On a few statistical applications of determinantal point processes

Rémi Bardenet, Frédéric Lavancier, Xavier Mary, Aurélien Vasseur

ESAIM: PROCEEDINGS AND SURVEYS, Vol. ?, 2017, 1-10
Editors: Will be set by the publisher

ON A FEW STATISTICAL APPLICATIONS OF DETERMINANTAL POINT PROCESSES

Rémi Bardenet[1], Frédéric Lavancier[2], Xavier Mary[3] and Aurélien Vasseur[4]

Abstract. Determinantal point processes (DPPs) are a repulsive distribution over configurations of points. The 2016 conference *Journées Modélisation Aléatoire et Statistique* (MAS) of the French society for applied and industrial mathematics (SMAI) featured a session on *statistical* applications of DPPs. This paper gathers contributions by the speakers and the organizer of the session.

Résumé. Les processus ponctuels déterminantaux (DPP) sont des distributions répulsives sur des configurations de points. Au cours des journées *Modélisation Aléatoire et Statistique* (MAS) 2016 de la Société française de Mathématiques Appliquées et Industrielles (SMAI), nous avons organisé une session sur les applications statistiques des DPP. Cet article rassemble des contributions des orateurs et de l'organisateur.

1. Introduction

Determinantal point processes (DPPs) are distributions over configurations of points that encode repulsiveness in a kernel function. Since their formalization in [**?**] as models for fermions in particle physics, specific instances of DPPs have half-mysteriously appeared in fields such as probability [**?**], number theory [**?**], or statistical physics [**?**]. More recently, they have been used as models for repulsiveness in statistics [**?**] and machine learning [**?**]. During the 2016 *Journées Modélisation Aléatoire et Statistique* (MAS) of the French society for applied and industrial mathematics (SMAI), we organized a session specifically on *statistical* applications of DPPs. This paper gathers contributions by the speakers (FL, XM, AV) and the organizer (RB), which cover and extend the talks of that session. Jamal Najim also contributed a talk to the session, on characterizing the distribution of eigenvalues of large covariance matrices. The content of his talk has already been the topic of a recent paper in the MAS proceedings [**?**].

[1] CNRS & CRIStAL, Université de Lille, France. remi.bardenet@gmail.com

[2] Jean Leray Mathematics Institute, University of Nantes, and Inria, Centre Rennes Bretagne Atlantique, France. frederic.lavancier@univ-nantes.fr

[3] Université Paris-Nanterre, laboratoire Modal'X, Paris, France. xavier.mary@u-paris10.fr

[4] Institut Mines-Télécom, Télécom ParisTech, LTCI, Paris, France. aurelien.vasseur@telecom-paristech.fr

The paper follows the schedule of the MAS session. Section **??** gives a tutorial introduction to DPPs, Section **??** investigates DPPs as a model in spatial statistics. In Section **??**, DPPs are shown to have desireable properties in survey sampling design. In Section **??**, DPPs are used as a model for the spatial distribution of antennas in cellular networks, and asymptotic results are obtained to justify the observed loss of repulsiveness in the superposition of independent DPPs. Finally, in Section **??**, Monte Carlo methods are built on DPPs that provide a stochastic version of Gaussian quadrature.

2. Determinantal point processes

DPPs can be defined on any abstract locally compact Polish space E, see [**?**,**?**]. For statistical applications the two important cases, detailed in this section, are E being a (finite) discrete space, say $E = \{1, \dots, N\}$, and $E = \mathbb{R}^d$.

2.1. DPPs on a finite discrete space

In this section we consider a finite discrete state space $E = \{1, \dots, N\}$. In this case, a point process on E is simply a probability measure on the set 2^E of all subsets of E. As demonstrated in the sequel, DPPs on E form a flexible family of processes, that generate subsets of E exhibiting diversity. Most properties of DPPs in this case, including those detailed in the following, can be found in [**?**, **?**, **?**].

We say that X is a DPP on E if there exists a $N \times N$ matrix K such that for any $A \subset E$

$$P(X \supset A) = \det K_A, \tag{1}$$

where K_A is the sub-matrix with entries $(K_{ij})_{i,j \in A}$, with the convention $K_\emptyset = 1$. If K is a Hermitian matrix, then the existence of X is ensured if and only if all eigenvalues of K are in $[0, 1]$. In the remainder of this section, we restrict ourselves to real symmetric matrices.

Equation (**??**) already leads to insightful interpretations. First, the diagonal of K encodes the marginal probability for each element of E to belong to X. Indeed, from (**??**), for any $i \in E$, $P(i \in X) = K_{ii}$. Second for $i \neq j$, (**??**) implies

$$P(i, j \in X) = K_{ii}K_{jj} - K_{ij}^2.$$

Thus, if K_{ij} measures similarity between i and j, then X favours diversity. In particular, the correlation between the events $\{i \in X\}$ and $\{j \in X\}$ is always negative, and all the more negative that i and j are similar. In this sense, X generates sets having diverse elements.

If all eigenvalues of K are strictly less than one, or equivalently if $(I - K)$ is invertible, then we can deduce from the inclusion-exclusion principle that for any $A \subset E$

$$P(X = A) = \frac{\det L_A}{\det(L + I)}, \tag{2}$$

where $L = K(I - K)^{-1}$. As a side note, Equation (**??**) provides an alternative way to define a DPP: if X satisfies (**??**) where L is a positive semidefinite matrix, then it is called an L-ensemble. Note that all L-ensembles are DPPs in the sense of (**??**) where $K = L(L + I)^{-1}$, but the converse is not true (if $(I - K)$ is not invertible). L-ensembles are a popular point of view in the machine learning community, see [**?**]. It allows to define easily a DPP through a positive

semidefinite matrix L, without the constraint that all eigenvalues must be less than or equal to 1, unlike K. The main drawback though is that L does not encode a clear interpretation of the process. Another limitation is the fact that the probability of the empty set is necessarily positive for L-ensembles, which is unrealistic for some applications, see e.g. survey sampling in Section **??**.

The use of DPPs on finite sets in machine learning or for survey sampling is motivated by the flexibility of this family of processes, through the choice of K (or L), and also by its many appealing properties, some of which are detailed below.

i) [**unicity**] Given a symmetric matrix K with all eigenvalues in $[0,1]$, X in (**??**) is unique [**?**]. However the converse is not true: two symmetric matrices K and $\tilde{K}$ have the same principal minors if and only if $K = D\tilde{K}D^{-1}$ where D is a diagonal matrix with entries ± 1, see [**?**, **?**]. Therefore in this case, K and $\tilde{K}$ generate the same DPP.

ii) [**cardinality**] In general, the number of elements in X, denoted by $|X|$, is random. Specifically, if we denote by $\lambda_1, \ldots, \lambda_N$ the eigenvalues of K, the law of $|X|$ corresponds to the sum of N independent Bernoulli random variables with respective mean λ_i. In particular

$$\mathbb{E}(|X|) = tr(K), \qquad \mathbb{V}(|X|) = \sum_{i=1}^{N} \lambda_i(1-\lambda_i) = tr(K - K^2),$$

where tr means trace. An important particular case occurs when K is an orthogonal projection matrix. Then X is called a determinantal *projection* process and $|X| = rank(K)$ is deterministic. This property is of particular relevance in survey sampling where fixed-size sampling designs play a major role, see Section **??**.

iii) [**stability by restriction**] The restriction of X to a subset F of E remains a DPP with kernel matrix K_F.

iv) [**stability by complement**] The complement of X in E, i.e. $\bar{X} = E \setminus X$, is also a DPP and its kernel matrix is $I - K$.

v) [**stability by conditioning**] If we condition X to contain all elements of a subset F of E and/or to exclude all elements of a subset G of E (where $F \cap G = \emptyset$), then the resulting process on $E \setminus (F \cup G)$ remains a DPP with explicit kernel, see [**?**, **?**].

vi) [**simulation**] The last property makes possible exact simulation of DPPs. The procedure, detailed below, basically amounts to generating a first element for X, then a second element given the first one, then a third element given the first two, and so on.

A generic algorithm to sample from a DPP is given in [**?**]. We start from an orthonormal eigendecomposition of K, namely $K = \sum_{i=1}^{N} \lambda_i v_i v_i^*$, where v_i^* denotes the conjugate transpose of v_i, that for generality we assume to possibly be complex ($v_i \in \mathbb{C}^N$). The procedure presented in [**?**] exploits the fact that a DPP is in fact a mixture of determinantal projection processes. Specifically, the DPP X with kernel K has the same distribution as the DPP with kernel $\sum_{i=1}^{N} B_i v_i v_i^*$, where the B_i's are Bernoulli random variables with mean λ_i [**?**] (whence property ii) above). The algorithm thus consists in first selecting a determinantal projection process composing this mixture, and second sequentially generating a realization of this process using the conditional properties of DPPs. The algorithm is summed up as Algorithm **??**.

The projection in the second step of Algorithm **??** is the composition of successive projections, each on the orthogonal complement of $V(k)$, for all element k selected in the previous steps. In

Input: $K = \sum_{i=1}^{N} \lambda_i v_i v_i^*$.
for $i = 1, \dots, N$, **do**
| select v_i with probability λ_i.
end
Now

- Denote by V the matrix whose columns are the selected vectors, by n their number, and by $V(k)$ the vector corresponding to the k-th row of V.
- Set $H = \{0\}$ and $X = \emptyset$. Let $H^\perp$ be the orthogonal complement of H in $\mathbb{C}^N$.

while $dim(H^\perp) > 0$ **do**

- Sample k in $E \setminus X$ according to the discrete distribution $(\|P_{H^\perp} V(i)\|^2 / dim(H^\perp))_{i \in E \setminus X}$, where $P_{H^\perp}$ denotes the orthogonal projection matrix onto $H^\perp$,
- Let $X \leftarrow X \cup \{k\}$, $H \leftarrow \text{span}(H, V(k))$ (the vector space of all linear combinations of $V(k)$ and the elements of H).

end

Algorithm 1: Sampling from a finite DPP

other words, at an intermediate step of the algorithm we have $P_{H^\perp} = \prod_{k \in X} \left(I - \frac{V(k)V(k)^*}{V(k)^* V(k)} \right)$. So $P_{H^\perp}$ can simply be updated at each step by multiplication. However, this leads to some numerical instabilities since after some multiplications $P_{H^\perp}$ can differ from a proper projection. A more stable alternative is to use a Gram-Schmidt procedure to sequentially construct an orthonormal basis of H, from which $P_{H^\perp}$ is easily deduced. The details are given for the continuous case in Section **??** and its adaptation to the discrete case is straightforward.

Further properties of DPPs on discrete sets are described in Section **??**.

2.2. DPPs on a continuous space

We assume in this section that $E = \mathbb{R}^d$. This is in fact the historical setting where DPPs have been developed, see the seminal paper [**?**]. The initial motivation was to characterize the probability distribution of the locations of particles in physics, specifically for particles in repulsion (so-called *fermions*). We will come back to this interpretation later. For a detailed presentation of DPPs in a continuous setting, we refer to [**?**, **?**, **?**].

A general background on point processes on $\mathbb{R}^d$ can be found in [**?**, **?**]. In brief, a point configuration is a subset of $\mathbb{R}^d$ that has a finite number of elements on any bounded subset of $\mathbb{R}^d$. A point process in $\mathbb{R}^d$ is a probability law on the set of all point configurations in $\mathbb{R}^d$. It is said simple if there is at most one point at each location, almost surely.

Let X be a simple point process on $\mathbb{R}^d$. For a bounded set $D \subset \mathbb{R}^d$, denote by $X(D)$ the number of points of X in D. Let μ be a reference measure on $\mathbb{R}^d$ that we assume for simplicity to be absolutely continuous with respect to the Lebesgue measure. If there exists a function $\rho^{(k)} : \mathbb{R}^{dk} \to \mathbb{R}^+$, for $k \geq 1$, such that for any family of mutually disjoint subsets $D_1, \dots, D_k$ in $\mathbb{R}^d$

$$\mathbb{E} \prod_{i=1}^{k} X(D_i) = \int_{D_1} \dots \int_{D_k} \rho^{(k)}(x_1, \dots, x_k) \, \mu(\mathrm{d}x_1) \dots \mu(\mathrm{d}x_k), \tag{3}$$

then this function is called the joint intensity of order k of X, with respect to μ. From its definition, the joint intensity of order k is unique up to a μ-nullset. Henceforth, for ease of presentation, we ignore nullsets. In particular we will say that a function is continuous whenever there exists a continuous version of it. The joint intensity $\rho^{(k)}(x_1, \dots, x_k)$ can be understood as the probability that X contains a point in each neighborhood of x_i for $i = 1, \dots, k$. In this sense joint intensities are continuous analogues of the probabilities $P(X \supset A)$ involved in the discrete case in (**??**).

We say that X is a DPP on $\mathbb{R}^d$ if there exists a function $C : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$, called the kernel of X, such that for all $k \geq 1$

$$\rho^{(k)}(x_1, \dots x_k) = \det[C](x_1, \dots, x_k) \tag{4}$$

for every $(x_1, \dots, x_k) \in \mathbb{R}^{dk}$, where $[C](x_1, \dots, x_k)$ denotes the matrix with entries $C(x_i, x_j)$, $1 \leq i, j \leq k$. In view of the interpretation of $\rho^{(k)}$, this definition is the continuous counterpart of (**??**).

Let us briefly explain why the particular form in (**??**) arises naturally for the distribution of fermions in quantum mechanics. A similar description can be found in [**?**], see also [**?**, Section 5.4] for the physical aspects. Let us consider for simplicity the case of a bounded set where the number of points (hereafter particles) n is fixed, in which case $\rho^{(n)}$ reduces to the joint distribution of the location of these points, up to a constant. In quantum mechanics, the quantum state of a system of n particles is described by a complex-valued n-body wave function, say $\psi(x_1, \dots, x_n)$ where $x_1, \dots, x_n$ stand for the position of each particle. The probability distribution of the position of particles then admits the density $|\psi|^2$, which corresponds in our notation to $\rho^{(n)}$ (up to a constant). As a naive proposition, we might assume that this n-body wave function is simply the tensorial product of all individual wave functions $\prod \psi_i(x_i)$. But since the particles are indistinguishable, this could equally well be $\prod \psi_i(x_{\pi(i)})$, where π is any permutation of $\{1, \dots, n\}$. Moreover, fermions repel and for this reason, in virtue of the so-called Pauli exclusion principle, the n-body wave function must satisfy the anti-symmetric property $\psi(\dots, x_i, \dots, x_j, \dots) = -\psi(\dots, x_j, \dots, x_i, \dots)$ for any i, j. These requirements (indistinguishability and repulsiveness) led physicists to construct the wave function as an anti-symmetric version of the tensorial product, namely $\psi(x_1, \dots, x_n) \propto \sum_\pi \operatorname{sgn}(\pi)\psi_1(x_{\pi(1)}) \dots \psi_n(x_{\pi(n)})$, which is the determinant of the matrix Ψ with entries $\psi_i(x_j)$. Finally we get $|\psi|^2$ proportional to $\det(\Psi)\det(\overline{\Psi}) = \det(\Psi')\det(\overline{\Psi}) = \det(\Psi'\overline{\Psi}) = \det C[x_1, \dots, x_n]$ where $C(x, y) = \sum \psi_i(x)\overline{\psi}_i(y)$.

Beyond quantum mechanics, DPPs surprisingly arise in many examples of probability, the most famous of them being the law of the eigenvalues of fundamental random matrix models [**?**]. For this reason, they have received a lot of attention from a theoretical point of view. From a statistical perspective, their increasing popularity to encode negative dependencies, as demonstrated in the following sections, is due as in the discrete case to their flexibility through the choice of the kernel C, and to their nice properties, mainly the same as in the discrete case, as discussed now.

A particular case of DPP is the Poisson point process with intensity $\rho(x)$, which is associated to $C(x, x) = \rho(x)$ and $C(x, y) = 0$ if $x \neq y$. This is in some sense the extreme case of a DPP without interaction, whereas a DPP implies in general repulsiveness. Assume that C is Hermitian, i.e. $C(x, y) = \overline{C}(y, x)$. It is not very restrictive for statistical applications to further assume that C is continuous, which only excludes from useful models the peculiar example of

the Poisson point process. On any compact set S, C then admits a spectral decomposition

$$C(x,y) = \sum_{i\geq 0} \lambda_i^S \phi_i^S(x)\overline{\phi_i^S}(y) \tag{5}$$

where (ϕ_i^S) forms an orthonormal basis of $L^2(S)$ with respect to μ. In this setting the existence of X satisfying (**??**) is equivalent to $0 \leq \lambda_i^S \leq 1$ for any $i \geq 0$ and any S [**?**, **?**]. In the particular case where μ is the Lebesgue measure and C is invariant by translation, i.e. there exists C_0 such that $C(x,y) = C_0(x-y)$, leading to a stationary DPP, this condition for existence takes a simpler form. Namely, it boils down to $C_0 \in L^2(\mathbb{R}^d)$ and $0 \leq \varphi \leq 1$, where φ denotes the Fourier transform of C_0, see also Theorem **??** in Section **??**. For a general kernel C, existence is ensured if C is a continuous covariance function and there exists C_0 as before such that $C_0(x-y) - C(x,y)$ remains a covariance function (see e.g. Corollary 3.2.6 in [**?**]).

As in the discrete case, DPPs on $\mathbb{R}^d$ enjoy a lot of appealing properties, see [**?**, **?**]. Given C and a bounded set S, a DPP with kernel C is unique. The distribution of $X(S)$ is known (this is the sum of Bernoulli variables with mean λ_i^S). All joint intensities of X are known by the very definition. The density of X on S with respect to the standard Poisson point process on S is explicitly known. The restriction of a DPP to a subset S remains a DPP and its kernel is simply the restriction of C to $S \times S$. If we condition X to contain a finite set of points, then the resulting process is a DPP with an explicit kernel.

By the last property, we can simulate X on any bounded set S, using the same algorithm by [**?**] as in the discrete case. The first step consists in sampling the indexes involved in the sum of the spectral decomposition (**??**) according to Bernoulli variables with mean λ_i. Up to a re-ordering, let us denote by $\{1,\dots,n\}$ the selected indexes and $V(x) = (\phi_1^S(x),\dots,\phi_n^S(x))'$. The second step is akin to the Gram-Schmidt procedure, and is presented in Algorithm **??**.

Initialize by

- sampling x_n according to the distribution with density $(||V(x)||^2/n)$,
- setting $e_1 = V(x_n)/||V(x_n)||$.

for $i = (n-1)$ *to* 1 **do**

- sample x_i in S according to the distribution with density

$$\frac{1}{i}\left(||V(x)||^2 - \sum_{j=1}^{n-i} |e_j^* V(x)|^2\right),$$

- set $w_i = V(x_i) - \sum_{j=1}^{n-i} e_j^* V(x_i) e_j$ and $e_{n-i+1} = w_i/||w_i||$.

end

return $X = \{x_1, \cdots, x_n\}$.

Algorithm 2: Sampling from a continuous DPP: 2nd step. See main text for details about the complete procedure.

The densities involved above have to be understood with respect to μ and simulations from them can be done by rejection sampling.

The algorithm of simulation and some of the properties of DPPs discussed above require to know the spectral decomposition (**??**). Unlike the finite case where the kernel is simply a matrix, this decomposition is in general unknown in the continuous case. To overcome this issue in practice, we can either define the kernel C directly through its spectral form (see Section **??**), or apply some approximation as discussed in Section **??**. Note finally that it is possible to do statistical inference on C without this spectral knowledge (see Sections **??** and **??**).

3. DPPs in spatial statistics

In spatial statistics, a common concern is the analysis of spatial point patterns observed on a bounded subset of $\mathbb{R}^d$, the most common situation being $d = 2$. In presence of this kind of data, a first step usually consists in the estimation of the (first order) intensity ρ, to decide if it is constant (or equivalently homogeneous), in which case the points are uniformly distributed in the observation window, or if the intensity is inhomogeneous, in which case the points are more dense in certain regions of the observation window than elsewhere. In a second step, we may be interested in the interaction between the points. Three main situations may occur: independence (the case of the Poisson point process), aggregation (equivalently clustering), or inhibition. Due to their repulsiveness property, DPPs are well adapted models for inhibition. We explain below how parametric DPP models can be constructed, in which extent they constitute a flexible family of models, and how inference can be conducted.

In the spatial point process community, see for instance [**?**], the two first order moments of a point process are generally summarized by the intensity ρ and by the pair correlation function (pcf)

$$g(x,y) = \frac{\rho^{(2)}(x,y)}{\rho(x)\rho(y)}, \quad x,y \in \mathbb{R}^d. \tag{6}$$

If $g(x,y) = 1$, there is no (second order) interaction between two points located in a vicinity of x and y. If $g(x,y) > 1$, there is attraction, meaning that it is more likely than in the independent case to find a point nearby y if a point is already located nearby x. If $g(x,y) < 1$, there is inhibition. Following Section **??**, a DPP with a Hermitian kernel C satisfies

$$\rho(x) = C(x,x) \quad \text{and} \quad g(x,y) = 1 - \frac{|C(x,y)|^2}{C(x,x)C(y,y)}.$$

The expression of g confirms the inhibitive property of a DPP. These formulas also give a clear interpretation of the kernel C, that helps for the specification of C in a modeling purpose. Some conditions for existence are nonetheless necessary.

Let us first assume that the point process is stationary, which implies that ρ is constant and $g(x,y)$ only depends on the difference $y-x$. For a DPP with a real valued kernel C, stationarity is equivalent to that $C(x,y)$ only depends on $y-x$. If C is complex valued, the latter condition also implies stationarity of the associated DPP, but this condition is only sufficient, an example being the Ginibre DPP studied in Section **??**, which is stationary while its kernel is not invariant by translation. The following result, proved in [**?**], gives simple sufficient conditions for existence of the DPP when $C(x,y)$ only depends on $y-x$.

Theorem 1. *Assume $C(x,y) = C_0(y-x)$ where C_0 is a continuous Hermitian function in $L^2(\mathbb{R}^d)$. Then the existence of a DPP with kernel C is equivalent to that the Fourier transform of C_0 is less than 1.*

The construction of parametric families of DPPs thus becomes straightforward. One only needs to choose a parametric family of covariance functions C_0 and to compute its Fourier transform in order to check the condition for existence. This condition then becomes a constraint on the parameters of the family. Note that the advent of a constraint is expected for repulsive models, as the range of repulsiveness has to be somehow balanced by the mean number of points. As an extreme example, there is an upper bound on the number of hard balls with a given radius that one can include in a bounded region. Many examples of parametric covariance functions having an explicit Fourier transform are already known: Gaussian, Whittle-Matèrn, generalized Cauchy, Bessel-type, ... We refer to [**?**, **?**] for the expression of these covariance functions and of their Fourier transform, and the associated constraint on the parameters for existence.

Remark. An alternative parametric family is the β-Ginibre DPP, studied in Section **??**, where the kernel takes complex values. Note that this family has exactly the same second order properties (same ρ, same g) as DPPs with a Gaussian kernel. However, due to its central role in random matrix theory, some properties are known for the β-Ginibre DPP that are not available for the Gaussian DPP family, as for instance the expression of the J-function (see Section **??**). This opens new possibilities for inference, see Section **??**.

As we are interested in parametric DPPs to model point patterns exhibiting inhibition, it is natural to wonder whether DPPs form a flexible class of models. The least repulsive DPP is the Poisson point process, a peculiar case of DPP without interaction, see Section **??**. The most repulsive stationary DPP with intensity ρ has been determined in [**?**], where the repulsiveness is quantified through the behavior of the pair correlation. It corresponds to the DPP whose kernel has a Fourier transform equal to 1 on the ball centered at the origin with volume ρ. In dimension 1, this corresponds to the *sinc* kernel, $C_0(x) = \sin(\pi\rho|x|)/(\pi|x|)$, while in dimension 2, it is sometimes called the *jinc* kernel given by

$$C_0(x) = \sqrt{\rho}\frac{J_1(2\sqrt{\pi\rho}||x||)}{\sqrt{\pi}||x||}, \quad x \in \mathbb{R}^2, \tag{7}$$

where J_1 denotes the Bessel function of the first kind of order 1. The expression in any dimension can be found in [**?**]. Figure **??** shows the pcf of DPPs with kernels from the Whittle-Matèrn, the generalized Cauchy and the Bessel-type families. This plot illustrates that the most repulsive DPP within the first two families corresponds to the Gaussian kernel, whose pcf is represented in bold dashed line, and a realisation of which is shown on the top left plot of Figure **??**. Even if the point pattern clearly exhibits some regularity (or inhibition), it is less regular than a realisation from the *most repulsive DPP* shown in the bottom left plot of Figure **??**. The pcf of the latter DPP is represented in bold solid line. It is actually part of the Bessel-type family, which appears as a more flexible parametric family than the Whittle-Matèrn and the generalized Cauchy families.

The previous analysis demonstrates that DPPs cover a large range of repulsiveness, from the Poisson point process to the most repulsive DPP with kernel (**??**), opening promising possibilities for modeling. However, it also reveals that DPPs cannot be extremely repulsive. In particular, a DPP cannot include a hardcore distance δ (this a consequence of [**?**], Corollary 1.4.13), a situation where no pairs of points can be at a distance less than δ.

Concerning inhomogeneous models, a common practice in spatial statistics is to assume the point process to be inhomogeneous at the first order only, meaning that the intensity ρ is not

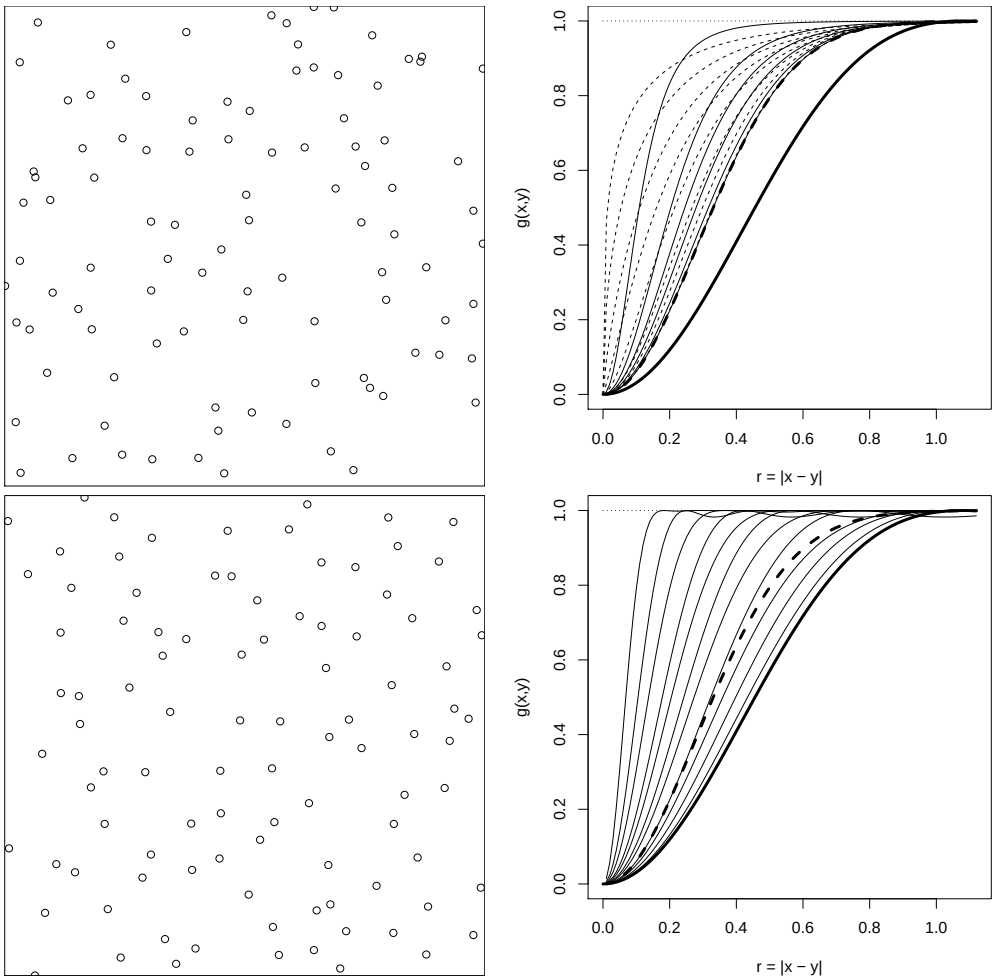


FIGURE 1. Left: realisations on $[0, 100]^2$ of the DPP with intensity $\rho = 1$ having a Gaussian kernel with the maximal possible scale parameter (top) and of the most repulsive DPP with intensity $\rho = 1$ given by (**??**) (bottom). Right: pcf of the two previous DPPs in bold dashed line and in bold solid line, respectively, along with the pcf of DPPs with intensity $\rho = 1$ having a Whittle-Matèrn kernel (dashed lines on the top right plot), a generalized Cauchy kernel (solid black lines on the top right plot) or a Bessel-type kernel (solid black lines on the bottom right plot) for different values of their shape and scale parameters.

constant while the pcf g is invariant by translation. For DPPs, this can be done by taking

$$C(x, y) = \sqrt{\rho(x)} C_0(x - y) \sqrt{\rho(y)} \tag{8}$$

with $C_0(0) = 1$. Then the intensity is $\rho(x)$ and the pcf is $1 - |C_0(x-y)|^2$. The existence of the associated DPP is ensured whenever $\rho(.)$ is bounded by $\tilde{\rho} > 0$ and the DPP with kernel $\tilde{\rho}\, C_0(x-y)$ satisfies the condition of Theorem **??**. Parametric models can thus be considered by assuming a parametric form for $\rho(.)$, possibly depending on covariates, and to choose a parametric covariance model for C_0 as discussed earlier.

Assuming a parametric form like (**??**) where $\rho(x) = \rho_\beta(x)$ and $C_0(x) = C_{0,\theta}(x)$ depend on two different parameters β and θ, we now discuss how to conduct statistical inference from an observation of X on a bounded set W. The standard procedure is to first estimate β by Poisson likelihood, that is

$$\hat{\beta} = \mathrm{argmax}_\beta \sum_{x \in X \cap W} \log \rho_\beta(x) - \int_W \rho_\beta(x) \mathrm{d}x.$$

In the homogeneous case of a constant intensity ρ (in which case $\beta = \rho$), this gives $\hat{\rho} = X(W)/|W|$. Second, the estimation of θ can be carried out by contrast estimating functions or estimating equations, typically based on second order characteristics of X like the pcf g. Note that because of (**??**), g depends on θ only and not on β. For instance, given a non-parametric estimate $\hat{g}$ of g (see [**?**] for an expression), we can estimate θ by

$$\hat{\theta} = \mathrm{argmax}_\theta \int_{r_{\min}}^{r_{\max}} (\hat{g}(t) - g_\theta(t))^2 \mathrm{d}t,$$

where $r_{\min}$ and $r_{\max}$ are user-specified parameters. A standard choice in practice is to set $r_{\min} = 0.01$ and $r_{\max}$ as one quarter of the side length of the observation window. The theoretical justifications of this two step procedure to estimate β and θ can be found in [**?**], in the particular case of stationary DPPs. Other contrast estimating functions are possible, where g is replaced by another characteristic of the point process, see Section **??** for an example with the J function.

As an alternative to contrast estimation, likelihood inference can be conducted in some cases. This method is expected to be more efficient than the other methods of inference, at least asymptotically (that is when W is big enough). Although no theoretical justifications for DPPs support this claim so far (neither any justification of the consistency of likelihood estimation), this has been confirmed by some simulations, see [**?**, **?**]. However, to get the expression of the likelihood from (**??**), one needs to know the spectral representation of C on W, which is rarely available. If W is a rectangular set and $\rho(x)$ is constant in (**??**), an approximation based on the Fourier transform of C_0 is proposed in [**?**]. If W is the unit square, this gives

$$C(x,y) \approx \rho \sum_{k \in \mathbb{Z}^d} \varphi(k) e^{2\pi \mathrm{i} k.(x-y)},$$

for any $x, y \in W$, where φ denotes the Fourier transform of C_0 and $k.(x-y)$ is the inner product between k and $(x-y)$. The case of a general rectangular set W can easily been deduced, see [**?**]. This approximation turns out to be quite accurate when ρ is not too small, see [**?**] for some justifications and a simulation study.

Several functions are available in the `spatstat` library [**?**] of R [**?**] to manipulate DPP models on a bounded subset of $\mathbb{R}^2$. They allow to define parametric models of homogeneous and inhomogeneous DPPs, to simulate them (using the spectral approximation presented above and the Gram-Schmidt algorithm described in Section **??**), and to fit them to a real dataset.

4. Determinantal survey sampling

Survey sampling considers the following problem. One wants to acquire knowledge of a parameter of interest θ, function of $\{y_k, k \in E\}$, where $E = \{1, \cdots, N\}$ is a finite population. A typical parameter θ is the sum t_y(or the mean m_y) of y. Unfortunately, the values y_i are inaccessible, except for a small part of the population. One therefore uses an estimator of θ constructed on a randomly chosen subpopulation. This choice is called the sampling design, whose properties are of crucial importance to get "good" estimators. In this section, we show that sampling designs based on DPPs, coined Determinantal Sampling Designs (DSDs) afterwards, enjoy must of the properties usually required for a sampling design. A good design should in particular meet the following requirements :

- simplicity of the design and simple algorithmic construction,
- control of the size of the sample,
- statistical amenability (consistency, central limit theorem,...),
- low mean square error of the estimator,

see for instance [**?**].

4.1. **Determinantal processes as sampling designs**

A sampling design can always be defined as the law of a vector of Bernoulli variables. Let $E = \{1, \cdots, N\}$ denote the population. Then a sampling design is the law of $B = (B_i, 1 \leq i \leq N)$, vector of Bernoulli variables, and the set $X = \{i \in E | B_i = 1\}$ is called the sample. This approach by Bernoulli variables is for instance used in [**?**]. Usually, except for very simple sampling designs, only the marginal law of each variable B_i is known, and other quantities remain unknown, or have to be estimated. This is not the case for determinantal sampling designs, whose description is given below.

Consider $X \sim DPP(K)$ a DPP with kernel matrix K on E, as defined in Section **??**, and let $B = (1_{i \in X}, 1 \leq i \leq N)$. Then $B_i = 1_{i \in X}$ is a Bernoulli variable, and (the law of) B is a sampling design. We call this design a determinantal sampling design with kernel K. We also note $X \sim DSD(K)$ to insist on the sampling interpretation. It then follows from the definition (**??**) of DPPs that

$$P(\Pi_{i \in A} B_i = 1) = \det K_A.$$

Therefore, the whole law of the design is known through the principal minors of its kernel.

As DPPs are indexed by symmetric matrices with eigenvalues in $[0, 1]$, DSDs form a parametric family of sampling designs. The existence of a perfect simulation algorithm as described in Section **??** allows to implement these designs in practice.

We interpret below some other properties of DPPs described in Section **??** in terms of sampling theory. Let $X \sim DSD(K)$.

(1) The first order inclusion probabilities are the diagonal terms of K, $P(i \in X) = K_{ii}$.
(2) For all $i \neq j \in E$, $\Delta_{i,j} = P(i, j \in X) - P(i \in X)P(j \in X) = -K_{ij}^2 \leq 0$. This is known in sampling theory as the Sen-Yates-Grundy Condition [**?**].
(3) The size of the sample X is known through the eigenvalues of K. In particular, X is of fixed size n (a property usually required in practice) iff K is a projection matrix of rank n. Also, $DSD(K)$ samples at least one point iff 1 is an eigenvalue of K.
(4) The restriction X_F of X to a domain $F \subset E$ is also a DSD, $X_F \sim DSD(K_F)$.

(5) Recall that a sampling design is stratified if the population E can be decomposed as $E = \bigcup_{k\in K} E_k$ (the E_k are the distinct strata) such that

$$i \in E_k, j \in E_l, k \neq l \Rightarrow B_i, B_j \text{ are independent.}$$

It holds that a $DSD(K)$ is stratified if and only if K admits a block matrix decomposition (after reordering of the population by strata).

(6) Poisson Sampling is determinantal with K diagonal. But Simple Random Sampling (SRS) of size n, defined by $P((i_1, \cdot, i_m) = X) = 1/\binom{N}{n}$ if $m = n$ and 0 otherwise is not determinantal (unless, $n = 0, 1, N-1$ or N).

(7) However, it is shown in [**?**] that for some couples (n, N) there exists DSDs with the same first and second order inclusion probabilities as SRS.

We finally address in this section a typical requirement of statisticians regarding sampling designs: the construction of fixed size sampling designs with prescribed unequal first order inclusion probabilities. Let us explain briefly the reason behind this requirement. In survey samplings, one usually knows the values of a variable x (on the whole population) correlated with y. If you choose the π_i proportional to x_i, then one can check that the variance of the Horvitz-Thompson estimator is proportional to the variance of the number of points of the design. Thus a fixed size sampling design with $\pi_i \propto x_i$ will achieve perfect estimation of x, and hopefully produce a low variance estimator of y. The Schur-Horn Theorem [**?**] allows to derive the following result:

Theorem 2. *Let Π be a vector of first order inclusion probabilities such that $\sum_{i=1}^{N} P_i = n$. Then there exists $X \sim DSD(K)$ of fixed size n such that $P(i \in X) = K_{ii} = \Pi_i$.*

Moreover, an explicit solution can be constructed. Simpler solutions for equal probability sampling designs ($P(i \in X) = K_{ii} = \frac{n}{N}$ for all $i \in E$) are also constructed in [**?**].

4.2. **Statistical properties**

In this section we sum up the statistical properties of the Horvitz-Thompson estimator $\hat{t}_y$ of a total $t_y = \sum_{i\in E} y_i$ constructed upon those DSDs. The central limit theorem is based on [**?**], and the concentration inequality on [**?**]. Let $X \sim DSD(K)$. We pose

$$\text{(Horvitz-Thompson estimator) } \hat{t}_y = \sum_{i\in X} K_{ii}^{-1} y_i = \sum_{i\in E} K_{ii}^{-1} 1_{i\in X} y_i$$

and $\hat{m}_y = \frac{1}{N}\hat{t}_y$ ($m_y = \frac{1}{N} t_y$). Then:

(1) $\mathbb{E}(\hat{t}_y) = t_y$ (the estimator is unbiased).

(2) $\mathbb{V}(\hat{t}_y) = z^T(I_N - K) \star Kz$ where $z_i = y_i K_{ii}^{-1}$ and $\star$ is the Schur-Hadamard (entrywise) product.

(3) (Consistency) If $\frac{1}{N^2} \sum_{i=1}^{N} K_{ii}^{-1} y_i^2 \underset{N\to\infty}{\longrightarrow} 0$, then $\hat{m}_y - m_y$ towards 0 in mean square.

(4) (CLT) Under technical assumptions, $\frac{\hat{m}_y - m_y}{\sqrt{\mathbb{V}(\hat{m}_y)}} \overset{law}{\to} \mathcal{N}(0, 1)$.

(5) (Concentration (1)) Pose $\mu = tr(K)$ and $M = \max_{1\leq i\leq N} |y_i|$. Then

$$P(|\hat{m}_y - m_y)| > a) \leq 5\exp\left(-\frac{N^2a^2}{16^2\left(NaM + 2\mu M^2\right)}\right).$$

(6) (Concentration (2)) If $DSD(K)$ is of fixed size n, we can improve the previous inequality in

$$P(|\hat{m}_y - m_y)| > a) \leq 2\exp\left(-\frac{N^2a^2}{8nM^2}\right).$$

The last three results allow to derive asymptotic as well as finite distance confidence intervals.

We conclude the section by stating three interesting consequences of the knowledge of the exact expression of the variance. As the Horvitz-Thompson estimator is unbiased, the variance of the estimator is a good indicator of the precision of the estimator, and one would tend to choose (if possible) a sampling design with low variance to increase the precision of the estimation.

(1) The quantity $y^T y - \mathbb{V}(\hat{t}_y)$ can be interpreted as a ponderated measure of global repulsiveness for point processes on a discrete space (see [**?**] in the continuous setting). As DPPs are repulsive, we then expect DSDs to achieve small variance within all sampling designs. This is validated by our empirical studies, see also Section **??** for results in the continuous setting.
(2) Imagine y is known. Is there a "best" way to estimate its total by a determintal sampling design ? That is, can we minimize (at least theoretically) the variance of $\hat{t}_y$?
Assume the diagonal $\mathrm{diag}(K)$ of K is fixed. Then

$$\mathbb{V}(\hat{t}_y) = y^T y - (y/(\mathrm{diag}(K))^T K \star K(y/(\mathrm{diag}(K))$$

and minimizing the variance is equivalent to maximizing $z^T K \star Kz$ with z a fixed vector. This can be interpreted as a quadratic semidefinite optimization problem. Semidefinite optimization has attracted a lot of interest recently (mainly in the linear setting, see for instance [**?**], [**?**]). But even in the linear case, problems are challenging.
(3) Nevertheless, we can solve the problem of exact estimation ($\mathbb{V}(\hat{t}_y) = 0$). The following result can be found in [**?**]:

Theorem 3. *The total t_y is perfectly estimated by $\hat{t}_y$ ($\hat{t}_y = t_y$) iff $DSD(K)$ is a fixed size stratified determinantal sampling design with $K_{ii}^{-1}y_i$ constant on each stratum.*

Equivalently, t_y is perfectly estimated by $\hat{t}_y$ iff K is a block diagonal matrix, with each block a projection matrix with diagonal $K_{ii} = \alpha y_i^{-1}$ (the value α depends on the block).

5. The β-Ginibre point process and design of a cellular network

This section aims to validate the β-Ginibre point process as a model for the distribution of base station locations in a cellular network, from real data collected in Paris, France.

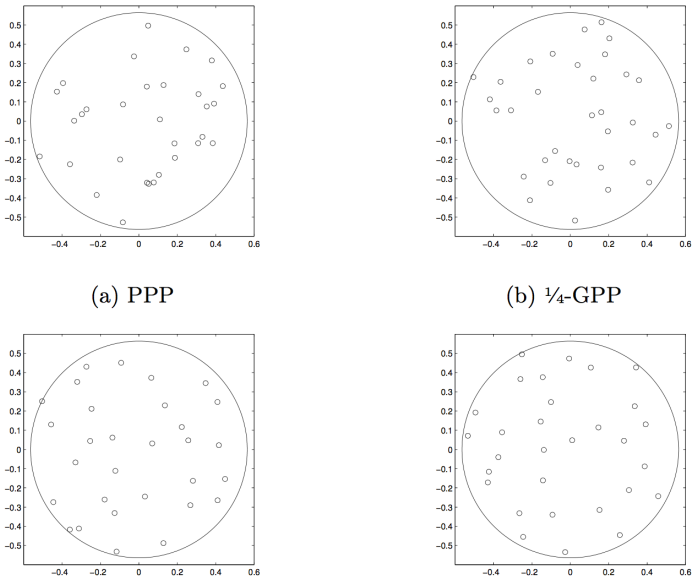
5.1. **Theoretical model: the β-Ginibre point process**

The Ginibre point process (GPP) with intensity $\rho = \frac{\gamma}{\pi}$ (with $\gamma > 0$) is a determinantal point process on $\mathbb{C}$ whose kernel C_γ is given for any $x, y \in \mathbb{C}$ by:

$$C_\gamma(x,y) = \frac{\gamma}{\pi} e^{-\frac{\gamma}{2}(|x|^2+|y|^2)} e^{\gamma x \overline{y}}.$$

If β is a real number between 0 and 1, the β-Ginibre point process (β-GPP) with intensity $\rho = \frac{\gamma}{\pi}$ is a determinantal point process on $\mathbb{C}$ whose kernel $C_{\gamma,\beta}$ is given for any $x, y \in \mathbb{C}$ by:

$$C_{\gamma,\beta}(x,y) = \frac{\gamma}{\pi} e^{-\frac{\gamma}{2\beta}(|x|^2+|y|^2)} e^{\frac{\gamma}{\beta} x \overline{y}}.$$



(a) PPP (b) ¼-GPP

(c) ¾-GPP (d) 1-GPP

FIGURE 2. Realizations of PPP and β-GPP for $\beta \in \{\frac{1}{4}; \frac{3}{4}; 1\}$.

A β-GPP may be built by combining two operations on a GPP: a thinning with parameter β (one keeps each point independently with probability β) then a rescaling with parameter $\sqrt{\beta}$, such that we keep the same intensity. Hence, the parameter β provides an information concerning the degree of repulsiveness of the point process: the smaller β is, the less repulsive the β-GPP is. Note that such a point process is not defined for $\beta > 1$. One can observe in Figure **??** some realizations of a Poisson point process (PPP) and β-GPPs for different values of β.

These point processes were investigated in the wireless communication field : they were at first introduced by Shirai et al. [**?**] in quantum physics to model fermion interactions. Works of Miyoshi et al. [**?**] and Deng et al. [**?**] have derived a computable representation for the coverage

probability in cellular networks - the probability that the signal-to-interference-plus-noise ration (SINR) for a mobile user achieves a target threshold.

5.2. **Statistical analysis**

In this subsection, we use real data from the mobile network in Paris to show that base station locations can be fitted with a β-GPP [**?**]. We introduce the fitting method that allows to obtain the parameter β and also present the results from the fitting of each deployment and operator. In order to fit the mobile network deployment on the β-Ginibre model, we use the J-function. The J-function of a stationary point process X on $\mathbb{R}^d$ is defined for any $r \in \mathbb{R}_+$ by:

$$J(r) = \frac{1 - G(r)}{1 - F(r)},$$

where F is the empty space function of X and G its nearest-neighbor distance distribution function, defined for some $u \in \mathbb{R}^d$ and any $r \in \mathbb{R}_+$ by:

$$F(r) = \mathbb{P}(\|u - X\| \leq r)$$

and

$$G(r) = \mathbb{P}(\|u - X \setminus \{u\}\| \leq r).$$

The J-function provides both a characterization of the point process and a direct information about its attractiveness or repulsiveness. More precisely, when $J < 1$, X is attractive, otherwise X is repulsive. The equality $J \equiv 1$ characterizes the PPP, where there is no interaction between the particles. For the case of the β-GPP, we get from [**?**] the following proposition.

Proposition 5.1. *The J-function of the β-GPP with intensity $\frac{\gamma}{\pi}$ is given for any $r \in \mathbb{R}_+$ by:*

$$J(r) = \frac{1}{1 - \beta + \beta e^{-\frac{\gamma}{\beta} r^2}}.$$

Note that for any β this J-function is bigger than one, which confirms that the β-GPP is a repulsive point process. When β tends to 0, this expression tends to 1, which corresponds to the J-function of a PPP.

This J-function allows to validate the β-GPP as a distribution model of the repartition of the base stations for each operator and each technology.

We use the `spatstat` package to obtain an estimate of the J-function from the raw data. Since we consider only a finite set of antennas, edge effects might appear on the J-function estimate. We then have to keep a subset of the data to perform the estimation. Figure **??** gives the window we considered for extracting data in Paris, France.

This window is chosen such that it covers about 60 % of the city and that its shape matches the geographical borders. The values of the J-function estimate are computed for $r \leq 600$ m. Above 600 m, the estimation is not relevant due to the edge-effect. J is then directly fitted on the estimate and the parameter β is deduced. An example of fitting is given in Fig. **??**. Visual inspection reveals a clearly repulsive behaviour of the base stations locations and a good fit to the theoretical model, especially when compared to the unit J-function of a PPP.

Numerical values of the parameter β and the intensity ρ from the fitting are given in Table 1 for each operator and each technology, and in Table 2 for each operator. Each intensity ρ is simply computed using the number of corresponding base stations in the window. The

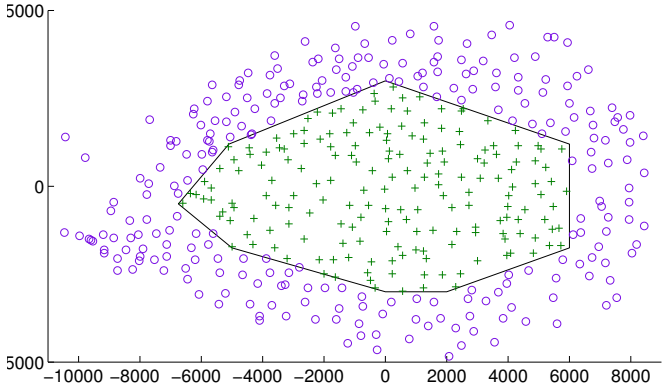


FIGURE 3. Example of data sample for one GSM operator. The J-function is fitted on the points within the polygonal window.

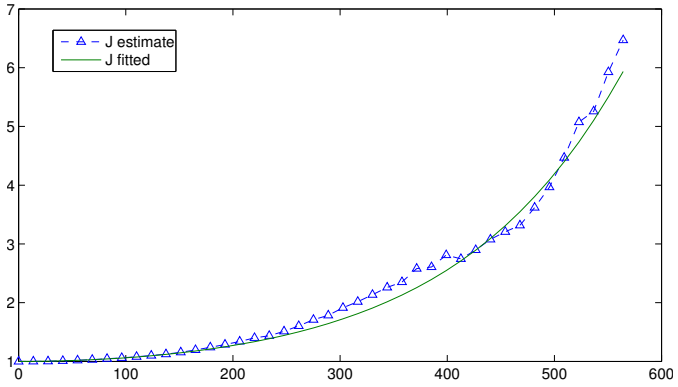


FIGURE 4. Example of J-function fitting for SFR on the 3G 900 MHz band. $\rho = 1.92$ base stations per kilometer square $\beta = 0.97$. Residuals: 0.74.

parameter β is then computed by the method of least squares applied to the J-function of the β-GPP and its estimation. Data analysis also shows that the superposition of all sites is tending to a PPP as β is equal to 0.17, which will be confirmed by the asymptotic results mentioned in the part **??**.

We hence show that base stations distribution for an operator and for a technology can be fitted with a β-GPP in the Paris area. The distribution of all base stations of all operators may therefore be considered as an independent superposition of β-GPP, but can actually be fitted with a PPP.

5.3. **Asymptotics**

In order to justify in a theoretical way the Poisson behavior of the independent superposition of the β-GPPs, we put in this subsection the corresponding asymptotic results.

TABLE 1. Numerical values of β per technology and operator.

	Orange	SFR	Bouygues	Free
GSM 900	0.81	0.76	0.65	NA
GSM 1800	0.84	0.85	0.71	NA
UMTS 900	NA	0.97	0.53	0.89
UMTS 2100	1.04	0.65	0.82	0.89
LTE 800	1.02	0.93	0.67	NA
LTE 1800	NA	NA	0.75	NA
LTE 2600	0.93	0.67	0.63	0.89

TABLE 2. Numerical values of β and ρ per operator and for the superposition of all the sites.

	Orange	SFR	Bouygues	Free	**Superposition**
β	0,94	0,70	0,81	0,89	**0,17**
ρ	3,48	3,70	4,23	1,05	**10,28**
Number of sites	185	197	225	56	**547**

Assume that the space E is endowed with its Borel space $\mathcal{B}(E)$ and denote by N_E the space of configurations on E. For a point process X on E, the application $c : E \times N_E \to \mathbb{R}_+$ is a Papangelou intensity [**?**] of X if, for any measurable function $u : E \times N_E \to \mathbb{R}_+$,

$$\mathbb{E}\Big[\sum_{x\in X} u(x, X \setminus x)\Big] = \int_E \mathbb{E}[c(x, X)u(x, X)]\mu(\mathrm{d}x).$$

Intuitively, if $x \in E$ and $\xi \in N_E$, $c(x, \xi)$ represents the conditional probability of finding a particle in the location x given the configuration ξ. It allows to propose a new definition for repulsiveness: a point process X on E with Papangelou intensity c is said to be repulsive (in the Papangelou sense) if, for any $\omega, \xi \in N_E$ such that $\omega \subset \xi$ and any $x \in E$,

$$c(x, \xi) \leq c(x, \omega),$$

and weakly repulsive (in the Papangelou sense) if, for any $\xi \in N_E$ and any $x \in E$,

$$c(x, \xi) \leq c(x, \varnothing).$$

The next proposition gives an explicit expression for the Papangelou intensity of a DPP [**?**]. If C is the kernel of a DPP, denote by T_C the functional operator defined for any $f \in L^2(E, \mu)$ and any $x \in E$ by

$$T_C f(x) = \int_E C(x, y)f(y)\mu(\mathrm{d}y).$$

Proposition 5.2. *Let X be a DPP on E with kernel C such that the operator $T_H = (Id - T_C)^{-1}T_C$ is well-defined. Then, its Papangelou intensity c is given for any $x_0, x_1, \ldots, x_k$ by:*

$$c(x_0, \{x_1, \cdots, x_k\}) = \frac{\det(H(x_i, x_j),\ 0 \leq i, j \leq k)}{\det(H(x_i, x_j),\ 1 \leq i, j \leq k)}.$$

Moreover, X is repulsive in the Papangelou sense.

We now introduce the topology on point processes which is used in this part [**?**]. The total variation distance $\mathrm{d_{TV}}$ is defined for any measures ν_1, ν_2 by:

$$\mathrm{d_{TV}}(\nu_1, \nu_2) := \sup_{\substack{A \in \mathcal{B}(E) \\ \nu_1(A), \nu_2(A) < \infty}} |\nu_1(A) - \nu_2(A)|,$$

we say that a map $h : N_E \to \mathbb{R}$ is 1-Lipschitz according to $\mathrm{d_{TV}}$ if for any $\omega_1, \omega_2 \in N_E$,

$$|h(\omega_1) - h(\omega_2)| \leq \mathrm{d_{TV}}(\omega_1, \omega_2),$$

and denote by $\mathrm{Lip}_1(\mathrm{d_{TV}})$ the set of all such maps which are measurable.

The Kantorovich-Rubinstein distance $\mathrm{d^*_{TV}}$ associated to $\mathrm{d_{TV}}$ between two point processes X_1 and X_2 is defined as:

$$\mathrm{d^*_{TV}}(X_1, X_2) = \sup \Big| \mathbb{E}[h(X_1)] - \mathbb{E}[h(X_2)] \Big|, \tag{9}$$

where the supremum is over all $h \in \mathrm{Lip}_1(\mathrm{d_{TV}})$ that are integrable with respect to the distributions of X_1 and X_2.

This distance provides a strong topology on the space of point processes: in particular, it is strictly stronger than convergence in law. A counterpart of this use is that we can only consider sequences of finite point processes. The following theorem [**?**] gives an upper bound for this distance taken between a finite PPP and an other point process.

Theorem 5.3. *Let Z be a Poisson point process on E with finite control measure $M(\mathrm{d}x) = m(x)\mathrm{d}x$ and X a second point process on E with Papangelou intensity c. Then,*

$$\mathrm{d^*_{TV}}(X, Z) \leq \int_E \mathbb{E}[|m(x) - c(x, X)|]\mathrm{d}x.$$

Using Laplace transforms, we are able to establish the convergence of a superposition of independent β-GPPs to a PPP and that a β-GPP tends to a PPP as β tends to 0, where these two results are given for convergence in law. The previous upper bound theorem allows to give more precise results [**?**] provided by the following propositions.

Proposition 5.4. *For any $n \in \mathbb{N}$, let X_n the superposition of n independent, finite and weakly repulsive point processes $X_{n,1}, \ldots, X_{n,n}$, with respective joint intensities $\rho^{(k)}_{n,1}, \ldots, \rho^{(k)}_{n,n}$ ($k \geq 1$) and let Z be a Poisson point process with control measure $M(\mathrm{d}x) = m(x)\mu(\mathrm{d}x)$. Then,*

$$\mathrm{d^*_{TV}}(X_n, Z) \leq R_n + 2n\Big(\max_{i \in \{1, \ldots, n\}} \int_E \rho^{(1)}_{n,i}(x)\mu(\mathrm{d}x) \Big)^2,$$

where

$$R_n := \int_E \Big| \sum_{i=1}^{n} \rho^{(1)}_{n,i}(x) - m(x) \Big| \mu(\mathrm{d}x).$$

Corollary 5.5. *Under the assumptions and notations of the Proposition* **??**, *and assuming moreover that there exists a real constant A such that for any $n \in \mathbb{N}$,*

$$\max_{i\in\{1,\dots,n\}} \int_E \rho_{n,i}^{(1)}(x)\mu(\mathrm{d}x) \leq \frac{A}{n},$$

one has for any $n \in \mathbb{N}$,

$$\mathrm{d}_{\mathrm{TV}}^*(X_n, Z) \leq R_n + \frac{2A^2}{n}.$$

Proposition 5.6. *Let C be the kernel of a stationary determinantal point process X on $\mathbb{R}^d$ with intensity $\rho \in \mathbb{R}$, Λ be a compact subset of $\mathbb{R}^d$, $(\beta_n)_{n\in\mathbb{N}} \subset (0,1)^{\mathbb{N}}$ and $Z_{\Lambda,\rho}$ designs the homogeneous Poisson point process with intensity ρ reduced to Λ. For any $n \in \mathbb{N}$, X_n is the point process on $\mathbb{R}^d$ obtained by combining a β_n-thinning with a β_n-rescaling on the point process X that one reduces to Λ. More precisely, X_n is the determinantal point process with kernel C_n defined by*

$$C_n : (x,y) \in E \times E \mapsto C\Big(\frac{x}{{\beta_n}^{\frac{1}{d}}}, \frac{x}{{\beta_n}^{\frac{1}{d}}}\Big)\mathbf{1}_{\Lambda\times\Lambda}(x,y).$$

Then,

$$\mathrm{d}_{\mathrm{TV}}^*(X_n, Z_{\Lambda,\rho}) \leq \frac{2\beta_n}{1-\beta_n}\rho|\Lambda|.$$

6. DPPs for Monte Carlo integration

DPPs have also been used in the design of Monte Carlo numerical integration methods [**?**]. In this section, we give an informal description of [**?**], insisting on the link with Gaussian quadrature.

For the purpose of this section, given a probability density π over $\mathbb{R}^d$, a numerical integration –or *quadrature*– method is defined as:

- a rule to build nodes $x_1, \dots x_N \in \mathbb{R}^d$,
- a rule to compute weights w_i for $1 \leq i \leq N$,

such that

$$\sum_{i=1}^{N} w_i f(x_i) \approx \int_{\mathbb{R}^d} f(x)\pi(x)dx \tag{10}$$

for a large class of functions f. *Monte Carlo* methods are defined as quadrature methods where the nodes are the realizations of a collection of N random variables. In traditional Monte Carlo approaches [**?**], nodes (x_i) are drawn independently, such as in importance sampling, or using a Markov chain, such as in Markov chain Monte Carlo (MCMC) algorithms. Laws of large numbers and central limit theorems then guarantee (**??**), leading to asymptotic confidence intervals for the integral of width $\mathcal{O}(N^{-1/2})$. This rate is often dubbed the *Monte Carlo rate*. The question naturally arises whether we could obtain a faster rate, i.e. smaller confidence intervals, by introducing more structure in the rule used to sample the nodes (x_i).

Intuitively, forcing the nodes to spread evenly across $\mathbb{R}^d$ should reduce the variance of the LHS of (**??**), as for each draw of the nodes (x_i) the whole support of π is evenly covered. In contrast, independent or MCMC draws will tend to leave more "holes" in the support of π,

and these holes will be different for each draw, resulting in a large variance of the LHS of (**??**). Simultaneously, the nodes should concentrate where π puts a lot of mass, as failing to capture rapid variations of f in the modes of π will cost a lot in quadrature error. In a nutshell, a good quadrature method should solve the tradeoff *sampling nodes in the modes of* π vs. *spreading nodes across* $\mathbb{R}^d$ *to avoid holes*. In this section, we explain how DPPs can satisfyingly solve this tradeoff. To understand how, we first make a detour by a two-century-old quadrature method.

6.1. Gaussian quadrature

For deterministic quadrature methods, Gauss [**?**] first realized that setting a regular grid over $\mathbb{R}$ was not the most economical way to integrate polynomials. More precisely, consider for a moment $d = 1$, and let $(p_k)_{k\geq 0}$ be the result of applying Gram-Schmidt orthonormalization in $L^2(\pi)$ to the sequence of monomials $q_k : x \mapsto x^k$. We call (p_k) the orthonormal polynomials with respect to π [**?**]. Note that in particular, the degree of p_k is k. Finally, let

$$K_N(x, y) = \sum_{k=0}^{N-1} p_k(x)p_k(y) \tag{11}$$

be the so-called Christoffel-Darboux kernel [**?**]. Then Gaussian quadrature is defined by setting $(x_i)_{i=1,\dots,N}$ to be the N zeros of p_N, and the weights w_i to be $1/K_N(x_i, x_i)$ for each $1 \leq i \leq N$. Gaussian quadrature has the property to be exact whenever f is a polynomial function of degree up to $2N - 1$, that is, (**??**) is an equality. Gaussian quadrature is thus expected to be accurate on functions that are limits of sequences of polynomials: Jackson's approximation theorem for algebraic polynomials, for instance, says that the error in (**??**) is $\mathcal{O}(1/N)$ for f continuously differentiable.

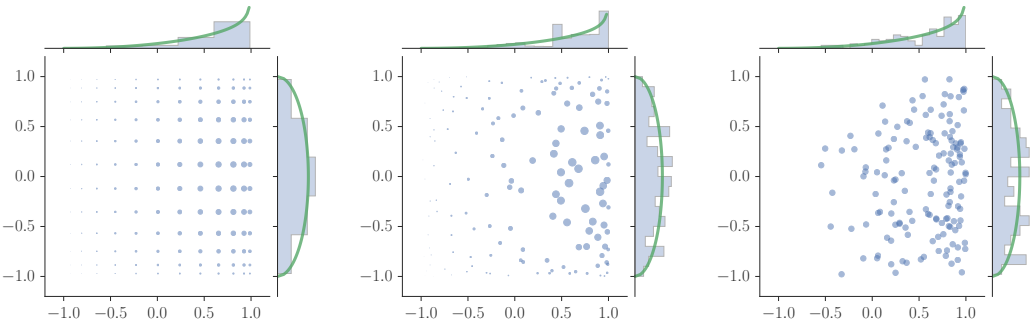
Although deterministic, Gaussian quadrature has the property we are looking after: nodes are well spread across the support of π, while more densely present in the modes of π. To understand why, denote by c_k the leading coefficient of p_k. Among all monic polynomials of degree k, p_k/c_k is the one with the smallest norm in $L^2(\pi)$, see e.g. [**?**, Section 3]. Consequently, the zeros of p_k will tend to be far from each other, to make p_k as flat and close to zero as possible. Simultaneously, there will be more zeros where π puts a lot of mass, to make p_k even closer to zero in the areas of $\mathbb{R}^d$ that contribute a lot to squared norms in $L^2(\pi)$. These intuitions are demonstrated on Figure **??**.

The main disadvantages of Gaussian quadrature are that

(1) there is no generic way to adapt Gaussian quadrature to $d \geq 2$. Outside very specific cases, one has to make the Cartesian product of sets of one-dimensional nodes, thus raising any bound on the quadrature error to the power $1/d$ [**?**].
(2) tight bounds on the error are scarce [**?**], specific to particular choices of π, and do not scale well with the dimension d.
(3) orthonormal polynomials are rarely known beforehand and computing them requires computing the moments of π. This limits the applicability of the method, with most applications focusing on Jacobi measures $\pi : x \mapsto (1 - x)^a(1 + b)^b$ for some $a, b > -1$.

6.2. Monte Carlo with DPPs

Looking back at Definitions (**??**) and (**??**), DPPs also have the potential to solve the tradeoff of "sampling in the modes of π but make points spread regularly". Choosing $\mu = \pi$ in (**??**), any



(a) Gaussian quadrature (b) DPP (c) i.i.d. sampling

FIGURE 5. Comparison of quadrature rules. In each plot, π is the same product Jacobi measure, depicted in green on the marginal plots. The area of each marker is proportional to its weight w_i in the corresponding estimator.

kernel K in (**??**) that makes the corresponding DPP exist, and denoting (x_i) a sample from the corresponding DPP, (**??**) implies that

$$\sum_i \frac{f(x_i)}{K(x_i, x_i)} \tag{12}$$

is an unbiased estimator of the RHS of (**??**). In particular, choosing $K = K_N$ the Christoffel-Darboux kernel (**??**) yields DPP samples of size N almost surely, and the estimator (**??**) takes the same form as Gaussian quadrature in Section **??**, except the nodes are a sample from a DPP and not the zeros of p_N. Actually, these two sets of nodes are asymptotically very similar [**?**], further confirming the intuition that Monte Carlo with a DPP using kernel (**??**) is a stochastic version of Gaussian quadrature. It turns out that stochasticity partly solves the issues of Gaussian quadrature:

(1) the above DPP easily generalizes to general d, as long as $\pi(x) = \pi^1(x_1) \dots \pi^d(x_d)$ is separable. For $1 \leq \ell \leq d$, take $(p_k^\ell)_k$ to be the orthonormal polynomials with respect to π^ℓ, fix $\mathfrak{b} : \mathbb{N} \to \mathbb{N}^d$ a bijection that orders d-uplets of integers, and consider the kernel on $\mathbb{R}^d \times \mathbb{R}^d$ defined by

$$K_N(x, y) = \sum_{k=1}^{N} p_{k_1}(x_1) \dots p_{k_d}(x_d) p_{k_1}(y_1) \dots p_{k_d}(y_d),$$

where $(k_1, \dots, k_d) = \mathfrak{b}(k)$. Then taking $\mu = \pi$ and $K = K_N$ in (**??**) and (**??**) leads to a DPP that exists and strictly generalizes the above one-dimensional construction. Figure **??** depicts a draw from this DPP. The weighted set of points in the DPP sample can be seen to leave fewer holes than i.i.d. sampling with the same marginals in Figure **??**.

(2) Most importantly, for any dimension d, one can prove a central limit theorem [**?**, Theorem 2.7] for (**??**) that leads to confidence intervals of size $\mathcal{O}(N^{-\frac{1}{2}-\frac{1}{2d}})$. The proof is technically involved and requires assumptions on π, f, and $\mathfrak{b}$. In particular, f should

be continuously differentiable. But the reward is an asymptotic quantification of the quadrature error at a higher level of generality than Gaussian quadrature. Furthermore, the limiting variance in the central limit theorem [**?**, Theorem 2.7] then has a strikingly simple form, and measures how fast the Fourier coefficients of $f\pi$ decrease. In other terms, non-smoothness of the integrand in (**??**) is paid in limiting variance.

(3) if the moments of π are not known, or π is not separable, one can still draw from an instrumental DPP that satisfies the assumptions, and then change the weights w_i accordingly. The *same* central limit theorem holds for this new estimator [**?**, Theorem 2.9].

Acknowledgments

We thank the organizers of the conference for a great meeting, scientifically and socially. We thank the publication committee for the opportunity to extend the session through this paper. RB acknowledges support from ANR grant BoB ANR-16-CE23-0003.