Maximum Likelihood Estimation for Discrete Determinantal Point Processes

V.-E. Brunel[1,2], A. Moitra[1,3], P. Rigollet[1,2], J. Urschel[1]

[1]Department of Mathematics, MIT, Cambridge, Massachusetts, [2]Institute for Data, Systems and Society, MIT, Cambridge, Massachusetts, [3]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, Massachusetts

I. INTRODUCTION

Definitions: DPP

- A **determinantal point process** (DPP) $Y \subseteq [N]$ is a random subset s.t.

$$\mathbb{P}[J \subseteq Y] = \det(K_J), \forall J \subseteq [N]$$

for some symmetric matrix $K \in \mathbb{R}^{N\times N}$ with $0 \leq K \leq I_N$.

- Ex: $\mathbb{P}[i \in Y] = K_{i,i}$, $\mathbb{P}[i,j \in Y] = K_{i,i}K_{j,j} - K_{i,j}^2$.
- K is the **kernel** of the DPP(K).
- $\mathbb{P}[Y = J] = |\det(K - I_{\bar{J}})|$, $\forall J \subseteq [N]$.
- If $K < I_N$, the DPP(K) is also an **L-ensemble**:

$$\mathbb{P}[Y = J] = \frac{\det(L_J)}{\det(I_N + L)}, \qquad \forall J \subseteq [N]$$

where $L = K(I_N - K)^{-1}$ ($\Leftrightarrow K = L(I_N + L)^{-1}$).

- Alternative representation: $(X_1, X_2, \dots, X_N) \in \{0,1\}^N$, where $X_j \in Y \Leftrightarrow j \in Y$.
- DPPs can model repulsive interactions: $(X_1, X_2, \dots, X_N)$ are **negatively associated** ($\gg$ negative correlation), i.e.,

$$cov\left(f(X_i, i \in S), g(X_j, j \in T)\right) \leq 0,$$

for all disjoint $S, T \subseteq [N]$ and coordinatewise nondecreasing functions f, g. E.g., $cov(X_i, X_j) = -K_{i,j}^2 \leq 0$.

Learning objective

Given i.i.d. copies $Y_1, Y_2, \dots, Y_n \sim$ DPP(K^*) with unknown kernel, estimate K^*.

Identifiability of K

$$\text{DPP}(K) = \text{DPP}(K') \Leftrightarrow \det(K_J) = \det(K'_J), \forall J \subseteq [N]$$
$$\Leftrightarrow K' = DKD, \text{ for some } D = \text{Diag}(\pm 1, \dots, \pm 1).$$

⇝ **Principal minor assignment problem** [RKT15]: Find all symmetric matrices that have a prescribed list of principal minors.

Approach: Maximum Likelihood Estimation

(Empirical) log-likelihood: $\hat{\Psi}(K) = \sum_{J \subseteq [N]} \hat{p}_J \, log|\det(K - I_{\bar{J}})|$

(Population) log-likelihood: $\Psi(K) = \sum_{J \subseteq [N]} p_J^* \, log|\det(K - I_{\bar{J}})|$

with $\hat{p}_J = \frac{1}{n}\#\{i: Y_i = J\}$ and $p_J^* = \mathbb{P}[Y = J]$.

MLE: $\hat{K} \in argmax\, \hat{\Psi}(K) \Rightarrow \sum_{J \subseteq [N]} \hat{p}_J \left(\hat{K} - I_{\bar{J}}\right)^{-1} = 0$.

Key point: statistical performance and computation of $\hat{K}$ are bound to the geometry of Ψ.

Our primary focus:

Behavior of Ψ around the global maximum K^*.
↪ statistical performance of the MLE (e.g., the Hessian $-\nabla^2\Psi(K^*)$ is the Fisher information operator)

Find the other critical points of Ψ and study their nature.
↪ computation of the MLE

II. GLOBAL MAXIMA OF Ψ

Assumption: $0 < K^* < I_N$: K^* is an interior point of the parameter space.

Theorem 1.1: K^* is a global maximum and a critical point of Ψ. The kernels $DK^*D, D = \text{Diag}(\pm 1, \dots, \pm 1)$, are the only global maxima of Ψ.

Indeed, for all kernel K, $\Psi(K^*) - \Psi(K) = \text{KL}(\text{DPP}(K^*), \text{DPP}(K))$.

In addition, $\nabla\Psi(K^*) = \sum_{J \subseteq [N]} p_J^* \left(K^* - I_{\bar{J}}\right)^{-1} = 0$.

Theorem 1.2: The Hessian $\nabla^2\Psi(K^*)$ is negative semi-definite.

Definition 1.1: The *determinantal graph* $G_K = ([N], E_K)$ of a DPP with kernel K is the unweighted, undirected graph with edge set $E_K = \{\{i,j\}: K_{i,j} \neq 0\}$. For $i, j \in [N]$, we write $i \sim_K j$ iff there is a path in G_K which connects i and j.

Definition 1.2: A kernel K is called *irreducible* if it is not block diagonal up to a permutation of its rows and columns.

Fact: K is irreducible $\Leftrightarrow G_K$ is connected. Otherwise, the blocks of K correspond to the connected components of G_K.

Theorem 1.3: The null space of $\nabla^2\Psi(K^*)$ is

$$\mathcal{N}(K^*) = \{H \in \mathcal{S}_N: H_{i,j} = 0 \text{ for all } i,j \in [N] \text{ with } i \sim_{K^*} j\}.$$

The second order derivative of Ψ vanishes in the directions H that are off blocks of K^*.

Corollary 1.1: $\nabla^2\Psi(K^*)$ is negative definite iff K^* is irreducible.

Proposition 1.1: Let $a \in (0,1), |b| < \frac{\min(a, 1-a)}{2}$ and let K^* be the tridiagonal kernel

$$K^* = \begin{pmatrix} a & b & 0 & \dots & 0 \\ b & a & b & \ddots & \vdots \\ 0 & b & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & a & b \\ 0 & \dots & 0 & b & a \end{pmatrix}; \quad G_{K^*} =$$

Then, $$0 < \inf_{\substack{H \in \mathcal{S}_N, \\ \|H\|_F = 1}} -\nabla^2\Psi(K^*) \leq c_1 e^{-c_2 N}.$$

Theorem 1.4: Let $H \in \mathcal{N}(K^*)$. Then,

- $\nabla^3\Psi(K^*)(H,H,H) = 0$,
- $\nabla^4\Psi(K^*)(H,H,H,H) \leq 0$,
- $\nabla^4\Psi(K^*)(H,H,H,H) = 0 \Leftrightarrow H = 0$.

III. STATISTICAL CONSEQUENCES

The performance of the MLE $\hat{K}$ is measured with the loss function

$$\ell(\hat{K}, K^*) = \inf_D \|\hat{K} - DK^*D\|_F$$

Theorem 2.1: $\ell(\hat{K}, K^*) \xrightarrow[n\to\infty]{} 0$ in probability.

Theorem 2.2: If K^* is irreducible, then $\ell(\hat{K}, K^*) = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)$.

Theorem 2.3 : (Central Limit Theorem) Let K^* be irreducible and $\tilde{K} = \tilde{D}\hat{K}\tilde{D}$ with $\|\tilde{K} - K^*\|_F = \ell(\hat{K}, K^*)$. Then,

$$\sqrt{n}\left(\tilde{K} - K^*\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \nabla^2\Psi(K^*)^{-1}),$$

Remark: The hidden constants in Theorem 2.2 depend on N and can be arbitrarily large (they behave like the inverse of the smallest eigenvalue of the Fisher information), see Proposition 1.1.

Theorem 2.4: Let K^* be any kernel. Then,

$$\ell(\hat{K}, K^*) = O_{\mathbb{P}}\left(n^{-\frac{1}{6}}\right)$$

Theorem 2.5: Let K^* be block diagonal and S, T be two disjoint blocks. Then,

- $\inf_D \|\hat{K}_{S,T} - (DK^*D)_{S,T}\|_F = O_{\mathbb{P}}\left(n^{-\frac{1}{6}}\right)$,
- $\inf_D \|\hat{K}_S - (DK^*D)_S\|_F = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$.

IV. OTHER CRITICAL POINTS

Question: The kernels $DK^*D, D = \text{Diag}(\pm 1, \dots, \pm 1)$, are critical points or Ψ. Is it possible to describe the other critical points (explicit form and nature)?

Definition 3.1: Let $\mathcal{P} = (B_1, B_2, \dots, B_p)$ be a partition of $[N]$. A *partial decoupling* of the DPP Y with respect to $\mathcal{P}$ is a random subset Y' of $[N]$ s.t. $Y' \cap B_j, j = 1, \dots, p$ are independent and $Y' \cap B_j$ has the same distribution as $Y \cap B_j$ for all $j = 1, \dots, p$.

Proposition 3.1: A partial decoupling of the DPP Y with respect to a partition $\mathcal{P} = (B_1, B_2, \dots, B_p)$ is a DPP with block diagonal kernel (up to permutation of the rows and columns)

$$K^{(\mathcal{P})} = \begin{pmatrix} K_{B_1}^* & & & \\ & K_{B_2}^* & & \\ & & \ddots & \\ & & & K_{B_p}^* \end{pmatrix}$$

Theorem 3.1: For all partitions $\mathcal{P}$ of $[N]$, $K^{(\mathcal{P})}$ is a critical point of Ψ, i.e., it solves the equation $\sum_{J \subseteq [N]} p_J^* \left(K - I_{\bar{J}}\right)^{-1} = 0$. Unless $K^{(\mathcal{P})} = DK^*D$ for some $D = \text{Diag}(\pm 1, \dots, \pm 1)$, $K^{(\mathcal{P})}$ is a saddle point of Ψ.

Proposition 3.2: Let K be a critical point of Ψ. Then, $K_{j,j} = K_{j,j}^*, \forall j \in [N]$.

Conjecture 3.1: All critical points of Ψ are of the form $DK^{(\mathcal{P})}D$, for some partition $\mathcal{P}$ of $[N]$ and some $D = \text{Diag}(\pm 1, \dots, \pm 1)$.

Remark: For all partitions $\mathcal{P}$ of $[N]$ and all $D_0 = \text{Diag}(\pm 1, \dots, \pm 1)$, $D_0 K^{(\mathcal{P})} D_0$ is the average (with equal weights) of kernels of the form $DK^*D, D = \text{Diag}(\pm 1, \dots, \pm 1)$.

V. CONCLUSION AND OPENING REMARKS

- The rate of estimation of the MLE is never worse than $n^{-1/6}$.
- If K^* is irreducible, the MLE achieves the rate $n^{-1/2}$. Otherwise, the MLE estimates the diagonal blocks of K^* at the rate $n^{-1/2}$ and the off diagonal blocks at the rate $n^{-1/6}$.
- However, the rates may be affected by very large constants in high dimensions.
- Computation of the MLE is a non convex optimization problem. However, it seems that the only critical points of the population log-likelihood are the global maxima and saddle points, which may partially solve computational issues.
- Another estimator, obtained by a method of moments, consists of estimating principal minors of K^* and follow the principal minor assignment problem to reconstruct K^* : It is shown to have good computational and statistical properties [UBMR17].

REFERENCES

- [BMRU17-a] Brunel, V.-E., Moitra, A., Rigollet, P., Urschel, J. (2017): *Rates of Estimation for Determinantal Point Processes*, COLT 2017
- [BMRU17-b] Brunel, V.-E., Moitra, A., Rigollet, P., Urschel, J. (2017): *Maximum Likelihood Estimation for Determinantal Point Processes*, arXiv:1701.06501
- [RKT15] Rising, J., Kulesza, A., Taskar, B. (2015): *An Efficient Algorithm for the Symmetric Principal Minor Assignment Problem*, Linear Algebra and its Applications, 473: 126-144
- [UBMR17] Urschel, J., Brunel, V.-E., Moitra, A., Rigollet, P. (2017): *Learning Determinantal Point Processes with Moments and Cycles*, ICML 2017

Notation lexicon

$[N]$: set of integers from 1 to N.

M_J: submatrix of M with rows and columns indexed in J.

I_N: identity matrix of size N.

$I_{\bar{J}}$: diagonal matrix with j-th diagonal entry 1 if $j \notin J$, 0 otherwise.

$\|\cdot\|_F$: Frobenius norm

$M_{S,T}$: submatrix of M with rows indexed in S and columns indexed in T.