Bayesian Statistics Exam Exercise

CID: 01278675
Computational Methods in Ecology and Evolution

March 27, 2017

1 Introduction

Aim. In this exercise we want to estimate the divergence time between polar and brown bears (notated as T in our code), in order to familiarize our selves with bayesian inference methods and computation.

The basic step to conduct our analysis are:
-Prepare R Environment
-Define parameters
-Perform Simulations
-Choose Summary Statistics
-Compute ABC
-Get Summary Results and Plots

2 Methods

First, we loaded the packages needed for our analysis.

```
############################## Packages ###################################
# Install:
install.packages(c('coda','abc', 'grid', 'maps', 'spam',
                   'fields', 'HDInterval', 'corrgram'))
# Load:
library(coda)
library(abc)
library(grid)
library(maps)
library(spam)
library(fields)
library(ggplot2)
library(HDInterval)
library(stats)
library(ggplot2)
```

```
library(corrgram)

# Load all Rfunctions and data:
source('functions.R')
load('polar.brown.sfs.Rdata')
```

Then we assigned all the variables needed for our analysis. To obtain our posterior distribution we run 10,000 simulations. The migration parameter was set to 0.

```
############################ Parameters ###################################
# Number of analysed sites
nrSites <- sum(polar.brown.sfs, na.rm = T)
nrSites

# Number of chromosomes
nChroms.polar <- nrow(polar.brown.sfs) -1
nChroms.polar
nChroms.brown <- ncol(polar.brown.sfs) -1
nChroms.brown

# The observed summary statistics
obsSummaryStats <- calcSummaryStats(polar.brown.sfs)

# Define how many simulations we want to perform
srSimul <- 1e4

# Set path to ms software needed to perform our simulations
msDir <- '~/Downloads/msdir/ms'

# and set the name of the output text file
fout <- 'ms.txt'

# Set migration parameter to 0
M <- 0

# Finally assign empty vectors and matrix to store our simulated
# parameteres
SimParamVec <- c() # for our draws from the prior distribution
SimStatsMatrix <- matrix(ncol= 9) # for our simulated statisticd
```

After having assigned all parameters needed, we run a loop, with each circle being one value simulation, and storing everything in the created vectors. For our prior distribution, we chose a discrete bounded uniform distribution, between 200k and 350k years for our T estimates. This interval was chosen after first running simulations for bounds 200k-700k and 200k-500k, but given that the accepted estimates where lower than 400k, we decided to keep a lower upper bound for our final simulations.

```
########################## Simmulations ###################################
# Run my simulations
while (length(SimParamVec) < srSimul)
{
        # while the stored posterior values are less
        # that the desired number of simulations continue

        # 1. Draw from prior (discrete, bounded, uniform)
        sampledT <- runif(1, min = 2e5, max = 3.5e5)
        SimParamVec <- c(SimParamVec, sampledT) # store in vector

        # 2. Simulate observations
        simulate(T=sampledT, M=M, nrSites, msDir, fout)

        # 3. Calculate summary statistics
        simulatedSFS <-fromMStoSFS(fout, nrSites, nChroms.polar,
                        nChroms.brown)
        simStats <- calcSummaryStats(simulatedSFS)

        # Store in matrix
        SimStatsMatrix <- rbind(SimStatsMatrix, simStats)
}

# delete first matrix row cause empty
SimStatsMatrix <-SimStatsMatrix[-1,]
```

After our simulations have been completed, we need to chose the summary simulated statistics we are going to use for our abc. First, we scale our statistics separately, and then run a correlation test. Given that most of them were highly correlated (something that we were expecting given their physical meaning), we need to keep only the most informative. In order to achieve that, we conducted a pca analysis. Based on its output (see Figure 2 in Results session), we decided to keep only the population genetic differentiation (fst), given that explains the biggest proportion of our variance. Additionally, this is the one of highest importance for the question we want to answer.

```
################ Scale simmulated statistics ##########################
# create empty matrix to store new values
ScaledSimStats <-matrix(ncol= 9, nrow = 10000)

# run a scaling for each statistic separately (for each column)
for (i in 1:length(SimStatsMatrix[1,]))
{
ScaledSimStats[,i] <- scale(SimStatsMatrix[,i], center = TRUE,
                            scale = TRUE)
}
```

```
# add summary statistics names to our new matrix
colnames(ScaledSimStats)<- names(obsSummaryStats)

### Correlation test

cor(ScaledSimStats, method = 'spearman')

### PCA

# applying principal component analysis on data
pca = prcomp(ScaledSimStats)

# plot to show variable importance
par(mar = rep(2, 4))
plot(pca)

# change the directions of the biplot
pca$rotation=-pca$rotation
pca$x=-pca$x

# plot pca components using biplot
biplot(pca ,scale =0)

# plot correlation diagram
corrgram(ScaledSimStats, order= TRUE, upper.panel = panel.pie)

#
summary(pca)
```

Now, we are ready to use the abc package to run a local-linear regression ABC to decide which of our drawn values we are going to keep.

```
################################ ABC ####################################
# Run abc (For the 1st summary statistic that assigns
# to the fst measure)

my_abc <- abc(target=obsSummaryStats[1], param= SimParamVec,
sumstat=SimStatsMatrix[,1], tol=0.1, method = "loclinear")

# Ask for a summary with all our distribution properties
summary(my_abc)

# Plot abc output with our parameter estimations
plot(my_abc, SimParamVec, cex = 0.5)
```

Finally, we wanted to visualize the result posterior distribution along with some of each properties, like the highest density interval (HDI).

```
###################### Posterior Plot ##################################
# calculate density of posterior distribution
postdensity <- density(my_abc$adj.values)

# use package HDInterval to calculate the HDI
HD <- hdi(postdensity)

# plot posterior distribution
hist(my_abc$adj.values, freq=FALSE,
          main = 'Posterior Distribution Histogram',
          xlab = 'Divergence Time')

# add density function
lines(postdensity, col = 'darkgreen')

# finally add HDI
ht <- attr(HD, 'height') # assign 'height' to a variable
segments(x0 = mean(my_abc$adj.values), y0=0,
x1=mean(my_abc$adj.values), y1 = 0.000163, col='pink',
 lwd = 3, lty = 2)

# add mean line
segments(x0= HD[1], x1=HD[2], y0=ht, y1=ht, col = 'red', lwd = 3,
                  lty =2)
# add legend
legend('topright', legend=c('Density','Mean', 'HDI'), lty=1:3,
 col = c('darkgreen', 'pink',  'red'))
```

3 Results

Quick question. For the test simulation we run for T=2e5, we received an fst value of approximately 0.43, lower than 0.57, which is the fst value of the observed summary statistics. The lower the population genetic differentiation is, the more genetically similar are the examined species. Our observed fst is higher than the simulated one for a time of 2e5 years ago, meaning that the species are more different that estimated. Hence their expected divergence time should be greater than 2e5, for the species to have more time to genetically differentiate from one another.

The Table 1 and Figure 1 demonstrate the summary statistics of our posterior distribution. The mean value of T is approximately 295,358 years ago, with a high density interval of almost 10,000 years (from 290,432 to 300,316). The results from the pca analysis are demonstrated in Figure 2.

Posterior Properties								
Mean	Median	Mode	Minimum	Maximum	HDI		Quantiles	
					Lower	Upper	2.5%	97.5%
295357.7	295415.9	295699.2	288499.1	302660.0	290432.2	300316.4	290693.8	300047.7

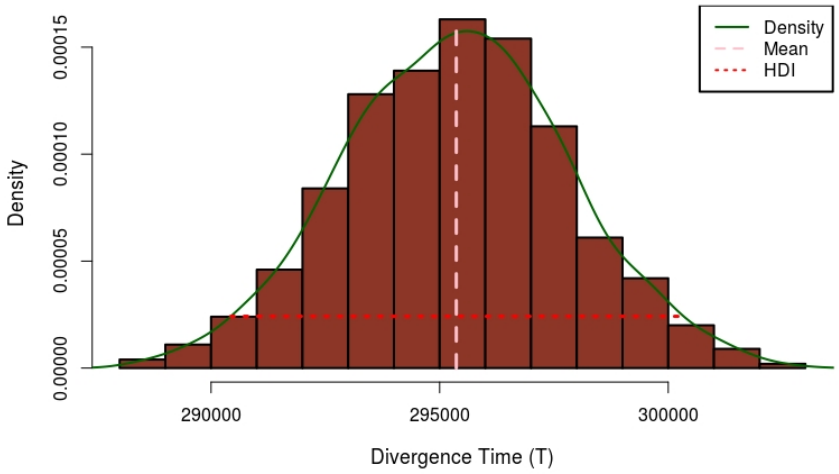

Figure 1: The posterior distribution, with the higher density interval, the mean value and the density function illustrated.

Detailed results of our abc are gathered in Figure 3. To conclude, we can say that there are evidence that the divergence time between polar and brown bears is most likely to be between 290k and 300k years ago.

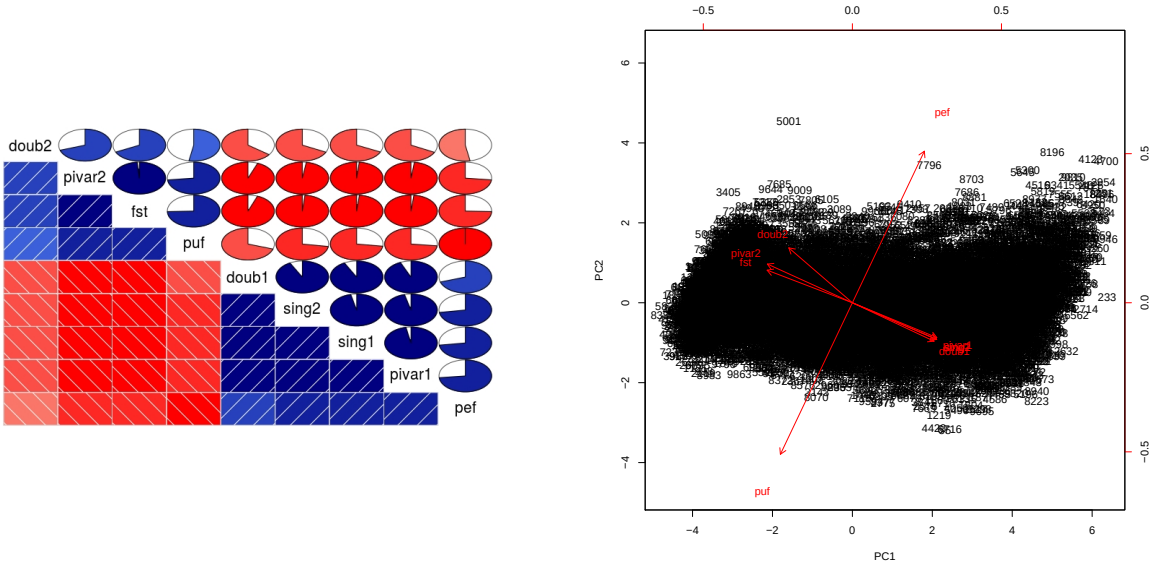

Figure 2: PCA.: Left: correlation diagram, Right: pca

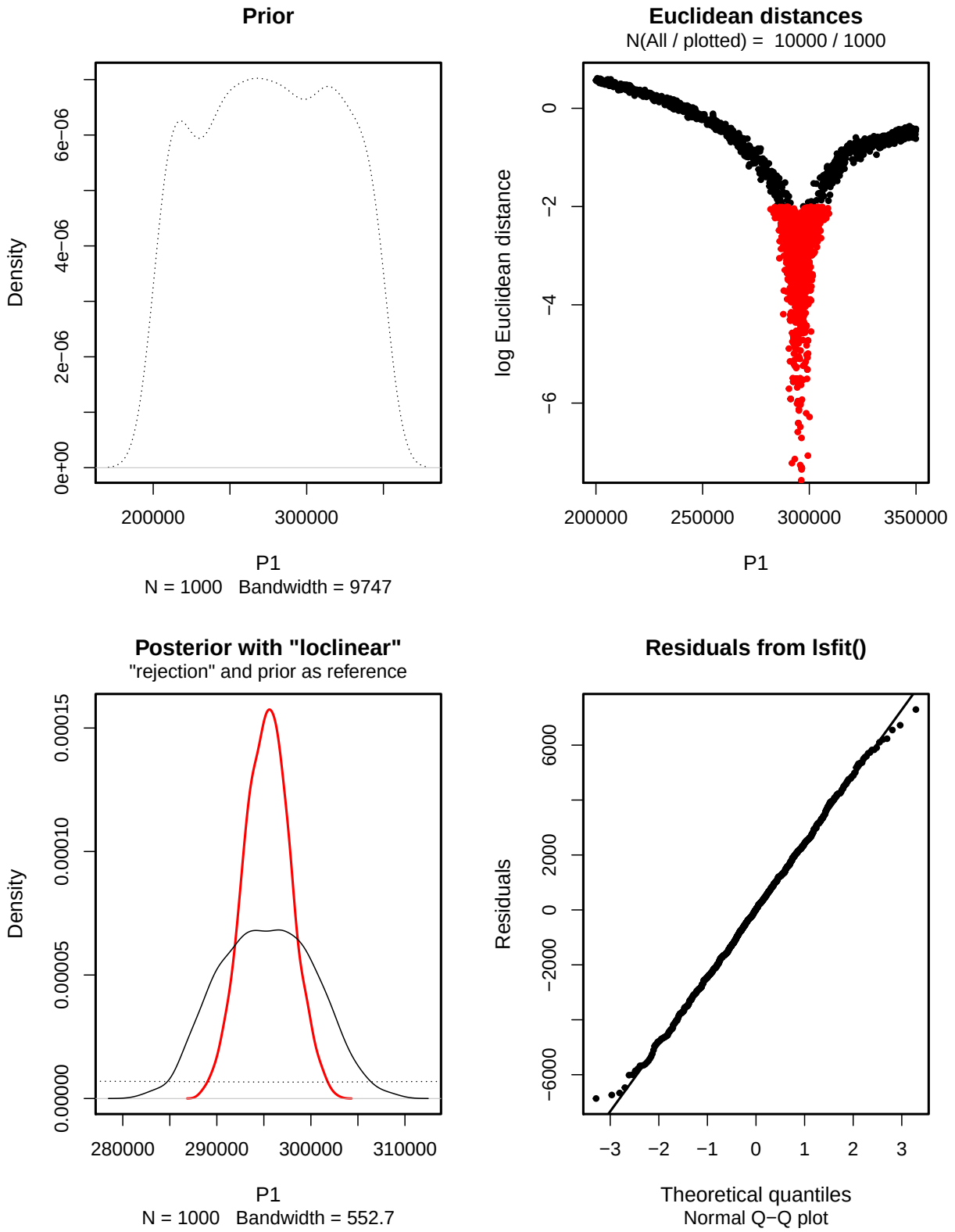

Figure 3: Diagnostics of our ABC.