Supervision Meeting Demo

J.P. Meagher

14 November 2017

Here I will develop more fully the pipeline for ancestral reconstruction of Bat Echolocation call spectrograms. This will go through creating a bootstrapped sample for analysis, identification of components and component weights, hyperparameter estimation, Leave-Species-Out validation of reconstructions, to finally producing reconstructions of Bat Echolocation Call Spectrograms.

Packages

Packages used in this analysis are as follows:

```
library(signal) # Produce Spectrogram
library(tidyverse) # see http://r4ds.had.co.nz/
library(magrittr) # pipe operator and alaises
library(batwork) # my own package, long form call data
library(sdsBAT) # my own package, tree data and functions for ancestral reconstruction
library(ape) # Analyses of Phylogenetics and Evolution
library(RColorBrewer) # for pretty pictures
library(ggridges) # Joy plots
library(ggtree) ## ggplot for phylogenetic trees
```

Data

Preprocessing of echolocation call spectrograms was performed in Matlab.

Mean Spectrograms

Mean spectrograms for the sample to the bat, species, and family levels may all be of interest in this analysis.

```
## # A tibble: 1 × 4
##      bat family species             full
##   <fctr> <fctr>  <fctr>           <list>
## 1    228    Mor    Mome <dbl [104 × 50]>
```

Standard Deviation of Spectrograms

What may also be of interest when considering these spectrograms is the standard deviation at each pixel over various levels of the sample.

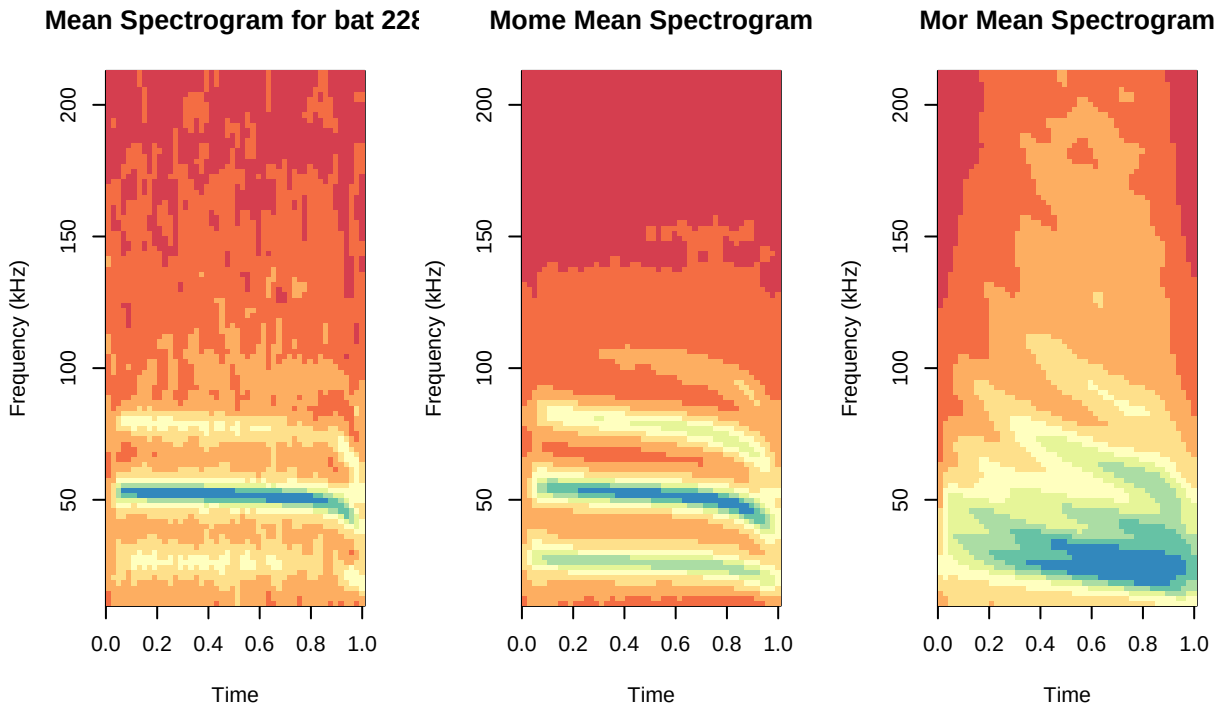


Figure 1: Plotted above the mean spectrogram for a randomly selected individual bat along with the species and family level mean spectrograms for that individual. These plots serve as a sense check on the data.

Bootstrapped Sample

As there is variation in the number of bats per species and calls per bat in this sample I think that it is appropriate to create a bootstrapped sample of calls upon which to run the analysis. Note also that it is the mean spectrogram per each species that we are particularl interested in and so the bootstrapped sample will contain estimates of the mean spectrogram per each species.

Principal Components as Evolutionary Features

A Principal Components Analysis is performed on the data, yielding a dimension reduction along modes of variation. Each principal component can be considered to be a suite of evolutionary features, with each suite being orthogonal to every other suite.

Component Scores

See Figure 2.

Score Distribution

See Figure 3.

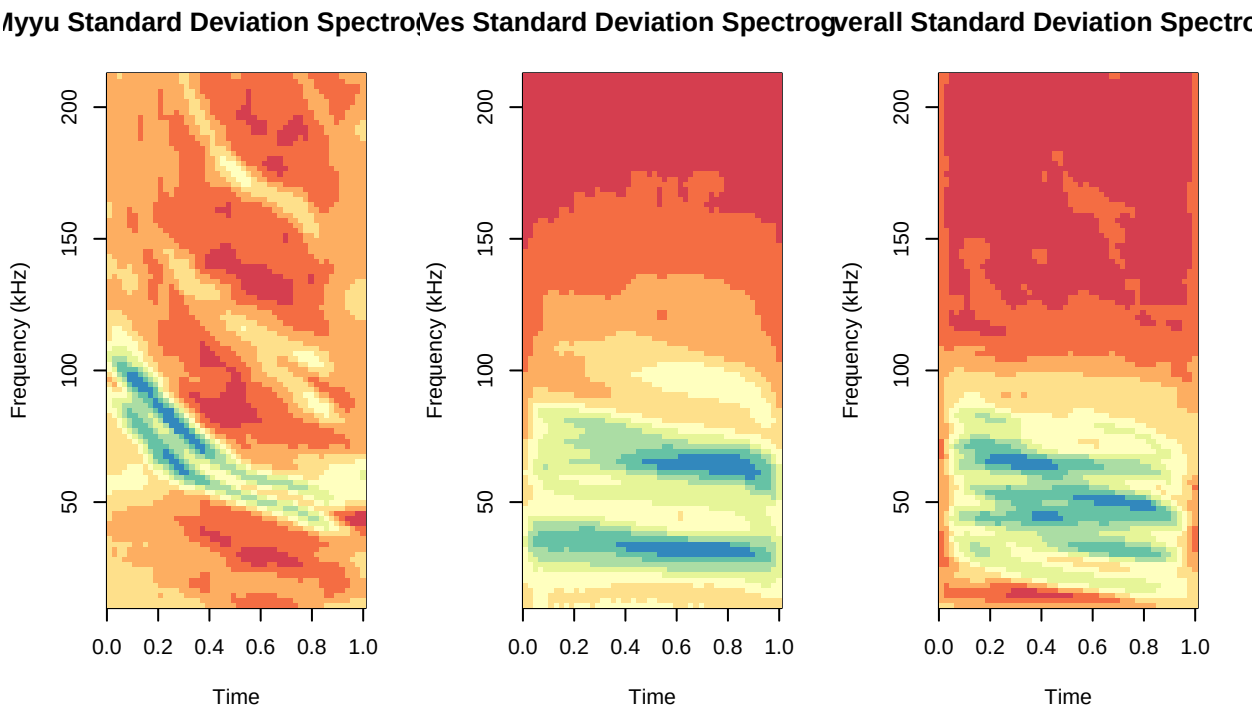


Figure 2: Plotted above the sdample standard deviation over the spectrogram for a randomly selected species of bat along with the family level and overall standard deviation spectrograms for that individual. These plots serve as a sense check on the data.

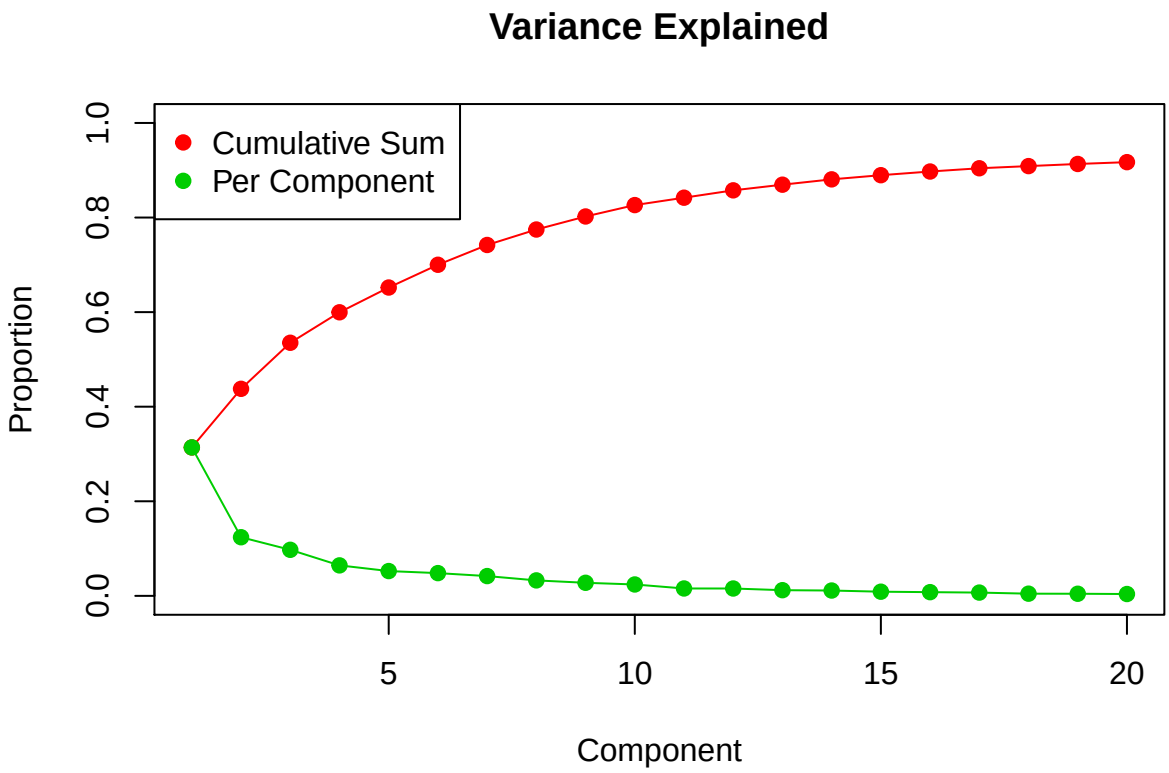


Figure 3: The scores associated with each component provide insight into the proportion of the sample variance captured by the corresponding principal component. This PCA was performed on a sample of bootstrapped mean estimates, so some of the bat level variation should have been stripped from the data. The bootstrapped sample was also weighted with the same number of samples from each species, an attempt at ensuring no one species dominated modes of variation identified. The scores above suggest that the dataset of echolocation call spectrograms is very high dimensional, requiring 13 components to explain 75% of the variance, while 16 are required to explain 90%. It can be shown that 9 components capture more than 2.5% of the variation, and 14 capture more than 1% of the variation.

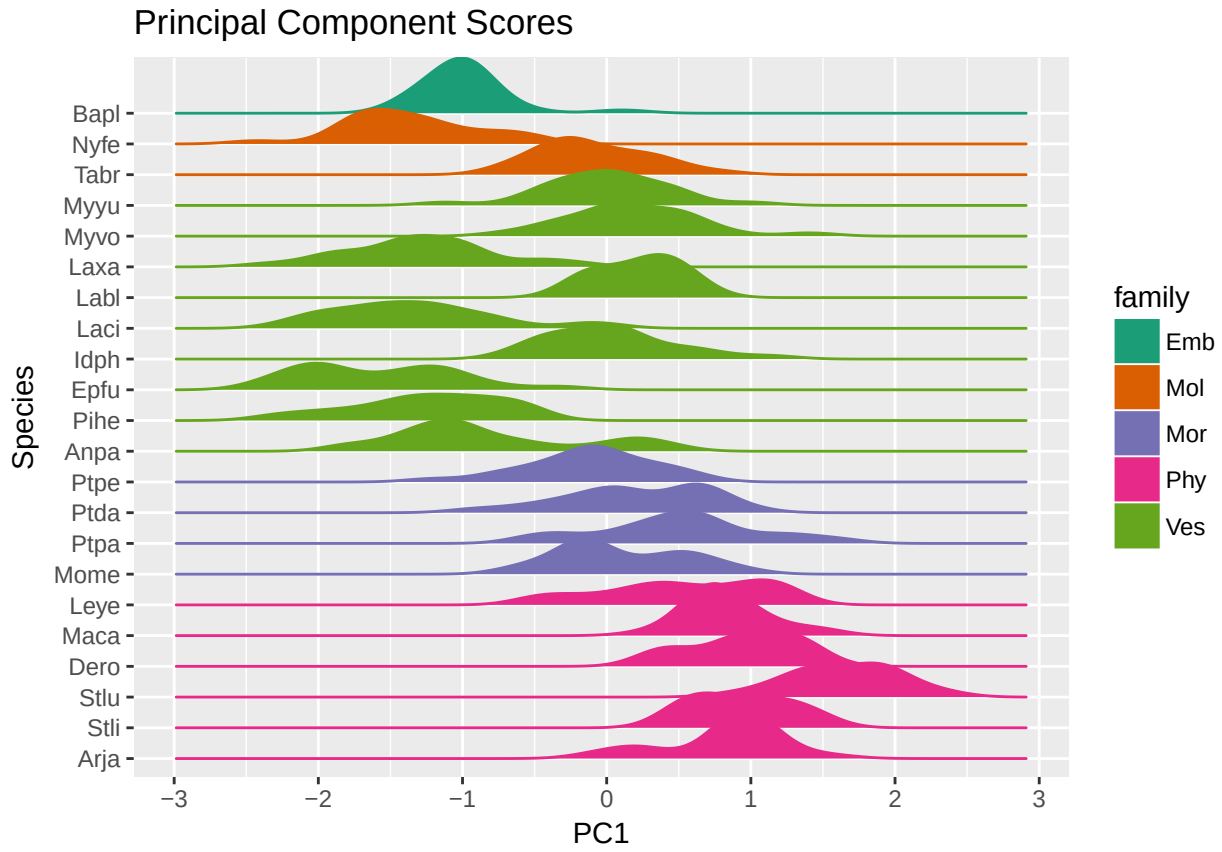


Figure 4: Distribution of component scores for a selected Principal component loading, separated out by species.

It seems reasonable to think that those principal components which explain most variation would be capturing variation between species rather than within species, but is there some clever way to check this?

Component PGP Hyperparameters

Hyperparameters for the phylogenetic Gaussian Processes can be estimated from these principal components

```
## [1] "BOOTSTRAPPED PC MLE HYP"
```

```
##         PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8    PC9   PC10
## s_p    0.87   0.94   0.99   0.99   0.99   0.93   0.98   0.96   0.96   0.91
## l     42.85   0.05  45.79   7.34  16.85  28.61   0.39   0.00   0.42   0.38
## s_n    0.45   0.29   0.49   0.31   0.31   0.39   0.32   0.30   0.36   0.40
## logl 389.24 160.15 434.52 197.60 196.03 309.43 205.76 183.18 279.91 325.95
##        PC11   PC12   PC13   PC14   PC15
## s_p    0.85   1.10   0.73   0.97   0.85
## l      5.59  76.65   0.27   0.34   0.39
## s_n    0.55   0.52   0.68   0.35   0.55
## logl 498.81 463.72 605.50 259.10 496.67
```

The components should be scaled by the square root of the associated eigen value anyway, this will lead to approx $\mathcal{N}(0, 1)$ distributed scores, I think, although I don't have a reference for this result, thus scaling in this way is appropriate. Scaling also makes the optimisation of the likelihood more straightforward as it has not been 'streched' over such a large region and has an easire time finding a maximum

Approximation of Species level Mean Spectrograms

One approach to investigating the quality of the components identified is to consider how much the projection of a spectrogram into a component space differs from the original.

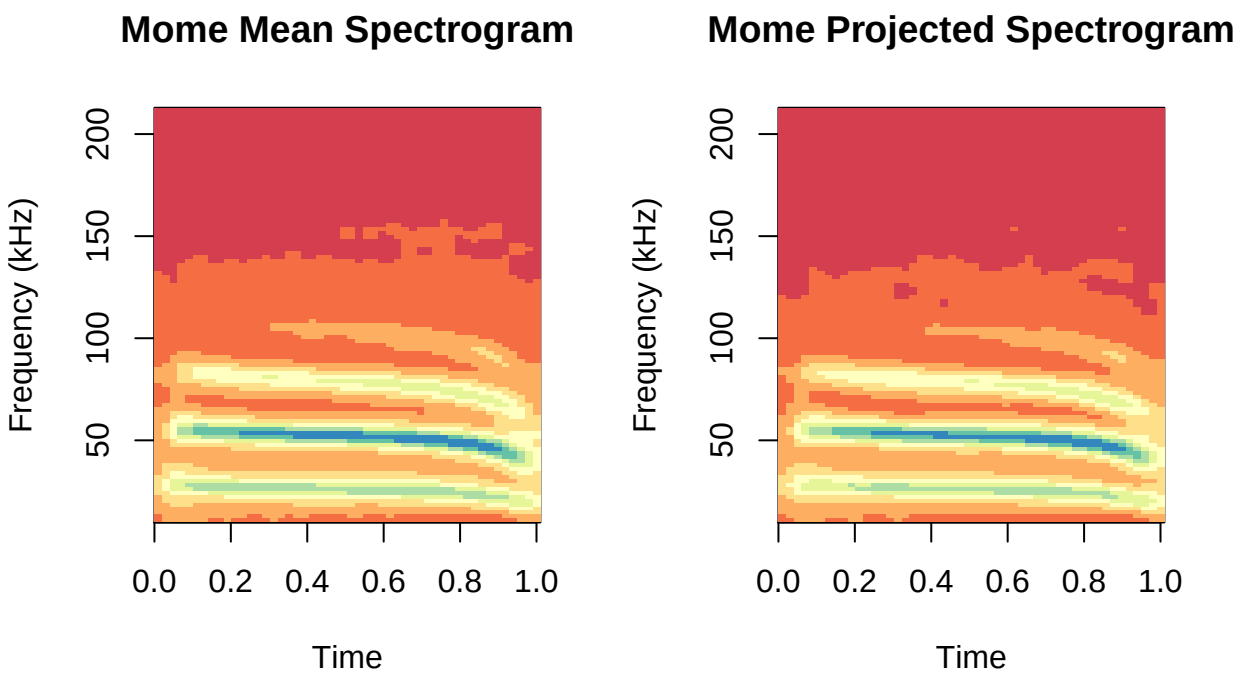
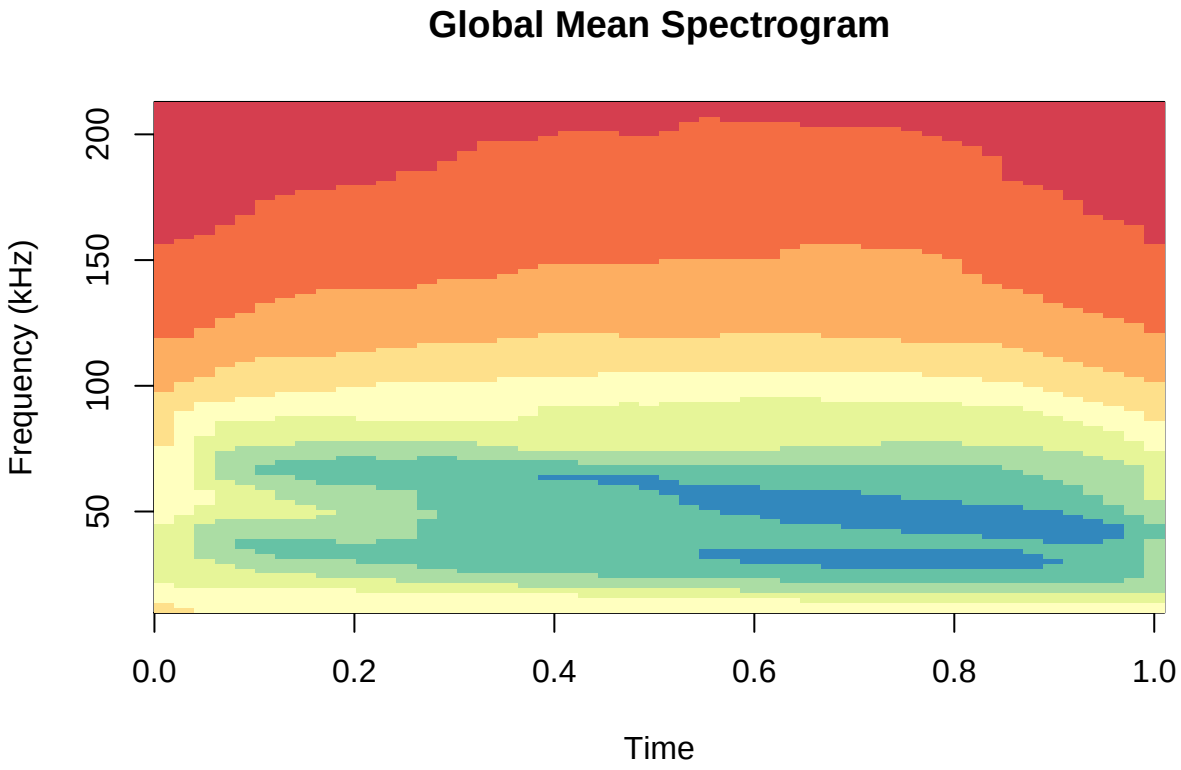


Figure 5: Plotted above the mean spectrogram for a randomly selected species alongside the call reconstructed from the component space. Based on this plot taking 15 components seems to result in excellent reconstructions of the mean species level calls, where all the salient features are preserved.

Exploration of Components




```
#dir.create("animation")
setwd("animation")

n <- 3

lambda <- seq(from = -2, to  = 2, length.out = 201)

component <- pca %>%
  use_series(rotation) %>%
  extract(, n) %>%
  array(dim = c(104, 50)) %>%
  multiply_by(pca %>% use_series(sdev) %>% extract(n))

for(i in seq_along(lambda)){
  png(file=paste('example', sprintf("%03d", i), '.png', sep = ''), width=800, height=600)
  component %>%
    multiply_by(lambda[i]) %>%
    add(global) %>%
    t %>%
    image(t, restricted_f, .,
      col = brewer.pal(9, 'Spectral'),
      xlab = 'Time (ms)', ylab = 'Frequency (kHz)',
      main = paste('Global Mean Spectrogram plus\n', lambda[i] %>% round(1),'times component' ,n))
  dev.off()
```

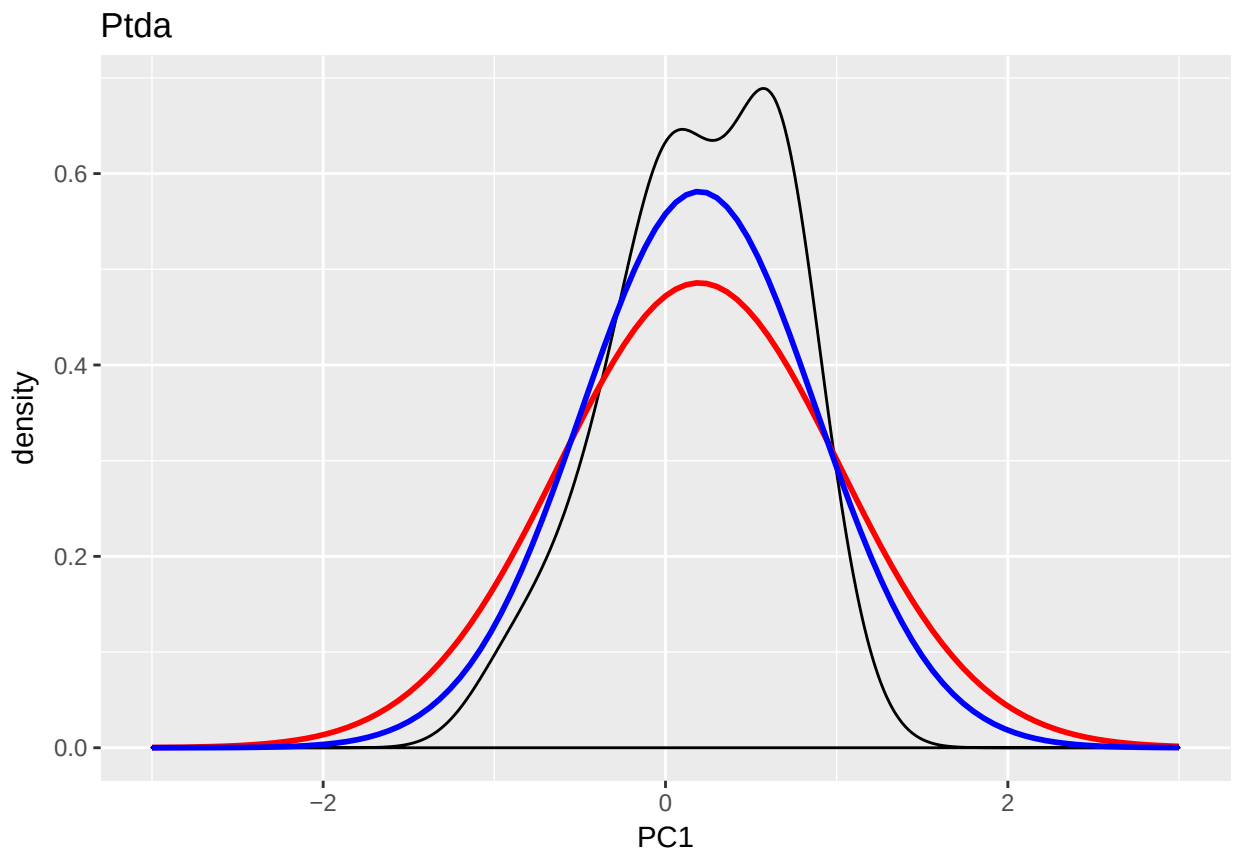


Figure 6: Shown above is a comparison of the empirical mean component score distribution (black) against the posterior predictive distribution for that score (red) alongside the posterior predictive distribution with the non-phylogenetic nois stripped out (blue). Plots of this nature demonstrate that while the empirical distribution tends to fall into the posterior predictive distribution, this is largely due to the posterior supporting a very wide region.

```
}

system('"C:/Program Files/ImageMagick-7.0.7-Q16/magick.exe" *.png -delay 50 pc3.gif')
file.remove(list.files(pattern=".png"))
```

Ancestral Reconstruction from Principal Components

The goal of this analysis is Ancestral Reconstruction of Echolocation Call spectrograms. Verifying ancestral calls is impossible, however perhaps we can get a feel for the effectiveness of the method by attempting an Ancestral reconsturction style posterior predictive spectrogram for exisiting species. This amounts to a leave one species out cross validation.

Ancestral Reconstruction

The final stage of this analysis is to perform an ancestral reconstruction for internal nodes of the tree.

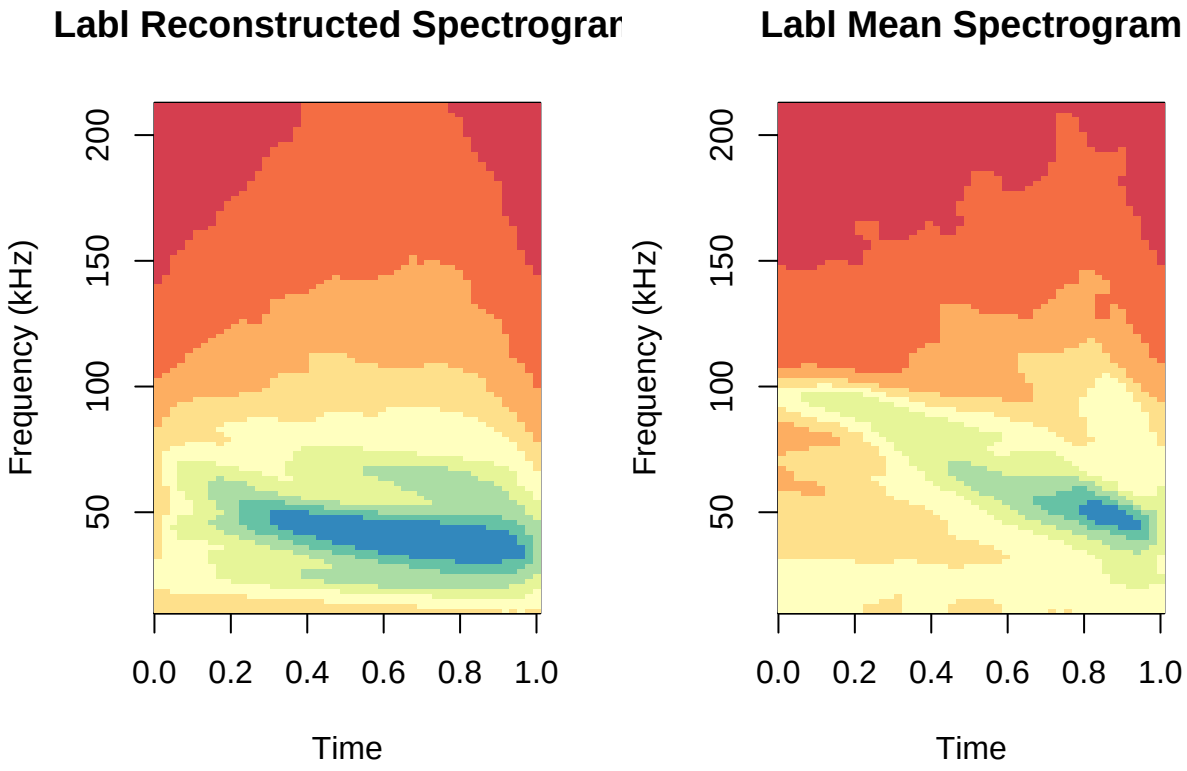


Figure 7: Illustrated here is a LOSO reconstructed spectrogram alongside the actual mean spectrogram for that species.

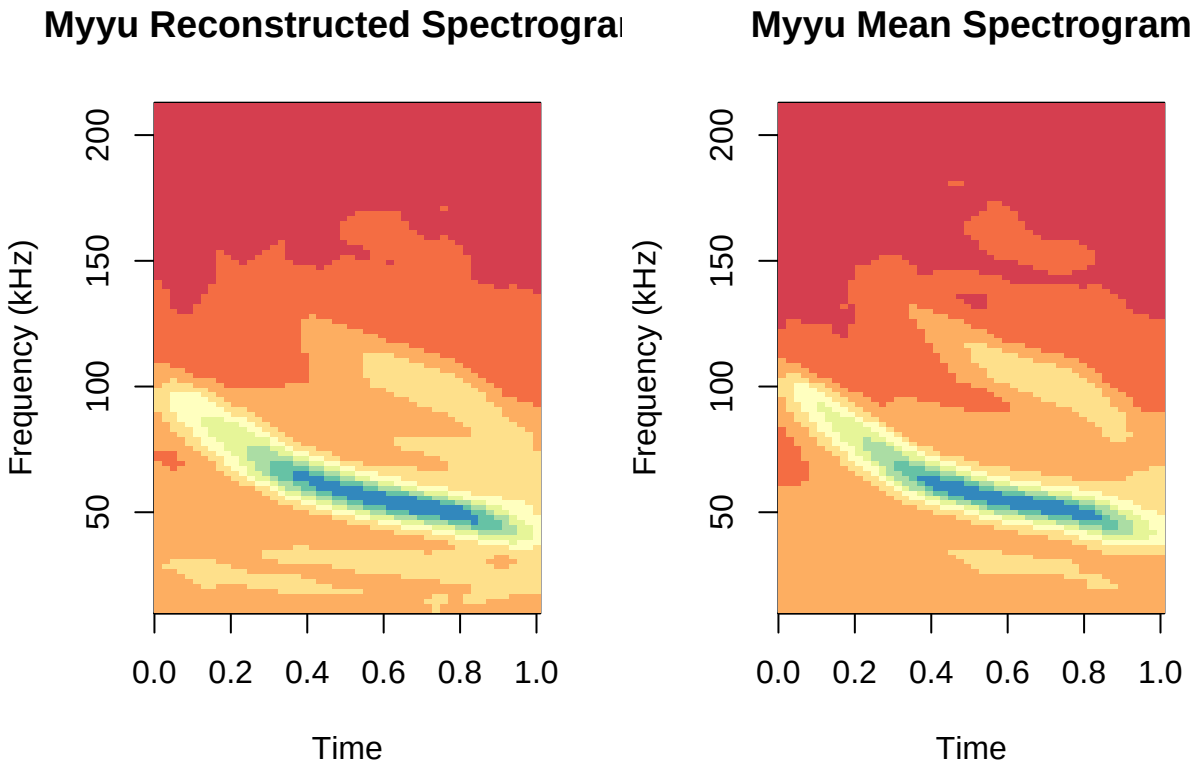


Figure 8: Tis is a comparison of the ancestrally reconstructed species level spectrogram, given observations from that species. This plot serves only as a sense check.

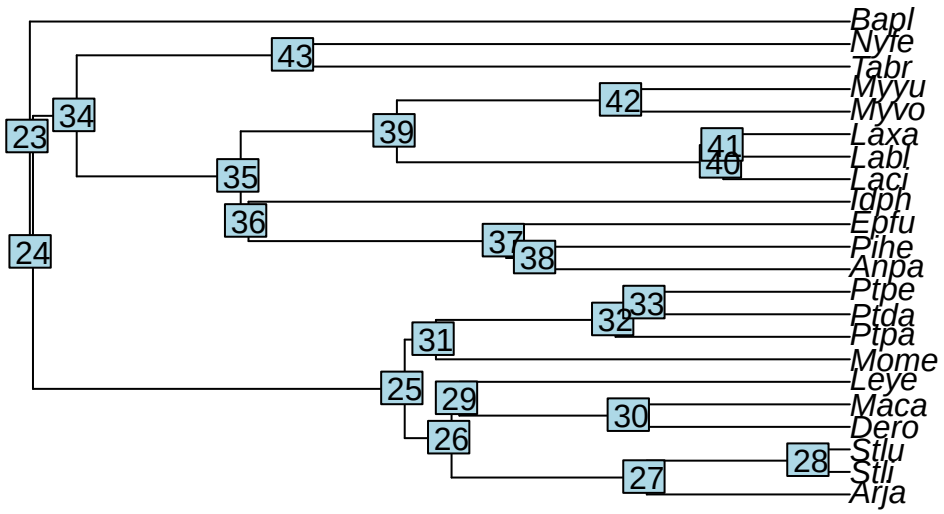


Figure 9: Figure showing node labels on the tree to figure out ancestral species labels

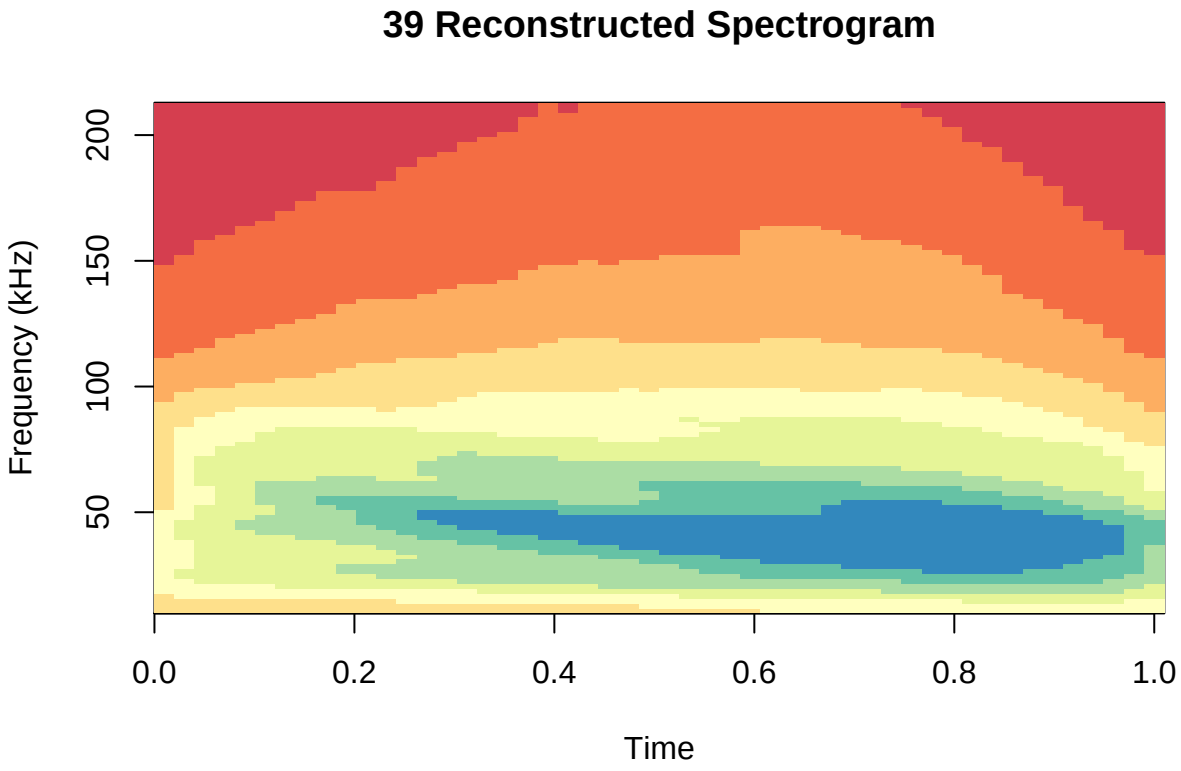


Figure 10: The Principal components reconstructed spectrograms can be produced for any point on the assumed Phylogeny.

Approximation of Spectrograms in the Time Domain

For the communication of my findings it will be important to have time domain representations of these spectrograms