Journal of the Royal Statistical Society Series B: Statistical Methodology, 2025, **00**, 1–27
https://doi.org/10.1093/jrsssb/qkaf027

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Original Article

Bayesian mixture models with repulsive and attractive atoms

Mario Beraha[1], Raffaele Argiento[2], Federico Camerlenghi[1] and Alessandra Guglielmi[3]

[1]Department of Economics, Management and Statistics, University of Milano-Bicocca, Milano, Italy
[2]Department of Economics, University of Bergamo, Bergamo, Italy
[3]Department of Mathematics, Politecnico di Milano, Milano, Italy

Address for correspondence: Mario Beraha, Department of Economics, Management and Statistics, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano, Italy. Email: mario.beraha@unimib.it

Abstract

The study of almost surely discrete random probability measures is an active line of research in Bayesian nonparametrics. The idea of assuming interaction across the atoms of the random probability measure has recently spurred significant interest in the context of Bayesian mixture models. This allows the definition of priors that encourage well-separated and interpretable clusters. In this work, we provide a unified framework for the construction and the Bayesian analysis of random probability measures with interacting atoms, encompassing both repulsive and attractive behaviours. Specifically, we derive closed-form expressions for the posterior distribution, the marginal and predictive distributions, previously unavailable except for the case of measures with i.i.d. atoms. We show how these quantities are fundamental for both prior elicitation and developing new posterior simulation algorithms for hierarchical mixture models. Our results are obtained without any assumption on the finite point process governing the atoms of the random measure. Their proofs rely on analytical tools borrowed from Palm calculus theory, which might be of independent interest. We specialize our treatment to the classes of Poisson, Gibbs, and determinantal point processes, as well as in the case of shot-noise Cox processes. Finally, we illustrate different modelling strategies on simulated and real datasets.

Keywords: distribution theory, mixture of mixtures, Palm calculus, repulsive point processes, shot-noise Cox process

1 Introduction

Clustering is a fundamental problem in statistics and machine learning, being one of the workhorses of unsupervised learning, aiming at dividing datapoints into similar groups. The Bayesian approach to clustering offers substantial advantages over traditional algorithms, allowing for straightforward uncertainty quantification around point estimates. See Wade (2023) and Grazian (2023) for two recent and comprehensive reviews. The most common formulations of Bayesian clustering revolve around mixture models (Fruhwirth-Schnatter et al., 2019). Mixture models assume that observations belong to one of a potentially infinite number of groups or components, and each group is suitably modelled by a parametric density $f(\cdot \mid Y)$ for some parameter Y; their relative prevalence is specified by random weights $(p_j)_{j\geq 1}$ (such that $p_j > 0$ and $\sum_{j\geq 1} p_j = 1$ almost surely).

Recent contributions focused on the robustness of the cluster estimate to model misspecification, establishing a lack thereof in commonly adopted models, both empirically (Miller & Dunson, 2019) and theoretically (Cai et al., 2021; Guha et al., 2021). In particular, Cai et al. (2021) showed how, if the mixture kernel $f(\cdot \mid \cdot)$ does not agree with the true data generating process, the estimated number of clusters by Bayesian nonparametric (BNP) mixtures diverges as the

Received: October 24, 2023. **Revised:** April 28, 2025. **Accepted:** May 2, 2025

sample size increases. The root of this issue can be traced back to a fundamental trade-off between density estimation and clustering in the misspecified regime. To exemplify this, consider the case when data are generated by a mixture of Student t distributions, but the model assumes a mixture of Gaussian distributions. The consistency of Gaussian mixture models for the density estimates (Ghosal et al., 1999) necessarily entails that the posterior distribution will identify an ever increasing number of clusters, thus producing an inconsistent cluster estimate.

Repulsive mixture models (Beraha et al., 2022; Bianchini et al., 2020; Cremaschi et al., 2025; Petralia et al., 2012; Quinlan et al., 2021; Xie & Xu, 2019; Xu et al., 2016) offer a practical solution to this problem by forcing the mixture components to be well-separated by assuming a repulsive point process prior for the mixture locations. However, despite their recent popularity, the mathematical properties of repulsive mixtures have yet to be thoroughly investigated. As a result, prior elicitation strategies and posterior inference algorithms have been derived on a case-by-case basis, often based on heuristics in the case of prior elicitation and inefficient Markov chain Monte Carlo (MCMC) updates for posterior inference. In sharp contrast, a unified treatment of the main mathematical properties of traditional (non-repulsive) mixtures can be found in the cornerstone paper by James et al. (2009), which led to a dramatic increase in the popularity of Bayesian nonparametrics and fostered the development of novel methods, algorithms, and applications.

The aim of this article is twofold. First, we provide a framework for the Bayesian analysis of mixture models encompassing repulsiveness as well as other forms of dependence in the locations, such as attractiveness. In particular, we fill a gap in the literature by providing a unified framework allowing for the analysis of the associated mixing measure and establishing general results characterizing the distribution (both a priori and a posteriori) of several functionals of interest. Our results do not merely generalize existing theory, but they shed light on large classes of processes which are used in applications without appropriate theoretical investigation. Secondly, we propose to sidestep the aforementioned trade-off between density and cluster estimation by relying on a novel definition of cluster that naturally arises when assuming a shot-noise Cox process prior (Møller, 2003) for the mixture locations, whereby a cluster can consist of multiple mixture components with similar parameters. In particular, this is the first time such a prior is considered in mixture modelling.

1.1 Mixture models and almost surely discrete random probability measures

To formalize the notation, let $Z = (Z_1, \ldots, Z_n)$ be the observed sample, which is assumed to be distributed as

$$Z_i \mid \tilde{p} \overset{\text{iid}}{\sim} \int f(\cdot \mid y)\, \tilde{p}(\mathrm{d}y), \tag{1}$$

with the mixing distribution $\tilde{p}$ being an almost surely discrete random probability measure (RPM), $\tilde{p} = \sum_{j\geq 1} p_j \delta_{X_j}$. From the seminal work of Ferguson (1973) on the Dirichlet process, various approaches for constructing RPMs have been introduced. A fruitful strategy is based on normalising completely random measures with infinite activity, i.e. whose number of support points $(X_j)_{j\geq 1}$ is countably infinite. This idea, systematically introduced in Regazzini et al. (2003) for measures on $\mathbb{R}$ with the name of normalized random measures with independent increments (NRMIs), has been extended later to more general spaces (see, e.g. James et al., 2009). More recently, Argiento and De Iorio (2022) have exploited the same ideas to construct random probability measures with a random finite number of support points, named normalized independent finite point processes (IFPPs). Both approaches build a discrete RPM by normalising a marked point process where the points of the process define the jumps $(p_j)_{j\geq 1}$ of the RPM and i.i.d. marks $(X_j)_{j\geq 1}$ define the atoms. See also Lijoi et al. (2022) for an allied approach.

The direct study of model (1) is challenging, and it is customary to augment the parameter space via the introduction of latent parameters Y_i, $i = 1, \ldots, n$ such that

$$Z_i \mid Y_i \overset{\text{ind}}{\sim} f(\cdot \mid Y_i), \quad Y_i \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}, \quad i = 1, \ldots, n. \tag{2}$$

Then, a fundamental preliminary step for Bayesian inference in the mixture model is to study the distributional properties of the latent sample $Y_1, \ldots, Y_n$. Specifically, for prior elicitation

purposes, it is interesting to investigate the moments of $\tilde{p}$, the prior distribution of the distinct values in the sample, and the marginal distribution of the Y_i's. For inferential purposes instead, both from a methodological and computational perspective, it is central to derive the conditional distribution of $\tilde{p}$ given $Y_1, \ldots, Y_n$, as well as the predictive distribution of Y_{n+1} given the latent sample. See, e.g. James et al. (2009) and Argiento and De Iorio (2022) for the expression of such quantities in the case of normalized completely random measures and normalized independent finite point processes, respectively.

In this article, we consider model (2) and propose a construction of RPMs via normalization of marked point processes with inverted roles with respect to Regazzini et al. (2003) and Argiento and De Iorio (2022). In our construction, indeed, the atoms $(X_j)_{j\geq 1}$ are the points of a general simple point process, while the weights $(p_j)_{j\geq 1}$ are obtained by normalising i.i.d. positive marks associated with the atoms. Through the law of the point process, we are able to encourage different behaviours among the support points of the random probability measure, such as independence (when the point process is Poisson or the class of independent finite point processes), separation (i.e. the support points are well separated, when the point process is repulsive, such as Gibbs or determinantal point processes), and also random aggregation (i.e. the support points are clustered together, when the point process is of Cox type).

1.2 Our contributions and outline of the article

We begin by introducing the model for the Y_i's and the general construction for normalized random measures based on marked point processes in Section 2. In Section 3, we study the properties of the latent sample $Y_1, \ldots, Y_n$, providing closed-form expressions for several quantities of interest, allowing for a universal theory of very different dependence behaviours such as repulsion or attractiveness. This is possible thanks to the introduction of technical tools, mainly based on Palm calculus, which might be of independent interest. Palm calculus, beyond notable exceptions, is not typically known in the Bayesian nonparametric literature. Though the closed-form expressions that we obtain can be generally challenging to evaluate, they drastically simplify in the case of well-known point processes. The manuscript illustrates the general theory using determinantal and shot-noise Cox point processes. In the 3Appendix, we also discuss the case of Poisson and Gibbs point processes. Then, we resume the study of the Bayesian mixture model (2) in Section 4 and show that our analyses can be used as the building block for two Markov chain Monte Carlo algorithms to approximate the posterior distribution. In Section 5, we discuss two simulation studies and an application to real data with large sample size, highlighting the main difference between traditional, repulsive, and attractive mixtures. In particular, we show that in a setting with *corrupted* observations, repulsive mixtures are useful to recover the underlying signal, while traditional and attractive mixtures tend to overfit. On the other hand, in more classical misspecification settings where data exhibit heavy tails, we find that attractive mixtures provide a better clustering and density estimate than repulsive mixtures. We conclude with a discussion in Section 6. In the appendix, we report the proofs of the main results, as well as a discussion on the properties of the prior distribution of our random measures and more details on the examples using Poisson, Gibbs, determinantal, and shot-noise point processes. Moreover, online supplementary material, Appendix S9 provides further numerical illustrations, allowing us to explain how we fixed the hyperparameters. We also include a quantitative discussion on the sensitivity of these hyperparameters on density and cluster estimation.

From a technical point of view, our results are based on applying Palm calculus, a fundamental tool in studying point processes. Essentially, Palm calculus can be regarded as an extension of Fubini's theorem. It allows the exchange of the expectation and integral signs when both are with respect to a point process, with the subtle difference that (in general) the expectation is now taken with respect to the law of another point process, i.e. the *reduced Palm version* of the original process. In the Poisson process case (and therefore in the case of completely random measures), the reduced Palm version coincides with the law of the original Poisson process, for which computations are usually easily manageable. This property characterizes Poisson processes (Last & Penrose, 2017).

1.3 Related works

This work extends the treatment of Argiento and De Iorio (2022) and Beraha et al. (2022), who consider finite mixtures with a random number of components. Contrary to our work, Argiento and De Iorio (2022) focus on the case of independent and identically distributed atoms of the mixing distribution, while we allow for dependence (general interaction, e.g. repulsiveness or attraction) among the atoms. Beraha et al. (2022) consider only repulsive mixtures. The approach there lacks the thorough theoretical background that is instead provided in this article, though they present a general framework for repulsive mixture models, and, more importantly, derive an efficient MCMC algorithm which avoids the difficulties of the reversible jump MCMC computation.

Another degree of novelty of the present article concerns the use of Palm calculus for Bayesian analysis. In the context of BNP models, Palm calculus was introduced in James (2002, 2005) for the analysis of random probability measures built from Poisson processes. See also James (2006) and James et al. (2009) for applications to the Bayesian analysis of neutral to the right processes and mixtures of normalized completely random measures, respectively. In particular, Section 8 in James (2002) deals with the broad class of weighted Poisson random measures, which are functionals of Gibbs point processes using the terminology of point processes. Our approach is substantially different. First, we do not require that our processes are absolutely continuous with respect to the Poisson one, being able to deal with important processes, such as a large class of determinantal point processes and the shot-noise Cox processes. Secondly, our approach has the merit of being more direct than James (2002) because we work directly on the process of interest rather than on the dominating Poisson process. This is possible since we rely on novel technical tools for the analysis of Bayesian nonparametric models borrowed from point process theory. Consequently, in the mathematical expressions we obtain, it is possible to recognize well-known objects in the study of point processes, which might be more challenging to see in James (2002). This also allows us to borrow ideas from the literature on the simulation of spatial point processes (Møller & Waagepetersen, 2003) to address posterior inference.

2 A general construction for normalized random measures

We consider a sequence of random variables $(Y_i)_{i\geq 1}$ defined on the probability space $(\Omega, \mathcal{A}, P)$ and taking values in the Polish space $(\mathbb{X}, \mathcal{X})$, endowed with its Borel σ-algebra. Moreover, we denote by $\mathbb{P}(\mathbb{X})$ the space of all probability measures over $(\mathbb{X}, \mathcal{X})$, and $\mathcal{P}(\mathbb{X})$ stands for the corresponding Borel σ-algebra. We suppose that

$$\begin{aligned} Y_i \mid \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p} \quad i \geq 1 \\ \tilde{p} &\sim Q, \end{aligned} \tag{3}$$

where Q is a distribution over $(\mathbb{P}(\mathbb{X}), \mathcal{P}(\mathbb{X}))$. Note that, by de Finetti's theorem (de Finetti, 1937), (3) is equivalent to assuming that the Y_i's are exchangeable, which justifies the Bayesian approach to inference. A sample of size n from (3) consists of the first n terms of the sequence $(Y_i)_{i\geq 1}$. We will denote this sample by $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. In this and the following sections, $\boldsymbol{Y}$ represents the observed data points.

2.1 Random measures as functionals of point processes

In this section, we define the class of priors Q in (3) that we will deal with throughout the article. As standard in Bayesian non-parametrics (see, e.g. Lijoi & Prünster, 2010) we assume that $\tilde{p}$ is (almost surely) a purely atomic probability measure, $\tilde{p} = \sum_{j\geq 1} w_j \delta_{X_j}$, where $(w_j)_{j\geq 1}$ is a collection of positive random variables summing to one and the random atoms X_j's take value in $\mathbb{X}$. To define such a $\tilde{p}$, we follow the general approach set forth in Regazzini et al. (2003) and assume that $\tilde{p}$ arises as the normalization of a finite random measure $\tilde{\mu} = \sum_{j\geq 1} S_j \delta_{X_j}$ built as follows. Start by considering a *simple point process* Φ on $\mathbb{X}$, which will define the atoms of $\tilde{\mu}$. That is, Φ is a random counting measure of the kind

$$\Phi(B) = \sum_{j\geq 1} \delta_{X_j}(B), \quad B \in \mathcal{X}, \tag{4}$$

where $(X_j)_{j\geq 1}$, the *points* of the process, form a random countable subset of $\mathbb{X}$ such that $P(X_i = X_j) = 0$ for all $i \neq j$. We denote by $\mathbf{P}_\Phi$ the distribution of Φ. Then, we associate with each point of Φ independent *marks* $S_j \in \mathbb{R}_+$, such that each S_j has marginal distribution H, to define the *marked* point process $\Psi = \sum_{j\geq 1} \delta_{(X_j,S_j)}$ and set

$$\tilde{\mu}(B) = \int_{B\times\mathbb{R}_+} s\Psi(\mathrm{d}x\,\mathrm{d}s) = \sum_{j\geq 1} S_j \delta_{X_j}(B), \quad B \in \mathcal{X}. \tag{5}$$

Our construction of $\tilde{\mu}$ allows us to recover the random measures built from the independent finite point processes (also known as mixed binomial processes) analysed in Argiento and De Iorio (2022). In addition, equation (5) defines a random measure as a functional of a point process Ψ, similarly as in the case of CRMs. More importantly, our construction allows us to consider measures with *interacting* atoms that result in more informative priors in Bayesian mixture models, for instance by assuming that Φ is a repulsive Gibbs point process (so that the support points $(X_j)_{j\geq 1}$ are encouraged to be well separated) or a shot-noise Cox process (which, instead, results in the X_j's being randomly clustered together).

Defining $\tilde{p}$ via the normalization of $\tilde{\mu}$ in (5) requires some care to ensure that $\tilde{p}$ is well-defined. In particular, we assume that Φ in (4) is a finite point process, i.e. $\Phi(\mathbb{X}) < +\infty$ almost surely (a.s.), which clearly implies $\tilde{\mu}(\mathbb{X}) < +\infty$ a.s. Then, we set

$$\tilde{p}(B) = \begin{cases} \dfrac{\tilde{\mu}(B)}{\tilde{\mu}(\mathbb{X})} & \text{if } \tilde{\mu}(\mathbb{X}) > 0 \\ 0 & \text{if } \tilde{\mu}(\mathbb{X}) = 0 \end{cases}, \quad B \in \mathcal{X} \tag{6}$$

where, when $\tilde{p} \equiv 0$, we intend that the model does not generate any observation Y_i. A similar agreement is adopted in Zhou et al. (2017). If we do observe datapoints, then $P(\tilde{\mu}(\mathbb{X}) = 0 \mid Y_1, \ldots, Y_n) = 0$, as shown in Theorem 1 below. Hence, a posteriori, the usual assumption $\tilde{\mu}(\mathbb{X}) > 0$ a.s. is guaranteed. Alternatively, one can assume Φ is always non-empty, which also fits the definition in (6). However, most well-studied (repulsive) point processes in the literature assume that $P(\Phi(\mathbb{X}) = 0) > 0$. Moreover, we point out that the requirement $\tilde{p} = 0$ when $\tilde{\mu}(\mathbb{X}) = 0$ makes no issue in any of the proofs; see, e.g. the proof of online supplementary material, Proposition S8. We write $\tilde{p} \sim \mathrm{nRM}(\mathbf{P}_\Phi, H)$ to denote the distribution of the normalized random measure, while $\tilde{\mu} \sim \mathrm{RM}(\mathbf{P}_\Phi, H)$ stands for the distribution of the associated unnormalized random measure. With a little abuse of notation, we will also denote by $\mathbf{P}_\Psi$ the law of $\tilde{\mu}$, where Ψ is Φ with marks from H. Note that, as it is clear from the definition, $\tilde{p} \sim \mathrm{nRM}(\mathbf{P}_\Phi, H)$ does not lie in the class of *species sampling models* (Pitman, 1996), since, for instance, the support points are not i.i.d.

2.2 Background on point processes

We now give the necessary background material on point processes and introduce the two processes that will be discussed throughout the article. To keep the discussion light, we will present here the main mathematical objects in an intuitive and non-formal way. A more technical treatment is deferred to the Appendix.

As mentioned in Section 1, our analyses are based on Palm calculus. A central concept of our treatment is the Palm kernel, which may be regarded as an extension of regular conditional distributions to the case of point processes (Baccelli et al., 2020; Kallenberg, 2021). Informally speaking, the Palm version Φ_x of Φ at $x \in \mathbb{X}$ is the random measure Φ conditionally to the event 'Φ has an atom in x'. Its law is denoted by $\mathbf{P}^x_\Phi$. The latter is referred to as the Palm kernel of Φ at x. Since x is a *trivial* atom of Φ_x, we can safely discard it and consider the *reduced* Palm version of Φ, $\Phi^!_x := \Phi_x - \delta_x$ with the associated reduced Palm kernel denoted by $\mathbf{P}^x_{\Phi^!}$. The argument outlined above can be extended to the case of multiple pairwise different points $\boldsymbol{x} = (x_1, \ldots, x_k)$, leading to the kth Palm distribution $\{\mathbf{P}^{\boldsymbol{x}}_\Phi\}_{\boldsymbol{x}\in\mathbb{X}^k}$. Again, $\Phi_{\boldsymbol{x}}$ can be understood as the law of Φ conditional to Φ having atoms at $\{x_1, \ldots, x_k\}$ and removing the trivial atoms yields the reduced Palm distribution, that is, the law of

$$\Phi^{!}_{\boldsymbol{x}} := \Phi_{\boldsymbol{x}} - \sum_{j=1}^{k} \delta_{x_j}.$$

The other central quantities needed for our subsequent discussion are the so-called moment measures. The moment measure of order one, M_Φ, is defined as $M_\Phi(B) = \mathrm{E}[\Phi(B)]$ for all $B \in \mathcal{X}$. The *factorial moment measure* $M_{\Phi^{(k)}}$ of order k, instead, is the only measure such that

$$\mathrm{E}\left[\sum_{(x_1,\ldots,x_k)\in\Phi}^{\neq} g(x_1, \ldots, x_k) \right] = \int_{\mathbb{X}^k} g(x_1, \ldots, x_k) M_{\Phi^{(k)}}(\mathrm{d}x_1 \cdots \mathrm{d}x_k)$$

for any measurable function $g : \mathbb{X}^k \to \mathbb{R}_+$, where the symbol $\neq$ over the summation sign means that $(x_1, \ldots, x_k)$ are pairwise distinct. Observe that when $k = 1$, $M_{\Phi^{(1)}} \equiv M_\Phi$, i.e. the factorial moment measure coincides with the mean measure of Φ.

For a marked point process Ψ, with independent marks, the Palm distribution $\Psi_{\boldsymbol{x},\boldsymbol{s}}$, with $\boldsymbol{x} = (x_1, \ldots, x_k)$ and $\boldsymbol{s} = (s_1, \ldots, s_k)$, does not depend on $\boldsymbol{s}$. Moreover, $\Psi^{!}_{\boldsymbol{x},\boldsymbol{s}}$ has the same law of the point process obtained by considering $\Phi^{!}_{\boldsymbol{x}}$ and marking it with i.i.d. marks. See online supplementary material, Lemma S4 in Appendix S2. We write $\Psi^{!}_{\boldsymbol{x}}$ in place of $\Psi^{!}_{\boldsymbol{x},\boldsymbol{s}}$. Since $\Psi^{!}_{\boldsymbol{x}}$ is a marked point process, we can define a random probability measure on $\mathbb{X}$ as in the general construction (5). Specifically,

$$\tilde{\mu}^{!}_{\boldsymbol{x}}(A) := \int_{A\times\mathbb{R}_+} s\Psi^{!}_{\boldsymbol{x}}(\mathrm{d}s\,\mathrm{d}x). \tag{7}$$

We write $\tilde{\mu}^{!}_{\boldsymbol{x}} \sim \mathbf{P}^{\boldsymbol{x}}_{\Psi^{!}}$. Note that we can interpret $\mathbf{P}^{\boldsymbol{x}}_{\Psi^{!}}$ as the law of a random measure obtained as follows: (i) take the random measure $\tilde{\mu}$ as in (5), (ii) condition to $x_1, \ldots, x_k$ being atoms of $\tilde{\mu}$ and then (iii) remove $x_1, \ldots, x_k$ from the support.

We conclude this section by defining the two classes of point processes we will deal with in the rest of the article. In the online supplementary material, we also report the case of the general Gibbs point and Poisson processes.

Example 1 (Determinantal point processes). Determinantal point processes (DPPs Hough et al., 2009; Lavancier et al., 2015; Macchi, 1975) are a class of repulsive point processes. We will restrict our focus to finite DPPs defined as follows. Let ω be a finite measure on $\mathbb{X}$. Usually, it is the case that $\mathbb{X}$ is a compact subset of $\mathbb{R}^q$ and ω is the Lebesgue measure, but one could consider more general spaces endowed, e.g. with a Gaussian measure. Consider a complex-valued covariance function $K : \mathbb{X} \times \mathbb{X} \to \mathbb{C}$, such that $\int_{\mathbb{X}} K(x, x)\omega(\mathrm{d}x) < +\infty$, with spectral representation

$$K(x, y) = \sum_{h\geq 1} \lambda_h \varphi_h(x)\overline{\varphi_h(y)}, \qquad x, y \in X \tag{8}$$

where $(\varphi_h)_{h\geq 1}$ form an orthonormal basis for the space $L^2(\mathbb{X}; \omega)$ of complex-valued functions, $\lambda_h \geq 0$ with $\sum_{h\geq 1} \lambda_h < +\infty$. By Mercer's theorem, the series on the right-hand side of (8) converges absolutely and uniformly on $\mathbb{X}^2$. If $0 \leq \lambda_h \leq 1$, K defines a DPP Φ on $\mathbb{X}$ (Macchi, 1975; Soshnikov, 2000). In particular, the kth factorial moment measure $M_{\Phi^{(k)}}$ admits a density with respect to the product measure ω^k satisfying

$$M_{\Phi^{(k)}}(\mathrm{d}x_1 \cdots \mathrm{d}x_k) = \det\left\{K(x_i, x_j)\right\}_{i,j=1}^{n} \omega(\mathrm{d}x_1) \cdots \omega(\mathrm{d}x_k),$$

where $\{K(x_i, x_j)\}_{i,j=1}^{n}$ is the $k \times k$ matrix with (i, j)th entry $K(x_i, x_j)$.

We state below a few remarks on the definition of Determinantal point processes in Example 1.

Remark 1 (Number of points in a DPP). The condition $\int K(x,x)\omega(\mathrm{d}x) < +\infty$ is equivalent to $M_\Phi(\mathbb{X}) = \mathrm{E}[\Phi(\mathbb{X})] < +\infty$, which clearly implies $\Phi(\mathbb{X}) < +\infty$ almost surely. That is, we deal only with finite point processes. This is needed to ensure that $\tilde{p}$ in (6) is well-defined.

Remark 2 (Projection DPPs). Let $m \in \mathbb{N}$ (the set of natural numbers); if the kernel K is such that $\lambda_h = 1$ for $h \leq m$ and $\lambda_h = 0$ for $h > m$, then the resulting DPP is a *projection* DPP. In particular, it consists of exactly m points, with a joint density with respect to ω^m given by

$$f_\Phi(v) \propto \det\big\{K(x,y)\big\}_{x,y\in\{x_1,\ldots,x_m\}}, \quad v = (x_1, \ldots, x_m).$$

Remark 3 (DPP densities). If the kernel K is such that $\lambda_h < 1$ for all h, then Φ is absolutely continuous with respect to the Poisson process on $\mathbb{X}$ with intensity measure ω, with density

$$f_\Phi(v) = \mathrm{e}^{\omega(\mathbb{X})-D}\det\big\{C(x,y)\big\}_{x,y\in v},$$

where $D := -\sum_{h\geq 1}\log(1-\lambda_h)$ and $C(x,y) := \sum_{h\geq 1}\frac{\lambda_h}{1-\lambda_h}\varphi_h(x)\overline{\varphi_h(y)}$ for $x, y \in \mathbb{X}$. Such an expression generalizes the one given by Lavancier et al. (2015), which deals with the case of $\mathbb{X}$ compact set in $\mathbb{R}^q$ and ω the Lebesgue measure. The proof of such an expression is derived straightforwardly by adapting results from Shirai and Takahashi (2003), and it is provided in online supplementary material, Appendix S7.

Remark 4 (Non-finite measures on $\mathbb{X}$). If ω is not finite, then Mercer's theorem does not apply and the spectral representation (8) might not hold. In such cases, the existence of Φ requires further conditions on K, which are rather technical and beyond the purpose of this article. See, e.g. Ferreira and Menegatto (2009) for further details.

Example 2 (Shot-Noise Cox Processes). Introduced in Møller (2003), shot-noise Cox processes (SNCPs) are a prominent example of Cox processes (Cox, 1955). For simplicity, we assume that $\mathbb{X} \subset \mathbb{R}^q$. A SNCP Φ is defined as

$$\Phi \mid \Lambda \sim \mathrm{PP}(\omega_\Lambda), \quad \omega_\Lambda(\mathrm{d}x) = \gamma\big(\textstyle\int_{\mathbb{X}} k_\alpha(x-v)\Lambda(\mathrm{d}v)\big)\mathrm{d}x, \quad \Lambda \sim \mathrm{PP}(\omega) \qquad (9)$$

where $\mathrm{PP}(\omega_\Lambda)$ and $\mathrm{PP}(\omega)$ denote Poisson processes with intensity ω_Λ and ω, respectively. Here, k_α is a probability density with respect to the q-dimensional Lebesgue measure and $\gamma > 0$. In particular, by assuming that $x \mapsto k_\alpha(x-v)$ is continuous for any v, we get that the point process Φ is simple. We write $\Phi \sim \mathrm{SNCP}(\gamma, k_\alpha, \omega)$, and refer to ω as the base intensity of the process Φ. We also introduce the following notation that will be useful in later examples, $\eta(x_1, \ldots, x_l) := \int \prod_{i=1}^{l} k_\alpha(x_i - v)\omega(\mathrm{d}v)$. Observe that if $\omega(\mathbb{X}) = 1$, η takes the interpretation of the marginal distribution of a model of the kind $x_1, \ldots, x_k \mid v \overset{\mathrm{iid}}{\sim} k_\alpha(\,\cdot - v)$, $v \sim \omega$.

Thanks to the colouring theorem of Poisson processes (Kingman, 1992), (9) is equivalent to the following model:

$$\Phi \mid \Lambda = \sum_{h=1}^{n(\Lambda)} \Phi_h, \quad \Phi_h \sim \mathrm{PP}(\gamma k_a(x-\zeta_h)\mathrm{d}x), \quad \Lambda \sim \mathrm{PP}(\omega), \tag{10}$$

where $\Lambda = \sum_{h=1}^{n(\Lambda)} \delta_{\zeta_h}$. Then, a shot-noise Cox process $\Phi = \sum_{j\geq 1} \delta_{X_j}$ admits the double-summation formulation $\Phi = \sum_{h=1}^{n(\Lambda)} \sum_{j\geq 1} \delta_{X_{h,j}}$. Clearly, it is possible to map the latter representation to the former by ordering the $X_{h,j}$'s arbitrarily (thanks to exchangeability), e.g. via their lexicographical order. From (10), it is also clear why shot-noise Cox processes are often referred to as *cluster processes* (see, e.g. Daley & Vere-Jones, 2003). Indeed, introducing an indicator variables T_j for each X_j, such that X_j belongs to Φ_h if and only if $T_j = h$ (i.e. $\Phi_h(X_j) = 1$, or, equivalently, $X_j = X_{h,j'}$ for some j'), it is possible to group together the different support points of Φ. In this context, instead of referring to the Φ_h's as *clusters*, we denote them as *groups* to avoid misunderstandings with the usual BNP terminology.

3 Bayesian analysis of normalized random measures

This section contains the most relevant results of the article: posterior, marginal, and predictive distributions for the statistical model (3), when $\tilde{p} \sim \mathrm{nRM}(\mathbf{P}_\Phi; H)$. All the results are available in closed form, and they constitute the backbone to develop computational procedures for the mixture models in Section 4.

All our proofs are based on the Laplace functional of the random measure $\tilde{\mu}$, defined as

$$L_{\tilde{\mu}}(f) = \mathrm{E}\left[\exp\left(-\textstyle\int_{\mathbb{X}} f(x)\tilde{\mu}(\mathrm{d}x)\right)\right]$$

for any bounded non-negative function $f : \mathbb{X} \to \mathbb{R}_+$. We now introduce useful notation and an additional variable U_n, which helps us to describe Bayesian inference for model (3). The joint distribution of $(\boldsymbol{Y}, \tilde{\mu})$ is given by

$$P(\boldsymbol{Y} \in \mathrm{d}\boldsymbol{y}, \tilde{\mu} \in \mathrm{d}\mu) = \frac{1}{\mu(\mathbb{X})^n} \prod_{i=1}^{n} \mu(\mathrm{d}y_j)\mathbf{P}_\Psi(\mathrm{d}\mu), \tag{11}$$

where $\mathbf{P}_\Psi$ is the law of $\tilde{\mu}$. Posterior inference in (11) is complex due to the term $\mu(\mathbb{X})^{-n}$, which does not allow to separate the atoms of $\tilde{\mu}$ that have been observed in the sample $\boldsymbol{Y}$ from the atoms that have not been observed. A popular workaround to this issue, originally noted in James et al. (2009), is to introduce an auxiliary variable $U_n \mid \tilde{\mu} \sim \mathrm{Gamma}(n, \tilde{\mu}(\mathbb{X}))$ and, by a suitable augmentation of the underlying probability space, we consider the joint distribution of $(\boldsymbol{Y}, U_n, \tilde{\mu})$

$$P(\boldsymbol{Y} \in \mathrm{d}\boldsymbol{y}, U_n \in \mathrm{d}u, \tilde{\mu} \in \mathrm{d}\mu) = \frac{u^{n-1}}{\Gamma(n)} \mathrm{e}^{-\mu(\mathbb{X})u}\mathrm{d}u \prod_{i=1}^{n} \mu(\mathrm{d}y_j)\mathbf{P}_\Psi(\mathrm{d}\mu), \tag{12}$$

where we have gotten rid of the term $\mu(\mathbb{X})^{-n}$. By considering (12), we can study the conditional law of $\tilde{\mu}$ given $(\boldsymbol{Y}, U_n)$. Together with the posterior of $U_n \mid \boldsymbol{Y}$, this allows us to obtain a disintegration of the posterior of $\tilde{\mu}$ in Theorem 1 below, which is amenable to analytical and numerical computations.

Since $\tilde{\mu}$ is almost surely discrete, with positive probability, there will be ties within the sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. For this reason, $\boldsymbol{Y}$ is equivalently characterized by the couple $(\boldsymbol{Y}^*, \tilde{\pi})$, where $\boldsymbol{Y}^* = (Y_1^*, \ldots, Y_{K_n}^*)$ is the vector of distinct values and $\tilde{\pi}$ is the random partition of $[n] := \{1, \ldots, n\}$ of size K_n, identified by the equivalence relation $i \sim j$ if and only if $Y_i = Y_j$. Given $K_n = k$, we indicate by $\boldsymbol{n} = (n_1, \ldots, n_k)$ the vector of counts, i.e. n_j is the cardinality of the set $\{i \in [n] : Y_i = Y_j^*\}$, as $j = 1, \ldots, k$. As a consequence, we may write

$$P(\boldsymbol{Y} \in \mathrm{d}\boldsymbol{y},\, U_n \in \mathrm{d}u,\, \tilde{\mu} \in \mathrm{d}\mu) = \frac{u^{n-1}}{\Gamma(n)} \mathrm{e}^{-\mu(\mathbb{X})u} \prod_{j=1}^{k} \mu(\mathrm{d}y_j^*)^{n_j} \mathbf{P}_\Psi(\mathrm{d}\mu).$$

The augmentation of (11) through U_n is made possible since the marginal distribution of U_n exists and has a density, with respect to the Lebesgue measure, that is

$$f_{U_n}(u) = \frac{u^{n-1}}{\Gamma(n)} \int_0^{+\infty} t^n \mathrm{e}^{-tu} f_\mu(t) \mathrm{d}t,$$

where f_μ is the density or the random variable $\tilde{\mu}(\mathbb{X})$.

3.1 Posterior characterization

We first characterize the posterior distribution of $\tilde{p}$ in model (3) when a priori $\tilde{p} \sim \mathrm{nRM}(\mathbf{P}_\Phi; H)$. Since $\tilde{p}$ is obtained by the normalization of the random measure $\tilde{\mu}$, it is sufficient to describe the posterior distribution of $\tilde{\mu}$, which is provided in the following theorem.

Theorem 1 Assume that $H(\mathrm{d}s) = h(s)\mathrm{d}s$ where $\mathrm{d}s$ is the Lebesgue measure. The distribution of $\tilde{\mu}$ conditionally on $\boldsymbol{Y} = \boldsymbol{y}$ and $U_n = u$ is equal to the distribution of

$$\sum_{j=1}^{k} S_j^* \delta_{y_j^*} + \tilde{\mu}', \tag{13}$$

where:

(i) $\boldsymbol{S}^* := (S_1^*, \ldots, S_k^*)$ is a vector of independent random variables, with density

$$f_{S_j^*}(s) \propto \mathrm{e}^{-us} s^{n_j} h(s) \quad \text{for } j = 1, \ldots, k;$$

(ii) $\tilde{\mu}'$ is a random measure with Laplace functional

$$\mathrm{E}\left[\exp \textstyle\int_{\mathbb{X}} -f(z)\tilde{\mu}'(\mathrm{d}z)\right] = \frac{\mathrm{E}\left[\exp\left\{-\int_{\mathbb{X}} (f(z)+u)\tilde{\mu}^!_{\boldsymbol{y}^*}(\mathrm{d}z)\right\}\right]}{\mathrm{E}\left[\exp\left\{-\int_{\mathbb{X}} u\tilde{\mu}^!_{\boldsymbol{y}^*}(\mathrm{d}z)\right\}\right]}, \tag{14}$$

and $\tilde{\mu}^!_{\boldsymbol{y}^*}$ is as in (7) for $\boldsymbol{x} = \boldsymbol{y}^*$.

In (13) above, $\boldsymbol{S}^*$ and $\tilde{\mu}'$ are independent. Moreover, the conditional distribution of U_n, given $\boldsymbol{Y} = \boldsymbol{y}$, has a density with respect to the Lebesgue measure, which satisfies

$$f_{U_n \mid \boldsymbol{Y}}(u) \propto u^{n-1} \mathrm{E}\left[\mathrm{e}^{-u\tilde{\mu}^!_{\boldsymbol{y}^*}(\mathbb{X})}\right] \prod_{j=1}^{k} \kappa(u, n_j), \quad u > 0, \tag{15}$$

with $\kappa(u, n) := \int_{\mathbb{R}_+} \mathrm{e}^{-us} s^n H(\mathrm{d}s)$.

First, we point out that the Laplace functional of $\tilde{\mu}'$ is an exponential tilting of the one of $\tilde{\mu}^!_{\boldsymbol{y}^*}$, that is to say the distribution of $\tilde{\mu}'$ is absolutely continuous with respect to the distribution of $\tilde{\mu}^!_{\boldsymbol{y}^*}$, with density given by $f_{\tilde{\mu}'}(\mu) \propto \mathrm{e}^{-u\mu(\mathbb{X})}$. Theorem 1 resembles the posterior characterization of $\tilde{p}$ in the case of NRMIs (James et al., 2009) and normalized IFPPs (Argiento & De Iorio, 2022). Interestingly, unlike in the previous cases, here the law of $\tilde{\mu}'$ depends on the observed $\boldsymbol{y}^*$. Note that the expression (14) is obtained without any specific assumption on the law of the point process Φ. Specific

expressions for the posterior distribution of $\tilde{\mu}$ are obtained when considering particular classes of point processes, as we showcase in the following examples. For the sequel, we remind that $\psi(u) := \mathrm{E}[\mathrm{e}^{-uS}]$ is the Laplace transform of a random variable S, with distribution H, and we also define

$$f_{S'}(s; u) := \frac{\mathrm{e}^{-su}h(s)}{\psi(u)} = \frac{\mathrm{e}^{-su}h(s)}{\int_{\mathbb{R}^+}\mathrm{e}^{-su}h(s)\mathrm{d}s} \tag{16}$$

to be the density of the exponential tilting of S.

We now specialize Theorem 1 to our examples, whose posterior distribution was not previously available in the literature.

Example 3 (Determinantal point process). Assume that Φ is a DPP on $\mathbb{X}$ as defined in Example 1, whose kernel K has eigenvalues λ_j in (8) all strictly smaller than one. Then, the random measure $\tilde{\mu}'$ in Theorem 1 is such that

$$\tilde{\mu}' \overset{d}{=} \sum_{j\geq 1} S'_j \delta_{X'_j}$$

where $\Psi' := \sum_{j\geq 1}\delta_{(X'_j,S'_j)}$ is a marked point process whose unmarked version $\Phi' := \sum_{j\geq 1}\delta_{X'_j}$ is a DPP with density with respect to PP(ω) given by $f_{\Phi'}(v) \propto \det\{C'(x_i, x_j)\}_{(x_i,x_j)\in v}$ and the marks S'_j are i.i.d. with distribution (16). The operator C' is defined as

$$C'(x, y) = \psi(u)\left[C(x, y) - \sum_{i,j=1}^{k}\left(C_{\boldsymbol{y}^*}^{-1}\right)_{i,j} C(x, y_i^*)C(y, y_j^*)\right], \quad x, y \in R,$$

where $C_{\boldsymbol{y}^*}$ is the $k \times k$ matrix with entries $C(y_i^*, y_j^*)$. See online supplementary material, Theorem S15 in Section S7 for a proof. Note that for simulation purposes, it is useful to know or approximate the eigendecomposition of the kernel C':

$$C'(x, y) = \sum_j \gamma_j \varphi'_j(x)\overline{\varphi'}_j(y).$$

The γ_j's and φ'_j's can be approximated numerically using the Nyström method (S. Sun et al., 2015).

In order to specialize Theorem 1 to the Shot-Noise Cox process, we introduce $\boldsymbol{T}' = (T'_1, \ldots, T'_k)$, the indicator variables describing a partition of $[k] = \{1, \ldots, k\}$, k being the number of distinct values in $\boldsymbol{y}$, namely $T'_j = l$ if and only if j belongs to the lth element of partition introduced in online supplementary material, Lemma S16. In practice, T'_j, $j = 1, \ldots, k$, describe the points of the Poisson process Λ associated to the unique value $y_1^*, \ldots, y_k^*$.

Example 4 (Shot-Noise Cox process). Assume that $\Phi \sim \mathrm{SNCP}(\gamma, k_\alpha, \omega)$. Then, the random measure $\tilde{\mu}'$ in (13) has a mixture representation. Let $\boldsymbol{T}' = (T'_1, \ldots, T'_k)$ be latent indicator variables introduced above and denote by $|\boldsymbol{T}'|$ the number of distinct values in $\boldsymbol{T}'$. We have

$$\tilde{\mu}' \mid \boldsymbol{T}' \overset{\mathrm{d}}{=} \tilde{\mu}_0 + \sum_{\ell=1}^{|\boldsymbol{T}'|}\tilde{\mu}_\ell,$$

where $\tilde{\mu}_0 = \sum_{j\geq 1}\tilde{S}_j\delta_{\tilde{X}_j}$ and $\tilde{\mu}_\ell = \sum_{j\geq 1}\tilde{S}_{\ell,j}\delta_{\tilde{X}_{\ell,j}}$, as $\ell = 1, \ldots, |\boldsymbol{T}'|$, are independent, and all their weights are i.i.d. with distribution (16). The unmarked point

processes $\Phi_0 = \sum_{j\geq 1} \delta_{\tilde{X}_j}$ is distributed as $\Phi_0 \sim \text{SNCP}(\gamma\psi(u), k_a, \mathrm{e}^{-\gamma(1-\psi(u))}\omega)$. Each $\Phi_\ell = \sum_{j\geq 1} \delta_{\tilde{X}_{\ell,j}}$ is a Poisson point process with random intensity $\omega_\ell(\mathrm{d}x) = \gamma\psi(u)k_a(\zeta_\ell - x)\mathrm{d}x$, where

$$\zeta_\ell \sim f_{\zeta_\ell}(v)\mathrm{d}v \propto \prod_{j:\, T'_j=\ell} k_a(y^*_j - v)\omega(\mathrm{d}v).$$

Finally, $P(\boldsymbol{T}' = \boldsymbol{t}') \propto \exp\{\gamma|\boldsymbol{t}'|(\psi(u)-1)\}\prod_{\ell=1}^{|\boldsymbol{t}'|}\eta((y^*_j : t'_j = \ell))$. For the proof and the formal statement, see online supplementary material, Theorem S19 in Appendix H.2.

3.2 Marginal and predictive distributions

In the present section, we describe the marginal and predictive distributions for a sample from $\tilde{p}$ as in (3), when the prior distribution is defined as $\tilde{p} \sim \text{nRM}(\mathbf{P}_\Phi; H)$. As mentioned at the beginning of the present section, the almost sure discreteness of $\tilde{p}$ entails that the sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is equivalently characterized by $(\boldsymbol{Y}^*, \tilde{\pi})$, where $\boldsymbol{Y}^* = (Y^*_1, \ldots, Y^*_k)$ is the vector of distinct values and $\tilde{\pi}$ is the random partition of $[n] := \{1, \ldots, n\}$ of size K_n, identified by the equivalence relation $i \sim j$ if and only if $Y_i = Y_j$.

Theorem 2 The marginal distribution of a sample $\boldsymbol{Y}$ with $K_n = k$ distinct values $\boldsymbol{Y}^*$ and associated counts $n_1, \ldots, n_k$ is

$$P(\boldsymbol{Y} \in \mathrm{d}\boldsymbol{y}) = P(\boldsymbol{Y}^* \in \mathrm{d}\boldsymbol{y}^*, \tilde{\pi} = \pi) = \int_{\mathbb{R}_+} \frac{u^{n-1}}{\Gamma(n)} \mathrm{E}\left[\mathrm{e}^{-u\tilde{\mu}^!_{\boldsymbol{y}^*}(\mathbb{X})}\right] \prod_{j=1}^{k} \kappa(u, n_j)\mathrm{d}u\, M_{\Phi^{(k)}}(\mathrm{d}\boldsymbol{y}^*),$$

where $\tilde{\mu}^!_{\boldsymbol{y}^*}$ is defined in (7) and $M_{\Phi^{(k)}}$ is the kth factorial moment measure of Φ defined in Section 2.2.

Theorem 2 resembles the marginal characterization of a sample from an NRMI (James et al., 2009) and a normalized IFPP (Argiento & De Iorio, 2022). However, there are key differences: in the case of NRMIs, $\tilde{\mu}^!_{\boldsymbol{y}^*} \overset{d}{=} \tilde{\mu}$ (i.e. it is distributed as the prior and, in particular, it does not depend on $\boldsymbol{y}^*$) so that $\mathrm{E}[\exp(-u\tilde{\mu}^!_{\boldsymbol{y}^*}(\mathbb{X}))]$ can be evaluated explicitly via the Lévy-Khintchine representation; in the case of IFPPs, $\tilde{\mu}^!_{\boldsymbol{y}^*}$ depends on $\boldsymbol{y}^*$ only through its cardinality and the latter expectation can be computed analytically. In both cases, the kth factorial moment measure has a product form: $M_{\Phi^{(k)}}(\mathrm{d}\boldsymbol{y}^*) = \prod_{j=1}^{k} \omega(\mathrm{d}y^*_j)$, where ω is a finite measure on $\mathbb{X}$. This allows for the marginalization of the $\boldsymbol{Y}^*$'s, also obtaining an analytical expression for the prior induced on the random partition $\tilde{\pi}$, the so-called exchangeable partition probability function. In our case, such a marginalization is possible only from a formal point of view, and the resulting expression is generally intractable. We now specialize Theorem 2 to important examples.

Example 5 (Determinantal point process, cont'd). If Φ is a DPP, for all $x_1, \ldots, x_n$ such that $K(x_i, x_j) > 0$, $\Phi^!_{\boldsymbol{x}}$ is a DPP with kernel

$$K^!_{\boldsymbol{x}}(x, y) = K(x, y) - \sum_{i,j=1}^{k} (K_{\boldsymbol{x}}^{-1})K(x, x_i)K(y, x_j)$$

where $K_{\boldsymbol{x}}$ is the $n \times n$ matrix with entries $K(x_i, x_j)$ (Lavancier et al., 2015). Denote by $\lambda^{\boldsymbol{x}}_j$ the eigenvalues of $K^!_{\boldsymbol{x}}$. Then, $\Phi^!_{\boldsymbol{x}}(\mathbb{X})$ is a Poisson-Binomial random variable with parameter $(\lambda^{\boldsymbol{x}}_j)_{j\geq 1}$, which entails $\mathrm{E}[\exp(-u\tilde{\mu}^!_{\boldsymbol{y}^*}(\mathbb{X}))] = \prod_{j\geq 1}(1 - \lambda^{x^*_j}_j + \lambda^{x^*_j}\psi(u))$. Moreover, the factorial moment measure equals

$$M_{\Phi^{(k)}}(\mathrm{d}\boldsymbol{y}^*) = \det\{K(y_i^*, y_j^*)\}_{i,j=1}^{k}\omega^k(\mathrm{d}\boldsymbol{y}^*).$$

As an alternative to numerically computing the $\lambda_j^{\boldsymbol{x}}$'s, Beraha, Camerlenghi et al. (2025) proposed using Le Cam's approximation (Steele, 1994), i.e. approximating $\Phi^!_{\boldsymbol{x}}(\mathbb{X})$ with a Poisson random variable of mean $\lambda_{\boldsymbol{x}^*} := \int_{\mathbb{X}} K^!_{\boldsymbol{x}^*}(x, x)\mathrm{d}x$, showing that the approximation error is small. In such a case $\mathrm{E}[\exp(-u\tilde{\mu}^!_{\boldsymbol{y}^*}(\mathbb{X}))] \approx \exp(\lambda_{\boldsymbol{x}^*}(\psi(u) - 1))$.

Example 6 (Shot-Noise Cox process, cont'd). If Φ is a SNCP, then $\Phi^!_{\boldsymbol{x}}$ can be written as in online supplementary material, Lemma S17 of Appendix S8, while the moment measure is described in online supplementary material, Lemma S16, which lead to the formula of the marginal distribution. However, a more transparent expression is obtained by considering also the variables $\boldsymbol{T}'$ introduced in Example 4. Indeed, we have

$$P(\boldsymbol{Y} \in \mathrm{d}\boldsymbol{y}, \boldsymbol{T}' = \boldsymbol{t}' \mid U_n = u) \propto \gamma^k \prod_{j=1}^{k} \kappa(u, n_j) \prod_{\ell=1}^{|\boldsymbol{t}'|} \eta(\boldsymbol{y}^*_\ell)\mathrm{e}^{\gamma|\boldsymbol{t}'|(\psi(u)-1)}\mathrm{d}\boldsymbol{y}^*,$$

where $\boldsymbol{y}^*_\ell = (y_j^* : t'_j = \ell)$.

We now focus on the predictive distribution, i.e. the distribution of Y_{n+1}, conditionally to the observable sample $\boldsymbol{Y}$.

Theorem 3 Assume that the factorial moment measure $M_{\Phi^{(k)}}$ is absolutely continuous with respect to the product measure P_0^k, where P_0 is a non-atomic probability on $(\mathbb{X}, \mathcal{X})$ and let m_{Φ^k} be the associated Radon-Nikodym derivative. Let U_n be a random variable with density $f_{U_n \mid \boldsymbol{Y}}$ as in (15). Then, conditionally on $\boldsymbol{Y} = \boldsymbol{y}$ and $U_n = u$, the predictive distribution of Y_{n+1} is such that

$$\begin{aligned} P(Y_{n+1} \in A \mid \boldsymbol{Y} = \boldsymbol{y}, U_n = u) \propto & \sum_{j=1}^{k} \frac{\kappa(u, n_j + 1)}{\kappa(u, n_j)} \delta_{y_j^*}(A) \\ & + \int_A \kappa(u, 1) \frac{\mathrm{E}\left[\mathrm{e}^{-u\tilde{\mu}^!_{(\boldsymbol{y}^*, y)}(\mathbb{X})}\right]}{\mathrm{E}\left[\mathrm{e}^{-u\tilde{\mu}^!_{\boldsymbol{y}^*}(\mathbb{X})}\right]} \frac{m_{\Phi^{k+1}}(\boldsymbol{y}^*, y)}{m_{\Phi^k}(\boldsymbol{y}^*)} P_0(\mathrm{d}y), \end{aligned} \tag{17}$$

where $A \in \mathcal{X}$.

The predictive distribution from Theorem 3 can be interpreted using a generalized Chinese restaurant metaphor. As in the traditional Chinese Restaurant process, customers sitting at the same table eat the same dish, whereas the same dish cannot be served at different tables. The first customer arrives at the restaurant and sits at the first table, eating dish $Y_1 = Y_1^*$ such that $P(Y_1^* \in \mathrm{d}y) \propto M_\Phi(\mathrm{d}y)$. The second customer sits at the same table as the first customer with probability proportional to $\kappa(U_1, 2)/\kappa(U_1, 1)$, or sits at a new table and eats a new dish with probability proportional to

$$P(Y_2 \in \mathbb{X} \setminus \{y_1^*\} \mid Y_1^* = y_1^*) \propto \frac{\kappa(U_1, 1)}{\mathrm{E}\left[\mathrm{e}^{-U_1\tilde{\mu}^!_{y_1^*}(\mathbb{X})}\right] m_\Phi(y_1^*)} \int_{\mathbb{X}} \mathrm{E}\left[\mathrm{e}^{-U_1\tilde{\mu}^!_{(y_1^*, y)}(\mathbb{X})}\right] m_{\Phi^2}(y_1^*, y) P_0(\mathrm{d}y),$$

where $U_1 \sim f_{U_1 \mid y_1}$ is defined in (15). The distribution of the new dish is proportional to

$$P(Y_2 \in \mathrm{d}y | Y_1^* = y_1^*) \propto \mathrm{E}\left[\mathrm{e}^{-U_1 \tilde{\mu}^1_{(y_1^*, y)}(\mathbb{X})}\right] m_{\Phi^2}(y_1^*, y) P_0(\mathrm{d}y).$$

The metaphor proceeds as usual for a new customer entering the restaurant at time $n+1$: they either sit at one of the $K_n = k$ previously occupied tables, with probability proportional to $\kappa(U_n, n_j + 1)/\kappa(U_n, n_j), j = 1, \ldots, k$ or sit at a new table eating dish $Y^*_{k+1} = y$ with probability proportional to

$$\kappa(U_n, 1) \frac{\mathrm{E}\left[\mathrm{e}^{-U_n \tilde{\mu}^1_{(\boldsymbol{y}^*, y)}(\mathbb{X})}\right]}{\mathrm{E}\left[\mathrm{e}^{-U_n \tilde{\mu}^1_{\boldsymbol{y}^*}(\mathbb{X})}\right]} \frac{m_{\Phi^{k+1}}(\boldsymbol{y}^*, y)}{m_{\Phi^k}(\boldsymbol{y}^*)} P_0(\mathrm{d}y).$$

Observe that the original formulation of the Chinese restaurant process (Aldous, 1985) is stated only in terms of the seating arrangements in a restaurant with infinite tables, i.e. the random partition. Then, K_n 'dishes' are sampled i.i.d. from a distribution on $\mathbb{X}$ and assigned to each of the occupied tables. Instead, in our process, the dishes are not i.i.d. and are chosen as soon as a new table is occupied. In particular, note that the probability of occupying a new table depends on n, K_n, and the unique values $\boldsymbol{Y}^*$. This also sheds light on how our model differs from *Gibbs type priors* (De Blasi et al., 2013), where the probability of occupying a new table depends only on n and K_n.

We now specialize the predictive distribution in some examples of interest.

Example 7 ((Determinantal point process, cont'd)). Let K_x be defined as in Example 5. Following Example 5, the ratio of the expected values in (17) equals

$$\prod_{j \geq 1} \frac{1 - \lambda_j^{x,y} + \lambda_j^{x,y} \psi(u)}{1 - \lambda_j^{x} + \lambda_j^{x} \psi(u)},$$

which can be numerically computed via Le Cam's approximation. Moreover, by the Schur determinant identity, denoting by $K_{y,y} := K(y, y)$, $K_{\boldsymbol{y}^*,y} := (K(y_1^*, y), \ldots, K(y_k^*, y))^\top$, and by $K_{\boldsymbol{y}^*}$ the $k \times k$ matrix with entries $(K(y_i^*, y_j^*))_{i,j=1}^k$, we have

$$\frac{m_{\Phi^{k+1}}(\boldsymbol{y}^*, y)}{m_{\Phi^k}(\boldsymbol{y}^*)} P_0(\mathrm{d}y) = \left(K_{y,y} - K^\top_{\boldsymbol{y}^*,y} K^{-1}_{\boldsymbol{y}^*,\boldsymbol{y}^*} K_{\boldsymbol{y}^*,y}\right) \mathrm{d}y.$$

Example 8 (Shot-Noise Cox process, cont'd.). As in Example 6, we consider the auxiliary latent vector $\boldsymbol{T}'$, and its corresponding realization $\boldsymbol{t}'$, to describe the predictive distribution. In particular, as discussed in online supplementary material, Appendix S8.3, we can think of a restaurant metaphor whereby tables are divided into rooms, such that $t'_j = \ell$ if table j, serving dish y_j^*, is located in room ℓ. If, after n customers have entered the restaurant, having occupied k tables in $|\boldsymbol{T}'| = |\boldsymbol{t}'|$ rooms, customer $n+1$ can do one of the following choices.

(i) Customer $n+1$ sits at one of the previously occupied tables, say table j, with probability proportional to $\kappa(u, n_j + 1)/\kappa(u, n_j)$.
(ii) Customer $n+1$ sits at a new table in room $\ell \in \{1, \ldots, |\boldsymbol{t}'_k|\}$ with probability proportional to $\kappa(u, 1) \int_{\mathbb{X}} \eta(\boldsymbol{y}^*_\ell, y)/\eta(\boldsymbol{y}^*_\ell) \mathrm{d}y$, eating a dish y generated from a probability density proportional to $\eta(\boldsymbol{y}^*_\ell, y)/\eta(\boldsymbol{y}^*_\ell) \mathrm{d}y$.

(iii) Customer $n+1$ enters an empty room (sitting, a fortiori, in a new table) with probability proportional to $\kappa(u, 1)e^{\gamma(\psi(u)-1))}\int_{\mathbb{X}}\eta(y)\mathrm{d}y$. The customer eats a new dish y generated from a probability density function proportional to $\eta(y)\mathrm{d}y$.

The process described above shares similarity with the Chinese restaurant franchise metaphor of Teh et al. (2006) in the case of a single restaurant (see Camerlenghi et al. (2018), further details), the main difference being that if a new table is chosen, the new dish will be different from previously eaten dishes almost surely.

3.3 Distribution of the distinct values

From the marginal characterization in Theorem 2, it is easy to derive the joint distribution of $(K_n, \boldsymbol{Y}^*)$, that is, the joint distribution of the number of the distinct values and their position in a sample of size n.

Proposition 4 Given a set of distinct points $\boldsymbol{y}^* = (y_1^*, \ldots, y_k^*)$, let $(q_r)_{r\geq 0}$ be the probability mass function of the number of points in $\Phi^!_{\boldsymbol{y}^*}$, i.e. $q_r := P(\Phi^!_{\boldsymbol{y}^*}(\mathbb{X}) = r)$. Define

$$V(n_1, \ldots, n_k; r) := \int_{\mathbb{R}_+} \frac{u^{n-1}}{\Gamma(n)}\psi(u)^r \prod_{j=1}^{k}\kappa(u, n_j)\mathrm{d}u.$$

Then, the joint distribution of K_n and $\boldsymbol{Y}^*$ equals

$$P(K_n = k, \boldsymbol{Y}^* \in \mathrm{d}\boldsymbol{y}^*) = \frac{1}{k!}\sum_{r=0}^{\infty}\left(\sum_{(n_1,\ldots,n_k)\in\mathbb{S}_n^k}\binom{n}{n_1\cdots n_k}V(n_1, \ldots, n_k; r)\right) q_r\, M_{\Phi^{(k)}}(\mathrm{d}\boldsymbol{y}^*),$$

where $\mathbb{S}_n^k$ denotes the set of k-compositions, i.e. $\mathbb{S}_n^k = \{(n_1, \ldots, n_k): n_i \geq 1, n_1 + \cdots + n_k = n\} \subset \mathbb{N}^k$.

Note that the joint distribution of K_n and $\boldsymbol{Y}^*$ does not factorize. Moreover, it is not easy to marginalize $\boldsymbol{Y}^*$ since each q_r in general depends on $\boldsymbol{Y}^*$. Now, it is interesting to specialize Proposition 4 for a special choice of the distribution of the jumps S_j.

Corollary 5 Under the same assumptions as in Proposition 4, if $S_j \overset{\text{iid}}{\sim} \text{Gamma}(a, 1)$, then

$$P(K_n = k, \boldsymbol{Y}^* \in \mathrm{d}\boldsymbol{y}^*) = (-1)^n\mathcal{C}(n, k; -a)\left(\sum_{r\geq 0} q_r \frac{\Gamma(k+r)a)}{\Gamma((k+r)a+n)}\right)M_{\Phi^{(k)}}(\mathrm{d}\boldsymbol{y}^*)$$

where for any non-negative integer $n \geq 0$, $0 \leq k \leq n$, $\mathcal{C}(n, k; a)$ denotes the central generalized factorial coefficient. See Charalambides (2002).

Observe that these $\mathcal{C}(n, k; a)$ can be computed using the recursive formula

$$\mathcal{C}(n, k; a) = a\mathcal{C}(n-1, k-1; a) + (ka - n + 1)\mathcal{C}(n-1, k; a)$$

with the initial conditions $\mathcal{C}(0, 0, a) = 1$, $\mathcal{C}(n, 0, a) = 0$ for any $n > 0$ and $\mathcal{C}(n, k, a) = 0$ for $k > n$. See Charalambides (2002) for further details.

Using Corollary 5, we now consider a concrete example highlighting the difference between a repulsive and a non-repulsive point process to show the great flexibility of our prior. To this end, we compute $P(K_n = k, \boldsymbol{Y}^* \in \mathrm{d}\boldsymbol{y}^*)$ for $n = 5$ and $a = 1$ under two possible priors for Φ, i.e. a

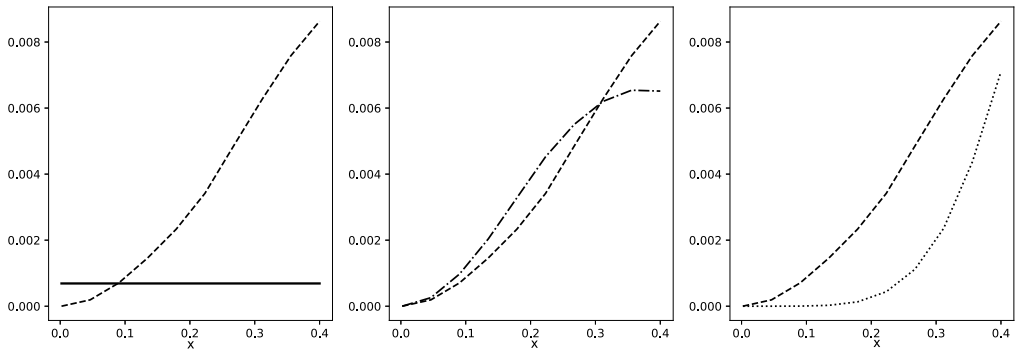

Figure 1. $P(K_n = k,\ \boldsymbol{Y}^* \in \mathrm{d}\boldsymbol{y}^*)$ when $n = 5$, $a = 1$, as a function of the parameter x under different settings. Left plot, setting (I) under the Poisson process (—) and DPP (– – –) prior. Middle plot: setting (I) (– – –) and setting (II) (– · – ·) under the DPP prior. Right plot setting (I) (– – –) and setting (III) (·····) under the DPP prior.

Poisson process and a DPP. In particular, the Poisson process prior has intensity $\omega(\mathrm{d}x) = \mathbb{1}_D(x)\mathrm{d}x$, where $D = (-1/2, 1/2)$. The DPP prior is defined on D as well and is characterized by a Gaussian covariance function $K(x, y) = 5\exp(-(x-y)^2/0.3)$. We consider different settings: in the first (I) $\boldsymbol{y}^* = (-x, x)$, in the second (II) $\boldsymbol{y}^* = (-0.3,\ -0.3 + 2x)$ and in the third (III) $\boldsymbol{y}^* = (-x, 0, x)$. Moreover, we assume that x varies in the interval $(0, 0.4)$. Figure 1 shows the joint probability of K_n and $\boldsymbol{Y}^*$ in the different scenarios. Note that under the Poisson process prior (solid line, left plot), the probability does not depend on $\boldsymbol{y}^*$. Instead, under the DPP prior, the three panels show that the probability increases when the points in $\boldsymbol{y}^*$ are well separated. The central panel shows the comparison between the probabilities under setting (I) (– – –) and setting (II) (– · – ·) for the DPP prior. This shows that these probabilities depend not only on the distance between the $\boldsymbol{y}^*$, but also on their position in D. The right panel displays setting (I) (– – –) and setting (III) (·····) under the DPP prior, showing the effect of observing 0 beyond observing $-x, x$.

4 Bayesian hierarchical mixture models

Discrete random probability measures are commonly employed as mixing measures in Bayesian mixture models to address clustering and density estimation. In a mixture model, instead of modelling observations via (3), we use the de Finetti measure Q in (3) as a prior for latent variables $Y_1, \ldots, Y_n$.

It is convenient to consider random measures on an extended space $\mathbb{X} \times \mathbb{W}$ instead of $\mathbb{X}$ to encompass location-scale mixtures. Hence, we deal with the random measure

$$\tilde{\mu}(\cdot) = \sum_{j \geq 1} S_j \delta_{(X_j, W_j)}, \tag{18}$$

obtained by marking the points in the point process Ψ with i.i.d. marks $(W_j)_{j \geq 1}$ from an absolutely continuous probability distribution over $\mathbb{W}$, whose density we will denote by $f_W(\cdot)$.

To formalize the mixture model, consider $\mathbb{Z}$-valued observations $Z_1, \ldots, Z_n$ and a probability kernel $f : \mathbb{Z} \times \mathbb{X} \times \mathbb{W} \to \mathbb{R}_+$, such that $z \mapsto f(z \mid y, v)$ is a probability density over $\mathbb{Z}$ for any $(y, v) \in \mathbb{X} \times \mathbb{W}$. We assume both $\mathbb{Z}$ and $\mathbb{W}$ to be Polish spaces, endowed with the associated Borel σ-algebras. More precisely, the statistical model we are dealing with is the following:

$$\begin{aligned} Z_i \mid Y_i, V_i &\overset{\text{ind}}{\sim} f(\cdot \mid Y_i, V_i), \qquad i = 1, \ldots, n \\ Y_i, V_i \mid \tilde{\mu} &\overset{\text{iid}}{\sim} \frac{\tilde{\mu}}{\tilde{\mu}(\mathbb{X} \times \mathbb{W})} \\ \tilde{\mu} &\sim \mathbf{P}_\mu. \end{aligned} \tag{19}$$

Usually, the kernel $f(\cdot \mid Y_i, V_i)$ is the Gaussian density with mean Y_i and variance (if data are univariate) or covariance matrix (if data are multivariate) V_i, where the points X_j's are the

component-specific means and the points W_j's the component-specific variances in a Gaussian mixture model. Note that, in accordance with previous literature on repulsive mixtures, we assume that the component-specific variances V_i's are i.i.d. and that interaction is on the locations Y_i's.

Let M be the number of support points in $\tilde{\mu}$. It is convenient to introduce auxiliary component indicator variables C_i, $i = 1, \ldots, n$ taking values in $\{1, \ldots, M\}$, such that $P(C_i = h \mid \tilde{\mu}) \propto S_h$. By noticing that $(Y_i, V_i)_{i\geq 1} = (X_{C_i}, W_{C_i})_{i\geq 1}$, (19) is equivalent to

$$\begin{aligned} Z_i \mid \tilde{\mu}, C_i &\overset{\text{ind}}{\sim} f(\cdot \mid X_{C_i}, W_{C_i}), \quad i = 1, \ldots, n \\ C_i \mid \tilde{\mu} &\overset{\text{iid}}{\sim} \text{Categorical}(S_1/S_\bullet, \ldots, S_M/S_\bullet) \\ \tilde{\mu} &\sim \mathbf{P}_\mu \end{aligned} \tag{20}$$

where $S_\bullet := \sum_{j=1}^{M} S_j$. Following the standard terminology of Argiento and De Iorio (2022) and Griffin and Walker (2011), we define $\boldsymbol{S}^{(a)} = \{S_1^{(a)}, \ldots, S_{K_n}^{(a)}\}$ as the distinct values in $\{S_{C_i}\}_{i\geq 1}$ and refer to them as *active* jumps. Moreover, we set the *non-active* jumps $\boldsymbol{S}^{(na)} := \{S_1, \ldots, S_M\} \setminus \boldsymbol{S}^{(a)}$. Analogously, we define $\boldsymbol{X}^{(a)}$, $\boldsymbol{X}^{(na)}$ and $\boldsymbol{W}^{(a)}$, $\boldsymbol{W}^{(na)}$ and refer to them as the active and non-active atoms, respectively. Note that, according to our notation, $\boldsymbol{X}^{(a)}$ coincide with the unique values in the latent variables $Y_1, \ldots, Y_n$ in (19).

In the rest of this section, we describe two Markov chain Monte Carlo algorithms for posterior inference under model (19). These are based on the posterior characterization of $\tilde{\mu}$ given $Y_1, \ldots, Y_n$ in Theorem 1 and the predictive distribution of Y_{n+1} in Theorem 3. Following Papaspiliopoulos and Roberts (2008), we term them *conditional* and *marginal* algorithms, respectively. In fact, in the former, the random measure $\tilde{\mu}$ is part of the state space of the algorithm, while $\tilde{\mu}$ is integrated out in the latter. We assume that, conditionally to the number of points M, the distribution of the vector $(X_1, \ldots, X_M) \in \mathbb{X}^M$ defining the support of Φ has a density. With an abuse of notation, we denote this density by f_Φ; the existence of the density of the point process Φ guarantees that the f_Φ above is indeed proportional to the density of the point process itself (see Møller Waagepetersen (2003), for more details).

4.1 A conditional MCMC algorithm

Theorem 1 can be easily rephrased to encompass the case of a point process Φ with i.i.d. marks $(W_j)_{j\geq 1}$: this is useful to derive the full-conditional of $\tilde{\mu}$ given $C_1, \ldots, C_n$, $\boldsymbol{X}^{(a)}$, $\boldsymbol{W}^{(a)}$ and U_n.

Corollary 6 Consider the model (20), and define $n_h = \sum_{i=1}^{n} \mathbb{1}_h(C_i)$. Then, conditionally on $C_1, \ldots, C_n$, $K_n = k$, $\boldsymbol{X}^{(a)} = \boldsymbol{x}^{(a)}$, $\boldsymbol{W}^{(a)} = \boldsymbol{w}^{(a)}$ and $U_n = u$, $\tilde{\mu}$ is equal in distribution to

$$\sum_{h=1}^{k} S_h^{(a)} \delta_{(x_h^{(a)}, w_h^{(a)})} + \tilde{\mu}'$$

where $S_h^{(a)} \sim f_{S_h^{(a)}}(s) \propto s^{n_h} \mathrm{e}^{-us} H(\mathrm{d}s)$, and $\tilde{\mu}' := \sum_{h\geq 1} S_h^{(na)} \delta_{(X_h^{(na)}, W_h^{(na)})}$ is such that $\tilde{\mu}'_X := \sum_{h\geq 1} S_h^{(na)} \delta_{X_h^{(na)}}$ has Laplace functional

$$\mathrm{E}\left[\exp \textstyle\int_{\mathbb{X}} -f(z)\tilde{\mu}'_X(\mathrm{d}z)\right] = \frac{\mathrm{E}\left[\exp\left\{-\int_{\mathbb{X}} (f(z)+u)\tilde{\mu}^{!}_{\boldsymbol{x}^{(a)}}(\mathrm{d}z)\right\}\right]}{\mathrm{E}\left[\exp\left\{-\int_{\mathbb{X}} u\tilde{\mu}^{!}_{\boldsymbol{x}^{(a)}}(\mathrm{d}z)\right\}\right]},$$

where $\tilde{\mu}^{!}_{\boldsymbol{x}^{(a)}}$ is as in (7) for $\boldsymbol{x} = \boldsymbol{x}^{(a)}$, and $W_h^{(na)} \overset{\text{iid}}{\sim} f_{\mathbb{W}}$.

Corollary 6 leads to the algorithm below, where we write '$\cdot \mid$ rest' to mean that we are conditioning with respect to all the variables but the ones appearing on the left-hand side of the conditioning symbol.

1. Sample $U_n \mid \text{rest} \sim \text{Gamma}(n, S_\bullet)$, where $S_\bullet := \sum_{j=1}^{M} S_j$.
2. Sample each C_i independently from a discrete distribution over $\{1, \ldots, M\}$ such that

$$P(C_i = h \mid \text{rest}) \propto S_h f(Z_i \mid X_h, W_h).$$

   Let k be the number of unique values in $\boldsymbol{X}^{(a)} := \{X_{C_i},\ i = 1, \ldots, n\}$. Define $\boldsymbol{W}^{(a)}$ and $\boldsymbol{S}^{(a)}$ analogously.
3. Sample $\tilde{\mu}$ using the distribution in Corollary 6. For the *active* part, sample each $S_h^{(a)}$ independently from a distribution on $\mathbb{R}_+$ with density

$$f_{S_h}(s) \propto \mathrm{e}^{-U_n s} s^{n_h} H(\mathrm{d}s).$$

   For the *non-active* part, first sample $\mu' = \sum_{j \geq 1} S_j^{(na)} \delta_{X_j^{(na)}}$ from the law of a random measure with Laplace transform (14). See below for some guidelines on how to perform this step. Then, sample the corresponding marks $W_j^{(na)} \overset{\text{iid}}{\sim} f_W$.
4. Let $\boldsymbol{X}^{(a)} = \{X_1^{(a)}, \ldots, X_k^{(a)}\}$ where $\tilde{X}^{(a)} = (X_1^{(a)}, \ldots, X_k^{(a)})$ is sampled from the density on $\mathbb{X}^k$ given by

$$P(\tilde{X}^{(a)} \in \mathrm{d}\boldsymbol{x} \mid \text{rest}) \propto f_\Phi(\boldsymbol{x}, \boldsymbol{X}^{(na)}) \prod_{h=1}^{k} \prod_{i:\, C_i = h} f(Z_i \mid x_h, W_h), \quad \boldsymbol{x} \in \mathbb{X}^k.$$

5. Sample each entry in $\boldsymbol{W}^{(a)}$ independently from a density on $\mathbb{R}_+$ proportional to

$$f_W(w) \prod_{i:\, C_i = h} f(Z_i \mid X_h^{(a)}, w).$$

6. Set $\boldsymbol{X} = \boldsymbol{X}^{(a)} \cup X^{(na)}$, $\boldsymbol{W} = \boldsymbol{W}^{(a)} \cup \boldsymbol{W}^{(na)}$ $\boldsymbol{S} = \boldsymbol{S}^{(a)} \cup \boldsymbol{S}^{(na)}$ and m equal to the cardinality of these sets.

Among the steps in the algorithm, the most complex is the third, which involves sampling the random measure $\tilde{\mu}'$ with a given Laplace transform. All the remaining ones can be handled either by closed-form full-conditionals (depending, for instance, on the law of the jumps $H(\mathrm{d}s)$ and the prior f_W) or by simple Metropolis–Hastings steps, yielding a Metropolis-within-Gibbs algorithm. To sample $\tilde{\mu}'$, we need to sample sequentially as follows: (3.i) the support points $\Phi' = \sum_{j \geq 1} \delta_{X_j^{(na)}}$, (3.ii) the jumps $\{S_j^{(na)}\}$, and (3.iii) the marks $W_j^{(na)} \overset{\text{iid}}{\sim} f_W$, which is trivial. Step (3.i) requires simulating a point process, and we refer to the literature on this topic, e.g. via perfect sampling (see Algorithm 11.7 in Møller Waagepetersen (2003)) or via a birth–death Metropolis–Hastings step (see Algorithm 11.3 in Møller Waagepetersen (2003)). A particular case of our Algorithm 1–6, when Φ is a Gibbs point process with a density, has been designed in Beraha et al. (2022). Step (3.ii) needs to iid sample from the tilted density $f_{S_j} \propto \mathrm{e}^{-Us} H(\mathrm{d}s)$. When $H(\mathrm{d}s)$ is a gamma density, this is straightforward; for different $H(\mathrm{d}s)$, we refer to Argiento and De Iorio (2022). On the other hand, when considering a DPP prior, we can use Algorithm 1 in Lavancier et al. (2015) (adapted from Hough et al., 2006) to obtain a perfect sample from the law of $\boldsymbol{X}^{(na)}$.

4.2 A marginal MCMC algorithm

When integrating out $\tilde{\mu}$ from (19), Theorem 3 can be used to devise a marginal MCMC strategy, i.e. an instance of *collapsed* Gibbs sampler. In such an algorithm, we alternate between updating the observation-specific parameters $\{(Y_i, V_i)\}_{i=1}^n \mid U_n, Z_1, \ldots, Z_n$ and sampling $U_n \mid \{(Y_i, V_i)\}_{i=1}^n,\ Z_1, \ldots, Z_n$, which is as (15). To sample $\{(Y_i, V_i)\}_{i=1}^n \mid U_n, Z_1, \ldots, Z_n$ we

perform a Gibbs scan, whereby we sample (Y_i, V_i) from its full conditional distribution, i.e. $(Y_i, V_i) \,|\, (\boldsymbol{Y}, \boldsymbol{V})_{-i}, U_n, Z_1, \ldots, Z_n$, where $(\boldsymbol{Y}, \boldsymbol{V})_{-i} = (Y_1, V_1), \ldots, (Y_{i-1}, V_{i-1}), (Y_{i+1}, V_{i+1}), (Y_n, V_n)$. It can be easily seen (Neal, 2000) that $P((Y_i, V_i) \in \mathrm{d}y\,\mathrm{d}v \,|\, (\boldsymbol{Y}, \boldsymbol{V})_{-i}, U_n, Z_1, \ldots, Z_n) \propto P(Y_i \in \mathrm{d}y \,|\, \boldsymbol{Y}_{-i}, U_n) f_{\mathbb{W}}(v)\mathrm{d}v f(Z_i \,|\, y, v)$ where $P(Y_i \in \mathrm{d}y \,|\, \boldsymbol{Y}_{-i}, U_n)$ is the conditional prior law of Y_i. In particular, by exchangeability, for any $\boldsymbol{y}_{-i} \in \mathbb{X}^{n-1}$ the conditional prior for Y_i satisfies $P(\boldsymbol{Y}_i \in \mathrm{d}y \,|\, \boldsymbol{Y}_{-i} = \boldsymbol{y}_{-i}, U_n) = P(\boldsymbol{Y}_n \in \mathrm{d}y \,|\, \boldsymbol{Y}_{-n} = \boldsymbol{y}_{-i}, U_n)$, where the latter is the law described Theorem 3. The plain application of this result yields the following MCMC algorithm.

1. Sample $U_n \,|\, \text{rest}$ from the full conditional distribution in (15).
2. For each observation, sample (Y_i, V_i) from

$$P((Y_i, V_i) \in (\mathrm{d}y\,\mathrm{d}v) \,|\, \text{rest}) \propto \sum_{j=1}^{k} \frac{\kappa(u, n_j^{(-i)} + 1)}{\kappa(u, n_j^{(-i)})} f(Z_i \,|\, Y_j^*, V_j^*) \delta_{(Y_j^*, V_j^*)}(\mathrm{d}y\,\mathrm{d}v) + \kappa(u, 1) \frac{\mathrm{E}\left[\mathrm{e}^{-u\tilde{\mu}^!_{\boldsymbol{y}^{*(-i)}, y}(\mathbb{X})}\right]}{\mathrm{E}\left[\mathrm{e}^{-u\tilde{\mu}^!_{\boldsymbol{y}^{*(-i)}}(\mathbb{X})}\right]} \frac{m_{\Phi^{k+1}}(\boldsymbol{y}^{*(-i)}, y)}{m_{\Phi^k}(\boldsymbol{y}^{*(-i)})} f(Z_i \,|\, y, v) P_0(\mathrm{d}y) f_{\mathbb{W}}(v)\mathrm{d}v,$$

   where the superscript $(-i)$ means that the ith observation is removed from the state for the computations. Here, $\boldsymbol{Y}^*$ and $\boldsymbol{V}^*$ are the vectors of unique values in $(Y_1, \ldots, Y_n)$ and $(V_1, \ldots, V_n)$, respectively.
3. Sample the unique values $\boldsymbol{Y}^*$ from a joint distribution on $\mathbb{X}^k$ proportional to

$$P(\boldsymbol{Y}^* \in \mathrm{d}\boldsymbol{y}^*) \propto m_{\Phi^k}(\boldsymbol{y}^*) \prod_{h=1}^{k} \prod_{\{i:\ Y_i = Y_h^*\}} f(Z_i \,|\, y_h^*, V_h^*) P_0^k(\mathrm{d}\boldsymbol{y}^*), \quad \boldsymbol{y}^* \in \mathbb{X}^k.$$

4. Sample each of unique values $\boldsymbol{V}^*$ independently from

$$P(V_h^* \in \mathrm{d}v^*) \propto f_{\mathbb{W}}(v^*) \prod_{i:\, C_i = h} f(Z_i \,|\, Y_h^*, v^*)\mathrm{d}v^*, \quad v^* \in \mathbb{W}.$$

Steps 3 and 4 of the above are not strictly necessary to produce an ergodic MCMC sampler that targets the posterior distribution. However, it is common practice in MCMC for mixture models to add these steps as it has been empirically demonstrated that they speed up the convergence of the algorithm and improve its mixing, see, e.g. Neal (2000). Step 2 of the algorithm above requires the computation of

$$\int_{\mathbb{X}\times\mathbb{W}} \frac{m_{\Phi^{k+1}}(\boldsymbol{y}^{*(-i)}, y)}{m_{\Phi^k}(\boldsymbol{y}^{*(-i)})} f(Z_i \,|\, y, v) P_0(\mathrm{d}y) f_{\mathbb{W}}(v)\mathrm{d}v$$

which might be challenging in situations where data are multidimensional. However, Xie and Xu (2019) shows how this can be overcome using numerical quadrature. Moreover, in the higher dimensional setting, we can adapt the strategy devised by Neal in his Algorithm 8 by introducing L auxiliary variables and replacing Step 2 with:

2'.a For $\ell = 1, \ldots, L$, sample $Y_{k+\ell}^*$ from

$$P(Y_{k+\ell}^* \in \mathrm{d}y \,|\, \text{rest}) \propto \frac{m_{\Phi^{k+1}}(\boldsymbol{y}^{*(-i)}, y)}{m_{\Phi^k}(\boldsymbol{y}^{*(-i)})} P_0(\mathrm{d}y)$$

and $V_{k+\ell}^* \overset{\text{iid}}{\sim} f_{\mathbb{W}}$.

2'.b Set (Y_i, V_i) equal to (Y_h^*, V_h^*) with probability proportional to

$$\begin{cases} \frac{\kappa(u, n_j^{(-i)}+1)}{\kappa(u, n_j^{(-i)})} f(Z_i \mid Y_h^*, V_h^*), & \text{for } h = 1, \ldots, k \\ \frac{1}{L}\kappa(u, 1) \frac{\mathrm{E}\left[\exp\left(-u\tilde{\mu}^{\dagger}_{y^{*(-i)}, Y_k^*}(\mathbb{X})\right)\right]}{\mathrm{E}\left[\exp\left(-u\tilde{\mu}^{\dagger}_{y^{*(-i)}}(\mathbb{X})\right)\right]} f(Z_i \mid Y_h^*, V_h^*), & \text{for } h = k+1, \ldots, k+L. \end{cases}$$

The computation of the ratio of expected values can be challenging if Φ is a DPP; see Example 5 for further details, while it poses no problems when Φ is a SNCP.

4.3 Shot-noise Cox process mixtures as mixtures of mixtures

Let us draw an interesting connection between a mixture model driven by a SNCP and the *mixtures of mixtures* model, whereby the density of each mixture component is approximated by a mixture model. This model was previously considered in Malsiner-Walli et al. (2017), who proposed a Bayesian methodology and Aragam et al. (2020), which studied identifiability and estimation within a frequentist setting.

First, recall from Example 2 that we can equivalently write $\Phi \mid \Lambda = \sum_{h=1}^{n(\Lambda)} \Phi_h$, where, with our notation, each Φ_h is a *group*. Observe that, for appropriate choices of k_α, the points in a *group* will be closer than points belonging to different *groups*. When we embed a random probability measure built from (9) in a Bayesian mixture model as done at the beginning of Section 4, we obtain that the atoms of the mixture are randomly *grouped* together. Hence, we rewrite the random population density as

$$f(z) = \frac{1}{S_\bullet} \sum_{j \geq 1} S_j f(z \mid X_j, W_j) \equiv \frac{1}{S_\bullet} \sum_{h=1}^{n(\Lambda)} \sum_{j \geq 1} S_{h,j} f(z \mid X_{h,j}, W_{h,j}).$$

On the right-hand side, we first sum over the atoms of Λ and then over the atoms of each of the *groups* Φ_j's. With an abuse of notation, we have introduced a second subscript to S_h, X_h, and W_h to keep track that each point of X_h (and its marks) is assigned to a point in Λ. It is possible to pass from the single-index summation to the double-index thanks to the variables T_j introduced in Example 2. Let us define

$$\tilde{f}_h(z) := \frac{1}{P_h} \sum_{j \geq 1} S_{h,j} f(z \mid X_{h,j}, W_{h,j}), \quad \text{where } P_h := \sum_{j \geq 1} S_{h,j}, \tag{21}$$

which is a (random) probability density function over $\mathbb{X}$. We clearly see that the random population density is equal to $f(z) = S_\bullet^{-1} \sum_{h=1}^{n(\Lambda)} P_h \tilde{f}_h(z)$. Hence, a SNCP mixture model can be written as a *mixture of mixtures*, where each component $\tilde{f}_h(z)$ is expressed as a mixture model itself. Therefore, we can consider the SNCP mixture model as a nonparametric generalization of the *mixtures of mixtures* model in Malsiner-Walli et al. (2017).

Observe that the SNCP mixture induces a two-level clustering. At the first level, the parameters (Y_i, V_i) in (19) induce the standard partition of the observations by the equivalence relation $i \sim j$ if and only if $(Y_i, V_i) = (Y_j, V_j) = (X_h, W_h)$ for some h. Let $C_i = h$ if and only if $(Y_i, V_i) = (X_h, W_h)$. Then, at the second level, each point (X_h, W_h) can be referred to as one of the *groups* Φ_j thanks to the variables T_j introduced in Example 2. Hence, we can refer observations to the *groups* using the variables $(T_{C_1}, \ldots, T_{C_n})$. In the next subsection, we call the partition induced by the T_{C_i}'s as the *grouping* of the observations.

5 Numerical illustrations

Here, we consider two simulated scenarios and a real dataset to highlight the core differences between a mixture model with repulsive, i.i.d., or 'attractive' atoms. We focus on mixtures of

Gaussian kernels, i.e. f in (19) is the Gaussian density where parameters (Y_i, V_i) are the mean and the variance, respectively. We compare posterior inference under three prior specifications for the measure $\tilde{\mu}$ in (18); in all cases, the unnormalized weights are $S_j \overset{\text{iid}}{\sim} \text{Gamma}(2, 2)$. Under the first model, the cluster centres $\{X_j\}_{j\geq 1}$ follow a DPP with kernel $K(x, y) = \rho \exp(-(x-y)^2/\alpha^2)$ for $\rho = 2$ and $\alpha = \sqrt{\pi}/2$, paired with inverse-Gamma density f_W for the prior of the W_j's with both shape and scale parameters equal to two. See online supplementary material, Appendix S9 for further details on the DPP prior considered and online supplementary material, Appendix S9.1.1 for details on the computation of the DPP density with respect to a suitable Poisson process. The second model we consider is the normalized IFPP mixture model by Argiento and De Iorio (2022), which corresponds to assuming that $\{(X_j, W_j)\}_{j=1}^{K} \mid K$ are i.i.d. from a Normal-inverse-Gamma density (i.e. $X_j | W_j \sim \mathcal{N}(0, 10W_j)$ and $1/W_i \sim \text{Gamma}(2, 2)$), and we further assume $K - 1 \sim \text{Poi}(1)$. Our third choice of the marginal prior for the X_j's is the SNCP process, where $k_\alpha(\cdot)$ is the Gaussian density centred at the origin with standard deviation α. Unless otherwise specified, we fix $\alpha = 1$. We also fix $\gamma = 1$ and assume that the W_j's are distributed as in the DPP case. Prior elicitation for the SNCP mixture is reported in greater detail in online supplementary material, Appendix S9. Moreover, we also discuss there the robustness of posterior inference with respect to hyperparameters in the case of a DPP and a SNCP prior, as well as the inherent difficulty of learning the repulsive parameters via a fully Bayesian approach, i.e. assuming the parameters controlling the kernel of the DPP random and updating them as part of the MCMC algorithm.

Posterior inference for the model in Argiento and De Iorio (2022) is addressed via the `BayesMix` library (Beraha, Guindani et al., 2025). To fit the SNCP mixture, we use the conditional algorithm in Section 4.1, where we further add to the MCMC state the points of Λ; see online supplementary material, Section S8.4 for further details. As for the repulsive mixture based on DPP, we use the conditional algorithm described in Section 4.1. See also Beraha et al. (2022) for further details. A Python implementation of the different MCMC algorithms is available at https://github.com/mberaha/interacting_mixtures. We run the MCMC algorithms for a total of 100,000 iterations, discarding the first 50,000 and keeping one every five iterations for a final size of 1,000. In the sequel, we compare the *grouping* of the datapoints induced by the SNCP mixture, as defined above, with the standard clustering resulting from the use of DPP and normalized IFPP mixtures; we also compare density estimation under the three models.

5.1 Data from a mixture of t distributions

In the first scenario, we generate 200 data points from a two-component mixture of univariate Student's t distributions with three degrees of freedom, centred respectively in -5 and $+5$. Posterior inference is summarized in Figure 2. Both the DPP and SNCP mixtures identify two clusters, with the difference that the SNCP mixture activates between 25 and 40 Gaussian components, while the DPP only has two. On the other hand, the IFPP mixture has between 3 and 8 active components. The co-clustering matrices for the IFPP and SNCP mixtures are shown in

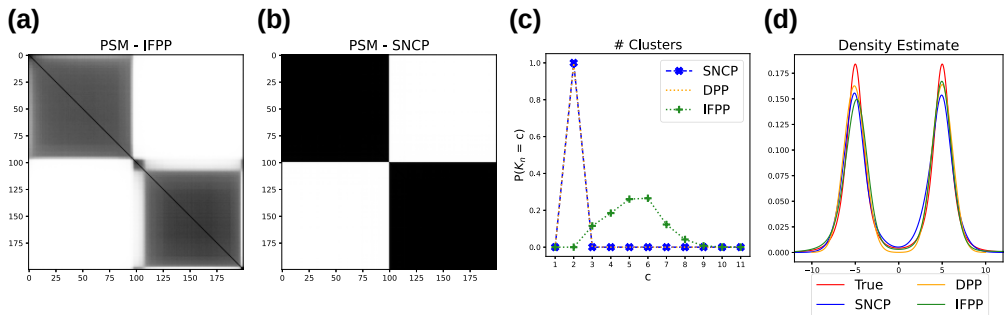

Figure 2. From left to right: posterior co-clustering matrix under the normalized IFPP (a) and SNCP (b) mixtures, posterior distribution of the number of clusters (c), density estimates (d) for the simulated example in Section 5.1. In the posterior co-clustering matrices, data are sorted from the smallest to the largest.

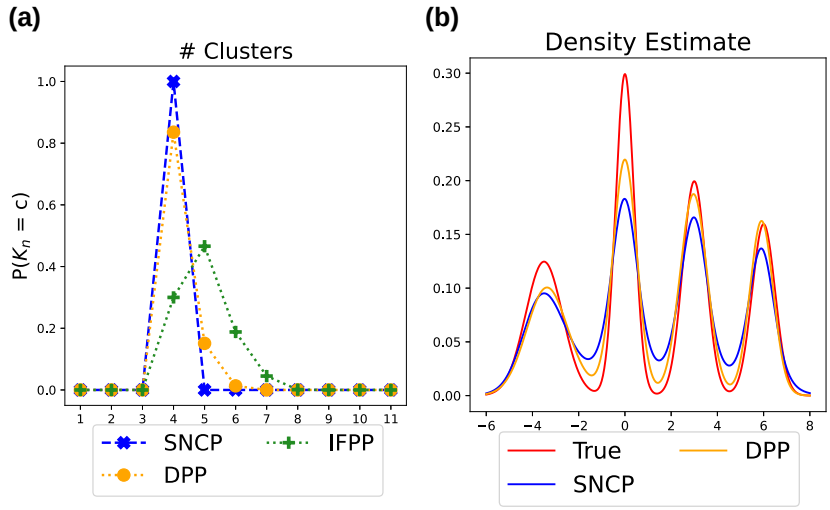

Figure 3. Posterior distribution of the number of clusters (a) and density estimates (b) for the simulated example in Section 5.2.

Figure 2, and the one associated with the DPP mixture is identical to that of the SNCP. We observe that the IFPP model erroneously identifies three clusters: the two large clusters, we would expect, and a third cluster, which collects the observations at the centre of the domain. Surprisingly, this last cluster also contains some observations at the farthest right of the domain: this can be explained by the presence of a mixture component located near zero and with an extremely large variance, which produces an overabundant estimated cluster that groups together data that are far apart. All the models produce a satisfactory density estimate. However, as it is to be expected, the DPP mixture fails to capture the heavy tails of the t distributions both at the centre of the domain and at the left and right extremities, resulting in a slightly poorer fit.

5.2 Data from a contaminated mixture

Our second simulation follows the setup in Miller and Dunson (2019). We assume that a 'true' data generating process with density $f_0 = 0.25\mathcal{N}(-3.5, 0.8^2) + 0.3\mathcal{N}(0, 0.4^2) + 0.25\mathcal{N}(3, 0.5^2) + 0.2\mathcal{N}(6, 0.5^2)$ has been corrupted by some noise, and we observe data distributed as

$$Z_i \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{f}(\cdot) = \int_{\mathbb{R}} \mathcal{N}(\cdot \mid \theta, 0.25^2)\tilde{p}(\mathrm{d}\theta), \quad \tilde{p} \sim DP(500 f_0),$$

where $DP(500f_0)$ denotes the Dirichlet process with total mass parameter equal to 500 and centreing probability measure induced by f_0. We interpret $\tilde{f}$ as a perturbation of f_0 so that the goal is then to recover f_0 from the observed data.

Figure 3 shows the posterior distribution for the number of clusters and the density estimates when $n = 500$. The density estimate under the IFPP mixture is essentially identical to the one under the SNCP and is, therefore, omitted. The SNCP mixture correctly identifies 4 clusters, while the DPP mixture chooses between 4 and 5 clusters (the point estimate of the partition agrees on 4 clusters) and the IFPP mixture between 4 and 7 clusters. Hence, although there is no 'true' number of clusters in the data-generating process, we can argue that the SNCP model produces the most parsimonious and interpretable clustering. On the other hand, it is clear that the density estimate under SNCP is somewhat worse than under the DPP prior. Indeed, the SNCP mixture picks up all the noise induced by the perturbation.

5.3 Shapley galaxy data

We consider now a real dataset containing the velocities of $n = 4{,}206$ galaxies (measured in 10^5 km/s) in the Shapley supercluster, available in the `R` package `spatstat`. This can be seen as a

modern and improved version of the popular galaxy dataset (Roeder, 1990). We compare the density and cluster estimates when using a subsample of $n = 100, 500, 2000$ datapoints as well as the whole dataset.

We set the prior distribution via an empirical Bayes approach; further details and the numerical values chosen are reported in online supplementary material, Appendix S9.2.2. Posterior inference is summarized in Figure 4. We notice how under all three models, the estimated number of clusters grows with the sample size; however, while under the SNCP and DPP mixtures, it stabilizes around 7 clusters at most, under the IFPP prior, when $n = 4{,}206$, the model induces more than 16 clusters for some iterations of the MCMC algorithm. However, remember that for the SNCP mixtures we report the estimated *groups*. The density estimates are consistent with the discussion in the examples above. The IFPP and SNCP mixtures induce very similar density estimates, while the DPP mixture tends to oversmooth the density in some regions to obtain well-separated components. In this particular example, the density estimate under a SNCP prior reflects the empirical histogram carefully, picking up even very subtle 'bumps' in the density, such as the one near the value 33 in the x-axis. Since the measurements of galaxy velocities are typically accurate (especially at the scale we are considering), we can recommend using the SNCP mixture over the IFPP model for density estimation in this case.

We remark here that a crucial parameter for the SNCP mixture is the scale α, which controls how spread a component $\tilde{f}_j$ (defined in (21)) can be. Indeed, if α is too large, the model will tend to group all the normal components in one single large $\tilde{f}_j$, thereby estimating only one cluster. On the other hand, if α is too small, each $\tilde{f}_j$ is more likely to correspond to a single normal component, thus producing too many clusters. In our examples, we did not notice a particular sensibility to this choice for both density estimation and clustering, as long as the values are chosen in an appropriate range. In online supplementary material, Appendix S9.2.3, we report a sensitivity analysis for the choice of α using the Shapley galaxy data.

6 Discussion and future works

In this work, we have investigated discrete random probability measures with interaction across support points. Our study is motivated by the recent popularity of repulsive mixtures (Beraha et al., 2022; Bianchini et al., 2020; Cremaschi et al., 2025; Petralia et al., 2012; Quinlan et al., 2021; Xu et al., 2016) for Bayesian model-based clustering.

The first main contribution of the article is to propose a general construction of RPMs via the normalization of marked point processes with repulsive or attractive support points, which can be embedded in a hierarchical mixture model, as shown in Section 4. The second contribution of this article is to present a unified mathematical theory by encompassing both repulsive and attractive mixtures, as well as the IFPP mixtures by Argiento and De Iorio (2022). Our results also extend the well-known distributional properties of NRMIs (Regazzini et al., 2003). Due to the generality of our framework, all the proofs rely on new arguments based on Palm calculus; see Baccelli et al. (2020). Thirdly, we discuss the use of shot-noise Cox processes, which were not previously considered in connection with Bayesian mixture models beyond Wang et al. (2024). As a further contribution, we include the applicability of the MCMC algorithms proposed in Section 4 to any choice of the point process prior. However, they might lack efficiency, especially for moderate or high-dimensional data. More adequate algorithms could be devised for specific classes of point processes; see, for instance, H. Sun et al. (2022) for an alternative MCMC scheme tailored to the class of Matérn point processes.

Our work paves the way for several contributions with both theoretical and computational flavour. We discuss some ideas below.

6.1 The trade-off between density and cluster estimates, and related asymptotics

Consider the setting of a well-specified mixture model, i.e. the data come from a finite mixture model whose kernel agrees with our modelling choices. Then, for standard mixtures whose atoms are i.i.d. from a base distribution, a well-established theory ensures convergence of the latent mixing measure (in an appropriate sense) to the true data-generating process under suitable conditions. See, e.g. Guha et al. (2021). In this setting, it is easy to see that the posterior of repulsive mixtures also converges to the true mixing measure under the additional requirement that the

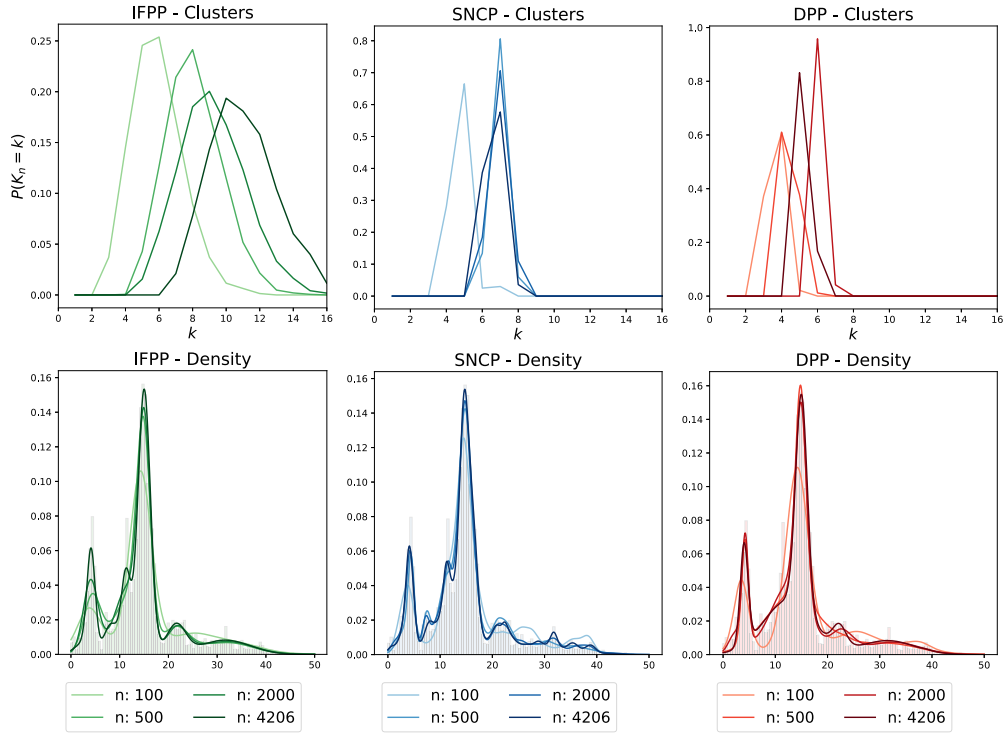

Figure 4. Analysis of the Shapley galaxy data under the three different models.

true cluster centres are sufficiently separated. In particular, a minor modification of Theorem 3.1 in Guha et al. (2021) ensures that the posterior of K_n is consistent and the mixing measure converges to the true one at an optimal rate, provided that the point process has a density that is bounded from below when evaluated at sufficiently separated points.

A more interesting scenario occurs when the mixture model is misspecified. This brings to light a fundamental trade-off between cluster and density estimates as shown in two simulation studies of Section 5. The general asymptotic theory for misspecified models outlined in Kleijn and van der Vaart (2006), specialized to the case of mixtures by Guha et al. (2021), ensures that, in the large sample limit, the mixing measure $\tilde{p}$ contracts to a measure p^* that minimizes the Kullback–Leibler divergence between the true data generating density and the support of the prior. As shown in Cai et al. (2021), for traditional mixtures with i.i.d. atoms, p^* has infinite support. We expect that also in the case of SNCP mixtures, the posterior of the number of components diverges, but, thanks to the decoupling of the notion of cluster (i.e. 'grouping') from that of allocated mixture component, the SNCP might estimate the correct number of groups of data even in the large sample limit. Things are less clear when assuming a repulsive point process prior. To the best of our knowledge, the only result in these directions is Corollary 1 by Xie and Xu (2019), which establishes that, under a specific prior, the number of clusters grows sub-linearly with the sample size. Of course, we expect different priors to yield different asymptotics. Consider, for instance, the case of a hardcore process, for which the probability of having two points closer than a certain radius r is exactly zero. Then, suppose the data generating process is, e.g. a truncated t distribution over a certain interval and r is sufficiently large. In that case, it is reasonable to expect that $K_n = 1$ for any n (since the prior makes it impossible to add more atoms to the mixture).

In a sense, we might argue that repulsive mixtures are the 'safe option', meaning that if the model is well-specified, we expect the same posterior inference as with traditional mixtures. In contrast, if the model is misspecified, empirical works suggest that inference is more robust. The asymptotic study of repulsive mixtures under misspecification presents a 'dual' challenge with respect to the

typical study of mixture models. Indeed, the challenge here is to explicitly characterize the limiting p^*, a problem that needs to be more noticed in the current literature.

6.2 Repulsive priors or more flexible kernels?

From the discussion above, it is clear that posterior inference in traditional mixture models is troublesome in misspecified regimes. Several approaches have been proposed to fix this issue, including generalized Bayes frameworks (Rigon et al., 2023) and distance-based clustering (Duan & Dunson, 2021; Natarajan et al., 2024). These offer robust cluster estimates but not density estimates, and obtaining a robust density estimate might be an equally important problem (Miller & Dunson, 2019). As shown in Section 5.2, repulsive mixtures do offer robust density estimates in addition to cluster estimates.

Another natural way to address model misspecification issues is to improve the flexibility of mixture kernels. See the recent contributions of Mukhopadhyay et al. (2020); Paez and Walker (2018); Rodríguez and Walker (2014). The SNCP mixture proposed in this work constitutes another contribution in this direction, whereby we model a mixture kernel via a mixture of Gaussians. The numerical illustrations of Section 5 show that the choice of the prior distribution (i.e. whether it has repulsive, independent, or attractive atoms) must be carried out on a case-by-case basis. Indeed, the use of an extremely flexible kernel might result in noisy density estimates if the data-generating process is thought to be corrupted. At the same time, a repulsive mixture might produce unsatisfactory density estimates in other settings. Ultimately, it is up to the practitioner to choose the right model for the job.

Acknowledgments

We wish to thank two anonymous referees for their valuable remarks, suggestions, and constructive criticism that led to a substantial improvement of the article. The research was (partially) completed while R.A., F.C., and A.G. were visiting the Institute for Mathematical Sciences, National University of Singapore, in 2024.

Conflicts of interest: None declared.

Funding

M.B., F.C., and A.G. gratefully acknowledge support from the Italian Ministry of Education, University and Research (MUR), 'Dipartimenti di Eccellenza' grant 2023-2027. F.C. is supported by the European Union—Next Generation EU funds, component M4C2, investment 1.1., PRIN-PNRR 2022 (P2022H5WZ9). F.C. is a member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). R.A. and A.G. have been partially supported by MUR - Prin 2022 - Grant no. 2022CLTYP4, funded by the European Union—Next Generation EU.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

Data availability

The data and code underlying the article are available at: https://github.com/mberaha/interacting_mixtures.

References

Aldous D. J. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*. Lecture notes in math (Vol. 1117, pp. 1–198). Springer.

Aragam B., Dan C., Xing E. P., & Ravikumar P. (2020). Identifiability of nonparametric mixture models and Bayes optimal clustering. *Annals of Statistics*, *48*(4), 2277–2302. https://doi.org/10.1214/19-AOS1887

Argiento R., & De Iorio M. (2022). Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *Annals of Statistics*, *50*(5), 2641–2663. https://doi.org/10.1214/22-AOS2201

Baccelli F., Błaszczyszyn B., & Karray M. (2020). Random measures, point processes, and stochastic geometry. HAL preprint available at https://hal.inria.fr/hal-02460214/

Beraha M., Argiento R., Møller J., & Guglielmi A. G. (2022). MCMC computations for Bayesian mixture models using repulsive point processes. *Journal of Computational and Graphical Statistics*, *31*(2), 422–435. https://doi.org/10.1080/10618600.2021.2000424

Beraha M., Camerlenghi F., & Ghilotti L. (2025). 'Bayesian calculus and predictive characterizations of extended feature allocation models', arXiv, arXiv:2502.10257, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2502.10257

Beraha M., Guindani B., Gianella M., & Guglielmi A. (2025b). Bayesmix: Bayesian mixture models in C++. *Journal of Statistical Software*, *112*(9), 1–40. https://doi.org/10.18637/jss.v112.i09

Bianchini I., Guglielmi A., & Quintana F. A. (2020). Determinantal point process mixtures via spectral density approach. *Bayesian Analysis*, *15*(1), 187–214. https://doi.org/10.1214/19-BA1150

Cai D., Campbell T., & Broderick T. (2021). Finite mixture models do not reliably learn the number of components. In *International Conference on Machine Learning* (pp. 1158–1169). PMLR.

Camerlenghi F., Lijoi A., & Prünster I. (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scandinavian Journal of Statistics*, *45*(4), 1062–1091. https://doi.org/10.1111/sjos.v45.4

Charalambides C. A. (2002). *Enumerative combinatorics*. CRC press series on discrete mathematics and its applications. Chapman & Hall/CRC.

Cox D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, *17*(2), 129–157. https://doi.org/10.1111/j.2517-6161.1955.tb00188.x

Cremaschi A., Wertz T. M., & De Iorio M. (2025). Repulsion, chaos, and equilibrium in mixture models. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *87*(2), 389–432. https://doi.org/10.1093/jrsssb/qkae096

Daley D. J., & Vere-Jones D. (2003). *An introduction to the theory of point processes. Vol. I: Elementary theory and methods* (2nd ed.). Probability and its applications (New York) (Springer).

De Blasi P., Favaro S., Lijoi A., Mena R. H., Prünster I., & Ruggiero M. (2013). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(2), 212–229. https://doi.org/10.1109/TPAMI.2013.217

de Finetti B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, *7*(1), 1–68. https://www.numdam.org/item/AIHP_1937__7_1_1_0.pdf

Duan L. L., & Dunson D. B. (2021). Bayesian distance clustering. *The Journal of Machine Learning Research*, *22*(1), 10228–10254.

Ferguson T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*(2), 209–230. https://doi.org/10.1214/aos/1176342360

Ferreira J., & Menegatto V. (2009). Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations and Operator Theory*, *64*(1), 61–81. https://doi.org/10.1007/s00020-009-1680-3

Fruhwirth-Schnatter S., Celeux G., & Robert C. P. (2019). *Handbook of mixture analysis*. CRC Press.

Ghosal S., Ghosh J. K., & Ramamoorthi R. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, *27*(1), 143–158. https://doi.org/10.1214/aos/1018031105

Grazian C. (2023). 'A review on Bayesian model-based clustering', arXiv, arXiv:2303.17182, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2303.17182

Griffin J. E., & Walker S. G. (2011). Posterior simulation of normalized random measure mixtures. *Journal of Computational and Graphical Statistics*, *20*(1), 241–259. https://doi.org/10.1198/jcgs.2010.08176

Guha A., Ho N., & Nguyen X. (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli*, *27*(4), 2159–2188. https://doi.org/10.3150/20-BEJ1275

Hough J. B., Krishnapur M., Peres Y., & Viràg B. (2006). Determinantal processes and independence. *Probability Surveys*, *3*(none), 206–229. https://doi.org/10.1214/154957806000000078

Hough J. B., Krishnapur M., Peres Y., & Virág B. (2009). *Zeros of Gaussian analytic functions and determinantal point processes*. American Mathematical Society.

James L. F. (2002). 'Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics', arXiv, arXiv:math/0205093, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.math/0205093

James L. F. (2005). Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *Annals of Statistics*, *33*(4), 1771–1799. https://doi.org/10.1214/009053605000000336

James L. F. (2006). Poisson calculus for spatial neutral to the right processes. *Annals of Statistics*, *34*(1), 416–440. https://doi.org/10.1214/009053605000000732

James L. F., Lijoi A., & Prünster I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, *36*(1), 76–97. https://doi.org/10.1111/sjos.2009.36.issue-1

Kallenberg O. (2021). *Foundations of modern probability*. Probability theory and stochastic modelling (3rd ed., Vol. *99*), Springer.

Kingman J. F. C. (1992). *Poisson processes* (Vol. *3*). Clarendon Press.

Kleijn B. J. K., & van der Vaart A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*, *34*(2), 837–877. https://doi.org/10.1214/009053606000000029
Last G., & Penrose M. (2017). *Lectures on the Poisson process* (Vol. 7). Cambridge University Press.
Lavancier F., Møller J., & Rubak E. (2015). Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *77*(4), 853–877. https://doi.org/10.1111/rssb.12096
Lijoi A., & Prünster I. (2010). Models beyond the Dirichlet process. *Bayesian Nonparametrics*, *28*(80), 342. https://doi.org/10.1017/cbo9780511802478.004
Lijoi A., Prünster I., & Rigon T. (2022). Finite-dimensional discrete random structures and Bayesian clustering. *Journal of the American Statistical Association*, *119*(546), 929–941. https://doi.org/10.1080/01621459.2022.2149406
Macchi O. (1975). The coincidence approach to stochastic point processes. *Advances in Applied Probability*, *7*(1), 83–122. https://doi.org/10.2307/1425855
Malsiner-Walli G., Frühwirth-Schnatter S., & Grün B. (2017). Identifying mixtures of mixtures using Bayesian estimation. *Journal of Computational and Graphical Statistics*, *26*(2), 285–295. https://doi.org/10.1080/10618600.2016.1200472
Miller J. W., & Dunson D. B. (2019). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, *114*(527), 1113–1125. https://doi.org/10.1080/01621459.2018.1469995
Møller J. (2003). Shot noise Cox processes. *Advances in Applied Probability*, *35*(3), 614–640. https://doi.org/10.1239/aap/1059486821
Møller J., & Waagepetersen R. P. (2003). *Statistical inference and simulation for spatial point processes*. CRC Press.
Mukhopadhyay M., Li D., & Dunson D. B. (2020). Estimating densities with non-linear support by using Fisher–Gaussian kernels. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *82*(5), 1249–1271. https://doi.org/10.1111/rssb.12390
Natarajan A., De Iorio M., Heinecke A., Mayer E., & Glenn S. (2024). Cohesion and repulsion in Bayesian distance clustering. *Journal of the American Statistical Association*, *119*(546), 1374–1384. https://doi.org/10.1080/01621459.2023.2191821
Neal R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265. https://doi.org/10.1080/10618600.2000.10474879
Paez M. S., & Walker S. G. (2018). Modeling with a large class of unimodal multivariate distributions. *Journal of Applied Statistics*, *45*(10), 1823–1845. https://doi.org/10.1080/02664763.2017.1396296
Papaspiliopoulos O., & Roberts G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, *95*(1), 169–186. https://doi.org/10.1093/biomet/asm086
Petralia F., Rao V., & Dunson D. (2012). Repulsive mixtures. In F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25). Curran Associates, Inc.
Pitman J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, probability and game theory*, IMS lecture notes monogr. ser. (Vol. 30, pp. 245–267). Inst. Math. Statist..
Quinlan J. J., Quintana F. A., & Page G. L. (2021). Parsimonious hierarchical modeling using repulsive distributions. *Test*, *30*(2), 445–461. https://doi.org/10.1007/s11749-020-00726-y
Regazzini E., Lijoi A., & Prünster I. (2003). Distributional results for means of normalized random measures with independent increments. *Annals of Statistics*, *31*(2), 560–585. https://doi.org/10.1214/aos/1051027881
Rigon T., Herring A. H., & Dunson D. B. (2023). A generalized Bayes framework for probabilistic clustering. *Biometrika*, *110*(3), 559–578. https://doi.org/10.1093/biomet/asad004
Rodríguez C. E., & Walker S. G. (2014). Univariate Bayesian nonparametric mixture modeling with unimodal kernels. *Statistics and Computing*, *24*(1), 35–49. https://doi.org/10.1007/s11222-012-9351-7
Roeder K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, *85*(411), 617–624. https://doi.org/10.1080/01621459.1990.10474918
Shirai T., & Takahashi Y. (2003). Random point fields associated with certain Fredholm determinants I: Fermion, Poisson and boson point processes. *Journal of Functional Analysis*, *205*(2), 414–463. https://doi.org/10.1016/S0022-1236(03)00171-X
Soshnikov A. (2000). Determinantal random point fields. *Russian Mathematical Surveys*, *55*(5), 923. https://doi.org/10.1070/RM2000v055n05ABEH000321
Steele J. M. (1994). Le cam's inequality and Poisson approximations. *The American Mathematical Monthly: The Official Journal of the Mathematical Association of America*, *101*(1), 48–54. https://doi.org/10.1080/00029890.1994.11996904
Sun H., Zhang B., & Rao V. (2022). 'Bayesian Repulsive Mixture Modeling with Matern Point Processes', arXiv, arXiv:2210.04140, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2210.04140
Sun S., Zhao J., & Zhu J. (2015). A review of Nyström methods for large-scale machine learning. *Information Fusion*, *26*, 36–48. https://doi.org/10.1016/j.inffus.2015.03.001

Teh Y. W., Jordan M. I., Beal M. J., & Blei D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581. https://doi.org/10.1198/016214506000000302

Wade S. (2023). Bayesian cluster analysis. *Philosophical Transactions of the Royal Society A*, *381*(2247), Paper No. 20220149, 20. https://doi.org/10.1098/rsta.2022.0149

Wang Y., Degleris A., Williams A. H., & Linderman S. W. (2024). Spatiotemporal clustering with Neyman-Scott processes via connections to Bayesian nonparametric mixture models. *Journal of the American Statistical Association*, *119*(547), 2382–2395. https://doi.org/10.1080/01621459.2023.2257896

Xie F., & Xu Y. (2019). Bayesian repulsive Gaussian mixture model. *Journal of the American Statistical Association*, *115*(529), 187–203. https://doi.org/10.1080/01621459.2018.1537918

Xu Y., Müller P., & Telesca D. (2016). Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics*, *72*(3), 955–964. https://doi.org/10.1111/biom.12482

Zhou M., Favaro S., & Walker S. G. (2017). Frequency of frequencies distributions and size-dependent exchangeable random partitions. *Journal of the American Statistical Association*, *112*(520), 1623–1635. https://doi.org/10.1080/01621459.2016.1222290