# Determinantal Point Processes as Priors in Bayesian Inference

*Fred Boehm*

*February 11, 2016*

1. Scientific motivation & scientific context

We present a family of stochastic processes, which are known as *determinantal point processes*, for use as priors in Bayesian inference. We focus on their uses in two distinct scientific settings:

1. Using DNA-sequence data to identify subclones within tumors, and
2. Using text data to identify topics among text corpora.

Other researchers have studied both of these questions using Bayesian nonparametric methods. In such studies, the Bayesian prior distributions are specified to be Dirichlet process mixtures. The number of components may be known *a priori* or may be learned from the data. The components are topics (in the case of topic modeling) and subclones (in the setting of tumor heterogeneity).

a. TH, cancer therapy define TH Why does it, TH, matter? What have others done to study TH? What do I propose to do? Why is my approach good? Novel?

b. Topic Modeling (with different priors)

2. Statistical approaches

- DPP

To understand DPPs, it helps to recognize that they are a subclass of *point processes*. Point processes can be viewed as collections of probability distributions on finite subsets of the integers. [check this definition]

The feature that distinguishes DPP from other point processes is that one can describe a DPP with a finite, symmetric, and positive-definite matrix C. Each row (and, similarly, each column) is indexed by a subset of the *ground set*. The number of rows *is* the number of (nonempty??) subsets of the finite ground set. The $(i, j)$th entry of the matrix C is the correlation between sets $i$ and $j$.

A simple example may help to illustrate the $C$ matrix. Consider a ground set that consists of two elements, 0 and 1. Counting the empty set, there are 4 subsets of this ground set:

1. $\emptyset$
2. $\{1\}$
3. $\{2\}$
4. $\{1, 2\}$

We can thus write a 4 x 4 matrix C:

$$\begin{pmatrix} 1 & b & c & d \\ b & 1 & f & g \\ c & f & 1 & h \\ d & g & h & 1 \end{pmatrix}$$

- Dirichlet process mixture

3. Details of DPP prior in statistical modeling

What is the model?
What are assumptions of model?

4. Using the DPP as prior in Cancer genomics
5. Using the DPP as prior in Topic Modeling

```
git2r::summary(git2r::repository("."))
```

```
## Local:    master /Users/fjboehm/Box Sync/Documents/wisconsin-stat-grad-school/spring2016/prelim/
## Remote:   master @ origin (https://github.com/fboehm/prelim.git)
## Head:     [85fd93a] 2016-02-11: format edits to glossary.Rmd
##
## Branches:         1
## Tags:             0
## Commits:         10
## Contributors:     2
## Stashes:          0
## Ignored files:    8
## Untracked files:  0
## Unstaged files:   2
## Staged files:     0
##
## Latest commits:
## [85fd93a] 2016-02-11: format edits to glossary.Rmd
## [8f2fc99] 2016-02-11: minor edits
## [9d08efa] 2016-02-11: minor edits to .gitignore
## [41d52dc] 2016-02-11: added glossary.Rmd
## [748b826] 2016-02-11: added git repo information to manuscript pdf
```