Learning Determinantal Processes with Moments and Cycles

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

ICML 2017, Sydney

Determinantal Point Processes (DPPs)

DPP: Random subset of $[N]$

- For all $J \subseteq [N]$,

$$\mathbb{P}[J \subseteq Y] = \det \boldsymbol{K}_J$$

- $\boldsymbol{K} \in \mathbb{R}^{N \times N}$, symmetric, $0 \preccurlyeq K \preccurlyeq I_N$: parameter (*kernel*) of the DPP
- $K_J = \left(K_{i,j}\right)_{i,j \in J}$
- E.g. $\mathbb{P}[1 \in Y] = K_{1,1}$, $\mathbb{P}[1,2 \in Y] = K_{1,1}K_{2,2} - K_{1,2}^2 \leq \mathbb{P}[1 \in Y]\mathbb{P}[2 \in Y]$.
- A.k.a. L-ensembles if $0 \prec K \prec I_N$: $\mathbb{P}[Y = J] \propto \det L_J$, $L = K(I_N - K)^{-1}$.

Binary representation

DPP $\longleftrightarrow$ Random binary vector of size N, represented as a subset of $[N]$.

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

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

$$(X_1, \ldots, X_N) \in \{0,1\}^N \quad \leftrightarrow \quad Y \subseteq [N]$$

$$\boxed{X_i = 1 \iff i \in Y}$$

Model for correlated Bernoulli r.v.'s (such as Ising, …) featuring repulsion.

Applications of DPP’s

DPPs have become popular in various applications:

- Quantum physics (*fermionic processes*) [Macchi ‘74]
- Document and timeline summarization [Lin, Bilmes ‘12; Yao *et al.* ‘16]
- Image search [Kulesza, Taskar ‘11; Affandi *et al.* ‘14]
- Bioinformatics [Batmanghelich *et al.* ‘14]
- Neuroscience [Snoek *et al.* ‘13]
- Wireless or cellular networks modelization [Miyoshi, Shirai ‘14; Torrisi, Leonardi ‘14; Li *et al.* ‘15; Deng *et al.* ‘15]

And they remain an elegant and important tool in probability theory [Borodin ‘11]

Learning DPPs

- Given $Y_1, Y_2, \dots, Y_n \overset{\text{iid}}{\sim} \mathrm{DPP}(K)$, estimate K.

- **Approach:** Method of moments

- Problem: Is K identified ?

Identification: $\mathcal{D}$-similarity

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

[Oeding '11]

$$\Leftrightarrow K' = DKD \quad \text{for some } D = \begin{pmatrix} \pm 1 & & & 0 \\ & \pm 1 & & \\ & & \ddots & \\ 0 & & & \pm 1 \end{pmatrix}.$$

- E.g.: $K = \begin{pmatrix} + & + & + & + \\ + & + & + & + \\ + & + & + & + \\ + & + & + & + \end{pmatrix} \rightsquigarrow DKD = \begin{pmatrix} + & - & - & + \\ - & + & + & - \\ - & + & + & - \\ + & - & - & + \end{pmatrix}$

- K and DKD are called **$\mathcal{D}$-similar**.

Method of moments

- **Diagonal entries:** $\boldsymbol{K_{i,i}} = \mathbb{P}[i \in Y]$ ⟹ $\hat{K}_{i,i} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{i \in Y_k}$

- **Magnitude of the off-diagonal entries:**

$$K_{i,j}^2 = K_{i,i}K_{j,j} - \mathbb{P}[i,j \in Y] \quad \Longrightarrow \quad \widehat{K_{i,j}^2} = \left(\widehat{K_{i,i}}\widehat{K_{j,j}} - \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{i,j \in Y_k}\right)^+$$

- **Signs (up to $\mathcal{D}$-similarity) ?**

Use estimates of higher moments: $\widehat{\det K_J} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{J \in Y_k}$

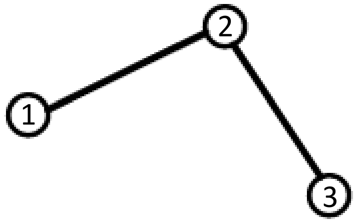
Determinantal Graphs

Definition

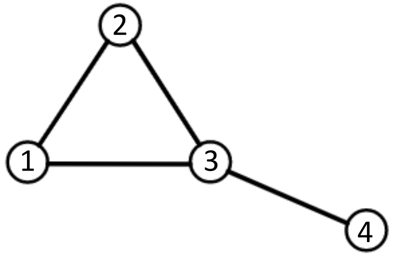
$G = ([N], E)$: $\{i, j\} \in E \Leftrightarrow K_{i,j} \neq 0.$

Examples:

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



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



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]

Cycle sparsity

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

Key: Signs of $\prod_{\{i,j\}\in C} K_{i,j}$ for each cycle of length $\leq \ell$.

Learning the signs

- **Assumption:** $K \in \mathcal{K}_\alpha$, i.e., either $K_{i,j} = 0$ or $\left|K_{i,j}\right| \geq \alpha > 0$
- All $K_{i,i}$'s and $\left|K_{i,j}\right|$'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 $C \in \mathcal{B}$

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

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

Main result

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

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

Near-optimal rate in a minimax sense.

Conclusions

- 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.