Rates of Estimation for Discrete Determinantal Point Processes

V.-E. Brunel, A. Moitra, P. Rigollet, J. Urschel

COLT 2017, Amsterdam

Discrete DPPs

Random variables on the hypercube $\{0,1\}^N$, represented as subsets of $[N]$.

$$1\,0\,0\,1\,1\,0\,1\,0\,1\,1\,0\,1\,0\,0\,1\,0\,0\,0\,1\,0 \quad \leftrightarrow \quad \{1,4,5,7,9,10,12,15,19\}$$

$$0\,0\,1\,1\,0\,1\,0\,1\,1\,0\,0\,1\,0\,0\,1\,0\,0\,0\,1\,0 \quad \leftrightarrow \quad \{3,4,6,8,9,12,15,19\}$$

$$1\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,1\,0\,1\,0\,0\,1\,1\,0\,1 \quad \leftrightarrow \quad \{1,4,8,12,14,17,18,20\}$$

…

$$0\,0\,1\,0\,0\,1\,0\,1\,1\,0\,0\,0\,0\,0\,1\,1\,0\,1\,0\,0 \quad \leftrightarrow \quad \{3,6,8,9,15,16,18\}$$

Discrete DPPs

- Probabilistic model for correlated Bernoulli r.v.
- Feature repulsion (negative association)

Definition Random subset $Y \subseteq [N]$,

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

$K \in \mathbb{R}^{N \times N}$, symmetric, $0 \preccurlyeq K \preccurlyeq I$

- $K_{i,j} \nleftrightarrow$ repulsion between items i and j.
- PMF: $\mathbb{P}[Y = J] = \left|\det(K - I_{\bar{J}})\right|$

Goal

- Given $Y_1, Y_2, \dots, Y_n \overset{\text{iid}}{\sim} \mathrm{DPP}(K^*)$, estimate K^*.
- **Approach:** Maximum Likelihood Estimator.
- **Question:** Rate of convergence of the MLE ?

Identification

- $\mathrm{DPP}(K) = \mathrm{DPP}(K^*) \Leftrightarrow \det(K_J) = \det(K_J^*), \forall J \subseteq [N]$

$$\Leftrightarrow K = DK^*D \quad \text{for some } \mathrm{D} = \begin{pmatrix} \pm 1 & & & 0 \\ & \pm 1 & & \\ & & \ddots & \\ 0 & & & \pm 1 \end{pmatrix}.$$

- E.g.: $\mathrm{K}^* = \begin{pmatrix} + & + & + & + \\ + & + & + & + \\ + & + & + & + \\ + & + & + & + \end{pmatrix} \rightsquigarrow DK^*D = \begin{pmatrix} + & - & - & + \\ - & + & + & - \\ - & + & + & - \\ + & - & - & + \end{pmatrix}$

Measure of the error of an estimator $\hat{K}$:

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

Maximum likelihood estimation

- **Log-likelihood:** $\hat{\Psi}(K) = \sum_{J \subseteq [N]} \hat{p}_J \ln\left|\det\left(\mathrm{K} - \mathrm{I}_{\bar{J}}\right)\right|$

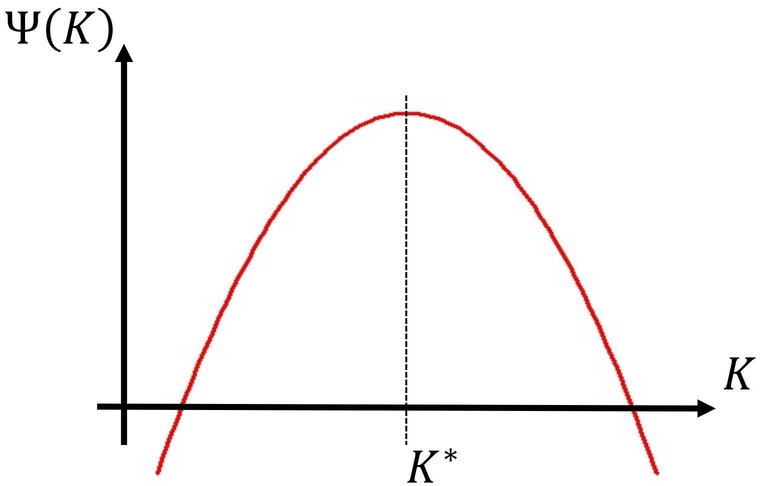
- **MLE:** $\hat{K} \in \operatorname{argmax} \hat{\Psi}(K)$

$$\Psi(K) \triangleq \mathbb{E}\left[\hat{\Psi}(K)\right] = \sum_{J \subseteq [N]} p_J^* \ln\left|\det\left(\mathrm{K} - \mathrm{I}_{\bar{J}}\right)\right|$$

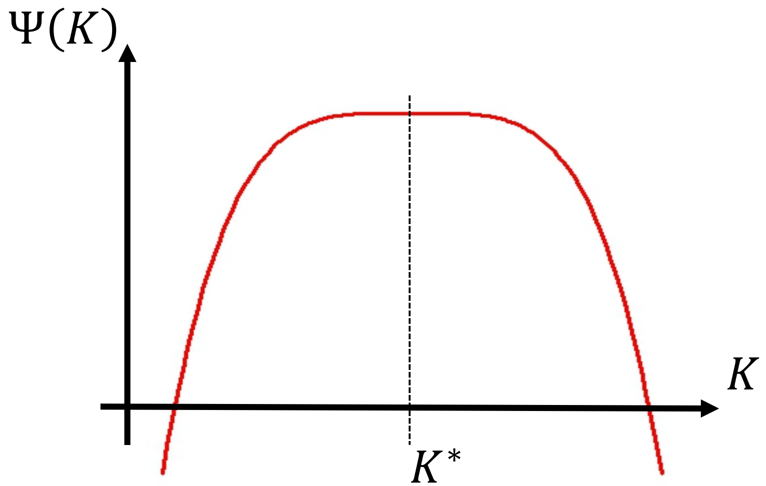
$$= \Psi(K^*) - KL\left(DPP(K^*), DPP(K)\right)$$

Likelihood geometry

Fisher information: $-\nabla^2\Psi(K^*)$



$\nabla^2\Psi(K^*) < 0$



$\nabla^2\Psi(K^*) = 0$

What is the order of the first non degenerate derivative of $\boldsymbol{\Psi}$ at $\boldsymbol{K = K^*}$?

Determinantal Graphs & Irreducibility

Definition

$G = ([N], E)$: $\{i, j\} \in E \Leftrightarrow K^*_{i,j} \neq 0$.

- K^* is ***irreducible*** iff G is connected.
- Otherwise, K^* is block diagonal.
- Rk: K^* is block diagonal $\Rightarrow$ Y = union of independent DPPs
- Write $i \sim j$ when i and j are connected in G.

Main Results: Irreducible case

Theorem 1

K^* irreducible $\Leftrightarrow$ $\nabla^2\Psi(K^*)$ is definite negative

Statistical consequences:

- $\ell\left(\widehat{K}, K^*\right) = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$
- CLT

Main Results: Block diagonal case (1)

Theorem 2

$$\mathrm{Ker}\left(\nabla^2\Psi(K^*)\right) = \left\{H \in \mathbb{R}^{N\times N}: H_{i,j} = 0, \forall i \sim j\right\}$$

$\boldsymbol{\nabla^2\Psi(K^*)}$ is negative definite along directions supported on the blocks of $\boldsymbol{K^*}$.

Theorem 3

For $H \in \mathrm{Ker}\left(\nabla^2\Psi(K^*)\right) \setminus \{0\}$: $\begin{cases} \nabla^3\Psi(K^*)\left(H^{\otimes 3}\right) = 0 \\ \nabla^4\Psi(K^*)\left(H^{\otimes 4}\right) < 0 \end{cases}$

Main Results: Block diagonal case (2)

Statistical consequences:

- $\ell\left(\widehat{K}, K^*\right) = O_{\mathbb{P}}\left(n^{-\frac{1}{6}}\right)$
- $\ell\left(\widehat{K}_S, K_S^*\right) = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$ for all blocks S of K^*.

Conclusions

- Rates of convergence of the MLE:

$$\begin{cases} n^{-1/2} & \text{if } K^* \text{ is irreducible} \\ n^{-1/6} & \text{otherwise} \end{cases}$$

- Rate only determined by connectedness of the **determinantal graph**
- Hidden constants can be arbitrarily large in N: e.g., if G is a path graph
- In another paper* we show that the sample complexity of a method-of-moment estimator is determined by the *cycle sparsity* of G.

* *Learning Determinantal Point Processes from Moments and Cycles*, J. Urschel, V.-E. Brunel, A. Moitra, P. Rigollet, ICML 2017