Learning Determinantal Point Processes Using Moments and Cycles

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 $K \in \mathbb{R}^{N\times N}$ s.t. $0 \leq K \leq I_N$.

- Ex: $\mathbb{P}[1 \in Y] = K_{1,1}$, $\mathbb{P}[1,2 \in Y] = K_{1,1}K_{2,2} - K_{1,2}^2$.

- 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

$$\text{DPP}(K) = \text{DPP}(K') \Leftrightarrow \det(K_J) = \det(K'_J), \forall J \subseteq [N]$$

$$\Leftrightarrow K' = DKD, \text{ for some } D = \text{Diag}(\pm 1, \dots, \pm 1).$$

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

II. Method of Moments

- ➢ First step: $K_{i,i} = \mathbb{P}[i \in Y] \rightsquigarrow \hat{K}_{i,i} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{i \in Y_k}$
- ➢ Second step: $K_{i,j}^2 = K_{i,i}K_{j,j} - \mathbb{P}[i,j \in Y]$

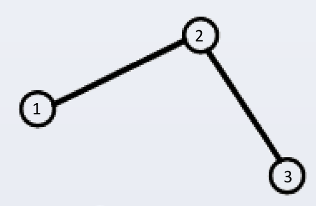
$$\rightsquigarrow \widehat{K_{i,j}^2} = \left(\widehat{K_{i,i}}\widehat{K_{j,j}} - \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{i,j \in Y_k}\right)^+$$

- ➢ Third step: Recover the signs

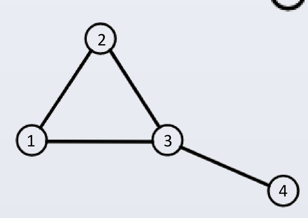
Determinantal graph:

Definition 1.1: $G = ([N], E)$, with $E_K = \left\{\{i,j\}: K_{i,j} \neq 0\right\}$.

$$K = \begin{pmatrix} * & * & 0 \\ * & * & * \\ 0 & * & * \end{pmatrix}$$

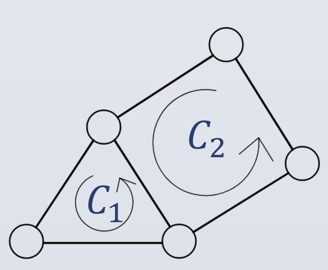


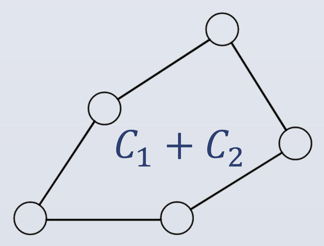
$$K = \begin{pmatrix} * & * & * & 0 \\ * & * & * & 0 \\ * & * & * & * \\ 0 & 0 & * & * \end{pmatrix}$$



Cycle Sparsity

Cycle basis: family of *induced cycles* that span the cycle space





Cycle sparsity: length ℓ of the largest cycle needed to span the cycle space

Horton's algorithm: Find a cycle basis with cycle lengths $\leq \ell$ in $O(|E|^2 N(\ln N)^{-1})$ steps [Horton '87; Amaldi *et al.* '10]

Theorem: K is completely determined, up to $\mathcal{D}$-similarity, by its principal minors of order $\leq \ell$.

Key idea: Find the sign of $\prod_{\{i,j\} \in C} K_{i,j}$ for each cycle of length $\leq \ell$, using the corresponding principal minors.

III. ALGORITHM

Assumption: $K \in \mathcal{K}_\alpha$, i.e., either $K_{i,j} = 0$ or $|K_{i,j}| \geq \alpha$, for some *known* $\alpha \in (0,1)$.

- All $K_{i,i}$'s and $|K_{i,j}|$'s are estimated within $\boldsymbol{n^{-1/2}}$**-rate**

G is recovered exactly w.h.p.

- **Horton's algorithm** outputs a minimum basis $\mathcal{B}$
- For all induced cycle $C \in \mathcal{B}$

$$\det K_C = F_C(K_{i,i}, K_{i,j}^2) + 2(-1)^{|C|} \prod_{\{i,j\} \in C} K_{i,j}$$

$\Rightarrow$ **Recover the sign** of $\prod_{\{i,j\} \in C} K_{i,j}$ w.h.p.

Remark: $\left|\prod_{\{i,j\} \in C} K_{i,j}\right|$ may be as small as α^ℓ $\hookrightarrow$ finding its sign requires $\Omega(\alpha^{-2\ell})$ samples.

IV. MAIN RESULTS

Upper bound:

Theorem: Let $K \in \mathcal{K}_\alpha$ with cycle sparsity ℓ and let $\varepsilon > 0$. Then, the following holds with probability at least $1 - n^{-A}$:

There is an algorithm that outputs $\hat{K}$ in $O(|E|^3 + nN^2)$ steps for which

$$n \gtrsim \left(\frac{1}{\alpha^2 \varepsilon^2} + \ell\left(\frac{2}{\alpha}\right)^{2\ell}\right) \ln N \quad \Rightarrow \quad \min_D \left|\hat{K} - DKD\right|_\infty \leq \varepsilon$$

Lower bound:

Theorem:

Let $0 < \varepsilon \leq \alpha \leq 1/8$ and $3 \leq \ell \leq N$. There exists $C > 0$ such that if

$$n < C\left(\frac{8^\ell}{\alpha^{2\ell}} + \frac{\ln\binom{N}{\ell}}{(6\alpha)^\ell} + \frac{\ln N}{\varepsilon^2}\right)$$

then the following holds:

For any estimator $\tilde{K}_n$, there exists $K \in \mathcal{K}_\alpha$ with cycle sparsity ℓ and for which $\min_D \left|\hat{K} - DKD\right|_\infty > \varepsilon$ with probability at least $1/3$.

V. CONCLUSION AND OPENING REMARKS

- Estimation of K by a method of moments in *polynomial time*
- Rates of estimation characterized by the topology of the determinantal graph through its *cycle sparsity* ℓ.
- These rates are provably *optimal* (up to logarithmic factors)
- *Adaptation* to ℓ.
- Another estimator, obtained by a maximum likelihood approach, does not assume separation of the nonzero entries of K and achieves the rates $n^{-1/2}$ or $n^{-1/6}$, depending on the connectedness of G [BMRU17-a,b].

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