Hierarchical Random Measures without Tables

Marta Catalano[1] and Claudio Del Sole[2]

[1] *Luiss University, Rome, Italy*, mcatalano@luiss.it
[2] *University of Milano-Bicocca, Milan, Italy*, claudio.delsole@unimib.it

Abstract

The hierarchical Dirichlet process is the cornerstone of Bayesian nonparametric multilevel models. Its generative model can be described through a set of latent variables, commonly referred to as *tables* within the popular restaurant franchise metaphor. The latent tables simplify the expression of the posterior and allow for the implementation of Gibbs sampling algorithms to approximately draw posterior samples. However, managing their assignments can become computationally expensive, especially as the size of the dataset and the number of levels increase. In this work, we identify a prior for the concentration parameter of the hierarchical Dirichlet process that (i) induces a quasi-conjugate posterior distribution, and (ii) removes the need for tables, leading to more interpretable expressions for the posterior, with both a scalable and an exact algorithm to sample from it. Remarkably, this construction extends beyond the Dirichlet process, leading to a new framework for defining normalized hierarchical random measures and a new class of algorithms to sample from their posteriors. The key analytical tool is the independence of multivariate increments, that is, their representation as *completely random vectors*.

Keywords: Bayesian nonparametrics, completely random measure, Dirichlet process, multilevel model, partial exchangeability.

1 Introduction

Historically, Bayesian nonparametric and hierarchical models have addressed data complexity in different ways. Nonparametric models ensure full flexibility to the marginal distribution of the observations by increasing the dimensionality of the parameter space, while hierarchical models focus on the interactions between the observations, grouping them and modeling their dependencies through shared parameters or latent features. The interaction between parameters is recursively modeled in a similar fashion, defining a hierarchical structure that enables pooling of information across different groups while

preserving their distinct characteristics – all within the principled framework of Bayesian inference. The hierarchical Dirichlet process (Teh et al., 2006) represented a significant breakthrough, demonstrating the advantages of combining the two approaches through the sharing of infinite-dimensional parameters. Since then, it has proved effective in an impressive number of contexts, including natural language processing (Teh et al., 2006; Zavitsanos et al., 2011), genomics (Sohn and Xing, 2009; Elliott et al., 2019; Liu et al., 2024), computer vision (Sudderth et al., 2008; Haines and Xiang, 2011), music segmentation and speaker diarization (Ren et al., 2008; Fox et al., 2011), cognitive science (Griffiths et al., 2007), robotics (Nakamura et al., 2011; Taniguchi et al., 2018), network analysis (Durante et al., 2025).

The computational feasibility of the hierarchical Dirichlet process in Teh et al. (2006) is strictly linked to a compelling posterior representation via latent variables, often referred to as *tables* in the restaurant franchise metaphor. The need for tables arises from the nature of the infinite-dimensional parameter, which is an almost surely discrete random probability $\tilde{P} = \sum_{i \geq 1} J_i \delta_{\theta_i}$, characterized by a countably infinite number of jumps J_i and atoms θ_i. In a Bayesian nonparametric setting, a single group of observations is often modeled as conditionally independent and identically distributed from $\tilde{P}$, where $\tilde{P}$ is e.g. a Dirichlet process (Ferguson, 1973) with diffuse mean measure P_0. In this case, since the atoms θ_i are independent and identically distributed from P_0, two observations coincide if and only if they share the same atom θ_i. Conversely, in the hierarchical Dirichlet process, the distribution P_0 of the atoms is itself a latent parameter, which is shared across different groups and likewise modeled as an almost surely discrete random probability. Hence, contrary to the exchangeable case, the atoms display ties with positive probability, and two observations coincide either if they share the same atom or if their respective atoms have identical values. This complicates the posterior representation, unless one keeps track of the atom associated to each observation, which is precisely the role of the latent tables (Catalano et al., 2024). For n observations, the tables induce a distribution on the space of partitions of n elements; the dimensionality of this space increases dramatically with the number of observations, and one typically needs a Gibbs sampling algorithm to sample from it.

The most popular implementation for the hierarchical Dirichlet process is the Gibbs sampler based on the restaurant franchise metaphor (Teh et al., 2006), often addressed as *marginal* or *collapsed* Gibbs sampler. Despite the convenient expressions of its full conditionals and its remarkable flexibility, it is not an exact algorithm, since it samples from a Markov chain and thus approximates the posterior distribution only asymptotically. In addition, there are other well-known drawbacks: it involves a considerable amount of bookkeeping, and scales poorly as the number of observations n increases, as it relies on a sequential updating scheme for the n latent tables. Moreover, each table allocation depends on all the other allocations, inducing high autocorrelation in the Markov chain, slow

mixing, and preventing parallelization. Such limitations have been recognized by several works; see e.g. Teh et al. (2006); Teh and Jordan (2010); Williamson et al. (2013); Lijoi et al. (2020); Das et al. (2025). Accordingly, a plethora of sampling-based strategies have been proposed to reduce computational time and improve mixing and scalability of the standard implementation. These are typically Gibbs samplers adopting a *conditional* or *blocked* approach, that is, instantiating the random measures with some finite-dimensional approximation of the posterior. Direct assignment schemes instantiate the jumps of the common random probability, often addressed as *global weights*, through the stick-breaking construction (Teh et al., 2006) or a finite-dimensional Dirichlet approximation (Fox et al., 2011), while variational methods (Teh et al., 2007; Wang et al., 2011; Bryant and Sudderth, 2012) construct mean field approximations of stick-breaking ratios at both levels of the hierarchy. Nevertheless, their full conditional distributions still depend on the nested partition induced by the tables, and thus do not completely avoid some of its inherent drawbacks. Interestingly, the conditional Gibbs samplers proposed in Lijoi et al. (2020) and Das et al. (2025) obviate the need for tables by considering a finite-dimensional approximation of the model, i.e. by truncating the common random probability a priori.

In this work we pursue a different strategy that eliminates the need for tables while preserving the infinite-dimensionality of the model, thanks to a specific gamma hyperprior for the shared concentration parameter of the Dirichlet processes. Intuitively, the hierarchical Dirichlet process defines a vector of dependent random probabilities with independent jumps and common atoms with ties. Our hyperprior allows for an alternative representation as dependent random probabilities with dependent jumps and common atoms without ties, thereby making the tables superfluous. A fundamental result of this work shows that this representation is the normalization of a specific *completely random vector* (Catalano et al., 2021), that is, a vector of dependent random measures with jointly independent increments. This is the natural multivariate extension of a completely random measure (Kingman, 1967), whose normalization is very popular in the Bayesian nonparametric literature (Regazzini et al., 2003; James et al., 2006, 2009). In particular, James et al. (2009) derive an almost conjugate representation of the posterior, conditionally on a real-valued latent variable U amenable to standard approximate sampling schemes, such as random walk Metropolis-Hastings (Barrios et al., 2013). In this work, we extend the posterior representation to any normalized completely random vector, which in principle could be applied to many other well-established models, including the normalization of GM-dependent measures (Lijoi et al., 2014), Lévy copulas (Epifani and Lijoi, 2010), compound random measures (Griffin and Leisen, 2017), and thinned random measures (Lau and Cripps, 2022).

The characterization as a normalized completely random vector allows to identify a novel posterior representation for the hierarchical Dirichlet process with hyperprior that does not require the latent tables. The price to pay for casting off the tables is the

introduction of a latent vector and k jump vectors, where k is the total number of distinct observations, typically much smaller than the number of observations n. In principle, the dimension of the latent and jump vectors coincides with the number of groups d, which could be potentially large, thus slowing down the posterior inference algorithm. However, through an in-depth analysis of their distributions, we reduce the non-standard sampling steps to sampling a random vector supported on $[0, \infty)^{2k+1}$. This approach offers three fundamental advantages: (i) the random vector lives in a standard space, and thus can be approximately sampled with standard Markov Chain Monte Carlo (MCMC) techniques, which show good mixing and can be analyzed with a well-established set of diagnostics; (ii) its dimension does not increase with the number of observations n nor with the number of groups d, which leads to scalable algorithms whenever n or d increase but k remains only moderately large; (iii) when n is moderately large, we devise an exact i.i.d. sampling algorithm, thus avoiding the approximation error of Gibbs sampling procedures. In summary, our novel posterior representation allows for the development of a scalable algorithm when n or d are large compared to k, and of an exact i.i.d. sampling algorithm when n is only moderately large. Note that k consistently smaller than n is the most common setting where the model should be applied, since, e.g., for the hierarchical Dirichlet process, k is of the same order as $\log(\log(n))$.

The hierarchical Dirichlet process with hyperprior is only a particular case of the more general class of nonparametric hierarchical models discussed in this work. These models arise as the normalization of a vector $(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ of conditionally independent completely random measures, given another completely random measure $\tilde{\mu}_0$. We show that these dependent random measures are completely random vectors and, for this reason, we term them *hierarchical completely random vectors*. Such hierarchical structures have been used to model dependent hazards (Camerlenghi et al., 2021; Del Sole et al., 2026) and beta processes (Thibaux and Jordan, 2007; Masoero et al., 2018; James et al., 2024), but have never been considered for normalization. This novel specification merges two compelling properties for the first time: the naturalness of the hierarchical construction and the analytical tractability of independent multivariate increments. The former leads to simple representations and draws interesting parallels with popular hierarchical models in the literature (Teh et al., 2006; Camerlenghi et al., 2018). The latter enhances the theoretical investigation of the model, leading to the almost-conjugate posterior representation we have already discussed, and playing a crucial role in deriving closed-form expressions for the prior moments and dependence structure.

The work is structured as follows. Section 2 introduces and characterizes hierarchical completely random vectors (hCRVs). Properties of their normalization are studied in Section 3, where we clarify the connections with other hierarchical models in the literature. Section 4 provides a characterization of the posterior distribution for normalized hCRVs. These results are specialized to the gamma-gamma model in Section 5, where we derive

our new posterior sampling algorithms and perform relevant comparisons in terms of computational complexity on simulated datasets. Finally, Section 6 sets rules for a fair comparison with the hierarchical Dirichlet process. Using both simulated and real datasets, we argue that a meaningful comparison between dependent priors should entail matching the first two marginal moments and measure of dependence. Proofs and technical details for the sampling algorithms are deferred to the Supplementary Material.

Notation The Cartesian product of d copies of a set A is denoted as $A^d = A \times \cdots \times A$. The product measure of d probabilities $P_1, \ldots, P_d$ is $\prod_{i=1}^d P_i$, or P^d if $P_1 = \cdots = P_d = P$. Bold symbols indicate vectors, such as $\boldsymbol{s} = (s_1, \ldots, s_d)$, and $\mathrm{d}\boldsymbol{s} = \mathrm{d}s_1 \cdot \cdots \cdot \mathrm{d}s_d$ is the Lebesgue measure on $\mathbb{R}^d$. The normal distribution with mean μ and variance σ^2 is denoted by $N(\mu, \sigma^2)$. If $f : \mathbb{X} \to \mathbb{Y}$ is a measurable map and ρ is a measure on $\mathbb{X}$, then $f_{\#}\rho$ is the pushforward measure on $\mathbb{Y}$ defined as $(f_{\#}\rho)(B) = \rho(f^{-1}(B))$, for any measurable set B. The symbol $\sim$ underlines the randomness of a random measure ($\tilde{\mu}$). The abbreviation a.s. stands for *almost surely*; we repeatedly use $\Omega_d = [0, +\infty)^d \setminus \{\mathbf{0}\}$. Contents in the Supplementary Material are labeled with an 'S' prefix, e.g. Section S1.

2 Hierarchical completely random vectors

In this section, we recall the definition of hierarchical random measures, highlight that they are homogeneous completely random vectors, and recover both their multivariate Laplace exponent and their Lévy measure. The proofs follow a structure similar to the *subordination* of Lévy processes, first defined in Bochner (1955), and beautifully described in Bertoin (1996) and Sato (1999). Their joint Lévy measure displays some similarities with that of compound random measures (Griffin and Leisen, 2017), but we show that there is no intersection between the two classes, at least for the same choice of the subordinating random measure. We note that hierarchical completely random vectors can be approximately sampled with standard techniques, exploiting their hierarchical structure. The specification of their law involves an outer and an inner Lévy measure, whose identifiability is studied with relevant examples.

A vector of random measures $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ is a measurable function on $M_{\mathbb{X}}^d$, where $M_{\mathbb{X}}$ denotes the space of boundedly finite measures on a Polish space $\mathbb{X}$. A completely random vector (Catalano et al., 2021) is the natural multivariate generalization of a completely random measure (CRM), defined in Kingman (1967). For a Borel set A of $\mathbb{X}$, we use the notation $\tilde{\boldsymbol{\mu}}(A) = (\tilde{\mu}_1(A), \ldots, \tilde{\mu}_d(A))$, which is a random vector in $\mathbb{R}^d$.

Definition 2.1. A vector of random measures $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ is a completely random vector (CRV) if, given pairwise disjoint Borel sets $A_1, \ldots, A_k$ of $\mathbb{X}$, the random vectors $\tilde{\boldsymbol{\mu}}(A_1), \ldots, \tilde{\boldsymbol{\mu}}(A_k)$ are mutually independent.

Refer to Section S1 for a brief and self-contained account on completely random measures, Lévy measures, Lévy intensities, Laplace exponents, and their multivariate extension to completely random vectors. Henceforth, $\tilde{\mu} \sim \mathrm{CRM}(\rho \otimes P_0)$ denotes a CRM with product Lévy intensity $\mathrm{d}\rho(s)\,\mathrm{d}P_0(x)$, and $\mathrm{ID}(\rho)$ indicates a pure-jump infinitely divisible distribution with Lévy measure ρ; its expression in integrals is $\mathrm{d}P_{\mathrm{ID}(\rho)}$ and its probability density function (p.d.f), if it exists, is denoted by $f_{\mathrm{ID}(\rho)}(s)$.

Definition 2.2. Let ρ_0 and ρ be Lévy measures on $(0, +\infty)$ and let P_0 be an atomless measure. We say that $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_d) \sim \mathrm{hCRV}(\rho, \rho_0, P_0)$ is a hierarchical CRV with idiosyncratic Lévy measure ρ, base Lévy measure ρ_0, and base measure P_0 if

$$\tilde{\mu}_1, \ldots, \tilde{\mu}_d \mid \tilde{\mu}_0 \overset{\text{iid}}{\sim} \mathrm{CRM}(\rho \otimes \tilde{\mu}_0); \qquad \tilde{\mu}_0 \sim \mathrm{CRM}(\rho_0 \otimes P_0).$$

The next theorem shows that the vector of random measures $\tilde{\boldsymbol{\mu}} \sim \mathrm{hCRV}(\rho, \rho_0, P_0)$ is in fact a completely random vector in the sense of Definition 2.1. Moreover, the theorem provides the expression of its multivariate Laplace functional through the Laplace exponent and determines its multivariate Lévy intensity; since $\tilde{\boldsymbol{\mu}}$ is a CRV, such intensity automatically satisfies the integrability conditions (i) and (ii) recalled in Section S1. This construction admits a natural interpretation in terms of subordination. Indeed, the proof follows the same structure as the classical subordination of Lévy processes (Bochner, 1955; Bertoin, 1996; Sato, 1999), and hierarchical CRVs may be regarded as the corresponding random-measure analogue. A related construction, based on the subordination of CRMs by infinitely divisible random measures, is also developed in Brück (2026).

Theorem 2.3. *Let $\tilde{\boldsymbol{\mu}} \sim \mathrm{hCRV}(\rho, \rho_0, P_0)$ with P_0 a probability measure. Then $\tilde{\boldsymbol{\mu}}$ is a homogeneous* CRV *with Laplace exponent $\psi_h : [0, +\infty)^d \mapsto [0, +\infty)$ and Lévy intensity $\nu_h = \rho_h \otimes P_0$ such that, for every $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d) \in [0, +\infty)^d$ and $\boldsymbol{s} = (s_1, \ldots, s_d) \in \Omega_d$,*

$$\psi_h(\boldsymbol{\lambda}) = \psi_0\Big(\sum_{i=1}^{d} \psi(\lambda_i)\Big), \qquad \mathrm{d}\rho_h(\boldsymbol{s}) = \int_0^{+\infty} \prod_{i=1}^{d} \mathrm{d}P_{\mathrm{ID}(t\rho)}(s_i)\,\mathrm{d}\rho_0(t).$$

In particular, if $\mathrm{ID}(t\rho)$ has p.d.f. $f_{\mathrm{ID}(t\rho)}$, then ρ_h has a Lévy density

$$\rho_h(s_1, \ldots, s_d) = \int_0^{+\infty} \prod_{i=1}^{d} f_{\mathrm{ID}(t\rho)}(s_i)\,\mathrm{d}\rho_0(t).$$

A sufficient condition for $\mathrm{ID}(t\rho)$ to have a probability density function, for every $t > 0$, is that the Lévy measure ρ has infinite mass and is absolutely continuous (Sato, 1999, Theorem 27.7).

Remark 2.4. The structure of the Lévy density above resembles that of *compound random measures* introduced in Griffin and Leisen (2017),

$$\rho_{\mathrm{co}}(s_1,\dots,s_d) = \int_0^{+\infty} \frac{1}{t^d}\, H\left(\frac{s_1}{t},\dots,\frac{s_d}{t}\right) \mathrm{d}\rho_0(t),$$

where H is a p.d.f. on $[0,+\infty)^d$. However, we show in Section S2.2 that there is no intersection between the two classes, at least for the same choice of base Lévy measure ρ_0.

Remark 2.5. Sampling from $\tilde{\boldsymbol{\mu}} \sim \mathrm{hCRV}(\rho, \rho_0, P_0)$ can be easily performed exploiting its hierarchical structure. Firstly, the base random measure $\tilde{\mu}_0$ can be sampled with different standard techniques; notably, when ρ_0 has infinite mass, $\tilde{\mu}_0$ comprises a countably infinite number of jumps, and we may only obtain an approximate sample. The most common strategies for infinitely active CRMs are based on the Ferguson-Klass representation (Ferguson and Klass, 1972), and sequentially sample their jumps in decreasing order, up to a certain truncation level L, yielding the discrete approximation

$$\tilde{\mu}_0^{\mathrm{approx}} = \sum_{\ell=1}^{L} \omega_{0\ell}\, \delta_{\phi_\ell},$$

with each $\phi_\ell \sim P_0$ independently. The practical implementations of such algorithms rely on the inversion of the tail of the Lévy measure (Wolpert and Ickstadt, 1998; Walker and Damien, 2000) or rejection sampling from a dominating Lévy density (Rosiński, 2001); we refer to Campbell et al. (2019); Zhang and Dassios (2025) for further details and alternative sampling techniques. Secondly, exploiting this approximation for $\tilde{\mu}_0$, the Lévy intensity of each $\tilde{\mu}_i \mid \tilde{\mu}_0^{\mathrm{approx}}$ can be decomposed as $\nu_i = \sum_{\ell=1}^{L} \omega_{0\ell}\rho\, \delta_{\phi_\ell}$. Since the sum of Lévy measures corresponds to a sum of independent CRMs, and a CRM $\tilde{\mu}$ with atomic base measure $P_0 = \delta_\phi$ satisfies $\tilde{\mu} = \tilde{\mu}(\mathbb{X})\delta_\phi$, then

$$\tilde{\mu}_i \overset{\mathrm{d}}{=} \sum_{\ell=1}^{L} \omega_{i\ell}\, \delta_{\phi_\ell},$$

where each $\omega_{i\ell} \sim \mathrm{ID}(\omega_{0\ell}\rho)$ independently. In other words, conditionally on a discrete approximation of the measure $\tilde{\mu}_0$, the vector $\tilde{\boldsymbol{\mu}}$ can be sampled exactly, provided that an exact simulation strategy for the infinitely divisible distribution $\mathrm{ID}(t\rho)$ for $t > 0$ is available. A detailed description of our sampling algorithms for the gamma-gamma hCRV is given in Sections S3.4 and S3.5, tailored to posterior sampling thanks to conditional conjugacy (Proposition 4.2).

The next result studies the identifiability of the parameters of a hierarchical CRV. For $c \in \mathbb{R}$, denote by $c_\#$ the pushforward measure of the multiplication map $s \mapsto cs$.

Theorem 2.6. *Let* $\tilde{\boldsymbol{\mu}}^{(\ell)} \sim \mathrm{hCRV}(\rho^{(\ell)}, \rho_0^{(\ell)}, P_0^{(\ell)})$, *for* $\ell = 1, 2$. *Then* $\tilde{\boldsymbol{\mu}}^{(1)} = \tilde{\boldsymbol{\mu}}^{(2)}$ *in distribution if and only if there exists* $c > 0$ *such that*

$$\rho_0^{(2)} = (c^{-1})_\# \rho_0^{(1)}, \qquad \rho^{(2)} = c\rho^{(1)}, \qquad P_0^{(1)} = P_0^{(2)}.$$

In particular, if $\rho^{(\ell)}$ and $\rho_0^{(\ell)}$ have Lévy densities, for $\ell = 1, 2$, this is equivalent to

$$\rho_0^{(2)}(s) = c\rho_0^{(1)}(cs), \qquad \rho^{(2)}(s) = c\rho^{(1)}(s), \qquad P_0^{(1)} = P_0^{(2)}.$$

The last condition can be easily checked on specific classes of models; for example, if we restrict to $\rho_0 = \alpha\rho$ for some $\alpha > 0$, then ρ and P_0 are identifiable. To clarify the interpretation of the identifiability conditions, note that $\rho_0^{(2)}(s) = c\rho_0^{(1)}(cs)$ if and only if $\tilde{\mu}_0^{(1)} = c\tilde{\mu}_0^{(2)}$.

We conclude this section with two leading examples of hierarchical CRVs.

Example 2.7. We term $\tilde{\boldsymbol{\mu}}$ a gamma-gamma hierarchical CRV if there exist shape parameters $\alpha, \alpha_0 > 0$, rate parameters $b, b_0 > 0$, and P_0 a base measure such that

$$\tilde{\mu}_1, \ldots, \tilde{\mu}_d \mid \tilde{\mu}_0 \overset{\text{iid}}{\sim} \text{CRM}\left(\alpha \frac{e^{-bs}}{s} \mathrm{d}s \otimes \tilde{\mu}_0\right); \qquad \tilde{\mu}_0 \sim \text{CRM}\left(\alpha_0 \frac{e^{-b_0 s}}{s} \mathrm{d}s \otimes P_0\right).$$

By applying Theorem 2.3 and the expression of the 1-dimensional Laplace exponent in Definition S1.3, the multivariate Laplace exponent of $\tilde{\boldsymbol{\mu}}$, for P_0 a probability measure, is

$$\psi_h(\lambda_1, \ldots, \lambda_d) = \alpha_0 \log\left(1 + \frac{\alpha}{b_0} \sum_{i=1}^{d} \log\left(1 + \frac{\lambda_i}{b}\right)\right).$$

Therefore, different gamma-gamma hCRVs coincide in distribution if they have the same ratio α/b_0. Moreover, since $t\rho(s) = t\alpha\, s^{-1} e^{-bs}$ is the Lévy density of a gamma CRM with shape parameter $t\alpha$ and scale parameter b, by Theorem 2.3 the multivariate Lévy density of $\tilde{\boldsymbol{\mu}}$ is

$$\rho_h(s_1, \ldots, s_d) = \alpha_0\, e^{-b\sum_{i=1}^{d} s_i} \int_0^{+\infty} \frac{b^{dt}}{\Gamma(t)^d} \prod_{i=1}^{d} s_i^{t-1}\, \frac{e^{-(b_0/\alpha)\, t}}{t}\, \mathrm{d}t.$$

Example 2.8. We term $\tilde{\boldsymbol{\mu}}$ a stable-stable hierarchical CRV if there exist shape parameters $\alpha, \alpha_0 > 0$, discount parameters $\sigma, \sigma_0 \in (0, 1)$, and a base measure P_0 such that

$$\tilde{\mu}_1, \ldots, \tilde{\mu}_d \mid \tilde{\mu}_0 \overset{\text{iid}}{\sim} \text{CRM}\left(\frac{\alpha\,\sigma}{\Gamma(1-\sigma)} \frac{1}{s^{1+\sigma}} \mathrm{d}s \otimes \tilde{\mu}_0\right);$$
$$\tilde{\mu}_0 \sim \text{CRM}\left(\frac{\alpha_0 \sigma_0}{\Gamma(1-\sigma_0)} \frac{1}{s^{1+\sigma_0}} \mathrm{d}s \otimes P_0\right).$$

Theorem 2.3 and Definition S1.4 imply that the multivariate Laplace exponent of $\tilde{\boldsymbol{\mu}}$ is

$$\psi_h(\lambda_1, \ldots, \lambda_d) = \alpha_0\, \alpha^{\sigma_0}\, (\lambda_1^{\sigma} + \cdots + \lambda_d^{\sigma})^{\sigma_0}.$$

Therefore, different stable-stable hCRVs coincide in distribution if they have the same value for $\alpha_0\, \alpha^{\sigma_0}$. For $d = 1$, we retrieve the Laplace functional of the marginal $\tilde{\mu}_i$, namely $\psi(\lambda) = \alpha_0\, \alpha^{\sigma_0}\, \lambda^{\sigma\sigma_0}$, which is the Laplace exponent of a stable CRM with shape $\alpha_0\, \alpha^{\sigma_0}$ and discount

parameter $\sigma\sigma_0$. Remarkably, we recover the well-known fact that the subordination of a stable Lévy process with a stable process is again a stable process (Bertoin, 1996; Sato, 1999; Camerlenghi et al., 2018). An explicit expression for the multivariate Lévy density is available only for $\sigma = 1/2$, as it requires the density of the stable infinitely divisible distribution. In such case, the Lévy density $t\rho(s) = t\alpha\,(4\pi)^{-1/2}s^{-3/2}$ is that of a Lévy distribution, and

$$\rho_h(s_1,\ldots,s_d) = \frac{\alpha_0\,\alpha^{\sigma_0}\sigma_0}{\pi^{d/2}\,2^{\sigma_0+1}}\frac{\Gamma(\frac{d-\sigma_0}{2})}{\Gamma(1-\sigma_0)}\prod_{i=1}^{d} s_i^{-3/2}\left(\sum_{i=1}^{d}\frac{1}{s_i}\right)^{-\frac{d-\sigma_0}{2}}.$$

3 Normalized hierarchical completely random vectors

One of the most common uses of completely random measures in Bayesian statistics is their normalization (Regazzini et al., 2003), which defines random probabilities whose law can act as nonparametric priors. The same construction can be extended to vectors of dependent random measures, such as hierarchical CRVs. In this section, we provide conditions for the normalization to be well-defined and investigate connections with popular models, such as the hierarchical Dirichlet process (Teh et al., 2006; Camerlenghi et al., 2019) and the hierarchical normalized σ-stable process (Camerlenghi et al., 2019). Moreover, we discuss general techniques to measure the dependence of normalized hierarchical CRVs.

For $\tilde{\boldsymbol{\mu}} \sim \mathrm{hCRV}(\rho, \rho_0, P_0)$, we derive a vector of dependent random probabilities as

$$\frac{\tilde{\boldsymbol{\mu}}}{\tilde{\boldsymbol{\mu}}(\mathbb{X})} := \left(\frac{\tilde{\mu}_1}{\tilde{\mu}_1(\mathbb{X})},\ldots,\frac{\tilde{\mu}_d}{\tilde{\mu}_d(\mathbb{X})}\right), \tag{1}$$

which is well-defined if $0 < \tilde{\mu}_i < +\infty$ a.s., for $i = 1,\ldots,d$. The upper bound forces P_0 to be a finite measure; thus, we can assume without loss of generality that P_0 is a probability measure. The lower bound forces each $\tilde{\mu}_i$ to be infinitely active, that is, the corresponding Lévy measures to have infinite mass.

Lemma 3.1. *Let $\tilde{\boldsymbol{\mu}} \sim \mathrm{hCRV}(\rho, \rho_0, P_0)$. Then each $\tilde{\mu}_i$ is infinitely active if and only if*

$$\int_0^{+\infty} \mathrm{d}\rho_0(t) = \int_0^{+\infty} \mathrm{d}\rho(t) = +\infty.$$

Therefore, $\tilde{\mu}_i$ is infinitely active if and only if both $\tilde{\mu}_0$ and $\tilde{\mu}_i \mid \tilde{\mu}_0$ are infinitely active.

Remark 3.2. The main subtlety of the proof of Lemma 3.1 is that ρ has a finite mass if and only if $\mathrm{ID}(t\rho)$ gives positive probability to $\{0\}$, as shown in Regazzini et al. (2003); in this case,

$$\int_{\Omega_d} \mathrm{d}P^d_{\mathrm{ID}(t\rho)} < \int_{[0,+\infty)^d} \mathrm{d}P^d_{\mathrm{ID}(t\rho)} = 1.$$

Hence, we need ρ to have infinite mass to conclude that $\int_{\Omega_d} \mathrm{d}\rho_h(\boldsymbol{s}) = \int_0^{+\infty} \mathrm{d}\rho_0(t)$ by Theorem 2.3 and Fubini-Tonelli theorem.

The construction in (1) is similar to the normalized hierarchical model in Camerlenghi et al. (2019); Catalano et al. (2024), where, however, the base random measure is normalized as well, that is

$$\tilde{\mu}_1, \ldots, \tilde{\mu}_d \mid \tilde{\mu}_0 \overset{\text{iid}}{\sim} \text{CRM}\left(\rho \otimes \frac{\tilde{\mu}_0}{\tilde{\mu}_0(\mathbb{X})}\right); \qquad \tilde{\mu}_0 \sim \text{CRM}(\rho_0 \otimes P_0). \tag{2}$$

This slight modification has crucial implications on the overall law of $\tilde{\boldsymbol{\mu}}$, which is no longer a CRV, and whose marginals $\tilde{\mu}_i$'s are not CRMs. Interestingly, at least two popular hierarchical specifications can be expressed in terms of a normalized hierarchical CRV, as discussed in the following.

Recall that $\tilde{\boldsymbol{P}} = (\tilde{P}_1, \ldots, \tilde{P}_d) \sim \text{HDP}(\alpha, \alpha_0, P_0)$ is a hierarchical Dirichlet process (Teh et al., 2006) with concentration parameters $\alpha, \alpha_0 > 0$ and base probability P_0 if

$$\tilde{P}_1, \ldots, \tilde{P}_d \mid \tilde{P}_0 \overset{\text{iid}}{\sim} \text{DP}(\alpha \tilde{P}_0); \qquad \tilde{P}_0 \sim \text{DP}(\alpha_0 P_0), \tag{3}$$

where $\text{DP}(\alpha_0 P_0)$ denotes a Dirichlet process (Ferguson, 1973) with base measure $\alpha_0 P_0$. In fact, the normalization of the gamma-gamma hCRV in Example 2.7 recovers the HDP with a specific gamma prior on the concentration parameter. Here, $\text{Gamma}(a, b)$ denotes the gamma distribution with shape a and rate b.

Proposition 3.3. *For parameters $\alpha, \alpha_0 > 0$ and $b, b_0 > 0$, and base measure P_0, let $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{P}}$ be vectors of random measures such that*

$$\begin{aligned} \tilde{\mu}_1, \ldots, \tilde{\mu}_d \mid \tilde{\mu}_0 &\overset{\text{iid}}{\sim} \text{CRM}\left(\alpha \frac{e^{-bs}}{s} \mathrm{d}s \otimes \tilde{\mu}_0\right); \qquad \tilde{\mu}_0 \sim \text{CRM}\left(\alpha_0 \frac{e^{-b_0 s}}{s} \mathrm{d}s \otimes P_0\right), \\ \tilde{P}_1, \ldots, \tilde{P}_d \mid \tilde{\alpha} &\overset{\text{iid}}{\sim} \text{HDP}(\tilde{\alpha}, \alpha_0, P_0); \qquad \tilde{\alpha} \sim \text{Gamma}(\alpha_0, b_0/\alpha). \end{aligned}$$

Then, with the notation in (1), it holds that $\tilde{\boldsymbol{\mu}}/\tilde{\boldsymbol{\mu}}(\mathbb{X}) \overset{d}{=} \tilde{\boldsymbol{P}}$.

The distribution of a normalized gamma-gamma hCRV depends in fact only on α_0 and α/b_0. The role of the ratio α/b_0 for identifiability is highlighted in Example 2.7, while b is a scale parameter for $\tilde{\boldsymbol{\mu}}$ and disappears with normalization; see also Figure 1 and Section S3.6. In practice, one may restrict to $b = b_0 = 1$, without loss of generality.

Additionally, when the idiosyncratic component is a stable CRM, the normalized hierarchical model of Camerlenghi et al. (2019) can be expressed as a normalized hierarchical CRV. This result generalizes, with a different technique, a result of Camerlenghi et al. (2018) which assumes the base CRM to be stable as well; this same fact is also observed for Lévy processes in Bertoin (1996). The key property is that, for a stable CRM $\tilde{\mu}$ with Lévy

measure ρ, any proportional measure $c\tilde{\mu}$ with $c > 0$ is a stable CRM with proportional Lévy measure $c'\rho$, for some $c' > 0$.

Proposition 3.4. *Let $\tilde{\boldsymbol{\mu}}^{(1)}$ and $\tilde{\boldsymbol{\mu}}^{(2)}$ be two vectors of random measures defined by*

$$\tilde{\mu}_1^{(1)}, \ldots, \tilde{\mu}_d^{(1)} \mid \tilde{\mu}_0 \overset{\text{iid}}{\sim} \text{CRM}\left(\frac{\alpha\,\sigma}{\Gamma(1-\sigma)}\frac{1}{s^{\sigma+1}}\mathrm{d}s \otimes \tilde{\mu}_0\right);$$
$$\tilde{\mu}_1^{(2)}, \ldots, \tilde{\mu}_d^{(2)} \mid \tilde{\mu}_0 \overset{\text{iid}}{\sim} \text{CRM}\left(\frac{\alpha\,\sigma}{\Gamma(1-\sigma)}\frac{1}{s^{\sigma+1}}\mathrm{d}s \otimes \frac{\tilde{\mu}_0}{\tilde{\mu}_0(\mathbb{X})}\right),$$

where $\alpha > 0$, $\sigma \in (0,1)$, and $\tilde{\mu}_0$ is an infinitely active CRM*. Then,*

$$\frac{\tilde{\boldsymbol{\mu}}^{(1)}}{\tilde{\boldsymbol{\mu}}^{(1)}(\mathbb{X})} \overset{d}{=} \frac{\tilde{\boldsymbol{\mu}}^{(2)}}{\tilde{\boldsymbol{\mu}}^{(2)}(\mathbb{X})}.$$

Remark 3.5. There is a strong connection between hCRVs and the hierarchical models in (2), even beyond the two specifications discussed in Proposition 3.3 and Proposition 3.4. Indeed, conditionally on $\tilde{\mu}_0$, a hierarchical CRV satisfies

$$\tilde{\mu}_1, \ldots, \tilde{\mu}_d \mid \tilde{\mu}_0 \overset{\text{iid}}{\sim} \text{CRM}(\rho \otimes \tilde{\mu}_0) \equiv \text{CRM}\left(\tilde{\mu}_0(\mathbb{X})\,\rho \otimes \frac{\tilde{\mu}_0}{\tilde{\mu}_0(\mathbb{X})}\right);$$

therefore, a hierarchical CRV can be regarded as a hierarchical model in (2) whose idiosyncratic Lévy measure is endowed with a hyperprior on the concentration parameter, that is,

$$\tilde{\mu}_1, \ldots, \tilde{\mu}_d \mid \tilde{\mu}_0 \overset{\text{iid}}{\sim} \text{CRM}(\tilde{\alpha}\rho \otimes \tilde{P}_0), \qquad (\tilde{\alpha}, \tilde{P}_0) = \left(\tilde{\mu}_0(\mathbb{X}), \frac{\tilde{\mu}_0}{\tilde{\mu}_0(\mathbb{X})}\right).$$

The subtlety of this hyperprior with respect to standard proposals is that, in general, it is dependent on the base measure $\tilde{P}_0$, since $\tilde{\mu}_0(\mathbb{X})$ and $\tilde{\mu}_0/\tilde{\mu}_0(\mathbb{X})$ are dependent. The only exception is the gamma CRM, which leads to Proposition 3.3.

In the following, we provide closed form expression for summary statistics encoding the mean, the variance, and the dependence structure of hCRVs. The simplest and most widely used measure of dependence between random probabilities $\tilde{P} = \tilde{\boldsymbol{\mu}}/\tilde{\boldsymbol{\mu}}(\mathbb{X})$ is the pairwise linear correlation $\text{Corr}(\tilde{P}_i(A), \tilde{P}_j(A))$, for a Borel set A. Note that its expression does not depend on A, since normalized hCRVs are normalized homogeneous CRVs and thus belong to the class of multivariate species sampling processes (Franzolini et al., 2025). At the level of the random measures $\tilde{\boldsymbol{\mu}}$, its computation is a slight modification of Catalano et al. (2024, Section 8), where the expressions are derived by leveraging on the conditional independence structure. We report these results in Proposition S2.1, and present an alternative proof that builds on their joint infinite divisibility. Interestingly, this second technique greatly simplifies the derivation in the normalized case.

Proposition 3.6. *Let $\tilde{\boldsymbol{\mu}} \sim \mathrm{hCRV}(\rho, \rho_0, P_0)$, and let ψ and ψ_0 denote the Laplace exponents of ρ and ρ_0, respectively. For any Borel set A and every $i \neq j$, the normalization $\tilde{\boldsymbol{P}} = \tilde{\boldsymbol{\mu}}/\tilde{\boldsymbol{\mu}}(\mathbb{X})$ satisfies $\mathbb{E}(\tilde{P}_i(A)) = P_0(A)$ and*

$$\mathrm{Var}(\tilde{P}_i(A)) = -P_0(A)(1-P_0(A)) \int_0^{+\infty} u\, e^{-\psi_0(\psi(u))} (\psi_0 \circ \psi)''(u)\, \mathrm{d}u,$$
$$\mathrm{Cov}(\tilde{P}_i(A), \tilde{P}_j(A)) = \mathrm{Var}\left(\frac{\tilde{\mu}_0}{\tilde{\mu}_0(\mathbb{X})}\right) = -P_0(A)(1-P_0(A)) \int_0^{+\infty} u\, e^{-\psi_0(u)} \psi_0''(u)\, \mathrm{d}u,$$

where $(\psi_0 \circ \psi)''(u) = \psi_0''(\psi(u))\, \psi'(u)^2 + \psi_0'(\psi(u))\, \psi''(u)$. Moreover, if $P_0(A) \notin \{0, 1\}$,

$$\mathrm{Corr}(\tilde{P}_i(A), \tilde{P}_j(A)) = \frac{\int_0^{+\infty} u\, e^{-\psi_0(u)} \psi_0''(u)\, \mathrm{d}u}{\int_0^{+\infty} u\, e^{-\psi_0(\psi(u))} (\psi_0 \circ \psi)''(u)\, \mathrm{d}u}.$$

Example 3.7. Let $\tilde{\boldsymbol{\mu}}$ be a gamma-gamma hCRV as in Example 2.7 with $b = b_0 = 1$, and let A be such that $P_0(A) \notin \{0, 1\}$. The marginal and mixed moments of $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{P}} = \tilde{\boldsymbol{\mu}}/\tilde{\boldsymbol{\mu}}(\mathbb{X})$ are the following.

Moments of the unnormalized measures

$$\mathbb{E}\big(\tilde{\mu}_i(A)\big) = \alpha_0 \alpha\, P_0(A), \qquad \mathrm{Var}\big(\tilde{\mu}_i(A)\big) = \alpha_0 \alpha(1+\alpha)\, P_0(A),$$
$$\mathrm{Cov}\big(\tilde{\mu}_i(A), \tilde{\mu}_j(A)\big) = \alpha_0 \alpha^2\, P_0(A),$$
$$\mathrm{Corr}\big(\tilde{\mu}_i(A), \tilde{\mu}_j(A)\big) = \frac{\alpha}{1+\alpha}.$$

Moments of the normalized random probabilities

$$\mathbb{E}\big(\tilde{P}_i(A)\big) = P_0(A), \qquad \mathrm{Var}\big(\tilde{P}_i(A)\big) = \Big(1 + \frac{\alpha_0}{\alpha} e^{1/\alpha} E_{\alpha_0}\big(\tfrac{1}{\alpha}\big)\Big) \frac{P_0(A)(1-P_0(A))}{1+\alpha_0},$$
$$\mathrm{Cov}\big(\tilde{P}_i(A), \tilde{P}_j(A)\big) = \frac{P_0(A)(1-P_0(A))}{1+\alpha_0},$$
$$\mathrm{Corr}\big(\tilde{P}_i(A), \tilde{P}_j(A)\big) = \Big(1 + \frac{\alpha_0}{\alpha} e^{1/\alpha} E_{\alpha_0}\big(\tfrac{1}{\alpha}\big)\Big)^{-1},$$

where $E_\eta(x) = \int_1^{+\infty} t^{-\eta} e^{-tx} \mathrm{d}t = x^{\eta-1} \Gamma(1-\eta, x)$ is the generalized exponential integral.

4 Posterior representation for normalized hCRVs

Vectors of dependent random probability measures are commonly employed in Bayesian statistics to model partially exchangeable observations. Indeed, any infinitely active CRV is suitable for this scope through normalization (1). Many models in the literature fall within this framework, including GM-dependent measures (Lijoi and Nipoti, 2014; Lijoi et al., 2014), compound random measures (Griffin and Leisen, 2017), Lévy copulas (Epifani and Lijoi, 2010), and thinned random measures (Lau and Cripps, 2022). In this section,

we derive the expression of the posterior distribution for a generic normalized CRV; this can be seen as the multivariate extension of James et al. (2009) and a special case of FuRBI random measures (Ascolani et al., 2024) with shared atoms. These results are later specialized to hierarchical normalized CRVs, and their posterior structure is explored in greater detail.

Let $\boldsymbol{X}_i = (X_{i1}, \dots, X_{in_i})$ be the i-th group of observations, for $i = 1, \dots, d$, and consider the model

$$\boldsymbol{X}_1, \dots, \boldsymbol{X}_d \mid \tilde{\boldsymbol{\mu}} \sim \left(\frac{\tilde{\mu}_1}{\tilde{\mu}_1(\mathbb{X})}\right)^{n_1} \times \cdots \times \left(\frac{\tilde{\mu}_d}{\tilde{\mu}_d(\mathbb{X})}\right)^{n_d}; \qquad \tilde{\boldsymbol{\mu}} \sim \mathrm{CRV}(\nu), \tag{4}$$

where P^m denotes the m-fold product measure and $\mathrm{d}\nu(\boldsymbol{s}, x) = \mathrm{d}\rho_x(\boldsymbol{s})\,\mathrm{d}P_0(x)$ is a multivariate Lévy intensity with P_0 a diffuse probability. In the following, we often use the compact notation $\boldsymbol{X}_{1:d} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_d)$.

The almost-sure discreteness and the dependence structure of the random measures $\tilde{\boldsymbol{\mu}}$ imply that the observations $\boldsymbol{X}_{1:d}$ display tied values with positive probability, both within and across groups. Let $\boldsymbol{x}^* = (x_1^*, \dots, x_k^*)$ be the distinct values taken by the observed values $\boldsymbol{x}_{1:d}$, that is, for every $j = 1, \dots, k$, there exist x_{ih} such that $x_{ih} = x_j^*$, and $x_j^* \neq x_\ell^*$ for every $j \neq \ell$. Moreover, for every $i = 1, \dots, d$ and $j = 1, \dots, k$, denote by n_{ij} the number of observed values in $\boldsymbol{x}_i$ equal to x_j^*. This implies that, for every $i = 1, \dots, d$, the decomposition $n_i = n_{i1} + \cdots + n_{ik}$ holds. Finally, for each $\boldsymbol{m} \in \mathbb{N}^d$, $x \in \mathbb{X}$ and $\boldsymbol{u} \in \Omega_d$, define the cumulant

$$\tau_{\boldsymbol{m}|x}(\boldsymbol{u}) = \int_{\Omega_d} e^{-\boldsymbol{u}\cdot\boldsymbol{s}} \prod_{i=1}^{d} s_i^{m_i} \mathrm{d}\rho_x(\boldsymbol{s}). \tag{5}$$

The next theorem provides a general expression for the posterior of $\tilde{\boldsymbol{\mu}} \mid \boldsymbol{X}_{1:d}$, showing that it preserves the CRV property, conditionally on a set of dependent latent variables. For this purpose, define a vector of latent variables $\boldsymbol{U} = (U_1, \dots, U_d)$ with joint p.d.f.

$$f_{\boldsymbol{U}}(\boldsymbol{u}) \propto e^{-\psi(\boldsymbol{u})} \prod_{i=1}^{d} u_i^{n_i - 1} \prod_{j=1}^{k} \tau_{n_{1j}, \dots, n_{dj}|x_j^*}(\boldsymbol{u}), \tag{6}$$

where ψ denotes the Laplace exponent of $\tilde{\boldsymbol{\mu}}$.

Theorem 4.1. *Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_d$ follow the model (4) with $\mathrm{d}\nu(\boldsymbol{s}, x) = \mathrm{d}\rho_x(\boldsymbol{s})\,\mathrm{d}P_0(x)$, where P_0 is a diffuse probability and ρ_x is an infinitely active Lévy measure P_0-a.s. Then, there exist $\boldsymbol{U}$ with p.d.f. (6) such that*

$$\mathcal{L}(\tilde{\boldsymbol{\mu}} \mid \boldsymbol{X}_{1:d} = \boldsymbol{x}_{1:d}) = \mathcal{L}\left(\tilde{\boldsymbol{\mu}}^* + \sum_{j=1}^{k} \boldsymbol{J}_j \delta_{x_j^*}\right),$$

where $\tilde{\boldsymbol{\mu}}^*$ and $(\boldsymbol{J}_1, \dots, \boldsymbol{J}_k)$ are conditionally independent given $\boldsymbol{U}$, and such that

(i) the vector $\tilde{\boldsymbol{\mu}}^* \mid \boldsymbol{U}$ *is a* CRV *with Lévy intensity* $\mathrm{d}\nu^*_{\boldsymbol{U}}(\boldsymbol{s}, x) = e^{-\boldsymbol{U}\cdot\boldsymbol{s}}\,\mathrm{d}\nu(\boldsymbol{s}, x)$*;*

(ii) for $j = 1, \ldots, k$*, the vector* $\boldsymbol{J}_j = (J_{1j}, \ldots, J_{dj})$ *contains the jumps at the shared fixed point of discontinuity* x^*_j*, whose conditional distribution satisfies*

$$\mathrm{d}P_{\boldsymbol{J}_j|\boldsymbol{U}}(\boldsymbol{s}) \propto e^{-\boldsymbol{U}\cdot\boldsymbol{s}} \prod_{i=1}^{d} s_i^{n_{ij}}\,\mathrm{d}\rho_{x^*_j}(\boldsymbol{s}). \tag{7}$$

Note that the Lévy measure ρ_x does not need a density on Ω_d. This may be seen as a technical detail on $(0, +\infty)$, but it is relevant on Ω_d, as some popular models, such as GM-dependent measures, do not admit a density; see Catalano et al. (2024, Lemma 6).

We now specialize Theorem 4.1 to hierarchical CRVs. For simplicity, we assume that ρ and ρ_0 have Lévy densities on $(0, +\infty)$, denoted with the same notation; this implies that both $\mathrm{ID}(\rho)$ and $\mathrm{ID}(\rho_0)$ have a p.d.f. (Sato, 1999, Theorem 27.7). A first result shows the conditional quasi-conjugacy of the model. Indeed, conditionally on $\boldsymbol{U}$, we can interpret the vector $\tilde{\boldsymbol{\mu}}^*$ as a hierarchical CRV with heterogeneous marginal distributions.

Proposition 4.2. *Let* ρ *and* ρ_0 *have Lévy densities and let* $\tilde{\boldsymbol{\mu}} \sim \mathrm{hCRV}(\rho, \rho_0, P_0)$*. Conditionally on* $\boldsymbol{U}$*, the* CRV $\tilde{\boldsymbol{\mu}}^*$ *in Theorem 4.1 satisfies*

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}^* \mid \tilde{\mu}^*_0, \boldsymbol{U} &\sim \prod_{i=1}^{d} \mathrm{CRM}\left(e^{-U_i s}\rho(s)\,\mathrm{d}s \otimes \tilde{\mu}^*_0\right); \\
\tilde{\mu}^*_0 \mid \boldsymbol{U} &\sim \mathrm{CRM}\left(e^{-\sum_{i=1}^{d}\psi(U_i)\,s}\rho_0(s)\,\mathrm{d}s \otimes P_0\right).
\end{aligned}$$

In order to sample from the posterior, one has to sample the jumps $\boldsymbol{J}_j$'s at fixed locations and the latent variables $\boldsymbol{U}$, which are d-dimensional random variables. When d is small, the sampling task can be performed via d-dimensional rejection sampling or approximated by a Metropolis-Hastings scheme. However, when d is moderate or large, reducing the dimension of the proposal becomes essential to preserve efficiency. For hierarchical CRVs, the sampling of jumps can be reduced to the sampling of 1-dimensional random variables. To this end, define the quantity

$$\bar{\tau}_m(u, t) = \int_0^{+\infty} s^m e^{-us} f_{\mathrm{ID}(t\rho)}(s)\,\mathrm{d}s. \tag{8}$$

Remarkably, the quantity above may be computed without integration. Indeed, whenever an explicit expression for the Laplace transform of $\mathrm{ID}(t\rho)$ is available, one may instead compute its derivatives, since

$$\bar{\tau}_m(u, t) = (-1)^m \frac{\mathrm{d}^m}{\mathrm{d}u^m} \int_0^{+\infty} e^{-us} f_{\mathrm{ID}(t\rho)}(s)\,\mathrm{d}s.$$

Proposition 4.3. *Let ρ and ρ_0 have Lévy densities and let $\tilde{\boldsymbol{\mu}} \sim \mathrm{hCRV}(\rho, \rho_0, P_0)$. Conditionally on the latent variables $\boldsymbol{U}$ and for every $j = 1, \dots, k$, the jumps $\boldsymbol{J}_j = (J_{1j}, \dots, J_{dj})$ in Theorem 4.1 satisfy*

$$J_{1j}, \dots, J_{dj} \mid \boldsymbol{U}, J_{0j} \;\sim\; f_{\boldsymbol{J}_j|\boldsymbol{U},J_{0j}}(\boldsymbol{s}) = \prod_{i=1}^{d} \frac{s_i^{n_{ij}} e^{-U_i s_i} f_{\mathrm{ID}(J_{0j}\rho)}(s_i)}{\bar{\tau}_{n_{ij}}(U_i, J_{0j})},$$

where J_{0j} is a random variable having p.d.f. $f_{J_{0j}|\boldsymbol{U}}(t) \propto \prod_{i=1}^{d} \bar{\tau}_{n_{ij}}(U_i, t)\, \rho_0(t)$.

Therefore, Proposition 4.3 reduces the sampling of each d-dimensional jumps vector to the easier task of sampling d conditionally independent 1-dimensional jumps, given an additional variable. These random variables can be sampled exactly in the case of gamma-gamma hCRVs, as detailed in Section 5.

As for the latent variables $\boldsymbol{U}$, the evaluation of their p.d.f. (6) up to a normalizing constant requires to compute the cumulants $\tau_{\boldsymbol{m}|x}$ in (5). Although this task can be tackled on a case-by-case basis, the d-dimensional integral in their definition can be reduced to a 1-dimensional integral for hierarchical CRVs. Given their homogeneity, the cumulant $\tau_{\boldsymbol{m}|x} = \tau_{\boldsymbol{m}}$ does not depend on $x \in \mathbb{X}$.

Lemma 4.4. *Let ρ and ρ_0 have Lévy densities on $(0, +\infty)$ and let $\tilde{\boldsymbol{\mu}} \sim \mathrm{hCRV}(\rho, \rho_0, P_0)$. For $\boldsymbol{m} \in \mathbb{N}^d$ and $\boldsymbol{u} \in \Omega_d$, the cumulants in (5) are expressed as*

$$\tau_{\boldsymbol{m}}(\boldsymbol{u}) = \int_0^{+\infty} \prod_{i=1}^{d} \bar{\tau}_{m_i}(u_i, t) \rho_0(t)\, \mathrm{d}t.$$

Interestingly, the cumulant $\tau_{n_{1j},\dots,n_{dj}}(\boldsymbol{U})$ is the normalizing constant for the density of the jump J_{0j} in Proposition 4.3. These quantities represent the building blocks for any computational method approximating the d-dimensional distribution of latent variables $\boldsymbol{U}$. However, when d is large, reducing to lower-dimensional distributions becomes essential. Therefore, we propose an alternative representation of these latent variables as conditionally independent gamma random variables with random rate parameters; this result may also be of interest in the univariate case. For $\boldsymbol{m} \in \mathbb{N}^k$ and $\boldsymbol{t} \in (0, +\infty)^{k+1}$, set $m_\bullet = \sum_{j=1}^{k} m_j$ and define the integral

$$C(\boldsymbol{m}; \boldsymbol{t}) = \int_{(0,+\infty)^{k+1}} (s_0 + s_1 + \cdots + s_k)^{-m_\bullet} f_{\mathrm{ID}(t_0\rho)}(s_0) \prod_{j=1}^{k} s_j^{m_j} f_{\mathrm{ID}(t_j\rho)}(s_j)\, \mathrm{d}\boldsymbol{s}. \tag{9}$$

Proposition 4.5. *Let ρ and ρ_0 have Lévy densities and let $\tilde{\boldsymbol{\mu}} \sim \mathrm{hCRV}(\rho, \rho_0, P_0)$. Then,*

$$U_1, \dots, U_d \mid \beta_1, \dots, \beta_d \sim \prod_{i=1}^{d} \mathrm{Gamma}(n_i, \beta_i),$$

where $\beta_i = S_{i0} + S_{i1} + \ldots, S_{ik}$, *for each* $i = 1, \ldots, d$. *Moreover, the* $\boldsymbol{S}_i = (S_{i0}, \ldots, S_{ik})$ *are conditionally independent vectors given* $\boldsymbol{T} = (T_0, T_1, \ldots, T_k)$, *and their p.d.f.s satisfy*

$$f_{\boldsymbol{S}_i|\boldsymbol{T}}(s_{i0}, \ldots, s_{ik}) \propto (s_{i0} + s_{i1} + \cdots + s_{ik})^{-n_i} f_{\mathrm{ID}(T_0\rho)}(s_{i0}) \prod_{j=1}^{k} s_{ij}^{n_{ij}} f_{\mathrm{ID}(T_j\rho)}(s_{ij}),$$

$$f_{\boldsymbol{T}}(t_0, \ldots, t_k) \propto \prod_{i=1}^{d} C(n_{i1}, \ldots, n_{ik}; \boldsymbol{t}) f_{\mathrm{ID}(\rho_0)}(t_0) \prod_{j=1}^{k} \rho_0(t_j).$$

Proposition 4.5 reduces the sampling of the d-dimensional vector of dependent latent variables $\boldsymbol{U}$ to the sampling of d conditionally independent 1-dimensional random variables $\beta_1, \ldots, \beta_d$, which represent the scale parameters of $U_1, \ldots, U_d$. This sampling task can be approached through standard computational methods or further simplified, as described in the following section.

5 The normalized gamma-gamma hCRV

5.1 Posterior representation

This section specifies results in Section 4 to the fundamental example of gamma-gamma hCRVs, and provides details for the subsequent implementation of posterior sampling algorithms. In the following, $((t))_n = \Gamma(t+n)/\Gamma(t)$ denotes the ascending factorial.

Proposition 5.1. *Let* $\tilde{\boldsymbol{\mu}}$ *be a gamma-gamma hCRV, as in Example 2.7, and define*

$$\lambda(\boldsymbol{U}) = \frac{b_0}{\alpha} + \sum_{i=1}^{d} \log\left(1 + \frac{U_i}{b}\right).$$

Then, conditionally on the latent variables $\boldsymbol{U}$, *and for each* $j = 1, \ldots, k$,

(a) the CRV $\tilde{\boldsymbol{\mu}}^*$ *in Proposition 4.2 is a hierarchy of conditionally gamma CRMs,*

$$\tilde{\mu}_1^*, \ldots, \tilde{\mu}_d^* \mid \tilde{\mu}_0^*, \boldsymbol{U} \sim \prod_{i=1}^{d} \mathrm{CRM}\left(\alpha\, s^{-1}\, e^{-b(1+U_i/b)\, s} \mathrm{d}s \otimes \tilde{\mu}_0^*\right);$$
$$\tilde{\mu}_0^* \mid \boldsymbol{U} \sim \mathrm{CRM}\left(\alpha_0\, s^{-1}\, e^{-\alpha\, \lambda(\boldsymbol{U})\, s} \mathrm{d}s \otimes P_0\right);$$

(b) the jumps $\boldsymbol{J}_j$ *in Proposition 4.3 are conditionally independent and, for* $i = 1, \ldots, d$,

$$J_{ij} \mid U_i, J_{0j} \sim \mathrm{Gamma}\big(\alpha J_{0j} + n_{ij}, b\,(1 + U_i/b)\big);$$

(c) the density of the rescaled random variable αJ_{0j} is proportional to

$$f_{\alpha J_{0j}|\boldsymbol{U}}(t) \propto t^{-1} e^{-\lambda(\boldsymbol{U})\, t} \prod_{i=1}^{d} ((t))_{n_{ij}}. \tag{10}$$

Details on sampling algorithms for the hierarchy in (a) are given in Sections S3.4 and S3.5, where we describe an efficient implementation of Newton's method for inverting the exponential integral, based on a logarithmic transformation that guarantees convergence for each starting point. Notably, this same algorithm can be readily employed to sample *a priori* from a gamma-gamma hCRV; see Remark 2.5 and Section S3.4. Moreover, the nontrivial task of sampling the jumps $\boldsymbol{J}_j$'s is here reduced to sampling variables αJ_{0j}'s in (c), whose densities can be computed up to normalizing constants. A convenient option is to resort to computational methods that generate approximate samples, such as Metropolis-Hastings algorithms. Details are provided in Section S3.1, where we empirically show that the random-walk Metropolis-Hastings scheme on the log-scale with Gaussian increments outperforms the approach proposed in Barrios et al. (2013), based on gamma proposals. Remarkably, we also develop an exact sampler in Section 5.2.

The remaining step to sample from the posterior is sampling the latent variables $\boldsymbol{U}$, whose characterization in Proposition 4.5 can be simplified through a change of variables.

Proposition 5.2. *Let $\tilde{\boldsymbol{\mu}}$ be a gamma-gamma hierarchical* CRV*. The distribution of latent variables $\boldsymbol{U} = (U_1, \ldots, U_d)$ can be decomposed as follows:*

(a) for each $i = 1, \ldots, d$, one has $U_i \mid \beta_i \sim \text{Gamma}(n_i, \beta_i)$, with $\beta_i \mid T \sim \text{Gamma}(\alpha T, b)$;

(b) the density of αT is proportional to

$$f_{\alpha T}(t) \propto t^{\alpha_0 - 1} e^{-(b_0/\alpha)\, t} \prod_{i=1}^{d} \frac{1}{((t))_{n_i}} \int_{\Delta^k} v_0^{\alpha_0 - 1} \prod_{j=1}^{k} \left(v_j^{-1} \prod_{i=1}^{d} ((t v_j))_{n_{ij}} \right) \mathrm{d}\boldsymbol{v}, \tag{11}$$

where $\boldsymbol{v} = (v_0, \ldots, v_k)$ is an auxiliary vector on k-dimensional unit simplex Δ^k.

For implementation convenience, one can directly sample U_i/b. Indeed, these are the relevant quantities in Proposition 5.1 for sampling the jumps $J_{i1}, \ldots, J_{ik}$ and the random measure $\tilde{\mu}_i^*$, and computing $\lambda(\boldsymbol{U})$. The non-standard step for sampling $\boldsymbol{U}$ is the marginal sampling of the variable αT in (11), whose joint density with the auxiliary vector $\boldsymbol{V}$ is known up to a normalizing constant. In Section S3.1 we describe a Metropolis-Hastings within Gibbs algorithm to sample from the marginal distribution of αT. Alternative procedures to obtain exact samples from αT are considered in Section 5.2.

5.2 Exact sampling

As highlighted by Propositions 5.1-5.2, posterior sampling from the normalized gamma-gamma hCRV mainly consists in sampling gamma random variables, except for two critical steps, namely the sampling of random variables $\alpha J_{01}, \ldots, \alpha J_{0k}$ in (10) and the marginal sampling of random variable αT in (11). In both cases, one may resort to MCMC algorithms, based on Metropolis-Hastings steps, to obtain approximate samples from such distributions. In the following, we propose alternative algorithms based on analytical manipulations of their density functions, which instead allow for exact sampling. For this purpose, introduce the coefficients $S(q_1, \ldots, q_d; h)$, defined by the recursive relation

$$S(q_1, \ldots, q_\ell + 1, \ldots, q_d; h) = q_\ell S(q_1, \ldots, q_d; h) + S(q_1, \ldots, q_d; h-1), \tag{12}$$

with boundary conditions $S(0, \ldots, 0; 0) = 1$ and $S(q_1, \ldots, q_d; 0) = S(0, \ldots, 0; h) = 0$ for $q_\bullet > 0$ or $h > 0$. Remarkably, these coefficients are the natural generalization of the unsigned Stirling numbers of the first kind to blocked permutations. Indeed, $S(q_1, \ldots, q_d; h)$ represents the number of permutations with h cycles in the Young subgroup of $\mathcal{S}_{q_\bullet}$ (group of permutations of $q_\bullet$ elements) induced by the integer partition $(q_1, \ldots, q_d)$. Henceforth, we refer to the coefficients $S(q_1, \ldots, q_d; h)$ as multivariate Stirling numbers. For convenience, let $m_{ij} = \min(1, n_{ij}) \in \{0, 1\}$, and define $m_{\bullet j} = \sum_{i=1}^{d} m_{ij}$.

Proposition 5.3. *For each $j = 1, \ldots, k$, the density of αJ_{0j} in (10) can be written as*

$$f_{\alpha J_{0j}|\boldsymbol{U}}(t) \propto \sum_{h_j = m_{\bullet j}}^{n_{\bullet j}} S(n_{1j}, \ldots, n_{dj}; h_j)\; t^{h_j - 1} e^{-\lambda(\boldsymbol{U})\, t}.$$

Therefore, the random variable αJ_{0j} is in fact a finite mixture of gamma distributions, with mixing weights $p_{jh_j} \propto \lambda(\boldsymbol{U})^{-h_j}\, \Gamma(h_j)\, S(n_{1j}, \ldots, n_{dj}; h_j)$, for $h_j = m_{\bullet j}, \ldots, n_{\bullet j}$. The evaluation of the $S(n_{1j}, \ldots, n_{dj}; h_j)$'s via the recursive relation in (12) has computational cost $\mathcal{O}\big(n_{\bullet j}^2\big)$; see Remark S2.3.

Proposition 5.4. *The density function of random variable αT in (11) can be written as*

$$f_{\alpha T}(t) \propto t^{\alpha_0 - 1}\, e^{-(b_0/\alpha)\, t} \prod_{i=1}^{d} \frac{1}{((t))_{n_i}} \left(\sum_{h=m}^{n} \frac{c_h}{((\alpha_0))_h}\, t^h \right), \tag{13}$$

where $m = \sum_{j=1}^{k} m_{\bullet j}$ and each coefficient c_h is defined, for $S(q_1, \ldots, q_d; h)$ in (12), by

$$c_h = \sum_{\substack{h_1 + \cdots + h_k = h \\ m_{\bullet j} \leq h_j \leq n_{\bullet j}}} \prod_{j=1}^{k} \Gamma(h_j)\, S(n_{1j}, \ldots, n_{dj}; h_j).$$

The evaluation of the c_h's may seem computationally expensive when k is large. Indeed, a

naive approach would involve k nested cycles, with a computational cost of $\mathcal{O}\big(\prod_{j=1}^{k} n_{\bullet j}\big)$. However, these coefficients can be obtained through a sequence of discrete convolutions and computed at cost $\mathcal{O}\big(\sum_{j<\ell} n_{\bullet j} n_{\bullet \ell}\big)$, as detailed in Section S3.2. Exact samples from the distribution of αT can be obtained resorting to rejection sampling algorithms. Indeed, considering a real parameter r, (13) can be rewritten as

$$f_{\alpha T}(t) \propto t^{\alpha_0+r-1}\, e^{-(b_0/\alpha)\, t}\, (t^{-r} R(t)), \qquad R(t) = \prod_{i=1}^{d} \frac{1}{((t))_{n_i}} \left(\sum_{h=m}^{n} \frac{c_h}{((\alpha_0))_h}\, t^h \right),$$

where $R(t)$ is a ratio of polynomials, and continuous for $t \geq 0$. The distribution of αT is absolutely continuous with respect to a Gamma$(\alpha_0 + r, b_0/\alpha)$, with Radon-Nikodym derivative proportional to $t^{-r} R(t)$, from which we construct a rejection sampling algorithm. A necessary condition for the Radon-Nikodym derivative is to be bounded above: this is guaranteed when $0 \leq r \leq m - d$. Within this interval, we choose the value of r that maximizes the acceptance probability, which is equivalent to maximizing

$$r \log t^*(r) - \log R(t^*(r)) + (\alpha_0 + r) \log(b_0/\alpha) - \log \Gamma(\alpha_0 + r),$$

where $t^*(r)$ is the value of t that maximizes $t^{-r} R(t)$. Details on the optimal choice of r can be found in Section S3.3. Alternatively, the adaptive rejection sampling algorithm of Gilks and Wild (1992) may considerably improve the acceptance rate. Although computationally convenient, theoretical guarantees are only available if the density of αT is logarithmically concave, which we were unable to prove in general. Therefore, for numerical experiments, we consider the standard rejection sampler outlined above.

5.3 Posterior sampling algorithms

Collecting the distributional results described above, as well as the details in Section S3, we obtain algorithms to sample jumps $\boldsymbol{J}$ at fixed locations in the posterior distribution (Proposition 4.2) induced by the gamma-gamma hCRV. The Julia implementation of these algorithms is available at github.com/claudiodelsole/hCRV.jl, together with an interface to allow their integration within the R environment. The common structure of conditional dependencies within the sampling procedures is summarized in Figure 1, where the computational bottlenecks are highlighted by red circles. Remarkably, the graph of dependencies has no proper cycles, i.e. it is a directed acyclic graph (DAG): variables are sampled in topological order, with each variable depending only on (some of) its ancestors. Such dependence structure may enhance mixing. Furthermore, the proposed algorithms allow for parallelization at different stages. Indeed, after the initial sampling of αT from its marginal, one can parallelize across groups to sample the latent variables $\boldsymbol{U}$ (Proposition 5.2). Likewise, after the computation of $\lambda(\boldsymbol{U})$, one can parallelize across

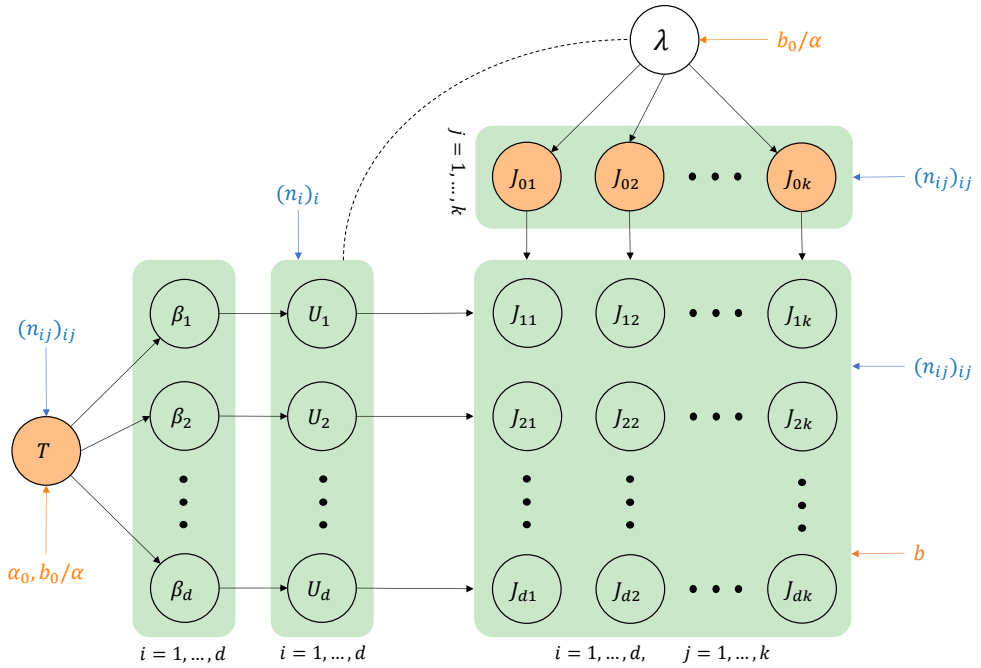

Figure 1: DAG of conditional dependencies between random variables in Algorithms 1 and 2. Red circles represent computational bottlenecks; quantities in empty circles are sampled from gamma distributions. Some variables are reported up to scaling w.r.t. model parameters, e.g. T instead of αT.

distinct values to sample the conditionally independent variables $\alpha J_{01}, \dots, \alpha J_{0k}$ and further parallelize across groups and distinct values to sample jumps $\boldsymbol{J}$ (Proposition 5.1).

The MCMC scheme for posterior sampling the jumps $\boldsymbol{J}$ is summarized in Algorithm 1. The state of the Markov chain consists of the latent variable αT and jumps $\alpha J_{01}, \dots, \alpha J_{0k}$, plus auxiliary variables $\boldsymbol{V}$, thus having dimension $2k + 1$. The exact posterior sampling procedure developed in Section 5.2 is instead summarized in Algorithm 2. This scheme outputs i.i.d. samples from the posterior, but requires a non-trivial initialization. The main computational bottleneck is represented by the rejection sampling step for the latent variable αT, which may suffer from low acceptance rates. Numerical illustrations proving the effectiveness of the proposed algorithms are discussed in Section S3.8.

As highlighted in Proposition 3.3, the normalized gamma-gamma hCRV model coincides with the hierarchical Dirichlet process (HDP) with a particular gamma prior on the concentration parameter. Therefore, the algorithms described above can be compared with standard samplers for posterior analysis of the HDP. Among the various alternatives and their countless variations, we focus on marginal algorithms, and consider (i) the Gibbs sampler based on the restaurant franchise metaphor (Teh et al., 2006), in which the allocation to dishes is observed (and thus fixed), and (ii) a collapsed Gibbs sampler for the number of tables serving each dish, derived from the pEPPF; cfr. Section S3.7 for details. The latter algorithm does not appear in the existing literature, and may be of independent interest. In fact, it is related to a sampler considered in Bacallado et al.

Algorithm 1: MCMC posterior sampling algorithm for jumps $\boldsymbol{J}$.

Current state: $\alpha T, (\alpha J_{01}, \dots, \alpha J_{0k}), (\boldsymbol{V})$;
$(\alpha T) \leftarrow$ Metropolis-Hastings step from its marginal (via auxiliary variables $\boldsymbol{V}$) ;
for $i = 1, \dots, d$ **do**
  $(b\beta_i) \mid (\alpha T) \sim \text{Gamma}(\alpha T, 1)$; $(U_i/b) \mid (b\beta_i) \sim \text{Gamma}(n_i, b\beta_i)$;
compute $\lambda(\boldsymbol{U}) = b_0/\alpha + \sum_{i=1}^{d} \log(1 + U_i/b)$;
for $j = 1, \dots, k$ **do**
  $(\alpha J_{0j}) \leftarrow$ Metropolis-Hastings step from the distribution of $(\alpha J_{0j}) \mid \lambda(\boldsymbol{U})$;
for $i = 1, \dots, d$ **do**
  for $j = 1, \dots, k$ **do**
    $J_{ij} \mid (\alpha J_{0j}), (U_i/b) \sim \text{Gamma}(n_{ij} + \alpha J_{0j}, b\,(1 + U_i/b))$;

(2022) for the hierarchical Pitman-Yor process (see also Camerlenghi et al., 2019), but relies on a further marginalization of the latent tables through the multivariate Stirling numbers in (12), reducing from kd to k latent variables. The marginal sampler (i) involves n latent variables, one for each observation's allocation to a table; hence, its computational complexity rapidly increases with the sample size. In this respect, the collapsed Gibbs sampler (ii) provides a substantial improvement, since its state space has dimension k, which is typically much smaller than n. However, updating each of the k variables (namely the number of tables serving each dish) requires the evaluation of the multivariate Stirling numbers in (12). These grow extremely fast and often cause numerical overflows even for moderately large sample sizes. For this reason, when n is (moderately) large, the marginal sampler (i) remains the practical choice.

In fact, our proposals present some interesting advantages. Firstly, in presence of a (moderately) large number of observations, the MCMC approach in Algorithm 1 relies on a state of dimension $2k + 1$, thereby reducing the computational burden of the marginal sampler (i), without requiring the evaluation of multivariate Stirling numbers. Moreover, the structure of conditional (in)dependencies enables parallelization and potentially speeds up mixing. Finally, running a Markov chain on the Euclidean space $(0, \infty)^{k+1} \times \Delta^k$ is more convenient than sampling on the constrained space of partitions: many standard tools for MCMC diagnostics are designed for Euclidean spaces, and acceptance rates of Metropolis-Hastings steps can be optimized by tuning the variance of the proposal; see Section S3.1. On the other hand, when n is small or moderately large, so that the computation of multivariate Stirling numbers in the collapsed sampler (ii) remains feasible, Algorithm 2 allows for exact i.i.d. sampling from the posterior. This approach avoids any potential issue inherent to MCMC schemes, without increasing the computational complexity.

Algorithm 2: Exact posterior sampling algorithm for jumps $\boldsymbol{J}$.

Initialize: compute coefficients in Propositions 5.3-5.4, the optimal r and the upper bound $t^*(r)$.
while *proposal is not accepted* **do**
  propose $\alpha T \sim \text{Gamma}(\alpha_0 + r, b_0/\alpha)$;
  accept proposal with probability proportional to $(\alpha T)^{-r} R(\alpha T)$ (Section 5.2);
for $i = 1, \dots, d$ **do**
  $(b\beta_i) \mid (\alpha T) \sim \text{Gamma}(\alpha T, 1)$; $(U_i/b) \mid (b\beta_i) \sim \text{Gamma}(n_i, b\beta_i)$;
compute $\lambda(\boldsymbol{U}) = b_0/\alpha + \sum_{i=1}^{d} \log(1 + U_i/b)$;
for $j = 1, \dots, k$ **do**
  $H_j \mid \lambda \sim \text{Categorical}(p_{jm_{\bullet j}}, \dots, p_{jn_{\bullet j}})$ from rescaled weights (Section 5.2);
  $(\alpha J_{0j}) \mid H_j, \lambda \sim \text{Gamma}(H_j, \lambda)$;
for $i = 1, \dots, d$ **do**
  for $j = 1, \dots, k$ **do**
    $J_{ij} \mid (\alpha J_{0j}), (U_i/b) \sim \text{Gamma}(n_{ij} + \alpha J_{0j}, b\,(1 + U_i/b))$;

5.4 Simulation studies

In the following, we carry out several simulation studies to compare the algorithms in terms of their execution time, as the dimensions of the input data grow. Specifically, we compare four algorithms: (i) the MCMC sampler in Algorithm 1 with symmetric random-walk Metropolis-Hastings steps on the log-scale (mcmc); (ii) the exact sampler in Algorithm 2; (iii) the marginal Gibbs sampler of Teh et al. (2006) based on the restaurant franchise metaphor (CRF); (iv) the collapsed Gibbs sampler for the number of tables serving each dish, detailed in Section S3.7. Algorithms are compared in terms of execution time per effective sample, averaging over 100 simulated datasets for each experimental setting. The burn-in time for algorithms (i), (iii) and (iv) and the initialization time for algorithm (ii) are deducted from the total execution times. The exact and collapsed Gibbs samplers are stopped when the computations of the multivariate Stirling numbers in (12) encounter a numerical overflow. The corresponding execution times are plotted when at least 75 out of 100 experiments are completed without errors.

The left panel of Figure 2 compares the algorithms for increasing number of groups d, while the number of observations per group $n_i = 50$ is fixed. Similarly, the right panel compares the algorithms for increasing number of observations per group n_i, while the number of groups $d = 20$ is fixed. In both cases, observations are sampled from a hierarchical Dirichlet process with concentration parameters $\alpha = 5$ and $\alpha_0 = 3$. Note that parameter α_0 impacts the number of columns in the counts matrix $(n_{ij})_{ij}$, while parameter α controls its sparsity. For the CRF-based Gibbs sampler, the time per effective sample grows more than linearly in the total number of observations, as expected given its sequential allocation structure. The other algorithms show a linear growth rate in the number of groups (Figure 2, left panel), with the exact and collapsed samplers displaying

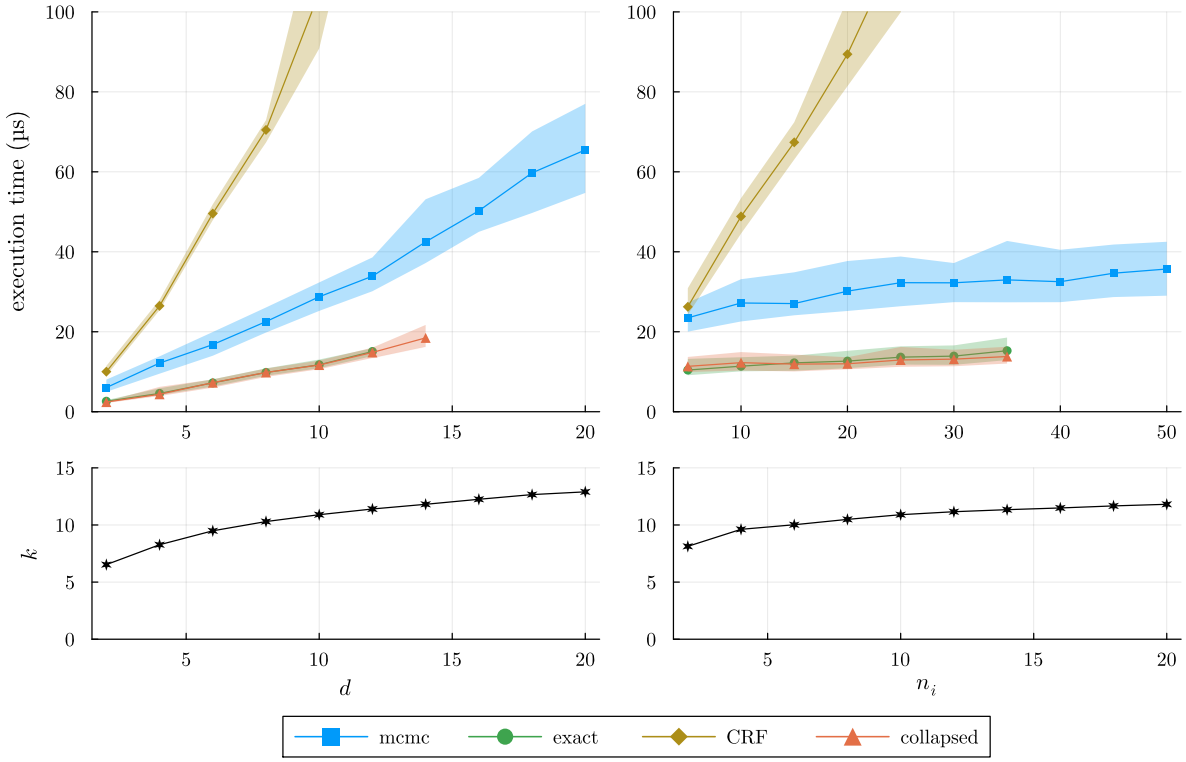

Figure 2: Execution times per effective sample for different algorithms with increasing number of groups d (left) and increasing number of observations per group n_i (right). Results are averaged over 100 simulated datasets per experimental setting. Solid curves represent median values, with shaded areas between the first and third quartiles. Plots at the bottom display the mean number of distinct values k for each setting.

a smaller slope compared with the MCMC approach. These algorithms appear less affected by the number of observations per group, showing a nearly constant behavior as n_i grows (Figure 2, right panel). We argue that their execution times may instead mostly depend on the number of distinct values k, which determines the dimension of the state space for both the MCMC and collapsed samplers. In fact, k is slowly increasing in the experimental settings considered above, as displayed in the bottom panels of Figure 2. Refer to Section S3.9 for a further simulation study supporting this claim.

In conclusion, the time complexity of both proposed sampling schemes scales linearly in the number of groups d and distinct values k, while it is essentially unaffected by the number of observations. In our implementation, the exact sampler shows better performances, yielding around twice as many independent posterior samples as the MCMC approach, in the same computation time. This is comparable with the collapsed Gibbs sampler for the HDP, which however does not output i.i.d. samples. Despite their efficiency, both algorithms suffer numerical issues as the sample size grows: in our experimental setting, their applicability is limited to 500-700 observations. On the contrary, the MCMC algorithm runs efficiently regardless of the sample size. In particular, it proves much faster than the CRF-based Gibbs sampler for moderate to large sample sizes.

5.5 Extensions to other hCRVs

The normalized gamma-gamma hCRV represents both the most relevant and the analytically simplest hCRV construction, since Proposition 3.3 establishes its tight connection with the HDP, arguably the cornerstone of Bayesian nonparametric dependent models, and several latent variables involved in its posterior characterization collapse to gamma distributions. Nevertheless, some analytical tractability is preserved beyond the gamma-gamma case, with potential advantages in terms of model flexibility, e.g. in the asymptotic behavior of the number of distinct values. We argue that the tractability of the gamma-gamma hCRV is primarily due to the idiosyncratic component ρ being the Lévy measure of a gamma CRM. Indeed, the results in Propositions 4.3 and 4.5 heavily depend on the distributional form of $\mathrm{ID}(t\rho)$, while the choice of ρ_0 only affects the distribution of the J_{0j}'s and of the vector $\boldsymbol{T}$. Therefore, one may consider different choices of the base Lévy measure ρ_0 to enhance model flexibility, at a reduced cost in terms of tractability; for example, ρ_0 can be taken to be the Lévy measure of a generalized gamma or stable CRM. The main limitation is the need for pointwise evaluations of the density of $\mathrm{ID}(\rho_0)$, whose analytical form is rarely available; however, this task can be performed through well-established numerical methods. Instead, the extension of the idiosyncratic measure ρ to other choices is less straightforward; in particular, within the present posterior representation, the availability of a closed-form expression for the density of $\mathrm{ID}(t\rho)$ for $t > 0$ is essential to sample the jumps J_{ij}'s in Proposition 4.3, as well as the vectors $\boldsymbol{S}_i$'s in Proposition 4.5. As a result, reasonable fully tractable choices for ρ are limited to few special cases, namely the inverse Gaussian process and σ-stable process with $\sigma = 0.5$. Nonetheless, the possibility to directly sample the β_i's in Proposition 4.5 is limited to the gamma case. The development of alternative techniques to handle $\mathrm{ID}(t\rho)$ beyond these specifications would considerably enlarge the class of tractable hCRVs, and represents an interesting direction for future work. In conclusion, we point out that there are no particular reasons to constrain ρ and ρ_0 to take the same analytical form, neither from the modelling nor the computational perspectives. Instead, their choice should be informed by the desired flexibility in the marginal and dependence structures; see Proposition 3.6 and the discussion in Catalano et al. (2024, Section 8).

6 Application: a fair comparison with the HDP

Section S4 investigates the impact of the hyperparameters on the borrowing of information and shrinkage induced by the normalized gamma-gamma hCRV, and discusses the proper elicitation of its prior dependence structure. In this section, we compare such model with the hierarchical Dirichlet process. We consider a simulated dataset with $d = 3$ groups of independent Poisson observations, each of size $n_i = 10$, with means equal to 2, 3

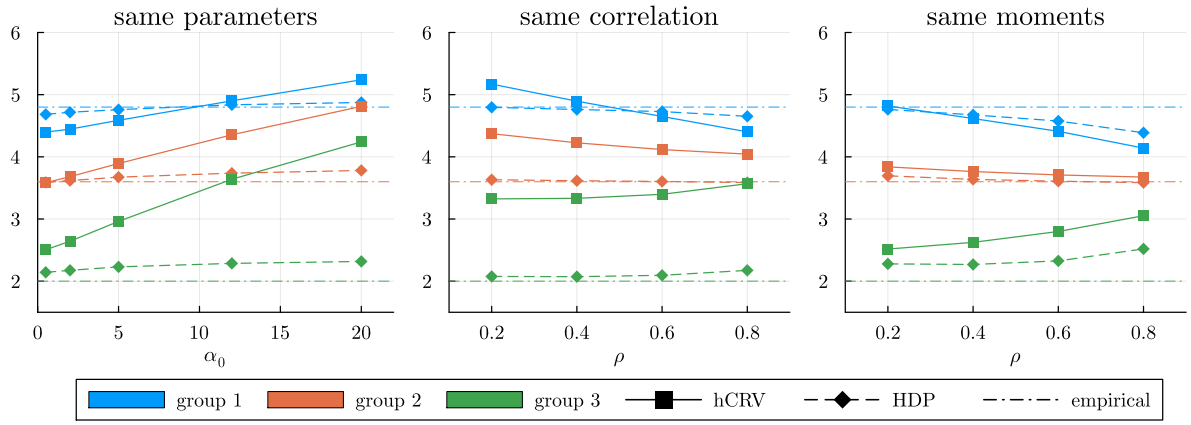

Figure 3: Comparison between predictive means for the normalized gamma-gamma hCRV and the HDP fixing the same hyperparameters (α, α_0), with $\alpha = 1$ (left); the same correlation and different variances, with σ^2 equal to 0.2 and 0.8 respectively (middle); the same variance and correlation, with $\sigma^2 = 0.5$ (right). Data are 3 groups of independent Poisson observations of size $n_i = 10$. The prior mean is $P_0 = N(8, 1)$.

and 4. Proposition 3.3 shows that the normalized gamma-gamma hCRV is equivalent to the HDP with a particular gamma prior on the concentration parameter α. Hence, it may seem natural to compare the normalized gamma-gamma hCRV with the HDP by fixing the same parameters (α, α_0, P_0), as in Figure 3 (left plot), where $P_0 = N(8, 1)$. However, this approach may alter the comparison, since fixing the same parameters can entail very different marginal variance and correlation structures. In particular, for the HDP(α, α_0, P_0), following Camerlenghi et al. (2019),

$$E(\tilde{P}_i(A)) = P_0(A), \qquad \mathrm{Var}(\tilde{P}_i(A)) = \frac{1+\alpha+\alpha_0}{(1+\alpha_0)(1+\alpha)} P_0(A)(1-P_0(A)),$$
$$\mathrm{Corr}(\tilde{P}_i(A), \tilde{P}_j(A)) = \frac{1+\alpha}{1+\alpha+\alpha_0}.$$

Table S2 reports some limiting behaviors, which notably differ from those of the normalized gamma-gamma hCRV, derived from Example 3.7. Therefore, rather than considering the same values of (α, α_0), a fair comparison should consider the same values of the mean, variance and correlation. Specifically, one should choose $\sigma^2, \rho \in (0, 1)$ such that $\mathrm{Var}(\tilde{P}_i(A)) = \sigma^2 P_0(A)(1 - P_0(A))$ and $\mathrm{Corr}(\tilde{P}_i(A), \tilde{P}_j(A)) = \rho$, and set the hyperparameters of the models accordingly. For the normalized gamma-gamma hCRV, the corresponding values of α and α_0 are obtained by solving the system of non-linear equations

$$\rho\left(1 + \alpha_0/\alpha\, e^{1/\alpha} E_{\alpha_0}(1/\alpha)\right) - 1 = 0, \qquad \sigma^2(1+\alpha_0) - 1/\rho = 0,$$

using standard numerical methods. The same task is achieved for the HDP by setting

$$\alpha = \frac{1}{1-\rho}\left(\frac{1}{\sigma^2} - 1\right), \qquad \alpha_0 = \frac{1}{\rho\sigma^2} - 1.$$

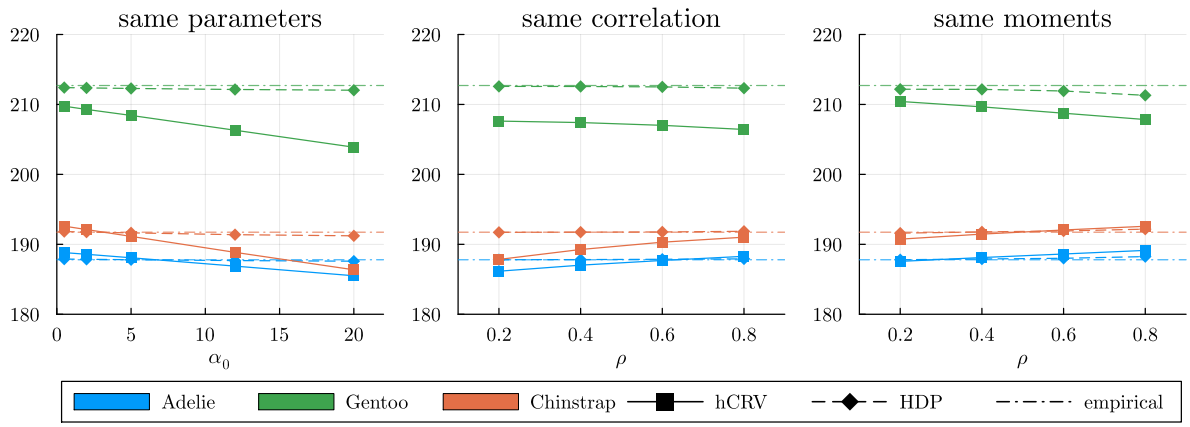

Figure 4: Comparison between predictive means for the normalized gamma-gamma hCRV and the HDP fixing the same hyperparameters (α, α_0), with $\alpha = 1$ (left); the same correlation and different variances, with σ^2 equal to 0.2 and 0.8 respectively (middle); the same variance and correlation, with $\sigma^2 = 0.5$ (right). Data are flipper lengths for 3 species of female penguins of the Palmer Archipelago (Horst et al., 2020).

Figure S9 displays the values of (α, α_0) for the two models as functions of (σ^2, ρ). The right plot of Figure 3 compares the normalized gamma-gamma hCRV with the HDP fixing the same mean, variance, and correlation. In contrast, the middle plot illustrates a common approach in which the correlation is fixed without accounting for variance. In conclusion, when fixing the same parameters (α, α_0), one might mistakenly conclude that the gamma-gamma hCRV borrows much more information than the HDP. A similar misinterpretation arises when fixing the correlation without adjusting for variance. Ultimately, by matching variance and correlation, it becomes clear that the gamma-gamma hCRV exhibits only slightly more borrowing of information. This increase can be attributed to the hierarchical CRV leveraging not only the common base random probability but also information contained in the random concentration parameter.

A similar analysis is reproduced on the Palmer Archipelago (Antarctica) penguin data (Horst et al., 2020), with the goal of predicting the flipper length for female penguins. The dataset includes three species of penguins (Adelie, Chinstrap, Gentoo) with many ties ($n = 165$, $k = 41$). The prior guess is $P_0 = N(100, 10)$, which provides a significant bias towards lower values of the flipper length. As shown in Figure 4, fixing the same value of the hyperparameters (α, α_0) could incorrectly suggest that the gamma-gamma hCRV gives too much weight to the prior with respect to the HDP (left plot). This effect is mitigated by fixing the same correlation but different variance (middle plot) and disappears by fixing the same variance and correlation (right plot). Under this fair model comparison, the gamma-gamma hCRV borrows slightly more information than the HDP. This behavior is the same observed for the simulated data, and can be ascribed to the information contained in the common concentration parameter.

7 Discussion

This work introduces normalized hierarchical completely random vectors (hCRVs), a new way of building dependent priors that combines the naturalness of hierarchical structures with a convenient posterior representation due to multivariate infinite divisibility. On the one hand, they provide an alternative sampling strategy for existing models, such as the hierarchical Dirichlet process (Teh et al., 2006); on the other hand, they represent a general recipe for constructing dependent priors with a convenient posterior representation. As such, normalized hCRVs enter the flourishing context of partially exchangeable models, which includes, among others, dependent stick-breaking constructions (MacEachern, 1999, 2000; Dunson and Park, 2008; Horiguchi et al., 2025), additive random measures (Müller et al., 2004; Lijoi et al., 2014), nested structures (Rodríguez et al., 2008; Camerlenghi et al., 2019; Beraha et al., 2021; Lijoi et al., 2023), Lévy copulas (Epifani and Lijoi, 2010; Riva-Palacio and Leisen, 2018), compound random measures (Griffin and Leisen, 2017), dependent Pólya trees (Christensen and Ma, 2020), FuRBI priors (Ascolani et al., 2024); see Quintana et al. (2022) for a recent review.

Future work will explore additional theoretical properties and methodological advantages of normalized hCRVs in different settings and for different families of CRMs. A first natural development is the derivation of the induced random partition through the pEPPF and of the predictive distribution for each group of observations, in the spirit of Camerlenghi et al. (2019). Moreover, in the context of multigroup data, the combination of normalized hCRVs with probability kernels in mixture models could lead to a novel class of priors for density estimation and clustering (Wade and Inácio, 2025). This raises interesting questions concerning the contraction rate to the true densities for partially exchangeable models (Catalano et al., 2022), the behavior of the mixing measure (Nguyen, 2016), and that of the clustering structure (Ascolani et al., 2022). Additionally, normalized hCRVs could be employed in stochastic block models for multi-layer network data (Durante et al., 2025), and, in the one-dimensional setting ($d = 1$), they could represent a viable alternative as priors beyond Gibbs type (Camerlenghi et al., 2018) or as building blocks for multiview clustering (Dombowsky and Dunson, 2025). Finally, from a numerical perspective, it would be of interest to evaluate the performance of hCRVs in more structured application domains, such as topic modelling, genomics, or sequential data, in order to assess the regimes in which the table-free representation yields the largest practical gains in terms of mixing, scalability, and computational cost.

Funding

Claudio Del Sole was partially supported by the European Union - Next Generation EU PRIN-PNRR (project P2022H5WZ9).

References

Andrieu, C. and J. Thoms (2008). A tutorial on adaptive MCMC. *Statistics and Computing 18*(4), 343–373.

Ascolani, F., B. Franzolini, A. Lijoi, and I. Prünster (2024). Nonparametric priors with full-range borrowing of information. *Biometrika 111*(3), 945–969.

Ascolani, F., A. Lijoi, G. Rebaudo, and G. Zanella (2022). Clustering consistency with Dirichlet process mixtures. *Biometrika 110*(2), 551–558.

Bacallado, S., S. Favaro, S. Power, and L. Trippa (2022). Perfect sampling of the posterior in the hierarchical Pitman–Yor process. *Bayesian Analysis 17*(3), 685–709.

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.

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

Bertoin, J. (1996). *Lévy Processes*. Cambridge Tracts in Mathematics. Cambridge: Cambridge University Press.

Bochner, S. (1955). *Harmonic Analysis and the Theory of Probability*. Berkeley: University of California Press.

Bryant, M. and E. Sudderth (2012). Truly nonparametric online variational inference for hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*, Volume 25, pp. 2699–2707.

Brück, F. (2026). Infinitely divisible priors for multivariate survival functions.

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 (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scandinavian Journal of Statistics 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.

Campbell, T., J. H. Huggins, J. P. How, and T. Broderick (2019). Truncated random measures. *Bernoulli 25*(2), 1256–1288.

Catalano, M., P. De Blasi, A. Lijoi, and I. Prünster (2022). Posterior asymptotics for boosted hierarchical Dirichlet process mixtures. *Journal of Machine Learning Research 23*(80), 1–23.

Catalano, M., C. Del Sole, A. Lijoi, and I. Prünster (2024). A unified approach to hierarchical random measures. *Sankhya A 86*(S1), 255–287.

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 (2021). Measuring dependence in the Wasserstein distance for Bayesian nonparametric models. *The Annals of Statistics 49*(5), 2916–2947.

Christensen, J. and L. Ma (2020). A Bayesian hierarchical model for related densities by using Pólya trees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(1), 127–153.

Constantine, G. M. and T. H. Savits (1996). A multivariate Faà di Bruno formula with applications. *Transactions of the American Mathematical Society 348*(2), 503–520.

Daley, D. and D. Vere-Jones (2002). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Probability and Its Applications. Springer.

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

Das, S., Y. Niu, Y. Ni, B. K. Mallick, and D. Pati (2025). Blocked Gibbs sampler for hierarchical Dirichlet processes. *Journal of Computational and Graphical Statistics 34*(2), 519–529.

Del Sole, C., A. Lijoi, and I. Prünster (2026). Principled estimation and prediction with competing risks: a Bayesian nonparametric approach. *Journal of the American Statistical Association*.

Dombowsky, A. and D. B. Dunson (2025). Product centred Dirichlet processes for Bayesian multiview clustering. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 87*(5), 1331–1352.

Dunson, D. B. and J.-H. Park (2008). Kernel stick-breaking processes. *Biometrika 95*(2), 307–323.

Durante, D., F. Gaffi, A. Lijoi, and I. Prünster (2025). Partially exchangeable stochastic block models for (node-colored) multilayer networks. *Journal of the American Statistical Association 120*(551), 1811–1827.

Elliott, L. T., M. De Iorio, S. Favaro, K. Adhikari, and Y. W. Teh (2019). Modeling population structure under hierarchical Dirichlet processes. *Bayesian Analysis 14*(2), 313–339.

Epifani, I. and A. Lijoi (2010). Nonparametric priors for vectors of survival functions. *Statistica Sinica 20*, 1455–1484.

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

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.

Fox, E. B., E. B. Sudderth, M. I. Jordan, and A. S. Willsky (2011). A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics 5*(2A), 1020–1056.

Franzolini, B., A. Lijoi, I. Prünster, and G. Rebaudo (2025). Multivariate species sampling models.

Gelman, A., G. O. Roberts, and W. R. Gilks (1996). Efficient Metropolis jumping rules. In *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting*. Oxford University Press.

Gilks, W. R. and P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 41*(2), 337–348.

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.

Griffiths, T., K. Canini, A. Sanborn, and D. Navarro (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*.

Haines, T. S. and T. Xiang (2011). Delta-dual hierarchical Dirichlet processes: A pragmatic abnormal behaviour detector. In *Proceedings of the 2011 International Conference on Computer Vision*, pp. 2198–2205.

Hardy, M. (2006). Combinatorics of partial derivatives. *The Electronic Journal of Combinatorics 13*(1), 1–13.

Horiguchi, A., C. Chan, and L. Ma (2025). A tree perspective on stick-breaking models in covariate-dependent mixtures (with discussion). *Bayesian Analysis 20*(3), 1139–1230.

Horst, A. M., A. P. Hill, and K. B. Gorman (2020). *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. R package version 0.1.0.

James, L. F., J. Lee, and A. Pandey (2024). Bayesian analysis of generalized hierarchical Indian buffet processes for within and across group sharing of latent features.

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.

Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics 21*, 59–78.

Lau, J. W. and E. Cripps (2022). Thinned completely random measures with applications in competing risks models. *Bernoulli 28*(1), 638–662.

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*, 1260–1291.

Lijoi, A., I. Prünster, and G. Rebaudo (2023). Flexible clustering via hidden hierarchical Dirichlet priors. *Scandinavian Journal of Statistics 50*(1), 213–234.

Lijoi, A., I. Prünster, and T. Rigon (2020). Sampling hierarchies of discrete random structures. *Statistics and Computing 30*, 1591–1607.

Liu, J., S. Wade, and N. Bochkina (2024). Shared differential clustering across single-cell RNA sequencing datasets with the hierarchical Dirichlet process. *Econometrics and Statistics*.

MacEachern, S. N. (1999). Dependent nonparametric processes. *In ASA Proceedings of the Section on Bayesian Statistical Science*.

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

Masoero, L., F. Camerlenghi, S. Favaro, and T. Broderick (2018). Posterior representations of hierarchical completely random measures in trait allocation models. In *NeurIPS Workshop on All of Bayesian Nonparametrics*.

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

Nakamura, T., T. Nagai, and N. Iwahashi (2011). Multimodal categorization by hierarchical Dirichlet process. In *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1520–1525.

Nguyen, X. (2016). Borrowing strength in hierarchical Bayes: Posterior concentration of the Dirichlet base measure. *Bernoulli 22*, 1535–1571.

Quintana, F. A., P. Müller, A. Jara, and S. N. MacEachern (2022). The dependent Dirichlet process and related models. *Statistical Science 37*, 24–41.

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

Ren, L., D. B. Dunson, and L. Carin (2008). The dynamic hierarchical Dirichlet process. In *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, pp. 824–831.

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.

Roberts, G. O. and J. S. Rosenthal (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science 16*(4), 351–367.

Rodríguez, A., D. B. Dunson, and A. E. Gelfand (2008). The nested Dirichlet process. *Journal of the American Statistical Association 103*(483), 1131–1154.

Rosiński, J. (2001). Series representations of Lévy processes from the perspective of point processes. In O. E. Barndorff-Nielsen, S. I. Resnick, and T. Mikosch (Eds.), *Lévy Processes: Theory and Applications*, pp. 401–415. Boston: Birkhäuser.

Sato, K.-I. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge: Cambridge University Press.

Schilling, R. L., R. Song, and Z. Vondracek (2012). *Bernstein Functions: Theory and Applications*. Berlin, Boston: De Gruyter.

Smith, A. (2014). A Gibbs sampler on the n-simplex. *The Annals of Applied Probability 24*(1), 114–130.

Sohn, K.-A. and E. P. Xing (2009). A hierarchical Dirichlet process mixture model for haplotype reconstruction from multi-population data. *The Annals of Applied Statistics 3*(2), 791–821.

Sudderth, E. B., A. Torralba, W. T. Freeman, and A. S. Willsky (2008). Describing visual scenes using transformed objects and parts. *International Journal of Computer Vision 77*, 291–330.

Taniguchi, T., R. Yoshino, and T. Takano (2018). Multimodal hierarchical Dirichlet process-based active perception by a robot. *Frontiers in Neurorobotics 12*.

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: 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.

Teh, Y. W., K. Kurihara, and M. Welling (2007). Collapsed variational inference for HDP. In *Advances in Neural Information Processing Systems*, Volume 20.

Thibaux, R. and M. I. Jordan (2007). Hierarchical beta processes and the Indian buffet process. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, pp. 564–571.

Vershik, A., M. Yor, and N. Tsilevich (2004). On the Markov–Krein identity and quasi-invariance of the gamma process. *Journal of Mathematical Sciences 121*, 2303–2310.

Wade, S. and V. Inácio (2025). Bayesian dependent mixture models: A predictive comparison and survey. *Statistical Science 40*(1), 81–108.

Walker, S. and P. Damien (2000). Representations of Lévy processes without Gaussian components. *Biometrika 87*(2), 477–483.

Wang, C., J. Paisley, and D. M. Blei (2011). Online variational inference for the hierarchical Dirichlet process. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pp. 752–760.

Williamson, S., A. Dubey, and E. Xing (2013). Parallel Markov chain Monte Carlo for nonparametric mixture models. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 98–106.

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. New York: Springer.

Zavitsanos, E., G. Paliouras, and G. A. Vouros (2011). Nonparametric estimation of topic hierarchies from texts with hierarchical Dirichlet processes. *Journal of Machine Learning Research 12*(83), 2749–2775.

Zhang, J. and A. Dassios (2025). Posterior sampling from truncated Ferguson-Klass representation of normalised completely random measure mixtures. *Bayesian Analysis 20*(3), 795–825.

Supplementary Material to: Hierarchical Random Measures without Tables

Abstract

Section S1 presents a brief and self-contained account on completely random measures, Lévy measures, Lévy intensities, Laplace exponents, and their multivariate extension to completely random vectors. Proofs of the results in the main manuscript are presented in Section S2. Section S3 provides additional technical details on the implementation of posterior sampling algorithms for the normalized gamma-gamma hCRV introduced in Section 5, and contains further numerical illustrations supporting their effectiveness. Finally, Section S4 investigates the impact of the hyperparameters on the borrowing of information and shrinkage induced by the normalized gamma-gamma hCRV, and discusses the proper elicitation of its prior dependence structure.

S1 Background on completely random measures

This appendix contains a concise account of completely random measures (Kingman, 1967) and their multivariate extension to vectors of random measures. In particular, we introduce the key notions of Lévy measure and Laplace exponent.

Let $(\mathbb{X}, d_{\mathbb{X}})$ be a Polish space endowed with a distance $d_{\mathbb{X}}$. The space $M_{\mathbb{X}}$ of boundedly finite measures on $\mathbb{X}$ is a Borel space with the topology of weak$^\sharp$ convergence (Daley and Vere-Jones, 2002). A random measure is a measurable function $\tilde{\mu} : \Omega \to M_{\mathbb{X}}$ from some probability space Ω.

Definition S1.1. A random measure $\tilde{\mu} : \Omega \to M_{\mathbb{X}}$ is a *completely random measure* (CRM) if, given a finite collection of pairwise disjoint and bounded Borel sets $A_1, \cdots, A_k$ of $\mathbb{X}$, the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_k)$ are mutually independent.

Kingman (1967) provides a remarkable representation theorem that decomposes any CRM $\tilde{\mu}$ into a unique sum $\tilde{\mu} = \mu_0 + \tilde{\mu}_f + \tilde{\mu}_c$ of three independent components: (i) a deterministic measure μ_0, (ii) a random measure with fixed atoms $\tilde{\mu}_f$, and (iii) an a.s. discrete random measure without fixed atoms $\tilde{\mu}_c$. The use of CRMs as priors in Bayesian nonparametric models usually restricts to the third component. For this reason, unless differently specified, we only focus on this class of CRMs and henceforth assume $\tilde{\mu} = \tilde{\mu}_c$. Moreover, Kingman (1967) shows that for any CRM $\tilde{\mu}$ there exists a Poisson random measure $\tilde{N}$ on $(0, +\infty) \times \mathbb{X}$ such that, for any Borel set A of $\mathbb{X}$,

$$\tilde{\mu}(A) = \iint_{(0,\infty)\times A} s \, \mathrm{d}\tilde{N}(s, x). \tag{S1}$$

The mean measure ν of the Poisson random measure $\tilde{N}$ characterizes the law of $\tilde{\mu}$ and is termed the *Lévy intensity* of the CRM, justifying the notation $\tilde{\mu} \sim \mathrm{CRM}(\nu)$. This measure on $(0, +\infty) \times \mathbb{X}$ can have infinite mass on sets of the form $(0, \varepsilon) \times A$. However, for every $\varepsilon > 0$ and for every bounded Borel set A, it must satisfy the following constraints:

(a) the jump component is bounded out of the origin, $\nu((\varepsilon, +\infty) \times A) < +\infty$;

(b) the jump component is integrable in the origin, $\iint_{(0,\varepsilon)\times A} s \, \mathrm{d}\nu(s, x) < +\infty$;

(c) the atom component has no point masses: for every $x \in \mathbb{X}$, $\nu((0, +\infty) \times \{x\}) = 0$.

Details can be found in Daley and Vere-Jones (2007, Theorem 10.1.III). The identity (S1) is used in Kingman (1967) to derive a Lévy-Khintchine representation of the Laplace transform

$$\log \mathbb{E}(e^{-\lambda \tilde{\mu}(A)}) = -\iint_{(0,+\infty)\times A} (1 - e^{-\lambda s}) \, \mathrm{d}\nu(s, x), \tag{S2}$$

for every $\lambda \geq 0$. Therefore $\tilde{\mu}(A)$ has a pure-jump infinitely divisible distribution with Lévy measure $\mathrm{d}\rho_A(s) = \int_A \mathrm{d}\nu(s, x)$, which we compactly write as $\tilde{\mu}(A) \sim \mathrm{ID}(\rho_A)$; refer to Sato (1999, Theorem 8.1) for further details. For $A = \mathbb{X}$, the expression in (S2) provides the *Laplace exponent* of the CRM, namely the function $\psi : [0, +\infty) \to [0, +\infty)$ defined as

$$\psi(\lambda) = -\log \mathbb{E}(e^{-\lambda \tilde{\mu}(\mathbb{X})}).$$

Remarkably, thanks to the independence of the evaluations on disjoint sets, the Laplace transform characterizes the law of the CRM. The usual specification of Lévy intensities is through disintegration

$$\mathrm{d}\nu(s, x) = \mathrm{d}\rho_x(s) \, \mathrm{d}P_0(x), \tag{S3}$$

where P_0 is a σ-finite atomless measure on $\mathbb{X}$ for condition (c), and ρ_x is a σ-finite measure on $(0, +\infty)$, P_0-a.e., such that, for every $\varepsilon > 0$,

$$\text{(a)} \quad \rho_x((\varepsilon, +\infty)) < +\infty, \qquad \text{(b)} \quad \int_{(0,\varepsilon)} s \, \mathrm{d}\rho_x(s) < +\infty. \tag{S4}$$

Note that P_0 does not have to be a probability measure nor a measure with finite mass, unless we extend condition (a) to unbounded Borel sets. In this case, we can assume P_0 to be a probability measure, and the disintegration (S3) is unique. Conditions (S4) imply that ρ_x is a Lévy measure on $(0, +\infty)$; when ρ_x is absolutely continuous with respect to the Lebesgue measure, we term its Radon-Nikodym derivative *Lévy density*. In our context, we are interested in two additional conditions. Firstly, we consider Lévy measures ρ_x that have infinite mass near the origin, that is, such that $\rho_x((0, +\infty)) = \infty$; we term such Lévy measures *infinitely active*. Secondly, we consider disintegration in product form, which leads to the definition of homogeneous CRM.

Definition S1.2. Let $\tilde{\mu} \sim \text{CRM}(\nu)$ such that ν satisfies (S3); we refer to $\tilde{\mu}$ as a *homogeneous* CRM if $\rho_x = \rho$, P_0-a.e, and write $\tilde{\mu} \sim \text{CRM}(\rho \otimes P_0)$.

For a homogeneous $\tilde{\mu} \sim \text{CRM}(\rho \otimes P_0)$ the Lévy-Khintchine representation in (S2) simplifies consistently. Indeed, under homogeneity and assuming P_0 to be a probability measure, the Laplace exponent is

$$\psi(\lambda) = \int_{(0,+\infty)} (1 - e^{-\lambda s}) \, \mathrm{d}\rho(s),$$

and $\log \mathbb{E}(e^{-\lambda \tilde{\mu}(A)}) = -P_0(A) \, \psi(\lambda)$, for $\lambda \geq 0$. We observe that ψ is a non-negative and infinitely differentiable function whose derivative is completely monotone. Moreover, it vanishes at 0 and its derivative vanishes at $+\infty$. Thanks to the Lévy-Khintchine representation of Bernstein functions (see, e.g. Schilling et al. (2012), Theorem 3.2) any such function is the Laplace exponent of a CRM.

The two fundamental examples are the gamma CRM and the stable CRM. We recall their definitions, which are used as building blocks throughout the manuscript.

Definition S1.3. A random measure $\tilde{\mu} \sim \text{CRM}(\rho \otimes P_0)$ is a gamma CRM of shape $\alpha > 0$ and rate $b > 0$ if ρ has Lévy density and Laplace exponent equal to, respectively,

$$\rho(s) = \alpha \frac{e^{-bs}}{s}, \qquad \psi(\lambda) = \alpha \log\left(1 + \frac{\lambda}{b}\right).$$

Definition S1.4. A random measure $\tilde{\mu} \sim \text{CRM}(\rho \otimes P_0)$ is a stable CRM of shape $\alpha > 0$ and discount parameter $\sigma \in (0, 1)$ if ρ has Lévy density and Laplace exponent equal to, respectively,

$$\rho(s) = \frac{\alpha\sigma}{\Gamma(1-\sigma)} \frac{1}{s^{1+\sigma}}, \qquad \psi(\lambda) = \alpha\lambda^{\sigma}.$$

The notion of completely random measure can be naturally extended to vectors $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ of random measures, as in Definition 2.1. Following Catalano et al. (2021), we term them *completely random vectors* (CRVs). The representation in (S1) can be extended to CRVs by considering a multivariate Lévy intensity ν on $\Omega_d \times \mathbb{X}$, where $\Omega_d = [0, +\infty)^d \setminus \{\mathbf{0}\}$. Similarly, the Laplace transform of the random vector $\tilde{\boldsymbol{\mu}}(A)$ is characterized by

$$\log \mathbb{E}(e^{-\boldsymbol{\lambda} \cdot \tilde{\boldsymbol{\mu}}(A)}) = -\iint_{\Omega_d \times A} (1 - e^{-\boldsymbol{\lambda} \cdot \boldsymbol{s}}) \, \mathrm{d}\nu(\boldsymbol{s}, x),$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d) \in [0, +\infty)^d$, $\boldsymbol{s} = (s_1, \ldots, s_d) \in \Omega_d$ and $\cdot$ denotes the scalar product. The multivariate Laplace exponent is defined as

$$\psi(\boldsymbol{\lambda}) = -\log \mathbb{E}(e^{-\boldsymbol{\lambda} \cdot \tilde{\boldsymbol{\mu}}(\mathbb{X})}) = \iint_{\Omega_d \times \mathbb{X}} (1 - e^{-\boldsymbol{\lambda} \cdot \boldsymbol{s}}) \, \mathrm{d}\nu(\boldsymbol{s}, x), \tag{S5}$$

Consistently with the univariate case, a CRV is homogeneous if $\nu = \rho \otimes P_0$, where P_0 is an atomless measure on $\mathbb{X}$ and ρ is d-dimensional Lévy measure on Ω_d such that

$$\text{(i)} \quad \rho_x((\varepsilon, +\infty)^d) < +\infty, \qquad \text{(ii)} \quad \int_{[0,\varepsilon)^d \setminus \{\mathbf{0}\}} \|\boldsymbol{s}\| \, \mathrm{d}\rho_x(\boldsymbol{s}) < +\infty.$$

Under homogeneity, if P_0 is a probability measure, the multivariate Laplace exponent is

$$\psi(\boldsymbol{\lambda}) = \int_{\Omega_d} (1 - e^{-\boldsymbol{\lambda}\cdot\boldsymbol{s}}) \, \mathrm{d}\rho(\boldsymbol{s}),$$

and $\log \mathbb{E}(e^{-\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{\mu}}(A)}) = -P_0(A)\,\psi(\boldsymbol{\lambda})$, for $\boldsymbol{\lambda} \in [0, +\infty)^d$. Finally, if $\tilde{\boldsymbol{\mu}} \sim \mathrm{CRV}(\nu)$, it easily follows from the definition that the marginal random measures $\tilde{\mu}_i$ are CRMs. Their Lévy intensities can be obtained by marginalization of ν as $\mathrm{d}\nu_i(s_i, x) = \int_{\boldsymbol{s}_{-i} \in [0,+\infty)^{d-1}} \mathrm{d}\nu(\boldsymbol{s}, x)$, where $\boldsymbol{s}_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_d)$.

S2 Proofs

S2.1 Proof of Theorem 2.3

Step 1. Show that $\tilde{\boldsymbol{\mu}}$ is a CRV. Let $\tilde{\boldsymbol{\mu}}(A) = (\tilde{\mu}_1(A), \ldots, \tilde{\mu}_d(A))$ be the random vector of evaluations of $\tilde{\boldsymbol{\mu}}$ on the Borel set A of $\mathbb{X}$. First, we observe that $\tilde{\boldsymbol{\mu}}$ takes values in the space of boundedly finite measures. Indeed, by Definition 2.2, for every bounded Borel set A and conditionally on $\tilde{\mu}_0$,

$$\tilde{\mu}_i(A) \mid \tilde{\mu}_0 \sim \mathrm{ID}(\tilde{\mu}_0(A)\rho),$$

where $\mathrm{ID}(\eta)$ denotes the infinitely divisible distribution with Lévy measure η. Since $\tilde{\mu}_0$ is a completely random measure, $\tilde{\mu}_0(A) < +\infty$ almost surely; hence, $\tilde{\mu}_0(A)\rho$ is a Lévy measure, and $\tilde{\mu}_i(A) < +\infty$ a.s. for every $i = 1, \ldots, d$. At this point, we prove that the random vectors $\tilde{\boldsymbol{\mu}}(A_1), \ldots, \tilde{\boldsymbol{\mu}}(A_k)$ are mutually independent, for every $A_1, \cdots, A_k$ mutually disjoint sets of $\mathbb{X}$. Specifically, we show that all linear combinations are mutually independent, that is, for coefficients $\lambda_{ij} > 0$, with $i = 1, \ldots, d$ and $j = 1, \ldots, k$,

$$\mathbb{E}\Big(e^{-\sum_{j=1}^{k}\sum_{i=1}^{d}\lambda_{ij}\tilde{\mu}_i(A_j)}\Big) = \prod_{j=1}^{k} \mathbb{E}\Big(e^{-\sum_{i=1}^{d}\lambda_{ij}\tilde{\mu}_i(A_j)}\Big).$$

This identity is proved exploiting the following properties:

(i) $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ are conditionally independent given $\tilde{\mu}_0$;

(ii) since $\tilde{\mu}_i \mid \tilde{\mu}_0$ is a CRM, its evaluations on disjoint sets are independent;

(iii) Since $\tilde{\mu}_i(A_j) \mid \tilde{\mu}_0 \sim \mathrm{ID}(\tilde{\mu}_0(A_j)\rho)$, then $\mathcal{L}(\tilde{\mu}_i(A_j) \mid \tilde{\mu}_0) = \mathcal{L}(\tilde{\mu}_i(A_j) \mid \tilde{\mu}_0(A_j))$, where $\mathcal{L}(X)$ is the probability law of the random object X;

(iv) since $\tilde{\mu}_0$ is a CRM, $\tilde{\mu}_0(A_1), \ldots, \tilde{\mu}_0(A_k)$ are independent;

(v) again, $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ are conditionally independent given $\tilde{\mu}_0$.

This entails

$$\begin{aligned}
\mathbb{E}\Big(e^{-\sum_{j=1}^k \sum_{i=1}^d \lambda_{ij}\tilde{\mu}_i(A_j)}\Big) &\overset{\text{(i)}}{=} \mathbb{E}\bigg(\prod_{i=1}^d \mathbb{E}\Big(e^{-\sum_{j=1}^k \lambda_{ij}\tilde{\mu}_i(A_j)} \mid \tilde{\mu}_0\Big)\bigg) \\
&\overset{\text{(ii)}}{=} \mathbb{E}\bigg(\prod_{i=1}^d \prod_{j=1}^k \mathbb{E}\Big(e^{-\lambda_{ij}\tilde{\mu}_i(A_j)} \mid \tilde{\mu}_0\Big)\bigg) \\
&\overset{\text{(iii)}}{=} \mathbb{E}\bigg(\prod_{i=1}^d \prod_{j=1}^k \mathbb{E}\Big(e^{-\lambda_{ij}\tilde{\mu}_i(A_j)} \mid \tilde{\mu}_0(A_j)\Big)\bigg) \\
&\overset{\text{(iv)}}{=} \prod_{j=1}^k \mathbb{E}\bigg(\prod_{i=1}^d \mathbb{E}\Big(e^{-\lambda_{ij}\tilde{\mu}_i(A_j)} \mid \tilde{\mu}_0(A_j)\Big)\bigg) \\
&\overset{\text{(v)}}{=} \prod_{j=1}^k \mathbb{E}\Big(e^{-\sum_{i=1}^d \lambda_{ij}\tilde{\mu}_i(A_j)}\Big).
\end{aligned}$$

Step 2. Derive the Laplace transform. From the specification in Definition 2.2, it follows that $\tilde{\mu}_i(A) \mid \tilde{\mu}_0 \overset{\text{iid}}{\sim} \mathrm{ID}(\tilde{\mu}_0(A)\rho)$. Therefore,

$$\begin{aligned}
\log \mathbb{E}\Big(e^{-\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{\mu}}(A)}\Big) &= \log \mathbb{E}\Big(\mathbb{E}\Big(e^{-\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{\mu}}(A)} \mid \tilde{\mu}_0\Big)\Big) = \log \mathbb{E}\bigg(\prod_{i=1}^d e^{-\psi(\lambda_i)\tilde{\mu}_0(A)}\bigg) \\
&= \log \mathbb{E}\Big(e^{-\tilde{\mu}_0(A)\sum_{i=1}^d \psi(\lambda_i)}\Big) = -P_0(A)\,\psi_0\bigg(\sum_{i=1}^d \psi(\lambda_i)\bigg).
\end{aligned}$$

Step 3. Determine homogeneity and Laplace exponent. The multivariate Lévy intensity ν_h and the Laplace exponent ψ_h of $\tilde{\boldsymbol{\mu}}$ are uniquely characterized by the Laplace transform. Given the product form of Step 2, it follows that $\nu_h = \rho_h \otimes P_0$ for some multivariate ρ_h and the Laplace exponent satisfies

$$\psi_h(\boldsymbol{\lambda}) = \psi_0\bigg(\sum_{i=1}^d \psi(\lambda_i)\bigg).$$

Step 4. Determine the Lévy measure. The Lévy measure ρ_h is characterized by

$$\psi_h(\boldsymbol{\lambda}) = \int_{\Omega_d} (1 - e^{-\boldsymbol{\lambda}\cdot\boldsymbol{s}})\,\mathrm{d}\rho_h(\boldsymbol{s}).$$

We use the notation $\mathscr{L}_\rho(\lambda) = \mathbb{E}(e^{-\lambda X}) = e^{-\psi(\lambda)}$, where $\lambda \in [0, +\infty)$ and $X \sim \mathrm{ID}(\rho)$, to

denote the Laplace transform of an infinitely divisible random variable with Lévy measure ρ and Laplace exponent ψ; note that $\mathscr{L}_\rho(\lambda)^t = \mathscr{L}_{t\rho}(\lambda)$. Defining $X_i \overset{\text{iid}}{\sim} \mathrm{ID}(t\rho)$, one has

$$\begin{aligned}
\psi_h(\boldsymbol{\lambda}) &= \psi_0\bigg(\sum_{i=1}^d \psi(\lambda_i)\bigg) \\
&= \int_0^{+\infty} \Big(1 - e^{-\sum_{i=1}^d \psi(\lambda_i)\, t}\Big)\, \mathrm{d}\rho_0(t) = \int_0^{+\infty} \bigg(1 - \prod_{i=1}^d e^{-\psi(\lambda_i)\, t}\bigg)\, \mathrm{d}\rho_0(t) \\
&= \int_0^{+\infty} \bigg(1 - \prod_{i=1}^d \mathscr{L}_\rho(\lambda_i)^t\bigg)\, \mathrm{d}\rho_0(t) = \int_0^{+\infty} \bigg(1 - \prod_{i=1}^d \mathscr{L}_{t\rho}(\lambda_i)\bigg)\, \mathrm{d}\rho_0(t) \\
&= \int_0^{+\infty} \bigg(1 - \prod_{i=1}^d \mathbb{E}_{X_i}\Big(e^{-\lambda_i X_i}\Big)\bigg)\, \mathrm{d}\rho_0(t) = \int_0^{+\infty} \mathbb{E}_{\boldsymbol{X}}\Big(1 - e^{-\sum_{i=1}^d \lambda_i X_i}\Big)\, \mathrm{d}\rho_0(t) \\
&= \int_0^{+\infty} \int_{[0,+\infty)^d} \Big(1 - e^{-\sum_{i=1}^d \lambda_i s_i}\Big) \prod_{i=1}^d \mathrm{d}P_{\mathrm{ID}(t\rho)}(s_i)\, \mathrm{d}\rho_0(t) \\
&= \int_0^{+\infty} \int_{\Omega_d} \Big(1 - e^{-\sum_{i=1}^d \lambda_i s_i}\Big) \prod_{i=1}^d \mathrm{d}P_{\mathrm{ID}(t\rho)}(s_i)\, \mathrm{d}\rho_0(t) \\
&= \int_{\Omega_d} (1 - e^{-\boldsymbol{\lambda}\cdot\boldsymbol{s}}) \int_0^{+\infty} \prod_{i=1}^d \mathrm{d}P_{\mathrm{ID}(t\rho)}(s_i)\, \mathrm{d}\rho_0(t),
\end{aligned}$$

by exchanging the order of the integrals thanks to Fubini-Tonelli theorem, and observing that the integrand vanishes in $\boldsymbol{s} = \mathbf{0}$. If $\mathrm{ID}(t\rho)$ has density $f_{\mathrm{ID}(t\rho)}$ with respect to the Lebesgue measure, then this is equal to

$$\int_{\Omega_d} (1 - e^{-\boldsymbol{\lambda}\cdot\boldsymbol{s}}) \int_0^{+\infty} \prod_{i=1}^d f_{\mathrm{ID}(t\rho)}(s_i)\, \mathrm{d}\rho_0(t)\, \mathrm{d}\boldsymbol{s}.$$

S2.2 Proof of Remark 2.4

Assume that the Lévy measure ρ_{co} of a compound random measure coincides with ρ_h in Theorem 2.3 for some base Lévy measure ρ_0. Since ρ_{co}, ρ_h and ρ_0 are σ-finite measures, by the uniqueness of the disintegration of measures ρ_0-a.e. it holds that

$$\frac{1}{t^d}\, H\bigg(\frac{s_1}{t}, \ldots, \frac{s_d}{t}\bigg)\mathrm{d}\boldsymbol{s} = \prod_{i=1}^d \mathrm{d}P_{\mathrm{ID}(t\rho)}(s_i).$$

This forces H to be in product form and $\mathrm{ID}(t\rho)$ to have a density. Therefore, there exists a density H_1 on $[0, +\infty)$ such that

$$\frac{1}{t}\, H_1\bigg(\frac{s}{t}\bigg) = f_{\mathrm{ID}(t\rho)}(s), \qquad \rho_0\text{-a.e.}$$

Let X be a random variable with density H_1; then tX has density $f(s) = H_1(s/t)/t$. From the identity above, tX is an infinitely divisible distribution with Lévy density $t\rho(s)$. Therefore, X has (strictly) stable distribution (Sato, 1999, Definition 13.1 and Theorem 14.3); in particular, it has stability index 1, corresponding to a Lévy density $\rho(s) \propto s^{-2}$. However, this choice of $\rho(s)$ does not satisfy the integrability condition (b) in (S4), and thus is not a valid Lévy density for a CRM.

S2.3 Proof of Theorem 2.6

Step 1. Express the condition in terms of Laplace exponents. Since hierarchical CRVs are CRVs, their law is uniquely determined by their multivariate Laplace exponent. The 1-dimensional Laplace exponents $\psi_0^{(1)}$ and $\psi^{(2)}$, associated to $\rho_0^{(1)}$ and $\rho^{(2)}$ respectively, are strictly increasing by definition, which implies that they are invertible. Therefore, from the expression of Theorem 2.3, $\tilde{\boldsymbol{\mu}}_1$ and $\tilde{\boldsymbol{\mu}}_2$ have the same multivariate Laplace exponent if and only if

$$\sum_{i=1}^{d} \psi^{(1)} \circ (\psi^{(2)})^{-1}(s_i) = (\psi_0^{(1)})^{-1} \circ \psi_0^{(2)}\left(\sum_{i=1}^{d} s_i\right), \qquad P_0^{(1)} = P_0^{(2)},$$

for every $(s_1, \ldots, s_d) \in \Omega_d$. Define $f = \psi^{(1)} \circ (\psi^{(2)})^{-1}$ and $g = (\psi_0^{(1)})^{-1} \circ \psi_0^{(2)}$. Since the inverse of a continuous function is continuous, $f, g : (0, +\infty) \to (0, +\infty)$ are continuous functions such that $f(0) = g(0) = 0$ and

$$f(s_1) + \cdots + f(s_d) = g(s_1 + \cdots + s_d). \tag{S6}$$

Step 2. Show that $f = g$ and are linear functions. Let $s_1 = s$ and $s_2 = \cdots = s_d = 0$. Then, since $f(0) = 0$, we have $g = f$. To prove linearity, we show that $f(s_1 + s_2) = f(s_1) + f(s_2)$ for every $s_1, s_2 > 0$ and $f(cs) = cf(s)$ for every $c, s > 0$. The first property follows by taking $s_3 = \cdots = s_d = 0$ in (S6). As for the second, we prove it first for c a natural number and then for c a rational number. We conclude by density thanks to the continuity of f. For $n \in \mathbb{N}$ natural number, by (S6), $f(ns) = f(s + \cdots + s) = nf(s)$, while for $q = n/m$ rational number, with $n, m \in \mathbb{N}$, $nf(s) = f(ns) = f(qms) = mf(qs)$, which implies $f(qs) = nf(s)/m = qf(s)$.

Step 3. Conclusion. We have proved that there exists $c > 0$ such that, for every $s > 0$,

$$\psi^{(1)} \circ (\psi^{(2)})^{-1}(s) = (\psi_0^{(1)})^{-1} \circ \psi_0^{(2)}(s) = cs.$$

Thus for every $s > 0$,

$$\psi_0^{(2)}(s) = \psi_0^{(1)}(cs), \qquad \psi^{(2)}(s) = \frac{1}{c}\,\psi^{(1)}(s), \qquad P_0^{(1)} = P_0^{(2)},$$

Using a change of variable, this condition can be equivalently expressed in terms of the Lévy measures as

$$\rho_0^{(2)} = c_{\#}\rho_0^{(1)}; \qquad \rho^{(2)}(s) = \frac{1}{c}\,\rho^{(1)}(s); \qquad P_0^{(1)} = P_0^{(2)}.$$

If the Lévy measures have densities, this is equivalent to

$$\rho_0^{(2)}(s) = \frac{1}{c}\,\rho_0^{(1)}\left(\frac{s}{c}\right); \qquad \rho^{(2)}(s) = \frac{1}{c}\,\rho^{(1)}(s); \qquad P_0^{(1)} = P_0^{(2)},$$

which coincides with the statement of the proposition by substituting c with $1/c$.

S2.4 Proof of Lemma 3.1

Step 1. Find the Lévy measure ρ_1 of $\tilde{\mu}_i$. By Theorem 2.3, the multivariate Lévy measure of $\tilde{\boldsymbol{\mu}}$ is the measure on Ω_d satisfying

$$\mathrm{d}\rho_h(s_1, \ldots, s_d) = \int_0^{+\infty} \prod_{i=1}^{d} \mathrm{d}P_{\mathrm{ID}(t\rho)}(s_i)\,\mathrm{d}\rho_0(t).$$

The marginal Lévy measure of each $\tilde{\mu}_i$ coincides with the one of $\tilde{\mu}_1$, that is,

$$\mathrm{d}\rho_1(s) = \int_{[0,+\infty)^{d-1}} \int_0^{+\infty} \mathrm{d}P_{\mathrm{ID}(t\rho)}(s) \prod_{i=2}^{d} \mathrm{d}P_{\mathrm{ID}(t\rho)}(s_i)\,\mathrm{d}\rho_0(t) = \int_0^{+\infty} \mathrm{d}P_{\mathrm{ID}(t\rho)}(s)\,\mathrm{d}\rho_0(t),$$

where we have exchanged the order of integration thanks to Fubini-Tonelli's theorem, since measures are positive and σ-finite.

Step 2. Show that if ρ and ρ_0 are infinitely active, then ρ_1 is infinitely active. If ρ is infinitely active, then $\mathrm{ID}(t\rho)$ gives zero probability to the origin, thus

$$\int_{\Omega_1} \mathrm{d}\rho_1(s) = \int_{\Omega_1} \int_0^{+\infty} \mathrm{d}P_{\mathrm{ID}(t\rho)}(s)\,\mathrm{d}\rho_0(t) = \int_0^{+\infty} \mathrm{d}\rho_0(t),$$

where again we have used Fubini-Tonelli's theorem to exchange the order of the integrals. Since ρ_0 is infinitely active, the last integral is $+\infty$.

Step 3. If ρ_0 is finitely active, then ρ_1 is finitely active. Since $\mathrm{ID}(t\rho)$ is a probability measure on $[0, +\infty)$, its mass on $(0, +\infty)$ is smaller than or equal to 1. Thus, by Fubini-Tonelli,

$$\int_{\Omega_1} \mathrm{d}\rho_1(s) = \int_0^{+\infty} \left(\int_{\Omega_1} \mathrm{d}P_{\mathrm{ID}(t\rho)}(s) \right) \mathrm{d}\rho_0(t) \leq \int_0^{+\infty} \mathrm{d}\rho_0(t) < +\infty,$$

where the last inequality is due to the finite activity of ρ_0.

Step 4: If ρ is finitely active, then ρ_1 is finitely active. If ρ is finitely active with

total mass a, then $Y \sim \mathrm{ID}(t\rho)$ has positive mass in 0, namely $\mathbb{P}(Y = 0) = e^{-ta}$. This follows by observing that Y is a compound Poisson distribution, and $\mathbb{P}(Y = 0)$ coincides with the probability mass in 0 of a Poisson random variable with rate $\lambda = at$. Thus, by Fubini-Tonelli's theorem,

$$\int_{\Omega_1} \mathrm{d}\rho_1(s) = \int_0^{+\infty} \int_{\Omega_1} \mathrm{d}P_{\mathrm{ID}(t\rho)}(s)\ \mathrm{d}\rho_0(t) = \int_0^{+\infty} (1 - e^{-at})\,\mathrm{d}\rho_0(t) = \psi_0(at) < +\infty.$$

S2.5 Proof of Proposition 3.3

The Dirichlet process arises as a normalization of a gamma CRM (Ferguson, 1973), thus

$$\frac{\tilde{\mu}_1}{\tilde{\mu}_1(\mathbb{X})}, \ldots, \frac{\tilde{\mu}_d}{\tilde{\mu}_d(\mathbb{X})} \;\Big|\; \tilde{\mu}_0 \overset{\text{iid}}{\sim} \mathrm{DP}(\alpha\tilde{\mu}_0) = \mathrm{DP}\left(\alpha\tilde{\mu}_0(\mathbb{X})\,\frac{\tilde{\mu}_0}{\tilde{\mu}_0(\mathbb{X})}\right).$$

For a gamma CRM, the total mass $\tilde{\mu}_0(\mathbb{X})$ is independent from the normalization (Vershik et al., 2004, Lemma 1). Moreover, $\tilde{\mu}_0(\mathbb{X}) \sim \mathrm{Gamma}(\alpha_0, b_0)$ and thus $\alpha\tilde{\mu}_0(\mathbb{X}) \sim \mathrm{Gamma}(\alpha_0, b_0/\alpha)$ is independent of $\tilde{P}_0 = \tilde{\mu}_0/\tilde{\mu}_0(\mathbb{X})$.

S2.6 Proof of Proposition 3.4

We prove that, conditionally on $\tilde{\mu}_0$, it holds $\tilde{\mu}^{(1)} \overset{\mathrm{d}}{=} c(\tilde{\mu}_0)\,\tilde{\mu}^{(2)}$ for some positive value $c(\tilde{\mu}_0)$ depending on $\tilde{\mu}_0$. This implies that the normalizations of $\tilde{\mu}^{(1)}$ and $\tilde{\mu}^{(2)}$ are equal in distribution. Conditionally on $\tilde{\mu}_0$, both $\tilde{\mu}^{(1)}$ and $\tilde{\mu}^{(2)}$ are CRMs, and we can express the condition $\tilde{\mu}^{(1)} \overset{\mathrm{d}}{=} c(\tilde{\mu}_0)\,\tilde{\mu}^{(2)}$ in terms of their conditional Lévy measures $\rho^{(1)}$ and $\rho^{(2)}$ as

$$\rho^{(2)}(s) \otimes \frac{\tilde{\mu}_0}{\tilde{\mu}_0(\mathbb{X})} \overset{\mathrm{d}}{=} c(\tilde{\mu}_0)\,\rho^{(1)}(c(\tilde{\mu}_0)s) \otimes \tilde{\mu}_0.$$

Plugging in the expressions for the stable Lévy measures, we need to find $c(\tilde{\mu}_0)$ such that

$$\frac{\alpha\,\sigma}{\Gamma(1-\sigma)}\frac{1}{s^{\sigma+1}} \otimes \frac{\tilde{\mu}_0}{\tilde{\mu}_0(\mathbb{X})} \overset{\mathrm{d}}{=} \frac{\alpha\,\sigma}{\Gamma(1-\sigma)}\frac{1}{c(\tilde{\mu}_0)^{\sigma}}\frac{1}{s^{\sigma+1}} \otimes \tilde{\mu}_0.$$

The proof is concluded by choosing $c(\tilde{\mu}_0) = \tilde{\mu}_0(\mathbb{X})^{1/\sigma}$.

S2.7 Proof of Proposition 3.6

Firstly, we state and prove a related result concerning the unnormalized random measures. For this purpose, denote by $M_r(\rho)$ the r-th moment of a Lévy measure ρ,

$$M_r(\rho) = \int_{[0,\infty)} s^r\,\mathrm{d}\rho(s).$$

Proposition S2.1. *Let $\tilde{\boldsymbol{\mu}} \sim \mathrm{hCRV}(\rho, \rho_0, P_0)$ and let A be a Borel set s.t. $P_0(A) \neq 0, 1$. Then, for every $i \neq j$,*

$$\begin{aligned}\mathbb{E}(\tilde{\mu}_i(A)) &= P_0(A)M_1(\rho_0)M_1(\rho),\\ \mathrm{Var}(\tilde{\mu}_i(A)) &= P_0(A)M_1(\rho_0)M_2(\rho) + P_0(A)M_2(\rho_0)M_1(\rho)^2,\\ \mathrm{Cov}(\tilde{\mu}_i(A), \tilde{\mu}_j(A)) &= P_0(A)M_2(\rho_0)M_1(\rho)^2.\end{aligned}$$

In particular, for every $i \neq j$,

$$\mathrm{corr}(\tilde{\mu}_i(A), \tilde{\mu}_j(A)) = \frac{M_2(\rho_0)M_1(\rho)^2}{M_1(\rho_0)M_2(\rho) + M_2(\rho_0)M_1(\rho)^2}.$$

Proof. The expressions can be derived (i) through the hierarchical structure using the tower property and the law of total (co)variance, or (ii) using the expression of the moments of (jointly) infinitely divisible distributions in terms of their Lévy measures. We provide a proof exploiting both techniques. Recall that, by Campbell's theorem, the mean and variance of an infinitely divisible random variable $X \sim \mathrm{ID}(\rho)$ satisfy

$$\mathbb{E}(X) = \int_0^{+\infty} s \,\mathrm{d}\rho(s) = M_1(\rho), \qquad \mathrm{Var}(X) = \int_0^{+\infty} s^2 \,\mathrm{d}\rho(s) = M_2(\rho).$$

Proof through hierarchical structure. Since $\tilde{\mu}_i(A) \mid \tilde{\mu}_0 \sim \mathrm{ID}(\tilde{\mu}_0(A)\rho)$, by the tower property,

$$\mathbb{E}(\tilde{\mu}_i(A)) = \mathbb{E}(\mathbb{E}(\tilde{\mu}_i(A) \mid \tilde{\mu}_0)) = \mathbb{E}\left(\tilde{\mu}_0(A)\int_0^{+\infty} s \,\mathrm{d}\rho(s)\right) = P_0(A)M_1(\rho_0)M_1(\rho).$$

Similarly, by the law of total variance,

$$\begin{aligned}\mathrm{Var}(\tilde{\mu}_i(A)) &= \mathbb{E}(\mathrm{Var}(\tilde{\mu}_i(A) \mid \tilde{\mu}_0)) + \mathrm{Var}(\mathbb{E}(\tilde{\mu}(A) \mid \tilde{\mu}_0))\\ &= \mathbb{E}\left(\tilde{\mu}_0(A)\int_0^{+\infty} s^2 \,\mathrm{d}\rho(s)\right) + \mathrm{Var}\left(\tilde{\mu}_0(A)\int_0^{+\infty} s \,\mathrm{d}\rho(s)\right)\\ &= P_0(A)M_1(\rho_0)M_2(\rho) + P_0(A)M_2(\rho_0)M_1(\rho)^2.\end{aligned}$$

Finally, by the law of total covariance and thanks to the conditional independence of the random measures $\tilde{\mu}_i$ and $\tilde{\mu}_j$ for $i \neq j$, given $\tilde{\mu}_0$,

$$\begin{aligned}\mathrm{Cov}(\tilde{\mu}_i(A), \tilde{\mu}_j(A)) &= \mathrm{Cov}(\mathbb{E}(\tilde{\mu}_i(A) \mid \tilde{\mu}_0), \mathbb{E}(\tilde{\mu}_j(A) \mid \tilde{\mu}_0))\\ &= \mathrm{Var}\left(\tilde{\mu}_0(A)\int_0^{+\infty} s \,\mathrm{d}\rho(s)\right) = P_0(A)M_2(\rho_0)M_1(\rho)^2.\end{aligned}$$

Proof through joint infinite divisibility. Since $\tilde{\mu}_i$ is a CRM, by considering its Lévy measure in Theorem 2.3,

$$
\begin{aligned}
\mathbb{E}(\tilde{\mu}_i(A)) &= P_0(A) \int_0^{+\infty} s \int_0^{+\infty} \mathrm{d}P_{\mathrm{ID}(t\rho)}(s)\, \mathrm{d}\rho_0(t) \\
&= P_0(A) \int_0^{+\infty} \int_0^{+\infty} s\, \mathrm{d}P_{\mathrm{ID}(t\rho)}(s)\, \mathrm{d}\rho_0(t) \\
&= P_0(A) \int_0^{+\infty} \mathbb{E}_{X\sim\mathrm{ID}(t\rho)}(X)\, \mathrm{d}\rho_0(t) \\
&= P_0(A) M_1(\rho) \int_0^{+\infty} t\, \mathrm{d}\rho_0(t) = P_0(A) M_1(\rho) M_1(\rho_0).
\end{aligned}
$$

Similarly, for the variance,

$$
\begin{aligned}
\mathrm{Var}(\tilde{\mu}_i(A)) &= P_0(A) \int_0^{+\infty} s^2 \int_0^{+\infty} \mathrm{d}P_{\mathrm{ID}(t\rho)}(s)\, \mathrm{d}\rho_0(t) \\
&= P_0(A) \int_0^{+\infty} \int_0^{+\infty} s^2\, \mathrm{d}P_{\mathrm{ID}(t\rho)}(s)\, \mathrm{d}\rho_0(t) \\
&= P_0(A) \int_0^{+\infty} \mathbb{E}_{X\sim\mathrm{ID}(t\rho)}(X^2)\, \mathrm{d}\rho_0(t) \\
&= P_0(A) \int_0^{+\infty} (\mathrm{Var}_{X\sim\mathrm{ID}(t\rho)}(X) + \mathbb{E}_{X\sim\mathrm{ID}(t\rho)}(X)^2)\, \mathrm{d}\rho_0(t) \\
&= P_0(A) \int_0^{+\infty} (tM_2(\rho) + t^2 M_1(\rho)^2)\, \mathrm{d}\rho_0(t) \\
&= P_0(A) M_2(\rho) M_1(\rho_0) + P_0(A) M_1(\rho)^2 M_2(\rho_0).
\end{aligned}
$$

For the covariance, we observe that, for any $(X_1, X_2) \sim \mathrm{ID}(\rho)$ jointly infinitely divisible random variables with a multivariate Lévy measure ρ, it holds $\mathrm{Cov}(X_1, X_2) = \int_{\Omega_2} s_1 s_2\, \mathrm{d}\rho(s_1, s_2)$; see, e.g., Sato (1999, 25.8). Thus,

$$
\begin{aligned}
\mathrm{Cov}(\tilde{\mu}_i(A), \tilde{\mu}_j(A)) &= P_0(A) \int_{\Omega_2} s_1 s_2\, \mathrm{d}\rho_h(s_1, s_2) \\
&= P_0(A) \int_0^{+\infty} \int_{\Omega_2} s_1 s_2\, \mathrm{d}P_{\mathrm{ID}(t\rho)}(s_1)\, \mathrm{d}P_{\mathrm{ID}(t\rho)}(s_2)\, \mathrm{d}\rho_0(t) \\
&= P_0(A) \int_0^{+\infty} \mathbb{E}_{X\sim\mathrm{ID}(t\rho)}(X)^2\, \mathrm{d}\rho_0(t) \\
&= M_1(\rho)^2 \int_0^{+\infty} t^2 \mathrm{d}\rho_0(t) = P_0(A) M_1(\rho)^2 M_2(\rho_0). \qquad \square
\end{aligned}
$$

We now prove Proposition 3.6. Similarly to the proof of Proposition S2.1, we could derive the desired results in two ways: (i) exploiting the hierarchical structure, or (ii) using the fact that we are normalizing a CRV. When focusing on the normalization, exploiting the CRV structure is particularly convenient for deriving the mean and variance, whereas using the hierarchical structure brings to a straightforward calculation for the covariance. The univariate results in James et al. (2006) show that if $\tilde{\mu} \sim \mathrm{CRM}(\rho \otimes P_0)$, where P_0 is a probability measure and ψ denotes its Laplace exponent, the normalization $\tilde{P} = \tilde{\mu}/\tilde{\mu}(\mathbb{X})$

satisfies

$$\mathbb{E}(\tilde{P}(A)) = P_0(A), \qquad \mathrm{Var}(\tilde{P}(A)) = -P_0(A)(1-P_0(A)) \int_0^{+\infty} u e^{-\psi(u)} \psi''(u)\, \mathrm{d}u.$$

Since $\tilde{P}_i(A)$ is the normalization of $\tilde{\mu}_i \sim \mathrm{CRM}(\rho_h \otimes P_0)$, this immediately implies that $\mathbb{E}(\tilde{P}_i(A)) = P_0(A)$. Moreover, by considering the Laplace exponent of $\tilde{\mu}_i$ in Theorem 2.3,

$$\mathrm{Var}(\tilde{P}_i(A)) = -P_0(A)(1-P_0(A)) \int_0^{+\infty} u e^{-\psi_0(\psi(u))} (\psi_0 \circ \psi)''(u)\, \mathrm{d}u.$$

By the law of total covariance and the conditional independence of the random probability measures $\tilde{P}_i$ and $\tilde{P}_j$ for $i \neq j$, given $\tilde{\mu}_0$,

$$\begin{aligned}\mathrm{Cov}(\tilde{P}_i(A), \tilde{P}_j(A)) &= \mathrm{Cov}(\mathbb{E}(\tilde{P}_i(A) \mid \tilde{\mu}_0), \mathbb{E}(\tilde{P}_j(A) \mid \tilde{\mu}_0)) \\ &= \mathrm{Var}\left(\frac{\tilde{\mu}_0(A)}{\tilde{\mu}_0(\mathbb{X})}\right) = -P_0(A)(1-P_0(A)) \int_0^{+\infty} u e^{-\psi_0(u)} \psi_0''(u)\, \mathrm{d}u.\end{aligned}$$

S2.8 Proof of Theorem 4.1

Since the random probabilities $\tilde{\mu}_i/\tilde{\mu}_i(\mathbb{X})$ are a.s. discrete with random atoms from a continuous distribution, the model for the observations $\boldsymbol{X}_{1:d} \mid \tilde{\boldsymbol{\mu}}$ is non-dominated. Therefore, we cannot rely on Bayes theorem to find the posterior distribution, and should use an alternative strategy, based on the multivariate Laplace functional. This proof can be seen as the multivariate extension of the proof in James et al. (2009). The multivariate Laplace functional

$$\mathbb{E} \exp\left(-\sum_{i=1}^{d} \int f_i \, \mathrm{d}\tilde{\mu}_i\right)$$

is defined for any non-negative measurable functions $f_1, \dots f_d$ and characterizes the law of any vector of random measures. In particular, when $\tilde{\boldsymbol{\mu}} \sim \mathrm{CRV}(\nu)$, it satisfies

$$\log \mathbb{E} \exp\left(-\sum_{i=1}^{d} \int f_i \, \mathrm{d}\tilde{\mu}_i\right) = -\int_{\Omega_d \times \mathbb{X}} (1 - e^{-\sum_{i=1}^{d} s_i f_i(x)})\, \mathrm{d}\nu(\boldsymbol{s}, x). \tag{S7}$$

Our goal is to find an appropriate expression of $\mathbb{E}(e^{-\sum_{i=1}^{d} \int f_i \mathrm{d}\tilde{\mu}_i} \mid \boldsymbol{X}_{1:d} = \boldsymbol{x}_{1:d})$, for any non-negative measurable functions $f_1, \dots f_d$. Throughout the proof, we exploit the following properties of the conditional expectation, which hold for any random variable X, Y and for any Borel set A such that $\mathbb{P}(Y \in A) \geq 0$:

(a) conditional expectation w.r.t. events: $\mathbb{E}(X \mid Y \in A) = \mathbb{E}(X \mathbb{1}_A(Y))/\mathbb{P}(Y \in A)$;

(b) tower property: $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X \mid Y))$ and $\mathbb{P}(A) = \mathbb{E}(\mathbb{1}_A) = \mathbb{E}(\mathbb{P}(A \mid Y))$.

Step 1. Express the conditional expectation in terms of events. By the dominated convergence theorem and the exchangeability of the observations

$$\mathbb{E}\Big(e^{-\sum_{i=1}^d \int f_i \mathrm{d}\tilde{\mu}_i} \mid \boldsymbol{X}_{1:d} = \boldsymbol{x}_{1:d}\Big) = \lim_{\varepsilon\to 0} \mathbb{E}\Bigg(e^{-\sum_{i=1}^d \int f_i \mathrm{d}\tilde{\mu}_i} \;\Big|\; \boldsymbol{X}_{1:d} \in \prod_{i=1}^d \prod_{j=1}^k B_\varepsilon^{n_{ij}}(x_j^*)\Bigg),$$

where $\boldsymbol{x}^* = (x_1^*, \dots, x_k^*)$ are the unique values in the observations, with multiplicities $n_{i1}, \dots, n_{ik}$ for each group $i = 1, \dots, d$; see Section 4. Here, $B_\varepsilon(x) = \{\omega : d(x,\omega) \le \varepsilon\}$ denotes the ball of radius $\varepsilon > 0$ centered in x, and we use $B_\varepsilon^m(x) = B_\varepsilon(x) \times \cdots \times B_\varepsilon(x)$ for their m-cartesian product. Without loss of generality, we always consider ε sufficiently small for the balls $\{B_\varepsilon(x_j^*)\}_j$ to be pairwise disjoint.

Step 2. Condition with respect to events. Using property (a),

$$\begin{aligned}
&\mathbb{E}\Big(e^{-\sum_{i=1}^d \int f_i \,\mathrm{d}\tilde{\mu}_i} \mid \boldsymbol{X}_{1:d} = \boldsymbol{x}_{1:d}\Big) \\
&\qquad = \lim_{\varepsilon\to 0} \frac{\mathbb{E}\Big(e^{-\sum_{i=1}^d \int f_i \,\mathrm{d}\tilde{\mu}_i} \mathbb{1}_{\prod_{i=1}^d \prod_{j=1}^k B_\varepsilon^{n_{ij}}(x_j^*)}(\boldsymbol{X}_{1:d})\Big)}{\mathbb{P}\big(\boldsymbol{X}_{1:d} \in \prod_{i=1}^d \prod_{j=1}^k B_\varepsilon^{n_{ij}}(x_j^*)\big)} =: \lim_{\varepsilon\to 0} \frac{N_\varepsilon(\boldsymbol{x}^*)}{D_\varepsilon(\boldsymbol{x}^*)},
\end{aligned}$$

where we have introduced the notation $N_\varepsilon(\boldsymbol{x}^*)$ for the numerator and $D_\varepsilon(\boldsymbol{x}^*)$ for the denominator. In the next steps (3–6) we show that both numerator and denominator decrease at the same speed as $\varepsilon \to 0$, namely

$$\begin{aligned}
N_\varepsilon(\boldsymbol{x}^*) &= C_N \prod_{j=1}^k P_0(B_\varepsilon(x_j^*)) + o\Big(\prod_{j=1}^k P_0(B_\varepsilon(x_j^*))\Big), \\
D_\varepsilon(\boldsymbol{x}^*) &= C_D \prod_{j=1}^k P_0(B_\varepsilon(x_j^*)) + o\Big(\prod_{j=1}^k P_0(B_\varepsilon(x_j^*))\Big),
\end{aligned}$$

for some constants $C_N, C_D > 0$, where P_0 is the diffuse base probability of the CRV. It then follows that the limit above coincides with C_N/C_D. Since the denominator is a special case of the numerator when $f_1 = \cdots = f_d = 0$, we focus on finding C_N and then specialize the result to C_D.

Step 3. Express the numerator in terms of $\tilde{\boldsymbol{\mu}}$. By the tower property (b),

$$\begin{aligned}
N_\varepsilon(\boldsymbol{x}^*) &= \mathbb{E}\Bigg(e^{-\sum_{i=1}^d \int f_i \,\mathrm{d}\tilde{\mu}_i}\, \mathbb{P}\Bigg(\boldsymbol{X}_{1:d} \in \prod_{i=1}^d \prod_{j=1}^k B_\varepsilon^{n_{ij}}(x_j^*) \;\Big|\; \tilde{\boldsymbol{\mu}}\Bigg)\Bigg) \\
&= \mathbb{E}\Bigg(e^{-\sum_{i=1}^d \int f_i \,\mathrm{d}\tilde{\mu}_i} \prod_{i=1}^d \prod_{j=1}^k \frac{\tilde{\mu}_i(B_\varepsilon(x_j^*))^{n_{ij}}}{\tilde{\mu}(\mathbb{X})^{n_{ij}}}\Bigg) \\
&= \mathbb{E}\Bigg(\prod_{i=1}^d \frac{1}{\tilde{\mu}_i(\mathbb{X})^{n_i}}\, e^{-\int f_i \,\mathrm{d}\tilde{\mu}_i} \prod_{j=1}^k \tilde{\mu}_i(B_\varepsilon(x_j^*))^{n_{ij}}\Bigg).
\end{aligned}$$

Step 4. Use the gamma trick to separate the integrand into independent components. Using the density of a gamma with shape n_i and rate $\tilde{\mu}_i(\mathbb{X})$, we rewrite

$$\frac{1}{\tilde{\mu}_i(\mathbb{X})^{n_i}} = \frac{1}{\Gamma(n_i)} \int_0^{+\infty} u_i^{n_i-1} e^{-\tilde{\mu}_i(\mathbb{X})u_i} \, \mathrm{d}u_i.$$

Henceforth, we adopt the compact notation $B_j = B_\varepsilon(x_j^*)$ and $B_0 = \mathbb{X} \setminus \{B_1 \sqcup \cdots \sqcup B_k\}$, with the convention $n_0 = n_{i0} = 0$ for $i = 1, \ldots, d$. Using Fubini-Tonelli's theorem and the independence property of a CRV on disjoint set-wise evaluations,

$$\begin{aligned} N_\varepsilon(\boldsymbol{x}^*) &= \mathbb{E}\bigg(\prod_{i=1}^d \frac{1}{\Gamma(n_i)} \int_0^{+\infty} u_i^{n_i-1} \prod_{j=0}^k e^{-\tilde{\mu}_i(B_j)u_i} \, \mathrm{d}u_i \prod_{j=0}^k e^{-\int_{B_j} f_i \, \mathrm{d}\tilde{\mu}_i} \tilde{\mu}_i(B_j)^{n_{ij}} \bigg) \\ &= \frac{1}{\prod_{i=1}^d \Gamma(n_i)} \int_{(0,+\infty)^d} \bigg(\prod_{i=1}^d u_i^{-1} \bigg) \, n_\varepsilon(u; \boldsymbol{x}^*) \, \mathrm{d}\boldsymbol{u}, \end{aligned}$$

where

$$n_\varepsilon(u; \boldsymbol{x}^*) = \prod_{j=0}^k \mathbb{E}\bigg(\prod_{i=1}^d e^{-\int_{B_j} (f_i(x)+u_i) \, \mathrm{d}\tilde{\mu}_i(x)} (u_i \, \tilde{\mu}_i(B_j))^{n_{ij}} \bigg).$$

We now study the asymptotic behaviour of the quantity $n_\varepsilon(u; \boldsymbol{x}^*)$.

Step 5. Express the integrand $n_\varepsilon(\boldsymbol{x}^*)$ in terms of the derivative of the multivariate Laplace functional. Let $\eta_1, \ldots, \eta_d \geq 1$ be auxiliary quantities, such that $n_\varepsilon(u; \boldsymbol{x}^*)$ can be written as

$$\begin{aligned} n_\varepsilon(u; \boldsymbol{x}^*) &= \prod_{j=0}^k \mathbb{E}\bigg(\prod_{i=1}^d \lim_{\eta_i \to 1^+} e^{-\int_{B_j} (f_i(x)+\eta_i u_i) \, \mathrm{d}\tilde{\mu}_i(x)} (u_i \, \tilde{\mu}_i(B_j))^{n_{ij}} \bigg) \\ &= \lim_{\boldsymbol{\eta} \to 1^+} \prod_{j=0}^k \mathbb{E}\bigg(\prod_{i=1}^d e^{-\int_{B_j} (f_i(x)+\eta_i u_i) \, \mathrm{d}\tilde{\mu}_i(x)} (u_i \, \tilde{\mu}_i(B_j))^{n_{ij}} \bigg), \end{aligned}$$

where $\boldsymbol{\eta} \to 1^+$ is a compact notation for $\eta_i \to 1^+$ for each $i = 1, \ldots, d$, and we have exchanged limit and expectation by monotone convergence theorem. We observe that, for some $\ell \in \{1, \ldots, d\}$,

$$\frac{\partial}{\partial \eta_\ell} \prod_{i=1}^d e^{-\int_{B_j} (f_i(x)+\eta_i u_i) \, \mathrm{d}\tilde{\mu}_i(x)} = -u_\ell \, \tilde{\mu}_\ell(B_j) \prod_{i=1}^d e^{-\int_{B_j} (f_i(x)+\eta_i u_i) \, \mathrm{d}\tilde{\mu}_i(x)}.$$

This formula can be applied recursively, using the convention $d^0/du^0 = \mathrm{Id}$, for Id the identity function. Specifically,

$$n_\varepsilon(u; \boldsymbol{x}^*) = \lim_{\boldsymbol{\eta} \to 1^+} \prod_{j=0}^k (-1)^{n_{\bullet j}} \, \mathbb{E}\bigg(\frac{\partial^{n_{\bullet j}}}{\partial \eta_1^{n_{1j}} \ldots \partial \eta_d^{n_{dj}}} \prod_{i=1}^d e^{-\int_{B_j} (f_i(x)+\eta_i u_i) \, \mathrm{d}\tilde{\mu}_i(x)} \bigg),$$

where $n_{\bullet j} = n_{1j} + \cdots + n_{dj}$. Since $\eta_i \geq 1$ and $f_i \geq 0$, the derivative is bounded above by the product

$$\prod_{i=1}^{d} (u_i\, \tilde{\mu}_i(B_j))^{n_{ij}} \exp(-u_i \tilde{\mu}_i(B_j)),$$

which has finite mean since the exponential decay of $\exp(-\sum_{i=1}^{d} u_i \tilde{\mu}_i(B_j))$ is not compromised by the slower polynomial growth of $\prod_{i=1}^{d} (u_i\, \tilde{\mu}_i(B_j))^{n_{ij}}$. Therefore, the derivative is uniformly integrable in $\eta_1, \ldots, \eta_d$ and we can exchange derivative and expectation. Remarkably, we need to introduce $\boldsymbol{\eta}$ and cannot derive the expressions directly with respect to $\boldsymbol{u}$: otherwise, the derivative would not be uniformly integrable for CRMs with unbounded moment measures, such as the σ-stable CRM. Using the expression (S5) of the multivariate Laplace functional of a CRV, we obtain

$$\begin{aligned} n_\varepsilon(u; \boldsymbol{x}^*) &= \lim_{\boldsymbol{\eta} \to 1^+} \prod_{j=0}^{k} (-1)^{n_{\bullet j}} \frac{\partial^{n_{\bullet j}}}{\partial \eta_1^{n_{1j}} \ldots \partial \eta_d^{n_{dj}}} \mathbb{E}\left(e^{-\sum_{i=1}^{d} \int_{B_j} (f_i(x) + \eta_i u_i)\, \mathrm{d}\tilde{\mu}_i(x)} \right) \\ &= \lim_{\boldsymbol{\eta} \to 1^+} \prod_{j=0}^{k} (-1)^{n_{\bullet j}} \frac{\partial^{n_{\bullet j}}}{\partial \eta_1^{n_{1j}} \ldots \partial \eta_d^{n_{dj}}} \exp\left(- \int_{\Omega_d \times B_j} \left(1 - e^{-\boldsymbol{s} \cdot (\boldsymbol{f}(x) + \boldsymbol{\eta u})}\right) \mathrm{d}\nu(\boldsymbol{s}, x) \right), \end{aligned}$$

where we have used the notation $\boldsymbol{f}(x) = (f_1(x), \ldots, f_d(x))$ and $\boldsymbol{\eta u} = (\eta_1 u_1, \ldots, \eta_d u_d)$. For $j = 0, \ldots, k$, define the function

$$g_j(\boldsymbol{\eta}, \boldsymbol{u}) = - \int_{\Omega_d \times B_j} \left(1 - e^{-\boldsymbol{s} \cdot (\boldsymbol{f}(x) + \boldsymbol{\eta u})}\right) \mathrm{d}\nu(\boldsymbol{s}, x).$$

Step 6. Determine the asymptotic behaviour of the partial derivatives of e^{-g_j}. For $j = 1, \ldots, k$, the partial derivatives of g_j satisfy, as $\varepsilon \to 0$,

$$\begin{aligned} &\frac{\partial^{n_{\bullet j}}}{\partial \eta_1^{n_{1j}} \ldots \partial \eta_d^{n_{dj}}} g_j(\boldsymbol{\eta}, \boldsymbol{u}) = \\ &\qquad = (-1)^{n_{\bullet j}} \left(\prod_{i=1}^{d} u_i^{n_{ij}} \right) \int_{\Omega_d \times B_j} \prod_{i=1}^{d} s_i^{n_{ij}} e^{-\boldsymbol{s} \cdot (\boldsymbol{f}(x) + \boldsymbol{\eta u})}\, \mathrm{d}\nu(\boldsymbol{s}, x) \\ &\qquad = (-1)^{n_{\bullet j}} \left(\prod_{i=1}^{d} u_i^{n_{ij}} \right) \int_{\Omega_d} \prod_{i=1}^{d} s_i^{n_{ij}} e^{-\boldsymbol{s} \cdot (\boldsymbol{f}(x_j^*) + \boldsymbol{\eta u})} \mathrm{d}\rho_{x_j^*}(\boldsymbol{s})\, P_0(B_j) + o(P_0(B_j)). \end{aligned}$$

The first line comes from uniform integrability of the derivative, which allows exchanging the integral and derivative operators; the second line follows from Lebesgue differentiation theorem for atomless measures on Polish spaces.

The multivariate Faà di Bruno formula (Constantine and Savits, 1996) allows to express multiple partial derivatives of a function. We introduce the notation as in Hardy (2006), who observed that partial derivatives of the type $\partial^{n_\bullet} / \partial \eta_1^{n_1} \ldots \partial \eta_d^{n_d}$ can be treated as *maximally mixed* partial derivatives of the type $\partial^{n_\bullet} / \partial v_1 \cdot \ldots \partial v_{n_\bullet}$ by allowing for ties

among the variables, which brings to more compact expressions. In particular, define $v_1 = \cdots = v_{n_{1j}} = \eta_1$, and $v_{\sum_{i'=1}^{i-1} n_{i'j}+1} = \cdots = v_{\sum_{i'=1}^{i} n_{i'j}} = \eta_i$ for $i = 2, \ldots, d$, so that,

$$(-1)^{n_{\bullet j}} \frac{\partial^{n_{\bullet j}}}{\partial \eta_1^{n_{1j}} \ldots \partial \eta_d^{n_{dj}}} e^{-g_j(\boldsymbol{\eta},\boldsymbol{u})} = e^{-g(\boldsymbol{\eta},\boldsymbol{u})} \sum_{\pi} \prod_{A\in\pi} \frac{\partial^{|A|}}{\prod_{i\in A} \partial v_i} g_j(\boldsymbol{\eta},\boldsymbol{u}), \tag{S8}$$

where the sum is over all partitions π of the numbers $\{1, \ldots, n_{\bullet j}\}$. From this expression, one can derive a formula in terms of $\eta_1, \ldots, \eta_d$ through appropriate combinatorial coefficients, retrieving the one in Constantine and Savits (1996). However, in our case, the combinatorial formulation above is not necessary, as we are only interested in the asymptotic behaviour as $\varepsilon \to 0$. From the previous discussion, all terms $\partial^{|A|}/(\prod_{i\in A} \partial v_i) g(\boldsymbol{\eta},\boldsymbol{u})$ are asymptotically equivalent up to a constant. Hence, from (S8), the asymptotically slowest term is the summand corresponding to the partition π with a minimal number of sets, that is, $\pi = \{1, \ldots, n_{\bullet j}\}$. Therefore, as $\varepsilon \to 0$, (S8) reads

$$\begin{aligned}(-1)^{n_{\bullet j}} \frac{\partial^{n_{\bullet j}}}{\partial \eta_1^{n_{1j}} \ldots \partial \eta_d^{n_{dj}}} e^{-g_j(\boldsymbol{\eta},\boldsymbol{u})} &= \exp\left(- \int_{\Omega_d \times B_j} (1 - e^{-\boldsymbol{s}\cdot(\boldsymbol{f}(x)+\boldsymbol{\eta}\boldsymbol{u})}) \,\mathrm{d}\nu(\boldsymbol{s},x)\right) \\ &\quad \times \left(\prod_{i=1}^d u_i^{n_{ij}}\right) \int_{\Omega_d} \prod_{i=1}^d s_i^{n_{ij}} e^{-\boldsymbol{s}\cdot(\boldsymbol{f}(x_j^*)+\boldsymbol{\eta}\boldsymbol{u})} \mathrm{d}\rho_{x_j^*}(\boldsymbol{s})\, P_0(B_j) + o(P_0(B_j)).\end{aligned}$$

Step 7. Determine the asymptotic behaviour of $n_\varepsilon(\boldsymbol{x}^*)$. Considering the term e^{-g_0}, which does not vanish as $\varepsilon \to 0$, we obtain that $n_\varepsilon(u; \boldsymbol{x}^*)$ is asymptotically equal to

$$\begin{aligned}n_\varepsilon(\boldsymbol{u}; \boldsymbol{x}^*) &= \exp\left(- \int_{\Omega_d \times \mathbb{X}} (1 - e^{-\boldsymbol{s}\cdot(\boldsymbol{f}(x)+\boldsymbol{u})}) \,\mathrm{d}\nu(\boldsymbol{s},x)\right) \\ &\quad \times \left(\prod_{i=1}^d u_i^{n_i}\right) \prod_{j=1}^k \int_{\Omega_d} \prod_{i=1}^d s_i^{n_{ij}} e^{-\boldsymbol{s}\cdot(\boldsymbol{f}(x_j^*)+\boldsymbol{u})} \mathrm{d}\rho_{x_j^*}(\boldsymbol{s}) \prod_{j=1}^k P_0(B_j) + o\Big(\prod_{j=1}^k P_0(B_j)\Big),\end{aligned}$$

where we have computed the limit for $\boldsymbol{\eta} \to 1^+$ under the integration using monotone convergence.

Step 8. Determine the asymptotic behaviour of the numerator $N_\varepsilon(\boldsymbol{x}^*)$. From the relation between $n_\varepsilon(u; \boldsymbol{x}^*)$ and $N_\varepsilon(\boldsymbol{x}^*)$ in Step 4, by monotone convergence theorem, we have that $N_\varepsilon(\boldsymbol{x}^*) = C_N \prod_{j=1}^k P_0(B_j) + o(\prod_{j=1}^k P_0(B_j))$, where C_N equals

$$\begin{aligned}\frac{1}{\prod_{i=1}^d \Gamma(n_i)} \int_{(0,+\infty)^d} \left(\prod_{i=1}^d u_i^{n_i-1}\right) &\exp\left(- \int_{\Omega_d \times \mathbb{X}} (1 - e^{-\boldsymbol{s}\cdot(\boldsymbol{f}(x)+\boldsymbol{u})}) \,\mathrm{d}\nu(\boldsymbol{s},x)\right) \\ &\times \prod_{j=1}^k \left(\int_{\Omega_d} \prod_{i=1}^d s_i^{n_{ij}} e^{-\boldsymbol{s}\cdot(\boldsymbol{f}(x_j^*)+\boldsymbol{u})} \mathrm{d}\rho_{x_j^*}(\boldsymbol{s})\right) \mathrm{d}\boldsymbol{u}.\end{aligned}$$

Step 9. Determine the expression of the Laplace functional a posteriori. By specializing the formula in Step 8 for C_N to $f = 0$, we determine the value of C_D, which can be conveniently expressed in terms of the multivariate Laplace exponent ψ and cumulants $\tau_{n_{1j},\dots,n_{dj}|x_j^*}$ as

$$C_D = \frac{1}{\prod_{i=1}^d \Gamma(n_i)} \int_{(0,+\infty)^d} \left(\prod_{i=1}^d u_i^{n_i-1} \right) e^{-\psi(\boldsymbol{u})} \prod_{j=1}^k \tau_{n_{1j},\dots,n_{dj}|x_j^*}(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u}.$$

It follows that the Laplace functional a posteriori is equal to

$$\begin{aligned}\mathbb{E}\Big(e^{-\sum_{i=1}^d \int f_i \mathrm{d}\tilde{\mu}_i} \mid \boldsymbol{X}_{1:d} = \boldsymbol{x}_{1:d}\Big) \\ = \frac{\int_{(0,+\infty)^d} \left(\prod_{i=1}^d u_i^{n_i-1} \right) e^{-\int_{\mathbb{X}} \psi(\boldsymbol{f}(x)+\boldsymbol{u}) \, \mathrm{d}P_0(x)} \, \mathrm{d}\boldsymbol{u}}{\int_{(0,+\infty)^d} \left(\prod_{i=1}^d u_i^{n_i-1} \right) e^{-\psi(\boldsymbol{u})} \prod_{j=1}^k \tau_{n_{1j},\dots,n_{dj}|x_j^*}(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u}}.\end{aligned}$$

Step 10. Interpret the denominator as the normalizing constant of latent random variables $U_1, \dots, U_d$. The multivariate Laplace exponent $\psi(\boldsymbol{u})$ can be retrieve at the numerator by multiplying and dividing by $e^{-\psi(\boldsymbol{u})}$. Indeed,

$$\begin{aligned}\int_{\Omega_d \times \mathbb{X}} (1 - e^{-\boldsymbol{s}\cdot(\boldsymbol{f}(x)+\boldsymbol{u})}) \, \mathrm{d}\nu(\boldsymbol{s}, x) + \psi(\boldsymbol{u}) &= \int_{\Omega_d \times \mathbb{X}} (e^{-\boldsymbol{s}\cdot\boldsymbol{u}} - e^{-\boldsymbol{s}\cdot(\boldsymbol{f}(x)+\boldsymbol{u})}) \, \mathrm{d}\nu(\boldsymbol{s}, x) \\ &= \int_{\Omega_d \times \mathbb{X}} (1 - e^{-\boldsymbol{s}\boldsymbol{f}(x)}) e^{-\boldsymbol{s}\cdot\boldsymbol{u}} \, \mathrm{d}\nu(\boldsymbol{s}, x).\end{aligned}$$

Therefore, the posterior Laplace functional is rewritten as

$$\begin{aligned}\mathbb{E}\Big(e^{-\sum_{i=1}^d \int f_i \mathrm{d}\tilde{\mu}_i} \mid \boldsymbol{X}_{1:d} = \boldsymbol{x}_{1:d}\Big) \\ = \mathbb{E}\left(e^{-\int_{\Omega_d \times \mathbb{X}} (1 - e^{-\boldsymbol{s}\boldsymbol{f}(x)}) e^{-\boldsymbol{s}\cdot\boldsymbol{U}} \mathrm{d}\nu(\boldsymbol{s},x)} \prod_{j=1}^k \frac{\int_{\Omega_d} \prod_{i=1}^d s_i^{n_{ij}} e^{-\boldsymbol{s}\cdot(\boldsymbol{f}(x_j^*)+\boldsymbol{U})} \mathrm{d}\rho_{x_j^*}(\boldsymbol{s})}{\tau_{n_{1j},\dots,n_{dj}|x_j^*}(\boldsymbol{U})} \right),\end{aligned}$$

where $\boldsymbol{U} = (U_1, \dots, U_d)$ is a vector of random variables with joint p.d.f.

$$f_{\boldsymbol{U}}(\boldsymbol{u}) \propto \prod_{i=1}^d u_i^{n_i-1} e^{-\psi(\boldsymbol{u})} \prod_{j=1}^k \tau_{n_{1j},\dots,n_{dj}|x_j^*}(\boldsymbol{u}).$$

This implies that there exist latent variables $\boldsymbol{U} = (U_1, \dots, U_d)$ such that the posterior Laplace functional of $\tilde{\boldsymbol{\mu}}$, conditionally on $\boldsymbol{U}$, satisfies

$$\mathbb{E}\Big(e^{-\sum_{i=1}^d \int f_i \mathrm{d}\tilde{\mu}_i} \mid \boldsymbol{X}_{1:d} = \boldsymbol{x}_{1:d}, \boldsymbol{U}\Big) = \mathbb{E}\Big(e^{-\sum_{i=1}^d \int f_i \mathrm{d}\tilde{\mu}_i^*} \mid \boldsymbol{U}\Big) \prod_{j=1}^k \mathbb{E}\Big(e^{-\boldsymbol{f}(x_j^*)\cdot\boldsymbol{J}_j} \mid \boldsymbol{U}\Big),$$

where, conditionally on $\boldsymbol{U}$, the measure $\tilde{\boldsymbol{\mu}}^*$ is a CRV with Lévy intensity $e^{-\boldsymbol{s}\cdot\boldsymbol{U}} \mathrm{d}\nu(\boldsymbol{s}, x)$

and $\boldsymbol{J}_j$ is a vector of jumps with distribution

$$\mathrm{d}P_{\boldsymbol{J}_j|\boldsymbol{U}}(\boldsymbol{s}) \propto \prod_{i=1}^{d} s_i^{n_{ij}} e^{-\boldsymbol{s}\cdot\boldsymbol{U}} \mathrm{d}\rho_{x_j^*}(\boldsymbol{s}).$$

If we define the random elements $\tilde{\boldsymbol{\mu}}^*$ and $\boldsymbol{J}_j$ to be conditionally independent given $\boldsymbol{U}$, the Laplace functional of their sum is the product of their Laplace functions, and thus

$$\mathbb{E}\Big(e^{-\sum_{i=1}^{d}\int f_i \mathrm{d}\tilde{\mu}_i} \mid \boldsymbol{X}_{1:d} = \boldsymbol{x}_{1:d}, \boldsymbol{U}\Big) = \mathbb{E}\Big(e^{-\sum_{i=1}^{d}\int f_i \mathrm{d}(\tilde{\mu}_i^* + \sum_{j=1}^{k} J_{ij}\delta_{x_j^*})} \mid \boldsymbol{U}\Big).$$

By uniqueness of the Laplace functional, this implies that

$$\mathcal{L}(\tilde{\boldsymbol{\mu}} \mid \boldsymbol{X}_{1:d} = \boldsymbol{x}_{1:d}) = \mathcal{L}\Big(\tilde{\boldsymbol{\mu}}^* + \sum_{j=1}^{k} \boldsymbol{J}_j \delta_{x_j^*}\Big).$$

S2.9 Proof of Proposition 4.2

We first prove a preliminary Lemma on the exponential tilting of a Lévy measure.

Lemma S2.2. *Let ρ be a Lévy measure on $(0,\infty)$ with Laplace exponent ψ such that $\mathrm{ID}(\rho)$ has a p.d.f. denoted by $f_{\mathrm{ID}(\rho)}$. For $u>0$, define $\mathrm{d}\rho_u(s) = e^{-us}\mathrm{d}\rho(s)$ the exponential tilting of ρ. Then for $s,t,u>0$,*

$$e^{-us} f_{\mathrm{ID}(t\rho)}(s) = e^{-t\psi(u)} f_{\mathrm{ID}(t\rho_u)}(s).$$

Proof. Let $X \sim \mathrm{ID}(t\rho)$ and let $X_u \sim \mathrm{ID}(t\rho_u)$. By the uniqueness of the Laplace transform, it is enough to show that, for every $\lambda > 0$,

$$\begin{aligned}\mathbb{E}\big(e^{-\lambda X_u}\big) &= \int_0^\infty e^{-\lambda s} f_{\mathrm{ID}(t\rho_u)}(s)\,\mathrm{d}s \\ &= e^{t\psi(u)} \int_0^\infty e^{-(\lambda+u)\,s} f_{\mathrm{ID}(t\rho)}(s)\,\mathrm{d}s = e^{t\psi(u)}\,\mathbb{E}\big(e^{-(\lambda+u)X}\big).\end{aligned}$$

Indeed, the Laplace transform of X_u is equal to

$$\begin{aligned}\mathbb{E}\big(e^{-\lambda X_u}\big) &= \int_0^\infty e^{-\lambda s} f_{\mathrm{ID}(t\rho_u)}(s)\,\mathrm{d}s = e^{-\int_0^{+\infty}(1-e^{-\lambda s})\,te^{-us}\,\mathrm{d}\rho(s)} \\ &= e^{-t\int_0^{+\infty}(e^{-us}-e^{-(\lambda+u)s})\mathrm{d}\rho(s)} = e^{t\psi(u)-t\psi(\lambda+u)} \\ &= e^{t\psi(u)} \int_0^\infty e^{-(\lambda+u)\,s} f_{\mathrm{ID}(t\rho)}(s)\,\mathrm{d}s = e^{t\psi(u)}\,\mathbb{E}\big(e^{-(\lambda+u)X}\big).\end{aligned}$$

□

Considering $\tilde{\boldsymbol{\mu}} \sim \mathrm{hCRV}(\rho, \rho_0, P_0)$, the expression of the Lévy intensity in Theorem 4.1(i) is given by

$$\mathrm{d}\nu_U^*(\boldsymbol{s}, x) = e^{-\boldsymbol{U}\cdot\boldsymbol{s}}\mathrm{d}\nu_h(\boldsymbol{s}, x) = e^{-\boldsymbol{U}\cdot\boldsymbol{s}}\rho_h(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s}\,\mathrm{d}P_0(x),$$

where we have used that $\mathrm{ID}(t\rho)$ has a p.d.f. on $(0,+\infty)$. Exploiting Lemma S2.2 for the exponential tilting, we obtain

$$\begin{aligned} e^{-\boldsymbol{U}\cdot\boldsymbol{s}}\rho_h(\boldsymbol{s}) &= \int_0^{+\infty} \prod_{i=1}^{d} \left(e^{-U_i s_i} f_{\mathrm{ID}(t\rho)}(s_i)\right) \rho_0(t)\,\mathrm{d}t \\ &= \int_0^{+\infty} \prod_{i=1}^{d} \left(e^{-t\psi(U_i)} f_{\mathrm{ID}(t\rho_{U_i})}(s_i)\right) \rho_0(t)\,\mathrm{d}t \\ &= \int_0^{+\infty} \prod_{i=1}^{d} f_{\mathrm{ID}(t\rho_{U_i})}(s_i)\, e^{-t\sum_{i=1}^{d}\psi(U_i)} \rho_0(t)\,\mathrm{d}t. \end{aligned}$$

In analogy with Theorem 2.3, this expression can be interpreted as the Laplace exponent of a hierarchical CRV with heterogeneous marginal distributions, characterized by the Lévy measures

$$\mathrm{d}\rho_i^*(s) = \mathrm{d}\rho_{U_i}(s) = e^{-U_i s}\rho(s)\,\mathrm{d}s, \qquad \mathrm{d}\rho_0^*(t) = e^{-t\sum_{i=1}^{d}\psi(U_i)}\rho_0(t)\,\mathrm{d}t.$$

S2.10 Proof of Proposition 4.3

Since ρ and ρ_0 have Lévy densities, the distribution of the vector of jumps $\boldsymbol{J}_j$ in (7), given the latent variables $\boldsymbol{U}$, has p.d.f. proportional to

$$\begin{aligned} \prod_{i=1}^{d} s_i^{n_{ij}}\, e^{-\boldsymbol{U}\cdot\boldsymbol{s}}\rho_h(\boldsymbol{s}) &= \prod_{i=1}^{d} s_i^{n_{ij}} e^{-U_i s_i} \int_0^{+\infty} \prod_{i=1}^{d} f_{\mathrm{ID}(t\rho)}(s_i)\,\rho_0(t)\,\mathrm{d}t \\ &= \int_0^{+\infty} \prod_{i=1}^{d} s_i^{n_{ij}} e^{-U_i s_i} f_{\mathrm{ID}(t\rho)}(s_i)\,\rho_0(t)\,\mathrm{d}t \\ &= \int_0^{+\infty} \prod_{i=1}^{d} \frac{s_i^{n_{ij}} e^{-U_i s_i} f_{\mathrm{ID}(t\rho)}(s_i)}{\bar{\tau}_{n_{ij}}(U_i,t)} \prod_{i=1}^{d} \bar{\tau}_{n_{ij}}(U_i,t)\,\rho_0(t)\,\mathrm{d}t, \end{aligned}$$

where $\bar{\tau}_m(u,t)$ is defined in (8). Therefore, the distribution of $\boldsymbol{J}_j \mid \boldsymbol{U}$ is a mixture of conditionally independent random variables $J_{1j},\ldots,J_{dj}$ with densities

$$f_{J_{ij}|U_i,J_{0j}}(s) = \frac{s^{n_{ij}} e^{-U_i s} f_{\mathrm{ID}(J_{0j}\rho)}(s)}{\bar{\tau}_{n_{ij}}(U_i,J_{0j})}, \qquad i=1,\ldots,d,$$

given a mixing random variable J_{0j} having p.d.f. proportional to $\prod_{i=1}^{d}\bar{\tau}_{n_{ij}}(U_i,t)\,\rho_0(t)$.

S2.11 Proof of Lemma 4.4

Substituting the expression for ρ_h as in the proof of Proposition 4.3, we use Fubini-Tonelli Theorem as in Remark 3.2 to obtain

$$\begin{aligned}\tau_{\boldsymbol{m}}(\boldsymbol{u}) &= \int_{\Omega_d} \prod_{i=1}^{d} s_i^{m_i} e^{-\boldsymbol{u}\cdot\boldsymbol{s}} \rho_h(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s} = \int_0^{+\infty} \int_{\Omega_d} \prod_{i=1}^{d} s_i^{m_i} e^{-u_i s_i} f_{\mathrm{ID}(t\rho)}(s_i)\,\mathrm{d}\boldsymbol{s}\,\rho_0(t)\,\mathrm{d}t \\ &= \int_0^{+\infty} \prod_{i=1}^{d} \int_0^{+\infty} s_i^{m_i} e^{-u_i s_i} f_{\mathrm{ID}(t\rho)}(s_i)\,\mathrm{d}s_i\,\rho_0(t)\,\mathrm{d}t = \int_0^{+\infty} \prod_{i=1}^{d} \bar{\tau}_{m_i}(u_i,t)\rho_0(t)\,\mathrm{d}t.\end{aligned}$$

S2.12 Proof of Proposition 4.5

Starting from the expression of the distribution of the latent variables $\boldsymbol{U}$ in (6), specialized for $\tilde{\boldsymbol{\mu}} \sim \mathrm{hCRV}(\rho, \rho_0, P_0)$,

$$\begin{aligned}f_U(\boldsymbol{u}) &\propto \prod_{i=1}^{d} u_i^{n_i-1} e^{-\psi_0(\sum_{i=1}^{d}\psi(u_i))} \prod_{j=1}^{k} \tau_{n_{1j},\ldots,n_{dj}}(\boldsymbol{u}) \\ &= \prod_{i=1}^{d} u_i^{n_i-1} e^{-\psi_0(\sum_{i=1}^{d}\psi(u_i))} \prod_{j=1}^{k} \int_0^{+\infty} \prod_{i=1}^{d} \bar{\tau}_{m_i}(u_i,t_j)\rho_0(t_j)\,\mathrm{d}t_j,\end{aligned}$$

where we have substituted the expression of $\tau_{\boldsymbol{m}}(\boldsymbol{u})$ obtained in Lemma 4.4. For each $j = 1,\ldots,k$, exploiting the definition of $\bar{\tau}_m(u,t)$ in (8), we obtain

$$\begin{aligned}\int_0^{+\infty} \prod_{i=1}^{d} \bar{\tau}_{m_i}(u_i,t_j)\rho_0(t_j)\,\mathrm{d}t_j &= \int_0^{+\infty} \prod_{i=1}^{d} \int_0^{\infty} s_{ij}^{n_{ij}} e^{-u_i s_{ij}} f_{\mathrm{ID}(t_j\rho)}(s_{ij})\,\mathrm{d}s_{ij}\,\rho_0(t_j)\,\mathrm{d}t_j \\ &= \int_{\Omega_d} \prod_{i=1}^{d} e^{-u_i s_{ij}} \int_0^{+\infty} \prod_{i=1}^{d} s_{ij}^{n_{ij}} f_{\mathrm{ID}(t_j\rho)}(s_{ij})\,\rho_0(t_j)\,\mathrm{d}t_j\,\mathrm{d}\boldsymbol{s}_{\bullet j},\end{aligned}$$

where we have exchanged the integrals thanks to Fubini-Tonelli's theorem. Recall that, conditionally on $\tilde{\mu}_0$, the distribution of $\tilde{\mu}_i(\mathbb{X}) \sim \mathrm{ID}(\tilde{\mu}_0(\mathbb{X})\rho)$, while at the root of the hierarchy $\tilde{\mu}_0(\mathbb{X}) \sim \mathrm{ID}(\rho_0)$. Therefore, exploiting the definition of multivariate Laplace exponent,

$$\begin{aligned}\prod_{i=1}^{d} u_i^{n_i-1} e^{-\psi_0(\sum_{i=1}^{d}\psi(u_i))} &= \prod_{i=1}^{d} u_i^{n_i-1} \mathbb{E}\big(e^{-\boldsymbol{u}\cdot\tilde{\boldsymbol{\mu}}(\mathbb{X})}\big) \\ &= \int_{\Omega_d} \prod_{i=1}^{d} u_i^{n_i-1} e^{-u_i y_i} \int_0^{\infty} \prod_{i=1}^{d} f_{\mathrm{ID}(z\rho)}(y_i)\, f_{\mathrm{ID}(\rho_0)}(z)\,\mathrm{d}z.\end{aligned}$$

In conclusion, the density of the vector of latent variables $\boldsymbol{U}$ is proportional to

$$f_{\boldsymbol{U}}(\boldsymbol{u}) \propto \int_{\Omega_d^{k+1}} \prod_{i=1}^{d} u_i^{n_i-1} e^{-u_i\,(s_{i0}+s_{i1}+\cdots+s_{ik})} \times \int_{(0,+\infty)^{k+1}} \prod_{i=1}^{d} f_{\mathrm{ID}(t_0\rho)}(s_{i0})\, f_{\mathrm{ID}(\rho_0)}(t_0) \prod_{j=1}^{k} s_{ij}^{n_{ij}} f_{\mathrm{ID}(t_j\rho)}(s_{ij})\, \rho_0(t_j)\, \mathrm{d}\boldsymbol{t}\, \mathrm{d}\boldsymbol{s},$$

where we have renamed integration variables y_i as s_{i0} and z as t_0. Therefore, conditionally on a vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$ of dependent random variables, the latent variables $U_1, \ldots, U_d$ are independent and gamma distributed, with $U_i \sim \mathrm{Gamma}(n_i, \beta_i)$, for $i = 1, \ldots, d$. Moreover, each $\beta_i = S_{i0} + S_{i1} + \cdots + S_{ik}$, where the joint density of $\boldsymbol{S} = (S_{ij})_{ij}$ is proportional to

$$f_{\boldsymbol{S}}(\boldsymbol{s}) \propto \prod_{i=1}^{d} (s_{i0} + s_{i1} + \cdots + s_{ik})^{-n_i} \times \int_{(0,+\infty)^{k+1}} \prod_{i=1}^{d} f_{\mathrm{ID}(t_0\rho)}(s_{i0})\, f_{\mathrm{ID}(\rho_0)}(t_0) \prod_{j=1}^{k} s_{ij}^{n_{ij}} f_{\mathrm{ID}(t_j\rho)}(s_{ij})\, \rho_0(t_j)\, \mathrm{d}\boldsymbol{t}\, \mathrm{d}\boldsymbol{s}.$$

This entails that, conditionally on $\boldsymbol{T} = (T_0, \ldots, T_k)$, $\boldsymbol{S}_i = (S_{i0}, S_{i1}, \cdots, S_{ik})$ are independent for each $i = 1, \ldots, d$, with density proportional to

$$f_{\boldsymbol{S}_i|\boldsymbol{T}}(\boldsymbol{s}_i) \propto (s_{i0} + s_{i1} + \cdots + s_{ik})^{-n_i} f_{\mathrm{ID}(T_0\rho)}(s_{i0}) \prod_{j=1}^{k} s_{ij}^{n_{ij}} f_{\mathrm{ID}(T_j\rho)}(s_{ij}).$$

Finally, the density of vector $\boldsymbol{T}$ is proportional to

$$f_{\boldsymbol{T}}(\boldsymbol{t}) \propto \prod_{i=1}^{d} C(n_{i1}, \ldots, n_{ik}; \boldsymbol{t})\, f_{\mathrm{ID}(\rho_0)}(t_0) \prod_{j=1}^{k} \rho_0(t_j),$$

where C, defined in (9), represents the normalizing constant for the distribution of $\boldsymbol{S}_i \mid \boldsymbol{T}$.

S2.13 Proof of Proposition 5.1

From Example 2.7, the gamma-gamma hCRV is characterized by

$$\rho(s) = \alpha\, \frac{e^{-bs}}{s}, \qquad \rho_0(s) = \alpha_0\, \frac{e^{-b_0 s}}{s},$$

where $\alpha, \alpha_0 > 0$ are shape parameters and $b, b_0 > 0$ rate parameters. The Laplace exponent in Definition S1.3 is $\psi(\lambda) = \alpha \log(1 + \lambda/b)$. The rest of the proof follows from Propositions 4.2 and 4.3.

(a) From Proposition 4.2, and substituting the expressions for ρ, ρ_0 and ψ,

$$e^{-U_i s}\rho(s) = \alpha s^{-1} e^{-bs-U_i s} = \alpha s^{-1} e^{-b(1+U_i/b)s},$$
$$e^{-\sum_{i=1}^d \psi(U_i)s}\rho_0(s) = \alpha_0 s^{-1} e^{-b_0 s - \alpha\sum_{i=1}^d \log(1+U_i/b)s} = \alpha_0 s^{-1} e^{-\alpha\lambda(\boldsymbol{U})s},$$

where $\lambda(\boldsymbol{U}) = b_0/\alpha + \sum_{i=1}^d \log(1+U_i/b)$.

(b) From Proposition 4.3, for each $j = 1, \ldots, k$, the jumps $J_{1j}, \ldots, J_{dj}$ are conditionally independent, given $\boldsymbol{U}$ and J_{0j}. Moreover, from the specification of ρ above, the random variable $\mathrm{ID}(t\rho)$ has gamma distribution with shape parameter αt and rate parameter b. Hence, for each $i = 1, \ldots, d$, the jump J_{ij} has density proportional to

$$s^{n_{ij}} e^{-U_i s} f_{\mathrm{ID}(J_{0j}\rho)}(s) \propto s^{\alpha J_{0j} + n_{ij} - 1} e^{-bs - U_i s},$$

which is the density of a gamma random variable with shape $\alpha J_{0j} + n_{ij}$ and rate $b + U_i$. The normalizing constant $\bar{\tau}_{n_{ij}}(U_i, t)$ in (8) is

$$\bar{\tau}_{n_{ij}}(U_i, t) = \int_0^{+\infty} s^{n_{ij}} e^{-U_i s} f_{\mathrm{ID}(t\rho)}(s)\,\mathrm{d}s = \frac{b^{\alpha t}}{(b+U_i)^{n_{ij}+\alpha t}} \frac{\Gamma(n_{ij}+\alpha t)}{\Gamma(\alpha t)}.$$

(c) Again from Proposition 4.3, for $j = 1, \ldots, k$, the density of J_{0j}, given $\boldsymbol{U}$, is proportional to

$$f_{J_{0j}|\boldsymbol{U}}(t) \propto \prod_{i=1}^d \bar{\tau}_{n_{ij}}(U_i, t)\,\rho_0(t) \propto \prod_{i=1}^d \left(\left(\frac{1}{1+U_i/b} \right)^{\alpha t} ((\alpha t))_{n_{ij}} \right) \alpha_0 \frac{e^{-b_0 t}}{t}\,\mathrm{d}t$$
$$\propto t^{-1}\, e^{-b_0 t - \alpha t \sum_{i=1}^d \log(1+U_i/b)} \prod_{i=1}^d ((\alpha t))_{n_{ij}} \propto t^{-1}\, e^{-\alpha\lambda(\boldsymbol{U})t} \prod_{i=1}^d ((\alpha t))_{n_{ij}},$$

where $((\alpha t))_n = \Gamma(\alpha t + n)/\Gamma(\alpha t)$ denotes the ascending factorial. The result is obtained computing the density of the linear transformation αJ_{0j}.

S2.14 Proof of Proposition 5.2

Recall that, from the specification of ρ in Example 2.7, the random variable $\mathrm{ID}(t\rho)$ has gamma distribution with shape parameter αt and rate parameter b. The rest of the proof follows from Proposition 4.5.

(a) For each $i = 1, \ldots, d$, the density of $\boldsymbol{S}_i = (S_{i0}, \ldots, S_{ik})$, given $\boldsymbol{T} = (T_0, \ldots, T_k)$, is proportional to

$$f_{\boldsymbol{S}_i|\boldsymbol{T}}(s_{i0}, \ldots, s_{ik}) \propto (s_{i0} + s_{i1} + \cdots + s_{ik})^{-n_i} f_{\mathrm{ID}(T_0\rho)}(s_{i0}) \prod_{j=1}^k s_{ij}^{n_{ij}} f_{\mathrm{ID}(T_j\rho)}(s_{ij})$$

$$\propto (s_{i0}+s_{i1}+\cdots+s_{ik})^{-n_i} s_{i0}^{\alpha T_0-1} e^{-bs_{i0}} \prod_{j=1}^{k} s_{ij}^{\alpha T_j+n_{ij}-1} e^{-bs_{ij}}$$

$$\propto (s_{i0}+s_{i1}+\cdots+s_{ik})^{-n_i} e^{-b(s_{i0}+s_{i1}+\cdots+s_{ik})} s_{i0}^{\alpha T_0-1} \prod_{j=1}^{k} s_{ij}^{\alpha T_j+n_{ij}-1}.$$

Applying the change of variables $\beta_i = S_{i0}+\cdots+S_{ik}$ and $W_{ij} = S_{ij}/\beta_i$, for $j = 0,\ldots,k$, the joint density of $\beta_i \geq 0$ and $\boldsymbol{W}_i = (W_{i0},\ldots,W_{ik}) \in \Delta^k$, where Δ^k is the k-dimensional unit simplex, is

$$f_{\beta_i,\boldsymbol{W}_i|\boldsymbol{T}}(z_i,\boldsymbol{w}_i) \propto z_i^{\alpha(T_0+T_1+\cdots+T_k)-1} e^{-bz_i}\, w_{i0}^{\alpha T_0-1} \prod_{j=1}^{k} w_{ij}^{\alpha T_j+n_{ij}-1}.$$

Therefore $\beta_i = S_{i0}+\cdots+S_{ik}$ is independent from $\boldsymbol{W}_i$ and has gamma distribution with shape $\alpha(T_0+\cdots+T_k)$ and rate b. Moreover, the quantity $C(\boldsymbol{m};\boldsymbol{t})$ in (9) is given by

$$\begin{aligned} C(\boldsymbol{m};\boldsymbol{t}) &= \int_{(0,+\infty)^{k+1}} (s_0+\cdots+s_k)^{-m_\bullet} f_{\mathrm{ID}(t_0\rho)}(s_0) \prod_{j=1}^{k} s_j^{m_j} f_{\mathrm{ID}(t_j\rho)}(s_j)\mathrm{d}\boldsymbol{s} \\ &= \frac{b^{\alpha\sum_{j=0}^k t_j}}{\Gamma(\alpha t_0)\prod_{j=1}^k \Gamma(\alpha t_j)} \int_{(0,+\infty)} z^{\alpha\sum_{j=0}^k t_j-1} e^{-bz}\mathrm{d}z \int_{\Delta^k} w_0^{\alpha t_0-1} \prod_{j=1}^{k} w_j^{\alpha t_j+m_j-1}\mathrm{d}\boldsymbol{w} \\ &= \frac{\Gamma(\alpha(t_0+\cdots+t_k))}{\Gamma(\alpha(t_0+\cdots+t_k)+m_\bullet)} \prod_{j=1}^{k} \frac{\Gamma(\alpha t_j+m_j)}{\Gamma(\alpha t_j)}. \end{aligned}$$

(b) From the specification of ρ_0 in Example 2.7, the random variable $\mathrm{ID}(\rho_0)$ has gamma distribution with shape parameter α_0 and rate parameter b_0. Therefore, the density of $\boldsymbol{T} = (T_0,\ldots,T_k)$ is proportional to

$$\begin{aligned} f_{\boldsymbol{T}}(\boldsymbol{t}) &\propto \prod_{i=1}^{d} C(n_{i1},\ldots,n_{ik};\boldsymbol{t}) f_{\mathrm{ID}(\rho_0)}(t_0) \prod_{j=1}^{k} \rho_0(t_j) \\ &\propto \prod_{i=1}^{d} \left(\frac{\Gamma(\alpha(t_0+\cdots+t_k))}{\Gamma(\alpha(t_0+\cdots+t_k)+n_i)} \prod_{j=1}^{k} \frac{\Gamma(\alpha t_j+n_{ij})}{\Gamma(\alpha t_j)} \right) t_0^{\alpha_0-1}\, e^{-b_0 t_0} \prod_{j=1}^{k} t_j^{-1}\, e^{-b_0 t_j} \\ &\propto \prod_{i=1}^{d} \left(\frac{1}{((\alpha(t_0+\cdots+t_k)))_{n_i}} \right) t_0^{\alpha_0-1}\, e^{-b_0 t_0} \prod_{j=1}^{k} t_j^{-1}\, e^{-b_0 t_j} \left(\prod_{i=1}^{d} ((\alpha t_j))_{n_{ij}} \right), \end{aligned}$$

where $((\alpha t))_n = \Gamma(\alpha t+n)/\Gamma(\alpha t)$ is the ascending factorial. Since we only need the distribution of $\alpha T = \alpha(T_0+\cdots+T_k)$ to sample the random variables $\beta_1,\ldots,\beta_d$, we

could apply the change of variables

$$\alpha T = \alpha(T_0 + \cdots + T_k), \qquad V_j = T_j/T, \quad j = 0, \ldots, k,$$

and obtain the joint density of αT and the vector $\boldsymbol{V} = (V_0, \ldots, V_k) \in \Delta^k$ of auxiliary latent variables, supported on the k-dimensional unit simplex Δ^k:

$$f_{\alpha T, \boldsymbol{V}}(t, \boldsymbol{v}) \propto t^{\alpha_0 - 1} e^{-(b_0/\alpha)\, t} \prod_{i=1}^{d} \frac{1}{((t))_{n_i}} \, v_0^{\alpha_0 - 1} \prod_{j=1}^{k} \left(v_j^{-1} \prod_{i=1}^{d} ((t v_j))_{n_{ij}} \right).$$

Therefore, the marginal density of αT is

$$f_{\alpha T}(t) \propto t^{\alpha_0 - 1} e^{-(b_0/\alpha)\, t} \prod_{i=1}^{d} \frac{1}{((t))_{n_i}} \int_{\Delta^k} v_0^{\alpha_0 - 1} \prod_{j=1}^{k} \left(v_j^{-1} \prod_{i=1}^{d} ((t v_j))_{n_{ij}} \right) \mathrm{d}\boldsymbol{v}.$$

Note that this function is integrable in t; indeed, it holds that $((s))_q \sim s$ for $s \to 0$ when q is a positive integer. Therefore, for $t \to 0$,

$$\begin{aligned} f_{\alpha T}(t) &\sim t^{\alpha_0 - 1} \prod_{i=1}^{d} \frac{1}{((t))_{n_i}} \int_{\Delta^k} v_0^{\alpha_0 - 1} \prod_{j=1}^{k} \left(v_j^{-1} \prod_{i=1}^{d} ((t v_j))_{n_{ij}} \right) \mathrm{d}\boldsymbol{v} \\ &\sim t^{\alpha_0 - 1} \prod_{i=1}^{d} \left(t^{-1} \prod_{j=1}^{k} t^{m_{ij}} \right) \sim t^{\alpha_0 - 1} \prod_{i=1}^{d} t^{m_{i\bullet} - 1} \sim t^{\alpha_0 + m - d - 1}, \end{aligned}$$

and since we are assuming $n_i > 0$, then $m_{i\bullet} \geq 1$ and therefore $m \geq d$. See also the proof of Proposition 5.4.

S2.15 Proof of Proposition 5.3

Recall that the ascending factorial $((s))_q$, for integer q, can be written as

$$((s))_q = \sum_{h=0}^{q} S(q, h)\, s^h, \tag{S9}$$

where $S(q, h)$ are the unsigned Stirling numbers of the first kind, defined through the recursive relation $S(q+1, h) = q\, S(q, h) + S(q, h-1)$, with initial conditions $S(0, 0) = 1$ and $S(q, 0) = S(0, h) = 0$ for $q > 0$ or $h > 0$. Since $S(q, 0) = 0$ whenever $q > 0$, the summation above can start from 1 if q is strictly positive.

From Proposition 5.1, the density of αJ_{0j}, for each $j = 1, \ldots, k$, can be rewritten as

$$f_{\alpha J_{0j} | \boldsymbol{U}}(t) \propto t^{-1}\, e^{-\lambda(\boldsymbol{U})\, t} \prod_{i=1}^{d} \left(\sum_{h_{ij} = m_{ij}}^{n_{ij}} S(n_{ij}, h_{ij})\, t^{h_{ij}} \right)$$

$$\propto t^{-1}\, e^{-\lambda(\boldsymbol{U})\, t} \sum_{h_j = m_{\bullet j}}^{n_{\bullet j}} \sum_{\substack{h_{1j}+\cdots+h_{dj}=h_j \\ m_{ij} \le h_{ij} \le n_{ij}}} \prod_{i=1}^{d} S(n_{ij}, h_{ij})\, t^{h_j}$$

$$\propto \sum_{h_j = m_{\bullet j}}^{n_{\bullet j}} \left(\sum_{\substack{h_{1j}+\cdots+h_{dj}=h_j \\ m_{ij} \le h_{ij} \le n_{ij}}} \prod_{i=1}^{d} S(n_{ij}, h_{ij}) \right) t^{h_j - 1} e^{-\lambda(\boldsymbol{U})\, t}$$

where $m_{ij} \in \{0, 1\}$ is the indicator for $n_{ij} > 0$, that is, $m_{ij} = \min(1, n_{ij})$, and $m_{\bullet j} = \sum_{i=1}^{d} m_{ij}$. Note that since $n_{\bullet j} > 0$, then $m_{\bullet j} > 0$, and the density of αJ_{0j} is properly defined. Moreover, for each $j = 1, \ldots, k$, define the coefficients $S(n_{1j}, \ldots, n_{dj}; h_j)$ where

$$S(q_1, \ldots, q_d; h) = \sum_{\substack{h_1+\cdots+h_d=h \\ 0 \le h_i \le q_i}} \prod_{i=1}^{d} S(q_i, h_i). \tag{S10}$$

Exploiting the recursive relation for unsigned Stirling numbers of the first kind, we obtain

$$\begin{aligned}
&S(q_1, \ldots, q_\ell + 1, \ldots, q_d; h) = \\
&\quad = \sum_{\substack{h_1+\cdots+h_d=h \\ 0 \le h_i \le q_i \\ 0 \le h_\ell \le q_\ell + 1}} S(q_\ell + 1, h_\ell) \prod_{\substack{i=1 \\ i \ne \ell}}^{d} S(q_i, h_i) \\
&\quad = \sum_{\substack{h_1+\cdots+h_d=h \\ 0 \le h_i \le q_i \\ 0 \le h_\ell \le q_\ell + 1}} q_\ell S(q_\ell, h_\ell) \prod_{\substack{i=1 \\ i \ne \ell}}^{d} S(q_i, h_i) + \sum_{\substack{h_1+\cdots+h_d=h \\ 0 \le h_i \le q_i \\ 0 \le h_\ell \le q_\ell + 1}} S(q_\ell, h_\ell - 1) \prod_{\substack{i=1 \\ i \ne \ell}}^{d} S(q_i, h_i) \\
&\quad = q_\ell \sum_{\substack{h_1+\cdots+h_d=h \\ 0 \le h_i \le q_i \\ 0 \le h_\ell \le q_\ell}} S(q_\ell, h_\ell) \prod_{\substack{i=1 \\ i \ne \ell}}^{d} S(q_i, h_i) + \sum_{\substack{h_1+\cdots+h_d=h-1 \\ 0 \le h_i \le q_i \\ 0 \le h_\ell \le q_\ell}} S(q_\ell, h_\ell) \prod_{\substack{i=1 \\ i \ne \ell}}^{d} S(q_i, h_i) \\
&\quad = q_\ell\, S(q_1, \ldots, q_d; h) + S(q_1, \ldots, q_d; h - 1).
\end{aligned}$$

Notably, in the third line we have used the fact that $S(q_\ell, q_\ell + 1) = 0$ (first term) and applied the change of index $h_\ell \mapsto h_\ell + 1$ (second term). Therefore, the coefficients above satisfy the recurrence relation in (12) defining the multivariate Stirling numbers.

Remark S2.3. A naive approach to the evaluation of the coefficients $S(q_1, \ldots, q_d; h)$ from the relation (S10) would involve d nested cycles, with a computational cost of $\mathcal{O}\big(\prod_{i=1}^{d} n_{ij}\big)$, for each $j = 1, \ldots, k$. However, recognizing the recursive structure (12) substantially reduces the cost to quadratic in the number of observations $n_{\bullet j}$.

S2.16 Proof of Proposition 5.4

The density of αT in Proposition 5.2 can be rewritten using the equality in (S9) and the definition of multivariate Stirling numbers in (S10):

$$
\begin{aligned}
f_{\alpha T}(t) &\propto \\
&\propto t^{\alpha_0-1}e^{-(b_0/\alpha)\,t}\prod_{i=1}^{d}\frac{1}{((t))_{n_i}}\int_{\Delta^k} v_0^{\alpha_0-1}\prod_{j=1}^{k}\left(v_j^{-1}\prod_{i=1}^{d}\left(\sum_{h_{ij}=m_{ij}}^{n_{ij}} S(n_{ij},h_{ij})\,v_j^{h_{ij}}t^{h_{ij}}\right)\right)\mathrm{d}\boldsymbol{v} \\
&\propto t^{\alpha_0-1}e^{-(b_0/\alpha)\,t}\prod_{i=1}^{d}\frac{1}{((t))_{n_i}}\int_{\Delta^k} v_0^{\alpha_0-1}\prod_{j=1}^{k}\left(\sum_{h_j=m_{\bullet j}}^{n_{\bullet j}} S(\boldsymbol{n}_j;h_j)\,v_j^{h_j-1}t^{h_j}\right)\mathrm{d}\boldsymbol{v} \\
&\propto t^{\alpha_0-1}e^{-(b_0/\alpha)\,t}\prod_{i=1}^{d}\frac{1}{((t))_{n_i}}\int_{\Delta^k} v_0^{\alpha_0-1}\sum_{h=m}^{n}\left(\sum_{\substack{h_1+\cdots+h_k=h\\ m_{\bullet j}\le h_j\le n_{\bullet j}}}\prod_{j=1}^{k} S(\boldsymbol{n}_j;h_j)\,v_j^{h_j-1}t^{h_j}\right)\mathrm{d}\boldsymbol{v},
\end{aligned}
$$

where $\boldsymbol{n}_j=(n_{1j},\ldots,n_{dj})$. The density of αT is then obtained computing the integral over the k-dimensional simplex:

$$
\begin{aligned}
f_{\alpha T}(t) &\propto \\
&\propto t^{\alpha_0-1}e^{-(b_0/\alpha)\,t}\prod_{i=1}^{d}\frac{1}{((t))_{n_i}}\left(\sum_{h=m}^{n} t^h\sum_{\substack{h_1+\cdots+h_k=h\\ m_{\bullet j}\le h_j\le n_{\bullet j}}}\prod_{j=1}^{k} S(\boldsymbol{n}_j;h_j)\int_{\Delta^k} v_0^{\alpha_0-1}\prod_{j=1}^{k} v_j^{h_j-1}\,\mathrm{d}\boldsymbol{v}\right) \\
&\propto t^{\alpha_0-1}e^{-(b_0/\alpha)\,t}\prod_{i=1}^{d}\frac{1}{((t))_{n_i}}\left(\sum_{h=m}^{n}\frac{t^h}{((\alpha_0))_h}\sum_{\substack{h_1+\cdots+h_k=h\\ m_{\bullet j}\le h_j\le n_{\bullet j}}}\prod_{j=1}^{k}\Gamma(h_j)\,S(\boldsymbol{n}_j;h_j)\right) \\
&\propto t^{\alpha_0-1}e^{-(b_0/\alpha)\,t}\prod_{i=1}^{d}\frac{1}{((t))_{n_i}}\left(\sum_{h=m}^{n}\frac{c_h}{((\alpha_0))_h}\,t^h\right),
\end{aligned}
$$

where the coefficients c_h for $h=m,\ldots,n$ are defined as

$$
c_h=\sum_{\substack{h_1+\cdots+h_k=h\\ m_{\bullet j}\le h_j\le n_{\bullet j}}}\prod_{j=1}^{k}\Gamma(h_j)\,S(n_{1j},\ldots,n_{dj};h_j).
$$

S3 Posterior sampling for the gamma-gamma hCRV

This section provides additional details on the practical implementation of posterior sampling algorithms for the normalized gamma-gamma hCRV described in Section 5, and contains numerical illustrations supporting their effectiveness. Section S3.1 analyzes the Metropolis-Hastings steps of Algorithm 1. Specifically, a random walk Metropolis-Hastings scheme on the log-scale with Gaussian increments is proposed and compared with the

approach of Barrios et al. (2013), based on gamma proposals. The initialization of Algorithm 2 is detailed in Sections S3.2 and S3.3, which respectively discuss the computation of coefficients c_h's in Proposition 5.4 and the optimization of parameter r within the rejection sampling scheme. Section S3.4 describes the inverse Lévy measure algorithm to obtain a truncated sample from the hierarchy of gamma CRMs in Proposition 5.1(a). For this purpose, an efficient procedure to sequentially invert the exponential integral function is outlined in Section S3.5. This same simulation strategy remains valid *a priori* to sample from gamma-gamma hCRVs. Section S3.6 characterizes the posterior random probabilities $\tilde{p}_i = \tilde{\mu}_i/\tilde{\mu}_i(\mathbb{X}) \mid \boldsymbol{X}_{1:d}$ as conditionally normalized gamma CRMs, i.e. conditionally Dirichlet processes, and thus provides an alternative approach to directly sample their posterior normalized jumps. This allows for a straightforward comparison with standard samplers for the HDP with gamma prior on the concentration parameter, which target the same posterior distributions (Proposition 3.3). The marginal Gibbs sampler of Teh et al. (2006), based on the restaurant franchise metaphor, is tailored to our setting in Section S3.7. This section also proposes an alternative collapsed Gibbs sampler for the HDP, which directly samples the number of tables serving each dish, and may be of independent interest. Numerical illustrations of the effectiveness of the proposed algorithms in terms of mixing properties and posterior accuracy are provided in Section S3.8. Finally, Section S3.9 extends the simulation study of Section 5.4 comparing the different algorithms in terms of execution time per effective sample, as the number k of distinct values increases.

S3.1 Metropolis-Hastings steps in Algorithm 1

The non-standard steps in the posterior sampling algorithms of Section 5.3 are:

(i) the marginal sampling of random variable αT in (11), whose joint density with the auxiliary vector $\boldsymbol{V}$ is known up to a normalizing constant;

(ii) the sampling of random variables $\alpha J_{01}, \ldots, \alpha J_{0k}$ in (10), whose densities are again known up to normalizing constants.

A natural approach to obtain samples from non-standard densities is resorting to MCMC schemes. In Algorithm 1, we consider a blocked Gibbs sampler with Metropolis-Hastings steps. The full conditional distributions are derived from Propositions 5.1 and 5.2, namely

$$(V_j, V_\ell) \mid \boldsymbol{V}^{-j,\ell}, (\alpha T) = t \;\sim\; v_j^{-1}\, v_\ell^{-1} \prod_{i=1}^{d} ((tv_j))_{n_{ij}} ((tv_\ell))_{n_{i\ell}} \qquad (j, \ell = 1, \ldots, k),$$

$$(V_j, V_0) \mid \boldsymbol{V}^{-j,0}, (\alpha T) = t \;\sim\; v_0^{\alpha_0 - 1}\, v_j^{-1} \prod_{i=1}^{d} ((tv_j))_{n_{ij}} \qquad (j = 1, \ldots, k),$$

$$(\alpha T) \mid \boldsymbol{V} \;\sim\; t^{\alpha_0 - 1}\, e^{-(b_0/\alpha)\, t} \prod_{i=1}^{d} \frac{1}{((t))_{n_i}} \prod_{j=1}^{k} \prod_{i=1}^{d} ((tv_j))_{n_{ij}},$$

$$(\alpha J_{0j}) \mid \boldsymbol{U} \;\sim\; t^{-1}\, e^{-\lambda(\boldsymbol{U})\, t} \prod_{i=1}^{d} ((t))_{n_{ij}} \qquad (j = 1, \ldots, k),$$

where $\boldsymbol{V}^{-j,\ell}$ denotes the vector $\boldsymbol{V} \in \Delta^k$ with components V_j and V_ℓ removed. We note that variables in $\boldsymbol{V} \in \Delta^k$ are subject to the constraint $\sum_{j=0}^{k} V_j = 1$, and one cannot update variables independently; on the other hand, variables αT and αJ_{0j} can take every positive value. Throughout the Gibbs sampling procedure, we propose new values for each variable, and accept or reject the proposal according to the Metropolis-Hastings ratio. The rest of the section outlines some of the possible proposals, focusing on rather straightforward and widely known options. Clearly, more sophisticated alternatives can be considered without undermining the validity of our results.

A simple and effective symmetric proposal for the pair of auxiliary variables (V_j, V_ℓ), for each $j, \ell = 0, \ldots, k$, is

$$v_j^* = \varepsilon\,(v_j + v_\ell), \qquad v_\ell^* = (1 - \varepsilon)\,(v_j + v_\ell),$$

where v_j and v_ℓ are the current values of V_j and V_ℓ and $\varepsilon \sim \mathcal{U}(0, 1)$; see e.g. Smith (2014). The proposal is accepted with log-probability

$$\log(r) = \min\left\{0,\; q_j(v_j^*) - q_j(v_j) + q_\ell(v_\ell^*) - q_\ell(v_\ell)\right\},$$

where $q_0(v) = (\alpha_0 - 1)\, v$, and $q_j(v) = \sum_{i=1}^{d} \log((tv))_{n_{ij}} - v$, for $j = 1, \ldots, k$. At each iteration of the Gibbs sampler, we randomly select k pairs of indexes (j, ℓ) and implement the Metropolis-Hastings step detailed above. Note that selecting k pairs is a good compromise between choosing a single pair and scanning each of the $k(k-1)/2$ possible pairs.

For the positive variables αT and αJ_{0j}'s, we consider random walks with two alternative proposals, which we first describe for a generic positive random variable with density $f(x)$.

(a) *Gamma proposal.* Following the approach discussed in Barrios et al. (2013), we propose a new value x^* from a gamma distribution centered at the current value x,

$$x^* \sim \text{Gamma}(\delta, \delta/x),$$

where $\delta > 0$ controls the variance of the proposal. The proposed value x^* is accepted with log-probability

$$\begin{aligned}\log(r) &= \min\left\{0,\; \log f(x^*) - \log f(x) + \log g(x; \delta, \delta/x^*) - \log g(x^*; \delta, \delta/x)\right\} \\ &= \min\left\{0,\; \log f(x^*) - \log f(x) + (2\delta - 1)(\log x - \log x^*) + \delta\left(\frac{x^*}{x} - \frac{x}{x^*}\right)\right\},\end{aligned}$$

where $x \mapsto \log g(x; a, b)$ is the log-density of a gamma distribution with shape $a > 0$ and rate $b > 0$. For the practical implementation, Barrios et al. (2013) suggest to restrict to $\delta \geq 1$.

(b) *Random walk on log-transform.* For a positive random variable, we can target the density of its log-transform, which is $t \mapsto f(e^t)\, e^t$. In this case, we resort to a random walk on the log-scale, and propose a new value $\log x^*$ from a normal distribution centered at the current value $\log x$,

$$\log x^* \sim \mathcal{N}(\log x, \sigma^2),$$

where $\sigma^2 > 0$ controls the variance of the proposal. The proposed value x^* is accepted with log-probability

$$\log(r) = \min\big\{0,\, \log f(x^*) - \log f(x) + \log x^* - \log x\big\}.$$

Interestingly, both proposals can be interpreted as multiplicative perturbations of the current value, with different distributions for the perturbation factor. Indeed, they both propose a new value $x^* = x\,\varepsilon$, where x is the current value and ε is distributed as either (a) a gamma random variable $\varepsilon \sim \text{Gamma}(\delta, \delta)$ with variance $1/\delta$, or (b) a log-normal random variable $\log \varepsilon \sim \mathcal{N}(0, \sigma^2)$ with variance σ^2 on the log-scale. Therefore, the acceptance rates can be rewritten as

$$\begin{aligned}
&\text{(a)} \qquad \log(r) = \min\left\{0,\, \log g(x\,\varepsilon) - \log g(x) + \delta(\varepsilon - 1/\varepsilon) - 2\delta\, \log \varepsilon\right\}, \\
&\text{(b)} \qquad \log(r) = \min\left\{0,\, \log g(x\,\varepsilon) - \log g(x)\right\},
\end{aligned}$$

where $\log g(x) = \log f(x) + \log x$. The practical implementation of these approaches requires the evaluation of the logarithm of target density at the current and proposed values, up to normalizing constants. For our purposes, we have

$$\log f_{\alpha T|\boldsymbol{V}}(t) = (\alpha_0 - 1)\log t - (b_0/\alpha)\, t + \sum_{j=1}^{k}\sum_{i=1}^{d} \log((tv_j))_{n_{ij}} - \sum_{i=1}^{d} \log((t))_{n_i},$$

$$\log f_{\alpha J_{0j}|\boldsymbol{U}}(t) = \sum_{i=1}^{d} \log((t))_{n_{ij}} - \lambda(\boldsymbol{U})\, t - \log t \qquad (j = 1, \ldots, k).$$

Note that the random variables αJ_{0j}'s are gamma distributed whenever $n_{1j}, \ldots, n_{dj} \leq 1$, and can thus be sampled exactly as

$$\alpha J_{0j} \mid \boldsymbol{U} \sim \text{Gamma}\left(\sum_{i=1}^{d} n_{ij},\, \lambda(\boldsymbol{U})\right). \tag{S11}$$

Integrating this feature into the algorithm leads to a substantial reduction in computational times in many scenarios. A crucial aspect of random walk Metropolis-Hastings algorithms is the choice of a suitable variance parameter for the proposal, in order for the resulting Markov chain to mix efficiently (Roberts and Rosenthal, 2001). For this purpose, we resort to adaptive MCMC techniques, which allow to automatically tune the proposals to optimize their performances. Specifically, we implement a simple but effective updating rule, based on the Robbins-Monro recursion (see e.g. Andrieu and Thoms, 2008). The explicit updates for the two proposals discussed above are

$$\text{(a)} \quad \log \delta_{s+1} \leftarrow \log \delta_s - \gamma_s(r_s - r^*), \qquad \text{(b)} \quad \log \sigma^2_{s+1} \leftarrow \log \sigma^2_s + \gamma_s(r_s - r^*),$$

where δ_s and σ^2_s are the parameters of the proposals at step s, r_s is the probability of acceptance at step s, $r^* = 0.44$ is the target acceptance rate (see Gelman et al., 1996) and $(\gamma_s)_{s\geq 1}$ is a sequence of positive and non-increasing numbers. In our implementation, we employ the deterministic sequence $\gamma_s = (10 + s)^{-1/2}$, for $s \geq 1$; see Andrieu and Thoms (2008) for a discussion on the desirable properties of such sequence. This adaptive procedure is employed at each iteration during the burn-in phase, and the values of variance parameters reached at the end of the burn-in are kept fixed for the actual posterior sampling. Our implementation allows for different variance parameters for αT and each of the αJ_{0j}'s.

We conclude this section with a comparison of the gamma and log-normal proposals, carried out through a simple numerical experiment. We consider $1,000$ synthetic datasets, each consisting of a random number of groups $d \sim \text{Poisson}(5)$, with $n_i = 100$ observations per group. Observations are sampled from a hierarchical Dirichlet process with random concentration parameters $\alpha \sim \text{Gamma}(5)$ and $\alpha_0 \sim \text{Gamma}(3)$. We draw $10,000$ posterior samples for each dataset, with a burn-in of $1,000$ steps, during which variance parameters are tuned. Table S1 reports the effective sample size (ESS) and acceptance rate for each target variable and proposal. Figures for the αJ_{0j}'s are the means computed over j, excluding those indexes for which the exact sampling in (S11) is available. The log-normal proposal outperforms the gamma proposal in terms of ESS for both αT and the αJ_{0j}'s; the difference is particularly evident for the latter. Moreover, the adaptive MCMC procedure systematically reaches the target acceptance rate of 0.44. Clearly, the evidence provided by this experiment is not conclusive in any respect, as we are restricting ourselves to a particular algorithm and sampling scenario. However, the synthetic datasets cover a wide range of possible count matrices $(n_{ij})_{ij}$ that one may encounter in applications. Therefore, we recommend using the log-normal proposal, which we also consider for the simulations in Section 5.4 of the main manuscript.

variable(s)	proposal	ESS	accept. rate
latent αT	gamma	992 (432)	0.440 (0.022)
	log-normal	1114 (479)	0.440 (0.021)
jumps αJ_{0j}'s	gamma	966 (225)	0.440 (0.009)
	log-normal	1779 (232)	0.440 (0.008)

Table S1: Effective sample size (ESS) and acceptance rate for each target variable and proposal. Results are averaged over 1,000 synthetic datasets; standard deviations are reported in parentheses. After 1,000 burn-in steps, 10,000 posterior samples are drawn for each dataset. The target acceptance rate for the adaptive MCMC scheme is 0.44 (see Gelman et al., 1996).

S3.2 Computing the coefficients in Proposition 5.4

In this section, we show that the coefficients c_h, for $h = m, \dots, n$, defined in Proposition 5.4, can be computed through a sequence of discrete convolutions. For each $\ell = 1, \dots, k$, define the vector $\boldsymbol{c}(\ell) = (c(\ell, h))_h$ with

$$c(\ell; h) = \sum_{\substack{h_1+\cdots+h_\ell=h \\ m_{\bullet j} \le h_j \le n_{\bullet j}}} \prod_{j=1}^{\ell} a(j; h_j), \qquad h = \sum_{j=1}^{\ell} m_{\bullet j}, \dots, \sum_{j=1}^{\ell} n_{\bullet j},$$

where $a(j; h) = \Gamma(h)\, S(n_{1j}, \dots, n_{dj}; h)$, for $h = m_{\bullet j}, \dots, n_{\bullet j}$ and $j = 1, \dots, \ell$. Thus, the c_h's are such that $c_h = c(k; h)$, that is, coincide with $\boldsymbol{c}(k)$. The entries of vector $\boldsymbol{c}(\ell)$ can be computed from vector $\boldsymbol{c}(\ell - 1)$ as follows:

$$\begin{aligned} c(\ell; h) &= \sum_{\substack{h_1+\cdots+h_\ell=h \\ m_{\bullet j} \le h_j \le n_{\bullet j}}} \prod_{j=1}^{\ell} a(j; h_j) = \sum_{h_\ell = m_{\bullet \ell}}^{n_{\bullet \ell}} \left(\sum_{\substack{h_1+\cdots+h_{\ell-1}=h-h_\ell \\ m_{\bullet j} \le h_j \le n_{\bullet j}}} \prod_{j=1}^{\ell-1} a(j; h_j) \right) a(\ell; h_\ell) \\ &= \sum_{h_\ell = m_{\bullet \ell}}^{n_{\bullet \ell}} c(\ell - 1; h - h_\ell)\, a(\ell; h_\ell). \end{aligned}$$

In other words, the vector $\boldsymbol{c}(\ell)$ is obtained by convolution between vector $\boldsymbol{c}(\ell - 1)$ and vector $\boldsymbol{a}(\ell) = (a(\ell; h), h)$. The computational cost for this operation is $\mathcal{O}\left(n_{\bullet \ell} \sum_{j=1}^{\ell-1} n_{\bullet j}\right)$. Therefore, the coefficients $(c_h)_h = \boldsymbol{c}(k)$ can be computed through the recursive relation

$$\boldsymbol{c}(0) = (1), \qquad \boldsymbol{c}(\ell) = \boldsymbol{c}(\ell - 1) * \boldsymbol{a}(\ell) \quad (\ell = 1, \dots, k),$$

or, equivalently, $\boldsymbol{c}(k) = \boldsymbol{a}(1) * \cdots * \boldsymbol{a}(k)$, where $*$ is the convolution operator. The total computational cost is $\mathcal{O}\left(\sum_{j<\ell} n_{\bullet j} n_{\bullet \ell}\right)$.

S3.3 Optimal choice of parameter r

Section 5.2 outlines a rejection sampling algorithm for sampling the latent variable αT from its density (13). Specifically, values are proposed from Gamma$(\alpha_0 + r, b_0/\alpha)$ and accepted with probability proportional to $t^{-r}R(t)$, where r is a real parameter and

$$R(t) = \prod_{i=1}^{d} \frac{1}{((t))_{n_i}} \left(\sum_{h=m}^{n} \frac{c_h}{((\alpha_0))_h} t^h \right).$$

A necessary condition for the rejection sampling scheme is $t^{-r}R(t)$ to be bounded above for $t \geq 0$. Since $R(t)$ is a ratio of polynomials, both having degree n and non-negative coefficients, it is a continuous function for $t > 0$. Moreover, when q is a positive integer, $((s))_q \sim s$ for $s \to 0$ while $((s))_q \sim s^q$ for $s \to \infty$. Hence, $R(t) \sim t^{m-d}$ for $t \to 0$ and $R(t) \sim c_n/((\alpha_0))_n$ for $t \to \infty$, which implies that $t^{-r}R(t)$ is continuous and bounded for $t \geq 0$ when $0 \leq r \leq m - d$. Our goal is choosing the value of r within this interval such that the acceptance probability is maximized.

For this purpose, let $t^*(r)$ be the value of t that maximizes $t^{-r}R(t)$ for $t \geq 0$. The overall acceptance probability is

$$\begin{aligned}
&\mathbb{E}_T \left[\left(\frac{t^*(r)}{t} \right)^r \frac{R(t)}{R(t^*(r))} \right] = \\
&= \frac{t^*(r)^r}{R(t^*(r))} \int_0^\infty t^{-r} R(t) \left(\frac{b_0}{\alpha} \right)^{\alpha_0 + r} \Gamma(\alpha_0 + r)^{-1} t^{\alpha_0 + r - 1} e^{-(b_0/\alpha)\, t} \mathrm{d}t \\
&= \frac{t^*(r)^r}{R(t^*(r))} \left(\frac{b_0}{\alpha} \right)^{\alpha_0 + r} \Gamma(\alpha_0 + r)^{-1} \int_0^\infty t^{\alpha_0 - 1} e^{-(b_0/\alpha)\, t} R(t)\, \mathrm{d}t.
\end{aligned}$$

Finding the value of $0 \leq r \leq m - d$ that maximizes the quantity above is equivalent to finding the value maximizing its logarithm, discarding terms not depending on r, that is

$$r^* = \arg\max_r \left\{ r \log t^*(r) - \log R(t^*(r)) + (\alpha_0 + r) \log(b_0/\alpha) - \log \Gamma(\alpha_0 + r) \right\}.$$

This maximization problem can be further simplified by restricting to a finite set of potentially maximizing values. Indeed, for $0 < r < m - d$, the function $R(t)$ is continuous and differentiable for $t > 0$, and such that $t^{-r}R(t) \to 0$ for both $t \to 0$ and $t \to \infty$. Hence, $t^*(r)$ is a stationary point for $t^{-r}R(t)$, which implies $R'(t^*(r))\, t^*(r) = r\, R(t^*(r))$. Moreover, by the implicit function theorem, $t^*(r)$ is a continuous and differentiable function in r. Therefore, the objective function is continuous and differentiable for $0 < r < m - d$, and the set of potentially maximizing points in $(0, m - d)$ may be restricted to the stationary points (if any), satisfying

$$\log t^*(r) + \log(b_0/\alpha) - \Psi(\alpha_0 + r) = 0,$$

where Ψ denotes here the digamma function $\Psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \log \Gamma(x)$, and we have used that $t^*(r)$ is a stationary point for $t^{-r} R(t)$ and thus $R'(t^*(r))\, t^*(r) = r\, R(t^*(r))$. Remarkably, this stationarity condition is also satisfied at the boundary of the maximization set. Indeed, for $r = 0$, the optimal $t^*(r)$ may be either a stationary point or $+\infty$, for which $R'(t^*(r)) = 0$ necessarily holds. On the other hand, if $r = m - d$, the optimal $t^*(r)$ may be either a stationary point or 0, for which $R(t^*(r)) = 0$ holds.

S3.4 Sampling from the hierarchy of gamma CRMs

From Proposition 5.1(a), the residual component $\tilde{\boldsymbol{\mu}}^*$ of the posterior distribution of $\tilde{\boldsymbol{\mu}}$ retains a hierarchical structure, conditionally on latent variables $\boldsymbol{U}$:

$$\tilde{\mu}_1^*, \ldots, \tilde{\mu}_d^* \mid \tilde{\mu}_0^*, \boldsymbol{U} \sim \prod_{i=1}^{d} \mathrm{CRM}\left(\alpha\, s^{-1}\, e^{-b(1+U_i/b)\, s} \mathrm{d}s \otimes \tilde{\mu}_0^*\right),$$
$$\tilde{\mu}_0^* \mid \boldsymbol{U} \sim \mathrm{CRM}\left(\alpha_0\, s^{-1}\, e^{-\alpha\, \lambda(\boldsymbol{U})\, s} \mathrm{d}s \otimes P_0\right).$$

Therefore, one needs to sample from a hierarchy of gamma CRMs in order to obtain complete samples from the posterior.

At the root of the hierarchy, approximate posterior samples from the rescaled gamma random measure $\alpha\tilde{\mu}_0^* \mid \boldsymbol{U}$ can be obtained through the Ferguson-Klass representation (Ferguson and Klass, 1972). This amounts to sampling the largest L jumps of an infinitely active random measure in decreasing order, and thus provides its best finite-dimensional approximation: for a fixed truncation level, the approximation error is minimized. A straightforward approach uses the inverse Lévy measure algorithm (Wolpert and Ickstadt, 1998; Walker and Damien, 2000). For $\ell = 1, \ldots, L$, let $\omega_{0\ell} \geq 0$ be the value solving the equation

$$\frac{\xi_\ell}{\alpha_0} = \int_{\omega_{0\ell}}^{\infty} s^{-1}\, e^{-\alpha\, \lambda(\boldsymbol{U})\, s} \mathrm{d}s = E_1(\alpha\, \lambda(\boldsymbol{U})\, \omega_{0\ell}) \qquad (\ell = 1, \ldots, L),$$

where E_1 is the exponential integral function and $\xi_1 < \cdots < \xi_L$ a.s. are the first L jump times of a unit rate Poisson process, that is, $\xi_0 = 0$ and the inter-arrival times are $\xi_\ell - \xi_{\ell-1} \overset{\text{iid}}{\sim} \mathrm{Exp}(1)$, for $\ell = 1, \ldots, L$. Then the random measure $\alpha\tilde{\mu}_0^* \mid \boldsymbol{U}$ can be approximated as

$$\alpha\tilde{\mu}_0^* \mid \boldsymbol{U} \approx \sum_{\ell=1}^{L} (\alpha\omega_{0\ell})\, \delta_{\phi_\ell},$$

where $\phi_1, \ldots, \phi_L \overset{\text{iid}}{\sim} P_0$ are independent from the Poisson process $(\xi_1, \ldots, \xi_L)$. This algorithm requires to sequentially invert the exponential integral function numerically, which is a nontrivial albeit much investigated task. Details about our implementation are provided in the following Section S3.5. Alternative approaches for sampling the largest L jumps of a random measure in decreasing order, and their specifications for the

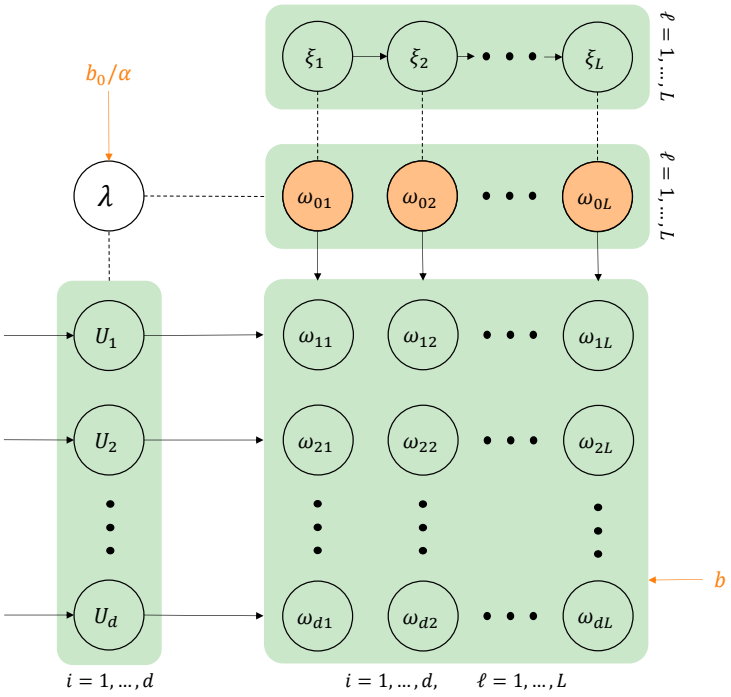

Figure S1: Conditional dependencies between random variables involved in sampling the hierarchy of gamma CRMs. Red circles represent the computational bottlenecks. The sampling scheme for the latent variables $\boldsymbol{U}$ is outlined in Section 5.1 and Figure 1. For simplicity, variables are reported up to scaling w.r.t. model parameters.

gamma process, are explored in Campbell et al. (2019); Zhang and Dassios (2025); see also Rosiński (2001).

The random measures in $\tilde{\boldsymbol{\mu}}^*$ are conditionally independent given $\alpha\tilde{\mu}_0^*$ and $\boldsymbol{U}$, and each component $\tilde{\mu}_i^* \mid \tilde{\alpha}\mu_0^*, \boldsymbol{U}$ can be approximately sampled from a gamma CRM having Lévy intensity

$$\mathrm{d}\nu_i^*(s,x) = \sum_{\ell=1}^{L} (\alpha\omega_{0\ell})\, s^{-1} e^{-b(1+U_i/b)\, s}\, \mathrm{d}s\, \mathrm{d}\delta_{\phi_\ell}(x).$$

The additive components of the Lévy measure represent independent summands for $\tilde{\mu}_i^*$ concentrated at different fixed location. Hence, $\tilde{\mu}_i^* \mid \alpha\tilde{\mu}_0^*, \boldsymbol{U}$ can be approximated as

$$\tilde{\mu}_i^* \mid \alpha\tilde{\mu}_0^*, \boldsymbol{U} \approx \sum_{\ell=1}^{L} \omega_{i\ell}\, \delta_{\phi_\ell},$$

where $\omega_{i1}, \dots \omega_{iL}$ are independent random variables with $\omega_{i\ell} \sim \mathrm{Gamma}(\alpha\omega_{0\ell}, b(1+U_i/b))$, for $\ell = 1, \dots, L$. The sampling procedure to obtain an approximation of $\tilde{\boldsymbol{\mu}}$ by truncation of its infinite sequence of jumps is summarized in Figure S1; the similarities with the sampling algorithms for jumps $\boldsymbol{J}$, depicted in Figure 1, are evident. Note that the total mass of each random measure $\tilde{\mu}_i^*$, that is, the mass of the posterior random measure $\tilde{\mu}_i \mid \boldsymbol{X}_{1:d}$ not assigned to fixed locations, can instead be sampled exactly from a hierarchy of gamma random variables.

This same sampling procedure can be also employed to sample *a priori* from the gamma-gamma hCRV, with minimal adjustments. Indeed, the prior construction introduced in Example 2.7 coincides with the hierarchy in Proposition 5.1(a) displayed above when $U_i = 0$ for each $i = 1, \dots, d$, and thus $\lambda(\boldsymbol{U}) = b_0/\alpha$. Specifically, approximate samples from the rescaled gamma random measure $\alpha\tilde{\mu}_0$ can be obtained via the inverse Lévy measure algorithm, while increments of each random measure $\tilde{\mu}_i \mid \alpha\tilde{\mu}_0$ can be sampled independently from gamma distributions, as discussed in Remark 2.5.

S3.5 Inverting the exponential integral

The implementation of the inverse Lévy measure algorithm of Walker and Damien (2000) for the gamma CRM requires to invert the tail integrals of its Lévy density (Section S3.4). This amounts to sequentially invert the exponential integral function E_1, defined for $x > 0$,

$$E_1(x) = \int_x^{+\infty} s^{-1} e^{-s} \mathrm{d}s.$$

Note that E_1 is a strictly decreasing function, and thus is invertible. A convenient approach to find its inverse $E_1^{-1}(y)$, for a given value $y > 0$, is to determine the unique root of the function $x \mapsto E_1(x) - y$, exploiting root-finding algorithms such as Newton's method. This method requires the evaluation of the derivative E_1' of the exponential integral, which can be computed in closed form, as it equals the opposite of the integrand function, $E_1'(x) = -x^{-1}e^{-x}$.

A well-known limitation of Newton's method is the possibility to obtain iteration values that fall outside the domain of the function, where its evaluation is not possible. This situation is particularly relevant to inverse Lévy measure algorithms, as the tail integrals of infinitely divisible Lévy densities diverge to $+\infty$ as the lower bound of the integration interval goes to 0. A simple workaround that typically solves this issue is choosing a starting point x_0 for Newton's algorithm which falls on the left of the solution, that is $E_1(x_0) \geq y$. This can be achieved by iteratively halving an initial guess. Remarkably, within the sequential approach required by the algorithm in Section S3.4, the standard choice for the initial guess is the solution at the previous step. However, this falls on the right of the current solution, and thus halving is always necessary. Convergence of this algorithm is guaranteed whenever the Lévy density is a decreasing function, and thus its tail integral is decreasing and convex. Figure S2 shows an illustration of the different behaviours of Newton's method, depending on the starting point x_0.

Instead, we consider an alternative approach that improves the efficiency of Newton's method. Specifically, we redefine the exponential integral through the logarithm of its argument; in other words, we invert the function

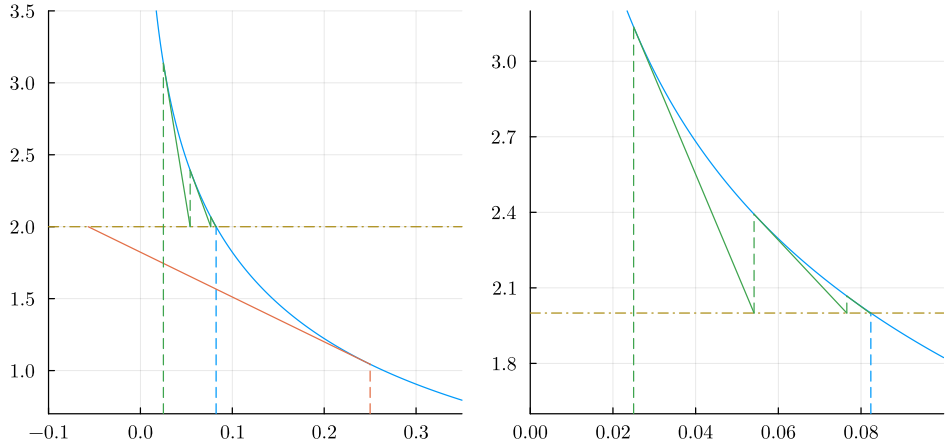

Figure S2: Iterations of Newton's method for solving the equation $E_1(x) = 2.0$, with different starting points. Starting on the right of the solution may lead the algorithm outside the domain of the function (orange); starting on the left guarantees convergence (green). The right panel zooms on converging iterations.

$$f(z) = E_1(e^z) = \int_{e^z}^{+\infty} s^{-1} e^{-s} \mathrm{d}s,$$

whose derivative is given by $f'(z) = E_1'(e^z)\, e^z = -\exp(-e^z)$. This formulation has the advantage of not being restricted to positive values, thus preventing iterations to fall outside the domain. Note that f is decreasing and convex, which guarantees convergence of Newton's method for every starting point x_0.

A decisive advantage of this approach is the asymptotic linearity of function f as z diverges to $-\infty$. Specifically, $f(z) \approx -\gamma - z$ for $z \to -\infty$, where γ is the Euler-Mascheroni constant. This property is crucial for speeding up convergence of the numerical scheme, as the rate of convergence of Newton's method is proportional to the second derivative around the solution. The asymptotic expansion is also useful to improve the numerical stability of function evaluations. We argue that redefining the tail integral as a function on the whole real line, removing the constraint to positive values, may be a general technique to enhance the performances of Newton's method for computing the inverse Lévy measure, beyond the gamma process case.

S3.6 Posterior distribution of normalized random measures

The posterior random probabilities arising from model (4) are the normalization of the posterior random measures characterized in Theorem 4.1 and specialized in Proposition 5.1 for the gamma-gamma hCRV. The simplest approach to obtain posterior samples is thus normalizing the posterior samples from the corresponding random measures. In this section, we show that such posterior random probabilities are distributed as conditionally Dirichlet processes with discrete base measures. Therefore, their probability weights can be alternatively sampled from a Dirichlet distribution. This allows for a straightforward

comparison with alternative samplers for the HDP, such as the marginal Gibbs samplers of Teh et al. (2006); see the following Section S3.8.

For this purpose, denote by $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^*$ the normalized jumps of $\tilde{\mu}_i \mid \boldsymbol{X}_{1:d}$, for $i = 1, \dots, d$,

$$\pi_{ij} = \frac{J_{ij}}{\sum_{j=1}^{k} J_{ij} + \tilde{\mu}_i^*(\mathbb{X})}, \qquad \pi_{i\ell}^* = \frac{\omega_{i\ell}}{\sum_{j=1}^{k} J_{ij} + \tilde{\mu}_i^*(\mathbb{X})},$$

where the J_{ij}'s are jumps at fixed locations and the $\omega_{i\ell}$'s are the jumps of $\tilde{\mu}_i^*$, arranged in decreasing order; see Section S3.4 for details on sampling from $\tilde{\mu}_i^*$. From Theorem 4.1 it follows that the random probabilities $\tilde{\boldsymbol{P}} = \tilde{\boldsymbol{\mu}}/\tilde{\boldsymbol{\mu}}(\mathbb{X})$ are distributed, a posteriori, as

$$\tilde{P}_i = \frac{\tilde{\mu}_i}{\tilde{\mu}_i(\mathbb{X})} \mid \boldsymbol{X}_{1:d} \overset{d}{=} \sum_{j=1}^{k} \pi_{ij}\, \delta_{X_j^*} + \sum_{\ell \geq 1} \pi_{i\ell}^*\, \delta_{Y_\ell} \quad (i = 1, \dots, d). \tag{S12}$$

For the normalized gamma-gamma hCRV, the conditional distribution of each posterior random probability is a Dirichlet process with discrete base measure.

Proposition S3.1. *Let $\tilde{\boldsymbol{P}}$ be a normalized gamma-gamma hCRV. A posteriori, the random probabilities $\tilde{\boldsymbol{P}} \mid \boldsymbol{X}_{1:d}$ are conditionally independent, given variables $\alpha J_{01}, \dots, \alpha J_{0k}$ and the random measure $\tilde{\mu}_0^*$ in Proposition 5.1, and distributed as*

$$\tilde{P}_i = \frac{\tilde{\mu}_i}{\tilde{\mu}_i(\mathbb{X})} \mid \boldsymbol{X}_{1:d}, \alpha\boldsymbol{J}_0, \tilde{\mu}_0^* \overset{\text{ind}}{\sim} \text{DP}\left(\alpha\tilde{\mu}_0^* + \sum_{j=1}^{k} (n_{ij} + \alpha J_{0j})\, \delta_{X_j^*}\right) \qquad (i = 1, \dots, d).$$

Proof. By Proposition 5.1, independently for each $i = 1, \dots, d$, the random measure $\tilde{\mu}_i^* \mid \tilde{\mu}_0^*, U_i$ is a gamma CRM with shape $\alpha\tilde{\mu}_0^*(\mathbb{X})$ and rate $b + U_i$. Moreover, for each $j = 1, \dots, k$, the random jump $J_{ij} \mid \alpha J_{0j}, U_i$ is independently gamma distributed with shape $n_{ij} + \alpha J_{0j}$ and rate $b + U_i$. Therefore, the posterior distribution of $\tilde{\mu}_i \mid \alpha\boldsymbol{J}_0, \tilde{\mu}_0^*, U_i$ is that of a gamma CRM, being a superposition of independent gamma processes with same rate. Indeed, its Lévy intensity is

$$s^{-1} e^{-b(1+U_i/b)} \left(\alpha\tilde{\mu}_0^* + \sum_{j=1}^{k} (n_{ij} + \alpha J_{0j})\, \delta_{X_j^*}\right) \qquad (i = 1, \dots, d).$$

The normalization of a gamma CRM is then a Dirichlet process (Ferguson, 1973). □

Remarkably, the posterior distribution of $\tilde{\boldsymbol{P}}$ does not depend on the prior rate parameter b, as already discussed right after Proposition 3.3. Moreover, it is conditionally independent of latent variables $\boldsymbol{U}$, given $\alpha\boldsymbol{J}_0$ and $\tilde{\mu}_0^*$.

This result is particularly relevant from the algorithmic point of view, as anticipated above. Indeed, after obtaining samples from variables $\alpha J_{01}, \dots, \alpha J_{0k}$ and an approximation by truncation of $\alpha\tilde{\mu}_0^*$, as described in Section S3.4, the probability weights in (S12) can

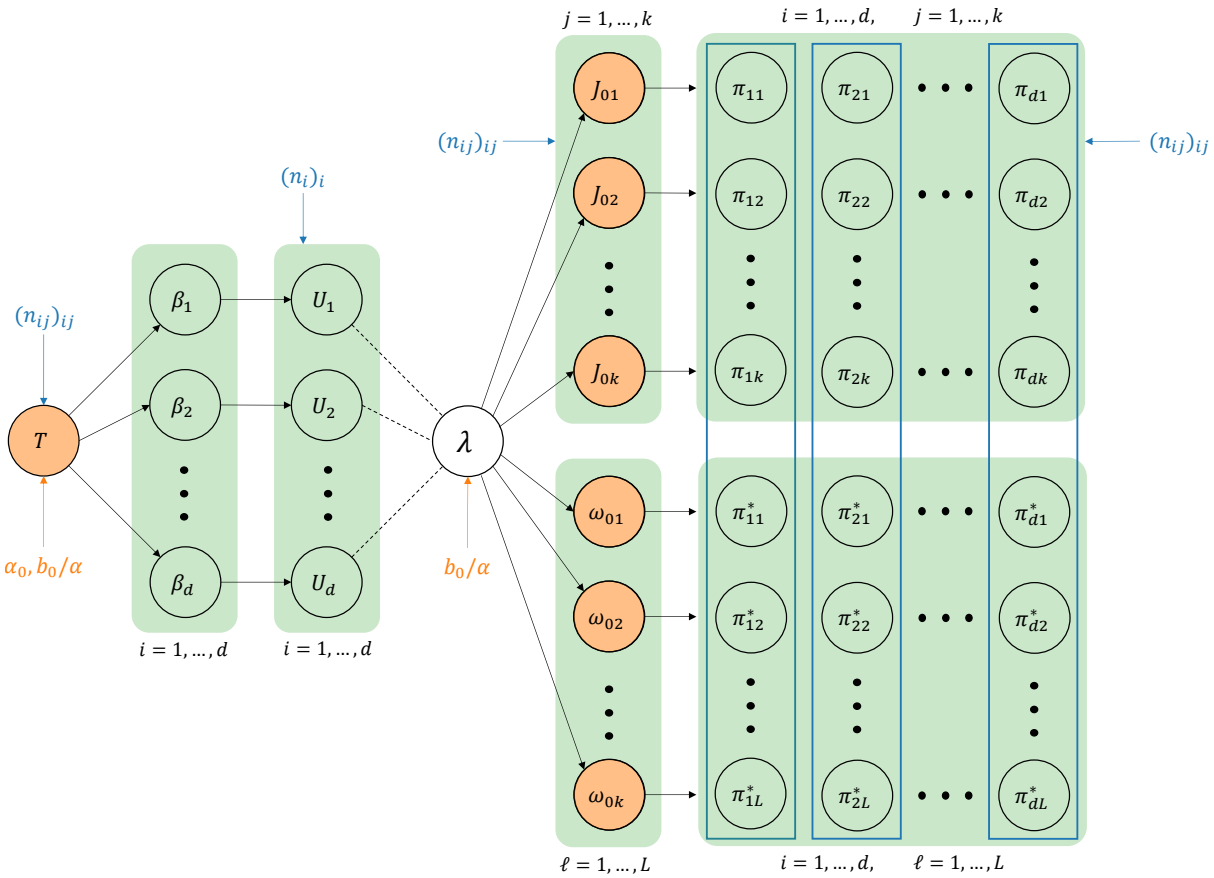

Figure S3: Conditional dependencies between random variables in the sampling algorithms for posterior normalized random measures. Red circles represent the computation bottlenecks; quantities enclosed in blue boxes are sampled from Dirichlet distributions. Variables are reported up to scaling w.r.t. model parameters.

be sampled from a $(k + L)$-dimensional Dirichlet distribution. In particular, for each $i = 1, \dots, d$ and $L \in \mathbb{N}$,

$$\begin{aligned}\left(\pi_{i1}, \dots, \pi_{ik}, \pi^*_{i1}, \dots, \pi^*_{iL}\right) \mid \boldsymbol{X}_{1:d}, \boldsymbol{J}_0, \tilde{\mu}^*_0 \\ \sim \text{Dirichlet}\left(n_{i1} + \alpha J_{01}, \dots, n_{ik} + \alpha J_{0k}, \alpha\omega_{01}, \dots, \alpha\omega_{0L}\right).\end{aligned}$$

The conditional independence from $\boldsymbol{U}$ allows further parallelization of the posterior sampling schemes. The resulting structure of conditional dependencies within the complete sampling algorithms for the normalized posterior random measures is summarized in Figure S3.

S3.7 Marginal Gibbs samplers for the HDP

This section provides the details for the derivation of the marginal posterior samplers for the HDP considered in Section 5.3, namely the CRF-based Gibbs sampler and the collapsed Gibbs sampler. The CRF-based Gibbs sampler is based on existing work (Teh et al.,

2006; Camerlenghi et al., 2019), which we recall here using our notation. Proposition S3.2 establishes a new result, which determines our collapsed Gibbs sampler; this can be seen as an extension of the sampling schemes proposed in Teh et al. (2006) and Bacallado et al. (2022).

The hierarchical Dirichlet process was first described by Teh et al. (2006), where the restaurant franchise metaphor is introduced. According to this metaphor, customers in each restaurant eat dishes selected from a countably infinite menu that is shared across restaurants. Moreover, in each restaurant, customers eating the same dish may or may not seat at the same table, while customers eating different dishes are necessarily seated at different tables. In other words, observations are organized within a nested partition structure. Remarkably, this structure is not specific to the hierarchical Dirichlet process, but characterizes every hierarchical structure of normalized random measures considered in Camerlenghi et al. (2019); see also Catalano et al. (2024). Such partition is typically described through two collections of variables, displaying ties among each collection: for each restaurant $i = 1, \dots, d$, variables $X_{i\ell}$ and $Z_{i\ell}$ denote, respectively, the dish and table of the ℓ-th customer. The distinct values taken by variables $\boldsymbol{X}_{1:d} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_d)$ are denoted by $X_1^*, \dots, X_k^*$, and represent the distinct dishes from the common menu. Similarly, for each restaurant $i = 1, \dots, d$, the distinct values taken by the variables $\boldsymbol{Z}_i = (Z_{i1}, \dots, Z_{in_i})$, are denoted by $Z_{ij1}^*, \dots, Z_{ijh_{ij}}^*$ for $j = 1, \dots, k$, and represent the different tables serving each dish in the i-th restaurant. The particular choice of hierarchical random measures determines the joint probability distribution of $\boldsymbol{X}_{1:d}$ and $\boldsymbol{Z}_{1:d} = (\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_d)$, namely the dishes and tables allocation, or equivalently the sequence of their predictive distributions. For a hierarchical Dirichlet process with concentration parameters $\alpha, \alpha_0 > 0$, the joint probability of tables and dishes takes the form

$$\mathbb{P}(\boldsymbol{X}_{1:d}, \boldsymbol{Z}_{1:d}) = \frac{\alpha_0^k}{((\alpha_0))_h} \prod_{j=1}^{k} \Gamma(h_{\bullet j}) \prod_{i=1}^{d} \left(\frac{\alpha^{h_{i\bullet}}}{((\alpha))_{n_i}} \prod_{j=1}^{k} \prod_{r=1}^{h_{ij}} \Gamma(q_{ijr}) \right). \tag{S13}$$

In the expression above, h_{ij} denotes the (random) number of tables in which the n_{ij} customers eating dish X_j^* in restaurant i are partitioned, while q_{ijr}, for $r = 1, \dots, h_{ij}$, denotes the (random) number of customers seated at each of those tables.

Contrary to the approach of Teh et al. (2006), which considers kernel mixtures, our model (4) assumes that the dishes $\boldsymbol{X}_{1:d}$ eaten by each customer are actually observed, and thus fixed. Therefore, one only has to sample the tables $\boldsymbol{Z}_{1:d}$, in order to obtain a complete description of the nested partition. Remarkably, the full conditional distribution of each $Z_{i\ell}$, given $\boldsymbol{X}_{1:d}$ and the other $\boldsymbol{Z}_{1:d}$, can be easily derived from the joint probability (S13). According to the restaurant franchise metaphor, having observed that the ℓ-th customer from restaurant i eats dish X_j^*, they may either sit at tables $Z_{ij1}^*, \dots, Z_{ijh_{ij}}^*$ or open a new table. More precisely, given $X_{i\ell} = X_j^*$, then

$$\mathbb{P}(Z_{i\ell} = Z^*_{ijr} \mid \cdots) \propto \frac{q^{-\ell}_{ijr}}{\alpha + n_i}, \qquad r = 1, \ldots, h_{ij},$$
$$\mathbb{P}(Z_{i\ell} = \text{'new'} \mid \cdots) \propto \frac{\alpha}{\alpha + n_i} \frac{h^{-\ell}_{\bullet j}}{\alpha_0 + h},$$

where $q^{-\ell}_{ijr}$ and $h^{-\ell}_{\bullet j}$ denote, respectively, the number of customers at table Z^*_{ijr} and the number of tables serving dish X^*_j across all restaurants, once customer ℓ is removed from the partition structure. As a result, a straightforward Gibbs sampler consists in sequentially resampling each $Z_{i\ell}$, for $i = 1, \ldots, d$ and $\ell = 1, \ldots, n_i$, from its conditional distribution. In Section 5.3, we refer to this sampling scheme as CRF-based Gibbs sampler. Note that, since one sequentially updates n latent variables, the computational cost is at least linear in the number of observations.

The posterior distribution of the hierarchical Dirichlet process (3), given the allocations of the observations within the nested partition, preserves the hierarchical structure (e.g., Teh and Jordan, 2010),

$$\tilde{P}_i \mid \boldsymbol{X}_{1:d}, \boldsymbol{Z}_{1:d}, \tilde{P}_0 \sim \text{DP}\left(\alpha \tilde{P}_0 + \sum_{j=1}^{k} n_{ij} \delta_{X^*_j}\right); \qquad \tilde{P}_0 \sim \text{DP}\left(\alpha_0 P_0 + \sum_{j=1}^{k} h_{\bullet j} \delta_{X^*_j}\right).$$

Interestingly, these distributions depend only on the numbers of customers n_{ij}'s for each restaurant and dish, and on the number of tables $h_{\bullet j}$'s serving each dish across all restaurants. In our setting, the counts $(n_{ij})_{ij}$ are observed, and thus the posterior distribution is fully characterized by sampling the $h_{\bullet j}$'s from their conditional distribution.

Proposition S3.2. *Let $\tilde{\boldsymbol{P}}$ be a hierarchical Dirichlet process, characterized by the probability distribution (S13). Conditionally on the counts $(n_{ij})_{ij}$, the probability of the discrete variables $h_j = h_{\bullet j}$ for $j = 1, \ldots, k$ is*

$$\mathbb{P}(h_1, \ldots, h_k) \propto \frac{\alpha^h}{((\alpha_0))_h} \prod_{j=1}^{k} \Gamma(h_j)\, S(n_{1j}, \ldots, n_{dj}; h_j),$$

where coefficients $S(q_1, \ldots, q_d; h)$ are the generalized Stirling numbers defined in (12).

Proof. For a hierarchical Dirichlet process with concentration parameters $\alpha, \alpha_0 > 0$, the joint probability of $\boldsymbol{X}_{1:d}$ and the tables counts $(h_{ij})_{ij}$ is

$$\frac{\alpha_0^k}{((\alpha_0))_h} \prod_{j=1}^{k} \Gamma(h_{\bullet j}) \prod_{i=1}^{d} \left(\frac{\alpha^{h_{i\bullet}}}{((\alpha))_{n_i}} \prod_{j=1}^{k} S(n_{ij}; h_{ij}) \right).$$

This expression can be obtained from (S13) by summing over the set of unordered partitions of $\{1, \ldots, n_{ij}\}$ into h_{ij} non-empty cycles, for each restaurant $i = 1, \ldots, d$ and dish $j = 1, \ldots, k$. The same result is derived in Camerlenghi et al. (2019, Example 3),

where the probability above is stated with a further marginalization over tables $(h_{ij})_{ij}$. Furthermore, for each dish j, we sum over the set of values $(h_{ij})_i$ such that $h_{1j}+\cdots h_{dj} = h_j$ and $0 \leq h_{ij} \leq n_{ij}$, for $i = 1, \ldots, d$. Exploiting the definition of multivariate Stirling numbers (S10), the probability distribution of $\boldsymbol{X}_{1:d}$ and the h_j's is

$$\frac{\alpha_0^k \, \alpha^h}{((\alpha_0))_h} \prod_{i=1}^{d} \frac{1}{((\alpha))_{n_i}} \prod_{j=1}^{k} \Gamma(h_j) \, S(n_{1j}, \ldots, n_{dj}; h_j).$$

The conditional distribution of the h_j's is proportional to the expression above. □

The direct sampling of the number of tables serving each dish in each restaurant was first suggested in Teh et al. (2006), although conditionally on the random probability $\tilde{P}_0$. Instead, Bacallado et al. (2022) adopt a marginal approach and rely on the conditional distribution of the h_{ij}'s for the hierarchical Pitman-Yor process, as derived in Camerlenghi et al. (2019, Theorem 4). The further marginalization over restaurants in Proposition S3.2 is particularly suitable for HDP, since its posterior structure depends only on the $h_{\bullet j}$'s; moreover, it conveniently reduces the dimension of the sampling space. As a result, one easily devises a Gibbs sampler for the $h_{\bullet j}$'s based on their joint distribution. In Section 5.3, we refer to this sampling approach as collapsed Gibbs sampler.

Finally, for a full comparison with the normalized gamma-gamma hCRV, we often consider a gamma prior on the concentration parameter $\alpha \sim \text{Gamma}(\alpha_0, \beta)$; see Proposition 3.3 for the distributional equivalence result. The update of α within both Gibbs samplers outlined above is performed via a Metropolis-Hastings step, targeting the full conditional distribution

$$\alpha \mid \boldsymbol{X}_{1:d}, \boldsymbol{Z}_{1:d} \sim f(s) \propto s^{\alpha_0+h-1} \, e^{-\beta \, s} \prod_{i=1}^{d} \frac{1}{((s))_{n_i}}.$$

An alternative sampling scheme based on an augmentation with auxiliary beta and Bernoulli variables is detailed in the Appendix of Teh et al. (2006). The Julia implementation of the CRF-based and the collapsed Gibbs samplers is available at github.com/claudiodelsole/hCRV.jl, where we also provide an interface to allow their integration within the R environment.

S3.8 Numerical illustrations

This section contains numerical illustrations of the posterior sampling algorithms introduced in Section 5.3. Specifically, we consider the MCMC sampler in Algorithm 1 with symmetric random-walk Metropolis-Hastings steps on the log-scale (mcmc), detailed in Section S3.1, and the exact sampler in Algorithm 2 (exact). As a reference to evaluate their effectiveness, we take the marginal Gibbs sampler of Teh et al. (2006) for the HDP,

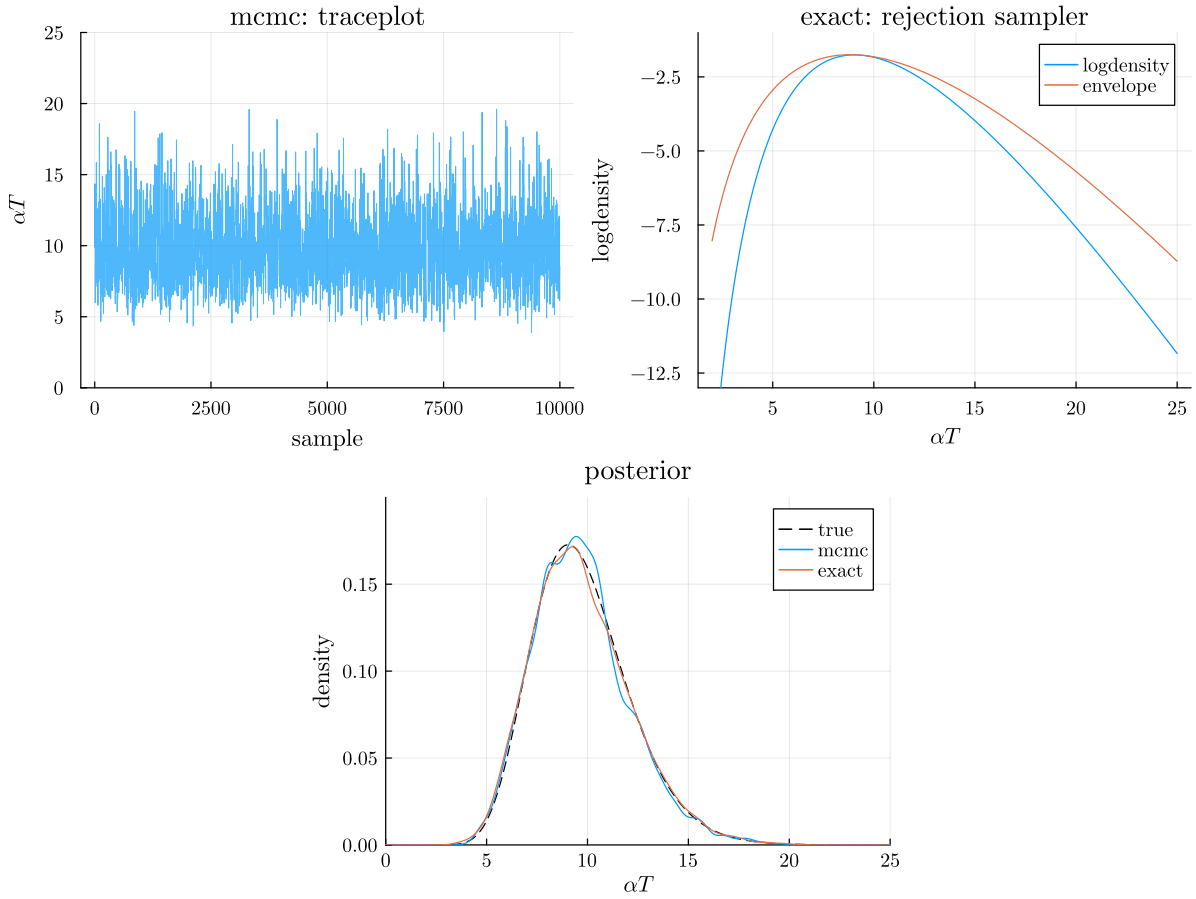

Figure S4: Diagnostics and posterior distribution for the random variable αT. Top left: traceplot for the MCMC algorithms. Top right: envelope for the rejection sampler. Bottom: estimates of the posterior distribution via Gaussian kernel smoothing, compared with the true density.

based on the restaurant franchise metaphor, with gamma prior on the concentration parameter; see Section S3.7 above for details. These algorithms target the same posterior distribution for the random probabilities (Proposition 3.3), hence their posterior estimates should coincide. We consider $d = 4$ groups of observations, each of size $n_i = 50$, sampled from independent Poisson distributions with means $2, 3, 4$ and 5. The number of distinct values in the simulated dataset is $k = 11$. Model parameters are fixed at $\alpha_0 = \alpha = 1$ and $b_0 = b = 1$. We draw $10,000$ posterior samples for each algorithm; for the MCMC schemes, we consider a burn-in of $1,000$ steps.

Figure S4 provides insights into the posterior sampling of random variable αT. The MCMC algorithm shows good mixing, with an effective sample size of 1996. The envelope for the rejection sampling in the exact algorithm is tight, and entails an acceptance rate of 0.765; the optimal value for r is 8.84 (see Section S3.3). Instead, Figure S5 displays the MCMC traceplot and posterior distributions for some of the J_{0j}'s. Specifically, we focus on the latent jumps at values $X_5^* = 4$ and $X_8^* = 7$, with counts $n_{\bullet 5} = 36$ and $n_{\bullet 8} = 7$. Again, the MCMC algorithms show good mixing, with effective sample sizes above $1,600$

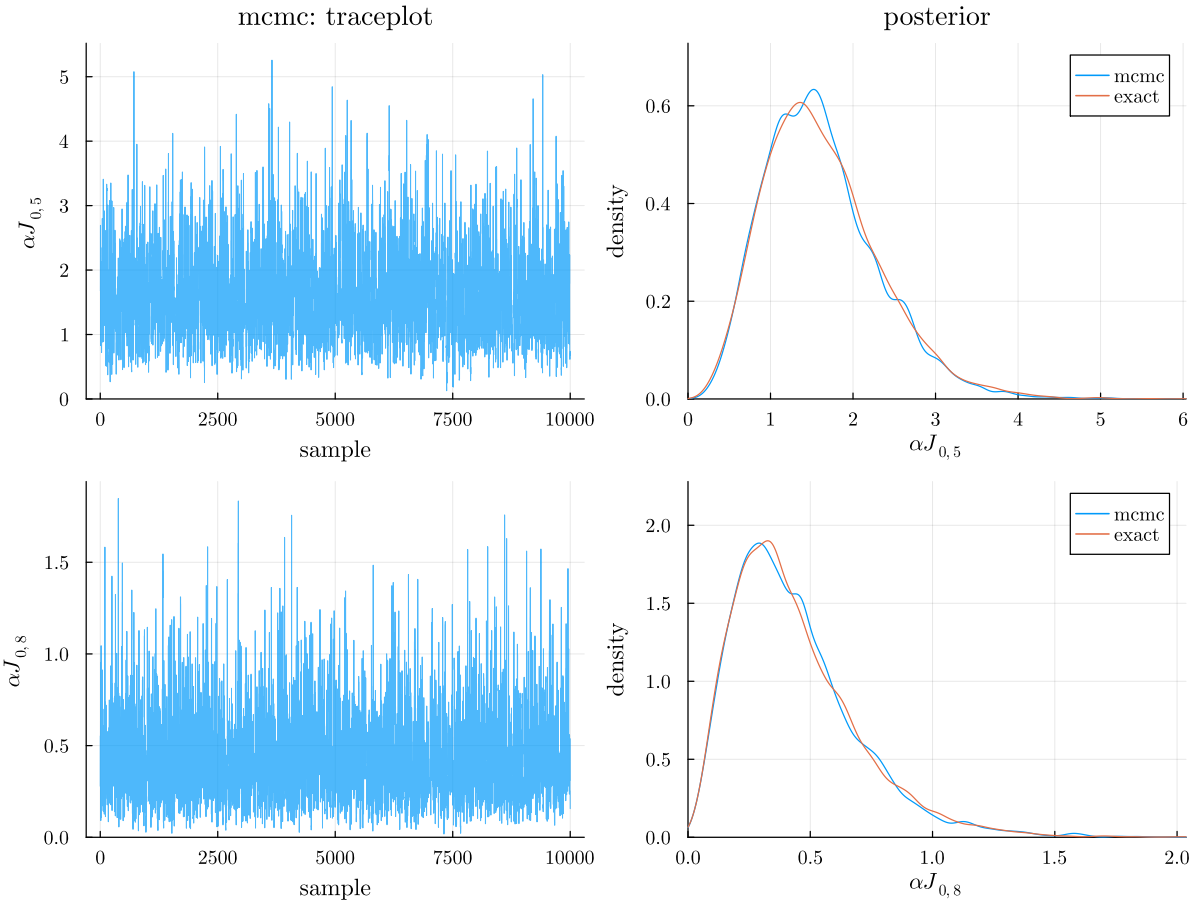

Figure S5: Diagnostics and posterior distributions for the latent jumps variables J_{0j} at values $X_5^* = 4$ and $X_8^* = 7$. Left: traceplots for the MCMC algorithm. Right: estimates of the posterior distributions via Gaussian kernel smoothing.

for all J_{0j}'s. Finally, the posterior distributions for the jumps J_{ij} at values $X_5^* = 4$ and $X_8^* = 7$ for the $d = 4$ groups are displayed in Figure S6. The effective sample sizes are above $3,300$ for all J_{ij}'s.

Furthermore, we conduct a visual comparison with the marginal Gibbs sampler for the HDP in terms of the accuracy of posterior distributions. Since this algorithm outputs posterior probability weights, we sample from the posterior random probabilities according to the procedure outlined in S3.6. Considering the same experimental setting described above, we compare the MCMC sampler, the exact sampler, and the marginal Gibbs sampler for HDP, with or without gamma prior on the concentration parameter (HDP and HDPpr, respectively). The posterior distributions for the probability weights π_{ij} at the same values $X_5^* = 4$ and $X_8^* = 7$ are displayed in Figure S7. The effective sample sizes consistently exceed $4,000$. As expected, estimates of the posterior distributions for the HDP model without prior visibly differ from those obtained with the other algorithms, which instead target the same posterior distributions. This difference may be likewise observed in the right plot of Figure 3.

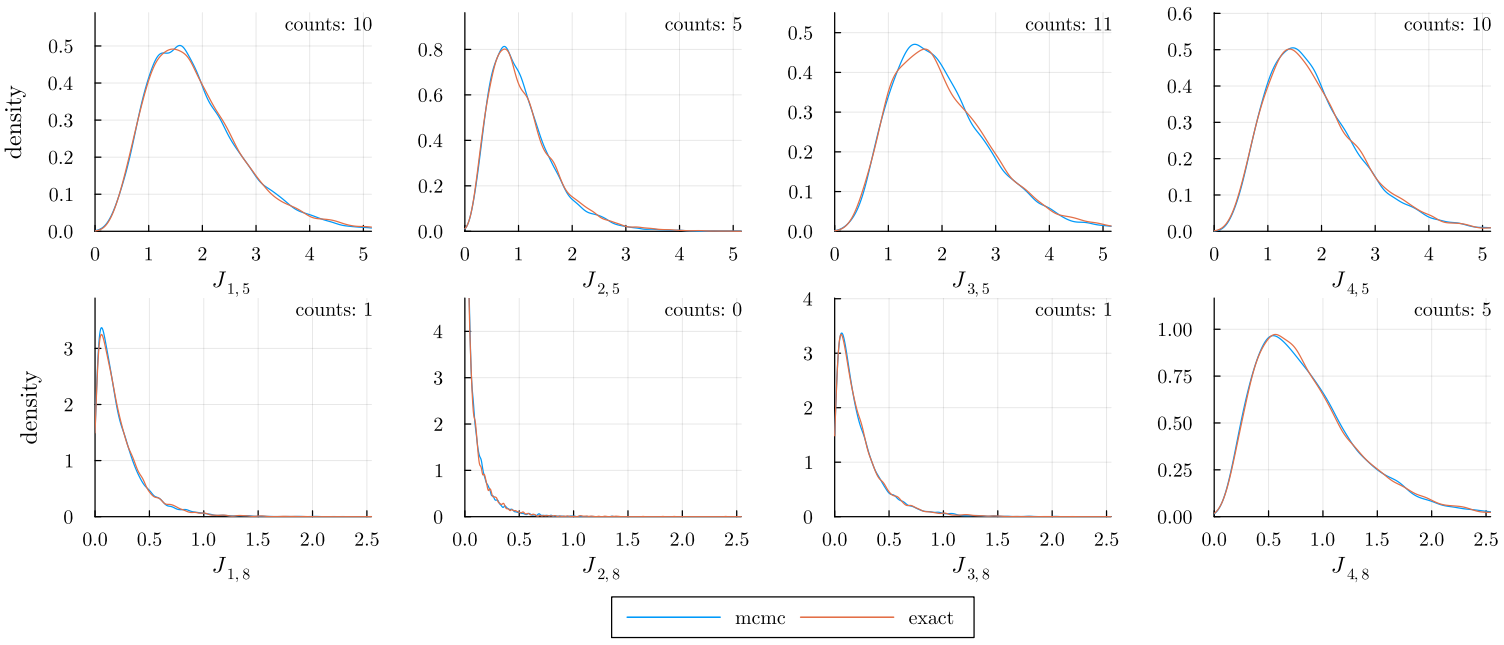

Figure S6: Posterior distributions for the jumps J_{ij} at values $X_5^* = 4$ and $X_8^* = 7$, for the $d = 4$ groups, obtained via Gaussian kernel smoothing.

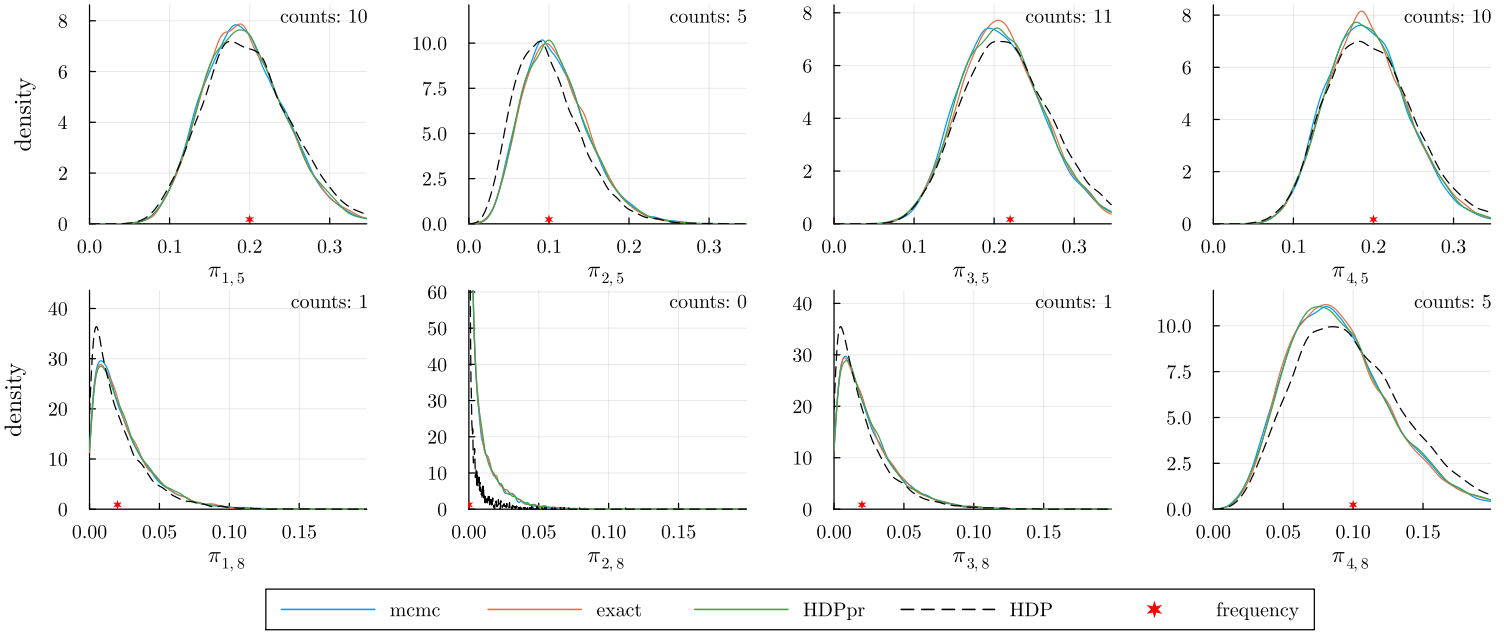

Figure S7: Posterior distributions for the probability weights π_{ij} at values $X_5^* = 4$ and $X_8^* = 7$, for the $d = 4$ groups. The HDP model (without prior) targets different posterior distributions.

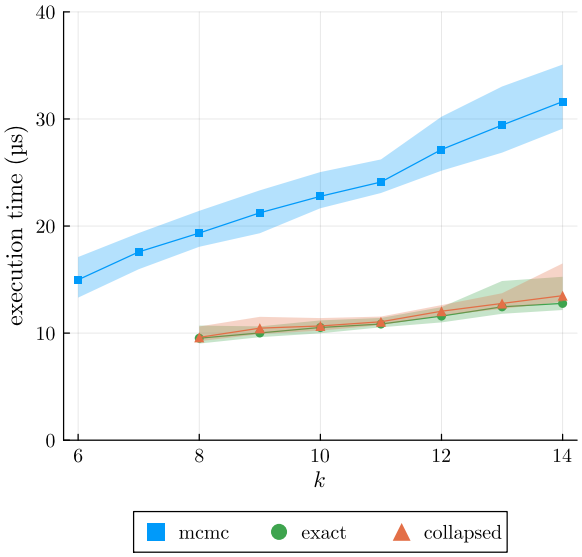

Figure S8: Execution times per effective sample for the different algorithms, with increasing number of distinct values k. The number of groups $d = 10$ and observations per group $n_i = 25$ are fixed. Results are averaged over 100 simulated datasets per experimental setting. Solid curves represent median values, with shaded areas between the first and third quartiles. Times for the CRF-based sampler are around 120μs.

S3.9 Comparison for increasing number of distinct values

Section 5.4 of the main manuscript compares the different algorithms for posterior sampling through simulation studies. The comparison is performed in terms of computational time per effective sample, as the number of groups d or the number of observations per group n_i increase. In this section, we conduct a further simulation study fixing both the number of groups $d = 10$ and the number of observations per group $n_i = 25$, while allowing the number of distinct values k to vary. To enhance the variability of k across simulated datasets, observations are sampled from a hierarchical Dirichlet process with random concentration parameters $\alpha \sim \text{Gamma}(5)$ and $\alpha_0 \sim \text{Gamma}(3)$. Consistently with Section 5.4, results are averaged over 100 simulated datasets for each value of k, and execution times are plotted whenever at least 75 experiments are completed without errors. Remarkably, numerical overflows in the computation of the multivariate Stirling numbers appear more frequently for smaller values of k; indeed, since the number of observations $n = 250$ is fixed, a smaller k entails a larger number of observations for each distinct value. As displayed in Figure S8, the execution time for the MCMC algorithm grows linearly with k: this is in line with the fact that the state of the Markov chain has dimension $2k + 1$. The exact sampler and collapsed Gibbs sampler also display a nearly linear dependence on the number of distinct values, with a smaller slope compared with the MCMC approach. In contrast, the execution time for the CRF-based Gibbs sampler (not shown) is essentially constant.

S4 Eliciting the dependence structure

Normalized hierarchical CRVs induce dependence between the marginal random probabilities, which in turn regulates the borrowing of information across different groups. Hierarchical models are a natural way to induce positive dependence, especially in a Bayesian setting. However, as highlighted in Catalano et al. (2024), the elicitation of the dependence is more difficult with respect to other proposals because model parameters typically affect both the marginal distribution and the dependence across groups. For this reason, Catalano et al. (2024) propose two kinds of (weak) flexibility for partially exchangeable models: (i) for every $\rho \in [0,1]$, there exists a specification of the parameters such that $\text{Corr}(\tilde{P}_i(A), \tilde{P}_j(A)) = \rho$, where A is a fixed set s.t. $E(\tilde{P}_i(A)) \neq 0, 1$; (ii) for every $\rho \in [0,1]$ and for every fixed value of the marginal mean $E(\tilde{P}_i(A))$ and variance $\text{Var}(\tilde{P}_i(A))$, there exists a specification of the parameters s.t. $\text{Corr}(\tilde{P}_i(A), \tilde{P}_j(A)) = \rho$. The expressions in Example 3.7 show that the normalized gamma-gamma hCRV achieves the flexibility of second kind. In this section, we investigate the role of parameter values in the borrowing of information, for fixed values of the marginal mean and variance.

Let $\tilde{\boldsymbol{\mu}}$ be a gamma-gamma hCRV as in Example 2.7 and let A be a Borel set s.t. $P_0(A) \neq 0, 1$. From Example 3.7, it is apparent that correlation is not affected by the mean $E(\tilde{P}_i(A)) = P_0(A)$ but is deeply related to the variance. Indeed, one has

$$\begin{aligned}
\text{Corr}(\tilde{P}_i(A), \tilde{P}_j(A)) &= \Big(1 + \frac{\alpha_0}{\alpha} e^{1/\alpha} E_{\alpha_0}\big(1/\alpha\big)\Big)^{-1}, \\
\text{Var}(\tilde{P}_i(A)) &= \Big(1 + \frac{\alpha_0}{\alpha} e^{1/\alpha} E_{\alpha_0}\big(1/\alpha\big)\Big) \frac{P_0(A)(1 - P_0(A))}{1 + \alpha_0}.
\end{aligned}$$

We refer to Table S2 for some limiting behaviours, which are compared with those of the HDP. Since α and α_0 impact both the variance and the dependence structure, it may be difficult to elicit them in practice. From the point of view of the practitioner, it is certainly more intuitive to elicit mean, variance and correlation independently. Hence, one can choose $\sigma^2, \rho \in (0,1)$ such that $\text{Var}(\tilde{P}_i(A)) = \sigma^2\, P_0(A)(1 - P_0(A))$ and $\text{Corr}(\tilde{P}_i(A), \tilde{P}_j(A)) = \rho$, and then find the corresponding values of α and α_0 by solving the system of non-linear equations in Section 6, namely

$$\rho(1 + \alpha_0/\alpha\, e^{1/\alpha} E_{\alpha_0}(1/\alpha)) - 1 = 0, \qquad \sigma^2(1 + \alpha_0) - 1/\rho = 0,$$

Figure S9 displays the values of α, α_0 that correspond to a range of values for σ^2 and ρ, along with the corresponding analysis for the HDP.

To visualize the effect of the borrowing, we consider the same simulated dataset as in Section 6, namely $d = 3$ groups of independent Poisson observations, each of size $n_i = 10$, with means equal to 2, 3 and 4. Figure S10 displays the expected values of the posterior random means $E\big(\int x \,\mathrm{d}\tilde{P}_i(x) \,\big|\, \boldsymbol{X}\big)$, which coincide with the means of the

parameters	fixed $\alpha > 0$		fixed $\alpha_0 > 0$	
	$\alpha_0 \to 0$	$\alpha_0 \to +\infty$	$\alpha \to 0$	$\alpha \to +\infty$
$\sigma^2 = \sigma^2(\alpha, \alpha_0)$	1	0	1	1
$\rho = \rho(\alpha, \alpha_0)$	1	$\alpha/(1+\alpha)$	$1/(1+\alpha_0)$	1

parameters	fixed $\alpha > 0$		fixed $\alpha_0 > 0$	
	$\alpha_0 \to 0$	$\alpha_0 \to +\infty$	$\alpha \to 0$	$\alpha \to +\infty$
$\sigma^2 = \sigma^2(\alpha, \alpha_0)$	1	$1/(1+\alpha)$	1	$1/(1+\alpha_0)$
$\rho = \rho(\alpha, \alpha_0)$	1	0	$1/(1+\alpha_0)$	1

Table S2: Limiting behaviours of variance and correlation parameters for the normalized gamma-gamma hCRV (top) and HDP (bottom), as functions of their respective parameters $\alpha, \alpha_0 > 0$, with $\mathrm{Var}(\tilde{P}_i(A)) = \sigma^2 P_0(A)(1 - P_0(A))$ and $\mathrm{Corr}(\tilde{P}_i(A), \tilde{P}_j(A)) = \rho$.

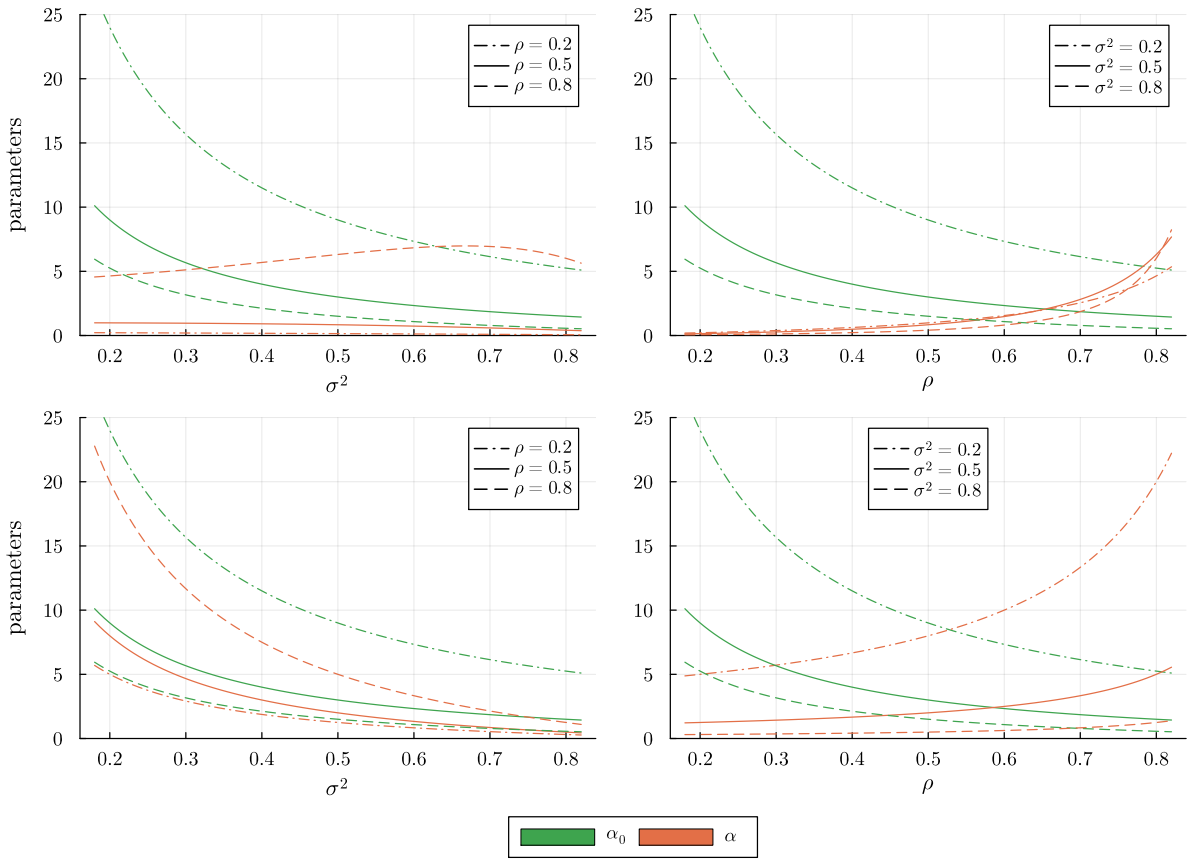

Figure S9: Values of parameters α and α_0 for the normalized gamma-gamma hCRV (top) and HDP (bottom) corresponding to fixed values of variance σ^2 and correlation ρ.

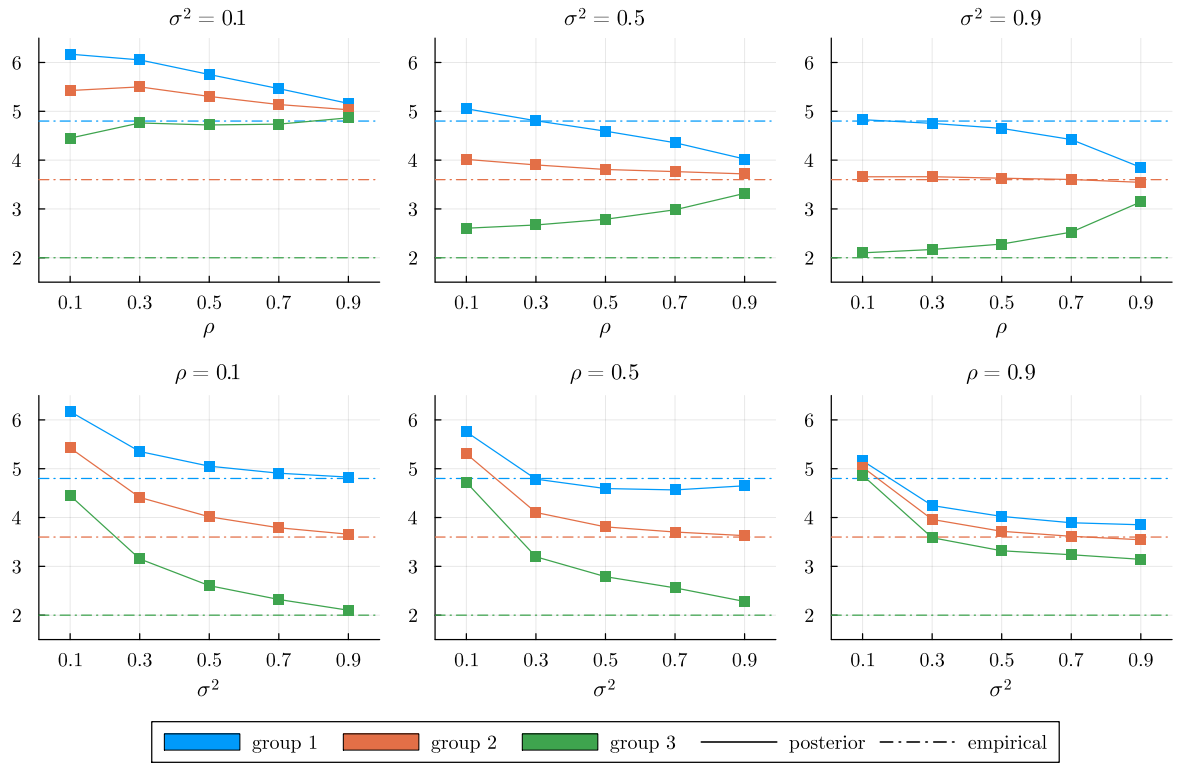

Figure S10: Posterior expected random means of the normalized gamma-gamma hCRV model, for three groups of independent Poisson observations with means equal to 2, 3 and 4, each of size $n_i = 10$, and prior mean $P_0 = N(10, 1)$. Top: fixed variance and increasing correlation. Bottom: fixed correlation and increasing variance.

predictive distributions for the three groups, as we vary the prior pairwise correlation coefficient $\rho \in (0, 1)$ and the prior variance through $\sigma^2 \in (0, 1)$, keeping the base measure $P_0 = N(10, 1)$ fixed. Posterior samples are obtained using the exact algorithm. For a fixed variance σ^2, a higher correlation ρ induces more borrowing, and thus the posterior means are closer to each other. Depending on σ^2, the estimates are closer to the prior (which is pushing them towards higher values, being centered at 10), or to their empirical means. On the other hand, for a fixed correlation ρ, a higher variance σ^2 reduces the weight of the prior, and thus pushes the estimates towards their empirical means. Interestingly, lower values of ρ can induce estimates that are closer to each other than higher values of ρ, depending on the value of σ^2. Indeed, low values of σ^2 will force estimates to be closer to the prior, and thus also closer to each other; e.g., compare $\rho = \sigma^2 = 0.1$ vs. $\rho = 0.5$ and $\sigma^2 = 0.9$. This is the effect of the shrinkage, which can sometimes be difficult to distinguish from the borrowing of information, especially if the prior mean is close to the grand mean of the observations.