Determinantal point processes for image processing

Claire Launay, Agnès Desolneux, Bruno Galerne

Determinantal point processes for image processing

Claire Launay claire.launay@parisdescartes.fr
Laboratoire MAP5
Université de Paris
Paris, 75006, FRANCE

Bruno Galerne bruno.galerne@univ-orleans.fr
Institut Denis Poisson
Université d'Orléans, Université de Tours
Orléans, 45100, FRANCE

Agnès Desolneux agnes.desolneux@math.cnrs.fr
CNRS, Centre Borelli
ENS Paris Saclay
Cachan, 94230, FRANCE

Abstract

Determinantal point processes (DPPs) are probabilistic models of configurations that favour diversity or repulsion. They have recently gained influence in the machine learning community, mainly because of their ability to elegantly and efficiently subsample large sets of data. In this paper, we consider DPPs from an image processing perspective, meaning that the data we want to subsample are pixels or patches of a given image. To this end, our framework is discrete and finite. First, we adapt their basic definition and properties to DPPs defined on the pixels of an image, that we call determinantal pixel processes (DPixPs). We are mainly interested in the repulsion properties of such a process and we apply DPixPs to texture synthesis using shot noise models. Finally, we study DPPs on the set of patches of an image. Because of their repulsive property, DPPs provide a strong tool to subsample discrete distributions such as that of image patches.

Keywords: Determinantal point processes, repulsion, subsampling, image, pixels, patches, stationarity, shot noise, inference

1. Introduction

Determinantal point processes (DPPs) are models of random sets of points that are currently of great interest in several mathematical fields. They were initially studied in probability, in particular for the modeling of fermions in quantum mechanics [42] or as a process describing the spectrum of random matrices [55], and recently gained influence in the machine learning community [34], mainly due to the necessity to handle high amounts of information [59], [14] and often high-dimensional data.

The main feature of DPPs is that they provide a family of models of random configurations that favour diversity or repulsion, in the sense that the probability of observing two points close or similar to each other is lower than in the case of the Poisson process whose points are independent [8]. DPPs are completely determined by their correlation function. Unlike Gibbs' point processes, the moments of the DPPs are all known from their correlation function. That is the reason why these processes are an elegant model to reduce the dimensionality of data or to select representative samples from huge sets of points.

The amount of images and video content available is overwhelming. To be handled, to be processed, it needs to be sorted and summarized. That is the purpose of recommendation systems. Some methods using DPPs have been developed to cope with this issue and to enforce diverse subsets, for images selection [35], [14] or video recommendation [62]. Moreover, images and videos are now in very high resolution, but remain intrinsically redundant. The strategies for video summarization intend to extract meaningful and representative frames using sequential DPPs. This is a type of DPP taking into account the temporal dependencies of video frames [26], [43]. Besides, Chen et al. [15] prove that DPPs can be an appropriate tool to reduce the dimensionality of hyperspectral images, to select representative pixels from these images and be able to process such large-scale data.

In this paper, we will consider DPPs from an image processing perspective, meaning that the data we want to subsample are pixels or patches of a given image. To this end, our framework is discrete and finite. In this setting, the correlation function associated to a DPP is a matrix K that will be called its kernel. First, we need to introduce some notation. The initial data set is denoted by $\mathcal{Y}$ and the data items are assimilated with their index, meaning that for a set of N elements, we consider $\mathcal{Y} = \{1, \dots, N\}$. The cardinality of a set A is denoted by $|A|$. When M is a $N \times N$ matrix, we denote by $M_{A\times B}$, for all subsets $A, B \subset \mathcal{Y}$, the matrix $(M(x,y))_{(x,y)\in A\times B}$ and we use the short notation $M_A = M_{A\times A}$. When focusing on a specific couple of points, for instance $x, y \in \mathcal{Y}$, we sometimes identify $M(x,y)$ and M_{xy} for clarity purpose. If A and B are subsets of $\mathcal{Y}$ such that $|A| = |B|$, the determinant $\det(M_{A\times B})$ is called a minor of M and in case $B = A$, $\det(M_A)$ is called a principal minor of M.

Let us define DPPs in the general discrete case. Consider the set $\mathcal{Y} = \{1, \dots, N\}$ and a Hermitian matrix K of size $N \times N$ such that

$$0 \preceq K \preceq I,$$

meaning that the eigenvalues of K are in $[0, 1]$. Then a random set $X \subset \mathcal{Y}$ is called a determinantal point process with kernel K if it is defined by

$$\mathbb{P}(A \subset X) = \det(K_A), \quad \forall A \subset \mathcal{Y}. \tag{1}$$

We will denote $X \sim \mathrm{DPP}(K)$. For a detailed presentation of discrete DPPs, their properties and some applications to machine learning, we recommend the article of Kulesza and Taskar [34].

Notice that a DPP is simple: two points of the point process can't coincide. The diagonal coefficients of K define the marginal probabilities of any singleton:

$$\forall x \in \mathcal{Y}, \quad \mathbb{P}(x \in X) = K(x,x), \tag{2}$$

and the off-diagonal coefficients of K give the similarity between points. The repulsion property becomes clear when observing the marginal probabilities of pairs of points. The more similar two points are, the less likely they belong to the DPP simultaneously:

$$\forall \{x, y\} \subset \mathcal{Y}, \quad \mathbb{P}(\{x, y\} \subset X) = K(x,x)K(y,y) - |K(x,y)|^2. \tag{3}$$

Moreover, let $\lambda_1, \dots, \lambda_N$ be the eigenvalues of K.

Proposition 1 *The cardinality $|X|$ of the DPP X is distributed as the sum of N independent Bernoulli random variables: $|X| \sim \sum_{x\in\mathcal{Y}} \mathcal{B}\mathrm{er}(\lambda_x)$, where the Bernoulli variables take the value 1 with probability λ_x. Then $\mathbb{E}(|X|) = \sum_{x\in\mathcal{Y}} \lambda_x = \mathrm{Tr}(K)$ and $\mathrm{Var}(|X|) = \sum_{x\in\mathcal{Y}} \lambda_x(1-\lambda_x)$.*

In this paper, we will adapt these definitions and properties to particular DPPs, defined on the pixels of an image.

One of the key contributions of this paper is the definition and the study of this new kind of DPP over pixels, that we present in the following three sections and that we call determinantal pixel process (DPixP). The second section presents the setting adapted to the pixel of an image for the use of DPPs. In the third section, we apply DPixPs to texture synthesis through shot noise models and we present how to adapt a DPixP kernel to a given spot function. The fourth section investigates the estimation of DPixP kernels from one or several samples. We also gather answers to the issue of equivalence classes of DPPs. As two different matrices may generate the same DPP, we give conditions for DPP kernels to be equivalent, in particular the DPixP kernels. In these sections, we provide several algorithms to use DPixPs: sampling, inference, adapted shot noise sampling. Finally in the last section, we study DPPs over the set of patches of an image. We detail how to use DPPs in this setting and how to define an appropriate kernel, depending on the purpose of the subsampling.

2. Determinantal pixel processes: DPPs on pixels

2.1 Notations and definitions

In the following sections, we will consider DPPs defined on the pixels of an image. Let us first define any image as a function $u : \Omega \to \mathbb{R}^d$ ($d = 1$ for gray-scale images and $d = 3$ for color images), where $\Omega = \{0, ..., N_1 - 1\} \times \{0, ..., N_2 - 1\} \subset \mathbb{Z}^2$ is a finite grid representing the image domain. We use a common matrix convention for the notation of the indices, meaning that $u(0, 0)$ is in the top left corner and N_1 and N_2 are respectively the height and the width of the image. The cardinality of Ω, that is the number of pixels in the image, is denoted by $N = |\Omega| = N_1N_2$. Note that, if necessary, the pixels of an image are ordered and they are considered column by column. For any image $u : \Omega \mapsto \mathbb{R}^d$, and $y \in \mathbb{Z}^2$, the translation $\tau_y u$ of u by the vector y is defined using periodic boundary conditions by

$$\forall\ x = (x_1, x_2) \in \Omega,\ \tau_y u(x_1, x_2) := u(x_1 - y_1 \bmod N_1, x_2 - y_2 \bmod N_2).$$

In the following, we consider the Fourier domain $\widehat{\Omega} = \left\{-\frac{N_1}{2}, \dots, \frac{N_1}{2} - 1\right\} \times \left\{-\frac{N_2}{2}, \dots, \frac{N_2}{2} - 1\right\}$ if N_1 and N_2 are even (otherwise, we consider $\left\{-\frac{N_i-1}{2}, \dots, \frac{N_i-1}{2}\right\}$ if N_i is odd), so that the frequency 0 is centered. We define the discrete Fourier transform of a function $f : \Omega \mapsto \mathbb{C}$ by, for all $\xi \in \widehat{\Omega}$,

$$\widehat{f}(\xi) = \mathcal{F}(f)(\xi) = \sum_{x\in\Omega} f(x)e^{-2i\pi\langle x,\xi\rangle},\ \text{with } \langle x, \xi\rangle = \tfrac{x_1\xi_1}{N_1} + \tfrac{x_2\xi_2}{N_2}.$$

This transform is inverted using the inverse discrete Fourier transform:

$$\forall x \in \Omega, \quad f(x) = \mathcal{F}^{-1}\left(\widehat{f}\right)(x) = \frac{1}{N}\sum_{\xi\in\widehat{\Omega}} \widehat{f}(\xi) e^{2i\pi\langle x,\xi\rangle}.$$

Besides, the Parseval formula asserts that for any function $f : \Omega \to \mathbb{C}$,

$$\|f\|_2^2 = \sum_{x\in\Omega} |f(x)|^2 = \frac{1}{N}\sum_{\xi\in\widehat{\Omega}} |\widehat{f}(\xi)|^2 = \frac{1}{N}\|\widehat{f}\|_2^2.$$

Following these conventions, note that given a function f defined on Ω, we consider it is extended by periodicity to $\mathbb{Z}^2$. For any function f defined on Ω, we set $f_-(x) := f(-x)$. The convolution of two functions f and g defined on Ω is given by

$$\forall x \in \Omega, \quad f * g(x) = \sum_{y\in\Omega} f(x-y)g(y),$$

where again the boundary conditions are considered periodic. Then, $f * g$ can be computed in the Fourier domain, since

$$\forall \xi \in \widehat{\Omega}, \quad \widehat{f * g}(\xi) = \widehat{f}(\xi)\widehat{g}(\xi).$$

The autocorrelation of a function f is denoted by R_f. It is defined for all $x \in \Omega$ by $R_f(x) = f * f_-(x) = \sum_{y\in\Omega} f(x-y)f(-y)$.

Let us consider a DPP with kernel K defined on $\Omega \times \Omega$. In this work, we will focus on the modeling of textures, which are often characterized by the repetition of a pattern, or small objects which may be indistinguishable individually. Their homogeneous aspect can be naturally modeled by a stationary random field. Thus we will suppose that the point processes under study are stationary and periodic. This hypothesis amounts to consider that the correlation between two pixels x and y only depends on the difference $x - y$: the distribution is invariant by translation, while assuming periodic boundary conditions. The pixels are ordered column by column so that the ordered index of a pixel $x = (x_1, x_2) \in \Omega$ is $x_1 + 1 + x_2 N_1$. Thus the kernel matrix K is a block-circulant matrix with circulant blocks, entirely characterized by its first column.

Definition 2 *A block-circulant matrix with circulant blocks K verifies for all $x = (x_1, x_2)$, $y = (y_1, y_2) \in \Omega$, for all $\tau = (\tau_1, \tau_2) \in \Omega$,*

$$K(x + \tau, y + \tau) = K(x, y),$$

where we still consider periodic boundary conditions.

Let us define a correlation function $C : \Omega \to \mathbb{C}$, extended to $\mathbb{Z}^2$ by periodicity, such that $K(x, y) = C(x - y), \forall\ x, y \in \Omega$. As it entirely characterizes K, it also characterizes the associated DPP. As circulant matrices, block-circulant matrices with circulant blocks are diagonalized in the Fourier basis [19, 16], and the eigenvalues of K are the Fourier coefficients of C. See Appendix A for an illustration and some details on the diagonalization of DPixP kernels.

In this new framework, we can define DPPs from their correlation function C, they are now called determinantal pixel processes (DPixP). A DPixP kernel has two representations: C defined on Ω or the initial matrix K defined on $\Omega \times \Omega$ which corresponds to the block-circulant matrix with circulant blocks whose first column is C.

Definition 3 (Stationary DPixP) *Let $C : \Omega \to \mathbb{C}$ be a function defined on Ω, extended by periodicity to $\mathbb{Z}^2$, such that*

$$\forall \xi \in \widehat{\Omega},\ \ \widehat{C}(\xi) \text{ is real and } 0 \leq \widehat{C}(\xi) \leq 1. \tag{4}$$

Such a function is called an admissible kernel. Any random subset $X \subset \Omega$ is called a (stationary) DPixP with kernel C and denoted $X \sim \mathrm{DPixP}(C)$ if

$$\forall A \subset \Omega,\ \ \mathbb{P}(A \subset X) = \det(K_A),$$

where $K_A = (C(x-y))_{x,y \in A}$ is a $|A| \times |A|$ matrix.

2.2 Properties

The next proposition is directly deduced from properties of general DPPs (Proposition 1).

Proposition 4 (Distribution of the cardinality) *The cardinality $|X|$ of a DPixP is distributed as the sum $\sum_{\xi \in \widehat{\Omega}} \mathcal{B}\mathrm{er}(\widehat{C}(\xi))$, where for all $\xi \in \widehat{\Omega}$, $\mathcal{B}\mathrm{er}(\widehat{C}(\xi))$ are independent Bernoulli random variables with parameters $\widehat{C}(\xi)$. In particular,*

$$\mathbb{E}(|X|) = \sum_{\xi \in \widehat{\Omega}} \widehat{C}(\xi) = NC(0) \text{ and } \mathrm{Var}(|X|) = \sum_{\xi \in \widehat{\Omega}} \widehat{C}(\xi)(1 - \widehat{C}(\xi)).$$

One can notice that it is easy to know and control the expected number of points in the point process. In the following, when comparing different DPixP kernels, we will consider a fixed expected cardinality n, meaning that we will fix $C(0) = \frac{n}{N}$.

Proposition 5 (Separable kernel) *Let C_1 and C_2 be two discrete kernels, of dimension 1, defined respectively on $\{0, ..., N_1 - 1\}$ and $\{0, ..., N_2 - 1\}$, both verifying Equation (4) (for the 1D discrete Fourier transform). Then the kernel C given by $\forall x = (x_1, x_2) \in \Omega$, $C(x) = C(x_1)C(x_2)$ is an admissible DPixP kernel. Such a kernel C will be called separable.*

Proof Notice that for all $\xi = (\xi_1, \xi_2) \in \widehat{\Omega}$, $\widehat{C}(\xi) = \sum_{x_1=0}^{N_1-1} \sum_{x_2=0}^{N_2-1} C_1(x_1)C_2(x_2)e^{-2i\pi\left(\frac{x_1\xi_1}{N_1} + \frac{x_2\xi_2}{N_2}\right)}$ $= \widehat{C}_1(\xi_1)\widehat{C}_2(\xi_2)$. Thus, clearly, for all $\xi \in \widehat{\Omega}$, $\widehat{C}(\xi)$ is real and $0 \leq \widehat{C}(\xi) \leq 1$. C is an admissible kernel. ■

Let us consider two fundamental examples of DPixPs. The first one is the Bernoulli process. It corresponds to the discrete analogous of the Poisson point process: points are

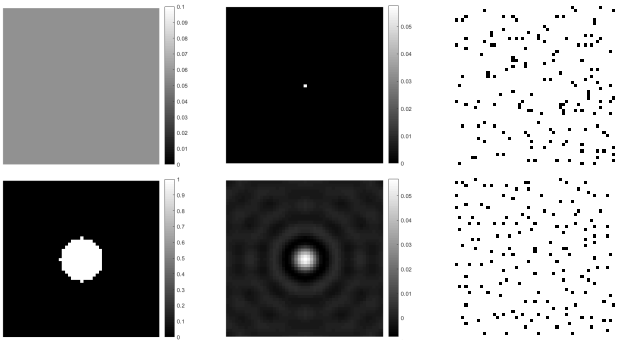


Figure 1: Comparison between samples (both have 148 points) of two DPixPs: a Bernoulli process (first line) and a projection DPixP defined by the kernel C such that $\widehat{C}$ is the indicator function of a discrete circle (second line). For both DPixPs, from left to right, the Fourier coefficients $\widehat{C}$ of the kernel, the real part of the kernel $\mathrm{Re}(C)$ and one associated sample.

drawn independently and following a Bernoulli distribution of parameter $p \in [0,1]$. This point process is the DPixP characterized by the kernel C such that $C(0) = p$ and for all $x \neq 0$, $C(x) = 0$, or equivalently for all $\xi \in \widehat{\Omega}$, $\widehat{C}(\xi) = p \in [0,1]$. The second main example is the family of projection DPixPs, that are determinantal processes defined by a kernel C which verifies for all $\xi \in \widehat{\Omega}$, $\widehat{C}(\xi)(1 - \widehat{C}(\xi)) = 0$. Thus, from Proposition 4, the number of points of projection DPixPs is fixed and equal to the number of non-zero Fourier coefficients of C.

Notice that in the general discrete case, the first example corresponds to the case where K is diagonal and the second one corresponds to the case where the eigenvalues of K are either equal to 0 or to 1. It is also called a projection DPP and the cardinality of the point process is equal to the number of non-zero eigenvalues, i.e. the rank of K.

Figure 1 presents two samples of these particular cases. Clearly, the projection DPixP enables a more "regular" distribution of the points in the square, tends to avoid both holes and clusters.

The common algorithm to sample exactly general determinantal processes is the spectral algorithm [30]. This is a two steps strategy which relies on an eigendecomposition $\{(\lambda_x, v_x)\}_{1 \leq x \leq N}$ of the matrix K. Indeed, define $(B_x)_{1 \leq x \leq N}$, N independent random variables such that $B_x \sim \mathcal{B}\mathrm{er}(\lambda_x)$ and $K_B = \sum_{x \in \Omega} B_x v_x^* v_x$, where v_x^* denotes the conjugate transpose of the vector v_x. Such a matrix K_B is a random version of K and Hough and al. [31] proved that $\mathrm{DPP}(K) = \mathrm{DPP}(K_B)$. Hence, the spectral algorithm consists in first drawing N independent Bernoulli random variables of parameters λ_x: these variables select n active eigenvalues and eigenvectors, where n is distributed as $\sum_{1 \leq x \leq N} B_x$. Then, it samples the n points from a projection DPP, obtained from the selected eigenvectors, thanks to a Gram-Schmidt procedure.

Recall that in our discrete stationary periodic framework, the eigenvalues of the matrix K are the Fourier coefficients of C and its eigenvectors are the elements of the Fourier basis

(Appendix A). Then an eigendecomposition of a DPixP of kernel C is computed using the 2D Fast Fourier Transform (FFT2) algorithm. Algorithm 1 presents the classic spectral algorithm [30, 38], adapted to sample a DPixP. In this algorithm, $(\varphi_\xi)_{\xi\in\widehat{\Omega}}$ denotes the columns of the unitary discrete Fourier transform matrix, the eigenvectors of K:

$$\forall\ \xi \in \widehat{\Omega},\ \forall\ x \in \Omega,\ \varphi_\xi(x) = \frac{1}{\sqrt{N}} e^{-2i\pi\langle x,\xi\rangle}. \tag{5}$$

Algorithm 1 Spectral simulation of $X \sim \mathrm{DPixP}(C)$

- Sample a random field $U = (U_\xi)_{\xi\in\widehat{\Omega}}$ where the U_ξ are i.i.d. uniform on $[0, 1]$.
- Define the "active frequencies" $\{\xi_1, \dots, \xi_n\} = \{\xi \in \widehat{\Omega}; U(\xi) \leq \widehat{C}(\xi)\}$, and denote

$$\forall x \in \Omega,\ \ v(x) = (\varphi_{\xi_1}(x), \dots, \varphi_{\xi_n}(x)) \in \mathbb{C}^n.$$

- Sample X_1 uniformly on Ω, and define $e_1 = v(X_1)/\|v(X_1)\| = \sqrt{\frac{N}{n}} v(X_1)$.
- For $k = 2$ to n do:
  - Sample X_k from the probability density p_k on Ω, defined by

$$\forall x \in \Omega,\ \ p_k(x) = \frac{1}{n-k+1}\left(\frac{n}{N} - \sum_{j=1}^{k-1} |e_j^* v(x)|^2\right)$$

  - Define $e_k = w_k/\|w_k\|$ where $w_k = v(X_k) - \sum_{j=1}^{k-1} e_j^* v(X_k) e_j$.
- Return $X = (X_1, \dots, X_n)$.

Note that the first point is supposed to be sampled from the distribution with density $p(x) = \frac{\|v(x)\|^2}{n}, \forall x \in \Omega$. Yet, in this framework, for all $x \in \Omega$, $\|v(x)\|^2 = \frac{n}{N}$, thus the first point of the realization is chosen uniformly. This is in accordance with the stationarity of $\mathrm{DPixP}(C)$. Because of the eigendecomposition of a matrix of size $|\Omega| \times |\Omega|$ the initial spectral algorithm runs in $\mathcal{O}(|\Omega|^3)$, yet thanks to the use of the Fast Fourier Transform algorithm to compute the Fourier coefficients of C, sampling DPixPs costs $\mathcal{O}(|\Omega| \log |\Omega|)$. Whereas in general the spectral algorithm is heavy when dealing with a huge data set, in this setting, it is very efficient. This allows us to handle large images. Thus, in addition to the explicit computation of marginals and of moments of a DPixP from its kernel, this exact sampler is one more asset of this family of point processes with respect to Gibbs processes, which are another common repulsive model.

In spatial statistics, the pair-correlation function (p.c.f.) g_X associated to a point process X is used to describe interactions between pairs of points. It characterizes the local repulsiveness of X [8]. For any discrete stationary point process on Ω, it is defined for all

$x \in \Omega$ by

$$g_X(x) = \frac{\mathbb{P}\left(\{0, x\} \subset X\right)}{\rho^2},$$

where ρ is the intensity of the point process, $\rho = \frac{\mathbb{E}(|X|)}{|\Omega|} = \mathbb{P}(0 \in X)$. It quantifies the degree of interaction between two points separated by a gap x: the closest g is to 1, the less correlated they are. If $g(x) > 1$, the points are considered to attract each other, whereas if $g(x) < 1$ the points are considered to repel each other. Notice that if $X \sim \mathrm{DPixP}(C)$,

$$g_X(x) = \frac{C(0)^2 - |C(x)|^2}{C(0)^2} = 1 - \frac{|C(x)|^2}{|C(0)|^2}.$$

Thus, if X is a Bernoulli point process, for all $x \neq 0$, $g_X(x) = 1$: there is no interaction between the points. Note also that for any DPixP, $g_X \leq 1$. During the sequential step of the sampling, each time a pixel is selected, a "repulsion zone" appears around it, where the probability for a pixel to be selected is low and whose shape depends on the kernel function C. This local "repulsion zone" is clearly retrieved in the pair correlation function computation.

Lavancier et al. [38] have also studied determinantal point processes in a stationary framework. Their work is based on a continuous setting, with kernels defined on $\mathbb{R}^d$, and due to the stationarity, it is also strongly related to the Fourier transform. They study diverse statistics to compare several stationary isotropic kernels. Similarly to our study in the next subsections, they are interested in the quantification of the repulsiveness of different kernels. In particular, they obtain that the most repulsive DPP kernels are projection kernels.

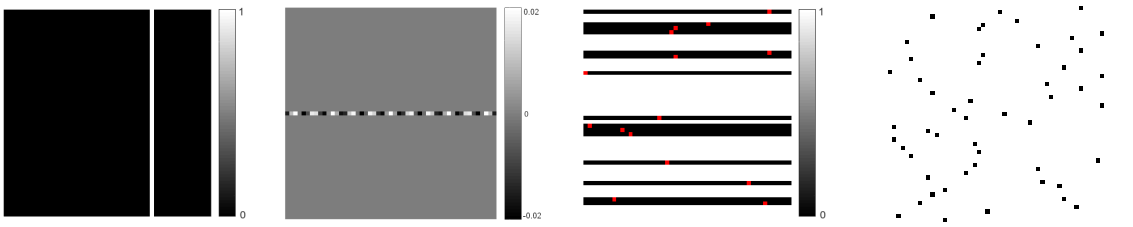
2.3 Hardcore repulsion

Gibbs processes are often used as their definition enables to precisely characterize the repulsion. Besides, they can allow us for hard core repulsion, meaning that the points are prohibited from being closer than a certain distance. To compare with this family of point processes, we investigate the possibility of hard core repulsion in the case of DPixPs. First, we study a hard core repulsion for pairs of points. Specifically, if $x \in \Omega$ and $e \in \Omega$ (for instance $e = (1, 0)$ or $(0, 1)$), is there a DPixP kernel such that x and $x + e$ can't belong simultaneously to the sample?

Proposition 6 *Let us consider $X \sim \mathrm{DPixP}(C)$ on Ω and $e \in \Omega$. Then the following propositions are equivalent:*

1. *For all $x \in \Omega$, the probability that x and $x + e$ belong simultaneously to X is zero.*
2. *For all $x \in \Omega$, the probability that x and $x + \lambda e$ belong simultaneously to X is zero, for $\lambda \in \mathbb{Q}$ such that $\lambda e \in \Omega$.*
3. *There exists $\theta \in \mathbb{R}$ such that the only frequencies $\xi \in \widehat{\Omega}$ such that $\widehat{C}(\xi)$ is non-zero are located on the discrete line defined by $\langle e, \xi \rangle = \theta$.*
4. *X contains almost surely at most one point on every discrete line of direction e.*

This is called directional repulsion.



(a) Fourier coefficients $\widehat{C}$ (b) Real part of the kernel C (c) Capture during the sampling (d) Resulting sample

Figure 2: Example of a kernel associated with hard core repulsion in the horizontal direction. From left to right, the Fourier coefficients of C, the real part of the kernel C, a capture of the conditional density during the simulation, the associated final sample.

Proof Let X be a DPixP defined on Ω with kernel C. First, let us prove that $1 \Leftrightarrow 3$. Recall that for all $x \in \Omega$, $\mathbb{P}(\{x, x+e\} \subset X) = C(0)^2 - |C(e)|^2$. We deduce from the triangle inequality that

$$|C(e)| = \left| \frac{1}{|\Omega|} \sum_{\xi \in \widehat{\Omega}} \widehat{C}(\xi) e^{2i\pi\langle e, \xi \rangle} \right| \leq \frac{1}{|\Omega|} \sum_{\xi \in \widehat{\Omega}} \widehat{C}(\xi) = C(0),$$

and the equality holds if and only if all non-zero elements of the left-hand side sum have equal argument. Thus, $\mathbb{P}(\{x, x+e\} \subset X) = 0$ if and only if there exists $\theta \in \mathbb{R}$ such that for all $\xi \in \widehat{\Omega}$, either $\widehat{C}(\xi) = 0$, or $\langle e, \xi \rangle = \theta$. Hence, for all $x \in \Omega$, the probability that x and $x+e$ belong simultaneously to X is zero if and only if the only non-zeros Fourier coefficients of C are aligned in the orthogonal direction of e. Second, let us prove that $2 \Leftrightarrow 3$. Consider $\lambda \in \mathbb{Q}$ such that $\lambda e \in \Omega$. Similarly, $\mathbb{P}(\{x, x+\lambda e\} \subset X) = 0$ if and only if there exists $\theta \in \mathbb{R}$ such that for all $\xi \in \widehat{\Omega}$, either $\widehat{C}(\xi) = 0$, or $\langle \lambda e, \xi \rangle = \theta$, meaning that $\langle e, \xi \rangle = \frac{\theta}{\lambda}$, which also is the equation of a discrete line orthogonal to e. Finally, suppose that X contains almost surely at most one point on every discrete line of direction e. Then, for all $x \in \Omega$, the probability that x and $x+e$ belong to X is zero so $4 \Rightarrow 1 \Leftrightarrow 3$. Now assume that the only non-zero Fourier coefficients of C are aligned on a discrete line that is orthogonal to e. As $2 \Leftrightarrow 3$ for all $\lambda \in \mathbb{Q}$ such that $\lambda e \in \Omega$, $\mathbb{P}(\{x, x+\lambda e\} \subset X) = 0$. Hence, X contains at most one point on any line of direction e, which can be described as a hard core repulsion of direction e. ■

Figure 2 illustrates this proposition: all non-zero Fourier coefficients are vertically aligned. The third figure presents a capture of the conditional density while the simulation is in progress, after 15 pixels already sampled. In each pixel, the probability that it is the next point selected is represented by the gray scale: the lighter a pixel is, the greater its probability of being the next point sampled. One can see that as soon as a pixel x is sampled, all the pixels belonging to the horizontal line passing through x have a zero probability of being sampled next. Note that this proposition is not limited to the horizontal or vertical cases.

Proposition 7 *Let $X \sim \mathrm{DPixP}(C)$ verifying the properties of Proposition 6, with $e = (1, 0)$, meaning that X contains at most one point on any horizontal line and all non-zero Fourier coefficients of C are aligned on a vertical line. Then C is separable in the sense of Proposition 5. Besides, the associated vertical point process is a DPixP of dimension 1 and conditionally to the drawn ordinates, the associated horizontal point process consists of a single point chosen uniformly and independently from the other horizontal point processes. The same proposition holds for $e = (0, 1)$ and vertical hardcore repulsion (inverting the terms horizontal and vertical).*

Proof Consider an admissible DPixP kernel C such that all its Fourier coefficients are either zero either aligned on a vertical line, positioned in $c \in \left\{-\frac{N_1}{2}, \dots, \frac{N_1}{2} - 1\right\}$ (here we assume that N_1 is even, the proof is similar if N_1 is odd). Thus we can define two functions $\widehat{C}_1 = \mathbb{1}_c$ and $\widehat{C}_2 = \widehat{C}(c, .)$ such that for all $\xi = (\xi_1, \xi_2) \in \widehat{\Omega}$, $\widehat{C}(\xi) = \widehat{C}_1(\xi_1)\widehat{C}_2(\xi_2) = \widehat{C}_2(\xi_2)\mathbb{1}_c(\xi_1)$. Notice that $C = \mathcal{F}^{-1}(\widehat{C}_1)\mathcal{F}^{-1}(\widehat{C}_2) = C_1C_2$. Such a function C_1 corresponds to an admissible DPixP projection kernel defined in one dimension, drawing one point and remember that the first point of a DPixP is drawn uniformly. Furthermore, C is a separable kernel. ■

Note that as soon as a pair of points configuration is prohibited, the whole direction is prohibited. As imposing a minimum distance between points is equivalent to prohibiting pair of points configurations in all directions, we deduce that the only DPixP imposing a minimum distance between the points is the degenerate DPixP, consisting of a single pixel. Hence, there is no DPixP with hard core repulsion in the broad sense, as it exists for Gibbs processes.

3. Shot noise models based on DPPs

3.1 Shot noise models and micro-textures

In the following section, we study discrete shot noise models driven by a DPixP. Shot noise models naturally appear to model phenomena such as the superposition of impulses occurring at independent and random times or positions. These models have been introduced in the computer graphics domain with the work of van Wijk [61]. Notice that van Wijk uses the expression spot noise texture as the spatial counterpart of 1D shot noise models yet the term shot noise is commonly employed for general models. Thus, in the rest of the section, we use this more general expression. Shot noise models are frequently used to approximate Gaussian textures as they are well-defined and simple mathematical models that allows us for fast synthesis [36], [23], [24]. Here, we are interested in the discrete version of these models on the finite grid $\Omega = \{0, \dots, N_1 - 1\} \times \{0, \dots, N_2 - 1\} \subset \mathbb{Z}^2$.

Definition 8 (Shot noise models based on a discrete point process) *Consider X a discrete point process with intensity ρ and g a (deterministic) function defined on Ω, periodically extended to $\mathbb{Z}^2$. Then, the shot noise random field S based on the points X and the spot g is defined by*

$$\forall x \in \Omega,\ S(x) = \sum_{x_i \in X} g(x - x_i).$$

In general, discrete shot noise models are based on a set of n i.i.d. random variables: it amounts to summing n randomly shifted versions of the spot. These models are particularly interesting for Gaussian texture synthesis as they have a Gaussian limit [22]. Indeed, in that case, the shot noise is the sum of n i.i.d. random images so that thanks to the Central Limit Theorem, we obtain a Gaussian limit. We study here shot noise models based on DPixPs. At the end of the section, we prove that there is a similar central limit theorem for shot noise models based on DPixPs that needs a modified framework but that ensures a Gaussian limit.

From now on, we consider an admissible kernel C and we suppose that X is the DPixP of kernel C. We study the interactions between the kernel C and the spot function g. To compute the moments of a shot noise model S based on X and a given spot, we need a moment formula ([44], [4]), also known as the Campbell or Slivnyak-Mecke formula, adapted to our discrete setting in the following proposition.

Proposition 9 (Moments formula for DPixPs) *Let X be a DPixP of kernel C defined on Ω, let us consider $k \geq 1$ an integer and f a function defined on Ω^k. We have*

$$\mathbb{E}\left(\sum_{x_1,\dots,x_k \in X}^{\neq} f(x_1,\dots,x_k)\right) = \sum_{y_1,\dots,y_k \in \Omega} f(y_1,\dots,y_k)\det((C(y_i-y_j)_{1\leq i,j\leq k}), \qquad (6)$$

where $\sum_{x_1,\dots,x_k \in X}^{\neq}$ means that the (x_i) are all different. In particular, for $k=1$, we have $\mathbb{E}\left(\sum_{x\in X} f(x)\right) = C(0)\sum_{y\in\Omega} f(y)$.

Proof By definition of the DPixP of kernel C, for any $y_1,\dots,y_k$ in Ω, we have

$$\mathbb{P}(\{y_1,\dots,y_k\}\subset X) = \det((C(y_i-y_j)_{1\leq i,j\leq k}).$$

Therefore, by the Slivnyak-Mecke formula [4], as we have

$$\mathbb{E}\left(\sum_{x_{i_1},\dots,x_{i_k}\in X}^{\neq} f(x_{i_1},\dots,x_{i_k})\right) = \sum_{y_1,\dots,y_k\in\Omega} f(y_1,\dots,y_k)\mathbb{P}(\{y_1,\dots,y_k\}\subset X),$$

we obtain the formula of the proposition. ■

Since $X \sim \mathrm{DPixP}(C)$ is stationary, S as defined in 8 is also stationary, so that $\mathbb{E}(S(x)^k) = \mathbb{E}(S(0)^k)$ for all $x\in\Omega$ and for all $k\geq 1$.

Proposition 10 (First and second order moments) *Let S be a shot noise model based on $X \sim \mathrm{DPixP}(C)$ and the spot g. Recall that R_g is the autocorrelation of g. We have* $\mathbb{E}(S(0)) = C(0)\sum_{y\in\Omega} g(y)$, *and for all $x\in\Omega$, $\Gamma_S(x) := \mathrm{Cov}(S(0),S(x)) = C(0)R_g(x) - (R_g * |C|^2)(x)$. In particular,*

$$\mathrm{Var}(S(0)) = C(0)\sum_{y\in\Omega} g(y)^2 - (R_g * |C|^2)(0)$$

and $\forall \xi \in \widehat{\Omega}, \widehat{\Gamma}_S(\xi) = |\widehat{g}(\xi)|^2 (C(0) - \widehat{|C|^2}(\xi))$.

Proof First, let us compute the mean value of such a shot noise model S. Using the periodicity of g,

$$\mathbb{E}(S(0)) = \mathbb{E}\left(\sum_{x \in X} g(-x)\right) = \sum_{y \in \Omega} g(-y) C(0) = C(0) \sum_{y \in \Omega} g(y).$$

Second, let us compute the covariance function of S for all $x \in \Omega$,

$$\begin{aligned}
&\Gamma_S(x) = \mathrm{Cov}\left(S(0), S(x)\right) = \mathbb{E}\left((S(0)S(x)\right) - \mathbb{E}\left(S(0)\right)^2 \\
&= \mathbb{E}\left(\sum_{x_1 \in X} g(-x_1) \sum_{x_2 \in X} g(x - x_2)\right) - \mathbb{E}\left(S(0)\right)^2 \\
&= \mathbb{E}\left(\sum_{x_1, x_2 \in X}^{\neq} g(-x_1) g(x - x_2)\right) + \mathbb{E}\left(\sum_{x_1 \in X} g(-x_1) g(x - x_1)\right) - \mathbb{E}\left(S(0)\right)^2 \\
&= \sum_{y_1, y_2 \in \Omega} g(-y_1) g(x - y_2) \left(C(0)^2 - |C(y_2 - y_1)|^2\right) + \sum_{y \in \Omega} g(-y) g(x - y) C(0) - \mathbb{E}\left(S(0)\right)^2 \\
&= C(0) g * g_-(x) - (g * g_- * |C|^2)(x)
\end{aligned}$$

■

3.2 Extreme cases of variance

We set $N = |\Omega| = N_1 N_2 \in \mathbb{N}$, the integer $n \leq N$ and $\mathcal{C}_n$ the set of admissible kernels such that $C(0) = \frac{n}{N}$. If $X \sim \mathrm{DPixP}(C)$, with $C \in \mathcal{C}_n$, notice that $\mathbb{E}(|X|) = |\Omega| C(0) = n$. Given a spot function g, we are looking for admissible kernels $C \in \mathcal{C}_n$ that generate shot noise models S of maximal and minimal variance. Indeed, the value $\mathrm{Var}(S(0))$ quantifies a repulsion "in the sense of g" or the regularity of the shot noise. It is related to the superposition of the spot and thus to a particular spatial arrangement of the points which can be adapted to the spot g. The case of a shot noise S based on a spot function g defined as an indicator function gives one some intuition into this idea. If $\mathrm{Var}(S(0))$ is low, the values taken by S are close to its mean value: there are few zones with no spot and few zones with many superpositions of the spot. This means that the points sampled from $\mathrm{DPixP}(C)$ tend to be far from one another, according to the shape of the function g and S appears more homogeneous. The repulsion is maximal. On the contrary, when $\mathrm{Var}(S(0))$ is high, S may take high values, so there can be many points in the same area. In that case, the repulsion is minimal.

Proposition 11 (Extreme cases of variance) *Fix $g : \Omega \to \mathbb{R}^+$ and an integer $n \leq N$. The variance of the shot noise model S is maximal if it is based on the Bernoulli DPixP that belongs to $\mathcal{C}_n$, meaning that its kernel C is such that $C(0) = \frac{n}{N}$ and for all $x \neq 0$, $C(x) = 0$. The variance of the shot noise model S is minimal when it is based on the projection DPixP of n points, such that the n frequencies $\{\xi_1, ..., \xi_n\}$ associated with the non-zero Fourier coefficients of its kernel maximize*

$$\sum_{\xi, \xi' \in \{\xi_1, ..., \xi_n\}} |\widehat{g}(\xi - \xi')|^2.$$

Proof Given a fixed integer $n \leq N$, let us consider $C \in \mathcal{C}_n$ that maximizes or minimizes

$$\begin{aligned}\mathrm{Var}(S(0)) &= C(0) g * g_-(0) - (g * g_- * |C|^2)(0) \\ &= \frac{n}{|\Omega|^2} \sum_{\xi} |\widehat{g}(\xi)|^2 - \frac{1}{|\Omega|^2} \sum_{\xi,\xi'} |\widehat{g}(\xi-\xi')|^2 \widehat{C}(\xi)\widehat{C}(\xi').\end{aligned}$$

If we identify the function $\widehat{C}$ to a vector of $\mathbb{R}^N$, the question becomes finding $C \in \mathcal{C}_n$ that maximizes or minimizes $F : \mathbb{R}^N \to \mathbb{R}$, where $F(\widehat{C}) = \sum_{\xi,\xi'} |\widehat{g}(\xi-\xi')|^2 \widehat{C}(\xi)\widehat{C}(\xi')$.

Maximal variance: We define a scalar product associated to g for all $v, w \in \mathbb{R}^N$, by $\langle v, w\rangle_g = \sum_{\xi,\xi' \in \Omega} |\widehat{g}(\xi-\xi')|^2 v_\xi w_{\xi'} = v^T G w$ where G is the $N \times N$ matrix such that $G = (\,|\widehat{g}(\xi-\xi')|^2)_{\xi,\xi' \in \widehat{\Omega}}$. This scalar product is well defined as it is bilinear, symmetric and for all $v \in \mathbb{R}^N$, $\sum_{\xi,\xi'=1}^{N} |\widehat{g}(\xi-\xi')|^2 v_\xi v_{\xi'} = R_g * |\widehat{v}|^2(0) \geq 0$, where R_g is the autocorrelation of g, and $\langle v, v\rangle_g = 0 \Leftrightarrow v = \mathbf{0}$. Notice that since G is symmetric positive definite then $F : \widehat{C} \mapsto \langle \widehat{C}, \widehat{C}\rangle_g$ is strictly convex. The case of maximal variance is achieved for the vector $\widehat{C}$ that minimizes this strictly convex function on the convex set $\mathcal{C}_n$: the problem has at most one solution [11].

According to the Cauchy-Schwarz inequality, we have for all $v, w \in \mathbb{R}^N$, $|\langle v, w\rangle_g| \leq \|v\|_g \|w\|_g$. Let us pick $v = \widehat{C}$, the vector whose components are the Fourier coefficients of a kernel $C \in \mathcal{C}_n$ and $w = \mathbf{1}$ $(= (1, 1, \ldots, 1)$ the constant vector of size $N)$. We have $\|v\|_g^2 = F(\widehat{C})$ and $\|w\|_g^2 = \sum_{\xi,\xi'} |\widehat{g}(\xi-\xi')|^2 = \sum_{\xi,\xi'} \widehat{R_g}(\xi-\xi') = N^2 R_g(0)$. Hence $\|v\|_g \|w\|_g = \sqrt{N^2 F(\widehat{C})(g * g_-)(0)}$ and

$$|\langle v, w\rangle_g| = \sum_{\xi,\xi'} |\widehat{g}(\xi-\xi')|^2 \widehat{C}(\xi) = \sum_{\xi} \widehat{C}(\xi) \sum_{\xi'} |\widehat{g}(\xi-\xi')|^2 = n\, N\, R_g(0).$$

Thus, $F(\widehat{C}) \geq n^2 R_g(0)$ and $F(\widehat{C})$ is minimal if and only if $\widehat{C}$ is proportional to w: necessarily, for all $\xi \in \widehat{\Omega}$, $\widehat{C}(\xi) = \frac{n}{N}$. Hence, C is a Bernoulli process. This kernel maximizes the variance of any shot noise S, independently of the spot g. It it the least repulsive DPixP.

Minimal variance: Let us characterize the kernel C that maximizes the function F on the convex set $\mathcal{C}_n$. F is quadratic so that solutions are on the boundaries of $\mathcal{C}_n$, meaning that for all kernel $\widehat{C}^* \in \widehat{\mathcal{C}}_F^* := \{\underset{\widehat{C}}{\mathrm{argmax}}(F(\widehat{C}))\}$, $\sum_{\xi} \widehat{C}^*(\xi) = n$ and $\forall \xi \in \widehat{\Omega}$, $\widehat{C}^*(\xi)(1 - \widehat{C}^*(\xi)) = 0$. Thus, the solutions are the projection DPixP kernels C^* with exactly n frequencies $\{\xi_1, ..., \xi_n\} \subset \widehat{\Omega}$ such that $\widehat{C}^*(\xi_i) = 1$ chosen so that $\sum_{\xi,\xi' \in \{\xi_1,\ldots,\xi_n\}} |\widehat{g}(\xi-\xi')|^2$ is maximal. ■

In the end, to determine the maximal repulsion kernel, one needs to maximize a quadratic function, which is NP-hard in general. In practice, it amounts to solve a combinatorial problem. It is possible to approximate the solution thanks to a greedy algorithm: first, one chooses two frequencies ξ_1, ξ_2 maximizing $|\widehat{g}(\xi_1 - \xi_2)|^2$ then, recursively, one chooses the kth frequency $\xi_k, 2 < k \leq N$, such that it maximizes $\sum_{\xi \in \{\xi_1,\ldots,\xi_{k-1}\}} |\widehat{g}(\xi - \xi_k)|^2$.

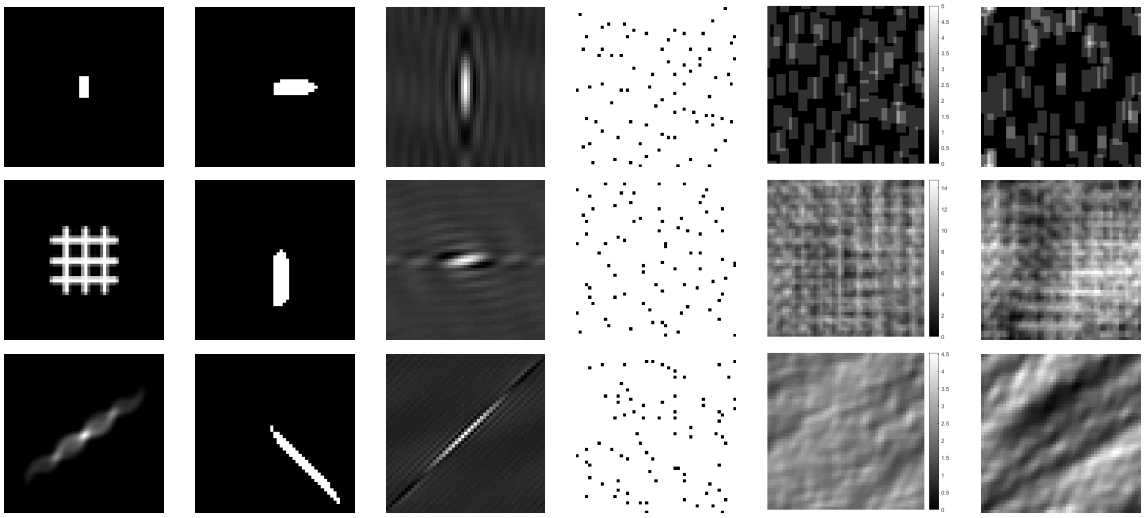


Figure 3: Realizations of the shot noise model driven by several spot functions and the most repulsive DPixP adapted to this spot. From left to right: the spot function, the Fourier coefficients obtained by our glutton algorithm, the real part of the associated kernel C, a sample of this most repulsive DPixP and a sample of the associated shot noise model and finally a Bernoulli shot noise model, both having the same expected number of points ($n = 80$).

Figure 3 presents some results of this algorithm. This figure shows that a projection DPixP adapted to g generates shot noise models with very few spot superpositions. Recall that in section 2, we proved that it was impossible to completely prevent superpositions. Yet, it is possible to characterize the least and the most repulsive DPixPs according to a specific desired repulsion. These extreme cases are coherent with the results of Biscio and Lavancier [8] who quantified the repulsion of stationary DPPs defined on $\mathbb{R}^d$ and stated that the least repulsive DPP is the Poisson point process whereas the most repulsive family of DPP contains the kernels C such that their Fourier transform $\mathcal{F}(C)$ is the indicator function of a Borel set, an analog to the projection DPixPs defined here.

3.3 Convergence to Gaussian processes

Shot noise models driven by DPixP enable more diverse types of textures than the usual shot noise models, based on points drawn uniformly and independently. As with usual shot noise models based on discrete Poisson processes, it is appealing to study the behavior of the model when the density of the point process increases and tends to infinity. The model presented here takes into account the case where the points are sampled independently, when the shot noise is based on a Bernoulli process. Yet, usual discrete shot noise models, as defined in [22], are based on a point process that is not simple, meaning that the points can coincide. Hence it is possible to study the asymptotic behavior of the shot noise model when the intensity of the point process grows. On the contrary, in this work, the DPixPs we study are by definition simple: the points can't coincide. Thus, the framework needs to be adapted so that Ω expands to $\mathbb{Z}^2$. In spatial statistics, this study is called increasing-domain asymptotics [17] and is commonly used when the data is defined on a lattice. In the following subsection, we will consider a fixed DPP kernel and spot function expanding with the image domain. A related configuration is to consider Ω as a grid in $[0,1]^2 = \mathbb{T}^2$, the torus of dimension 2, that is refined. In that setting, called infill asymptotics [17], the kernel of the DPP is rescaled so that the points are allowed to be increasingly close and the number of points inside $[0,1]^2$ tends to infinity. In these configurations, it is possible to characterize asymptotic behaviors and to derive limit theorems such as a Law of Large Numbers or a Central Limit Theorem. To this end, let us consider stationary determinantal point processes on $\mathbb{Z}^2$ [52, 41], that we will also call determinantal pixel processes. This point process is defined by a discrete bounded operator K on $\ell^2(\mathbb{Z}^2)$. That means that $K : \ell^2(\mathbb{Z}^2) \to \ell^2(\mathbb{Z}^2), f \mapsto Kf$ such that $\forall\, t \in \mathbb{Z}^2, Kf(t) = \sum_{s\in\mathbb{Z}^2} K(t,s)f(s)$. We suppose that this DPP is stationary: we define a kernel function $C : \mathbb{Z}^2 \to \mathbb{C}$, such that $K(t,s) = C(s-t)$ and $C \in \ell^2(\mathbb{Z}^2)$. Then for all $t \in \mathbb{Z}^2, Kf(t) = \sum_{s\in\mathbb{Z}^2} C(s-t)f(s)$: such a K is a convolution operator.

As C belongs to $\ell^2(\mathbb{Z}^2)$, there exists a function $\widehat{C} \in L^2(\mathbb{T}^2)$ such that $\widehat{C} : \mathbb{T}^2 \mapsto [0,1]$, $\forall t \in \mathbb{Z}^2$, $C(t) = \displaystyle\int_{\mathbb{T}^2} \widehat{C}(x) e^{2i\pi t\cdot x} dx$ and $\widehat{C}$ is equal to $x \mapsto \displaystyle\sum_{t\in\mathbb{Z}^2} C(t) e^{-2i\pi t\cdot x}$ in the sense of $L^2(\mathbb{T}^2)$. Finally, the point process $X \sim \mathrm{DPixP}(C)$ is defined by $\forall A \subset \mathbb{Z}^2$, a finite subset,

$$\mathbb{P}(A \subset X) = \det(C_A), \text{ where } C_A = \left(C(x_i - x_j)\right)_{x_i,x_j\in A}.$$

This new definition of DPixPs on $\mathbb{Z}^2$ is an extension of the point process defined on Ω. The main properties of DPixPs are preserved and it allows us to study the asymptotic behavior of shot noise models driven by DPixPs, when the grid is refined or equivalently

when the support of the spot is spread out. To do so, we need to consider spot functions defined on $\mathbb{R}^2$.

Shirai and Takahashi [54] state the following limit theorems. Some guidelines for the proofs can be found in [54] for the $\mathbb{Z}^2$ case and in [52] and [53] for its continuous counterpart.

Proposition 12 (Limit theorems for shot noise models and DPixPs [54]) *Let f be a bounded measurable function on $\mathbb{R}^2$ with compact support, and $X \sim \mathrm{DPixP}(C)$ with C some admissible kernel on $\mathbb{Z}^2$. Then, we have the following Law of Large Numbers*

$$\frac{1}{M^2}\sum_{x\in X} f\left(\frac{x}{M}\right) \xrightarrow[M\to\infty]{} C(0)\int_{\mathbb{R}^2} f(x)dx, \text{ a.e and in } L^1.$$

Moreover, assume that f is continuous and $\int_{\mathbb{R}^2} f(x)dx = 0$. Then, $\forall t \in \mathbb{R}$,

$$\lim_{M\to\infty} \mathbb{E}\left(\exp\left(\frac{i}{\sqrt{M^2}}\sum_{x\in X} f\left(\frac{x}{M}\right)\right)\right) = \exp\left(-\frac{1}{2}\sigma(C)^2\|f\|_2^2\right)$$

where $\sigma(C)^2 = C(0) - \sum_{x\in\mathbb{Z}^2} |C(x)|^2$, and consequently,

$$\frac{1}{\sqrt{M^2}}\sum_{x\in X} f\left(\frac{x}{M}\right) \xrightarrow[M\to\infty]{D} \mathcal{N}(0, \sigma(C)^2\|f\|_2^2).$$

In the following, let g be a spot function, that we assume continuous, with compact support, and $M > 0$. Denote the M-normalized shot noise S_M associated to g defined for all $y \in \mathbb{Z}^2$ by $S_M(y) = \frac{1}{M^2}\sum_{x\in X} g\left(y - \frac{x}{M}\right)$. We obtain a Law of Large Numbers for the shot noise driven by DPixPs:

$$S_M(0) = \frac{1}{M^2}\sum_{x\in X} g\left(-\frac{x}{M}\right) \xrightarrow[M\to\infty]{} C(0)\int_{\mathbb{R}^2} g(x)dx, \text{ a.e and in } L^1. \tag{7}$$

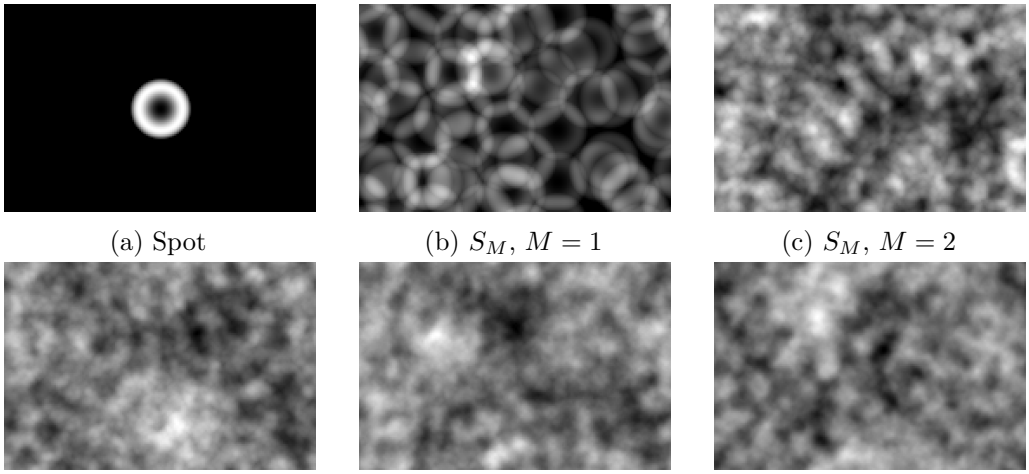
Finally, it is also possible to obtain a multidimensional central limit theorem thanks to the previous formulations.

Proposition 13 (Central limit theorem for shot noise models) *Let g be a continuous function on $\mathbb{R}^2$ with zero mean and compact support, $X \sim \mathrm{DPixP}(C)$ and the related shot noise S_M: $S_M(y) = \frac{1}{M^2}\sum_{x\in X} g\left(y - \frac{x}{M}\right), \forall y \in \mathbb{Z}^2$.*
Then, $\forall x_1, ..., x_m \in \mathbb{Z}^2$,

$$\sqrt{M^2}\left(S_M(x_1), \cdots, S_M(x_m)\right) \xrightarrow[M\to\infty]{D} \mathcal{N}(0, \Sigma(C))$$

where for all $k, l \in \{1, \cdots, m\}$,

$$\Sigma(C)(k,l) = \left(C(0) - \|C\|_2^2\right)\int_{\mathbb{R}^2} g(x_k - t)g(x_l - t)dt = \left(C(0) - \|C\|_2^2\right)R_g(x_l - x_k).$$



(a) Spot (b) S_M, $M = 1$ (c) S_M, $M = 2$

(d) S_M, $M = 3$ (e) S_M, $M = 6$ (f) $\mathcal{N}(0, \Sigma(C))$

Figure 4: Determinantal shot noise realizations S_M as defined in Theorem 12 with various $M = 1, 2, 3, 6$ and a comparison with their associated limit Gaussian random field $\mathcal{N}(0, \Sigma(C))$ (f). The shot noise is based on the spot (a) and the DPixP with kernel C whose Fourier coefficients form an isotropic 2D Gaussian function (Figure 1, bottom).

Proof Consider the M-normalized shot noise S_M associated to g: $\forall y \in \mathbb{Z}^2$, $S_M(y) = \frac{1}{M^2} \sum_{x \in X} g\left(y - \frac{x}{M}\right)$. By setting $\forall u \in \mathbb{R}^m, \forall x_1, ..., x_m \in \mathbb{Z}^2, \forall x \in \mathbb{R}^2$,

$$f(x) = u_1 g(x_1 - x) + u_2 g(x_2 - x) + \cdots + u_m g(x_m - x),$$

f is continuous on $\mathbb{R}^2$, with compact support such that $\int_{\mathbb{R}^2} f(x)dx = 0$ so it is possible to apply the limit theorem 12 and the Levy's continuity theorem. ■

Thus, shot noise models driven by a DPixP also converge to a Gaussian limit whose covariance is related to the spot and to the kernel C of the point process. Note that, in the previous proposition, the limit variance $\Sigma(C)$ is equal to the product of a constant depending on the kernel C and the autocorrelation of the spot g. Similarly, a normalized Poisson shot noise associated to the spot g converges towards the distribution $\mathcal{N}(0, R_g)$, where R_g is the autocorrelation of g [22]. As the Bernoulli case corresponds to the kernel function $C = \delta_0$, we retrieve the same result here.

Notice the similar work in a continuous framework of Poinas et al. on the limit distribution of sums of functionals of DPPs defined on $\mathbb{R}^d$ [48]. Figure 4 presents the asymptotic behavior of shot noise models driven by the continuous spot function shown in (a), whose integral is zero, and a DPixP on $\mathbb{Z}^2$ with a kernel whose Fourier coefficients are given by an isotropic 2D Gaussian function. When the grid is refined, the shot noise as defined in this section tends to a Gaussian texture associated to the spot and the kernel of the DPixP.

There are several perspectives to extend these results. Note that there is no more interaction between the spot and the kernel in the limit and the higher the repulsion is, in

the sense of high kernel coefficients, the lower the variance is. In particular, for projection kernels C, the asymptotic variance $\Sigma(K)$ vanishes, as $\|C\|_2^2 = \int_{\mathbb{T}^2} |\widehat{C}(x)|^2 dx = \int_{\mathbb{T}^2} \widehat{C}(x) dx = C(0)$. This suggests that the convergence rate must be faster for projection DPixPs. The convergence of determinantal shot noise models towards a Gaussian process has also been proven by Soshnikov [56], with some relaxed assumptions. In his work, the spot function g does not need to be continuous or to have a zero mean but it needs to be bounded with precompact support and the expectation and variance of the shot noise are constrained. We plan to adapt these results to our framework to lighten the assumptions on the spot function.

4. Identifiability and Inference for Determinantal Pixel processes

One of the purposes of statistical inference is to fit a predetermined model to data that can be represented by points, using information on their global or local behaviour. When the data are assumed independent and well represented by a homogeneous point process, one can use Poisson point processes. Yet, some data may present attraction or repulsion, they may also have an anisotropic structure. DPixP models can be suitable for representing 2-dimensional discrete data points with repulsion. For instance, the positions of plant seeds [44] or trees in a forest [38] often exhibit repulsion because of limited shared supply, but also anisotropy due to environmental factors as wind orientation or ground steepness. In [47], Perrin et al. study statistical inference on repulsive point processes, Strauss processes, to detect tree crowns on aerial images of forests. Similarly, Descamps et al. [20] use Gibbs point processes to automatically detect flamingos on images of colonies. Once one has inferred the parameters of an appropriate model, it is possible to reproduce similar data, to detect objects or distinguish different regions by statistical testing.

Learning the parameters of a determinantal point process, either the whole underlying kernel K as in [32, 1] or a few parameters encoding the kernel as in [5, 9], is still considered as a difficult task, first because the likelihood is often non-convex, and most of all because it is complex to compute as it uses the determinant of a huge matrix. Most papers studying inference for DPPs overcome this difficult computation by using restrictive hypothesis on the kernel such as in the papers [33] or [1]. Bardenet and Titsias [5] develop bounds on the likelihood and use Markov Chain Monte Carlo methods to infer the parameters of the kernel. On the other hand, using descriptive statistics to fit the models to the data enables to cope with this difficult computation and to obtain more efficient inference algorithm. It is the approach that we choose in this paper. Some authors try to infer first order characteristics such as the intensity of the point process [10], which provides the average number of points in a given area. In our finite and discrete setting, we can obtain a direct estimation of the intensity, as the ratio between the number of points and the size of the domain. Several second order characteristics are used to describe a sample, for example the empty space distance, the cumulative nearest-neighbor function, the pair correlation function (p.c.f. in short), presented above, or the Ripley's K function, closely related to the p.c.f (see [44] for a detailed presentation). These statistics provide information on the interactions between points. Møller and Waagepetersen [44] present these different statistics and state that higher order characteristics may be less stable if the number of points is low. In the following, we choose to focus on a quantity related to the p.c.f. It has several advantages: it is easy to

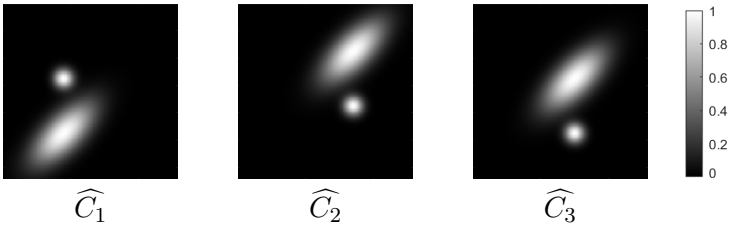


Figure 5: Three DPixP kernel functions, defined by their Fourier coefficients, generating the same DPixP.

interpret, it is easy to compute and it provides insights on local interactions. Biscio and Lavancier [9] also use the p.c.f for a minimum contrast estimation in continuous settings.

The purpose of this section is to derive a DPixP kernel function C from one or several samples of points on a finite and discrete domain. This estimation is non-parametric as we focus on general DPixP even though it can be seen as a parametric estimation of a DPP kernel matrix K of size $|\Omega| \times |\Omega|$ that we suppose block-circulant and determined by $|\Omega|$ parameters, the values of C. Before we investigate this question, it is necessary to characterize the identifiability of DPixP models.

4.1 Equivalence classes of DPP and DPixP kernels

A model is not identifiable if two different parametrizations are equivalent. Here, it would correspond to several different kernel functions generating the same DPixP. Lavancier et al. [39] proved that in a continuous setting, assuming that its intensity function is positive, a DPP kernel is uniquely defined. Yet, in the discrete case, DPPs and DPixP are not identifiable, as illustrates Figure 5. It is crucial, in particular for estimation purposes, to characterize these equivalence classes of kernels. Of course this question is also decisive in more general cases, when the kernel matrix K is Hermitian, with real or complex coefficients. We propose here a brief synthesis of what is known on this question, and we add a study on DPixP kernels.

The distribution of a DPP is entirely defined by all its principal minors (see Equation 1), thus characterizing DPP kernel equivalences classes is equivalent to understanding the consequences of equal principal minors on matrices, in the symmetric or Hermitian cases, and in the DPixP framework where the matrix is Hermitian circulant.

Notice that the characteristic polynomial of a matrix can be written as a function of its principal minors: $\det(tI + K) = \sum_{k=0}^{N} (-1)^k \left(\sum_{A \subseteq \mathcal{Y}, |A|=k} \det K_A \right) t^{n-k}$. Hence, two matrices with equal principal minors have equal characteristic polynomial so they have the same eigenvalues, with the same algebraic multiplicity. Thus, two kernel matrices generating the same DPP have the same spectrum. A key notion here is the diagonal similarity between two matrices: two square matrices M_1, M_2 are called diagonally similar if there exists a diagonal matrix D such that $M_2 = D^{-1} M_1 D$. In the following, we also need the notion of the directed graph associated to a matrix ([21], [27], [32]). Consider a matrix M of size $N \times N$. Its associated directed graph G_M contains the N vertices $\mathcal{Y} = \{1, \dots, N\}$ and an edge between

the vertices x and y if and only if $M(x,y) \neq 0$. The matrix M is called irreducible if G_M is strongly connected, meaning that there exists a sequence a path from any vertex to any other one. In the opposite case, the matrix is called reducible, which is equivalent to being permutation-similar to a block upper triangular matrix. Besides, it is called completely reducible if it is permutation-similar to a block diagonal matrix with irreducible blocks, meaning that there exists a permutation matrix P such that $P^T M P = \begin{pmatrix} M_1 & & 0 \\ & \ddots & \\ 0 & & M_r \end{pmatrix}$, $M_1, \dots, M_r$ irreducible. Notice that a Hermitian matrix is either irreducible or completely reducible.

Let us consider two general admissible DPP kernels K_1 and K_2, admissible meaning that they are Hermitian and their eigenvalues are in $[0,1]$. Thanks to basic determinant properties, notice that if there exists a diagonal matrix D such that $K_2 = D^{-1} K_1 D$ or $K_2^T = D^{-1} K_1 D$, then K_1 and K_2 have same principal minors, that is, the equivalence class of a DPP kernel contains all the admissible matrices of which the kernel matrix itself or its transpose is diagonally similar. In the case where the DPP kernel is real, Kulesza [32] proved the following proposition.

Proposition 14 (Equivalence classes of real symmetric kernels [32]) *Let K_1 and K_2 be two real positive symmetric $N \times N$ matrices with eigenvalues bounded by 1. Then $\mathrm{DPP}(K_1) = \mathrm{DPP}(K_2)$ if and only if there exists a $N \times N$ diagonal matrix D such that $K_2 = D^{-1} K_1 D$, where the coefficients of D are either 1 or -1.*

The proof of this proposition is in two parts. First, the author demonstrates the relation when all coefficients of the matrices are non-zero. Then, using graph theory, Kulesza extend this proof to matrices associated to a connected graph and finally to a disconnected graph, when the matrix is reducible. This equivalence property for real DPP kernels has impacted serveral learning strategies as in [49], [13], [60] or [12] which try to estimate real DPP kernels from several i.i.d. samples. In particular, the first two papers intend to solve the so-called principal minor assignment problem for symmetric matrices, and Brunel et al. [12] maximize a log-likelihood depending on the equivalence class of DPP kernels. Urschel et al. [60] obtain a bound on a distance between the estimated kernels L^* and the equivalence class of the original kernel: $\min_D \|L^* - D^{-1} L D\|_F$, on diagonal matrices D with coefficients only equal to 1 or -1.

In the paper [57], Stevens characterizes equivalence classes of real or complex symmetric DPP kernels. We would like to characterize DPP equivalence classes in a more general setting, where the DPP kernels are no longer real or symmetric but complex and Hermitian. Schneider, Saunders and Engel ([51],[21]) worked on the relation between equal principal minors and diagonal similarity through graph theory: see for instance [51] for links between equality of cyclic products and diagonal similarity, or [21] where they deal with real symmetric matrices. In 1986, Loewy [40] gives several sufficient conditions ensuring that if two square matrices have equal principal minors, one is diagonally similar to the other one or to the conjugate of the other one. We adapt these conditions to Hermitian DPP kernels in Theorem 16. In the following, we define $\mathcal{D}_N \subset \mathcal{M}_N(\mathbb{C})$ as the set of diagonal matrices of size $N \times N$ such that its coefficients are of modulus one.

Lemma 15 *Let K_1 and K_2 be two irreducible Hermitian matrices and assume that there exists an invertible diagonal matrix D such that $K_2 = D^{-1}K_1D$ or $K_2^T = D^{-1}K_1D$. Then all the coefficients of D have the same modulus so one can choose D in $\mathcal{D}_N$.*

Proof Assume that K_1 and K_2 are two irreducible Hermitian matrices and there exists a diagonal matrix D such that $K_2 = D^{-1}K_1D$ or $K_2^T = D^{-1}K_1D$. First, let us suppose that $K_2 = D^{-1}K_1D$. For all $x, y \in \mathcal{Y}$ such that $K_1(x, y) \neq 0$, we have also $K_2(x, y) \neq 0$ and

$$K_2(x, y) = \frac{1}{d_x} K_1(x, y) d_y.$$

As K_2 is Hermitian, $K_2(x, y) = \overline{K_2(y, x)} = \overline{\frac{1}{d_y} K_1(y, x) d_x} = \frac{\overline{d_x}}{\overline{d_y}} K_1(x, y)$. Then $\frac{d_y}{d_x} = \frac{\overline{d_x}}{\overline{d_y}}$, hence for all $x, y \in \mathcal{Y}$ such that $K_1(x, y) \neq 0$, $|d_x| = |d_y|$. Now recall that K_1 is irreducible. Its associated graph is connected and every node is reachable from any other node so it is possible to propagate this equality so that for all $x, y \in \mathcal{Y}$, $|d_x| = |d_y| = \lambda$. Then without loss of generality, changing if necessary to $\frac{1}{\lambda}D$, we can choose D as the matrix such that $K_2 = D^{-1}K_1D$ with diagonal coefficients of modulus equal to 1. The proof is similar if $K_2^T = D^{-1}K_1D$. ■

Now we can prove the following theorem on the equivalence classes of Hermitian DPP kernels.

Theorem 16 (Identifiability for Hermitian DPP kernels) *Let N be a positive integer and let $\mathcal{Y} = \{1, \dots, N\}$. Suppose that $K_1, K_2 \in \mathcal{M}_N(\mathbb{C})$ are two Hermitian admissible DPP kernels and that K_1 is irreducible. If $N \geq 4$, suppose furthermore that, for every partition of $\mathcal{Y}$ into subsets α, β such that $|\alpha| \geq 2, |\beta| \geq 2$, $\operatorname{rank}(K_1)_{\alpha\times\beta} \geq 2$. Then, the following propositions are equivalent:*

(i) $\mathrm{DPP}(K_1) = \mathrm{DPP}(K_2)$,

(ii) There exists a diagonal matrix D such that $K_2 = D^{-1}K_1D$ or $K_2^T = D^{-1}K_1D$,

(iii) There exists a diagonal matrix $D \in \mathcal{D}_N$ such that $K_2 = D^{-1}K_1D$ or $K_2^T = D^{-1}K_1D$.

Proof Define K_1 and K_2 two admissible DPP kernels, such that K_1 verifies the hypothesis of Theorem 16. By definition, $\mathrm{DPP}(K_1) = \mathrm{DPP}(K_2)$ is equivalent to K_1 and K_2 having equal principal minors. In the papers [27] (Theorem 7) and [40] (Theorem 1), Hartfiel and Loewy prove that if K_1 is irreducible and for every partition of $\mathcal{Y}$ into two subsets, α and β such that $|\alpha| \geq 2$ and $|\beta| \geq 2$, rank $(K_1)_{\alpha\times\beta} \geq 2$, then K_1 and K_2 have equal principal minors if and only if there exists a diagonal matrix D such that $K_2 = D^{-1}K_1D$ or $K_2^T = D^{-1}K_1D$. Notice that these two theorems, making the distinction between $\operatorname{rank}(K_1)_{\alpha\times\beta}$ and $\operatorname{rank}(K_1)_{\beta\times\alpha}$, are equivalent in this Hermitian setting. Then *(i)* is equivalent to *(ii)*. Besides, clearly *(iii)* implies *(ii)* and under these assumptions, by Lemma 15, *(ii)* implies *(iii)*. ■

In this general setting, assuming that K_1 is irreducible is crucial. Indeed, Hartfiel and Loewy [27] provide counterexamples of two admissible kernels generating the same DPP distribution without being diagonally similar.

We now turn to the special case of DPixP defined on Ω, the image domain of size $N_1 \times N_2$. Their kernel matrices are Hermitian block-circulant with circulant blocks. Recall that matrices generating DPixPs have all the same eigenvectors, the vectors of the Fourier basis. We also know that two matrices generating the same DPixP distribution have the same eigenvalues, so there is at most $(N_1N_2)!$ different kernels associated to one DPixP model. In the following proposition and remark, we prove that in most cases, the class of equivalence is much more constrained.

Proposition 17 (Identifiability for DPixP) *Let Ω be a finite grid of size $N_1 \times N_2$, and C_1, C_2 be two admissible DPixP kernels on Ω in the sense of Definition 3, generating the block-circulant matrices K_1 and K_2 that satisfy the hypothesis of Theorem 16. Then,* $\mathrm{DPixP}(C_1) = \mathrm{DPixP}(C_2)$ *if and only if there exists a translation mapping the Fourier coefficients of C_2 to the Fourier coefficient of C_1 or to their symmetry with respect to $(0,0)$, meaning that*

$$\begin{aligned}\mathrm{DPixP}(C_1) = \mathrm{DPixP}(C_2) \iff \exists\ \tau \in \Omega \text{ s.t. either } \forall \xi \in \Omega,\ \widehat{C}_2(\xi) = \widehat{C}_1(\xi - \tau)\\ \text{or } \forall \xi \in \Omega,\ \widehat{C}_2(\xi) = \widehat{C}_1(-\xi - \tau).\end{aligned} \tag{8}$$

Proof

As K_1 and K_2 satisfy the hypothesis of Therorem 16, there exists an invertible diagonal matrix D such that $K_2 = D^{-1}K_1D$ or $K_2^T = D^{-1}K_1D$, where $D \in \mathcal{D}_N$, meaning that D is a diagonal matrix with coefficients of modulus equal to one. First, assume that $K_2 = D^{-1}K_1D$. Define for all $x \in \Omega, \theta_x \in [0, 2\pi[$ such that $D(x,x) = e^{i\theta_x}$. The goal is to prove that there exists τ such that $\theta_x = 2\pi\langle x, \tau\rangle$, for all $x \in \Omega$. Notice that, by changing D into $\frac{1}{D(0,0)}D$, we can assume that $\theta_0 = 0$, that is $D(0,0) = 1$. By assumption, we obtain

$$\forall x, y \in \Omega,\ K_2(x,y) = C_2(y-x) = e^{-i\theta_x}K_1(x,y)e^{i\theta_y} = e^{i(\theta_y - \theta_x)}C_1(y-x),$$

$$\text{and } C_2(x) = C_2(x-0) = e^{i\theta_x}C_1(x).$$

Recall, thanks to Equations (2) and (3), that $C_1(0) = C_2(0)$ and that, for all $x \in \Omega$, $|C_1(x)| = |C_2(x)|$. As $C_2(x) = 0$ if and only if $C_1(x) = 0$, for such $x \in \Omega$, any value θ_x is valid. Consider the set $\Omega_C^* = \{x \in \Omega;\ C_1(x) \neq 0\}$. For all $z \in \Omega$, and all $x \in \Omega$, we have

$$C_2(z) = e^{i\theta_z}C_1(z) = C_2(z+x-x) = e^{i(\theta_{z+x} - \theta_x)}C_1(z+x-x) = e^{i(\theta_{z+x} - \theta_x)}C_1(z).$$

Denote for all $x \in \Omega$, $\alpha(x) = e^{i\theta_x}$. Thus, for all $z \in \Omega_C^*$, for all $x \in \Omega$, $\alpha(z) = \alpha(z+x)\overline{\alpha(x)}$, meaning that $\alpha(x) = \alpha(z+x)\overline{\alpha(z)}$. For all $\xi \in \widehat{\Omega}$, for all $z \in \Omega_C^*$, we have

$$\hat{\alpha}(\xi) = \sum_{x\in\Omega} \alpha(x)e^{-2i\pi\langle x,\xi\rangle} = \sum_{x\in\Omega} \overline{\alpha(z)}\alpha(z+x)e^{-2i\pi\langle x,\xi\rangle} = \overline{\alpha(z)}e^{2i\pi<z,\xi>}\hat{\alpha}(\xi).$$

As α is not the zero function, consider $\tau \in \widehat{\Omega}$ such that $\hat{\alpha}(\tau)$ is non-zero. Then, for all $z \in \Omega_C^*$, $\alpha(z) = e^{2i\pi<z,\tau>}$. Thus, for all $z \in \Omega_C^*$, $C_2(z) = e^{2i\pi<z,\tau>}C_1(z)$, which is also true for z such that $C_1(z) = 0$. To conclude, for all $z \in \Omega$, $C_2(z) = e^{2i\pi<z,\tau>}C_1(z)$. In the second case when $K_2^T = D^{-1}K_1D$, the proof is identical. ■

Remark 18 *Notice that when we consider two equivalent DPixP kernels C_1 and C_2, generating the block-circulant matrices K_1 and K_2, there are three possible configurations. The first one is when K_1 verifies the assumptions of Theorem 16, it leads to Proposition 17. In the second case, K_1 is irreducible, but $N = N_1N_2 \geq 4$ and there exists a partition α, β of $\mathcal{Y}$ such that $|\alpha| \geq 2$, $|\beta| \geq 2$ and $rank(K_1)_{\alpha\times\beta} < 2$. In the third case, K_1 is not irreducible. Let us characterize the second and third cases. It appears that these configurations are "rare" in practice.*

Case 2: *Assume that K_1 is irreducible, $N = N_1N_2 \geq 4$ and that there exists a partition α, β of $\mathcal{Y}$ such that $|\alpha| \geq 2$, $|\beta| \geq 2$ and $rank(K_1)_{\alpha\times\beta} < 2$. If $rank(K_1)_{\alpha\times\beta} = 0$, that is $(K_1)_{\alpha\times\beta} = 0$. There exists a permutation matrix such that K_1 is permutation similar to a block diagonal matrix, which is in contradiction with the irreducible hypothesis. Hence, $rank(K_1)_{\alpha\times\beta} = 1$. This means that there exist two vectors $u \in \mathbb{C}^{|\alpha|} \setminus \{0\}$ and $v \in \mathbb{C}^{|\beta|} \setminus \{0\}$ such that $(K_1)_{\alpha\times\beta} = u^T v$. In practice, as K_1 is Hermitian and the Fourier coefficients of C are real, the coefficients of the matrix K_1 are tightly constrained. The matrix is determined by a small number of modulus and arguments. Then, when assuming that K_1 and K_2 are equivalent, as DPixP kernels, the matrices are even more constrained. See Appendix B.1 for a simple example of this configuration. Notice that in the 1D case of dimension 5, two equivalent DPixP kernels K_1 and K_2 in this configuration still verify that there exists a diagonal matrix $D \in \mathcal{D}_N$ such that $K_2 = D^{-1}K_1D$ or $K_2^T = D^{-1}K_1D$. Our conjecture is that this is always the case, whatever the dimension of Ω. Thus, this assumption on the rank of the submatrix $(K_1)_{\alpha\times\beta}$ leads to degenerate kernels that are numerically "rare".*

Case 3: *K_1 is not irreducible. Then, as a Hermitian or circulant matrix, K_1 is necessarily completely reducible, meaning that there exists a permutation matrix P such that K_1 is permutation similar to a block diagonal matrix with irreducible blocks. We prove in Appendix B.2 that these blocks are copies of one Hermitian block-circulant sub-matrix, that we can call the canonical block: they all have equal size and the coefficients are identical. Note that restricting DPP to a subset A define also a DPP on this subset A [34, Section 2.3]. Furthermore, as each block matrix is still circulant, each one defines a sub-DPixP defined on the associated subset of pixels. By assumption, these blocks are irreducible so they are either in the first or in the second configuration. Let us consider K_2 a DPixP kernel equivalent to K_1. Thanks to the modulus equality, K_2 is similar to a block diagonal matrix with blocks of same size, using the same permutation matrix. If the canonical block is in the first configuration, verifying the rank hypothesis of Theorem 16, the final diagonal matrix D is simply the concatenation and rearrangement of all the diagonal sub-matrices D_i associated to its respective i-th block. Notice that as the block submatrices are identical to the canonical block and each one concerns a different set of pixels, all submatrices are in the same configuration, meaning that either for all submatrices K_{1i} of K_1, $K_{1i} = \overline{D_i}K_{2i}D_i$ or for all submatrices K_{1i}, $\overline{K}_{1i} = \overline{D_i}K_{2i}D_i$. On the other side, if the canonical block is in the second configuration, we can't conclude on the similarity of both matrices K_1 and K_2 in the general case yet. Notice that this completely reducible hypothesis is quite degenerate. It corresponds to a DPixP defined on an image domain that can be partitioned in groups of pixels evenly spaced with independence from one group to the other: that means that the pixels are independent to their immediate neighbors. A typical example of this model would be image domain partitioned following a grid. As DPixPs deals with spatial repulsion, there seems to be few applications of such models.*

It is important to notice that the size of the equivalence classes we characterized in Proposition 17 is small and known: given a DPixP kernel verifying the appropriate hypothesis, it admits at most $2|\Omega|$ equivalent kernels, generating the same DPixP distribution. Moreover, we have shown previously how a kernel that does not verify the hypothesis of the proposition is quite degenerate: in practice, when dealing with kernels adapted to a given problem, these hypothesis are always verified. In fact, we were not able to find an example of DPixP kernels C_1 and C_2 such that $\mathrm{DPixP}(C_1) = \mathrm{DPixP}(C_2)$ and which does not verify the right hand side of Equivalence (8). We conjecture that Equivalence (8) holds for all DPixP kernels, regardless of the hypotheses of Theorem 16. Characterizing equivalence classes of DPPs and DPixPs is crucial for the estimation of DPixP kernels from point process samples. This is what we investigate in the next subsection.

4.2 Learning a DPixP kernel from one realization

First, we address the question of inference from one single realization. Consider one set of points Y on Ω, the finite and discrete grid of size $N_1 \times N_2 = N$ and assume that Y has been sampled from a certain DPixP of kernel C_0. Note that in general, one realization does not provide enough information to characterize a model. Yet, due to the stationarity of the kernels we consider, all the translations of Y can also be seen as samples drawn by the same DPixP kernel C_0.

Let $n = |Y|$ denote the cardinality of Y. The problem is to find C_e an admissible DPixP kernel that estimates C_0, the original one. Equivalently, we want to find the Fourier coefficients $\widehat{C}_\mathrm{e} \in [0,1]^N$ the closest to $\widehat{C}_0$, in a sense defined below. In the following, we will work in the Fourier domain.

Let C be any admissible kernel on Ω and $X \sim \mathrm{DPixP}(C)$. As before, we will consider $\widehat{C}$ either as a function from $\widehat{\Omega}$ to $[0,1]$, or as a vector in $[0,1]^N$. Recall that the intensity of the point process is given by $\dfrac{\mathbb{E}(|X|)}{\Omega} = \dfrac{1}{\Omega}\displaystyle\sum_{\xi\in\widehat{\Omega}} \widehat{C}(\xi) = C(0)$. In case of a kernel estimation from one sample, it is natural to consider that the expected cardinality of the point process to be estimated is the cardinality of this unique sample. Thus, a straightforward estimation of the intensity of the point process is

$$C_e(0) = \frac{n}{N} \tag{9}$$

or equivalently $\displaystyle\sum_{\xi\in\widehat{\Omega}} \widehat{C}_e(\xi) = n$. Now, we want to determine the estimator $C_e(x)$, for all $x \in \Omega \setminus \{0\}$ denoted Ω^*. Let us consider

$$p_C(x) = \begin{cases} \mathbb{P}(x \in X |\, 0 \in X) = \dfrac{\mathbb{P}(\{0,x\} \subset X)}{\mathbb{P}(0 \in X)} = C(0) - \dfrac{|C(x)|^2}{C(0)} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases} \tag{10}$$

Now, from the realization Y, we can obtain $\theta(x)$ the empirical estimator of $p_C(x)$ by

$$\theta(x) = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{y\in\Omega} 1_Y(y) 1_Y(y+x) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases} \tag{11}$$

For optimization purposes, we express all the quantities in function of $\widehat{C}_e$. In the following computations, we consider that the vectors are column vectors. Let us denote the set of admissible functions by

$$\widehat{\mathcal{C}}_n = \{\widehat{C} \in \mathbb{R}^N \text{ such that } \sum_{\xi\in\widehat{\Omega}} \widehat{C}(\xi) = n \text{ and } \forall\ \xi \in \widehat{\Omega},\ 0 \leq \widehat{C}(\xi) \leq 1\}.$$

We are looking for $\widehat{C}_{\mathrm{e}}$ such that

$$\begin{aligned}
\widehat{C}_{\mathrm{e}} &= \operatorname*{argmin}_{\widehat{C}\in\widehat{\mathcal{C}}_n} \|p_C - \theta\|_2^2 \\
&= \operatorname*{argmin}_{\widehat{C}\in\widehat{\mathcal{C}}_n} \sum_{x\in\Omega^*} \left(\frac{n}{N} - \frac{N}{n}|\mathcal{F}^{-1}(\widehat{C})(x)|^2 - \frac{1}{n}\sum_{y\in Y} 1_Y(y)1_Y(y+x) \right)^2 \\
&= \operatorname*{argmin}_{\widehat{C}\in\widehat{\mathcal{C}}_n} \sum_{x\in\Omega^*} \left(\frac{n^2}{N^2} - \frac{1}{N}\sum_{y\in Y} 1_Y(y)1_Y(y+x) - |\mathcal{F}^{-1}(\widehat{C})(x)|^2 \right)^2 \\
&= \operatorname*{argmin}_{\widehat{C}\in\widehat{\mathcal{C}}_n} \sum_{x\in\Omega^*} \Big(b(x) - g(\widehat{C})(x) \Big)^2 = \operatorname*{argmin}_{\widehat{C}\in\widehat{\mathcal{C}}_n} E(\widehat{C}),
\end{aligned}$$

where, for all $\widehat{C} \in \mathbb{R}^N$, and for all $x \in \Omega^*$,

$$g(\widehat{C})(x) = |\mathcal{F}^{-1}(\widehat{C})(x)|^2 \ \text{ and } \ b(x) = \frac{n^2}{N^2} - \frac{1}{N}\sum_{y\in\Omega} 1_Y(y)1_Y(y+x). \tag{12}$$

We want to minimize E on $\widehat{\mathcal{C}}_n$ a non empty closed convex set so we can use the projected gradient algorithm. To project on the set of constraints $\widehat{\mathcal{C}}_n$, we use a classic adapted version of the algorithm to project onto the simplex [28, p. 77-78], integrating a maximum bound constraint, denoted "proj". From a vector $u \in \mathbb{R}^N$, this mapping returns the vector $\mathrm{proj}(u) \in \mathbb{R}^N$ such that $\mathrm{proj}(u)_x = \min(\max(u_x - a, 0), 1)$, where a is chosen such that $\sum_{x=1}^N \mathrm{proj}(u)_x = n$. Let us compute the gradient of the energy E we want to minimize.

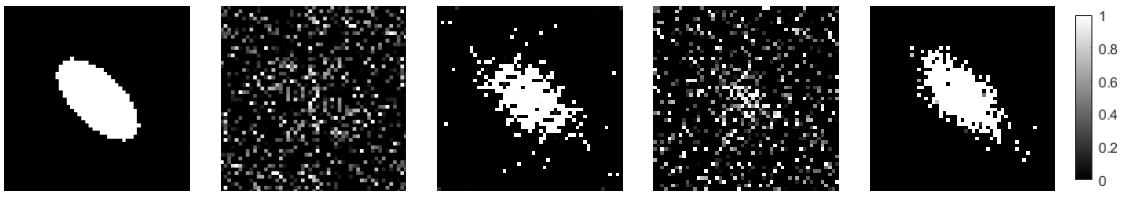
As $g : \mathbb{R}^N \to \mathbb{R}^{N-1}, \widehat{C} \mapsto \left(|\mathcal{F}^{-1}(\widehat{C})(x)|^2\right)_{x\in\Omega^*}$, we have

$$\begin{aligned}
\forall x \in \Omega^*, \forall \xi \in \widehat{\Omega}, \quad \frac{\partial g(\widehat{C})(x)}{\partial \widehat{C}(\xi)} &= \frac{1}{N}\overline{\mathcal{F}^{-1}(\widehat{C})}(x)e^{2i\pi\langle x,\xi\rangle} + \frac{1}{N}\mathcal{F}^{-1}(\widehat{C})(x)e^{-2i\pi\langle x,\xi\rangle} \\
&= \frac{2}{N}\mathrm{Re}\left(\mathcal{F}^{-1}(\widehat{C})(x)e^{-2i\pi\langle x,\xi\rangle}\right),
\end{aligned}$$

and moreover $\nabla E(\widehat{C}) = Dg(\widehat{C})^T\, 2\left(g(\widehat{C}) - b\right)$.

Notice that given a vector $u = (u_0, \ldots, u_{N-1})^T \in \mathbb{R}^\Omega$, we let u^* be $(u_1, \ldots, u_{N-1})^T$ the restriction of u to Ω^*. For all $\xi \in \widehat{\Omega}$,

$$\begin{aligned}
\left(Dg(\widehat{C})^T u^*\right)_\xi &= \frac{2}{N}\sum_{x\in\Omega^*} u_x \mathrm{Re}\left(\mathcal{F}^{-1}(\widehat{C})(x)e^{-2i\pi\langle x,\xi\rangle}\right) \\
&= \frac{2}{N}\mathrm{Re}\left(\sum_{x\in\Omega}\left(u_x\mathcal{F}^{-1}(\widehat{C})(x)\right)e^{-2i\pi\langle x,\xi\rangle} - u_0 C(0)\right).
\end{aligned}$$



a. b. c. d. e.

Figure 6: Two examples of initialization of our estimation algorithm. From left to right: the Fourier coefficients of the original kernel to retrieve (a), the initialization from one realization (b), the resulting estimated kernel (c), the initialization from 10 realizations (d), the resulting estimation from 10 realizations (e).

$$\text{Then } Dg(\widehat{C})^T u^* = \frac{2}{N}\mathrm{Re}\left(\mathcal{F}\left(u \odot \mathcal{F}^{-1}(\widehat{C})\right)\right) - \frac{2n}{N^2}u_0,$$

where $\odot$ refers to the componentwise product of vectors. Finally we obtain

$$\nabla E(\widehat{C}) = \frac{4}{N}\mathrm{Re}\left(\mathcal{F}\left(\left(|\mathcal{F}^{-1}(\widehat{C})|^2 - b\right)\mathcal{F}^{-1}(\widehat{C})\right)\right), \text{ by setting } b(0) = \frac{n^2}{N^2}. \tag{13}$$

In particular, computing $\nabla E(\widehat{C})$ only requires two FFT calls. The projected gradient descent algorithm is recalled and adapted to this problem in Algorithm 2. Note that the energy we want to minimize is not convex and it has several local minima: the initialization of the algorithm is crucial. Indeed, if the algorithm is initialized with a random matrix $\widehat{C}_{\text{init}}$, the results can be far from the original target. We propose to initialize the algorithm with

$$\widehat{C}_{\text{init}} = \text{proj}\left(\mathcal{F}\left(\sqrt{b}\right)\right), \tag{14}$$

which is believed to be quite close to a solution of the optimization and provides good results, as observed in the experiments. Note that b can be negative, so applying a square root to b may produce complex coefficients to which we apply the Fourier transform. This enables the initialization kernel $\widehat{C}_{\text{init}}$ to be asymmetric. The step of the gradient descent is chosen as a constant depending on the order of magnitude of the gradient of the energy E. Figure 6 presents two initialization kernels computed from a given realization.

Figures 7 and 8 (column 3) provide some results of this algorithm, from one realization generated by different DPixP kernels. One realization seems enough to retrieve the Fourier coefficients of a simple symmetric projection kernel (see Figure 7, a, b whose non-zero Fourier coefficients form a convex set). Even though for most projection kernels a predominant shape appears in the estimation, as soon as the kernel is more complex, one sample does not provide enough information.

4.3 Learning the kernel of a DPixP from several realizations

A unique realization does not provide enough information for our proposed algorithm to estimate the Fourier coefficients of a DPixP kernels but if several realizations are available, combining them provides better results. Assume that we have J realizations, $J \in \mathbb{N}^*$, each of cardinality n_j, that we suppose generated by the same DPixP kernel.

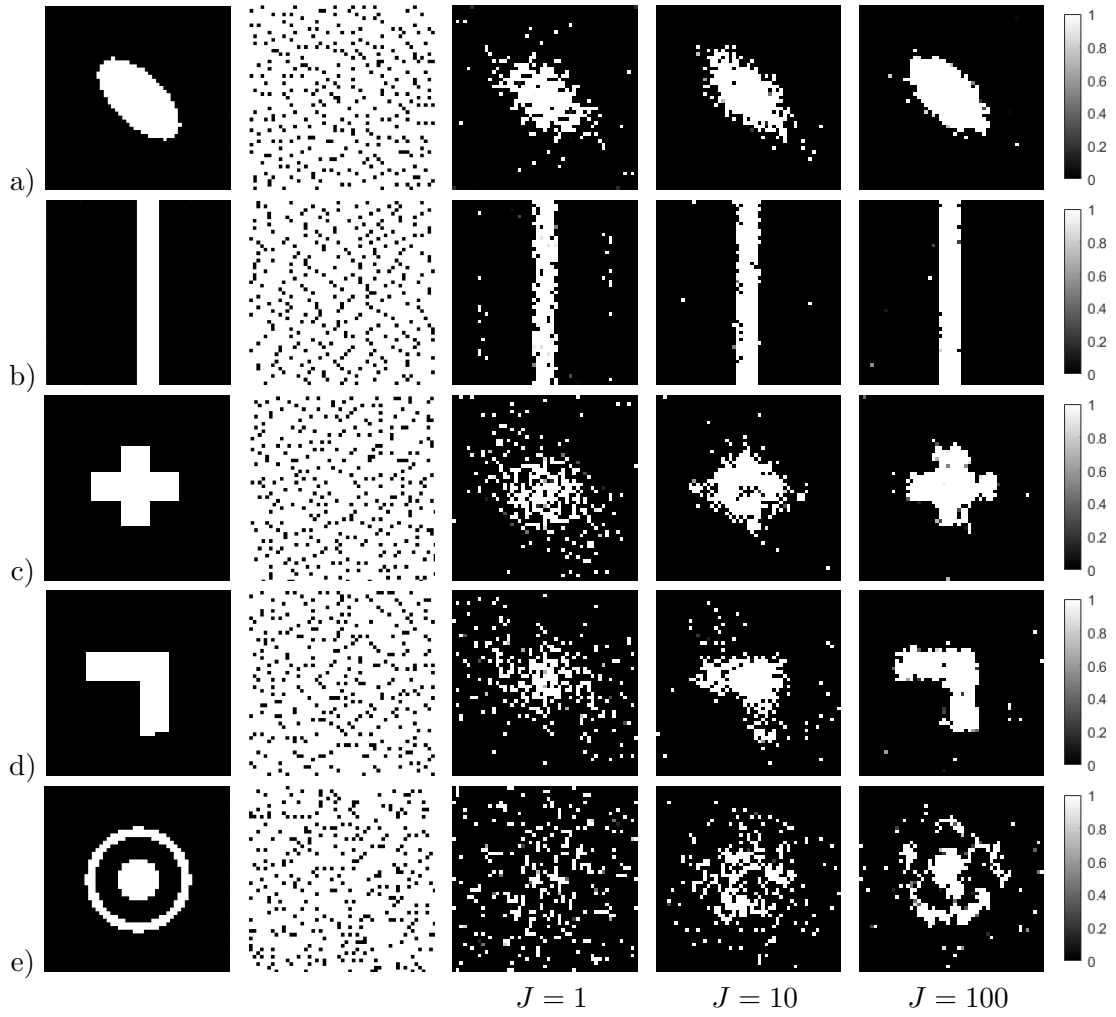


Figure 7: Experiments on several projection kernels. From left to right: the target Fourier coefficients of the kernel we want to recover, one realization of this DPixP, the estimation of the Fourier coefficients from one, from 10 and from 100 realizations, with $k_{\max} = 2000$.

Algorithm 2 Projected gradient descent algorithm used to minimize E.

Input: Y the input realization, step size t, $k_{\max}$,

- Compute for all $x \in \Omega^*$, $b(x) = \frac{n^2}{N^2} - \frac{1}{N}\sum_{y\in Y} 1_Y(y)1_Y(y+x), b(0) = 0$ (12).
- Set $\widehat{C}_0 = \widehat{C}_{\text{init}}$ (14).
- for $k = 1,\dots, k_{\max}$
  - Compute $\nabla E(\widehat{C}_{k-1})$ (13).
  - Set $\widehat{C}_k = \text{proj}\left(\widehat{C}_{k-1} - t\nabla E(\widehat{C}_{k-1})\right)$.

Output: $\widehat{C}_K$.

We propose to combine all the realizations to produce a better empirical estimator θ_J of p_C. First, the expected number of points is approximated by the mean number of points in the realizations, $n = \dfrac{n_1 + \cdots + n_J}{J}$.

If we have J realizations $(Y_i)_{i\in\{1,\dots,J\}}$, Equation (11) is replaced by:

$$\forall x \in \Omega,\ \theta_J(x) = \begin{cases} \dfrac{1}{nJ}\displaystyle\sum_{i=1}^{J}\sum_{y\in\Omega} 1_{Y_i}(y)1_{Y_i}(y+x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

The rest of the procedure remains similar as we want to minimize the function $\|p_C - \theta_J\|_2^2$, in particular, the initialization kernel is

$$\widehat{C}_{\text{init}} = \text{proj}\left(\mathcal{F}\left(\sqrt{\frac{n^2}{N^2} - \frac{1}{NJ}\sum_{i=1}^{J}\sum_{y\in\Omega} 1_{Y_i}(y)1_{Y_i}(y+x)}\right)\right).$$

Figures 7 and 8 present some experiments on several DPixP kernels, using the second strategy presented here and combining all the samples in one estimation process. First, Figure 7 presents the results of this estimation procedure with projection kernels, meaning that the Fourier coefficients of these kernels are zero or one. It shows how 10 realizations provide enough information to retrieve a kernel close to the original one. Using 100 realizations enables to obtain satisfying results. We have seen in the previous subsection that any translation of the estimated Fourier coefficients or a symmetry with respect to $(0,0)$ of the estimated Fourier coefficients generate the same DPixP. Thus, in Figures 7 and 8, we display a centered version of the estimation.

Figure 8 presents some results of this algorithm for non-projection DPixP kernels. Kernel a) is a Bernoulli kernel: all the Fourier coefficients are equal to $\frac{n}{N}$. As expected, no specific structure appears from the estimation, regardless of the number of samples used. The estimations b) and c) are much noisier than their projection equivalent (Figure 7(a,e)) even if the shape formed by the Fourier coefficients (which directly impacts the local repulsion of the point process) seems retrieved.

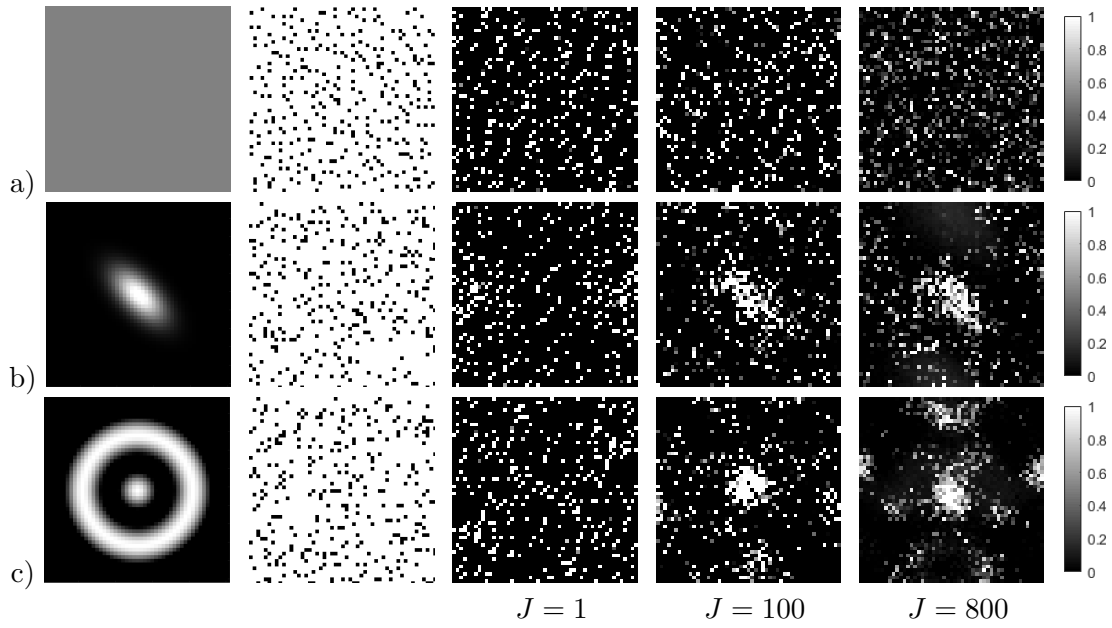


Figure 8: Experiments on general DPixP kernels. From left to right: the target Fourier coefficient of the kernel we want to recover, one realization of this DPixP, the estimation of the Fourier coefficients from one, from 100 and from 800 realizations, with $k_{\max} = 2000$.

To conclude, the algorithm presented in this subsection provides satisfying estimations if the original kernel is a projection DPixP kernel, in particular when we have more than 10 samples. Indeed, as we have seen in Section 3.2 and as the authors of [8] noted, projection determinantal processes can be seen as the most repulsive DPPs. Thus, within a sample, the characteristics of the repulsion, and of the kernel, are more accessible. Nevertheless, if we deal with a general complex kernel, the algorithm fails to retrieve meaningful information.This question of DPixP kernel estimation is complex and could be the subject of its own paper. Note that considering the quantity p_C related to the pair correlation function imposes that we have only access to $|C|$ thus different admissible kernels may be solution to this optimization problem. It would be necessary to study the theoretical properties of convergence, the bias and the variance of the estimator, the influence of the initialization or the existence of local minima, for instance.

5. Determinantal patch processes: DPPs on patches

As datasets to analyze and to process keep being larger and more complex, strategies to subsample these sets or to reduce the dimension of data have recently flourished. DPP subsampling is part of these approaches, as they enable to capture the structure of data and produce a representative subset of the whole initial set, taking into account its inner diversity. DPPs have been used in different applications such as text summarization [34], feature selection in high dimensional data [7] or approximation of a dataset adapted to a given learning task [58]. In image processing and computer vision, DPPs have raised interest through video summarization ([26], [63]). The authors of [26] introduced sequential DPP to take into account both the diversity of the frames and the chronology of the video. To represent the diversity of the frames they use a decomposition similar to the quality-diversity decomposition that is introduced in [34] and that we recall below. Furthermore, the paper [63] proposes a strategy enhanced by DPPs which makes it one of the state of the art methods for video summarization. This method also uses a decomposition similar to a quality-diversity decomposition to describe the diversity in the video.

In this section, we focus on subsampling the set of patches $\mathcal{P}$ of an image. This procedure can be useful for compression purpose for instance. It can also be necessary in order to fit a model on the patch set using only a proportion of the set, to increase the efficiency of the algorithm. For example, several patch-based denoising methods represent the patch distribution as a Gaussian mixture model ([64], [29]). These methods rely on the estimation of the parameters of such models thanks to the Expectation-Maximization (EM) algorithm. To do so, in general, they randomly and uniformly select a subset of patches. This random selection is fast but the subset needs to be large enough so that it captures the patches diversity. The size of this selection impacts the running time of the estimation process, so a smaller selection, representative of the patches of the image, would ensure a faster and more accurate estimation. DPPs offer the opportunity to select a small subset of patches that captures the whole image. For instance, the authors of [37] have used DPPs to subsample the patches of an image and approximate the patch distribution to speed up their texture synthesis algorithm.

Tremblay et al. [59] compare several k-means initializations with one using DPP sampling in the context of coresets: the authors sample an appropriate DPP to select the initial

centroids for the clustering strategy. Similarly, Agarwal et al. [2] propose to adapt the k-means algorithm by using a DPP initialization and they prove that this initialization compares favorably with k-means++, the most popular adaptation of the k-means algorithm, whose initialization uses another negatively correlated sampling strategy, called D^2 sampling. One advantage of an algorithm using DPPs over the second is its adaptability concerning the number of clusters. Thus, in the previous example with denoising methods, DPPs could also provide a satisfying initialization to the EM algorithm. We investigate here the possible choices of DPP kernels for such applications, in order to subsample the patch space of an image.

5.1 DPP kernels to sample in the space of image patches

When considering determinantal processes on patches, the framework is more general than before: We are no longer dealing with stationary periodic point processes. We consider a Hermitian kernel K adapted to select diverse subsets of patches from an image, as set in Definition 1. The definition of this diversity depends on the problem we want to solve: for instance, compression, reconstruction of the image or initialization of the centroids of a clustering or the EM algorithm.

First, we recall that there exists a second characterization of DPPs, using a positive semi-definite matrix L [34].

Definition 19 *We consider $\mathcal{Y} = \{1, \dots, N\}$ and L a Hermitian matrix of size $N \times N$ such that $L \succeq 0$, then the random set $X \subset \mathcal{Y}$ defined by*

$$\forall A \subset \mathcal{Y}, \quad \mathbb{P}(X = A) = \frac{\det(L_A)}{\det(I + L)}$$

is a DPP with L-ensemble kernel L. We will denote $X \sim \mathrm{DPP}_L(L)$.

Defined in this manner, these DPPs are called L-ensembles. Recall that the initial definition (1), using the kernel denoted by K, requires that $0 \preceq K \preceq I$. This L-ensemble definition doesn't need the constraint of bounding the eigenvalues of the kernel by one. This property is convenient to define a kernel, and a diversity model adapted to a specific problem. So this characterization is increasingly used in the machine learning community. That is also the definition we use in this subsection. However, note that this definition excludes the case of projection DPPs, which have a fixed cardinality. In subsection 5.2, we consider such a projection DPP kernel. Consider the following spectral decomposition of a DPP kernel K, $K = \sum_{k=1}^{N} \lambda_k v_k v_k^*$. Note that the definitions using the kernels K and the L-ensemble kernel L characterize the same DPP if and only if for all $k \in \{1, \dots, N\}$, $0 \leq \lambda_n < 1$ and if

$$K = L(L + I)^{-1} = I - (I + L)^{-1} \text{ and conversely } L = K(I - K)^{-1}.$$

Hence, in this case, $L = \displaystyle\sum_{k=1}^{N} \frac{\lambda_k}{1 - \lambda_k} v_k v_k^*$. Note that if K has any eigenvalue equal to 1, the DPP can't be associated to an L-ensemble.

In the following, consider an image u and the initial set $\mathcal{P} = \{P_i, i = 1, \dots, N\}$, the set of its patches of size $(2\rho + 1) \times (2\rho + 1) \times d$, where $\rho \in \mathbb{N}$ and d is the number of color

channel. Let us present some kernels that can be used to subsample the patches of this image.

Consider an expected cardinality n, so that in average, the number of selected patches is equal to n. In practice, this target cardinality can be imposed by normalizing the kernel matrices with a constant so note that in the following formulas, the matrices are defined up to a normalization constant, precised in the experiments section. A first type of L-ensembles that are commonly used ([59],[37]) is the class of Gaussian kernels. Let us consider a Gaussian kernel based on the intensity of the patches, that we call the intensity Gaussian kernel, defined by

$$\forall\, P_i, P_j \in \mathcal{P}, \quad L_{ij} = \exp\left(-\frac{\|P_i - P_j\|_2^2}{s^2}\right), \tag{15}$$

where s is called the bandwidth or scale parameter. This kernel depends on the squared Euclidean distance between the intensity values of pairs of patches. It is often used as a similarity measure on patches. Despite its natural limitations, this similarity measure provides good results. The value of the parameter s has an impact on how repulsive the DPP is. Notice that if s is small, due to the exponential function, L_{ij} converges very quickly to zero as soon as $i \neq j$ and the distinction between patches is not very subtle. Thus, if s is small, L is close to the identity matrix and the DPP selection of patches is similar to a random uniform selection. On the contrary, for the same reason, the larger s is, the more repulsive the DPP is. However, this scale parameter should not be set too large because this would cause high numerical instability. As noticed in [3] and [59], the median of the distances of intensities between the patches is a satisfying choice for setting the value of s.

We propose to compare this kernel with another Gaussian kernel that we call the PCA kernel, which depends on the squared distance between patches in the space given by keeping only the k principal components after a Principal Component Analysis (PCA). Set A the matrix gathering all the patches of the image reshaped in column so that the size of A is $d(2\rho+1)^2 \times N$. We assume that A has been centered, by subtracting the average patch to all the patches. It has not been reduced, meaning that patches with high variance, for instance patches with edges, will highly influence the decomposition. Thanks to a singular value decomposition, consider U, V two unitary matrices and Σ a diagonal matrix storing the principal values of A such that $A = U\Sigma V^T$. We choose to keep only k principal components and we obtain the matrix $A^k = V_k A$, where we kept only the k first rows of the matrix V in V_k of size $k \times d(2\rho+1)^2$ and the matrix $A^k = \{P_i^k,\, i = 1, \ldots, N\}$ is $k \times N$. Every initial patch $P_i \in A$ is associated with a projected vector $P_i^k \in A^k$. Thus, the PCA kernel is defined by

$$\forall\, P_i, P_j \in \mathcal{P}, \quad L_{ij} = \exp\left(-\frac{\|P_i^k - P_j^k\|_2^2}{s^2}\right). \tag{16}$$

This method discards principal vectors associated to small singular values and projects the patches on a low-dimensional space associated with the large singular values. This enables to find the components that best represent the variance of the patches and ignore mainly noise (depending on the number of dimension discarded). Thus, comparing patches in this low-dimensional space seems relevant to capture more precisely their dissimilarity.

A second type of common L-ensemble kernels uses a quality-diversity decomposition of the data. Kulesza and Taskar present in [34] this decomposition that uses a given quality

measure computed on each element of the set and a dissimilarity computed between pairs of elements. In this paper, each patch P_i is associated with a quality measure, which is a non-negative number $q_i = q(P_i, \mathcal{P}) \in \mathbb{R}^+$, depending on the patch itself and on the other patches. Each patch P_i is also associated with a feature vector $\phi_i = \phi(P_i) \in \mathbb{R}^D$, such that $\|\phi_i\|_2 = 1$, which depends only on the patch itself. The quality/diversity kernel L is defined by

$$\forall\, P_i, P_j \in \mathcal{P}, \quad L_{ij} = q_i \phi_i^T \phi_j q_j. \tag{17}$$

This class of kernels presents several advantages. The first advantage of this definition is its interpretability. Each patch is associated with a quality measure, that one can adapt depending on the characteristics one wants to favor. The comparison between patches is also accessible and adjustable to obtain the most adapted kernel. This decomposition has a second advantage: the kernel becomes a low-rank matrix, with a rank equal at most to D, the number of features. In case of low-rank kernels, Kulesza and Taskar [33] propose a dual representation and a dual sampling algorithm. This sampling scheme is equivalent to the original algorithm but it takes advantage of the low-rank kernel and becomes much faster. Note that, whatever the DPP L-ensemble kernel, the cardinality of a sample generated from $\mathrm{DPP}_L(L)$ will necessarily be lower than the rank of L. This low-rank definition imposes to sample subsets of smaller size than D, the number of features computed from the patches. Thus, this kernel is adapted when small and very small subsets of patches are needed. In these cases, it is very important to precisely control the selection process so such kernels are particularly relevant.

For the kernel that we call Qual-div kernel, we associate each patch with a feature vector given by a discrete cosine transform of the patch. Thus, each feature vector is of size $d(2\rho+1)^2$. Note that in the experiments, we use color images (with 3 color channels) and patches of size 7×7 (meaning that $\rho = 3$) so the feature vectors of length 147. We define the quality measure such that it attributes a high value to patches whose intensity is far from that of its neighbors in the pixel grid. For each pixel, we count the number of neighboring pixels whose intensity is further in Euclidean distance than a certain constant. This constant is chosen equal to half the maximum Euclidean distance between the intensities in the image. This choice gives further priority to singular patches, that can be seen as the outliers of the set of patches. As experiences will show, it highly favors textures and edges.

5.2 Minimizing the selection error

The question is to choose the best kernel, such that the sampled DPP on the patches minimizes an error computed as a distance between the selected patches and the initial set of patches $\mathcal{P}$. This last kernel will be called the optimized kernel. This problem is similar to discrete optimal quantization problems [46] where the aim is to find the best subset of patches $\mathcal{Q}$ such that $\mathbb{E}_{\mathcal{Q}\sim\mu}(d(\mathcal{Q}, \mathcal{P}))$ is minimal, for a given distance d. Yet, this computation is often costly and hardly tractable. In the following, we suppose that the patches are of size $(2\rho+1) \times (2\rho+1)$ for some positive integer ρ and we denote by $\omega \subset \mathbb{Z}^2$ the patch domain $\{-\rho, \dots, \rho\}^2$.

First, the error, or the distance between the sample and the initial set of points, we want to minimize depends on the application. The mean square error (MSE in short) is commonly used to compare an image and its reconstruction. Here, we use a similar distance,

the squared L^2 norm between the patches of the image and their nearest neighbor in the selection given by the DPP sampling on the patches. Consider $\mathcal{Q}$ a subset of patches. This error is defined by

$$E_1 = \frac{1}{N}\sum_{i=1}^{N} d_{L^2}(P_i, \mathcal{Q})^2 = \frac{1}{N}\sum_{i=1}^{N} \min_{Q\in\mathcal{Q}} \sum_{x\in\omega} (P_i(x) - Q(x))^2, \tag{18}$$

where ω is the patch domain. One hopes that using a DPP to generate $\mathcal{Q}$ will prevent from concentrating only on the most common patches and select singular patches. Given an expected cardinality $n \leq N$ and a kernel K_n, we would consider $\mathcal{Q} \sim \mathrm{DPP}(K_n)$. The following error can be useful to verify this property:

$$E_2 = \max_{i\in\{1,\dots,N\}} d_{L^2}(P_i, \mathcal{Q})^2 = \max_{i\in\{1,\dots,N\}} \min_{Q\in\mathcal{Q}} \sum_{x\in\omega} \left(P_i(x) - Q(x)\right)^2. \tag{19}$$

A low error value asserts that the outlier patches (non redundant) are selected.

Given an expected cardinality $n \leq N$ and a kernel K_n, we will consider $\mathcal{Q} \sim \mathrm{DPP}(K_n)$. We would like to find the DPP kernel minimizing the expectation of the errors: $\mathbb{E}_{\mathcal{Q}\sim\mathrm{DPP}(K_n)}(E_1)$ and $\mathbb{E}_{\mathcal{Q}\sim\mathrm{DPP}(K_n)}(E_2)$. Yet, this optimization problem depending on a DPP matrix K_n is intractable. As in the papers by Kulesza and Taskar [34] and Affandi [1], we would like to have a closed-form minimization problem to obtain optimal parameters. These strategies are based on the quality-diversity decomposition of an L-ensemble kernel described in the previous section. Given predetermined features vectors, they determine an appropriate quality measures from the data. Here, we use a similar parametrization, using the first definition of DPPs, with a kernel matrix K. We suppose that its eigenvectors are fixed (given by features computed from the patches of the image) and we want to determine the optimal spectrum so that the associated matrix K minimizes a tractable error. Furthermore, thanks to the Campbell Formula (6), we know that the expectation of some functionals defined on point processes are tractable. That is what we use in the following.

Suppose we select a subset of patches using a DPP of kernel K: $\mathcal{Q} \sim \mathrm{DPP}(K)$. We would like to study the following measure:

$$R(\mathcal{Q}) = \sum_{P\in\mathcal{P}} \sum_{Q\in\mathcal{Q}} f_P(Q). \tag{20}$$

It can be seen as a reconstruction evaluation, if the function f_P involves a distance between the input patch and the patch P. With the appropriate function f_P, R can represent how well a patch $P \in \mathcal{P}$ is represented by the selection $\mathcal{Q}$. For instance, by considering the functions $f_{\alpha,P}(Q) = \mathbf{1}_{\|P-Q\|^2\leq\alpha}$ or $f_P(Q) = e^{-\|P-Q\|^2}$, R will return a high value if the selection $\mathcal{Q}$ encompasses the set of patches. Notice that if we use a function f_p which depends on the L^2 distance between patches, maximizing R will favor selections similar to the ones minimizing the MSE. Thus, contrary to the previous error quantities, E_1 and E_2, we want to generate a subset $\mathcal{Q}$ such that R is large. From the Campbell Formula (6) adapted to general discrete DPPs, we have

$$\mathbb{E}(R(\mathcal{Q})) = \mathbb{E}\left(\sum_{P\in\mathcal{P}}\sum_{Q\in\mathcal{Q}} f_P(Q)\right) = \sum_{j=1}^{N} \mathbb{E}\left(\sum_{Q\in\mathcal{Q}} f_{P_j}(Q)\right) = \sum_{j=1}^{N}\sum_{i=1}^{N} f_{P_j}(P_i)K(P_i, P_i).$$

Assume that K admits the eigendecomposition

$$K(P_i, P_j) = \sum_{k=1}^{D} \lambda_k \phi_k(P_i) \phi_k^*(P_j), \tag{21}$$

with $D \leq N$, and unknown eigenvalues $(\lambda_k)_{k\in\{1,\cdots,D\}}$. We suppose that the eigenvectors are fixed. In the experiments, $(\phi_k)_{k\in\{1,\cdots,D\}}$ are the same feature vectors as defined in the previous subsection, given by the discrete cosine transform of each patch. Then the previous expectation becomes

$$\mathbb{E}(R(\mathcal{Q})) = \sum_{k=1}^{D} \lambda_k \sum_{i=1}^{N} |\phi_k(P_i)|^2 \sum_{j=1}^{N} f_{P_j}(P_i) \tag{22}$$

The maximization of this quantity with respect to $(\lambda_1, \ldots, \lambda_D)$ is a linear problem, under the linear constraints: $\sum_{P\in\mathcal{P}} K(P,P) = \sum_{k=1}^{D} \lambda_k = n$, and for all $k \in \{1, \ldots, D\}$, $0 \leq \lambda_k \leq 1$. The advantage of solving such a problem is that the solution $(\lambda_k^*)_{k\in\{1,\cdots,D\}}$ is explicit. It is on the boundary of the constraints, meaning that is a kernel K with only n non-zero eigenvalues, each one equal to 1: K is a projection DPP. Given any function f_p, any integer $n \leq D$, let us consider I_n the set of the indices associated to the n largest coefficients of the vector ψ of size D defined by $\psi_k = \sum_{i=1}^{N} |\phi_k(P_i)|^2 \sum_{j=1}^{N} f_{P_j}(P_i)$. The solution of the problem

$$\underset{(\lambda_k)}{\operatorname{argmax}} \mathbb{E}\left(R(\mathcal{Q})\right) \text{ such that } \sum_{k=1}^{D} \lambda_k = n \text{ and } \forall k,\ 0 \leq \lambda_k \leq 1, \tag{23}$$

is the set of eigenvalues $(\lambda_k^*)_{k=1,\ldots,D}$ defined by

$$\lambda_k^* = \begin{cases} 1 & \text{if } k \in I_n \\ 0 & \text{otherwise} \end{cases}. \tag{24}$$

For instance, if we choose $f_{\alpha,P_i}(P_j) = \mathbf{1}_{\|P_i - P_j\|^2 \leq \alpha}$, then we need to maximize the function

$$\mathbb{E}(R(\mathcal{Q})) = \sum_{k=1}^{D} \lambda_k \sum_{i=1}^{N} |\phi_k(P_i)|^2 \sum_{j=1}^{N} \mathbf{1}_{\|P_i - P_j\| \leq \alpha} = \sum_{k=1}^{D} \lambda_k \sum_{i=1}^{N} |\phi_k(P_i)|^2 |\mathcal{B}(P_i, \alpha)|, \tag{25}$$

where $\mathcal{B}(P, \alpha)$ is the ball with center P and radius α for the Euclidean distance between patch intensities, and $|A|$ is the cardinality of the subset A. Thus, $|\mathcal{B}(P_i, \alpha)|$ denotes the number of patches in the image that are within a distance of P_i smaller than α. In the experiments, we use this function and we choose α to be half the median of interdistances between patches. Note that this maximization problem generates a kernel called the optimized kernel that will favor patches similar to many others. This creates an interesting compromise: the DPP will tend to select diverse subsets of redundant patches. As anticipated, we will see in the experiments that this method tends to miss singular patches.

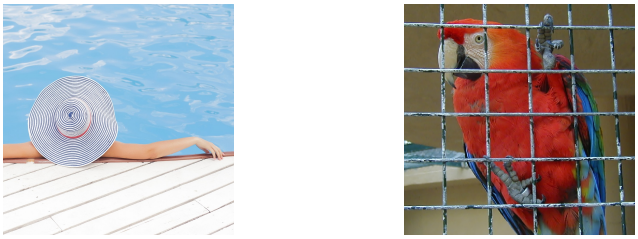


Figure 9: Original images considered in Figures 10 to 12.

5.3 Experiments

The following figures present some results of subsampling in the space of image patches, for different cardinality. First notice that the cardinality is fixed for the uniform sampling. It is also fixed for the last optimized kernel, as we obtain a projection kernel from the maximization problem. Concerning the three other kernels, they are defined using the L-ensemble definition in Equations (15), (16) and (17). We used a common normalization strategy, formalized in [6], using a L-ensemble kernel L whose eigenvalues are denoted $(\lambda_k)_{k\in\{1,\dots,N\}}$. Given a desired expected cardinality n, we normalize L to obtain a kernel $L_c = cL$, where c is chosen such that $\sum_{k=1}^{N} \frac{c\lambda_k}{1+c\lambda_k} = n$. Note also that the Qual-div kernel and the optimized kernel are low-rank, with a rank equal at most to the number of features that we use to defined the kernels. In these experiments, the feature vector associated to each patch (ϕ in Equations (17) and (5.2)) is obtained from the discrete cosine transform of the patch. Note that a DPP kernel can't generate samples with more items than its rank and in the following experiments, we use patches of size $7 \times 7 \times 3$. Thus, the rank of the two previous kernels is 147 and we can observe the results, with a step of 50, up to a cardinality equal to 100 in Figure 12.

Figures 10 and 11 show images reconstructed using the associated selected patches presented below the reconstruction. Each patch in the initial image is replaced by its nearest neighbor in the DPP selection. The final image is obtained by average: given a pixel, all the overlapping patches containing this pixel are averaged. This is a common strategy to aggregate the patches. Several other methods are proposed in the literature, such as using a weighted average [18, 50] or implicitly including the reconstruction in a global variational problem [64]. An average considering uniform weights on all the patches is often used as it does not require any other computation or information to store. Thus, after subsampling the set of patches, the initial image can be represented by its size $N_1 \times N_2 = N$, the small set of patches of size $(2\rho+1) \times (2\rho+1) \times d$ and a vector of indices of length N, associating each initial patch to its nearest neighbor in the selection. Figure 12 compares the errors E_1 (18), E_2 (19) and the peak signal-to-noise ratio (PSNR) of the reconstruction images generated from samples given by the different kernels. The PSNR is a metric commonly used to evaluate the quality of the reconstruction of an image. Consider an initial image I_0 and a reconstruction I_1, both having d color channels and N pixels with a value between 0

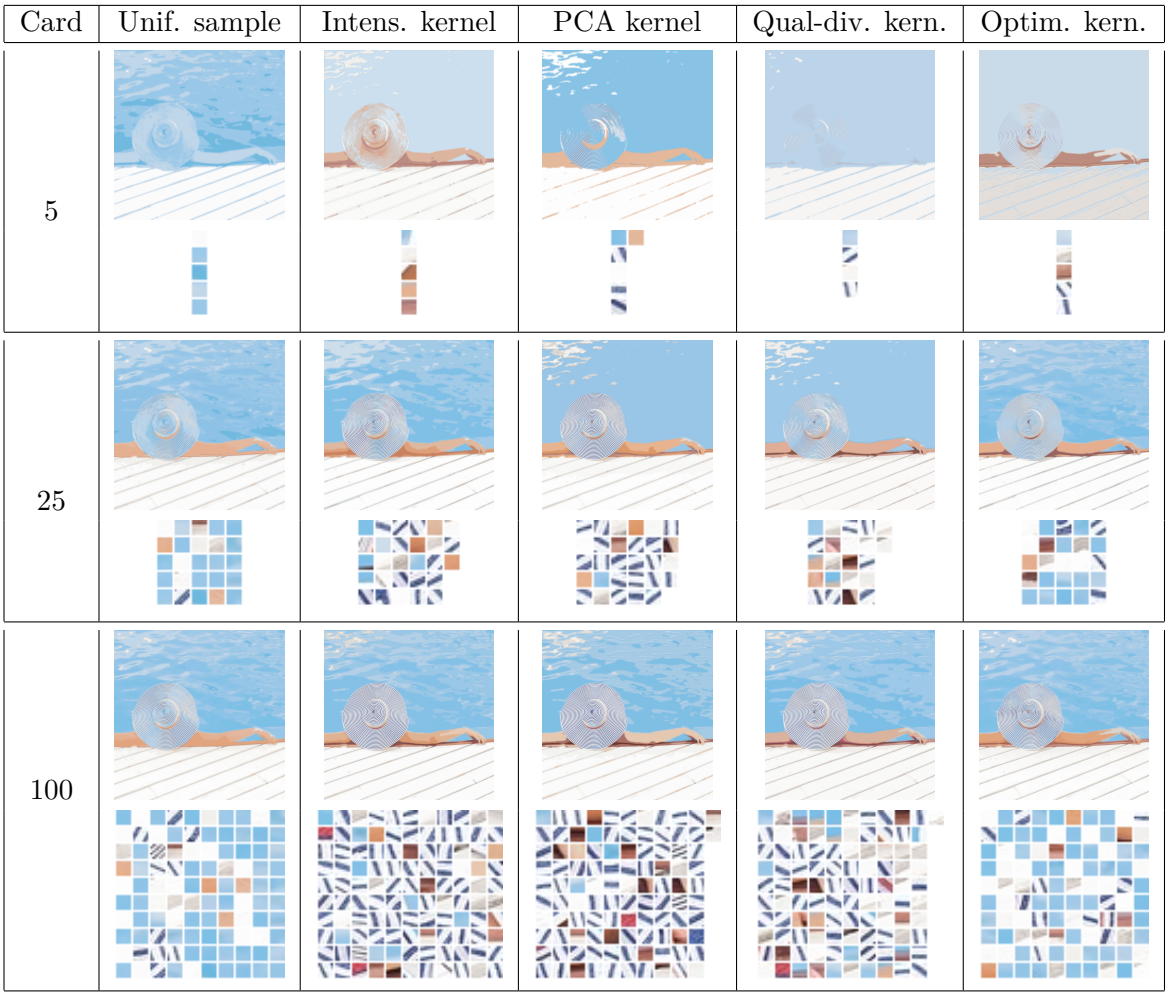


Figure 10: Image reconstruction comparing different expected cardinality and the DPP kernels presented in the previous subsections. For each cardinality, the first row presents the reconstruction of the image using only the patches selected by the corresponding kernel, given in the second row.

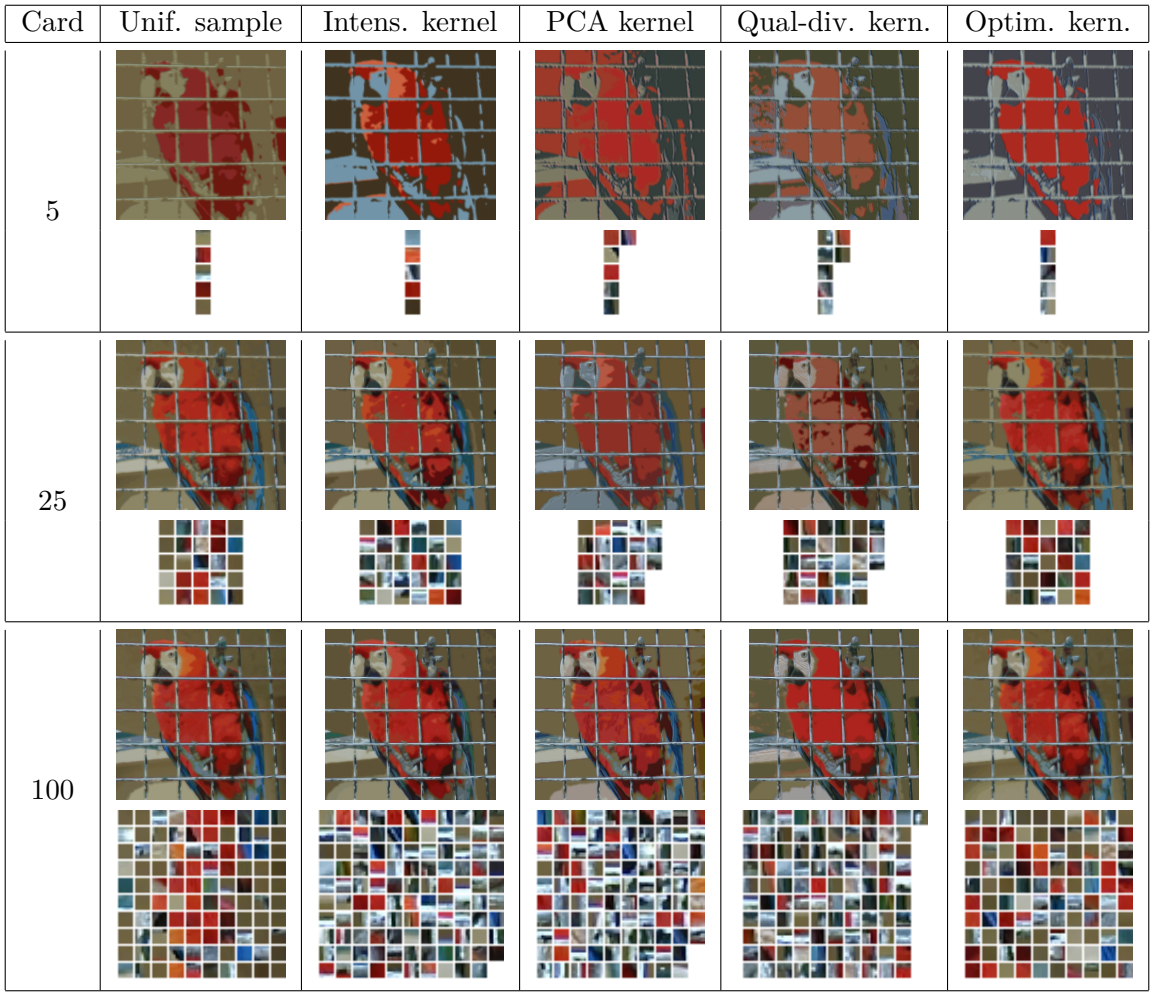
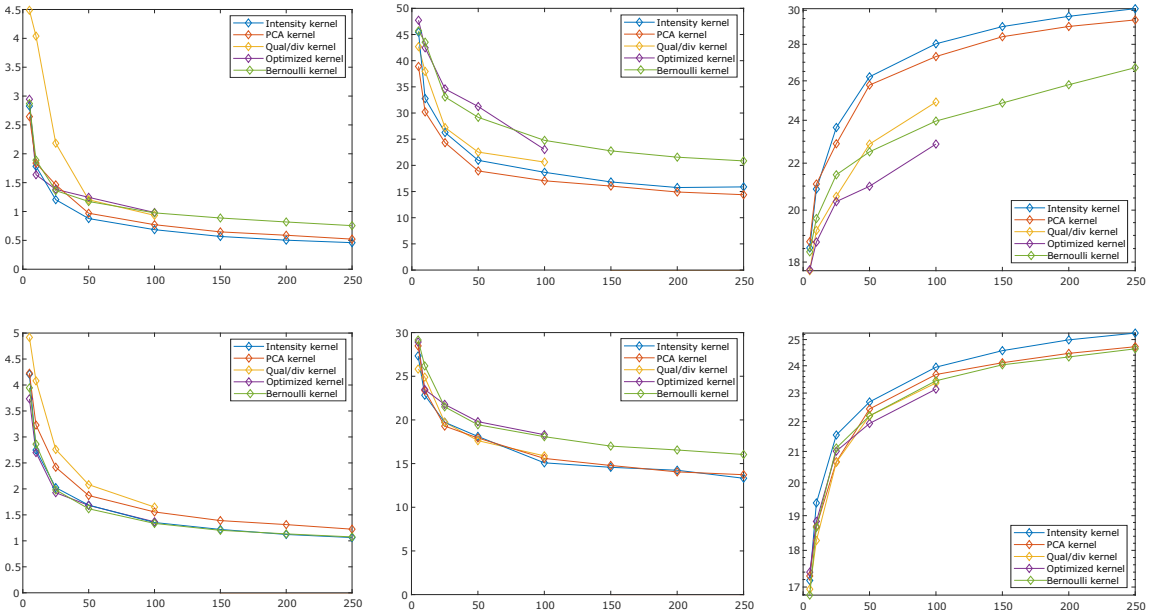


Figure 11: Same as Figure 10 for the Parrot image.



(a) $E_1 = \frac{1}{N}\sum_{i=1}^{N} \|P_i - \mathcal{Q}\|_2^2$ (b) $E_2 = \max_{i\in\{1,\dots,N\}} \|P_i - \mathcal{Q}\|_2^2$ (c) PSNR

Figure 12: Reconstruction errors E_1 and E_2 and the PSNR for the Pool image (top) and the Parrot image (bottom), comparing several DPP kernels and a uniform selection (Bernoulli kernel) in function of different expected cardinality, from 5 to 250. Note that the curves associated to the "Qual/div" and the optimized kernels stop at an expected cardinality equal to 100 selected patches since, for these two kernels, the number of features chosen to describe the patches (equal to 147 in the experiments) determines the maximal size of the generated subset of patches.

and 1. Then,

$$\mathrm{PSNR} = 10 \log_{10} \frac{Nd}{\sum_{c=1}^{d} \sum_{i=1}^{N} (I_0(i,c) - I_1(i,c))^2}.$$

First, as expected, a uniform sampling can produce samples which contains many similar patches. The first image (Pool) has several large and regular regions that could be represented by a few patches and these regions are often over-represented in the results. Note that when we compare the kernels using the error E_1, in particular for the second image (Parrot), the uniform selection provides satisfying results. On the contrary, small and rare details are often missed by the uniform sampling, and the second graph of Figure 12 shows that this sampling strategy compares badly with the others when considering this criteria. Furthermore, the graph presenting the PSNR results illustrates how this uniform strategy provides overall poorer reconstructed images.

Note that the optimized kernel, making a compromise between the diversity induced by DPPs and the redundancy imposed by maximizing the chosen reconstruction error (20), produces quantitative results similar to a uniform sampling. When observing the patches selected by this kernel in Figures 10 and 11, one can see that this kernel tends to select slightly more diverse patches than a uniform sampling.

Second, the PCA kernel and the Qual-div kernel behave rather similarly. They tend to favor singular patches and patches containing edges, even sometimes over-representing them. Thus, they provide good results when looking at the second error measuring the distance between the selection and the furthest patch, especially the PCA kernel. Yet, they can provide even worst results than the uniform selection when we look at the average distance between the selection and the initial set of patches (Error E_1 (18)).

Finally, the Intensity kernel, using only the squared Euclidean distance between intensities, seems to be the most stable kernel. It provides small average error and tends to include singular patches in the selection. For both images, whatever the expected cardinality, the samples generated by this kernel produce visually satisfying reconstructions.

Thus, the choice of subsampling strategy in the patch space of an image highly depends on the purpose of the generated selection. The most stable strategy seems to be using the Intensity kernel (15), which provides a selection close in average to the initial patches and which selects also singular patches. If the priority of the application is efficiency, the best strategy may remain to use a uniform selection with a high number of selected patches. Yet, if the size of the selection needs to be low or if the selection needs to contain mainly structure and texture information, the good choice may be to use a PCA kernel or a kernel using the quality-diversity decomposition. Moreover, note that when the kernel is low-rank, it is also possible to use a dual representation of the kernel and to speed up the simulation of the process [34, 25]. In this paper, we present several low-rank kernels: the quality-diversity kernel and the optimized kernel. In these cases, the computational time can be reduced.

6. Conclusion

In this paper, we presented several applications of determinantal point processes to image processing. We defined these models on the pixels of an image and we call such point

processes determinantal pixel processes (DPixP). In this setting, we showed that the only possible hard core repulsion for DPixP is directional. Given a direction, it is possible to impose to select at most one pixel on any discrete line in the image, but any further hard core constraint leads to a degenerate kernel. We studied shot noise models based on DPixP as a method to sample homogeneous micro-textures and adapted the choice of DPixP kernel in function of a given spot function of the shot noise and of the homogeneity we look for. We also developed an algorithm to infer the Fourier coefficients of a DPixP kernel from a sample or from a set of samples. As a future work, we plan to investigate the estimation, from a texture image, of the spot function and of the DPixP kernel associated to a shot noise that could have generated the texture. We also considered DPPs defined on the set of patches of an image, in order to subsample this set while retaining the information necessary to appreciate the image. It would be interesting to use more features to describe the patches of an image, in order to improve the results of the Quality-diversity kernel and nuance the selection it generates, which for now tends to over-represent patches with textures or edges. We also would like to explore different functions used in the minimization of the selection error, depending on the purpose of the subsampling of the set of patches. At last, we intend to apply these selection strategies to patch-based methods and evaluate the gain in quality and computation time generated by the use of DPPs.

Appendix A. DPixP kernels as block-circulant matrices with circulant blocks

Recall that a $N \times N$ circulant matrix K is entirely characterized by its generating elements $c_0, \dots, c_{N-1}$ such that $K = \text{circ}(c_0, \dots, c_{N-1})$, in the sense that $K(x, y) = c_{(x-y \mod N)}$:

$$K = \text{circ}(c_0, \dots, c_{N-1}) = \begin{pmatrix} c_0 & c_{N-1} & \dots & c_1 \\ c_1 & c_0 & \dots & c_{N-2} \\ \vdots & & \ddots & \vdots \\ c_{N-1} & c_{N-2} & \dots & c_0 \end{pmatrix}. \tag{26}$$

Similarly, a $N_1N_2 \times N_1N_2$ block-circulant matrix K is entirely characterized by its N_2 generating matrices of size $N_1 \times N_1$, $K = \text{circ}(K_1, K_2, \dots, K_{N_2})$. If each generating matrix K_i is circulant, K is called a block-circulant matrix with circulant blocks. In that case, the matrix K can be characterized by the generating elements of all the generating matrices.

In our framework, the pixels are considered column by column so that the ordered index $\underline{x}$ of a pixel $x = (x_1, x_2) \in \Omega$ is $\underline{x} = x_1 + 1 + x_2 N_1$. Moreover, we suppose that the process is stationary and periodic. Hence we can define a function $C : \Omega \to \mathbb{C}$ such that $\forall x, y \in \Omega$, $K(\underline{x}, \underline{y}) = C(x - y)$, extended by periodicity to $\mathbb{Z}^2$. In this configuration, the matrix K is block-circulant with circulant blocs. To simplify the notations in the paper, we chose to identify the pixel's position in the image and its ordering.

A.1 Illustration on a small image domain

Let us present a simple illustration.

Example 1 *Consider* $\Omega = \{0,1\} \times \{0,1,2\}$*, with* $N_1 = 2$ *and* $N_2 = 3$*. The image domain is*

$$\Omega = \begin{array}{|c|c|c|} \hline (0,0) & (0,1) & (0,2) \\ \hline (1,0) & (1,1) & (1,2) \\ \hline \end{array}.$$

This image domain is associated to the following ordering of pixels

$$\underline{\Omega} = \begin{array}{|c|c|c|} \hline 1 & 3 & 5 \\ \hline 2 & 4 & 6 \\ \hline \end{array}.$$

A kernel matrix K *defined on such image domain will be defined using the previous ordering. Consider the function* C *defined on* Ω *and extended by periodicity to* $\mathbb{Z}^2$ *as*

$$\forall x, y \in \Omega, \quad K(\underline{x}, \underline{y}) = C(x - y).$$

Moreover, due to the periodicity assumption, we know for instance that $C((-1,-2)) = C((1,1))$ *or* $C((-1,-1)) = C((1,2))$*. We obtain the following kernel, which is indeed a* 6×6 *circulant matrix with* 2×2 *circulant blocks:*

$$K = \begin{pmatrix} C((0,0)) & C((1,0)) & C((0,2)) & C((1,2)) & C((0,1)) & C((1,1)) \\ C((1,0)) & C((0,0)) & C((1,2)) & C((0,2)) & C((1,1)) & C((0,1)) \\ C((0,1)) & C((1,1)) & C((0,0)) & C((1,0)) & C((0,2)) & C((1,2)) \\ C((1,1)) & C((0,1)) & C((1,0)) & C((0,0)) & C((1,2)) & C((0,2)) \\ C((0,2)) & C((1,2)) & C((0,1)) & C((1,1)) & C((0,0)) & C((1,0)) \\ C((1,2)) & C((0,2)) & C((1,1)) & C((0,1)) & C((1,0)) & C((0,0)) \end{pmatrix}.$$

It can be entirely characterized by its first column.

A.2 Diagonalization of a DPixP kernel

The following results on the diagonalization of circulant matrices and its generalization on block-circulant matrices with circulant blocks can be found in the book of Davis, *Circulant matrices* [19]. See [19, 45] for more details and illustrations. First, it is well known that any circulant matrix $K = \mathrm{circ}(c_0, \dots, c_{N-1})$ is diagonalized in the Fourier basis and its eigenvalues $\lambda_0, \dots, \lambda_{N-1}$ are given by

$$\forall\, \xi = 0, \dots, N-1, \quad \lambda_\xi = \sum_{x=0}^{N-1} c_x e^{-2i\pi \frac{x\xi}{N}}.$$

Moreover, if the matrix K is Hermitian, its eigenvalues are real and most of them are constrained [19]: λ_0 is distinct but the remaining eigenvalues are such that $\lambda_\xi = \lambda_{N-\xi}$, if N is odd. If N is even, $\lambda_{N/2}$ is also distinct.

One can obtain a similar result concerning block circulant matrices with circulant blocks. Consider $\omega_N = e^{\frac{2i\pi}{N}}$, the primitive N-th roots of unity, and the following $N \times N$ unitary matrix

$$U_N = \frac{1}{\sqrt{N}} \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \omega_N & \omega_N^2 & \dots & \omega_N^{N-1} \\ 1 & \omega_N^2 & \omega_N^4 & \dots & \omega_N^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_N^{N-1} & \omega_N^{2(N-1)} & \dots & \omega_N^{(N-1)^2} \end{pmatrix}. \tag{27}$$

This matrix is associated to the inverse discrete Fourier transform.

Proposition 20 (Diagonalization of a DPixP kernel) *Consider a DPixP defined on Ω with kernel K, the block-circulant matrix with circulant blocks whose first column is $C : \Omega \to \mathbb{C}$. Then, K is diagonalized in the Fourier basis, in the sense that*

$$(U_{N_2}^* \otimes U_{N_1}^*) K (U_{N_2} \otimes U_{N_1}) = \Lambda,$$

where $\otimes$ is the Kronecker product, and where Λ is diagonal such that

$$\forall \xi \in \Omega, \quad \Lambda_{\xi,\xi} = \widehat{C}(\xi) = \sum_{x \in \Omega} C(x) e^{-2i\pi \left(\frac{x_1 \xi_1}{N_1} + \frac{x_2 \xi_2}{N_2} \right)}.$$

Proof Let us prove that we have $K(U_{N_2} \otimes U_{N_1}) = (U_{N_2} \otimes U_{N_1})\Lambda$. Note that $\forall x, \xi \in \Omega$, $(U_{N_2} \otimes U_{N_1})_{x,\xi} = \frac{1}{\sqrt{N_1 N_2}} e^{2i\pi \left(\frac{x_1 \xi_1}{N_1} + \frac{x_2 \xi_2}{N_2} \right)} = \frac{1}{\sqrt{N}} e^{2i\pi \langle x, \xi \rangle}$, with the notation $\langle \cdot \rangle$ that we use throughout the paper. Given $x, \xi \in \Omega$, we have

$$\begin{aligned} \Big[K(U_{N_2} \otimes U_{N_1}) \Big]_{x,\xi} &= \frac{1}{\sqrt{N}} \sum_{y \in \Omega} C(x-y) e^{2i\pi \langle y, \xi \rangle} = \frac{1}{\sqrt{N}} \sum_{y \in \Omega} C(y) e^{2i\pi \langle x-y, \xi \rangle} \\ &= (U_{N_2} \otimes U_{N_1})_{x,\xi} \, \widehat{C}(\xi) = \Big[(U_{N_2} \otimes U_{N_1}) \Lambda \Big]_{x,\xi}, \end{aligned}$$

where $\Lambda = \operatorname{diag}\Big((\widehat{C}(\xi), \xi \in \Omega) \Big)$. ■

A necessary and sufficient condition for the DPixP to be defined is that the eigenvalues of the kernel are in $[0, 1]$. In particular, the Fourier coefficients of C are real. Thus we have

$$\forall x \in \mathbb{Z}^2, C(-x) = \overline{C(x)}.$$

Example 2 *Let us use the same illustration as in Example 1. We can deduce from the previous remark and from the periodic boundaries that the kernel is of the shape*

$$K = \begin{pmatrix} C((0,0)) & C((1,0)) & \overline{C((0,1))} & \overline{C((1,1))} & C((0,1)) & C((1,1)) \\ C((1,0)) & C((0,0)) & \overline{C((1,1))} & \overline{C((0,1))} & C((1,1)) & C((0,1)) \\ C((0,1)) & C((1,1)) & C((0,0)) & C((1,0)) & \overline{C((0,1))} & \overline{C((1,1))} \\ C((1,1)) & C((0,1)) & C((1,0)) & C((0,0)) & \overline{C((1,1))} & \overline{C((0,1))} \\ \overline{C((0,1))} & \overline{C((1,1))} & C((0,1)) & C((1,1)) & C((0,0)) & C((1,0)) \\ \overline{C((1,1))} & \overline{C((0,1))} & C((1,1)) & C((0,1)) & C((1,0)) & C((0,0)) \end{pmatrix}.$$

Note that any DPixP kernel defined on a grid Ω of size 2×3 is entirely characterized by four coefficients.

Appendix B. Identifiability of a DPixP

B.1 Case 2: K_1 is irreducible and doesn't verify the rank hypothesis (Theorem 16)

Let us study the equivalence class of a DPixP of kernel C_1 such that its associated matrix K_1 is irreducible and there exists a partition $\alpha,\ \beta$ of $\mathcal{Y}$ such that $\text{rank}(K_1)_{\alpha\times\beta} = 1$, in the case Ω of size 1×5:

$$K_1 = \text{circ}\left(C_1(0), C_1(1), C_1(2), \overline{C_1(2)}, \overline{C_1(1)}\right)$$

Define $r_{11}, \theta_{11}, r_{12}, \theta_{12}$ the respective modulus and argument of $C_1(1)$ and $C_1(2)$. Whatever α, β, the partition of $\mathcal{Y}$ such that $\text{rank}(K_1)_{\alpha\times\beta} = 1$, due to rows proportionality, one obtains $r_{11} = r_{12}$ and $\theta_{12} = -3\theta_{11} \mod 2\pi$. Now, assume that C_2 is an admissible DPixP kernel such that $\text{DPixP}(C_2) = \text{DPixP}(C_1)$. Then the matrices K_1 and K_2 have equal principal minors. Necessarily, K_2 is irreducible and there exists a partition such that $\text{rank}(K_2)_{\alpha\times\beta} = 1$, otherwise K_2 would verify the assumptions of Theorem 16 and so would K_1. Then, as C_1, C_2 is fully determined by $C_2(0)$, one modulus r_{21} and one argument θ_{21}. Once again, we know that $C_1(0) = C_2(0) = C_0$ and thanks to the equality of principal minors of size 2, the modulus are equal so $r_{21} = r_{11} = r$. One of the principal minors of size 3 for C_1 is equal to

$$C_0^3 + \overline{C_1(1)C_1(1)}C_1(2) + C_1(1)C_1(1)\overline{C_1(2)} - C_0C_1(2)\overline{C_1(2)} - 2C_0C_1(1)\overline{C_1(1)},$$

so by equality of principal minors, we obtain

$$\begin{aligned}
&\text{Re}\left(C_1(1)C_1(1)\overline{C_1(2)}\right) = \text{Re}\left(C_2(1)C_2(1)\overline{C_2(2)}\right)\\
&\Leftrightarrow \text{Re}\left(r^3 e^{2i\theta_{11}+3i\theta_{11}}\right) = \text{Re}\left(r^3 e^{2i\theta_{21}+3i\theta_{21}}\right)\\
&\Leftrightarrow r^3 cos(5\theta_{11}) = r^3 cos(5\theta_{21})\\
&\Leftrightarrow \exists\ k \in \mathbb{Z} \text{ s.t. } \theta_{11} = \begin{cases}\theta_{21} + \frac{2}{5}k\pi \text{ (case 1)}\\ -\theta_{21} + \frac{2}{5}k\pi \text{ (case 2).}\end{cases}
\end{aligned}$$

Finally, let us assume we are in the first case, K_1 can be written

$$K_1 = \text{circ}\left(C_0, re^{i(\theta_{21}+\frac{2}{5}k\pi)}, re^{-3i(\theta_{21}+\frac{2}{5}k\pi)}, re^{3i(\theta_{21}+\frac{2}{5}k\pi)}, re^{-i(\theta_{21}+\frac{2}{5}k\pi)}\right) = DK_2D^{-1}$$

with $D = \text{diag}\left(1, e^{i\frac{2}{5}k\pi}, e^{i\frac{4}{5}k\pi}, e^{-i\frac{4}{5}k\pi}, e^{-i\frac{2}{5}k\pi}\right)$, which corresponds to a translation of the Fourier coefficients of C of k pixels. The second case yields to $K_1 = D\overline{K_2}D^{-1}$ which corresponds to the symmetry and the translation of k pixels of the Fourier coefficients of C.

B.2 Case 3: K_1 is not irreducible

In this appendix, we consider a Hermitian block-circulant matrix K of size $N \times N$ that is completely reducible, meaning that it is permutation similar to a block diagonal matrix with irreducible blocks. We want to prove that in that case, the blocks are identical, that is they are of equal size and they are composed of the same coefficients. Moreover, we prove that these blocks are not only irreducible but also Hermitian and circulant.

First, let us study the 1D case, meaning that K is a kernel defined on the points of $\mathcal{Y} = \{0, \dots, N-1\}$ (to be consistant with our 2D representation) and it is circulant. Therefore, for all $i, j \in \mathcal{Y}$, $K(i,j) = c_{j-i} = \overline{c_{i-j}}$. As K is not irreducible, there exist $i, j \in \mathcal{Y}$, such that $K(i,j) = c_{j-i} = 0$. Let us denote $k = \inf\{l > 0 \text{ such that } c_l \neq 0\}$, hence $c_1 = \cdots = c_{k-1} = 0 = c_{-1} = \cdots = c_{-k+1}$. Notice that k is necessarily larger or equal to 2, otherwise by the circulant property, K would not have any zero coefficient and it would be irreducible. Similarly, k necessarily divides N and the only non-zero coefficients c_m are multiples of k, as otherwise the non-zero elements of K would be located such that it would be possible to access to any index from any other by traveling only through non-zero coefficients: K would be irreducible. Then, if we define l such that $N = k \times l$, there are k cycles of size l in the graph associated to K, each block with the same l coefficients $\{c_k, c_{2k}, \dots, c_{lk}\}$, or equivalently,

$$\forall i_0 = 0, \dots, N-1, \quad K(i_0,\, j) = \begin{cases} c_{kp}, & \text{if } j = kp + i_0 \mod N, \text{ with } p = 0, \dots, l-1, \\ 0, & \text{otherwise.} \end{cases}$$

Thus it is possible to define the permutation matrix P which gathers the cycles, and which associates K with a block diagonal matrix:

$$\forall p = 0, \dots, l-1, \quad \forall r = 0, \dots, k-1, \quad P(p + lr, r + pk) = 1$$

meaning that P moves the coefficient K_{r+pk} to the coordinate $p + lr$. Moreover, these blocks $(B_r)_{r \in \{0,\dots,k-1\}}$ are circulant: for all $r = 0, \dots, k-1$, for all $i, i' = 0, \dots, l-1$,

$$B_r(i, i') = K(r + ik, r + i'k) = c_{(i-i')k},$$

for all $\tau \in \mathcal{Y}$ such that $(i + \tau \mod N)$ and $(i' + \tau \mod N)$ are in the r-th cycle,

$$B_r(i + \tau, i' + \tau) = K(r + (i+\tau)k, r + (i'+\tau)k) = c_{(i-i')k} = B_r(i, i').$$

To conclude, K is permutation similar to a block-diagonal matrix, which is the repetition of one irreducible, circulant and Hermitian block.

Now let us consider the 2D case, when K is a kernel matrix defined on $\Omega = \{0, \dots, N_1 - 1\} \times \{0, \dots, N_2 - 1\}$ and assume that K is Hermitian, block-circulant with circulant blocks and completely reducible. Define C the function such that for all $(i,j), (i',j') \in \Omega$, $K\left((i,j),(i',j')\right) = C(i'-i, j'-j)$. As in the 1D case, define $(e_1, e_2) \in \mathbb{Z}^2 \cap \Omega$ the two generating vectors such that $C(r,s) = 0, \forall (r,s)$ inside the elementary cell generated by (e_1, e_2). These two vectors generate a subgroup of $\mathbb{Z}^2$ and it contains $\mathbb{Z}(0, N_2) + \mathbb{Z}(N_1, 0)$, as K is not irreducible and similarly as in the 1D case. Then e_1 divides N_1, e_2 divides N_2. As before, the only non-zero coefficients of C belong to $\{\mathbb{Z}e_1 + \mathbb{Z}e_2\} \cap \Omega$. The size of the elementary cell determines the number of cycles (and future blocks) and $l = \sharp\{\mathbb{Z}e_1 + \mathbb{Z}e_2\} \cap \Omega$ defines the size of each cycle. It is possible to define the permutation matrix that transforms K into a block-diagonal matrix with irreducible blocks. For all $(i,j) \in \Omega$, let us define (r,s) its representative element in the elementary cell such that there exists p, q such that $(i,j) = (pe_1 + qe_2) + (r,s)$ mod (N_1, N_2). Thus, P associates the coefficient $(i,j) = (pe_1 + qe_2) + (r,s) \bmod (N_1, N_2)$ with the coordinate $(p,q) + (r,s)$ (block (r,s), coefficient (p,q)). As before, the blocks

$(B_{(r,s)})$ have the same size and have an identical structure. Let us consider the block (r,s), consider $(i,j),(i',j') \in \Omega$,

$$\begin{aligned} B_{(r,s)}((i,j),(i',j')) &= K((pe_1+qe_2)+(r,s) \bmod (N_1,N_2),(p'e_1+q'e_2)+(r,s) \bmod (N_1,N_2)) \\ &= C((p'-p)e_1+(q'-q)e_2) \end{aligned}$$

Let $(\tau_1,\tau_2) \in \Omega$ be such that $(i+\tau_1,j+\tau_2),(i'+\tau_1,j'+\tau_2)$ both belong to the cycle (r,s). Then $(\tau_1,\tau_2) \in \mathbb{Z}e_1+\mathbb{Z}e_2$, we can write $(\tau_1,\tau_2)=t_1e_1+t_2e_2$.

$$\begin{aligned} &B_{(r,s)}((i+\tau_1,j+\tau_2),(i'+\tau_1,j'+\tau_2)) \\ &\quad = K((pe_1+qe_2)+(r,s)+(t_1e_1+t_2e_2) \bmod (N_1,N_2), \\ &\qquad (p'e_1+q'e_2)+(r,s)+(t_1e_1+t_2e_2) \bmod (N_1,N_2)) \\ &\quad = C((p'-p)e_1+(q'-q)e_2) = B_{(r,s)}((i,j),(i',j')). \end{aligned}$$

Thus, for all (r,s), the associated bloc $B_{(r,s)}$ is block circulant with circulant blocks. Similarly, it is Hermitian. To conclude, K is permutation similar to a block diagonal matrix with one repeated irreducible, circulant, Hermitian block.

Acknowledgments

This work was supported by grants from Région Ile-de-France. We thank the reviewers for their valuable comments and suggestions that helped us to improve the paper.

References

[1] R. H. Affandi, E. B. Fox, R. P. Adams, and B. Taskar. Learning the parameters of determinantal point process kernels. In ICML, volume 32 of JMLR Workshop and Conference Proceedings, pages 1224–1232. JMLR.org, 2014.

[2] A. Agarwal, A. Choromanska, and K. Choromanski. Notes on using determinantal point processes for clustering with applications to text clustering. CoRR, abs/1410.6975, 2014.

[3] C. C. Aggarwal. Outlier Analysis. Springer Publishing Company, Incorporated, 2nd edition, 2016.

[4] F. Baccelli and B. Blaszczyszyn. Stochastic Geometry and Wireless Networks, Volume I - Theory, volume 1 of Foundations and Trends in Networking Vol. 3: No 3-4, pp 249-449. NoW Publishers, 2009. Stochastic Geometry and Wireless Networks, Volume II - Applications; see http://hal.inria.fr/inria-00403040.

[5] R. Bardenet and M. Titsias. Inference for determinantal point processes without spectral knowledge. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 3393–3401. Curran Associates, Inc., 2015.

[6] S. Barthelmé, P.-O. Amblard, and N. Tremblay. Asymptotic equivalence of fixed-size and varying-size determinantal point processes. Bernoulli, 25(4B):3555–3589, 11 2019.

[7] A. Belhadji, R. Bardenet, and P. Chainais. A determinantal point process for column subset selection. CoRR, abs/1812.09771, 2018.

[8] C. Biscio and F. Lavancier. Quantifying repulsiveness of determinantal point processes. Bernoulli, 22(4):2001–2028, 11 2016.

[9] C. Biscio and F. Lavancier. Contrast estimation for parametric stationary determinantal point processes. Scandinavian Journal of Statistics, 44(1):204–229, 2017.

[10] Christophe A.N. Biscio and Jean-François Coeurjolly. Standard and robust intensity parameter estimation for stationary determinantal point processes. Spatial Statistics, 18:24 – 39, 2016. Spatial Statistics Avignon: Emerging Patterns.

[11] S. Boyd and L. Vandenberghe. Convex Optimization. Cambridge University Press, March 2004.

[12] V. Brunel, A. Moitra, P. Rigollet, and J. Urschel. Rates of estimation for determinantal point processes. In COLT, volume 65 of Proceedings of Machine Learning Research, pages 343–345. PMLR, 2017.

[13] V.-E. Brunel. Learning signed determinantal point processes through the principal minor assignment problem. In Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada., pages 7376–7385, 2018.

[14] Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth Vishnoi. Fair and diverse DPP-based data summarization. In Jennifer Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 716–725. PMLR, 10–15 Jul 2018.

[15] W. Chen, Z. Yang, F. Cao, Y. Yan, M. Wang, C. Qing, and Y. Cheng. Dimensionality reduction based on determinantal point process and singular spectrum analysis for hyperspectral images. IET Image Processing, 13(2):299–306, 2019.

[16] M. Combescure. Block-circulant matrices with circulant blocks, Weil sums, and mutually unbiased bases. II. The prime power case. Journal of mathematical physics, 50(3):032104, 2009.

[17] N. Cressie. Statistics for spatial data. John Wiley & Sons, 2015.

[18] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Transactions on image processing, 16(8):2080–2095, 2007.

[19] Philip J Davis. Circulant matrices. American Mathematical Soc., 2013.

[20] S. Descamps, X. Descombes, A. Béchet, and J. Zerubia. Automatic flamingo detection using a multiple birth and death process. In 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 1113–1116. IEEE, 2008.

[21] G. M. Engel and H. Schneider. Matrices diagonally similar to a symmetric matrix. Linear Algebra and its Applications, 29:131–138, February 1980.

[22] B. Galerne, Y. Gousseau, and J.-M. Morel. Random phase textures: Theory and synthesis. IEEE Trans. Image Process., 20(1):257 – 267, 2011.

[23] B. Galerne, A. Lagae, S. Lefebvre, and G. Drettakis. Gabor noise by example. ACM Trans. Graph., 31(4):73:1–73:9, jul 2012.

[24] B. Galerne, A. Leclaire, and L. Moisan. Texton noise. In Computer Graphics Forum, volume 36, pages 205–218. Wiley Online Library, 2017.

[25] M. Gartrell, U. Paquet, and N. Koenigstein. Low-rank factorization of determinantal point processes. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, pages 1912–1918. AAAI Press, 2017.

[26] B. Gong, W.-L. Chao, K. Grauman, and F Sha. Diverse sequential subset selection for supervised video summarization. In Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pages 2069–2077, 2014.

[27] D. J. Hartfiel and R. Loewy. On matrices having equal corresponding principal minors. j-LINEAR-ALGEBRA-APPL, 58:147–167, April 1984.

[28] M. Held, P. Wolfe, and H. P. Crowder. Validation of subgradient optimization. Mathematical Programming, 6(1):62–88, Dec 1974.

[29] A. Houdard, C. Bouveyron, and J. Delon. High-dimensional mixture models for unsupervised image denoising (HDMI). SIAM Journal on Imaging Sciences, 2018.

[30] J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. Determinantal processes and independence. Probability Surveys, pages 206–229, 2006.

[31] J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. Zeros of Gaussian Analytic Functions and Determinantal Point Processes, volume 51 of University Lecture Series. American Mathematical Society, Providence, RI, 2009.

[32] A. Kulesza. Learning with Determinantal Point Processes. PhD thesis, University of Pennsylvania, 2012.

[33] A. Kulesza and B Taskar. Learning determinantal point processes. In Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, pages 419–427, 2011.

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

[35] Alex Kulesza and Ben Taskar. k-DPPs: Fixed-size determinantal point processes. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, pages 1193–1200, USA, 2011. Omnipress.

[36] A. Lagae, S. Lefebvre, G. Drettakis, and P. Dutré. Procedural noise using sparse Gabor convolution. ACM Transactions on Graphics, 28(3):54–64, 2009.

[37] Claire Launay and Arthur Leclaire. Determinantal patch processes for texture synthesis. In GRETSI 2019, Lille, France, Aug 2019.

[38] F. Lavancier, J. Møller, and E. Rubak. Determinantal point process models and statistical inference. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77(4):853–877, 2015.

[39] F. Lavancier, A. Poinas, and R. Waagepetersen. Adaptive estimating function inference for nonstationary determinantal point processes. Scandinavian Journal of Statistics, n/a(n/a), 2020.

[40] R. Loewy. Principal minors and diagonal similarity of matrices. Linear Algebra and its Applications, 78:23–64, June 1986.

[41] R. Lyons and J. E. Steif. Stationary determinantal processes: Phase multiplicity, Bernoullicity, entropy, and domination. Duke Math. J., 120(3):515–575, 12 2003.

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

[43] B. Mahasseni, M. Lam, and S. Todorovic. Unsupervised video summarization with adversarial LSTM networks. In CVPR, pages 2982–2991. IEEE Computer Society, 2017.

[44] J. Møller and R. Waagepetersen. Statistical Inference and Simulation for Spatial Point Process, volume 100. Chapman and Hall/CRC, Boca Raton, 2003.

[45] Brian J Olson, Steven W Shaw, Chengzhi Shi, Christophe Pierre, and Robert G Parker. Circulant matrices and their application to vibration analysis. Applied Mechanics Reviews, 66(4), 2014.

[46] G. Pagès. Introduction to vector quantization and its applications for numerics. ESAIM: Proceedings and Surveys, 48(1):29–79, 2015. Proceedings of CEMRACS 2013 - Modelling and simulation of complex systems: stochastic and deterministic approaches. : T. Lelièvre et al. Editors.

[47] G. Perrin, X. Descombes, and J. Zerubia. Point processes in forestry : an application to tree crown detection. Research Report RR-5544, INRIA, 2006.

[48] A. Poinas, B. Delyon, and F. Lavancier. Mixing properties and central limit theorem for associated point processes. Bernoulli, 25(3):1724–1754, 2019.

[49] Justin Rising, Alex Kulesza, and Ben Taskar. An efficient algorithm for the symmetric principal minor assignment problem. Linear Algebra and its Applications, 473:126–144, May 2015.

[50] Joseph Salmon and Yann Strozecki. From patches to pixels in non-local methods: Weighted-average reprojection. In 2010 IEEE International Conference on Image Processing, pages 1929–1932. IEEE, 2010.

[51] B. D. Saunders and H. Schneider. Flows on graphs applied to diagonal similarity and diagonal equivalence for matrices. Discrete Mathematics, 24(2):205 – 220, 1978.

[52] T. Shirai and Y. Takahashi. Fermion Process and Fredholm Determinant, pages 15–23. Springer US, Boston, MA, 2000.

[53] 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(2):414–463, 2003.

[54] T. Shirai and Y. Takahashi. Random point fields associated with certain Fredholm determinants II: Fermion shifts and their ergodic and Gibbs properties. Ann. Probab., 31(3):1533–1564, 07 2003.

[55] A. Soshnikov. Determinantal random point fields. Russian Mathematical Surveys, (55):923–975, 2000.

[56] A. Soshnikov. Gaussian limit for determinantal random point fields. The Annals of Probability, 30(1):171–187, 2002.

[57] Marco Stevens. Equivalent symmetric kernels of determinantal point processes. arXiv e-prints, page arXiv:1905.08162, May 2019.

[58] N. Tremblay, S. Barthelmé, and P.-O. Amblard. Optimized algorithms to sample determinantal point processes. CoRR, abs/1802.08471, 2018.

[59] N. Tremblay, S. Barthelmé, and P.-O. Amblard. Determinantal point processes for coresets. Journal of Machine Learning Research, November 2019.

[60] J. Urschel, V. Brunel, A. Moitra, and P. Rigollet. Learning determinantal point processes with moments and cycles. In ICML, volume 70 of Proceedings of Machine Learning Research, pages 3511–3520. PMLR, 2017.

[61] J. J. van Wijk. Spot noise texture synthesis for data visualization. In SIGGRAPH '91, pages 309–318, New York, NY, USA, 1991. ACM.

[62] M. Wilhelm, A. Ramanathan, A. Bonomo, S. Jain, E. H. Chi, and J. Gillenwater. Practical diversified recommendations on Youtube with determinantal point processes. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, pages 2165–2173, New York, NY, USA, 2018. ACM.

[63] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. In Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII, pages 766–782, 2016.

[64] D. Zoran and Y. Weiss. From learning models of natural image patches to whole image restoration. In Proceedings of the 2011 International Conference on Computer Vision, ICCV '11, page 479–486, USA, 2011. IEEE Computer Society.