Stochastic Processes: Lecture 5

Andrew J. Holbrook

UCLA Biostatistics 270

Spring 2022

Determinantal point processes

- A determinantal point process (DPP) on $\mathbb{R}^D$ is determined by a kernel $K(x, x')$.
- The joint intensities can be written

$$\det \begin{pmatrix} K(x_i, x_i) & K(x_i, x_j) \\ K(x_i, x_j) & K(x_j, x_j) \end{pmatrix}$$

- The kernel defines an integral operator $\mathcal{K}$ acting on $L^2(\mathbb{R}^D)$ that is self-adjoint, positive semidefinite and trace class.

Joint intensities of a DPP

Definition 1

The joint intensities of a point process N are functions (if any exist) $\rho_k : (\mathbb{R}^D)^k \to [0, \infty)$ for $k \geq 1$, such that for any family of disjoint sets $D_1, \ldots, D_k \subset \mathbb{R}^D$,

$$E\left(\prod_{i=1}^{k} N(D_i)\right) = \int_{\prod D_i} \rho_k(x_1, \ldots, x_k) dx_1 \ldots dx_k \, .$$

Definition 2

A point process N on $\mathbb{R}^D$ is said to be a DPP with kernel K if its joint intensities satisfy

$$\rho_k(x_1, \ldots, x_k) = \det \left(K(x_i, x_j)\right)_{1 \leq i,j \leq k}$$

for every $k \geq 1$ and $x_1, \ldots, x_k \in \mathbb{R}^D$.

Permanental point processes

Leibniz' formula for the determinant of a $k \times k$ matrix M is

$$\det(M) = \sum_{\sigma \in S_k} \left(\operatorname{sgn}(\sigma) \prod_{i=1}^{k} M_{i,\sigma(i)} \right) .$$

We denote the *permanent* of a $k \times k$ matrix M

$$\operatorname{per}(M) = \sum_{\sigma \in S_k} \prod_{i=1}^{k} M_{i,\sigma(i)} .$$

Definition 3

A point process N on $\mathbb{R}^D$ is said to be a permanental point process with kernel K if its joint intensities satisfy

$$\rho_k(x_1, \ldots, x_k) = \mathit{per}(K(x_i, x_j))_{1 \le i,j \le k}$$

for every $k \geq 1$ and $x_1, \ldots, x_k \in \mathbb{R}^D$.

Poisson processes, DPPs and PPPs

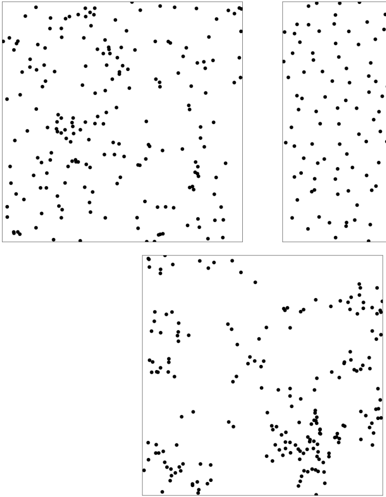


FIG 1. *Samples of translation invariant point processes in the plane: Poisson (left), determinantal (center) and permanental for* $K(z,w) = \frac{1}{\pi} e^{z\overline{w} - \frac{1}{2}(|z|^2+|w|^2)}$. *Determinantal processes exhibit repulsion, while permanental processes exhibit clumping.*

Hough et al., 2006

DPP results

Lemma 1

Suppose $\{\phi_k\}_{k=1}^n$ *is an orthonormal set in* $L^2(\mathbb{R}^D)$*. Then there exists a DPP with kernel*

$$K(x, y) = \sum_{k=1}^{n} \phi_k(x)\overline{\phi}_k(y) .$$

Theorem 1

Let K *determine a self-adjoint integral operator* $\mathcal{K}$ *on* $L^2(\mathbb{R}^D)$ *that is locally trace-class. Then* K *defines a DPP on* $\mathbb{R}^D$ *iff all the eigenvalues of* $\mathcal{K}$ *are in* $[0, 1]$*.*

DPP results

Theorem 2
Suppose N is a DPP with kernel $K(x, y)$. Write

$$K(x, y) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) \overline{\phi}_k(y) \,,$$

where ϕ_k are normalized eigenfunctions with eigenvalues $\lambda_k \in [0, 1]$. Let $I_k \overset{\perp}{\sim}$ Bernoulli(λ_k) and define K's random analogue

$$K_I(x, y) = \sum_{k=1}^{\infty} I_k \phi_k(x) \overline{\phi}_k(y) \,.$$

Let N_I be a DPP with kernel K_i. Then

$$N \overset{d}{=} N_I \,.$$

In particular, the total number of points in N follows the distribution of the sum of independent Bernoulli(λ_k) r.v.s.

DPP example: non-intersecting random walks

Consider n independent simple symmetric walks on $\mathbb{Z}$ started from $i_1 < \cdots < i_n$, all even. Let $P_{ij}(t)$ be the t-step transition probabilities. The probability the r.w.s are at $j_1 < \cdots < j_n$ at time t and have non-intersecting paths is

$$\det \begin{pmatrix} P_{i_1 j_1}(t) & \dots & P_{i_1 j_n}(t) \\ \vdots & \ddots & \\ P_{i_n j_1}(t) & & P_{i_n j_n}(t) \end{pmatrix} .$$

If t is even and we condition the walks to return to $i_1, \dots, i_n$ at time t, then the positions at time $t/2$ follow a DPP with Hermitian kernel.

DPP example: Ginibre ensemble

Let Q be an $n \times n$ matrix with i.i.d. complex standard normal entries. The eigenvalues of Q form a DPP on $\mathbb{C}$ with the kernel

$$K_n(z,w) = \frac{1}{\pi} e^{-\frac{1}{2}(|z|^2+|w|^2)} \sum_{k=0}^{n-1} \frac{(z\overline{w})^k}{k!} .$$

As $n \to \infty$, we have a DPP on $\mathbb{C}$ with kernel

$$\begin{aligned} K(z,w) &= \frac{1}{\pi} e^{-\frac{1}{2}(|z|^2+|w|^2)} \sum_{k=0}^{\infty} \frac{(z\overline{w})^k}{k!} \\ &= \frac{1}{\pi} e^{-\frac{1}{2}(|z|^2+|w|^2)+z\overline{w}} . \end{aligned}$$

Zero set of a Gaussian analytic function

The power series $f(z) = \sum_{n=0}^{\infty} a_n z^n$, where a_n are i.i.d. standard complex normals defines a random analytic function on the unit disk (a.s.). The zero set of f is a determinantal process in the disk with the Bergman kernel

$$K(z,w) = \frac{1}{\pi(1 - z\overline{w})^2} = \frac{1}{\pi}\sum_{k=0}^{\infty}(k+1)(z\overline{w})^k \,.$$

DPPs on discrete sets

Let $\mathcal{Y}$ be a discrete set with n items. A point process N on $\mathcal{Y}$ is a probability distribution on the power set $2^{\mathcal{Y}}$.

Definition 4
A point process N is a determinantal point process if for $Y \subseteq \mathcal{Y}$ randomly sampled according to N we have for every $S \subseteq \mathcal{Y}$

$$Pr(S \subseteq Y) = \det K_S$$

for some similarity matrix $K \in \mathbb{R}^{n \times n}$ that is symmetric and positive semidefinite.

Let S be a two-element set with elements i and j. Then

$$\Pr(S \subset Y) = K_{ii}K_{jj} - K_{ij}^2 = \Pr(i \subset Y)\Pr(j \subset Y) - K_{ij}^2 \,.$$

Conditioning

DPPs are closed under conditioning:

$$
\begin{aligned}
\Pr(A \subseteq Y | B \subseteq Y) &= \Pr(A \cup B \subseteq Y) / \Pr(A \subseteq Y) \\
&= \frac{\det K_{A \cup B}}{\det K_A} \\
&= \frac{\det(K_A) \det\left(K_B - K_{BA} K_A^{-1} K_{AB}\right)}{\det(K_A)} \\
&= \det\left(K_B - K_{BA} K_A^{-1} K_{AB}\right) \\
&= \det\left(\left[K - K_{\mathcal{Y}A} K_A^{-1} K_{A\mathcal{Y}}\right]_B\right) .
\end{aligned}
$$

Restrictions on K

- Because marginal probabilities of any set $S \subseteq \mathcal{Y}$ must be in $[0, 1]$, all $\det(K_S) \geq 0$ and hence K must be positive semidefinite.

- Moreover, all eigenvalues of K must inhabit $[0, 1]$, i.e. $0 \preceq K \preceq 1$.

- Any K satisfying $0 \preceq K \preceq 1$ defines a DPP.

L-ensembles

- L-ensembles provide a convenient way to avoid dealing with $K \preceq 1$ constraints.
- An L-ensemble is defined using a symmetric matrix $L \succeq 0$ that defines the *atomic* probability of an event set S thus:

$$\mathrm{Pr}_L(S) = \mathrm{Pr}(S = Y) \propto \det\left(L_Y\right)$$

- Conveniently, the normalizing constant is known:

$$\sum_{S \subseteq \mathcal{Y}} \det\left(L_S\right) = \det(L + I)\,.$$

L-ensembles

Theorem 3
For any $S \subseteq \mathcal{Y}$

$$\sum_{S \subseteq Y \subseteq \mathcal{Y}} \det(L_Y) = \det(L + I_{S^c})$$

Corollary 1

$$\sum_{Y \subseteq \mathcal{Y}} \det(L_Y) = \det(L + I)$$

Proof.
Let S from Theorem 3 equal the empty set. □

L-ensembles

Theorem 4

An L-ensemble is a DPP and its marginal kernel is

$$K = L(L+I)^{-1} = I - (L+I)^{-1}$$

Proof.

The marginal probability of a set S under the L-ensemble is

$$\begin{aligned}
\Pr_L(S \subseteq Y) &= \frac{\sum_{S \subseteq Y \subseteq \mathcal{Y}} \det(L_Y)}{\sum_{Y \subseteq \mathcal{Y}} \det(L_Y)} = \frac{\det(L + I_{S^c})}{\det(L+I)} \\
&= \det\left((L + I_{S^c})(L+I)^{-1}\right) \\
&= \det\left(I_{S^c}(L+I)^{-1} + I - (L+I)^{-1}\right) \\
&= \det\left(I_{S^c}(L+I)^{-1} + (I_S + I_{S^c})\left(I - (L+I)^{-1}\right)\right) \\
&= \det(I_{S^c} + I_S K) = \begin{vmatrix} I_{|S^c| \times |S^c|} & 0 \\ K_{S,S^c} & K_S \end{vmatrix} = \det(I_{|S^c| \times |S^c|}) \det(K_S) \\
&= \det(K_S) \,.
\end{aligned}$$

□

L-ensembles

- Given a marginal kernel, we may construct an L-ensemble by setting $L = K(I - K)^{-1}$.

- The inverse of $I - K$ might not exist, so DPPs are a larger class than L-ensembles.

- If $L = \sum_k \lambda_k v_k v_k^T$, then $K = \sum_k \frac{\lambda_k}{1+\lambda_k} v_k v_k^T$.

- Linear kernel. Let X be an $n \times p$ design matrix (set of feature vectors). Taking $L = XX^T$, we have

$$\Pr_L(S) \propto \det(L_S) = Vol^2(\{x_i\}_{i \in S})$$

If $p < n$, the DPP will only have p points.

Working with DPPs

- Complements: if $Y \sim DPP(K)$, then $Y^c \sim DPP(I - K)$
- Conditioning:

$$\Pr_L(Y = S_{in} \cup B | S_{in} \subseteq Y, S_{out} \cap Y = \emptyset) = \frac{\det(L_{S_{in} \cup B})}{\det(L_{S_{out}^c} + I_{S_{in}^c})}$$

- Marginalization:

$$\Pr(B \subseteq Y | S \subseteq Y) = \det\left(\left[I - \left[(L + I_{S^c})^{-1}\right]_{S^c}\right]_B\right)$$

- Scaling: if $K' = \gamma K$ for $\gamma \in [0, 1]$, then for all $S \subseteq \mathcal{Y}$

$$\Pr_{K'}(S \subseteq Y) = \det(K'_S) = \gamma^{|S|} K_S\,.$$

Elementary DPPs

- A DPP is elementary if every eigenvalue of K is 0 or 1.
- N^V denotes an elementary DPP with marginal kernel $K^V = \sum_{v \in V} vv^T$ if V is a set of orthonormal vectors.
- The expected total count for a DPP is

$$E(|Y|) = E(\sum_{i=1}^{n} 1\{i \in Y\}) = \sum_{i=1}^{n} \Pr(i \in Y) = \sum_{i=1}^{n} K_{ii} = \mathrm{tr}(K)\,.$$

- For an elementary DPP this is

$$E(|Y|) = \mathrm{tr}(K^V) = \mathrm{tr}\left(\sum_{v \in V} vv^T\right) = \sum_{v \in V} v^T v = |V|\,.$$

- Furthermore, $|Y| = |V|$ a.s. because $\det(K_Y^V) = 0$ when $|Y| > |V|$.

DPPs as mixtures of elementary DPPs

Lemma 2

A DPP with kernel $L = \sum_{i=1}^{n} \lambda_i v_i v_i^T$ *is a mixture of elementary DPPs:*

$$Pr_L = \frac{1}{\det(L+I)} \sum_{J \subseteq \{1,2,\ldots,n\}} Pr^{V_J} \prod_{i \in J} \lambda_i$$

where $V_J = \{v_i\}_{i \in J}$

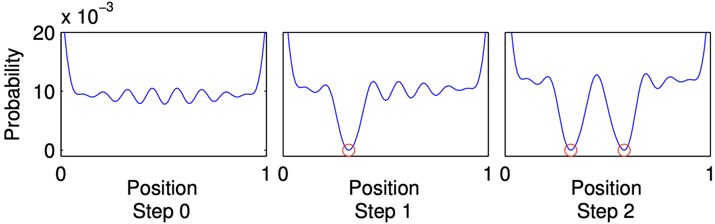
Sampling DPPs

Algorithm 1 Sampling from a DPP

Input: eigendecomposition $\{(\boldsymbol{v}_n, \lambda_n)\}_{n=1}^N$ of L
$J \leftarrow \emptyset$
for $n = 1, 2, \ldots, N$ **do**
  $J \leftarrow J \cup \{n\}$ with prob. $\frac{\lambda_n}{\lambda_n + 1}$
end for
$V \leftarrow \{\boldsymbol{v}_n\}_{n \in J}$
$Y \leftarrow \emptyset$
while $|V| > 0$ **do**
  Select i from $\mathcal{Y}$ with $\Pr(i) = \frac{1}{|V|} \sum_{\boldsymbol{v} \in V} (\boldsymbol{v}^\top \boldsymbol{e}_i)^2$
  $Y \leftarrow Y \cup i$
  $V \leftarrow V_\perp$, an orthonormal basis for the subspace of V orthogonal to $\boldsymbol{e}_i$
end while
Output: Y

Kulesza and Taskar, 2013

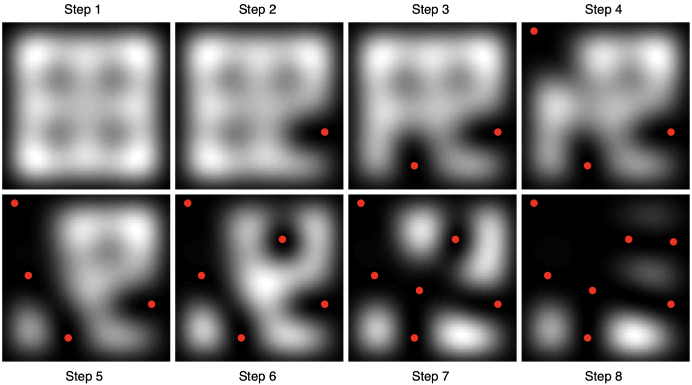
Sampling DPPs



(a) Sampling points on an interval

Kulesza and Taskar, 2013

Sampling DPPs



(b) Sampling points in the plane

Kulesza and Taskar, 2013

Sampling DPPs

- Finding the eigendecomposition of L is $O(n^3)$.
- Sampling algorithm is $O(n|V|^3)$ for V the set of eigenvectors selected in phase 1 and each repeated Gram-Schmidt to compute $V_\perp$ is $O(n|V|^2)$.

Dual representation

- Let B be the $D \times N$ matrix with columns $B_i = q_i \phi_i$ such that $L = B^T B$. Consider the $D \times D$ matrix

$$C = BB^T .$$

- Here, D is the dimension of the diversity feature function ϕ.
- D is often fixed by design, whereas N may grow as more items are modeled.

Dual representation

Proposition 1

The non-zero eigenvalues of L and C are identical, and the corresponding eigenvectors are related by the matrix B. That is,

$$C = \sum_{d=1}^{D} \lambda_d \hat{v}_d \hat{v}_d^T$$

is an eigendecomposition of C if and only if

$$L = \sum_{d=1}^{D} \lambda_d \left(\frac{1}{\sqrt{\lambda_d}} B^T \hat{v}_d \right) \left(\frac{1}{\sqrt{\lambda_d}} B^T \hat{v}_d \right)^T$$

is an eigendecomposition of L.

Dual representation

Proof.

First, assume $\{\lambda_d, \hat{v}_d\}_{d=1}^D$ is an eigendecomposition of C. Then,

$$\sum_{d=1}^{D} \lambda_d \left(\frac{1}{\sqrt{\lambda_d}} B^T \hat{v}_d \right) \left(\frac{1}{\sqrt{\lambda_d}} B^T \hat{v}_d \right)^T = B^T \left(\sum_{d=1}^{D} \hat{v}_d \hat{v}_d^T \right) B = B^T B = L \, .$$

Furthermore, we have

$$\begin{aligned} \left|\left| \frac{1}{\sqrt{\lambda_d}} B^T \hat{v}_d \right|\right|^2 &= \frac{1}{\lambda_d} (B^T \hat{v}_d)^T (B^T \hat{v}_d) = \frac{1}{\lambda_d} \hat{v}_d^T C \hat{v}_d \\ &= \frac{1}{\lambda_d} \lambda_d \hat{v}_d^T \hat{v}_d = 1 \, , \end{aligned}$$

and

$$\begin{aligned} \left(\frac{1}{\sqrt{\lambda_d}} B^T \hat{v}_d \right)^T \left(\frac{1}{\sqrt{\lambda_{d'}}} B^T \hat{v}_{d'} \right) &= \frac{1}{\sqrt{\lambda_d \lambda_{d'}}} \hat{v}_d^T C \hat{v}_{d'} \\ &= \frac{\sqrt{\lambda_{d'}}}{\sqrt{\lambda_d}} \hat{v}_d^T \hat{v}_{d'} = 0 \, . \end{aligned}$$

A similar argument holds in the other direction when one accounts for the fact $L = B^T B$ and has rank at most D. □

Dual representation and computing

- Normalization: the normalization constant is

$$\det(L + I) = \prod_{d=1}^{D} (\lambda_d + 1) = \det(C + I)\,,$$

  which only takes $O(D^3)$ time.

- Marginalization: get entries of K using C. First get the eigendecomposition $C = \sum_{d=1}^{D} \lambda_d \hat{v}_d \hat{v}_d^T$. Then

$$K_{ij} = \sum_{d=1}^{D} \frac{\lambda_d}{\lambda_d + 1} \left(\frac{1}{\sqrt{\lambda_d}} B_i^T \hat{v}_d \right)^T \left(\frac{1}{\sqrt{\lambda_d}} B_j^T \hat{v}_d \right) .$$

  One may therefore obtain the marginal probability of an event in time $O(D^2)$. For a k event, this becomes $O(D^2 k^2 + k^3)$. This beats the usual $O(n^3)$ to translate from L to K.

Dual representation and computing

In general, one may represent the orthonormal set V in $\mathbb{R}^n$ using the set $\hat{V}$ in $\mathbb{R}^D$ with the mapping

$$V = \{B^T \hat{v} | \hat{v} \in \hat{V}\}.$$

One may implicitly obtain linear combinations of vectors in V by performing actions on their preimages: $v_1 + v_2 = B^T(\hat{v}_1 + \hat{v}_2)$. Moreover,

$$v_1^T v_2 = (B^T \hat{v}_1)^T (B^T \hat{v}_2) = \hat{v}_1^T C \hat{v}_2,$$

so we can compute dot products of elements in V in time $O(D^2)$. We can implicitly normalize the elements of V by updating

$$\hat{v} \longleftarrow \frac{\hat{v}}{\hat{v}^T C \hat{v}}.$$

Sampling DPPs

Algorithm 1 Sampling from a DPP

Input: eigendecomposition $\{(\boldsymbol{v}_n, \lambda_n)\}_{n=1}^N$ of L
$J \leftarrow \emptyset$
for $n = 1, 2, \ldots, N$ **do**
  $J \leftarrow J \cup \{n\}$ with prob. $\frac{\lambda_n}{\lambda_n + 1}$
end for
$V \leftarrow \{\boldsymbol{v}_n\}_{n \in J}$
$Y \leftarrow \emptyset$
while $|V| > 0$ **do**
  Select i from $\mathcal{Y}$ with $\Pr(i) = \frac{1}{|V|} \sum_{\boldsymbol{v} \in V} (\boldsymbol{v}^\top \boldsymbol{e}_i)^2$
  $Y \leftarrow Y \cup i$
  $V \leftarrow V_\perp$, an orthonormal basis for the subspace of V orthogonal to $\boldsymbol{e}_i$
end while
Output: Y

Kulesza and Taskar, 2013

Can we use the dual representation to speed up the sampling of *i* and Gram-Schmidt steps?

Dual representation and computing

The sampling step is handled thus:

$$\Pr(i) = \frac{1}{|V|}\sum_{v\in V}(v^T e_i)^2 = \frac{1}{|\hat{V}|}\sum_{\hat{v}\in\hat{V}}((B^T\hat{v})^T e_i)^2$$
$$= \frac{1}{|\hat{V}|}\sum_{\hat{v}\in\hat{V}}(B_i^T\hat{v})^2$$

The entire distribution may be computed in time $O(nD|\hat{V}|)$ instead of $O(n^3)$.

Sampling DPPs

Algorithm 3 Sampling from a DPP (dual representation)

Input: eigendecomposition $\{(\hat{\boldsymbol{v}}_n, \lambda_n)\}_{n=1}^N$ of C

$J \leftarrow \emptyset$

for $n = 1, 2, \ldots, N$ **do**

  $J \leftarrow J \cup \{n\}$ with prob. $\frac{\lambda_n}{\lambda_n + 1}$

end for

$\hat{V} \leftarrow \left\{ \frac{\hat{\boldsymbol{v}}_n}{\sqrt{\hat{\boldsymbol{v}}_n^\top C \hat{\boldsymbol{v}}_n}} \right\}_{n \in J}$

$Y \leftarrow \emptyset$

while $|\hat{V}| > 0$ **do**

  Select i from $\mathcal{Y}$ with $\Pr(i) = \frac{1}{|\hat{V}|} \sum_{\hat{\boldsymbol{v}} \in \hat{V}} (\hat{\boldsymbol{v}}^\top B_i)^2$

  $Y \leftarrow Y \cup i$

  Let $\hat{\boldsymbol{v}}_0$ be a vector in $\hat{V}$ with $B_i^\top \hat{\boldsymbol{v}}_0 \neq 0$

  Update $\hat{V} \leftarrow \left\{ \hat{\boldsymbol{v}} - \frac{\hat{\boldsymbol{v}}^\top B_i}{\hat{\boldsymbol{v}}_0^\top B_i} \hat{\boldsymbol{v}}_0 \mid \hat{\boldsymbol{v}} \in \hat{V} - \{\hat{\boldsymbol{v}}_0\} \right\}$

  Orthonormalize $\hat{V}$ with respect to the dot product $\langle \hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_2 \rangle = \hat{\boldsymbol{v}}_1^\top C \hat{\boldsymbol{v}}_2$

end while

Output: Y

Kulesza and Taskar, 2013

Quality-diversity representation

In addition to the Gram matrix representation $L = B^T B$, we can factor each column B_i as the product of a 'quality' term $q_i > 0$ and a normalized 'diversity feature' $\phi_i \in \mathbb{R}^D$. Thus,

$$L_{ij} = q_i \phi_i^T \phi_j q_j \,.$$

If q_i communicates the 'goodness' of item i, then

$$S_{ij} = \frac{L_{ij}}{\sqrt{L_{ii} L_{jj}}} \,.$$

This representation allows one to independently model quality and diversity using the model

$$\Pr_L(Y) \propto \left(\prod_{i \in Y} q_i^2 \right) \det(S_Y)$$

Conditional DPPs

- A conditional DPP takes the form of an L-ensemble

$$\Pr_L(Y|X) \propto \det(L_Y(X))\,.$$

- L is a positive semi-definite kernel matrix.
- The normalizing constant takes the form $\det(L(X) + I)$.
- Using the quality-diversity decomposition, we have

$$L_{ij}(X) = q_i(X)\phi_i(X)^T\phi_j(X)q_j(X)$$

for $q_i > 0$, $\phi_i \in \mathbb{R}^D$ and $||\phi_i|| = 1$.

Supervised learning

We observe $\{Y_t, X_t\}_{t=1}^{T}$ and assume individual Y_ts generated independently with probabilities

$$\Pr(Y|X,\theta) = \frac{\det(L_Y(X,\theta))}{\det(L(X,\theta)+I)}\,.$$

Then the log-likelihood takes the form

$$\begin{aligned}\ell(\theta) &= \log\left(\prod_{t=1}^{T}\Pr(Y_t|X_t,\theta)\right)\\ &= \sum_{t=1}^{T}\Big(\log\det\left(L_{Y_t}\left(X_t,\theta\right)\right) - \log\det\left(L\left(X_t,\theta\right)+I\right)\Big)\,.\end{aligned}$$

Supervised learning

Suppose one keeps the feature functions $\phi_i(X)$ fixed but models the quality scores with the log-linear model

$$q_i(X, \theta) = e^{f_i(X)^T \theta} .$$

Then the probability of a single sample can be written

$$\Pr(Y|X, \theta) = \frac{\det S_Y \prod_{i \in Y} e^{f_i(X)^T \theta}}{\sum_{Y' \subseteq \mathcal{Y}} \det S_{Y'} \prod_{i \in Y'} e^{f_i(X)^T \theta}} .$$

The resulting log-likelihood is convex in θ:

$$\ell(\theta) \propto \theta^T \sum_{i \in Y} f_i(X) - \log \sum_{Y' \subseteq \mathcal{Y}} \exp\left(\theta^T \sum_{i \in Y'} f_i(X)\right) \det S_{Y'}(X) .$$

k-DPPs

- A k-DPP on a discrete set $\mathcal{Y} = \{1, 2, \ldots, N\}$ is a distribution over all sets $Y \subseteq \mathcal{Y}$ with cardinality k.

- A k-DPP is obtained by conditioning a standard DPP on the event that the set Y has cardinality k.

- The k-DPP N_L^k has probabilities

$$\Pr_L^k(Y) = \frac{\det(L_Y)}{\sum_{|Y'|=k} \det(L_{Y'})}.$$

k-DPPs: normalization

Define the kth elementary symmetric polynomial on $\lambda_1, \ldots, \lambda_N$

$$e_k(\lambda_1, \ldots, \lambda_N) = \sum_{\substack{J \subseteq \{1,\ldots,N\} \\ |J|=k}} \prod_{n \in J} \lambda_n \,.$$

For example,

$$\begin{aligned} e_1(\lambda_1, \lambda_2, \lambda_3) &= \lambda_1 + \lambda_2 + \lambda_3 \\ e_2(\lambda_1, \lambda_2, \lambda_3) &= \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3 \\ e_3(\lambda_1, \lambda_2, \lambda_3) &= \lambda_1\lambda_2\lambda_3 \,. \end{aligned}$$

Proposition 2

The normalizing constant for a k-DPP is

$$Z_k = \sum_{|Y'|=k} \det(L_{Y'}) = e_k(\lambda_1, \ldots, \lambda_N) \,,$$

where λ_n are the eigenvalues of L.

k-DPPs: normalization

Proof.
Recalling that

$$\sum_{Y \subseteq \mathcal{Y}} \det(L_Y) = \det(L + I)\,,$$

we know

$$\sum_{|Y'|=k} \det(L_{Y'}) = \det(L + I) \sum_{|Y'|=k} \Pr_L(Y')\,.$$

Then, because every DPP is a mixture of elementary DPPs:

$$\begin{aligned}
\det(L+I) \sum_{|Y'|=k} \Pr_L(Y') &= \frac{\det(L+I)}{\det(L+I)} \sum_{|Y'|=k} \sum_{J \subseteq \{1,\ldots,N\}} \Pr^{V_J}(Y') \prod_{n \in J} \lambda_n \\
&= \sum_{|J|=k} \sum_{|Y'|=k} \Pr^{V_J}(Y') \prod_{n \in J} \lambda_n \\
&= \sum_{|J|=k} \prod_{n \in J} \lambda_n\,.
\end{aligned}$$

□

Computing elementary symmetric polynomials

Use the shorthand $e_k^N = e_K(\lambda_1, \ldots, \lambda_N)$, we have the recursion

$$e_k^N = e_k^{N-1} \lambda_N e_{k-1}^{N-1} \,.$$

Thus, the following algorithm computes e_k^N in time $O(Nk)$.

Algorithm 7 Computing the elementary symmetric polynomials

Input: k, eigenvalues $\lambda_1, \lambda_2, \ldots \lambda_N$
$e_0^n \leftarrow 1 \quad \forall\, n \in \{0, 1, 2, \ldots, N\}$
$e_l^0 \leftarrow 0 \quad \forall\, l \in \{1, 2, \ldots, k\}$
for $l = 1, 2, \ldots k$ **do**
  for $n = 1, 2, \ldots, N$ **do**
    $e_l^n \leftarrow e_l^{n-1} + \lambda_n e_{l-1}^{n-1}$
  end for
end for
Output: $e_k(\lambda_1, \lambda_2, \ldots, \lambda_N) = e_k^N$

Kulesza and Taskar, 2013

k-DPPs: sampling

- One may use a (slow) rejection sampling approach, sampling DPPs and discarding those for which $|Y| \neq k$.

- It is more efficient to first recognize that, when $|Y| = k$

$$\Pr_L^k(Y) = \frac{\det(L+I)}{e_k^N} \Pr_L(Y)$$

  and therefore

$$\Pr_L^k(Y) = \frac{1}{e_k^N} \sum_{|J|=k} \Pr^{V_J}(Y) \prod_{n \in J} \lambda_n \,.$$

- A k-DPP is also a mixture of elementary DPPs! So *if* we can sample k eigenvalues, we can then use the mixture of elementary DPPs to generate samples.

k-DPPs: sampling

The following $O(Nk)$ algorithm samples sets of k eigenvalues according to desired probabilities

$$\Pr(J) = \frac{1\{|J| = k\}}{e_k^N} \prod_{n \in J} \lambda_n \,.$$

Algorithm 8 Sampling k eigenvectors

Input: k, eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$
compute e_l^n for $l = 0, 1, \ldots, k$ and $n = 0, 1, \ldots, N$ (Algorithm 7)
$J \leftarrow \emptyset$
$l \leftarrow k$
for $n = N, \ldots, 2, 1$ **do**
  if $l = 0$ **then**
    break
  end if
  if $u \sim U[0,1] < \lambda_n \frac{e_{l-1}^{n-1}}{e_l^n}$ **then**
    $J \leftarrow J \cup \{n\}$
    $l \leftarrow l - 1$
  end if
end for
Output: J

Kulesza and Taskar, 2013

k-DPPs: marginalization

Recall that for a general L-ensemble, we have

$$\begin{aligned}\Pr_L(B \subseteq Y | A \subseteq Y) &= \det\left(\left[I - \left[(L + I_{A^c})^{-1}\right]_{A^c}\right]_B\right) \\ &= \det(L_B^A)\,.\end{aligned}$$

k-DPPs: marginalization

k-DPPs are not DPPs and do not have a marginal kernel. But for $|A| \leq k$, we have:

$$
\begin{aligned}
\Pr_L^k(A \subseteq Y) &= \sum_{\substack{|Y'|=k-|A| \\ Y' \cap A = \emptyset}} \Pr_L^k(Y' \cup A) \\
&= \frac{\det(L+I)}{Z_k} \sum_{\substack{|Y'|=k-|A| \\ Y' \cap A = \emptyset}} \Pr_L(Y' \cup A) \\
&= \frac{\det(L+I)}{Z_k} \sum_{\substack{|Y'|=k-|A| \\ Y' \cap A = \emptyset}} \Pr_L(Y = Y' \cup A | A \subseteq Y) \Pr_L(A \subseteq Y) \\
&= \frac{Z_{k-|A|}^A}{Z_k} \frac{\det(L+I)}{\det(L^A+I)} \Pr_L(A \subseteq Y)\,,
\end{aligned}
$$

where

$$
Z_{k-|A|}^A = \det(L^A + I) \sum_{\substack{|Y'|=k-|A| \\ Y' \cap A = \emptyset}} \Pr_L(Y = Y' \cup A | A \subseteq Y) = \sum_{\substack{|Y'|=k-|A| \\ Y' \cap A = \emptyset}} \det(L_{Y'}^A)
$$

is the normalizing constant for the $(k - |A|)$-DPP with kernel L^A.

k-DPPs: marginalization

Thus, the marginal probabilities for a k-DPP are the same as those of the DPP with the same kernel but properly renormalized. By observing that

$$\frac{\det(L^A)}{\det(L+I)} = \frac{\Pr_L(A \subseteq Y)}{\det(L^A + I)},$$

(since $1/\det(L^A + I)$ is the probability of observing nothing else conditioned on A), the equation simplifies further:

$$\begin{aligned} \Pr_L^k(A \subseteq Y) &= \frac{Z^A_{k-|A|}}{Z_k} \frac{\det(L+I)}{\det(L^A+I)} \Pr_L(A \subseteq Y) \\ &= \frac{Z^A_{k-|A|}}{Z_k} \det(L^A) = Z^A_{k-|A|} \Pr_L^k(A). \end{aligned}$$

Computing such a probability is $O((N - |A|)^3)$ and very inefficient for $|A|$ small.

k-DPPs: singleton marginals

First, write the marginal probability of an item i using elementary DPPs:

$$\Pr_L^k(i \in Y) = \frac{1}{e_k^N} \sum_{|J|=k} \Pr^{V_J}(i \in Y) \prod_{n' \in J} \lambda_{n'} \, .$$

But the marginal kernel of an elementary DPP is $\sum_{n \in J} v_n v_n^T$, so this becomes:

$$\begin{aligned}
\Pr_L^k(i \in Y) &= \frac{1}{e_k^N} \sum_{|J|=k} \left(\sum_{n \in J} (e_i^T v_n)^2 \right) \prod_{n' \in J} \lambda_{n'} \\
&= \frac{1}{e_k^N} \sum_{n=1}^{N} (e_i^T v_n)^2 \sum_{\substack{J \supset \{n\} \\ |J|=k}} \prod_{n' \in J} \lambda_{n'} \\
&= \sum_{n=1}^{N} (e_i^T v_n)^2 \lambda_n \frac{e_{k-1}^{-n}}{e_k^N}
\end{aligned}$$

If we have the eigendecomposition of L and know the values e_{k-1}^{-n}/e_k^N, then we can obtain all singleton marginals in time $O(N^2)$. e_k^N can be computed in time $O(Nk)$ and *all* e_{k-1}^{-n} can be computed in time $O(N^2 k)$. This can be improved to $O(N \log(N) k)$.

k-DPPs: conditioning

For $|A| + |B| = k$,

$$\begin{aligned}
\Pr_L^k(Y = A \cup B | A \subseteq Y) &\propto \Pr_L^k(Y = A \cup B) \\
&\propto \Pr_L(Y = A \cup B) \\
&\propto \Pr_L(Y = A \cup B | A \subseteq Y) \\
&\propto \det(L_B^A)\,.
\end{aligned}$$

So the conditional k-DPP is a (k$-|A|$)-DPP.