Bayesian Statistics
Exercise 3 Submission

CID numbers:
01125912
01138251
01249014
01310808

1 Introduction

Polar bears (*Ursur maritimus*) and brown bears (*Ursur arctos*) are sister species, with their divergence being attributed to the specialisation of the polar bear to the extreme environment of the northern polar region. Whilst there have been a number of attempts at inferring the time at which the two species diverged no general consensus has yet been reached, with genetic data and the fossil record often suggesting very different dates [1]. The aim of this exercise was to estimate the divergence time between polar bears and brown bears using current genetic data and a Bayesian framework.

2 Methods

In order to estimate the divergence time we utilised approximate bayesian computing. This involved the following procedure (which can also be seen in the R code below):
-Calculating summary statistics of the observed genetic data available for polar and brown bears.
-Defining a prior distribution for possible divergence times.
-Sampling from the prior.
-Using the sample divergence time to simulate genetic data for polar and brown bears.
-Calculating the summary statistics of the simulated data.
-Repeating the sampling and simulation steps until N samples were obtained.
-Scaling the summary statistics.

-Applying local linear regression to the specified scaled summary statistic/s using the 'abc' function in R to generate a posterior distribution.

The prior we used for divergence time was uniform, continuous and bounded at 100,000 and 700,000 years ago. We used a uniform continuous prior because we do not possess any previous knowledge concerning the expected divergence time between the polar and brown bear species to suggest that a prior in the form of another distribution would be preferred. In addition, we retained such a broad range to ensure that the prior did not inadvertently limit our estimation of the posterior distribution by not including the true divergence time.

In order to ascertain which summary statistics/s are the most informative we performed principal components analysis. The scree plot (Figure 1) generated suggests that PC1 and PC2 retain 96.3% of the total variation in the genetic data.

We subsequently analysed the contributions of variables to PC1 and PC2, as shown in the contribution plot (Figure 2). This demonstrates that Fst, pivar1, sing1, doub1 and doub2 contribute significantly to PC1 and PC2 (bars higher than dahed red line which indicates greater than average contribution). Then we computed the correlations between these summary statistics, and found that except for doub2, all other summary statistics are highly correlated, with absolute correlation coefficients exceeding 0.98. This is visualized on the factor map (Figure 3), with the variables grouping together when highly positively correlated or opposite to each other suggesting negative correlations. As doub2 only concerns the genetic information for brown bears, it is unlikely to be strongly influenced by the speciation event. Instead we believe it will be mostly affected by changes in the size of the brown bear population.

Consequently, we thought it best to choose Fst as our chosen summary statistic for assessing the most likely divergence time between polar and brown bears. This is because Fst represents the proportion of genetic diversity caused by allele frequency differences between populations and it is used extensively in population and evolutionary genetics [2, 1]. Crucially, from our preliminary simulation runs we were also able to demonstrate that Fst is highly correlated with divergence time, having a correlation coefficient of 0.93 and with an AIC of 1546.11, the lowest score for any summary statistic, again indicating that it is a suitable summary statistic to use in our analysis.

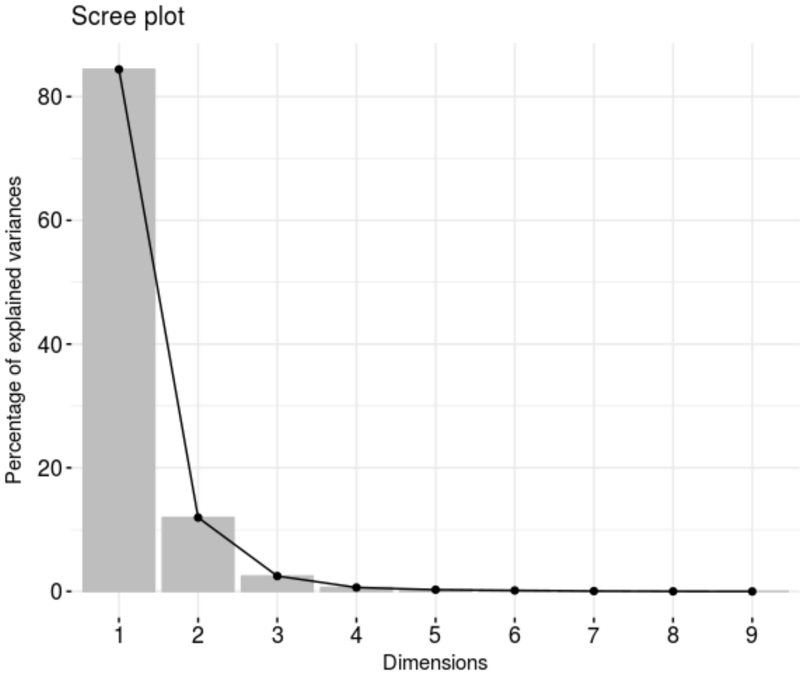

Figure 1: A scree plot depicting the percentage of variance explained by each dimension in a PCA of the summary statistics.

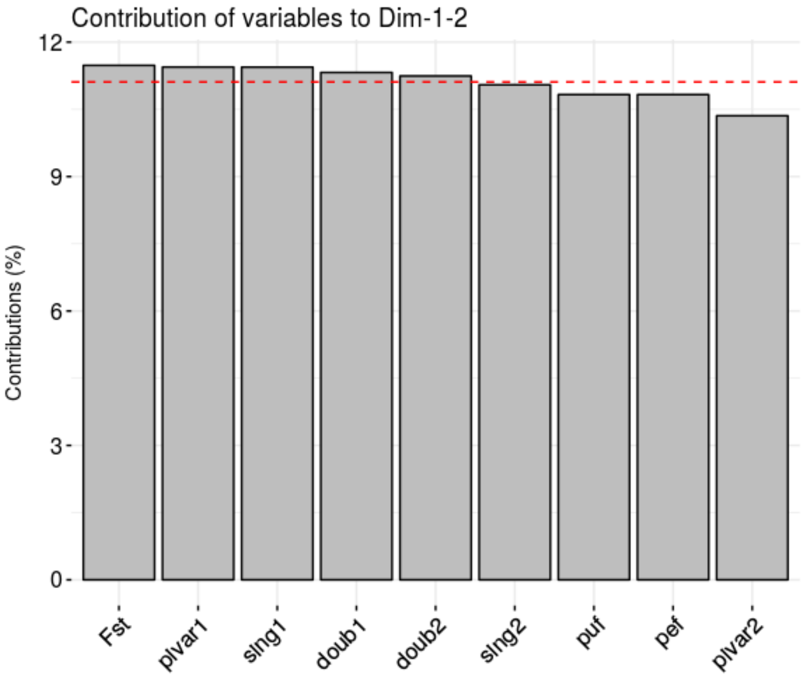

Figure 2: A contribution plot showing how the individual summary statistics contribute to the first and second principle component.

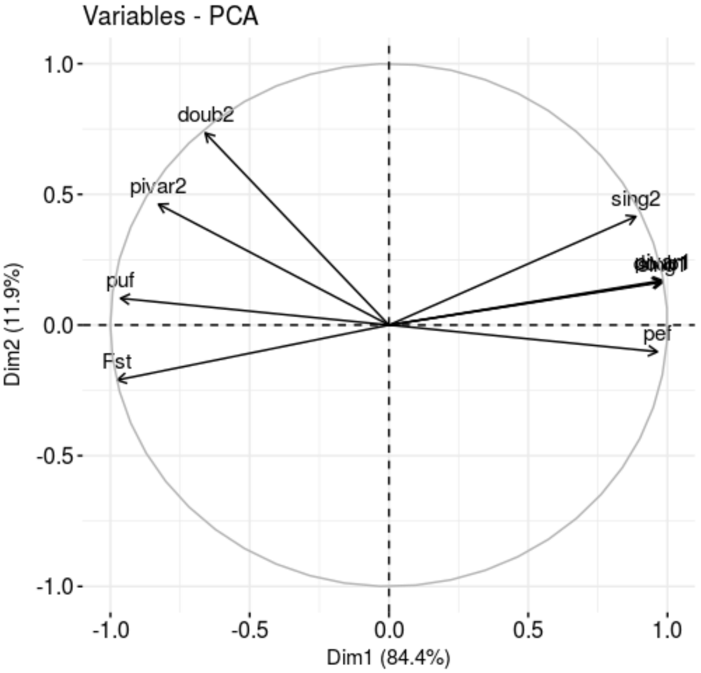

Figure 3: A factor map showing how the summary statistics correlate with each other.

Having generated the posterior we calculated a number of summary quantities, including the mean, median, mode and highest posterior density (HPD) interval. These were either obtained from the 'abc' result object or calculated as described in the code below.

The following R code was used to perform the Approximate Bayesian Computation and generate our results.

```
##########
# Bayesian Statistics Practical 3 Code
##########

# Clear the working environment.
rm(list = ls())

# Load the abc function from the R library.
library(abc)
# Load the HDInterval package.
library(HDInterval)
# Load the packages for PCA.
library(”FactoMineR”)
library(”factoextra”)

# Load the files required for the analysis;
# functions.R contains the R functions used and
# polar.brown.sfs .Rdata contains the genetic data.
source(”functions.R”)
load(”../Data/polar.brown.sfs.Rdata”)

# Plot site frequency spectrum of polar vs brown bear.
plot2DSFS(polar.brown.sfs, xlab = ”Polar.bear”, ylab = ”Brown.bear”,
           main = ”2D-SFS”)

# Calculate the number of chromosomes present for each species.
nChroms.polar <- nrow(polar.brown.sfs)-1
nChroms.brown <- ncol(polar.brown.sfs)-1

# Calculate the total number of sites analysed.
# Required for simulations.
nrSites <- sum(polar.brown.sfs, na.rm = T)

# Calculate summary statistics of the observed data.
# (Fst, pivar1, pivar2, ing1, sing2, doub1, doub2, pef, puf)
obsSummaryStats <- calcSummaryStats(polar.brown.sfs)


#####
# PCA
#####

# Calculate the PCA values from a simulation with 1000 iterations.
pca <- PCA(scaled.results[1:nrSims,], graph = FALSE)

# Generate a scree plot to visualise the importance of PCs
fviz_screeplot(pca, ncp = 10, barfill = ”gray”, barcolor = ”gray”)

# Plot the variables as points in the component space using correlation between
# a variable and a PC as coordinates.
fviz_pca_var(pca)

# Create a contribution plot for PC1 and PC2
fviz_contrib(pca, choice = ”var”,axes = 1:2, fill = ”gray”, color = ”black”)


#####
# Running simulations
#####

# Define number of simulations to perform.
```

```
nrSims <- 1e5

# Specify path for ms program.
msDir <- "~/Documents/Bayesian␣Statistics/Code/ms"

# Specify output file.
fout <- "../Results/ms.txt"


# Create a matrix to store divergence time and simulated summary statistics.
# rows 1 to nrSims will contain data obtained from an individual simulation.
results.matr <- matrix(nrow = nrSims+1, ncol = 10)

# Columns of matrix are named according to the data they contain.
colnames(results.matr) <- c("T", "Fst", "pivar1", "pivar2", "sing1", "sing2",
                            "doub1", "doub2", "pef", "puf")

# Input observed summary statistic values into the final row of the matrix.
results.matr[nrSims+1,] <- c(NA, obsSummaryStats[[1]], obsSummaryStats[[2]],
                             obsSummaryStats[[3]], obsSummaryStats[[4]],
                             obsSummaryStats[[5]], obsSummaryStats[[6]],
                             obsSummaryStats[[7]], obsSummaryStats[[8]],
                             obsSummaryStats[[9]])

# Run the number of simulations specified by nrSims.
for (i in 1:nrSims){
  # Sample divergence time from the prior
  # (continuous, bounded, uniform distribution)
  theta <- runif(1, min = 1e5, max = 7e5)

  # Simulate genetic data for polar and brown bears;
  # use sampled divergence time, migration rate as zero and set number of sites
  # to simulate as nrSites.
  simulate(T = theta, M = 0, nrSites, msDir, fout)
  # Convert simulated results results to site frequency spectrum format.
  # Enables calculation of summary statistics.
  simulatedSFS <- fromMStoSFS(fout, nrSites, nChroms.polar, nChroms.brown)
  # Calculate summary statistics of the simulated genetic data.
  simSummaryStats <- calcSummaryStats(simulatedSFS)

  # Add the sampled divergence time and summary statistics of the simulated
  # data to row i in matrix.
  results.matr[i,] <- c(theta, simSummaryStats[[1]], simSummaryStats[[2]],
                        simSummaryStats[[3]], simSummaryStats[[4]],
                        simSummaryStats[[5]], simSummaryStats[[6]],
                        simSummaryStats[[7]], simSummaryStats[[8]],
                        simSummaryStats[[9]])
}


#####
# Analysing results
#####

# Scale the summary statistics for analysis.
# Allows for consistent comparisons irrespective of the range of the
# summary statistics.
scaled.results <- scale(results.matr[,2:10])

# Save the results matrix and scaled results matric to Rda files.
# Ensures data can be anaylsed again without re-running the simulations.
save(results.matr, file = "../Data/results.matr1.Rda")
save(scaled.results, file = "../Data/scaled.results1.Rda")

# Plot sampled divergence time against summary statistics.
# Use to identify summary statistics that may be informative regarding the
# estimation of the polar bear - brown bear speciation date.
# For a thorough discussion on the choice of summary statistic used for the
# approximate bayesian computation, please see the main text.

# plot(results.matr[1:nrSims,1], scaled.results[1:nrSims,1]) # Fst, +ve saturating relationship
# plot(results.matr[1:nrSims,1], scaled.results[1:nrSims,2]) # Pivar1 -ve correlation
# plot(results.matr[1:nrSims,1], scaled.results[1:nrSims,3]) # Pivar2 humped pattern
# plot(results.matr[1:nrSims,1], scaled.results[1:nrSims,4]) # Sing1 -ve relationship
# plot(results.matr[1:nrSims,1], scaled.results[1:nrSims,5]) # Sing2 initial +ve then -ve relationship
# plot(results.matr[1:nrSims,1], scaled.results[1:nrSims,6]) # Dub1 -ve relationship
# plot(results.matr[1:nrSims,1], scaled.results[1:nrSims,7]) # Dub2 humped pattern
# plot(results.matr[1:nrSims,1], scaled.results[1:nrSims,8]) # PEF -ve exponential appearance
# plot(results.matr[1:nrSims,1], scaled.results[1:nrSims,9]) # PUF +ve saturating curve

# Run the abc function.
# Use the observed, scaled Fst value as the target.
# Use the column of sampled divergence times as the parameter samples.
# Use the simulated scaled Fst values as the summary statistic.
# Set tol to 0.1 to retain the 10% of simulations which generated Fst values
# closest to the observed Fst value.
# Set method to loclinear for local weighted regression.
result <- abc(target = scaled.results[nrSims+1,1], param = results.matr[1:nrSims,1],
              sumstat = scaled.results[1:nrSims,1], tol = 0.1, method = "loclinear")

# Print a summary of the result object
summary(result)
# Provides a summary of the posterior distribution.


#####
# Plotting of results
#####

# Plot a number of informative graphs generated by the abc function.
# 1. A plot of the sampled prior distribution.
# 2. A plot of the log Euclidian distance for simulated Fst values at
# different divergence times.
# 3. A plot of the approximated posterior distribution.
# 4. A plot of the residuals from the linear regression.

# Open a pdf file to save plot to.
pdf("../Results/abc_plot.pdf")

# Plot the graphs.
plot(result, results.matr[1:nrSims,1], subsample = 10000)

# Turn off the graphics.
dev.off()


# Calculating the HPD of the posterior.
# Calculate the density of the posterior using the adjusted values from the
# abc object.
dens.time.matr <- density(result$adj.values)

# Calculate the 95% HPD from the density values.
hdi.time.matr.dens <- hdi(dens.time.matr, credMas = 0.95)

# Open a pdf file to save plot to.
pdf("../results/posterior_HPD.pdf")

# Create a histogram of the posterior disribution for divergence time.
hist(result1$adj.values, freq=FALSE,
     main = "Posterior_Distribution_of_Divergence_Time",
     xlab = "Divergence_Time", col = "lightgrey")

# Plot the probability density line over the histogram.
lines(dens.time.matr, col = "blue")

# Add a legend to the plot to provide the HPD interval.
legend(276000, 0.00010, legend = c("Lower_HPD_=_266094",
                                   "Upper_HPD_=_276086",
                                   "at_95%_Cred._Mass"),
       cex = 0.6)

# Set the height of the horizontal line indicating the range of the HPD.
ht <- attr(hdi.time.matr.dens, "height")

# Plot the HPD as calculated above.
segments(hdi.time.matr.dens[1], ht, hdi.time.matr.dens[2], ht,
         lwd=3, lty = 2, col='red')

# Turn off the graphics.
dev.off()

# Obtaining the quantiles of the posterior distribution.
```

```
postq <- quantile(result$adj.values)

# Open a pdf file to save boxplot to.
pdf("../Results/boxplot.pdf")

#  Generate boxplot depicting quantiles of posterior.
boxplot(result$adj.values, main = "Posterior's Quantiles")
abline(h = 272727.4, lty = 3, lwd = 3, col = "red")
abline(h = 269392.6, lty = 3,  lwd = 3, col = "red")

# Turn off the graphics.
dev.off()
```

3 Results

Figure 4 shows that the posterior distribution is well within the bounds of the prior but is very different in shape, confirming that Fst conveys information about the time of divergence between polar and brown bears. This can be seen when comparing the upper and lower left panels. The upper right panel of Figure 4 depicts the distance between the simulated and observed summary Fst values as a function of divergence time. This plot also confirms that Fst conveys information about divergence time because the distances corresponding to accepted values (those coloured red) are clustered in a narrow region of the prior distribution. The lower right panel displays a standard Q-Q plot of the residuals of the regression.

From the posterior distribution, we also obtained a number of key summary quantities which can be found in Table 1 and are depicted in Figures 5 and 6.

Summary of posterior	Values (years before present)
Mean	270,998.4
Median	270,995.7
Mode	271251.3
HPD interval	266,094 - 276,086
25% quantile	269392.6
75% quantile	272727.4

Table 1: A table containing values summarising the posterior distribution of estimated divergence time between polar and brown bears.

Our results therefore suggest that polar bears and brown bears diverged from each other around 271,000 years ago with a 95% probability that this occurred between 266,094 and 276,086 years ago.

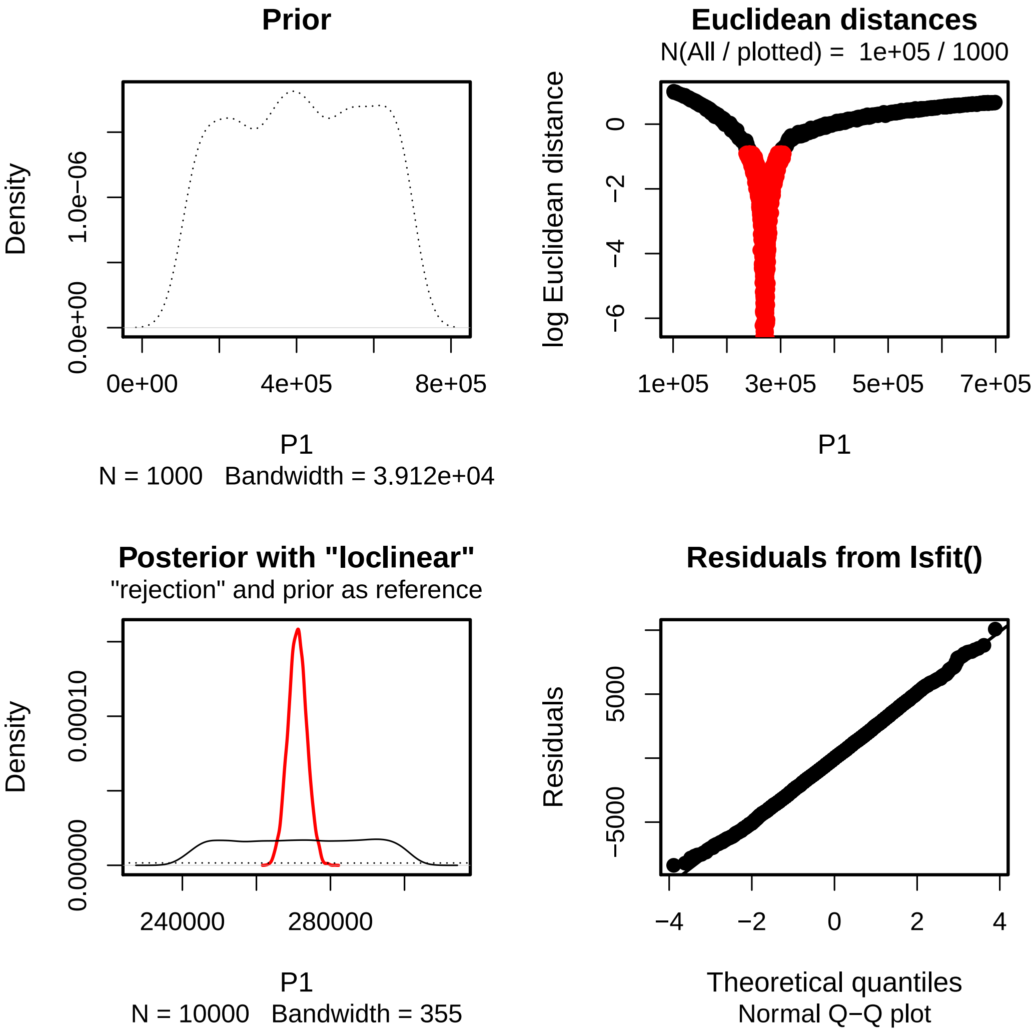

Figure 4: A multi-panel plot summarising the output of the abc function, with P1 representing divergence time.

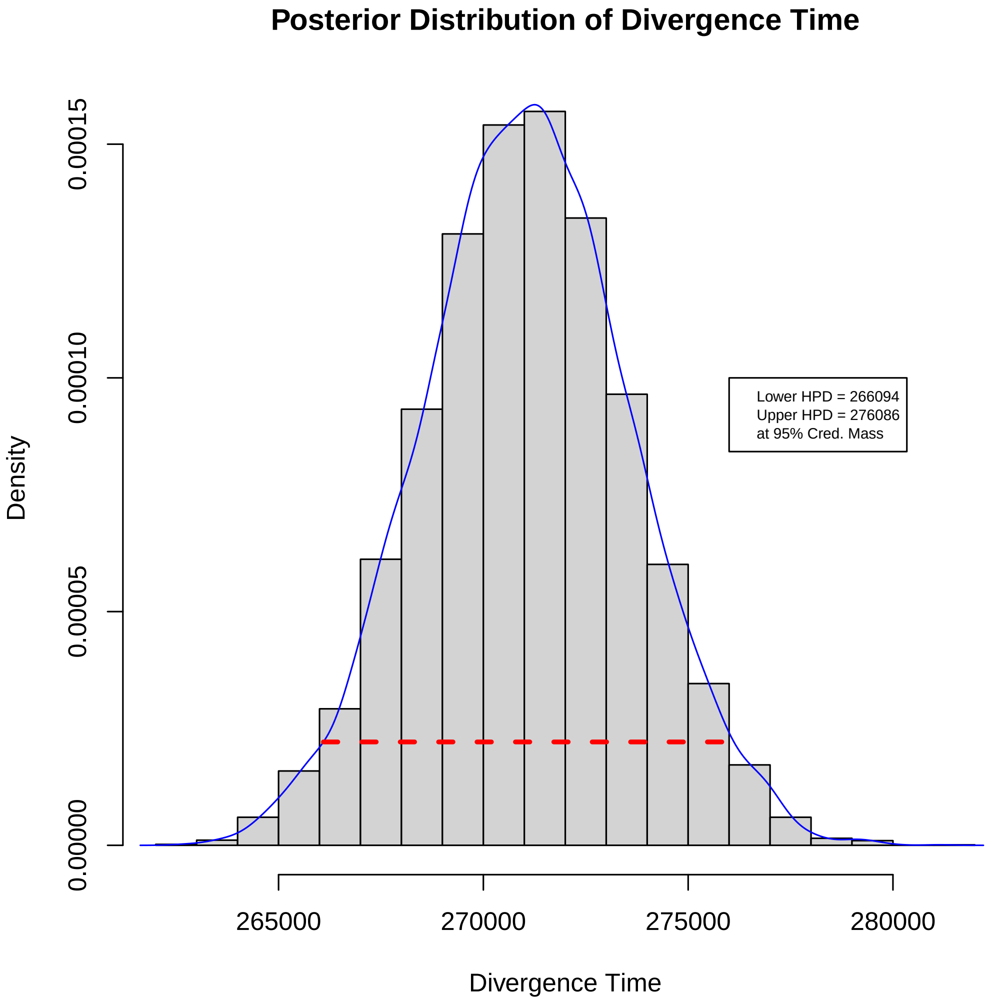

Figure 5: A graph showing the posterior distribution of divergence time with the HPD interval indicated.

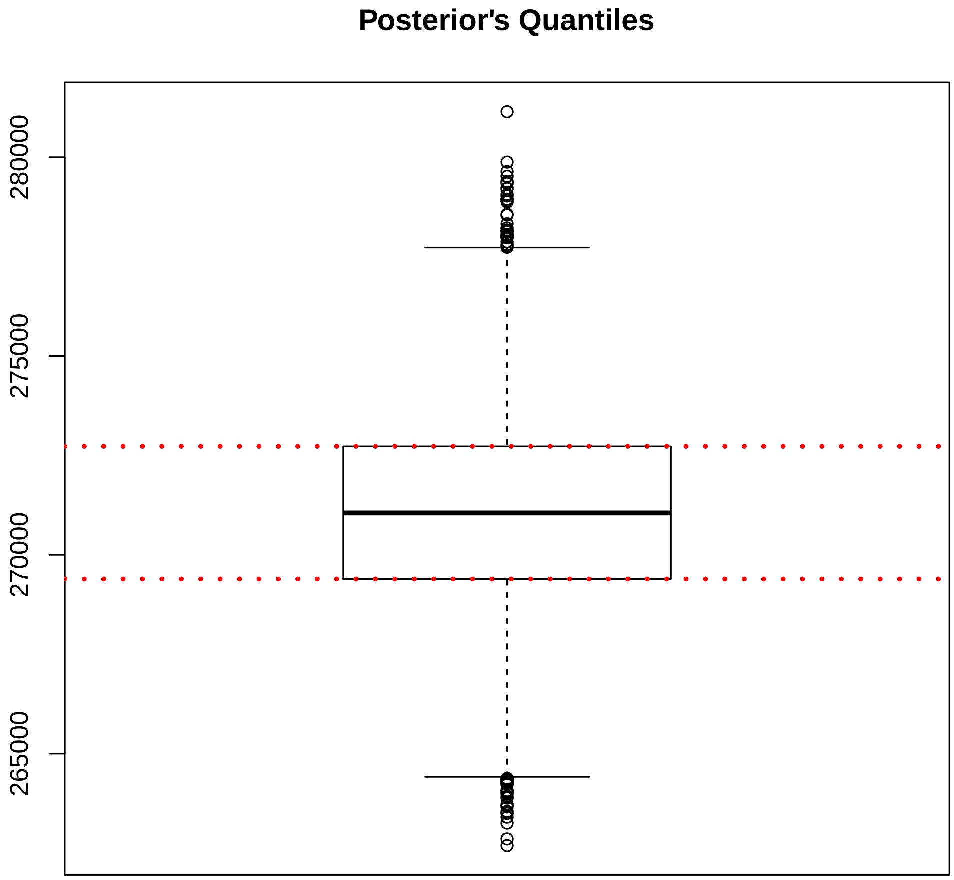

Figure 6: A boxplot highlighting the 25% and 75% quantiles of the posterior distribution.

References

[1] Shiping Liu, Eline D Lorenzen, Matteo Fumagalli, Bo Li, Kelley Harris, Zijun Xiong, Long Zhou, Thorfinn Sand Korneliussen, Mehmet Somel, Courtney Babbitt, et al. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, 157(4):785–794, 2014.

[2] Kent E Holsinger and Bruce S Weir. Genetics in geographically structured populations: defining, estimating and interpreting fst. *Nature Reviews Genetics*, 10(9):639–650, 2009.