Parameter estimation for discrete determinantal point processes

Dissertation submitted for the degree of
MASTER OF SCIENCE IN INTERDISCIPLINARY MATHEMATICS

Johannes Müller

July 27, 2018

SUPERVISED BY
PROFESSOR NIKOLAOS ZYGOURAS AND DR THEODOROS DAMOULAS

UNIVERSITY OF WARWICK
DEPARTMENT OF MATHEMATICS

Acknowledgements

I would like to thank the dreamteam for being such an eggcelent friend-group and for all the pun we had during the last year.

Abstract

Determinantal point processes are random subsets that exhibit a diversifying behaviour in the sense that the randomly selected points tend to be not similar in some way. This repellent structure first arrose in theortical physics and pure mathematics, but they have recently been used to model a variety of many real world scenarios in a machine learning setup. We aim to give an overview over the main ideas of this approach which is easily accessible even without prior knowledge in the area of machine learning and sometimes omit technical calculations in order to keep the focus on the concepts.

Contents

Chapter I

Introduction and motivating examples

I.1 Motivation

I.2 Previous work

I.3 Aim and outline of the dissertation

Chapter II

Determinantal points processes: Basic notions and properties

II.1 Definitions and properties

We begin by presenting general frame we will work in. This means that we will keep the notation introduced now and will use those objects throughout the thesis without further explanation. Further we will present all the important properties of determinantal point processes that we will need and postpone some calculations to the last section of this chapter. A much more ... survey of properties of determinantal point processes including extensive comparisons to several other point processes can be found in the report [Kulesza et al., 2012].

2.1 Setting. Let in the following $\mathcal{Y}$ be a finite set, which we call the *ground set* and $N := |\mathcal{Y}|$ its cardinality. For the sake of easy notation we will assume $\mathcal{Y} = \{1, \dots, N\}$ unless otherwise specified. A *point process* on $\mathcal{Y}$ is a random subset of $\mathcal{Y}$, i.e. a random variable with values in the powerset $2^{\mathcal{Y}}$. We will identify this random variable with its law $\mathbb{P}$ and thus refer to probability measures $\mathbb{P}$ on $2^{\mathcal{Y}}$ as point processes and will not distinguish between those objects. Let further $\mathbf{Y}$ denote a random subset drawn according to $\mathbb{P}$.

2.2 Definition (Determinantal point process). We call $\mathbb{P}$ a *determinantal point process*, or in short a DPP, if we have

$$\mathbb{P}(A \subseteq \mathbf{Y}) = \det(K_A) \quad \text{for all } A \subseteq \mathcal{Y} \tag{2.1}$$

where K is a symmetric matrix indexed by the elements in $\mathcal{Y}$ and K_A denotes the submatrix of K to indexed by the elements of A. We call K the *marginal kernel* of the DPP. If the marginal kernel K is diagonal, then we call $\mathbb{P}$ a *Poisson point process*.

We note that all principal minors[1] of K are non negative and Sylvester's criterion implies that K is non negative definite[2]. Further it can be shown (cf. page 3 in [Borodin, 2009]) that also the complement of a DPP is a DPP with marginal kernel $I - K$ where I is the identity matrix, i.e.

have a look at this and maybe explain it!

$$\mathbb{P}(A \subseteq \mathbf{Y}^c) = \det(I_A - K_A).$$

[1] The *principle minors* of K are the determinants of the submatrices K_A for $A \subseteq \mathcal{Y}$.

[2] K is called *non negative definite* if $x^T K x \geq 0$ for all $x \in \mathbb{R}^{\mathcal{Y}}$. The Sylvester criterion states that a matrix is non negative definite if and only if all principle minors are non negative.

Thus we conclude $I - K \geq 0$ and obtain $0 \leq K \leq I$. This actually turns out to be sufficient for K to define a DPP through (2.1) which we will see in the fourth section of this chapter.

2.3 Repulsive behaviour of DPPs. We call the elements of $\mathcal{Y}$ *items* and by choosing $A = \{i\}$ and $A = \{i, j\}$ for $i, j \in \mathcal{Y}$ and using (2.1) we obtain the probabilities of their occurrence

$$\begin{aligned} \mathbb{P}(i \in \mathbf{Y}) &= K_{ii} \quad \text{and} \\ \mathbb{P}(i, j \in \mathbf{Y}) &= K_{ii}K_{jj} - K_{ij}^2 = \mathbb{P}(i \in \mathbf{Y}) \cdot \mathbb{P}(j \in \mathbf{Y}) - K_{ij}^2, \end{aligned} \tag{2.2}$$

Thus the appearances of the two items i and j are always negatively correlated. This negative correlation is exactly what causes the diversifying behaviour of determinantal point processes. In practice one usually models the negative correlations to be high between items that are similar in some notion. For example in a spatial setting being similar could mean being close together and in this case the selected items would tend to be not very close together. This is repulsive behaviour can be seen in Figure. Note that Poisson point processes are exactly the DPPs without correlations of the points.

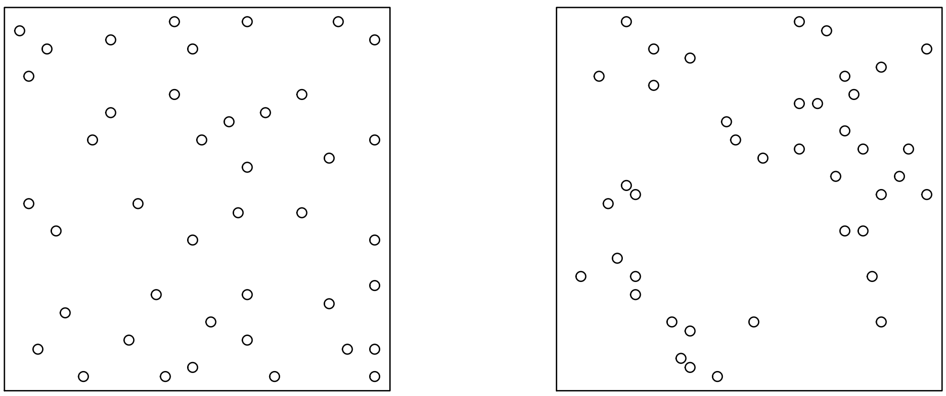

Figure II.1.: A DPP with negative correlations of close points on a 40×40 grid in the unit square on the left and a Poisson point process on the same grid on the right with the same expected cardinality. The – in this case spatially – repellent structure of the DPP is clearly visible.

In this light the fact that also $\mathbf{Y}^c$ exhibits negative correlations becomes less surprising. Since the set $\mathbf{Y}$ tends to spread out due to the repulsion in (2.2), the complement, which is nothing but the gaps that are left after eliminating the elements in $\mathbf{Y}$, tend to show a non clustering behaviour as well.

2.4 L-ensembles. Let us now introduce an important subclass of DPPs, namely the ones where not only the marginal probabilities can be expressed through a suitable kernel, but also the elementary probabilities. This will be convenient for us and lead to some explicit expression. If we have even $K < I$, then we define the *elementary kernel*

$$L := K(I - K)^{-1} \tag{2.3}$$

which specifies the elementary probabilities since one can check

$$\mathbb{P}(A = \mathbf{Y}) = \frac{\det(L_A)}{\det(I + L)} \quad \text{for all } A \subseteq \mathcal{Y}. \tag{2.4}$$

Conversely for any $L \geq 0$ a DPP can be defined via (2.2) and the corresponding marginal kernel is given by the inversion of (2.3)

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

and we have again $K < I$. We call DPPs which arise this way L *ensembles*. Later we will see that the cardinality of a DPP distributed like the sum of N Bernoulli experiments with expectation $(\lambda_n)_{n=1,\dots,N}$ where λ_n are the eigenvalues of K. Being an L-ensemble is equivalent to $K < I$ which again is equivalent to $\lambda_n < 1$ for all $n = 1, \dots, N$ and hence equivalent to

$$\mathbb{P}(\mathbf{Y} = \varnothing) = \mathbb{P}(|\mathbf{Y}| = 0) > 0.$$

The quality diversity decomposition

Note that any symmetric, positive semidefinite matrix L can be written as a Gram matrix

$$L = B^T B$$

where $B \in \mathbb{R}^{D\times N}$ whenever D is larger than the rank $\mathrm{rk}(L)$ of L. For example one could take the spectral decomposition $L = U^T C U$ of L and set $B := \sqrt{C}U$ and eventually drop some zero rows from $\sqrt{C}$. Let B_i denote the i-th column of B and write it as the product $q_i \cdot \phi_i$ where $q_i \geq 0$ and $\phi_i \in \mathbb{R}^D$ such that $\|\phi_i\| = 1$. This yields the representation

$$L_{ij} = q_i \phi_i^T \phi_j q_j =: q_i S_{ij} q_j$$

and we call q_i the *quality* of the item $i \in \mathcal{Y}$ and ϕ_i the *diversity feature vector* of i and S the *similarity matrix*. Since we will use this decomposition multiple times, we fix its properties.

2.5 Proposition (Quality diversity parametrisation). *Let $D \in \mathbb{N}$ and let $\mathbb{S}_D$ denote the sqhere in $\mathbb{R}^D$. Further let $\mathbb{R}^{N\times N}_{sym,+}$ be the set of symmetric positive semidefinite $N \times N$ matrices. The quality diversity parametrisation is a continuous and surjective mapping*

$$\Psi\colon \mathbb{R}^N_+ \times \mathbb{S}^N_D \to \left\{L \in \mathbb{R}^{N\times N}_{sym,+} \mid \mathrm{rk}(L) \leq D\right\}, \quad (q, \phi) \mapsto \left(q_i \phi_i^T \phi_j q_j\right)_{1\leq i,j\leq N}.$$

2.6 Remark. (*i*) In the case $D = N$ the quality diversity decomposition gives a parametrisation of the whole symmetric positive definite $N \times N$ matrices.

(*ii*) Note that this parametrisation is not unique, i.e. Ψ is not injective. For example the identity matrix I can be parametrised by any orthonormal system ϕ and $q = (1, \dots, 1)^T$.

(*iii*) One can without any problems consider diversity features ϕ_i in an abstract Hilbert space $\mathcal{H}$. However we will not need this in the remainder and thus restrict ourselves to the easier case Euklidean diversity features.

(*iv*) We call every preimage (q, ϕ) of L under Ψ *quality diversity decomposition* of L. Further we call the tupel $\phi \in \mathbb{S}_D^N$ of normalised vectors *diversity feature matrix*.

The quality diversity decomposition will provide some useful expressions. For example the elementary probabilities take the form

$$\mathbb{P}(A = \mathbf{Y}) \propto \det\big((B^T B)_A\big) = \left(\prod_{i\in A} q_i^2\right) \cdot \det(S_A) \quad \text{for all } A \subseteq \mathcal{Y}. \tag{2.5}$$

An intuitive understanding of the quality diversity decomposition will play a central role in the modelling process of real world phenomena through DPPs. To get this we can think of $q_i \geq 0$ as a measure of how important or high in quality the item is and the diversity feature vector $\phi_i \in \mathbb{R}^D$ can be thought of as some kind of state vector that consists of internal quantities that describe the item i in some way. Further we interpret the scalar product $\phi_i^T \phi_j \in [0, 1]$ as a measure of similarity between the items i and j which justifies the name similarity matrix for S. Note that if i and j are perfectly similar or antisimilar, i.e. $\phi_i^T \phi_j = \pm 1$, then they can not occur at the same time, since

$$\mathbb{P}(i, j \in \mathbf{Y}) = \det \begin{pmatrix} 1 & \pm 1 \\ \pm 1 & 1 \end{pmatrix} = 0.$$

If we identify i with the vector $B_i = q_i \phi_i \in \mathbb{R}^D$, we can obtain a geometric interpretation of (2.5) since $\det\big((B^T B)_A\big)$ is the volume that is spanned by the columns $B_i, i \in A$, which is visualised in II.2. This volume increases if the lengths of the edges that correspond to the quality increase and decrease when the similarity feature vectors point into more similar directions.

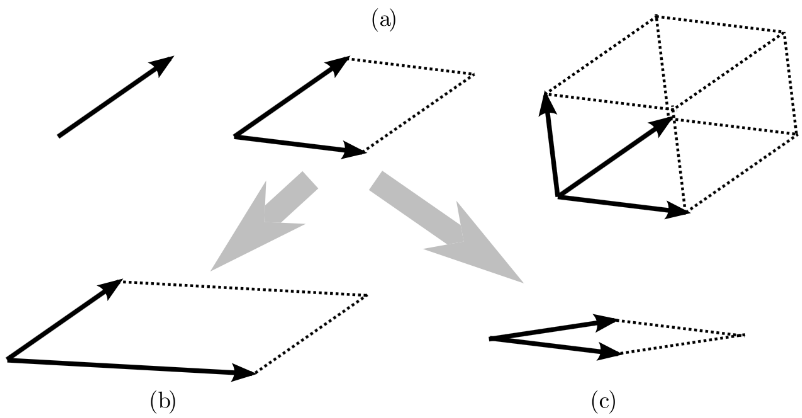

Figure II.2.: Taken from [Kulesza et al., 2012]; the first line (a) illustrates the volumes spanned by vectors, and in the second line it can be seen how this volume increases if the length – associated with the quality – increases (b) and decreases if they become more similar in direction which we interpret as two items becoming more similar (c)

2.7 Modelling diversity over distance. Since we will use one form of diversity features multiple times, we will now give a short general formulation of it. Let $\mathcal{R} = \{r_1, \ldots, r_D\}$ be a finite set which we will call the *reference set* and its elements the *reference points*. Further let

$$d : \mathcal{Y} \times \mathcal{R} \to \mathbb{R}_+, \quad f : \mathbb{R}_+ \to \mathbb{R}$$

mappings. Usually $d(i,r)$ will be interpreted as a measure of distance between an item $i \in \mathcal{Y}$ and a reference point $r \in \mathcal{R}$ and will typically be given by a metric on a larger space that contains both $\mathcal{Y}$ and $\mathcal{R}$. One can now model $\phi_i \in \mathbb{R}^{\mathcal{R}}$ via

$$(\phi_i)_r \propto f\big(d(i,r)\big) \quad \text{for } r \in \mathcal{R}$$

The function f will typically be decreasing and thus $(\phi_i)_r$ can be seen as a measure of how similar item i is to the reference point $r \in \mathcal{R}$. Thus the diversity feature vector ϕ_i stores how similar the item i is to all reference points and the scalar product $\phi_i^T \phi_j$ will be close to one, if the items i and j have approximately the same degrees of similarity to the reference points. It shall be noted that the choice of the D, the number of reference points bounds the rank of the kernel L and therefore of the largest subset that occurs with positive probability. Indeed we have $\mathrm{rk}(L) \le D$ and for $A \subseteq \mathcal{Y}$ with more than D elements $\det(L_A) = 0$ and therefore $\mathbb{P}(A) = 0$. In the fourth chapter we will see that there is a natural choice for the mapping d in most cases, at least in the ones where $\mathcal{Y}$ consists of points in a metric space. On the other hand the choice of f is crucial since it determines the structure and strength of the repulsion.

2.8 Transitivity of repulsion. One last property of DPPs that we shall mention is the fact that the negative correlations of the DPP posses a transient property in the sense, that if i and j and j and k are similar, then i and k are also similar. This is due to the fact

$$\|\phi_i - \phi_j\|^2 = \|\phi_i\|^2 + \|\phi_j\|^2 - 2\phi_i^T\phi_j = 2(1 - \phi_i^T\phi_j)$$

and thus

$$\sqrt{1-\phi_i^T\phi_k} = \frac{1}{2}\|\phi_i - \phi_k\| \le \frac{1}{2}\big(\|\phi_i - \phi_j\| + \|\phi_j - \phi_k\|\big) = \sqrt{1-\phi_i^T\phi_j} + \sqrt{1-\phi_j^T\phi_k}.$$

reformulate that part!

2.9 Comparison to other point processes.

II.2 Variations of DPPs

In this section we will present some useful variations of determinantal point processes. They serve different purposes and we will shortly explain their individual benefits.

2.10 Conditional DPPs. A *conditional DPP* is a collection of DPPs indexed by $X \in \mathcal{X}$, where X is called the *input* of the conditional DPP. Thus for every $X \in \mathcal{X}$ we get a finite set $\mathcal{Y}(X)$ and a determinantal point process $\mathbb{P}(\cdot \mid X)$ on $\mathcal{Y}(X)$ which is given by the elementary kernel $L(X)$, i.e.

$$\mathbb{P}(A|X) \propto \det\big(L_A(X)\big) \quad \text{for all } A \subseteq \mathcal{Y}(X).$$

Further we denote the quality and diversity features of the conditional DPP by $q_i(X)$ and $\phi_i(X)$ respectively.

It is not immediately clear why one would want to model a family of DPPs as a conditional DPP rather than as seperate DPPs. The reason for this is that one wants to estimate the kernels $L(X)$ for every $X \in \mathcal{X}$. However if we would do this naively we would need to observe each of the DPPs $\mathbb{P}(\cdot \mid X)$ individually which is often not possible. Thus one hopes to not only memorise the kernels $L(X)$ for every single input $X \in \mathcal{X}$ but rather to learn the mapping L that assigns every input X its elementary kernel $L(X)$. If one achieved this task, one would be able to simulate and predict a DPP that one has not observed so far just by the knowledge about some DPPs that belong to the same conditional DPP. Of course this can only work if we assume some regularity or a certain structure of the function L which we will do in the third chapter where we put those consideration into a precise framework.

2.11 Fixed size or k-DPPs.

2.12 Structured DPPs. We call a DPP *structured DPP* or short sDPP if the ground set is the cartesian product of some other set $\mathcal{M}$, which we will call the *set of parts*, i.e. if we have

$$\mathcal{Y} = \mathcal{M}^R = \left\{ y_i = (y_i^r)_{r=1,\dots,R} \mid i = 1, \dots, N \right\}$$

where R is a natural number, $M = |\mathcal{M}|$ and $N = M^R$. The quality diversity decomposition of L take the form

$$L_{ij} = q(y_i)\phi(y_i)^T \phi(y_j)q(y_j)$$

and since $N = M^R$ is typically very big, it is impractical to define or store the quality and diversity features for every item $y_i \in \mathcal{Y}$. To deal with this problem we will assume that they admit factorisations and are thus a combination of only a few qualities and diversities.

More precisely we call $F \subseteq 2^{\{1,\dots,R\}}$ a *set of factorisations* and for a *factor* $\alpha \in F$, y_α denotes the subtupel of $y \in \mathcal{Y}$ that is indexed by α. Further we will work with the decompositions

$$\begin{aligned} q(y) &= \prod_{\alpha \in F} q_\alpha(y_\alpha) \\ \phi(y) &= \sum_{\alpha \in F} \phi_\alpha(y_\alpha) \end{aligned} \tag{2.6}$$

for a suitable set of factorisations F and qualities and diversities q_α and ϕ_α for $\alpha \in F$. Note that so far this is neither a restriction of generality – we could simply choose $F = \big\{\{1, \dots, R\}\big\}$ – nor a simplification – in that case we have the exact same number of qualities and diversities. However we are interested in the case where F consists only of small subsets of $\{1, \dots, R\}$. For example suppose that F is the set of all subsets with one or two elements, then we only have

$$R \cdot M + \binom{R}{2} \cdot M^2 = O(R^2 M^2)$$

quality and diversity features instead of

$$M^R = O(M^R).$$

This reduction of variables will make modelling, storing and estimating them feasible again in a lot of cases where naive approaches are foredoomed because of their shear size.

II.3 Simulation and Existence of DPPs

One of the main difficulties that arrises in the theory of discrete point processes is that they are probability measures on an exponentially large set, namely the powerset $2^{\mathcal{Y}}$ which has cardinality 2^N. Determinantal point processes have the benefit that they describe this distribution through the matrix K which consists of only N^2 parameters. This reduction of the number of parameters plays a central role in making a lot of operations possible in an computationally efficient way.

However it is not only the relatively small amount of parameters that lead to this, but also the structure of the determinant itself that leads to closed expressions for a lot of quantities like the normalisation constant in (2.4). In this section we will focus on the efficient simulation of DPPs and give a short overview of further techniques that can improve the performance of this algorithm.

But before we can do this we will present a famous identity for integrals – or sums – of the product of determinants.

Two Cauchy-Binet type identities

The result we present is a little technical, but we will need it later in the proof of existence and also in the proof of the sampling algorithm, so we will give the proof of it here, although it should be mentioned that all the major ideas can be found in [Hough et al., 2006]. In this section write $[n]$ for the set $\{1, \dots, n\}$ where n is a natural number and A_{IJ} for the submatrix of A where the first index is in I and the second one in J. Further we keep the notation $A_I = A_{II}$.

2.13 Proposition (Cauchy-Binet). *Let $m, n \in \mathbb{N}, m \leq n$ be two natural numbers and $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times m}$ two matrices. Then we have*

$$\det(AB) = \sum_{\substack{I \subseteq \{1,\dots,n\} \\ |I|=m}} \det\left(A_{[m]I}\right) \det\left(B_{I[m]}\right).$$

Proof. Let without loss of generality A and B have full rank, otherwise both sides are zero. First we note that both sides are multilinear in the rows of A and columns of B. Hence we can assume by Gaussian row, respectively column elimination that each row of A and each column of B has exactly one non zero entry which is 1. Hence there is $J_1 = \{i_1, \dots, i_m\}$ and $J_2 = \{j_1, \dots, j_m\}$ such that $A_{ki_k} = 1$ and $B_{kj_k} = 1$ and all other entries are empty. Note that the right hand side is only non trivial if $I = J_1$ and $I = J_2$. If $J_1 \neq J_2$ then this case does not occur in the sum, but then at least one row of AB consists of zeroes and hence also the left hand side is equal to 0. If however $J = J_1 = J_2$ we get

really?

$$\det\left(A_{[m]J}\right) \det\left(B_{J[m]}\right) = \det\left(A_{[m]J} B_{J[m]}\right) = \det(AB)$$

since $A_{ij} = B_{ji} = 0$ whenever $j \notin J$. □

2.14 Proposition (Variation of Cauchy-Binet). *Let $m \leq n$ be two natural numbers and $B \in \mathbb{R}^{n \times m}$ be a matrix such that the columns $B_i \in \mathbb{R}^n$ of B form an orthonormal system. Further let $I \subseteq [m]$, then we have*

$$\det\left((B^T B)_I\right) = \sum_{\substack{I \subseteq J \subseteq [n] \\ |J|=m}} \det\left(B_{J[m]}\right)^2.$$

Proof. This can be proved in analogue fashion to the result above. □

check this, I think the statement is slightly wrong...

Sampling and Existence

We roughly follow the approaches taken in [Hough et al., 2006] and [Kulesza et al., 2012] and will start by showing that every determinantal point process can be seen as a mixture of a smaller class of determinantal point processes.

2.15 Theorem (Mixture representation of DPPs). *Let $\mathbb{P}$ be a DPP and*

$$K = \sum_{k=1}^{N} \lambda_k v_k v_k^T$$

be the spectral decomposition of its marginal kernel. Let now $\{B_k\}_{k=1,\dots,N}$ be a collection of independent Bernoulli random variables with mean λ_i. Define now the random kernel

$$K_B = \sum_{k=1}^{N} B_k v_k v_k^T. \tag{2.7}$$

Finally define a second point process $\tilde{\mathbb{P}}$ on $\mathcal{Y}$ that is obtain by first drawing the Bernoulli variables B_k and then a DPP according to K_B. Then we have $\tilde{\mathbb{P}} = \mathbb{P}$ and thus $\tilde{\mathbb{P}}$ is also a DPP with marginal kernel K.

We will postpone the proof and first discuss its consequences which will be the existence of DPPs for a given marginal kernel as well as the construction of a sampling algorithm.

2.16 Remark. Since it is fairly easy to simulate Bernoulli experiments, it remains to know how we can sample from DPPs with marginal kernels of the form $K = \sum_{k=1}^{m} v_k v_k^T$ for some $m \leq N$. We call DPPs of this type *elementary* and note that this corresponds to the class of DPPs where the eigenvalues of the marginal kernel are contained in $\{0, 1\}$.

Now we study the existence and simulation of elementary DPPs first and will be able to generalise those results to general DPPs without much effort.

2.17 Proposition (Existence of elementary DPPs). *Let $K = \sum_{k=1}^{m} v_k v_k^T$ for some orthonormal set $V = \{v_k\}_{k=1,\dots,m} \subseteq \mathbb{R}^{\mathcal{Y}}$. Further define the measure on $2^{\mathcal{Y}}$ through*

$$\mathbb{P}(A) := \begin{cases} \det(K_A) & \text{if } |A| = m \\ 0 & \text{else} \end{cases}. \tag{2.8}$$

Then $\mathbb{P}$ is a DPP on $\mathcal{Y}$ with marginal kernel K. In particular elementary DPPs exist.

Proof. First we have to show that (2.8) defines a probability measure. For this let $B \in \mathbb{R}^{m\times N}$ be the matrix with rows v_k for $k = 1, \dots, m$. By definition we have $K = B^T B$ and hence

$$\begin{aligned}\sum_{A\subseteq\mathcal{Y}, |A|=m} \det(K_A) &= \sum_{\substack{I\subseteq\mathcal{Y}\\|I|=m}} \det(K_I) = \sum_{\substack{I\subseteq\mathcal{Y}\\|I|=m}} \det\left(B_{[m]I}\right)^2 \\ &= \det\left(B^T B\right) = \det\left(v_k^T v_l\right)_{1\leq k,l\leq m} = 1\end{aligned}$$

where we have used the Cauchy-Binet identity and the fact that V is orthonormal. It remains to check that all marginal probabilities satisfy

$$\mathbb{P}(A \subseteq \mathbf{Y}) = \det(K_A).$$

For $|A| \geq m$ this follows immediately, so let $A = \{i_1, \dots, i_r\}$ for $r < m$. Then we obtain the marginal probability of A through integration over the other $m - r$ points. Namely we have

$$\mathbb{P}(A \subseteq \mathbf{Y}) = \sum_{\substack{A\subseteq J\subseteq[n]\\|J|=m}} \mathbb{P}(J \subseteq \mathbf{Y}) = \sum_{\substack{A\subseteq J\subseteq[n]\\|J|=m}} \det\left(B_{[m]J}\right)^2 = \det\left((B^T B)_A\right) = \det(K_A)$$

where we used Proposition 2.14. $\square$

Now we can turn towards the simulation of elementary DPPs where we will make use of the previous result.

Algorithm 1 Sampling from an elementary DPP

Input: Marginal kernel $K = \sum_{k=1}^{m} v_k v_k^T$ for $\{v_k\}_{k=1,\dots,m}$ orthonormal

```
V ← {v_k}_{k=1,...,m}
Y ← ∅
while |V| > 0 do
    p_i ← P e_i the projection of e_i onto span(V) for i ∈ 𝒴
    Select i ∈ 𝒴 with probability 1/|V| · ‖p_i‖²
    Y ← Y ∪ {i}
    V ← V_⊥ an orthonormal basis of the subspace of V perpendicular to p_i
end while
return Y
```

2.18 Proposition (Sampling from elementary DPPs). *Let $K = \sum_{k=1}^{m} v_k v_k^T$, then Algorithm 1 produces a random variable $\mathbf{Y}$ with values in $\mathcal{Y}$ which is an elementary DPP with marginal kernel K.*

reread!

Proof. We note that we only have to check that (2.8) holds and for this we fix $A \subseteq \mathcal{Y}$. First we note that the output $\mathbf{Y}$ has cardinality $|m|$ since no element can be selected twice in the while loop and the size of V decreases by exactly one in each iteration. Hence it remains to show

$$\mathbb{P}(A = \mathbf{Y}) = \det(K_A)$$

if $|A| = m$. Let for the sake of convenience $A = \{1, \dots, m\}$ and $\mathcal{Y} = \{1, \dots, N\}$. Note that it suffices to show that the while loop selects $1, \dots, m$ in this exact order with probability $\frac{1}{m!}\det(K_A)$.

Let V_k denote the orthonormal set V in the k-th step of the while loop and let P_{k-1} be the projection onto $\operatorname{span}(V_k)$ and set $b_i := P_0 e_i$ for $i = 1, \dots, N$. We note that if $1, \dots, k-1$ were selected in the first steps, then P_{k-1} is exactly the projection to the subspace of $\operatorname{span}(V_{k-1})$ that is orthogonal to $b_1, \dots, b_{k-1}$. Since the spaces $\operatorname{span}(V_k)$ are decreasing we have $P_k P_j = P_k$ for $k \geq j$ and thus $P_{k-1} e_k = P_{k-1} P_0 e_k = P_{k-1} b_k$. Suppose now that we have selected $1, \dots, k-1$ in the first $k-1$ steps of the while loop. The probability to select k in the next iteration is

$$\frac{1}{|V_k|} \cdot \|P_{k-1} e_k\|^2 = \frac{1}{m-k} \cdot \|P_{k-1} b_k\|^2 .$$

Thus the probability to sample $1, \dots, m$ in this order is equal to

$$\frac{1}{m!} \cdot \|b_1\|^2 \cdot \dots \cdot \|P_{m-1} b_m\|^2 .$$

Since P_{k-1} is the projection onto the subspace orthogonal to $b_1, \dots, b_k$, the product is equal to the squared m-dimensional surface measure of the parallel epiped spanned by $b_1, \dots, b_m$. It is well known from measure and integration theory that the squared surface is given by the determinant of the Gram matrix

$$\det \begin{pmatrix} b_1^T b_1 & \cdots & b_1^T b_m \\ \vdots & \ddots & \vdots \\ b_m^T b_1 & \cdots & b_m^T b_m \end{pmatrix} = \det\big((B^T B)_A\big)$$

where $B \in \mathbb{R}^{N\times N}$ is the matrix which rows are equal to b_k. Therefore it remains to show $B^T B = K^V$. However by definition B is the projection onto the span of V and thus $B = K^V$. Because K^V is symmetric like every projection, we have $B^T = B$ and hence can conclude $B^T B = B^2 = B = K^V$ where we used that B is a projection. $\square$

We will use Theorem 2.15 to prove that the following algorithm samples from a DPP. This will also show the existence of DPPs to a given marginal kernel since it gives an explicit construction.

Algorithm 2 Sampling from a DPP

Input: Eigendecomposition $\{v_k, \lambda_k\}_{k=1,\dots,N}$ of K

```
J ← ∅
for k = 1, ..., N do
    J ← J ∪ {k} with probability λ_k
end for
V ← {v_k}_{k∈J}
Y ← ∅
while |V| > 0 do
    p_i ← P e_i the projection of e_i onto span(V) for i ∈ 𝒴
    Select i ∈ 𝒴 with probability 1/|V| · ‖p_i‖²
    Y ← Y ∪ {i}
    V ← V_⊥ an orthonormal basis of the subspace of V perpendicular to p_i
end while
return Y
```

2.19 Theorem (Sampling algorithm). *Let $K \in \mathbb{R}^{N\times N}$ be any symmetric and positive definite matrix such that $K \leq I$. Then the distribution of the output Y of Algorithm 2 is a DPP with marginal kernel K.*

Proof. Theorem 2.15 states that an arbitrary DPP is the mixture of elementary DPPs and the for loop in the algorithm represents exactly this mixing with the respective weights. Further the sampling result for elementary DPPs yields that the output of the second part of the algorithm, namely the while loop, is distributed according to a DPP with marginal kernel $K^V := \sum_{v\in V} vv^T$. $\square$

comment on the intuition one can get from this!

2.20 Corollary (Existence of DPPs). *Let K be a symmetric $N \times N$ matrix. Then K is the marginal kernel of a DPP if and only if $0 \leq K \leq I$.*

2.21 Corollary (Cardinality of DPPs). *Let $\mathbb{P}$ be a DPP with kernel*

$$K = \sum_{k=1}^{N} \lambda_k v_k v_k^T.$$

Then the cardinality of the DPP is distributed like the sum of the Bernoulli variables $\{B_k\}_{k=1,\dots,N}$ from theorem 2.15.

Proof. To proof this, we only have to convince ourselves that after the Bernoulli experiments the cardinality of a DPP with kernel (2.7) has size $m := \sum_{k=1}^N B_k$ almost surely. Since K_B has rank at most k, the cardinality is almost surely smaller than m. On the other hand we have

$$\mathbb{E}\big[|\mathbf{Y}|\big] = \sum_{i \in \mathcal{Y}} \mathbb{P}(i \in \mathbf{Y}) = \sum_{i \in \mathcal{Y}} (K_B)_{ii} = \operatorname{Tr}(K_B) = m. \tag{2.9}$$

In the last step we used that the trace of a symmetric matrix is the sum over its eigenvalues, which are B_k in our case. This computation lets us conclude $|\mathbf{Y}| = m$ almost surely. □

We close this section with the proof of 2.15 given in [Kulesza et al., 2012].

reread!

Proof of Theorem 2.15. Let $A \subseteq \mathcal{Y}, k := |A|$. Further set $W_n := (v_n v_n^T)_A$ and $W_J := \sum_{n \in J} W_n$. Then we have

$$\tilde{\mathbb{P}}(A \subseteq \mathbf{Y}) = \sum_{J \subseteq \mathcal{Y}} \det(W_J) \cdot \tilde{\mathbb{P}}\big(B_i = 1 \text{ for } i \in J\big).$$

Let $\big((W_{n_1})_1 (W_{n_2})_2 \dots (W_{n_k})_k\big)$ denote the $k \times k$ matrix with i-th row equal to the i-th row of W_{n_i}. Using the multilinearity of the determinant we obtain that the marginal probability above is equal to

$$\begin{aligned}
&\sum_{J \subseteq \mathcal{Y}} \sum_{n_1,\dots,n_k \in J} \det\big((W_{n_1})_1 (W_{n_2})_2 \dots (W_{n_k})_k\big) \cdot \tilde{\mathbb{P}}\big(B_i = 1 \text{ for } i \in J\big) \\
&= \sum_{n_1,\dots,n_k \in \mathcal{Y}} \det\big((W_{n_1})_1 (W_{n_2})_2 \dots (W_{n_k})_k\big) \sum_{J \supseteq \{n_1,\dots,n_k\}} \tilde{\mathbb{P}}\big(B_i = 1 \text{ for } i \in J\big) \\
&= \sum_{n_1,\dots,n_k \in \mathcal{Y}} \det\big((W_{n_1})_1 (W_{n_2})_2 \dots (W_{n_k})_k\big) \cdot \tilde{\mathbb{P}}\big(B_{n_i} = 1 \text{ for } i = 1,\dots,k\big) \\
&= \sum_{n_1,\dots,n_k \in \mathcal{Y}} \det\big((\lambda_{n_1} W_{n_1})_1 (\lambda_{n_2} W_{n_2})_2 \dots (\lambda_{n_k} W_{n_k})_k\big) \\
&= \det\Big(\sum_{n \in \mathcal{Y}} W_n\Big) = \det(K_A).
\end{aligned}$$

This computation shows that $\tilde{\mathbb{P}}$ is a DPP with marginal kernel K. □

Possible improvements

2.22 DUAL SAMPLING.

2.23 DIMENSION REDUCTION.

II.4 The mode problem

One general motivation for modelling is the hope that predictions can be made from the selected model. If the model is of stochastic nature, like in our case, and if one wants to predict its outcome, there are a few possible approaches. The first one would be to sample from this model. This relies on the intuition that a realisation of our random variable will be a rather typical example for the random event. Going one step further one could try to find the most likely outcome of the random variable, which is known as the mode problem.

2.24 The mode problem. Let X be a random variable with values in some space $\mathcal{X}$ and let f be the density of the distribution of X with respect to some reference measure. Then the *mode* is the maximiser

$$\hat{x} = \arg\max_{x \in \mathcal{X}} f(x)$$

of the density if it exists. The search for the mode is called the *mode problem*.

Our motivation for finding the mode of a random variable was to make better predictions for it. This is justified by the assumption that the mode should be a typical realisation of the random variable. However this is not generally the case and therefore one should be cautious with this intuition. Consider for example the mixture of two independent Gaussian random variables

look for better word

$$0.1 \cdot X + 0.9 \cdot Y$$

where X is centered with variance 10 and Y has mean 5 with variance 1, the densities are shown in Figure .... It is clear that mode is 0 in this example, but it is not a very typical outcome of the random variable, since the majority of events is centered around 10.

check whether this gives the desired effect and plot the density!

The mode problem is rather well behaved if the density f is a smooth function defined on a subset of $\mathbb{R}^d$, but in the case of DPPs we have to deal with the probability measure on a finite set. Thus this turns into a discrete optimisation problem over the exponentially large powerset $2^{\mathcal{Y}}$. This is in general very hard to solve and it has been shown in that it is NP hard to do so or even approximate it upto a factor of $\frac{8}{9}$. However there were still different strategies proposed and we will present some of them including their main ideas.

cite

do this!

Chapter III

Point estimators and parametric models

Parameter estimation is one of the central components of every theory of real world phenomena. In a nutshell one could split the process of the construction of a descriptive model into two parts. The first one being the selection of the model which is done by a scientist and the second being the determination of the constants that belong to the model.

To make this more clear we will consider one of the most famous advances in the natural sciences namely the law of universal gravitation that was discovered by Sir Isaac Newton and published in one of the most famous books in the history of science, the *Philosophiæ Naturalis Principia Mathematica*. More precisely Newton discovered that the gravitational force acting between two massive objects is given by

$$F = G \cdot \frac{m_1 m_2}{r^2}$$

where m_1, m_2 are the masses of the two objects, r is the distance of the centers of masses and G is the gravitational constant. This constant can not be deduced from the theory itself and needs to be estimated based on some empirical data.

If we want to describe, simulate and predict the occurence of diverse subsets we can take a similar approach and impose the model of a determinantal point process. This will usually be an assumption that will not strictly hold, but will often lead to reasonable, sometimes even impressive results. We will not be concerned to measure how suitable this model selection is, although this is a highly interesting question. Leaving that aside we are left with the second step, namely the estimation of the parameters of the model, which are in the case of a DPP over a set of cardinality N exactly $N(N-1)/2$. Because of the rather large amount of parameters and also the complicated structure of the DPPs it will in practice only be possible to perform those estimations through the use of computational tools. The task of computer based parameter or density estimation is an important field in the discipline of *machine learning* and thus we will sometimes speak of the parameters being learned instead of estimated. Actually the interest of parameter estimation for DPPs arose from the machine learning community at the beginning of this decade. However we will phrase things in a way that no prior knowledge in this field is required.

In this chapter we will be concerned in how we can make point estimates for either the marginal or the elementary kernels K and L. Point estimators are the most basic type of estimators and consist of the suggestion of one possible parameter set, for example in the case of the gravitational constant

$$6.674 \cdot 10^{-11} \mathrm{N\,kg^{-2}\,m^2}.$$

This is in contrast to the Bayesian approach to parameter estimation that we will present in the next chapter where the philosophy is to estimate a distribution over all possible parameter sets that indicates how likely they are given some the empirical data. We will discuss two essentially different methods of point estimators, the first one provides a way to reconstruct a marginal kernel for the empirical marginal distributions at least in the case where the empirical distribution is essentially a DPP. The other type of methods are all maximum likelihood estimators in different variations.

But before we can proceed we want to remind the reader of two desirable properties of point estimators. For this we will assume that we want to estimate the distribution of a random variable X from a parametric family of probability measures

$$\left\{\mathbb{P}_\theta \mid \theta \in \Theta\right\}.$$

This means we want to estimate θ out of a possible set of parameters Θ such that X is distributed according to $\mathbb{P}_\theta$ which we will based upon some data $x_1, \dots, x_n$. Further we assume that those points are actually generated by $\mathbb{P}_\theta$ for one $\theta \in \Theta$ and denote the estimator by $\hat{\theta}_n$. We call *unbiased* if we have

$$\mathbb{E}[\hat{\theta}_n] = \theta$$

and *consistent* if we have

$$\hat{\theta}_n \to \theta \quad \text{in probability}.$$

It shall be noted that although those properties are beneficial, they are not crucial for an estimator to be reasonable. First they both assume that the data generating process, i.e. the process one wants to describe actually follows one of the laws $\mathbb{P}_\theta$ which will typically be not the case in real world examples. Further the asymptotic property of consistency is rather of theoretical nature since in practice it is not possible to create large sets of empirical data and certainly not infinitely large ones.

III.1 Kernel reconstruction from the empirical measures

Now we will display the first way how one can estimate the marginal kernel K of a DPP based on some samples drawn from it.

3.1 SETTING. Let $\mathcal{Y}$ be a finite set of cardinality N and let $K \in \mathbb{R}^{N\times N}_{\text{sym}}$ satisfy $0 \leq K \leq I$. Let further $\mathbf{Y}_1, \mathbf{Y}_2, \dots$ be distributed according to the DPP with marginal kernel K.

In order to perform an approximate reconstruction of the marginal kernel we will need to consider the *empirical measure*

$$\hat{\mathbb{P}}_n := \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{Y}_i}.$$

The interest in $\hat{\mathbb{P}}_n$ lies in the fact that they quite natural estimates for the actual underlying distribution. More precisely they are unbiased estimators for $\mathbb{P}$, i.e. they agree in expectation with $\mathbb{P}$. This can be seen by evaluating it at $A \subseteq \mathcal{Y}$

$$\mathbb{E}_{\mathbb{P}}\big[\hat{\mathbb{P}}_n(A)\big] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbb{P}}\big[\delta_{\mathbf{Y}_i}(A)\big] = \mathbb{P}(A).$$

And even stronger by the strong law of large numbers they converge to $\mathbb{P}$ almost surely if the sequence $(\mathbf{Y}_k)_{k\in\mathbb{N}}$ of observations is independent. This can be seen by identifying the probability measures on $2^{\mathcal{Y}}$ with the probability simplex

$$\left\{\mu \in \mathbb{R}^{2^{\mathcal{Y}}} \;\middle|\; \mu_A \in [0,1] \text{ for all } A \subseteq \mathcal{Y} \text{ and } \sum_{A\subseteq\mathcal{Y}} \mu_A = 1\right\}$$

and using the strong law of large numbers in $\mathbb{R}^{2^{\mathcal{Y}}}$.

Therefore the empirical measures are reasonable approximations of the actual probability distribution. Assume now for one moment that the empirical measures $\hat{\mathbb{P}}_n$ are also determinantal point processes with marginal kernel $\hat{K}_n$, then $\hat{K}_n$ would be a quite intuitive estimate for the actual marginal kernel K. Thus we are interested in the question whether we can reconstruct the kernel marginal of a DPP if we know the DPP itself. Since the marginal density of a DPP corresponds to the principal minors of the marginal kernel, we first investigate whether we can reconstruct a matrix from its principal minors. For the answer to this problem we follow the main ideas presented in [Urschel et al., 2017] and [Rising et al., 2015] although we modify their arguments to make them shorter and hopefully more accessible.

3.2 The principal minor assignment problem. Let $K \in \mathbb{R}^{N\times N}$ be a symmetric matrix. We want to investigate whether K uniquely specified by its principal minors

$$\Delta_S := \det(K_S) \quad \text{where } S \subseteq \{1,\dots,N\}\,.$$

We call this the *symmetric principal minor assignment problem* and it will turn out that the matrix K can be reconstructed up to an equivalence relation.

Before we present the general procedure we want to see how this would work in the case of a symmetric 3×3 matrix $K = (K_{ij})_{1\le i,j\le 3}$. First we note that we can regain the diagonal elements as the determinant of the 1×1 principal minors

$$\det(K_{\{i\}}) = K_{ii} \quad \text{for } i = 1,2,3.$$

Further the squares of the off diagonal are determined by the 2×2 principal minors since

$$\det(K_{\{i,j\}}) = K_{ii}K_{jj} - K_{ij}^2 \quad \text{for } i,j = 1,2,3.$$

Therefore we only need to reconstruct the signs off diagonal entries. To do this, we consider the determinant of the matrix itself

$$\det(K) = K_{11}K_{22}K_{33} + 2K_{12}K_{13}K_{23} - K_{11}K_{23}^2 - K_{22}K_{13}^2 - K_{33}K_{12}^2. \tag{3.1}$$

Rearranging this yields

$$K_{12}K_{13}K_{23} = \frac{1}{2}\left(\det(K) + K_{11}K_{23}^2 + K_{22}K_{13}^2 + K_{33}K_{12}^2 - K_{11}K_{22}K_{33}\right).$$

Since we know all of the expressions on the right side, we can determine the sign of the product on the left side. Now we assign the signs of the off diagonal elements in such a way, that the above equation holds. More precisely if the product is negative, we assign a minus to one or all three elements, if it is positive, then we assign a minus to none or two elements. If the product is zero, every configuration of signs satisfy the desired property. It is now straight forward to check that this assignment actually leads to the desired principal minors.

III.1.1 Graph theoretical concepts

One main part in the general procedure will be to obtain a generalisation of the formula (3.1) for larger principal minors that will allow the reconstruction of the signs. For this we will need the following graph theoretical concepts.

3.3 Notions from graph theory. Let $G = (V, E)$ be a finite graph, i.e. V is a finite set, called the *vertex set* and E consists of subsets of V with two elements, the *edges*. Sometimes we will be sloppy in notation and not distinguish between the graph and the edge set. We will need the following notions:

(*i*) *Degree:* For a vertex $v \in V$ the *degree* is the number of edges that contains v.

(*ii*) *Subgraph:* A graph $\tilde{G} = (\tilde{V}, \tilde{E})$ is called a *subgraph* of G if $\tilde{V} \subseteq V$ and $\tilde{E} \subseteq E$.

(*iii*) *Induced graph:* For a subset $S \subseteq V$ of vertices the *induced graph* $G(S) = (S, E(S))$ is formed of all edges $e \in E$ of G that are subsets of S.

(*iv*) *Path:* A *path* in G is a sequence $v_0 v_1 \cdots v_k$ of vertices such that $\{v_{i-1}, v_i\} \in E$ for all $i = 1, \ldots, k$.

(*v*) *Connected graph:* A graph is called *connected* if for every pair of vertices $v, w \in V$ there is a path from v to w.

(*vi*) *Cycle:* A *cycle* C is a connected subgraph such that every vertex has even degree in C.

(*vii*) *Cycle space:* Each cycle C can be identified with a vector $x = x(C) \in \mathbb{F}_2^E$ such that

$$x_e := \begin{cases} 1 & \text{if } e \in C \\ 0 & \text{if } e \notin C \end{cases}$$

indicates whether the edge $e \in E$ belongs to the cycle C. The *cycle space* $\mathcal{C}$ is the span of $\{x(C) \mid C \text{ is a cycle}\}$ in $\mathbb{F}_2^E$. Note that the sum of two cycles in the cycle space corresponds to the symmetric difference of the edges.

(*viii*) *Chordless cycle:* A cycle C is called *chordless* if two verteces $v, w \in C$ form an edge in G if and only they form an edge in C. This is equivalent to the statement that C is an induced subgraph that is a cycle.

(*ix*) *Cycle sparsity:* The cycle sparsity is the minimal number l such that a basis of the cycle space consisting of chordless simple cycles exists. Such a basis is called *shortest maximal cycle basis* or short *SMCB*. If the cycle space is trivial we define the cycle sparsity to be 2.

(*x*) *Pairings:* Let $S \subseteq V$ be a set of of vertices. Then a *pairing* P of S is a subset of edges of $G(S)$ such that two different edges of P are disjoint. The vertices contained in the edges of P are denoted by $V(P)$ and the set of all pairings by $\mathcal{P}(S)$.

It is highly recommended to study the examples in Figure III.1.1 in order to get more familiar with the definitions above. To see that the above definition of the cycle sparsity is well defined, we have need to show that shortest maximal cycle basis exist. This might be well known to people that are familiar with graph theory, but we will present an elementary proof here. The first part of the statement, namely the existence of cycle basis consisting of simple cycles is known as Veblen's theorem and can be found in its original form in [Veblen, 1912], however we will rather follow the approach in [Bondy and Murty, 2011].

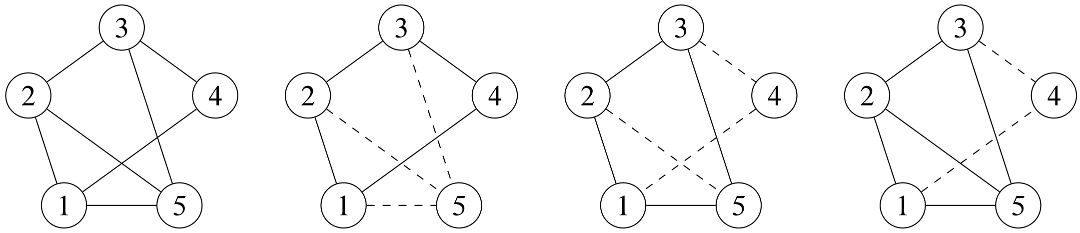

Figure III.1.: Some examples of graphs and cycles. The first sketch shows a graph and the three other ones subsgraphs of it where the edges not belonging to the subgraph are depicted dashed. The first one is a symple chordless cycle, the second one a simple but not chordless cycle and the last one is not a cycle at all.

3.4 Proposition (Existence of SMCBs). *There always exists a basis $\{x(C_1), \dots, x(C_k)\}$ of the cycle space where $C_1, \dots, C_k$ are chordless simple cycles.*

Proof. First we prove that the set of simple cycles generates the whole cycle space which we can then improve to show that the simple chordless cycles already generate the cycle space. A shortest maximal cycle basis is then attained by successively dropping simple chordless cycles.

We show that every cycle $x(C)$ can be written as the sum of simple cycles $x(C_1), \dots, x(C_k)$ where $C_i \subseteq C$. This is equivalent to the statement that the edges of every cycle are the disjoint union of the edges of simple cycles. Take now a maximal non intersecting path $v_0 v_1 \cdots v_k$. Since v_k has degree at least 2, there is an edge $\{v_k, v_{k+1}\}$ such that $v_{k+1} \neq v_{k-1}$. Since the

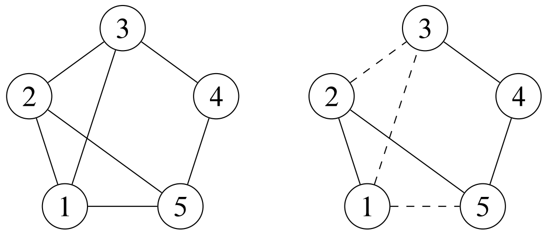

Figure III.2.: Illustration of the search for a simple cycle in a graph with degrees greater than two. Once a maximal non intersecting path like 12543 is selected, every continuation of the path – in this case 2 or 1 – is already present in the path and therefore induces a simple cycle.

path is maximal, v_{k+1} has to agree with one a vertex $v_i \in \{v_0, \dots, v_{k-2}\}$, because otherwise we could add v_{k+1} to the path which is a contradiction to the maximality. Now $v_i v_{i+1} \cdots v_k v_i$ corresponds to a simple cycle C_1 and $C_2 := C \setminus C_1$ is again a cycle. Thus we can write C as the disjoint union $C = C_1 \cup C_2$ where C_1 is a simple cycle. By repeating this procedure we get the desired expression for C in terms of simple cycles.

To prove that already the simple chordless cycles generate the cycle space we have to prove that we can write every simple cycle $x(C)$ as a sum of simple chordless cycles $x(C_1), \dots, x(C_k)$. Let $\big\{\{v_0, v_1\}, \dots, \{v_k, v_0\}\big\}$ be the edge set of C and assume that C is not chordless like in Figure III.1.1, otherwise the statement would be trivial. Thus there is are indices $1 \leq i < j-1 \leq k-1$ such that $\{v_i, v_j\} \in E$. Let now C_1 and C_2 be the two cycles associated with the paths

$$v_0 v_1 \cdots v_i v_j v_{j+1} \cdots v_k v_0 \quad \text{and } v_i v_{i+1} \cdots v_{j-1} v_j v_i .$$

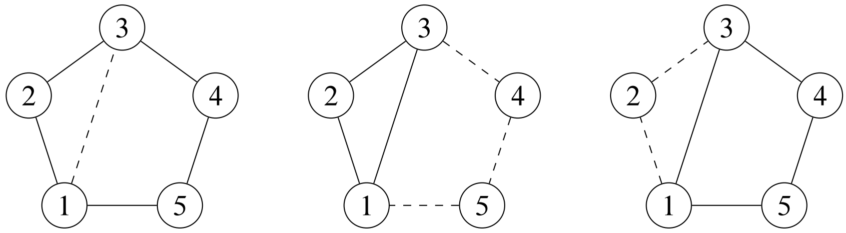

Figure III.3.: The simple cycle 123451 on the left is not chordless but the symmetric difference of the two simple chordless cycles 1231 and 13451 on the right.

Then we have $x(C) = x(C_1) + x(C_2)$. By iterating this procedure as long as the cycles are not chordless the desired decomposition can be achieved in finitely many steps.

□

III.1.2 The solution of the principal minor assignment problem

Now we have all the graph theoretical prerequisites to show how one can reconstruct a matrix with preassigned principal minors. However the matrix that arises from this reconstruction is not unique and thus we need to identify matrices with the same principal minors with each other.

3.5 Definition (Determinantal equivalence). Two symmetric matrices $A, B \in \mathbb{R}^{N\times N}$ are called *determinantally equivalent* if the have the same principal minors and we write $A \sim B$.

It is obvious that we can only hope to reconstruct a symmetric matrix up to determinantal equivalence. However this would be satisfactory, because determinantally equivalent matrices are exactly those that give rise to the same DPP. Let us in the following denote the principal minor $\det(K_S)$ by Δ_S for $S \subseteq \{1, \dots, N\}$. To come back to our original problem, we notice that the principal minors up to size two immediately determine the diagonal and the absolute values of the off diagonal of K since we have

$$K_{ii} = \Delta_{\{i\}} \quad \text{and } K_{ij}^2 = K_{ii}K_{jj} - \Delta_{\{i,j\}}.$$

Thus it only remains to regain the signs $\operatorname{sgn}(K_{ij})$ of the off diagonal entries. For this we use the following object.

3.6 The adjacency graph and sign function. The adjacency graph $G_K = (V_K, E_K)$ associated with K consists of the vertex set $\{1, \dots, N\}$ and $\{i, j\}$ form an edge if and only if $K_{ij} \neq 0$. Further we introduce some *weights* on the edges. This means we consider a mapping $w\colon E_K \to \mathbb{R}$ and we set

$$w_{ij} := w(\{i, j\}) := \operatorname{sgn}(K_{ij})$$

where we call w_{ij} the weight of the edge $\{i, j\}$. This graph together with the weights determines the signs of the off diagonal elements, so we are interested in reconstruction the weights from the principal minors. Finally we define the sign of a cycle and for a cycle $C = (S, \tilde{E})$ we set $\operatorname{sgn}(C) := \prod_{e\in\tilde{E}} w_e$. It will become important later to consider this sign function on the cycle space and thus we note that this definition corresponds to

$$\operatorname{sgn}(x(C)) := \prod_{e\in E} w_e^{x(C)_e}.$$

Note that this is a group homomorphism from the cycle space $\mathcal{C}$ to $\{\pm 1\}$ and therefore it is uniquely determined by its value on a generator, for example on a shortest maximal cycle basis.

3.7 Proposition (Principal minors of simple chordless cycles). *Let $C = (S, E(S))$ be a simple and chordless cycle. Then the principal minor of K with respect to S is given by*

$$\Delta_S = \sum_{P \in \mathcal{P}(S)} (-1)^{|P|} \cdot \prod_{\{i,j\} \in P} K_{ij}^2 \cdot \prod_{i \notin V(P)} K_{ii} + 2 \cdot (-1)^{|S|+1} \cdot \prod_{\{i,j\} \in E(S)} K_{ij}. \tag{3.2}$$

Proof. Let $k := |S|$. Then by Leibniz formula we have

$$\Delta_S = \sum_{\sigma \in S_k} \operatorname{sgn}(\sigma) \prod_{i \in S} K_{i\sigma(i)}$$

where S_k is the set of permutations of S. Note that since the cycle is chordless, the product is only non trivial if $\{i, \sigma(i)\} \in E(S)$ for all $i \in S$. Since C is a simple cycle, those permutations consist exactly of the pairing of S or the two shifts of the set S along the cycle in both directions. Those correspond exactly to the summands in (3.2).

To see this, we fix a permutation σ such that $\{i, \sigma(i)\}$ always forms an edge in $(S, E(S))$. We note that every vertex $i \in S$ has two possible images which are exactly the endpoint of its two edges, c.f. Figure III.1.2. Lets assume it is mapped to $j \in S$, then j has again two possible

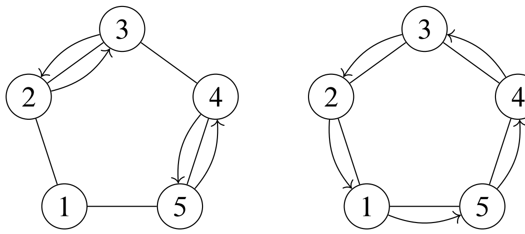

Figure III.4.: An easy example for the two kinds of permutations of a chordless simple cycle that maps vertices to neighbors.

images under σ namely i and a second vertex $k \in \mathcal{Y}$. If $j \mapsto i$, no other vertex can be mapped to i or j, however some other items can be swapped in the same way. The permutations of this form correspond exactly to the pairings of S and are represented in the first sum in (3.2). If however j is not mapped back to i but rather to its other neighbor k, then k can't get mapped back to j since σ is injective. Thus it has to be mapped to its other neighbor $l \in \mathcal{Y}$. Through a repetition of this argument shows that this induces until i is reached again. Since the cycle is simple this path exhausts the entire cycle. The factor 2 is due to the fact that this shift of the indices can be done into either direction. □

3.8 Proposition (Sign determines principals minors). *The knowledge of all principal minors up to size two and the sign function*

$$\operatorname{sgn} \colon \mathcal{C} \to \{\pm 1\}$$

completely determines all principal minors of K.

Proof. Let $S \subseteq \{1, \dots, N\}$ be arbitrary. We will again work with the expression (3.2) of the principal minor Δ_S and fix one permutation σ. We can assume without loss of generality that

$\{i, \sigma(i)\} \in E_K$ because the product it trivial otherwise. Since we know the absolute values of the off diagonal elements and the diagonal elements from the principal minors up to size two, it suffices to express the sign

$$\prod_{i \in S} \operatorname{sgn}(K_{i\sigma(i)}) \tag{3.3}$$

of the product through the sign function. For this we write σ as the product of disjoint cycles

$$\sigma = \sigma_1 \circ \cdots \circ \sigma_m \tag{3.4}$$

where $\sigma_k : D_k \to D_k$ for $k = 1, \ldots, m$ and the domains D_k are pairwise disjoint. The sign (3.3) can be written as the product of

$$\prod_{i \in D_k} \operatorname{sgn}(K_{i\sigma_k(i)})$$

so it suffices to give expressions for those. Note that we could assume $\{i, \sigma_k(i)\} \in E_K$ and therefore $C_k = (D_k, E_k)$ with

$$E_k = \big\{\{i, \sigma_k(i)\} \mid i \in D_k\big\}$$

is a cycle and therefore (3.4) is equal to $\operatorname{sgn}(C_k)$. □

3.9 THEOREM. *Let $K \in \mathbb{R}^{N \times N}$ be a symmetric matrix and l be the sparsity of its adjacency graph. Then the principal minors up to size l uniquely determine all principal minors of K and therefore the matrix K up to determinantal equivalence.*

Proof. In the light of the previous proposition it suffices to show that the sign function is uniquely specified by the principal minors up to size l. Recall that the sign function is determined by its values on a shortest maximal cycle basis, which consists by definition of simple chordless cycles of length at most l. However under the knowledge of the diagonal elements and the absolute values of the off diagonal ones, the sign of those simple chordless cycle is uniquely determined by the principal minors up to size l using the euqality (3.2) . □

3.10 REMARK. One can even show that this result is optimal in the sense that if one only has access to the principal minors up to size $l - 1$, then the equivalence class is not uniquely determined. To see this, we note that the sign function is not uniquely specified through the principal minors up to size $l - 1$ and thus there is more than one extension of the sign function onto the shortest maximal cycle basis. The equation (3.2) shows that those different extensions give rise to different principal minors.

3.11 CONSTRUCTION OF THE EQUIVALENCE CLASS. We have shown that the determinantal equivalence class of a symmetric matrix is uniquely specified by its principal minors up to size l. Now we want to investigate how this equivalence class can be computed and we will see that we can reduce this task to the solution of a system of linear equations over the finite field $\mathbb{F}_2$.

Let us assume that we have knowledge of the principal minors Δ_S for every $S \subseteq \{1, \ldots, N\}$ with size at most l and we want to construct a matrix $\tilde{K}$ that is determinantally equivalent to K. We have seen that we only need to reconstruct the signs of the off diagonal entries of K which is equivalent to reconstructing the edge weight w_{ij}. To do this fix a shortest maximal cycle basis $\{C_1, \ldots, C_m\}$ with vertex sets $S_1, \ldots, S_m$. Let us now rewrite (3.2) in the form

$$H_k := \Delta_{C_k} - \sum_{P \in \mathcal{P}(C_k)} (-1)^{|P|} \cdot \prod_{\{i,j\} \in P} K_{ij}^2 \cdot \prod_{i \notin V(P)} K_{ii} = 2 \cdot (-1)^{|C_k|+1} \operatorname{sgn}(C_k) \cdot \prod_{\{i,j\} \in C_k} \left|K_{ij}\right|.$$

Given the principal minors, we can determine the value on the right side and taking the sign on both sides yields

$$(-1)^{|C_k|+1} \cdot \operatorname{sgn}(H_k) = \operatorname{sgn}(C_k) = \prod_{\{i,j\} \in E(S_k)} w_{ij}$$

which we seek to solve for w. However this multiplicative equation is hard to solve and therefore we use the canonical group isomorphism ϕ between $\{\pm 1\}$ and $\{0, 1\}$ to turn it into a linear equation. Setting $x_{ij} := \phi(w_{ij})$ we get that the condition above is equivalent to

$$b_k := \phi(\operatorname{sgn}(H_k)) + |\hat{S}_k| + 1 = \sum_{\{i,j\} \in E(S_k)} x_{ij} = (Ax)_k \quad \text{in } \mathbb{F}_2$$

where A is the matrix with the rows $x(C_k)^T$. Now we can fix any such solution $x \in \mathbb{F}_2^E$ of

$$Ax = b \tag{3.5}$$

and we know that at least one exists, namely the one given by $x_{ij} = \phi(\operatorname{sgn}(K_{ij}))$. Let now $w_{ij} := x_{ij}$, then it is straight forward to see that $\tilde{K}$ defined through

$$\tilde{K}_{ii} := \Delta_{\{i\}} \quad \text{and } \tilde{K}_{ij} = w_{ij} \cdot \sqrt{\tilde{K}_{ii} \tilde{K}_{jj} - \Delta_{\{i,j\}}}$$

is determinantally equivalent to K.

It shall be noted that there are algorithms with much better computational performance for the construction of the determinantal equivalence class. For some examples of efficient algorithms we refer to [Urschel et al., 2017] and [Rising et al., 2015].

III.1.3 Definition of the estimator and consistency

So far we have seen that the principal minors determine a symmetric matrix up to determinantal equivalence. However the empirical marginal densities do not in general need to be the principal minors of any symmetric matrix, in other words the empirical measures are not necessarily determinantal. Therefore the definition of the estimator is till not quite straight forward and we will follow [Urschel et al., 2017] for this and make the following assumption.

3.12 Assumption. Let $\alpha > 0$ and assume that

$$\min \left\{ \left| K_{ij} \right| \mid K_{ij} \neq 0 \right\} \geq \alpha.$$

Note that such an α can always be found, however it is not a priori known. For example if we want to make a statement about the speed of approximation of the estimators , which depends on α, we have to make the assumption above.

3.13 Definition of the estimator. The straight forward estimators of the principal minors are

$$\hat{\Delta}_S := \hat{\mathbb{P}}_n(S \subseteq \mathbf{Y}) \quad \text{for } S \subseteq \{1, \ldots, N\}.$$

The resulting estimates for the diagnoal elements and the squares of the off diagonals are

$$\hat{K}_{ii} := \hat{\Delta}_{\{i\}} \quad \text{and } \hat{B}_{ij} := \hat{K}_{ii} \hat{K}_{jj} - \hat{\Delta}_{\{i,j\}}.$$

Next we will introduce an estimate $\hat{G}$ for the adjacency graph and will then try to choose the signs of the estimated matrix $\hat{K}$ such that the its principal minors are the estimates for the principal minors. For this define the edge set $\hat{E}$ of $\hat{G}$ to consist of all sets $\{i, j\}$ such that $\hat{B}_{ij} \geq \frac{1}{2}\alpha^2$. This truncation yields the desired effect that by the strong law of large numbers the estimator for the graph will converge to the actual adjacency graph almost surely. In analogy to the previous paragraph we define $\{\hat{C}_1, \dots, \hat{C}_{\hat{m}}\}, \hat{H}_1, \dots, \hat{H}_{\hat{m}}, \hat{A}$ and $\hat{b}$ exactly the same way. If there is a solution $\hat{x} \in \mathbb{F}_2^E$ to the linear equation

$$\hat{A}\hat{x} = \hat{b}, \tag{3.6}$$

then we estimate the signs to be $\hat{w}_{ij} := \phi^{-1}(\hat{x}_{ij})$ and define

$$\hat{K}_{ij} := \hat{w}_{ij}\sqrt{\hat{B}_{ij}}.$$

If there is no such solution $\hat{x}$ then we simply set the signs of the off diagonal elements to be positive, i.e. we define

$$\hat{K}_{ij} := \sqrt{\hat{B}_{ij}}.$$

This choice is completely arbitrary, but we will see in the consistency result 3.15 that the probability for this case tends to zero as the sample size increases. In fact we will see that the two linear equations (3.5) and 3.6 agree with increasing probability.

In order to talk about consistency of the estimator that we constructed above, it is necessary to define a metric on the marginal kernels of DPPs. However the usual operator norm is clearly not right for this job, since we already know that we can only hope to reconstruct the determinantal equivalence class but not the exact marginal kernel. Thus we will work with the usual choice of pseudometric if one has to deal with equivalence classes.

3.14 PSEUDOMETRIC ON THE MARGINAL KERNELS. We define the distance between two marginal kernels $A, B \in \mathbb{R}^{N\times N}$ through

$$d(A, B) := \inf_{C\sim A} \|B - C\|_\infty$$

where $\|A\|_\infty := \max_{1\leq i,j\leq N} |A_{ij}|$ denotes the uniform norm on the space of matrices.

3.15 THEOREM (CONSISTENCY). *Let K be the marginal kernel of a DPP that satisfy the assumption 3.12. Let further l be the cycle sparsity of G_K and $\varepsilon > 0$.*

$$\mathbb{P}\left(d(\hat{K}, K) \leq \varepsilon\right) \to 1 \quad \textit{for } n \to \infty.$$

Proof. We will keep the notations from the paragraphs 3.11 and 3.13. We have already seen in the motivation of this section that the empirical measures converge almost surely which directly implies

$$\hat{K}_{ii} \to K_{ii} \quad \text{and } \hat{K}_{ij}^2 \to K_{ij}^2 \quad \text{almost surely.} \tag{3.7}$$

Note that almost surely convergence implies convergence in probability and thus we have

$$\mathbb{P}(\hat{G} = G_K) = \mathbb{P}\left(\hat{K}_{ij}^2 \geq \alpha^2/2 \text{ for } K_{ij} \neq 0\right) \to 1 \quad \text{for } n \to \infty.$$

In this case the two shortest cycle basis can be chosen the same and so $\hat{A}$ and A agree. Because of (3.7) we also have $\hat{H}_k \to H_k$ almost surely and thus $\hat{b}_k \to b_k$ almost surely for all k. This yields

$$\mathbb{P}\left(\hat{A} = A \text{ and } \hat{b} = b\right) \to 1 \quad \text{for } n \to \infty. \tag{3.8}$$

In this case the two linear quations (3.5) and (3.6) agree, then $\tilde{K} \in \mathbb{R}^{N\times N}$ defined through $\tilde{K}_{ij} := \hat{w}_{ij} \left|K_{ij}\right|$ is determinantally equivalent to K. Further we know that for any $\delta > 0$

$$\mathbb{P}\left(\left|\hat{K}_{ij}^2 - K_{ij}^2\right| < \delta \text{ for all } i,j\right) \to 1 \quad \text{for } n \to \infty. \tag{3.9}$$

If this is true, then we have

$$d(\hat{K}, K) \le \|\hat{K} - \tilde{K}\|_\infty = \sup_{i,j} \left||\hat{K}_{ij}| - |\tilde{K}_{ij}|\right|$$

where we used, that the entries of $\hat{K}$ and $\tilde{K}$ have equal signs. Further we have

$$\left||\hat{K}_{ij}| - |\tilde{K}_{ij}|\right| = \frac{|\hat{K}_{ij}^2 - \tilde{K}_{ij}^2|}{|\hat{K}_{ij}| + |\tilde{K}_{ij}|} < \frac{\delta}{\alpha} \le \varepsilon$$

if $\delta \le \alpha\varepsilon$. In conclusion we have seen that if

$$\hat{A} = A, \quad \hat{b} = b \quad \text{and} \quad \left|\hat{K}_{ij}^2 - K_{ij}^2\right| < \delta \quad \text{for all } i,j$$

then we have

$$d(\hat{K}, K) < \varepsilon.$$

However (3.8) and (3.9) shows that the probability for this tends to one. □

3.16 Remark (Speed of convergence). Although the result above states that the estimators $\hat{K}$ converges to K in probability, it does give no information about the speed of convergence. This problem is addressed in [Urschel et al., 2017], but it turns out that the convergence is very slow. For example for the very moderate case $\alpha = 0.4$ and $l = 3$ one already needs more than 10^6 samples to get some theoretical guarantees from their result. This is not due to careless estimates since they even show that this bound is optimal. However since this result is beyond practical relevance, we will keep away from those calculations.

is this true?

explain how the algorithm for the reconstruction works

III.1.4 Computation of the estimator

is this estimator unbiased?

III.2 Maximum likelihood estimation using optimisation techniques

rewrite introduction to MLE

The method of maximum likelihood estimation is a very well established procedure to estimate parameters. The philosophy of MLE is that one selects the parameter under which the given data would be the most likely to be observed and to motivate this in more detail we roughly follow the corresponding section in [Rice, 2006].

Suppose that we want to estimate a parameter $\theta \in \Theta$ based on some realisations $x_1, \dots, x_n$ of some random variables $X_1, \dots, X_n$. We have given some candidates $f(x_1, \dots, x_n|\theta)$ for the joint density of $X_1, \dots, X_n$ with respect to some reference measure $\prod_{i=1}^n \mu(\mathrm{d}x_i)$ and we want to decide which parameter $\theta \in \Theta$ describes the realisations, which we will also call data or observation, best. Hence it is reasonable to pick that θ under which the observations $x_1, \dots, x_n$

are the most likely. In other words we want to find the parameter θ that maximises the density $f(x_1, \dots, x_n|\theta)$. If additionally the random variables are indepent and identically distributed, their joint density factorises and thus we obtain

$$f(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

where $f(x|\theta)$ is the density with respect to μ of the X_i. In practice it is often easier to maximise the logarithm of the density

$$\mathcal{L}(\theta) = \log\big(f(x_1, \dots, x_n|\theta)\big) = \sum_{i=1}^{n} \log\big(f(x_i|\theta)\big)$$

since this transforms the product over functions into a sum. However this is clearly equivalent to maximising the density since the logarithm is strictly monotone.

3.17 Definition of the MLE. Let Θ be a set, which we call the *parameter set* and let

$$\mathcal{F} = \Big\{ f(\cdot|\theta) \colon X \to [0, \infty) \;\Big|\; \theta \in \Theta \Big\}$$

be a family of probability densities with respect to some measure μ on some measurable space X. We call the function

$$\mathcal{L} \colon \Theta \to [-\infty, 0]$$

the *log likelihood function* and its maximiser

$$\hat{\theta}_n := \arg\max_{\theta \in \Theta} \mathcal{L}(\theta) \tag{3.10}$$

the *maximum likelihood estimator* or short MLE.

Very short reminder on optimisation

Since the calculation of the MLE is a maximisation task, it is suitable to review some general properties of optimisation problems. It shall be noted that optimisation problems are usually stated as minimisation tasks, but we will stick to the maximisation, which is clearly equivalent up to a sign. For this let $U \subseteq \mathbb{R}^M$ and $f : U \to \mathbb{R}$ be a function. In practice the maximisation

$$\hat{x} := \arg\max_{x \in U} f(x)$$

will not be explicitly solvable and therefore one usually has to exploit numerical algorithms.

Those work particularly well if the function f is concave and possibly smooth and one powerful method is the given by the so called gradient descent. To quickly explain the philosophy of those methods, we note that ∇f points into the direction of the steepest ascent of the function f and thus an intuitive approach the maximise f would be to follow the gradient, i.e. to take a solution γ of the gradient flow $\gamma' = \nabla f(\gamma)$ and work out its limit. However if the function is not concave one can not even guarantee that the gradient flow reaches a local minimum, since one can construct examples where γ gets stuck in a critical point. However in the concave case this suffices since critical points and global minima agree for convex functions. The gradient descent is an algorithm derived from this observation and is essentially a discretisations of the gradient flow meaning that it iteratively takes small steps into the direction of the gradient and thus lowers the value of the function. Some more sophisticated versions of gradient descent methods usually even consider higher order derivatives and use the information they provide over the geometry of the graph. Generally speaking those algorithms work extremely well even in high dimensions and thus their efficiency and stability have been studied broadly and we refer to the extensive monograph [Boyd and Vandenberghe, 2004]. All together we note that concavity is an extremely favourable property for a function that shall be maximised, which will be the log likelihood function later on.

A second property which is important in the existence theory of maximisers is *coercivity* in the sense that

$$f(x) \to -\infty \quad \text{for } |x| \to \infty.$$

In fact every (upper semi-) continuous and coercive function defined on a closed set $U \subseteq \mathbb{R}^M$ attains its minimum. To see this one can fix $x_0 \in U$ and use the coercivity to obtain $f < f(x_0)$ outside of a compact set K and thus the supremum of f agrees with the supremum of f over $K \cap U$ which is compact again and thus it is attained. We will later introduce some abstract theory about the consistency of estimators and for this we will need this result in a ore general setting. However the version above is enough in the case of the maximum likelihood estimators for parameters of DPPs and therefore readers that are not familiar with elementary notions of topology are advised to neglect following statement.

3.18 Proposition (Existence of maximisers). *Let $\mathcal{X}$ be a topological Hausdorff space and $f\colon \mathcal{X} \to [-\infty, \infty)$ be an* upper semicontinuous *function, i.e.*

$$L_f(\alpha) := \{x \in \mathcal{X} \mid f(x) \geq \alpha\}$$

is closed for all $\alpha \in \mathbb{R}$. Further we will assume that f is coercive*, meaning that for any $\alpha \in \mathbb{R}$ the set $L_f(\alpha)$ is compact. Then f attains its maximum in at least one point, i.e. there is $\hat{x} \in \mathcal{X}$ such that*

$$f(\hat{x}) = \sup_{x \in \mathcal{X}} f(x).$$

Proof. Let without loss of generality f be not identical to $-\infty$ because otherwise the statement is trivial. Then we have

$$\alpha := \sup_{x \in \mathcal{X}} f(x) > -\infty.$$

If we choose (α_n) to be strictly increasing towards α, then we get $L_f(\alpha_{n+1}) \subseteq L_f(\alpha_n)$ for all $n \in \mathbb{N}$ and further none of the sets $L_f(\alpha_n)$ is empty. By the Cantor intersection theorem[1] we get that also the intersection is non empty, i.e there is

$$\hat{x} \in \bigcap_{n \in \mathbb{N}} L_f(\alpha_n).$$

This implies

$$f(\hat{x}) \geq \alpha_n \xrightarrow{n \to \infty} \alpha = \sup_{x \in \mathcal{X}} f(x).$$

□

[1] A precise formulation can be found in the appendix.

III.2.1 Presentation of different models

Assume again that we have a set of observations $(\mathbf{Y}_n)_{n\in\mathbb{N}} \subseteq \mathcal{Y}$ drawn independently and according to the DPP. This time we want to find the maximum likelihood estimator for the elementary kernel and in order to do this we need to be able to express the density of the DPP which is nothing but the values of the elementary probabilities. Thus we will assume that we are dealing with L-ensembles in this section. Since the observations $(\mathbf{Y}_n)$ are defined on some common probability space which we will denote by $(\Omega, \mathbb{P})$ we will change the notation in this section and write

$$f(A|\theta) \propto \det(L(\theta)_A)$$

for the elementary probabilities of the DPP that arises from the parameter θ. Note that the elementary probabilities are nothing than the density with respect to the counting measure. We will now present the maximum likelihood estimators for different parametric classes, i.e. different families $\mathcal{F}$ of DPPs.

MLE OF THE ELEMENTARY KERNEL L

The most intuitive parameter that one can estimate is the elementary kernel L itself since it parametrises the entire class of L-ensembles.

3.19 MAXIMUM LIKELIHOOD ESTIMATOR FOR L. We consider the parameter space $\Theta = \mathbb{R}^{N\times N}_{\text{sym},+}$ of positive definite symmetric matrices and the parametric family

$$\mathcal{F} = \left\{ f(\cdot, L) \;\middle|\; L \in \mathbb{R}^{N\times N}_{\text{sym},+} \right\}$$

where $f(A, L) \propto \det(L_A)$ is the elementary probability of DPP with elementary kernel L. We seek to find the MLE

$$\hat{L}_n := \operatorname*{arg\,max}_{L\in\mathbb{R}^{N\times N}_{\text{sym},+}} \mathcal{L}(L).$$

The log likelihood function is now given by

$$\mathcal{L}\colon \mathbb{R}^{N\times N}_{\text{sym},+} \to [-\infty, 0], \qquad L \mapsto \log\left(\prod_{i=1}^{n} f(\mathbf{Y}_i|L)\right).$$

Using (2.4) we get the expression

$$\mathcal{L}(L) = \sum_{i=1}^{n} \log\left(\det(L_{\mathbf{Y}_i})\right) - n\log\left(\det(L+I)\right). \tag{3.11}$$

Although the parametric family of that arises from the elementary kernels L gives a high variety of different associated L-ensembles, it will also make the computation of the MLE more complex. Therefore we will consider some smaller classes of L-ensembles, which will decrease the flexibility of the model, but make computation more efficient.

MLE of the qualities

Unlike earlier we will not try to estimate the whole kernel L but only the qualities q_i of the items $i \in \mathcal{Y}$. More precisely we recall that we can parametrise the positive definite symmetric matrices L using the quality diversity parametrisation

$$(q, \phi) \mapsto \Psi(q, \phi) = L \quad \text{where } L_{ij} = q_i \phi_i^T \phi_j q_j.$$

Now we fix a diversity feature matrix $\hat{\phi}$, that we will usually model according to some perceptions we might have and set $\hat{S}_{ij} := \phi_i^T \phi_j$. We will now try to estimate the quality vector $q \in \mathbb{R}_+^N$ instead of the whole kernel L. This means that we optimise the likelihood function over a smaller set of kernels, namely the ones of the form $\Psi(q, \hat{\phi})$ for $q \in \mathbb{R}_+^N$. Obviously the maximal likelihood that can be achieved using this more restrictive model decreases since we consider less positive definite matrices and we have

$$\max_{q \in \mathbb{R}_+^N} \mathcal{L}(\Psi(q, \hat{\phi})) \leq \max_{L \in \mathbb{R}_{\mathrm{sym},+}^{N \times N}} \mathcal{L}(L).$$

Although we can only expect a worse descriptive power of the observation, the hope is that the task of estimating only the qualities $q \in \mathbb{R}_+^N$ is more feasible which actually turn out to be true in certain cases. But before we investigate this, we clearly state our goal.

3.20 Maximum likelihood estimator for the quality. This time we work with the parameter set $\Theta = \mathbb{R}_+^N$ and the parametric family

$$\mathcal{F} = \left\{ f(\cdot|q) \;\middle|\; q \in \mathbb{R}_+^N \right\}$$

where $f(A, q) \propto \det(\Psi(q, \hat{\phi})_A)$ is the elementary probability of DPP with elementary kernel $\Psi(q, \hat{\phi})$. We aim to find the MLE of the quality vector $q \in \mathbb{R}_+^N$, in other words we set

$$\hat{q}_n := \arg\max_{q \in \mathbb{R}_+^N} \mathcal{L}(q)$$

where we perceive the likelihood function as a function of q.

Using (2.5) we obtain the following expression for the single summands of the log likelihood function

$$\log\left(\prod_{j \in Y_i} q_j^2\right) + \log(\det(\hat{S}_{Y_i})) - \log\left(\sum_{A \subseteq \mathcal{Y}} \prod_{j \in A} q_j^2 \det(\hat{S}_A)\right) \tag{3.12}$$

and note that it is upper semicontinuous.

Log linear model for the qualities

The motivation for restricting our ambitions of estimation to the qualities q_i rather than the whole elementary kernel $L \in \mathbb{R}_{\mathrm{sym},+}^{N \times N}$ was to obtain a more tractable optimisation problem. Unfortunately we can tell from (3.12) that the log likelihood still isn't concave in q and in order to achieve this, we will introduce the following model for the qualities.

3.21 Log linear model for the qualities and MLE. From now on we will fix vectors $f_i \in \mathbb{R}^M$ for $i \in \mathcal{Y}$ and call them *feature vectors*. Further we set

$$q_i = \exp\left(\theta^T f_i\right) \quad \text{for } \theta \in \mathbb{R}^M$$

and will only consider quality vectors $q \in \mathbb{R}_+^N$ that have this form. To formulate the maximum likelihood estimator for θ we set $\Theta := \mathbb{R}^M$ and consider the parametric family

$$\mathcal{F} = \left\{ f(\cdot|\theta) \mid \theta \in \mathbb{R}^M \right\}$$

where $f(\cdot|\theta)$ is the density of the DPP with similarity kernel $\hat{S}$ and qualities $q_i = \exp\left(\frac{1}{2}\theta^T f_i\right)$. Further we will consider the maximum likelihood estimator

$$\hat{\theta}_n := \underset{\theta \in \mathbb{R}^M}{\arg\max}\, \mathcal{L}(\theta)$$

where we regard $\mathcal{L}$ again as function of θ.

3.22 Remark. It shall be noted that although this log linear model seems to be a harsh restriction, it isn't a restriction at all, at least theoretically. If we take $M = N$ and choose f_i to be the unit vectors in $\mathbb{R}^N$, then this just a logarithmic transformation of the parameters and thus the maximal likelihood that can be achieved with this model does not change. In practice however it will be of interest to work with rather low dimensional parameters θ, because if the ground set $\mathcal{Y}$ gets large, optimisation in $\mathcal{R}^N$ can be inefficient. In this case of course the maximal likelihood under the optimal parameter may decrease. However the approximation of the optimal parameter might become possible again which justifies this sacrifice.

Under the assumption of a log linear model for the qualities the individual terms of the log likelihood function take the form

$$2 \cdot \theta^T \sum_{i \in Y} f_i + \det(\hat{S}_Y) - \log\left(\sum_{A \subseteq \mathcal{Y}} \exp\left(2 \cdot \theta^T \sum_{i \in A} f_i\right) \det(\hat{S}_A)\right). \tag{3.13}$$

MLE of the repulsiveness parameter

III.2.2 Coercivity and existence of the maximum likelihood estimators

A priori it is not clear that the maximum likelihood estimators exist and we will actually see that they do not exist in general. However one can still save this approach because the probability that they exist tends to 1 if the sample size increases. We begin by showing this for the MLE of the qualities and then we will adapt this proof to the other models.

MLE of the qualities

The MLE $\hat{q}_n$ does not exist for all realisations $(Y_n)_{n\in\mathbb{N}}$ of $(\mathbf{Y}_n)_{n\in\mathbb{N}}$. To see this we suppose that we have only one sample $Y_1 = \mathcal{Y}$ which is the whole set. The higher the qualities of the items are, the more likely this observation gets and therefore the maximum of the log likelihood function – which is 0 in this case – is not obtained. This can also be made rigorous in the following computation. Under the assumption of constant qualities the log likelihood function takes the form

$$\log\left(q^{2N}\det(\hat{S}_{\mathcal{Y}})\right) - \log\left(\sum_{A\subseteq\mathcal{Y}} q^{2|A|}\det(\hat{S}_A)\right) = \log\left(\frac{q^{2N}\det(\hat{S}_{\mathcal{Y}})}{\sum_{A\subseteq\mathcal{Y}} q^{2|A|}\det(\hat{S}_A)}\right) \xrightarrow{q\to\infty} 0.$$

However this maximum is never attained, since for every L-ensemble we have $\mathbb{P}_L(\varnothing) > 0$ and therefore

$$\mathcal{L}(q) = \log\left(\mathbb{P}_{\Psi(q,\hat{S})}(\mathcal{Y})\right) < 0 \quad \text{for every } q \in \mathbb{R}_+^N.$$

The thing that goes wrong in this case is, that under the observation of the whole set $\mathcal{Y}$ we would estimate a deterministic model that always selects the whole set, namely the DPP with marginal kernel I. Since all of the eigenvalues are 1 in this case, this DPP is not a L ensemble and therefore we can not describe it with the quality diversity decomposition. However if we assume that the data is actually generated by a L-ensemble, then such a scenario becomes unlikely as the sample size increases. We will fix this in the following result.

3.23 Proposition (Coercivity and existence of the MLE). *Let $\mathbf{Y}_1, \mathbf{Y}_2, \dots$ be a sequence of independent and identically distributed point processes that fall in the class of L-ensembles. Then we have*

$$\mathbb{P}\left(\hat{q}_n \in \mathbb{R}_+^N \text{ exists}\right) \geq \mathbb{P}\big(\mathcal{L} \text{ is coercive}\big) \xrightarrow{n\to\infty} 1.$$

Proof. The first inequality is obvious since the log likelihood function is upper semicontinuous. We will show that $\mathcal{L}$ is coercive if one of the observations is the emptyset. Then the claim follows from

$$\begin{aligned}\mathbb{P}\big(\mathcal{L}\text{ is coercive}\big) &\geq \mathbb{P}\left(\bigcup_{i=1}^n \{\mathbf{Y}_i = \varnothing\}\right) = 1 - \mathbb{P}\left(\bigcap_{i=1}^n \{\mathbf{Y}_i \neq \varnothing\}\right) \\ &= 1 - \mathbb{P}(\mathbf{Y}_1 \neq \varnothing)^n \xrightarrow{n\to\infty} 1\end{aligned}$$

since we have $\mathbb{P}(\mathbf{Y}_1 \neq \varnothing) < 1$ for every L-ensemble.

So let $Y_1, \dots, Y_n$ be some observations with $Y_i = \varnothing$ for at least one $i \in \{1, \dots, n\}$ and let $(q^k)_{k\in\mathbb{N}} \subseteq \mathbb{R}_+^N$ be a sequence such that $|q^k| \to \infty$. Note that it suffices to show that every subsequence of (q^k) contains a subsubsequence (q^l) such that

$$\mathcal{L}(q^l) \to -\infty \quad \text{for } l \to \infty.$$

Hence we fix a subsequence of (q^k) which we denote by (q^k) again in slightly abusive notation. Let (q^l) be a subsequence of (q^k) such that one coordinate diverges to infinity, i.e.

$$q^l_{j_0} \xrightarrow{l\to\infty} \infty \quad \text{for one } j_0 \in \{1, \dots, N\}.$$

The i-th summand of $\mathcal{L}$ takes the form

$$-\log\left(\sum_{A\subseteq\mathcal{Y}}\prod_{j\in A}(q^l_j)^2\det(\hat{S}_A)\right) \leq -\log\left((q^l_{j_0})^2\right) \xrightarrow{l\to\infty} -\infty$$

where we used $\hat{S}_{\{j_0\}} = 1$. Because the other summands are non positive this implies

$$\mathcal{L}(q^l) \xrightarrow{l\to\infty} -\infty$$

which we had to show. □

3.24 Remark. The proof above should be read in the following way. The statement $q^l_{j_0} \to \infty$ is equivalent to a model that would always select the item j_0. However since we have observed the empty set, the observations would be impossible under this model and thus the log likelihood function takes the value $-\infty$ for this model. An analogue argument shows that the estimated qualities are strictly positive with high probability if the actual qualities are strictly positive. This will be of interest for us if we consider the log linear model.

3.25 Proposition (Positivity of the MLE). *Assume that $\mathbf{Y}_1, \mathbf{Y}_2, \dots$ is a sequence of independent and identically distributed point processes that are distributed according to a L-ensemble with strictly positive qualities. Then we have*

$$\mathbb{P}\left(\hat{q}_n \in \mathbb{R}^N_+ \textit{ exists and } \hat{q}_n \in (0,\infty)^N\right) \xrightarrow{n\to\infty} 1.$$

Proof. We have already seen that the probability that the MLE exists tends to one, so we only have to show that the probability that the estimated qualities are strictly positive tends to one. The philosophy to prove this is exactly the same than in the proof of existence. Indeed we note that once j occurs in one of the observations $Y_1, \dots, Y_n$ we have $\mathcal{L}(q) = -\infty$ for every $q \in \mathbb{R}^N_+$ with $q_j = 0$. Therefore we have $(\hat{q}_n)_j > 0$ if $j \in Y_i$ for at least one $j \in \{1, \dots, n\}$. Finally we note that the probability that j occurs in the i-th sample is strictly positive since we have

$$\mathbb{P}(j \in \mathbf{Y}_i) \geq \mathbb{P}(\{j\} = \mathbf{Y}_i) = q_j^2 > 0.$$

□

MLE of the elementary kernel

We can quite easily adapt the proof for the existence of MLEs of the qualities to the case of MLEs for the whole elementary kernel L.

3.26 Proposition (Coercivity and existence of MLE). *Let $\mathbf{Y}_1, \mathbf{Y}_2, \dots$ be an sequence of independent and identically distributed point processes that fall in the class of L-ensembles. Then we have*

$$\mathbb{P}\left(\hat{L}_n \in \mathbb{R}^{N\times N}_{sym,+} \textit{ exists}\right) \geq \mathbb{P}\big(\mathcal{L} \textit{ is coercive}\big) \xrightarrow{n\to\infty} 1.$$

Proof. Again it suffices to show $\mathcal{L}(L) \to -\infty$ for $|L| \to \infty$ once we have observed the empty set once. To see this, we use the quality diversity parametrisation

$$\Psi\colon \mathbb{R}^N_+ \times \mathbb{S}^N_N \to \mathbb{R}^{N\times N}_{\text{sym},+}, \quad (q,\phi) \mapsto \left(q_i \phi_i^T \phi_j q_j\right)_{1\leq i,j\leq N}.$$

Note that since Ψ is continuous and therefore bounded on bounded sets and $\mathbb{S}^N_N$ is bounded, $|\Psi(q,\phi)| \to \infty$ implies $|q| \to \infty$. The exact same calculations as in the previous proof show

$$\mathcal{L}(L) = \mathcal{L}(\Psi(q,\phi)) \to -\infty \quad \text{for } |L| \to \infty.$$

□

The log linear model

We have seen that the log linear model can provide a parametrisation of the whole space $(0,\infty)^N$ of possible qualities. However it can also be very ristrictive, for example if all feature vectors are trivial, i.e. $f_i = 0$ for all items i. Hence we need to convince ourselves that we do not loose too much information through the transformation

$$F\colon \mathbb{R}^M \to (0,\infty)^N, \quad \theta \mapsto \big(\exp(\theta^T f_1), \dots, \exp(\theta^T f_N)\big)^T.$$

In order to do this, let $U \subseteq \mathbb{R}^M$ be the span of $f_1, \dots, f_N$ and let write $\theta = \theta_1 + \theta_2$ such that $\theta_1 \in U$ and $\theta_2 \in U^\perp$. We note that $F(\theta) = F(\tilde{\theta})$ if and only if $\theta_1 = \tilde{\theta}_1$.

3.27 Proposition (Coercivity and existence of MLE). *Assume that* $\mathbf{Y}_1, \mathbf{Y}_2, \dots$ *is a sequence of independent and identically distributed point processes that are distributed according to a L-ensemble with strictly positive qualities. Then we have*

$$\mathbb{P}\left(\hat{\theta}_n \in \mathbb{R}^M \textit{ exists}\right) \ge \mathbb{P}\Big(\mathcal{L} \textit{ is coercive as a function on } U\Big) \xrightarrow{n\to\infty} 1.$$

Proof. First we note, that it suffices to show that $\mathcal{L}$ has a maximiser on U, since $F(U) = F(\mathbb{R}^M)$. To do this we show – just like in the previous cases – that the $\mathcal{L}$ is coercive on U whenever we have observed the emptyset as well as every item at least once. Let now $(\theta^k)_{k\in\mathbb{N}} \subseteq U$ be a sequence such that $|\theta^k| \to \infty$. Then there is at least one index $i \in \{1, \dots, N\}$ and a subsequence $(\theta^l)_{l\in\mathbb{N}}$ such that

$$f_i^T \theta^l \to \infty \quad \text{or } f_i^T \theta^l \to -\infty \quad \text{for } l \to \infty$$

since otherwise all sequences $\big(f_i^T \theta^l\big)$ therefore also (θ^l) would be bounded. However this is equivalent to

$$\exp(f_i^T \theta^l) \to \infty \quad \text{or } \exp(f_i^T \theta^l) \to 0 \quad \text{for } l \to \infty$$

and we have seen in the proof of 3.25 that the log likelihood function tends to $-\infty$ in this case. □

MLE for the repulsiveness parameter

III.2.3 Consistency of the maximum likelihood estimators

We will now turns towards the question of consistency of the maximum likelihood estimators introduced earlier in this section. For this we will first give a formal proof of the consistency of the MLE and present a rather general framework that will allow us to turn the formal proof into a rigorous one.

3.28 Formal proof of consistency. We will consider a general MLE like in (3.10) and we will assume that the observations (X_n) are independent and have density $f(x|\theta_0)$ with respect to some measure μ. By the law of large number we have

$$\frac{1}{n}\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \log(f(X_i|\theta)) \xrightarrow{n\to\infty} \mathbb{E}\big[\log(f(X|\theta))\big]. \tag{3.14}$$

Hence the maximiser of the left hand side should be close to the maximiser of the right hand. Differentiating the left hand side yields

$$\begin{aligned}\partial_\theta \mathbb{E}\big[\log(f(X|\theta))\big] &= \mathbb{E}\big[\partial_\theta \log(f(X|\theta))\big] = \mathbb{E}\left[\frac{\partial_\theta f(X|\theta)}{f(X|\theta)}\right] \\ &= \int \frac{\partial_\theta f(x|\theta)}{f(x|\theta)} f(x|\theta_0)\mu(\mathrm{d}x).\end{aligned}$$

Evaluating this at $\theta = \theta_0$ yields

$$\int \partial_\theta f(x|\theta)\mu(\mathrm{d}x) = \partial_\theta \int f(x|\theta)\mu(\mathrm{d}x) = \partial_\theta(1) = 0.$$

Hence θ_0 is a critical point and under mild conditions the left hand side is concave and thus θ_0 is the unique maximiser. In conclusion the estimator $\hat{\theta}$ should be close to θ_0.

Although rough structure of the rigorous proof is present in the argument above it is highly formal. For example we argue that if a sequence $(f_n)_{n\in\mathbb{N}}$ of functions converges towards f pointwise, then the maximisers $(x_n)_{n\in\mathbb{N}}$ should converge to the maximiser x of f. The major tool to make this rigorous will be to use some kind of uniform convergence. Namely we have the following result where we will omit the proof since it is very easy and we proof a similar but stronger version of it later.

3.29 Lemma (Swapping limit and maximisation). *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of real functions with maximisers $(x_n)_{n\in\mathbb{N}}$ that are bounded from above and converge uniformly towards f. Further assume that f is coercive and continuous and let x be the unique maximiser of f. Then we have $x_n \to x$ for $n \to \infty$.*

Unfortunately the convergence in (3.14) does only hold uniformly on a compact set K. To fix this one has to ensure that the maximisers (x_n) lie in this compact set K for large n. We will do this in a general setup in the next paragraph.

A general consistency result for extremal estimators

We will prove a general consistency result for a rather broad class of estimators which is taken from [Newey and McFadden, 1994] and slightly adapted to our needs. Although it would be possible to prove the consistency of the MLEs directly we present this general procedure since this can easily be adjusted to other cases.

3.30 Setting. Let in the following Θ be a topological Hausdorff space and $F_n\colon \Theta \to [-\infty, \infty)$ be a sequence of random functions with maximisers

$$\hat{\theta}_n := \arg\max_{\theta\in\Theta} F_n(\theta).$$

If no maximiser exists, we choose $\hat{\theta}_n \in \Theta$ arbitrary. Further let $F\colon \Theta \to [-\infty, \infty)$ be a deterministic function with maximiser θ_0.

The maximisers $\hat{\theta}_n$ are called *extremal* estimators since they are the extremal points of the functions F_n. We now turn towards the question whether the extremal estimators converge to the maximiser θ_0.

3.31 Theorem (Consistency of extremal estimators). *Let the setting be as above and assume that the following conditions hold.*

(*i*) *Assume that there is $\varepsilon_0 > 0$ and a compact set K_0 containing θ_0, such that with probability tending to one*

$$F_n(\theta) \le F(\theta_0) - \varepsilon_0 \quad \textit{for all } \theta \notin K_0. \tag{3.15}$$

(*ii*) *Let F_n converge to F uniformly on K_0 in probability, i.e. for any $\varepsilon > 0$ we have with probability tending to one*

$$\left|F_n(\theta) - F(\theta)\right| \leq \varepsilon \quad \text{for all } \theta \in K_0. \tag{3.16}$$

(*iii*) *Let F have a unique maximum at $\theta_0 \in \Theta$.*

(*iv*) *Assume that F is upper semicontinuous in the sense that*

$$\left\{\theta \in \Theta \mid F(\theta) \geq \alpha\right\} \subseteq \Theta$$

is closed for all $\alpha \in \mathbb{R}$.

(*v*) *With probability tending to one F_n admits a maximiser.*

Then we have $\hat{\theta}_n \to \theta_0$ in probability, i.e.

$$\mathbb{P}\left(\hat{\theta}_n \in U\right) \xrightarrow{n\to\infty} 1$$

for any open subset $U \subseteq \Theta$ containing θ_0.

Proof. Note that it suffices to show $\hat{\theta}_n \in U$ whenever (3.15) and (3.16) hold and F_n admits a maximiser. From here on the proof is of purely analytic content.

Fix now an open set $U \subseteq \Theta$ that contains θ_0. Choosing $\varepsilon < \varepsilon_0$ in (*ii*) and using (*i*) yields

$$F_n(\theta_0) \geq F(\theta_0) - \varepsilon > F(\theta_0) - \varepsilon_0 \geq F_n(\theta) \quad \text{for all } \theta \notin K_0.$$

Hence the maximum of F_n is attained in K_0 and we have $\hat{\theta}_n \in K_0$. Thus if $K_0 \subseteq U$ we are done. If this is not the case, 3.18 together with (*iv*) implies that F attains its maximum α on $K_0 \setminus U$ which is strictly smaller than $F(\theta_0)$ because of (*iii*). Thus we have

$$K_0 \cap \left\{\theta \in \Theta \mid F(\theta) > \alpha\right\} \subseteq U.$$

So in order to show $\hat{\theta}_n \in U$, it suffices to show $\hat{\theta}_n \in K_0$ and $F(\hat{\theta}_n) > \alpha$. Since we have already seen that the first statement holds, it remains to show the second one. However (*ii*) implies

$$F(\hat{\theta}_n) \geq F_n(\hat{\theta}_n) - \varepsilon \geq F_n(\theta_0) - \varepsilon \geq F(\theta_0) - 2\varepsilon > \alpha$$

for ε small enough. □

The proof above can be read in the following way. The two conditions (*i*), (*ii*) and (*v*) force $\hat{\theta}_n \in K_0$ and hence we can restrict our considerations to a compact set. This is actually the setting the consistency results are usually presented in. Then we argue that if $F(\theta)$ is close to $F(\theta_0)$, then θ has to be close to θ_0, where we use that we are on a compact subset as well as (*iii*) and (*iv*). However since F_n is close to F on K_0 we obtain that $F(\hat{\theta}_n)$ is close to $F_n(\hat{\theta}_n)$ which is close to $F(\theta_0)$ since the maximum values converge to each other.

If we want to apply the previous result to the case of maximum likelihood estimation we need to set

$$F_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \log(f(X_i|\theta)).$$

Note that the factor $\frac{1}{n}$ does not change the maximum. However (3.14) already gives the almost surely pointwise limit of those functions and if condition (*ii*) of the previous statement should hold, we have to define

$$F(\theta) := \mathbb{E}\big[\log(f(X|\theta))\big].$$

The quantity F is known as the *entropy* and plays an important role in statistical mechanics, applied statistics, information theory and many other fields. For more information on this we refer to [Martin and England, 2011], [MacKay and Mac Kay, 2003], [Volkenstein, 2009] and [Gray, 1990].

Information inequality and locally uniform convergence

The third requirement of the previous consistency result can be proven in a general setting and without quantitative assumption and we adapt an argument from [Newey and McFadden, 1994] to fit our needs. In order to do this we will work with the following assumptions.

3.32 Setting. Let in the following Θ be a set and let

$$\mathcal{F} = \Big\{ f(\cdot|\theta) \colon \mathcal{X} \to [0,\infty) \mid \theta \in \Theta \Big\}$$

be a family of probability densities on some measurable space $\mathcal{X}$ with respect to some measure μ. Further fix $\theta_0 \in \Theta$ and denote the expectation with respect to $f(\cdot|\theta_0)\mathrm{d}\mu$ by

$$\mathbb{E}[h(X)] := \int h(x) f(x|\theta_0)\mu(\mathrm{d}x).$$

Let (X_n) be a sequence of independent random variables distributed according to $f(\cdot|\theta_0)\mathrm{d}\mu$. Finally define

$$F_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \log(f(X_i|\theta)) \quad \text{and } F(\theta) := \mathbb{E}\big[\log(f(X|\theta))\big].$$

3.33 Proposition (Information inequality). *Let the setting be as above and assume that the parameter $\theta_0 \in \Theta$ is identifiable, i.e. we have $f(\cdot|\theta) \neq f(\cdot|\theta_0)$ whenever $\theta \neq \theta_0$. Let further*

show $F(\theta_0) > -\infty$

$$\sup_{x\in\mathcal{X}|\theta\in\Theta} f(x|\theta) < \infty \quad \textit{and } F(\theta_0) > -\infty.$$

Then the entropy

$$F(\theta) = \mathbb{E}\big[\log(f(X|\theta))\big]$$

has a unique maximum in θ_0.

Proof. Let $\theta \neq \theta_0$, then we either have $F(\theta) = -\infty < F(\theta_0)$ or

$$F(\theta) = \mathbb{E}\big[\log(f(X|\theta))\big] > -\infty. \tag{3.17}$$

In this case we want to exploit the strict Jensen inequality (c.f. [Lehmann and Casella, 2006]) that yields for any positive random variable Y with finite expectation that is not constant

$$\mathbb{E}[\log(Y)] < \log(\mathbb{E}[Y]).$$

We set $Y := \frac{f(X|\theta)}{f(X|\theta_0)}$. This is positive $f(\cdot|\theta_0)\mathrm{d}\mu$ almost everywhere because otherwise (3.17) could not hold. Since θ_0 is identifiable, the random variable Y is not constant and we will see in the following computation that the expectation is finite. Now we obtain

$$\begin{aligned} F(\theta) - F(\theta_0) &= \mathbb{E}\big[\log(f(X|\theta))\big] - \mathbb{E}\big[\log(f(X|\theta_0))\big] = \mathbb{E}\left[\log\left(\frac{f(X|\theta)}{f(X|\theta_0)}\right)\right] \\ &< \log\left(\mathbb{E}\left[\frac{f(X|\theta)}{f(X|\theta_0)}\right]\right) = \log\left(\int f(x|\theta)\mu(\mathrm{d}x)\right) = 0. \end{aligned}$$

□

Next we take care of the first requirement of the consistency result. Namely we will show that the functions F_n associated with the MLE almost surely converge to F locally uniformly under fairly mild conditions. For this we modify the proof of a more general convergence result in [Tauchen, 1985].

3.34 Lemma (Locally uniform convergence). *Let the setting be as above, but let Θ be a metric space and we assume that the following conditions hold.*

(*i*) *Let $K \subseteq \Theta$ be such that*

$$\mathbb{E}\left[\sup_{\theta\in K}\big|\log(f(X|\theta))\big|\right] < \infty.$$

(*ii*) *For every $\theta \in K$ we have $\log(f(\cdot,\gamma)) \to \log(f(\cdot|\theta))$ almost surely with respect to $f(\cdot|\theta_0)\mathrm{d}\mu$ for $\gamma \to \theta$.*

Then we almost surely have $F_n \to F$ uniformly on K, i.e. almost surely we have

$$\sup_{\theta\in K}\big|F_n(\theta) - F(\theta)\big| \xrightarrow{n\to\infty} 0.$$

Proof. Fix $\varepsilon > 0$ and a compact set $K \subseteq \Theta$. Define for $\rho > 0$

$$u(x,\theta,\rho) := \sup_{d(\gamma|\theta)\le\rho}\big|\log(f(x,\gamma)) - \log(f(x|\theta))\big| \xrightarrow{\rho\to\infty} 0$$

almost surely for θ fixed where we used condition (*ii*).

This in combination with (*i*) and the dominated convergence theorem imply that the convergence also holds in expectation and therefore we have

$$\mathbb{E}\big[u(X,\theta,\rho)\big] \le \varepsilon \quad \text{for } \rho \le \delta(\theta).$$

The open balls $B_{\delta(\theta)}(\theta)$ with center θ and radius $\delta(\theta)$ cover the compact set K and hence we can select a finite subcover

$$K \subseteq \bigcup_{k=1}^{m} B_{\delta(\theta_k)}(\theta_k).$$

Further we set

$$\mu_k := \mathbb{E}\big[u(X,\theta_k,\delta(\theta_k))\big] \le \varepsilon.$$

Let $\theta \in K$ and choose k such that $\theta \in B_{\delta(\theta_k)}(\theta_k)$, then we can conclude

$$
\begin{aligned}
|F_n(\theta) - F(\theta)| &\leq \frac{1}{n}\sum_{i=1}^{n} \big| \log(f(X_i|\theta)) - \log(f(X_i|\theta_k)) \big| \\
&\quad + \left| \frac{1}{n}\sum_{i=1}^{n} \log(f(X_i|\theta_k)) - F(\theta_k) \right| + \big| F(\theta_k) - F(\theta) \big| \\
&\leq \left(\frac{1}{n}\sum_{i=1}^{n} u(X_i, \theta_k, \delta(\theta_k)) - \mu_k \right) + \mu_k + 2\varepsilon \\
&\leq 4\varepsilon
\end{aligned}
$$

almost surely for $n \geq N(\varepsilon)$ where we used the strong law of large numbers twice. □

Consistency of the MLEs for the quality and elementary kernel

In this part we will – for the first time – make use of the specific structure of the model. Since we have already taken care of condition (*ii*) and (*iii*) and condition (*iv*) will be fairly straight forward, we dedicate ourselves to proving the first requirement of Theorem 3.31. For this we keep the setting of the previous section although we now consider the case that

$$
\mathcal{F} = \Big\{ f(\cdot|\theta) \colon X \to [0, \infty) \mid \theta \in \Theta \Big\}
$$

is one of the parametric families introduced in III.2.1.

3.35 Lemma (Control outside of a compact set). *The requirement* (*i*) *from Theorem 3.31 is satisfied for the three kinds of parametric families for the kernel estimation. Further the compact set K_0 can be chosen as follows. Let $\mathcal{A}$ be the family of subsets $A \subseteq \mathcal{Y}$ with positive probability $f(A|\theta_0) > 0$ and let $c(A) > 0$ such that*

$$
-c(A) < 2 \cdot \log(f(A|\theta_0)).
$$

Then we set

$$
K_0 := \Big\{ \theta \in \Theta \;\Big|\; \log(f(A|\theta)) \geq -c(A) \text{ for all } A \in \mathcal{A} \Big\}.
$$

Proof. Let

$$
\hat{\mathbb{P}}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{Y}_i}
$$

be the empirical measure. We have by the law of large numbers

$$
\mathbb{P}\left(\hat{\mathbb{P}}_n(A) \geq \frac{f(A|\theta_0)}{2} \right) \xrightarrow{n\to\infty} 1
$$

and we can assume $\hat{\mathbb{P}}_n(A) \geq \frac{f(A|\theta_0)}{2}$, since we are only interested in proving a statement with probability tending to one. Whenever $f(A|\theta_0) > 0$ we choose $c(A) > 0$ such that

$$
-c(A) < 2 \cdot \log(f(A|\theta_0)).
$$

For exactly those A we define now

$$K_A := \Big\{\theta \in \Theta \;\Big|\; \log(f(A|\theta)) \geq -c(A)\Big\}$$

which is closed since $f(A|\theta)$ is upper semicontinuous. Further K_A is compact for $A = \varnothing \in \mathcal{A}$ since $\log(f(\varnothing|\theta))$ is coercive in θ as been shown in the coercivity proofs earlier in this chapter. Further it contains θ_0 as

$$\log(f(A|\theta_0)) \geq 2 \cdot \log(f(A|\theta_0)) > -c(A)$$

because $f(A|\theta) \leq 1$. Set now

$$K_0 := \bigcap_{A \in \mathcal{A}} K_A$$

which is compact because $K_\varnothing$ is compact. Finally we have for $\theta \notin K_0$

$$\begin{aligned} F_n(\theta) = \int \log(f(x|\theta))\hat{\mathbb{P}}_n(\mathrm{d}x) &= \sum_{A \in \mathcal{A}} \hat{\mathbb{P}}_n(A) \cdot \log(f(A|\theta)) < -\sum_{A \in \mathcal{A}} \frac{f(A|\theta_0)}{2} c(A) \\ &< \sum_{A \in \mathcal{A}} \log(f(A|\theta_0)) f(A|\theta_0) = F(\theta_0). \end{aligned}$$

□

Now we have all the auxiliary results to prove the desired consistency result.

3.36 Theorem (Consistency). (*i*) *The maximum likelihood estimator $\hat{L}_n$ for the elementary kernel is consistent. Namely if the observations $(\mathbf{Y}_n)$ follow the law of a L-ensemble with kernel L_0, then we have*

$$\mathbb{P}\left(d(\hat{L}_n, L_0) \leq \varepsilon\right) \xrightarrow{n\to\infty} 1 \quad \textit{for all } \varepsilon > 0.$$

(*ii*) *The maximum likelihood estimator $\hat{q}_n$ for the quality vector is consistent. Namely if the observations $(\mathbf{Y}_n)$ follow the law of a L-ensemble with kernel $\Psi(p_0, \hat{S})$, then we have*

$$\mathbb{P}\left(\left\|\hat{q}_n - q_0\right\| \leq \varepsilon\right) \xrightarrow{n\to\infty} 1 \quad \textit{for all } \varepsilon > 0.$$

(*iii*) *Suppose that the observations $(\mathbf{Y}_n)$ follow the law of a L-ensemble with kernel $\Psi(p_0, \hat{S})$ where $(p_0)_i = \exp(\theta^T f_i)$ and let P denote the projection onto the subspace U. Then we have*

$$\mathbb{P}\left(\left\|P\hat{\theta}_n - P\theta_0\right\| \leq \varepsilon\right) \xrightarrow{n\to\infty} 1 \quad \textit{for all } \varepsilon > 0.$$

Proof. We will only sketch the main parts of the proof of the second statement, since all other arguments will be mostly analogue and therefore redundant.

Obviously we want to exploit the machinery we have introduced and thus we will check the requirements of Theorem 3.31. First we note that (*v*) holds because of the section of the existence of the maximum likelihood estimators.

We can express the entropy function

$$F(q) = \mathbb{E}\big[\log(f(\mathbf{Y}|q))\big] = \sum_{A \subseteq \mathcal{Y}} \log(f(A|q)) f(A|q_0) \tag{3.18}$$

where the elementary probabilities are given by

$$f(A|q) = \frac{\prod_{i\in A} q_i^2 \det(\hat{S}_A)}{\sum_{B\subseteq\mathcal{Y}} \prod_{i\in B} q_i^2 \det(\hat{S}_B)} \tag{3.19}$$

which is continuous in q. Hence the entropy function F is upper semicontinuous and thus condition (*iv*) holds.

To check that (*iii*) holds we will use the information inequality 3.33. First we note that because of

$$f(\{i\}\,|q) \propto q_i^2$$

the parameter q_0 is identifiable and further we have

$$\sup_{A\subseteq\mathcal{Y}, q\in\mathbb{R}_+^N} f(A|q) \le 1$$

since the densities are elementary probabilities. Finally $F(q_0) > -\infty$ is clear from (3.18) and hence the third requirement is satisfied.

Since the previous lemma already takes care of condition (*i*) it suffices to show the second condition for which we will use 3.34. Hence it remains to check the two conditions of this lemma, but the second one – the continuity condition – obviously holds as can be seen from (3.19). To see that the first one also holds we note that for $A \subseteq \mathcal{Y}$ with $f(A|q_0) > 0$ and $\theta \in K_0$ we have

$$0 \ge \log(f(A|q)) \ge -c(A) > -\infty.$$

Hence the random variable

$$\sup_{q\in K_0} \left|\log(f(X|q))\right|$$

is almost surely finite and since the probability space $2^{\mathcal{Y}}$ is finite, the second condition holds. $\square$

3.37 Remark. Obviously in the proof of the consistency of the whole elementary kernel L and the log linearity constant θ one runs into the problem of unidentifiability. This is why one has to identify the parameters with each other that give rise to the same probability densities. In the case of the elementary kernel this is just the determinantal equivalence once again and in the case of the log linearity constant two parameters give rise to the same densities if and only if their projection onto U agrees.

III.2.4 Approximation of the MLE

Likelihood maximisation for L

We note that $\mathcal{L}$ is smooth and that its gradient can be expressed explicitly, at least on the domain $\{\mathcal{L} > -\infty\}$. This is due to the fact that the determinants of the submatrices are polynomials in the entries of L and the composition of those with the smooth function $\log\colon (0,\infty) \to \mathbb{R}$ stays smooth. This property allows the use of gradient methods but they face the problem that the loss function is non concave and thus those algorithms will generally not converge to a global maximiser. To see that the log linear likelihood function is not concave, we may consider the span $\{qI \mid q \in \mathbb{R}\}$ of the identity matrix. On this subspace $\mathcal{L}$ takes the form

$$\mathcal{L}(qI) = \sum_{i=1}^n \log(q^{|Y_i|}) - n\log((1+q)^N) = \sum_{i=1}^n |Y_i|\log(q) - nN\log(1+q)$$

which is not concave in general.

This obviously causes substantial computational problems in the calculation of the MLE let alone it exists. In fact it is NP hard to maximise a general non concave function and it is also conjectured to be NP hard to maximise the log likelihood function $\mathcal{L}$ in the case of L-ensembles. However there are still efficient maximising techniques for such functions that will eventually converge to local maximiser and that also work in very high dimensional spaces and thus this approach was taken by . Nevertheless we will not present this approach here, but rather favour a maximisation technique that is based on a fixed point iteration and was proposed in .

explain this term

cite

cite

cite

Fixed point iteration based maximisation

read, understand and summarise the paper

Computation for the log linear model

The first two terms are affine linear in θ and thus concave. To see that the last expression is also concave, it is convenient to to introduce the notion of log concavity and give a fundamental result.

3.38 Definition (Log concavity). We call a function f *log concave*, *log convex* or *log (affine) linear* if $\log(f)$ has the respective property.

3.39 Proposition (Additivity of log concavity). *The sum of log concave functions is again log concave.*

Proof. □

Give of cite proof.

As an immediate consequence we obtain that the expression in (3.4) is log concave which we will fix in a separate statement.

3.40 Corollary (Concavity of the likelihood function). *Under the log linear model for the qualities, the log likelihood function is concave in the log linearity parameter $\theta \in \mathbb{R}^M$.*

III.2.5 Learning for conditional DPPs

III.2.6 Estimating the mixture coefficients of k-DPPs

Chapter IV

Bayesian learning for DPPs

So far we have seen two different estimation techniques for the parameters of DPPs. Although we proved that they provide reasonable estimators in the sense that they are consistent, they have some drawbacks. For example we have seen that the MLEs for the different parameters do not exist in general, let alone that they are impossible to compute in reality. Further all of the estimators presented so far are point estimators, i.e. they return a single value for the desired parameter. Obviously this does not allow to capture any uncertainties and we have already seen in that the selection of the most possible outcome – in this case the MLE – might not yield a very typical one for a given random variable. Those are some reasons to consider the Bayesian approach of parameter estimation where the goal is to give a distribution – called the posterior – of the parameter that should be estimated instead of a single value. This can also help to overcome some – maybe even all of the problems presented above.

cite

At first we will present the general idea of Bayesian parameter estimation and then we will turn towards the question of computability. For this we will follow the approach of [Affandi et al., 2014a] and turn towards the popular Markov chain Monte Carlo (MCMC) methods and quickly explain their philosophy and how they can be used to approximate the posterior distribution of the parameter that is to be estimated.

IV.1 Bayesian approach to parameter estimation

For the introduction of the general Bayesian setup we pursue like in [Rice, 2006]. We are – just like in the case of MLE – in the setting that we want to estimate a parameter $\theta \in \Theta$ based on some relisations $x = (x_1, \dots, x_n)$ of some random variables $X = (X_1, \dots, X_n)$ where we have given a family

$$\mathcal{F} = \left\{ f_{X|\Theta}(\cdot|\theta) \mid \theta \in \Theta \right\}$$

of densities with respect to $\mu^n := \prod_{i=1}^n \mu(\mathrm{d}x_i)$. This time however we are not interested in returning a single value θ because this would be a vast simplification of the stochastic nature of the estimator. Thus we want to obtain a probability distribution over whole Θ that indicates how the parameters are to have caused the observed data. In order to present the procedure we will introduce the frame we will work in.

4.1 Setting. Let Θ be a measurable space and ν be a measure on Θ. Further let $f_\Theta \colon \Theta \to$

$[0, \infty]$ be a probability density with respect to ν, i.e.

$$\int f_\Theta(\theta)\nu(\mathrm{d}\theta) = 1$$

which we will call the *prior* distribution of the parameter θ.

Usually the prior distribution will encode some perceptions we have of the parameter. For example if we are trying to estimate a physical constant that we know has to be positive, then it is reasonable to select a prior that has its whole mass on the positive real line. However there is no clear set of rules how one can select a suitable prior to a given problem.

The density $f(x|\theta)$ describes how likely the observations are under the parameter theta and we want to find an expression of how likely the parameter theta is under the observations x. In order to obtain this, we will work with the joint density

$$f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x|\theta)f_\Theta(\theta) \quad \text{with respect to } \mu^n \times \nu$$

and condition this onto x. This yields

$$f_{\Theta|X}(\theta|x) = \frac{f_{X,\Theta}(x,\theta)}{\int f_{X,\Theta}(x,\theta)\nu(\mathrm{d}\theta)} = \frac{f_{X|\Theta}(x|\theta)f_\Theta(\theta)}{\int f_{X,\Theta}(x,\theta)\nu(\mathrm{d}\theta)} \tag{4.1}$$

4.2 Definition (Posterior distribution). The density $f_{\Theta|X}$ is called the *posterior distribution* of the parameter θ given the data x.

This posterior density is already the object we are interested in which is supposed to give us the information of the distribution of the parameter given the data x. It is proportional to the likelihood $f_{X|\Theta}(x|\theta)$ of the occuring data times the prior $f_\Theta(\theta)$ which can be understood in the way, that

4.3 Comparison to MLE. Maybe one feels slightly uncomfortable with the need of a choice for the prior distribution and it turns out that this is in fact a difficult step that has to be taken with a certain amount of care. However we could pretend for one moment to be completely ignorant in the sense that we do not know anything about the parameter and hence we don't feel in the position to propose a reasonable prior. Then we could simply choose the uniform distribution as a prior – given it exists – and would obtain

$$f_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta).$$

Hence we can regain the MLE from our posterior distribution since it is just the mode, i.e. the maximiser of the posterior density. This relation to the MLE can be seen in Figure IV.1. Hence the Bayesian approach is a a more powerful tool than MLE and allows also to capture the random uncertainty of the parameter θ and we have seen that the mode is not always a vey typical outcome of a random variable.

cite

A second advantage over the MLE presented in the third chapter is, that it might be possible to computationally approximate the posterior density but not the MLE. This is typically the case if the log likelihood function is not concave, like in the setting of the MLE of the whole elementary kernel L. In fact only hard step in the calculation of the posterior (4.1) is the computation of the normalisation constant

$$\int f_{X,\Theta}(x,\theta)\nu(\mathrm{d}\theta).$$

This step can actually also not be performed efficiently for the case of the estimation of L, however we will introduce the methods of Markov chain Monte Carlo simulation that allow an approximation of the posterior without the need to compute the normalisation contant.

4.4 Regularisation through the prior.

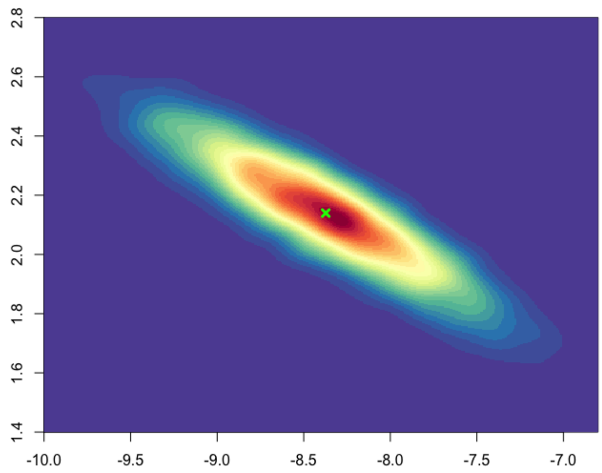

Figure IV.1.: Approximated posterior density of the two dimensional log linearity constant of a two dimensional DPP. The MLE estimator is marked green.

IV.2 Markov chain Monte Carlo methods

IV.2.1 Reminder on Markov chains

(i) Definition

(ii) irreducibility

(iii) existence of stationary distributions

(iv) reversibility

(v) detail-balance

(vi) Ergodicity

(vii) idea of MCMC

Irreducibility and existence of stationary distributions

Ergodicity

Idea of Markov chain Monte Carlo methods

IV.2.2 Metropolis-Hastings random walk

Presentation of the model

Acceptance rate, effective sample size and Tuning

IV.2.3 Slice sampling

Presentation of the model

Realisation of the model

IV.3 The variational approach

Chapter V

Toy examples and experiments

V.1 Minimal example?

V.2 Points on the line

The first example we present is a selection of points on a (discretised) line. More precisely we will assume that we have 100 points on a line that are equally spaced and we aim to model a spacial repulsion between the selected points. For this we will use the method 2.7 of reference points the diversity features. In this case we will use the set $\mathcal{Y}$ itself as reference set and use a

5.1 Setup of the example. Let $\mathcal{Y} := \{1, \dots, 100\}$ and for $i \in \mathcal{Y}$. Then we will let $\phi_i \in \mathbb{R}^{100}$ be given up to scaling by

$$(\phi_i)_j \propto f\left(\frac{|i-j|}{???}\right)$$

where f is the density of the standard nomal distribution. Further we choose the qualities to be constant and so that the expected cardinatlity is 10.

check

5.2 Remark. (*i*) describe scaling including choice of cardinality

(*ii*) describe rank of the kernel?

(*iii*) describe choice of ‚repulsiveness', plot density around a point; make comment to kernel methods? comment on the qualitative properties of f and why they are suitable here

To make the difference to an uncorrelated point pattern more apparent we also defined a Poisson process, i.e. a DPP without correlations between the points with the same expected cardinality. The sampling results are compared in Figure V.2.

make comment on zeta function!

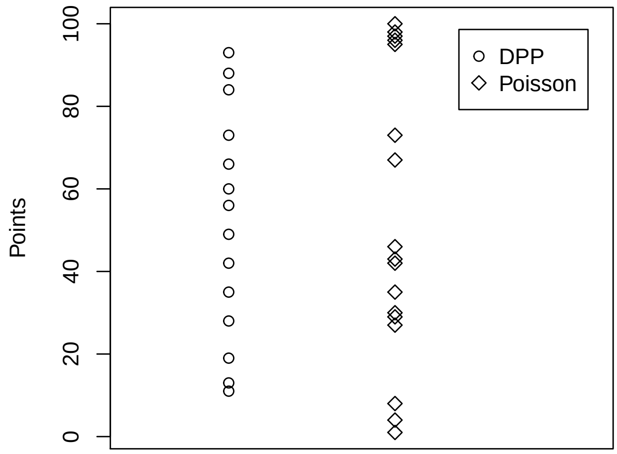

Figure V.1.: Comparison of a DPP with negative correlations on the left and no correlations, i.e. a Poisson point process on the right.

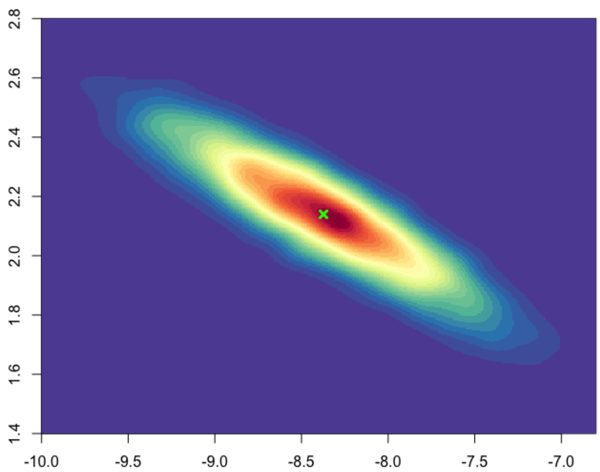

Figure V.2.: Comparison of a DPP with negative correlations on the left and no correlations, i.e. a Poisson point process on the right.

5.3 Representation as binary sequence.

comment on problems and findings?

V.3 Points in the square

V.4 Toy example for quality learning

Chapter VI

Summary and conclusion

Chapter A

Calculations

(*i*) Complement of DPPs.

(*ii*) Normalisation constant of L-ensembles.

(*iii*) Cantor's intersection theorem for Hausdorff spaces.

(*iv*)

•

Chapter B

Generated code

All my coding was done in R and I will provide the code for sampling, my examples and also the learning algorithm of my toy example here. During my coding I mostly followesd Google's R Style Guide (https://google.github.io/styleguide/Rguide.xml).

B.1 Sampling algorithm

```
# Implementation of the sampling algorithm as a function

SamplingDPP <- function (lambda, eigenvectors) {
  # First part of the algorithm, doing the selection of the eigenvectors
  N = length(lambda)
  J <- runif(N) <= lambda/(1 + lambda)
  k <- sum(J)
  V <- matrix(eigenvectors[, J], nrow=N)
  Y <- rep(0, k)

  # Second part of the algorithm, the big while loop
  while (k > 0) {
    # Calculating the weights and selecting an item i according to them
    wghts <- k^(-1) * rowSums(V^2)
    i <- sample(N, 1, prob=wghts)
    Y[k] <- i
    if (k == 1) break

    # Projecting e_i onto the span of V
    help <- V %*% V[i, ]
    help <- sum(help^2)^(-1/2) * help

    # Projecting the elements of V onto the subspace orthogonal to help
    V <- V - help %*% t(t(V) %*% help)

    # Orthonormalize V and set near zero entries to zero
    V[abs(V) < 10^(-9)] <- 0
    j <- 1
    while(j <= k) {
      help2 <- rep(0, N)
     m <- 1
        while (m <= j - 1) {
        help2 <- help2 + sum(V[, j] * V[, m]) * V[, m]
```

```
        m <- m + 1
      }
      V[, j] <- V[, j] - help2
      if (sum(V[, j]^2) > 0) {
        V[, j] <- sum(V[, j]^2)^(-1/2) * V[, j]
      }
      j <- j + 1
    }
    V[abs(V) < 10^(-9)] <- 0

    # Selecting a linear independent set in V
    k <- k - 1
    q <- qr(V)
    V <- matrix(V[, q$pivot[seq(k)]], ncol=k)
  }
  return(Y)
}
```

B.2 Points on the line

```
# NEEDS: sampling algorithm

# In this example we sample points on a (discrete) line according to a DPP
# We model L directly and via the quality-diversity decomposition. We plot and
# compare the patterns to uncorrelated points i.e. to a Poisson point process.

# Minimal example ______________________________________________________________
n <- 3
L <- matrix(c(2,1,0,1,2,0,0,0,2), nrow=n)

# Points on a line _____________________________________________________________
n <- 100
L <- rep(0, n^2)
for (i in 1:n) {
  for (j in 1:n) {
    L[(i - 1) * n + j] <- dnorm((i-j) * n^(-1/4))
  }
}
L <- matrix(L, nrow=n)

# Modelling phi and q __________________________________________________________
# Points on the line.
m <- 99  # 29
n <- m + 1
q <- rep(10, n)  # 0-1 sequences: rep(10^2, n)
phi <- rep(0, n^2)
for (i in 1:n) {
  for (j in 1:n) {
    phi[(i - 1) * n + j] <- dnorm((i - j) / 10)  # 0-1 sequences: devide by 2
  }
}
phi <- matrix(phi, ncol=n)

# Log linear quality for the points on the line ________________________________
m <- 99
n <- m + 1
```

```
q <- rep(0, n)
for (i in 1:n) {
  q[i] <- 10^2 * sqrt(m) * exp(-0.2 * abs(i - 50.5))
}
phi <- rep(0, n^2)
for (i in 1:n) {
  for (j in 1:n) {
    phi[(i - 1) * n + j] <- dnorm(2 * (i - j) / sqrt(m))
  }
}
phi <- matrix(phi, ncol=n)

# General part, define L ____________________________________________________
for (i in 1:n) {
  phi[, i] <- sum(phi[, i]^2)^(-1/2) * phi[, i]
}
S <- t(phi) %*% phi
time <- proc.time()
L <- t(q * S) * q
proc.time() - time

# Compute the eigendecomposition, set near zero eigenvalues to zero and
# set up poisson point process with same expected cardinality _______________
time <- proc.time()
edc <- eigen(L)
lambda <- edc$values
lambda[lambda < 10^(-9)] <- 0
mean <- sum(lambda / (1 + lambda))
eigenvectors <- edc$vectors
lambda2 <- rep(mean / n / (1 - mean / n), n)
eigenvectors2 <- diag(rep(1, n))
proc.time() - time

# Sample and plot things ____________________________________________________
# Minimal example

# 0-1 sequences
x <- sort(SamplingDPP(lambda, eigenvectors))
as.integer(1:n %in% x)
y <- sort(SamplingDPP(lambda2, eigenvectors2))
as.integer(1:n %in% y)

# Sample from both point processes and plot the points on the line
pointsDPP <- SamplingDPP(lambda, eigenvectors)
pointsPoisson <- SamplingDPP(lambda2, eigenvectors2)
plot(rep(1, length(pointsDPP)), pointsDPP,
     ylim=c(1, n), xlim=c(.4, 3.2), xaxt='n', ylab="Points", xlab="")
points(rep(2, length(pointsPoisson)), pointsPoisson, pch=5)
legend("topright", inset=.05, legend=c("DPP", "Poisson"), pch=c(1, 5))

# Remove all objects apart from functions
rm(list = setdiff(ls(), lsf.str()))
```

B.3 Points in the square

```
# NEEDS: sampling algorithm
```

```
# In this example we sample points on a two dimensional grid according to a DPP
# We model L directly and via the quality-diversity decomposition including
# different dimensions D for the feature vectors phi. We plot and compare the
# patterns to uncorrelated points i.e. to a Poisson point process.

# Define the coordinates of a point ___________________________________________
CoordinatesNew <- function(i, n) {
  y1 <- floor((i - 1) / (n + 1))
  x1 <- i - 1 - (n + 1) * y1
  return (t(matrix(c(x1, y1)/n, nrow=length(i))))
}
DistanceNew <- function (i, j, n, d) {
  return (sqrt(colSums((CoordinatesNew(i, n) - CoordinatesNew(j, d))^2)))
}

# Direct modelling of L ________________________________________________________
m <- 19
n <- (m + 1)^2
L <- rep(0, n^2)
for (i in 1:n) {
  for (j in 1:n) {
    L[(i - 1) * n + j] = n^2 * dnorm(Distance(i, j, m))
  }
}
L <- matrix(L, nrow=n)

# Modelling phi and q __________________________________________________________
# Points in the square.
m <- 19
n <- (m + 1)^2
q <- rep(sqrt(m), n)
x <- ceiling(1:n^2 / n)
y <- rep(1:n, n)
time <- proc.time()
phi <- dnorm(sqrt(m) *matrix(DistanceNew(x, y, m, m), n))
proc.time() - time

# Quality diversity decomposition with small D _________________________________
d <- 25
q <- rep(10^5 * sqrt(m), n)
x <- ceiling(1:(n*d) / d)
y <- rep(1:d, n)
time <- proc.time()
phi <- dnorm(2 * sqrt(m) * matrix(DistanceNew(x, y, m, sqrt(d) - 1), ncol=n))
proc.time() - time

# Log linear quality for the points in the square ______________________________
m <- 39
n <- (m + 1)^2
q <- exp(-6 * DistanceNew(rep(5, n), 1:n, 2, m) + log(sqrt(m)))
x <- ceiling(1:n^2 / n)
y <- rep(1:n, n)
time <- proc.time()
phi <- dnorm(2 * sqrt(m) * matrix(DistanceNew(x, y, m, m), n))
proc.time() - time
```

```
# General part, defining L ____________________________________________________
# d <- length(phi) / n
for (i in 1:n) {
  phi[, i] <- sum(phi[, i]^2)^(-1/2) * phi[, i]
}
S <- t(phi) %*% phi
# B <- t(phi) * q
time <- proc.time()
L <- t(t(q * S) * q)  # B %*% t(B)
proc.time() - time

# Compute the eigendecomposition, set near zero eigenvalues to zero and
# set up poisson point process with same expected cardinality _________________
time <- proc.time()
edc <- eigen(L)
lambda <- edc$values
lambda[abs(lambda) < 10^(-9)] <- 0
mean <- sum(lambda / (1 + lambda))
eigenvectors <- edc$vectors
lambda2 <- rep(mean / n / (1 - mean / n), n)
eigenvectors2 <- diag(rep(1, n))
proc.time() - time

# Sample from both point processes and plot the points in the square __________
# par(mfrow = c(1,1))
time <- proc.time()
dataDPP <- sort(SamplingDPP(lambda, eigenvectors))
pointsDPP <- t(CoordinatesNew(dataDPP, m))
plot(pointsDPP, xlim=0:1, ylim=0:1, xlab="", ylab="", xaxt='n', yaxt='n', asp=1)
proc.time() - time
dataPoisson <- sort(SamplingDPP(lambda2, eigenvectors2))
pointsPoisson <- t(CoordinatesNew(dataPoisson, m))
plot(pointsPoisson, xlim=0:1, ylim=0:1, xlab="", ylab="",
                                        xaxt='n', yaxt='n', asp=1)

# Remove all objects apart from functions
rm(list = setdiff(ls(), lsf.str()))
```

B.4 Toy learning example

```
# NEEDS: Sampling algorithm, declaration of the points in the square
# TODO: Maybe do the gradient descent directly over the representation
# od the gradient

# With this toy example we aim to perform the first learning of paramters
# associated to a kernel of a DPP. More precisely we will generate our own
# data of points on a two dimensional grid with a log linear quality model
# and aim to estimate the log linearity parameter.

# Generation of data
time <- proc.time()
T <- 30
data <- rep(list(0), T)
for (i in 1:T) {
  data[[i]] <- sort(SamplingDPP(lambda, eigenvectors))
}
proc.time() - time
```

```r
# Define the quality q, L, the feature sum and the loss in dependency of the
# parameter theta
Quality <- function(theta) {
  return(exp(theta[1] * DistanceNew(rep(5, n), 1:n, 2, m) + theta[2]))
}
LFunction <- function(theta) {
  return(t(t(Quality(theta) * S) * Quality(theta)))
}
Feature <- function(A) {
  # return(sum(DistanceNew(rep(5, length(A)), A, 2, m)))
  return(c(sum(DistanceNew(rep(5, length(A)), A, 2, m)), length(A)))
}
Loss <- function(theta) {
  T <- length(data)
  # Sum this over all data entries
  x <- 0
  for (i in 1:T) {
    A <- data[[i]]
    x <- x + 2 * sum(theta * Feature(A)) + log(det(matrix(S[A, A], length(A))))
  }
  return(- x + T * log(det(diag(rep(1, n)) + LFunction(theta))))
}

# Parameter estimations
time <- proc.time()
sol <- nlm(Loss, c(-3, 0))
proc.time() - time
sol$estimate

# Remove all objects apart from functions
rm(list = setdiff(ls(), lsf.str()))
```

Bibliography

[Affandi et al., 2014a] Affandi, R. H., Fox, E., Adams, R., and Taskar, B. (2014a). Learning the parameters of determinantal point process kernels. In *International Conference on Machine Learning*, pages 1224–1232.

[Affandi et al., 2014b] Affandi, R. H., Fox, E., Adams, R., and Taskar, B. (2014b). Learning the parameters of determinantal point process kernels. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1224–1232, Bejing, China. PMLR.

[Benard and Macchi, 1973] Benard, C. and Macchi, O. (1973). Detection and"emission"processes of quantum particles in a"chaotic state". *Journal of Mathematical Physics*, 14(2):155–167.

[Bondy and Murty, 2011] Bondy, A. and Murty, U. (2011). *Graph Theory*. Graduate Texts in Mathematics. Springer London.

[Borodin, 2009] Borodin, A. (2009). Determinantal point processes. *arXiv preprint arXiv:0911.1153*.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[Gray, 1990] Gray, R. M. (1990). Entropy and information. In *Entropy and information theory*, pages 21–55. Springer.

[Griffin and Tsatsomeros, 2006] Griffin, K. and Tsatsomeros, M. J. (2006). Principal minors, part ii: The principal minor assignment problem. *Linear Algebra and its applications*, 419(1):125–171.

[Higham, 1990] Higham, N. J. (1990). Exploiting fast matrix multiplication within the level 3 blas. *ACM Transactions on Mathematical Software (TOMS)*, 16(4):352–368.

[Hough et al., 2006] Hough, J. B., Krishnapur, M., Peres, Y., Virág, B., et al. (2006). Determinantal processes and independence. *Probability surveys*, 3:206–229.

[Kulesza and Taskar, 2010] Kulesza, A. and Taskar, B. (2010). Structured determinantal point processes. In *Advances in neural information processing systems*, pages 1171–1179.

[Kulesza et al., 2012] Kulesza, A., Taskar, B., et al. (2012). Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286.

[Lehmann and Casella, 2006] Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

[MacKay and Mac Kay, 2003] MacKay, D. J. and Mac Kay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

[Magen and Zouzias, 2008] Magen, A. and Zouzias, A. (2008). Near optimal dimensionality reductions that preserve volumes. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 523–534. Springer.

[Martin and England, 2011] Martin, N. F. and England, J. W. (2011). *Mathematical theory of entropy*, volume 12. Cambridge university press.

[Neal, 2003] Neal, R. M. (2003). Slice sampling. *Annals of statistics*, pages 705–741.

[Newey and McFadden, 1994] Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

[Newton and Halley, 1744] Newton, I. and Halley, E. (1744). *Philosophiae naturalis principia mathematica*, volume 62. Jussu Societatis Regiae ac typis Josephi Streater, prostant venales apud Sam. Smith.

[Rice, 2006] Rice, J. (2006). *Mathematical statistics and data analysis*. Nelson Education.

[Rising et al., 2015] Rising, J., Kulesza, A., and Taskar, B. (2015). An efficient algorithm for the symmetric principal minor assignment problem. *Linear Algebra and its Applications*, 473:126–144.

[Robert and Casella, 1999] Robert, C. P. and Casella, G. (1999). The metropolis-hastings algorithm. In *Monte Carlo Statistical Methods*, pages 231–283. Springer.

[Samuel, 1959] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229.

[Tauchen, 1985] Tauchen, G. (1985). Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics*, 30(1-2):415–443.

[Urschel et al., 2017] Urschel, J., Brunel, V.-E., Moitra, A., and Rigollet, P. (2017). Learning determinantal point processes with moments and cycles. *arXiv preprint arXiv:1703.00539*.

[Veblen, 1912] Veblen, O. (1912). An application of modular equations in analysis situs. *Annals of Mathematics*, 14(1/4):86–94.

[Volkenstein, 2009] Volkenstein, M. V. (2009). *Entropy and information*, volume 57. Springer Science & Business Media.