Proceedings of Machine Learning Research vol 65:1–3, 2017

Rates of estimation for determinantal point processes

Victor-Emmanuel Brunel VEBRUNEL@MIT.EDU
Massachusetts Institute of Technology, Department of Mathematics

Ankur Moitra MOITRA@MIT.EDU
Massachusetts Institute of Technology, Department of Mathematics

Philippe Rigollet RIGOLLET@MIT.EDU
Massachusetts Institute of Technology, Department of Mathematics

John Urschel URSCHEL@MIT.EDU
Massachusetts Institute of Technology, Department of Mathematics

Keywords: Determinantal point processes, statistical estimation, maximum likelihood, L-ensembles.

1. Introduction

Determinantal point processes (DPPs) describe a family of repulsive point processes; they induce probability distributions that favor configurations of points that are far away from each other. They have played a central role in various corners of probability, algebra, combinatorics, and machine learning (following the seminal work of Kulesza and Taskar (2012)), where their repulsive character has been used to enforce the notion of diversity in subset selection problems.

Even though many applications necessitate estimation of the parameters of a DPP, statistical inference for DPPs has received little attention. In this context, maximum likelihood estimation is a natural method, but generally leads to a non-convex optimization problem. This problem has been addressed by various heuristics, including Expectation-Maximization (Gillenwater et al. (2014)), MCMC (Affandi et al. (2014)), and fixed point algorithms (Mariet and Sra (2015)). None of these methods come with global guarantees, however. In this paper, we take an information geometric approach to understand the asymptotic properties of the maximum likelihood estimator for discrete DPPs. First, we study the curvature of the expected log-likelihood around its maximum. Our main result is an exact characterization of when the maximum likelihood estimator converges at a parametric rate. Moreover, we give quantitative bounds on the strong convexity constant that translate into lower bounds on the asymptotic variance. This shed light on what combinatorial parameters of a DPP control the variance.

2. Definitions

A (discrete) *determinantal point process* (DPP) on the finite space $[N] = \{1, 2, \dots, N\}$ is a random set $Z \subseteq [N]$ that satisfies

$$\mathbb{P}[J \subseteq Z] = \det(K_J), \quad \forall\, J \subseteq [N], \tag{2.1}$$

. Extended abstract. Full version appears as arXiv:1706.00961, v1 for some symmetric matrix $K \in \mathbb{R}^{N\times N}$ with all its eigenvalues between 0 and 1. Here, we denote by K_J the submatrix of K obtained from K by keeping the columns and rows indexed by J.

If it holds further that $I - K$ is invertible, then Z is called *L-ensemble* and

$$\mathbb{P}[Z = J] = \frac{\det(L_J)}{\det(I + L)}, \quad \forall\, J \subseteq [N], \tag{2.2}$$

where $L = K(I - K)^{-1}$ is called the *kernel* of Z.

In this work, we only consider DPPs that are L-ensembles. In that setup, we can identify L-ensembles and DPPs, and the kernel L and correlation kernel K are related by the identities

$$L = K(I - K)^{-1}, \qquad K = L(I + L)^{-1}. \tag{2.3}$$

Note that we only consider kernels L that are positive definite. We denote by $\mathsf{DPP}(L)$ the probability distribution associated with the DPP with kernel L and refer to L as the *parameter* of the DPP in the context of statistical estimation.

The probability mass function (2.2) of $\mathsf{DPP}(L)$ depends only on the principal minors of L and on $\det(I + L)$. In particular, L is not fully identified by $\mathsf{DPP}(L)$ and the lack of identifiability of L has been characterized exactly (Kulesza, 2012, Theorem 4.1). Denote by $\mathcal{D}$ the collection of $N \times N$ diagonal matrices with ± 1 diagonal entries. Then, for a symmetric matrix L',

$$\mathsf{DPP}(L') = \mathsf{DPP}(L) \iff \exists D \in \mathcal{D}, L' = DLD. \tag{2.4}$$

Hence, the parameter of a DPP is only indentified up to a flip of the signs of its columns and rows.

A DPP with kernel L is called irreducible whenever L is irreducible, i.e., if L does not have a block diagonal structure. The following graph associated to L naturally describes its block structure.

Definition 1 *The* determinantal graph $\mathcal{G}_L = ([N], E_L)$ *of a DPP with kernel L is the undirected graph with vertices $[N]$ and edge set $E_L = \big\{\{i, j\} : i \neq j, L_{i,j} \neq 0\big\}$. If $i, j \in [N]$ with $i \neq j$, write $i \sim_L j$ if there exists a path in $\mathcal{G}_L$ that connects i and j.*

It is not hard to see that a DPP with kernel L is irreducible if and only if its determinantal graph $\mathcal{G}_L$ is connected. If L is block diagonal, then its blocks correspond to the connected components of $\mathcal{G}_L$. Moreover, it follows directly from (2.2) that if $Z \sim \mathsf{DPP}(L)$ and L has blocks $J_1, \ldots, J_k$, then $Z \cap J_1, \ldots, Z \cap J_k$ are mutually independent DPPs with kernels $L_{J_1}, \ldots, L_{J_k}$ respectively.

3. Information geometry

Our goal is to estimate an unknown kernel L^* from n independent copies of $Z \sim \mathsf{DPP}(L^*)$. In this paper, we study the statistical properties of what is arguably the most natural estimation technique: maximum likelihood estimation.

First, we prove important properties of the Fisher information $\mathcal{I}(L^*)$. Our main result is the following.

Theorem 2

The nullspace of $\mathcal{I}(L^*)$ is given by

$$\big\{H \in \mathbb{R}^{N\times N} : H_{i,j} = 0 \text{ for all } i, j \in [N] \text{ such that } i \sim_{L^*} j\big\}\ . \tag{3.1}$$

In particular, $\mathcal{I}(L^*)$ is positive definite if and only if L^* is irreducible.

Theorem 2 is a qualitative result. In particular, we provide examples of irreducible kernels L^* for which $\mathcal{I}(L^*)$ have eigenvalues that are exponentially small in N.

4. Maximum likelihood estimation

Let $Z_1, \ldots, Z_n$ be n independent copies of $Z \sim \mathsf{DPP}(L^*)$, where the kernel L^* is unknown. We assume that L^* is positive definite. Let $\hat{L}$ be the maximum likelihood estimator (*MLE*) of L^* (unique up to a flip of the sign of its columns and its rows).

We measure the performance of the MLE using the *loss* ℓ defined by

$$\ell(\hat{L}, L^*) = \min_{D \in \mathcal{D}} \|\hat{L} - DL^*D\|_F$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Our statistical results establish asymptotic properties of the MLE. We use $O_{\mathbb{P}}$ for big-O notation in probability. For $L \in \mathbb{R}^{N \times N}$ and $J, J' \subseteq [N]$, we denote by $L_{J,J'}$ the submatrix of L obtained by keeping the rows indexed in J and the columns indexed in J'.

Theorem 3

- $\ell(\hat{L}, L^*) \xrightarrow[n \to \infty]{} 0\,,$ *in probability.*
- *If* L^* *is irreducible, then,* $\tilde{L}$ *is asymptotically normal. In particular,* $\ell(\hat{L}, L^*) = O_{\mathbb{P}}(n^{-1/2})$.
- *If* L^* *is block diagonal, then, for any pair of distinct blocks* J, J' *of* L^**,*

$$\min_{D \in \mathcal{D}} \|\hat{L}_{J,J'} - DL^*_{J,J'}D\|_F = O_{\mathbb{P}}(n^{-1/4}) \tag{4.1}$$

and

$$\min_{D \in \mathcal{D}} \|\hat{L}_J - D_J L^*_J D_J\|_F = O_{\mathbb{P}}(n^{-1/2}). \tag{4.2}$$

References

Raja Hafiz Affandi, Emily B. Fox, Ryan P. Adams, and Benjamin Taskar. Learning the parameters of determinantal point process kernels. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1224–1232, 2014.

Jennifer Gillenwater, Alex Kulesza, Emily Fox, and Ben Taskar. Expectation-maximization for learning determinantal point processes. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 3149–3157, Cambridge, MA, USA, 2014. MIT Press.

A. Kulesza. *Learning with determinantal point processes*. PhD thesis, University of Pennsylvania, 2012.

Alex Kulesza and Ben Taskar. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc., Hanover, MA, USA, 2012. ISBN 1601986289, 9781601986283.

Zelda Mariet and Suvrit Sra. Fixed-point algorithms for learning determinantal point processes. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2389–2397, 2015.