Bayesian Analysis (2020) **15**, Number 1, pp. 187–214

Determinantal Point Process Mixtures Via Spectral Density Approach

Ilaria Bianchini*, Alessandra Guglielmi† and Fernando A. Quintana‡

Abstract. We consider mixture models where location parameters are a priori encouraged to be well separated. We explore a class of determinantal point process (DPP) mixture models, which provide the desired notion of separation or repulsion. Instead of using the rather restrictive case where analytical results are partially available, we adopt a spectral representation from which approximations to the DPP density functions can be readily computed. For the sake of concreteness the presentation focuses on a power exponential spectral density, but the proposed approach is in fact quite general. We later extend our model to incorporate covariate information in the likelihood and also in the assignment to mixture components, yielding a trade-off between repulsiveness of locations in the mixtures and attraction among subjects with similar covariates. We develop full Bayesian inference, and explore model properties and posterior behavior using several simulation scenarios and data illustrations. Supplementary materials for this article are available online (Bianchini et al., 2019).

Keywords: density estimation, nonparametric regression, repulsive mixtures, reversible jumps.

1 Introduction

1.1 Mixture models

Mixture models are an extremely popular class of models, that have been successfully used in many applications. For a review, see, e.g. Frühwirth-Schnatter (2006). Such models are typically stated as

$$y_i \mid k, \boldsymbol{\theta}, \boldsymbol{\pi} \overset{\text{iid}}{\sim} \sum_{j=1}^{k} \pi_j f(y_i \mid \theta_j), \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ are constrained to be nonnegative and sum up to 1, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, and $1 \leq k \leq \infty$, with $k = \infty$ corresponding to a nonparametric model. A common prior assumption is that $\boldsymbol{\pi} \sim \text{Dirichlet}(\delta_1, \ldots, \delta_k)$ and that the components of $\boldsymbol{\theta}$ are drawn i.i.d. from some suitable prior p_0. However, the weights $\boldsymbol{\pi}$ may be constructed differently, e.g. using a stick-breaking representation (finite or infinite), which poses a well-known connection with more general models, including nonparametric ones. See, e.g., Ishwaran and James (2001) and Miller and Harrison (2017). A popular class

*Politecnico di Milano, ilaria.bianchini@polimi.it
†Politecnico di Milano, alessandra.guglielmi@polimi.it
‡Pontificia Universidad Católica de Chile, quintana@mat.uc.cl

 https://doi.org/10.1214/19-BA1150

of Bayesian nonparametric models is the Dirichlet process mixture (DPM) model, introduced in Ferguson (1983) and Lo (1984). It is well-known that this class of mixtures usually overestimates the number of clusters, mainly because of "the rich gets richer" property of the Dirichlet process. By this we mean that both prior and posterior distributions are concentrated on a relatively large number of clusters, but a few are very large while the rest have very small sizes. Mixture models may even be inconsistent; see Rousseau and Mengersen (2011), where concerns about over-fitted mixtures are illustrated, and Miller and Harrison (2013), for inconsistency features of DPMs.

Despite their success, models like (1) tend to use excessively many mixture components. Xu et al. (2016) point out that this is due to the fact that the component-specific parameters are a priori i.i.d., and therefore, free to move. This motivated Petralia et al. (2012), Fúquene et al. (2016) and Quinlan et al. (2017) to explicitly define joint distributions for $\boldsymbol{\theta}$ having the property of *repulsion* among its components, i.e. $p(\theta_1, \ldots, \theta_k)$ puts higher mass on configurations with well separated components. See also Quinlan et al. (2018) for an application to density estimation. For a different approach, via sparsity in the prior, see Malsiner-Walli et al. (2016).

Xu et al. (2016) explored a similar way to accomplish separation of mixture components, by means of a Determinantal Point Process (DPP) acting on the parameter space. DPPs have recently received increased attention in the statistical literature (Lavancier et al., 2015). DPPs are point processes having a product density function expressed as the determinant of a certain matrix constructed using a covariance function, evaluated at the pairwise distances among points, in such a way that higher mass is assigned to configurations of well-separated points. We give details below. DPPs have been used to make inference mostly on spatial data. Bardenet and Titsias (2015) and Affandi et al. (2014) applied DPPs to model spatial patterns of nerve fibers in diabetic patients, a basic motivation being that such fibers become more clustered as diabetes progresses. The latter discussed also applications to image search, showing how such processes could be used to study human perception of diversity in different image categories. Similarly, Kulesza and Taskar (2012) show how DPPs can be applied to various problems that are relevant to the machine learning community, such as finding diverse sets of high-quality search results, building informative summaries by selecting diverse sentences from documents, modeling non-overlapping human poses in images or videos, and automatically building timelines of important news stories. More recently, Shirota and Gelfand (2017) have described an approximate Bayesian computation method to fit DPPs to spatial point pattern data. Historically, the first paper where DPPs were adopted as a prior for statistical inference in mixture models is Affandi et al. (2013). The statistical literature also includes a number of papers illustrating theoretical properties for estimators of DPPs from a non-Bayesian viewpoint; see, for instance, Biscio and Lavancier (2016, 2017) and Bardenet and Titsias (2015).

1.2 Main contributions of this work

We discuss full Bayesian inference for a class of mixture densities where the locations follow stationary DPPs. Our first contribution is the introduction of an approach that generalizes and extends the model studied in Xu et al. (2016) who base their analysis

on a special, finite state space case of DPPs called L-ensembles, and with a particular closed-form expression for the determinant, in terms of a Gaussian covariance function. Instead, we resort to the spectral representation of the covariance function defining the determinant as the joint distribution of component-specific parameters. Our methods can thus be used with any such valid spectral representation, as described by Lavancier et al. (2015), which implies great generality of the proposal, that goes well beyond the Gaussian covariance function with a squared exponential kernel case of Xu et al. (2016). The extensions considered here are stated in the context of both uni- and multi-dimensional responses, and with different spectral densities. For brevity of presentation, we defer details of computational implementation to Section 1 of the on-line Supplementary Material file. For the sake of concreteness, our illustrations focus on the case of power exponential spectral representation; see examples with different spectral densities in Section 2 of the on-line Supplementary Material. This particular specification allows for flexible repulsion patterns, and we discuss how to set up different types of prior behavior, shedding light on the practical use of our approach in that particular scenario. Although we limit ourselves to the case of isotropic DPPs, inhomogeneous DPPs can be obtained by transforming or thinning a stationary process (Lavancier et al., 2015). A crucial point in our models and algorithms is the DPP density expression with respect to the unit rate Poisson process. This density is only defined for DPPs restricted to compact subsets S of the state space, and if it exists, it explicitly depends on S. A sufficient condition for the existence of the density is that all the eigenvalues of (Mercer's Theorem representation of) covariance function, restricted to S, are smaller than 1. We follow the spectral approach and assume that the covariance function defining the DPP has a spectral representation. One basic motivation for our choice is that conditions for the existence of a density become easier to check. Another motivation is the flexibility in assuming different degrees of repulsion in the DPP by way of the spectral density. We review here the basic theory on DPPs, making an effort to be as clear and concise as possible in the presentation of our subsequent models. We discuss applications in the context of synthetic and real data applications.

A second contribution of this work is the extension of the proposed spectral DPP model to incorporate covariate information in the likelihood and also in the assignment to mixture components. In particular, subjects with similar covariates are a priori more likely to co-cluster, just as in mixtures of experts models (see, e.g., McLachlan and Peel, 2005), where weights are defined as normalized exponential functions. From a computational viewpoint, a third contribution of our work is the generalization of the reversible jump (RJ) MCMC posterior simulation scheme proposed by Xu et al. (2016) to the general spectral approach and also to the covariate-dependent extensions we consider. We discuss two RJ MCMC versions, one for each of uni- and multi-variate responses. In all cases the algorithms require computing the DPP density with respect to the unit rate Poisson process. We explain how to carry out the calculations, and discuss the required restriction of the process to compact subsets. When extending the model to incorporate covariate information in both likelihood and prior assignment to mixture components, we discuss how to modify the RJ MCMC algorithms; see Section 3 of the on-line Supplementary Material.

Although not our primary target, a sub-product of the approach is the estimation of a partition in the sample. Specifically, we consider the partition minimizing the posterior

expectation of Binder's loss function (Binder, 1978) under equal misclassification costs, a common choice in the applied Bayesian nonparametric literature (Lau and Green, 2007). Nevertheless, we emphasize one conceptual advantage of the separation induced by the prior assumption, namely, the reduction in the estimated number of clusters a posteriori compared to the usual mixture models without this feature. The fact that our prior generates components from a repulsive process, automatically penalizes redundancy, thus inducing fewer clusters than if they were allowed to be independently selected. Reducing the effective number of clusters a posteriori helps scaling up our model – better than alternatives with no separation – when the sample size grows. We discuss this particular point in our data illustrations.

We also provide comparison with alternative models, via predictive goodness-of-fit tools. We specifically consider some "natural competitors", namely DPMs and the repulsive mixtures such as those introduced in Petralia et al. (2012) and Quinlan et al. (2017), as well as finite mixture models, as implemented in the package `Mclust` (Fraley et al., 2012). We find that our model is a competent alternative to those previously listed.

1.3 Outline

The rest of this article is organized as follows. Section 2 presents notation and theoretical background necessary to understand how DPP mixture models are constructed; we also sketch alternative repulsive models already introduced in the literature, later to be used as a comparison in the context of specific illustrative examples. Section 3 illustrates the covariate-dependent extension. Here we build on regular mixture models, incorporating covariate dependence in the mixture weights and optionally in the likelihood, which still allows for repulsion among components after correcting for the regression effect. Section 4 presents results from a simulation study and for unidimensional and bidimensional reference datasets, and an application to Biopic data is discussed in Section 5. We conclude in Section 6 with final comments and discussion. The on-line Supplementary Material contains a description of the two RJ MCMC algorithms, additional illustrative examples based on the well-known Galaxy dataset, supplemental figures, and additional application to Air Quality Index data.

2 Using DPPs to induce repulsion

We review here the basic theory on DPPs to the extent required to explain our mixture model. We use the same notation as in Lavancier et al. (2015), where further details on this theory may be found.

2.1 Basic theory on DPPs

Let $B \subseteq \mathbb{R}^d$; we mainly consider the cases $B = \mathbb{R}^d$ and $B = S$, a compact subset in $\mathbb{R}^d$. Let X be a simple locally finite spatial point process defined on B, i.e. the number of points of the process in any bounded region is a finite random variable, and

there is at most one point at any location. See Daley and Vere-Jones (2003; 2007) for a general presentation on point processes. The class of DPPs we consider is defined in terms of their moments, expressed by their product density functions $\rho^{(n)} : B^n \to [0, +\infty)$, $n = 1, 2, \ldots$. Intuitively, for any pairwise distinct points $x_1, \ldots, x_n \in B$, $\rho^{(n)}(x_1, \ldots, x_n)dx_1 \cdots dx_n$ is the probability that X has a point in an infinitesimal small region around x_i of volume dx_i, for each $i = 1, \ldots, n$. More formally, X has n-th order product density function $\rho^{(n)} : B^n \to [0, +\infty)$ if this function is locally integrable (i.e. $\int_S |\rho^{(n)}(x)|\mathrm{d}x < +\infty$ for any compact S) and, for any Borel-measurable function $h : B^n \to [0, +\infty)$,

$$\mathbb{E}\left(\sum_{x_1,\ldots,x_n \in X}^{\neq} h(x_1, \ldots, x_n)\right) = \int_{B^n} \rho^{(n)}(x_1, \ldots, x_n)h(x_1, \ldots, x_n)dx_1 \cdots dx_n,$$

where the $\neq$ sign over the summation means that $x_1, \ldots, x_n$ are pairwise distinct. See also Møller and Waagepetersen (2007).

Let $C : B \times B \to \mathbb{R}$ denote a covariance function. A simple locally finite spatial point process X on B is called a determinantal point process with kernel C if its product density functions are

$$\rho^{(n)}(x_1, \ldots, x_n) = det[C](x_1, \ldots, x_n), \ (x_1, \ldots, x_n) \in B^n,$$

for $n = 1, 2, \ldots$, where $[C](x_1, \ldots, x_n)$ is the $n \times n$ matrix with entries $C(x_i, x_j)$. We write $X \sim DPP_B(C)$; when $B = \mathbb{R}^d$ we write $X \sim DPP(C)$. Note that, if A is a Borel subset of B, then the restriction $X_A := X \cap A$ of X to A is a DPP with kernel given by the restriction of C to $A \times A$.

By Theorem 1 in Lavancier et al. (2015), first proved by Macchi (1975), such DPP's exist under the two following conditions:

- C is a continuous covariance function, hence, by Mercer's Theorem, for any compact subset S,

$$C(x, y) = \sum_{k=1}^{+\infty} \lambda_k^S \phi_k(x)\phi_k(y), \quad (x, y) \in S \times S,$$

  where $\{\lambda_k^S\}$ and $\{\phi_k(x)\}$ are the eigenvalues and eigenfunctions in Mercer's Theorem representation of C restricted to $S \times S$, respectively;

- $\lambda_k^S \leq 1$ for all compact S in $\mathbb{R}^d$ and all k.

Note that, since $\rho^{(n)}(x_1, \ldots, x_n)$ defined above has the interpretation of the density of a point $(x_1, \ldots, x_n)$, and the determinant of the matrix $[C](x_1, \ldots, x_n)$ is the hypervolume of the parallelepiped spanned by the columns of this matrix, the density grows as the columns become more separated. Continuity of the covariance function C implies that $\rho^{(n)}(x_1, \ldots, x_n)$ converges to 0 if $x_i \to x_j$ for some $i \neq j$. Moreover, since the determinant of $[C](x_1, \ldots, x_n)$ is less than or equal to the product of its diagonal

elements, then $\rho^{(n)}(x_1,\ldots,x_n) \leq \rho(x_1)\cdots\rho(x_n)$, where $\rho(x) = \rho^{(1)}(x) = C(x,x)$ is the intensity function; this can be interpreted as follows: the joint probability of any configuration of points is less than if they were placed independently from each other placement. All this motivates the use of the term *repulsive* processes for determinantal point processes; see Lavancier et al. (2015), Section 2.2.

Formula (2.10) in Lavancier et al. (2015) reports the distribution of the number $N(S)$ of points of X in S, for any compact S:

$$N(S) \stackrel{d}{=} \sum_{k=1}^{+\infty} B_k, \qquad \mathbb{E}(N(S)) = \sum_{k=1}^{+\infty} \lambda_k^S, \qquad \mathrm{Var}(N(S)) = \sum_{k=1}^{+\infty} \lambda_k^S(1-\lambda_k^S), \tag{2}$$

where $B_k \stackrel{\text{ind}}{\sim} Be(\lambda_k^S)$, i.e. the Bernoulli random variable with mean λ_k^S. In practice, all these summations have to be numerically evaluated. When restricted to any compact subset S, the DPP has a density with respect to the unit rate Poisson process which, when $\lambda_k^S < 1$ for all $k = 1, 2, \ldots$, has the following expression:

$$f(\{x_1,\ldots,x_n\}) = \mathrm{e}^{|S|-D_S} det[\widetilde{C}](x_1,\ldots,x_n), \tag{3}$$

for $n = 1, 2, \ldots$, where $|S| = \int_S \mathrm{d}x$, $D_S = -\sum_1^{+\infty} \log(1-\lambda_k^S)$ and

$$\widetilde{C}(x,y) = \sum_1^{+\infty} \frac{\lambda_k^S}{1-\lambda_k^S} \phi_k(x)\phi_k(y), \qquad x, y \in S. \tag{4}$$

When $n = 0$ the density (as well as the determinant) is defined to be equal to 0. See Møller and Waagepetersen (2007) for a thorough definition of absolute continuity of a spatial process with respect to the unit rate Poisson process. However, note that from the first part of (2) we have $\mathbb{P}(N(S) = 0) = \prod_{k=1}^{+\infty}(1-\lambda_k^S)$; this probability could be positive due to the assumption $\lambda_k^S < 1$ for all $k = 1, 2, \ldots$.

From now on we restrict our attention to stationary DPP's, that is, when $C(x,y) = C_0(x-y)$, where $C_0 \in L^2(\mathbb{R}^d)$ is such that its spectral density φ exists, i.e.

$$C_0(x) = \int_{\mathbb{R}^d} \varphi(y)\cos(2\pi x \cdot y)\mathrm{d}y, \quad x \in \mathbb{R}^d$$

and $x \cdot y$ is the scalar product in $\mathbb{R}^d$. If $\varphi \in L^1(\mathbb{R}^d)$ and $0 \leq \varphi \leq 1$, then the $DPP(C)$ process exists. Summing up, the distribution of a stationary DPP can be assigned by its spectral density; see Corollary 1 in Lavancier et al. (2015). Notice that C_0, which defines the covariance function to be isotropic, is the Fourier transform of φ, so that they are in one-to-one correspondence. Therefore, we can indistinctly identify one by the other, which explains why we can define the DPP by φ.

To explicitly evaluate (3) over $S = \left[-\frac{1}{2}, \frac{1}{2}\right]^d$, we approximate $\widetilde{C}$ as suggested in Lavancier et al. (2015). In other words, we approximate the density of X on S by

$$f^{app}(\{x_1,\ldots,x_n\}) = \mathrm{e}^{|S|-D_{app}} det[\widetilde{C}_{app}](x_1,\ldots,x_n), \tag{5}$$

where $\{x_1, \ldots, x_n\} \subset S$ and

$$\widetilde{C}_{app}(x,y) = \widetilde{C}_{app,0}(x-y) = \sum_{k\in\mathbb{Z}^d} \left[\frac{\varphi(k)}{1-\varphi(k)}\right] \cos(2\pi k\cdot(x-y)),\ x,y\in S, \tag{6}$$

$$D_{app} = \sum_{k\in\mathbb{Z}^d} \log\left(1+\frac{\varphi(k)}{1-\varphi(k)}\right). \tag{7}$$

To understand why the approximation $C(x,y) \approx C_{app,0}(x-y)$ for $x-y\in S$ follows, as well as the corresponding approximation for the tilted versions of these functions, we observe that the exact Fourier expansion of $C_0(x-y)$ in S is as in (6) with the real part of $\int_S C_0(y)\mathrm{e}^{-2\pi i k\cdot y}\mathrm{d}y$ instead of $\varphi(k)$; assuming C_0 is such that $C_0(t)\approx 0$ for $t\notin S$, then

$$Re\left(\int_S C_0(y)\mathrm{e}^{-2\pi i k\cdot y}\mathrm{d}y\right) \approx \varphi(k) := Re\left(\int_{\mathbb{R}^d} C_0(y)\mathrm{e}^{-2\pi i k\cdot y}\mathrm{d}y\right).$$

See also Lavancier et al. (2015), Section 4.1. See Figure 1 in the Supplementary Material, where we display an example of the function C_0. Note that the above approximation implies that λ_k^S is numerically approximated by $\varphi(k)$. Indeed, all the quantities in formula (2) can be obtained in this way.

When $S = R$ is a rectangle in $\mathbb{R}^d$, we can always find an affine transformation T such that $T(R) = S = \left[-\frac{1}{2},\frac{1}{2}\right]^d$. Define $Y = T(X)$. If f_Y^{app} is the approximate density of Y as in (5), we can then approximate the density of X_R by

$$f^{app}(\{x_1,\ldots,x_n\}) = |R|^{-n}\mathrm{e}^{|R|-|S|} f_Y^{app}(T(\{x_1,\ldots,x_n\})), \tag{8}$$

for $\{x_1,\ldots,x_n\}\subset R$. In practice, the summation over $\mathbb{Z}^d$ in (6) above is truncated to $\mathbb{Z}_N^d$, where $\mathbb{Z}_N := \{-N,-N+1,\ldots,0,\ldots,N-1,N\}$ (see Section 4.3 in Lavancier et al., 2015).

As mentioned in the Introduction, one particular example of spectral density that we found useful is

$$\varphi(x;\rho,\nu) = s^d \exp\left\{-\left(\frac{s}{\sqrt{\pi}}\right)^{\nu}\left(\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{\nu}+1)}\right)^{\nu d}\rho^{\nu d}\|x\|^{\nu}\right\}, \rho,\nu>0, \tag{9}$$

for fixed $s\in(0,1)$ (e.g. $s=\frac{1}{2}$) and $\|x\|$ is the Euclidean norm of $x\in\mathbb{R}^d$. This function is the spectral density of a *power exponential spectral model*. See (3.7) in Lavancier et al. (2015) setting $\alpha = s\ \alpha_{max}(\rho,\nu)$. In this case, we write $X\sim PES-DPP(\rho,\nu)$. The corresponding spatial process is isotropic. When $\nu=2$, the spectral density is

$$\varphi(x;\rho,\nu) = s^d\exp\left\{-\frac{s^2\rho^{2d}}{\sqrt{\pi}}\|x\|^2\right\}, \qquad \rho>0,$$

corresponding to the Gaussian spectral density. We discuss more specifically the choice of (9) later in Section 4.

2.2 The mixture model with repulsive means

To deal with limitations of model (1) or DPMs, we consider repulsive mixtures. Our aim is to estimate a random partition of the available subjects, and we want to do so using "few" groups. By repulsion we mean that cluster locations are a priori encouraged to be well separated, thus inducing fewer clusters than if they were allowed to be independently selected. We start from parametric densities $f(\cdot;\theta)$, which we take to be Gaussian, and assume that the collection of location parameters follows a DPP. We specify a hierarchical model that achieves the goals previously described. Concretely, we propose:

$$y_i \mid s_i = k, \{\mu_k\}, \{\sigma_k^2\}, K \overset{\text{ind}}{\sim} \mathcal{N}\left(y_i; \mu_k, \sigma_k^2\right) \text{ for each } i \tag{10}$$

$$X = \{\mu_1, \mu_2, \ldots, \mu_K, K\} \sim PES - DPP(\rho, \nu) \tag{11}$$

$$(\rho, \nu) \sim \pi \tag{12}$$

$$p(s_i = k) = w_k, \ k = 1, \ldots, K \text{ for each } i \tag{13}$$

$$w_1, \ldots, w_K \mid K \sim \text{Dirichlet}(\delta, \delta, \ldots, \delta) \tag{14}$$

$$\sigma_k^{-2} \mid K \overset{\text{iid}}{\sim} \text{Gamma}(a_0, b_0), \tag{15}$$

where the PES-DPP(ρ, ν) assumption (11) is regarded as a default choice that could be replaced by any other valid DPP alternative. We stress the explicit role of K in formula (12), where K represents the (random) total number of points generated by the determinantal point process prior. By the approximation procedure described in Section 2.1, K represents the total number $N(S)$ of points in the compact set S, assumed to be large enough to contain all required latent cluster means (points); see (2). Recall also that in practice all we need is to work on the cube $[-1/2, 1/2]^d$ and then use a linear transformation of this cube onto a cube large enough to cover all required centers, as just described. The choice of π in (12) will be discussed below in Section 4. We note that, as stated, the prior model may assign a positive probability to the case $K = 0$. This case of course makes no sense from the viewpoint of the model described above. Nevertheless, we adopt the working convention of redefining the prior to condition on $K \geq 1$, i.e., truncating the DPP to having at least one point. In practice, the posterior simulation scheme later described simply ignores the case $K = 0$, which produces the desired result. Note also that we have assumed prior independence among blocks of parameters not involving the locations μ_k.

Model (10)-(15) is a DPP mixture model along the lines proposed in Xu et al. (2016). Indeed, we both use DPPs as priors for location points in the mixture of parametric densities. However, the specific DPP priors are different, as they restrict to a particular case of DPPs (L-ensembles) that require a finite state space, and choose a Gaussian covariance function for which eigenvalues and eigenfunctions in the corresponding Mercer's Theorem representation are analytically available. We adopt instead the spectral approach for assigning the prior (11), which implies great generality of the proposal. Similar to Xu et al. (2016), we carry out posterior simulation using a reversible jump step as part of the Gibbs sampler. However, when updating the location points $\mu_1, \ldots, \mu_K$

we refer to formulas (5)-(8). Xu et al. (2016) take advantage of the analytical expressions that we do not have for our case, and that are also unavailable in other possible specific choices of the spectral density.

As a general comment, we underline that the numerical evaluation of the DPP density, involving the computation of the determinant of a $K \times K$ matrix, is not particularly expensive, even in case of a large dataset; in this case, the repulsion property will favor a moderate number K of clusters. See Section 4.4, where we describe applications of this model to datasets, using the posterior simulation algorithms described in Section 1 in the Supplementary material. In our experience, the proposed model scales well compared to mixtures with independent components.

2.3 Competitor repulsive models

We briefly introduce the class of parsimonious mixture models in Quinlan et al. (2017), to be used as a competitor model for our applications. Quinlan et al. (2017) exploit the idea of repulsion, i.e. when any two mixture components are encouraged to be well separated, as we do. For the sake of comparison, we introduce their model for unidimensional data: similarly to our case, they consider a mixture of K Gaussian components, but assume a fixed value k for K in (10) and (13)-(15). The prior for the location parameters $\mu_1, \dots, \mu_k$ is called repulsive distribution, and denoted by $NRep_k(\mu, \Sigma, \tau)$, where $\mu \in \mathbb{R}$, $\Sigma, \tau > 0$; see (3.4)-(3.6) in Quinlan et al. (2017). This prior is characterized by a repulsion potential that assumes the following expression:

$$\phi_1(r; \tau) = -\log\left(1 - \mathrm{e}^{-\frac{1}{2\tau}r^2}\right) \mathbb{1}_{(0,+\infty)}(r), \quad \tau > 0 \ ;$$

Petralia et al. (2012) use a similar model, where the repulsion potential is

$$\phi_2(r; \tau) = \frac{\tau}{r^2} \mathbb{1}_{(0,+\infty)}(r), \quad \tau > 0 \ .$$

Potential ϕ_2 introduces a stronger repulsion than ϕ_1, in the sense that in Petralia et al. (2012), locations are encouraged to be further apart than in Quinlan et al. (2017). Note also that, by nature of the point process, our approach does not require an upper bound on the allowed number of mixture components (similar to DPM models), contrary to the approach in Quinlan et al. (2017) and Petralia et al. (2012).

All models under comparison for a specific application will be matched in such a way that they have the same prior expected number K of components in the mixtures.

3 Generalization to covariate-dependent models

The methods discussed in Section 2 were devised for density estimation-like problems. We now extend the previous modeling to the case where p-dimensional covariates $z_1, \dots, z_n$ are recorded as well. We do so by allowing the mixture weights to depend on such covariates. In this case, there is a trade-off between repulsiveness of locations in

the mixtures and attraction among subjects with similar covariates. We also entertain the case where covariate dependence is added to the likelihood part of the model. Our modeling choice here is akin to mixtures of experts models (see, e.g., McLachlan and Peel, 2005), i.e., the weights are defined by means of normalized exponential function.

Building on the model from Section 2.2, we assume the same likelihood (10) and the DPP prior for $X = \{\mu_1, \mu_2, \ldots, \mu_K, K\}$ in (11)-(12), but change (13) and (14) to

$$p(s_i = k) = w_k(z_i) \quad = \quad \frac{\exp\left(\beta_k^T z_i\right)}{\sum_{l=1}^{K} \exp\left(\beta_l^T z_i\right)}, \; k = 1, \ldots, K \tag{16}$$

$$\beta_2, \ldots, \beta_K \mid K \quad \overset{\text{iid}}{\sim} \quad \mathcal{N}_p\left(\beta_0, \Sigma_0\right), \quad \beta_1 = 0, \tag{17}$$

where the $\beta_1 = 0$ assumption is to ensure identifiability. To complete the model, we assume (15) as the conditional marginal for σ_k^2; the prior for (ρ, ν) in (12) is later specified. Here $\beta_0 \in \mathbb{R}^p$, and to choose Σ_0, we use a g-prior approach, namely $\Sigma_0 = \phi \times \left(Z^T Z\right)^{-1}$, where ϕ is fixed, typically of the same order of magnitude of the sample size (see Zellner, 1986). Here Z denotes the design matrix.

Assuming now (10) on top of (16)-(17) rules out the case of a likelihood explicitly depending on covariates, which instead would generally achieve a better fit than otherwise. Of course, there are many ways in which such dependence may be added. For the sake of concreteness, we assume here a Gaussian regression likelihood, where only the intercept parameters arise from the DPP prior. More precisely, we assume

$$y_i \mid s_i = k, z_i, \{\mu_k\}, \{\sigma_k^2\}, K \overset{\text{ind}}{\sim} \mathcal{N}\left(y_i; \mu_k + z_i^T \gamma_k, \sigma_k^2\right) \tag{18}$$

for all $i = 1, \ldots, n$, and

$$\gamma_i \mid K, \sigma_i^2 \overset{\text{ind}}{\sim} \mathcal{N}(\gamma_0, \sigma_i^2 \Lambda_0) \quad \text{and} \quad \sigma_i^{-2} \overset{\text{iid}}{\sim} \text{Gamma}(a_0, b_0), \tag{19}$$

where the γ_k's are p-dimensional regression coefficients, $\gamma_0 \in \mathbb{R}^p$ and Λ_0 is a covariance matrix. The prior for $\{s_i\}$ and β_j's is given in (16)-(17) as in the previous model. Note that (18) implies that only the intercept term is distributed according to the repulsive prior. Thus, we allow the response mean to be corrected by a linear combination of the covariates with cluster-specific coefficients, with the repulsion acting only on the residual of this regression. The result is a more flexible model than the repulsive mixture (10)-(15). Observe that there is no need to assume the same covariate vector in (18) and (16), but we do so for illustration purposes only.

The Gibbs sampler algorithm employed to carry out posterior inference for this model is detailed in Section 3 of the Supplementary material. However, it is worth noting that the reversible jump step related to updating the number of mixture components K and the update of the coefficients $\{\beta_2, \beta_3, \ldots, \beta_K\}$ are complicated by the presence of the covariates. For the β coefficients, we resort to a Metropolis-Hastings step, with a multivariate Gaussian proposal centered in the current value. For K, we employ an ad hoc Reversible Jump move.

4 Simulated data and reference datasets

Before illustrating the application of our models to specific datasets, we discuss some general choices that apply to all examples.

4.1 Spectral density and other general choices

As discussed earlier, our methods are quite general in the sense that they can be used with any valid spectral density. Practical implementation, however, requires choosing one such function, and the power exponential function in (9) is a convenient and flexible choice. In this case, note that $\varphi(x;\rho,\nu) < 1$ when $0 < s < 1$ for any $x \in \mathbb{R}^d$, $\rho, \nu > 0$, so that X_S has a density as described in (3). Figure 2 in the Supplementary material shows a plot of the power exponential spectral density (9) for different values of parameters ρ, ν. Note that ν controls the shape of $\varphi(x;\rho,\nu)$, which ranges from a slowly decreasing function of x when ν is small, to an indicator function when ν is large. On the other hand ρ plays the role that resembles a location parameter, with higher values retarding the decay speed of $\varphi(x;\rho,\nu)$. As discussed in Section 2.1, knowledge about the spectral density is all that is needed for the approximations to work. Moreover, even if the analytic expression of $C(x,y), (x,y) \in \mathbb{R}^d \times \mathbb{R}^d$ is known, as in, e.g. (9), we still need to compute the eigenvalues and eigenfunctions (of Mercer's Theorem representation) of C restricted to $S \times S$ for any compact S, and this may be analytically impossible. A potential disadvantage derived from this is that parameter interpretation in the spectral domain becomes unclear. We reckon this is a problem that arises for any particular parametrized version of φ one may choose.

Prior elicitation for (ρ,ν) is complicated due to the reasons stated above. Therefore, an extensive sensitivity analysis regarding $\pi(\rho,\nu)$ was carried out for the datasets considered below. See Sections 4.2 and 4.3. We point out that an initial prior independence assumption $\pi(\rho,\nu) = \pi(\rho)\,\pi(\nu)$ produced bad mixing of the chain. In particular, when ρ is small with respect to ν, the spectral function $\varphi(\cdot)$ has a very narrow support, concentrated near the origin, forcing the covariance function $\widetilde{C}_{app}(x,y)$ in (6) to become nearly constant for $x, y \in S$ and thus producing nearly singular matrices. We next investigated the case $\pi(\rho,\nu) = \pi(\rho \mid \nu) \times \pi_\nu(\nu)$, where

$$\rho \mid \nu \stackrel{d}{=} M(s,\varepsilon,\nu) + \rho_0, \quad \rho_0 \sim \text{Gamma}(a_\rho, b_\rho).$$

Here, $M(s,\varepsilon,\nu)$ is a constant that is the minimum value of ρ such that $\varphi(2) > \varepsilon$ (here $\varphi(2)$ is a reference value chosen to avoid a small support), and ε is a threshold value, assumed to be small (0.05, for instance). From Figure 2 in the Supplementary Material, it is clear that $\varphi(\cdot;\rho,\nu)$ goes to 0 too fast when ν is small relative to ρ. It follows that

$$M(s,\varepsilon,\nu) = \frac{2s\Gamma(1/\nu+1)\pi^{1/2}}{\Gamma(3/2)\left(\log\left(\frac{s}{\varepsilon}\right)\right)^{1/\nu}}.$$

We considered two different choices for π_ν: a gamma distribution, which gave a bad chain mixing, and a discrete distribution on $\mathcal{V}_2 = \{0.5, 1, 2, 3, 5, 10, 15, 20, 30, 50\}$ (or on

one of its subsets). In this case, the mixing of the chain was better, but the posterior for ν did not discriminate among the values in the support. For this reason, in Sections 4.5 and 5 and in Section 5 of the Supplementary Material, we assume $\nu = 2$, $s = 1/2$ and

$$\rho \stackrel{d}{=} \sqrt{\frac{\pi}{\log(\frac{1}{2\varepsilon})}} + \rho_0, \qquad \rho_0 \sim \text{Gamma}(a_\rho, b_\rho). \tag{20}$$

These choices guarantee that a reasonable number of nonzero terms are present when evaluating $\widetilde{C}_{app}(x, y)$, avoiding either too few, and hence a near singularity when evaluating (8), or too many, which will cause excessive computational burden.

In what follows, every run of the Gibbs sampler (implemented in `R`) produced a final sample size of 5,000 or 10,000 iterations (unless otherwise specified), after a thinning of 10 and initial burn-in of 5,000 iterations. In all cases, convergence was checked using both visual inspection of the chains and standard diagnostics available in the CODA package.

4.2 Data illustration on reference datasets without covariates

We illustrate our model via two datasets without covariates with unidimensional (Galaxy data) and bidimensional (Air Quality data) observations, both publicly available in `R` (`galaxy` from the `DPpackage` and `airquality` in the base version). For the latter data set we removed 42 incomplete observations.

The popular dataset Galaxy contains $n = 82$ measured velocities of different galaxies from six well-separated conic sections of space. Values are expressed in Km/s, scaled by a factor of 10^{-3}. We set the hyperparameters in this way: for the variance σ_k^2 of the components, $(a_0, b_0) = (3, 3)$ (such that the mean is 1.5 and the variance is 9/4) and for the weights $\{w_k\}$ the Dirichlet has parameter $(1, 1, \ldots, 1)$. The other hyperparameters are modified in the tests, as in Table 1, where we report summaries of interest, such as the prior and posterior mean and variance for the number of components K. In addition, we also display the mean squared error (MSE) and the log-pseudo marginal likelihood (LPML) as indices of goodness of fit, defined as $MSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $LPML = \sum_{i=1}^{n} \log\left(f(y_i \mid y^{(-i)})\right)$, where $\hat{y}_i$ is the posterior predictive mean and $f(y_i \mid y^{(-i)})$ is the $i-$th *conditional predictive ordinate*, that is the predictive distribution obtained using the dataset without the $i-$th observation. Figure 1 shows density estimates and the estimated partition of the data, obtained as the partition that minimizes the posterior expectation of Binder's loss function under equal misclassification costs (see Lau and Green, 2007). The points at the bottom of the plots represent observations, while colors refer to the corresponding cluster. See Figure 4 below for the posterior distribution of K for Test 4 and 6 in Table 1.

As a comparison, the same posterior quantities than in Table 1 were computed using the DPM, the Repulsive Gaussian Mixture Models (RGMM) by Quinlan et al. (2017), and also the proposal by Petralia et al. (2012). To make results comparable, we assumed the same prior information on hyperparameters common to all the mixture models. See Table 4. From these tables, it is clear that alternative repulsive models are

Test	ρ	ν	$\mathbb{E}(K)$	$Var(K)$	$\mathbb{E}(K\|data)$	$Var(K\|data)$	MSE	LPML
1	2	2	2	1.67	6.09	1.10	78.95	-171.72
2	5	10	5.00	7.12	6.07	1.09	78.33	-167.96
3	$a_\rho = 1, b_\rho = 1$	2	2.18	1.978	6.10	1.10	73.89	-164.47
4	$a_\rho = 1, b_\rho = 1$	10	2.73	2.15	6.11	1.12	74.93	-162.71
5	$a_\rho = 1, b_\rho = 1$	discr($\mathcal{V}_1$)	2.47	2.21	6.06	1.08	74.02	-172.54
6	$a_\rho = 1, b_\rho = 1$	discr($\mathcal{V}_2$)	2.51	2.27	6.10	1.13	76.64	-170.94

Table 1: Prior specification for (ρ, ν) and K and posterior summaries for the Galaxy dataset; (a_ρ, b_ρ) appear in (20); here $\mathcal{V}_1$ is $\{1, 2, 5, 10, 20\}$ and $\mathcal{V}_2 = \{0.5, 1, 2, 3, 5, 10, 15, 20, 30, 50\}$.

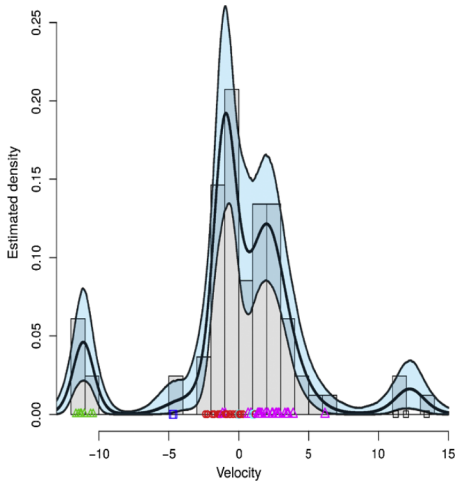

Figure 1: Density estimates and estimated partition for the Galaxy dataset under Test 4 in Table 1, including 90% credibility bands (light blue).

good competitors to ours, and that they generally achieve a better fit to the dataset. The tests showing the best indexes of goodness of fit are typically those overestimating the number of clusters. It is well-known that, in general, clustering in the context of DPMs and mixture models as those in Quinlan et al. (2017) and Petralia et al. (2012) is affected by the base measure. Our model, on the other hand, avoids the delicate choice of the base measure leading to more stable estimates of K. And in fact, we found that our model has better goodness-of-fit indexes than the competitor repulsive mixtures in the case of a simulated dataset with 8 well-separated mixture components; see Section 4.4. This shows that our proposal is indeed a competitive alternative.

Finally, we recall that in Section 2 in the Supplementary material we report some further tests on the Galaxy dataset to show the influence of various choices of spectral density on the inference. We conclude that there is evidence that the approach is not sensitive with respect to the choice of spectral density.

We have considered one further application, this time using the same variables from the dataset Air Quality (ozone and solar radiation) as considered in Quinlan et al. (2017). Instead of (10), we assume that our likelihood is a bidimensional Gaussian, with bidimensional mean vectors distributed according to the $PES - DPP(\rho, \nu)$ prior as before, and with covariance matrices Σ_k independent and identically distributed according to the inverse-Wishart distribution. See Section 1 in the Supplementary Material for changes in the Gibbs sampler with multidimensional data points, this time adapted from Dellaportas and Papageorgiou (2006). Table 2 reports summaries of interest for a few tests carried out, including the prior and posterior mean and variance for the number of components K, and the LPML. As usual in the context of other mixture models, we find that the inference depends on the chosen hyperparameters. If we compare with corresponding inference in Quinlan et al. (2017), we got lower estimates of K, and a better fit of the model to the data. The posterior predictive densities, not shown here, seem very similar to those in Quinlan et al. (2017), Fig. 9 (b).

Test	ρ	ν	$\mathbb{E}(K)$	$Var(K)$	$\mathbb{E}(K\|data)$	$Var(K\|data)$	LPML
7	3	2	3	2.62	2.18	0.39	-246.81
8	$\rho_0 \sim \text{Gamma}(1, 0.5)$	2	2.7	2.37	2.15	0.21	-257.66

Table 2: Prior specification for (ρ, ν) and K and posterior summaries for the `airquality` dataset; ρ_0 appear in (20).

4.3 Tests on data from a mixture with 8 components

We simulated a dataset with $n = 100$ observations from a mixture of 8 components. Each component is the Gaussian density with mean θ_k and $\sigma_k^2 = \sigma^2 = 0.05$: the means $\{\theta_k\}$ are evenly spaced in the interval $(-10, 10)$. In the model (10)-(15), we set $a_0 = 2.0025$, $b_0 = 0.050125$ so that $\mathbb{E}(\rho_0) = 0.05$ and $Var(\rho_0) = 1$; again, $s = 0.5$ and $\delta = 1$.

Table 3 reports hyperparameters values for different tests and posterior summaries of interest, as well as prior mean and variance of K. In particular, we show the posterior mean and variance for the number of components K (with which we assess the effectiveness of the model for clustering), the mean squared error (MSE) and the log-pseudo marginal likelihood (LPML) (that helps quantifying the goodness-of-fit). In all cases we obtain a pretty satisfactory estimate of the exact number of components, which is 8: the posterior is concentrated around the true value with a very small variance. In particular, we got that $\mathbb{P}(K = 8|data)$ is equal to 0.96 and 0.93 under Tests S_2 and S_7 in Table 3, respectively.

From the density estimation viewpoint, we have from Table 3 that both MSE and LPML are similar for all the tests, thus indicating insensitiveness with respect to the prior choice of parameters ρ and ν. However, preferable tests seem to be S_2 and S_7; see Figure 2, where density estimates and estimated partitions for these two cases are displayed. The posterior density of ρ under Tests S_2 and S_7 is shown in Figure 3.

Prior specification				
Test	ρ	ν	$\mathbb{E}(K)$	$V(K)$
S_0	9.00	1	8.98	45.12
S_1	9	10	9	23.05
S_2	$a_\rho = 1, b_\rho = 1$	1	1.94	1.99
S_3	$a_\rho = 1, b_\rho = 1$	2	2.18	1.99
S_4	$a_\rho = 1, b_\rho = 1$	10	2.74	2.17
S_5	$a_\rho = 1, b_\rho = 1$	discr(2,5,20)	2.52	2.11
S_6	$a_\rho = 1, b_\rho = 1$	discr($\mathcal{V}_1$)	2.45	2.18
S_7	$a_\rho = 1, b_\rho = 1$	discr($\mathcal{V}_2$)	2.5	2.25

Posterior summaries				
Test	$\mathbb{E}(K \mid data)$	$V(K \mid data)$	MSE	LPML
S_0	7.98	0.20	4.65	2.39
S_1	7.99	0.19	4.62	3.10
S_2	8.00	0.17	4.62	3.66
S_3	7.991	0.16	4.62	3.03
S_4	7.99	0.17	4.63	2.96
S_5	7.99	0.16	4.63	3.61
S_6	7.99	0.17	4.65	3.42
S_7	7.99	0.18	4.63	3.36

Table 3: Prior specification for (ρ, ν) and the corresponding mean and variance induced on K (top). Hyperparameters (a_ρ, b_ρ) appear in (20), while $\mathcal{V}_1 = \{1, 2, 5, 10, 20\}$ and $\mathcal{V}_2 = \mathcal{V}_1 \cup \{0.5, 3, 15, 30, 50\}$. Posterior summaries for the simulated dataset from a mixture with 8 components are in the bottom subtable.

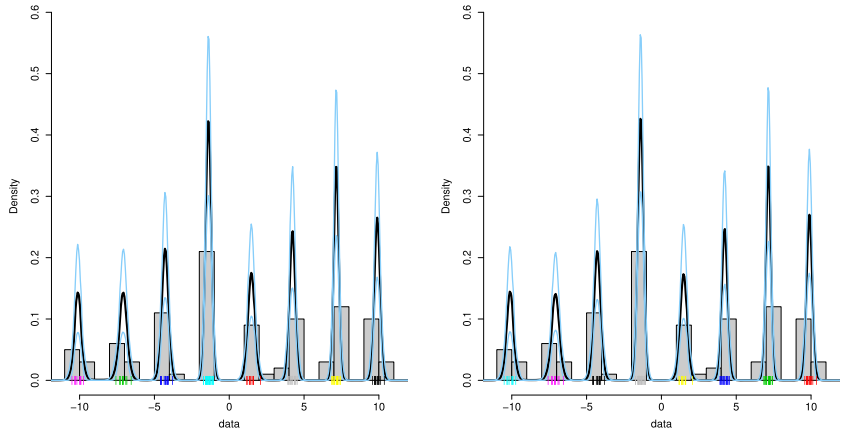

Figure 2: Density estimate and estimated partition for the simulated dataset from the mixture of 8 components under Tests S_2 (left) and S_7 (right) in Table 3. The points at the bottom of the density estimate represent the data, and each color represents one of the eight estimated clusters.

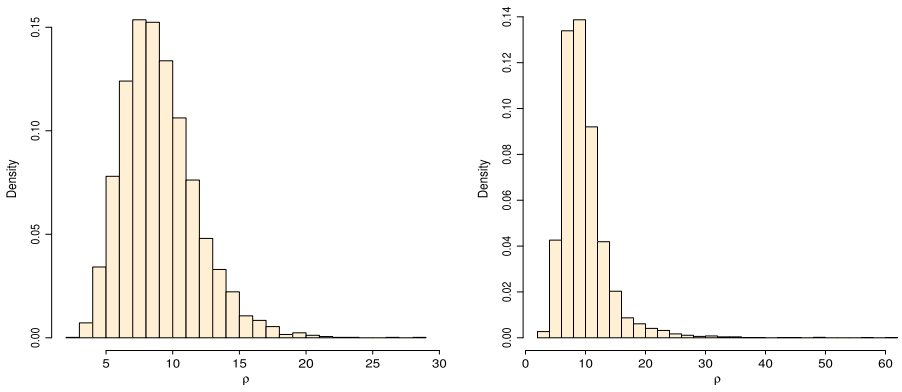

Figure 3: Posterior distribution of ρ for the simulated dataset from the mixture of 8 components under Tests S_2 (left) and S_7 (right) in Table 3.

DPM						
Test	α	$\mathbb{E}(K)$	$\mathbb{E}(K\|data)$	$Var(K\|data)$	MSE	LPML
D_1	Gamma$(0.5, 1)$	2.9	6.166	1.549	62.703	-151.797
D_2	0.8	4.3	5.936	1.25	61.255	-151.146
D_3	0.45	3	4.371	1.142	139.659	-169.978
D_4	Gamma$(4, 2)$	7.7	7.271	1.594	36.708	-149.258

Repulsive models				
Model	$\mathbb{E}(K\|data)$	$Var(K\|data)$	MSE	LPML
Quinlan et al. (2017)	6.462	0.440	38.122	-162.574
Petralia et al. (2012)	7.621	0.757	20.964	-156.522

Table 4: Prior specification for α and posterior summaries for the Galaxy dataset using the function DPdensity in `DPpackage` (top) and repulsive models (bottom).

As a comparison, the same posterior quantities than in Table 1 were computed; see Tables 4 and 5. The DPM was fitted via the function `DPdensity` available from `DPpackage` (Jara et al., 2011), while the code for the alternative repulsive models was gently provided by José Quinlan and Garritt Page.

4.4 Comparison to alternative models

We now consider fitting alternative models to the Galaxy and two simulated datasets, one from the mixture with 8 components introduced in the previous section, and the second consisting of 10,000 observations generated from a mixture of 20 components. We consider first the gold standard of Bayesian nonparametric models, the DPM, and

DPM						
Test	α	$\mathbb{E}(K)$	$\mathbb{E}(K\|data)$	$Var(K\|data)$	MSE	LPML
D_5	0.43	3	7.961	0.5	4.779	-11.246
D_6	Gamma(4, 2)	8.17	8.665	0.910	4.248	-10.116

Repulsive models				
Model	$\mathbb{E}(K\|data)$	$Var(K\|data)$	MSE	LPML
Quinlan et al. (2017)	10.73	1.407	3.121	-4.754
Petralia et al. (2012)	8.51	0.357	4.152	-4.022

Table 5: Posterior summaries for the simulated dataset from the mixture of 8 components using the function DPdensity in `DPpackage` (top) and repulsive models (bottom).

then the RGMM by Quinlan et al. (2017), and the similar specification in Petralia et al. (2012). The same prior information on hyperparameters common to all the mixture models was assumed, i.e. the same marginal prior for σ_k^2 and $(w_1, \ldots, w_k)$. Hyperparameter τ in the potentials ϕ_1 and ϕ_2 was set according to the suggestion in Quinlan et al. (2017) ($\tau = 5.54$).

Comparison of the tables above and Tables 1 and 3 show that the alternative repulsive models are good competitors to ours, and according to the dataset and hyperparameters specification, they may achieve a better (Galaxy) or worse (simulated data) fit to the data. The tests showing the best indexes of goodness of fit are typically those overestimating the number of clusters. It is well-known that, in general, clustering in the context of nonparametric mixture models as DPMs is strongly affected by the base measure (see, e.g. Miller and Harrison, 2017). The same disadvantage affects the mixture models in Quinlan et al. (2017) and Petralia et al. (2012). Our model, on the other hand, avoids the delicate choice of the base measure leading to more stable estimates of K.

As a further comparison, see also Figure 4 which displays the posterior distribution of K under the DPM mixture and our models for the Galaxy dataset.

For the second simulated dataset, we considered applicability for a moderately large sample size, generating 10,000 observations from a 20-component mixture, 10 of them being Gaussian, and the rest being skew-normal distributions with positive and negative skewness. The true density is showed in Figure 5. To estimate the true number of clusters, we fitted different alternative models to this dataset: our model, the repulsive mixture models by Quinlan et al. (2017) and Petralia et al. (2012), and the finite mixture model implemented in the `mclust` R package via the function `Mclust` (Fraley et al., 2012) with a number of components between 10 and 25. The same prior information on hyperparameters common to all the Bayesian mixture models was assumed. The `Mclust` function returns the estimates of the number of components corresponding to the "best three" models, in this case 11, 17 and 18. See Table 6 for posterior summaries of the inference under the three repulsive models.

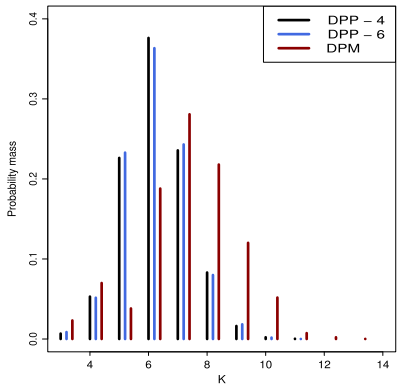

Figure 4: Posterior distribution of the number K of components under Test 4 (black) and 6 (blue) in Table 1 and under the DPM model (red) as in Test D1 in Table 4.

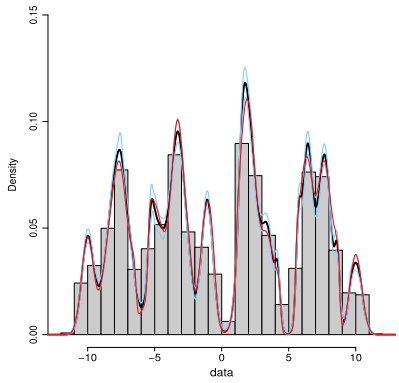

Figure 5: Histogram, true density (red) and density estimate (black) of the large simulated dataset, including 90% credibility bands (light blue).

Model	$\mathbb{E}(K\|data)$	$Var(K\|data)$	MSE	LPML
$PES-DPP$	16.41	1.38	1356.43	-13239.54
Quinlan et al. (2017)	14.13	0.146	1475.98	-13771.56
Petralia et al. (2012)	20.81	0.564	1002.05	-12940.49

Table 6: Posterior summaries for the large simulated dataset.

Though the run-time for this application is around 15 times longer than in the case of the Galaxy data, our algorithm reduces the effective number of clusters a posteriori, thus helping our model scaling up. Intuitively, the increase in the run-time is mostly due to the larger number of mixture components and the much larger sample size than in the case of other datasets illustrated here.

4.5 Simulated data with covariates

We consider the same simulated dataset as in Müller et al. (2011), Section 5.2; the simulation "truth" consists of 12 different distributions, corresponding to different covariate settings (see Figure 1 of that paper). Model (10)-(12), (15)-(17) was fitted to the dataset, assuming $\beta_0 = 0$, $\Sigma_0 = 400 \times \left(Z^T Z\right)^{-1}$, $a_\rho = 1, b_\rho = 1.2$, and a_0, b_0 such that the prior mean of σ_k^2 is 50 and variance is 300. Recall also that here we assume $\nu = 2$.

As an initial step, inference for the complete dataset (1000 observations) was carried out, yielding a posterior of K, not reported here, mostly concentrated over the set $\{8, 9, \ldots, 16\}$, with a mode at 11. Figure 3 in the Supplementary Material shows posterior predictive distributions for the 12 different reference covariate values, along with 90% credibility intervals. These are in good accordance with the simulation truth (compare with Figure 1 in Müller et al., 2011).

To replicate the tests in Müller et al. (2011), a total of $M = 100$ datasets of size 200 were generated by randomly subsampling 200 out of the 1000 available observations. Computational burden over multiple repetitions was controlled by limiting the posterior sample sizes to 2,000. Table 7 displays the root MSE for estimating $\mathbb{E}(y \mid z_1, z_2, z_3)$ for each of the 12 covariate combinations defining the different clusters for our model and for the PPM$_x$, as in Table 1 of Müller et al. (2011). The computations also include evaluation of the root MSE and LPML for all the 100 datasets for estimating the data used to train the model, with $MSE_{train} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is the expected value of the estimated predictive distribution, and for a test dataset of 100 new data, $MSE_{test} = \sum_{i=1}^{n} (y_i^{test} - \hat{y}_i)^2$. In addition, we report $LPML_{train}$, value of the Log Pseudo Marginal Likelihood for the training dataset. Table 8 shows the values compared to other competitor models, i.e the linear dependent Dirichlet process mixture (LDDP) defined in De Iorio et al. (2004), the product partition model (PPM$_x$) in Müller et al. (2011) and the linear dependent tailfree process model (LDTFP) in Jara and Hanson (2011). The best values are in bold: our model performs well according to the LPML, while the MSE suggests to use PPM$_x$ or LDTFP. In general, our model is competitive with respect to other popular models in the literature. Moreover, in the LDDP case, we have that the average number of clusters is 20.6 with variance 2.266, thus indicating a less parsimonious model compared to ours.

In summary, our extensive simulations suggest that the proposed approach tends to require less mixture components than non-repulsive counterpart, while still providing a reasonably good fit to the data. Note that, with respect to other repulsive competitors, our results are comparable.

5 Biopic movies dataset

For this illustrative example we consider the Biopics data available in the R package `fivethirtyeight` (Ismay and Chunn, 2017). This dataset is based on the IMDB database, related to biographical films released from 1915 through 2014. An interest-

z_1	z_2	z_3	DPP	PPM_X
-1	0	0	6.1	7.9
0	0	0	6.7	3.9
1	0	0	7.2	2.8
-1	1	0	6.5	5.4
0	1	0	6.5	4.6
1	1	0	6.8	4.0
-1	0	1	6.8	6.1
0	0	1	6.1	4.2
1	0	1	5.7	4.5
-1	1	1	5.9	9.5
0	1	1	6.6	8.3
1	1	1	5.8	6.2
avg			**6.4**	**5.6**

Table 7: Root MSE for estimating $\mathbb{E}(y \mid z_1, z_2, z_3)$ for 12 combinations of covariates (z_1, z_2, z_3) and PPM_x as competing model of reference (compare also the results in Table 1 of Müller et al., 2011).

	DPPx	LDDP	PPM_X	LDTFP
Root MSE_{train}	324.531	304.742	**278.395**	304.374
Root MSE_{test}	216.675	215.1694	217.2459	**212.761**
$LPML_{train}$	**-871.8**	-902.2295	-873.1671	-901.465

Table 8: Comparison with competitors for the simulated dataset with covariates: best values according to each index are in bold. DPPx denotes our model, while LDDP is the linear dependent Dirichlet process mixture, PPM_x is the product partition model with covariates, and LDTFP is the linear dependent tailfree process model.

ing explorative analysis of the data can be found in https://fivethirtyeight.com/features/straight-outta-compton-is-the-rare-biopic-not-about-white-dudes/.

We consider the logarithm of the gross earnings at US box office as a response variable, with the following covariates: (i) year of release of the movie (in a suitable scale, continuous); (ii) a binary variable that indicates whether the main character is a person of color; and (iii) a categorical variable that considers if the country of the movie is US, UK or other. After removing the missing data from the dataset, we were left with $n = 437$ observations and $p = 4$. We note that 76 biopics have a person of color as a subject and the frequencies of the category "origin" are $(256, 79, 64)$ for US, UK and "other", respectively; "other" means mixed productions (e.g. US and Canada, or US and UK). In what follows, the hyperparameters in model (18)-(19), (11)-(12), (16)-(17) are chosen as $\beta_0 = 0$, $(a_\rho, b_\rho) = (1, 1)$. The prior mean and variance of K induced by these hyperparameters are 2.162 and 1.978, respectively. The scale hyperparameter ϕ in the g-prior for β and (a_0, b_ρ) vary as determined in Table 9, where m and v denote the prior

mean $b_0/(a_0-1)$ and variance $b_0^2/((a_0-1)^2(a_0-2))$, respectively, of the inverse gamma distribution for σ_k^2 as in (19). We also assume γ_0 equal to the vector of all 0's, while Λ_0 is such that the marginal a priori variance of γ_k is equal to $diag(0.01, 0.1, 0.1, 0.1)$, in accordance to the variances of the corresponding frequentist estimators.

Test	ϕ	m	v	$\mathbb{E}(K \mid data)$	sd(K |data)	MSE	LPML
A	50	5	1	4.49	1.10	1126.32	-960.89
B	200	5	10	4.45	1.19	983.55	-954.55
C	50	3	$+\infty$	5.66	1.27	501.22	-918.74
D	200	10	5	4.21	1.33	1805.83	-980.61
E	100	2	1	5.31	1.21	564.26	-935.56
F	200	2	10	5.51	1.26	557.44	-925.22

Table 9: Prior specification for β_k's and σ_k^2's parameters and posterior summaries for the Biopics dataset; m and v are prior mean and variance, respectively, of σ_k^2. Posterior mean and variance of the number K of mixture components are in the fifth and sixth columns, respectively, while the last two columns report MSE and LPML, respectively.

We have found that the posterior of K is stable with respect to the choice of prior hyperparameters, i.e. it does not change significantly when changing the prior hyperparameters; on the other hand, our results show that by not including covariates in the likelihood, i.e. setting all γ_k's are equal to 0, inference on K is much more sensitive to the choice of (a_0, b_0) (results not shown here).

Predictive inference was also considered, by evaluating the posterior predictive distribution at the following combinations of covariate values: (i) (mean value for covariate year, US, white); (ii) (mean value for covariate year, US, color); (iii) (mean value for covariate year, UK, white); (iv) (mean value for covariate year, UK, color); (v) (mean value for covariate year, "other", white); and (vi) (mean value for covariate year, "other", color). Corresponding plots are shown in Figure 6. These distributions appear to be quite different in the six cases: in particular, we can observe that in cases (i) and (ii), the posterior is shifted towards higher values. This is quite easy to interpret, since the measurements are given by the earnings in the US box offices; therefore, we expect that in general US movies will be more profitable in that market. The difference due to the race is, on the other hand, less evident. However, the predictive densities show slightly higher earnings for movies where the subject is a person of color, if the origin is "other" ((v) and (vi)). Movies from the UK, on the other hand, exhibit the opposite behavior ((iii) and (iv)).

We report here the posterior cluster estimate for Test B in Table 9, that is, the partition that minimizes the posterior expectation of Binder's loss function under equal misclassification costs (see Lau and Green, 2007). We note that here, as it is also a common situation with point estimates of partitions, the number of clusters can be different from the posterior mean of K, 4.45 in this case. Indeed, we found three groups, with sizes 10, 193, 234, respectively; see Figure 7 for the estimated clusters and boxplots of the response. As a comparison, it can be useful to report the total average values for the response, 15.36, and for the covariates: 7.89 (year), 0.18 (UK), 0.15 ("other"),

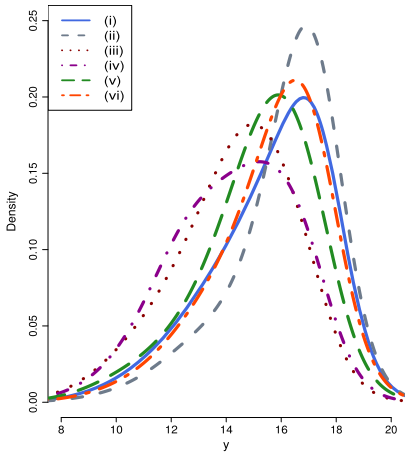

Figure 6: Predictive distribution for cases $(i) - (vi)$ under Test E in Table 9 for the Biopics dataset.

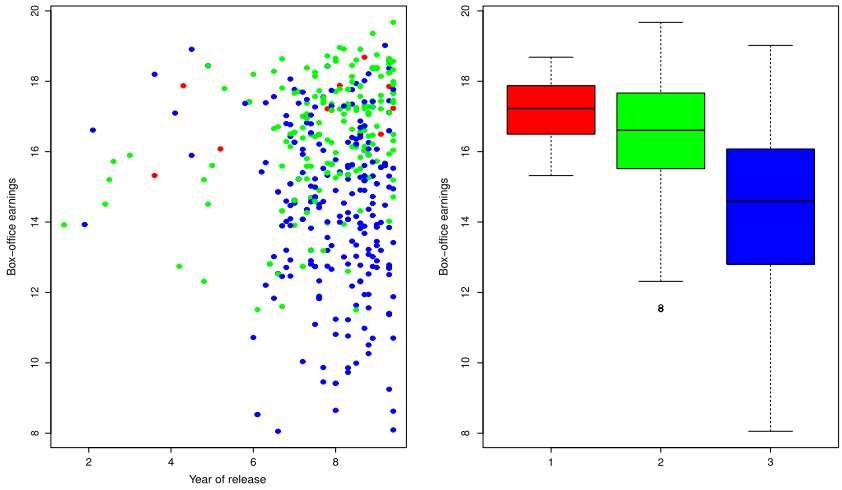

Figure 7: Cluster estimate (left) under our model (Test B in Table 9) for the Biopics dataset. Each color represents one of the three estimated clusters. Coordinate y is the response, i.e. box-office earning, while coordinate x is the covariate year of release. The boxplot of the response per group is in the right panel.

0.83 (white). These 3 groups have a nice interpretation in terms of covariates: group 1 is the smallest, with a high average response (17.18), and it is characterized by a high percentage of movies from "other" countries, with a person of color as its subject. Group 2 corresponds also to a high average response (16.42), but the average values of UK, "other" and person of color are similar to the total averages (0.14, 0.09, 0.84,

respectively). The average response in group 3 is smaller (14.40) than the total sample mean, while the average values of UK, "other" and person of color are 0.22, 0.17, 0.84, respectively.

To assess effectiveness of the proposed model, we compare the results with the linear dependent Dirichlet process mixture model introduced in De Iorio et al. (2004) and implemented in the `LDDPdensity` function of `DPpackage` (Jara et al., 2011). Prior information has been fixed as follows: for Test G the mass parameter of the Dirichlet process α is set equal to 0.3 such that $\mathbb{E}(K) = 2.87$ and $Var(K) = 1.81$, that approximately match the prior information we gave on the parameter K. Similarly, under Test H, α is distributed according to the Gamma$(1/4, 1/2)$, such that the prior mean on K is 3.6 and variance 22.18. The normal baseline distribution is a multivariate Gaussian with mean vector 0 and a random covariance matrix which is given a non-informative prior and the inverse-gamma distribution for the variances of the mixture components has parameters such that mean and variance are equal to 5, 1, respectively similarly as in Table 9. Posterior summaries can be found in Table 10.

Case	$\mathbb{E}(K \mid data)$	$\text{sd}(K \mid data)$	MSE	LPML
G	2.95	1.03	1282.49	-937.51
H	3.56	2.36	682.98	-914.00

Table 10: Posterior summaries for the tests on the Biopics dataset under a linear dependent Dirichlet process mixture.

As a comparison between the estimated partitions under our model (Figure 7) and the LDDP mixture model, Figure 4 in the Supplementary Material displays the estimated partition obtained under the LDDP model under Test G, that has 3 groups with sizes $\{300, 127, 10\}$.

6 Conclusion

This work deals with mixture models where the prior has the property of repulsion across location parameters. Specifically, the discussion is centered on mixtures built on determinantal point processes (DPPs), that can be constructed using a general spectral representation. The methods work with any valid spectral density, but for the sake of concreteness, illustrations were discussed in the context of the power exponential case. Implementing the posterior inference for our models requires the use of numerical approximations of the density function of the DPP.

Our approach has focused on mixture modeling. One may use essentially the same models for clustering purposes. In such case, it is quite simple to modify the cluster-specific covariance matrices to have any particular form, such as enforcing elliptically- or circularly-shaped contours. We have here used standard assumptions such as unstructured covariance matrices as a default choice in our discussion, but again, such assumptions, or others, such as stick-breaking mixture weights, can be easily incorporated to the model definitions.

Repulsiveness can be measured in a number of ways, as discussed in the online supplementary material in Lavancier et al. (2015), Appendix J. These include Ripley's K-function and the Pair Correlation Function. They both measure how dispersed the points appear to be and they are usually compared to the Poisson process, where for a given a number of points, they are randomly scattered. Some analytical results can be obtained, but in general, one must evaluate these functions numerically. The power exponential case includes the most repulsive type of stationary DPP when $\nu \to \infty$. When $d = 2$, this corresponds to a "jinc-like" covariance function, that is, the case in which the spectral density is of the form $\varphi(x) = I\{\|x\| \leq \tau\}$. We found out that in most applications a value of $\nu = 2$ is sufficient for reasonable inference results, but larger values of ν also provide stable enough inference.

Though we limit ourselves to the case of isotropic DPPs, inhomogeneous DPPs can be obtained by transforming or thinning a stationary process. However, we believe that this case is not very interesting, unless there is a strong reason to assume non-homogeneous locations a priori.

Our computational experiments and data illustrations show that the repulsion induced by the DPP priors indeed tends to eliminate the annoying case of many very small clusters that commonly arises when using models that do not constrain location/centering parameters. This happens with very small sacrifice of model fit compared to the usual mixture models. In this sense, since it is a *parsimonious* model with respect to the clustering structure, interpretation is facilitated because there are typically less clusters to interpret.

From a computational viewpoint, we found our model to scale reasonably well with respect to sample size, compared to similar alternatives. However, scaling with respect to data dimension is not a property we claim to have. In fact, this is a situation shared with mixture models either with a finite or infinite number of components.

Another advantage of our model over DPMs and natural competitors in the repulsive mixture models framework, such as those in Quinlan et al. (2017) and Petralia et al. (2012), is that we avoid the delicate choice of the base measure of the Dirichlet process, leading to more stable estimates on the number K of components in the mixture. Moreover, with respect to competitor repulsive mixtures, and just in the case of DPM models, our approach does not require an upper bound on the allowed number of mixture components. This particular aspect can also be considered as a con, because the competitor repulsive mixtures may produce slightly better performance in terms of goodness of fit. On the other hand, since K is finite and random, our algorithm requires a reversible jump implementation, which could be seen in practice as an extra computational price to pay when using the proposed model.

Another worth noting aspect of the model is the roles of subsets S and R. They are both required to be compact so that the prior joint density of latent cluster centers is well defined. Subset S, in practice fixed as the unit hypercube, is specifically needed to guarantee that approximations (6) and (7) work well. In turn, the role of the rectangle R is to cover the empirical data range, and the affine transformation T, which maps R back onto S, is used to ensure that all calculations are carried in S. The size of R has

little consequence in itself, beyond obvious changes in formula (8), because the prior distribution of number of points is still determined by (2), i.e. by S and C_0. And a very large set R would simply have large portions with very low posterior mass assigned, corresponding to places with no observed data.

Summing up, through extensive computational experience in various settings, we believe our model to be a valid and competitive alternative that, with little sacrifice in model fit, may help achieving a more parsimonious representation of mixtures, compared to other similar models.

Supplementary Material

Supplementary Material for "Determinantal Point Process Mixtures Via Spectral Density Approach"
(DOI: 10.1214/19-BA1150SUPP; .pdf). Supplementary materials for this article are available online. In particular, the document contains a description of the two Gibbs samplers used for posterior inference under the DPP mixture model (Sections 1 and 3), without and with covariate dependence, respectively. Moreover, a new test on the Galaxy dataset when different spectral densities are considered (Section 2). Finally, supplemental figures that did not fit in the paper and additional analysis of an Air Quality Index dataset are provided.

References

Affandi, R. H., Fox, E., and Taskar, B. (2013). "Approximate inference in continuous determinantal processes." In *Advances in Neural Information Processing Systems*, 1430–1438. 188

Affandi, R. H., Fox, E. B., Adams, R. P., and Taskar, B. (2014). "Learning the Parameters of Determinantal Point Process Kernels." In *ICML*, 1224–1232. 188

Bianchini, I., Guglielmi, A., and Quintana F. A. (2019). "Supplementary Material for "Determinantal Point Process Mixtures Via Spectral Density Approach"." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1150. 187

Bardenet, R. and Titsias, M. (2015). "Inference for determinantal point processes without spectral knowledge." In *Advances in Neural Information Processing Systems*, 3393–3401. 188

Binder, D. A. (1978). "Bayesian cluster analysis." *Biometrika*, 65: 31–38. MR0501592. doi: https://doi.org/10.1093/biomet/65.1.31. 190

Biscio, C. A. N. and Lavancier, F. (2016). "Quantifying repulsiveness of determinantal point processes." *Bernoulli*, 22: 2001–2028. MR3498021. doi: https://doi.org/10.3150/15-BEJ718. 188

Biscio, C. A. N. and Lavancier, F. (2017). "Contrast estimation for parametric stationary determinantal point processes." *Scandinavian Journal of Statistics*, 44: 204–229. MR3619702. doi: https://doi.org/10.1111/sjos.12249. 188

Daley, D. J. and Vere-Jones, D. (2003). "Basic Properties of the Poisson Process." *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*, 19–40. MR1950431. 191

Daley, D. J. and Vere-Jones, D. (2007). *An introduction to the theory of point processes: volume II: general theory and structure*. Springer. 191

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). "An ANOVA model for dependent random measures." *Journal of the American Statistical Association*, 99: 205–215. MR2054299. doi: https://doi.org/10.1198/016214504000000205. 205, 209

Dellaportas, P. and Papageorgiou, I. (2006). "Multivariate mixtures of normals with unknown number of components." *Statistics and Computing*, 16(1): 57–68. MR2224189. doi: https://doi.org/10.1007/s11222-006-5338-6. 200

Ferguson, T. S. (1983). "Bayesian density estimation by mixtures of normal distributions." In M. H. Rizvi, J. R. and Siegmund, D. (eds.), *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his Sixtieth Birthday*, 287–302. Academic Press. MR0736538. 188

Fraley, C., Raftery, A. E., Murphy, T. B., and Scrucca, L. (2012). *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*. 190, 203

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Series in Statistics. Springer, New York. MR2265601. 187

Fúquene, J., Steel, M., and Rossell, D. (2016). "On choosing mixture components via non-local priors." http://arxiv.org/abs/1604.00314v1. 188

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96: 161–173. MR1952729. doi: https://doi.org/10.1198/016214501750332758. 187

Ismay, C. and Chunn, J. (2017). *fivethirtyeight: Data and Code Behind the Stories and Interactives at 'FiveThirtyEight'*. R package version 0.1.0. https://CRAN.R-project.org/package=fivethirtyeight. 205

Jara, A. and Hanson, T. E. (2011). "A class of mixtures of dependent tail-free processes." *Biometrika*, 98: 553–566. MR2836406. doi: https://doi.org/10.1093/biomet/asq082. 205

Jara, A., Hanson, T. E., Quintana, F. A., Müller, P., and Rosner, G. L. (2011). "DPpackage: Bayesian semi-and nonparametric modeling in R." *Journal of Statistical Software*, 40: 1. MR3309338. doi: https://doi.org/10.1007/978-3-319-18968-0. 202, 209

Kulesza, A. and Taskar, B. (2012). "Determinantal point processes for machine learning." *Foundations and Trends in Machine Learning*, 5: 123–286. 188

Lau, J. W. and Green, P. J. (2007). "Bayesian model-based clustering proce-

dures." *Journal of Computational and Graphical Statistics*, 16: 526–558. MR2351079. doi: https://doi.org/10.1198/106186007X238855. 190, 198, 207

Lavancier, F., Møller, J., and Rubak, E. (2015). "Determinantal point process models and statistical inference." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77: 853–877. MR3382600. doi: https://doi.org/10.1111/rssb.12096. 188, 189, 190, 191, 192, 193, 210

Lo, A. Y. (1984). "On a class of Bayesian nonparametric estimates: I. Density estimates." *The Annals of Statistics*, 12: 351–357. MR0733519. doi: https://doi.org/10.1214/aos/1176346412. 188

Macchi, O. (1975). "The coincidence approach to stochastic point processes." *Advances in Applied Probability*, 83–122. MR0380979. doi: https://doi.org/10.2307/1425855. 191

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). "Model-based clustering based on sparse finite Gaussian mixtures." *Statistics and Computing*, 26: 303–324. MR3439375. doi: https://doi.org/10.1007/s11222-014-9500-2. 188

McLachlan, G. and Peel, D. (2005). *Finite Mixture Models*. John Wiley & Sons, Inc. MR1789474. doi: https://doi.org/10.1002/0471721182. 189, 196

Miller, J. W. and Harrison, M. T. (2013). "A simple example of Dirichlet process mixture inconsistency for the number of components." In *Advances in neural information processing systems*, 199–206. 188

Miller, J. W. and Harrison, M. T. (2017). "Mixture models with a prior on the number of components." *Journal of the American Statistical Association*. In Press. 187, 203

Møller, J. and Waagepetersen, R. P. (2007). "Modern statistics for spatial point processes." *Scandinavian Journal of Statistics*, 34: 643–684. MR2392447. doi: https://doi.org/10.1111/j.1467-9469.2007.00569.x. 191, 192

Müller, P., Quintana, F., and Rosner, G. L. (2011). "A product partition model with regression on covariates." *Journal of Computational and Graphical Statistics*, 20: 260–278. MR2816548. doi: https://doi.org/10.1198/jcgs.2011.09066. 205, 206

Petralia, F., Rao, V., and Dunson, D. B. (2012). "Repulsive Mixtures." In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems 25*, 1889–1897. Curran Associates, Inc. http://papers.nips.cc/paper/4589-repulsive-mixtures.pdf. 188, 190, 195, 198, 199, 202, 203, 204, 210

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. MR3832183. doi: https://doi.org/10.1080/00949655.2018.1491578. 188

Quinlan, J. J., Quintana, F. A., and Page, G. L. (2017). "Parsimonious Hierarchical Modeling Using Repulsive Distributions." https://arxiv.org/abs/1701.04457v1. 188, 190, 195, 198, 199, 200, 202, 203, 204, 210

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: 689–710. MR2867454. doi: https://doi.org/10.1111/j.1467-9868.2011.00781.x. 188

Shirota, S. and Gelfand, A. E. (2017). "Approximate Bayesian Computation and Model Assessment for Repulsive Spatial Point Processes." *Journal of Computational and Graphical Statistics*, 26(3): 646–657. MR3698674. doi: https://doi.org/10.1080/10618600.2017.1299627. 188

Xu, Y., Müller, P., and Telesca, D. (2016). "Bayesian Inference for Latent Biological Structure with Determinantal Point Processes (DPP)." *Biometrics*, 72: 955–964. MR3545688. doi: https://doi.org/10.1111/biom.12482. 188, 189, 194, 195

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6: 233–243. MR0881437. 196

Acknowledgments

The authors gratefully acknowledge grants FONDECYT 1141057 and 1180034; the first two authors thank people from the Departamento de Estadística at PUC, Chile, for their kind hospitality. This work was also partially supported by Iniciativa Científica Milenio - Minecon Núcleo Milenio MiDaS.