Getting Started with DPP

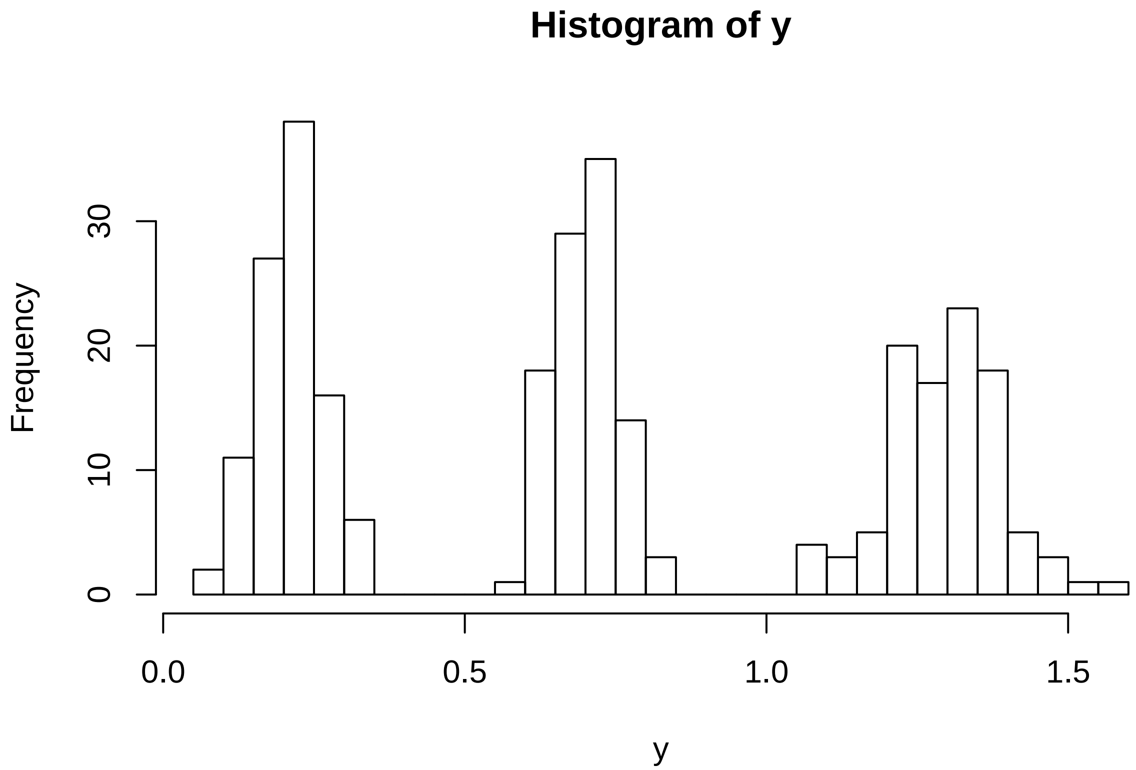
Introduction

DPP can be used to infer the number of categories or clusters in a one dimensional numeric vector. From a potentially infinite number of normal distributions the MCMC algorithm will try to find the most likely number of normal distributions (k) that describes the data.

Simulating data

For a very simple example, we generate data from three normal distributions

```
set.seed(12345)
y <- c(rnorm(100,mean=0.2,sd=0.05), rnorm(100,0.7,0.05), rnorm(100,1.3,0.1))
hist(y,breaks=30)
```



Setup

We load the DPP library and create a NormalModel object with the initial (prior) parameters for the potentially infinite number of normal distribuitions we will infer from the data.

```
library(DPP)
normal.model<-new(NormalModel,
                  mean_prior_mean=0.5,
                  mean_prior_sd=0.1,
                  sd_prior_shape=3,
                  sd_prior_rate=20,
                  estimate_concentration_parameter=TRUE,
                  concentration_parameter_alpha=10,
                  proposal_disturbance_sd=0.1)
```

Creating a dppMCMC_C object and running the MCMC

We setup some additional mcmc parameters and instantiate an object of the class dppMCMC_C. Note that we are passing the previously created NormalModel object as a paramter.

```
my_dpp_analysis <- dppMCMC_C(data=y,
                             output = "output_prefix_",
                             model=normal.model,
                             num_auxiliary_tables=4,
                             expected_k=1.5,
                             power=1)

#this might take a few minutes
my_dpp_analysis$run(generations=1000,auto_stop=TRUE,max_gen = 10000,min_ess = 500)
```

Results

The inferred number of categories/clusters/distributions

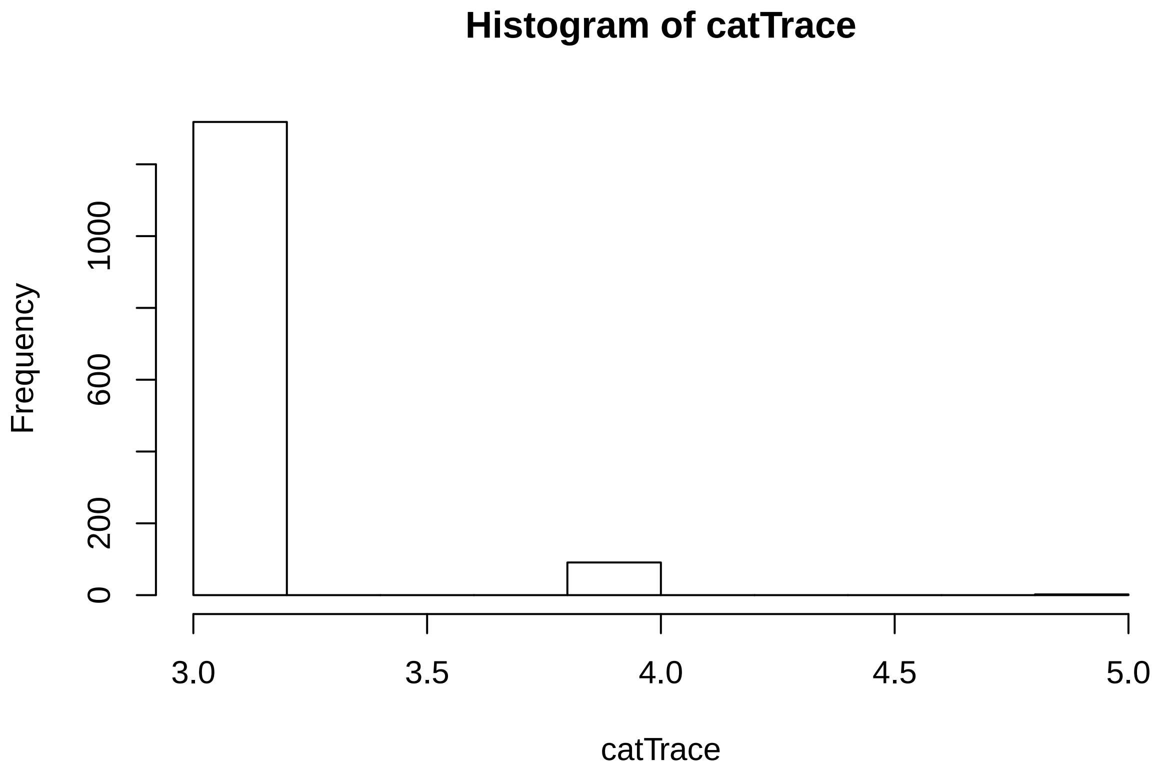
To infer of the number of categories we look at the actual posterior distribution of the parameter k or its MCMC trace .

A histogram of the trace

```
catTrace<-my_dpp_analysis$getNumCategoryTrace(0.25) # we discard the first 25% results
length(catTrace)
```

```
## [1] 1411
```

```
hist(catTrace)
```



The probabilities for k categories

```
category_probabilities<-my_dpp_analysis$getNumCategoryProbabilities(0.25)
category_probabilities
```

```
##           1           2           3           4           5
## 0.000000000 0.000000000 0.934089298 0.064493267 0.001417434
```

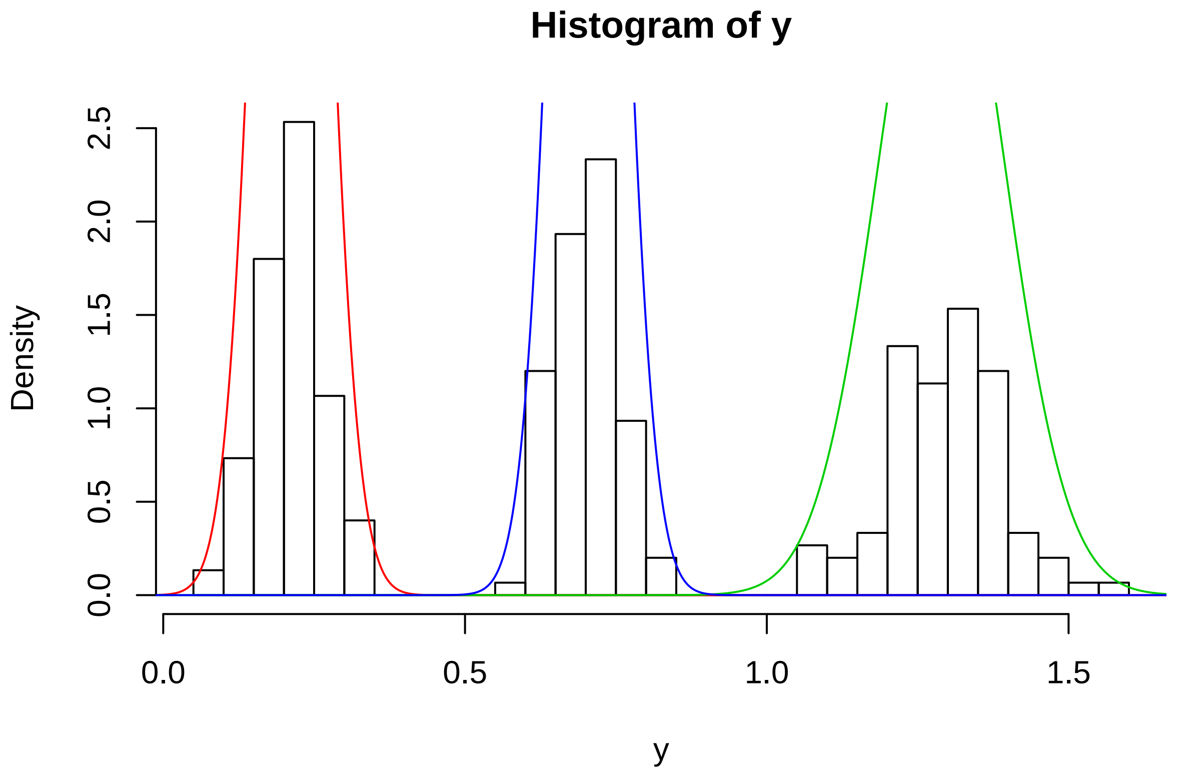
The most likely number of categories

```
topNumCat<-as.numeric(names(which.max(category_probabilities)))
topNumCat
```

```
## [1] 3
```

Plotting the inferred normal distributions

```
hist(y,breaks=30,prob=TRUE)
params<-my_dpp_analysis$dpp_mcmc_object$getParamVector()
for(j in 1:topNumCat) {
    curve(dnorm(x,
                mean=params[[1]][j],
                sd=params[[2]][j]),
                from=-10,
                to=10,
                col=1+j,
                add=TRUE,
                n=20001)
}
```



```
params
```

```
## $means
## [1] 0.2122830 1.2894813 0.7043112
##
## $sds
## [1] 0.05298712 0.10338958 0.05251156
```

And the allocation of the individual elements of the numeric vector as classified in one of the inferred normal distributions

```
allocations<- my_dpp_analysis$dpp_mcmc_object$getAllocationVector()
head(allocations)
```

```
## [1] 1 1 1 1 1 1
```

```
table(allocations)
```

```
## allocations
##   1   2   3
## 100 100 100
```