which is again an expectation with respect to the Beta distribution with parameters θ and n. Moreover, the marginal distribution in Corollary 5.3 can be rewritten, thanks to the change of variable $v \mapsto u = (1 - v)/v$, as

$$\frac{1}{\Gamma(n)} \int_0^{\infty} u^{n-1} \exp\left\{ - \int_{\mathbb{X}} \log(u + G(x))\, \theta P_0(dx) \right\} \prod_{j=1}^{k} \frac{\Gamma(n_j)}{(u + G(X_j^*))^{n_j}}\, \theta P_0(dX_j^*)\, du;$$

this is precisely the form proposed by James et al. (2009) for the marginal distribution of a normalized extended gamma CRM.

A fundamental aspect of Bayesian nonparametric priors is the characterization of the partition structures they induce on the observations, possibly at the latent level; to this end, denote by $\boldsymbol{\Pi}$

the random partition of the observations induced by the nonparametric prior. For each $n \geq 1$, the distribution of $\mathbf{\Pi}$, that is, the probability distribution on the space of partitions of n observations, is known as *exchangeable partition probability function* (EPPF), and can be obtained from the marginal distribution of the observations upon marginalization of their distinct values; moreover, as a consequence of the exchangeability assumption on the sequence of observations, its expression depends only on the number of partition groups k and their multiplicities $n_1 + \cdots + n_k = n$.

Proposition 5.4. *The exchangeable partition probability function for the partition* $\mathbf{\Pi}$ *induced by ties in the sequence of observations* $\boldsymbol{X}$ *from a nonhomogeneous Pitman-Yor process is*

$$\text{EPPF}(n_1, \ldots, n_k) = \frac{\prod_{\ell=1}^{k-1}(\theta + \ell\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j-1} \int_0^1 H(v)^{-\theta} \prod_{j=1}^{k} S(v; n_j)\, f_{\theta,n}(v)\, dv,$$

where the following function is defined, for each $v \in [0,1]$*:*

$$S(v; m) := \frac{\displaystyle\int_{\mathbb{X}} (1 + v\, G_0(x))^{\sigma-m} P_0(dx)}{\displaystyle\int_{\mathbb{X}} (1 + v\, G_0(x))^{\sigma} P_0(dx)}.$$

An interesting property of the EPPF introduced above is the possibility to regard its expression as a mixture, over the Beta distribution, of exchangeable partition functions in product form

$$\text{EPPF}(n_1, \ldots, n_k) = \int_0^1 \frac{\prod_{\ell=1}^{k-1}(\theta + \ell\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j-1}\, S(v; n_j)\, H(v)^{-\theta}\, f_{\theta,n}(v)\, dv.$$

Furthermore, the structure of such partition functions resembles the typical form of Gibbs-type priors (Gnedin and Pitman, 2006), to which a weighting function S is applied; similar structures also appear in Proposition 4.2 for the kernel-weighted Pitman-Yor process, and may be of interest by themselves.

Besides altering the distribution of the induced random partition, with respect to its homogeneous counterpart, another remarkable consequence of the nonhomogeneity function is the modification of the marginal probability distribution for a single observation; specifically, from Proposition 5.2,

$$\mathbb{P}\left(X_i \in dx\right) = \int_0^1 H(v)^{-(\theta+1)} \left(\frac{1 + v\, G_0(x)}{H(v)}\right)^{\sigma-1} f_{\theta,1}(v)\, dv\, P_0(dx).$$

This probability distribution is absolutely continuous with respect to the base probability measure P_0, which in turn coincides with the marginal distribution of observations from the corresponding

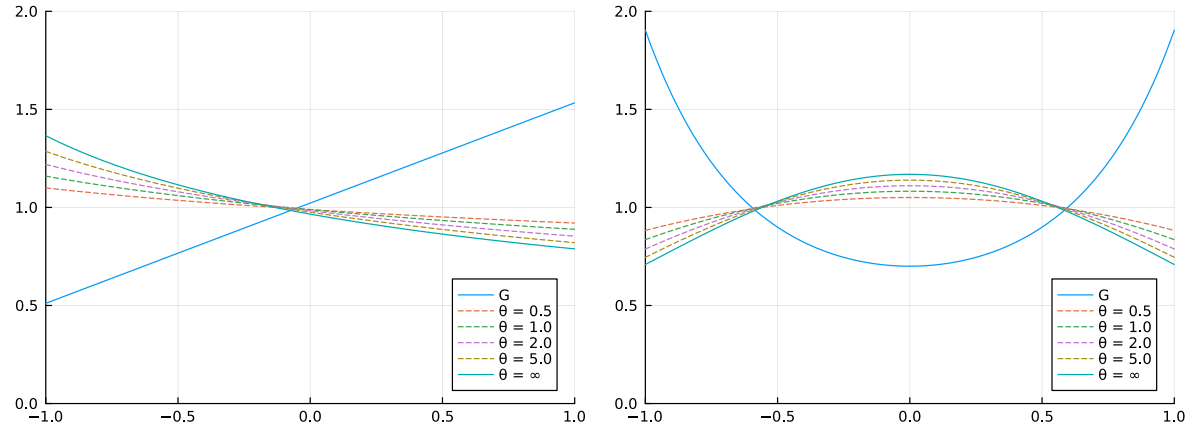

Figure 5.1: Radon-Nikodym derivative functions in (5.6) for the two specifications of the nonhomogeneity function G in (5.7), and different values of the parameter θ; the parameter $\sigma = 0.5$ is fixed.

homogeneous Pitman-Yor process: the Radon-Nikodym derivative

$$\frac{\mathbb{P}\left(X_i \in dx\right)}{P_0(dx)} = \int_0^1 H(v)^{-(\theta+1)} \left(\frac{1 + v\, G_0(x)}{H(v)}\right)^{\sigma-1} f_{\theta,1}(v)\, dv \tag{5.6}$$

represents a useful quantity to investigate the role of the nonhomogeneity function in the proposed model. In particular, its maximum coincides with the minimum of function G, and viceversa, that is, it is more likely for an observation to be in a region in which G takes lower values, compared to the homogeneous Pitman-Yor process; moreover, the impact on the marginal distribution increases with θ.

The possibility to compute the values of the Radon-Nikodym derivative in closed form is unfortunately precluded by the intractable form of the integrand function, and specifications of the function G leading to simple examples are not available at the moment; however, the integral in (5.6) can be evaluated numerically for arbitrary specifications of the function G. For simplicity, let the base probability measure P_0 be the uniform distribution over $[-1, 1]$, and consider

$$G_1(x) = \kappa_1(2 + x), \qquad G_2(x) = \kappa_2 \exp(x^2), \tag{5.7}$$

where $\kappa_i > 0$ is the appropriate constant such that,

$$\int_{\mathbb{X}} G_i(x)^\sigma P_0(dx) = \frac{1}{2} \int_{-1}^1 G_i(x)^\sigma dx = 1.$$

Numerical evaluations of the integral in (5.6) for such choices of function G and different values of the concentration parameter θ are displayed in Figure 5.1, confirming the analytical evidence: indeed, observations are more likely to be in regions corresponding to lower values of function G, and the effect increases with θ.

5.3 Latent variable and posterior distribution

A standard approach to the analysis of statistical models in presence of mixture terms is regarding the mixing parameter as an auxiliary latent variable. As for the marginal distribution in Proposition 5.2, the integration in v can be considered a marginalization with respect to a latent variable V_n, supported on the unit interval, so that the augmented marginal distribution is

$$\mathbb{P}(\boldsymbol{X}, V_n \in dv) = \frac{\prod_{\ell=1}^{k-1}(\theta+\ell\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k}(1-\sigma)_{n_j-1}\, P_0(dX_j^*) \\ \times H(v)^{-(\theta+n)} \prod_{j=1}^{k}\left(\frac{1+v\,G_0(X_j^*)}{H(v)}\right)^{\sigma-n_j} f_{\theta,n}(v)\, dv.$$

In this setting, the latent variable V_n admits the interpretation as relative weight given to the nonhomogeneity function G within the learning mechanism; indeed, function G enters the marginal distribution through the quantity $(1+V_nG_0(x))$. As a further insight on this interpretation, note that, for every $x \in \mathbb{X}$,

$$\lim_{V_n\to 1}(1+V_nG_0(x)) = G(x), \qquad \lim_{V_n\to 0}(1+V_nG_0(x)) = 1,$$

with the former situation entailing maximal nonhomogeneity effect, and the latter corresponding to the homogeneous case.

The augmentation introduced above is fundamental for the characterization of the posterior distribution of the nonhomogeneous Pitman-Yor process; a similar circumstance appears in Section 4.4 for the posterior characterization of the kernel-weighted Pitman-Yor process, and in the posterior analysis of normalized CRMs in James et al. (2009). For this purpose, consider the latent variable V_n on the unit interval $[0,1]$ having density function, given the observations $\boldsymbol{X}$, proportional to

$$f(v;\boldsymbol{X}) \propto H(v)^{-(\theta+n)} \prod_{j=1}^{k}\left(\frac{1+v\,G_0(X_j^*)}{H(v)}\right)^{\sigma-n_j} f_{\theta,n}(v); \tag{5.8}$$

where $f_{\theta,n}(v)$ is the density function of the Beta distribution with parameters θ and n. Moreover, let T be an auxiliary random variable which, given the number k of observed distinct values, has generalized gamma distribution with density function

$$f_T(t \mid k) = \frac{\sigma}{\Gamma(k+\theta/\sigma)}\, t^{k\sigma+\theta-1}\, e^{-t^\sigma}.$$

Proposition 5.5. *The posterior distribution of the random measure $\tilde{\mu}$, given the observations*

$\boldsymbol{X}$ and the latent variables V_n and T, coincides with the distribution of the random measure

$$\tilde{\mu}(dx) \mid \boldsymbol{X}, V_n, T \;\sim\; \tilde{\mu}^*(dx) + \sum_{j=1}^{k} W_j\, \delta_{X_j^*}(dx),$$

where the random elements in the sum are mutually independent, $\tilde{\mu}^*$ is a generalized gamma completely random measure, with Lévy intensity measure

$$\nu^*(ds, dx) = \frac{\sigma}{\Gamma(1-\sigma)}\, s^{-1-\sigma}\, \exp\left\{-\frac{1+V_n G_0(x)}{H(V_n)}\, T\, s\right\} ds\, P_0(dx),$$

and each W_j is a non-negative random variables having gamma distribution

$$W_j \;\sim\; \text{Gamma}\left(n_j - \sigma,\, \frac{1+V_n G_0(X_j^*)}{H(V_n)}\, T\right).$$

Therefore, the posterior distribution of $\tilde{\mu}$ is a mixture of completely random measures, with latent variables V_n and T playing the role of mixing parameters; moreover, the nonhomogeneity a priori is preserved a posteriori, as highlighted by the expression of the Lévy intensity characterizing $\tilde{\mu}^*$, as well as the dependence of jumps $W_1, \ldots, W_k$ from their locations $X_1^*, \ldots, X_k^*$.

A fundamental characterizing property of the standard Pitman-Yor process within the class of Gibbs-type priors is its quasi-conjugacy (Lijoi et al., 2008); see also Section 1.3 and Section 4.4 for further details and references on this concept. Remarkably, a form of conditional quasi-conjugacy is retained by the nonhomogeneous version of the Pitman-Yor process introduced in this chapter.

Proposition 5.6. *The distribution of the random measure $\tilde{\mu}^*$ in Proposition 5.5, given the number k of observed distinct values and the latent variable V_n, is absolutely continuous with respect to the distribution of a σ-stable completely random measure $\tilde{\mu}_\sigma$, with Radon-Nikodym derivative*

$$\frac{d\mathcal{L}(\tilde{\mu}^*)}{d\mathcal{L}(\tilde{\mu}_\sigma)}(m) = \frac{\sigma\,\Gamma(\theta+k\sigma)}{\Gamma(\theta/\sigma+k)} \left(\int_{\mathbb{X}} \frac{1+V_n G_0(x)}{H(V_n)}\, m(dx)\right)^{-(\theta+k\sigma)}$$

The result presented above can be equivalently formulated, in terms of quasi-conjugacy, as follows: *The nonhomogeneous Pitman-Yor process is conditionally quasi-conjugate, given the latent variable V_n.* In particular, the process generating new distinct values is the nonhomogeneous Pitman-Yor process

$$\tilde{p}^*(dx) = \frac{\tilde{\mu}^*(dx)}{\tilde{\mu}^*(\mathbb{X})} \;\sim\; \text{PY}(\sigma, \theta^*, P_0; G^*(\cdot\,; V_n)),$$

where the base probability measure P_0 and discount parameters σ coincide with the prior process, while the concentration parameter and nonhomogeneity function are updated as

$$\theta^* = \theta + k\sigma, \qquad G^*(x; V_n) = \frac{1+V_n\, G_0(x)}{H(V_n)}.$$

5.4 Predictive distribution

A further appealing property of the Pitman-Yor process is represented by its simple one-step predictive distribution:

$$\mathbb{P}(X_{n+1} \in dx \mid \boldsymbol{X}) = \frac{\theta + k\sigma}{\theta + n} P_0(dx) + \sum_{j=1}^{k} \frac{n_j - \sigma}{\theta + n} \delta_{X_j^*}(dx). \tag{4.15}$$

Similarly to the other marginal properties described in Section 5.2, such predictive structure is equally made more involved by the introduction of nonhomogeneity in the prior process. In order to described it, consider a sequence of m additional observations $\boldsymbol{X}^+ = (X_{n+1}, \ldots, X_{n+m})$ taking values in X, and assume they take h additional distinct values, besides possibly taking some of the previous distinct values; specifically, the previous values are taken with multiplicities $m_1, \ldots, m_k \geq 0$, while the additional values, denoted by $X_{k+1}^*, \ldots, X_{k+h}^*$, are taken with multiplicities $m_{k+1}, \ldots, m_{k+h} \geq 1$, such that $m_1 + \cdots + m_{k+h} = m$.

Proposition 5.7. *The predictive distribution for the new observations $\boldsymbol{X}^+$, given the previous observations $\boldsymbol{X}$ and the latent variable V_n, is*

$$\begin{aligned}\mathbb{P}(\boldsymbol{X}^+ \mid \boldsymbol{X}, V_n = v) = {} & \frac{\prod_{\ell=0}^{h-1}(\theta + k\sigma + \ell\sigma)}{(\theta + n)_m} \prod_{j=1}^{k} (n_j - \sigma)_{m_j} \prod_{j=k+1}^{k+h} (1-\sigma)_{m_j - 1} P_0(dX_j^*) \\ & \times H(v)^{\theta+n} \prod_{j=1}^{k} \left(\frac{1 + v\, G_0(X_j^*)}{H(v)} \right)^{n_j - \sigma} \int_0^1 H(wv)^{-(\theta+n+m)} \prod_{j=1}^{k} \left(\frac{1 + wv\, G_0(X_j^*)}{H(wv)} \right)^{\sigma - n_j - m_j} \\ & \qquad \times \prod_{j=k+1}^{k+h} \left(\frac{1 + wv\, G_0(X_j^*)}{H(wv)} \right)^{\sigma - m_j} f_{n+\theta, m}(w)\, dw.\end{aligned}$$

In the expression above, the m-step predictive distribution for the homogeneous Pitman-Yor process, that is

$$\frac{\prod_{\ell=0}^{h-1}(\theta + k\sigma + \ell\sigma)}{(\theta + n)_m} \prod_{j=1}^{k} (n_j - \sigma)_{m_j} \prod_{j=k+1}^{k+h} (1-\sigma)_{m_j - 1} P_0(dX_j^*),$$

can be recognized as a factor in the first line; this result closely parallels the predictive distribution for the kernel-weighted Pitman-Yor process in Proposition 4.6.

More interesting insights on the predictive structure can be gathered by inspecting the

one-step predictive distribution for an additional observation X_{n+1}, which is proportional to

$$\begin{aligned}
&\mathbb{P}(X_{n+1} \in dx \mid \boldsymbol{X}, V_n = v) \\
&\quad \propto \int_0^1 \left(\frac{\theta + k\sigma}{n+\theta} \left(\frac{H(wv)}{1 + wv\, G_0(x)} \right)^{1-\sigma} P_0(dx) + \sum_{j=1}^{k} \frac{n_j - \sigma}{n+\theta} \frac{H(wv)}{1 + wv\, G_0(X_j^*)} \delta_{X_j^*}(dx) \right) \\
&\qquad\qquad \times H(wv)^{-(\theta+n+1)} \prod_{j=1}^{k} \left(\frac{1 + wv\, G_0(X_j^*)}{H(wv)} \right)^{\sigma - n_j} f_{\theta+n,1}(w)\, dw.
\end{aligned}$$

In view of the construction of sampling algorithms based on urn schemes, it is convenient to regard the integration with respect to w as a marginalization with respect to a further auxiliary latent variables W_{n+1} on the unit interval, and consider the sequential decomposition of the augmented predictive distribution

$$\begin{aligned}
&\mathbb{P}(X_{n+1} \in dx, W_{n+1} \in dw \mid \boldsymbol{X}, V_n = v) \\
&\qquad\qquad = \mathbb{P}(W_{n+1} \in dw \mid \boldsymbol{X}, V_n = v) \times \mathbb{P}(X_{n+1} \in dx \mid \boldsymbol{X}, V_n = v, W_{n+1} = w)
\end{aligned}$$

In particular, the predictive distribution for the new observation X_{n+1}, given the additional latent W_{n+1}, is proportional to

$$\begin{aligned}
&\mathbb{P}(X_{n+1} \in dx \mid \boldsymbol{X}, V_n = v, W_{n+1} = w) \\
&\quad \propto \frac{\theta + k\sigma}{\theta + n} \frac{H(wv)}{1 + wv\, G_0(x)} \left(\frac{1 + wv\, G_0(x)}{H(wv)} \right)^{\sigma} P_0(dx) + \sum_{j=1}^{k} \frac{n_j - \sigma}{\theta + n} \frac{H(wv)}{1 + wv\, G_0(X_j^*)} \delta_{X_j^*}(dx),
\end{aligned}$$

which indeed represents a generalization of the predictive urn scheme (4.15) for the standard Pitman-Yor process.

5.5 Future developments

This chapter introduces a novel nonparametric prior process for Bayesian inference on exchangeable observations, which represents a nonhomogeneous version of the standard Pitman-Yor process (Perman et al., 1992; Pitman and Yor, 1997); see Section 1.3 for further references. The nonhomogeneity is modeled through a non-negative function G, satisfying the integrability condition (5.2), with the constant choice $G(x) = 1$ corresponding to the homogeneous Pitman-Yor process. The rest of the chapter proposes a preliminary overview on the main properties of such prior. Specifically, the marginal distribution and exchangeable partition probability function induced by the nonhomogeneous Pitman-Yor process are characterized in Section 5.2, where some insights on the role played by different choices of function G are explored. Moreover, its posterior distribution is explicitly characterized in Section 5.3, and a conditional quasi-conjugacy

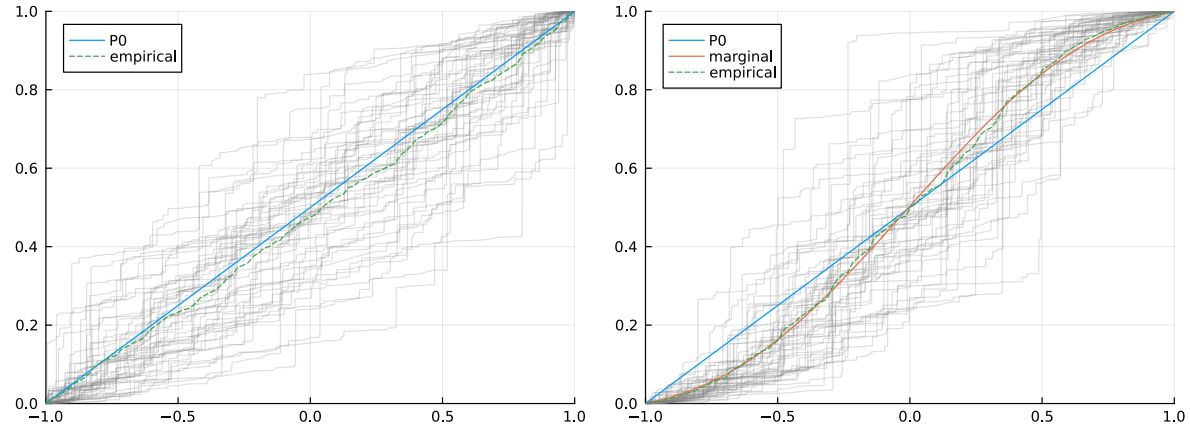

Figure 5.2: Trajectories from the homogeneous (left) and nonhomogeneous Pitman-Yor process (right), with $G(x) \propto \exp(8|x|^{1.5})$; the discount and concentration parameters are set to $\sigma = 0.5$ and $\theta = 5.0$, while the base measure is uniform on $[-1, 1]$.

property is shown to hold, given the additional latent variable V_n, supported on the unit interval. Finally, the predictive distribution is derived in Section 5.4, where a possible marginal sampling algorithm inspired by the classical urn schemes is sketched.

Nevertheless, some aspects require further investigations in order to provide a comprehensive understanding of the proposed process. A first glance at the impact of the nonhomogeneity function specification on the marginal distribution of each observation is provided in Figure 5.1. However, the probability of co-clustering for two or more observations is also modified by the function G; therefore, a deeper understanding of its implications should be sought. In this respect, the lack of simple and analytically tractable examples is currently a major drawback of this formulation, as most quantities of interest should rely on numerical integration.

As for the algorithmic developments, a first challenge is represented by the sampling of the latent variable V_n from the intractable density (5.8); drawing inspiration from the result in Lemma 4.8, a useful approximation for the distribution of the rescaled variable $n\,V_n$ can be obtained for $n \to \infty$, but its theoretical validity should be further assessed. Likewise, a viable algorithmic procedure for sampling the additional latent variable W_{n+1} should be devised and implemented. Moreover, theoretical investigations regarding the asymptotic properties of the nonhomogeneous Pitman-Yor process, such as the asymptotic distribution of the number of clusters, both *a priori* and conditionally on the first n observations, should be considered; in particular, the techniques developed in the inspiring book by Pitman (2006) for the standard Pitman-Yor process should be applicable directly to its nonhomogeneous version.

In conclusion, a preliminary comparison between the homogeneous and nonhomogeneous Pitman-Yor processes is proposed, in order to further highlight some of the main differences discussed in this chapter. In the following, let the base probability measure P_0 be the uniform

distribution over $[-1, 1]$, and consider the nonhomogeneity function

$$G(x) = \kappa_\sigma \, e^{8|x|^{1.5}},$$

where $\kappa_\sigma > 0$ is the appropriate normalizing constant such that condition (5.2) is satisfied; moreover, fix the discount and concentration parameters at $\sigma = 0.5$ and $\theta = 5.0$, respectively. Figure 5.2 compares the trajectories of the homogeneous Pitman-Yor process with the trajectories of the nonhomogeneous process, having the nonhomogeneity function G defined above. As extensively discussed in Section 5.2, the nonhomogeneity function impacts the marginal distribution of the process: in particular, observations are more likely to occur in regions where G takes lower values. This modification of the marginal distribution is a direct consequence of the nonhomogeneity of the process, that is, of the dependence between the locations and jumps of the random probability measure. Indeed, the sequence of random locations are sampled from the base measure P_0 for both the homogeneous and nonhomogeneous Pitman-Yor processes, as displayed in Figure 5.3 (top). On the other hand, conditionally on the jumps size, the random jumps of the nonhomogeneous process are not uniformly distributed over $[-1, 1]$, with largest jumps located in regions where G takes lower values, and thus altering the marginal distribution accordingly. Such phenomenon can be observed in Figure 5.3 (bottom), where the empirical distribution functions of the locations corresponding to the largest jumps are reported in both scenarios: indeed, the overlapping with the marginal distribution is evident.

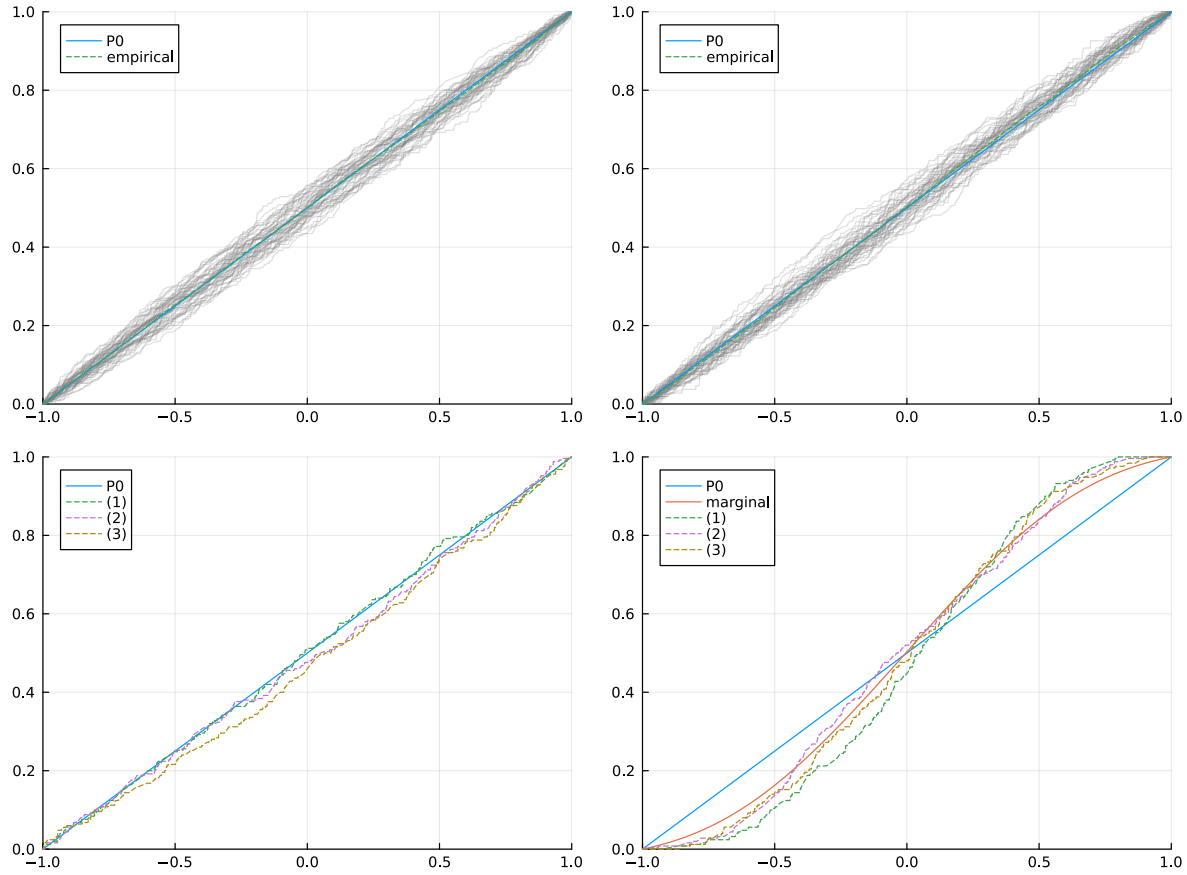

Figure 5.3: (top) empirical distribution functions of sequences of random locations (distinct values) sampled from the homogeneous (left) and nonhomogeneous (right) Pitman-Yor processes; (bottom) empirical distribution functions of the locations of the largest 3 jumps over 100 trajectories of the homogeneous (left) and nonhomogeneous (right) Pitman-Yor processes.

Appendix A

Proofs

A.1 Proofs of Chapter 2

Proof of the identity (2.9) This proof is based on the characterization of completely random measures through their Laplace functional transform, and does not require any further knowledge on calculus with Poisson point processes; however, additional assumptions on the function f (namely, continuity) are needed in order to conclude the argument. An alternative proof which does not require continuity can be derived using the tools described in James (2005).

Consider $\varepsilon > 0$ taking a sufficiently small value, such that the ε-balls $B_\varepsilon(x_1^*), \ldots, B_\varepsilon(x_k^*)$ are pairwise disjoint, and define the complementary subset $\mathbb{X}^* = \mathbb{X} \setminus \left(\cup_{j=1}^k B_\varepsilon(x_j^*)\right)$. By the independence of the evaluations of a completely random measure on pairwise disjoint sets,

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} f(x)\tilde{\mu}(dx)\right\}\prod_{j=1}^{k}\tilde{\mu}(B_\varepsilon(x_j^*))^{n_j}\right]$$
$$= \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}^*} f(x)\tilde{\mu}(dx)\right\}\right]\prod_{j=1}^{k}\mathbb{E}\left[\exp\left\{-\int_{B_\varepsilon(x_j^*)} f(x)\tilde{\mu}(dx)\right\}\tilde{\mu}(B_\varepsilon(x_j^*))^{n_j}\right].$$

The first factor coincides with the Laplace functional transform of the random measure $\tilde{\mu}$ computed at function $f(x)(x \in \mathbb{X}^*)$, therefore

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}^*} f(x)\tilde{\mu}(dx)\right\}\right] = \exp\left\{-\int_{\mathbb{X}^*}\left(1-e^{-f(x)\,s}\right)\rho(ds)\,\alpha(dx)\right\}$$
$$= \exp\left\{-\int_{\mathbb{X}^*}\psi(f(x))\,\alpha(dx)\right\};$$

in the limit for $\varepsilon \to 0$, by monotone convergence theorem,

$$\lim_{\varepsilon\to 0} \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}^*} f(x)\tilde{\mu}(dx)\right\}\right] = \lim_{\varepsilon\to 0} \exp\left\{-\int_{\mathbb{X}^*} \psi(f(x))\,\alpha(dx)\right\} = \exp\left\{-\int_{\mathbb{X}} \psi(f(x))\,\alpha(dx)\right\}.$$

On the other hand, for each $j = 1, \ldots, k$, by monotone convergence theorem,

$$\mathbb{E}\left[\exp\left\{-\int_{B_\varepsilon(x_j^*)} f(x)\tilde{\mu}(dx)\right\}\tilde{\mu}(B_\varepsilon(x_j^*))^{n_j}\right] = \lim_{u\to 0} \mathbb{E}\left[\exp\left\{-\int_{B_\varepsilon(x_j^*)} (f(x)+u)\tilde{\mu}(dx)\right\}\tilde{\mu}(B_\varepsilon(x_j^*))^{n_j}\right] = \lim_{u\to 0} \mathbb{E}\left[\exp\left\{-\int_{B_\varepsilon(x_j^*)} f(x)\tilde{\mu}(dx) - u\,\tilde{\mu}(B_\varepsilon(x_j^*))\right\}\tilde{\mu}(B_\varepsilon(x_j^*))^{n_j}\right].$$

Given that, for $n_j \geq 0$,

$$\frac{\mathrm{d}^{n_j}}{\mathrm{d}u^{n_j}} \exp\left\{-u\,\tilde{\mu}(B_\varepsilon(x_j^*))\right\} = (-1)^{n_j}\,\tilde{\mu}(B_\varepsilon(x_j^*))^{n_j} \exp\left\{-u\,\tilde{\mu}(B_\varepsilon(x_j^*))\right\},$$

the expression above becomes

$$= (-1)^{n_j} \lim_{u\to 0} \mathbb{E}\left[\frac{\mathrm{d}^{n_j}}{\mathrm{d}u^{n_j}} \exp\left\{-\int_{B_\varepsilon(x_j^*)} f(x)\tilde{\mu}(dx) - u\,\tilde{\mu}(B_\varepsilon(x_j^*))\right\}\right] = (-1)^{n_j} \lim_{u\to 0} \frac{\mathrm{d}^{n_j}}{\mathrm{d}u^{n_j}} \mathbb{E}\left[\exp\left\{-\int_{B_\varepsilon(x_j^*)} (f(x)+u)\tilde{\mu}(dx)\right\}\right].$$

Again, the expectation coincides with the Laplace functional transform of the random measure $\tilde{\mu}$ computed at function $(f(x)+u)(x \in B_\varepsilon(x_j^*))$, hence

$$= (-1)^{n_j} \lim_{u\to 0} \frac{\mathrm{d}^{n_j}}{\mathrm{d}u^{n_j}} \exp\left\{-\int_{B_\varepsilon(x_j^*)} \psi(f(x)+u)\,\alpha(dx)\right\}.$$

Since α is assumed to be finite and diffuse, the measure of the ε-ball $B_\varepsilon(x_j^*)$ vanishes for $\varepsilon \to 0$, and the expression above is rewritten as

$$= (-1)^{n_j+1} \lim_{u\to 0} \int_{B_\varepsilon(x_j^*)} \frac{\mathrm{d}^{n_j}}{\mathrm{d}u^{n_j}} \psi(f(x)+u)\,\alpha(dx) \exp\left\{-\int_{B_\varepsilon(x_j^*)} \psi(f(x)+u)\,\alpha(dx)\right\} + o\left(\alpha(B_\varepsilon(x_j^*))\right).$$

By definition of Laplace exponent of $\tilde{\mu}$, its derivatives coincide with the cumulants of $\tilde{\mu}$, that is

$$\begin{aligned}\frac{\mathrm{d}^m}{\mathrm{d}u^m}\,\psi(u) &= \frac{\mathrm{d}^m}{\mathrm{d}u^m}\int_{\mathbb{R}^+}\left(1-e^{-us}\right)\rho(ds) = \int_{\mathbb{R}^+}\frac{\mathrm{d}^m}{\mathrm{d}u^m}\left(1-e^{-us}\right)\rho(ds)\\ &= (-1)^{m+1}\int_{\mathbb{R}^+} s^m e^{-us}\,\rho(ds) = (-1)^{m+1}\,\tau(m;u);\end{aligned}$$

therefore, the quantity above becomes

$$\begin{aligned}&= \lim_{u\to 0}\int_{B_\varepsilon(x_j^*)}\tau(n_j;f(x)+u)\,\alpha(dx)\,\exp\left\{-\int_{B_\varepsilon(x_j^*)}\psi(f(x)+u)\,\alpha(dx)\right\} + o\left(\alpha(B_\varepsilon(x_j^*))\right)\\ &= \int_{B_\varepsilon(x_j^*)}\tau(n_j;f(x))\,\alpha(dx)\,\exp\left\{-\int_{B_\varepsilon(x_j^*)}\psi(f(x))\,\alpha(dx)\right\} + o\left(\alpha(B_\varepsilon(x_j^*))\right).\end{aligned}$$

In conclusion, taking the limit for $\varepsilon\to 0$, by monotone convergence theorem,

$$\begin{aligned}&\lim_{\varepsilon\to 0}\frac{1}{\alpha(B_\varepsilon(x_j^*))}\,\mathbb{E}\left[\exp\left\{-\int_{B_\varepsilon(x_j^*)}f(x)\tilde{\mu}(dx)\right\}\tilde{\mu}(B_\varepsilon(x_j^*))^{n_j}\right]\\ &\qquad= \lim_{\varepsilon\to 0}\frac{1}{\alpha(B_\varepsilon(x_j^*))}\int_{B_\varepsilon(x_j^*)}\tau(n_j;f(x))\,\alpha(dx)\,\exp\left\{-\int_{B_\varepsilon(x_j^*)}\psi(f(x))\,\alpha(dx)\right\}\\ &\qquad= \lim_{\varepsilon\to 0}\frac{1}{\alpha(B_\varepsilon(x_j^*))}\int_{B_\varepsilon(x_j^*)}\tau(n_j;f(x))\,\alpha(dx) = \tau(n_j;f(x_j^*));\end{aligned}$$

the last equality holds by a continuity argument if the non-negative measurable function f is further assumed to be continuous.

A.2 Proofs of Chapter 3

The proofs presented in this section rely on the key identity for completely random measures (2.10), introduced and discussed in Section 2.4 and briefly restated hereunder. Consider a homogeneous completely random measure $\tilde{\mu}$ with Lévy intensity measure $\nu(ds,dx)=\rho(ds)\,\alpha(dx)$, where α is a finite diffuse measure on $\mathbb{X}$; for any non-negative measurable function $f\colon\mathbb{X}\mapsto\mathbb{R}^+$, and distinct values $x_1^*,\ldots,x_k^*\in\mathbb{X}$, with multiplicities $n_1,\ldots,n_k\geq 1$,

$$\begin{aligned}&\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}}f(x)\,\tilde{\mu}(dx)\right\}\prod_{j=1}^{k}\tilde{\mu}(dx_j^*)^{n_j}\right]\\ &\qquad\qquad= \exp\left\{-\int_{\mathbb{X}}\psi(f(x))\,\alpha(dx)\right\}\prod_{j=1}^{k}\tau(n_j;f(x_j^*))\,\alpha(dx_j^*),\end{aligned}$$

where ψ and τ are the Laplace exponent and cumulants of random measure $\tilde{\mu}$, as defined in (2.7).

Proof of Proposition 3.1 Consider the augmented likelihood function in (3.7),

$$\mathcal{L}(\tilde{\mu}_1^e, \ldots, \tilde{\mu}_D^e; \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}) = Q(\boldsymbol{T}, \boldsymbol{X}) \exp\left\{ -\sum_{d=1}^{D} \int_{[0,1]\times\mathbb{X}} K_n(x)\, \tilde{\mu}_d^e(dz, dx) \right\} \prod_{d=1}^{D} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \tilde{\mu}_d^e(dZ_{djh}^*, dX_j^*)^{q_{djh}},$$

where the quantities $Q(\boldsymbol{T}, \boldsymbol{X})$ and $K_n(x)$ are defined as in (3.8). Conditionally on the random measure $\tilde{\mu}_0$ at the root of the hierarchical prior $\mathcal{Q}$, the extended random measures $\tilde{\mu}_1^e, \ldots, \tilde{\mu}_d^e$ are independent completely random measures on $[0,1] \times \mathbb{X}$, characterized by the homogeneous Lévy intensity

$$\tilde{\nu}^e(ds, dz, dx) = \rho(ds)\, H(dz)\, \tilde{\mu}_0(dx).$$

The conditional expectation of the augmented likelihood, given the random measure $\tilde{\mu}_0$, is

$$\mathbb{E}\left[\mathcal{L}(\tilde{\mu}_1^e, \ldots, \tilde{\mu}_D^e; \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}) \mid \tilde{\mu}_0\right] = Q(\boldsymbol{T}, \boldsymbol{X})\, \mathbb{E}\left[\exp\left\{ -\sum_{d=1}^{D} \int_{[0,1]\times\mathbb{X}} K_n(x)\, \tilde{\mu}_d^e(dz, dx) \right\} \prod_{d=1}^{D} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \tilde{\mu}_d^e(dZ_{djh}^*, dX_j^*)^{q_{djh}} \;\middle|\; \tilde{\mu}_0 \right];$$

by conditional independence of the extended random measures $\tilde{\mu}_1^e, \ldots, \tilde{\mu}_d^e$,

$$= Q(\boldsymbol{T}, \boldsymbol{X}) \prod_{d=1}^{D} \mathbb{E}\left[\exp\left\{ -\int_{[0,1]\times\mathbb{X}} K_n(x)\, \tilde{\mu}_d^e(dz, dx) \right\} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \tilde{\mu}_d^e(dZ_{djh}^*, dX_j^*)^{q_{djh}} \;\middle|\; \tilde{\mu}_0 \right],$$

and exploiting the identity in (2.10) with $\alpha(dz, dx) = H(dz)\, \tilde{\mu}_0(dx)$ and $f(z, x) = K_n(x)$ one has

$$= Q(\boldsymbol{T}, \boldsymbol{X}) \prod_{d=1}^{D} \exp\left\{ -\int_{\mathbb{X}} \psi(K_n(x))\, \tilde{\mu}_0(dx) \right\} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \tau(q_{djh}; K_n(X_j^*))\, H(dZ_{djh}^*)\, \tilde{\mu}_0(dX_j^*)$$

$$= Q(\boldsymbol{T}, \boldsymbol{X}) \exp\left\{ -\int_{\mathbb{X}} D\, \psi(K_n(x))\, \tilde{\mu}_0(dx) \right\} \prod_{j=1}^{k} \tilde{\mu}_0(dX_j^*)^{r_j} \prod_{d=1}^{D} \prod_{h=1}^{r_{dj}} \tau(q_{djh}; K_n(X_j^*))\, H(dZ_{djh}^*).$$

The random measure $\tilde{\mu}_0$ at the root of the hierarchical prior is itself a completely random measure with homogeneous Lévy intensity $\nu_0(ds, dx) = \rho_0(ds)\, \theta P_0(dx)$. Therefore, the joint marginal distribution of observations $(\boldsymbol{T}, \boldsymbol{\Delta})$ and latent variables $(\boldsymbol{X}, \boldsymbol{Z})$, that is, the expectation of the

augmented likelihood, is

$$\mathbb{P}(\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z}) = \mathbb{E}\left[\mathcal{L}(\tilde{\mu}_1^e,\ldots,\tilde{\mu}_D^e;\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z})\right] = \mathbb{E}\left[\,\mathbb{E}\left[\mathcal{L}(\tilde{\mu}_1^e,\ldots,\tilde{\mu}_D^e;\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z}) \mid \tilde{\mu}_0\right]\right]$$
$$= Q(\boldsymbol{T},\boldsymbol{X})\,\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} D\,\psi(K_n(x))\,\tilde{\mu}_0(dx)\right\}\prod_{j=1}^{k}\tilde{\mu}_0(dX_j^*)^{r_j}\right]$$
$$\times\prod_{j=1}^{k}\prod_{d=1}^{D}\prod_{h=1}^{r_{dj}}\tau(q_{djh};K_n(X_j^*))\,H(dZ_{djh}^*);$$

exploiting again the identity in (2.10) with $\alpha(dx) = \theta P_0(dx)$ and $f(x) = D\,\psi(K_n(x))$ one concludes that

$$= Q(\boldsymbol{T},\boldsymbol{X})\,\exp\left\{-\int_{\mathbb{X}}\psi_0\big(D\,\psi(K_n(x))\big)\,\theta P_0(dx)\right\}\prod_{j=1}^{k}\tau_0\big(r_j;D\,\psi(K_n(X_j^*))\big)\,\theta P_0(dX_j^*)$$
$$\times\prod_{j=1}^{k}\prod_{d=1}^{D}\prod_{h=1}^{r_{dj}}\tau(q_{djh};K_n(X_j^*))\,H(dZ_{djh}^*).$$

Proof of equations (3.12) and (3.13) and Proposition 3.3 The joint predictive distribution for the additional observations (T_{n+1},Δ_{n+1}) and the corresponding latent location X_{n+1} and latent mark Z_{n+1} is obtained from the joint marginal distribution in Proposition 3.1,

$$\mathbb{P}(T_{n+1},\Delta_{n+1},X_{n+1},Z_{n+1}\mid\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z}) = \frac{\mathbb{P}(T_{n+1},\Delta_{n+1},X_{n+1},Z_{n+1},\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z})}{\mathbb{P}(\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z})}.$$

where the numerator is the joint marginal distribution for $n+1$ observations and corresponding latent variables. The specific form of this quantity depends on the allocation of the additional observation within the latent partition structure, that is, on the specific values assumed by the latent variables X_{n+1} and Z_{n+1}. Three alternative cases can be identified:

(1) both the latent location X_{n+1} and latent mark Z_{n+1} display ties with values in $\boldsymbol{X}^*$ and $\boldsymbol{Z}^*$, respectively, say $X_{n+1} = X_j^*$ and $Z_{n+1} = Z_{djh}^*$, where necessarily $\Delta_{n+1} = d$;

(2) the latent location X_{n+1} displays a tie with values in $\boldsymbol{X}^*$, say $X_{n+1} = X_j^*$, while the latent mark Z_{n+1} assumes a value not included in $\boldsymbol{Z}^*$;

(3) both the latent location X_{n+1} and latent mark Z_{n+1} assume values not included in $\boldsymbol{X}^*$ and $\boldsymbol{Z}^*$, respectively.

These three cases are highlighted in the following expression for the numerator:

$$\begin{aligned}
&\mathbb{P}(T_{n+1} \in dt, \Delta_{n+1} = d, X_{n+1} \in dx, Z_{n+1} \in dz, \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}) \\
&\qquad = Q(\boldsymbol{T}, \boldsymbol{X})\, k(t;x)\, \exp\left\{ - \int_{\mathbb{X}} \psi_0\big(D\,\psi(K_{n+1}(y;t))\big)\, \theta P_0(dy) \right\} \\
&\quad \times \prod_{j=1}^{k} \left(\prod_{\ell=1}^{D} \prod_{h=1}^{r_{\ell j}} \tau(q_{\ell jh}; K_{n+1}(X_j^*;t))\, H(dZ_{\ell jh}^*) \right) \tau_0\big(r_j; D\,\psi(K_{n+1}(X_j^*;t))\big)\, \theta P_0(dX_j^*) \\
&\qquad \times \left\{ \sum_{j=1}^{k} \sum_{h=1}^{r_{dj}} \frac{\tau(q_{djh}+1; K_{n+1}(X_j^*;t))}{\tau(q_{djh}; K_{n+1}(X_j^*;t))}\, \delta_{X_j^*}(dx)\, \delta_{Z_{djh}^*}(dz) \qquad \leftarrow \text{ case (1)} \right. \\
&\text{case (2)} \;\rightarrow \qquad + \sum_{j=1}^{k} \tau(1; K_{n+1}(X_j^*;t))\, \frac{\tau_0(r_j+1; D\,\psi(K_{n+1}(X_j^*;t)))}{\tau_0(r_j; D\,\psi(K_{n+1}(X_j^*;t)))}\, \delta_{X_j^*}(dx)\, H(dz) \\
&\qquad\qquad \text{case (3)} \;\rightarrow \qquad \left. + \tau(1; K_{n+1}(x;t))\, \tau_0(1; D\,\psi(K_{n+1}(x;t)))\, \theta\, P_0(dx)\, H(dz) \right\} dt,
\end{aligned}$$

where $K_{n+1}(x;t) = K_n(x) + \int_0^t k(s,x)\, ds$ is the updated kernel term. Remarkably, these three cases coincide with the three cases considered in the description of the marginal sampling algorithm in Section 3.6. As a result, the joint predictive distribution is obtained by dividing the expression above by the joint marginal distribution in Proposition 3.1:

$$\begin{aligned}
&\mathbb{P}(T_{n+1} \in dt, \Delta_{n+1} = d, X_{n+1} \in dx, Z_{n+1} \in dz \mid \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}) \\
&\qquad = k(t;x)\, \exp\left\{ - \int_{\mathbb{X}} \big(\psi_0\big(D\,\psi(K_{n+1}(y;t))\big) - \psi_0\big(D\,\psi(K_n(y))\big)\big)\, \theta P_0(dy) \right\} \\
&\qquad\quad \times \prod_{j=1}^{k} \left(\prod_{\ell=1}^{D} \prod_{h=1}^{r_{\ell j}} \frac{\tau(q_{\ell jh}; K_{n+1}(X_j^*;t))}{\tau(q_{\ell jh}; K_n(X_j^*))} \right) \frac{\tau_0\big(r_j; D\,\psi(K_{n+1}(X_j^*;t))\big)}{\tau_0\big(r_j; D\,\psi(K_n(X_j^*))\big)} \\
&\qquad \times \left\{ \sum_{j=1}^{k} \sum_{h=1}^{r_{dj}} \frac{\tau(q_{djh}+1; K_{n+1}(X_j^*;t))}{\tau(q_{djh}; K_{n+1}(X_j^*;t))}\, \delta_{X_j^*}(dx)\, \delta_{Z_{djh}^*}(dz) \right. \\
&\qquad\quad + \sum_{j=1}^{k} \tau(1; K_{n+1}(X_j^*;t))\, \frac{\tau_0(r_j+1; D\,\psi(K_{n+1}(X_j^*;t)))}{\tau_0(r_j; D\,\psi(K_{n+1}(X_j^*;t)))}\, \delta_{X_j^*}(dx)\, H(dz) \\
&\qquad\quad \left. + \tau(1; K_{n+1}(x;t))\, \tau_0(1; D\,\psi(K_{n+1}(x;t)))\, \theta\, P_0(dx)\, H(dz) \right\} dt.
\end{aligned}$$

The expressions in equations (3.12) and (3.13) are derived from the quantity above by marginalization with respect to the latent mark Z_{n+1}; on the other hand, the predictive distribution for

the additional observation (T_{n+1}, Δ_{n+1}) is obtained by marginalizing both latent variables,

$$\begin{aligned}
&\mathbb{P}(T_{n+1} \in dt, \Delta_{n+1} = d \mid \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}) \\
&\qquad = \exp\left\{ - \int_{\mathbb{X}} \big(\psi_0\big(D\,\psi(K_{n+1}(y;t))\big) - \psi_0\big(D\,\psi(K_n(y))\big)\big)\,\theta P_0(dy) \right\} \\
&\qquad\quad \times \prod_{j=1}^{k} \left(\prod_{\ell=1}^{D} \prod_{h=1}^{r_{\ell j}} \frac{\tau(q_{\ell jh}; K_{n+1}(X_j^*;t))}{\tau(q_{\ell jh}; K_n(X_j^*))} \right) \frac{\tau_0\big(r_j; D\,\psi(K_{n+1}(X_j^*;t))\big)}{\tau_0\big(r_j; D\,\psi(K_n(X_j^*))\big)} \\
&\qquad \times \left\{ \sum_{j=1}^{k} k(t; X_j^*) \sum_{h=1}^{r_{dj}} \frac{\tau(q_{djh}+1; K_{n+1}(X_j^*;t))}{\tau(q_{djh}; K_{n+1}(X_j^*;t))} \right. \\
&\qquad\quad + \sum_{j=1}^{k} k(t; X_j^*)\, \tau(1; K_{n+1}(X_j^*;t)) \frac{\tau_0(r_j+1; D\,\psi(K_{n+1}(X_j^*;t)))}{\tau_0(r_j; D\,\psi(K_{n+1}(X_j^*;t)))} \\
&\qquad\quad \left. + \int_{\mathbb{X}} k(t;x)\, \tau(1; K_{n+1}(x;t))\, \tau_0(1; D\,\psi(K_{n+1}(x;t)))\, \theta\, P_0(dx) \right\} dt. \qquad \text{(A.1)}
\end{aligned}$$

Finally, the predictive distribution for the event type Δ_{n+1}, given the survival time T_{n+1} is proportional to the expression above, in which factors not depending on $\Delta_{n+1} = d$ can be ignored:

$$\begin{aligned}
\mathbb{P}(\Delta_{n+1} = d \mid T_{n+1} = t, \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}) &\propto \sum_{j=1}^{k} k(t; X_j^*) \sum_{h=1}^{r_{dj}} \frac{\tau(q_{djh}+1; K_{n+1}(X_j^*;t))}{\tau(q_{djh}; K_{n+1}(X_j^*;t))} \\
&\quad + \sum_{j=1}^{k} k(t; X_j^*)\, \tau(1; K_{n+1}(X_j^*;t)) \frac{\tau_0(r_j+1; D\,\psi(K_{n+1}(X_j^*;t)))}{\tau_0(r_j; D\,\psi(K_{n+1}(X_j^*;t)))} \\
&\quad + \int_{\mathbb{X}} k(t;x) \tau(1; K_{n+1}(x;t))\, \tau_0(1; D\,\psi(K_{n+1}(x;t)))\, \theta\, P_0(dx).
\end{aligned}$$

Proof of Proposition 3.5 The posterior distribution of the random measure $\tilde{\mu}_0$ at the root of the hierarchy, given the observations $(\boldsymbol{T}, \boldsymbol{\Delta})$ and the latent variables $(\boldsymbol{X}, \boldsymbol{Z})$, can be characterized through its conditional Laplace transform,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{ - \int_{\mathbb{X}} h_0(x)\, \tilde{\mu}_0(dx) \right\} \,\middle|\, \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z} \right] \\
&\qquad = \frac{\mathbb{E}\left[\exp\left\{ - \int_{\mathbb{X}} h_0(x)\, \tilde{\mu}_0(dx) \right\} \mathcal{L}(\tilde{\mu}_1^e, \ldots, \tilde{\mu}_D^e; \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}) \right]}{\mathbb{E}\left[\mathcal{L}(\tilde{\mu}_1^e, \ldots, \tilde{\mu}_D^e; \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}) \right]},
\end{aligned}$$

where $h_0 \colon \mathbb{X} \mapsto \mathbb{R}^+$ is any non-negative measurable function. The quantity at the denominator is the joint marginal distribution in Proposition 3.1, while the expectation at the numerator can be computed similarly; specifically, the conditional expectation of the numerator, given the random

measure $\tilde{\mu}_0$, is

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h_0(x)\,\tilde{\mu}_0(dx)\right\}\mathcal{L}(\tilde{\mu}_1^e,\ldots,\tilde{\mu}_D^e;\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z})\;\Big|\;\tilde{\mu}_0\right]\\
&\qquad=\exp\left\{-\int_{\mathbb{X}} h_0(x)\,\tilde{\mu}_0(dx)\right\}\mathbb{E}\left[\mathcal{L}(\tilde{\mu}_1^e,\ldots,\tilde{\mu}_D^e;\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z})\mid\tilde{\mu}_0\right]\\
&\qquad=Q(\boldsymbol{T},\boldsymbol{X})\,\exp\left\{-\int_{\mathbb{X}}\left(h_0(x)+D\,\psi(K_n(x))\right)\tilde{\mu}_0(dx)\right\}\prod_{j=1}^{k}\tilde{\mu}_0(dX_j^*)^{r_j}\\
&\qquad\qquad\times\prod_{j=1}^{k}\prod_{d=1}^{D}\prod_{h=1}^{r_{dj}}\tau(q_{djh};K_n(X_j^*))\,H(dZ_{djh}^*).
\end{aligned}$$

The expectation of the numerator is obtained by marginalization of the conditional expectation with respect to the random measure $\tilde{\mu}_0$,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h_0(x)\,\tilde{\mu}_0(dx)\right\}\mathcal{L}(\tilde{\mu}_1^e,\ldots,\tilde{\mu}_D^e;\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z})\right]\\
&\qquad=Q(\boldsymbol{T},\boldsymbol{X})\,\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}}\left(h_0(x)+D\,\psi(K_n(x))\right)\tilde{\mu}_0(dx)\right\}\prod_{j=1}^{k}\tilde{\mu}_0(dX_j^*)^{r_j}\right]\\
&\qquad\qquad\times\prod_{j=1}^{k}\prod_{d=1}^{D}\prod_{h=1}^{r_{dj}}\tau(q_{djh};K_n(X_j^*))\,H(dZ_{djh}^*);
\end{aligned}$$

exploiting the identity in (2.10) with $f(x)=h_0(x)+D\,\psi(K_n(x))$ one gets

$$\begin{aligned}
&=Q(\boldsymbol{T},\boldsymbol{X})\,\exp\left\{-\int_{\mathbb{X}}\psi_0\big(h_0(x)+D\,\psi(K_n(x))\big)\,\theta P_0(dx)\right\}\\
&\qquad\times\prod_{j=1}^{k}\tau_0\big(r_j;h_0(X_j^*)+D\,\psi(K_n(X_j^*))\big)\,\theta P_0(dX_j^*)\prod_{j=1}^{k}\prod_{d=1}^{D}\prod_{h=1}^{r_{dj}}\tau(q_{djh};K_n(X_j^*))\,H(dZ_{djh}^*).
\end{aligned}$$

Therefore, the conditional Laplace transform is

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h_0(x)\,\tilde{\mu}_0(dx)\right\}\;\Big|\;\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z}\right]\\
&\qquad=\exp\left\{-\int_{\mathbb{X}}\Big(\psi_0\big(h_0(x)+D\,\psi(K_n(x))\big)-\psi_0\big(D\,\psi(K_n(x))\big)\Big)\,\theta P_0(dx)\right\}\\
&\qquad\qquad\times\prod_{j=1}^{k}\frac{\tau_0\big(r_j;h_0(X_j^*)+D\,\psi(K_n(X_j^*))\big)}{\tau_0\big(r_j;D\,\psi(K_n(X_j^*))\big)}.
\end{aligned}$$

Thanks to simple algebra with Laplace exponents, the exponential term can be rewritten as

$$\exp\left\{-\int_{\mathbb{X}}\Big(\psi_0\big(h_0(x)+D\,\psi(K_n(x))\big)-\psi_0\big(D\,\psi(K_n(x))\big)\Big)\,\theta P_0(dx)\right\}$$
$$=\exp\left\{-\int_{\mathbb{X}}\psi_0^*(h_0(x)\mid x)\,\theta P_0(dx)\right\}=\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}}h_0(x)\,\tilde{\mu}_0^*(dx)\right\}\right],$$

where $\psi_0^*(\cdot\mid x)$ is the Laplace exponent corresponding to $\rho_0^*(ds\,|\,x)=\exp\{-D\,\psi(K_n(x))\,s\}\,\rho_0(ds)$, and $\tilde{\mu}_0^*$ is a completely random measure having Lévy intensity $\nu_0^*(ds,dx)=\rho_0^*(ds\,|\,x)\,\theta P_0(dx)$. Moreover, for each $j=1,\ldots,k$,

$$\tau_0\big(r_j;h_0(X_j^*)+D\,\psi(K_n(X_j^*))\big)=\int_{\mathbb{R}^+}s^{r_j}e^{-h_0(X_j^*)\,s}\,\exp\big\{-D\,\psi(K_n(X_j^*)\,s\big\}\,\rho_0(ds)$$
$$=\int_{\mathbb{R}^+}e^{-h_0(X_j^*)\,s}\,s^{r_j}\rho_0^*(ds\mid X_j^*);$$

therefore, the ratios in the expression of the conditional Laplace transform can be regarded as expectations,

$$\frac{\tau_0\big(r_j;h_0(X_j^*)+D\,\psi(K_n(X_j^*))\big)}{\tau_0\big(r_j;D\,\psi(K_n(X_j^*))\big)}=\frac{\displaystyle\int_{\mathbb{R}^+}e^{-h_0(X_j^*)\,s}\,s^{r_j}\rho_0^*(ds\mid X_j^*)}{\displaystyle\int_{\mathbb{R}^+}s^{r_j}\rho_0^*(ds\mid X_j^*)}=\mathbb{E}\left[\exp\big\{-h_0(X_j^*)\,V_j\big\}\right],$$

where each V_j is a non-negative random variable having density function proportional to

$$s^{r_j}\exp\big\{-D\,\psi(K_n(X_j^*))\,s\big\}\,\rho_0(ds)=s^{r_j}\rho_0^*(ds\mid X_j^*),\qquad j=1,\ldots,k.$$

In conclusion, the conditional Laplace transform can be written as

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}}h_0(x)\,\tilde{\mu}_0(dx)\right\}\;\Big|\;\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z}\right]$$
$$=\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}}h_0(x)\,\tilde{\mu}_0^*(dx)\right\}\right]\prod_{j=1}^{k}\mathbb{E}\left[\exp\big\{-h_0(X_j^*)\,V_j\big\}\right]$$
$$=\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}}h_0(x)\,\tilde{\mu}_0^*(dx)-\sum_{j=1}^{k}h_0(X_j^*)\,V_j\right\}\right],$$

which highlights the structure of the posterior distribution of $\tilde{\mu}_0$ as the sum of the non-homogeneous completely random measure $\tilde{\mu}_0^*$ and the random jumps $V_1,\ldots,V_k$ at fixed locations $X_1^*,\ldots,X_k^*$; moreover, these components are mutually independent (second line). A similar characterization is obtained for the posterior distribution of the extended random measures $\tilde{\mu}_1^e,\ldots,\tilde{\mu}_D^e$ at the lower level of the hierarchy; indeed, their joint conditional Laplace transform,

given the observations $(\boldsymbol{T}, \boldsymbol{\Delta})$, the latent variables $(\boldsymbol{X}, \boldsymbol{Z})$ and the root random measure $\tilde{\mu}_0$, is

$$\mathbb{E}\left[\exp\left\{-\sum_{d=1}^{D}\int_{[0,1]\times\mathbb{X}} h_d(z,x)\,\tilde{\mu}_d^e(dz,dx)\right\} \,\middle|\, \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}, \tilde{\mu}_0\right]$$
$$= \frac{\mathbb{E}\left[\exp\left\{-\sum_{d=1}^{D}\int_{[0,1]\times\mathbb{X}} h_d(z,x)\,\tilde{\mu}_d^e(dz,dx)\right\} \mathcal{L}(\tilde{\mu}_1^e,\ldots,\tilde{\mu}_D^e;\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z}) \mid \tilde{\mu}_0\right]}{\mathbb{E}\left[\mathcal{L}(\tilde{\mu}_1^e,\ldots,\tilde{\mu}_D^e;\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z}) \mid \tilde{\mu}_0\right]},$$

where $h_d\colon [0,1]\times\mathbb{X} \mapsto \mathbb{R}^+$, for $d = 1,\ldots,D$, are non-negative measurable functions. The quantity at the denominator is the conditional expectation of augmented likelihood, obtained as an intermediate step in the proof of Proposition 3.1, while the expectation at the numerator can be computed similarly; in particular,

$$\mathbb{E}\left[\exp\left\{-\sum_{d=1}^{D}\int_{[0,1]\times\mathbb{X}} h_d(z,x)\,\tilde{\mu}_d^e(dz,dx)\right\} \mathcal{L}(\tilde{\mu}_1^e,\ldots,\tilde{\mu}_D^e;\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z}) \mid \tilde{\mu}_0\right]$$
$$= Q(\boldsymbol{T},\boldsymbol{X})\,\mathbb{E}\left[\exp\left\{-\sum_{d=1}^{D}\int_{[0,1]\times\mathbb{X}} \big(h_d(z,x)+K_n(x)\big)\,\tilde{\mu}_d^e(dz,dx)\right\}\right.$$
$$\left.\times \prod_{d=1}^{D}\prod_{j=1}^{k}\prod_{h=1}^{r_{dj}} \tilde{\mu}_d^e(dZ_{djh}^*,dX_j^*)^{q_{djh}}\right];$$

by conditional independence of the extended random measures $\tilde{\mu}_1^e,\ldots,\tilde{\mu}_d^e$,

$$= Q(\boldsymbol{T},\boldsymbol{X}) \prod_{d=1}^{D}\mathbb{E}\left[\exp\left\{-\int_{[0,1]\times\mathbb{X}} \big(h_d(z,x)+K_n(x)\big)\,\tilde{\mu}_d^e(dz,dx)\right\}\prod_{j=1}^{k}\prod_{h=1}^{r_{dj}} \tilde{\mu}_d^e(dZ_{djh}^*,dX_j^*)^{q_{djh}}\right],$$

and exploiting the identity in (2.10) with $f(z,x) = h_d(z,x) + K_n(x)$, for $d = 1,\ldots,D$, one gets

$$= Q(\boldsymbol{T},\boldsymbol{X}) \prod_{d=1}^{D}\exp\left\{-\int_{[0,1]\times\mathbb{X}} \psi(h_d(z,x)+K_n(x))\,H(dz)\,\tilde{\mu}_0(dx)\right\}$$
$$\times \prod_{d=1}^{D}\prod_{j=1}^{k}\prod_{h=1}^{r_{dj}} \tau(q_{djh};h_d(Z_{djh}^*,X_j^*)+K_n(X_j^*))\,H(dZ_{djh}^*)\,\tilde{\mu}_0(dX_j^*).$$

Therefore, the conditional Laplace transform is

$$\mathbb{E}\left[\exp\left\{-\sum_{d=1}^{D}\int_{[0,1]\times\mathbb{X}} h_d(z,x)\,\tilde{\mu}_d^e(dz,dx)\right\}\;\middle|\;\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z},\tilde{\mu}_0\right]$$
$$=\prod_{d=1}^{D}\exp\left\{-\int_{[0,1]\times\mathbb{X}}\Big(\psi(h_d(z,x)+K_n(x))-\psi(K_n(x))\Big)H(dz)\,\tilde{\mu}_0(dx)\right\}$$
$$\times\prod_{d=1}^{D}\prod_{j=1}^{k}\prod_{h=1}^{r_{dj}}\frac{\tau(q_{djh};h_d(Z^*_{djh},X^*_j)+K_n(X^*_j))}{\tau(q_{djh};K_n(X^*_j))}.$$

For each $d=1,\dots,D$, the exponential term can be rewritten as

$$\exp\left\{-\int_{[0,1]\times\mathbb{X}}\Big(\psi(h_d(z,x)+K_n(x))-\psi(K_n(x))\Big)\,H(dz)\,\tilde{\mu}_0(dx)\right\}$$
$$=\mathbb{E}\left[\exp\left\{-\int_{[0,1]\times\mathbb{X}}h_d(z,x)\,\tilde{\mu}^*_d(dz,dx)\right\}\right]$$

where $\tilde{\mu}^*_d$ is a completely random measure having non-homogeneous Lévy intensity $\nu^*(ds,dz,dx)=\rho^*(ds\,|\,x)\,H(dz)\,\tilde{\mu}_0(dx)$, with jump component $\rho^*(ds\,|\,x)=e^{-K_n(x)\,s}\,\rho(ds)$. Moreover, for each $d=1,\dots,D$, $j=1,\dots,k$, and $h=1,\dots,r_{dj}$,

$$\tau(q_{djh};h_d(Z^*_{djh},X^*_j)+K_n(X^*_j))=\int_{\mathbb{R}^+}e^{-h_d(Z^*_{djh},X^*_j)\,s}\,s^{q_{djh}}\rho^*(ds\mid X^*_j);$$

therefore, the ratios in the expression of the conditional Laplace transform can be regarded as expectations

$$\frac{\tau(q_{djh};h_d(Z^*_{djh},X^*_j)+K_n(X^*_j))}{\tau(q_{djh};K_n(X^*_j))}=\mathbb{E}\left[\exp\left\{-h_d(Z^*_{djh},X^*_j)\,S_{djh}\right\}\right],$$

where each S_{djh} is a non-negative random variable having density function proportional to

$$s^{q_{djh}}\,e^{-K_n(X^*_j)\,s}\,\rho(ds)=s^{q_{djh}}\rho^*(ds\mid X^*_j),\qquad d=1,\dots,D,\ j=1,\dots,k,\ h=1,\dots,r_{dj}.$$

In conclusion, the conditional Laplace transform can be written as

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{-\sum_{d=1}^{D}\int_{[0,1]\times\mathbb{X}} h_d(z,x)\,\tilde{\mu}_d^e(dz,dx)\right\}\;\middle|\;\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z},\tilde{\mu}_0\right] \\
&\quad=\prod_{d=1}^{D}\mathbb{E}\left[\exp\left\{-\int_{[0,1]\times\mathbb{X}} h_d(z,x)\,\tilde{\mu}_d^*(dz,dx)\right\}\right]\prod_{j=1}^{k}\prod_{h=1}^{r_{dj}}\mathbb{E}\left[\exp\left\{-h_d(Z^*_{djh},X^*_j)\,S_{djh}\right\}\right] \\
&\quad=\prod_{d=1}^{D}\mathbb{E}\left[\exp\left\{-\int_{[0,1]\times\mathbb{X}} h_d(z,x)\,\tilde{\mu}_d^*(dz,dx)-\sum_{j=1}^{k}\sum_{h=1}^{r_{dj}}h_d(Z^*_{djh},X^*_j)\,S_{djh}\right\}\right],
\end{aligned}$$

which highlights the structure of the posterior distribution of each $\tilde{\mu}_d$, given the root random measure $\tilde{\mu}_0$, as the sum of the non-homogeneous completely random measure $\tilde{\mu}_d^*$ and the random jumps $(S_{djh})_{jh}$ at fixed locations; moreover, these components are mutually independent, and the random measures $\tilde{\mu}_1,\ldots,\tilde{\mu}_D$ are conditionally independent, given $\tilde{\mu}_0$. The posterior characterization in Proposition 3 is obtained by marginalization with respect to the first component of the random measure, that is $\tilde{\mu}_d(dx)=\tilde{\mu}_d^e([0,1],dx)$.

Proof of Proposition 3.7 The posterior estimate of the overall survival function $\tilde{S}(t)$, given the observations $(\boldsymbol{T},\boldsymbol{\Delta})$ and the latent variables $(\boldsymbol{X},\boldsymbol{Z})$, for $t>0$, is obtained as

$$\begin{aligned}
\mathbb{E}\left[\tilde{S}(t)\mid\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z}\right]&=\mathbb{E}\left[\exp\left\{-\sum_{d=1}^{D}\int_0^t\int_{\mathbb{X}}k(s;x)\,\tilde{\mu}_d(dx)\,ds\right\}\;\middle|\;\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z}\right] \\
&=\mathbb{E}\left[\exp\left\{-\sum_{d=1}^{D}\int_{\mathbb{X}}K_1(x;t)\,\tilde{\mu}_d(dx)\right\}\;\middle|\;\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z}\right],
\end{aligned}$$

where $K_1(x;t)=\displaystyle\int_0^t k(s;x)\,ds$ is the integrated kernel up to time $t>0$. Conditionally on the root random measure $\tilde{\mu}_0$, the specifications of the posterior distributions of random measures $\tilde{\mu}_1,\ldots,\tilde{\mu}_D$ are substituted into the expression above, so that

$$\begin{aligned}
&\mathbb{E}\left[\tilde{S}(t)\mid\boldsymbol{T},\boldsymbol{\Delta},\boldsymbol{X},\boldsymbol{Z},\tilde{\mu}_0\right] \\
&\quad=\mathbb{E}\left[\exp\left\{-\sum_{d=1}^{D}\int_{\mathbb{X}}K_1(x;t)\,\tilde{\mu}_d^*(dx)-\sum_{d=1}^{D}\sum_{j=1}^{k}\sum_{h=1}^{r_{dj}}K_1(X^*_j;t)\,S_{djh}\right\}\;\middle|\;\tilde{\mu}_0\right] \\
&\quad=\mathbb{E}\left[\prod_{d=1}^{D}\exp\left\{-\int_{\mathbb{X}}K_1(x;t)\,\tilde{\mu}_d^*(dx)\right\}\prod_{j=1}^{k}\prod_{h=1}^{r_{dj}}\exp\left\{-K_1(X^*_j;t)\,S_{djh}\right\}\;\middle|\;\tilde{\mu}_0\right];
\end{aligned}$$

by conditional independence of the random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$, and of the their components,

$$= \prod_{d=1}^{D} \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} K_1(x;t)\,\tilde{\mu}_d^*(dx)\right\} \;\middle|\; \tilde{\mu}_0\right] \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \mathbb{E}\left[\exp\left\{-K_1(X_j^*;t)\,S_{djh}\right\}\right].$$

The posterior characterization of the random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ implies that

$$\begin{aligned}
&= \prod_{d=1}^{D} \exp\left\{-\int_{\mathbb{X}} \psi^*(K_1(x;t) \mid x)\,\tilde{\mu}_0(dx)\right\} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \frac{\tau^*(q_{djh}; K_1(X_j^*;t) \mid X_j^*)}{\tau^*(q_{djh}; 0 \mid X_j^*)} \\
&= \exp\left\{-\int_{\mathbb{X}} D\,\psi^*(K_1(x;t) \mid x)\,\tilde{\mu}_0(dx)\right\} \prod_{d=1}^{D} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \frac{\tau^*(q_{djh}; K_1(X_j^*;t) \mid X_j^*)}{\tau^*(q_{djh}; 0 \mid X_j^*)},
\end{aligned}$$

where $\psi^*(\cdot \mid x)$ and $\tau^*(m;\cdot \mid x)$ are the Laplace exponent and cumulants, respectively, corresponding to $\rho^*(ds \mid x)$. The posterior estimate is obtained by substituting the specification of the posterior distribution of the root random measure $\tilde{\mu}_0$, and exploiting the mutual independence of its components,

$$\begin{aligned}
\mathbb{E}\left[\tilde{S}(t) \mid \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}\right] &= \mathbb{E}\left[\mathbb{E}\left[\tilde{S}(t) \mid \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}, \tilde{\mu}_0\right]\right] \\
&= \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} D\,\psi^*(K_1(x;t) \mid x)\,\tilde{\mu}_0^*(dx)\right\}\right] \prod_{j=1}^{k} \mathbb{E}\left[\exp\left\{-D\,\psi^*(K_1(X_j^*;t) \mid X_j^*)\,V_j\right\}\right] \\
&\qquad \times \prod_{d=1}^{D} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \frac{\tau^*(q_{djh}; K_1(X_j^*;t) \mid X_j^*)}{\tau^*(q_{djh}; 0 \mid X_j^*)}.
\end{aligned}$$

In conclusion, the posterior characterization of the random measure $\tilde{\mu}_0$ implies that

$$\begin{aligned}
\mathbb{E}\left[\tilde{S}(t) \mid \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}\right] &= \exp\left\{-\int_{\mathbb{X}} \psi_0^*\big(D\,\psi^*(K_1(x;t) \mid x) \mid x\big)\,\theta P_0(dx)\right\} \\
&\qquad \times \prod_{j=1}^{k} \frac{\tau_0^*\big(r_j; D\,\psi^*(K_1(X_j^*;t) \mid X_j^*) \mid X_j^*\big)}{\tau_0^*(r_j; 0 \mid X_j^*)} \prod_{j=1}^{k} \prod_{d=1}^{D} \prod_{h=1}^{r_{dj}} \frac{\tau^*(q_{djh}; K_1(X_j^*;t) \mid X_j^*)}{\tau^*(q_{djh}; 0 \mid X_j^*)},
\end{aligned}$$

where $\psi_0^*(\cdot \mid x)$ and $\tau_0^*(m;\cdot \mid x)$ are the Laplace exponent and cumulants, respectively, corresponding to $\rho_0^*(ds \mid x)$.

Proof of Proposition 3.8 The proof follows the same structure of the proof of Proposition 3.7. Indeed, the posterior estimate of the cause-specific incidence function $\tilde{p}(dt, \delta)$, given the

observations $(\boldsymbol{T}, \boldsymbol{\Delta})$ and the latent variables $(\boldsymbol{X}, \boldsymbol{Z})$, for $t > 0$, is obtained as

$$\mathbb{E}\big[\tilde{p}(dt, \delta) \mid \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}\big] = \mathbb{E}\left[\int_{\mathbb{X}} k(t; x)\, \tilde{\mu}_\delta(dx)\, \exp\left\{-\sum_{d=1}^{D} \int_{\mathbb{X}} K_t(y)\, \tilde{\mu}_d(dy)\right\} dt \;\middle|\; \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}\right],$$

where $K_1(x; t) = \displaystyle\int_0^t k(s; x)\, ds$ is the integrated kernel up to time $t > 0$. Conditionally on the root random measure $\tilde{\mu}_0$, the specifications of the posterior distributions of random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_D$ are substituted into the expression above, obtaining

$$\begin{aligned}
\mathbb{E}\big[\tilde{p}(dt, \delta) \mid \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}, \tilde{\mu}_0\big] = \mathbb{E}\Bigg[& \prod_{d \neq \delta} \exp\left\{-\int_{\mathbb{X}} K_1(y; t)\, \tilde{\mu}_d^*(dy)\right\} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \exp\left\{-K_1(X_j^*; t)\, S_{djh}\right\} \\
& \times \left(\int_{\mathbb{X}} k(t; x)\, \tilde{\mu}_\delta^*(dx)\, \exp\left\{-\int_{\mathbb{X}} K_1(y; t)\, \tilde{\mu}_\delta^*(dy)\right\} \prod_{j=1}^{k} \prod_{h=1}^{r_{\delta j}} \exp\left\{-K_1(X_j^*; t)\, S_{\delta jh}\right\} \right. \\
& \left. + \sum_{\ell=1}^{k} k(t; X_\ell^*) \sum_{\epsilon=1}^{r_{\delta\ell}} \exp\left\{-\int_{\mathbb{X}} K_1(y; t)\, \tilde{\mu}_\delta^*(dy)\right\} S_{\delta\ell\epsilon} \prod_{j=1}^{k} \prod_{h=1}^{r_{\delta j}} \exp\left\{-K_1(X_j^*; t)\, S_{\delta jh}\right\} \right) dt \;\middle|\; \tilde{\mu}_0 \Bigg];
\end{aligned}$$

by conditional independence of the random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$, and of the their components,

$$\begin{aligned}
= & \prod_{d \neq \delta} \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} K_1(y; t)\, \tilde{\mu}_d^*(dy)\right\} \;\middle|\; \tilde{\mu}_0\right] \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \mathbb{E}\left[\exp\left\{-K_1(X_j^*; t)\, S_{djh}\right\}\right] \\
& \times \left(\int_{\mathbb{X}} k(t; x)\, \mathbb{E}\left[\tilde{\mu}_\delta^*(dx)\, \exp\left\{-\int_{\mathbb{X}} K_1(y; t)\, \tilde{\mu}_\delta^*(dy)\right\} \;\middle|\; \tilde{\mu}_0\right] \prod_{j=1}^{k} \prod_{h=1}^{r_{\delta j}} \mathbb{E}\left[\exp\left\{-K_1(X_j^*; t)\, S_{\delta jh}\right\}\right] \right. \\
& \quad + \sum_{\ell=1}^{k} k(t; X_\ell^*) \sum_{\epsilon=1}^{r_{\delta\ell}} \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} K_1(y; t)\, \tilde{\mu}_\delta^*(dy)\right\} \;\middle|\; \tilde{\mu}_0\right] \mathbb{E}\left[S_{\delta\ell\epsilon} \exp\left\{-K_1(X_\ell^*; t)\, S_{\delta\ell\epsilon}\right\}\right] \\
& \qquad \left. \times \prod_{j \neq \ell} \prod_{h \neq \epsilon} \mathbb{E}\left[\exp\left\{-K_1(X_j^*; t)\, S_{\delta jh}\right\}\right] \right) dt.
\end{aligned}$$

The posterior characterization of the random measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ implies that

$$= \exp\left\{ - \int_{\mathbb{X}} (D-1)\, \psi^*(K_1(y;t) \mid y)\, \tilde{\mu}_0(dy) \right\} \prod_{d \neq \delta} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \frac{\tau^*(q_{djh}; K_1(X_j^*;t) \mid X_j^*)}{\tau^*(q_{djh}; 0 \mid X_j^*)}$$

$$\times \left(\int_{\mathbb{X}} k(t;x)\, \tau^*(1; K_1(x;t) \mid x)\, \tilde{\mu}_0(dx)\, \exp\left\{ - \int_{\mathbb{X}} \psi^*(K_1(y;t) \mid y)\, \tilde{\mu}_0(dy) \right\} \prod_{j=1}^{k} \prod_{h=1}^{r_{\delta j}} \frac{\tau^*(q_{\delta jh}; K_1(X_j^*;t) \mid X_j^*)}{\tau^*(q_{\delta jh}; 0 \mid X_j^*)} \right.$$

$$+ \sum_{\ell=1}^{k} k(t; X_\ell^*) \sum_{\epsilon=1}^{r_{\delta\ell}} \exp\left\{ - \int_{\mathbb{X}} \psi^*(K_1(y;t) \mid y)\, \tilde{\mu}_0(dy) \right\} \frac{\tau^*(q_{\delta\ell\epsilon} + 1; K_1(X_\ell^*;t) \mid X_\ell^*)}{\tau^*(q_{\delta\ell\epsilon}; 0 \mid X_\ell^*)}$$

$$\left. \times \prod_{j \neq \ell} \prod_{h \neq \epsilon} \frac{\tau^*(q_{\delta jh}; K_1(X_j^*;t) \mid X_j^*)}{\tau^*(q_{\delta jh}; 0 \mid X_j^*)} \right) dt;$$

in the second line, the identity in (2.10) is exploited with $f(y) = K_1(y;t)$, while in the last line

$$\mathbb{E}\left[S_{\delta\ell\epsilon} \exp\left\{ -K_1(X_j^*;t)\, S_{\delta\ell\epsilon} \right\} \right] = \frac{\displaystyle\int_{\mathbb{R}^+} s\, e^{-K_1(X_\ell^*;t)\, s}\, s^{q_{\delta\ell\epsilon}} \rho^*(ds \mid X_\ell^*)}{\displaystyle\int_{\mathbb{R}^+} s^{q_{\delta\ell\epsilon}} \rho^*(ds \mid X_\ell^*)} = \frac{\tau^*(q_{\delta\ell\epsilon} + 1; K_1(X_\ell^*;t) \mid X_\ell^*)}{\tau^*(q_{\delta\ell\epsilon}; 0 \mid X_\ell^*)}.$$

Rearranging the terms in the expression above, one obtains

$$\mathbb{E}\big[\tilde{p}(dt, \delta) \mid \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}, \tilde{\mu}_0\big]$$

$$= \exp\left\{ - \int_{\mathbb{X}} D\, \psi^*(K_1(y;t) \mid y)\, \tilde{\mu}_0(dy) \right\} \prod_{d=1}^{D} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \frac{\tau^*(q_{djh}; K_1(X_j^*;t) \mid X_j^*)}{\tau^*(q_{djh}; 0 \mid X_j^*)}$$

$$\times \left(\int_{\mathbb{X}} k(t;x)\, \tau^*(1; K_1(x;t) \mid x)\, \tilde{\mu}_0(dx) + \sum_{j=1}^{k} k(t; X_j^*) \sum_{h=1}^{r_{\delta j}} \frac{\tau^*(q_{\delta jh} + 1; K_1(X_j^*;t) \mid X_j^*)}{\tau^*(q_{\delta jh}; K_1(X_j^*;t) \mid X_j^*)} \right) dt.$$

The posterior estimate is obtained by substituting the specification of the posterior distribution of the root random measure $\tilde{\mu}_0$,

$$\mathbb{E}\big[\tilde{p}(dt, \delta) \mid \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}\big] = \mathbb{E}\left[\mathbb{E}\big[\tilde{p}(dt, \delta) \mid \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}, \tilde{\mu}_0\big] \right]$$

$$= \mathbb{E}\left[\exp\left\{ - \int_{\mathbb{X}} D\, \psi^*(K_1(y;t) \mid y)\, \tilde{\mu}_0^*(dy) \right\} \prod_{j=1}^{k} \exp\left\{ -D\, \psi^*(K_1(X_j^*;t) \mid X_j^*)\, V_j \right\} \right.$$

$$\times \prod_{d=1}^{D} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \frac{\tau^*(q_{djh}; K_1(X_j^*;t) \mid X_j^*)}{\tau^*(q_{djh}; 0 \mid X_j^*)} \left(\int_{\mathbb{X}} k(t;x)\, \tau^*(1; K_1(x;t) \mid x)\, \tilde{\mu}_0^*(dx) \right.$$

$$\left. \left. + \sum_{j=1}^{k} k(t; X_j^*)\, \tau^*(1; K_1(X_j^*;t) \mid X_j^*)\, V_j + \sum_{j=1}^{k} k(t; X_j^*) \sum_{h=1}^{r_{\delta j}} \frac{\tau^*(q_{\delta jh} + 1; K_1(X_j^*;t) \mid X_j^*)}{\tau^*(q_{\delta jh}; K_1(X_j^*;t) \mid X_j^*)} \right) \right] dt;$$

by rearranging the terms and exploiting the mutual independence of $\tilde{\mu}_0^*$ and $V_1, \ldots, V_k$, one obtains a sum of three components:

$$
\begin{aligned}
&= \int_{\mathbb{X}} k(t;x)\, \tau^*(1; K_1(x;t) \mid x)\, \mathbb{E}\left[\tilde{\mu}_0^*(dx)\, \exp\left\{-\int_{\mathbb{X}} D\,\psi^*(K_1(y;t) \mid y)\, \tilde{\mu}_0^*(dy)\right\}\right] \\
&\qquad \times \prod_{j=1}^{k} \mathbb{E}\left[\exp\left\{-D\,\psi^*(K_1(X_j^*;t) \mid X_j^*)\, V_j\right\}\right] \prod_{d=1}^{D} \prod_{h=1}^{r_{dj}} \frac{\tau^*(q_{djh}; K_1(X_j^*;t) \mid X_j^*)}{\tau^*(q_{djh}; 0 \mid X_j^*)}\, dt \\
&+ \sum_{\ell=1}^{k} k(t; X_\ell^*)\, \tau^*(1; K_1(X_\ell^*;t) \mid X_\ell^*)\, \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} D\,\psi^*(K_1(y;t) \mid y)\, \tilde{\mu}_0^*(dy)\right\}\right] \\
&\qquad \times \mathbb{E}\left[V_\ell\, \exp\left\{-D\,\psi^*(K_1(X_\ell^*;t) \mid X_\ell^*)\, V_\ell\right\}\right] \prod_{j \neq \ell} \mathbb{E}\left[\exp\left\{-D\,\psi^*(K_1(X_j^*;t) \mid X_j^*)\, V_j\right\}\right] \\
&\qquad\qquad \times \prod_{d=1}^{D} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \frac{\tau^*(q_{djh}; K_1(X_j^*;t) \mid X_j^*)}{\tau^*(q_{djh}; 0 \mid X_j^*)}\, dt \\
&+ \sum_{\ell=1}^{k} k(t; X_\ell^*) \sum_{\epsilon=1}^{r_{\delta\ell}} \frac{\tau^*(q_{\delta\ell\epsilon} + 1; K_1(X_\ell^*;t) \mid X_\ell^*)}{\tau^*(q_{\delta\ell\epsilon}; K_1(X_\ell^*;t) \mid X_\ell^*)}\, \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} D\,\psi^*(K_1(y;t) \mid y)\, \tilde{\mu}_0^*(dy)\right\}\right] \\
&\qquad \times \prod_{j=1}^{k} \mathbb{E}\left[\exp\left\{-D\,\psi^*(K_1(X_j^*;t) \mid X_j^*)\, V_j\right\}\right] \prod_{d=1}^{D} \prod_{h=1}^{r_{dj}} \frac{\tau^*(q_{djh}; K_1(X_j^*;t) \mid X_j^*)}{\tau^*(q_{djh}; 0 \mid X_j^*)}\, dt
\end{aligned}
$$

As for the first component, by the posterior characterization of the random measure $\tilde{\mu}_0$, and exploiting the identity in (2.10) with $f(y) = D\,\psi^*(K_1(y;t) \mid y)$, one obtains

$$
\begin{aligned}
&= \int_{\mathbb{X}} k(t;x)\, \tau^*(1; K_1(x;t) \mid x)\, \tau_0^*\big(1; D\,\psi^*(K_1(x;t) \mid x) \mid x\big)\, \theta P_0(dx) \\
&\qquad \times \exp\left\{-\int_{\mathbb{X}} \psi_0^*\big(D\,\psi^*(K_1(y;t) \mid y) \mid y\big)\, \theta P_0(dy)\right\} \\
&\qquad\qquad \times \prod_{j=1}^{k} \frac{\tau_0^*\big(r_j; D\,\psi^*(K_1(X_j^*;t) \mid X_j^*) \mid X_j^*\big)}{\tau_0^*(r_j; 0 \mid X_j^*)} \prod_{d=1}^{D} \prod_{h=1}^{r_{dj}} \frac{\tau^*(q_{djh}; K_1(X_j^*;t) \mid X_j^*)}{\tau^*(q_{djh}; 0 \mid X_j^*)}\, dt
\end{aligned}
$$

while for the second and third component, the expectation with respect to $\tilde{\mu}_0$ and $V_1, \dots, V_k$ reads

$$
\begin{aligned}
&+ \sum_{\ell=1}^{k} k(t; X_\ell^*)\, \tau^*(1; K_1(X_\ell^*; t) \mid X_\ell^*) \exp\left\{ - \int_{\mathbb{X}} \psi_0^*\big(D\,\psi^*(K_1(y;t) \mid y) \mid y\big)\, \theta P_0(dy) \right\} \\
&\qquad \times \frac{\tau_0^*\big(r_\ell + 1; D\,\psi^*(K_1(X_\ell^*; t) \mid X_\ell^*) \mid X_\ell^*\big)}{\tau_0^*(r_\ell; 0 \mid X_\ell^*)} \prod_{j \neq \ell} \frac{\tau_0^*\big(r_j; D\,\psi^*(K_1(X_j^*; t) \mid X_j^*) \mid X_j^*\big)}{\tau_0^*(r_j; 0 \mid X_j^*)} \\
&\qquad\qquad \times \prod_{d=1}^{D} \prod_{j=1}^{k} \prod_{h=1}^{r_{dj}} \frac{\tau^*(q_{djh}; K_1(X_j^*; t) \mid X_j^*)}{\tau^*(q_{djh}; 0 \mid X_j^*)}\, dt \\
&+ \sum_{\ell=1}^{k} k(t; X_\ell^*) \sum_{\epsilon=1}^{r_{\delta\ell}} \frac{\tau^*(q_{\delta\ell\epsilon} + 1; K_1(X_\ell^*; t) \mid X_\ell^*)}{\tau^*(q_{\delta\ell\epsilon}; K_1(X_\ell^*; t) \mid X_\ell^*)} \exp\left\{ - \int_{\mathbb{X}} \psi_0^*\big(D\,\psi^*(K_1(y;t) \mid y) \mid y\big)\, \theta P_0(dy) \right\} \\
&\qquad \times \prod_{j=1}^{k} \frac{\tau_0^*\big(r_j; D\,\psi^*(K_1(X_j^*; t) \mid X_j^*) \mid X_j^*\big)}{\tau_0^*(r_j; 0 \mid X_j^*)} \prod_{d=1}^{D} \prod_{h=1}^{r_{dj}} \frac{\tau^*(q_{djh}; K_1(X_j^*; t) \mid X_j^*)}{\tau^*(q_{djh}; 0 \mid X_j^*)}\, dt.
\end{aligned}
$$

In conclusion, by collecting and rearranging the terms in the expressions above,

$$
\begin{aligned}
\mathbb{E}\big[\tilde{p}(dt, \delta) \mid \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{X}, \boldsymbol{Z}\big] &= \exp\left\{ - \int_{\mathbb{X}} \psi_0^*\big(D\,\psi^*(K_1(y;t) \mid y) \mid y\big)\, \theta P_0(dy) \right\} \\
&\quad \times \prod_{j=1}^{k} \frac{\tau_0^*\big(r_j; D\,\psi^*(K_1(X_j^*; t) \mid X_j^*) \mid X_j^*\big)}{\tau_0^*(r_j; 0 \mid X_j^*)} \prod_{d=1}^{D} \prod_{h=1}^{r_{dj}} \frac{\tau^*(q_{djh}; K_1(X_j^*; t) \mid X_j^*)}{\tau^*(q_{djh}; 0 \mid X_j^*)} \\
&\quad \times \Bigg(\int_{\mathbb{X}} k(t; x)\, \tau^*(1; K_1(x; t) \mid x)\, \tau_0^*\big(1; D\,\psi^*(K_1(x; t) \mid x) \mid x\big)\, \theta P_0(dx) \\
&\qquad + \sum_{j=1}^{k} k(t; X_j^*)\, \tau^*(1; K_1(X_j^*; t) \mid X_j^*)) \frac{\tau_0^*\big(r_j + 1; D\,\psi^*(K_1(X_j^*; t) \mid X_j^*) \mid X_j^*\big)}{\tau_0^*\big(r_j; D\,\psi^*(K_1(X_j^*; t) \mid X_j^*) \mid X_j^*\big)} \\
&\qquad\qquad + \sum_{j=1}^{k} k(t; X_j^*) \sum_{h=1}^{r_{\delta j}} \frac{\tau^*(q_{\delta jh} + 1; K_1(X_j^*; t))}{\tau^*(q_{\delta jh}; K_1(X_j^*; t))} \Bigg)\, dt. \qquad \text{(A.2)}
\end{aligned}
$$

Note that the factors in the first two lines of this expression coincide with the posterior estimate of the survival function obtained in Proposition 3.7.

Proof of (A.1) from (A.2) The posterior estimate of cause-specific incidence functions detailed in Proposition 3.8, and more explicitly in (A.2), coincides with the predictive distribution for the additional observation (T_{n+1}, Δ_{n+1}) described in (A.1), as discussed in Section 3.4. This fact can be shown by clarifying the relationships of the Laplace exponents and cumulants for the posterior random measures $\tilde{\mu}_1^*, \dots, \tilde{\mu}_D^*$ with the corresponding quantities for the random measures $\tilde{\mu}_1, \dots, \tilde{\mu}_D$, defined in (3.14) and (3.9), respectively. In particular, at the lower level of

the hierarchy, the Laplace exponents are related as

$$\psi^*(u \,|\, x) = \psi(u + K_n(x)) - \psi(K_n(x)),$$

or equivalently, $\psi^*(u \,|\, x) + \psi(K_n(x)) = \psi(u + K_n(x))$, while at the root of the hierarchy

$$\psi_0^*(u \,|\, x) = \psi_0(u + D\psi(K_n(x))) - \psi_0(D\psi(K_n(x))).$$

Therefore, exploiting the expressions above and noticing that the updated kernel term is given by $K_{n+1}(x;t) = K_n(x) + K_1(x;t)$, the exponential term in (A.2) becomes

$$\begin{aligned}
&\exp\left\{-\int_{\mathbb{X}} \psi_0^*\big(D\psi^*(K_1(y;t) \mid y) \mid y\big)\, \theta P_0(dy)\right\} \\
&\qquad = \exp\left\{-\int_{\mathbb{X}} \left(\psi_0(D\psi^*(K_1(y;t) \mid y) + D\psi(K_n(y))) - \psi_0(D\psi(K_n(y)))\right) \theta P_0(dy)\right\} \\
&\qquad = \exp\left\{-\int_{\mathbb{X}} \left(\psi_0(D\psi(K_1(y;t) + K_n(y))) - \psi_0(D\psi(K_n(y)))\right) \theta P_0(dy)\right\} \\
&\qquad = \exp\left\{-\int_{\mathbb{X}} \left(\psi_0(D\psi(K_{n+1}(y;t))) - \psi_0(D\psi(K_n(y)))\right) \theta P_0(dy)\right\}.
\end{aligned}$$

On the other hand, for the cumulants one has

$$\tau^*(m; u \mid x) = \tau(m; u + K_n(x)), \qquad \tau_0^*(m; u \mid x) = \tau_0(m; u + D\psi(K_n(x))).$$

Therefore, the quantities appearing in (A.2) can be rewritten as

$$\begin{aligned}
\tau_0^*\big(m; D\psi^*(K_1(x;t) \mid x) \mid x\big) &= \tau_0\big(m; D\psi^*(K_1(x;t) \mid x) + D\psi(K_n(x))\big) \\
&= \tau_0\big(m; D\psi(K_1(x;t) + K_n(x))\big) = \tau_0\big(m; D\psi(K_{n+1}(x;t))\big),
\end{aligned}$$

while $\tau_0^*(m; 0 \mid x) = \tau_0(m; D\psi(K_n(x)))$; moreover,

$$\tau^*(m; K_1(x;t) \mid x) = \tau(m; K_1(x;t) + K_n(x)) = \tau(m; K_{n+1}(x;t)),$$

while $\tau^*(m; 0 \mid x) = \tau^*(m; K_n(x))$. In conclusion, the expression in (A.1) can be easily obtained from (A.2) using the identities above.

A.3 Proofs of Chapter 4

Proof of Proposition 4.1 Consider the augmented likelihood function in (4.8),

$$\mathcal{L}(\tilde{\mu}; \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\xi}, \boldsymbol{\beta}) = R(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})\, Q(\boldsymbol{X}, \boldsymbol{\xi}) \times \int_{(\mathbb{R}^+)^n} \exp\left\{ - \int_{\mathbb{X}\times\mathbb{B}} \left(\sum_{i=1}^{n} u_i\, k(X_i, \xi) \right) \tilde{\mu}(d\xi, d\beta) \right\} \prod_{j=1}^{k} \tilde{\mu}(d\xi_j^*, d\beta_j^*)^{n_j}\, d\boldsymbol{u},$$

where the kernel quantities $R(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})$ and $Q(\boldsymbol{X}, \boldsymbol{\xi})$ are defined in (4.9). The joint marginal distribution of responses $\boldsymbol{Y}$, latent locations $\boldsymbol{\xi}$ and latent regression parameters $\boldsymbol{\beta}$ is obtained by marginalizing the augmented likelihood with respect to the random measure $\tilde{\mu}$, that is

$$\mathbb{P}(\boldsymbol{Y}, \boldsymbol{\xi}, \boldsymbol{\beta} \mid \boldsymbol{X}) = \mathbb{E}\left[\mathcal{L}(\tilde{\mu}; \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\xi}, \boldsymbol{\beta})\right] = R(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})\, Q(\boldsymbol{X}, \boldsymbol{\xi}) \times \int_{(\mathbb{R}^+)^n} \mathbb{E}\left[\exp\left\{ - \int_{\mathbb{X}\times\mathbb{B}} \left(\sum_{i=1}^{n} u_i\, k(X_i, \xi) \right) \tilde{\mu}(d\xi, d\beta) \right\} \prod_{j=1}^{k} \tilde{\mu}(d\xi_j^*, d\beta_j^*)^{n_j} \right] d\boldsymbol{u}.$$

The distribution of the random measure $\tilde{\mu}$ is absolutely continuous with respect to the distribution of a σ-stable completely random measure $\tilde{\mu}_\sigma$, with Radon-Nikodym derivative (4.7),

$$\begin{aligned} d\mathcal{L}(\tilde{\mu}) &= \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)}\, \tilde{\mu}_\sigma(\mathbb{X}\times\mathbb{B})^{-\theta}\, d\mathcal{L}(\tilde{\mu}_\sigma) \\ &= \frac{\sigma}{\Gamma(\theta/\sigma)} \int_{\mathbb{R}^+} u_0^{\theta-1} \exp\left\{ - \int_{\mathbb{X}\times\mathbb{B}} u_0\, \tilde{\mu}_\sigma(d\xi, d\beta) \right\} du_0\, d\mathcal{L}(\tilde{\mu}_\sigma), \end{aligned}$$

where a standard analytical manipulation based on the density of a gamma random variable is used in the second line. Therefore, the expectation with respect to $\tilde{\mu}$ can be rewritten as an expectation with respect to $\tilde{\mu}_\sigma$ as

$$\mathbb{P}(\boldsymbol{Y}, \boldsymbol{\xi}, \boldsymbol{\beta} \mid \boldsymbol{X}) = R(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})\, Q(\boldsymbol{X}, \boldsymbol{\xi}) \times \frac{\sigma}{\Gamma(\theta/\sigma)} \int_{(\mathbb{R}^+)^{n+1}} \mathbb{E}\left[\exp\left\{ - \int_{\mathbb{X}\times\mathbb{B}} K_n(\xi, \boldsymbol{u})\, \tilde{\mu}_\sigma(d\xi, d\beta) \right\} \prod_{j=1}^{k} \tilde{\mu}_\sigma(d\xi_j^*, d\beta_j^*)^{n_j} \right] u_0^{\theta-1}\, d\boldsymbol{u},$$

where the following function is defined:

$$K_n(\xi, \boldsymbol{u}) = K_n(\xi, \boldsymbol{u}; \boldsymbol{X}) = u_0 + \sum_{i=1}^{n} u_i\, k(X_i, \xi).$$

The σ-stable completely random measure $\tilde{\mu}_\sigma$ is characterized by the Lévy intensity (4.6), for which the Laplace exponent and cumulants in (2.7) take the forms

$$\psi(u) = u^\sigma, \qquad \tau(m;u) = \sigma\,(1-\sigma)_{m-1}\; u^{\sigma-m}, \tag{A.3}$$

where $(a)_m$ denotes the ascending factorial; therefore, exploiting the identity in (2.10) with $\alpha(d\xi, d\beta) = P_0(d\xi)\,Q_0(d\beta)$ and $f(\xi,\beta) = K_n(\xi,\boldsymbol{u})$ one obtains

$$\begin{aligned}\mathbb{P}(\boldsymbol{Y},\boldsymbol{\xi},\boldsymbol{\beta}\mid\boldsymbol{X}) &= R(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta})\,Q(\boldsymbol{X},\boldsymbol{\xi})\;\sigma^k\prod_{j=1}^{k}(1-\sigma)_{n_j-1}\,P_0(d\xi_j^*)\,Q_0(d\beta_j^*)\\ &\times\frac{\sigma}{\Gamma(\theta/\sigma)}\int_{(\mathbb{R}^+)^{n+1}} u_0^{\theta-1}\,\exp\left\{-\int_{\mathbb{X}}K_n(\xi,\boldsymbol{u})^\sigma P_0(d\xi)\right\}\prod_{j=1}^{k}K_n(\xi_j^*,\boldsymbol{u})^{\sigma-n_j}\,d\boldsymbol{u}.\end{aligned} \tag{A.4}$$

Considering the change of integration variables $\boldsymbol{u}\mapsto(t,\boldsymbol{v})$ such that

$$t=\sum_{i=0}^{n}u_i, \qquad v_i=\frac{u_i}{t}, \quad i=1,\dots n-1, \qquad d\boldsymbol{u}=t^n\,dt\,d\boldsymbol{v},$$

the expression of the marginal distribution can be rewritten as

$$\begin{aligned}&= R(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta})\,Q(\boldsymbol{X},\boldsymbol{\xi})\;\sigma^k\prod_{j=1}^{k}(1-\sigma)_{n_j-1}\,P_0(d\xi_j^*)\,Q_0(d\beta_j^*)\\ &\quad\times\frac{\sigma}{\Gamma(\theta/\sigma)}\int_{\Delta^n}v_0^{\theta-1}\int_{\mathbb{R}^+}t^{\theta+k\sigma}\,\exp\left\{-t^\sigma\int_{\mathbb{X}}K_n(\xi,\boldsymbol{v})^\sigma P_0(d\xi)\right\}dt\prod_{j=1}^{k}K_n(\xi_j^*,\boldsymbol{v})^{\sigma-n_j}\,d\boldsymbol{v};\end{aligned}$$

computing the integral with respect to t, one gets

$$\begin{aligned}&= R(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta})\,Q(\boldsymbol{X},\boldsymbol{\xi})\;\frac{\sigma^k\,\Gamma(k+\theta/\sigma)}{\Gamma(\theta/\sigma)}\prod_{j=1}^{k}(1-\sigma)_{n_j-1}\,P_0(d\xi_j^*)\,Q_0(d\beta_j^*)\\ &\qquad\times\int_{\Delta^n}\left(\int_{\mathbb{X}}K_n(\xi,\boldsymbol{v})^\sigma P_0(d\xi)\right)^{-(k+\theta/\sigma)}\prod_{j=1}^{k}K_n(\xi_j^*,\boldsymbol{v})^{\sigma-n_j}\,v_0^{\theta-1}d\boldsymbol{v}.\end{aligned}$$

The result is obtained by rearranging the terms and defining the following function:

$$H_n(\boldsymbol{v}) := \left(\int_{\mathbb{X}}K_n(\xi,\boldsymbol{v})^\sigma P_0(d\xi)\right)^{1/\sigma}.$$

Proof of Proposition 4.2 The probability distribution of the induced random partition is obtained from the joint marginal distribution in Proposition 4.1 upon marginalization of the responses and of the distinct values assumed by latent locations and regression parameters. In

particular, the marginalization of the responses $\boldsymbol{Y}$ reduces to

$$\int_{\mathbb{R}^n} R(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})\, d\boldsymbol{Y} = \prod_{i=1}^{n} \int_{\mathbb{R}} \phi(Y_i; X_i, \beta_i)\, dY_i = 1,$$

where the last equivalence follows from the assumption of ϕ being a probability density function; therefore, the marginal distribution of latent locations $\boldsymbol{\xi}$ and regression parameters $\boldsymbol{\beta}$ becomes

$$\mathbb{P}(\boldsymbol{\xi}, \boldsymbol{\beta} \mid \boldsymbol{X}) = Q(\boldsymbol{X}, \boldsymbol{\xi})\, \frac{\prod_{\ell=1}^{k-1}(\theta + \ell\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1}\, P_0(d\xi_j^*)\, Q_0(d\beta_j^*) \\ \times \int_{\Delta^n} H_n(\boldsymbol{v})^{-(\theta+n)} \prod_{j=1}^{k} \left(\frac{K_n(\xi_j^*, \boldsymbol{v})}{H_n(\boldsymbol{v})} \right)^{\sigma - n_j} f_{\theta,1,\dots,1}(\boldsymbol{v})\, d\boldsymbol{v},$$

where $f_{\theta,1,\dots,1}(\boldsymbol{v})$ is the density function of the Dirichlet distribution with parameters $(\theta, 1, \dots, 1)$. Note that the expression above coincides with the joint distribution of the random partition $\boldsymbol{\Pi}$ and of the distinct values $\boldsymbol{\xi}^*$ and $\boldsymbol{\beta}^*$ assumed by the latent locations and regression parameters. The further marginalization with respect to the distinct regression parameters is straightforward; hence, the marginal distribution of the random partition and distinct latent locations is

$$\mathbb{P}(\boldsymbol{\Pi}, \boldsymbol{\xi}^* \mid \boldsymbol{X}) = \frac{\prod_{\ell=1}^{k-1}(\theta + \ell\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1} \\ \int_{\Delta^n} H_n(\boldsymbol{v})^{-\theta} \prod_{j=1}^{k} \left(\frac{K_n(\xi_j^*, \boldsymbol{v})^{\sigma}}{H_n(\boldsymbol{v})^{\sigma}} \prod_{i:\, \xi_i = \xi_j^*} \frac{k(X_i, \xi_j^*)}{K_n(\xi_j^*, \boldsymbol{v})}\, P_0(d\xi_j^*) \right) f_{\theta,1,\dots,1}(\boldsymbol{v})\, d\boldsymbol{v},$$

where the product of kernels $Q(\boldsymbol{X}, \boldsymbol{\xi})$ is made explicit and terms rearranged. In conclusion, upon marginalization of the distinct locations, the quantity in the parentheses above turns into

$$S_j(\boldsymbol{v}; \boldsymbol{X}) := \frac{\displaystyle\int_{\mathbb{X}} \prod_{\{i:\, A_i = j\}} \frac{k(X_i, \xi)}{K_n(\xi, \boldsymbol{v})}\, K_n(\xi, \boldsymbol{v})^{\sigma}\, P_0(d\xi)}{\displaystyle\int_{\mathbb{X}} K_n(\xi, \boldsymbol{v})^{\sigma}\, P_0(d\xi)}, \qquad j = 1, \dots, k,$$

where $\boldsymbol{A}$ is the vector of allocation variables, that is, $A_i = j$ if and only if $\xi_i = \xi_j^*$. Therefore, the marginal distribution of the induced random partition can be written as

$$\mathbb{P}(\boldsymbol{\Pi} \mid \boldsymbol{X}) = \frac{\prod_{\ell=1}^{k-1}(\theta + h\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1} \int_{\Delta^n} H_n(\boldsymbol{v})^{-\theta} \prod_{j=1}^{k} S_j(\boldsymbol{v}; \boldsymbol{X})\, f_{\theta,1,\dots,1}(\boldsymbol{v})\, d\boldsymbol{v}.$$

Proof of Proposition 4.4 The posterior distribution of the common random measure $\tilde{\mu}$, given the responses $\boldsymbol{Y}$, latent locations $\boldsymbol{\xi}$ and regression parameters $\boldsymbol{\beta}$, can be characterized through

its conditional Laplace transform,

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}} h(\xi,\beta)\,\tilde{\mu}(d\xi,d\beta)\right\}\,\middle|\,\boldsymbol{Y},\boldsymbol{\xi},\boldsymbol{\beta}\right] = \frac{\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}} h(\xi,\beta)\,\tilde{\mu}(d\xi,d\beta)\right\}\mathcal{L}(\tilde{\mu};\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\xi},\boldsymbol{\beta})\right]}{\mathbb{E}\left[\mathcal{L}(\tilde{\mu};\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\xi},\boldsymbol{\beta})\right]},$$

where $h\colon \mathbb{X}\times\mathbb{B}\mapsto\mathbb{R}^+$ is any non-negative measurable function. The quantity at the denominator is the joint marginal distribution in Proposition 4.1, while the expectation at the numerator can be computed similarly; specifically, from the augmented likelihood in (4.8),

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}} h(\xi,\beta)\,\tilde{\mu}(d\xi,d\beta)\right\}\mathcal{L}(\tilde{\mu};\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\xi},\boldsymbol{\beta})\right] = R(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta})\,Q(\boldsymbol{X},\boldsymbol{\xi})$$
$$\times\int_{(\mathbb{R}^+)^n}\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}}\left(h(\xi,\beta)+\sum_{i=1}^{n}u_i\,k(X_i,\xi)\right)\tilde{\mu}(d\xi,d\beta)\right\}\prod_{j=1}^{k}\tilde{\mu}(d\xi_j^*,d\beta_j^*)^{n_j}\right]d\boldsymbol{u},$$

where the kernel quantities $R(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta})$ and $Q(\boldsymbol{X},\boldsymbol{\xi})$ are defined in (4.9). The distribution of the random measure $\tilde{\mu}$ is absolutely continuous with respect to the distribution of a σ-stable completely random measure $\tilde{\mu}_\sigma$, with Radon-Nikodym derivative (4.7),

$$d\mathcal{L}(\tilde{\mu}) = \frac{\sigma}{\Gamma(\theta/\sigma)}\int_{\mathbb{R}^+} u_0^{\theta-1}\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}} u_0\,\tilde{\mu}_\sigma(d\xi,d\beta)\right\}du_0\,d\mathcal{L}(\tilde{\mu}_\sigma);$$

therefore, the expectation with respect to $\tilde{\mu}$ can be rewritten as an expectation with respect to the random measure $\tilde{\mu}_\sigma$ as

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}} h(\xi,\beta)\,\tilde{\mu}(d\xi,d\beta)\right\}\mathcal{L}(\tilde{\mu};\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\xi},\boldsymbol{\beta})\right] = R(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta})\,Q(\boldsymbol{X},\boldsymbol{\xi})\,\frac{\sigma}{\Gamma(\theta/\sigma)}$$
$$\times\int_{(\mathbb{R}^+)^{n+1}}\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}}(h(\xi,\beta)+K_n(\xi,\boldsymbol{u}))\,\tilde{\mu}_\sigma(d\xi,d\beta)\right\}\prod_{j=1}^{k}\tilde{\mu}_\sigma(d\xi_j^*,d\beta_j^*)^{n_j}\right]u_0^{\theta-1}\,d\boldsymbol{u},$$

where the function $K_n(\xi,\boldsymbol{u}) = K_n(\xi,\boldsymbol{u};\boldsymbol{X})$ is defined as in the proof of Proposition 4.1. This expectation can be computed exploiting the identity in (2.10) with $\alpha(d\xi,d\beta) = P_0(d\xi)\,Q_0(d\beta)$

and $f(\xi,\beta) = h(\xi,\beta) + K_n(\xi,\boldsymbol{u})$,

$$= R(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\beta})\, Q(\boldsymbol{X},\boldsymbol{\xi})\, \frac{\sigma^{k+1}}{\Gamma(\theta/\sigma)} \prod_{j=1}^{k} (1-\sigma)_{n_j-1}\, P_0(d\xi_j^*)\, Q_0(d\beta_j^*)$$
$$\times \int_{(\mathbb{R}^+)^{n+1}} u_0^{\theta-1} \exp\left\{ - \int_{\mathbb{X}\times\mathbb{B}} (h(\xi,\beta) + K_n(\xi,\boldsymbol{u}))^{\sigma}\, P_0(d\xi)\, Q_0(d\beta) \right\}$$
$$\times \prod_{j=1}^{k} \big(h(\xi_j^*,\beta_j^*) + K_n(\xi_j^*,\boldsymbol{u})\big)^{\sigma-n_j}\, d\boldsymbol{u};$$

where the expressions in (A.3) for the Laplace exponent and cumulants of the σ-stable completely random measure $\tilde{\mu}_\sigma$ are substituted. Therefore, considering the intermediate expression in (A.4) for the joint marginal distribution at the denominator, the conditional Laplace transform is

$$\mathbb{E}\left[\exp\left\{ - \int_{\mathbb{X}\times\mathbb{B}} h(\xi,\beta)\, \tilde{\mu}(d\xi, d\beta) \right\} \,\middle|\, \boldsymbol{Y}, \boldsymbol{\xi}, \boldsymbol{\beta} \right]$$
$$= \frac{\displaystyle\int_{(\mathbb{R}^+)^{n+1}} u_0^{\theta-1} \exp\left\{ - \int_{\mathbb{X}\times\mathbb{B}} (h(\xi,\beta) + K_n(\xi,\boldsymbol{u}))^{\sigma}\, P_0(d\xi)\, Q_0(d\beta) \right\} \prod_{j=1}^{k} \big(h(\xi_j^*,\beta_j^*) + K_n(\xi_j^*,\boldsymbol{u})\big)^{\sigma-n_j}\, d\boldsymbol{u}}{\displaystyle\int_{(\mathbb{R}^+)^{n+1}} u_0^{\theta-1} \exp\left\{ - \int_{\mathbb{X}} K_n(\xi,\boldsymbol{u})^{\sigma} P_0(d\xi) \right\} \prod_{j=1}^{k} K_n(\xi_j^*,\boldsymbol{u})^{\sigma-n_j}\, d\boldsymbol{u}}.$$

Note that this expression does not contain the responses $\boldsymbol{Y}$, and thus the posterior distribution of $\tilde{\mu}$ does not depend on them. Rearranging the terms of the integrand function at the numerator in order to match the those of the integrand at the denominator, one gets

$$= \int_{(\mathbb{R}^+)^{n+1}} \exp\left\{ - \int_{\mathbb{X}\times\mathbb{B}} \Big((h(\xi,\beta) + K_n(\xi,\boldsymbol{u}))^{\sigma} - K_n(\xi,\boldsymbol{u})^{\sigma} \Big) P_0(d\xi)\, Q_0(d\beta) \right\}$$
$$\times \prod_{j=1}^{k} \left(\frac{h(\xi_j^*,\beta_j^*) + K_n(\xi_j^*,\boldsymbol{u})}{K_n(\xi_j^*,\boldsymbol{u})} \right)^{\sigma-n_j} f(\boldsymbol{u})\, d\boldsymbol{u}, \quad \text{(A.5)}$$

where $f(\boldsymbol{u}) = f(\boldsymbol{u} \mid \boldsymbol{\xi})$ is a density function proportional to

$$u_0^{\theta-1} \exp\left\{ - \int_{\mathbb{X}} K_n(\xi,\boldsymbol{u})^{\sigma} P_0(d\xi) \right\} \prod_{j=1}^{k} K_n(\xi_j^*,\boldsymbol{u})^{\sigma-n_j}.$$

The exponential term in (A.5) coincides with Laplace transform of a generalized gamma completely random measure (see Section 2.7) evaluated at function $h(\xi, \beta)$, and thus can be rewritten as

$$\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}}\Big((h(\xi,\beta)+K_n(\xi,\boldsymbol{u}))^{\sigma}-K_n(\xi,\boldsymbol{u})^{\sigma}\Big)P_0(d\xi)\,Q_0(d\beta)\right\}$$
$$=\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}}\psi^*(h(\xi,\beta)\mid\xi)\,P_0(d\xi)\,Q_0(d\beta)\right\}=\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}}h(\xi,\beta)\,\tilde{\mu}^*(d\xi,d\beta)\right\}\right],$$

where $\tilde{\mu}^*$ is a generalized gamma completely random measure having Lévy intensity measure

$$\nu^*(ds,d\xi,d\beta)=\frac{\sigma}{\Gamma(1-\sigma)}\,s^{-1-\sigma}\exp\{-K_n(\xi,\boldsymbol{u})\,s\}\,ds\,P_0(d\xi)\,Q_0(d\beta).$$

Moreover, the product of ratios in (A.5) coincides with the product of Laplace transform of gamma random variables, computed at values $h(\xi_j^*,\beta_j^*)$, for $j=1,\ldots,k$, that is

$$\left(\frac{h(\xi_j^*,\beta_j^*)+K_n(\xi_j^*,\boldsymbol{u})}{K_n(\xi_j^*,\boldsymbol{u})}\right)^{\sigma-n_j}=\mathbb{E}\left[\exp\left\{-h(\xi_j^*,\beta_j^*)\,W_j\right\}\right],\qquad j=1,\ldots,k,$$

where each W_j is a gamma random variable having shape parameter $n_j-\sigma$ and rate parameter $K_n(\xi_j^*,\boldsymbol{u})$. As a result, the conditional Laplace transform can be written as

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}}h(\xi,\beta)\,\tilde{\mu}(d\xi,d\beta)\right\}\,\Big|\,\boldsymbol{Y},\boldsymbol{\xi},\boldsymbol{\beta}\right]$$
$$=\int_{(\mathbb{R}^+)^{n+1}}\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}}h(\xi,\beta)\,\tilde{\mu}^*(d\xi,d\beta)\right\}\right]\prod_{j=1}^{k}\mathbb{E}\left[\exp\left\{-h(\xi_j^*,\beta_j^*)\,W_j\right\}\right]f(\boldsymbol{u})\,d\boldsymbol{u}$$
$$=\int_{(\mathbb{R}^+)^{n+1}}\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}}h(\xi,\beta)\,\tilde{\mu}^*(d\xi,d\beta)-\sum_{j=1}^{k}h(\xi_j^*,\beta_j^*)\,W_j\right\}\right]f(\boldsymbol{u})\,d\boldsymbol{u};$$

therefore, the posterior distribution of the common random measure $\tilde{\mu}$ can be regarded as a mixture of completely random measures, having Lévy intensities depending on the mixing latent parameters vector $\boldsymbol{U}$, which in turn is distributed according to $f(\boldsymbol{u})$. By further conditioning on such auxiliary vector,

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}}h(\xi,\beta)\,\tilde{\mu}(d\xi,d\beta)\right\}\,\Big|\,\boldsymbol{Y},\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{U}\right]$$
$$=\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}}h(\xi,\beta)\,\tilde{\mu}^*(d\xi,d\beta)-\sum_{j=1}^{k}h(\xi_j^*,\beta_j^*)\,W_j\right\}\right];$$

which highlights the structure of the conditional posterior distribution of the common random measure $\tilde{\mu}$ as the sum of the non-homogeneous generalized gamma completely random measure

$\tilde{\mu}^*$ and the gamma-distributed random jumps $W_1, \ldots, W_k$ at fixed locations $(\xi_1^*, \beta_1^*), \ldots, (\xi_k^*, \beta_k^*)$; moreover, these components are mutually independent. The proof is concluded by a sequence of reparameterizations.

(1) Consider the change of variables $\boldsymbol{u} \mapsto (z, \boldsymbol{v})$ such that

$$z = \sum_{i=0}^{n} u_i, \qquad v_i = \frac{u_i}{z}, \quad i = 1, \ldots n-1, \qquad d\boldsymbol{u} = z^n \, dz \, d\boldsymbol{v};$$

the joint density function of Z and $\boldsymbol{V}$ is supported on the product space $\mathbb{R}^+ \times \Delta^n$ and proportional to

$$f(z, \boldsymbol{v}) \propto z^{k\sigma+\theta-1} \, v_0^{\theta-1} \, \exp\left\{-z^\sigma \int_{\mathbb{X}} K_n(\xi, \boldsymbol{v})^\sigma \, P_0(d\xi)\right\} \prod_{j=1}^{k} K_n(\xi_j^*, \boldsymbol{v})^{\sigma-n_j}.$$

(2) Consider the change of variables $t = z \, H_n(\boldsymbol{v})$, where the quantity $H_n(\boldsymbol{v})$ is defined in Proposition 4.1; the joint density function of T and $\boldsymbol{V}$ is again supported on the product space $\mathbb{R}^+ \times \Delta^n$ and proportional to

$$f(t, \boldsymbol{v}) \propto t^{k\sigma+\theta-1} \, e^{-t^\sigma} \, H_n(\boldsymbol{v})^{-(\theta+k\sigma)} \, v_0^{\theta-1} \prod_{j=1}^{k} K_n(\xi_j^*, \boldsymbol{v})^{\sigma-n_j}.$$

In conclusion, the auxiliary latent variables T and $\boldsymbol{V}$ are independent, the vector $\boldsymbol{V}$ has density function as in (4.12) and T has generalized gamma distribution. Moreover, the quantity $K_n(\xi, \boldsymbol{u})$ characterizing the random measure $\tilde{\mu}^*$ and the random jumps $W_1, \ldots, W_k$ is reparameterized as

$$K_n(\xi, \boldsymbol{u}) = z \, K_n(\xi, \boldsymbol{v}) = \frac{K_n(\xi, \boldsymbol{v})}{H_n(\boldsymbol{v})} \, t.$$

Proof of Proposition 4.5 The distribution of the random measure $\tilde{\mu}^*$ is characterized through its Laplace transform; specifically, given the latent locations $\boldsymbol{\xi}$ and regression parameters $\boldsymbol{\beta}$, and the latent variables $\boldsymbol{V}$ and T, it is a generalized gamma completely random measure,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}} h(\xi, \beta) \, \tilde{\mu}^*(d\xi, d\beta)\right\} \mid \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{V}, T\right] \\
&\quad = \exp\left\{-\int_{\mathbb{X}\times\mathbb{B}} \left(h(\xi, \beta) + \frac{K_n(\xi, \boldsymbol{V})}{H_n(\boldsymbol{V})} \, T\right)^\sigma P_0(d\xi) \, Q_0(d\beta) + \int_{\mathbb{X}} \left(\frac{K_n(\xi, \boldsymbol{V})}{H_n(\boldsymbol{V})} \, T\right)^\sigma P_0(d\xi)\right\}
\end{aligned}$$

which can be rewritten, exploiting the definition of $H_n(\boldsymbol{V})$, as

$$= \exp\left\{-\int_{\mathbb{X}\times\mathbb{B}} \left(h(\xi, \beta) + \frac{K_n(\xi, \boldsymbol{V})}{H_n(\boldsymbol{V})} \, T\right)^\sigma P_0(d\xi) \, Q_0(d\beta) + T^\sigma\right\}.$$

Therefore, upon marginalization with respect to the random variable T, the distribution of $\tilde{\mu}^*$ is characterized by the Laplace transform

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}} h(\xi,\beta)\,\tilde{\mu}^*(d\xi,d\beta)\right\} \mid \boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{V}\right] \\
&\quad= \int_{\mathbb{R}^+} \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}} h(\xi,\beta)\,\tilde{\mu}^*(d\xi,d\beta)\right\} \mid \boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{V},T=t\right] f_T(t\mid\boldsymbol{\xi})\,dt \\
&\quad= \frac{\sigma}{\Gamma(k+\theta/\sigma)}\int_{\mathbb{R}^+} t^{k\sigma+\theta-1}\,\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}}\left(h(\xi,\beta)+\frac{K_n(\xi,\boldsymbol{V})}{H_n(\boldsymbol{V})}\,t\right)^{\sigma} P_0(d\xi)\,Q_0(d\beta)\right\}dt.
\end{aligned}$$

The exponential term coincides with the Laplace transform of a σ-stable completely random measure, having Lévy intensity measure (4.6), evaluated at function $h(\xi,\beta)+t\,K_n(\xi,\boldsymbol{V})/H_n(\boldsymbol{V})$; therefore, one obtains

$$\begin{aligned}
&= \frac{\sigma}{\Gamma(k+\theta/\sigma)}\int_{\mathbb{R}^+} t^{k\sigma+\theta-1}\,\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}}\left(h(\xi,\beta)+\frac{K_n(\xi,\boldsymbol{V})}{H_n(\boldsymbol{V})}\,t\right)\tilde{\mu}_\sigma(d\xi,d\beta)\right\}\right]dt \\
&= \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}} h(\xi,\beta)\,\tilde{\mu}_\sigma(d\xi,d\beta)\right\}\frac{\sigma}{\Gamma(k+\theta/\sigma)}\right. \\
&\qquad\qquad\left.\int_{\mathbb{R}^+} t^{k\sigma+\theta-1}\,\exp\left\{-t\int_{\mathbb{X}\times\mathbb{B}}\frac{K_n(\xi,\boldsymbol{V})}{H_n(\boldsymbol{V})}\,\tilde{\mu}_\sigma(d\xi,d\beta)\right\}dt\right].
\end{aligned}$$

By computing the integral with respect to t,

$$= \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}} h(\xi,\beta)\,\tilde{\mu}_\sigma(d\xi,d\beta)\right\}\frac{\sigma\,\Gamma(k\sigma+\theta)}{\Gamma(k+\theta/\sigma)}\left(\int_{\mathbb{X}\times\mathbb{B}}\frac{K_n(\xi,\boldsymbol{V})}{H_n(\boldsymbol{V})}\,\tilde{\mu}_\sigma(d\xi,d\beta)\right)^{-(k\sigma+\theta)}\right];$$

in conclusion, the distribution of the random measure $\tilde{\mu}^*$ is absolutely continuous with respect to the distribution of $\tilde{\mu}_\sigma$, with Radon-Nikodym derivative

$$\frac{d\mathcal{L}(\tilde{\mu}^*)}{d\mathcal{L}(\tilde{\mu}_\sigma)}(m) = \frac{\sigma\Gamma(k\sigma+\theta)}{\Gamma(k+\theta/\sigma)}\left(\int_{\mathbb{X}\times\mathbb{B}}\frac{K_n(\xi,\boldsymbol{V})}{H_n(\boldsymbol{V})}\,m(d\xi,d\beta)\right)^{-(k\sigma+\theta)}.$$

Proof of Proposition 4.6 (marginal approach) The joint predictive distribution for the additional responses, latent locations and regression parameters is obtained from the joint marginal distribution in Proposition 4.1 as

$$\mathbb{P}(\boldsymbol{Y}^+,\boldsymbol{\xi}^+,\boldsymbol{\beta}^+\mid\boldsymbol{Y},\boldsymbol{\xi},\boldsymbol{\beta}) = \frac{\mathbb{P}(\boldsymbol{Y}^+,\boldsymbol{Y},\boldsymbol{\xi}^+,\boldsymbol{\xi},\boldsymbol{\beta}^+,\boldsymbol{\beta})}{\mathbb{P}(\boldsymbol{Y},\boldsymbol{\xi},\boldsymbol{\beta})},$$

where the numerator is the joint marginal distribution for the $n+m$ observations and corresponding latent variables. Specifically, the numerator takes the form

$$\begin{aligned}\mathbb{P}(\boldsymbol{Y}^+, \boldsymbol{Y}, \boldsymbol{\xi}^+, \boldsymbol{\xi}, \boldsymbol{\beta}^+, \boldsymbol{\beta}) &= R(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})\, R(\boldsymbol{Y}^+, \boldsymbol{X}^+, \boldsymbol{\beta}^+)\, Q(\boldsymbol{X}, \boldsymbol{\xi})\, Q(\boldsymbol{X}^+, \boldsymbol{\xi}^+) \\ &\times \frac{\prod_{\ell=1}^{k+h-1}(\theta+\ell\sigma)}{(\theta+1)_{n+m-1}} \prod_{j=1}^{k}(1-\sigma)_{n_j+m_j-1} \prod_{j=k+1}^{k+h}(1-\sigma)_{m_j-1} \prod_{j=1}^{k+h} P_0(d\xi_j^*)\, Q_0(d\beta_j^*) \\ &\times \int_{\Delta^{n+m}} H_{n+m}(\boldsymbol{z})^{-(\theta+n+m)} \prod_{j=1}^{k}\left(\frac{K_{n+m}(\xi_j^*, \boldsymbol{z})}{H_{n+m}(\boldsymbol{z})}\right)^{\sigma-n_j-m_j} \\ &\times \prod_{j=+1}^{k+h}\left(\frac{K_{n+m}(\xi_j^*, \boldsymbol{z})}{H_{n+m}(\boldsymbol{z})}\right)^{\sigma-m_j} f_{\theta,1,\ldots,1}(\boldsymbol{z})\, d\boldsymbol{z},\end{aligned}$$

where $f_{\theta,1,\ldots,1}(\boldsymbol{z})$ denotes the density function of the Dirichlet distribution over the $(n+m)$-dimensional simplex, with parameters $(\theta, 1, \ldots, 1)$, while the function $K_{n+m}(\xi, \boldsymbol{z})$ and $H_{n+m}(\boldsymbol{z})$ are the natural extensions of the function defined Proposition 4.1 for n observations. As a consequence, the joint predictive distribution is

$$\begin{aligned}\mathbb{P}(\boldsymbol{Y}^+, \boldsymbol{\xi}^+, \boldsymbol{\beta}^+ \mid \boldsymbol{Y}, \boldsymbol{\xi}, \boldsymbol{\beta}) &= R(\boldsymbol{Y}^+, \boldsymbol{X}^+, \boldsymbol{\beta}^+)\, Q(\boldsymbol{X}^+, \boldsymbol{\xi}^+) \\ &\times \frac{\prod_{\ell=0}^{h-1}(\theta+k\sigma+\ell\sigma)}{(\theta+n)_m} \prod_{j=1}^{k}(n_j-\sigma)_{m_j} \prod_{j=k+1}^{k+h}(1-\sigma)_{m_j-1}\, P_0(d\xi_j^*)\, Q_0(d\beta_j^*) \\ &\times \int_{\Delta^{n+m}} H_{n+m}(\boldsymbol{z})^{-(\theta+n+m)} \prod_{j=1}^{k}\left(\frac{K_{n+m}(\xi_j^*, \boldsymbol{z})}{H_{n+m}(\boldsymbol{z})}\right)^{\sigma-n_j-m_j} \\ &\times \prod_{j=+1}^{k+h}\left(\frac{K_{n+m}(\xi_j^*, \boldsymbol{z})}{H_{n+m}(\boldsymbol{z})}\right)^{\sigma-m_j} C(\boldsymbol{\xi})^{-1}\, f_{\theta,1,\ldots,1}(\boldsymbol{z})\, d\boldsymbol{z},\end{aligned}$$

where the quantity $C(\boldsymbol{\xi})$ is the normalizing constant for the density function of the auxiliary latent variables $\boldsymbol{V}$, that is

$$C(\boldsymbol{\xi}) = \int_{\Delta^n} H_n(\boldsymbol{v})^{-(\theta+n)} \prod_{j=1}^{k}\left(\frac{K_n(\xi_j^*, \boldsymbol{v})}{H_n(\boldsymbol{v})}\right)^{\sigma-n_j} f_{\theta,1,\ldots,1}(\boldsymbol{v})\, d\boldsymbol{v}.$$

Note that this expression for the predictive distribution does not depend on the responses $\boldsymbol{Y}$, which can thus be removed from the conditioning. Considering the change of variables $\boldsymbol{z} \mapsto (\boldsymbol{v}, \boldsymbol{w})$ such that

$$w_0 = \sum_{i=0}^{n} z_i, \qquad v_i = \frac{z_i}{w_0}, \quad i = 1, \ldots, n, \qquad w_i = z_{n+i}, \quad i = 1, \ldots, m-1,$$

the integral on the $(n+m)$-dimensional simplex can be split into a double integral

$$\int_{\Delta^n}\int_{\Delta^m} H_m^+(\boldsymbol{w};\boldsymbol{v})^{-(\theta+n+m)}\prod_{j=1}^{k}\left(\frac{K_m^+(\xi_j^*,\boldsymbol{w};\boldsymbol{v})}{H_m^+(\boldsymbol{w};\boldsymbol{v})}\right)^{\sigma-n_j-m_j} \times\prod_{j=+1}^{k+h}\left(\frac{K_m^+(\xi_j^*,\boldsymbol{w};\boldsymbol{v})}{H_m^+(\boldsymbol{w};\boldsymbol{v})}\right)^{\sigma-m_j} f_{\theta+n,1,\ldots,1}(\boldsymbol{w})\,d\boldsymbol{w}\,C(\boldsymbol{\xi})^{-1}\,f_{\theta,1,\ldots,1}(\boldsymbol{v})\,d\boldsymbol{v},$$

where the functions $K_{n+m}(\xi,\boldsymbol{z})$ and $H_{n+m}(\boldsymbol{z})$ are replaced by the functions

$$K_m^+(\xi,\boldsymbol{w};\boldsymbol{v}) := w_0\,K_n(\xi,\boldsymbol{v})+\sum_{i=1}^{m} w_i\,k(X_{n+i},\xi) = z_0+\sum_{i=1}^{k+h} z_i\,k(X_i,\xi) = K_{n+m}(\xi,\boldsymbol{z}),$$

$$H_m^+(\boldsymbol{w};\boldsymbol{v}) = \left(\int_{\mathbb{X}} K_m^+(\xi,\boldsymbol{w};\boldsymbol{v})^{\sigma}P_0(d\xi)\right)^{1/\sigma} = \left(\int_{\mathbb{X}} K_{n+m}(\xi,\boldsymbol{z})^{\sigma}P_0(d\xi)\right)^{1/\sigma} = H_{n+m}(\boldsymbol{z}).$$

Recalling that the density function of the auxiliary latent variables $\boldsymbol{V}$ in (4.12) is

$$f_{\boldsymbol{V}}(\boldsymbol{v}\mid\boldsymbol{\xi}) = C(\boldsymbol{\xi})^{-1}\,H_n(\boldsymbol{v})^{-(\theta+n)}\prod_{j=1}^{k}\left(\frac{K_n(\xi_j^*,\boldsymbol{v})}{H_n(\boldsymbol{v})}\right)^{\sigma-n_j} f_{\theta,1,\ldots,1}(\boldsymbol{v}),$$

the joint predictive distribution is rewritten as

$$\begin{aligned}\mathbb{P}(\boldsymbol{Y}^+,\boldsymbol{\xi}^+,\boldsymbol{\beta}^+\mid\boldsymbol{\xi},\boldsymbol{\beta}) &= R(\boldsymbol{Y}^+,\boldsymbol{X}^+,\boldsymbol{\beta}^+)\,Q(\boldsymbol{X}^+,\boldsymbol{\xi}^+)\\ &\quad\times\frac{\prod_{\ell=0}^{h-1}(\theta+k\sigma+\ell\sigma)}{(\theta+n)_m}\prod_{j=1}^{k}(n_j-\sigma)_{m_j}\prod_{j=k+1}^{k+h}(1-\sigma)_{m_j-1}\,P_0(d\xi_j^*)\,Q_0(d\beta_j^*)\\ \times\int_{\Delta^n} H_n(\boldsymbol{v})^{(\theta+n)}\prod_{j=1}^{k}\left(\frac{K_n(\xi_j^*,\boldsymbol{v})}{H_n(\boldsymbol{v})}\right)^{n_j-\sigma}&\int_{\Delta^m} H_m^+(\boldsymbol{w};\boldsymbol{v})^{-(\theta+n+m)}\prod_{j=1}^{k}\left(\frac{K_m^+(\xi_j^*,\boldsymbol{w};\boldsymbol{v})}{H_m^+(\boldsymbol{w};\boldsymbol{v})}\right)^{\sigma-n_j-m_j}\\ &\quad\times\prod_{j=+1}^{k+h}\left(\frac{K_m^+(\xi_j^*,\boldsymbol{w};\boldsymbol{v})}{H_m^+(\boldsymbol{w};\boldsymbol{v})}\right)^{\sigma-m_j} f_{\theta+n,1,\ldots,1}(\boldsymbol{u})\,d\boldsymbol{w}\,f_{\boldsymbol{V}}(\boldsymbol{v}\mid\boldsymbol{\xi})\,d\boldsymbol{v}.\end{aligned}$$

The result is obtained by conditioning on the auxiliary latent variables $\boldsymbol{V}$, playing the role of mixing parameters in the expression above.

Proof of Proposition 4.6 (conditional approach) Consider the augmented likelihood function in (4.8) for the additional responses, latent locations and regression parameters,

$$\begin{aligned}\mathcal{L}(\tilde{\mu};\boldsymbol{Y}^+,\boldsymbol{X}^+,\boldsymbol{\xi}^+,\boldsymbol{\beta}^+) &= R(\boldsymbol{Y}^+,\boldsymbol{X}^+,\boldsymbol{\beta}^+)\,Q(\boldsymbol{X}^+,\boldsymbol{\xi}^+)\\ &\quad\times\int_{(\mathbb{R}^+)^m}\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}}\left(\sum_{i=n+1}^{n+m}u_i\,k(X_i,\xi)\right)\tilde{\mu}(d\xi,d\beta)\right\}\prod_{j=1}^{k+h}\tilde{\mu}(d\xi_j^*,d\beta_j^*)^{m_j}\,d\boldsymbol{u},\end{aligned}$$

The joint predictive distribution is obtained by marginalizing the augmented likelihood with respect to the posterior distribution of the random measure $\tilde{\mu}$, given the previous latent locations, regression parameters and latent variables $\boldsymbol{V}$, that is

$$\begin{aligned}\mathbb{P}(\boldsymbol{Y}^+, \boldsymbol{\xi}^+, \boldsymbol{\beta}^+ \mid \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{V}) &= \mathbb{E}\left[\mathcal{L}(\tilde{\mu}; \boldsymbol{Y}^+, \boldsymbol{X}^+, \boldsymbol{\xi}^+, \boldsymbol{\beta}^+) \mid \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{V}\right] \\ &= R(\boldsymbol{Y}^+, \boldsymbol{X}^+, \boldsymbol{\beta}^+)\, Q(\boldsymbol{X}^+, \boldsymbol{\xi}^+) \int_{(\mathbb{R}^+)^m} \mathbb{E}\Bigg[\exp\left\{ - \int_{\mathbb{X}\times\mathbb{B}} \left(\sum_{i=n+1}^{n+m} u_i\, k(X_i, \xi) \right) \tilde{\mu}(d\xi, d\beta) \right\} \\ &\qquad \times \prod_{j=1}^{k+h} \tilde{\mu}(d\xi_j^*, d\beta_j^*)^{m_j} \;\Big|\; \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{V} \Bigg]\, d\boldsymbol{u}.\end{aligned}$$

The posterior distribution of $\tilde{\mu}$ is characterized in Proposition 4.4 as a completely random measure, conditionally on the auxiliary latent variable T, that is

$$\begin{aligned}&= R(\boldsymbol{Y}^+, \boldsymbol{X}^+, \boldsymbol{\beta}^+)\, Q(\boldsymbol{X}^+, \boldsymbol{\xi}^+)\, \frac{\sigma}{\Gamma(k + \theta/\sigma)} \int_{(\mathbb{R}^+)^{m+1}} t^{k\sigma+\theta-1}\, e^{-t^\sigma} \\ &\quad \times \mathbb{E}\left[\exp\left\{ - \int_{\mathbb{X}\times\mathbb{B}} \left(\sum_{i=n+1}^{n+m} u_i\, k(X_i, \xi) \right) \tilde{\mu}(d\xi, d\beta) \right\} \prod_{j=1}^{k+h} \tilde{\mu}(d\xi_j^*, d\beta_j^*)^{m_j} \;\Big|\; \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{V}, T \right] d\boldsymbol{u}\, dt.\end{aligned}$$

Substituting the specification of the posterior distribution of $\tilde{\mu}$ into the expression above, and rearranging the terms,

$$\begin{aligned}&= R(\boldsymbol{Y}^+, \boldsymbol{X}^+, \boldsymbol{\beta}^+)\, Q(\boldsymbol{X}^+, \boldsymbol{\xi}^+)\, \frac{\sigma}{\Gamma(k + \theta/\sigma)} \int_{\mathbb{R}^+} t^{k\sigma+\theta-1}\, e^{-t^\sigma} \\ &\quad \times \int_{(\mathbb{R}^+)^m} \mathbb{E}\Bigg[\exp\left\{ - \int_{\mathbb{X}\times\mathbb{B}} \left(\sum_{i=n+1}^{n+m} u_i\, k(X_i, \xi) \right) \tilde{\mu}^*(d\xi, d\beta) \right\} \prod_{j=k+1}^{k+h} \tilde{\mu}^*(d\xi_j^*, d\beta_j^*)^{m_j} \\ &\qquad \times \prod_{j=1}^{k} W_j^{m_j} \exp\left\{ - \left(\sum_{i=n+1}^{n+m} u_i\, k(X_i, \xi_j^*) \right) W_j \right\} \Bigg]\, d\boldsymbol{u}\, dt;\end{aligned}$$

by conditional independence of the random measure $\tilde{\mu}^*$ and random variables $W_1, \ldots, W_k$,

$$\begin{aligned}&= R(\boldsymbol{Y}^+, \boldsymbol{X}^+, \boldsymbol{\beta}^+)\, Q(\boldsymbol{X}^+, \boldsymbol{\xi}^+)\, \frac{\sigma}{\Gamma(k + \theta/\sigma)} \int_{\mathbb{R}^+} t^{k\sigma+\theta-1}\, e^{-t^\sigma} \\ &\quad \times \int_{(\mathbb{R}^+)^m} \mathbb{E}\left[\exp\left\{ - \int_{\mathbb{X}\times\mathbb{B}} \left(\sum_{i=n+1}^{n+m} u_i\, k(X_i, \xi) \right) \tilde{\mu}^*(d\xi, d\beta) \right\} \prod_{j=k+1}^{k+h} \tilde{\mu}^*(d\xi_j^*, d\beta_j^*)^{m_j} \right] \\ &\qquad \times \prod_{j=1}^{k} \mathbb{E}\left[W_j^{m_j} \exp\left\{ - \left(\sum_{i=n+1}^{n+m} u_i\, k(X_i, \xi_j^*) \right) W_j \right\} \right] d\boldsymbol{u}\, dt.\end{aligned}$$

The random measure $\tilde{\mu}^*$ is a generalized gamma completely random measure (see Section 2.7), for which the Laplace exponent and cumulants in (2.7) take the forms

$$\psi^*(u \mid \xi) = \left(u + t\,\frac{K_n(\xi, \boldsymbol{V})}{H_n(\boldsymbol{V})}\right)^{\sigma} - t^{\sigma}\left(\frac{K_n(\xi, \boldsymbol{V})}{H_n(\boldsymbol{V})}\right)^{\sigma},$$
$$\tau^*(m; u \mid \xi) = \sigma\,(1-\sigma)_{m-1}\,\left(u + t\,\frac{K_n(\xi, \boldsymbol{V})}{H_n(\boldsymbol{V})}\right)^{\sigma-m},$$

where $(a)_m$ denotes the ascending factorial; therefore, exploiting the identity in (2.10) with $\alpha(d\xi, d\beta) = P_0(d\xi)\,Q_0(d\beta)$ and $f(\xi, \beta) = \sum_i u_i\,k(X_i, \xi)$, the expectation with respect to $\tilde{\mu}^*$ is

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}\times\mathbb{B}}\left(\sum_{i=n+1}^{n+m} z_i\,k(X_i,\xi)\right)\tilde{\mu}^*(d\xi, d\beta)\right\}\prod_{j=k+1}^{k+h}\tilde{\mu}^*(d\xi_j^*, d\beta_j^*)^{m_j}\right] \\
&\qquad = \sigma^h\,\exp\left\{-\int_{\mathbb{X}}\left(t\,\frac{K_n(\xi,\boldsymbol{V})}{H_n(\boldsymbol{V})} + \sum_{i=n+1}^{n+m} u_i\,k(X_i,\xi)\right)^{\sigma} P_0(d\xi) + t^{\sigma}\right\} \\
&\qquad\qquad \times \prod_{j=k+1}^{k+h}(1-\sigma)_{m_j-1}\left(t\,\frac{K_n(\xi_j^*,\boldsymbol{V})}{H_n(\boldsymbol{V})} + \sum_{i=n+1}^{n+m} u_i\,k(X_i,\xi_j^*)\right)^{\sigma-m_j} P_0(d\xi_j^*)\,Q_0(d\beta_j^*).
\end{aligned}$$

Moreover, each random variable W_j has gamma distribution with shape parameter $n_j - \sigma$ and rate parameter $t\,K_n(\xi_j^*, \boldsymbol{V})/H_n(\boldsymbol{V})$, which entails

$$\begin{aligned}
&\mathbb{E}\left[W_j^{m_j}\exp\left\{-\left(\sum_{i=n+1}^{n+m} u_i\,k(X_i,\xi_j^*)\right)W_j\right\}\right] \\
&\qquad = (n_j-\sigma)_{m_j}\left(t\,\frac{K_n(\xi_j^*,\boldsymbol{V})}{H_n(\boldsymbol{V})}\right)^{n_j-\sigma}\left(t\,\frac{K_n(\xi_j^*,\boldsymbol{V})}{H_n(\boldsymbol{V})} + \sum_{i=n+1}^{n+m} u_i\,k(X_i,\xi_j^*)\right)^{\sigma-n_j-m_j}.
\end{aligned}$$

Therefore, the joint predictive distribution takes the form

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{Y}^+, \boldsymbol{\xi}^+, \boldsymbol{\beta}^+ \mid \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{V}) = R(\boldsymbol{Y}^+, \boldsymbol{X}^+, \boldsymbol{\beta}^+)\,Q(\boldsymbol{X}^+, \boldsymbol{\xi}^+)\,\frac{\sigma^{h+1}}{\Gamma(k+\theta/\sigma)} \\
&\quad \times \int_{\mathbb{R}^+} t^{k\sigma+\theta-1}\int_{(\mathbb{R}^+)^m}\exp\left\{-\int_{\mathbb{X}}\left(t\,\frac{K_n(\xi,\boldsymbol{V})}{H_n(\boldsymbol{V})} + \sum_{i=n+1}^{n+m} u_i\,k(X_i,\xi)\right)^{\sigma} P_0(d\xi)\right\} \\
&\quad \times \prod_{j=k+1}^{k+h}(1-\sigma)_{m_j-1}\left(t\,\frac{K_n(\xi_j^*,\boldsymbol{V})}{H_n(\boldsymbol{V})} + \sum_{i=n+1}^{n+m} u_i\,k(X_i,\xi_j^*)\right)^{\sigma-m_j} P_0(d\xi_j^*)\,Q_0(d\beta_j^*) \\
&\quad \times \prod_{j=1}^{k}(n_j-\sigma)_{m_j}\left(t\,\frac{K_n(\xi_j^*,\boldsymbol{V})}{H_n(\boldsymbol{V})}\right)^{n_j-\sigma}\left(t\,\frac{K_n(\xi_j^*,\boldsymbol{V})}{H_n(\boldsymbol{V})} + \sum_{i=n+1}^{n+m} u_i\,k(X_i,\xi_j^*)\right)^{\sigma-n_j-m_j} d\boldsymbol{u}\,dt.
\end{aligned}$$

Considering the change of integration variables $(t, \boldsymbol{u}) \mapsto (s, \boldsymbol{w})$ such that

$$s = \frac{t}{H_n(\boldsymbol{V})} + \sum_{i=1}^{n} u_i, \qquad w_i = \frac{u_i}{s}, \quad i = 1, \dots m-1, \qquad d\boldsymbol{u}\, dt = s^m\, H_n(\boldsymbol{V})\, ds\, d\boldsymbol{w},$$

and rearranging the terms, the expression above is rewritten as

$$\begin{aligned}
&= R(\boldsymbol{Y}^+, \boldsymbol{X}^+, \boldsymbol{\beta}^+)\, Q(\boldsymbol{X}^+, \boldsymbol{\xi}^+)\, H_n(\boldsymbol{V})^{\theta+n} \prod_{j=1}^{k} \left(\frac{K_n(\xi_j^*, \boldsymbol{V})}{H_n(\boldsymbol{V})} \right)^{n_j - \sigma} \\
&\quad \times \frac{\Gamma(\theta+n)}{\Gamma(\theta+n+m)}\, \frac{\sigma^{h+1}}{\Gamma(k+\theta/\sigma)} \prod_{j=1}^{k} (n_j - \sigma)_{m_j} \prod_{j=k+1}^{k+h} (1-\sigma)_{m_j - 1}\, P_0(d\xi_j^*)\, Q_0(d\beta_j^*) \\
&\quad \times \int_{\Delta^m} \int_{\mathbb{R}^+} s^{(k+h)\sigma + \theta} \exp\left\{ -s^\sigma \int_{\mathbb{X}} K_m^+(\xi, \boldsymbol{w}; \boldsymbol{V})^\sigma P_0(d\xi) \right\} ds \\
&\qquad \times \prod_{j=1}^{k} K_m^+(\xi_j^*, \boldsymbol{w}; \boldsymbol{V})^{\sigma - n_j - m_j} \prod_{j=k+1}^{k+h} K_m^+(\xi_j^*, \boldsymbol{w}; \boldsymbol{V})^{\sigma - m_j}\, f_{\theta+n,1,\dots,1}(\boldsymbol{w})\, d\boldsymbol{w},
\end{aligned}$$

where $f_{\theta+n,1,\dots,1}(\boldsymbol{w})$ denotes the density function of the Dirichlet distribution with parameters $(\theta+n, 1, \dots, 1)$, and the function $K_m^+(\xi, \boldsymbol{w}; \boldsymbol{V})$ is defined as

$$K_m^+(\xi, \boldsymbol{w}; \boldsymbol{V}) = w_0\, K_n(\xi, \boldsymbol{V}) + \sum_{i=1}^{m} w_i\, k(X_{n+i}, \xi).$$

Finally, computing the integral with respect to s, one gets

$$\begin{aligned}
&= R(\boldsymbol{Y}^+, \boldsymbol{X}^+, \boldsymbol{\beta}^+)\, Q(\boldsymbol{X}^+, \boldsymbol{\xi}^+)\, H_n(\boldsymbol{V})^{\theta+n} \prod_{j=1}^{k} \left(\frac{K_n(\xi_j^*, \boldsymbol{V})}{H_n(\boldsymbol{V})} \right)^{n_j - \sigma} \\
&\quad \times \frac{\prod_{\ell=0}^{h-1} (\theta + k\sigma + \ell\sigma)}{(\theta+n)_m} \prod_{j=1}^{k} (n_j - \sigma)_{m_j} \prod_{j=k+1}^{k+h} (1-\sigma)_{m_j - 1}\, P_0(d\xi_j^*)\, Q_0(d\beta_j^*) \\
&\quad \times \int_{\Delta^m} \left(\int_{\mathbb{X}} K_m^+(\xi, \boldsymbol{w}; \boldsymbol{V})^\sigma P_0(d\xi) \right)^{-(k+h+\theta/\sigma)} \prod_{j=1}^{k} K_m^+(\xi_j^*, \boldsymbol{w}; \boldsymbol{V})^{\sigma - n_j - m_j} \\
&\qquad \times \prod_{j=k+1}^{k+h} K_m^+(\xi_j^*, \boldsymbol{w}; \boldsymbol{V})^{\sigma - m_j}\, f_{\theta+n,1,\dots,1}(\boldsymbol{w})\, d\boldsymbol{w},
\end{aligned}$$

The result is obtained by rearranging the terms and defining the following function:

$$H_m^+(\boldsymbol{w}; \boldsymbol{V}) = \left(\int_{\mathbb{X}} K_m^+(\xi, \boldsymbol{w}; \boldsymbol{V})^\sigma P_0(d\xi) \right)^{1/\sigma}.$$

Proof of Corollary 4.7 The joint predictive distribution for the additional response Y_{n+1} and corresponding latent location ξ_{n+1} and regression parameter β_{n+1} is derived from Proposition 4.6 in case $m = 1$. The specific form of this distribution depends on the allocation of the additional observation within the random partition structure induced by the previous latent locations, that is, on the specific value assumed by the random variable ξ_{n+1}; indeed, two alternative cases can be identified, corresponding to $h = 0$ and $h = 1$.

(1) The latent location ξ_{n+1} displays a tie with values in $\boldsymbol{\xi}$, say $\xi_{n+1} = \xi_j^*$, that is, $h = 0$ and $m_j = 1$:

$$\mathbb{P}(Y_{n+1} \in dy, \xi_{n+1} = \xi_j^*, \beta_{n+1} = \beta_j^* \mid \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{V}) \propto \phi(y; X_{n+1}, \beta_j^*)\, dy\, k(X_{n+1}, \xi_j^*) \\ \times \frac{n_j - \sigma}{\theta + n} \int_0^1 \frac{H_1^+(w; \boldsymbol{V})^{-(\theta+n)}}{K_1^+(\xi_j^*, w; \boldsymbol{V})} \prod_{\ell=1}^{k} \left(\frac{K_1^+(\xi_\ell^*, w; \boldsymbol{V})}{H_1^+(w; \boldsymbol{V})} \right)^{\sigma - n_\ell} f_{1,\theta+n}(w)\, dw,$$

where $f_{1,\theta+n}(w)$ is the density function of a Beta distribution with parameters 1 and $\theta + n$; in this case, with reference to the general expression in Proposition 4.6, the integration variables (w_0, w_1) simplify to $(1 - w, w)$, and therefore

$$K_1^+(\xi, w; \boldsymbol{V}) = (1 - w)\, K_n(\xi, \boldsymbol{V}) + w\, k(X_{n+1}, \xi).$$

(2) The latent location ξ_{n+1} assumes a new value, not included in $\boldsymbol{\xi}$:

$$\mathbb{P}(Y_{n+1} \in dy, \xi_{n+1} \in d\xi, \beta_{n+1} \in d\beta \mid \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{V}) \propto \phi(y; X_{n+1}, \beta)\, dy\, k(X_{n+1}, \xi)\, P_0(d\xi)\, Q_0(d\beta) \\ \times \frac{\theta + k\sigma}{\theta + n} \int_0^1 \frac{H_1^+(w; \boldsymbol{V})^{-(\theta+n+\sigma)}}{K_1^+(\xi, w; \boldsymbol{V})^{1-\sigma}} \prod_{\ell=1}^{k} \left(\frac{K_1^+(\xi_\ell^*, w; \boldsymbol{V})}{H_1^+(w; \boldsymbol{V})} \right)^{\sigma - n_\ell} f_{1,\theta+n}(w)\, dw.$$

The result is obtained by unifying the two cases above into a unique expression and factorizing the common terms.

Proof of Lemma 4.8 (sketch) The predictive distribution of the latent variable W_{n+1}, given the previous locations $\boldsymbol{\xi}$ and latent variables $\boldsymbol{V}$, has density function proportional to (4.13),

$$\left(\frac{\theta + k\sigma}{\theta + n} \int_{\mathbb{X}} \frac{k(X_{n+1}, \xi)}{K_1^+(\xi, w; \boldsymbol{V})} \left(\frac{K_1^+(\xi, w; \boldsymbol{V})}{H_1^+(w; \boldsymbol{V})} \right)^{\sigma} P_0(d\xi) + \sum_{j=1}^{k} \frac{n_j - \sigma}{\theta + n} \frac{k(X_{n+1}, \xi_j^*)}{K_1^+(\xi_j^*, w; \boldsymbol{V})} \right) \\ \times H_1^+(w; \boldsymbol{V})^{-(\theta+n)} \prod_{j=1}^{k} \left(\frac{K_1^+(\xi_j^*, w; \boldsymbol{V})}{H^+(w; \boldsymbol{V})} \right)^{\sigma - n_j} f_{1,\theta+n}(w),$$

where $f_{1,\theta+n}(w)$ is the density function of a Beta distribution with parameters 1 and $\theta + n$. Consider the linear transformation $U = n\, W_{n+1}$; the random variable U takes non-negative values

and has density function proportional to

$$\left(\frac{\theta+k\sigma}{\theta+n}\int_{\mathbb{X}}\frac{k(X_{n+1},\xi)}{K_1^+(\xi,u/n;\boldsymbol{V})}\left(\frac{K_1^+(\xi,u/n;\boldsymbol{V})}{H_1^+(u/n;\boldsymbol{V})}\right)^{\sigma}P_0(d\xi)+\sum_{j=1}^{k}\frac{n_j-\sigma}{\theta+n}\frac{k(X_{n+1},\xi_j^*)}{K_1^+(\xi_j^*,u/n;\boldsymbol{V})}\right)$$
$$\times H_1^+(u/n;\boldsymbol{V})^{-(\theta+n)}\prod_{j=1}^{k}\left(\frac{K_1^+(\xi_j^*,u/n;\boldsymbol{V})}{H^+(u/n;\boldsymbol{V})}\right)^{\sigma-n_j}\left(1-\frac{u}{n}\right)^{\theta+n-1}.$$

In the limit for $n\to\infty$, exploiting standard first-order Taylor expansions,

$$\lim_{n\to\infty}\left(\frac{K_1^+(\xi,u/n;\boldsymbol{V})}{K_n(\xi,\boldsymbol{V})}\right)^{\alpha_n}=\lim_{n\to\infty}\left(1+\frac{u}{n}\frac{k(X_{n+1},\xi)}{K_n(\xi,\boldsymbol{V})}-\frac{u}{n}\right)^{\alpha_n}$$
$$=\lim_{n\to\infty}\exp\left\{u\,\frac{\alpha_n}{n}\,\frac{k(X_{n+1},\xi)}{K_n(\xi,\boldsymbol{V})}-u\,\frac{\alpha_n}{n}\right\}=\exp\left\{u\,\alpha\,\frac{k(X_{n+1},\xi)}{K_\infty(\xi,\boldsymbol{V})}-u\,\alpha\right\},$$

where $(\alpha_n)_n$ is a sequence of real values such that $\lim_n\alpha_n/n=\alpha$, and the function

$$K_\infty(\xi,\boldsymbol{V})=\lim_{n\to\infty}K_n(\xi,\boldsymbol{V})>0,\qquad\forall\,\xi\in\mathbb{X},$$

is assumed to be strictly positive for every $\xi\in\mathbb{X}$. Moreover, using similar arguments based on first-order Taylor expansions,

$$\lim_{n\to\infty}\left(\frac{H_1^+(u/n;\boldsymbol{V})}{H_n(\boldsymbol{V})}\right)^{\alpha_n}=\lim_{n\to\infty}\exp\left\{\frac{\alpha_n}{\sigma}\log\left(\int_{\mathbb{X}}K_1^+(\xi,u/n;\boldsymbol{V})^{\sigma}P_0(d\xi)\right)-\alpha_n\log H_n(\boldsymbol{V})\right\}$$
$$=\lim_{n\to\infty}\exp\left\{u\,\frac{\alpha_n}{n}\int_{\mathbb{X}}\frac{k(X_{n+1},\xi)}{K_n(\xi,\boldsymbol{V})}\left(\frac{K_n(\xi,\boldsymbol{V})}{H_n(\boldsymbol{V})}\right)^{\sigma}P_0(d\xi)-u\,\frac{\alpha_n}{n}\right\}$$
$$=\exp\left\{u\,\alpha\int_{\mathbb{X}}\frac{k(X_{n+1},\xi)}{K_\infty(\xi,\boldsymbol{V})}\left(\frac{K_\infty(\xi,\boldsymbol{V})}{H_\infty(\boldsymbol{V})}\right)^{\sigma}P_0(d\xi)-u\,\alpha\right\},$$

where $H_\infty(\boldsymbol{V})$ is defined as $H_n(\boldsymbol{V})$ replacing the integrand function $K_n(\xi,\boldsymbol{V})$ with its limit. Note that the terms in the first line of the density function above are characterized by bounded values for the sequence α_n, which implies $\alpha=0$: therefore, they can be disregarded, in the approximation for $n\to\infty$, as they do not depend on u. On the other hand, in the second line of the density function, for $n\to\infty$,

$$H_1^+(u/n;\boldsymbol{V})^{-(k\sigma+\theta)}$$
$$\approx H_n(\boldsymbol{V})^{-(k\sigma+\theta)}\exp\left\{u\,\frac{k\sigma+\theta}{n}-u\,\frac{k\sigma+\theta}{n}\int_{\mathbb{X}}\frac{k(X_{n+1},\xi)}{K_n(\xi,\boldsymbol{V})}\left(\frac{K_n(\xi,\boldsymbol{V})}{H_n(\boldsymbol{V})}\right)^{\sigma}P_0(d\xi)\right\},$$

while, for each $j=1,\ldots,k$ and $n\to\infty$,

$$K_1^+(\xi_j^*,u/n;\boldsymbol{V})^{\sigma-n_j}\approx K_n(\boldsymbol{V})^{\sigma-n_j}\exp\left\{u\,\frac{n_j-\sigma}{n}-u\,\frac{n_j-\sigma}{n}\,\frac{k(X_{n+1},\xi_j^*)}{K_n(\xi_j^*,\boldsymbol{V})}\right\}$$

Under the prior distribution $\mathcal{P}$ for the latent locations, and the induced partition structure, the number of partition groups is such that $\lim_n k/n = 0$, and therefore the term $H_1^+(u/n; \boldsymbol{V})^{-(k\sigma+\theta)}$ does not depend on u in the limit approximation. As a result, collecting the terms for which dependence on u is preserved in the limit, the approximated density function of the random variable U is proportional to

$$\exp\left\{-u\sum_{j=1}^{k}\frac{n_j}{n}\,\frac{k(X_{n+1},\xi)}{K_n(\xi,\boldsymbol{V})}\right\},$$

which in turn is proportional to the density function of an exponential random variable. Finally, the result is obtained by multiplying the random variable U by the rate of the exponential distribution above, so that the transformed random variable is exponential with unit rate.

A.4 Proofs of Chapter 5

Proof of Proposition 5.2 Consider the likelihood function, i.e. the conditional distribution of the observations $\boldsymbol{X}$, given the random measure $\tilde{\mu}$:

$$\begin{aligned}\mathcal{L}(\tilde{\mu};\boldsymbol{X}) = \mathbb{P}(\boldsymbol{X}\mid\tilde{\mu}) &= \prod_{i=1}^{n}\frac{\tilde{\mu}(dX_i)}{\tilde{\mu}(\mathbb{X})} = \tilde{\mu}(\mathbb{X})^{-n}\prod_{i=1}^{n}\tilde{\mu}(dX_i)\\ &= \frac{1}{\Gamma(n)}\int_{\mathbb{R}^+}u^{n-1}\exp\left\{-\int_{\mathbb{X}}u\,\tilde{\mu}(dx)\right\}\prod_{j=1}^{k}\tilde{\mu}(dX_j^*)^{n_j}\,du,\end{aligned} \tag{A.6}$$

where a standard analytical manipulation based on the density of a gamma random variable is used in the second line. The marginal distribution is obtained by marginalization with respect to $\tilde{\mu}$, that is

$$\mathbb{P}(\boldsymbol{X}) = \frac{1}{\Gamma(n)}\int_{\mathbb{R}^+}u^{n-1}\,\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}}u\,\tilde{\mu}(dx)\right\}\prod_{j=1}^{k}\tilde{\mu}(dX_j^*)^{n_j}\right]du;$$

The distribution of the random measure $\tilde{\mu}$ is absolutely continuous with respect to the distribution of a σ-stable completely random measure $\tilde{\mu}_\sigma$, with Radon-Nikodym derivative (5.1),

$$\begin{aligned}d\mathcal{L}(\tilde{\mu}) &= \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)}\left(\int_{\mathbb{X}}G(x)\,\mu_\sigma(dx)\right)^{-\theta}d\mathcal{L}(\tilde{\mu}_\sigma)\\ &= \frac{\sigma}{\Gamma(\theta/\sigma)}\int_{\mathbb{R}^+}u_0^{\theta-1}\exp\left\{-\int_{\mathbb{X}}u_0\,G(x)\,\tilde{\mu}_\sigma(dx)\right\}du_0\,d\mathcal{L}(\tilde{\mu}_\sigma);\end{aligned}$$

therefore, the expectation above can be rewritten as an expectation with respect to a σ-stable completely random measure,

$$\mathbb{P}(\boldsymbol{X}) = \frac{1}{\Gamma(n)} \frac{\sigma}{\Gamma(\theta/\sigma)} \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} u_0^{\theta-1}\, u^{n-1} \times \mathbb{E}\left[\exp\left\{ -\int_{\mathbb{X}} (u + u_0\, G(x))\, \tilde{\mu}_\sigma(dx) \right\} \prod_{j=1}^{k} \tilde{\mu}_\sigma(dX_j^*)^{n_j} \right] du\, du_0.$$

This expectation can be computed exploiting the identity in (2.10) with $\alpha(dx) = P_0(dx)$ and $f(x) = u + u_0\, G(x)$,

$$= \frac{1}{\Gamma(n)} \frac{\sigma^{k+1}}{\Gamma(\theta/\sigma)} \prod_{j=1}^{k} (1-\sigma)_{n_j-1}\, P_0(dX_j^*) \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} u_0^{\theta-1}\, u^{n-1} \times \exp\left\{ -\int_{\mathbb{X}} (u + u_0\, G(x))^\sigma P_0(dx) \right\} \prod_{j=1}^{k} (u + u_0\, G(X_j^*))^{\sigma-n_j} du\, du_0,$$

where the expressions in (A.3) for the Laplace exponent and cumulants of the σ-stable completely random measure $\tilde{\mu}_\sigma$ are substituted. Considering the change of variables $(u, u_0) \mapsto (t, v)$ such that

$$t = u + u_0, \qquad v = \frac{u_0}{u + u_0}, \qquad du\, du_0 = t\, dt\, dv.$$

the expression of the marginal distribution becomes

$$= \frac{1}{\Gamma(n)} \frac{\sigma^{k+1}}{\Gamma(\theta/\sigma)} \prod_{j=1}^{k} (1-\sigma)_{n_j-1} P_0(dX_j^*) \int_0^1 v^{\theta-1} (1-v)^{n-1} \times \int_{\mathbb{R}^+} t^{k\sigma+\theta-1} \exp\left\{ -t \int_{\mathbb{X}} (1 + v\, G(x))^\sigma P_0(dx) \right\} dt \prod_{j=1}^{k} (1 + v\, G(X_j^*))^{\sigma-n_j}\, dv.$$

The integral with respect to t can be computed analytically,

$$= \frac{\sigma^k}{\Gamma(n)} \frac{\Gamma(k+\theta/\sigma)}{\Gamma(\theta/\sigma)} \prod_{j=1}^{k} (1-\sigma)_{n_j-1}\, P_0(dX_j^*) \int_0^1 v^{\theta-1}\, (1-v)^{n-1} \left(\int_{\mathbb{X}} (1 + v\, G(x))^\sigma P_0(dx) \right)^{-(k+\theta/\sigma)} \prod_{j=1}^{k} (1 + v\, G_0(X_j^*))^{\sigma-n_j}\, dv;$$

the result is obtained by rearranging the terms and defining the following function:

$$H(v) := \left(\int_{\mathbb{X}} (1 + v\, G(x))^\sigma P_0(dx) \right)^{1/\sigma}.$$

Proof of Corollary 5.3 Considering the expression of the marginal distribution, one can safely substitute $\sigma = 0$ everywhere, except for the term $H(v)$, which can be rewritten as

$$H(v) = \exp\left\{\frac{1}{\sigma}\,\log\left(\int_{\mathbb{X}}(1+v\,G_0(x))^{\sigma}P_0(dx)\right)\right\} = \exp\left\{\frac{1}{\sigma}\,\log\left(\int_{\mathbb{X}}e^{\sigma\log(1+v\,G_0(x))}P_0(dx)\right)\right\}.$$

The limit for $\sigma \to 0$ is obtained by exploiting the Taylor expansions of the exponential and logarithm functions,

$$\begin{aligned}\lim_{\sigma\to 0} H(v) &= \lim_{\sigma\to 0}\exp\left\{\frac{1}{\sigma}\,\log\left(\int_{\mathbb{X}}e^{\sigma\log(1+v\,G_0(x))}P_0(dx)\right)\right\}\\ &= \lim_{\sigma\to 0}\exp\left\{\frac{1}{\sigma}\log\left(1+\sigma\int_{\mathbb{X}}\log(1+v\,G_0(x))\,P_0(dx)\right)\right\}\\ &= \exp\left\{\int_{\mathbb{X}}\log(1+v\,G_0(x))\,P_0(dx)\right\}.\end{aligned}$$

Therefore, the limit of the marginal distribution for $\sigma \to 0$ is

$$\frac{\Gamma(\theta)}{\Gamma(n+\theta)}\prod_{j=1}^{k}\Gamma(n_j)\,\theta P_0(dX_j^*)\int_0^1\exp\left\{-\int_{\mathbb{X}}\log(1+v\,G_0(x))\,\theta P_0(dx)\right\}\prod_{j=1}^{k}(1+v\,G_0(X_j^*))^{-n_j}\,f_{\theta,n}(v)\,dv.$$

Proof of Proposition 5.4 The exchangeable partition probability function, that is, the marginal distribution of the induced random partition, is obtained from the marginal distribution in Proposition 5.2 upon marginalization of the distinct values $X_1^*,\ldots,X_k^*$. Specifically, the marginal distribution can be rewritten as

$$\mathbb{P}(\boldsymbol{X}) = \frac{\prod_{\ell=1}^{k-1}(\theta+\ell\sigma)}{(\theta+1)_{n-1}}\prod_{j=1}^{k}(1-\sigma)_{n_j-1}\times\int_0^1 H(v)^{-\theta}\prod_{j=1}^{k}\left(\frac{(1+v\,G_0(X_j^*))^{\sigma-n_j}\,P_0(dX_j^*)}{H(v)^{\sigma}}\right)f_{\theta,n}(v)\,dv.$$

Marginalizing the distinct values $X_1^*,\ldots,X_k^*$, the quantity in the parentheses above turns into

$$S(v;n_j) := \frac{\displaystyle\int_{\mathbb{X}}(1+v\,G_0(x))^{\sigma-n_j}P_0(dx)}{\displaystyle\int_{\mathbb{X}}(1+v\,G_0(x))^{\sigma}P_0(dx)}, \qquad j=1,\ldots,k.$$

where the function $H(v)$ is made explicit; therefore, the exchangeable partition probability function can be written as

$$\mathbb{P}(\boldsymbol{\Pi}) = \frac{\prod_{\ell=1}^{k-1}(\theta+\ell\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k}(1-\sigma)_{n_j-1} \int_0^1 H(v)^{-\theta} \prod_{j=1}^{k} S(v;n_j)\, f_{\theta,n}(v)\, dv.$$

Proof of Proposition 5.5 The posterior distribution of the random measure $\tilde{\mu}$, given the observations $\boldsymbol{X}$, can be characterized through its conditional Laplace transform,

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h(x)\,\tilde{\mu}(dx)\right\} \,\middle|\, \boldsymbol{X}\right] = \frac{\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h(x)\,\tilde{\mu}(dx)\right\}\mathcal{L}(\tilde{\mu};\boldsymbol{X})\right]}{\mathbb{E}\left[\mathcal{L}(\tilde{\mu};\boldsymbol{X})\right]},$$

where $h\colon \mathbb{X}\mapsto\mathbb{R}^+$ is any non-negative measurable function. The quantity at the denominator is the marginal distribution in Proposition 5.2, while the expectation at the numerator can be computed similarly; specifically, from the augmented likelihood in (A.6),

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h(x)\,\tilde{\mu}(dx)\right\}\mathcal{L}(\tilde{\mu};\boldsymbol{X})\right]$$
$$= \frac{1}{\Gamma(n)}\int_{\mathbb{R}^+} u^{n-1}\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}}(h(x)+u)\,\tilde{\mu}(dx)\right\}\prod_{j=1}^{k}\tilde{\mu}(dX_j^*)^{n_j}\right] du.$$

According to the change of measure (5.1) defining the random measure $\tilde{\mu}$, the expectation above can be rewritten as an expectation with respect to a σ-stable completely random measure,

$$= \frac{1}{\Gamma(n)}\frac{\sigma}{\Gamma(\theta/\sigma)}\int_{\mathbb{R}^+}\int_{\mathbb{R}^+} u_0^{\theta-1}\,u^{n-1}$$
$$\times\,\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}}(h(x)+u+u_0\,G(x))\,\tilde{\mu}_\sigma(dx)\right\}\prod_{j=1}^{k}\tilde{\mu}_\sigma(dX_j^*)^{n_j}\right] du\,du_0;$$

exploiting the identity in (2.10) with $\alpha(dx)=P_0(dx)$ and $f(x)=h(x)+u+u_0\,G(x)$,

$$= \frac{1}{\Gamma(n)}\frac{\sigma^{k+1}}{\Gamma(\theta/\sigma)}\prod_{j=1}^{k}(1-\sigma)_{n_j-1}\,P_0(dX_j^*)\int_{\mathbb{R}^+}\int_{\mathbb{R}^+} u_0^{\theta-1}\,u^{n-1}$$
$$\times\exp\left\{-\int_{\mathbb{X}}(h(x)+u+u_0\,G(x))^\sigma P_0(dx)\right\}\prod_{j=1}^{k}(h(X_j^*)+u+u_0\,G(X_j^*))^{\sigma-n_j}\,du\,du_0,$$

Therefore, the conditional Laplace transform is

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h(x)\,\tilde{\mu}(dx)\right\}\,\middle|\,\boldsymbol{X}\right]$$
$$= \frac{\displaystyle\int_{\mathbb{R}^+}\int_{\mathbb{R}^+} u_0^{\theta-1}\,u^{n-1}\exp\left\{-\int_{\mathbb{X}}(h(x)+u+u_0\,G(x))^{\sigma}P_0(dx)\right\}\prod_{j=1}^{k}(h(X_j^*)+u+u_0\,G(X_j^*))^{\sigma-n_j}du\,du_0}{\displaystyle\int_{\mathbb{R}^+}\int_{\mathbb{R}^+} u_0^{\theta-1}\,u^{n-1}\exp\left\{-\int_{\mathbb{X}}(u+u_0\,G(x))^{\sigma}P_0(dx)\right\}\prod_{j=1}^{k}(u+u_0\,G(X_j^*))^{\sigma-n_j}du\,du_0}$$

Rearranging the terms of the integrand function at the numerator in order to match the those of the integrand at the denominator, one gets

$$= \int_{\mathbb{R}^+}\int_{\mathbb{R}^+}\exp\left\{-\int_{\mathbb{X}}\Big((h(x)+u+u_0\,G(x))^{\sigma}-(u+u_0\,G(x))^{\sigma}\Big)P_0(dx)\right\}$$
$$\times\prod_{j=1}^{k}\left(\frac{h(X_j^*)+u+u_0\,G(X_j^*)}{u+u_0\,G(X_j^*)}\right)^{\sigma-n_j} f(u_0,u)\,du\,du_0,$$

where $f(u_0,u)$ is a density function proportional to

$$u_0^{\theta-1}\,u^{n-1}\exp\left\{-\int_{\mathbb{X}}(u+u_0\,G(x))^{\sigma}P_0(dx)\right\}\prod_{j=1}^{k}(u+u_0\,G(X_j^*))^{\sigma-n_j}.$$

The exponential term coincides with Laplace transform of a generalized gamma completely random measure (see Section 2.7) evaluated at function $h(x)$, and thus can be rewritten as

$$\exp\left\{-\int_{\mathbb{X}}\Big((h(x)+u+u_0\,G(x))^{\sigma}-(u+u_0\,G(x))^{\sigma}\Big)P_0(dx)\right\}$$
$$=\exp\left\{-\int_{\mathbb{X}}\psi^*(h(x)\mid x)\,P_0(dx)\right\}=\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}}h(x)\,\tilde{\mu}^*(dx)\right\}\right],$$

where $\tilde{\mu}^*$ is a generalized gamma completely random measure having Lévy intensity measure

$$\nu^*(ds,dx)=\frac{\sigma}{\Gamma(1-\sigma)}\,s^{-1-\sigma}\exp\{-(u+u_0\,G(x))\,s\}\,ds\,P_0(dx).$$

Moreover, the product of ratios coincides with the product of Laplace transforms of gamma random variables, computed at values $h(X_j^*)$, for $j=1,\ldots,k$, that is

$$\left(\frac{h(X_j^*)+u+u_0\,G(X_j^*)}{u+u_0\,G(X_j^*)}\right)^{\sigma-n_j}=\mathbb{E}\left[\exp\left\{-h(X_j^*)\,W_j\right\}\right],\qquad j=1,\ldots,k,$$

where each W_j is a gamma random variable with shape $n_j - \sigma$ and rate $u + u_0\, G(X_j^*)$. Hence, the conditional Laplace transform can be written as

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h(x)\,\tilde{\mu}(dx)\right\} \,\middle|\, \boldsymbol{X}\right] \\
&\qquad = \int_{\mathbb{R}^+}\int_{\mathbb{R}^+} \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h(x)\,\tilde{\mu}^*(dx)\right\}\right] \prod_{j=1}^{k} \mathbb{E}\left[\exp\left\{-h(X_j^*)\,W_j\right\}\right] f(u_0,u)\,du\,du_0 \\
&\qquad = \int_{\mathbb{R}^+}\int_{\mathbb{R}^+} \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h(x)\,\tilde{\mu}^*(dx) - \sum_{j=1}^{k} h(X_j^*)\,W_j\right\}\right] f(u_0,u)\,du\,du_0;
\end{aligned}$$

the posterior distribution of the random measure $\tilde{\mu}$ can be regarded as a mixture of completely random measures, having Lévy intensities depending on the mixing latent parameters U_0 and U, distributed according to $f(u_0, u)$. By further conditioning on such auxiliary variables,

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h(x)\,\tilde{\mu}(dx)\right\} \,\middle|\, \boldsymbol{X}, U_0, U\right] = \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h(x)\,\tilde{\mu}^*(dx) - \sum_{j=1}^{k} h(X_j^*)\,W_j\right\}\right];$$

which highlights the structure of the conditional posterior distribution of the random measure $\tilde{\mu}$ as the sum of the non-homogeneous generalized gamma completely random measure $\tilde{\mu}^*$ and the gamma-distributed random jumps $W_1, \ldots, W_k$ at fixed locations $X_1^*, \ldots, X_k^*$; moreover, these components are mutually independent. The proof is concluded by the following reparameterizations.

(1) Consider the change of variables $(u_0, u) \mapsto (z, v)$ such that

$$z = u_0 + u, \qquad v = \frac{u_0}{u_0 + u}, \qquad du_0\,du = z\,dz\,dv;$$

the joint density function of Z and V is proportional to

$$f(z,v) \propto z^{k\sigma+\theta-1}\, v^{\theta-1}\, \exp\left\{-z^{\sigma}\int_{\mathbb{X}} (1 + v\,G(x))^{\sigma} P_0(dx)\right\} \prod_{j=1}^{k} (1 + v\,G(X_j^*))^{\sigma-n_j}.$$

(2) Consider the change of variables $t = z\,H(v)$, where the quantity $H(v)$ is defined in (5.5); the joint density function of T and V is proportional to

$$f(t,v) \propto t^{k\sigma+\theta-1}\, e^{-t^{\sigma}}\, H(v)^{-(\theta+k\sigma)}\, v_0^{\theta-1} \prod_{j=1}^{k} (1 + v\,G(X_j^*))^{\sigma-n_j}.$$

In conclusion, the auxiliary latent variables T and V are independent, the random variable V has density function as in (5.8) and T has generalized gamma distribution. Moreover, the

quantity $u + u_0\, G(x)$ characterizing the random measure $\tilde{\mu}^*$ and the random jumps $W_1, \ldots, W_k$ is reparameterized as

$$u + u_0\, G(x) = \frac{1 + v\, G(x)}{H(v)}\, t.$$

Proof of Proposition 5.6 The distribution of the random measure $\tilde{\mu}^*$ is characterized through its Laplace transform; specifically, given the observations $\boldsymbol{X}$ the latent variables V_n and T, it is a generalized gamma completely random measure,

$$\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h(x)\, \tilde{\mu}^*(dx)\right\} \mid \boldsymbol{X}, V_n, T\right] = \exp\left\{-\int_{\mathbb{X}} \left(h(x) + \frac{1 + V_n G_0(x)}{H(V_n)}\, T\right)^{\sigma} P_0(dx) + T^{\sigma}\right\}.$$

where the definition of $H(V_n)$ is exploited. Therefore, upon marginalization with respect to the random variable T, the distribution of $\tilde{\mu}^*$ is characterized by the Laplace transform

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h(x)\, \tilde{\mu}^*(dx)\right\} \mid \boldsymbol{X}, V_n\right] \\
&\qquad = \int_{\mathbb{R}^+} \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h(x)\, \tilde{\mu}^*(dx)\right\} \mid \boldsymbol{X}, V_n, T = t\right] f_T(t \mid \boldsymbol{X})\, dt \\
&\qquad = \frac{\sigma}{\Gamma(k + \theta/\sigma)} \int_{\mathbb{R}^+} t^{k\sigma+\theta-1} \exp\left\{-\int_{\mathbb{X}} \left(h(x) + \frac{1 + V_n G_0(x)}{H(V_n)}\, t\right)^{\sigma} P_0(dx)\right\} dt.
\end{aligned}$$

The exponential term coincides with the Laplace transform of a σ-stable completely random measure, having Lévy intensity measure (4.6), evaluated at function $h(x) + t\,(1 + V_n G_0(x))/H(V_n)$; therefore, one obtains

$$\begin{aligned}
&= \frac{\sigma}{\Gamma(k + \theta/\sigma)} \int_{\mathbb{R}^+} t^{k\sigma+\theta-1}\, \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} \left(h(x) + \frac{1 + V_n G_0(x)}{H(V_n)}\, t\right) \tilde{\mu}_\sigma(dx)\right\}\right] dt \\
&= \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h(x)\, \tilde{\mu}_\sigma(dx)\right\} \frac{\sigma}{\Gamma(k + \theta/\sigma)} \int_{\mathbb{R}^+} t^{k\sigma+\theta-1} \exp\left\{-t \int_{\mathbb{X}} \frac{1 + V_n G_0(x)}{H(V_n)}\, \tilde{\mu}_\sigma(dx)\right\} dt\right].
\end{aligned}$$

By computing the integral with respect to t,

$$= \mathbb{E}\left[\exp\left\{-\int_{\mathbb{X}} h(x)\, \tilde{\mu}_\sigma(dx)\right\} \frac{\sigma\, \Gamma(k\sigma + \theta)}{\Gamma(k + \theta/\sigma)} \left(\int_{\mathbb{X}} \frac{1 + V_n G_0(x)}{H(V_n)}\, \tilde{\mu}_\sigma(dx)\right)^{-(k\sigma+\theta)}\right];$$

in conclusion, the distribution of the random measure $\tilde{\mu}^*$ is absolutely continuous with respect to the distribution of $\tilde{\mu}_\sigma$, with Radon-Nikodym derivative

$$\frac{d\mathcal{L}(\tilde{\mu}^*)}{d\mathcal{L}(\tilde{\mu}_\sigma)}(m) = \frac{\sigma\Gamma(k\sigma + \theta)}{\Gamma(k + \theta/\sigma)} \left(\int_{\mathbb{X}} \frac{1 + V_n G_0(x)}{H(V_n)}\, m(dx)\right)^{-(\theta+k\sigma)}.$$

Proof of Proposition 5.7 (marginal approach) The predictive distribution for the additional observations is obtained from the marginal distribution in Proposition 5.2 as

$$\mathbb{P}(\boldsymbol{X}^+ \mid \boldsymbol{X}) = \frac{\mathbb{P}(\boldsymbol{X}^+, \boldsymbol{X})}{\mathbb{P}(\boldsymbol{X})},$$

where the numerator is the marginal distribution for $n+m$ observations. Specifically, the numerator takes the form

$$\begin{aligned}\mathbb{P}(\boldsymbol{X}^+, \boldsymbol{X}) = {} & \frac{\prod_{\ell=1}^{k+h-1}(\theta+\ell\sigma)}{(\theta+1)_{n+m-1}} \prod_{j=1}^{k}(1-\sigma)_{n_j+m_j-1} \prod_{j=k+1}^{k+h}(1-\sigma)_{m_j-1} \prod_{j=1}^{k+h} P_0(dX_j^*) \\ & \times \int_0^1 H(z)^{-(\theta+n+m)} \prod_{j=1}^{k}\left(\frac{1+z\,G(X_j^*)}{H(z)}\right)^{\sigma-n_j-m_j} \prod_{j=+1}^{k+h}\left(\frac{1+z\,G(X_j^*)}{H(z)}\right)^{\sigma-m_j} f_{\theta,n+m}(z)\,dz,\end{aligned}$$

where $f_{\theta,n+m}(z)$ denotes the density function of the Beta distribution with parameters θ and n. As a consequence, the joint predictive distribution is

$$\begin{aligned}\mathbb{P}(\boldsymbol{X}^+ \mid \boldsymbol{X}) = {} & \frac{\prod_{\ell=0}^{h-1}(\theta+k\sigma+\ell\sigma)}{(\theta+n)_m} \prod_{j=1}^{k}(n_j-\sigma)_{m_j} \prod_{j=k+1}^{k+h}(1-\sigma)_{m_j-1}\, P_0(dX_j^*) \\ & \times \int_0^1 H(z)^{-(\theta+n+m)} \prod_{j=1}^{k}\left(\frac{1+z\,G(X_j^*)}{H(z)}\right)^{\sigma-n_j-m_j} \\ & \times \prod_{j=+1}^{k+h}\left(\frac{1+z\,G(X_j^*)}{H(z)}\right)^{\sigma-m_j} C(\boldsymbol{X})^{-1}\, f_{\theta,n+m}(z)\,dz,\end{aligned}$$

where the quantity $C(\boldsymbol{X})$ is the normalizing constant for the density function of the latent variable V_n, that is

$$C(\boldsymbol{X}) = \int_0^1 H(v)^{-(\theta+n)} \prod_{j=1}^{k}\left(\frac{1+v\,G(X_j^*)}{H(v)}\right)^{\sigma-n_j} f_{\theta,n}(v)\,dv.$$

The integral above can be augmented by introducing an additional integration variable, and exploiting the density function of a rescaled Beta random variable, so that

$$\begin{aligned}f_{\theta,n+m}(z) &= \frac{\Gamma(\theta+n+m)}{\Gamma(\theta)\Gamma(n+m)}\, z^{\theta-1}(1-z)^{n+m-1} \\ &= \frac{\Gamma(\theta+n+m)}{\Gamma(\theta)\Gamma(n)\Gamma(m)}\, z^{\theta-1} \int_0^{1-z} z_0^{n-1}(1-z-z_0)^{m-1}\,dz_0.\end{aligned}$$

Considering the change of variables $(z_0, z) \mapsto (v, w)$ such that

$$w = z + z_0, \qquad v = \frac{z}{w}, \qquad dz_0\,dz = w\,dw\,dv,$$

the integral double integral can be rewritten as

$$\int_0^1 \int_0^1 H(wv)^{-(\theta+n+m)} \prod_{j=1}^{k} \left(\frac{1+wv\,G(X_j^*)}{H(wv)} \right)^{\sigma-n_j-m_j} \times \prod_{j=+1}^{k+h} \left(\frac{1+wv\,G(X_j^*)}{H(wv)} \right)^{\sigma-m_j} f_{\theta+n,m}(w)\,dw\,C(\boldsymbol{X})^{-1}\,f_{\theta,n}(v)\,dv.$$

Recalling that the density function of the auxiliary latent variables V_n in (5.8) is

$$f_{V_n}(v \mid \boldsymbol{X}) = C(\boldsymbol{X})^{-1}\,H(v)^{-(\theta+n)} \prod_{j=1}^{k} \left(\frac{1+v\,G(X_j^*)}{H(v)} \right)^{\sigma-n_j} f_{\theta,n}(v),$$

the joint predictive distribution is rewritten as

$$\mathbb{P}(\boldsymbol{X}^+ \mid \boldsymbol{X}) = \frac{\prod_{\ell=0}^{h-1}(\theta+k\sigma+\ell\sigma)}{(\theta+n)_m} \prod_{j=1}^{k}(n_j-\sigma)_{m_j} \prod_{j=k+1}^{k+h}(1-\sigma)_{m_j-1}\,P_0(dX_j^*) \times \int_0^1 H(v)^{\theta+n} \prod_{j=1}^{k} \left(\frac{1+v\,G(X_j^*)}{H(v)} \right)^{n_j-\sigma} \int_0^1 H(wv)^{-(\theta+n+m)} \prod_{j=1}^{k} \left(\frac{1+wv\,G(X_j^*)}{H(wv)} \right)^{\sigma-n_j-m_j} \times \prod_{j=+1}^{k+h} \left(\frac{1+wv\,G(X_j^*)}{H(wv)} \right)^{\sigma-m_j} f_{\theta+n,m}(w)\,dw\,f_{V_n}(v \mid \boldsymbol{X})\,dv.$$

The result is obtained by conditioning on the auxiliary latent variable V_n, playing the role of mixing parameters in the expression above.

Bibliography

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII – 1983*, pp. 1–198. Springer Berlin Heidelberg.

Allignol, A., M. Schumacher, C. Wanner, C. Drechsler, and J. Beyersmann (2011). Understanding competing risks: a simulation point of view. *BMC Medical Research Methodology 86*(11).

Andersen, P. K., S. Z. Abildstrom, and S. Rosthøj (2002). Competing risks as a multi-state model. *Statistical Methods in Medical Research 11*(2), 203–215.

Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer New York.

Andersen, P. K. and N. Keiding (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine 31*, 1074–1088.

Antoniano-Villalobos, I., S. Wade, and S. G. Walker (2014). A Bayesian nonparametric regression model with normalized weights: A study of hippocampal atrophy in Alzheimer's disease. *Journal of the American Statistical Association 109*(506), 477–490.

Arfé, A., S. Peluso, and P. Muliere (2019). Reinforced urns and the subdistribution beta-Stacy process for competing risks analysis. *Scandinavian Journal of Statistics 46*, 706–734.

Argiento, R., A. Cremaschi, and M. Vannucci (2020). Hierarchical normalized completely random measures to cluster grouped data. *Journal of the American Statistical Association 115*(529), 318–333.

Ascolani, F., A. Lijoi, and M. Ruggiero (2021). Predictive inference with Fleming–Viot-driven dependent Dirichlet processes. *Bayesian Analysis 16*(2), 371 – 395.

Barndorff-Nielsen, O. E. and N. Shephard (2001). Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(2), 167–241.

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

Barry, D. and J. A. Hartigan (1992). Product partition models for change point problems. *The Annals of Statistics 20*(1), 260–279.

Basu, D. and R. C. Tiwari (1982). A note on the Dirichlet process. In *Statistics and Probability: Essays in honor of C. R. Rao*, pp. 89–103. North-Holland, Amsterdam - New York.

Beyersmann, J., A. Latouche, A. Buchholz, and M. Schumacher (2009). Simulating competing risks data in survival analysis. *Statistics in Medicine 28*(6), 956–971.

Blackwell, D. and J. B. MacQueen (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics 1*, 353—-355.

Brix, A. (1999). Generalized Gamma measures and shot-noise Cox processes. *Advances in Applied Probability 31*(4), 929–953.

Camerlenghi, F., D. B. Dunson, A. Lijoi, I. Prünster, and A. Rodriguez (2019). Latent nested nonparametric priors. *Bayesian Analysis 14*(4), 1303–1356.

Camerlenghi, F., A. Lijoi, P. Orbanz, and I. Prünster (2019). Distribution theory for hierarchical processes. *The Annals of Statistics 47*(1), 67–92.

Camerlenghi, F., A. Lijoi, and I. Prünster (2017). Bayesian prediction with multiple-samples information. *Journal of Multivariate Analysis 156*, 18–28.

Camerlenghi, F., A. Lijoi, and I. Prünster (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scandinavian Journal of Statistics. Theory and Applications 45*(4), 1062–1091.

Camerlenghi, F., A. Lijoi, and I. Prünster (2021). Survival analysis via hierarchically dependent mixture hazards. *The Annals of Statistics 49*(2), 863–884.

Catalano, M., H. Lavenant, A. Lijoi, and I. Prünster (2024). A Wasserstein index of dependence for random measures. *Journal of the American Statistical Association 119*(547), 2396–2406.

Catalano, M., A. Lijoi, and I. Prünster (2020). Approximation of Bayesian models for time-to-event data. *Electronic Journal of Statistics 14*(2), 3366–3395.

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

Chen, C., V. Rao, W. Buntine, and Y. Whye Teh (2013). Dependent normalized random measures. In *Proceedings of the 30th International Conference on Machine Learning*, Volume 28 of *Proceedings of Machine Learning Research*, pp. 969–977. PMLR.

Chung, Y. and D. B. Dunson (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association 104*(488), 1646–1660.

Chung, Y. and D. B. Dunson (2011). The local Dirichlet process. *Annals of the Institute of Statistical Mathematics 63*(1), 59–80.

Cifarelli, D. M. and E. Regazzini (1978). Nonparametric statistical problems under partial exchangeability: The role of associative means. *Quaderni Istituto Matematica Finanziaria dell'Università di Torino: Serie III 12*, 1–36.

Cox, D. R. (1959). The analysis of exponentially distributed life-times with two types of failures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 21*, 411–421.

Crowder, M. (1991). On the identifiability crisis in competing risks analysis. *Scandinavian Journal of Statistics 18*(3), 223–233.

Crowder, M. J. (2012). *Multivariate Survival Analysis and Competing Risks*. Texts in Statistical Science. Chapman and Hall.

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

De Blasi, P., G. Peccati, and I. Prünster (2009). Asymptotics for posterior hazards. *The Annals of Statistics 37*(4), 1906–1945.

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

de Finetti, B. (1938). Sur la condition d'équivalence partielle. *Actualités Scientifique et Industrielles 739*, 5–18.

De Iorio, M., P. Müller, G. L. Rosner, and S. N. MacEachern (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association 99*(465), 205–215.

Devroye, L. (1981). On the computer generation of random variables with a given characteristic function. *Computers & Mathematics with Applications 7*(6), 547–552.

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability 2*(2), 183–201.

Drzewiecki, K. T., C. Ladefoged, and H. E. Christensen (1980). Biopsy and prognosis for cutaneous malignant melanomas in clinical stage I. *Scandinavian Journal of Plastic and Reconstructive Surgery 14*(2), 141–144.

Dunson, D. B. (2010). Nonparametric Bayes applications to biostatistics. In N. L. Hjort, C. C. Holmes, P. Müller, and S. G. Walker (Eds.), *Bayesian Nonparametrics*, pp. 223–273. Cambridge University Press.

Dunson, D. B. and J.-H. Park (2008). Kernel stick-breaking processes. *Biometrika 95*(2), 307–323.

Dunson, D. B., N. Pillai, and J.-H. Park (2007). Bayesian density regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology 69*(2), 163–183.

Dykstra, R. L. and P. Laud (1981). A Bayesian nonparametric approach to reliability. *The Annals of Statistics 9*(2), 356–367.

Epifani, I. and A. Lijoi (2010). Nonparametric priors for vectors of survival functions. *Statistica Sinica 20*(4), 1455–1484.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*(430), 577–588.

Escobar, M. D. and M. West (1998). Computing nonparametric hierarchical models. In D. Dey, P. Müller, and D. Sinha (Eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*, pp. 1–22. Springer New York.

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

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics 2*(4), 615–629.

Ferguson, T. S. and M. J. Klass (1972). A representation of independent increment processes without Gaussian components. *The Annals of Mathematical Statistics 43*(5), 1634–1643.

Fine, J. P. and R. J. Gray (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association 94*(446), 496–509.

Foti, N. and S. Williamson (2012). Slice sampling normalized kernel-weighted completely random measure mixture models. In *Advances in Neural Information Processing Systems*, Volume 25. Curran Associates, Inc.

Gelfand, A. E., A. Kottas, and S. N. MacEachern (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association 100*(471), 1021–1035.

Geskus, R. B. (2015). *Data Analysis with Competing Risks and Intermediate States*. CRC Biostatistics Series. Chapman and Hall.

Geskus, R. B. (2024). Competing risks: Concepts, methods, and software. *Annual Review of Statistics and Its Application 11*(1), null.

Ghosal, S. and A. van der Vaart (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Gnedin, A. and J. Pitman (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences 138*(3), 5674–5685.

Griffin, J. E. (2011). The Ornstein–Uhlenbeck Dirichlet process and other time-varying processes for Bayesian nonparametric inference. *Journal of Statistical Planning and Inference 141*(11), 3648–3664.

Griffin, J. E., M. Kolossiatis, and M. F. J. Steel (2013). Comparing distributions by using dependent normalized random-measure mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(3), 499–529.

Griffin, J. E. and F. Leisen (2017). Compound random measures and their use in bayesian nonparametrics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(2), 525–545.

Griffin, J. E. and F. Leisen (2018). Modelling and computation using NCoRM mixtures for density regression. *Bayesian Analysis 13*(3), 897 – 916.

Griffin, J. E. and M. Steel (2006). Order-based Dependent Dirichlet Processes. *Journal of the American Statistical Association 101*(473), 179–194.

Hartigan, J. A. (1990). Partition models. *Communications in Statistics - Theory and Methods 19*(8), 2745–2756.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data. *The Annals of Statistics 18*(3), 1259–1294.

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

Ishwaran, H. and L. F. James (2004). Computational methods for multiplicative intensity models using weighted gamma processes: proportional hazards, marked point processes, and panel count data. *Journal of the American Statistical Association 99*(465), 175–190.

James, L. F. (2005). Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *The Annals of Statistics 33*(4), 1771–1799.

James, L. F., A. Lijoi, and I. Prünster (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scandinavian Journal of Statistics 33*(1), 105–120.

James, L. F., A. Lijoi, and I. Prünster (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics 36*(1), 76–97.

Kalbfleisch, J. D. and R. L. Prentice (2002). *The Statistical Analysis of Failure Time Data.* Wiley Series in Probability and Statistics.

Kingman, J. F. (1967). Completely random measures. *Pacific Journal of Mathematics 21*(1), 59–78.

Kingman, J. F. (1975). Random discrete distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 37*(1), 1–22.

Kingman, J. F. (1993). *Poisson Processes.* Oxford Studies in Probability. Clarendon Press.

Lau, J. W. and E. Cripps (2022). Thinned completely random measures with applications in competing risks models. *Bernoulli 28*(1), 638–662.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data.* Wiley Series in Probability and Statistics.

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

Lijoi, A., R. H. Mena, and I. Prünster (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(4), 715–740.

Lijoi, A. and B. Nipoti (2014). A class of hazard rate mixtures for combining survival data from different experiments. *Journal of the American Statistical Association 109*(506), 802–814.

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

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

Lijoi, A., I. Prünster, and T. Rigon (2020). Sampling hierarchies of discrete random structures. *Statistics and Computing 30*, 1591–1607.

Lijoi, A., I. Prünster, and S. G. Walker (2008). Bayesian nonparametric estimators derived from conditional gibbs structures. *The Annals of Applied Probability 18*(4), 1519–1547.

Lijoi, A., I. Prünster, and G. Rebaudo (2023). Flexible clustering via hidden hierarchical Dirichlet priors. *Scandinavian Journal of Statistics 50*(1), 213–234.

Lo, A. and C.-S. Weng (1989). On a class of Bayesian nonparametric estimates: II. Hazard rate estimates. *Annals of the Institute of Statistical Mathematics 41*(2), 227–245.

Lo, A. Y. (1982). Bayesian nonparametric statistical inference for Poisson point processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 59*(1), 55–66.

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

MacEachern, S. N. (1998). Computational methods for mixture of Dirichlet process models. In D. Dey, P. Müller, and D. Sinha (Eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*, pp. 23–43. Springer New York.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA. American Statistical Association.

MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical report, Department of Statistics, The Ohio State University.

Müller, P., A. Erkanli, and M. West (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika 83*(1), 67–79.

Müller, P. and R. Mitra (2013). Bayesian nonparametric inference - Why and how. *Bayesian Analysis 8*(2), 269–302.

Müller, P. and F. Quintana (2010). Random partition models with regression on covariates. *Journal of Statistical Planning and Inference 140*(10), 2801–2808.

Müller, P., F. Quintana, and G. L. Rosner (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(3), 735–749.

Müller, P., F. Quintana, and G. L. Rosner (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics 20*(1), 260–278.

Nieto-Barajas, L. E. and S. G. Walker (2004). Bayesian nonparametric survival analysis via Lévy driven Markov processes. *Statistica Sinica 14*(4), 1127–1146.

Papaspiliopoulos, O. (2011). Monte Carlo probabilistic inference for diffusion processes: a methodological framework. In D. Barber, A. T. Cemgil, and S. Chiappa (Eds.), *Bayesian Time Series Models*, pp. 82—103. Cambridge University Press.

Park, J.-H. and D. B. Dunson (2010). Bayesian generalized product partition model. *Statistica Sinica 20*(3), 1203–1226.

Perman, M., J. Pitman, and M. Yor (1992). Size-biased sampling of poisson point processes and excursions. *Probability Theory and Related Fields 92*(1), 21–39.

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

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. In *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, Volume 30 of *Lecture Notes - Monograph Series*, pp. 245–267. Institute of Mathematical Statistics.

Pitman, J. (2003). Poisson-Kingman partitions. In D. R. Goldstein (Ed.), *Statistics and science: a Festschrift for Terry Speed*, Volume 40 of *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, pp. 1–34. Institute of Mathematical Statistics.

Pitman, J. (2006). *Combinatorial Stochastic Processes*. Lecture Notes in Mathematics. Springer Berlin, Heidelberg.

Pitman, J. and M. Yor (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability 25*(2), 855–900.

Putter, H., M. Fiocco, and R. B. Geskus (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine 26*(11), 2389–2430.

Quintana, F. A., P. Müller, A. Jara, and S. N. MacEachern (2022). The dependent Dirichlet process and related models. *Statistical Science 37*(1), 24–41.

Rao, V. and Y. W. Teh (2009). Spatial normalized gamma processes. In *Advances in Neural Information Processing Systems*, Volume 22. Curran Associates, Inc.

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

Ridout, M. (2009). Generating random numbers from a distribution specified by its Laplace transform. *Statistics and Computing 19*(4), 439–450.

Rigon, T. and D. Durante (2021). Tractable Bayesian density regression via logit stick-breaking priors. *Journal of Statistical Planning and Inference 211*, 131–142.

Riva-Palacio, A. and F. Leisen (2018). Bayesian nonparametric estimation of survival functions with multiple-samples information. *Electronic Journal of Statistics 12*(1), 1330–1357.

Riva-Palacio, A. and F. Leisen (2021). Compound vectors of subordinators and their associated positive Lévy copulas. *Journal of Multivariate Analysis 183*, 104728.

Rodriguez, A. and D. B. Dunson (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis 6*(1), 145–177.

Rodriguez, A., D. B. Dunson, and A. E. Gelfand (2008). The nested Dirichlet process. *Journal of the American Statistical Association 103*(483), 1131–1154.

Rodriguez, A. and E. ter Horst (2008). Bayesian dynamic density estimation. *Bayesian Analysis 3*(2), 339–365.

Schmitt, S., A. Buchholz, and A.-K. Ozga (2023). Systematic comparison of approaches to analyze clustered competing risks data. *BMC Medical Research Methodology 23*(86).

Scoggins, C. R., M. I. Ross, D. S. Reintgen, D. Noyes, J. S. Goydos, P. D. Beitsch, M. M. Urist, S. Ariyan, J. J. Sussman, M. J. Edwards, A. B. Chagpar, R. C. G. Martin, A. J. Stromberg, L. Hagendoorn, and K. M. McMasters (2006). Gender-related differences in outcome for melanoma patients. *Annals of Surgery 243*(5), 693–700.

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

Sparapani, R., B. R. Logan, R. E. McCulloch, and P. W. Laud (2020). Nonparametric competing risks analysis using Bayesian Additive Regression Trees. *Statistical Methods in Medical Research 29*(1), 57–77.

Teh, Y. W. and M. I. Jordan (2010). Hierarchical Bayesian nonparametric models with applications. In N. L. Hjort, C. C. Holmes, P. Muller, and S. G. Walker (Eds.), *Bayesian Nonparametrics*, pp. 158–207. Cambridge University Press.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association 101*(476), 1566–1581.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences of the United States of America 72*(1), 20–22.

Wade, S., V. Inacio, and S. Petrone (2023). Bayesian dependent mixture models: A predictive comparison and survey. *arXiv:2307.16298*.

Walker, S. and P. Damien (2000). Representations of Lévy processes without Gaussian components. *Biometrika 87*(2), 477–483.

Walker, S. and P. Muliere (1997). Beta-Stacy processes and a generalization of the Pólya-urn scheme. *The Annals of Statistics 25*(4), 1762–1780.

Wolpert, R. L. and K. Ickstadt (1998). Simulation of Lévy random fields. In D. Dey, P. Müller, and D. Sinha (Eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*, pp. 227–242. Springer New York.

Xu, Y., D. Scharfstein, P. Müller, and M. Daniels (2020). A Bayesian nonparametric approach for evaluating the causal effect of treatment in randomized trials with semi-competing risks. *Biostatistics 23*(1), 34–49.

Zhou, B., J. Fine, A. Latouche, and M. Labopin (2012). Competing risks regression for clustered data. *Biostatistics 13*(3), 371–383.

Zhu, W. and F. Leisen (2015). A multivariate extension of a vector of two-parameter Poisson–Dirichlet processes. *Journal of Nonparametric Statistics 27*(1), 89–105.