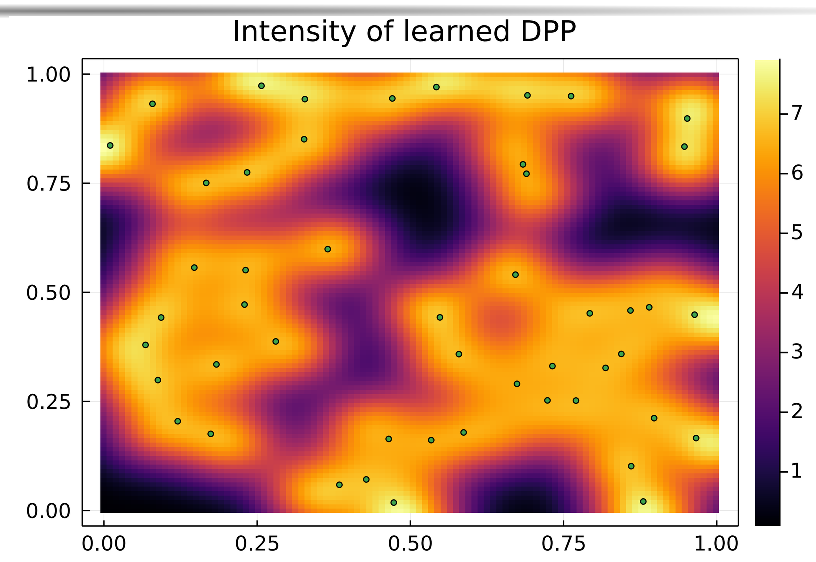
Nonparametric estimation of continuous DPPs with kernel methods

*Michaël Fanuel**, Rémi Bardenet*

*Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, 59000 Lille

What is a DPP ?

A DPP is a point process generating random repulsive point patterns. We learn the DPP from s point patterns $\mathcal{C}_1, \dots, \mathcal{C}_s$. Here, consider only one point pattern $\mathcal{C}$ (dots) and its intensity estimate (color). See →



Correlation functions and correlation kernel

Intuitively, the correlation function at $x_1, \dots, x_\ell$ is

$$\underbrace{\varrho_\ell(x_1, \dots, x_\ell)}_{\text{correlation function}} \approx \frac{\Pr(\text{one point in each } B(x_i, \delta), i = 1, \dots \ell)}{\operatorname{vol}(B(x_1, \delta)) \dots \operatorname{vol}(B(x_\ell, \delta))}.$$

We learn the correlation kernel of the DPP $\mathsf{k}(x, x')$

For a DPP, order-ℓ correlation function (or joint intensity)

$$\varrho_\ell(x_1, \dots, x_m) = \det[\mathsf{k}(x_i, x_j)]_{i,j} \text{ for all } \ell \geq 1.$$

Learning the integral kernel of operator

- Let $\mathcal{X}$ a compact set of $\mathbb{R}^d$. Integral kernels of (integral) operator

$$\mathsf{K}f(x) = \int_{\mathcal{X}} \underbrace{\mathsf{k}(x, y)}_{\text{correlation kernel}} f(y) \mathrm{d}\mu(y), \quad \text{with } \mu = \operatorname{unif}(\mathcal{X}).$$

- Hypothesis: $\mathsf{K} : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ *trace class symmetric.*

DPP exists iff the eigenvalues of K are in $[0, 1]$

Special case $\mathsf{K} = \mathsf{A}(\mathsf{A} + \mathbb{I})^{-1}$ with the *likelihood* operator

$$\mathsf{A}f(x) = \int_{\mathcal{X}} \mathsf{a}(x, y) f(y) \mathrm{d}\mu(y).$$

The full Maximum Likelihood Estimation problem (MLE)

$$\max_{\mathsf{A} \in \mathcal{S}_+(L^2(\mathcal{X}))} \log\det\left[\mathsf{a}(x_i, x_j)\right]_{i,j \in \mathcal{C}} - \log\det(\mathbb{I} + \mathsf{A}),$$

where $\mathcal{S}_+(H)$: *psd* trace class operators on Hilbert space H.

Kernelization and discretization

- Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ RKHS on $\mathcal{X}$ with bounded continuous $k(x, y)$.
- **Feature map**: $\phi(x) = k(x, \cdot) \in \mathcal{H}$.
- **Restriction operator** $S : \mathcal{H} \to L^2(\mathcal{X})$ as $(Sg)(x) = g(x)$.
  Define $\mathsf{A} : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ as $\mathsf{A} = SAS^*$ with $A \in \mathcal{S}_+(\mathcal{H})$.

Kernelized Maximum likelihood estimation (kMLE)

$$\min_{A \in \mathcal{S}_+(\mathcal{H})} f(A) = -\log\det\left[\langle \phi(x_i), A\phi(x_j) \rangle\right]_{i,j \in \mathcal{C}} + \log\det(\mathbb{I} + SAS^*).$$

- **Discretization**: Sample $\mathcal{I} = \{x'_1, \dots, x'_n\}$ i.i.d. $\sim \operatorname{unif}(\mathcal{X})$.

$$S_n : \mathcal{H} \to \mathbb{R}^n \text{ such that } S_n g = \frac{1}{\sqrt{n}} [g(x'_1), \dots, g(x'_n)]^\top.$$

Approximation of Fredholm determinant

With high probability,

$$\left| \log\det \underbrace{(\mathbf{I}_n + S_n A S_n^*)}_{\text{matrix}} - \log\det \underbrace{(\mathbb{I} + SAS^*)}_{\text{operator}} \right| \lesssim \operatorname{Tr}(A)/\sqrt{n}.$$

Discretized kernelized MLE + regularization

Define the **sample version** of negative log-likelihood $f(A)$ as

$$f_n(A) = -\log\det\left[\langle \phi(x_i), A\phi(x_j) \rangle\right]_{i,j \in \mathcal{C}} + \log\det(\mathbf{I}_n + S_n A S_n^*).$$

Solve problem with discrete and **penalized** objective

$$\min_{A \in \mathcal{S}_+(\mathcal{H})} f_n(A) + \underbrace{\lambda \operatorname{Tr}(A)}_{\text{penalization}}, \text{ with } \lambda > 0.$$

Define $\mathcal{Z} = \{x_1, \dots, x_{|\mathcal{C}|}, x'_1, \dots, x'_n\}$ and denote $m = |\mathcal{Z}|$.

Representer theorem Marteau-Ferey, Bach, Rudi [1]

$\exists$ partial isometry $V : \mathcal{H} \to \mathbb{R}^m$ such that $A = V^* \mathbf{B} V$.

Finite optimization problem

MLE reduces to *finite* non-convex problem (λ-kMLE):

$$\min_{\mathbf{B} \succeq 0} f_n(V^* \mathbf{B} V) + \lambda \operatorname{Tr}(\mathbf{B}),$$

where $f_n(V^* \mathbf{B} V) = -\log\det\left[\mathbf{\Phi}^\top \mathbf{B} \mathbf{\Phi}\right]_{\mathcal{CC}} + \log\det\left[|\mathcal{I}| \mathbf{I} + \mathbf{\Phi}^\top \mathbf{B} \mathbf{\Phi}\right]_{\mathcal{II}}$.

Statistical guarantee: approximate full MLE objective

Let $A_\star$ be a solution of (kMLE) . Let $\mathbf{B}_\star$ be a solution of (λ-kMLE). Let $\delta \in (0, 1/2)$. If $\lambda \geq 2c_n(\delta)$, w.p. at least $1 - 2\delta$, it holds

$$|f(A_\star) - f(V^* \mathbf{B}_\star V)| \leq 3\lambda \operatorname{Tr}(A_\star) \text{with } c_n \lesssim 1/\sqrt{n}.$$

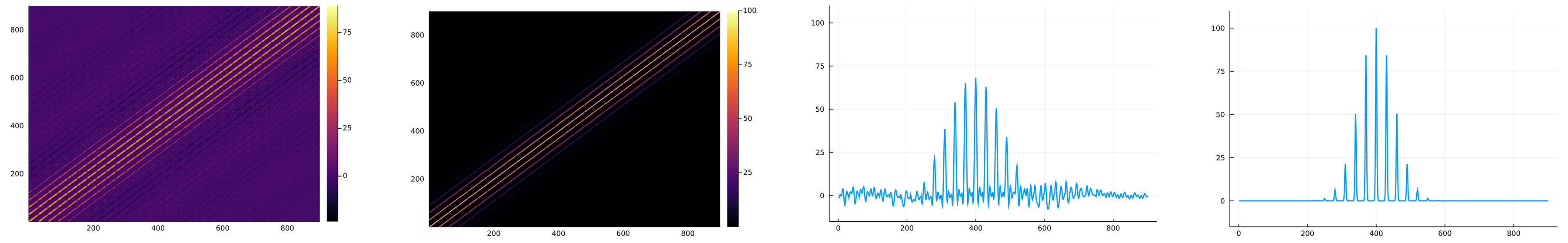
See [2] for the proof techniques.

- Numerical solution: regularized Picard iteration with monotone objective values; see [3] for proof techniques.

Estimation of $\mathsf{k}(x, y)$, i.e., kernel of $\mathsf{K} = \mathsf{A}(\mathsf{A} + \mathbb{I})^{-1}$.

Let $\delta \in (0, 1)$ and $\epsilon \in (0, 1)$. Let K be the correlation kernel associated to $\mathsf{A} = SAS^*$. We can compute $\hat{\mathsf{K}}(p)$ by using p points i.i.d. $\sim \operatorname{unif}(\mathcal{X})$, s.t. if $p \gtrsim \frac{\|A\|_{op}}{\epsilon^2} \log\left(\frac{4\operatorname{Tr}(\mathsf{K})}{\delta \|\mathsf{K}\|_{op}}\right)$, then, w.p. at least $1 - \delta$, we have $\frac{1}{1+\epsilon}\mathsf{K} \preceq \hat{\mathsf{K}}(p) \preceq \frac{1}{1-\epsilon}\mathsf{K}$.

Correlation kernel estimation: Gram matrix on a grid in $[0, 1]^2$.



(a) $[\hat{\mathsf{k}}(x, x')]_{x,x' \in \text{grid}}$ estimated Gram (b) $[\mathsf{k}(x, x')]_{x,x' \in \text{grid}}$ exact Gram (c) $[\hat{\mathsf{k}}((x_0, x)]_{x \in \text{grid}}$ Slice estim. Gram (d) $[\mathsf{k}(x_0, x)]_{x \in \text{grid}}$ Slice exact Gram

Ground truth $\mathsf{k}(x, y) = 100 \exp(-\|x - y\|_2^2 / 0.05^2)$ ($s = 10$ point patterns).

References:

[1] Marteau-Ferey, Bach, and Rudi, NeurIPS 2020.

[2] Rudi, Marteau-Ferey, and Bach, arXiv:2012.11978

[3] Mariet and Sra, ICML 2015

We acknowledge support from ERC grant Blackjack (ERC-2019-STG-851866) and ANR AI chair Baccarat (ANR-20-CHIA-0002).