Auxiliary Notes on Determinantal Point Processes

Wray Buntine

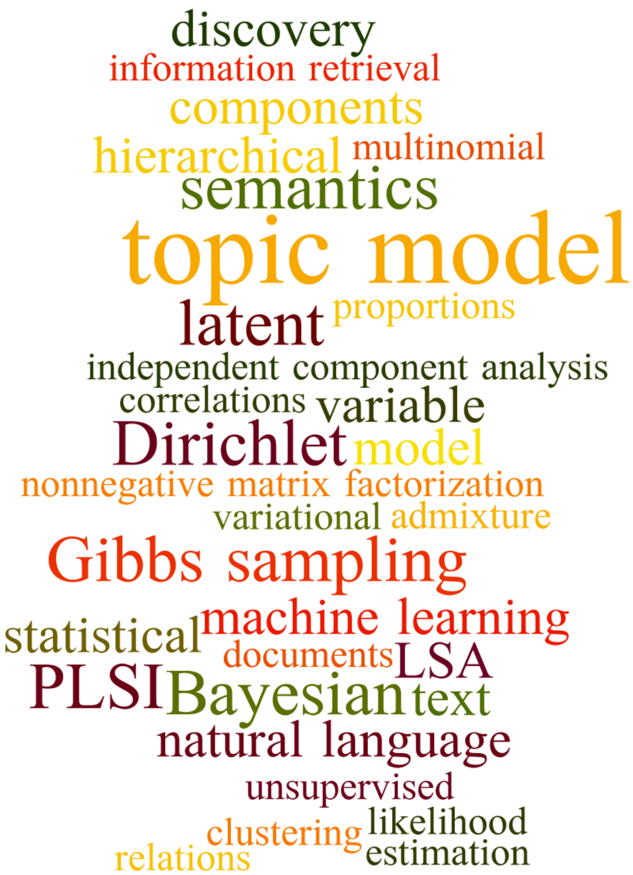
Professor in Faculty of IT
Director of Master of Data Science
Monash University
http://topicmodels.org

2017-09-14

Resources

- Alex Kulesza and Ben Taskar (2012), *"Determinantal Point Processes for Machine Learning"*, *Foundations and Trends in Machine Learning*: Vol. 5: No. 23, pp 123-286.
  - A long article with all the major results for the discrete case.
- The big set of slides are from *a UAI 2011 tutorial* up at *Taskar's website*.
- A compressed version of these are at *Videolectures.NET* with both slides and video, synced.
- Matlab sampling algorithms at *Alex Kulesza's website*.

Outline


discovery
information retrieval
components
hierarchical
multinomial
semantics
topic model
latent
proportions
independent component analysis
correlations
variable
Dirichlet
model
nonnegative matrix factorization
variational
admixture
Gibbs sampling
statistical
machine learning
documents
LSA
PLSI
Bayesian
text
natural language
unsupervised
clustering
likelihood
relations
estimation

Reminder: Volumes and Gram Matrices

- Consider $\det(L_Y) = \det\left(\vec{g}(i)^T \vec{g}(j) \,:\, i,j \in Y\right)$.
- By analytic geometry, the right determinant is interpreted as the volume of the parallelpiped formed by the vectors $\vec{g}(i) \,:\, i \in Y$.
- Details appearing in proof of Theorem 2.3 in L&T.

- Let $\vec{e}_i$ be the unit vector in the i-th dimension, so $\vec{e}_2$ is $(0,1,0,0,...,0)$. Then assuming $1 \in Y$,

$$\mathrm{Vol}\Big(\vec{g}(i) : i \in Y\Big) \;=\; \parallel \vec{g}(1) \parallel_2 \;\mathrm{Vol}\Big(\mathrm{Proj}_{\perp \vec{e}_1} \vec{g}(i) : i \in Y - \{1\}\Big)$$

Reminder: Volumes and Gram Matrices

- Consider $\det(L_Y) = \det\left(\vec{g}(i)^T \vec{g}(j) \,:\, i,j \in Y\right)$.
- By analytic geometry, the right determinant is interpreted as the volume of the parallelpiped formed by the vectors $\vec{g}(i) \,:\, i \in Y$.
- Details appearing in proof of Theorem 2.3 in L&T.
- Let $\vec{e}_i$ be the unit vector in the i-th dimension, so $\vec{e}_2$ is $(0, 1, 0, 0, ..., 0)$. Then assuming $1 \in Y$,

$$\mathrm{Vol}\Big(\vec{g}(i) : i \in Y\Big) \;=\; \parallel \vec{g}(1) \parallel_2 \;\mathrm{Vol}\Big(\mathrm{Proj}_{\perp \vec{e}_1} \vec{g}(i) : i \in Y - \{1\}\Big)$$

- This is equivalent to computing a determinant using the Schur complement formula on dimensions $\{1\}$ and $Y - \{1\}$.

Outline

Marginals

L-ensembles and Marginal Kernels

The representation L-ensembles uses a kernel L and samples Y from a DPP using

$$p(Y = A|L) = \frac{\det(L_A)}{\det(L+I)}$$

where L_Y denotes the principle minor of L by restricting it to the columns and rows Y.

Create a new representation called a marginal kernel K for which instead we compute the marginal for the DPP

$$p(Y \supseteq A|K) = \det(K_A)$$

But, how do we derive K from L?

Converting an L-ensemble to a Marginal Kernel

Given an L-ensemble with kernel L, what is the corresponding marginal kernel K? (Theorem 2.2 in L&T)

Now $p(A \subseteq Y) = \frac{1}{det(L+I)} \sum_{A \subseteq Y \subseteq \mathcal{Y}} det(L_Y) = \frac{\det(L+I_{\overline{A}})}{\det(L+I)}$ using similar arguments to those used in getting the normaliser. Simplify:

$$
\begin{aligned}
p(A \subseteq Y) &= \det\left((L + I_{\overline{A}})(L+I)^{-1}\right) \\
&\qquad\qquad\qquad \text{(noting } L(L+I)^{-1} = I - (L+I)^{-1}\text{)} \\
&= \det\left(I_{\overline{A}}(L+I)^{-1} + I - (L+I)^{-1}\right) \\
&= \det\left(I_{\overline{A}} + I_A(I - (L+I)^{-1})\right) \qquad \text{(rearranging)} \\
&= \det\left(I_{\overline{A}} + (I - (L+I)^{-1})_A\right) \qquad \text{(note zero block for } (\overline{A}, A)\text{)} \\
&= \det\left((I - (L+I)^{-1})_A\right) \qquad \text{(dropping } \overline{A} \text{ dimension)}
\end{aligned}
$$

So use $K = I - (L+I)^{-1}$ as a marginal kernel.

Determinant of Different Dimensions

$$
\begin{aligned}
\det(K_{\{1\}}) &= K_{1,1} \\
\det(K_{\{1,2\}}) &= K_{1,1}K_{2,2} - K_{1,2}^2 \\
\det(K_{\{1,2,3\}}) &= K_{1,1}K_{2,2}K_{3,3} \\
&\quad - K_{1,1}K_{2,3}^2 - K_{2,2}K_{1,3}^2 - K_{3,3}K_{1,2}^2 \\
&\quad + 2K_{1,2}K_{2,3}K_{1,3}
\end{aligned}
$$

Things get more complex in higher dimensions because there are many more potential cycles in a graph of size ≥ 4.

Determinant of Different Dimensions, cont.

In general for a symmetric matrix

$$\det(K_Y) = \sum_{\sigma \in \text{orders}(Y)} (-1)^{\text{sign}(\sigma)} \prod_{x \in Y} K_{x,\sigma(x)}$$

Consider the terms inside the sum, $(-1)^{\text{sign}(\sigma)} \prod_{x \in Y} K_{x,\sigma(x)}$.

This must only consist of multiples of

- singletons $K_{x,x}$,
- doubles, $-K^2_{x,y}$,
- products over pure cycles $x_1, x_2, ..., x_K$ of length $K > 2$, $(-1)^{K+1} K_{x_1,x_2} K_{x_2,x_3} ... K_{x_{K-1},x_K} K_{x_K,x_1}$

The leads to a theory of symmetric matrices based on the cycles induced by the matrix.

Outline

Conditionals

Conditioning on Excluded Points

For universal set $\mathcal{Y}$, have a set A to exclude so $Y \cap A = \emptyset$.

- Proportionally, things don't change, so we can use $L_{\overline{A}}$ as the L-ensemble kernel.
    - has zeroes in all rows/columns associated with A so has the right property
- We also need to renormalise, so the new normaliser is $\det(L_{\overline{A}} + I)$.
- Denote the marginal kernel conditioned on $\overline{A}$ as $K^{\overline{A}}$. From above

$$K^{\overline{A}} = I - (L_{\overline{A}} + I)^{-1}$$

- Its messy but we could represent this in terms of K.
    - The result is on the next page for the simple case of $A = \{x\}$.

Conditioning on Singleton Sets

We can take the previous formula and compute the change in K given A is a singleton, so $A = \{x\}$. This can be done for A included or excluded.

- Including $\{x\}$, where $\vec{k}_x$ is the x-th column of K:

$$K^x = \left[\begin{array}{cc} K_{\overline{x}} - \left(\frac{1}{K_{xx}} \vec{k}_x \vec{k}_x^T \right)_{\overline{x}} & \vec{0} \\ \vec{0}^T & 1 \end{array} \right]$$

- Excluding $\{x\}$ (proof in appendix):

$$K^{\overline{x}} = \left[\begin{array}{cc} K_{\overline{x}} - \left(\frac{1}{1-K_{xx}} \vec{k}_x \vec{k}_x^T \right)_{\overline{x}} & \vec{0} \\ \vec{0}^T & 0 \end{array} \right]$$

- Note the first kernel ensures x is always included (with $K^x_{xx} = 1$) and the second ensures x is always excluded (with $K^{\overline{x}}_{xx} = 0$).
- Therefore, conditioning a DPP on points to include or exclude always yields another DPP.

Outline

Representation: Summary

Summarising the Representations

There are four different representations:

- marginal kernel K,
  - positive semi-definite with all eigenvalues ≤ 1, square in N;
- the L-ensemble kernel L,
  - positive semi-definite, square in N;
- the gram matrix form $L = BB^T$ based on feature matrix B, which is
  - rectangular with N rows, and D columns (number of "features");

  this has an associated dual form $C = B^T B$ which is $D \times D$;
- the quality-diversity decomposition of the feature matrix $B = Q\Phi$, where Q is diagonal and rows of Φ are unit vectors,
  - though rows of Φ are not orthogonal.

Mapping Formulas

Mapping formulas are:

$$
\begin{aligned}
K &= L(I+L)^{-1} = I - (I+L)^{-1} \\
L &= K(I-K)^{-1} = (I-K)^{-1} - I \quad \text{(if } K\text{'s eigenvalues all} < 1) \\
L &= BB^T \\
L &= Q\Phi\Phi^T Q \\
K &= B(I+B^T B)^{-1} B^T \qquad \text{(from Woodbury identity on matrices)} \\
K &= Q\Phi(I+\Phi^T Q^2 \Phi)^{-1} \Phi^T Q
\end{aligned}
$$

Mapping Eigen Values/Vectors

Moreover mapping between eigenvalues and vectors of K, L is easily done:

$$\begin{aligned} K &= L(I+L)^{-1} = I-(I+L)^{-1} \\ L &= K(I-K)^{-1} = (I-K)^{-1} - I \end{aligned}$$

- If $(\lambda_k, \vec{v}_k)$ are eigen value-vector pairs for L, then $\left(\frac{\lambda_k}{1+\lambda_k}, \vec{v}_k\right)$ are eigen value-vector pairs for K.
- From (2), in reverse, If λ_k is eigenvalue of K, then $\frac{\lambda_k}{1-\lambda_k}$ is the eigenvalue for L.

Mapping Eigen Values/Vectors, cont.

Moreover mapping between eigenvalues and vectors of K, L and the dual form $C = B^T B$ is easily done:

$$\begin{aligned} L &= BB^T \\ K &= B(I + B^T B)^{-1} B^T \end{aligned}$$

Use the K form because it contains the dual form $C = B^T B$:

- if $B^T B = \sum_n \lambda_n \vec{u}_n \vec{u}_n^T$, then

$$K = B\left(I + \sum_n \lambda_n \vec{u}_n \vec{u}_n^T\right)^{-1} B^T = \sum_n \frac{1}{1+\lambda_n} (B\vec{u}_n)(B\vec{u}_n))^T$$

- since u_k is eigenvector, $\vec{u}_k B^T B \vec{u}_l = \lambda_k \vec{u}_k \vec{u}_l = \lambda_k \delta_{k,l}$, so $B\vec{u}_l$ are orthogonal

Mapping Eigen Values/Vectors, cont.

Results:

- If $(\lambda_k, \vec{v}_k)$ are eigen value-vector pairs for $C = B^T B$, then $\left(\frac{\lambda_k}{1+\lambda_k}, \frac{1}{\sqrt{\lambda_k}} B\vec{v}_k\right)$ are eigen value-vector pairs for K.
- Following on, $\left(\lambda_k, \frac{1}{\sqrt{\lambda_k}} B\vec{v}_k\right)$ are eigen value-vector pairs for L.

Note the gram matrix becomes efficient to use when $D \ll N$, since the eigen value-vectors can be found via $C = B^T B$ in $O(D^2 N)$: first for $C = B^T B$ and then for K, L.

Determining a Marginal Kernel from Data

If we have loads of data, samples $Y \sim \text{DPP}(K)$, how could we estimate K?

In what follows, expectations are over $Y \sim \text{DPP}(K)$.

- $K_{xx} = \mathcal{E}[p(x \in Y)]$
- $K_{x,y}^2 = K_{xx}K_{yy} - \mathcal{E}[p(x, y \in Y)]$ for $x \neq y$.
- Determining sign of $K_{x,y}$ is complicated:
  - If no $K_{x,y} = 0$, it can be found from 3rd order moments $\mathcal{E}[p(x, y, z \in Y)]$,
  - otherwise, analysis of cycles and a cycle basis is needed, as per Urschel et al.

The problem of estimating a marginal kernal K for a DPP from samples was studied by Urschel, Brunel, Moitra and Rigollet, ICML 2017.

Outline

Representation: Cycles

Special Cases: Simple Cycle

Definition of induced graph

The induced graph corresponding to a marginal kernel is built by connecting all pairs $x \neq y \in \mathcal{Y}$ where $K_{xy} \neq 0$.

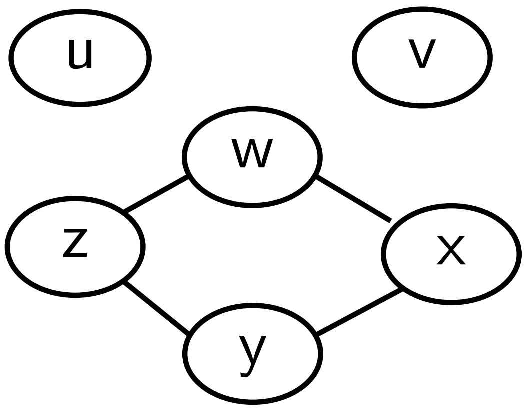
Definition of a simple cycle

In graph theory, the induced graph is a simple cycle when some nodes, say $A \subseteq \mathcal{Y}$, are singleton, with no connections, and the remaining set $\overline{A}$ has a single cycle: so all nodes $x \in \overline{A}$ have exactly two edges and the graph on $\overline{A}$ is connected.

Special Cases: Simple Cycle, cont.

Let $\mathcal{Y} = \{u, v, w, x, y, z\}$,
the kernel K be

$$\begin{bmatrix} \vec{k}_u^T \\ \vec{k}_v^T \\ \vec{k}_w^T \\ \vec{k}_x^T \\ \vec{k}_y^T \\ \vec{k}_z^T \end{bmatrix} = \begin{bmatrix} p & 0 & 0 & 0 & 0 & 0 \\ 0 & p & 0 & 0 & 0 & 0 \\ 0 & 0 & p & a & 0 & d \\ 0 & 0 & a & p & b & 0 \\ 0 & 0 & 0 & b & p & c \\ 0 & 0 & d & 0 & c & p \end{bmatrix}$$

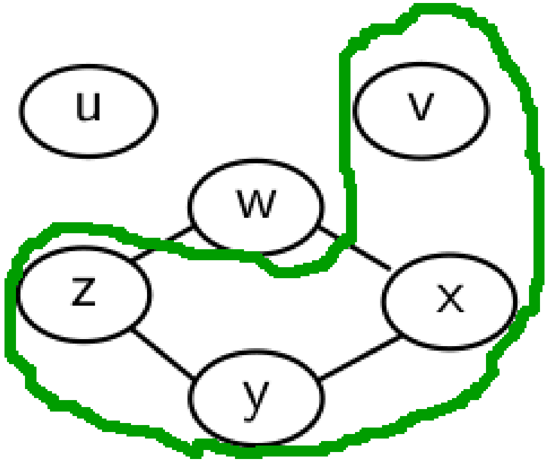


- There is a cycle of length 4 given by x, y, z, w.
- $A = \{u, v\}$ and $\overline{A} = \{x, y, z, w\}$.
- Induced graph has a simple cycle.

Special Cases: Simple Cycle, cont.

Let $\mathcal{Y} = \{u, v, w, x, y, z\}$,
the kernel K be

$$\begin{bmatrix} \vec{k}_u^T \\ \vec{k}_v^T \\ \vec{k}_w^T \\ \vec{k}_x^T \\ \vec{k}_y^T \\ \vec{k}_z^T \end{bmatrix} = \begin{bmatrix} p & 0 & 0 & 0 & 0 & 0 \\ 0 & p & 0 & 0 & 0 & 0 \\ 0 & 0 & p & a & 0 & d \\ 0 & 0 & a & p & b & 0 \\ 0 & 0 & 0 & b & p & c \\ 0 & 0 & d & 0 & c & p \end{bmatrix}$$

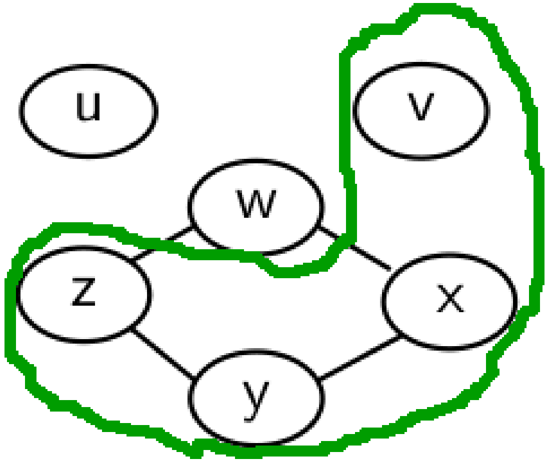


- Let $Y = \{v, w, x, y\}$, so $\overline{A} \not\subseteq Y$.
- $\det(K_{v,w,x,y}) = p^4 - p^2a^2 - p^2b^2$.

Special Cases: Simple Cycle, cont.

Let $\mathcal{Y} = \{u, v, w, x, y, z\}$,
the kernel K be

$$\begin{bmatrix} \vec{k}_u^T \\ \vec{k}_v^T \\ \vec{k}_w^T \\ \vec{k}_x^T \\ \vec{k}_y^T \\ \vec{k}_z^T \end{bmatrix} = \begin{bmatrix} p & 0 & 0 & 0 & 0 & 0 \\ 0 & p & 0 & 0 & 0 & 0 \\ 0 & 0 & p & a & 0 & d \\ 0 & 0 & a & p & b & 0 \\ 0 & 0 & 0 & b & p & c \\ 0 & 0 & d & 0 & c & p \end{bmatrix}$$

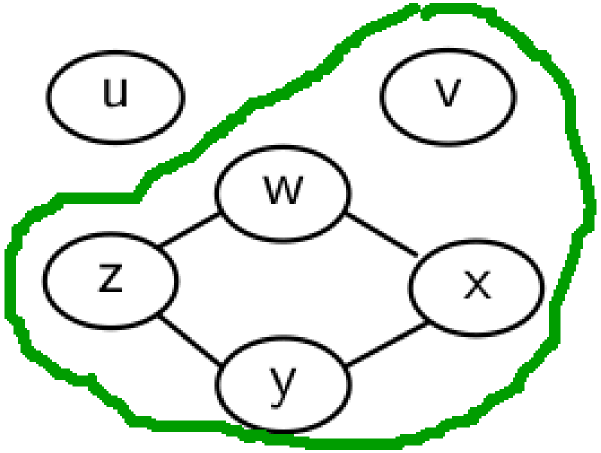


- Let $Y = \{v, w, x, y\}$, so $\overline{A} \not\subseteq Y$.
- $\det(K_{v,w,x,y}) = p^4 - p^2 a^2 - p^2 b^2$.
- If $\overline{A} \not\subseteq Y$ then $\det(K_Y)$ is composed of terms K_{xx} and $-K_{xy}^2$ only.

Special Cases: Simple Cycle, cont.

Let $\mathcal{Y} = \{u, v, w, x, y, z\}$,
the kernel K be

$$\begin{bmatrix} \vec{k}_u^T \\ \vec{k}_v^T \\ \vec{k}_w^T \\ \vec{k}_x^T \\ \vec{k}_y^T \\ \vec{k}_z^T \end{bmatrix} = \begin{bmatrix} p & 0 & 0 & 0 & 0 & 0 \\ 0 & p & 0 & 0 & 0 & 0 \\ 0 & 0 & p & a & 0 & d \\ 0 & 0 & a & p & b & 0 \\ 0 & 0 & 0 & b & p & c \\ 0 & 0 & d & 0 & c & p \end{bmatrix}$$

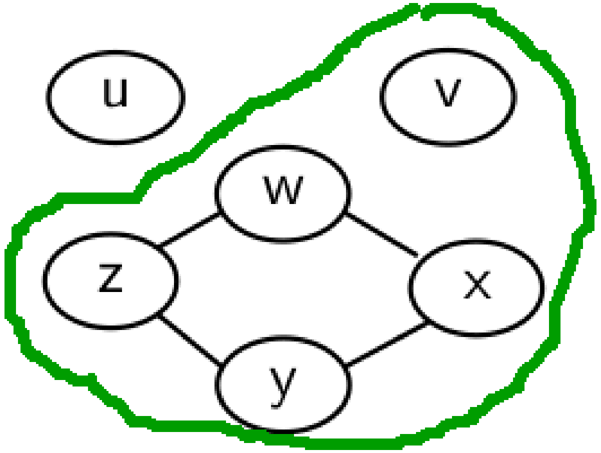


- Let $Y = \{v, w, x, y, z\}$, so $\overline{A} \subseteq Y$.
- $\det(K_{v,w,x,y,z}) = p^5 - p^3 a^2 - p^3 b^2 - p^3 c^2 - p^3 d^2 - 2pabcd$.

Special Cases: Simple Cycle, cont.

Let $\mathcal{Y} = \{u, v, w, x, y, z\}$,
the kernel K be

$$\begin{bmatrix} \vec{k}_u^T \\ \vec{k}_v^T \\ \vec{k}_w^T \\ \vec{k}_x^T \\ \vec{k}_y^T \\ \vec{k}_z^T \end{bmatrix} = \begin{bmatrix} p & 0 & 0 & 0 & 0 & 0 \\ 0 & p & 0 & 0 & 0 & 0 \\ 0 & 0 & p & a & 0 & d \\ 0 & 0 & a & p & b & 0 \\ 0 & 0 & 0 & b & p & c \\ 0 & 0 & d & 0 & c & p \end{bmatrix}$$



- Let $Y = \{v, w, x, y, z\}$, so $\overline{A} \subseteq Y$.
- $\det(K_{v,w,x,y,z}) = p^5 - p^3a^2 - p^3b^2 - p^3c^2 - p^3d^2 - 2pabcd$.
- If $\overline{A} \subseteq Y$ then $\det(K_Y)$ has a term corresponding to the big cycle.

Special Cases: Simple Cycle, cont.

The right matrix has multiplied rows and columns for x by -1.

$$\det\left(\begin{bmatrix} p & 0 & 0 & 0 & 0 & 0 \\ 0 & p & 0 & 0 & 0 & 0 \\ 0 & 0 & p & a & 0 & d \\ 0 & 0 & a & p & b & 0 \\ 0 & 0 & 0 & b & p & c \\ 0 & 0 & d & 0 & c & p \end{bmatrix}\right) = \det\left(\begin{bmatrix} p & 0 & 0 & 0 & 0 & 0 \\ 0 & p & 0 & 0 & 0 & 0 \\ 0 & 0 & p & -a & 0 & d \\ 0 & 0 & -a & p & -b & 0 \\ 0 & 0 & 0 & -b & p & c \\ 0 & 0 & d & 0 & c & p \end{bmatrix}\right)$$

Note this leaves the determinants for all minors unchanged too: for $A \subset \mathcal{Y}$, then $\det(K_A) = \det((DKD)_A)$, where D is a diagonal matrix affecting the scaling.

Special Cases: Scaling by -1

Given a marginal kernel K, if we rescale the x-th row and x-th column by -1 then by multi-linearity of determinants:

- K_{xx} is unchanged (the two -1s cancel)
- $\det(K)$ is unchanged,
- for all $A \subseteq \mathcal{Y}$, $\det(K_A)$ is unchanged.

Hence $\mathrm{DPP}(K) = \mathrm{DPP}(DKD^{-1})$ for D a diagonal matrix with 1s or -1s in the diagonals.

Thus you can arbitrarily fix the sign of uppermost non-zero entry in the 2nd, 3rd, ... N-th columns of K.

Special Cases: Simple Cycle, cont.

The determinant for kernels with induced graphs that are simple cycles of length > 2 have a particular form.

Let σ be an ordering of the elements in the cycle in $\overline{A}$ (the will always be 2 possible). Also, let pairs(Y) be the set of sets of (x, y) pairs from Y with (1) $x \neq y$ and $K_{xy} > 0$ and (2) no x occurs in more than one pair. Then:

$$\det(K_Y) = \sum_{P \in \text{pairs}(Y)} (-1)^{|P|} \prod_{x \in Y/\text{flatten}(P)} K_{xx} \prod_{(x,y) \in P} K_{xy}^2$$
$$+ \ 1_{\overline{A} \subseteq Y}\ 2\ (-1)^{|\overline{A}|+1} \left(\prod_{x \in Y \cap A} K_{xx} \right) K_{\sigma(|\sigma|),\sigma(1)} \prod_{i=1}^{|\sigma|-1} K_{\sigma(i),\sigma(i+1)}$$

More complex formulas apply for more general graphs.

Outline

Sampling

Special Cases: Rank 1

If K is rank 1, so $K = \lambda \vec{u}\vec{u}^T$ for unit vector $\vec{u}$.

Suppose $Y \sim DPP(K)$:

- Since all minors of size ≥ 2 of a rank one matrix have zero determinant: Y is singleton or empty.
- Probability $Y = \emptyset$ is $(1 - \lambda)$.
- Probability $Y = \{x\}$ is λu_x^2.

Sampling an Elementary DPP

Elementary DPPs can be sampled easily. Given marginal kernel K which is for an elementary DPP of rank J.

- Suppose $Y \sim \text{DPP}(K)$. $p(x \in Y) = K_{xx}$ is the probability one of the J points in Y is x. By symmetry $\frac{1}{J} K_{xx}$ is the probability that the first point is x.
- Thus we can sample the first point using $p(x) = \frac{1}{J} K_{xx}$.
- We then condition on the "including $\{x\}$" case given previously,

$$K^x = \begin{bmatrix} K_{\overline{x}} - \left(\frac{1}{K_{xx}} \vec{k}_x \vec{k}_x^T \right)_{\overline{x}} & \vec{0} \\ \vec{0}^T & 1 \end{bmatrix}$$

  Thus turns out to be converting each $\vec{u}_n$ to be perpendicular to the x-th dimension. Details in Kulesza and Taskar 2012.
- The resultant DPP, K^x must be rank $J-1$, so we recurse on K^x.

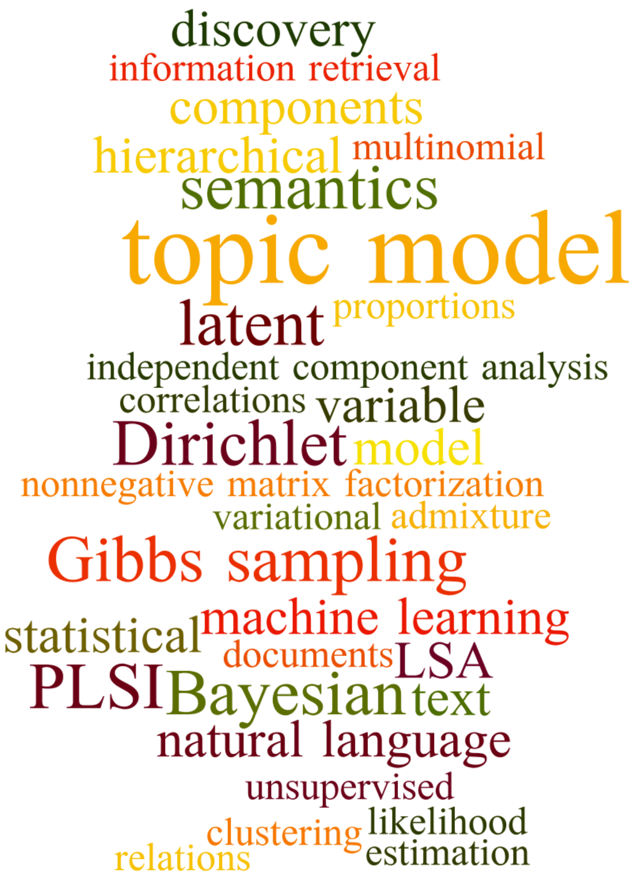
Outline

Conclusion

Other Issues

- Scaling up done with the gram matrix and the alternate $B^T B$ form which is D^2 size, not N^2.
- DPPs work well with random projections.
- In some cases want exactly K points in the sample, i.e., in presenting top-K results.
  - L&T show how to condition a DPP on wanting exactly K points.
- Finding then "best" sample, MAP inference on a DPP, is NP-hard but the search space is sub-modular, so good algorithms exist.

Outline


discovery
information retrieval
components
hierarchical multinomial
semantics
topic model
latent proportions
independent component analysis
correlations variable
Dirichlet model
nonnegative matrix factorization
variational admixture
Gibbs sampling
statistical machine learning
documents LSA
PLSI Bayesian text
natural language
unsupervised
clustering likelihood
relations estimation

Proving the Exclude Case for Conditioning on Singleton

- Have $K^x = I - (L_{\overline{x}} + I)^{-1}$. So arrange $L_{\overline{A}} + I$ as a block matrix with edge dimensions $N-1$ and 1, and arbitrarily make x the last dimension.

$$\begin{pmatrix} L_{\overline{x}} + I_{\overline{x}} & \vec{l}_x \\ \vec{l}_x^T & (L_{xx}+1) \end{pmatrix}^{-1} = \begin{pmatrix} I_{\overline{x}} - K_{\overline{x}} & -\vec{k}_x \\ -\vec{k}_x^T & (1-K_{xx}) \end{pmatrix}$$

- Match this up with the block matrix inversion formula:

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} C_1^{-1} & -A_{11}^{-1}A_{12}C_2^{-1} \\ -C_2^{-1}A_{12}^TA_{11}^{-1} & C_2^{-1} \end{pmatrix}$$

  where $C_1 = A_{11} - A_{22}^{-1}A_{12}A_{12}^T$ and the scalar $C_2 = A_{22} - A_{12}^TA_{11}^{-1}A_{12}$.

- We want $(L_{\overline{x}} + I)^{-1} = \begin{pmatrix} (L_{\overline{x}} + I_{\overline{x}})^{-1} & 0 \\ 0 & 1 \end{pmatrix}$. The embedded inverse is given by A_{11}^{-1}, given by the Woodbury identity:

$$A_{11}^{-1} = C_1^{-1} - \frac{C_1^{-1}A_{12}A_{12}^TC_1^{-1}}{A_{22} + A_{12}C_1^{-1}A_{12}^T} \ . \tag{1}$$

Proving the Exclude Case for Conditioning on Singleton

- Manipulating Equation (1) gives a number of useful intermediate forms:

$$
\begin{aligned}
A_{11}^{-1}A_{12} &= (C_1^{-1}A_{12})\frac{A_{22}}{A_{22}+A_{12}C_1^{-1}A_{12}^T} \\
C_1^{-1}A_{12} &= \vec{k}_x\frac{1}{1-K_{xx}}\frac{A_{22}+A_{12}C_1^{-1}A_{12}^T}{A_{22}} \\
A_{12}^TA_{11}^{-1}A_{12} &= \frac{A_{22}(A_{12}^TC_1^{-1}A_{12})}{A_{22}+A_{12}C_1^{-1}A_{12}^T} \\
\frac{A_{22}^2}{A_{22}+A_{12}C_1^{-1}A_{12}^T} &= A_{22}-A_{12}^TA_{11}^{-1}A_{12} = C_2 = \frac{1}{1-K_{xx}}
\end{aligned}
$$

  The 1st dots Equation (1) with A_{12} and simplifies. The 2nd comes by rearranging the 1st, noting $A_{11}^{-1}A_{12}C_2^{-1} = \vec{k}_x$. The 4th rearranges the 3rd and then uses the definition of C_2.

- Substituting in Equation (1) the above matched terms

$$
(L_{\overline{x}}+I_{\overline{x}})^{-1} = A_{11}^{-1} = I_{\overline{x}} - K_{\overline{x}} - \frac{\vec{k}_x\vec{k}_x^T}{(1-K_{xx})}
$$

  and the result for K^x follows directly using $I - \begin{pmatrix} (L_{\overline{x}}+I_{\overline{x}})^{-1} & 0 \\ 0 & 1 \end{pmatrix}$.