Bayesian Assignment

CIA:01282428

Introduction

The aim of this project was to estimate the divergence time of brown and polar bear populations. I had site frequency spectrum data available from brown and polar bear populations and then I simulated data from a uniform distribution. I chose a uniform distribution because I had no reliable information on what an appriopriate prior would be and a uniform prior doesn't favour any value over another. Using approximate Bayesian computation a rejection algorithm was implemented so that only data similar to the observations were kept, I used a 15% tolerance intevral so that only data points within 15% of the observed data were used. The loclinear method was used as can produce diagnostic plots of the results. This method also allows you to get a potentially wider set of accepted data points.

Rcode

```
rm(list=ls())
source("functions.R")
load("polar.brown.sfs.Rdata")

ls()
plot2DSFS(polar.brown.sfs, xlab="Polar", ylab="Brown", main="2D-SFS")


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

#run simulation
nrSimul <- 1e4 # can change later
```

```
#set path to ms software installed
msDir <- "~/Documents/msdir/ms"

#output file
fout <- "ms.txt"

#simulate data
N<-10000
#t <- c() #array
simsummarystats<-matrix(nrow=N, ncol=(length(obsSummaryStats)))
colnames(simsummarystats) <- paste(c("fst", "pivar1", "pivar2", "sing1", "
paramT<-matrix(nrow=N, ncol=1)

for (i in 1:N){
t<-runif(1, 0, 400000) # uniform prior

# simulate obervations, stores in fout
simulate(T=t, M=0, nrSites, msDir, fout)

#calculate summary stats from simulated observations
simulatedSFS <- fromMStoSFS(fout, nrSites, nChroms.polar, nChroms.brown)
simsummarystats[i,]<-calcSummaryStats(simulatedSFS)

###saving t separately###
paramT[i,]<-t

}

#save stats to rdata file
save(obsSummaryStats, simulatedSFS, simsummarystats, t, file=("stats.rda"))

#correlations (I changed the numbers wen testing for different correlations
#plot(simsummarystats[,2], simsummarystats[,3])
#cor(simsummarystats[,9], simsummarystats[,10])

#fst:2 =0.99 # INCLUDE, high correlation with T
#pivar1 =-0.94 #INCLUDE, high correlation with T
#pivar2 = 0.958 #INCLUDE, high correlation with T
#sing1 = -0.94 #don't include high corr with pivar1
```

```
#sing2 = 0.09 # no corr
#doub1 = -0.94 # INCLUDE, high correlation with T
#doub2 = 0.74 #no corr
#pef = -0.66 # no corr
#puf = 0.66 # no corr
#highly correlated: 2,4 3,5 3,7 5,7

#put observed and simulated stats together
summarystats<-rbind(obsSummaryStats, simsummarystats)
scaled_summarystats<-matrix(nrow=N+1, ncol=(ncol(summarystats)))
colnames(scaled_summarystats) <- paste(c("fst", "pivar1", "pivar2", "sing1"


##scale the summary stats together so that they are comparable
for (i in 1:ncol(scaled_summarystats))
{
scaled_summarystats[,i]<-scale(summarystats[,i])
}

#function to calculate the mode
Mode <- function(x) {
ux <- unique(x)
ux[which.max(tabulate(match(x, ux)))]
}


pdf("hist_prior_posterior.pdf")
hist(x)
hist(paramT, col=rgb(0, 0, 1, 0.5), add=T)
legend("bottomright", c("Prior", "Posterior"), fill=c("white", "blue"))
dev.off()

#run abc using chosen summary stats from earlier
require(abc)
abc<-abc(target = scaled_summarystats[1,1:3&6], param = paramT, sumstat =

#save plot to pdf
pdf("graph_1:3,6.pdf")
plot(abc, paramT)
dev.off()
```

```
#run abc just using Fst
abc<-abc(target = scaled_summarystats[1,1], param = paramT, sumstat = scale

pdf("graph_1.pdf")
plot(abc, paramT)
dev.off()

#calculate HDInterval
require(HDInterval)

dens.time.matr <- density(abc$adj.values)
hdi.time.matr.dens <- hdi(dens.time.matr, credMas = 0.975)
pdf("HDInterval.pdf")
hist(abc$adj.values, freq=FALSE)
lines(dens.time.matr)
ht <- attr(hdi.time.matr.dens, "height")
segments(hdi.time.matr.dens[1], ht, hdi.time.matr.dens[2], ht, lwd=3, col=
dev.off()
```

Results and Discussion

The mean parameter estimate in the posterior was 271, 009 meaning that this model predicted a divergence time 271 000 years ago. The mode result was 271 341, and the median estimate was 271 009. I chose to go with the mean as the posterior distribution was approximately normally distributed so wasn't affected by extreme values. The 95% confidence interval was between 266 190 years and 275 796 years, giving a range of 9,606 years.

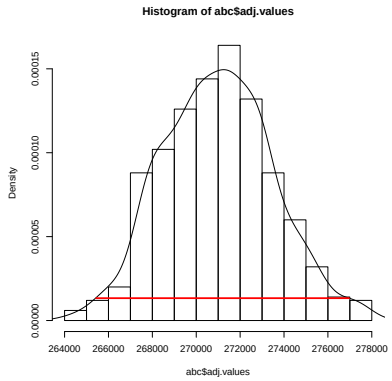

Figure 1: Histogram of adjusted values of the abc. The red line shows the 95% confidence interval

Choice of summary statistics

From looking at correlations between the summary statistics and the parameter T and also from looking at correlations between summary statistics I chose to include Fst, pivar1, pivar2 and doub1. As none of these were correlated with each other and they were strongly correlated with the parameter T.

Once diagnostic plots had been produced (Figure 3), I decided to just include F_{st} as this gave the most similar mean to the parameter estimate at the 15% tolerance interval. F_{st} looks at genetic differentiation between subpopulations so in this case F_{st} would identify the genes that are highly differentiated between brown and polar bears subpopulations and is therefore the most important sumary statistic when estimating the divergence time in this data. Due to the fact that the posterior distribution looks very different to the prior, this shows that the posterior distribution got some information from the MLE created by simulating the data and not the prior alone. So some information from the data must be driving this distribution.

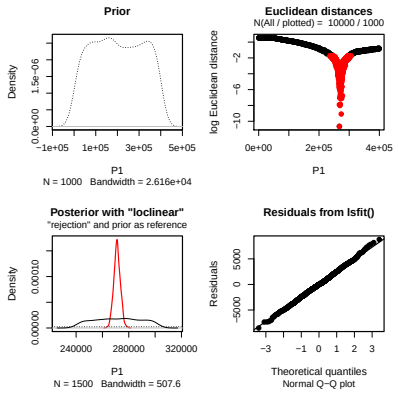

Figure 2: Diagnostic plots of the abc just including the Fst summary statistic