Biometrika (2025), **112**, 2, asae059 https://doi.org/10.1093/biomet/asae059
Advance Access publication 4 November 2024

Bayesian clustering of high-dimensional data via latent repulsive mixtures

By L. GHILOTTI

Department of Economics, Management and Statistics, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milan, Italy
l.ghilotti@campus.unimib.it

M. BERAHA and A. GUGLIELMI

Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy
mario.beraha@polimi.it alessandra.guglielmi@polimi.it

Summary

Model-based clustering of moderate- or large-dimensional data is notoriously difficult. We propose a model for simultaneous dimensionality reduction and clustering by assuming a mixture model for a set of latent scores, which are then linked to the observations via a Gaussian latent factor model. This approach was recently investigated by Chandra et al. (2023). The authors used a factor-analytic representation and assumed a mixture model for the latent factors. However, performance can deteriorate in the presence of model misspecification. Assuming a repulsive point process prior for the component-specific means of the mixture for the latent scores is shown to yield a more robust model that outperforms the standard mixture model for the latent factors in several simulated scenarios. The repulsive point process must be anisotropic to favour well-separated clusters of data, and its density should be tractable for efficient posterior inference. We address these issues by proposing a general construction for anisotropic determinantal point processes. We illustrate our model in simulations, as well as a plant species co-occurrence dataset.

Some key words: Anisotropic point process; Determinantal point process; Gaussian factor model; Markov chain Monte Carlo; Model-based clustering.

1. Introduction

1.1. *Motivation*

This paper concerns Bayesian cluster analysis for moderate- or high-dimensional data. Specifically, we focus on observations $y_1, \ldots, y_n \in \mathbb{R}^p$. Although a variety of models have been proposed in the literature (Neal, 2003; Teh et al., 2007; Duan & Dunson, 2021; Natarajan et al., 2023), Bayesian mixtures constitute a direct approach for model-based clustering; see Fruhwirth-Schnatter et al. (2019) for a recent review. In mixture models, it is assumed

that data are generated from m, either random or fixed, homogeneous populations. Typically, each population is assumed to be suitably modelled via a parametric density $f_\theta(\cdot)$ for some parameter $\theta \in \Theta$. Weights $w = (w_1, \ldots, w_m)$, $w_h \geqslant 0$, $\sum_{h=1}^m w_h = 1$, specify the relative frequency of each population. In summary, the conditional distribution of data, given parameters, under the mixture model, takes the form

$$y_1, \ldots, y_n \mid w, \theta \overset{\text{iid}}{\sim} \sum_{h=1}^{m} w_h f_{\theta_h}(\cdot). \tag{1}$$

Under the Bayesian approach, suitable priors are assumed for w, $\theta = (\theta_1, \ldots, \theta_m)$ and m.

The poor performance of Bayesian mixtures when data dimension p is larger than a moderate value, e.g., $p > 10$, is a long-standing issue. This is partly due to the poor scalability of the algorithms for posterior inference (see, e.g., Malsiner-Walli et al., 2016; Celeux et al., 2019), and partly due to the asymptotic properties of the model when p is large. Specifically, Theorem 1 together with Corollaries 1 and 2 of Chandra et al. (2023) shows the degeneracy of the cluster estimate when f_θ in (1) is the Gaussian distribution, the de-facto standard. As $p \to +\infty$ for a fixed n, if the covariance matrix is cluster specific then, with posterior probability one, all observations are clustered as singletons, while if the covariance matrix is shared among all the clusters, only one cluster is detected.

The most popular approach among practitioners to cluster high-dimensional data follows a two-step procedure. First, fitting a latent factor model (Lopes, 2014), a d-dimensional score η_i, where $d \ll p$, is associated with each observation. Then, traditional clustering algorithms are applied to the η_i. However, this two-step procedure does not allow propagation of the variability induced by the dimensionality reduction step into the cluster estimates. To overcome this critical limitation, the natural option is to consider a model for simultaneous dimensionality reduction and clustering, by assuming a mixture model for the latent scores, which are then linked to the observations via a Gaussian latent factor model. This approach was recently investigated by Chandra et al. (2023) within the Bayesian nonparametric framework.

However, mixture models might overestimate the number of clusters to recover the 'true' sampling model. Indeed, the typical postulate in Bayesian mixture models is that parameters θ_h in (1) are independent and identically distributed from a base distribution. Under this assumption, if the true data-generating density does not agree with (1), the number of clusters a posteriori diverges as n increases (Cai et al., 2021). Two sources of misspecification pertain to the Gaussian latent factor model with a standard mixture model for the latent scores: first, the factor-analytic representation entails that data lie close to a d-dimensional hyperplane; second, the deviation from such a hyperplane is Gaussian distributed. Both these assumptions can be questioned and are unlikely to hold in practice. Moreover, Chandra et al. (2023) assumed a Dirichlet process mixture prior for the latent scores, which might yield inconsistent estimates even for well-specified models (Miller & Harrison, 2014). Promising results towards consistency of Dirichlet process mixtures have been obtained by Ascolani et al. (2023), but their assumptions do not hold for the model of Chandra et al. (2023).

The present paper proposes the anisotropic repulsive point process latent mixture (APPLAM) model. We consider a model for simultaneous dimensionality reduction and clustering as above, but assume a repulsive point process as the prior for the component-specific location parameters of the mixture for the latent scores. Repulsive mixtures, pioneered by Petralia et al. (2012), offer a practical solution to the lack of robustness of

mixture models with independent and identically distributed component parameters. See also Xu et al. (2016), Fúquene et al. (2019), Xie & Xu (2019), Bianchini et al. (2020) and Beraha et al. (2022). In these works, repulsive priors are used to jointly model the component-specific location parameters, and, in some of them, their number. However, in the context of a latent mixture model, we argue that to have well-separated clusters of data, it is not sufficient to have well-separated clusters at the latent level, but the repulsion should also take into account the factor analytic model that links the latent variables to the observations. To this end, we propose an anisotropic determinantal point process as the prior for the component-specific parameters, where the matrix of factor loadings drives the anisotropy. We derive a general construction of anisotropic determinantal point processes that induces the desired repulsion.

1.2. *Bayesian clustering via latent mixtures*

Let $y = (y_1, \ldots, y_n)$, $y_i \in \mathbb{R}^p$, $\Lambda \in \mathbb{R}^{p \times d}$ be the matrix of factor loadings, $\eta_1, \ldots, \eta_n \in \mathbb{R}^d$ be a set of latent factors and $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, $\sigma_j > 0$, be a diagonal covariance matrix. Let $\mathcal{N}_p(a, B)$ denote the p-dimensional Gaussian distribution with mean a and covariance matrix B. A latent factor mixture model assumes that

$$y_i \mid \eta_i, \Lambda, \Sigma \overset{\text{ind}}{\sim} \mathcal{N}_p(\Lambda\eta_i, \Sigma), \qquad \eta_i \,\Big|\, w, \mu, \Delta \overset{\text{iid}}{\sim} \sum_{h=1}^{m} w_h \mathcal{N}_d(\mu_h, \Delta_h), \tag{2}$$

for $i = 1, \ldots, n$. Extensions to other kernel distributions for the mixture model for latent factors and for observations are straightforward. The mixture model is completed when assigning the prior for $w = (w_1, \ldots, w_m)$, $\mu = (\mu_1, \ldots, \mu_m)$, $\Delta = (\Delta_1, \ldots, \Delta_m)$, Λ and the σ_j^2. In the case of repulsive mixtures, as we propose here, m is also random, so the number of mixture components is learned from the data. See § 2.2 below for the full description of our prior.

Introducing a set of latent cluster indicator variables c_i such that $\mathrm{pr}(c_i = h \mid w) = w_h$, we can equivalently state the prior for the η_i in (2) as $\eta_i \mid c_i = h, \mu, \Delta \overset{\text{ind}}{\sim} \mathcal{N}_d(\mu_h, \Delta_h)$, $i = 1, \ldots, n$. Therefore, as is standard in Bayesian mixture models, we cluster data y_i through the latent variables η_i, i.e., y_i and y_j belong to the same cluster if $c_i = c_j$. Through the c_i we can identify the clusters with the allocated components, i.e., those components $h \in \{1, \ldots, m\}$ for which there exists i such that $c_i = h$.

The cluster estimate is interpretable and, hence, useful if observations belonging to different clusters are well separated. Repulsive mixture models encourage well-separated clusters by assuming a prior for the cluster centres μ_h that favours regular, i.e., well-separated, point configurations. A straightforward approach to defining such a prior is assuming a repulsive point process that governs both the cardinality m of the components and the locations of the μ_h. In particular, it is possible to define such a process by specifying a density with respect to a Poisson point process. For instance, Xie & Xu (2019) and Quinlan et al. (2021) assumed a pairwise interaction process whose density, with respect to a suitably defined Poisson process, is

$$p(\{\mu_1, \ldots, \mu_m\}) = \frac{1}{Z} \prod_{j=1}^{m} \phi_1(\mu_j) \prod_{1 \leqslant h < k \leqslant m} \phi_2(\|\mu_h - \mu_k\|), \tag{3}$$

where ϕ_1 is a bounded function, ϕ_2 is a nondecreasing function and Z is a normalizing constant that is usually intractable. See Møller & Waagepetersen (2003) and Daley & Vere-Jones (2008) for the definition of density with respect to a Poisson point process.

This choice of the prior would ensure that different clusters are associated with well-separated latent scores η_i, but is this enough to ensure well-separated clusters of data y_i? By the properties of the Gaussian distribution, we have

$$\{y_i \colon c_i = h\} \mid c, \mu, \Delta, \Lambda \overset{\text{iid}}{\sim} \mathcal{N}_p(\Lambda\mu_h, \Lambda\Delta_h\Lambda^{\mathrm{T}} + \Sigma).$$

Hence, it is clear that it is not sufficient to encourage a priori that $\|\mu_h - \mu_k\|$ is large to obtain well-separated clusters of data points, as the distance between cluster centres is $\|\Lambda\mu_h - \Lambda\mu_k\|$. Such intuition is confirmed by our numerical experiments, as discussed in the Supplementary Material, where we show that failing to induce the right repulsion across cluster centres produces extremely poor cluster estimates a posteriori. Of course, to produce repulsion across the $\Lambda\mu_h$ in a pairwise interaction point process, we could easily modify (3) by considering $\phi_2(\|x\|) \equiv \tilde{\phi}_2(\|\Lambda x\|)$, where now the normalizing constant Z depends on ϕ_1, ϕ_2 and, most importantly, Λ. The intractability of the normalizing constant in (3) poses a significant challenge to posterior simulation. Beraha et al. (2022) discussed how to sample from the posterior distribution of the parameters involved in (3) using the exchange algorithm (Møller et al., 2006; Murray et al., 2006), which requires perfect sampling from (3). However, Beraha et al. (2022) investigated the efficiency of the algorithm only when sampling a single real-valued parameter. Here, instead, the point process density does depend on Λ and our preliminary investigation showed that updating Λ via the exchange algorithm results in extremely poor mixing due to the high-dimensionality of matrix Λ itself.

1.3. *Our contributions*

We propose an anisotropic determinantal point process (DPP) in § 2.1 below as the prior for the component-specific latent factor centres $\{\mu_1, \ldots, \mu_m\}$, where the matrix of factor loadings drives the anisotropy. As a first main contribution, we derive a general construction of anisotropic DPPs together with novel existence conditions for our class of DPPs that are similar to those of Lavancier et al. (2015), and the associated density does not involve intractable terms. We show analytically that the anisotropic DPP induces the desired repulsion. Indeed, we prove that our process is equivalent to a standard isotropic DPP distribution for points $\{\Lambda\mu_1, \ldots, \Lambda\mu_m\}$ on the hyperplane spanned by the columns of Λ. Though one could, in principle, directly define a standard DPP on such a hyperplane, the resulting process would be analytically intractable, requiring one to perform the eigendecomposition of a kernel function defined on a random hyperplane, and thus of limited practical utility. We further provide an explicit expression for the spectral density of the DPP, which is essential for simulation purposes. This step also requires a standard approximation (see, e.g., Lavancier et al., 2015) of two infinite sums in the expression of the density of the DPP. We empirically show that the approximation error is negligible.

Our second contribution is to embed the anisotropic DPP in a latent factor mixture model (2) as a prior for the locations of the μ_h and their cardinalities. We discuss several aspects of the proposed model, focusing in particular on the role of the parameters governing the repulsiveness of the process. We also provide guidance on setting such parameters via a data-dependent procedure. Posterior inference presents nontrivial computational challenges that we address by proposing a Metropolis-within-Gibbs algorithm. In particular, the full conditional for Λ is updated using the well-known Metropolis-adjusted Langevin algorithm,

which, to be computationally efficient, requires the analytical expression for the gradient of the logarithm of the density of the DPP; we provide such an expression among our theoretical results. We illustrate the adequacy of our model by applying it to real and simulated datasets. Specifically, we consider data collecting the occurrence of plant species, $p = 123$, at different sites of the Bauges Natural Regional Park, France. The goal is to infer the clustering structure of the $n = 1139$ sites, such that sites belonging to the same cluster show similar patterns concerning the occurrence of the plant species. Further insights into our model are obtained via extensive simulation studies with high-dimensional and large datasets, e.g., p up to 1000, which is extremely large for traditional clustering methods, and $n = 1000$. Beyond justifying the need to assume an anisotropic repulsive prior, we compare our APPLAM model to the latent mixture for Bayesian (LAMB) model of Chandra et al. (2023). We find that the APPLAM model produces more robust cluster estimates in the presence of model misspecification.

2. Methodology

2.1. *A general construction for anisotropic DPPs*

A determinantal point process Φ on $\{\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\}$, where $\mathcal{B}(\mathbb{R}^d)$ is the Borel sigma algebra, is a random subset $\{\mu_1, \ldots, \mu_m\} \subset \mathbb{R}^d$. See Macchi (1975), Lavancier et al. (2015) and Baccelli et al. (2020). The probability distribution of a DPP is completely characterized by a continuous complex-valued covariance kernel $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ in terms of its m-factorial moment measure. Such a general definition might appear cumbersome to the unfamiliar reader and is reported in the Supplementary Material. We focus here on the DPP restricted on a compact region $R \subset \mathbb{R}^d$, so that we can write its density with respect to a Poisson process. However, our results hold for general DPPs defined on the whole $\mathbb{R}^d$. By Mercer's theorem, the restriction of K to $R \times R$ admits the spectral representation $K(x, y) = \sum_{j \geqslant 1} \gamma_j \xi_j(x) \overline{\xi}_j(y)$, where the ξ_j are a set of eigenfunctions of $L^2(R; \mathbb{C})$, forming an orthonormal basis, and the γ_j are the summable nonnegative sequence of eigenvalues; $\overline{\xi}_j$ denotes the conjugate value of ξ_j. In particular, when restricted to R, a DPP satisfying $\gamma_j < 1$ for all $j = 1, 2, \ldots$ admits a density with respect to the unit-rate Poisson process on R as

$$p(\{\mu_1, \ldots, \mu_m\}) = e^{|R|-D} \det\{C(\mu_h, \mu_k)\}_{h,k=1,\ldots,m}, \qquad \mu_1, \ldots, \mu_m \in R,\ m = 0, 1, 2, \ldots, \tag{4}$$

where $C(x, y) = \sum_{j \geqslant 1} \gamma_j/(1 - \gamma_j) \xi_j(x) \overline{\xi}_j(y)$, $|R| = \int_R \mathrm{d}x$, $D = -\sum_{j \geqslant 1} \log(1 - \gamma_j)$ and we adopt the convention that $\det\{C(\mu_h, \mu_k)\}_{h,k=1,\ldots,m} = 1$ if $m = 0$. See Lavancier et al. (2015) for a proof of such results. As noted by Lavancier et al. (2015), the continuity of C implies that the determinant in (4) tends to zero as the Euclidean distance $\|\mu_h - \mu_k\|$ goes to zero for some $h \neq k$, which shows that a DPP is a repulsive point process. The probability distribution of a DPP is uniquely characterized by its kernel $K(\cdot)$ on $\mathbb{R}^d$ or, equivalently, by its spectral representation. Compared to the expression in (3), the density of a DPP features an analytically tractable normalizing constant $e^{|R|-D}$, which significantly simplifies posterior inference for the hyperparameters of the process; see, e.g., Beraha et al. (2022) for a discussion on the computational complexities inherited from an intractable normalizing constant. On the other hand, to work with a DPP density (4), one needs to restrict the point process on a compact region R, which is a nuisance not shared by the pairwise interaction point processes. However, as we discuss in § 2.3 below, the choice of R plays a limited role in our model.

Analytic expressions for eigenvalues γ_j are crucial for inferential purposes. Following the so-called 'spectral approach' by Lavancier et al. (2015), Bianchini et al. (2020) and Beraha et al. (2022) assumed that $K(x, y) = K_0(x - y) = K_0(\|x - y\|)$, i.e., K is a stationary and isotropic function. Instead of modelling K, they fixed the ξ_j as the Fourier basis and assumed a parametric model for the γ_j. This approach ensures the positive definiteness of K and the existence of the DPP density, but is not flexible enough for the mixture model (2). In particular, isotropy of the DPP kernel K conflicts with our goal of forcing repulsion across the $\Lambda\mu_h$. In the following, we provide a new general construction for stationary anisotropic DPPs, an explicit expression for the Fourier transform of its kernel K_0 and easy-to-check conditions that guarantee the DPP's existence.

THEOREM 1. *Let Λ be a fixed $p \times d$ real matrix with full rank. Let W be a strictly positive random variable, and let $h(y)$ be the marginal density of the random variable Y defined as*

$$Y \mid W \sim \mathcal{N}_d\{0, W(\Lambda^T\Lambda)^{-1}\}.$$

Let $K_0(x) = \rho h(x)/h(0)$ for $x \in \mathbb{R}^d$ and $\rho > 0$. Then there exists a DPP Φ on $\mathbb{R}^d$ with kernel $K(x, y) = K_0(x - y)$ for $\rho \leqslant \rho_{\max}$ defined as

$$\rho_{\max} = |\Lambda^T\Lambda|^{1/2}(2\pi)^{-d/2}\,\mathbb{E}(W^{-d/2}),$$

and

$$K_0(x) = \frac{\rho}{\mathbb{E}(W^{-d/2})}\mathbb{E}\left\{W^{-d/2}\exp\left(-\frac{||\Lambda x||^2}{2W}\right)\right\}, \qquad x \in \mathbb{R}^d.$$

If $\varphi(x) = \mathcal{F}(K_0)(x)$ denotes the Fourier transform of K_0, we have

$$\varphi(x) = \frac{\rho}{h(0)}\mathbb{E}[\exp\{-2\pi^2 W x^T(\Lambda^T\Lambda)^{-1}x\}], \qquad x \in \mathbb{R}^d.$$

Moreover, for any compact $R \subset \mathbb{R}^d$, the restriction of Φ to R has a density with respect to the unit-rate Poisson point process on R if $\rho < \rho_{\max}$.

Kernel K_0 characterizing the DPP in Theorem 1 is not isotropic. Parameter ρ is the intensity of the process, i.e., it controls the distribution of the number of points in the process. In particular, the expected total number of points in R is equal to $\rho|R|$. We also emphasize that explicit knowledge of the Fourier transform, as described in Theorem 1, is essential for simulation purposes, as one typically approximates the density of the DPP using φ, as described in § 3.1 below.

The following result is an equivalent characterization of the anisotropic DPP defined above.

THEOREM 2. *Let Φ be an anisotropic DPP on $\mathbb{R}^d$ defined as in Theorem 1. Then $\tilde{\Phi} = \{\Lambda\mu : \mu \in \Phi\}$ is a stationary and isotropic DPP on $B = \Lambda\mathbb{R}^d$ with kernel*

$$\tilde{K}_0(y) = \frac{\rho|\Lambda^{\mathrm{T}}\Lambda|^{-1/2}}{\mathbb{E}(W^{-d/2})}\mathbb{E}\left\{W^{-d/2}\exp\left(-\frac{\|y\|^2}{2W}\right)\right\}.$$

Theorem 2 sheds light on the repulsiveness induced by our anisotropic DPP and why it is suited in the context of latent mixture models. As discussed in § 1.2, the model produces interpretable clusters if the points $\{\Lambda\mu_1, \ldots, \Lambda\mu_m\}$ are well separated. From the theorem above, it turns out that our new construction in Theorem 1 is equivalent to assuming an isotropic DPP prior for the points $\Lambda\mu_h$, thus matching our goal of inducing separation across clusters of data. However, as will be discussed in § 3.1, using the construction in Theorem 2 leads to much greater analytical and computational hurdles compared to the approach in Theorem 1 when computing the posterior inference.

The type or strength of repulsion is controlled by the random variable $W \sim p(w)$ in Theorem 1. In the rest of the paper, we consider one specific choice for the random variable W in Theorem 1, leading to the anisotropic counterpart of the Gaussian DPPs discussed by Lavancier et al. (2015), which we refer to as Gaussian-like DPPs in the following. Moreover, to prove the generality of our approach, in the Supplementary Material, we show how to construct an anisotropic counterpart of the Whittle–Matérn DPP, although we will not use it in our examples.

COROLLARY 1. *Using the same notation as in Theorem 1, let W be a degenerate random variable defined as $W = |\Lambda^T\Lambda|^{1/d}c^{-2/d}$ for $c > 0$, where Λ is fixed. Then kernel K_0, its Fourier transform $\varphi = \mathcal{F}(K_0)$ and $\rho_{\max}$ follow:*

$$K_0(x) = \rho \exp\left(-\frac{||\Lambda x||^2}{2|\Lambda^T\Lambda|^{1/d}c^{-2/d}}\right), \qquad x \in \mathbb{R}^d,$$

$$\varphi(x) = \rho\,\frac{(2\pi)^{d/2}}{c}\exp\{-2\pi^2|\Lambda^T\Lambda|^{1/d}c^{-2/d}x^T(\Lambda^T\Lambda)^{-1}x\}, \qquad x \in \mathbb{R}^d, \tag{5}$$

$$\rho_{\max} = c(2\pi)^{-d/2}.$$

We conclude this section by investigating the effect that Λ has on the repulsiveness of the DPP. First of all, it is clear that Λ induces anisotropy. To visualize this, we consider the pair correlation function $g(x)$ (Lavancier et al., 2015); for the Gaussian-like DPP, we have

$$g(x) = 1 - \frac{K_0(x)}{K_0(0)} = 1 - \exp\left(-\frac{||\Lambda x||^2}{2|\Lambda^T\Lambda|^{1/d}c^{-2/d}}\right)^2, \qquad x \in \mathbb{R}^d,$$

and set $p = d = 2$ for visual purposes. Figure 1 shows the pair correlation functions (PCFs) of two Gaussian-like DPPs with different $\Lambda \in \mathbb{R}^{2\times 2}$. In panel (a), Λ has eigenvectors $e_1 = (1,0)^{\mathrm{T}}$, $e_2 = (0,1)^{\mathrm{T}}$ and eigenvalues $\lambda_1 = 1$, $\lambda_2 = \lambda(= 4)$, which induces stronger repulsion along the horizontal axis than along the vertical one. In panel (b), Λ has eigenvectors $e_1 = (\sqrt{2}/2)(1,1)^{\mathrm{T}}$, $e_2 = (\sqrt{2}/2)(-1,1)^{\mathrm{T}}$ and eigenvalues $\lambda_1 = 1, \lambda_2 = \lambda(= 4)$, which induces stronger repulsion along the bisector of the first quadrant than along the orthogonal direction.

Although the pair correlation function is useful for visualizing the anisotropy of repulsiveness, it provides little information on how much repulsiveness is induced by the DPP. Several repulsiveness measures have been proposed for spatial point processes. See, e.g., Lavancier et al. (2015) and Biscio & Lavancier (2016) for a few quantitative indexes. Here, we focus in particular on coefficient p_0 introduced by Møller & O'Reilly (2021), which is a global index of repulsion quantifying the effect of a point on the expected number of points in the process. For a stationary DPP, $p_0 = \rho^{-1}\int |K_0(x)|^2\,\mathrm{d}x$, where $\rho := K_0(0)$. The

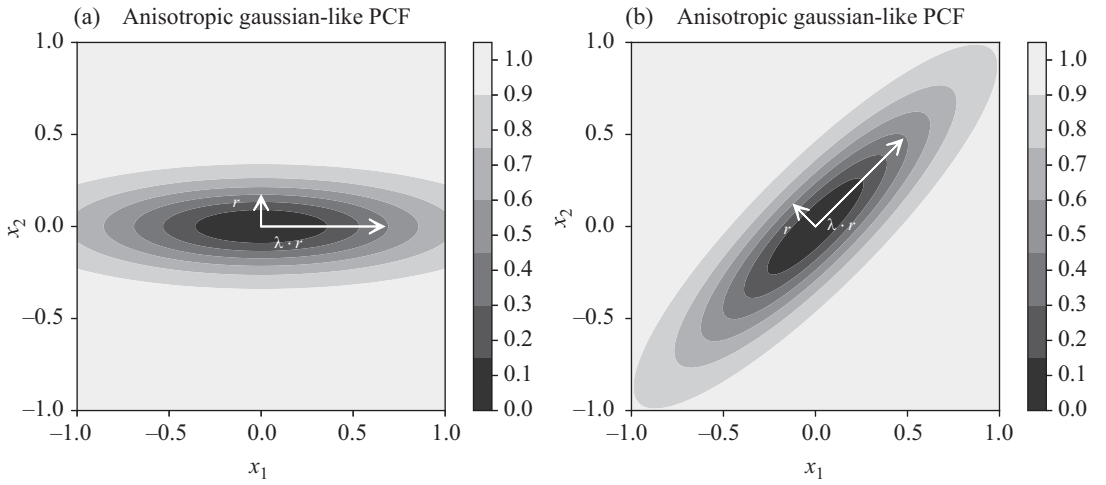

Fig. 1. Pair correlation function of two Gaussian-like DPPs, showing (a) strong repulsion along the horizontal direction and (b) strong repulsion along the bisector of the first quadrant.

larger p_0, the more repulsive the DPP. For our Gaussian-like DPP in Corollary 1, standard computations lead to $p_0 = \rho\pi^{d/2}c^{-1}$, which shows that the amount of repulsiveness in the process does not depend on Λ.

2.2. *The APPLAM model*

The anisotropic repulsive point process latent mixture model proposed here assumes likelihood (2). We complete the prior specification as follows. First, following common practice in mixture models, conditional on $\{\mu_1, \ldots, \mu_m\}$, indeed conditionally only on m, we assume that

$$w_1, \ldots, w_m \mid m \sim \text{Dir}(\alpha, \ldots, \alpha), \qquad \Delta_1, \ldots, \Delta_m \mid m \overset{\text{iid}}{\sim} iw_d(\nu_0, \Psi_0), \tag{6}$$

where $iw_d(\nu_0, \Psi_0)$ denotes the d-dimensional inverse Wishart distribution, with $\nu_0 > d - 1$ degrees of freedom and mean $\Psi_0/(\nu_0 - d - 1)$. Moreover, as is common in latent factor models, we assume $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ in (2) to be a diagonal matrix and

$$\sigma_j^2 \overset{\text{iid}}{\sim} ig(a_\sigma, b_\sigma) \qquad (j = 1, \ldots, p), \tag{7}$$

i.e., their marginal prior distribution is inverse gamma with mean $b_\sigma/(a_\sigma - 1)$. As far as matrix Λ is concerned, we impose sparsity through the prior. Indeed, sparse priors are often employed in latent factor models to avoid overparameterization. In particular, we assume the Dirichlet–Laplace prior (Bhattacharya et al., 2015) with parameter a as the prior for Λ, that is, denoting with λ_{jh} the elements of Λ,

$$\begin{aligned} &\lambda_{jh} \mid \phi, \tau, \psi \overset{\text{ind}}{\sim} \mathcal{N}(0, \psi_{jh}\phi_{jh}^2\tau^2), \qquad j = 1, \ldots, p,\ h = 1, \ldots, d, \\ &\text{vec}(\phi) \sim \text{Dir}(a, \ldots, a), \qquad \psi_{jh} \overset{\text{iid}}{\sim} \text{Exp}(1/2), \qquad \tau \sim \text{Ga}(pda, 1/2), \end{aligned} \tag{8}$$

where, for any $p \times d$ matrix A, $\text{vec}(A)$ denotes the vector of dimension $p \times d$ such that $\text{vec}(A)_{p(h-1)+j} = (A)_{j,h}$. This prior is very popular in the literature on Bayesian factor

models because of the ease of sampling and good frequentist properties (Bhattacharya et al., 2015). We have assumed prior independence among the blocks of parameters above.

The key point in our model is the prior specification for the latent factor centres $\{\mu_1, \ldots, \mu_m\}$. To introduce repulsiveness between the points $\{\Lambda\mu_h\}_{h=1}^m$, we first fix a compact $R \subset \mathbb{R}^d$, and assume that the locations $\{\mu_1, \ldots, \mu_m\}$ are distributed as an anisotropic DPP on R, conditioned on $m \geqslant 1$. As shown below, the choice of R plays a very limited role, and we discuss how to fix it in § 2.3 below. Conditioning on the DPP being nonempty is necessary for our formulation of the model, as (2) is not well defined if $m = 0$. Therefore, we assume the following density for the locations $\{\mu_1, \ldots, \mu_m\}$ and their cardinalities:

$$p(\{\mu_1, \ldots, \mu_m\} \mid \Lambda) = f_{\mathrm{DPP}}(\mu \mid \rho, \Lambda, K_0; R) = \frac{e^{|R|-D}}{1 - e^{-D}} \det\{C(\mu_h, \mu_k)\}_{h,k=1,\ldots,m}, \qquad m \geqslant 1,\ \mu_1, \ldots, \mu_m \in R, \quad (9)$$

and $p(\emptyset \mid \Lambda) = 0$, with D and C defined just after (4). The above notation $f_{\mathrm{DPP}}(\mu \mid \rho, \Lambda, K_0; R)$ makes explicit that the density depends on R, the intensity ρ, and the stationary kernel K_0 and that anisotropy is driven by Λ.

2.3. *Prior elicitation for the anisotropic DPP*

The prior distributions for Σ, the Δ_h and Λ are shared between the LAMB and APPLAM models, and we adopt the default values suggested by Chandra et al. (2023). See the Supplementary Material for further details. Here, we specifically focus on those hyperparameters defining the DPP density prior in (9) when K_0 is the anisotropic Gaussian-like kernel in Corollary 1, i.e., R, ρ and c.

First, we discuss the choice of the compact set R of $\mathbb{R}^d$, which is the support of the μ_h. Marginally, each μ_h is uniformly distributed on R. We argue that a reasonable choice is to assume that $R = [-r, r]^d$, i.e., a hypercube. Indeed, there is no reason to assume a priori that the latent factors exhibit a higher variance along any particular axis. Moreover, such a hypercube should be centred in the origin as in traditional latent factor models, whereby the latent factors are given a standard Gaussian prior. Clearly, there is nonidentifiability between Λ and r. Indeed, letting $t > 0$ and $\tilde{\Lambda} = t\Lambda$, $\tilde{R} = [-r/t, r/t]^d$ leaves the likelihood of the data invariant. Therefore, the choice of r plays a limited role in the model definition. In particular, our experience shows that posterior inference is robust if r is sufficiently large, i.e., $r \geqslant 5$. Therefore, we set $R = [-10, 10]^d$ in the rest of the paper.

Parameters ρ and c control the intensity and the repulsiveness of the process. As noted in § 2.1, these quantities are intimately related in DPPs. While $\rho|R|$ is the expected number of points in R of the DPP, in (9) we condition on the process being nonempty so that the role of ρ is not as clear as in (4). Therefore, we suggest fixing those parameters by looking only at the repulsiveness induced by the DPP. In the Supplementary Material, we exploit the equivalent representation of the anisotropic DPP given in Theorem 2 to obtain insights into how such parameters govern the so-called repulsion range, i.e., how far should two cluster centres be to influence each other, as well as the strength of the repulsion across close points. Such interpretation can be used for prior elicitation by users by specifying a value, or grid of values, for the repulsion range. Moreover, we also propose a heuristic approach based on the observed data to set reasonable ranges for hyperparameters ρ, c. In particular, we follow Lavancier et al. (2015) and Bianchini et al. (2020) and parameterize the DPP by (ρ, s) such that $\rho = s\rho_{\max} \equiv sc(2\pi)^{-d/2}$, $s \in (0, 1)$; cf. Corollary 1. In such a way, s takes the interpretation of the strength of repulsion for a fixed ρ.

3. POSTERIOR SIMULATION

3.1. *Approximation of the DPP density*

The point process density in (9) cannot be evaluated in closed form due to the infinite series in the expressions of D and C. We follow Lavancier et al. (2015) and approximate it with $f^{\text{app}}_{\text{DPP}}$, which is as in (9) where D is replaced by $D^{\text{app}} = -\sum_{k\in\mathbb{Z}^d_N} \log\{1-\varphi(k)\}$ and C by

$$C^{\text{app}}(x,y) = \frac{1}{|R|}\sum_{k\in\mathbb{Z}^d_N} \frac{\varphi(k)}{1-\varphi(k)} \exp\left(\frac{2\pi i}{|R|^{1/d}}\langle k, x-y\rangle\right) \tag{10}$$

with $\mathbb{Z}^d_N = \{-N,\dots,N\}^d$ and $\langle\cdot,\cdot\rangle$ denoting the scalar product. Note that $\mu \sim f^{\text{app}}_{\text{DPP}}(\rho,\Lambda,C^{\text{app}};R)$ is still a DPP on R, conditioned that it is nonempty. The truncation level N allows us to trade off the accuracy of the approximation and the computational complexity. In particular, observe that evaluating (10) scales with N^d. Moreover, the truncation in (10) sets an upper bound of $(2N+1)^d$ on the number of points in the DPP (Lavancier et al., 2015). This approximation can be seen as analogous to the truncation of stick-breaking priors commonly adopted in Bayesian nonparametric mixture models (Ishwaran & James, 2001). For the Dirichlet process, a standard choice is truncating the prior to 50 support points. Similarly, when $d = 2$, which is the smallest value we consider, and $N = 3$, we are setting here an upper bound of 49 points in the DPP, which increases exponentially with the dimension. In the Supplementary Material, we report a simulation showing that the approximation error introduced by the truncation is, in fact, negligible when Λ is either fixed to the identity matrix, or random and distributed according to prior (8) or the posterior.

From (10), it is also clear why the definition of the DPP in Theorem 2 is not operational. Indeed, to approximate the DPP density as done in (9) in combination with (10), one needs the Fourier transform of the DPP kernel. For the anisotropic DPP in Theorem 1, an analytical expression is available. In contrast, working with the construction in Theorem 2 requires finding the eigendecomposition of $\tilde{K}_0$, which is a challenging task as it is defined on the manifold $B = \Lambda\mathbb{R}^d$, where Λ is random. In the context of the Gibbs sampling algorithm described below, this would entail numerically performing such a decomposition anytime Λ needs to be sampled, resulting in a huge computational burden.

3.2. *The Gibbs sampling algorithm*

Including the mixture weights w, defined in the first line of (6), as finite-Dirichlet distributed in the state space is not convenient, as the sum-to-one constraint on w would lead to complex split-merge reversible jump moves with poor mixing of the chain. As Beraha et al. (2022), we consider the following equivalent characterization of the prior for w:

$$w = \left(\frac{S_1}{T},\dots,\frac{S_m}{T}\right), \qquad T = \sum_{h=1}^{m} S_h, \qquad S_h \overset{\text{iid}}{\sim} \text{Ga}(\alpha,1).$$

Conditional on c, we consider $\mu = \mu^{(a)} \cup \mu^{(na)}$, $S = (S^{(a)}, S^{(na)})$ and $\Delta = (\Delta^{(a)}, \Delta^{(na)})$, divided into allocated and nonallocated components, denoted with (a) and (na) superscripts, respectively. That is, $\mu^{(a)} = \{\mu_{c_i},\, i = 1,\dots,n\}$, $\mu^{(na)} = \mu \setminus \mu^{(a)}$, and analogously for the terms involving the S_h and Δ_h.

Combining likelihood (2) with the prior as in (6)–(9), we can see that the joint distribution of the data and parameters has a density, with respect to some dominating measure on the proper space; see the Supplementary Material for this density and the associated dominating measure. Normalization of the weights leads to a term $T^{-n} = (\sum S_{h_1}^{(a)} + \sum S_{h_2}^{(na)})^{-n}$ in the expression of the joint density, which makes it impossible to factorize the density according to the allocated and nonallocated components. To overcome this issue, as Beraha et al. (2022), we introduce an auxiliary random variable $u \mid T \sim \text{Ga}(n, T)$. We report the joint density of the data and parameters in the Supplementary Material.

We propose a Gibbs sampler algorithm, recurring to a Metropolis step when the corresponding full conditionals cannot be sampled directly. Most of the updates are straightforward, except for two steps, which are outlined below. The full description of the algorithm is given in the Supplementary Material.

To update the nonallocated variables $(\mu^{(na)}, S^{(na)}, \Delta^{(na)})$, we disintegrate the joint full conditional of the nonallocated variables as

$$p(\mu^{(na)}, s^{(na)}, \Delta^{(na)} \mid \text{rest}) = p(\mu^{(na)} \mid \text{rest}) p(s^{(na)} \mid \mu^{(na)}, \text{rest}) p(\Delta^{(na)} \mid \mu^{(na)}, \text{rest}),$$

where 'rest' identifies all the variables except for $(\mu^{(na)}, s^{(na)}, \Delta^{(na)})$. Then $\mu^{(na)} \mid$ rest is a Gibbs point process with density

$$p(\{\mu_1^{(na)}, \ldots, \mu_\ell^{(na)}\} \mid \text{rest}) \propto f_{\text{DPP}}^{\text{app}}(\{\mu_1^{(na)}, \ldots, \mu_\ell^{(na)}\} \cup \mu^{(a)} \mid \rho, \Lambda, K_0; R) \psi(u)^\ell,$$

where $\psi(u) = \mathbb{E}(e^{-uS})$. See also Beraha et al. (2022). We employ the birth-death Metropolis-Hastings algorithm of Geyer & Møller (1994) to sample from this point process density. Given $\mu^{(na)}$, it is straightforward to show that

$$\Delta_1^{(na)}, \ldots, \Delta_\ell^{(na)} \mid \text{rest} \overset{\text{iid}}{\sim} iw_d(\nu_0, \Psi_0), \qquad S_1^{(na)}, \ldots, S_\ell^{(na)} \mid \text{rest} \overset{\text{iid}}{\sim} \text{Ga}(\alpha, 1 + u).$$

To update the latent allocation variables c, we found it useful to marginalize over the η_i to get better mixing chains. Hence, we can sample each c_i independently from a discrete distribution over $\{1, \ldots, k+\ell\}$, where k is the number of allocated components, with weights ω_{ih}:

$$\omega_{ih} \propto S_h^{(a)} \mathcal{N}_p(y_i \mid \Lambda \mu_h^{(a)}, \Sigma + \Lambda \Delta_h^{(a)} \Lambda^{\mathrm{T}}) \qquad (h = 1, \ldots, k),$$
$$\omega_{ik+h} \propto S_h^{(na)} \mathcal{N}_p(y_i \mid \Lambda \mu_h^{(na)}, \Sigma + \Lambda \Delta_h^{(na)} \Lambda^{\mathrm{T}}) \qquad (h = 1, \ldots, \ell).$$

Each evaluation of the p-dimensional Gaussian density would require $\mathcal{O}(p^3)$ operations if care is not taken. However, we take advantage of the special structure of the covariance matrix. Using Woodbury's matrix identity, we have

$$(\Sigma + \Lambda \Delta \Lambda^{\mathrm{T}})^{-1} = \Sigma^{-1} - \Sigma^{-1} \Lambda (\Delta^{-1} + \Lambda^{\mathrm{T}} \Sigma^{-1} \Lambda)^{-1} \Lambda^{\mathrm{T}} \Sigma^{-1},$$

and hence we now need to compute the inverse of a $d \times d$ matrix. Therefore, evaluating the quadratic form in the exponential requires only $\mathcal{O}(p)$ computational cost. Moreover, using the matrix determinant lemma, the determinant of the covariance matrix can be computed as

$$\det(\Sigma + \Lambda \Delta \Lambda^{\mathrm{T}}) = \det(\Delta^{-1} + \Lambda^{\mathrm{T}} \Sigma^{-1} \Lambda) \det(\Delta) \det(\Sigma).$$

This is computed without additional cost by caching operations from the matrix inversion.

3.3. *Updating Λ using gradient-based Markov chain Monte Carlo algorithms*

As mentioned in § 1.2, sampling from Λ's full conditional is nontrivial. In particular, we found that the random walk Metropolis–Hastings algorithm led to very poor mixing of the Markov chain Monte Carlo (MCMC) chain, while the adaptive Metropolis–Hastings algorithm of Haario et al. (2001) is not feasible here due to the high dimensionality of Λ. Instead, we found the Metropolis-adjusted Langevin algorithm (MALA; Roberts & Tweedie, 1996) to be more adequate here. The target density is

$$p(\Lambda \mid \cdots) \propto p(y \mid \Lambda, \eta, \Sigma) p(\Lambda \mid \phi, \tau, \psi) \frac{e^{1-D^{\text{app}}}}{1-e^{-D^{\text{app}}}} \det\{C^{\text{app}}(\mu_h, \mu_k)\}_{h,k=1,\ldots,m}, \tag{11}$$

where $\cdots$ denotes all the variables except Λ and $\eta = (\eta_1, \ldots, \eta_n)$. Although not explicitly stated, D^{app} and C^{app} both depend on Λ.

To sample from (11) using MALA, we must evaluate $\nabla \log\{p(\Lambda \mid \cdots)\}$. In a preliminary investigation, we tried using automatic differentiation (AD; Griewank, 1989) to get $\nabla \log\{p(\Lambda \mid \cdots)\}$, as it requires only the implementation of a function evaluating $\log\{p(\Lambda \mid \cdots)\}$. Unfortunately, we found that this strategy is viable only in trivial scenarios, i.e., up to $p = 50$ and $d = 3$ due to RAM memory requirements. Therefore, in the following theorem, we provide the analytical expression of $\nabla \log\{p(\Lambda \mid \cdots)\}$ when the associated DPP is Gaussian like. See the Supplementary Material for the Whittle–Matérn DPP case.

THEOREM 3. *Under the Gaussian-like DPP prior, the gradient of the log-full-conditional density of Λ is given by*

$$\begin{aligned}\nabla \log p(\Lambda \mid \cdots) = \Sigma^{-1} \sum_{i=1}^{n} (y_i - \Lambda \eta_i) \eta_i^{\mathrm{T}} - \frac{1}{(\psi \odot \phi^2)\tau^2} \odot \Lambda \\ + (2\pi^2 c^{-2/d}) \sum_{k \in \mathbb{Z}_N^d} g^{(k)} \frac{\varphi(k)}{\{1-\varphi(k)\}^2} \left\{ \frac{1-\varphi(k)}{1-e^{-D^{\text{app}}}} - v_k^T (C^{\text{app}})^{-1} u_k \right\},\end{aligned}$$

where $\varphi(k)$ is defined in (5). Here $\odot$ denotes the elementwise Hadamard product,

$$g^{(k)} = 2|\Lambda^T \Lambda|^{1/d} \Lambda (\Lambda^T \Lambda)^{-1} \left[\frac{1}{d} k^T (\Lambda^T \Lambda)^{-1} k \mathbb{1}_d - k\{(\Lambda^T \Lambda)^{-1} k\}^T \right],$$

u_k and v_k are m-dimensional column vectors for each $k \in \mathbb{Z}^d$ with entries

$$(u_k)_j = e^{2\pi i \langle k, T\mu_j \rangle}, \qquad (v_k)_j = e^{-2\pi i \langle k, T\mu_j \rangle} \qquad (j = 1, \ldots, m),$$

and C^{app} is the $m \times m$ matrix with entries $C^{\text{app}}(\mu_h, \mu_k)$, $h, k = 1, \ldots, m$.

Figure S5 in the Supplementary Material reports a comparison of the memory requirements and the runtime execution using the AD gradients or our analytical expressions. In particular, using the AD gradients requires roughly $100\times$ more memory, which makes a significant difference in practice since this is the bottleneck of our algorithm. The runtimes are $10\times$ larger when using AD as well. The stepsize parameter of the MALA algorithm is tuned running short preliminary chains to get an acceptance rate of around 20%. In particular, we found that values between 10^{-8} and 10^{-10} usually give a good mixing of the MCMC chain.

4. Simulation studies

The Supplementary Material shows several simulated examples aimed at showcasing different aspects of our model. Beyond investigating the effect of truncating the DPP density, and the usefulness of the analytic expression of the gradients found in Theorem 3, we show the need of assuming the anisotropic DPP, as opposed to an isotropic DPP prior, in the context of latent mixture models and a procedure to fix hyperparameters. We also demonstrate the robustness of posterior inference to the hyperparameters of the anisotropic DPP prior for the latent cluster centres.

The natural competitor of our APPLAM model is the LAMB model by Chandra et al. (2023). Chandra et al. (2023) already made a compelling argument in favour of latent mixture models compared to more heuristic approaches based on a two-step procedure. Therefore, we limit ourselves to showing the robustness of the APPLAM model to model misspecification and the LAMB model's lack thereof. Specifically, we consider two simulation settings, where model misspecification occurs either at the latent level, i.e., the latent scores η_i are simulated from a mixture of t distributions, or at the likelihood level, i.e., the observations, conditional on all parameters, follow a t distribution. In both cases, for various choices of the latent dimension $d = \{4, 8\}$, data dimension $p = \{500, 1000\}$ and 100 independent and identically distributed replicates of the datasets, we demonstrate how the LAMB model significantly overestimates the number of clusters, contrary to our APPLAM model. Moreover, we show that the APPLAM model produces more reliable cluster estimates than the LAMB model.

In summary, our simulations empirically confirm the intuition that repulsive mixtures are more robust than nonrepulsive ones in the presence of model misspecification. Moreover, they also justify the introduction of the anisotropic DPPs presented here, as it is clear that forcing isotropic repulsion across cluster centres does not translate to well-separated clusters of data a posteriori. On the contrary, the isotropic models might greatly overestimate the number of clusters.

5. Joint species modelling

We apply our APPLAM model to data collecting the occurrence of plant species at different sites of the Bauges Natural Regional Park in France. The data are available from the Alpine Botanical Conservatory and have been previously investigated by Thuiller et al. (2018) and Bystrova et al. (2021). Specifically, the dataset we analyse refers to $n = 1139$ sites and, for each site, the presence or absence of $p = 123$ different plant species is reported. Our goal is to infer the clustering structure of the sites, such that sites belonging to the same cluster show similar patterns concerning the occurrence of the plant species. The data are binary; precisely, let $z_i \in \{0, 1\}^p$, $i = 1, \ldots, n$, such that $z_{i,j} = 1$ if plant species j occurs at site i, and $z_{i,j} = 0$ otherwise. We assume that $z_{i,j} = \mathbb{1}_{[0,\infty)}(y_{i,j})$ and apply the APPLAM model to the y_i, assuming the Gaussian-like DPP. To avoid nonidentifiability, we fix the Δ_h in (6) as the $d \times d$ identity matrices. Posterior simulation requires minor modifications to the Gibbs sampling algorithm, and we discuss them in the Supplementary Material.

We rely on model selection to select the hyperparameters. In particular, we examine a total of 144 models where we vary the degree of repulsiveness by tuning (ρ, c), as well as other hyperparameters; see the Supplementary Material for more details. From Figs. S2 and S3 within the Supplementary Material, it is clear that if $\rho|R|$ is large and/or the strength of

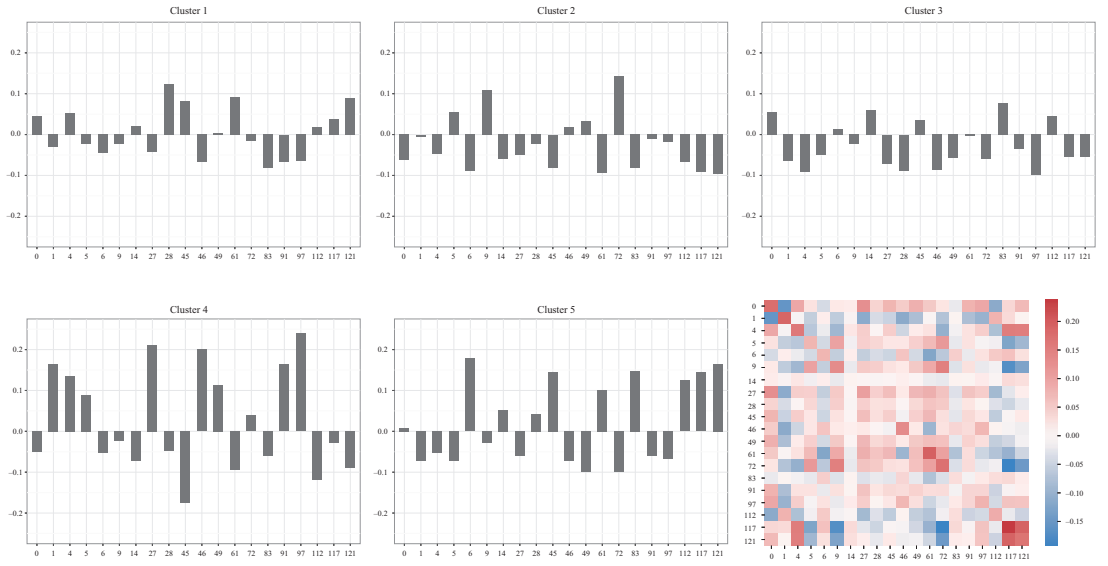

Fig. 2. Values of $\delta_{c,j}$ for the 20 species identifying the five clusters and (bottom right) posterior estimate of $\Lambda\Lambda^{\mathrm{T}}$ restricted to the 20 species.

repulsion s is mild, the number of estimated clusters increases, which confirms our insights from the simulation studies. In particular, the number of estimated clusters ranges between 3 and 15; see also Tables S1 and S2 within the Supplementary Material. We focus only on models for which the posterior mean of the number of clusters is less than 8 and select the best one by maximizing the Watanabe–Akaike information criterion index (Watanabe, 2013), which is standard in model selection. This yields the following choice of hyperparameters: $d = 3$, $\rho|R| = 0.1$, $s = 0.9$, $\alpha = 10^{-3}$ and $a_\sigma = 2, b_\sigma = 1$. When α is small, as in this case, it implies a sparse marginal prior for the mixture weights w in (6) (Rousseau & Mengersen, 2011).

We discuss here the estimated clusters, obtained by minimizing the posterior expectation of Binder's loss function (Binder, 1978), under this last hyperparameter selection. We consider five estimated clusters, with cardinalities between 200 and 250. To interpret the estimated clusters, we look at the patterns of presence and absence of species in each cluster. Specifically, for $j = 1, \ldots, p$, let $\bar{z}_j$ and $\bar{z}_{c,j}$ be the empirical frequencies of the jth species in the whole sample and in the cth cluster, respectively. For each cluster, we select six species that better represent it by choosing the three species that maximize, respectively minimize, $\delta_{c,j} = (\bar{z}_{c,j} - \bar{z}_j)$. In total, we find 20 species that best describe the different clusters. We report the corresponding $\delta_{c,j}$ in Fig. 2. Cluster 3 does not significantly depart from the sample averages, with $|\delta_{3,j}| < 0.1$ for all the selected species, but with an overall prevalence of lower-than-average species presence. Cluster 4 features the most extreme deviations from the mean compared to the other clusters, especially higher-than-average deviations. Indeed, species 27 (the common rock-rose), 46 (the great yellow gentian) and 97 (sweet vernal grass) have a $\delta_{4,j}$ around or higher than 0.2. However, this cluster also records the most extreme lower-than-average deviations, with the presence of species 45 (the common beech) lower than 0.15 compared to the associated sample average and species 112 (the silver fir) lower than 0.1. Cluster 5 differs from the other clusters mainly for the higher presence of species 6 (ivy), 45 (the common beech), 83 (sweet woodruff), 112 (the silver fir), 117

(field maple) and 121 (a blackberry species). The common beech helps explain the difference between clusters 4 and 5. Finally, cluster 2 is mainly characterized by the higher presence of species 72 (Saint Anthony's Turnip), while cluster 1 records a higher presence of species 28 (common hazel). The bottom right panel of Fig. 2 reports the posterior expectation of $\Lambda\Lambda^{\mathrm{T}}$ restricted to the 20 species explaining the five estimated clusters. A nonnegligible negative correlation value is estimated between species 72 (Saint Anthony's Turnip) and 117 (the field maple): this can be observed in the empirical proportions of these two species within each of the five estimated clusters.

6. Discussion

Model-based clustering of moderate- or large-dimensional data is notoriously difficult. We have proposed a model for simultaneous dimensionality reduction and clustering, by assuming a mixture model for the latent scores, which are then linked to the observations via a Gaussian latent factor model. This approach was recently investigated by Chandra et al. (2023). The authors used a factor-analytic representation and assumed a mixture model for the latent factors. However, performance can deteriorate in the presence of model misspecification. Assuming a repulsive point process prior for the component-specific means of the mixture for the latent scores is shown to yield a more robust model that outperforms the standard LAMB model in several simulated scenarios. To favour well-separated clusters of data, the repulsive point process must be anisotropic, and its density should be tractable for efficient posterior inference. We address these issues by proposing a general construction for anisotropic determinantal point processes.

The bottleneck in our sampling algorithm is the spectral approximation of the DPP density, which has a computational cost that scales exponentially in d, the dimension of the latent factors. It is common practice to set small values for d in latent factor models. Nonetheless, for moderate or large values of d, the approach of Bardenet & Titsias (2015) would be more efficient, at the price that the parameters in the model are harder to interpret. Another approximation of the DPP density is provided by Poinas & Lavancier (2021), also suitable for non-hyper-rectangular domains. In the case of projection DPPs, i.e., when the eigenvalues of the spectral decomposition are either zero or one, we could sample exactly from the full conditional of the nonallocated components using the methods of Lavancier & Rubak (2023) instead of employing a birth-death Metropolis–Hastings move.

We have not investigated the estimation of the factor-loading matrix Λ, which could be of interest to explain the correlation structure of the data. Our model inherits the rotational invariance property of classical factor models, but this issue can be dealt with via ex post procedures for the estimation of Λ. See, for instance, Papastamoulis & Ntzoufras (2022) and the references therein.

Acknowledgement

Beraha and Guglielmi acknowledge support from MUR, under grant Dipartimento di Eccellenza 2023–2027. Beraha received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme. Guglielmi was partially supported by MUR - Prin 2022, funded by the European Union - Next Generation EU. We gratefully acknowledge the DataCloud laboratory (https://datacloud.polimi.it).

SUPPLEMENTARY MATERIAL

The Supplementary Material contains further details on the MCMC algorithm, a comparison between the APPLAM and LAMB models on simulated datasets and proofs.

REFERENCES

ASCOLANI, F., LIJOI, A., REBAUDO, G. & ZANELLA, G. (2023). Clustering consistency with Dirichlet process mixtures. *Biometrika* **110**, 551–8.

BACCELLI, F., BŁASZCZYSZYN, B. & KARRAY, M. (2020). *Random Measures, Point Processes, and Stochastic Geometry*. *Inria:* https://hal.inria.fr/hal-02460214/.

BARDENET, R. & TITSIAS, M. (2015). Inference for determinantal point processes without spectral knowledge. In *Proc. 28th Int. Conf. Neural Info. Proces. Syst.*, Ed. C. Cortes, D. D. Lee, M. Sugiyama and R. Garnett, pp. 3393–401. Cambridge, MA: MIT Press.

BERAHA, M., ARGIENTO, R., MØLLER, J. & GUGLIELMI, A. (2022). MCMC computations for Bayesian mixture models using repulsive point processes. *J. Comp. Graph. Statist.* **31**, 422–35.

BHATTACHARYA, A., PATI, D., PILLAI, N. S. & DUNSON, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *J. Am. Statist. Assoc.* **110**, 1479–90.

BIANCHINI, I., GUGLIELMI, A. & QUINTANA, F. A. (2020). Determinantal point process mixtures via spectral density approach. *Bayesian Anal.* **15**, 187–214.

BINDER, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31–8.

BISCIO, C. A. N. & LAVANCIER, F. (2016). Quantifying repulsiveness of determinantal point processes. *Bernoulli* **22**, 2001–28.

BYSTROVA, D., POGGIATO, G., BEKTAŞ, B., ARBEL, J., CLARK, J. S., GUGLIELMI, A. & THUILLER, W. (2021). Clustering species with residual covariance matrix in joint species distribution models. *Front. Ecol. Evol.* **9**, 601384.

CAI, D., CAMPBELL, T. & BRODERICK, T. (2021). Finite mixture models do not reliably learn the number of components. In *Proc. 38th Int. Conf. Mach. Learn.*, pp. 1158–69. PMLR.

CELEUX, G., KAMARY, K., MALSINER-WALLI, G., MARIN, J.-M. & ROBERT, C. P. (2019). Computational solutions for Bayesian inference in mixture models. In *Handbook of Mixture Analysis*, Ed. S. Fruhwirth-Schnatter, G. Celeux and C. P. Robert, pp. 73–96. New York: Chapman and Hall/CRC.

CHANDRA, N. K., CANALE, A. & DUNSON, D. B. (2023). Escaping the curse of dimensionality in Bayesian model-based clustering. *J. Mach. Learn. Res.* **24**, 6884–925.

DALEY, D. J. & VERE-JONES, D. (2008). *An Introduction to the Theory of Point Processes*, vol. II, 2nd ed. New York: Springer.

DUAN, L. L. & DUNSON, D. B. (2021). Bayesian distance clustering. *J. Mach. Learn. Res.* **22**, 1–27.

FRUHWIRTH-SCHNATTER, S., CELEUX, G. & ROBERT, C. P. (2019). *Handbook of Mixture Analysis*. New York: Chapman and Hall/CRC.

FÚQUENE, J., STEEL, M. & ROSSELL, D. (2019). On choosing mixture components via non-local priors. *J. R. Statist. Soc.* B **81**, 809–37.

GEYER, C. J. & MØLLER, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Statist.* **21**, 359–73.

GRIEWANK, A. (1989). On automatic differentiation. In *Mathematical Programming: Recent Developments and Applications*, Ed. M. Iri and K. Tanabe, pp. 83–107. Dordrecht: Springer.

HAARIO, H., SAKSMAN, E. & TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7**, 223–42.

ISHWARAN, H. & JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Am. Statist. Assoc.* **96**, 161–73.

LAVANCIER, F., MØLLER, J. & RUBAK, E. (2015). Determinantal point process models and statistical inference. *J. R. Statist. Soc.* B **77**, 853–77.

LAVANCIER, F. & RUBAK, E. (2023). On simulation of continuous determinantal point processes. *arXiv:* 2301.11081v2.

LOPES, H. F. (2014). Modern Bayesian factor analysis. In *Bayesian Inference in the Social Sciences*, Ed. I. Jeliazkov and X.-S. Yang, pp. 115–53. Hoboken, NJ: John Wiley and Sons.

MACCHI, O. (1975). The coincidence approach to stochastic point processes. *Adv. App. Prob.* **7**, 83–122.

MALSINER-WALLI, G., FRÜHWIRTH-SCHNATTER, S. & GRÜN, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statist. Comp.* **26**, 303–24.

MILLER, J. W. & HARRISON, M. T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *J. Mach. Learn. Res.* **15**, 3333–70.

MØLLER, J. & O'REILLY, E. (2021). Couplings for determinantal point processes and their reduced palm distributions with a view to quantifying repulsiveness. *J. Appl. Prob.* **58**, 469–83.

Møller, J., Pettitt, A. N., Reeves, R. & Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* **93**, 451–8.

Møller, J. & Waagepetersen, R. P. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. Boca Raton, FL: Chapman and Hall/CRC Press.

Murray, I., Ghahramani, Z. & MacKay, D. J. C. (2006). MCMC for doubly-intractable distributions. In *Proc. 22nd Conf. Uncertainty Artif. Intel.*, pp. 359–66. Washington, DC: AUAI Press.

Natarajan, A., De Iorio, M., Heinecke, A., Mayer, E. & Glenn, S. (2023). Cohesion and repulsion in Bayesian distance clustering. *J. Am. Statist. Assoc.* **119**, 1374–84.

Neal, R. M. (2003). Density modeling and clustering using Dirichlet diffusion trees. In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, Ed. V. Lindley, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. Smith, and M West, pp. 619–29. Oxford, England: Oxford Academic.

Papastamoulis, P. & Ntzoufras, I. (2022). On the identifiability of Bayesian factor analytic models. *Statist. Comp.* **32**, 23.

Petralia, F., Rao, V. & Dunson, D. (2012). Repulsive mixtures. In *Proc. 25th Int. Conf. Neural Info. Proces. Syst.*, vol. 2, pp. 1889–97. Red Hook, NY: Curran Associates.

Poinas, A. & Lavancier, F. (2021). Asymptotic approximation of the likelihood of stationary determinantal point processes. *Scand. J. Statist.* **50**, 842–74.

Quinlan, J. J., Quintana, F. A. & Page, G. L. (2021). Parsimonious hierarchical modeling using repulsive distributions. *Test* **30**, 445–61.

Roberts, G. O. & Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2**, 341–63.

Rousseau, J. & Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Statist. Soc.* B **73**, 689–710.

Teh, Y., Daume III, H. & Roy, D. M. (2007). Bayesian agglomerative clustering with coalescents. In *Proc. 20th Int. Conf. Neural Info. Proces. Syst.*, pp. 1473–80. Red Hook, NY: Curran Associates.

Thuiller, W., Guéguen, M., Bison, M. & Duparc, E. A. (2018). Combining point-process and landscape vegetation models to predict large herbivore distributions in space and time—a case study of *Rupicapra rupicapra*. *Divers. Distrib.* **24**, 352–62.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.* **14**, 867–97.

Xie, F. & Xu, Y. (2019). Bayesian repulsive Gaussian mixture model. *J. Am. Statist. Assoc.* **115**, 187–203.

Xu, Y., Müller, P. & Telesca, D. (2016). Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics* **72**, 955–64.

[*Received on* 21 *March* 2023. *Editorial decision on* 3 *October* 2024]