Determinantal Point Processes and β-ensembles

Shawn Im

December 2020

1 Introduction

This report will explore the spacing of points when sampled from a determinantal point process on $n \times n$ grids as well as determinantal point processes with L-ensembles defined by random β-matrices.

2 Generalized β-ensembles

The commonly seem random matrix ensembles usually have a β value of 1, 2, or 4. For Hermitian ensembles, this would be the GOE, GUE, and the GSE respectively. However, we can define random matrix ensembles with any $\beta > 0$ and sample a random matrix from such an ensemble using tridiagonal matrices.

Hermite Matrix $n \in \mathbb{N}$	$H_\beta \sim \frac{1}{\sqrt{2}} \begin{pmatrix} N(0,2) & \chi_{(n-1)\beta} & & & \\ \chi_{(n-1)\beta} & N(0,2) & \chi_{(n-2)\beta} & & \\ & \ddots & \ddots & \ddots & \\ & & \chi_{2\beta} & N(0,2) & \chi_\beta \\ & & & \chi_\beta & N(0,2) \end{pmatrix}$
Laguerre Matrix $m \in \mathbb{N}$ $a \in \mathbb{R}$ $a > \frac{\beta}{2}(m-1)$	$L_\beta = B_\beta B_\beta^T$, where $B_\beta \sim \begin{pmatrix} \chi_{2a} & & & \\ \chi_{\beta(m-1)} & \chi_{2a-\beta} & & \\ & \ddots & \ddots & \\ & & \chi_\beta & \chi_{2a-\beta(m-1)} \end{pmatrix}$

Figure 1: Tridiagonal matrices for β-ensembles [1]

One element of β-ensembles is the distribution of their eigenvalue spacings, the differences between two consecutive eigenvalues when sorted, and this varies with β. Then, given a set of eigenvalues, we can estimate the β-ensemble that it was most likely sampled from, and this has been implemented by Cy Chan.

The beta estimator trains on eigenvalue spacings, normalized to have mean 1, of 10000 by 10000 matrices sampled from β-ensembles with β ranging from 0 to 10 in increments of 0.1. Then, for a set of eigenvalue spacings, it estimates the likelihood of the set coming from each β-ensemble and chooses the β with the highest likelihood [2].

We have implemented this beta estimator in Julia and will be using it in this report to analyze and compare distributions of spacings resulting from determinantal point processes.

```
A = randn(2000, 2000)
A = (A+A')/2
e = eigvals(A)
diffe = [e[i]-e[i-1] for i in 2:2000]
b, pl = beta_est(diffe)
display(b)
plot(pl)
```

```
1.1
```

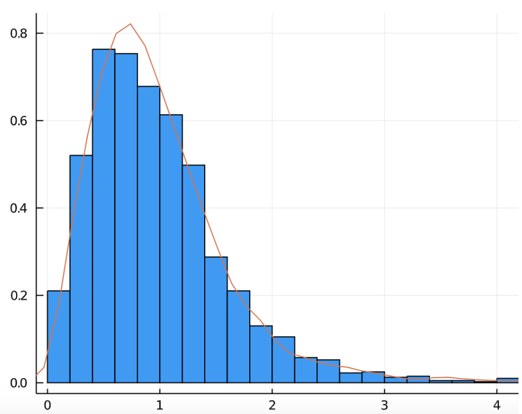

Figure 2: Beta Estimator implementation

The figure above shows our implementation of the beta estimator giving a reasonable approximation of $\beta = 1.1$ for a set of eigenvalue spacings sampled from the GOE ($\beta = 1$).

3 Determinantal Point Processes

3.1 Definitions

A point process $\mathcal{P}$ on a set $\mathcal{Y}$ is a probability measure over the subsets of $\mathcal{Y}$, and a point process is determinantal if Y is a random subset of $\mathcal{Y}$ chosen according to $\mathcal{P}$ and for every subset A of $\mathcal{Y}$

$$\mathcal{P}(A \subseteq Y) = \det(K_A)$$

where K is a positive semidefinite matrix indexed by the elements of $\mathcal{Y}$ with eigenvalues less than or equal to 1 and K_A is the submatrix of K with elements indexed by the elements of A. From this, we can see that $\det(K_A)$ acts a cumulative distribution function, and so K is called the marginal kernel.

An alternative to defining a DPP with the marginal kernel is defining a DPP through a real positive semidefinite matrix L indexed by the elements of $\mathcal{Y}$ which defines an L-ensemble. With L, we have

$$\mathcal{P}(A = Y) = \frac{\det(L_A)}{\det(L + I)}$$

and given a DPP defined by L, we can find its marginal kernel

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

In a general DPP, we can choose any subset of $\mathcal{Y}$ according to $\mathcal{P}$, but if we choose to condition the DPP on the event that the subset chosen has size k, we have a k-DPP.

One way of sampling points from a DPP or k-DPP is through an algorithm that takes the eigendecomposition of L as its input, and we will be using an implementation of such an algorithm from DeterminantalPointProcesses.jl.

3.2 Applications to Machine Learning

When sampling from a DPP over a set of points, choosing one point will reduce the chances of including other points and this negative correlation between points is stronger the closer or more similar the points are. This allows for a higher change of choosing a diverse set of points.

This negative correlation of DPPs can help cover or summarize a set of data such as associated topics for a query. It can be used to model repulsion such as the location of trees in a forest, and it can be used for filtering out similar or nearby data.

An example of a DPP being used to improve machine learning is with pose estimation. In [3], they explore a multiple pose estimation task and using a DPP to clean up clusters of predictions and create bias towards non-overlapping predictions.

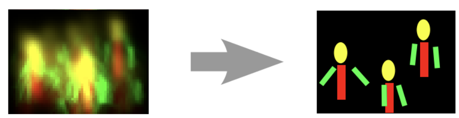

Figure 3: Pose Estimation with DPP [3]

4 Experiments and Results

4.1 k-DPPs in the plane

One simple example of DPPs covering a more diverse set of points can be seen when sampling points from a grid.

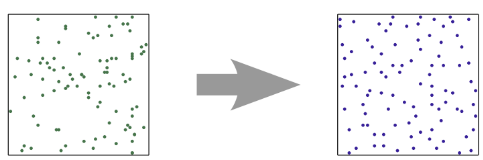

Figure 4: Independent samples, DPP on grid [3]

The figure above shows independently sampled points on the left and points sampled according to a DPP with a marginal kernel matrix K with K_{ij} inversely proportional to the distance between points i and j.

This leads to a question of how to measure the diversity of a set of points or the distribution of spacings between the points as well as how the spacings change as the marginal kernel changes. In order to begin exploring this question, we will define a way of measuring how the points are spaced, and we will start with a small class of k-DPPs.

For eigenvalues, the spacings are defined as the difference between two consecutive eigenvalues. Similarly, we will define the spacings for the sampled points to be the distance between two consecutive points, and two points are consecutive if they share an edge in the minimum spanning tree with edge weights being distance.

We will look at how the distribution of spacings between points change as we use L^p norms to define distance. In order to do so, we need to define the kernel that we will be using and we will use

$$K_{ij} = \frac{1}{1+ \parallel s_i - s_j \parallel_p}$$

where s_i and s_j are the coordinates for the corresponding points.

Now, we can compare the distributions of spacings between the sampled points as we change the L^p norm to use, and we will do so by using the Beta estimator.

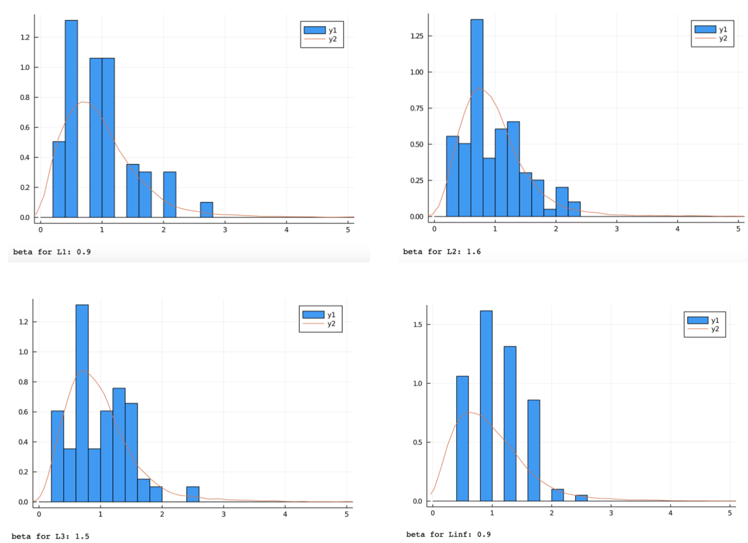

Figure 5: Distributions of Spacings for L^1, L^2, L^3, L^∞

Above, we can the distributions of spacings for the L^1, L^2, L^3, L^∞ norms and the β value for the distributions from the Beta estimator. The results come from sampling 100 points from a 40 by 40 grid. It seems that the distribution of point spacings actually resembles the distributions of the β-ensemble spacings, and changing the L^p norm changes the value of β it is closest to.

The plot below shows how the estimate for β changes as we go from L^1 to $L^1$0 for different grid sizes and a different number of samples, and in both cases, the β value seems to be greatest around $p = 2$ to $p = 3$, but it is difficult to draw any conclusions as to why this is the case simply from this data.

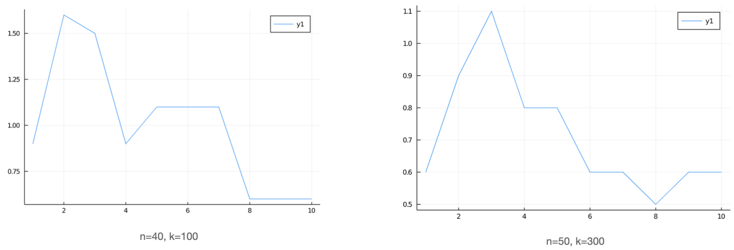

Figure 6: β estimate for L^p norms

One other thing we experiment with is how β changes as the ratio of samples to total points changes.

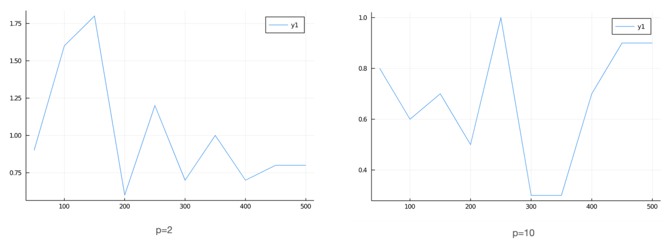

Figure 7: β estimate for sampling ratios

In the figure, we show how β changes as the number of samples increases from 50 to 500 out of 1600 points, and on the left the L^p norm used is L^2 and on the right is L^10. For L^2, as the ratio increased, the β seemed to be decreasing and converging to a value, but for L^10, the β seems to change with no pattern. It seems that again the $p = 2$ to $p = 3$ range has properties not seen at other values of p.

4.2 Randomized L-ensembles

Another question regarding the distribution of spacings of points sampled from a DPP is how it changes as L changes. If we let $L = HH^T$ where H is a random β-Hermitian matrix, and we sample k points from integers 1 through n and take the difference between consecutive points, we can see how the distribution of these differences change as β changes.

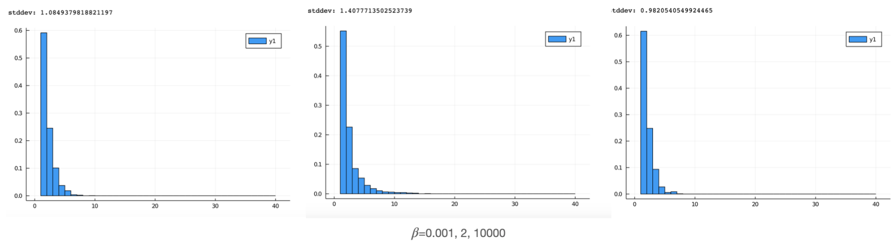

Figure 8: Distribution of spacings over different β

The figure above shows the resulting distributions for $\beta = 0.001, 2, 1000$ from left to right, and we can see they have a similar shape. Rather than using the β estimator to compare these as they do not resemble the distributions from the β estimator and all of them would give β near 0, we can compare them using standard deviation. We can see that the standard deviation is lower for β near 0 and large β and greater around $\beta = 2$. This is likely due to the transition in the distribution of eigenvalues of the β-ensembles as near 0, the distribution is near normal while for large β, the distribution gets closer to a semicircle.

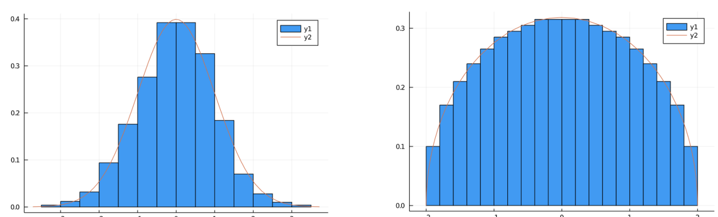

Figure 9: Eigenvalue distribution for small and large β

It seems that using a randomized L with $L = HH^T$ does little to diversify the sampled points as most of the spacings seem to be close to 0 over all β, and generating random L that have larger spacings and diversify the sampled set would require a different method.

References

[1] Ioana Dumitriu and Alan Edelman. Matrix models for beta ensembles. *Journal of Mathematical Physics*, 43(11):5830–5847, Nov 2002.

[2] Cy Chan. Cy's beta estimator. http://people.csail.mit.edu/cychan/BetaEstimator.html.

[3] Alex Kulesza. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2-3):123–286, 2012.