Asymptotic approximation of the likelihood of stationary determinantal point processes

Arnaud Poinas, Frédéric Lavancier

Asymptotic approximation of the likelihood of stationary determinantal point processes

Arnaud Poinas*,[1] and Frédéric Lavancier†,[2]

[1]Univ. Lille – CNRS – UMR 9189 – CRIStAL, 59651, Villeneuve d'Ascq, France

[2]Laboratoire de Mathématiques Jean Leray, Nantes, France.

March 3, 2021

Abstract

Continuous determinantal point processes (DPPs) are a class of repulsive point processes on $\mathbb{R}^d$ with many statistical applications. Although an explicit expression of their density is known, this expression is too complicated to be used directly for maximum likelihood estimation. In the stationary case, an approximation using Fourier series has been suggested, but it is limited to rectangular observation windows and no theoretical results support it. In this contribution, we investigate a different way to approximate the likelihood by looking at its asymptotic behaviour when the observation window grows towards $\mathbb{R}^d$. This new approximation is not limited to rectangular windows, is faster to compute than the previous one, does not require any tuning parameter, and some theoretical justifications are provided. The performances of the associated estimator are assessed in a simulation study on standard parametric models on $\mathbb{R}^d$ and compare favourably to common alternative estimation methods for continuous DPPs.

1 Introduction

Determinantal point processes (DPPs for short) are a type of repulsive point processes with statistical applications ranging from machine learning [17] to telecommunications [12, 22, 15], biology [1], forestry [19], signal processing [5] and computational statistics [6]. In this paper, we focus on likelihood estimation of parametric families of stationary DPPs on $\mathbb{R}^d$, but we will also include in our study stationary DPPs defined on $\mathbb{Z}^d$. From a theoretical point of view, we are specifically interested in an increasing domain setting, meaning that we assume

*arnaud.poinas@univ-lille.fr

†frederic.lavancier@univ-nantes.fr

to observe only one realisation of the DPP on $W \subset \mathbb{R}^d$, and our asymptotic results will concern the case where W grows towards $\mathbb{R}^d$, making the cardinality of the observed DPP tend to infinity. From this perspective, the likelihood is just the density of the DPP.

For a DPP on $\mathbb{R}^d$ with kernel K, the expression of its density on any compact set W (with respect to the unit rate Poisson point process) is known since the seminal paper of Macchi (1975) [21]. But this expression is hardly tractable. It requires the knowledge of another kernel, usually called L, that can only be obtained from K by solving an integral equation or by knowing the spectral representation of the integral operator associated to K on W. Both methods are not feasible in practice. Some approximations are then needed. A general numerical procedure to solve an integral equation is the Nyström method [2], but to our knowledge this approach has not been considered so far for parameter estimation of continuous DPPs. A likely reason is that this approximation can be very time consuming to run in the continuous case, while in view of likelihood parametric estimation, the density must be computed for different values of the parameter, and so the approximation must be repeated for each new proposed value. In the stationary case and when W is a rectangular window, an alternative approximation of the density has been proposed in [19] by considering a Fourier series approximation of K. This approximation has the pleasant feature to be explicit, but is restricted to rectangular windows and lacks theoretical justifications.

Our contribution is an (increasing domain) asymptotic approximation of the density as well as a way to correct the edge effects arising as a consequence of this approximation. This approach is not restricted to rectangular windows W, does not depend on any tuning parameter, and is faster to compute than the Fourier series approximation of [19].

The density of a DPP depends on the log-determinant of a random kernel matrix whose behaviour is difficult to control from a theoretical point of view, making challenging a theoretical study of our approximation. The situation simplifies slightly for stationary DPPs defined on a regular grid, typically $\mathbb{Z}^d$. We prove in this case that our approximation has the same asymptotic behaviour as the true density, under mild assumptions. The proof relies on an asymptotic control of the L kernel when W grows to $\mathbb{Z}^d$ and to concentration inequalities for DPPs established in [23]. For DPPs defined on $\mathbb{R}^d$, getting the same kind of results remains an open problem. However we prove that any DPP on $\mathbb{R}^d$ is arbitrarily close to a DPP defined on a small enough regular grid, the density approximation of which is consistent from the previous result.

Likelihood estimation of DPPs has been considered in other settings. For DPPs defined on a finite space, getting the expression of the density from K is not an issue (providing the space dimension is not too large), as it only requires the eigendecomposition of the kernel K, which reduces to a matrix in this case. Likelihood estimation in this setting, based on the observation of n i.i.d. discrete DPPs, has been studied in [10], who investigate asymptotic properties when n tends to infinity. In the continuous case, likelihood estimation based on n i.i.d. observations is considered in [7]. In this contribution, the DPP is directly defined through the kernel L, not K, avoiding the need to approximate its density from K. However, this comes at the cost of a loss of interpretability of the parameters, and more importantly, this approach does not allow to consider increasing domain asymptotic. Indeed, as detailed in

Section 2, it is extremely difficult to relate the kernel $L_{[W]}$ associated to the DPP defined on W, with the kernel $L_{[W']}$ for $W \subset W'$. For this reason, it is difficult to suggest a parametric family of kernels $L_{[W]}$ indexed by W. In contrast the kernel K of the DPP on any set W is just the restriction of K to W, and it suffices to define K on $\mathbb{R}^d$ in order to automatically get a consistent family of kernels on any subset W.

The remainder of the paper is organised as follows. We introduce our notations and basic definitions in Section 2. Our asymptotic approximation of the likelihood is presented in Section 3, along with some theoretical justifications. We show in Section 4 how this approximation applies to standard parametric families of DPPs in $\mathbb{R}^d$. Section 5 is devoted to a simulation study demonstrating the performances of our approach. Some concluding remarks are given in Section 6. Finally Section 7 includes the proof of our theoretical results, while some technical lemmas are gathered in the appendix.

2 Definitions and notation

We consider point processes on $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \nu)$ where $\mathcal{X}$ is either $\mathbb{R}^d$ or $\mathbb{Z}^d$ and the corresponding measure ν is either the Lebesgue measure on $\mathbb{R}^d$ or the counting measure on $\mathbb{Z}^d$, respectively. For any point process X and ν-measurable set $W \subset \mathcal{X}$ we write $N(W)$ for the number of points of $X \cap W$ and $|W|$ for the volume of W, i.e. $|W| = \nu(W)$ is either the Lebesgue measure of W if $\mathcal{X} = \mathbb{R}^d$ or its cardinality if $\mathcal{X} = \mathbb{Z}^d$. Moreover, for any finite set $X \subset \mathcal{X}$ and any function $F : \mathcal{X}^2 \to \mathbb{R}$, we write $F[X]$ for the matrix $(F(x,y))_{x,y \in X}$ where all $x \in X$ are arbitrarily ordered. We write $F(x,y) := F_0(y-x)$ if F is invariant by translation, in which case $F_0[X]$ will refer to the matrix $F[X]$, and we write $F(x,y) := F_{\text{rad}}(\|y-x\|)$ if F is a radial function. Here $\|.\|$ denotes the euclidean norm on $\mathcal{X}$ but we will also use the notation $\|.\|$ for the operator norm when applied to a linear operator, without ambiguity. We denote by $\hat{f}$ the Fourier transform of any function $f : \mathcal{X} \mapsto \mathbb{R}$, defined for any $x \in \mathcal{X}^*$ by

$$\hat{f}(x) := \int_{\mathcal{X}} f(t) \exp(-2i\pi \langle t, x \rangle) \mathrm{d}\nu(t),$$

where $\mathcal{X}^* = \mathbb{R}^d$ if $\mathcal{X} = \mathbb{R}^d$, $\mathcal{X}^* = [0,1]^d$ if $\mathcal{X} = \mathbb{Z}^d$ and $\langle ., . \rangle$ denotes the usual scalar product on $\mathcal{X}$. Finally, for any hermitian matrix M we write $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ for the highest and the lowest eigenvalue of M, respectively, and, for any two hermitian matrices (or operators on a Hilbert space) M and M', we use the Loewner order notation $M \leqslant M'$ when $M' - M$ is positive definite.

DPPs are commonly defined through their joint intensity functions.

Definition 2.1. *Let X be a point process on $(\mathcal{X}, \nu)$ and $n \geqslant 1$ be an integer. If there exists a non-negative function $\rho_n : \mathcal{X}^n \to \mathbb{R}$ such that*

$$\mathbb{E}\left[\sum_{x_1, \cdots, x_n \in X}^{\neq} f(x_1, \cdots, x_n) \right] = \int_{\mathcal{X}^n} f(x_1, \cdots, x_n) \rho_n(x_1, \cdots, x_n) \mathrm{d}\nu(x_1) \cdots \mathrm{d}\nu(x_n)$$

for all locally integrable functions $f : \mathcal{X}^n \to \mathbb{R}$, where the symbol $\neq$ means that the sum is done for distinct x_i, then ρ_n is called the n-th order joint intensity function of X.

DPPs are then defined the following way.

Definition 2.2. *Let $K : \mathcal{X}^2 \to \mathbb{R}$ be a locally square integrable, hermitian function such that its associated integral operator on $L^2(\mathcal{X}, \nu)$,*

$$\mathcal{K} : f \mapsto \left(\mathcal{K}f : x \mapsto \int_{\mathcal{X}} K(x,y) f(y) \mathrm{d}\nu(y) \right),$$

is locally of trace class with eigenvalues in $[0,1]$. X is said to be a determinantal point process on $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \nu)$ with kernel K if its joint intensity functions exist and satisfy

$$\rho_n(x_1, \cdots, x_n) = \det\left(K[x]\right) \tag{2.1}$$

for all integer n and for all $x = (x_1, \cdots, x_n) \in \mathcal{X}^n$.

When $\mathcal{X} = \mathbb{R}^d$ the DPP is said to be continuous and when $\mathcal{X} = \mathbb{Z}^d$ then the DPP is said to be discrete. In the latter case, the integral operator $\mathcal{K}$ can be seen as the infinite matrix $K[\mathbb{Z}^d] := (K(x,y))_{x,y \in \mathbb{Z}^d}$. Moreover, when K is translation invariant (resp. radial) then the associated DPP is stationary (resp. isotropic) and, in the discrete case, $\mathcal{K}$ writes as a multilevel Toeplitz matrix. Finally, we write $\mathcal{I}$ for the identity operator on $L^2(\mathcal{X}, \nu)$ and $\mathcal{I}_W$ for its restriction on $L^2(W, \nu)$ for any $W \subset \mathcal{X}$.

Let X be a DPP on $\mathcal{X}$ with kernel K and associated integral operator $\mathcal{K}$. If $\|\mathcal{K}\| < 1$, then X admits on any compact set $W \subset \mathcal{X}$ a density with respect to the unit rate homogenous Poisson point process on W, as described now. We recall that for any compact set W, the projection $\mathcal{K}_W$ of $\mathcal{K}$ on $L^2(W, \nu)$ is a compact operator whose kernel can be written by Mercer's theorem as

$$K_W(x,y) = \sum_i \lambda_i^W \phi_i^W(x) \bar{\phi}_i^W(y), \ \forall x, y \in W,$$

where the λ_i^W are the eigenvalues of $\mathcal{K}_W$ and the ϕ_i^W are the corresponding family of orthonormal eigenfunctions (see [16] for more details). When $\|\mathcal{K}\| < 1$, we define the operator $\mathcal{L} = \mathcal{K}(\mathcal{I} - \mathcal{K})^{-1}$ and denote by L its kernel. Similarly, we define the operator $\mathcal{L}_{[W]} = \mathcal{K}_W(\mathcal{I}_W - \mathcal{K}_W)^{-1}$ and denote by $L_{[W]}$ its kernel whose spectral decomposition writes

$$L_{[W]}(x,y) := \sum_i \frac{\lambda_i^W}{1 - \lambda_i^W} \phi_i^W(x) \bar{\phi}_i^W(y). \tag{2.2}$$

Note that contrary to $\mathcal{K}_W$ with $\mathcal{K}$, the operator $\mathcal{L}_{[W]}$ does not correspond to the restriction of $\mathcal{L}$ to $L^2(W, \nu)$. Another difference between $\mathcal{L}_{[W]}$ and $\mathcal{L}$ is that when X is a stationary (resp. isotropic) DPP, $L(x,y)$ only depends on $y - x$ (resp. $\|y - x\|$) but this is not true for $L_{[W]}$.

Theorem 2.3 ([21, 27]). *Let X be a DPP on $(\mathcal{X}, \nu)$ with kernel K whose eigenvalues lie in $[0, 1[$ and let W be a compact set of $\mathcal{X}$. Then $X \cap W$ is absolutely continuous with respect to the homogeneous Poisson point process on W with intensity 1 and has density*

$$f(x) = \exp(|W|) \det(\mathcal{I}_W - \mathcal{K}_W) \det(L_{[W]}[x])$$

for all $x \in \cup_n W^n$.

In the above expression, the first determinant corresponds to the Fredholm determinant of the operator $\mathcal{I}_W - \mathcal{K}_W$, which is equal to $\prod_i (1 - \lambda_i^W)$, while the second determinant is the standard matrix determinant.

3 Likelihood of DPPs

3.1 Likelihood estimation

Let X be a DPP on $(\mathcal{X}, \nu)$ with kernel K^{θ^*} belonging to a parametric family $\{K^\theta, \theta \in \Theta\}$, where Θ is the space of parameters. We consider the likelihood estimation of θ^*, as described below, from a unique observation of $X \cap W$ where W is a bounded subset of $\mathcal{X}$. We furthermore consider an increasing domain asymptotic framework, meaning that our asymptotic properties stand when $n \to \infty$ and $W = W_n$ is an increasing sequence of subsets of $\mathcal{X}$.

For the standard parametric families of continuous DPPs in $\mathbb{R}^d$, as those presented in Section 4.1, the parameter space Θ is a subset of $\mathbb{R}^p$ for some integer $p \geqslant 1$. However we do not need to make such an assumption for our purpose, and the likelihood approximation that we develop below is true whatever Θ is, provided the associated DPP is stationary. In particular the parameter θ in K^θ can be the kernel K itself. This last setting makes sense when $\mathcal{X} = \mathbb{Z}^d$ where the whole matrix K_W (which is multilevel Toeplitz in the stationary case) can be estimated from a realisation of $X \cap W$, as considered in image analysis in [18].

From Theorem 2.3, we get that the (normalized) log-likelihood of $X \cap W$ for any parametric family of DPPs writes:

$$l(\theta|X) = 1 + \frac{1}{|W|}\Big(\mathrm{logdet}(\mathcal{I}_W - \mathcal{K}_W^\theta) + \mathrm{logdet}(L_{[W]}^\theta[X \cap W])\Big) \tag{3.1}$$

where $\mathcal{K}^\theta$ is the integral operator associated to K^θ and $L_{[W]}^\theta$ is given by (2.2), the eigenvalues and eigenvectors then depending on θ. The maximum likelihood estimate of θ is then

$$\hat{\theta} \in \arg\max_{\theta \in \Theta} l(\theta|X).$$

Computing the log-likelihood (3.1) requires knowing the spectral decomposition of $\mathcal{K}_W^\theta$ for all θ. This is possible in the case of DPPs on a finite space whose kernels are finite matrices, provided the dimension of the space is not too large, but this spectral decomposition is usually not known for continuous DPPs. This motivates the following approximations.

3.2 Approximation of the likelihood for stationary DPPs

When $\mathcal{X} = \mathbb{R}^d$ and the observation window W is rectangular, an approximation of (3.1) for stationary kernels is proposed in [19], using a truncated Fourier series. For example, if $W = [-1/2, 1/2]^d$, this relies on the following approximation of the kernel:

$$K^\theta(x,y) = K_0^\theta(x-y) \approx \sum_{\substack{k\in\mathbb{Z}^d \\ \|k\|<N}} c_k e^{i2\pi\langle k, y-x\rangle}, \quad \text{where} \quad c_k := \int_W K_0^\theta(t) e^{-i2\pi\langle k,t\rangle} \mathrm{d}t \approx \hat{K}_0^\theta(k),$$

for some truncation constant N. Since the eigenvalues and eigenvectors of this kernel approximation are explicitly known, the log-likelihood (3.1) is then approximated in [19] by

$$1 + \frac{1}{|W|}\left(\sum_{\substack{k\in\mathbb{Z}^d \\ \|k\|<N}} \log(1-\hat{K}_0^\theta(k)) + \mathrm{logdet}(L^\theta_{app}[X\cap W]) \right), \tag{3.2}$$

where

$$L^\theta_{app}(x,y) := \sum_{\substack{k\in\mathbb{Z}^d \\ \|k\|<N}} \frac{\hat{K}_0^\theta(k)}{1-\hat{K}_0^\theta(k)} e^{i2\pi\langle k, y-x\rangle}. \tag{3.3}$$

The same kind of approximations can be carried out when $\mathcal{X} = \mathbb{Z}^d$, still for rectangular windows W, in which case $\hat{K}_0^\theta(k)$ in (3.2) has to be replaced by the discrete Fourier transform of $K_0(x)$, $x \in W$, and no truncation is needed since the series become a finite sum. This approximation in $\mathbb{Z}^d$ amounts to consider a periodic extension of the stationary DPP outside W, see [18] for details.

Our new approximation is based on a different expression of (3.1) in terms of the self-convolution products of the function $(x,y) \mapsto \mathbb{1}_W(x) K^\theta(x,y) \mathbb{1}_W(y)$ through the following identities (see [26] for example). For all $W \subset \mathcal{X}$,

$$\mathrm{logdet}(\mathcal{I}_W - \mathcal{K}_W^\theta) = -\sum_{k=1}^\infty \frac{1}{k} \int_{W^k} K^\theta(x_1,x_2)\cdots K^\theta(x_{k-1},x_k) K^\theta(x_k,x_1) \mathrm{d}\nu^k(x) \tag{3.4}$$

and for all $x, y \in W$,

$$L^\theta_{[W]}(x,y) = K^\theta(x,y) + \sum_{k=1}^\infty \int_{W^k} K^\theta(x,z_1) K^\theta(z_1,z_2)\cdots K^\theta(z_{k-1},z_k) K^\theta(z_k,y) \mathrm{d}\nu^k(z), \tag{3.5}$$

$$L^\theta(x,y) = K^\theta(x,y) + \sum_{k=1}^\infty \int_{\mathcal{X}^k} K^\theta(x,z_1) K^\theta(z_1,z_2)\cdots K^\theta(z_{k-1},z_k) K^\theta(z_k,y) \mathrm{d}\nu^k(z). \tag{3.6}$$

These convolution products are too difficult to be computed in the general case, but for stationary DPPs satisfying $\|\mathcal{K}^\theta\| < 1$ then $\hat{L}_0^\theta = \hat{K}_0^\theta/(1-\hat{K}_0^\theta)$ as a consequence of (3.6). Accordingly, as justified later in Proposition 3.1, an asymptotic approximation when the

observation window W is large enough gives

$$L^{\theta}_{[W]}(x,y) \approx L^{\theta}_0(y-x) = \int_{\mathcal{X}^*} \frac{\hat{K}^{\theta}_0(t)}{1-\hat{K}^{\theta}_0(t)} \exp(2i\pi\langle t, y-x\rangle)\mathrm{d}t, \tag{3.7}$$

$$\frac{1}{|W|}\mathrm{logdet}(\mathcal{I}_W - \mathcal{K}^{\theta}_W) \approx \int_{\mathcal{X}^*} \log(1-\hat{K}^{\theta}_0(x))\mathrm{d}x. \tag{3.8}$$

This motivates our following approximation of the log-likelihood:

$$\widetilde{l}(\theta|X) := 1 + \int_{\mathcal{X}^*} \log(1-\hat{K}^{\theta}_0(x))\mathrm{d}x + \frac{1}{|W|}\mathrm{logdet}(L^{\theta}_0[X\cap W]), \tag{3.9}$$

where L^{θ}_0 is given in (3.7). This approximation, like (3.2), can be computed whenever we know the expression of $\hat{K}^{\theta}_0$, which is the case for all classical families of stationary DPPs built from covariance functions, as those presented in Section 4.1. The main advantage of (3.9) compared to the Fourier approximation (3.2) is that it is not limited to rectangular windows W but can be used with any window shape. This estimation method also has the advantage that it does not require any tuning parameter of any kind compared to the choice of N in (3.2) or alternative moment methods [9, 20].

The idea to use a convolution approximation was actually briefly suggested in [19, Appendix L] but the associated approximation was given under a more restrictive form that required knowing an exact expression of the iterative self-convolution products of K^{θ}_0 for all θ. Moreover, an important drawback was pointed out in [19] concerning the presence of possible edge effects, which may affect the quality of estimation of strongly repulsive DPPs. As shown in Section 5, this problem also occurs with our approximation: while it works really well with DPPs with low repulsion, and therefore minimal edge effects, some edge corrections are needed for more repulsive DPPs. The next section deals with this aspect.

3.3 Periodic edge-corrections

In order to alleviate the possible edge-effects mentioned above, we suggest to introduce a periodic approximation. We assume in this section that the observation window $W \subset \mathcal{X}$ is rectangular. Without loss of generality, we set $W = ([-l_1/2, l_1/2] \times \cdots \times [-l_d/2, l_d/2]) \cap \mathcal{X}$. Using a periodic approximation amounts to consider the observation window as the flat torus $\mathbb{T}_W := \mathcal{X}\backslash l_1\mathbb{Z} \times \cdots \times \mathcal{X}\backslash l_d\mathbb{Z}$. This way, points close to the border of the window W are brought close to each other in order to compensate edge effects.

More precisely, we replace all instances of $L^{\theta}_0(y-x)$ in the stochastic part $L^{\theta}_0[X\cap W]$ of (3.9) by

$$L^{\theta,\mathbb{T}}_0(y-x) := L^{\theta}_0 \begin{pmatrix} y_1 - x_1 \bmod(l_1) \\ \vdots \\ y_d - x_d \bmod(l_d) \end{pmatrix}.$$

This is equivalent to replacing L_0^θ by a periodic version of itself on W. The approximate likelihood then writes for any parameter θ:

$$\widetilde{l}^{\,\mathbb{T}}(\theta|X) := 1 + \int_{\mathcal{X}^*} \log(1 - \hat{K}_0^\theta(x))\mathrm{d}x + \frac{1}{|W|}\mathrm{logdet}\left(L_0^{\theta,\mathbb{T}}[X \cap W]\right). \tag{3.10}$$

Note that, since we consider a periodic version of L_0^θ on W then it can be approximated by its Fourier series, which corresponds to the idea of the approximation (3.2) of [19]. This is why both (3.10) and (3.2) are nearly equal, see Figure 1 for an example. But approximating $L_{[W]}^\theta$ as in (3.10) instead of using a truncation of its Fourier series leads to a smoother likelihood and overall slightly better results, as well as a more computationally efficient method. Indeed, as explained in [19], the Fourier approximation (3.3) of $L_{[W]}^\theta$ is a sum of $(2N)^d$ terms where the truncation parameter N is chosen such that

$$\sum_{n\in\mathbb{Z}^d\cap[-N,N]^d} \hat{K}_0^\theta(n) > 0.99 \sum_{n\in\mathbb{Z}^d} \hat{K}_0^\theta(n).$$

For important parametric models, including the Whittle-Matern and the Bessel families (see Section 4.1), $\hat{K}_0^\theta(n)$ has a polynomial decay with respect to n, leading to a large choice of N in (3.3). In comparison, as detailed in section 4.2, depending on the parametric model, we either have an analytic expression of L_0^θ or, when the self convolution products of K_0^θ are known, we can express L_0^θ as the infinite sum

$$L_0^\theta(x) = \sum_{n\geqslant 0} (K_0^\theta)^{*n}(x) \tag{3.11}$$

where

$$|(K_0^\theta)^{*n}(x)| = \left|\int_{\mathcal{X}^*} (\hat{K_0^\theta})^n(t) e^{-i2\pi\langle x,t\rangle}\mathrm{d}t\right| \leqslant K_0^\theta(0)\|\hat{K}_0^\theta\|_\infty^{n-1}$$

has an exponential decay with respect to n. The approximation of $L_{[W]}^\theta$ by (3.11) will then require much fewer terms than the approximation by (3.3).

Despite the appealing of the approximation (3.10), there is one possible issue in that the determinant of $L^{\theta,\mathbb{T}}[X \cap W]$ is not guaranteed to be positive. Remember that this positivity is guaranteed for any $X \cap W$ whenever the kernel $L^{\theta,\mathbb{T}}$ is positive, or equivalently whenever its associated integral operator has positive eigenvalues. But due to the periodicity of $L_0^{\theta,\mathbb{T}}$, these eigenvalues correspond to the coefficients of its Fourier series that write for any $k = (k_1, \ldots, k_d)$

$$\frac{1}{|W|}\int_W L_0^\theta(x) \exp\left(-2i\pi \sum_{i=1}^d \frac{k_i x_i}{l_i}\right) \mathrm{d}\nu(x). \tag{3.12}$$

When W is large, the above integral is approximately equal to $\hat{L}_0^\theta(k_1/l_1, \cdots, k_d/l_d)$ which is positive. This shows that we can expect the determinant of $L^{\theta,\mathbb{T}}[X \cap W]$ to be positive when W is large enough. In our simulations displayed in Section 5, this determinant was positive in all runs, except a few times with the Bessel-type kernel associated to high values of the repulsion parameter α.

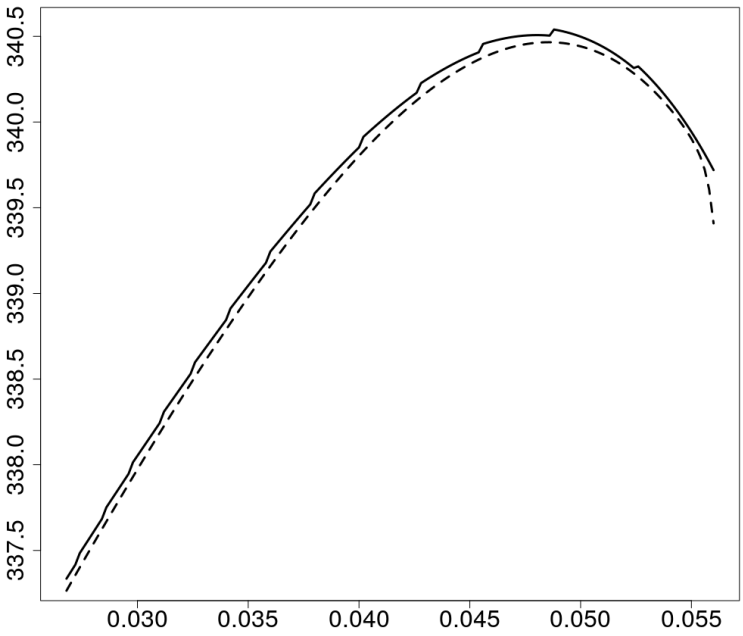


Figure 1: Comparison between the two approximations (3.2) (solide line) and (3.10) (dashed line) of $\alpha \mapsto l(\rho^*, \alpha|X)$ where X is a realisation of a DPP with Gaussian-type kernel (see Table 1) with true parameters $\rho^* = 100$ and $\alpha^* = 0.05$ on the window $W = [0, 1]^2$.

Finally, note that extending the above edge correction to non rectangular windows is not straightforward and we do not provide a general solution. We however introduce in the simulation example of Section 5.3 a procedure that can be adapted to any isotropic DPP model.

3.4 Theoretical Results

In order to verify the theoretical soundness of the asymptotic log-likelihood approximation (3.9) we want to show that $|\widetilde{l}(\theta|X) - l(\theta|X)| \xrightarrow{a.s.} 0$ for all $\theta \in \Theta$ when the observation window W grows towards $\mathcal{X}$. For this purpose, we consider a sequence of increasing observation windows W_n satisfying the following assumptions.

Condition $(\mathcal{W})$: W_n is an increasing sequence of compact subsets of $\mathcal{X}$ such that $\bigcup_{n\geqslant 0} W_n = \mathcal{X}$ and there exists an increasing non-negative sequence $r_n \in \mathbb{R}_+^{\mathbb{N}}$ such that $r_n \underset{n\to\infty}{\longrightarrow} \infty$ and

$$|(\partial W_n \oplus r_n) \cap W_n| = o(|W_n|), \tag{3.13}$$

where $\oplus$ denotes the Minkowski sum. Moreover,

$$\forall \delta > 0, \ \sum_{n\geqslant 0} \exp(-\delta |W_n|) < \infty. \tag{3.14}$$

The first assumption (3.13) means that the boundary of W_n must not be too irregular. This is not an issue in most practical applications. For example, if $\mathcal{X} = \mathbb{R}^d$ and $(W_n)_{n\geqslant 0}$ is

a sequence of spheres with radius $R_n \underset{n\to\infty}{\longrightarrow} \infty$, then (3.13) is satisfied with $r_n = \sqrt{R_n}$. As another example, assume that $(W_n)_{n\geqslant 0}$ is a sequence of rectangular windows $[-l_{1,n}/2, l_{1,n}/2]\times \cdots \times [-l_{d,n}/2, l_{d,n}/2]$ such that $l_{i,n} \underset{n\to\infty}{\longrightarrow} \infty$ for each i, then

$$(\partial W_n \oplus r_n) \cap W_n \subset \Big([-l_{1,n}/2, -l_{1,n}/2 + r_n] \cup [l_{1,n}/2 - r_n, l_{1,n}/2]\Big) \times \cdots \times \Big([-l_{d,n}/2, -l_{d,n}/2 + r_n] \cup [l_{d,n}/2 - r_n, l_{d,n}/2]\Big)$$

and

$$\frac{|(\partial W_n \oplus r_n) \cap W_n|}{|W_n|} \leqslant \prod_{i=1}^{d} \left(\frac{2r_n}{l_{i,n}}\right)$$

which vanishes when n goes to infinity with the choice $r_n = \sqrt{\min_i l_{i,n}}$. The second hypothesis (3.14) is a technical assumption needed to get the almost sure convergence in Proposition 3.2. Without this assumption, the convergence remains true but in probability instead of almost surely.

We first consider the deterministic part of (3.1), which is the Fredholm log-determinant. Its asymptotic behaviour given below is justified in Section 7.1 and was already proved in a slightly different setting in [26, Proposition 5.9].

Proposition 3.1. *Let $K_0 : \mathcal{X} \mapsto \mathbb{R}$ be a function in $L^2(\mathcal{X}, \nu)$ with an integrable Fourier transform $\hat{K}_0$ taking values in $[0, 1[$ and let $(W_n)_{n\geqslant 0}$ satisfy Condition $(\mathcal{W})$. We denote by $\mathcal{K}_{W_n}$ the projection on $L^2(W_n)$ of the integral operator associated with the kernel $(x, y) \mapsto K_0(x - y)$. Then,*

$$\frac{1}{|W_n|}\mathrm{logdet}(\mathcal{I}_{W_n} - \mathcal{K}_{W_n}) \underset{n\to\infty}{\longrightarrow} \int_{\mathcal{X}^*} \log(1 - \hat{K}_0(x))\mathrm{d}x.$$

Concerning the stochastic part of the log-likelihood (3.1), that is $\mathrm{logdet}(L^\theta_{[W]}[X \cap W])$, its behaviour is much more difficult to control in general. The main issue is that the determinant vanishes when two points of $X \cap W$ gets arbitrarily close to each other, but no relationship between how close these points are from each other and the value of the determinant is known, making the likelihood difficult to control. To our knowledge, the only related result is that, in most cases, the lowest eigenvalue of $L^\theta_{[W]}[X]$ is non zero iff $\inf_{x,y\in X} \|y - x\| > 0$ [3]. The latter condition is automatically satisfied if X is supported on a lattice but not when $\mathcal{X} = \mathbb{R}^d$. The next result focuses on the first case.

Proposition 3.2. *Let $(W_n)_{n\in\mathbb{N}}$ satisfy Condition $(\mathcal{W})$ and let $\{K^\theta, \theta \in \Theta\}$ be a family of translation-invariant DPP kernels on $\mathbb{Z}^d$ such that for all $\theta \in \Theta$, there exists constants $A_\theta, \tau_\theta > 0$ and $M_\theta < 1$ such that*

$$\forall x \in \mathbb{Z}^d, \quad |K^\theta_0(x)| \leqslant \frac{A_\theta}{1 + \|x\|^{d+\tau_\theta}} \text{ and } 0 < \hat{K}^\theta_0(x) \leqslant M_\theta.$$

Let X be the realization of a DPP on $\mathbb{Z}^d$ with kernel K^{θ^*}, $\theta^* \in \Theta$. Then, for all $\theta \in \Theta$,

$$\frac{1}{|W_n|}\Big|\mathrm{logdet}(L^\theta_0[X \cap W_n]) - \mathrm{logdet}(L^\theta_{[W_n]}[X \cap W_n])\Big| \xrightarrow{a.s.} 0.$$

The only restrictive assumptions in Proposition 3.2 is the need for K_0^θ to decay faster than $\|x\|^{-d}$ and the fact that $\hat{K}_0^\theta$ never vanishes. In the usual setting where the kernels are parametric covariance functions (see Propositon 4.1), these assumptions are generally satisfied. That includes the Gaussian, Cauchy and Whittle-Matern kernels. The only exception amongst standard kernels is the Bessel-type kernel, that will be examined by simulations in Section 5.2. Based on Propositions 3.1 and 3.2, we thus obtain the consistency of the likelihood approximation (3.9) when $\mathcal{X} = \mathbb{Z}^d$.

Corollary 3.3. *Let $\{K^\theta, \theta \in \Theta\}$ be a family of translation-invariant DPP kernels on $\mathbb{Z}^d$ satisfying the assumptions of Proposition 3.2, then $|\widetilde{l}(\theta|X) - l(\theta|X)| \xrightarrow{a.s.} 0$ for all $\theta \in \Theta$.*

Getting the same result for DPPs on $\mathbb{R}^d$ is still an open problem. However the next proposition shows that a DPP on $\mathbb{R}^d$ can be approximated by a discrete DPP on an arbitrarily small regular grid of $\mathbb{R}^d$, for which Corollary 3.3 applies. Note that the assumptions on $\hat{K}_0$ below are satisfied for all standard parametric families, see Section 4.1.

Proposition 3.4. *Let X be a stationary DPP on $\mathbb{R}^d$ with kernel $K(x,y) = K_0(y-x)$, where K_0 is a square integrable function such that $\hat{K}_0$ takes values in $[0,1[$ and*

$$\forall x \in \mathbb{R}^d, \quad 0 \leqslant \hat{K}_0(x) \leqslant \frac{A}{1 + \|x\|^{d+\tau}}$$

for some constant $A, \tau > 0$. For all integer $\varepsilon \geqslant 0$, define X_ε as the DPP on $\mathbb{Z}^d$ with kernel $K_\varepsilon(x,y) := \varepsilon^d K_0(\varepsilon(y-x))$. Then, X_ε is well-defined for small enough ε and the distribution of $\varepsilon X_\varepsilon$, the DPP X_ε rescaled by a factor ε, weakly converges to the distribution of X when ε tends to 0.

In the end, Corollary 3.3 tells us that the asymptotic approximation of the log-likelihood (3.9) is theoretically sounded for most classical parametric families of stationary DPPs on $\mathbb{Z}^d$ and, as a consequence of Proposition 3.4, also theoretically sounded for any discrete approximation of continuous DPPs on an arbitrarily small regular grid of $\mathbb{R}^d$.

4 Application to standard parametric families

4.1 Classical parametric families of stationary DPPs

A classical way of generating parametric families of stationary DPPs is the following result.

Proposition 4.1. *Let $K_0 : \mathcal{X} \mapsto \mathbb{R}$ be a bounded square integrable symmetric function on $\mathbb{R}^d$ such that its Fourier transform $\hat{K}_0$ takes values in $[0,1]$. Then, the function $K(x,y) := K_0(y-x)$ is a DPP kernel on $(\mathcal{X}, \nu)$.*

This proposition is proved in [19] in the case $\mathcal{X} = \mathbb{R}^d$. Since symmetric functions K_0 with non negative Fourier transform are covariance functions, this result implies that we can consider as many parametric families of DPPs as there are parametric families of covariance

	$K_0(x)$	$\hat{K}_0(x)$	$\rho_{\max}$
Gauss	$\rho \exp\left(-\frac{\|x\|^2}{\alpha^2}\right)$	$\rho(\sqrt{\pi}\alpha)^d \exp(-\|\pi\alpha x\|^2)$	$(\sqrt{\pi}\alpha)^{-d}$
Bessel	$\rho 2^{d/2}\Gamma(d/2+1)\frac{J_{d/2}(\sqrt{2d}\|y-x\|/\alpha)}{(\sqrt{2d}\|y-x\|/\alpha)^{d/2}}$	$\frac{\rho}{\rho_{\max}}\mathbb{1}_{\|x\|\leqslant\sqrt{d/(2\pi^2\alpha^2)}}$	$\frac{d^{d/2}}{(2\pi)^{d/2}\alpha^d\Gamma(d/2+1)}$
Cauchy	$\rho\left(1+\left\|\frac{x}{\alpha}\right\|^2\right)^{-\frac{d+1}{2}}$	$\frac{\rho(\sqrt{\pi}\alpha)^d\sqrt{\pi}}{\Gamma((d+1)/2)}e^{-\|2\pi\alpha x\|}$	$\frac{\Gamma((d+1)/2)}{\pi^{(d+1)/2}\alpha^d}$
WM	$\rho\frac{2^{1-\sigma}}{\Gamma(\sigma)}\left\|\frac{x}{\alpha}\right\|^\sigma K_\sigma\left(\left\|\frac{x}{\alpha}\right\|\right)$	$\rho\frac{\Gamma(\sigma+d/2)}{\Gamma(\sigma)}\frac{(2\sqrt{\pi}\alpha)^d}{(1+\|2\pi\alpha x\|^2)^{\sigma+d/2}}$	$\frac{\Gamma(\sigma)}{\Gamma(\sigma+d/2)(2\sqrt{\pi}\alpha)^d}$

Table 1: Examples of parametric kernels K_0 on $\mathbb{R}^d$, along with their Fourier transform $\hat{K}_0$. For each family, the intensity is ρ and the range parameter is α. The existence condition $\hat{K}_0 \leqslant 1$ is equivalent to $\rho \leqslant \rho_{\max}$ where $\rho_{\max}$ is given in the last column. The Whittle-Matérn model (WM) also contains a shape parameter $\sigma > 0$. Here $J_{d/2}$ denotes the Bessel function of the first kind and K_σ the modified Bessel function of the second kind.

functions. The assumption that $\hat{K}_0 \leqslant 1$ simply adds a bound on the parameters of the family. Various examples are presented and studied in [8, 19]. We provide in Table 1 some examples in $\mathbb{R}^d$. Note that for simplification, we call in this table Bessel kernel the particular case of the Bessel kernel in [8] where the shape parameter is $\sigma = 0$, and Cauchy kernel the particular case in [19] where the shape parameter is $1/2$. If the shape parameter is different for these models, then closed formulas are available for K_0 and $\hat{K}_0$, but not for L_0 (see the next section and Table 2).

4.2 Expressions of L_0

When computing the approximate log-likelihood $\widetilde{l}(\theta|X)$ in (3.9) or its edge-corrected version (3.10), one has to compute $L_0(y-x)$ for each pair of points $(x,y) \in (X \cap W)^2$. It is thus important to find faster ways to compute values of L_0 than the d-dimensional integral (3.7). An important example arises when K_0 is a radial function, denoted by K_{rad}. In this case, the corresponding DPP is isotropic and L_0 is also a radial function, denoted by L_{rad}. The Fourier transform can then be expressed by a Hankel transform which gives

$$\hat{K}_{\text{rad}}(r) = \frac{2\pi}{r^{d/2-1}}\int_0^\infty s^{d/2}K_{\text{rad}}(s)J_{d/2-1}(2\pi sr)\mathrm{d}s$$

and

$$L_{\text{rad}}(r) = \frac{2\pi}{r^{d/2-1}}\int_0^\infty s^{d/2}\frac{\hat{K}_{\text{rad}}(s)}{1-\hat{K}_{\text{rad}}(s)}J_{d/2-1}(2\pi sr)\mathrm{d}s.$$

The expression of L_0 therefore simplifies into a unidimensional integral.

	$L_0(x)$
Gauss	$\sum_{n\geqslant 1} \rho^n \frac{(\sqrt{\pi}\alpha)^{d(n-1)}}{n^{d/2}} \exp\left(-\frac{\|x\|^2}{n\alpha^2}\right)$
Bessel	$\frac{\rho 2^{d/2}\Gamma(d/2+1)}{1-\rho\frac{(2\pi)^{d/2}\alpha^d\Gamma(d/2+1)}{d^{d/2}}} \frac{J_{d/2}(\sqrt{2d}\|x\|/\alpha)}{(\sqrt{2d}\|x\|/\alpha)^{d/2}}$
Cauchy	$\sum_{n\geqslant 1} \frac{\rho^n}{n^d} \left(\frac{\pi^{(d+1)/2}\alpha^d}{\Gamma((d+1)/2)}\right)^{n-1} \left(1+\left\|\frac{x}{n\alpha}\right\|^2\right)^{-(d+1)/2}$
WM	$\sum_{n\geqslant 1} \frac{\rho^n(\sqrt{\pi}\alpha)^{d(n-1)}\Gamma(\sigma+d/2)^n}{2^{n\sigma-1-(n-1)d/2}\Gamma(\sigma)^n\Gamma(n\sigma+nd/2)} \left\|\frac{x}{\alpha}\right\|^{n\sigma+(n-1)d/2} K_{n\sigma+(n-1)d/2}\left(\left\|\frac{x}{\alpha}\right\|\right)$

Table 2: Expression of L_0 defined in (3.7) for the parametric kernels given in Table 1.

Moreover, we may exploit the relation $\hat{L}_0 = \hat{K}_0/(1-\hat{K}_0) = \sum_{n\geqslant 1}(\hat{K}_0)^n$ and try to compute the inverse Fourier transform to express L_0 as a series with exponentially decreasing coefficients (see the discussion in Section 3.3) or even get an analytic expression. This strategy leads to closed-form formulas of L_0 for the classical parametric families displayed in Table 1. The results, obtained after straightforward calculus, are given in Table 2.

4.3 Estimation of the intensity by MLE

Assume that the parametric DPP kernel writes for some parameters ρ and θ

$$K^{\rho,\theta}(x,y) = \rho\widetilde{K}^{\theta}(x,y) \tag{4.1}$$

where $\widetilde{K}^{\theta}(x,x) = 1$ for all x. The parameter ρ corresponds here to the intensity of the DPP and θ to the other parameters of the model. This is the setting of all standard parametric models, including those presented in Table 1.

When jointly estimating (ρ,θ) from a realisation of the DPP X on W by the approximate MLE, simulations usually show that the estimate of ρ appears to be very close to $N(W)/|W|$. One explanation given in [19] is that, by doing a first order convolution approximation in (3.4) and (3.6), we get

$$l(\rho,\theta|X) \approx 1-\rho+\log(\rho)\frac{N(W)}{|W|} + \frac{1}{|W|}\mathrm{logdet}(\widetilde{K}^{\theta}_W[X\cap W])$$

and the maximum point of this approximation is $\hat{\rho} = N(W)/|W|$. We even show in Proposition A.3 that, in the case of Bessel type DPP kernels with parameters (ρ,α) as presented

in Table 1, $\hat{\rho} = N(W)/|W|$ is always the maximum point of $\rho \mapsto \widetilde{l}(\rho, \alpha|X)$ for any α. This result suggests that, instead of jointly estimating ρ and θ, it is more computationally efficient to directly estimate ρ by $\hat{\rho} = N(W)/|W|$ and then θ by an argument of the maximum of $\theta \mapsto \widetilde{l}(\hat{\rho}, \theta|X)$.

5 Simulation study

In this section we perform a simulation study to investigate the performance of our approximate MLE, with and without edge effect correction, and compare it to minimum contrast estimators (MCE for short) based on Ripley's K function and on the pair correlation function (pcf for short), both being common second-order moment estimators used in spatial statistics. We refer to [9] for more detailed information on these MCEs applied to DPPs. At the exception of the special case of Bessel-type DPPs considered in Section 5.2, we chose not to compare our estimators to the Fourier approximation (3.2) of [19] since, as explained in Section 3.3, this estimator yields nearly the same results as our corrected MLE, which we confirmed in our testings, with the notable difference of the Fourier approximation being about ten times longer to compute in our examples.

5.1 Whittle-Matérn, Cauchy and Gaussian-type DPPs

We consider in this section the parametric models in Table 1 that are covered by our theoretical assumptions in Section 3.4, that are the Whittle-Matérn, Cauchy and Gaussian-type DPPs. From this perspective, these are favourable models for our likelihood approximation approach. All these models are of the form (4.1), then following Section 4.3, we estimate ρ by $\hat{\rho} = N(W)/|W|$ for all methods, and the performances are evaluated on the estimation of α only. Note that for the Whittle-Matérn model, we do not consider the estimation of the shape parameter σ, which was assumed to be known. The joint estimation of (α, σ) for this model is known to be a poorly identifiable problem and it is customary to choose the best σ from a small finite grid by profile likelihood (see [19]). For the estimation of α, we have performed the same kind of simulations for the three models in $\mathbb{R}^2$. The results and conclusions are similar. In the following we only present the details for the Gaussian-type DPP.

We consider realizations of the Gaussian-type DPP with true parameters $\rho^* = 100$ and $\alpha^* \in \{0.01, 0.03, 0.05\}$, when the observation window W is either $[0,1]^2$, $[0,2]^2$ or $[0,3]^2$. When $\rho^* = 100$, α can take values in $]0, (10\sqrt{\pi})^{-1} \approx 0.056[$ since the process exists if and only if $\pi\rho\alpha^2 \leqslant 1$. Therefore, $\alpha^* = 0.01$ corresponds to a weakly repulsive point process, close to a Poisson point process, while $\alpha^* = 0.03$ corresponds to a mildly repulsive DPP and $\alpha^* = 0.05$ corresponds to a strongly repulsive DPP. Examples of realizations are shown in Figure 2. We estimate α^* by the approximate MLE defined in (3.9) and compare it to its edge-corrected version defined in (3.10) as well as MCEs based on the pcf or Ripley's K function. As mentioned before, ρ is replaced by $\hat{\rho} = N(W)/|W|$ in (3.9) and (3.10), and we truncate the series defining L_0 to $n = 50$ (see Table 2). All realisations have been generated

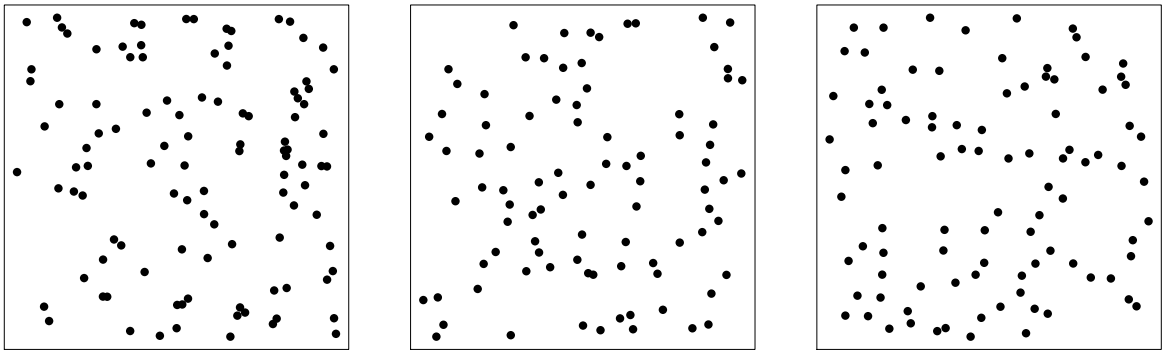


Figure 2: Examples of realizations of Gaussian-type DPPs on $[0,1]^2$ with parameters $\rho^* = 100$ and $\alpha^* = 0.01, 0.03, 0.05$ corresponding to three different degrees of repulsiveness.

in R [25] using the *spatstat* [4] package and both MCEs were computed by the function *dppm* of the same package. The tuning parameters for these MCEs were $r_{\min} = 0.01$, $r_{\max}$ being one quarter of the side length of the window and $q = 0.5$ as recommended in [13]. Boxplots of the difference between the four considered estimators and the true value α^* for 500 runs in all different cases are displayed in Figure 3 and the corresponding mean square errors are given in Table 3.

Window	$[0,1]^2$			$[0,2]^2$			$[0,3]^2$		
α^*	0.01	0.03	0.05	0.01	0.03	0.05	0.01	0.03	0.05
MLE based on $\widetilde{l}^{\,\mathbb{T}}$	**0.83**	0.81	**0.41**	**0.21**	**0.18**	**0.088**	**0.090**	**0.079**	**0.051**
MLE based on $\widetilde{l}$	1.25	1.75	0.54	0.24	0.23	0.28	0.095	0.10	0.20
MCE (pcf)	0.86	**0.77**	0.74	0.31	0.27	0.23	0.17	0.17	0.19
MCE (K)	1.81	1.17	0.51	0.74	0.46	0.21	0.48	0.23	0.12

Table 3: Estimated mean square errors (x10^4) of $\hat{\alpha}$ for Gaussian-type DPPs on different windows and with different values of α, each computed from 500 simulations.

From these results, we remark that when $\alpha^* = 0.01$ and $\alpha^* = 0.03$, inference based on the approximate likelihood $\widetilde{l}(\hat{\rho}, \alpha|X)$ outperforms moment based inference for windows bigger than $[0,2]^2$. This is expected from maximum likelihood based inference and shows that hundreds of points are enough for $\widetilde{l}(\hat{\rho}, \alpha|X)$ to be a good enough approximation of the true likelihood when the underlying DPP is not too repulsive. When $\alpha^* = 0.05$, that is when the negative dependence of the DPP is very strong, then $\widetilde{l}(\hat{\rho}, \alpha|X)$ suffers from edge effects and is heavily biased. In fact, as can be seen in Figure 4, $\widetilde{l}(\hat{\rho}, \alpha|X)$ is an increasing function of α in this case and the estimate is often the highest possible value for α, which is $1/\sqrt{\pi\hat{\rho}}$. The correction $\widetilde{l}^{\,\mathbb{T}}$ introduced in (3.10) gives more accurate values of the likelihood for high values of α, as shown in Figure 4. Finally this estimator outperforms the other ones in nearly every cases and especially the most repulsive ones.

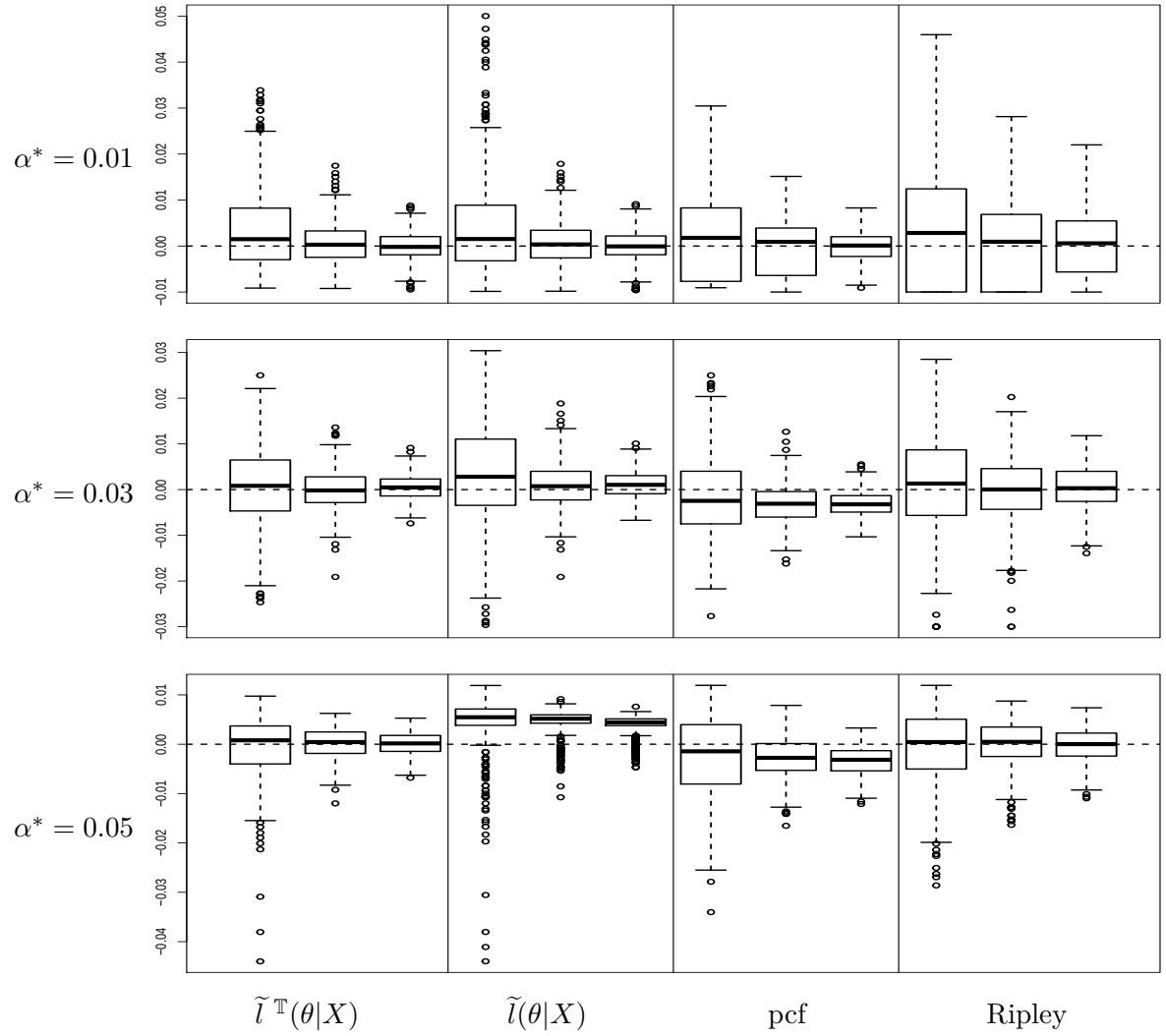


Figure 3: Boxplots of $\hat{\alpha} - \alpha^*$ generated from 500 simulations of Gaussian-type DPPs with true parameters $\rho^* = 100$ and, from top to bottom, $\alpha^* = 0.01$, 0.03 and 0.05. Each row shows the behaviour of the following 4 estimators when the simulation window is, from left to right in each box, $W = [0,1]^2$, $[0,2]^2$ and $[0,3]^2$: the approximate MLE with edge-corrections based on $\widetilde{l}^{\,\mathbb{T}}(\hat{\rho}, \alpha|X)$, the approximate MLE based on $\widetilde{l}(\hat{\rho}, \alpha|X)$, the MCE based on the pair correlation function and the MCE based on the Ripley's K function.

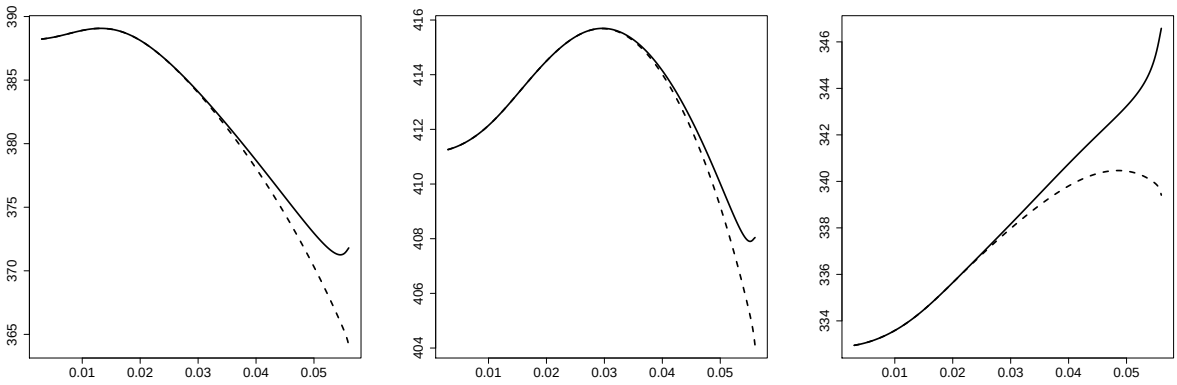


Figure 4: Comparison between $\widetilde{l}(100, \alpha|X)$ (solid lines) and $\widetilde{l}^{\,\mathbb{T}}(100, \alpha|X)$ (dashed lines) with respect to α where X is one realization of a DPP on $[0,1]^2$ with a Gaussian-type kernel with true parameters $\rho^* = 100$ and, from left to right, $\alpha^* = 0.01$, 0.03 and 0.05.

Concerning the computation time, even if our MLE approximation is much faster than the Fourier approximation (3.2), it can be heavy due to the need to optimize a function defined as the log-determinant of an $n \times n$ matrix, where n is the number of observed points. For comparison, each MCE took less than one second on a regular laptop in each case considered in Figure 3, while each computation of the approximate MLE took about a second when $W = [0,1]^2$, about 20 seconds when $W = [0,2]^2$ and about 100 seconds when $W = [0,3]^2$.

5.2 Performance for Bessel-type DPPs

In order to evaluate the possible limitations of our approach, we consider in this section the estimation of Bessel-type DPPs, see Table 1, whose kernels do not satisfy the theoretical assumptions in Section 3.4. As in the previous section, we set $\rho^* = 100$, $\alpha^* = 0.01, 0.03, 0.05$, corresponding to weak, medium and strong repulsiveness, and the observation window is $[0,1]^2$, $[0,2]^2$ and $[0,3]^2$. The results on 500 runs in each situation are shown in Figure 5 and in Table 4. They compare our edge-correction approximate MLE, the Fourier series approximation (3.2), the MCE based on the pair correlation function and the MCE based on the Ripley's K function. The performances are globally in line with the observations made in the previous section, showing that the approximate MLE outperforms MCEs, especially when the observation windows is large enough. Note that we have added the Fourier series approximation for comparison, because contrary to the models considered in the previous section, its behavior slightly differs from our edge-correction approximation for Bessel-type DPPs, as discussed in the following.

Despite the decent results of our approximation for Bessel-type DPPs, some issues appear with this model in the most repulsive case $\alpha^* = 0.05$. As noticed in Section 3.3, the determinant in (3.10) may be negative for high values of α, making the computation of the approximate likelihood impossible. This problem is illustrated in the rightmost plot of Figure 6, that shows an example of an approximated likelihood function as in (3.10) from one realization of a Bessel-type DPP on $W = [0,3]^2$ with $\rho^* = 100$ and $\alpha^* = 0.05$. The

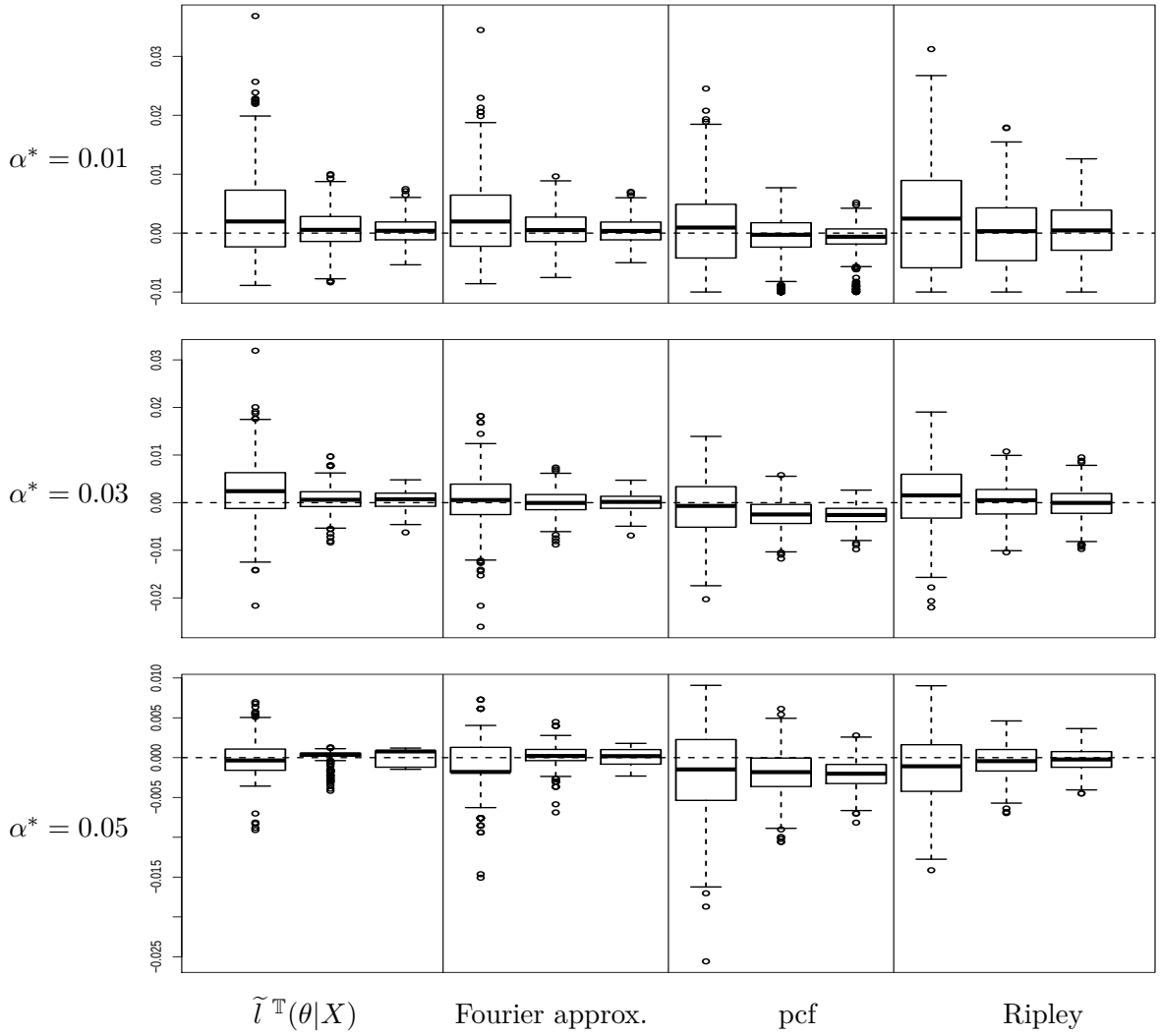


Figure 5: Boxplots of $\hat{\alpha} - \alpha^*$ generated from 500 simulations of Bessel-type DPPs with true parameters $\rho^* = 100$ and, from top to bottom, $\alpha^* = 0.01$, 0.03 and 0.05. Each row shows the behaviour of the following 4 estimators when the simulation window is, from left to right in each box, $W = [0,1]^2$, $[0,2]^2$ and $[0,3]^2$: the approximate MLE with edge-corrections based on $\widetilde{l}^{\,\mathbb{T}}(\hat{\rho}, \alpha|X)$, the Fourier series approximate MLE (3.2), the MCE based on the pair correlation function and the MCE based on the Ripley's K function.

Window	$[0,1]^2$			$[0,2]^2$			$[0,3]^2$		
α^*	0.01	0.03	0.05	0.01	0.03	0.05	0.01	0.03	0.05
MLE based on $\widetilde{l}^{\,\mathbb{T}}$	0.56	0.49	**0.04**	0.12	0.08	**0.01**	**0.05**	**0.03**	**0.01**
Fourier approx. MLE	**0.47**	**0.32**	0.09	**0.11**	**0.06**	0.02	**0.05**	**0.03**	**0.01**
MCE (pcf)	0.50	0.39	0.33	0.21	0.14	0.11	0.10	0.11	0.07
MCE (K)	0.95	0.46	0.19	0.41	0.15	0.04	0.27	0.10	0.02

Table 4: Estimated mean square errors (x10^4) of $\hat{\alpha}$ for Bessel-type DPPs on different windows and with different values of α, each computed from 500 simulations.

cross-type points on the right of this plot indicate the values of α where the determinant was negative. More generally, for the highest values of α, the approximate likelihood is clearly not trustable. Fortunately, the optimization procedure was not affected by this phenomena and succeeded to return a local maximum in the vicinity of α^*. However, another peculiar behaviour occurs in this situation, which is the small M-shape of the approximate likelihood in this vicinity. This feature was common to most of the approximate likelihoods in our simulations on $W = [0,3]^2$ with $\rho^* = 100$ and $\alpha^* = 0.05$, but we are not able to provide a clear explanation of this phenomena. The consequence is that the optimizer chooses one of the two local maxima from this M-shape, resulting in a bi-modal distribution of $\hat{\alpha}$ in this case, as showed in the leftmost plot of Figure 6. This also explains the shape of the boxplot associated to this case in Figure 5. In front of such peculiar M-shape of the contrast function, it might be natural to choose as the optimum the average of the two local maxima instead of one of them. Adopting this strategy decreases the estimation mean square error from 1 to 0.25 (x10^{-6}).

It is interesting to note that for Bessel-type DPPs, unlike the DPP models of Section 5.1, the Fourier series approximation (3.2) of the MLE has a more significative difference of behaviour than our approximate MLE with edge correction (3.10). As shown in Figure 6, it does not have undefined values and it does not follow a chaotic behavior for large values of α. Moreover, because the Fourier transform of the Bessel kernel only takes two different values (see Table 1), the terms in the Fourier approximation (3.2) when $d = 2$ simplify as:

$$\sum_{\substack{k\in\mathbb{Z}^2\\ \|k\|<N}} \log(1-\hat{K}_0^\theta(k)) = \log\left(1-\rho\pi\alpha^2\right) \sum_{k\in\{-N,\cdots,N\}} \left(2\left\lfloor\sqrt{\frac{1}{\pi^2\alpha^2}-k^2}\right\rfloor+1\right), \text{ and}$$

$$L_{app}^\theta(x,y) = \sum_{k\in\{-N,\cdots,N\}} \frac{\cos(2\pi kx)\sin\left(\pi y\left(2\left\lfloor\sqrt{\frac{1}{\pi^2\alpha^2}-k^2}\right\rfloor+1\right)\right)}{\sin(\pi y)},$$

where the truncation constant is $N = \left\lfloor\frac{1}{\pi\alpha}\right\rfloor$. This simplification makes it easier to compute than in the general case, and results in a more competitive computation time, similar to our approximation (3.10). As a result, we observe in Table 4 and Figure 5 that for $\alpha^* =$

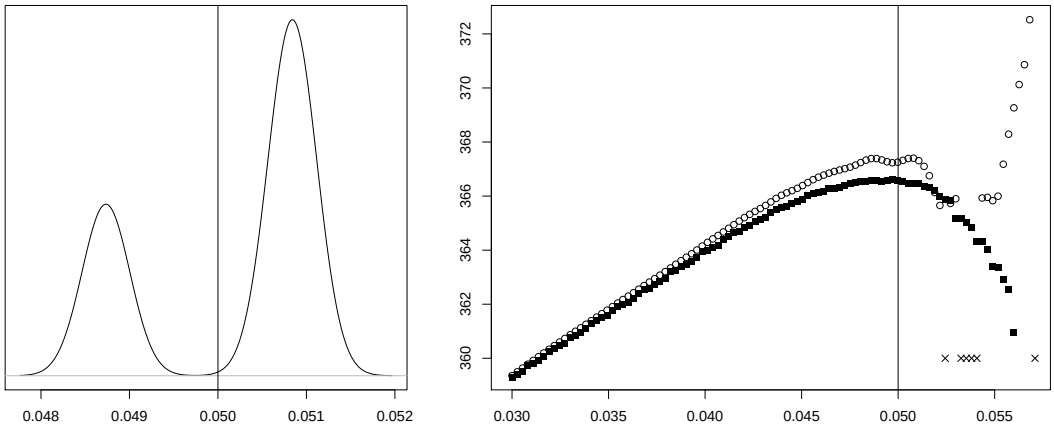


Figure 6: Left: distribution of $\hat{\alpha}$ obtained by the approximate MLE $\widetilde{l}^{\,\mathbb{T}}$, based on 500 simulations when $\rho^* = 100$, $\alpha^* = 0.05$ (represented by the vertical line) and $W = [0,3]^2$. Right: $\alpha \mapsto \widetilde{l}^{\,\mathbb{T}}(\hat{\rho}, \alpha | X)$ (circles) and Fourier series approximation (3.2) of the log-likelihood (black squares) from one realization X as before, where the vertical line shows the true parameter $\alpha^* = 0.05$ and the cross-type points indicate the values of α for which the determinant in $\widetilde{l}^{\,\mathbb{T}}(\hat{\rho}, \alpha | X)$ was negative.

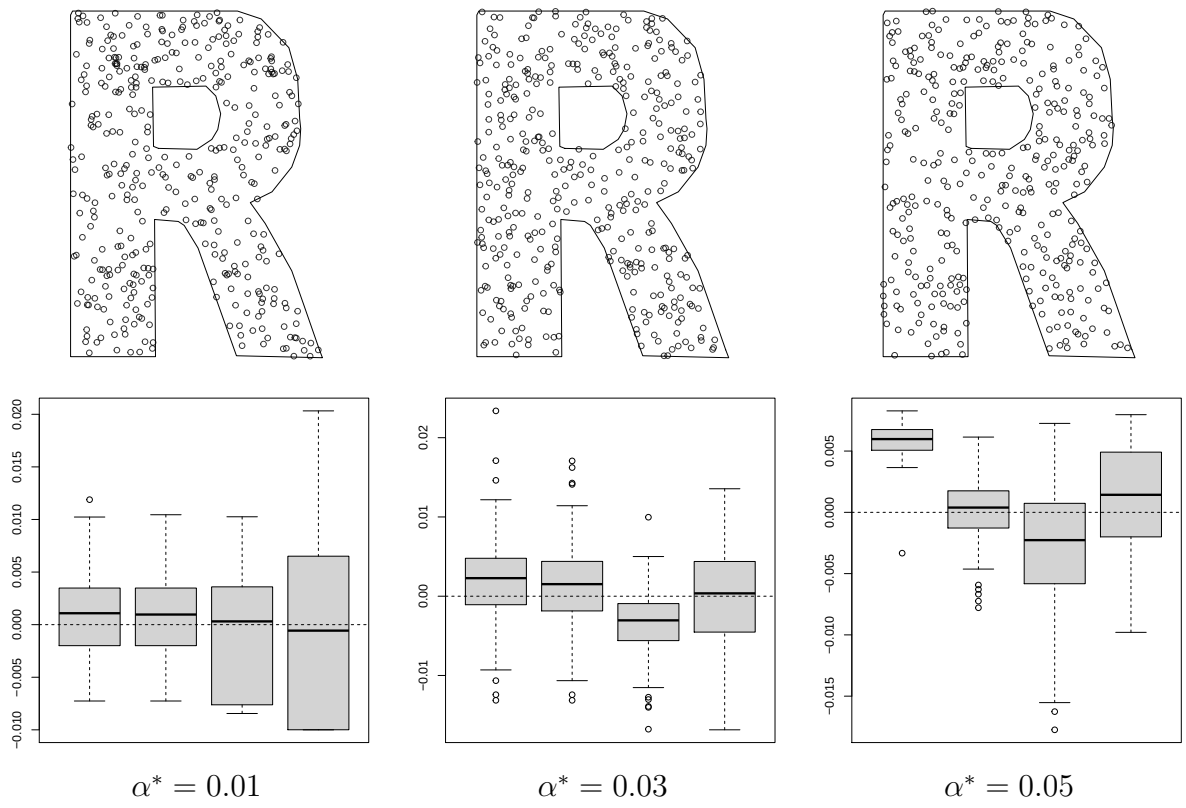
0.01 and 0.03, the Fourier approximation method has very similar performances than our approximation (3.10). When $\alpha^* = 0.05$, the Fourier approximation estimator has also a similar quadratic error, but the distribution of the estimator is more regular, for the reasons noticed above.

Finally, despite the fact that Bessel-type DPPs are not covered by our theory and the peculiar behaviour of $\widetilde{l}^{\,\mathbb{T}}$ for some values of α as described above, our approach still remains competitive in this case and outperforms standard MCE methods. Nevertheless, because the Fourier approximation (3.2) simplifies nicely in this setting and does not show the same chaotic behaviour as (3.10) for large values of α, it seems to be a slightly better choice for Bessel-type DPPs. However, we recall that this approach is limited to rectangular observation windows only.

5.3 Simulations on a non-rectangular window

We consider in this section the estimation of a Gaussian-type DPP on the (non-rectangular) R-shape window as in the simulations of Figure 7. The underlying parameters are $\rho^* = 100$, resulting in 370 points on average, and $\alpha^* = 0.01, 0.03$ and 0.05. The estimation of α^* is carried out by the MLE approximation (3.9) (without edge-corrections), the edge-corrected version described below, and the MCEs based on the pcf and the Ripley's K-function. Note that in this situation, the Fourier approximation (3.2) is not feasible.

We handle the edge-effects for this non-rectangular window in the following way. Note that the periodic edge-correction presented in Section 3.3 amounts to replace some zero-values of the matrix $L_0^\theta[X \cap W]$ by non negligible values. If we assume that the function L_0^θ



$\alpha^* = 0.01$ $\alpha^* = 0.03$ $\alpha^* = 0.05$

Figure 7: Top: Examples of realizations of Gaussian-type DPPs with parameters $\rho^* = 100$ and $\alpha^* = 0.01, 0.03, 0.05$ (from left to right) on a R-shape window. Bottom: distribution of $\hat{\alpha}$ from 100 simulations for each value of α^* and for the following estimators (from left to right in each plot): the approximate MLE (3.9), its edge-corrected version as detailed in the text, and the MCEs based on the pcf and the Ripley's K function.

is isotropic, as for the Gaussian-type DPPs considered in this section, then $L_0^\theta[X \cap W] = L_{\mathrm{rad}}^\theta(R)$ where R is the pairwise distance matrix of X, i.e. $R = (r_{ij} := \|x_i - x_j\|)_{1 \leqslant i,j \leqslant n}$ if $X = \{x_1, \dots, x_n\}$. Precisely, the replacements concern the entries involving a point x_i close to the border of the window and they amount to replace some of the largest distances r_{ij} by smaller ones. The idea is that for these points, we need to artificially increase the number of closed neighbours to account for edge-effects. Adopting this idea, we replace some entries of the matrix R as follows:

(i) We start by setting a maximal range of interaction $r_{\max}$. In our example we choose

$$r_{\max} = \arg\max_{r_{ij}} \{L_{\mathrm{rad}}^+(r_{ij}) > 0.001 L_{\mathrm{rad}}^+(0)\},$$

where $L_{\mathrm{rad}}^+ = L_{\mathrm{rad}}^\theta$ for $\theta = (\hat{\rho}, \alpha_{\max})$, $\hat{\rho} = N(W)/|W|$, $\alpha_{\max} = \sqrt{1/(\pi\hat{\rho})}$ and $L_{\mathrm{rad}}^+(0)$ is the maximal possible value of L_{rad}^+. This choice guarantees that for any $r > r_{\max}$ and any admissible $\theta = (\hat{\rho}, \alpha)$, $L_{\mathrm{rad}}^\theta(r)$ can be considered to be negligible.

(ii) For $i = 1, \dots, n$, we denote by d_i the Euclidean distance from x_i to ∂W, and by $n_i = \mathrm{card}\{j,\, r_{ij} < r_{\max}\}$ the number of neighbours of x_i in X. We further denote by $\mathcal{B} = \{x_i \in X,\, d_i < r_{\max}\}$ the set of "border" points of X in W and by $\bar{\mathcal{B}} = X \setminus \mathcal{B}$ the set of "interior" points of X in W. Finally, we consider $\mathcal{R}_{\bar{\mathcal{B}}} = \{r_{ij},\, x_i \in \bar{\mathcal{B}}\}$ the set of observed pairwise distances for the interior points of X, and $\mathcal{N}_{\bar{\mathcal{B}}} = \{n_i, x_i \in \bar{\mathcal{B}}\}$ the set of numbers of neighbours of the interior points.

(iii) For all $x_i \in \mathcal{B}$, we randomly pick out $\widetilde{n}_i$ in $\mathcal{N}_{\bar{\mathcal{B}}}$ and compare it to n_i. If $n_i \geqslant \widetilde{n}_i$, we do nothing. Else, for $j = (i+1), \dots, (i + \widetilde{n}_i - n_i) \wedge n$ and if $r_{ij} > r_{\max}$, we randomly pick out $\widetilde{r}_{ij}$ in $\mathcal{R}_{\bar{\mathcal{B}}} \cap \{r_{ij} > d_i\}$ and we replace r_{ij} and r_{ji} by $\widetilde{r}_{ij}$.

Note that the number of replacements in this edge-correction procedure is limited: they only concern the border points of X, there are a maximum of $\widetilde{n}_i - n_i$ of them for each border point x_i, and the replaced value $\widetilde{r}_{ij}$ of $r_{ij} > r_{\max}$ is necessarily greater than d_i, which in many cases (especially if α is small) entails $L_{\mathrm{rad}}^\theta(\widetilde{r}_{ij}) \approx 0$ and does not affect the initial value $L_{\mathrm{rad}}^\theta(r_{ij}) \approx 0$. With the resulting new matrix R, there is not guaranty that $L_{\mathrm{rad}}^\theta(R)$ is positive, a common issue with the periodic edge corrections of Section 3.3, but the restricted number of replacements limits the risk to encounter such a problem. In our experience, this happened only for very high values of α and did not affect the optimisation procedure.

The results displayed in Figure 7 show that the above edge-correction version of (3.9) provides the best results and clearly outperforms the MCE methods. They also confirm that this edge-correction is only necessary for the most repulsive DPPs, i.e. $\alpha^* = 0.05$ here, otherwise the approximation (3.9) and its edge-corrected version perform just as well.

6 Conclusion

In this paper, we have introduced an asymptotic approximation (3.9) of the log-likelihood of stationary determinantal point processes on $\mathbb{R}^d$ and $\mathbb{Z}^d$. While the true likelihood is

not numerically tractable, this approximation can be computed for stationary parametric families of DPPs based on correlation functions with a known Fourier transform, as the classical ones presented in Table 1. Compared to the Fourier approximation of [19] that only works for rectangular windows, our approximation can be computed for windows of any shape. However, due to edge effects, the resulting maximum likelihood estimators gets heavily biased for strongly repulsive DPPs, as shown in Figure 3. We have proposed to use the periodic correction (3.10) to fix this issue in the case of rectangular windows and showed that the resulting approximation is very close to the one in [19] (see Figure 1) but overall easier to compute. The idea to use a periodic correction has been detailed for rectangular windows, but a similar idea can be applied for a window with a different shape, as exemplified in Section 5.3. We showed in the simulation study of Section 5 that for standard parametric families of DPPs, the resulting approximate MLE outperforms classical moment methods based on the pair correlation function and Ripley's K function.

Finally, we proved in Propositions 3.1 and 3.2 that the difference between the true log-likelihood and our approximated log-likelihood converges almost surely towards 0 for classical parametric families of stationary DPPs on $\mathbb{Z}^d$. We also showed in Proposition 3.4 that DPPs on $\mathbb{R}^d$ can be arbitrarily approached by DPPs on a regular grid, which suggests that our approximation should also converge for DPPs on $\mathbb{R}^d$. A formal proof of such result is still a seemingly difficult open problem. Beyond the approximation of the likelihood, as proposed in this paper, a natural theoretical concern is the consistency of the maximum likelihood estimator, either based on the true likelihood or on the approximated one. This question is challenging and is not addressed in the present contribution. We however think that our findings are a step in the right direction towards such a result, because they allow to replace the true likelihood by an easier expression to deal with mathematically.

7 Proofs of Section 3

7.1 Proof of Proposition 3.1

In the case where $\mathcal{X} = \mathbb{R}^d$ and W_n is of the form $n \times W$ for some compact set W, then this proposition corresponds to [26, Proposition 5.9] with $\alpha = -1$ and $f(x) = \infty \times \mathbb{1}_W$. Our proof follows a similar idea.

Since all eigenvalues of $\mathcal{K}_{W_n}$ are in $[0, 1[$ then the logarithm of the Fredholm determinant of $\mathcal{I}_{W_n} - \mathcal{K}_{W_n}$ can be expanded into

$$\operatorname{logdet}(\mathcal{I}_{W_n} - \mathcal{K}_{W_n}) = -\sum_{k \geqslant 1} \frac{\operatorname{Tr}((\mathcal{K}_{W_n})^k)}{k} = -\sum_{k \geqslant 1} \frac{1}{k} \int_{W_n^k} K_0(x_2 - x_1) \cdots K_0(x_1 - x_k) \mathrm{d}\nu^k(x).$$

Now, we first assume that K_0 and $\hat{K}_0$ are integrable. Then, for any $x_1 \in \mathcal{X}$ the function

$$(x_2, \cdots, x_k) \mapsto K_0(x_2 - x_1) \cdots K_0(x_1 - x_k)$$

is integrable and its integral is equal to $K_0^{*k}(0)$ where K_0^{*k} is the k-th times self-convolution of K_0. Since we assumed that $(W_n)_{n\geqslant 0}$ satisfy (3.13), then by Lemma A.2 we get that

$$\frac{1}{|W_n|}\int_{W_n^k} K_0(x_2-x_1)\cdots K_0(x_1-x_k)\mathrm{d}\nu^k(x) \underset{n\to\infty}{\longrightarrow} (K_0)^{*k}(0).$$

Moreover, since

$$\frac{|\mathrm{Tr}((\mathcal{K}_{W_n})^k)|}{k|W_n|} \leqslant \frac{\|\mathcal{K}\|^{k-1}}{k}\times\frac{\mathrm{Tr}(\mathcal{K}_{W_n})}{|W_n|} = \frac{\|\mathcal{K}\|^{k-1}}{k}\times K_0(0)$$

which is summable with respect to k and does not depend on n, then we can conclude by the dominated convergence theorem that

$$\frac{1}{|W_n|}\mathrm{logdet}(\mathcal{I}_{W_n}-\mathcal{K}_{W_n}) \underset{n\to\infty}{\longrightarrow} -\sum_{k\geqslant 1}\frac{K_0^{*k}(0)}{k} = \int_{\mathcal{X}^*}\log(1-\hat{K}_0(x))\mathrm{d}x.$$

In the general case, we can always find a sequence K_m of integrable functions such that $\|\hat{K}_m-\hat{K}_0\|_{L^1(\mathcal{X}^*)} \underset{m\to\infty}{\longrightarrow} 0$ and conclude in the same way as in the end of the proof of [26, Proposition 5.9] using [26, Lemma 5.11]. Strictly speaking, to fit our setting, one has to replace in this Lemma $\mathbb{R}^d$ by $\mathcal{X}$ and their function ϕ_N by $\mathbb{1}_{W_n}$. The result of [26, Lemma 5.11] in this adapted setting remains true, the proof being exactly the same.

7.2 Proof of Proposition 3.2

For this proof, we consider X to be a DPP on $\mathbb{Z}^d$. Let $\theta\in\Theta$, we denote by λ_m^θ the lowest eigenvalue of $K^\theta[\mathbb{Z}^d]$. It is important to note that $\lambda_m^\theta>0$ as a consequence of [3, Theorem 4] and the assumptions on K^θ. We begin by proving the following lemma allowing us to control $L^\theta[X\cap W]-L^\theta_{[W]}[X\cap W]$ for any $W\subset\mathbb{Z}^d$ by controlling the difference between their associated operators.

Lemma 7.1. *Let $\mathcal{K}$ be an integral operator on $L^2(\mathcal{X},\nu)$ with kernel K such that $\|\mathcal{K}\|<1$. For any Borel set $W\subset\mathcal{X}$, we denote by $\mathcal{P}_W$ the projection on $L^2(W)$ and we define the operators $\mathcal{K}_W:=\mathcal{P}_W\mathcal{K}\mathcal{P}_W$ on $L^2(W)$, $\mathcal{L}_{[W]}:=\mathcal{K}_W(\mathcal{I}_W-\mathcal{K}_W)^{-1}$ on $L^2(W)$ and $\mathcal{L}:=\mathcal{K}(\mathcal{I}-\mathcal{K})^{-1}$ on $L^2(\mathcal{X})$. We denote by L the kernel of $\mathcal{L}$ and finally we define the operator $\mathcal{N}_W:=\mathcal{P}_W\mathcal{L}\mathcal{P}_{W^C}\mathcal{L}\mathcal{P}_W$ on $L^2(W)$ with kernel*

$$N_W(x,y)=\int_{W^C}L(x,z)L(z,y)\mathrm{d}\nu(z)\quad\forall x,y\in W. \tag{7.1}$$

Then,

$$0\leqslant\mathcal{P}_W\mathcal{L}\mathcal{P}_W-\mathcal{L}_{[W]}\leqslant\mathcal{N}_W.$$

Proof. We consider the following decomposition of the linear operators $\mathcal{I}-\mathcal{K}$ and $(\mathcal{I}-\mathcal{K})^{-1}$ on $L^2(W)\oplus L^2(W^C)$:

$$\mathcal{I}-\mathcal{K}=\begin{pmatrix}\mathcal{I}_W-\mathcal{K}_W & -\mathcal{P}_W\mathcal{K}\mathcal{P}_{W^C}\\ -\mathcal{P}_{W^C}\mathcal{K}\mathcal{P}_W & \mathcal{I}_{W^C}-\mathcal{K}_{W^C}\end{pmatrix}=\begin{pmatrix}(\mathcal{L}_{[W]}+\mathcal{I}_W)^{-1} & -\mathcal{P}_W\mathcal{K}\mathcal{P}_{W^C}\\ -\mathcal{P}_{W^C}\mathcal{K}\mathcal{P}_W & (\mathcal{L}_{[W^C]}+\mathcal{I}_{W^C})^{-1}\end{pmatrix}$$

and

$$(\mathcal{I}-\mathcal{K})^{-1}=\mathcal{I}+\mathcal{L}=\begin{pmatrix}\mathcal{P}_W\mathcal{L}\mathcal{P}_W+\mathcal{I}_W & \mathcal{P}_W\mathcal{L}\mathcal{P}_{W^C}\\ \mathcal{P}_{W^C}\mathcal{L}\mathcal{P}_W & \mathcal{P}_{W^C}\mathcal{L}\mathcal{P}_{W^C}+\mathcal{I}_{W^C}\end{pmatrix}.$$

A well-known result is that the $(1,1)$ block of $\mathcal{I}-\mathcal{K}$ is equal to the inverse of the Schur complement of $(\mathcal{I}-\mathcal{K})^{-1}$ relative to its $(2,2)$ block. This property is proved for 2×2 block matrices in [24, Theorem 1.2], and since the proof does not use any finite dimensionality argument, it works all the same for nonsingular operators on a Hilbert space, see [14] for example. As a consequence, we get

$$(\mathcal{L}_{[W]}+\mathcal{I}_W)^{-1}=(\mathcal{P}_W\mathcal{L}\mathcal{P}_W+\mathcal{I}_W-\mathcal{P}_W\mathcal{L}\mathcal{P}_{W^C}(\mathcal{P}_{W^C}\mathcal{L}\mathcal{P}_{W^C}+\mathcal{I}_{W^C})^{-1}\mathcal{P}_{W^C}\mathcal{L}\mathcal{P}_W)^{-1}$$

hence

$$\mathcal{P}_W\mathcal{L}\mathcal{P}_W-\mathcal{L}_{[W]}=\mathcal{P}_W\mathcal{L}\mathcal{P}_{W^C}(\mathcal{P}_{W^C}\mathcal{L}\mathcal{P}_{W^C}+\mathcal{I}_{W^C})^{-1}\mathcal{P}_{W^C}\mathcal{L}\mathcal{P}_W\geqslant 0.$$

Finally, since $(\mathcal{P}_{W^C}\mathcal{L}\mathcal{P}_{W^C}+\mathcal{I}_{W^C})^{-1}\leqslant\mathcal{I}_{W^C}$ this concludes the lemma. □

Now, we rewrite

$$\frac{1}{|W_n|}\Big|\mathrm{logdet}(L^\theta[X\cap W_n])-\mathrm{logdet}(L^\theta_{[W_n]}[X\cap W_n])\Big|$$

as

$$\frac{1}{|W_n|}\Big|\mathrm{logdet}\Big(Id+(L^\theta[X\cap W_n]-L^\theta_{[W_n]}[X\cap W_n])L^\theta_{[W_n]}[X\cap W_n]^{-1}\Big)\Big|.$$

By Lemma 7.1, we know that

$$0\leqslant L^\theta[X\cap W_n]-L^\theta_{[W_n]}[X\cap W_n]\leqslant N^\theta_{W_n}[X\cap W_n]$$

where $N^\theta_{W_n}$ is defined as in (7.1). Therefore, using Lemma A.1 we obtain the bound

$$0\leqslant\mathrm{logdet}(L^\theta[X\cap W_n])-\mathrm{logdet}(L^\theta_{[W_n]}[X\cap W_n])\leqslant\mathrm{Tr}(N^\theta_{W_n}[X\cap W_n]L^\theta_{[W_n]}[X\cap W_n]^{-1}).$$

Now, since $\mathcal{L}^\theta_{[W_n]}\geqslant\mathcal{K}^\theta_{W_n}$ by definition, then $\lambda_{min}(L^\theta_{[W_n]}[X\cap W_n])\geqslant\lambda_{min}(K^\theta[X\cap W_n])\geqslant\lambda^\theta_m$ where the last inequality is a consequence of $K^\theta[X\cap W_n]$ being a sub-matrix of $K^\theta[\mathbb{Z}^d]$. Therefore,

$$\mathrm{Tr}(N^\theta_{W_n}[X\cap W_n]L^\theta_{[W_n]}[X\cap W_n]^{-1})\leqslant(\lambda^\theta_m)^{-1}\mathrm{Tr}(N^\theta_{W_n}[X\cap W_n])=(\lambda^\theta_m)^{-1}\sum_{x\in X\cap W_n}N^\theta_{W_n}(x,x).$$

The function $X\mapsto|W_n|^{-1}\sum_{x\in X}N^\theta_{W_n}(x,x)$ is $\|N^\theta_{W_n}\|_\infty/|W_n|$-Lipschitz on $\bigcup_{k\geqslant 0}W_n^k$ with

$$\|N^\theta_{W_n}\|_\infty\leqslant\|L^\theta_0\|_2^2=\left\|\frac{\hat{K}^\theta_0}{1-\hat{K}^\theta_0}\right\|_2^2\leqslant\frac{\|\hat{K}^\theta_0\|_2^2}{1-M_\theta}=\frac{\|K^\theta_0\|_2^2}{1-M_\theta}<\infty.$$

By [23, Theorem 3.5], we get for all $a \in \mathbb{R}_+$

$$\mathbb{P}_{\theta^*}\left(\frac{1}{|W_n|}\left|\sum_{x\in X\cap W_n} N^\theta_{W_n}(x,x) - \mathbb{E}_{\theta^*}\left[\sum_{x\in X\cap W_n} N^\theta_{W_n}(x,x)\right]\right| > a\right) \leqslant 5\exp\left(-\frac{a^2|W_n|^2/\|N^\theta_{W_n}\|^2_\infty}{16(a|W_n|/\|N^\theta_{W_n}\|_\infty + 2\mathbb{E}_{\theta^*}[N(W_n)])}\right) \quad (7.2)$$

where $\mathbb{E}_{\theta^*}[N(W_n)] = |W_n|K_0^{\theta^*}(0)$ and

$$\begin{aligned}
&\frac{1}{|W_n|}\mathbb{E}_{\theta^*}\left[\sum_{x\in X\cap W_n} N^\theta_{W_n}(x,x)\right] \\
=&\frac{K_0^{\theta^*}(0)}{|W_n|}\int_{W_n}\int_{W_n^C} L_0^\theta(y-x)^2 \mathrm{d}\nu(x)\mathrm{d}\nu(y) \\
=&K_0^{\theta^*}(0)\left(\int_{\mathbb{Z}^d} L_0^\theta(y)^2\mathrm{d}\nu(y) - \frac{1}{|W_n|}\int_{W_n^2} L_0^\theta(y-x)^2\mathrm{d}\nu(x)\mathrm{d}\nu(y)\right).
\end{aligned}$$

But, as a consequence of Lemma A.2, we have

$$\frac{1}{|W_n|}\int_{W_n^2} L_0^\theta(y-x)^2\mathrm{d}\nu(x)\mathrm{d}\nu(y) \underset{n\to\infty}{\longrightarrow} \int_{\mathbb{Z}^d} L_0^\theta(y)^2\mathrm{d}\nu(y)$$

hence

$$\frac{1}{|W_n|}\mathbb{E}_{\theta^*}\left[\sum_{x\in X} N^\theta_{W_n}(x,x)\right] \underset{n\to\infty}{\longrightarrow} 0.$$

Finally, by (7.2) and the inequality $\|N^\theta_{W_n}\|_\infty \leqslant \|L_0^\theta\|_2^2$, we get that for all $a \in \mathbb{R}_+$,

$$\mathbb{P}_{\theta^*}\left(\frac{1}{|W_n|}\left|\sum_{x\in X\cap W_n} N^\theta_{W_n}(x,x)\right| > a\right) = O\left(\exp\left(-\frac{a^2|W_n|}{16\|L_0^\theta\|_2^2(a+2K_0^{\theta^*}(0)\|L_0^\theta\|_2^2)}\right)\right).$$

Since we assumed (3.14), then by the Borel–Cantelli Lemma,

$$\frac{1}{|W_n|}\sum_{x\in X\cap W_n} N^\theta_{W_n}(x,x) \xrightarrow{a.s.} 0$$

and therefore

$$\frac{1}{|W_n|}\left|\mathrm{logdet}(L^\theta[X\cap W_n]) - \mathrm{logdet}(L^\theta_{[W_n]}[X\cap W_n])\right| \xrightarrow{a.s.} 0.$$

7.3 Proof of Proposition 3.4

First, we need to show that X_ε is a well defined DPP for small enough ε by showing that its kernel, the infinite matrix $\varepsilon^d K[\varepsilon\mathbb{Z}^d]$, is hermitian with eigenvalues in $[0,1]$. Everything is

trivial except for showing that the eigenvalues become lower or equal to 1 as ε vanishes. For every $v = (v_j)_{j\in\mathbb{Z}^d}$ such that $\sum_j |v_j|^2 = 1$, we define the function

$$\phi(t) = \sum_{j\in\mathbb{Z}^d} v_j e^{2i\pi<j,t>}$$

such that the integral of $|\phi|^2$ on any unit cube is equal to 1. Therefore, we can write

$$\begin{aligned}
\langle v, \varepsilon^d K[\varepsilon\mathbb{Z}^d]v\rangle &= \sum_{j,k\in\mathbb{Z}^d} \varepsilon^d v_j v_k K_0(\varepsilon(k-j)) \\
&= \sum_{j,k\in\mathbb{Z}^d} v_j v_k \int_{\mathbb{R}^d} \hat{K}_0(t/\varepsilon) e^{2i\pi\langle k-j,t\rangle}\mathrm{d}t \\
&= \int_{\mathbb{R}^d} \hat{K}_0(t/\varepsilon)|\phi(t)|^2\mathrm{d}t \\
&\leqslant \sum_{i\in\mathbb{Z}^d} \sup_{x\in C_i} \hat{K}_0(x/\varepsilon)
\end{aligned}$$

where C_i is the unit cube defined as $[i_1 - 1/2, i_1 + 1/2] \times \cdots \times [i_d - 1/2, i_d + 1/2]$ for all $i = (i_1, \cdots, i_d) \in \mathbb{Z}^d$. By our assumptions on $\hat{K}_0$, we have $\sup_{x\in C_0} \hat{K}_0(x/\varepsilon) \leqslant \|\hat{K}_0\|_\infty < 1$ and

$$\begin{aligned}
\sup_{x\in C_i} \hat{K}_0(x/\varepsilon) &\leqslant \sup_{\forall j,\ x_j\in[i_j-1/2,i_j+1/2]} \frac{A}{1+\varepsilon^{-(d+\tau)}(\sum_{j=1}^d x_j^2)^{(d+\tau)/2}} \\
&= \frac{A}{1+\varepsilon^{-(d+\tau)}(\sum_{1\leqslant j\leqslant d,\ i_j\neq 0}(|i_j|-1/2)^2)^{(d+\tau)/2}}
\end{aligned}$$

hence, the sum of all $\sup_{x\in C_i} \hat{K}_0(x/\varepsilon)$ for i of the form $(i_1, \cdots, i_k, 0, \cdots, 0)$ where $i_1, \cdots, i_k \in \mathbb{Z}\backslash\{0\}$ and $k \in \{1, \cdots, d\}$ is bounded by

$$\begin{aligned}
&\sum_{i_1,\cdots,i_k\in(\mathbb{Z}\backslash\{0\})^k} \frac{A}{1+\varepsilon^{-(d+\tau)}(\sum_{j=1}^k(|i_j|-1/2)^2)^{(d+\tau)/2}} \\
&\leqslant \varepsilon^{d+\tau} \sum_{i_1,\cdots,i_k\in(\mathbb{Z}\backslash\{0\})^k} \frac{A}{(\sum_{j=1}^k |i_j|^2)^{(d+\tau)/2}} \underset{\varepsilon\to 0}{\longrightarrow} 0.
\end{aligned}$$

By symmetry, this is also true for the sum of all $\sup_{x\in C_i} \hat{K}_0(x/\varepsilon)$ for i with any k non-zero components and $d-k$ zero components when $k \in \{1, \cdots, d\}$. This shows that

$$\sum_{\substack{i_1,\cdots,i_d\in\mathbb{Z}^d \\ i\neq(0,\cdots,0)}} \sup_{x\in C_i} \hat{K}_0(x/\varepsilon) \underset{\varepsilon\to 0}{\longrightarrow} 0$$

and therefore

$$0 \leqslant \sup_{v:\ \sum_j |v_j|^2=1} \langle v, \varepsilon^d K[\varepsilon\mathbb{Z}^d]v\rangle \leqslant 1$$

for small enough values of ε, and in this case the DPP X_ε is then well defined.

Now, we prove the weak convergence of the discrete DPPs to the continuous one by showing the pointwise convergence of their Laplace functionals (see [11, Proposition 11.1.VIII]). We recall that the Laplace functional of a point process Y is defined as

$$L_Y(f) := \mathbb{E}_Y\left[\exp\left(-\sum_{x\in Y} f(x)\right)\right]$$

for all non-negative continuous function f vanishing outside a bounded set. Let D be a compact set of $\mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$ be a continuous function vanishing outside D. We define the kernel

$$K_f := (x,y) \mapsto \sqrt{1-e^{-f(x)}}K_0(y-x)\sqrt{1-e^{-f(y)}}$$

and call $\mathcal{K}_f$ its associated integral operator. Then, the Laplace transform of the continuous DPP X writes (see [26])

$$L_X(f) = \det(I-\mathcal{K}_f) = \exp\left(-\sum_{n\geqslant 1}\frac{1}{n}\mathrm{Tr}(\mathcal{K}_f^n)\right)$$

and for all ε, the rescaled DPP $\varepsilon X_\varepsilon$ has the same distribution as a DPP on $\varepsilon\mathbb{Z}^d$ with kernel $\varepsilon^d K_0(y-x)$ hence its Laplace transform writes

$$\begin{aligned}
L_{\varepsilon X_\varepsilon}(f) &= \det\left(I-\varepsilon^d K_f[\varepsilon\mathbb{Z}^d]\right)\\
&= \exp\left(-\sum_{n\geqslant 1}\frac{\varepsilon^{dn}}{n}\mathrm{Tr}\left(K_f\left[D\cap\varepsilon\mathbb{Z}^d\right]^n\right)\right)\\
&= \exp\left(-\sum_{n\geqslant 1}\frac{1}{n}\left(\varepsilon^{dn}\sum_{x_1,\cdots,x_n\in D\cap\varepsilon\mathbb{Z}^d} K_f(x_1,x_2)\cdots K_f(x_{n-1},x_n)K_f(x_n,x_1)\right)\right).
\end{aligned}$$

For all $n\geqslant 1$, we have the convergence of the following Riemann sum on the compact sets D^n:

$$\begin{aligned}
&\varepsilon^{dn}\sum_{x_1,\cdots,x_n\in D\cap\varepsilon\mathbb{Z}^d} K_f(x_1,x_2)\cdots K_f(x_{n-1},x_n)K_f(x_n,x_1)\\
&\qquad\qquad\xrightarrow[\varepsilon\to 0]{}\int_{D^n} K_f(x_1,x_2)\cdots K_f(x_{n-1},x_n)K_f(x_n,x_1)\mathrm{d}x = \mathrm{Tr}(\mathcal{K}_f^n).
\end{aligned}$$

Moreover, we have

$$\mathrm{Tr}\left(\left(\varepsilon^d K_f\left[D\cap\varepsilon\mathbb{Z}^d\right]\right)^n\right) \leqslant \lambda_{\max}\left(\varepsilon^d K_f\left[D\cap\varepsilon\mathbb{Z}^d\right]\right)^{n-1}\mathrm{Tr}\left(\varepsilon^d K_f\left[D\cap\varepsilon\mathbb{Z}^d\right]\right)$$

and since $\mathcal{K}_f\leqslant\mathcal{K}$ then $\lambda_{\max}(\varepsilon^d K_f[D\cap\varepsilon\mathbb{Z}^d])\leqslant\lambda_{\max}(\varepsilon^d K[D\cap\varepsilon\mathbb{Z}^d])\leqslant\lambda_{\max}(\varepsilon^d K[\varepsilon\mathbb{Z}^d])$ which we showed was arbitrary close to $\|K_0\|_\infty < 1$ for small enough ε, then by the dominated convergence theorem we get that

$$L_{\varepsilon X_\varepsilon}(f)\xrightarrow[\varepsilon\to 0]{} L_X(f)$$

which proves the weak convergence of the distributions of $\varepsilon X_\varepsilon$ towards the distribution of X when ε goes towards 0.

A Technical Lemmas

Lemma A.1. *Let $n \in \mathbb{N}$ and A, B be two $n \times n$ positive semi-definite matrices. Then,*

$$0 \leqslant \text{logdet}(I + AB) \leqslant Tr(AB).$$

Proof. We first assume that B is the identity matrix. Let $\lambda_1, \cdots, \lambda_n$ be the eigenvalues (with multiplicity) of A. Then,

$$0 \leqslant \text{logdet}(I + A) = \sum_{i=1}^{n} \log(1 + \lambda_i) \leqslant \sum_{i=1}^{n} \lambda_i = \text{Tr}(A).$$

In the general case, Sylvester's determinant identity gives us

$$0 \leqslant \text{logdet}(I + AB) = \text{logdet}(I + A^{1/2}BA^{1/2}) \leqslant \text{Tr}(A^{1/2}BA^{1/2}) = \text{Tr}(AB).$$

□

Lemma A.2. *Let $f : \mathcal{X}^k \mapsto \mathbb{R}$ be a translation invariant function such that*

$$(x_2, \cdots, x_k) \mapsto f(0, x_2, \cdots, x_k) \in L^1\Big(\mathcal{X}^{k-1}, \nu^{k-1}\Big).$$

Let W_n be a sequence of increasing compact subsets of $\mathcal{X}$ such that there exists an increase non-negative sequence $r_n \in \mathbb{R}_+^{\mathbb{N}}$ satisfying $r_n \underset{n\to\infty}{\longrightarrow} \infty$ and

$$|(\partial W_n \oplus r_n) \cap W_n| = o(|W_n|), \tag{A.1}$$

then

$$\frac{1}{|W_n|} \int_{W_n^k} f(x) \mathrm{d}\nu^k(x) \underset{n\to\infty}{\longrightarrow} \int_{\mathcal{X}^{k-1}} f(0, x_2, \cdots, x_k) \mathrm{d}\nu(x_2) \cdots \mathrm{d}\nu(x_k). \tag{A.2}$$

Proof. We write $W_n \ominus r_n$ for the set $W_n \backslash (\partial W_n \oplus r_n)$ of points in W_n at distance at least r_n from the boundary of W_n. Since f is translation invariant then the right term in (A.2) is equal to

$$\frac{1}{|W_n|} \int_{W_n \times \mathcal{X}^{k-1}} f(x) \mathrm{d}\nu^k(x).$$

As a consequence,

$$\begin{aligned}
&\left|\int_{\mathcal{X}^{k-1}} f(0,x_2,\cdots,x_k)\mathrm{d}\nu(x_2)\cdots\mathrm{d}\nu(x_k) - \frac{1}{|W_n|}\int_{W_n^k} f(x)\mathrm{d}\nu^k(x)\right| \\
&= \frac{1}{|W_n|}\left|\int_{W_n\times(\mathcal{X}^{k-1}\setminus W_n^{k-1})} f(x)\mathrm{d}\nu^k(x)\right| \\
&= \frac{1}{|W_n|}\left|\int_{W_n\ominus r_n}\left(\int_{\mathcal{X}^{k-1}\setminus W_n^{k-1}} f(x)\mathrm{d}\nu(x_2)\cdots\mathrm{d}\nu(x_k)\right)\mathrm{d}\nu(x_1)\right. \\
&\qquad\qquad \left. + \frac{1}{|W_n|}\left(\int_{(\partial W_n\oplus r_n)\cap W_n}\int_{\mathcal{X}^{k-1}\setminus W_n^{k-1}} f(x)\mathrm{d}\nu(x_2)\cdots\mathrm{d}\nu(x_k)\right)\mathrm{d}\nu(x_1)\right| \\
&\leqslant \frac{1}{|W_n|}\int_{W_n\ominus r_n}\left(\int_{\mathcal{X}^{k-1}} |f(0,y)|\mathbb{1}_{\{\forall i,\ \|y_i\|>r_n\}}\mathrm{d}\nu^{k-1}(y)\right)\mathrm{d}\nu(x) \\
&\qquad\qquad + \frac{1}{|W_n|}\int_{(\partial W_n\oplus r_n)\cap W_n}\left(\int_{\mathcal{X}^{k-1}} |f(0,y)|\mathrm{d}\nu^{k-1}(y)\right)\mathrm{d}\nu(x) \\
&\leqslant \int_{(\mathcal{B}(0,r_n)^C)^{k-1}} |f(0,y)|\mathrm{d}\nu^{k-1}(y) + \frac{|(\partial W_n\oplus r_n)\cap W_n|}{|W_n|}\|f(0,.)\|_{L^1}
\end{aligned}$$

where $\mathcal{B}(0,r_n)^C$ is the complement of the euclidian ball centered at the origin with radius r_n. This expression thus converges to 0 because f is integrable with respect to its last $k-1$ variables and by (A.1). □

Proposition A.3. *Let X be a DPP with Bessel-type kernel $K_0^{\rho,\alpha}$, as defined in Table 1, observed on a window $W\subset\mathbb{R}^d$. Recall that $\rho_{\max}$, given in Table 1, is the upper bound of ρ for which X is well-defined. Then, for all $\alpha>0$ such that $N(W)/|W|\leqslant\rho_{\max}$,*

$$\left\{\frac{N(W)}{|W|}\right\} = \underset{0\leqslant\rho\leqslant\rho_{\max}}{\arg\max}\,\widetilde{l}(\rho,\alpha|X). \tag{A.3}$$

Proof. By noticing that $\rho_{\max}$ is the volume of the d-dimensional ball with radius $\sqrt{d/(2\pi^2\alpha^2)}$, we get from the expression of $\hat{K}_0^{\rho,\alpha}$ in Table 1 that

$$\int_{\mathbb{R}^d}\log(1-\hat{K}_0^{\rho,\alpha}(x))\mathrm{d}x = \rho_{\max}\log(1-\rho/\rho_{\max}).$$

Moreover, $L_0^{\rho,\alpha}$ can be written as $\rho F^\alpha/(1-\rho/\rho_{\max})$, where F^α is a function not depending on ρ (see Table 2). Therefore, $\mathrm{logdet}(L_0^{\rho,\alpha}[X\cap W])$ can be expressed as the sum of

$$N(W)\log\left(\frac{\rho}{1-\rho/\rho_{\max}}\right)$$

and an expression not depending on ρ. As a consequence, $\widetilde{l}(\rho,\alpha|X)$ is twice differentiable with respect to ρ with derivative

$$\frac{-1}{1-\rho/\rho_{\max}} + \frac{N(W)}{|W|\rho(1-\rho/\rho_{\max})}.$$

It is easy to see that this expression vanishes only when $\rho=N(W)/|W|$ with the second derivative being negative at this point, concluding the proof. □

References

[1] R.H. Affandi, E. Fox, R. Adams, and B. Taskar. Learning the parameters of determinantal point process kernels. In *International Conference on Machine Learning*, pages 1224–1232, 2014.

[2] K. E. Atkinson. *The Numerical Solution of Integral Equations of the Second Kind*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 1997.

[3] F. Bachoc and R. Furrer. On the smallest eigenvalues of covariance matrices of multivariate spatial processes. *Stat*, 5:102–107, 2016.

[4] A. J. Baddeley, E. Rubak, and R. Turner. *Spatial Point Patterns: Methodology and Applications with R*. Interdisciplinary Statistics. Chapman & Hall/CRC, Boca Raton, Florida, 2015.

[5] R. Bardenet, J. Flamant, and P. Chainais. On the zeros of the spectrogram of white noise. *Appl. Comput. Harmon. Anal.*, 48(2):682–705, 2020.

[6] R. Bardenet and A. Hardy. Monte carlo with determinantal point processes. *Ann. Appl. Probab.*, 30(1):368–417, 02 2020.

[7] R. Bardenet and M. Titsias RC AUEB. Inference for determinantal point processes without spectral knowledge. In *Advances in Neural Information Processing Systems 28*, pages 3393–3401. Curran Associates, Inc., 2015.

[8] C.A.N. Biscio and F. Lavancier. Quantifying repulsiveness of determinantal point processes. *Bernoulli*, 22:2001–2028, 2016.

[9] C.A.N. Biscio and F. Lavancier. Contrast estimation for parametric stationary determinantal point processes. *Scandinavian Journal of Statistics*, 44:204–229, 2017.

[10] V.-E. Brunel, A. Moitra, P. Rigollet, and J. Urschel. Maximum likelihood estimation of determinantal point processes. *arXiv:1701.06501*, 2017. preprint.

[11] D.J. Daley and D. Vere-Jones. *An introduction to the theory of point processes, Volume II: General Theory and Structure*. Probability and Its Applications. Springer, 2nd ed edition, 2007.

[12] N. Deng, W. Zhou, and M. Haenggi. The ginibre point process as a model for wireless networks with repulsion. *IEEE Transactions on Wireless Communications*, 1:479–492, 2015.

[13] P. Diggle. *The Statistical Analysis of Spatial Point Patterns (2nd ed.)*. Hodder Arnold, London, 2003.

[14] T. Fujimoto, H. Hisamatsu, and R. Ranade. Schur complements in banach spaces. *Kagawa University economic review*, 77(2), Sep 2004.

[15] J. S. Gomez, A. Vasseur, A. Vergne, P. Martins, L. Decreusefond, and W. Chen. A case study on regularity in cellular network deployment. *IEEE Wireless Communications Letters*, 4(4):421–424, 2015.

[16] J.B. Hough, M. Krishnapur, Y. Peres, and B. Virag. *Zeros of Gaussian Analytic Functions and Determinantal Point Processes*. American Mathematical Society, 2009.

[17] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2-3):123–286, 2012.

[18] C. Launay, A. Desolneux, and B. Galerne. Determinantal point processes for image processing. *to appear in the SIAM Journal on Imaging Sciences*, 2021.

[19] F. Lavancier, J. Møller, and E. Rubak. Determinantal point process models and statistical inference. *Journal of Royal Statistical Society: Series B (Statistical Methodology)*, 77:853–877, 2015.

[20] F. Lavancier, A. Poinas, and R. Waagepetersen. Adaptive estimating function inference for nonstationary determinantal point processes. *Scandinavian Journal of Statistics*, 2020.

[21] O. Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7:83–122, 1975.

[22] N. Miyoshi and T. Shirai. A cellular network model with ginibre configured base stations. *Advances in Applied Probability*, 46:832–845, 2014.

[23] R. Pemantle and Y. Peres. Concentration of lipschitz functionals of determinantal and other strong rayleigh measures. *Combin. Probab. Comput.*, 23:140–160, 2014.

[24] S. Puntanen and F. Zhang. *The Schur Complement and Its Applications*. Numerical Methods and Algorithms 4. Springer US, 2005.

[25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[26] T. Shirai and Y. Takahashi. Random point fields associated with certain fredholm determinants i: fermion, poisson and boson point processes. *Journal of Functional Analysis*, 205:414–463, 2003.

[27] A. Soshnikov. Determinantal random point fields. *Russian Math. Surveys*, 55:923–975, 2000.