On a Class of Repulsive Mixture Models

José J. Quinlan
Pontificia Universidad Católica de Chile, Santiago, Chile
Fernando A. Quintana
Pontificia Universidad Católica de Chile, Santiago, Chile and
Millennium Nucleus Center for the Discovery of Structures in Complex Data
Garritt L. Page
Brigham Young University, USA

July 12, 2020

Abstract

Finite or infinite mixture models are routinely used in Bayesian statistical practice for tasks such as clustering or density esimation. Such models are very attractive due to their flexibility and tractability. However, a common problem in fitting these or other discrete models to data is that they tend to produce a large number of overlapping clusters. Some attention has been given in the statistical literature to models that include a repulsive feature, i.e. that encourage separation of mixture components. We study here a method that has been shown to achieve this goal without sacrificing flexibility or model fit. The model is a special case of Gibbs measures, with a parameter that controls the level of repulsion that allows construction of d-dimensional probability densities whose coordinates tend to repel each other. This approach was successfully used for density regression in Quinlan et al. (2018). We detail some of the global properties of the repulsive family of distributions and offer some further insight by means of a small simulation study.

Key words: Gibbs measures, mixture models, repulsive point processes, hierarchical modeling.

1 Introduction

Hierarchical mixture models have been very successfully employed in a myriad of applications of Bayesian modeling. A typical formulation for such models adopts the basic form

$$y_i \mid \theta_i \sim k(y_i;\theta_i), \quad \theta_1,\ldots,\theta_n \sim \sum_{\ell=1}^{N} w_\ell \delta_{\phi_\ell}, \quad \phi_1,\ldots,\phi_N \sim G_0, \tag{1}$$

where $k(y;\theta)$ is a suitable kernel density, $1 \leq N \leq \infty$, component weights $w_1,\ldots,w_N$ are nonnegative random variables such that $\sum_{\ell=1}^{N} w_\ell = 1$ with probability 1 (often assigned a Dirichlet or stick-breaking prior) and G_0 is a suitable nonatomic probability distribution. Conditional independence is typically assumed for $y_i \mid \theta_i$ and $\theta_1,\ldots,\theta_n$ are independent and identically distributed. Here N could be regarded as fixed or random and in the latter case a prior $p(N)$ would need to be specified. Depending on the modeling goals and data particularities, the model could have additional parameters and levels in the hierarchy. The generic model (1) includes, as special cases, finite mixture models (Frühwirth-Schnatter; 2006) and species sampling mixture models (Pitman; 1996; Quintana; 2006), in turn including several well-known particular examples such as the Dirichlet process (Ferguson; 1973) and the Pitman-Yor process (Pitman and Yor; 1997). There is also a substantial body of literature concerning important properties of these models such as wide support, posterior consistency, posterior convergence rates, and connections to finite point processes among others. See, for instance, Ghosal and van der Vaart (2007), Shen et al. (2013), and Argiento and De Iorio (2019).

A common feature of models like (1) is the use of independent and identically distributed (i.i.d.) atoms $\phi_1,\ldots,\phi_N$. This choice seems to have been largely motivated by the resulting tractability of the models, specially in the nonparametric case ($N = \infty$). While the use of i.i.d. atoms in (1) is technically (and practically) convenient, a typical summary of the induced posterior clustering will usually contain a number of redundant clusters (i.e., clusters that are very close to each other) or very small clusters or even some singletons. This is facilitated precisely by the use of i.i.d. atoms, which imply no

prior restriction on where these atoms may land. Motivated by similar considerations, the literature has developed some approaches to define probability models that feature atoms that mutually repel each other. We refer to this feature as *repulsion*. Colloquially, the concept of repulsion among a set of objects implies that the objects tend to separate rather than congregate. This notion of repulsion has been studied in the context of point processes. For example, determinantal point processes (Lavancier et al.; 2015), Strauss point processes (Mateu and Montes; 2000; Ogata and Tanemura; 1985) and Matérn-type point processes (Rao et al.; 2017) are all able to generate point patterns that exhibit more repulsion than that expected from a Poisson point process (Daley and Vere-Jones; 2002). From these, we are only aware of determinantal point processes being used in statistical modeling (Xu et al. 2016; Bianchini et al. 2020).

An alternative way to incorporate the notion of repulsion in modeling is to construct a multivariate probability distribution that contains a repulsion parameter that explicitly influences or dictates the level of repulsion. Along these lines, Fúquene et al. (2019) develop a family of probability densities called non-local priors that incorporates repulsion by penalizing small relative distances between location parameters in a mixture model. Quinlan et al. (2018) discussed density regression that featured the definition of an explicit family of probability distributions for mixture location parameters through potentials (functions that describe the ability to interact) as found in Gibbs measures. The idea also includes a penalization based on pairwise distances between locations. Their proposal is related to Xie and Xu (2020), who also consider repulsion based on penalizing pairwise distances among cluster location parameters. The main difference between these approaches is the ability to control the repulsion strength and the flexibility on the types of repulsion that can be considered.

Gibbs measures have been widely studied and used for describing phenomena from mechanical statistics (Daley and Vere-Jones; 2002). Essentially, they are used to model the average macroscopic behavior of particle systems through a set of probability and physical laws that are imposed over the possible microscopic states of the system. Through the action of potentials, Gibbs measures can induce attraction or repulsion between particles. For example, although no direct connection was made, Petralia et al. (2012)'s method uses a Gibbs measure with a Lennard-Jones type potential (Jones; 1924) to introduce repulsion. The works by Petralia et al. (2012) and Quinlan et al.

(2018) are both special cases of Gibbs measures, the later sharing also some connections with determinantal point processes via versions of Papangelou intensities (Papangelou; 1974). See Georgii and Yoo (2005) for more details.

The aim of this article is to develop properties of the modeling approach proposed in Quinlan et al. (2018). We do this considering the behavior of repulsive distributions and also their use in Gaussian mixture models. In particular, we show conditions for the class of repulsive distributions considered to be well defined, study its usage in the context of hierarchical repulsive mixture models, and state properties of the corresponding posterior distribution. Our strategy for modeling using mixtures consists of assuming a fixed and sufficiently large number of components, an approach that is referred to as *overfitted mixtures* in, e.g., Rousseau and Mengersen (2011). In this context, our approach focuses on the random induced number of *occupied* or *active* components (see Section 6 of Argiento and De Iorio 2019). In particular, we show that our approach of incorporating repulsion in modeling results in posterior convergence rates that are similar to the i.i.d. case.

The rest of this article is organized as follows. In Section 2 we recall and contextualize the definition of repulsive distributions that are the main subject of interest and discuss several of its properties. In Section 3, we detail how the repulsive probability distributions can be employed in hierarchical mixture modeling. Some aspects of the model are illustrated by way of a simulation study in Section 4, and Section 5 provides a brief discussion and points out to some potentially interesting extensions. An online Supplementary Material contains proofs of the main results and computational strategies.

2 A Probability Repulsive Distribution

We start by introducing some notation. Let the N-fold product space of $\mathbb{R}^d$ be denoted by $\mathbb{R}^d_N$ and let $\mathcal{B}(\mathbb{R}^d_N)$ be its associated Borel σ-algebra. They constitute the reference space on which the class of distributions we consider is defined, where $N, d \in \mathbb{N}$ ($N \geq 2$). Let $x_{N,d} = (x_1, \ldots, x_N)$ with $x_1, \ldots, x_N \in \mathbb{R}^d$. In the context of d-dimensional location-scale mixture models, the coordinates $x_1, \ldots, x_N$ of $x_{N,d}$ can be thought of as the N ordered location parameters jointly allocated in $\mathbb{R}^d_N$. To the measurable space

$(\mathbb{R}^d_N, \mathcal{B}(\mathbb{R}^d_N))$ we add λ_d^N, the N-fold product of the d-dimensional Lebesgue measure λ_d. To represent integrals with respect to λ_d^N, we will use $\mathrm{d}x_{N,d}$ instead of $\mathrm{d}\lambda_d^N(x_{N,d})$. Also, given two metric spaces Ω_1 and Ω_2 we will use $C(\Omega_1; \Omega_2)$ to denote the class of all continuous functions $f : \Omega_1 \to \Omega_2$. In what follows we use the term "repulsive distribution" to reference a distribution that formally incorporates the notion of repulsion.

2.1 Gibbs Measures

The repulsive distribution in Quinlan et al. (2018) is a member of the general class of Gibbs measures for which dependence (and hence repulsion) between the coordinates of $x_{N,d}$ is introduced via functions that model interactions between them. More formally, let $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ be the extended real line and $\mathcal{B}(\overline{\mathbb{R}})$ its associated Borel σ-algebra generated by the order topology. Consider $\varphi_1 : \mathbb{R}^d \to \overline{\mathbb{R}}$ a measurable function and $\varphi_2 : \mathbb{R}^d \times \mathbb{R}^d \to \overline{\mathbb{R}}$ a measurable and symmetric function. Define

$$\nu_{\mathrm{G}}(A_1 \times \cdots \times A_N) = \int_{A_1 \times \cdots \times A_N} \exp\left\{ -\sum_{i=1}^{N} \varphi_1(x_i) - \sum_{1 \le r < s \le N} \varphi_2(x_r, x_s) \right\} \mathrm{d}x_{N,d}, \quad (2)$$

where $A_1 \times \cdots \times A_N$ is the cartesian product of Borel sets $A_1, \ldots, A_N$ in $\mathbb{R}^d$. Here, φ_1 can be thought of as a physical force that controls the influence that the environment has on each coordinate x_i while φ_2 controls the interaction between pairs of coordinates x_r and x_s. The induced probability measure corresponding to the normalized version of (2), is called a (second-order) Gibbs measure. The normalizing constant (total mass of $\mathbb{R}^d_N$ under ν_{G}) is commonly known as the partition function (Pathria and Beale; 2011) and encapsulates important qualitative information about the interactions and the degree of disorder present in the coordinates of $x_{N,d}$. In general, $\nu_{\mathrm{G}}(\mathbb{R}^d_N)$'s tractability depends mainly on the presence of φ_2.

Note that symmetry of φ_2 means that ν_{G} defines a symmetric measure. If $\varphi_2 = 0$ then ν_{G} reduces to a structure where coordinates do not interact and are only subject to environmental influence through φ_1. When $\varphi_2 \neq 0$, it is common that $\varphi_2(x, y)$ only depends on the relative distance between x and y (Daley and Vere-Jones; 2002). More formally, let $\rho : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ be a metric on $\mathbb{R}^d$ and $\phi : [0, \infty) \to \overline{\mathbb{R}}$ a measurable function. To avoid pathological or degenerate cases, we consider metrics that do not treat singletons as open sets in the topology induced by ρ. Then letting

$\varphi_2(x,y) = \phi\{\rho(x,y)\}$, interactions will be smooth if $\phi \in C([0,\infty);\overline{\mathbb{R}})$. For example, Petralia et al. (2012) use $\phi(r) = \tau r^{-\nu}$ with $\tau, \nu > 0$ to construct repulsive probability densities, which is a particular case of the Lennard-Jones type potential (Jones; 1924) that appears in molecular dynamics. Another potential that can be used to define repulsion is the (Gibbs) hard-core potential $\phi(r) = \infty \mathbb{I}_{[0,b]}(r)$ with $b > 0$ (Illian et al.; 2008), which is a particular case of the Strauss potential (Strauss; 1975). Here, $\mathbb{I}_A(r)$ is the indicator function over a Borel set A in $\mathbb{R}$. This potential, used in the context of point processes, generates disperse point patterns whose points are all separated by a distance greater than b units. However, the threshold of separation b prevents the repulsion from being smooth (Daley and Vere-Jones; 2002). Other examples of repulsive potentials can be found in Ogata and Tanemura (1981, 1985). The key characteristic that differentiates the behavior of the potentials provided above is the action near 0; the faster the potential function goes to infinity as relative distance between coordinates goes to zero, the stronger the repulsion that the coordinates of $x_{N,d}$ will experiment when they are separated by small distances.

2.2 $\text{Rep}_{N,d}(f_0, C_0, \rho)$ Distribution

There are of course many potentials that could be considered in a Gibbs measure. The motivation in Quinlan et al. (2018) was to find one that permits modeling repulsion flexibly, i.e. that avoids forcing more separation among coordinates than required to satisfactorily model the available data. As noted by Daley and Vere-Jones (2002) and Ogata and Tanemura (1981) the potential

$$\phi(r) = -\log\left\{1 - \exp\left(-cr^2\right)\right\}, \quad c > 0, \tag{3}$$

produces smoother repulsion compared to other types of potentials in terms of "repelling strength". This is adopted in the construction below. Note first that connecting (3) with ν_{G} is straightforward: if we take $\varphi_2(x,y) = -\log[1 - C_0\{\rho(x,y)\}]$ for $x, y \in \mathbb{R}^d$ with $C_0(r) = \exp(-cr^2)$ then ν_{G} will have a "pairwise-interaction term" given by

$$\exp\left\{-\sum_{1\le r<s\le N} \varphi_2(x_r, x_s)\right\} = \prod_{1\le r<s\le N} [1 - C_0\{\rho(x_r, x_s)\}]. \tag{4}$$

The right-hand side of (4) induces a particular interaction structure that separates the coordinates of $x_{N,d}$, thus introducing a notion of repulsion. The degree of separation is

regulated by the speed at which C_0 decays to 0. In general, the focus is on functions $C_0 : [0, \infty) \to (0, 1]$ that satisfy the following four properties: (1) $C_0 \in C([0, \infty); (0, 1])$, (2) $C_0(0) = 1$; (3) $C_0(x) \downarrow 0$ when $x \to \infty$; and (4) for all $x, y \geq 0$, if $x < y$ then $C_0(x) > C_0(y)$. Continuity of the repulsion in terms of $x_{N,d}$ induced by these four properties is guaranteed next.

Lemma 1. *Given a metric $\rho : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ such that singletons are not open sets in the topology induced by ρ, the function $\mathrm{R_C} : \mathbb{R}_N^d \to [0, 1)$ defined by*

$$\mathrm{R_C}(x_{N,d}) = \prod_{1 \leq r < s \leq N} [1 - C_0\{\rho(x_r, x_s)\}] \tag{5}$$

belongs to $C(\mathbb{R}_N^d; [0, 1))$ for all $N, d \in \mathbb{N}$ $(N \geq 2)$.

We omit the corresponding proof. We call (5) the repulsive component. Let now $f_0 \in C(\mathbb{R}^d; (0, \infty))$ be a probability density function, so that under $\varphi_1(x) = -\log\{f_0(x)\}$, ν_{G} in (2) has a "baseline term" given by

$$\exp\left\{-\sum_{i=1}^{N} \varphi_1(x_i)\right\} = \prod_{i=1}^{N} f_0(x_i). \tag{6}$$

Incorporating (4) and (6) into (2) we get

$$\nu_{\mathrm{G}}(A_1 \times \cdots \times A_N) = \int_{A_1 \times \cdots \times A_N} \left\{\prod_{i=1}^{N} f_0(x_i)\right\} \mathrm{R_C}(x_{N,d}) \mathrm{d}x_{N,d}.$$

The repulsive probability measures just constructed are well defined as stated next.

Proposition 1. *Let $f_0 \in C(\mathbb{R}^d; (0, \infty))$ be a probability density function. The function*

$$g(x_{N,d}) = \left\{\prod_{i=1}^{N} f_0(x_i)\right\} \mathrm{R_C}(x_{N,d}) \tag{7}$$

is measurable and integrable for $N, d \in \mathbb{N}$ $(N \geq 2)$.

The proof of this result is straightforward and therefore omitted. The class of probability measures $\mathrm{Rep}_{N,d}(f_0, C_0, \rho)$ is defined next.

Definition 1. The probability distribution $\text{Rep}_{N,d}(f_0, C_0, \rho)$ has density function

$$\text{Rep}_{N,d}(x_{N,d}) = \frac{1}{c_{N,d}} \left\{ \prod_{i=1}^{N} f_0(x_i) \right\} \text{R}_\text{C}(x_{N,d}), \tag{8}$$

$$\text{where} \qquad c_{N,d} = \int_{\mathbb{R}_N^d} \left\{ \prod_{i=1}^{N} f_0(x_i) \right\} \text{R}_\text{C}(x_{N,d}) \text{d}x_{N,d}. \tag{9}$$

Here $x_{N,d} \in \mathbb{R}_N^d$, $f_0 \in C(\mathbb{R}^d; (0, \infty))$ is a probability density function, $C_0 : [0, \infty) \to (0, 1]$ is a function that satisfies the four above properties and $\rho : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ is a metric such that singletons are not open sets in the topology induced by it.

Proposition 1 guarantees that the class $\text{Rep}_{N,d}(f_0, C_0, \rho)$ is well defined. Notice that the support of (7) is determined by the shape of the "baseline" distribution f_0 and then subsequently distorted (i.e. contracted) by the repulsive component. We also note that the proposals studied in Xie and Xu (2020) have a similar form to (7), but with different choices of repulsive component. They use, for instance,

$$\min \left\{ g(||x_r - x_s||_2) : 1 \leq r < s \leq N \right\},$$

for some monotonically increasing function $g : [0, \infty) \to [0, 1]$ with $g(0) = 0$ and $|| \cdot ||_2$ being the Euclidean L_2-norm in $\mathbb{R}^d$, which results in a distribution that is not of Gibbs type. They also consider a variation of the form

$$\prod_{1 \leq r < s \leq N} \left\{ g(||x_r - x_s||_2) \right\}^{1/N},$$

which varies the penalty with N, something we avoid.

It is worth noting two properties related to the $\text{Rep}_{N,d}(f_0, C_0, \rho)$ distribution: (i) because of symmetry, it is an exchangeable distribution in $x_1, \ldots, x_N$; and (ii) it does not induce a sample-size consistent sequence of finite-dimensional distributions, meaning that

$$\int_{\mathbb{R}^d} \text{Rep}_{N+1,d}(x_{N+1,d}) \text{d}x_{N+1} \neq \text{Rep}_{N,d}(x_{N,d}).$$

This makes predicting locations of new coordinates problematic. In Section 3 we address how this may be accommodated in the context of modeling using mixtures.

3 Repulsive Mixture Models

A main application of the $\text{Rep}_{N,d}(f_0, C_0, \rho)$ class is as a prior distrbution for location parameters in a hierarchical mixture model. We now describe such a setting, focusing on Gaussian kernels. In applications one would need to specify all the parameters of the repulsive measure. In what follows, we study properties of a particular choice of these parameters, which can be thought of as the natural repulsive extension of standard mixture models with i.i.d parameters.

3.1 Mixtures of Repulsive Distributions

To simplify notation, we will use $[m] = \{1, \ldots, m\}$, with $m \in \mathbb{N}$. Consider $n \in \mathbb{N}$ experimental units whose d-dimensional responses $y_1, \ldots, y_n$ are assumed to be exchangeable. Let K denote the number of components (or the possible number of clusters) in the mixture model and assume that the jth cluster ($j \in [K]$) is modeled with a Gaussian density $\text{N}_d(\,\cdot\,; \theta_j, \Lambda_j)$ with location $\theta_j \in \mathbb{R}^d$ and scale $\Lambda_j \in \mathbb{S}^d$. Here, $\mathbb{S}^d$ is the space of real, symmetric and positive-definite matrices of dimension $d \times d$. We let $\theta_{K,d} = (\theta_1, \ldots, \theta_K) \in \mathbb{R}^d_K$ and $\Lambda_{K,d} = (\Lambda_1, \ldots, \Lambda_K) \in \mathbb{S}^d_K$ where $\mathbb{S}^d_K$ is the K-fold product space of $\mathbb{S}^d$. Next let $\pi_{K,1} = (\pi_1, \ldots, \pi_K)^\top \in \Delta_K$, where Δ_K is the $(K-1)$-dimensional simplex ($\Delta_1 = \{1\}$). The standard Gaussian mixture model is then

$$y_i \mid \pi_{K,1}, \theta_{K,d}, \Lambda_{K,d} \sim \sum_{j=1}^{K} \pi_j \text{N}_d(y_i; \theta_j, \Lambda_j), \tag{10}$$

which is commonly restated by introducing latent cluster membership indicators z_i, $i \in [n]$ such that y_i is drawn from the jth mixture component if and only if $z_i = j$:

$$y_i \mid z_i, \theta_{K,d}, \Lambda_{K,d} \sim \text{N}_d(y_i; \theta_{z_i}, \Lambda_{z_i}) \tag{11}$$

$$z_i \mid \pi_{K,1} \sim \mathbb{P}(z_i = j) = \pi_j. \tag{12}$$

The model is completed by assigning standard conjugate-style priors for all parameters.

In the above mixture model, the location parameters associated with each mixture component are typically assumed to be independent a priori. This is precisely the assumption that facilitates the presence of redundant mixture components. We instead consider employing $\text{Rep}_{K,d}(f_0, C_0, \rho)$ as a model for location parameters in (10). This

promotes reducing redundant mixture components (or singletons) without substantially sacrificing goodness-of-fit, i.e, more parsimony relative to alternatives with independent locations, and responses that are a priori encouraged to be allocated to a few well-separated clusters. A specification of the $\text{Rep}_{K,d}(f_0, C_0, \rho)$ parameters that achieves the desired goals is given by

$$\begin{aligned}
\theta_{K,d} &\sim \text{Rep}_{K,d}(f_0, C_0, \rho) \\
f_0(x) &= \text{N}_d(x; \mu, \Sigma), \quad \mu \in \mathbb{R}^d, \quad \Sigma \in \mathbb{S}^d && (13) \\
C_0(r) &= \exp(-0.5\tau^{-1} r^2), \quad \tau > 0 && (14) \\
\rho(x, y) &= \{(x-y)^\top \Sigma^{-1} (x-y)\}^{1/2}. && (15)
\end{aligned}$$

The specific forms of f_0, C_0 and ρ are admissible according to Definition 1. The repulsive distribution parameterized by (13)–(15) will be denoted by $\text{NRep}_{K,d}(\mu, \Sigma, \tau)$. Because $\text{NRep}_{K,d}(\mu, \Sigma, \tau)$ introduces dependence a priori (in particular, repulsion) between location parameters, they are no longer conditionally independent given $(y_{n,d}, z_{n,1}, \Lambda_{K,d})$, with $y_{n,d} = (y_1, \ldots, y_n) \in \mathbb{R}_n^d$ and $z_{n,1} = (z_1, \ldots, z_n)^\top \in [k]^n$. The parameter τ in (14) controls the strength of repulsion via (15): as τ approaches 0, the repulsion becomes weaker. To finish the model specification we employ the following independent prior distributions

$$\begin{aligned}
\pi_{K,1} &\sim \text{Dir}(\alpha_{K,1}), \quad \alpha_{K,1} = (\alpha_1, \ldots, \alpha_K), \quad \alpha_1, \ldots, \alpha_K > 0 && (16) \\
\theta_{K,d} &\sim \text{NRep}_{K,d}(\mu, \Sigma, \tau), \quad \mu \in \mathbb{R}^d, \quad \Sigma \in \mathbb{S}^d, \quad \tau > 0 && (17) \\
\Lambda_j &\sim \text{IW}_d(\Psi, \nu), \quad \Psi \in \mathbb{S}^d, \quad \nu > 0. && (18)
\end{aligned}$$

At first glance it may seem that the unknown number of clusters in our model specification is fixed. However, our construction induces a prior distribution on this quantity in the following way. Denote by $n_j = \sum_{i=1}^n \mathbb{I}_{\{j\}}(z_i)$ the number of units assigned to the jth mixture component. The quantity that regulates the number of clusters is not necessarily K, but rather

$$k = K - \sum_{j=1}^{K} \mathbb{I}_{\{0\}}(n_j).$$

Here K represents an upper bound on the number of clusters which we denote by k. This is a well-defined quantity and the $\text{NRep}_{K,d}(\mu, \Sigma, \tau)$ distribution for mixture component

centers along with a Dirichlet-Multinomial distribution for component labels induce a prior model on k (though its distribution is not easy to write down explicitly), with support on $[K]$. See Argiento and De Iorio (2019) for a nice discussion on the differences between mixture components and clusters. In what follows we use "active number of components" and "number of clusters" interchangeably.

In the remainder of the article we will refer to the model in (11)–(18) as the (Bayesian) repulsive Gaussian (finite) mixture model (RGMM).

3.2 Theoretical Properties

In this section we explore properties associated the support and posterior consistency based on the model detailed in (10) and (16)–(18). Theoretical results are guided by the derivations found in Petralia et al. (2012). Our contribution is that we explore rates of convergence for the modeling framework we employ in practice (i.e., fixing the upper bound K), as opposed to the approach taken by Petralia et al. (2012) in which they explore convergence rates when a prior is assigned to the number of components, but in practice they fixed an upper bound K.

Consider the family of probability density functions

$$\mathcal{G}_K = \left\{ f(\,\cdot\,;\xi_K) = \sum_{j=1}^{K} \pi_j \mathrm{N}(\,\cdot\,;\theta_j,\lambda) : \xi_K = (\pi_{K,1}, \theta_{K,1}, \lambda) \in \Theta_K \right\},$$

where $\pi_{K,1} = (\pi_1,\ldots,\pi_K)^\top$, $\theta_{K,1} = (\theta_1,\ldots,\theta_K)^\top$ and $\Theta_K = \Delta_K \times \mathbb{R}^K \times (0,\infty)$. The class $\mathcal{G}_K$ consists of location mixtures of Gaussian distributions with common variance λ and at most K atoms given by $\{\theta_1,\ldots,\theta_K\}$. Additionally, let $B_p(x,r)$ with $x \in \mathbb{R}^K$ and $r > 0$ denote an open ball centered on x, with radius r, and $D_p(x,r)$ its closure relative to the Euclidean L_p-metric ($p \geq 1$) on $\mathbb{R}^K$.

The results stated next are based on the following conditions:

- B1. The true data generating density $f_0 \in \mathcal{G}_K$, i.e. $f_0(\,\cdot\,) = f(\,\cdot\,;\xi_K^0)$ for some $\xi_K^0 = (\pi_{K,1}^0, \theta_{K,1}^0, \lambda_0) \in \Theta_K$. Here, f_0 has exactly $k_0 \in [K]$ atoms given by $\theta_1^0,\ldots,\theta_{k_0}^0$ with respective weights $\pi_1^0,\ldots,\pi_{k_0}^0$ jointly lying in the interior of Δ_{k_0}. If $k_0 < K$ then $\pi_{K,1}^0$ and $\theta_{K,1}^0$ are viewed as follow: choose $\{\theta_i^0 : i = (k_0{+}1),\ldots,K\}$ such that

$$\min\left\{|\theta_r^0 - \theta_s^0| : 1 \leq r < s \leq K\right\} \geq v_0$$

for some $v_0 > 0$ and $\pi_i^0 = 0$ for $i = (k_0 + 1), \ldots, K$.

- B2. The space Θ_K is equipped with the prior distribution

$$P_K = \text{Dir}(\alpha_{K,1}) \otimes \text{NRep}_{K,1}(\mu, \sigma, \tau) \otimes \text{IG}(a, b),$$

where $\mu \in \mathbb{R}$ and $a, b, \sigma, \tau > 0$. As for $\alpha_{K,1}$ each coordinate $\alpha_i : i \in [K]$ satisfies $A_0 \varepsilon_0^{a_0} \leq \alpha_i \leq D_0$ for some constants $a_0, A_0, D_0 > 0$ and $0 < \varepsilon_0 \leq (D_0 K)^{-1}$.

Condition B1 implies that the true cluster centers are separated by a minimum (Euclidean) distance which favors disperse cluster centroids within the range of the response. Condition B2 guarantees the existence of a prior probability measure Π_K defined on $\mathcal{G}_K$ through P_K.

We study the support of Π_K using the Kullback–Leibler divergence. We say that $f_0 \in \mathcal{G}_K$ belongs to the Kullback–Leibler support with respect to Π_K if, for all $\varepsilon > 0$

$$\Pi_K \left\{ \left(f \in \mathcal{G}_K : \int_{\mathbb{R}} \log \left\{ \frac{f_0(x)}{f(x)} \right\} f_0(x) \mathrm{d}x < \varepsilon \right) \right\} > 0. \tag{19}$$

Condition (19) can be understood as Π_K's ability to assign positive mass to arbitrarily small neighborhoods around the true density f_0. A fundamental step to proving that f_0 lies in the Kullback–Leibler support of Π_K is based on the following lemmas:

Lemma 2. *Under condition B1, let $0 < \varepsilon < \lambda_0$. Then there exists $\delta > 0$ such that*

$$\int_{\mathbb{R}} \log \left\{ \frac{f(x; \xi_K^0)}{f(x; \xi_K)} \right\} f(x; \xi_K^0) \mathrm{d}x < \varepsilon$$

for all $\xi_K \in B_1(\pi_{K,1}^0, \delta) \times B_1(\theta_{K,1}^0, \delta) \times (\lambda_0 - \delta, \lambda_0 + \delta)$.

Proof: See Section A of the Supplementary Material.

Lemma 3. *Under condition B1, let $\theta_{K,1} \sim \text{NRep}_{K,1}(\mu, \sigma, \tau)$. Then there exists $\delta_0 > 0$ such that*

$$\mathbb{P}\left\{ \theta_{K,1} \in B_1(\theta_{K,1}^0, \delta) \right\} > 0$$

for all $0 < \delta \leq \delta_0$. This result remains valid even when replacing $B_1(\theta_{K,1}^0, \delta)$ with $D_1(\theta_{K,1}^0, \delta)$.

Proof: See Section A of the Supplementary Material.

Using Lemmas 2 and 3 we are able to prove the following proposition:

Proposition 2. *Assume that conditions B1 and B2 hold. Then f_0 belongs to the Kullback–Leibler support of Π_K.*

Proof: See Section A of the Supplementary Material.

We next study the rate of convergence of the posterior distribution under the prior given by condition B2.. To do this, we show that the conditions specified in Ghosal and van der Vaart (2001) hold for the model specifications and priors we consider. Arguments are similar to those found in Scricciolo (2011) when considering univariate Gaussian mixture models and cluster-location parameters that follow conditions B1 and B2. First, we need the following lemma:

Lemma 4. *The coordinates of $\theta_{K,1} \sim \text{NRep}_{K,1}(\mu, \sigma, \tau)$ share the same functional form. Moreover, there exists $\gamma > 0$ such that*

$$\mathbb{P}(|\theta_i| > t) \leq \frac{2\sigma^{1/2}}{(2\pi)^{1/2} c_K} \exp\left(-\frac{t^2}{4\sigma}\right)$$

for all $t \geq \gamma$ and $i \in [K]$. Here, c_K is the normalizing constant of $\text{NRep}_{K,1}(\mu, \sigma, \tau)$.

Proof: See Section A of the Supplementary Material.

This result permits us to adapt certain arguments found in Scricciolo (2011) that are applicable when the location parameters of each mixture component are independent and follow a common distribution that is absolutely continuous with respect to the Lebesgue measure, whose support is $\mathbb{R}$ and with tails that decay exponentially. Using Lemma 4, we now state the following:

Proposition 3. *Assume that conditions B1 and B2 hold. Then, the posterior rate of convergence relative to the L_1-metric is $\varepsilon_n = n^{-1/2} \log(n)^{1/2}$.*

Proof: See Section A of the Supplementary Material.

4 Simulation Study

To numerically explore how τ, the upper bound K on the number of mixture components, and n impact performance of the repulsive mixture model, we conduct a

simulation study. This is done by treating (10) as a data generating mechanism such that

$$y \sim 0.2\mathrm{N}_2(\theta_1, \Lambda_1) + 0.3\mathrm{N}_2(\theta_2, \Lambda_2) + 0.3\mathrm{N}_2(\theta_3, \Lambda_3) + 0.2\mathrm{N}_2(\theta_4, \Lambda_4), \tag{20}$$

with

$$\theta_1 = (0,0)^\top, \quad \theta_2 = (4.5, 4.5)^\top, \quad \theta_3 = (-4.5, -4.5)^\top, \quad \theta_4 = (-3, 3)^\top,$$
$$\Lambda_1 = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}, \quad \Lambda_2 = \begin{pmatrix} 4 & 1.5 \\ 1.5 & 3 \end{pmatrix},$$
$$\Lambda_3 = \begin{pmatrix} 2 & -1.5 \\ -1.5 & 4 \end{pmatrix} \quad \text{and} \quad \Lambda_4 = \begin{pmatrix} 4 & -3 \\ -3 & 3 \end{pmatrix}.$$

This bivariate mixture distribution produces clusters with little overlap. For each data set generated, we fit four finite Gaussian mixtures based on the following prior distributions for component means:

1. M1: $(\theta_1, \ldots, \theta_K)^\top$ are i.i.d. according to a Gaussian distribution.
2. M2: $(\theta_1, \ldots, \theta_K)^\top$ follows a repulsive distribution with Ogata & Tanemura type potential.
3. M3: $(\theta_1, \ldots, \theta_K)^\top$ follows a repulsive distribution with potential used in Petralia et al. (2012).
4. M4: $(\theta_1, \ldots, \theta_K)^\top$ follows a repulsive distribution with a hard-core potential.

The exact functional form of each of the potential functions employed in M2–M4 are provided in Section E of the Supplementary Material. We nevertheless stress here that M2–M4 are all special forms of the proposed class, and M1 also follows as a limiting case when the repulsive component of the prior vanishes.

Before detailing results of the simulation, it is worth illustrating how the value of τ impacts the relative distance between the θs for each potential. We do this by sampling 10,000 draws from each of M2–M4 with $K = 4$. Further, we set $\mu = (0,0)^\top$, $\Sigma = I_2$, and consider a sequence of τ values between 0 and 5. For each draw we compute the minimum pairwise distances among the four θs. The average minimum pairwise

distances over the 10,000 draws can be found in Figure 1. Notice that the repulsion associated with M2 as τ increases is not as strong as that of M3. It is easy to see that the hard-core potential repels very strongly for values of τ much greater than 2.5. This is illustrated further in the simulation study.

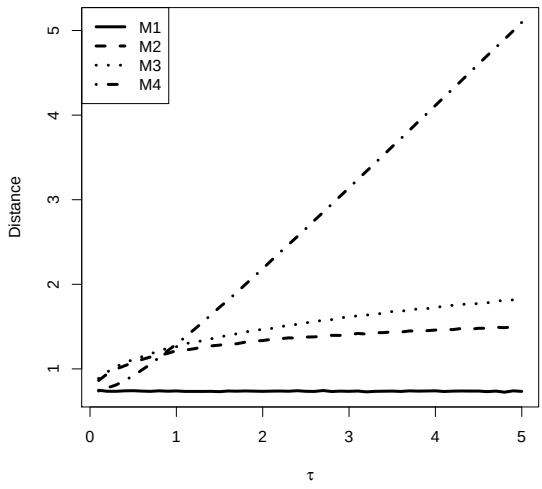

Figure 1: Minimum relative Euclidean distances averaged over 10000 draws sampled from the repulsive distribution of each of the four models considered in the simulation study.

Returning to the simulation study, after standardizing the data, each of the four models were fit to each data set by employing the MCMC algorithm detailed in Sections B and C of the Supplementary Material. We collected 1000 MCMC samples after discarding the first 10000 as burn-in and thinning by 2. The specific hyper-prior values we employed for parameters found in (16)–(18) are $\mu = (0, 0)^\top$, $\Sigma = I_2$, $\nu = 6$, $\Psi = \text{diag}(3, 2)$, and $\alpha_1 = \cdots = \alpha_K = K^{-1}$. The first two prior specifications are reasonable since we standardized the data, while the last three are specifically designed to make priors diffuse. In the simulation study we consider $K \in \{4, 7, 10\}$ and $\tau \in \{0.1, 1.0, 5.0\}$. Results based on 100 datasets of $n \in \{500, 1000, 5000\}$ observations are provided in Figures 2 and 3.

Note first that M1 does not depend on τ and as a result the metrics measured in the simulation study for M1 do not change as a function of τ (apart from Monte Carlo

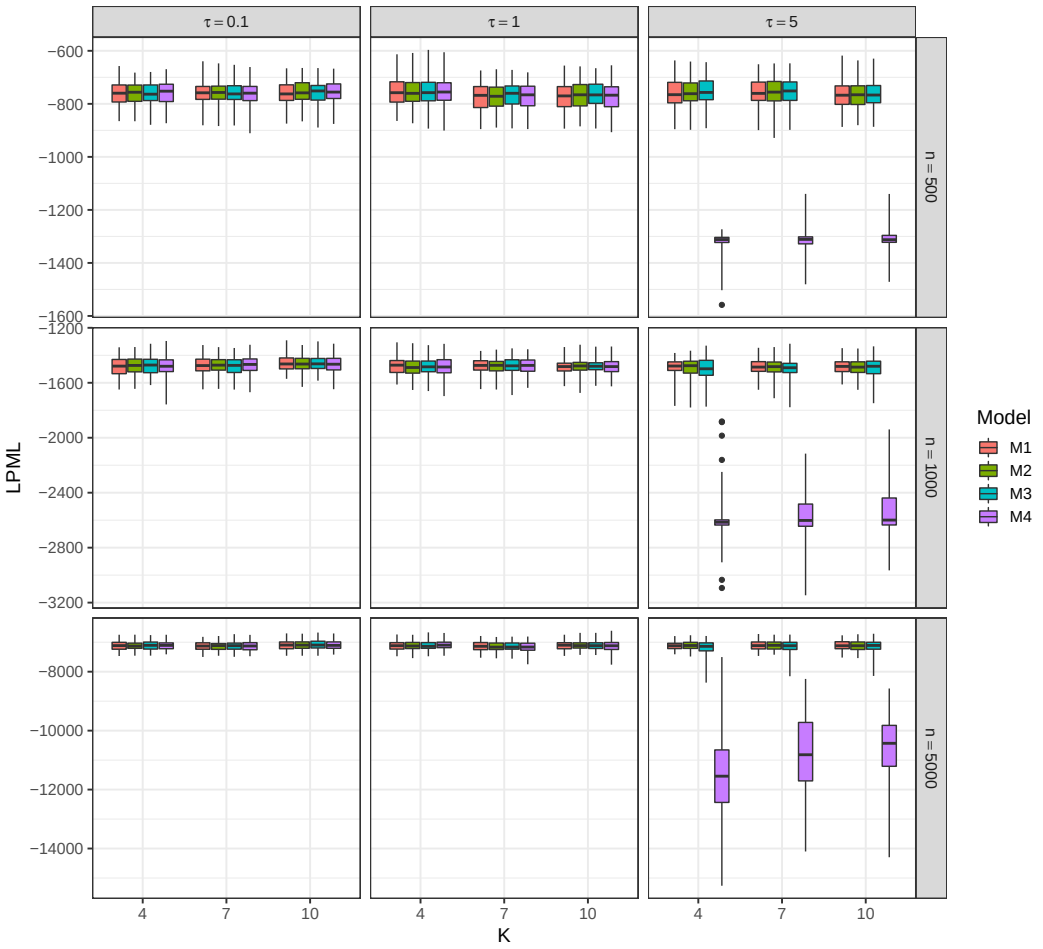

Figure 2: LPML values associated with each of the four models considered in the simulation study. Recall that model M1 does not depend on τ. For this reason LPML values for model M1 remain constant for different values of τ save for Monte Carlo error.

error). Now, notice that the four models provide very similar fits (see Figure 2) based on the logarithm of the pseudo-marginal likelihood (LPML) values (Geisser and Eddy; 1979). The lone exception is M4 where $\tau = 5$ clearly introduces too much repulsion for the hard-core potential. However, even though the repulsive mixtures have similar model fits, the relative distance between component centers is increased relative to the i.i.d. mixture model. This can be seen in Figure 3 where as K grows, the relative distance between component means increases, but remains relatively the same for the repulsive mixtures. Once again, $\tau = 5$ introduces quite a bit of repulsion for the hard-core potential which forces the mixture to be comprised of only one component in "smaller" sample sizes. Further, note that the repulsion employed in Petralia et al. (2012) is stronger than the "soft" repulsion that accompanies the potential of Ogata & Tanemura, resulting in larger relative distances between component means.

We consider next the sum of weights associated with the "extra" mixture components. Results are displayed in Figure 4. Note that for all models the sum of "extra" component weights tends towards zero as the sample size grows. However, the rate at which the sum converges to zero is much faster for M2–M4 with M4 showing the fastest rate (which is to be expected). Thus, Figure 4 seems to empirically corroborate that the theory found in Rousseau and Mengersen (2011) applies to our framework (with a faster convergence rate), even though we do not yet have an analogous result for our context.

Finally, we also considered a bivariate mixture with clusters having substantial overlap. We found that the relative differences between models remained the same, and so we omit these results.

5 Discussion

We have studied a general framework based on Gibbs measures under which a class of probability models that explicitly parameterizes repulsion can be constructed. We discuss how different types of repulsion can be accommodated in this class by suitably choosing potentials. We also argue that soft repulsion provides a nice balance between the desire to remove redundant (or singleton) clusters that often appear when modeling location parameters of a mixture model independently, and the forced parsimony that

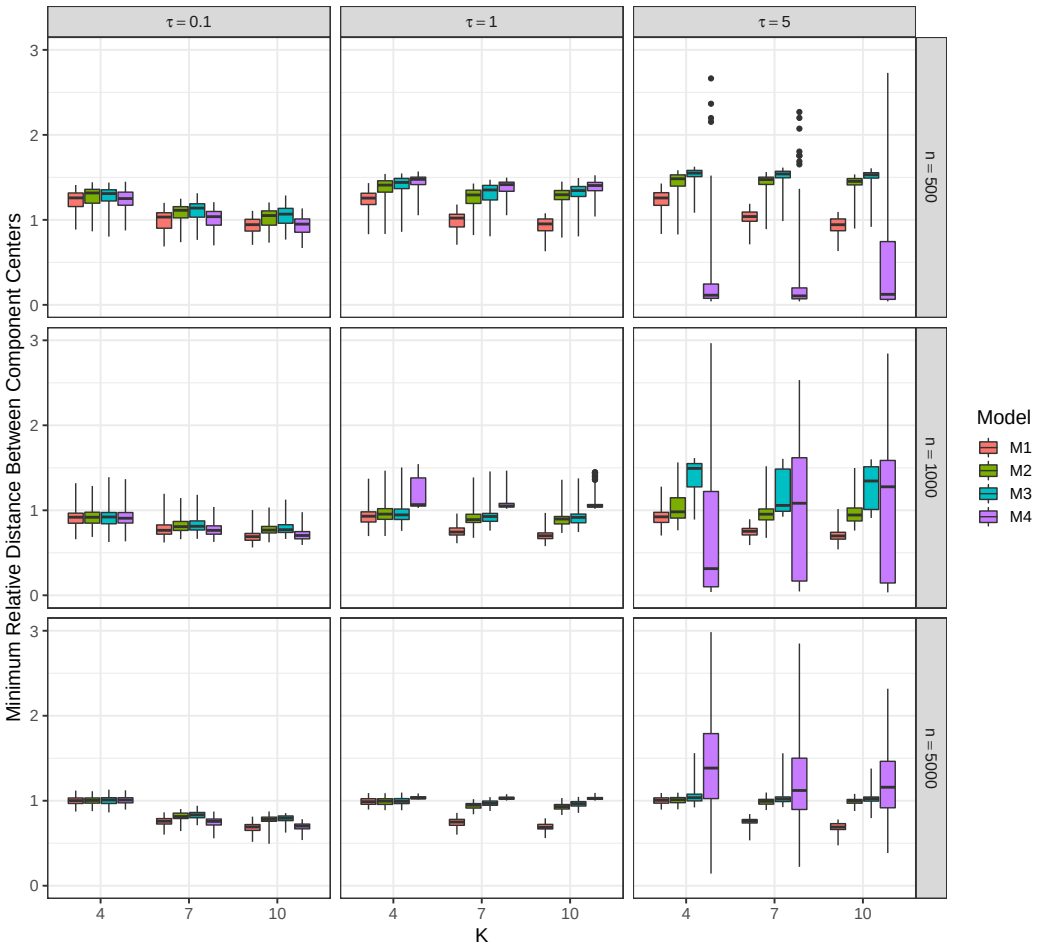

Figure 3: Minimum relative Euclidean distance between cluster centers associated with each of the four models considered in the simulation study. Recall that model M1 does not depend on τ. For this reason the minimum relative distances between two clusters remains constant for M1 for the different values of τ save for Monte Carlo error.

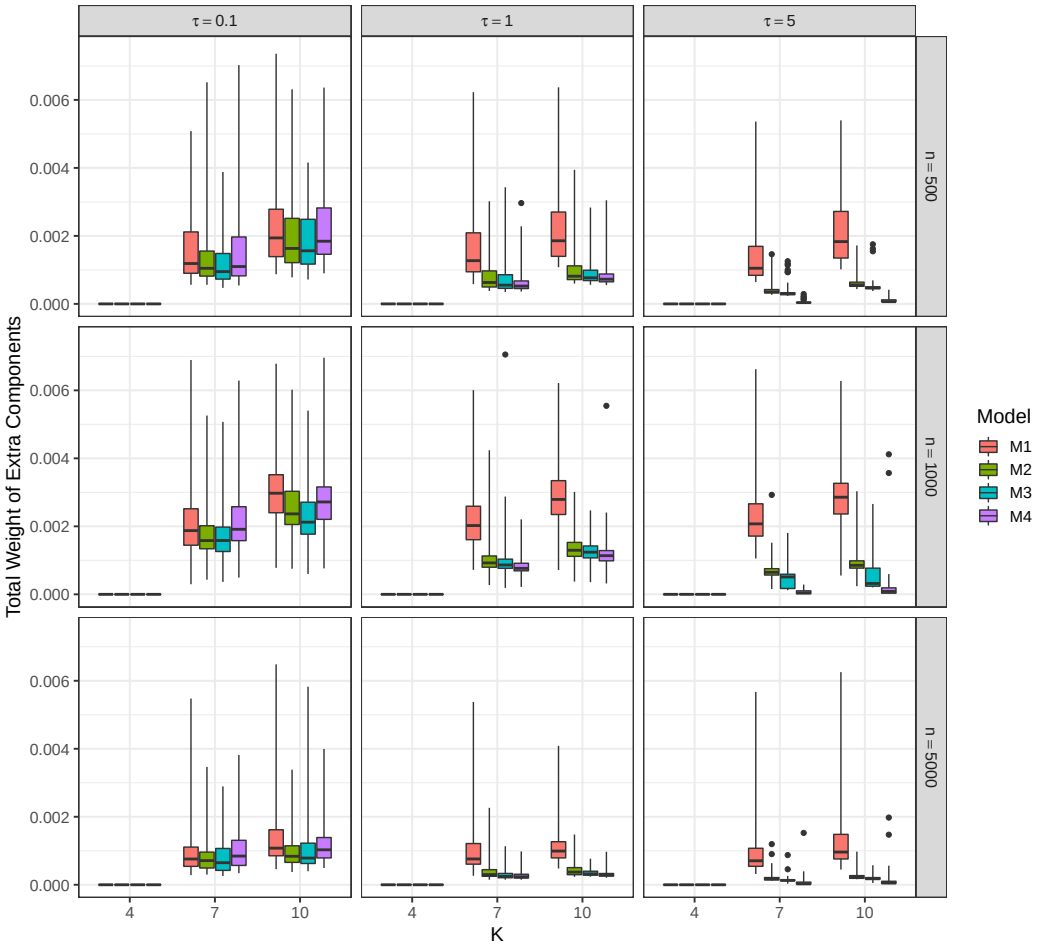

Figure 4: Sum of the $K - 4$ smallest component weights (recall $k_0 = 4$). Note that since model M1 does not depend on τ, the sum of "extra" component weights remains constant for M1 as a function of τ save for Monte Carlo error.

occurs with hard types of repulsion.

We studied properties of the models and developed theory in a similar way as Petralia et al. (2012) tailoring it to the potential function and model specifications we considered. Our approach shares the same modeling spirit (presence of repulsion) as in Xu et al. (2016), Fúquene et al. (2019), and Xie and Xu (2020). However, the specific mechanism we propose to model repulsion differs from these works. Xu et al. (2016) is based on determinantal point processes, which introduces repulsion through the determinant of a matrix driven by a Gaussian covariance kernel. Our approach introduces a parameter that directly influences repulsion strength which is easier to conceptualize. We emphasize that our approach exploits the randomness induced by the prior on the number of clusters, which is indeed random despite the fact that the maximum number of components is fixed. The work by Fúquene et al. (2019) defines a family of probability densities that promotes well-separated location parameters through a penalization function, that cannot be re-expressed as a (pure) repulsive potential. However, for small relative distances, the penalization function can be identified as an interaction potential that produces repulsion similar to that found in Petralia et al. (2012) (i.e., a hard type of repulsion). Finally, Xie and Xu (2020) consider a Bayesian nonparametric model that is computationally demanding and employs a different repulsive specification.

The Gaussian mixture model can be extended to other component-specific kernel, $q(\cdot;\theta)$, say, where $\theta \in \Theta$, a strict subset of $\mathbb{R}^d$. Indeed, by changing the parametrization to $\varphi = h(\theta)$, where $h : \Theta \to \mathbb{R}^d$ is a one-to-one function with continuous derivatives, we can adopt the same prior definition (13)-(15). In addition, it might prove beneficial to employ a prior distribution for the mixture component weights that encourages sparsity (Heiner et al. 2019, Blasi et al. 2020). However, studying properties of such a model extension is beyond the scope of this article.

In practice τ in (17) is unknown and needs to be specified or estimated. Because treating τ as an unknown and assigning it a prior renders the MCMC algorithm detailed in Sections B and C of the Supplementary Material doubly intractable, we suggest fixing τ at a specific value. To this end, we devised a straightforward procedure that permits calibrating τ. Details are provided in Section D of the Supplementary Material.

Finally, an application of this line of modeling to conditional density regression was already presented in Quinlan et al. (2018). In addition, we are currently studying the

application of similar strategies to the construction of clusters in situations where there is a specific motivation for considering well-separated clusters.

Acknowledgments

We thank the anonymous referees and the associated editor for valuable comments that greatly improved this work. Also, we'd like to thank Gregorio Moreno and Duvan Henao for helpful conversations and comments. José J. Quinlan gratefully recognizes the support provided by CONICYT through Fondecyt Grant 3190324 and CONACyT Grant 241195. Fernando A. Quintana was supported by Fondecyt Grant 1180034. This work was supported by Millennium Science Initiative of the Ministry of Economy, Development, and Tourism, grant "Millenium Nucleus Center for the Discovery of Structures in Complex Data".

References

Argiento, R. and De Iorio, M. (2019). Is infinity that far? A Bayesian nonparametric perspective of finite mixture models, *arXiv e-prints* .

Bianchini, I., Guglielmi, A. and Quintana, F. A. (2020). Determinantal point process mixtures via spectral density approach, *Bayesian Analysis* **15**(1): 187–214.

Blasi, P. D., Martínez, A. F., Mena, R. H. and Prünster, I. (2020). On the inferential implications of decreasing weight structures in mixture models, *Computational Statistics and Data Analysis* **147**(1): 106940.

Daley, D. and Vere-Jones, D. (2002). *An Introduction to the Theory of Point Processes*, Vol. I: Elementary Theory and Methods, second edn, Springer-Verlag, New York.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems, *The Annals of Statistics* **1**: 209–30.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*, Springer Series in Statistics, Springer, New York.

Fúquene, J., Steel, M. and Rossell, D. (2019). On choosing mixture components via non-local priors, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81**(5): 809–837.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection, *Journal of the American Statistical Association* **74**(365): 153–160.

Georgii, H.-O. and Yoo, H. J. (2005). Conditional intensity and gibbsianness of determinantal point processes, *Journal of Statistical Physics* **118**(1): 55–84.

Ghosal, S. and van der Vaart, A. (2007). Posterior convergence rates of dirichlet mixtures at smooth densities, *The Annals of Statistics* **35**(2): 697–723.

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities, *The Annals of Statistics* **29**(5): 1233–1263.

Heiner, M., Kottas, A. and Munch, S. B. (2019). Structured priors for sparse probability vectors with application to model selection in markov chains, *Statistics and Computing* pp. 1–17.

Illian, J., Penttinen, A., Stoyan, H. and Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*, Statistics in Practice, John Wiley & Sons, Ltd., Chichester.

Jones, J. E. (1924). On the determination of molecular fields. ii. from the equation of state of a gas, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **106**: 463–477.

Lavancier, F., Möller, J. and Rubak, E. (2015). Determinantal point processes models and statistical inference, *Journal of the Royal Statistical Society: Series B* **77**(4): 853–877.

Mateu, J. and Montes, F. (2000). Approximate maximum likelihood estimation for a spatial point pattern, *Questiio: Quaderns d'Estadistica, Sistemes, Informatica i Investigació Operativa* **24**(1): 3–25.

Ogata, Y. and Tanemura, M. (1981). Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure, *Annals of the Institute of Statistical Mathematics* **33**(2): 315–338.

Ogata, Y. and Tanemura, M. (1985). Estimation of interaction potentials of marked spatial point patterns through the maximum likelihood method, *Biometrics* **41**(2): 421–433.

Papangelou, F. (1974). The conditional intensity of general point processes and an application to line processes, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **28**(3): 207–226.

Pathria, R. K. and Beale, P. D. (2011). Statistical mechanics of interacting systems: The method of cluster expansions, *in* R. Pathria and P. D. Beale (eds), *Statistical Mechanics*, third edn, Academic Press, pp. 299–343.

Petralia, F., Rao, V. and Dunson, D. B. (2012). Repulsive mixtures, *in* F. Pereira, C. Burges, L. Bottou and K. Weinberger (eds), *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc., pp. 1889–1897.

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme, *Statistics, probability and game theory*, Vol. 30 of *IMS Lecture Notes Monogr. Ser.*, Inst. Math. Statist., Hayward, CA, pp. 245–267.

Pitman, J. and Yor, M. (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator, *The Annals of Probability* **25**(2): 855–900.

Quinlan, J. J., Page, G. L. and Quintana, F. A. (2018). Density regression using repulsive distributions, *Journal of Statistical Computation and Simulation* **88**(15): 2931–2947.

Quintana, F. A. (2006). A predictive view of bayesian clustering, *Journal of Statistical Planning and Inference* **136**(8): 2407–2429.

Rao, V., Adams, R. P. and Dunson, D. D. (2017). Bayesian inference for matérn repulsive processes, *Journal of the Royal Statistical Society: Series B* **79**(3): 877–897.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(5): 689–710.

Scricciolo, C. (2011). Posterior rates of convergence for dirichlet mixtures of exponential power densities, *Electronic Journal of Statistics* **5**: 270–308.

Shen, W., Tokdar, S. T. and Ghosal, S. (2013). Adaptive bayesian multivariate density estimation with dirichlet mixtures, *Biometrika* **100**(3): 623–640.

Strauss, D. J. (1975). A model for clustering, *Biometrika* **62**(2): 467–475.

Xie, F. and Xu, Y. (2020). Bayesian repulsive gaussian mixture model, *Journal of the American Statistical Association* **115**: 187–203.

Xu, Y., Müller, P. and Telesca, D. (2016). Bayesian inference for latent biological structure with determinantal point processes (dpp), *Biometrics* **72**(3): 955–964.