Anisotropic determinantal point processes and their application in Bayesian mixtures

Processi di punto anisotropici di tipo determinantal e loro applicazione nelle misture bayesiane

Lorenzo Ghilotti, Mario Beraha and Alessandra Guglielmi

Abstract Repulsive mixture models have recently gained visibility in Bayesian statistics. In such models, a finite repulsive point process is assumed as prior distribution for the number of components and component-specific parameters. We assume a determinantal point process as such prior, proposing a simple construction of anisotropic determinantal point processes, that can better characterize repulsion when data have different scales along the axes. In turn, this produces better cluster estimates. We discuss the model on simulated data.

Abstract *I modelli mistura repulsivi hanno avuto di recente un incremento di visibilità in statistica bayesiana. In tali modelli, si assume un processo di punto finito repulsivo come prior sul numero di componenti e sui parametri specifici di ogni componente. In particolare, noi assumiamo un processo di punto di tipo determinantal, proponendo una semplice costruzione per processi di punto di tipo determinantal anisotropi, che possano meglio caratterizzare la repulsione quando i dati hanno dispersioni differenti lungo gli assi. Di conseguenza, modelli di questo tipo producono stime dei clusters migliori. Discutiamo il nostro modello su dati simulati.*

Key words: repulsive mixture models, determinantal point processes, anisotropic covariance function, spectral density

1 Introduction

Mixture models are a popular framework in Bayesian inference, providing useful tools for density estimation problems and cluster detection; see [4].

Lorenzo Ghilotti[1], Mario Beraha[1, 2] and Alessandra Guglielmi[1]
[1]Department of Mathematics, Politecnico di Milano, Milano, Italy
[2]Department of Computer Science, Università degli Studi di Bologna, Bologna, Italy
e-mail: lorenzo1.ghilotti@mail.polimi.it, {mario.beraha, alessandra.guglielmi}@polimi.it

Mixture models assume that data arise from one of M homogeneous populations, each suitably modelled by a density $\{f_m\}_{m=1}^M$, henceforth denoted as *component*. A set of nonnegative weights specifies the probability of each population to be selected. In the Bayesian setting, a prior is assumed on the weights, on the parameters governing the densities f_m and possibly on M. The most common formulation assumes that the parameters of the components are a priori independent and identically distributed, because of mathematical tractability, but, specifically for clustering purposes, it often reveals to be an oversimplification. As shown in [3], if the mixture model is misspecified, assuming component-specific parameters iid leads to overestimating the number of components, so that inference may produce redundant clusters of the data.

Repulsive mixture models use the notion of repulsion between cluster-specific parameters specifying prior that encourages well separated components, see for instance [2, 5] and the references therein. In particular, [1] proposes a general framework for this family of models, by assuming a *repulsive point process* as joint prior distribution for the location centers and M. Within the spectrum of repulsive point processes, *determinantal point processes* (DPPs) (see [6]) are rather appealing since they do not carry intractable normalizing constants and are defined through a covariance function. Often DPPs in the literature assume *stationary* and *isotropic* covariance functions, but this might be a modelling limitation.

In this work, we propose a simple construction for anisotropic DPPs, that preserves the analytical tractability of isotropic DPPs. The structure of the paper is as follows. Section 2 covers background material on (determinantal) point processes, while in Section 3 we introduce our anisotropic DPP. In Section 4 we assume this DPP as a joint prior for location parameters and the number of components in a Bayesian mixture model. We show the advantages of introducing anisotropism in a simulation study in Section 5.

2 Background on Determinantal Point Processes

Let $R \subseteq \mathbb{R}^d$ be a compact set, a *finite point process* X on R is a finite random subset of R. Several choices are available to characterize X. For instance, we may assign the *product density functions* $\rho^{(n)} : R^n \to [0,\infty)$, $n = 1,2,\ldots$; see [7]. Intuitively, for any pairwise distinct points $x_1,\ldots,x_n$ in R, $\rho^{(n)}(x_1,\ldots,x_n)\,dx_1\cdots dx_n$ represents the probability that X has a point in an infinitesimal small region around x_i of volume dx_i , for each $i = 1,\ldots,n$.

To define a DPP X on R, we consider a covariance function $C : R \times R \to \mathbb{R}$ and define the product density functions $\rho^{(n)}$ as

$$\rho^{(n)}(x_1,\ldots,x_n) = det\{\,[C](x_1,\ldots,x_n)\,\}, \quad (x_1,\ldots,x_n) \in R^n,\ n = 1,2,\ldots$$

where $[C](x_1,\ldots,x_n)$ is the $n \times n$ matrix with elements $C(x_i,x_j)$ and det denotes the matrix determinant. Of course, some assumptions on C guarantee the DPP to exist.

Observe that $\rho^{(n)}(x_1,\ldots,x_n) = 0$ if $x_i = x_j$, for some $i \neq j$, since, in this case, the matrix $[C](x_1,\ldots,x_n)$ is not full rank. Consequently, if C is continuous, $\rho^{(n)}(x_1,\ldots,x_n) \to 0$ if $x_i \to x_j$, for some $i \neq j$. Hence, the probability of having two points in a given neighborhood tends to zero as the size of the neighborhood shrinks. Moreover, $\rho^{(n)}(x_1,\ldots,x_n) \leq \prod_{j=1}^{n} C(x_j,x_j)$, and $C(x,x)$ represents the *intensity function* of the process. Hence, DPPs are *repulsive* point processes: in fact, the joint probability of any points configuration is smaller than the case of independent point configurations. *Stationarity* is a common assumption for a point process, describing invariance under translations in $\mathbb{R}^d$. For DPPs, it is expressed by assuming $C(x,y) = C_0(x-y)$.

Under this assumption, conditions on the existence of the process are given as conditions on the spectral density $\varphi = \mathscr{F}(C_0)$, where $\mathscr{F}$ indicates the Fourier transform. Additionally, if $\varphi < 1$, then the DPP has a density f with respect to the unit rate Poisson point process Y_1 on R. That is, letting $I(\cdot)$ denote the indicator function, it holds that

$$P(X \in F) = \mathbb{E}\left[I(Y_1 \in F) f(Y_1)\right]$$

for any collection of point patterns F contained in R.

Using the spectral density approach, [2] derived a Markov chain Monte Carlo (MCMC) sampling scheme based on split-merge reversible jump moves, while [1] proposed a Metropolis-within-Gibbs sampler based on spatial birth-death processes. In both these papers, the authors consider modeling directly φ and using the approximation of the density f proposed in [6] when $R = [-1/2,1/2]^d$. In particular, [2, 1] work with an isotropic DPP on $[-1/2,1/2]^d$ and apply an affine transformation mapping the DPP onto the smallest rectangle containing all the data.

As shown in Section 4, isotropism might produce misleading results when such a process is adopted as a prior for Bayesian mixture modeling and a more complex (anisotropic) model should be preferred.

3 Anisotropic DPPs

Suppose to be modeling points $x_1,\ldots,x_M$ through a DPP. If we assume an isotropic DPP, this would result in $C_0(x)$ of the form $C_0(\|x\|)$ for $\|\cdot\|$ the standard Euclidean norm. If the x_i's represent spatial locations, then isotropy is likely to be a justifiable assumption. However, if the x_i's represent more complex kinds of data, such as medical measurements on a patient, isotropy can be an oversimplification and more complex models could be more suitable. For instance, if $x_i \in \mathbb{R}^2$, one might want to model different scales along the two axes, i.e. having a DPP that considers close two points such as $(x,0)$, $(x+d,0)$ and distant two points such as $(0,y)$, $(0,y+d)$ for the same value of d (or viceversa). In this section, we show how this behavior can be achieved by constructing a stationary but anisotropic DPP.

Note that the kind of anisotropy we are interested in can be well represented by employing a different metric on $\mathbb{R}^d$. In particular, we consider a $d \times d$ sym-

metric positive definite matrix Λ and define $\|x\|^2_\Lambda = x^T\Lambda x$, for $x \in \mathbb{R}^d$. Through Λ it is possible to define several kind of anisotropic behaviors: for instance, if $\Lambda = diag(\lambda_1, \dots, \lambda_d)$, we could well model data that have different scales along different axes, and, by considering a full-matrix Λ we could also have different scales along directions that are not parallel to the cardinal axes.

Hence, the problem is how to define a *valid* DPP with kernel $C(x,y) = C_0(x-y)$ such that $C_0(x) = C_0(\|x\|_\Lambda)$. We were able to prove that, if B is a $p \times d$ matrix with full rank, with $p \geq d$, and $k > 0$, $\rho > 0$, then the kernel

$$C_0(x) = \rho \exp\left(-\frac{||Bx||^2}{2k}\right), \ x \in \mathbb{R}^d \tag{1}$$

defines a valid DPP. Moreover, the resulting DPP has a density with respect to the unit rate Poisson point process if $\rho < \rho_{\max}$, where $\rho_{\max} = |B^TB|^{\frac{1}{2}}k^{-d/2}/(2\pi)^{d/2}$. In this case, the Fourier transform $\varphi = \mathscr{F}(C_0)$ has a closed form expression :

$$\varphi(x) = \rho \, \frac{(2\pi k)^{d/2}}{|B^TB|^{1/2}} \exp\left(-2\pi^2 k x^T (B^TB)^{-1}x\right), \ x \in \mathbb{R}^d \tag{2}$$

Note that the expression of C_0 in (1) recovers indeed the desired kind of anisotropy; since $\|Bx\|^2 = x^TB^TBx$, it is sufficient to let $B = \Lambda^{1/2}$. Moreover, since ρ controls the intensity of the DPP, one might want to fix the maximum admissible intensity ρ_{MAX} independently of B. By applying the change of variable $c := |B^TB|^{1/2}k^{-d/2}$, and substituting $k = |B^TB|^{1/d}c^{-2/d}$, we get a new parametrization of the covariance function

$$C_0(x) = \rho \cdot \exp\left(-c^{2/d}\|Bx\|^2/\left(2|B^TB|^{1/d}\right)\right) \qquad \rho_{MAX} = c/(2\pi)^{d/2}. \tag{3}$$

It is evident that, within this parametrization, parameter c tunes the maximum intensity allowed.

4 Bayesian repulsive DPP mixture model

In this section, we introduce the Bayesian mixture model with an anisotropic DPP as a prior for the centers of the components. Let data $y_1, \dots, y_n \in \mathbb{R}^d$; we assume

$$y_i \,|\, \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, M \overset{\text{iid}}{\sim} \sum_{h=1}^{M} w_h \mathscr{N}_d(\cdot \mid \mu_h, \Sigma) \tag{4}$$

where $\boldsymbol{w} = (w_1, \dots, w_M)$ are the weights, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_M)$ are the centers of the components, M denotes the total number of components, and Σ is a covariance matrix that we assume to be known and fixed.

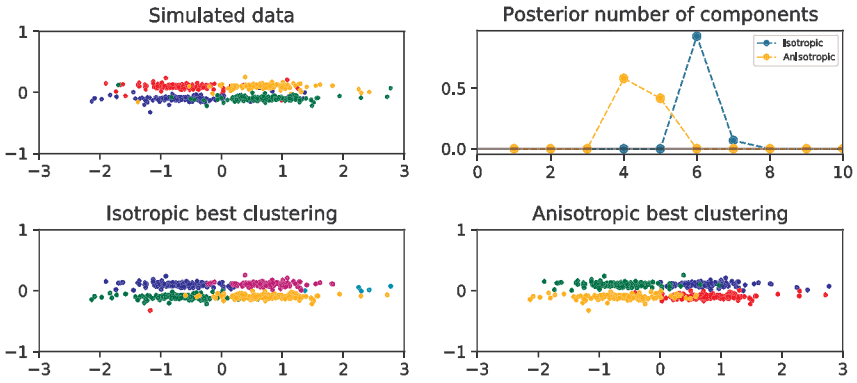

Fig. 1 Simulated dataset (top left), postiori distribution of M (top right), cluster estimate under the isotropic (bottom left) and anisotropic (bottom right) prior.

Prior assumptions. The model is completed assuming the same hierarchical prior as in [1, 2]:

$$\{\mu_1,\ldots,\mu_M;M\} \sim \text{DPP}(C_0;R), \qquad \boldsymbol{w}\,|\,M \sim \text{Dirichlet}_M(\alpha) \tag{5}$$

where $\text{DPP}(C_0;R)$ denotes the distribution of a stationary determinantal point process on the compact set $R \subset \mathbb{R}^q$ with kernel C_0 and $\text{Dirichlet}_M(\alpha)$ denotes the Dirichlet distribution on the $M-1$ dimensional simplex with parameters $(\alpha,\ldots,\alpha)$. Assuming a DPP as a prior on the locations $\mu_1,\ldots,\mu_M$ induces repulsion between them, favoring well separated components, see [1]. Note that it also determines the distribution of the number of components M.

Posterior inference. We have designed a Metropolis-within-Gibbs MCMC algorithm to sample from the posterior distribution of $(\boldsymbol{\mu},\boldsymbol{w},M)$ given $y_1,\ldots,y_n$, as in [1]. The code has been implemented in `C++`. A central building block of the proposed MCMC scheme is the approximation of the DPP density as described in [6], by means of the Fourier transform in (2).

5 Simulation study

We present a simple simulated scenario to highlight the difference between the proposed anisotropic DPP prior and previously considered (isotropic) priors and the corresponding posterior inferences. We generated $n = 600$ data from an equally weighted mixture of four bivariate Student-t distributions, with means $m_1 = [-0.7, 0.1]$, $m_2 = [-0.7, -0.1]$, $m_3 = [0.7, 0.1]$, $m_4 = [0.7, -0.1]$, the same covariance matrix $H = diag(0.1, 0.0005)$ and same degrees of freedom $\nu = 3$. Simulated data is shown in Figure 1, top left. Note that the dispersion is much more extreme along the horizontal axis than along the vertical one.

We consider two specifications for the DPP prior, fixing $R = [-2,2]^2$. The first one (isotropic) assumes B in (3) to be the identity matrix, while the second one (anisotropic) assumes $B = diag(1,5)$. We assume $c = 6$, $\rho = 0.9 \cdot \rho_{MAX}$ for both models and $\alpha = 3$; see (5). Moreover, we assume the covariance Σ in (4) as $\Sigma = \nu H$. Note that the two models differ just on the *shape* of the repulsion: while the first assumes isotropism, the second induces a stronger repulsion along the horizontal axis and a weaker one along the vertical direction.

MCMC chains were run for $20,000$ iterations, discarding the first $10,000$ and keeping one iteration every 10, for a final sample size of $1,000$. Figure 1(top right) shows the posterior distributions of M under the two priors. It is clear that the anisotropic DPP is more effective in recovering the true number of components. Moreover, Figure 1(bottom) shows cluster estimates obtained by minimizing the Binder loss function: the anisotropic DPP correctly recovers four clusters (bottom right), while the isotropic DPP (bottom left) estimates six of them.

6 Conclusion

In this paper, we have introduced a determinantal point process with anisotropism. Assuming this process as a prior in a Bayesian mixture model was shown to produce better cluster estimates in scenarios where data have different scales along different axes or directions.

The approach considered could be further extended to describe more complex models, such as an analogous of the Whittle-Matern DPP density in [6].

References

1. Beraha, M., Argiento, R., Møller, J., Guglielmi, A.: MCMC computations for Bayesian mixture models using repulsive point processes. arXiv preprint arXiv:2011.06444 (2020)
2. Bianchini, I., Guglielmi, A., Quintana, F.A.: Determinantal point process mixtures via spectral density approach. Bayesian Analysis **15**, 187–214 (2020)
3. Cai, D., Campbell, T., Broderick, T.: Finite mixture models do not reliably learn the number of components (2020)
4. Fruhwirth-Schnatter, S., Celeux, G., Robert, C.P.: Handbook of mixture analysis. CRC press (2019)
5. Fúquene, J., Steel, M., Rossell, D.: On choosing mixture components via non-local priors. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **81**(5), 809–837 (2019)
6. Lavancier, F., Møller, J., Rubak, E.: Determinantal point process models and statistical inference. Journal of the Royal Statistical Society: Series B: Statistical Methodology pp. 853–877 (2015)
7. Møller, J., Waagepetersen, R.P.: Statistical Inference and Simulation for Spatial Point Processes. CRC press (2004)