Discrete Determinantal Point Processes

Stats/ML in Saclay

Jan 27, 2020

Victor-Emmanuel Brunel
Department of Statistics
ENSAE/CREST

ENSAE

IP PARIS

Random selection models

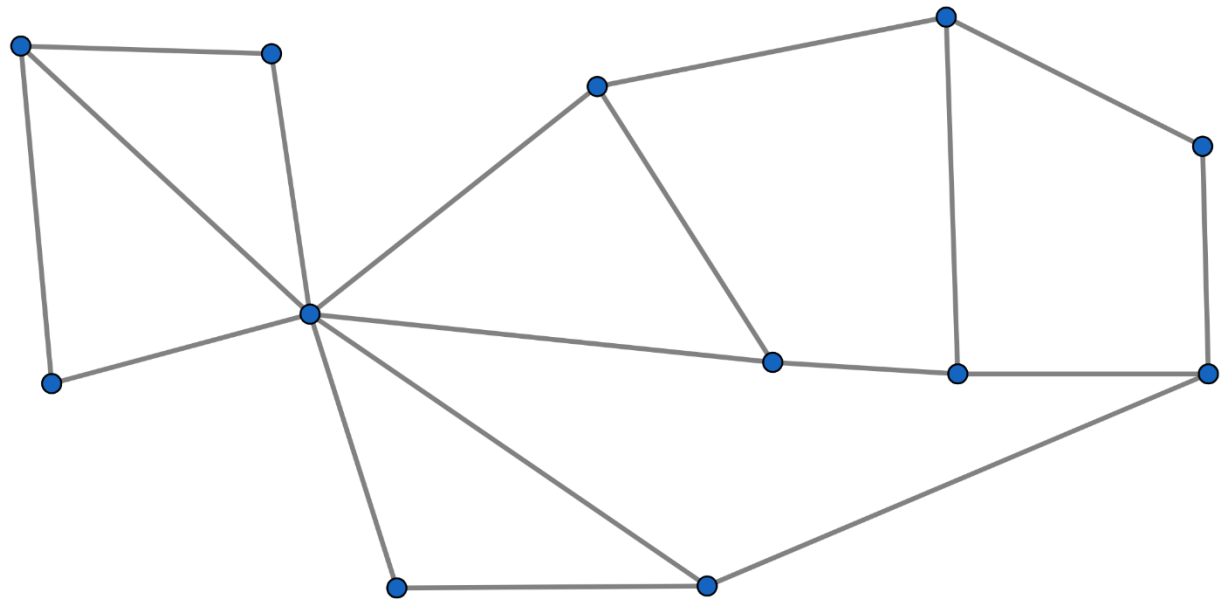
- Sampling
  - Independent Bernoulli sampling
  - with/without replacement

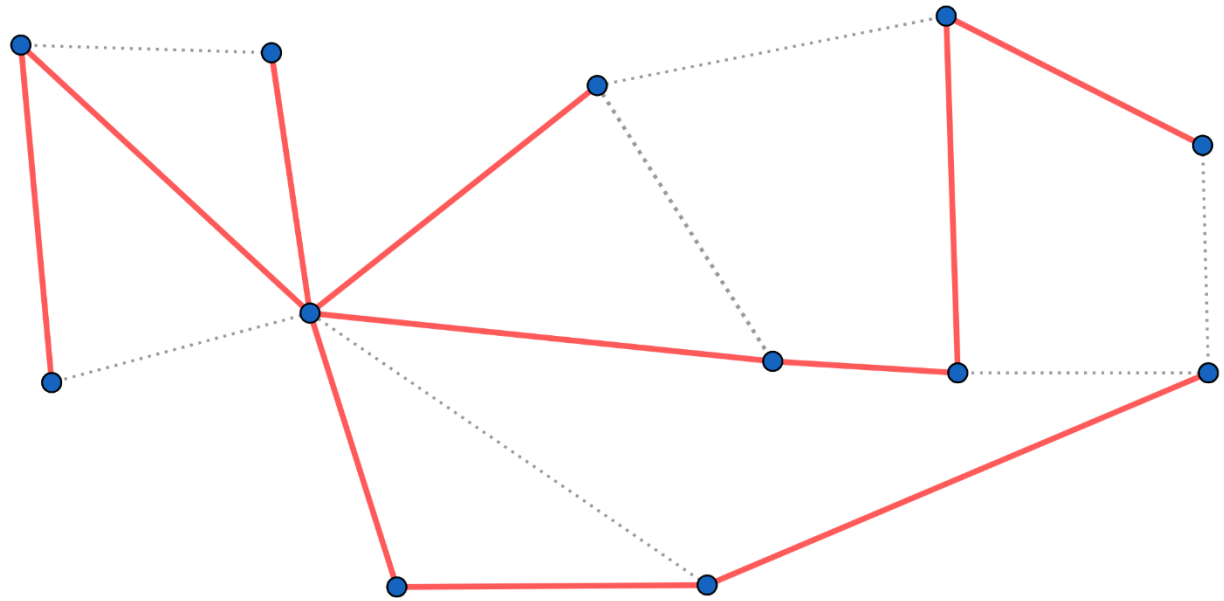
- Ising models
  - Ferromagnetic (attractive interactions)
  - Antiferromagnetic (repulsive interactions)

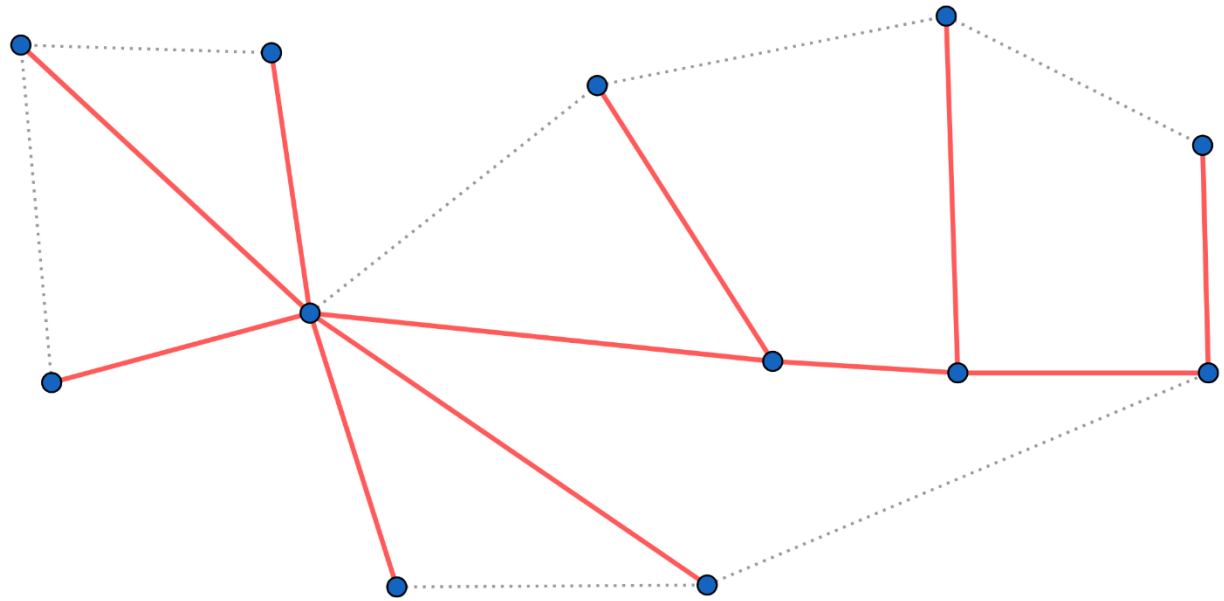
- DPPs

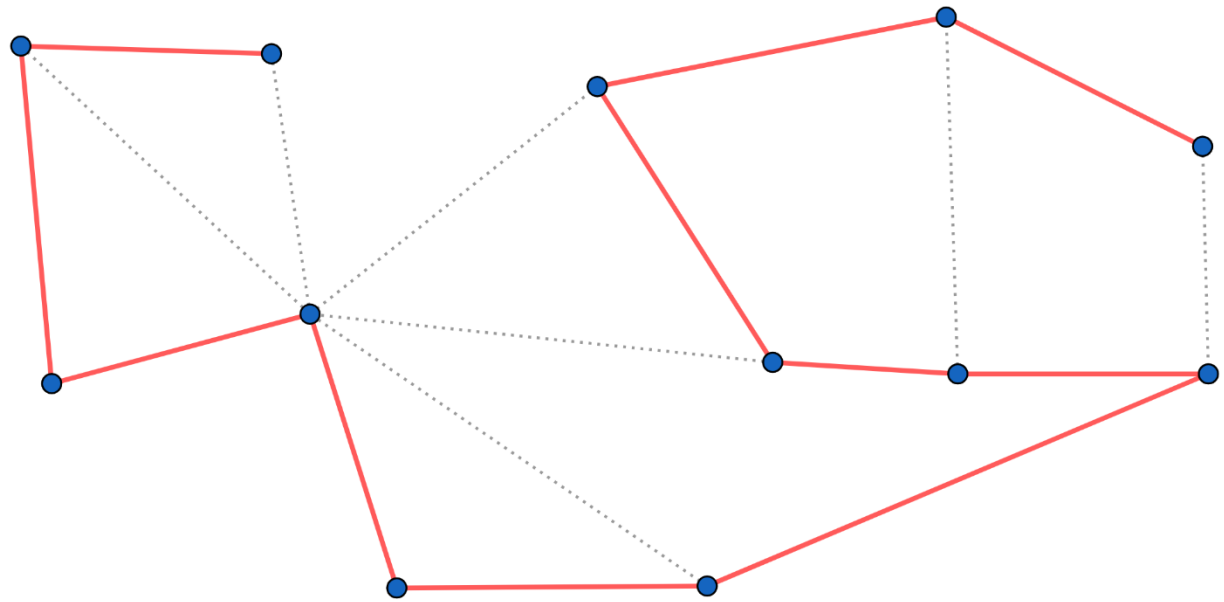
DPPs in probability

- Random spanning trees
- Increases in i.i.d. random sequences
- Non-intersecting random walks
- Eigenvalues of random matrices







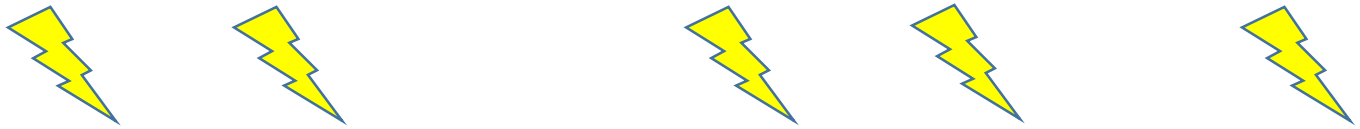


DPPs in probability

- Random spanning trees
- Increases in i.i.d. random sequences
- Non-intersecting random walks
- Eigenvalues of random matrices

- $X_1, X_2, \ldots, X_{N+1} \overset{\text{i.i.d.}}{\sim} P$
- $\{i = 1, \ldots, n : X_i < X_{i+1}\}$ is a DPP

3 2 7 3 6 1 1 0 8 5 9 6 3 8 5 5 3 1



$$Y = \{2,4,8,10,13\} \subseteq [17]$$

DPPs in probability

- Random spanning trees
- Increases in i.i.d. random sequences
- Non-intersecting random walks
- Eigenvalues of random matrices

DPPs in probability

- Random spanning trees
- Increases in i.i.d. random sequences
- Non-intersecting random walks
- Eigenvalues of random matrices

DPPs in ML

- Model for random selections of items within a catalog/dictionary
-  Tractable models: marginalizing, conditioning, are computationally simple
- In some cases, sampling is simple and fast (e.g., via spectral decomposition or MCMC)

Applications of DPPs

(Symmetric) DPPs have become popular in various applications:

- Quantum physics (*fermionic processes*) [Macchi ‘74]
- Experimental design [Derezinski et al. ’19]
- 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 network modelisation [Li *et al.* ’15; Deng *et al.* ‘15]
- Recommendation systems [Gartrell et al ‘16, ‘17]

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

Set representation of binary vectors

A binary vector of size N can be represented as a subset of $[N]$:

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

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

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

$$X_i = 1 \Leftrightarrow i \in Y$$

DPPs: Definition

DPP: Random subset $\boldsymbol{Y}$ of $[\boldsymbol{N}]$

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

$$\mathbb{P}[Y = J] \propto \det \boldsymbol{L}_J$$

  - $\boldsymbol{L} \in \mathbb{R}^{N \times N}$: parameter of the DPP
  - $L_J = \left(L_{i,j}\right)_{i,j \in J}$

- Normalization constant: $\det(I + L)^{-1}$.
- What are the admissible matrices L? [B. '18]

$\mathbf{P_0}$-matrices

- L must be a $\boldsymbol{P_0}$**-matrix** (all its principal minors are ≥ 0).

- Examples:
  - Any PSD matrix : **Symmetric DPPs**
  - If $L + L^\top$ is PSD

- More about P_0-matrices: *Convex sets of nonsingular and P-matrices*, C. R. Johnson, M. J. Tsatsomeros (1995)

DPPs: Alternative representation (1)

- For all $J \subseteq [N]$, $\mathbb{P}[J \subseteq Y] = \det \boldsymbol{K}_J$

$\boldsymbol{K} = L(I + L)^{-1}$: *kernel* of the DPP

- $\mathbb{P}[Y = J] = (-1)^{|\bar{J}|} \det\left(K - I_{\bar{J}}\right)$

DPPs: Alternative representation (2)

- $X_i \sim \mathrm{Ber}\left(K_{i,i}\right)$
- $\mathbb{P}[1,2 \in Y] = \begin{vmatrix} K_{1,1} & K_{1,2} \\ K_{2,1} & K_{2,2} \end{vmatrix} = K_{1,1}K_{2,2} - K_{1,2}K_{2,1}$

$$\Rightarrow \quad \boxed{\mathrm{cov}(X_1, X_2) = -K_{1,2}K_{2,1}}$$

- $\mathbb{E}[|Y|] = \mathbb{E}[\sum X_i] = \mathrm{Tr}\, K$
- $\mathrm{Var}[|Y|] = \mathrm{Tr}\big(K(I-K)\big)$

DPPs: Alternative representation (2)

- $X_i \sim \mathrm{Ber}\left(K_{i,i}\right)$
- $\mathbb{P}[1,2 \in Y] = \begin{vmatrix} K_{1,1} & K_{1,2} \\ K_{2,1} & K_{2,2} \end{vmatrix} = K_{1,1}K_{2,2} - K_{1,2}K_{2,1}$

$$\Rightarrow \quad \boxed{\mathrm{cov}(X_1, X_2) = -K_{1,2}K_{2,1}}$$

- $\mathbb{E}[|Y|] = \mathbb{E}[\sum X_i] = \mathrm{Tr}\, K$
- $\mathrm{Var}[|Y|] = \mathrm{Tr}\big(K(I-K)\big)$

DPPs: Alternative representation (2)

- $X_i \sim \mathrm{Ber}\left(K_{i,i}\right)$
- $\mathbb{P}[1,2 \in Y] = \begin{vmatrix} K_{1,1} & K_{1,2} \\ K_{2,1} & K_{2,2} \end{vmatrix} = K_{1,1}K_{2,2} - K_{1,2}K_{2,1}$

$$\Rightarrow \quad \boxed{\mathrm{cov}(X_1, X_2) = -K_{1,2}K_{2,1}}$$

- $\mathbb{E}[|Y|] = \mathbb{E}[\sum X_i] = \mathrm{Tr}\, K$
- $\mathrm{Var}[|Y|] = \mathrm{Tr}\big(K(I-K)\big)$

DPPs: Alternative representation (2)

- $X_i \sim \mathrm{Ber}\left(K_{i,i}\right)$
- $\mathbb{P}[1,2 \in Y] = \begin{vmatrix} K_{1,1} & K_{1,2} \\ K_{2,1} & K_{2,2} \end{vmatrix} = K_{1,1}K_{2,2} - K_{1,2}K_{2,1}$

$$\Rightarrow \quad \boxed{\mathrm{cov}(X_1, X_2) = -K_{1,2}K_{2,1}}$$

- $\mathbb{E}[|Y|] = \mathbb{E}[\sum X_i] = \mathrm{Tr}\, K$
- $\mathrm{Var}[|Y|] = \mathrm{Tr}\big(K(I-K)\big)$

DPPs: Symmetric case (1)

- L is symmetric and PSD ($\Leftrightarrow K$ is symmetric with $0 \preccurlyeq K \preccurlyeq I$)
- Write $K = V^\top V, \quad V \in \mathbb{R}^{r \times N}$
- Columns of V: $v_1, \ldots, v_N$ (v_i: vector of r features of item i)
- $K_{i,i} = \|v_i\|^2$: *popularity of item i*
- $\det K_J = (\text{Volume spanned by } v_i, i \in J)^2$

DPPs: Symmetric case (2)

- $\mathbb{P}[i, j \in Y] = \det K_{\{i,j\}} = \begin{vmatrix} K_{i,i} & K_{i,j} \\ K_{i,j} & K_{j,j} \end{vmatrix} = K_{i,i}K_{j,j} - K_{i,j}^2$
- $\operatorname{cov}(X_i, X_j) = \mathbb{P}[i, j \in Y] - \mathbb{P}[i \in Y]\mathbb{P}[j \in Y] = -K_{i,j}^2 \leq 0$
- More generally, if $S \cap T = \emptyset$:

$$\operatorname{cov}\left(\prod_{i \in S} X_i, \prod_{j \in T} X_j\right) = \mathbb{P}[S, T \subseteq Y] - \mathbb{P}[S \subseteq Y]\mathbb{P}[T \subseteq Y]$$

$$= \det(K_{S \cup T}) - \det(K_S)\det(K_T)$$

$$\leq 0$$

DPPs: Symmetric case (2)

- $\mathbb{P}[i, j \in Y] = \det K_{\{i,j\}} = \begin{vmatrix} K_{i,i} & K_{i,j} \\ K_{i,j} & K_{j,j} \end{vmatrix} = K_{i,i}K_{j,j} - K_{i,j}^2$
- $\mathrm{cov}(X_i, X_j) = \mathbb{P}[i, j \in Y] - \mathbb{P}[i \in Y]\mathbb{P}[j \in Y] = -K_{i,j}^2 \leq 0$
- More generally, if $S \cap T = \emptyset$:

$$\mathrm{cov}\left(\prod_{i \in S} X_i, \prod_{j \in T} X_j\right) = \mathbb{P}[S, T \subseteq Y] - \mathbb{P}[S \subseteq Y]\mathbb{P}[T \subseteq Y]$$

$$= \det(K_{S \cup T}) - \det(K_S)\det(K_T)$$

$$\leq 0$$

DPPs: Symmetric case (2)

- $\mathbb{P}[i, j \in Y] = \det K_{\{i,j\}} = \begin{vmatrix} K_{i,i} & K_{i,j} \\ K_{i,j} & K_{j,j} \end{vmatrix} = K_{i,i}K_{j,j} - K_{i,j}^2$
- $\mathrm{cov}\big(X_i, X_j\big) = \mathbb{P}[i, j \in Y] - \mathbb{P}[i \in Y]\mathbb{P}[j \in Y] = -K_{i,j}^2 \leq 0$
- More generally, if $S \cap T = \emptyset$:

$$\mathrm{cov}\left(\prod_{i \in S} X_i\,, \prod_{j \in T} X_j\right) = \mathbb{P}[S, T \subseteq Y] - \mathbb{P}[S \subseteq Y]\mathbb{P}[T \subseteq Y]$$

$$= \det(K_{S \cup T}) - \det(K_S)\det(K_T)$$

$$\leq 0$$

Negative Association (1)

Definition: Boolean rv's are *negatively associated** iff

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

for all $S, T \subseteq [N]$ and nondecreasing functions f, g.

*Remark: Hard to prove in general…

Negative Association (2)

Definition: Boolean rv's are *negatively associated** iff

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

for all $S, T \subseteq [N]$ and nondecreasing functions f, g.

*Remark: Hard to prove in general…

Strongly Rayleigh distributions (1)

- Let μ be the joint distribution of a Boolean vector $(X_1, \dots, X_N)$
- μ: Distribution on the cube $\{0,1\}^N$
- Generating polynomial: $g_\mu(z) = \mathbb{E}\left[z_1^{X_1} \dots z_N^{X_N}\right], \quad z = (z_1, \dots, z_N) \in \mathbb{C}^N$
- Ex:
  - Independent Bernoulli: $g_\mu(z) = \prod_{j=1}^{N} (p_j z_j + 1 - p_j)$
  - DPP: $g_\mu(z) = \sum_{J \subseteq [N]} \frac{\det L_J}{\det(I+L)} z^J = \frac{\det(I+LZ)}{\det(I+L)} = \det(I - K + KZ)$

Strongly Rayleigh distributions (1)

- Let μ be the joint distribution of a Boolean vector $(X_1, \dots, X_N)$
- μ: Distribution on the cube $\{0,1\}^N$
- Generating polynomial: $g_\mu(z) = \mathbb{E}\left[z_1^{X_1} \dots z_N^{X_N}\right], \quad z = (z_1, \dots, z_N) \in \mathbb{C}^N$
- Ex:
  - Independent Bernoulli: $g_\mu(z) = \prod_{j=1}^{N} \left(p_j z_j + 1 - p_j\right)$
  - DPP: $g_\mu(z) = \sum_{J \subseteq [N]} \frac{\det L_J}{\det(I+L)} z^J = \frac{\det(I+LZ)}{\det(I+L)} = \det(I - K + KZ)$

Strongly Rayleigh distributions (2)

Definition: μ is *Strongly Rayleigh* iff g_μ is real stable, i.e.,

$$\mathcal{I}m(z_j) > 0, \forall j \in [N] \Rightarrow g_\mu(z) \neq 0$$

Property: SR $\Rightarrow$ NA

See [Borcea et al. '08]

Strongly Rayleigh distributions (2)

Definition: μ is *Strongly Rayleigh* iff g_μ is real stable, i.e.,

$$\mathcal{I}m(z_j) > 0, \forall j \in [N] \Rightarrow g_\mu(z) \neq 0$$

Property: SR ⇒ NA

See [Borcea et al. ’08]

Strongly Rayleigh distributions (2)

Examples:

- Independent Bernoulli
- **Symmetric** DPPs
- **Symmetric** k-DPPs
- Sampling without replacement

Useful stability properties: Marginalization and conditioning.

Challenges in Stats and ML

Learning symmetric DPPs (1)

- Given $Y_1, Y_2, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{DPP}(L)$, estimate L

- Problem 1: Identifiability

Learning symmetric DPPs (1)

- Given $Y_1, Y_2, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{DPP}(L^*)$, estimate L^*
- Problem 1: Identifiability

Identification (symmetric case): $\mathcal{D}$-similarity

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

Identification (symmetric case): $\mathcal{D}$-similarity

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

[Oeding '11]

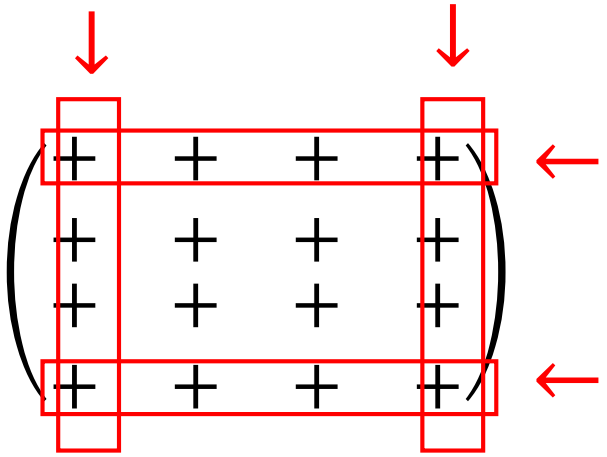
$\Leftrightarrow L' = DLD$ for some $\mathrm{D} = \begin{pmatrix} \pm 1 & & 0 \\ & \pm 1 & \\ 0 & & \ddots \\ & & & \pm 1 \end{pmatrix}$.

Identification (symmetric case): $\mathcal{D}$-similarity

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

[Oeding '11]

$\Leftrightarrow L' = DLD$ for some $\mathrm{D} = \begin{pmatrix} \pm 1 & & & 0 \\ & \pm 1 & & \\ & 0 & \ddots & \\ & & & \pm 1 \end{pmatrix}$.

- E.g.: $\mathrm{L} =$  $\rightsquigarrow \quad DLD = \begin{pmatrix} + & - & - & + \\ - & + & + & - \\ - & + & + & - \\ + & - & - & + \end{pmatrix}$

- L and DLD are called **$\mathcal{D}$-similar**.  Learn L up to $\mathcal{D}$-similarity.

Learning symmetric DPPs (2)

- Error of an estimator $\hat{L}$: $\min_{D \in \mathcal{D}} \left\| \hat{L} - DLD \right\|$

- Generic methods:

  - Method of moments (Urschel et al, 2017): Noisy principal minor assignment problem
  - Maximum likelihood (B. et al, 2017)
  - Penalised likelihood (only empirical results)

Maximum likelihood estimation

Samples: $Y_1, Y_2, \ldots, Y_n \sim \mathrm{DPP}(L^*), \quad \mathrm{L}^* \succ 0$

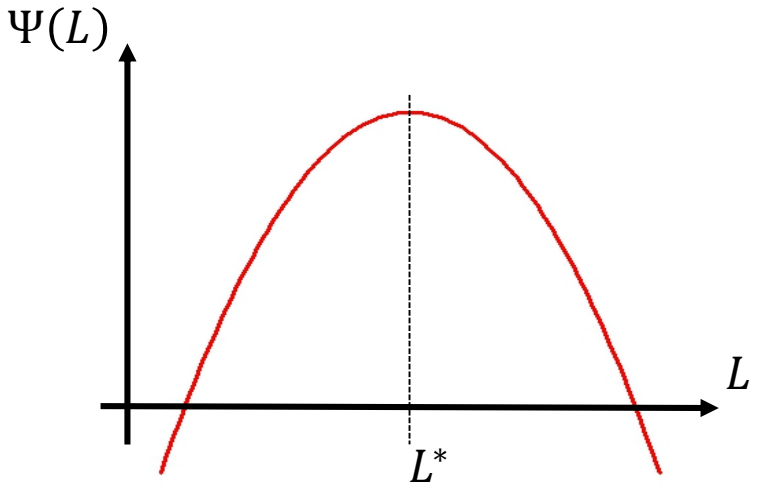
- **Log-likelihood:** $\widehat{\Psi}(L) = \sum_{J \subseteq [N]} \widehat{p_J} \log\det\left(\mathrm{L_J}\right) - \log\det(\mathrm{I} + \mathrm{L})$
- **MLE:** $\hat{L} \in \operatorname{argmax} \widehat{\Psi}(L)$

$$\Psi(L) \triangleq \mathbb{E}\left[\widehat{\Psi}(L)\right] = \sum_{J \subseteq [N]} p_J^* \log\det\left(\mathrm{L_J}\right) - \log\det(\mathrm{I} + \mathrm{L})$$

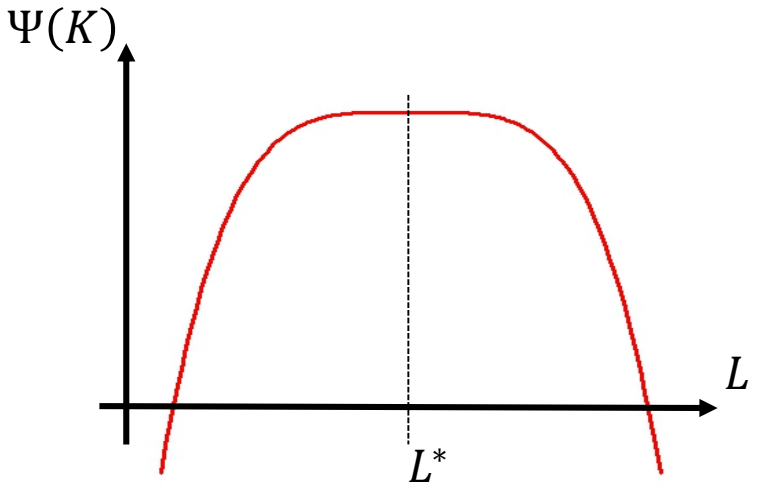
$$= \Psi(L^*) - \mathrm{KL}\left(\mathrm{DPP}(L^*), \mathrm{DPP}(L)\right)$$

Likelihood geometry

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



$\nabla^2\Psi(\mathrm{L}^*) < 0$



$\nabla^2\Psi(\mathrm{L}^*) = 0$

What is the order of the first non degenerate derivative of Ψ at $L = L^*$?

Asymptotic Results

- If L^* is irreducible:

$$\min_D \left\| \hat{L} - DL^*D \right\|_\infty = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$$

- If L^* is block diagonal:

  - $\min_D \left\| \hat{L} - DL^*D \right\|_\infty = O_{\mathbb{P}}\left(n^{-\frac{1}{4}}\right)$
  - $\min_D \left\| \hat{L}_S - (DL^*D)_S \right\|_\infty = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$ for all diagonal blocks S of L^*.

Non-asymptotic results in high dimension?

- Low rank assumption:

$$\mathrm{rk}(L^*) \leq r \Leftrightarrow |Y| \leq r \quad \text{a.s.}$$

- If $\mathrm{rk}(L^*) \leq r$, then $\widehat{p_J} = 0, \ \forall J \subseteq [N] \ s.t. \ |J| > r$
- However, $\mathrm{rk}(\hat{L})$ can still be very large.
- Minimize a penalized version of the log-likelihood?
- How to compute $\hat{L}$ (or $\hat{L}_{pen}$) efficiently?

Non-asymptotic results in high dimension?

- Low rank assumption:

$$\mathrm{rk}(L^*) \leq r \Leftrightarrow |Y| \leq r \quad \text{a.s.}$$

- If $\mathrm{rk}(L^*) \leq r$, then $\widehat{p_J} = 0, \ \forall J \subseteq [N] \ s.t. \ |J| > r$

- However, $\mathrm{rk}(\hat{L})$ can still be very large.

- Minimize a penalized version of the log-likelihood?

- How to compute $\hat{L}$ (or $\hat{L}_{pen}$) efficiently?

Non-asymptotic results in high dimension?

- Low rank assumption:

$$\mathrm{rk}(L^*) \leq r \Leftrightarrow |Y| \leq r \quad \text{a.s.}$$

- If $\mathrm{rk}(L^*) \leq r$, then $\widehat{p_J} = 0, \ \forall J \subseteq [N] \ s.t. \ |J| > r$

- However, $\mathrm{rk}\left(\hat{L}\right)$ can still be very large.

- Minimize a penalized version of the log-likelihood?

- How to compute $\hat{L}$ (or $\hat{L}_{pen}$) efficiently?

Non-asymptotic results in high dimension?

- Low rank assumption:

$$\mathrm{rk}(L^*) \leq r \Leftrightarrow |Y| \leq r \quad \text{a.s.}$$

- If $\mathrm{rk}(L^*) \leq r$, then $\widehat{p_J} = 0, \ \forall J \subseteq [N] \ s.t. \ |J| > r$

- However, $\mathrm{rk}\left(\hat{L}\right)$ can still be very large.

- Minimize a penalized version of the log-likelihood?

- How to compute $\hat{L}$ (or $\hat{L}_{pen}$) efficiently?

Non-asymptotic results in high dimension?

- Low rank assumption:

$$\mathrm{rk}(L^*) \leq r \Leftrightarrow |Y| \leq r \quad \text{a.s.}$$

- If $\mathrm{rk}(L^*) \leq r$, then $\widehat{p_J} = 0, \ \forall J \subseteq [N] \ s.t. \ |J| > r$

- However, $\mathrm{rk}\left(\hat{L}\right)$ can still be very large.

- Minimize a penalized version of the log-likelihood?

- How to compute $\hat{L}$ (or $\hat{L}_{pen}$) efficiently?

Thank you for your attention!