CBER Presentation

J.P. Meagher

6 November 2017

The following is an outline of my work towards the ancestral reconstruction of bat echolocation calls, modelling this phenomenon by phylogenetic Gaussian processes.

This outline is to be presented to Kate Jones' lab group at the Centre for Biodiversity and Environment Research, UCL on Monday November 6th.

Packages

Packages used in this analysis are as follows:

```r
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
```

Raw Data

Two datasets provide the basis on which this analysis is performed, the first being post processed echolocation call data accompanying Stathopoulos et al. (2017). Live bats were caught, identified, and recorded at a sampling frequency of 500 kHz. In total the dataset consists of 22 species from five families, 449 individual bats and 1816 individual echolocation call recordings.

This data was placed in a tidy dataset (long form) where each call recording was considered to be an observation.

```
## 'data.frame':    1816 obs. of  5 variables:
##  $ bat    : Factor w/ 449 levels "1","2","3","4",..: 1 1 2 2 3 3 4 4 5 5 ...
##  $ species: Factor w/ 22 levels "Anpa","Arja",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ family : Factor w/ 5 levels "Emb","Mol","Mor",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ sex    : Factor w/ 2 levels "F","M": 1 1 1 1 2 2 2 2 1 1 ...
##  $ calls  :List of 1816
##   ..$ : num  -0.0119 -0.0566 -0.05708 0.00354 0.03007 ...
##   ..$ : num  0.00604 -0.05864 0.00343 -0.02058 -0.03251 ...
##   ..$ : num  -0.00833 -0.00116 -0.01149 -0.00938 -0.00727 ...
##   ..$ : num  -0.00464 -0.00488 -0.00601 -0.00711 -0.00314 ...
##   ..$ : num  -0.0457 -0.0272 -0.0282 -0.0363 -0.0366 ...
##   .. [list output truncated]
```

The bat super-tree in Collen (2012) provided the phylogenetic tree of the recorded bat species.

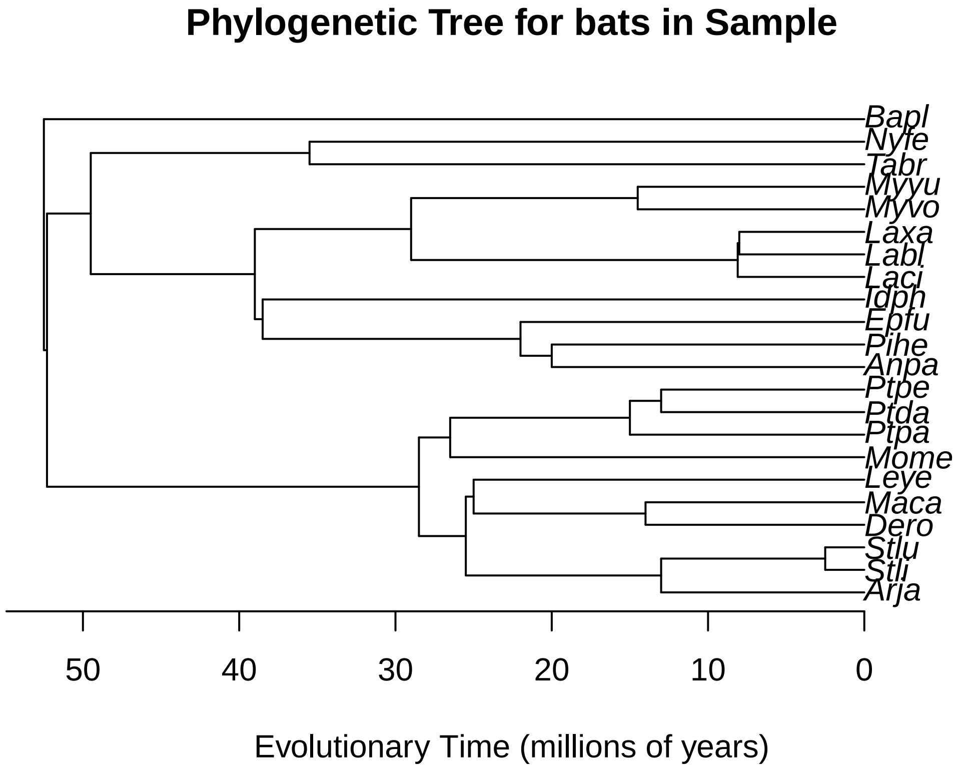


Figure 1: The phylogenetic tree describing the evolutionary relationships between bat species within the sample.

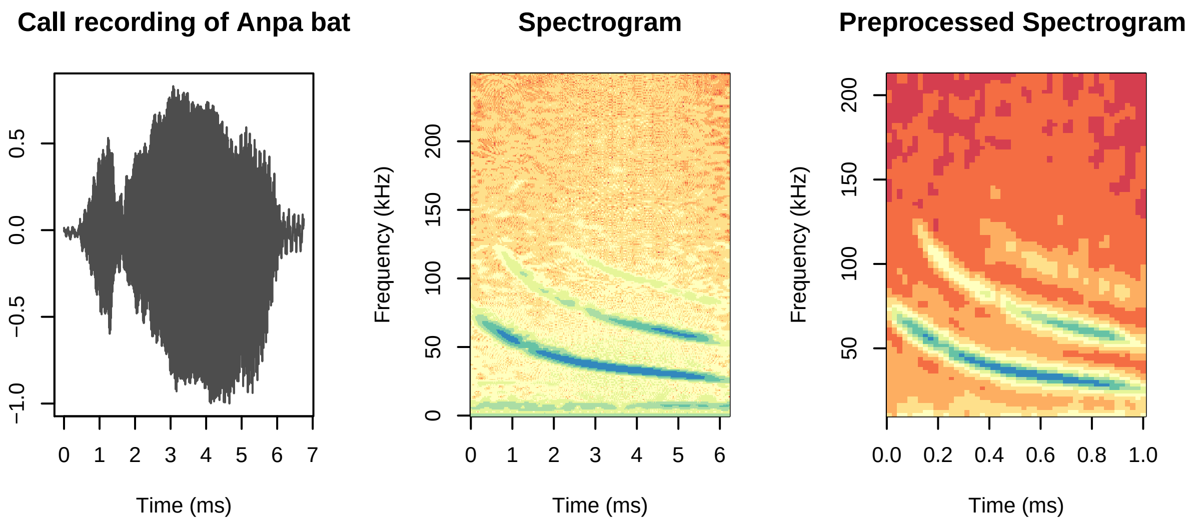


Figure 2: Plotted above is the call waveform, the raw spectrogram, and the preprocessed spectrogram surface for a randomly selected observation in the dataset.

Preprocessed Data

Echolocation call recordings must be preprocessed before further analysis. A standard, informative technique for the analysis of acoustic signals is to obtain a time-frequency representation, the spectrogram. The spectrogram is calculated by a Short Time Fourier transform of the signal with overlapping windows. Spectrograms were calculated by fourier transforms of size 512, with a Hamming window of size 256, and 96.875% overlap.

The range of frequencies used by bats for their echolocation calls is [9, 212] kHz, and so spectrograms can safely be restricted to this range for analysis (Stathopoulos et al. 2017).

This analysis then considers the spectrograms produced to be functional data objects, in this case surfaces. This involves smoothing the surfaces and then mapping them all on to an absolute time scale by a combination of dynamic time warping and parwise synchronisation. This stage of the analysis was performed in Matlab due to the availability of effective algorithms on this platform. Smoothing was done using a weighted robust spline smoothing algorithm for 1-D to N-D data (Garcia 2010) (Garcia 2011), where weighting was used to preserve detail at spectrogram peaks. Dynamic time warping (Damoulas et al. 2010) was used adapt pairwise curve synchronisation (Tang and Müller 2008) to the spectrograms.

The preprocessed data was then ported into R and added as a column to the raw call dataset.

Mean Spectrogram

The preprocessed spectrograms, or spectrogram surfaces, have been reported on a regular 104 × 50 grid. This makes the calculation of summary statistics relatively straightforward.

An important statistic is the mean spectrogram for each species. In order to estimate an unbiased mean, the mean spectrogram for each bat will be estimated first. The species mean spectrogram will be estimated from these individual bat mean spectrograms.

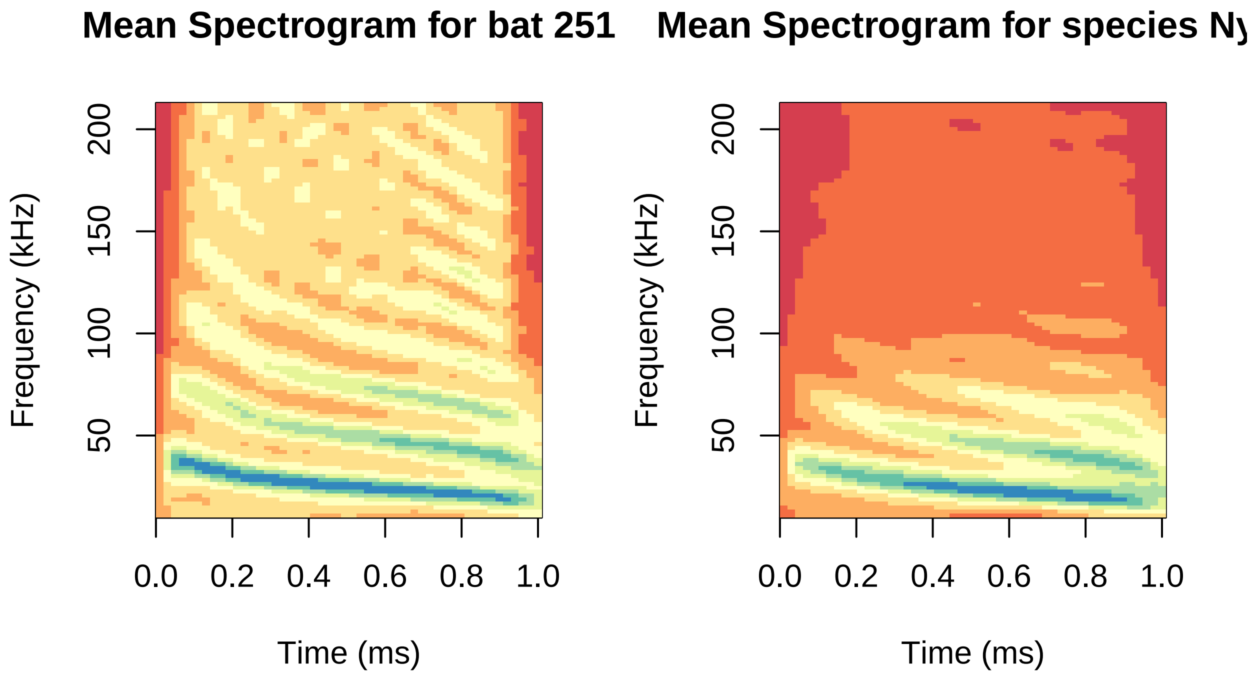


Figure 3: Plotted above the mean spectrogram for a randomly selected individual bat and a randomly selected species of bat. These plots serve as a sense check on the data.

Evolutionary Features

By making some assumptions about the nature of the data, namely that the call spectrograms and evolutionary process driving them can be considered Gaussian processes, spectrogram features, the weights of which evolved independently can be identified. This can be done by a principal components analysis of the call spectrogram data.

The data was tested to investigate whether or not it could be modelled as being separable in time and frequency. This assumption was found to be inappropriate and so a PCA was performed on the full, flattened, preprocessed spectrograms.

PCA Scores

See Figure 4.

PCA Loadings

These loadings provide the components by which we model echolocation as having evolved.

A visual inspection of these components allows them to be interpreted to some degree especially when examined in terms of its impact on the global mean spectrogram.

The model for evolution implies that there exists a global mean spectrogram for all bat species, see Figure 5. By by adding a weighted component to this surface some intuition on the impact of the component can be obtained.

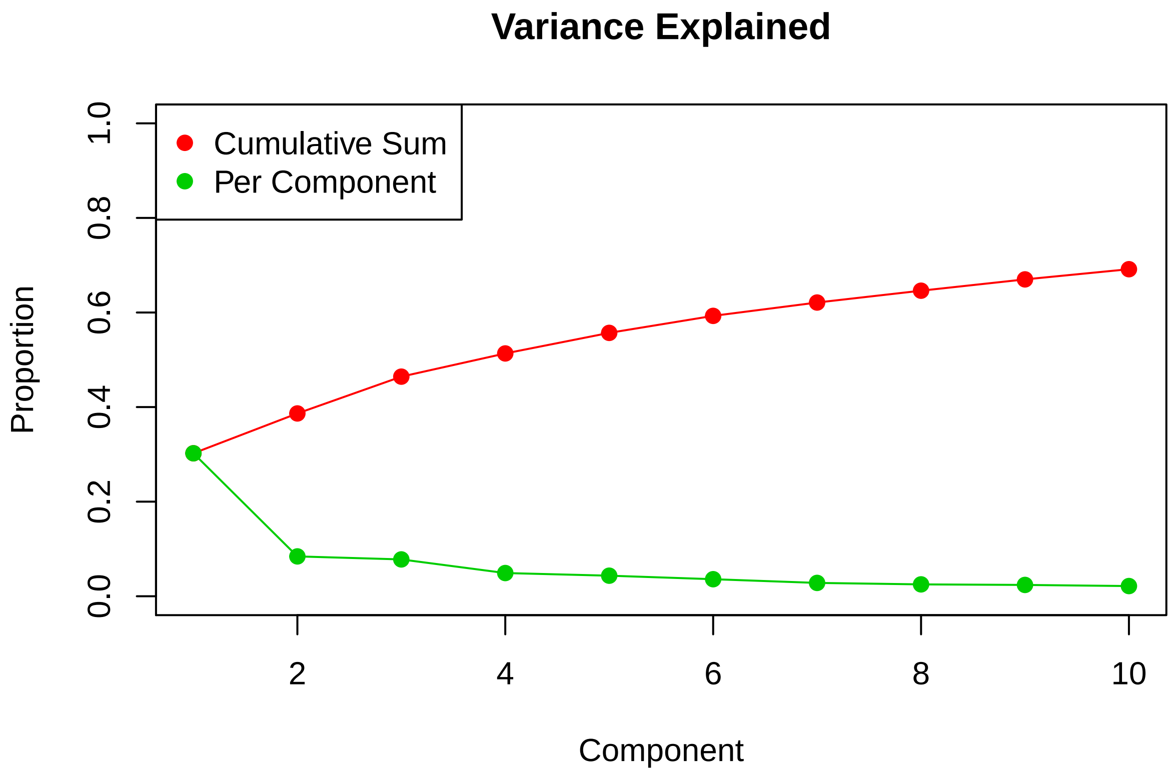


Figure 4: The scores associated with each component provide insight into the proportion of the sample variance captured by the corresponding principal component. The scores above suggest that the dataset of echolocation call spectrograms is a very high dimensional, requiring 3 principal components to capture even 50% of the sample variance. It can be shown that 8 components capture more than 2.5% of the variation, and 15 capture more than 1% of the variation. 13 components are required to explain 75% of the variance, while 51 are required to explain 90%

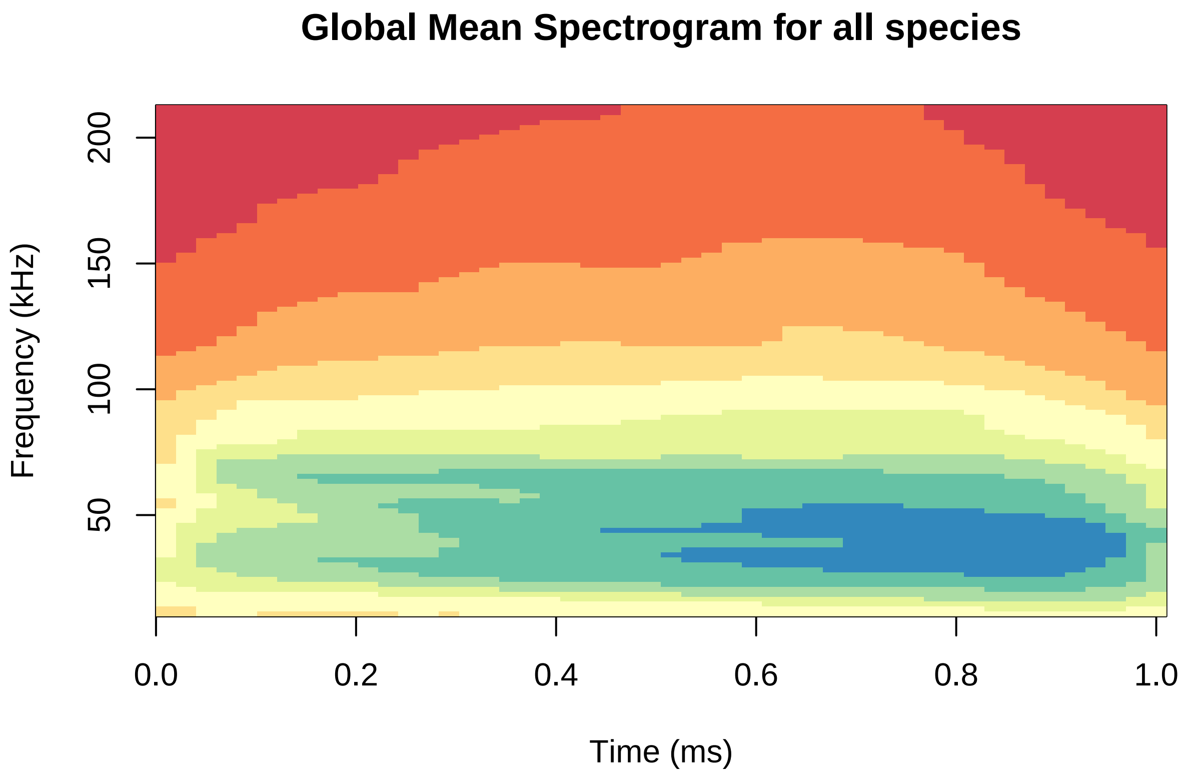


Figure 5: The global mean echlocation call spectrogram implied by the model. I believe that this will prove to be the 'best guess' at the ancestral bat lying a the root node of the phylogenetic tree.

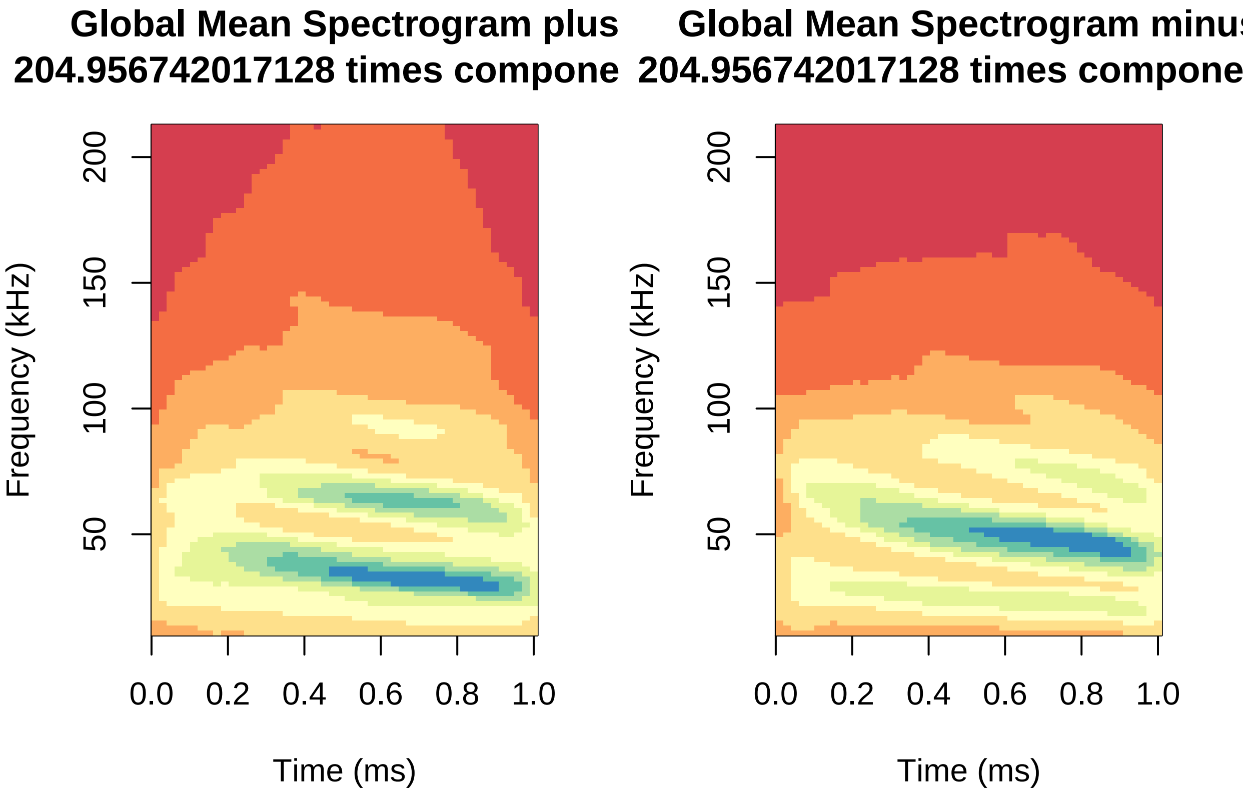


Figure 6: By adding or subtracting a component to the global mean spectrogram we gain insight into the effect of this component in evolutionary terms.

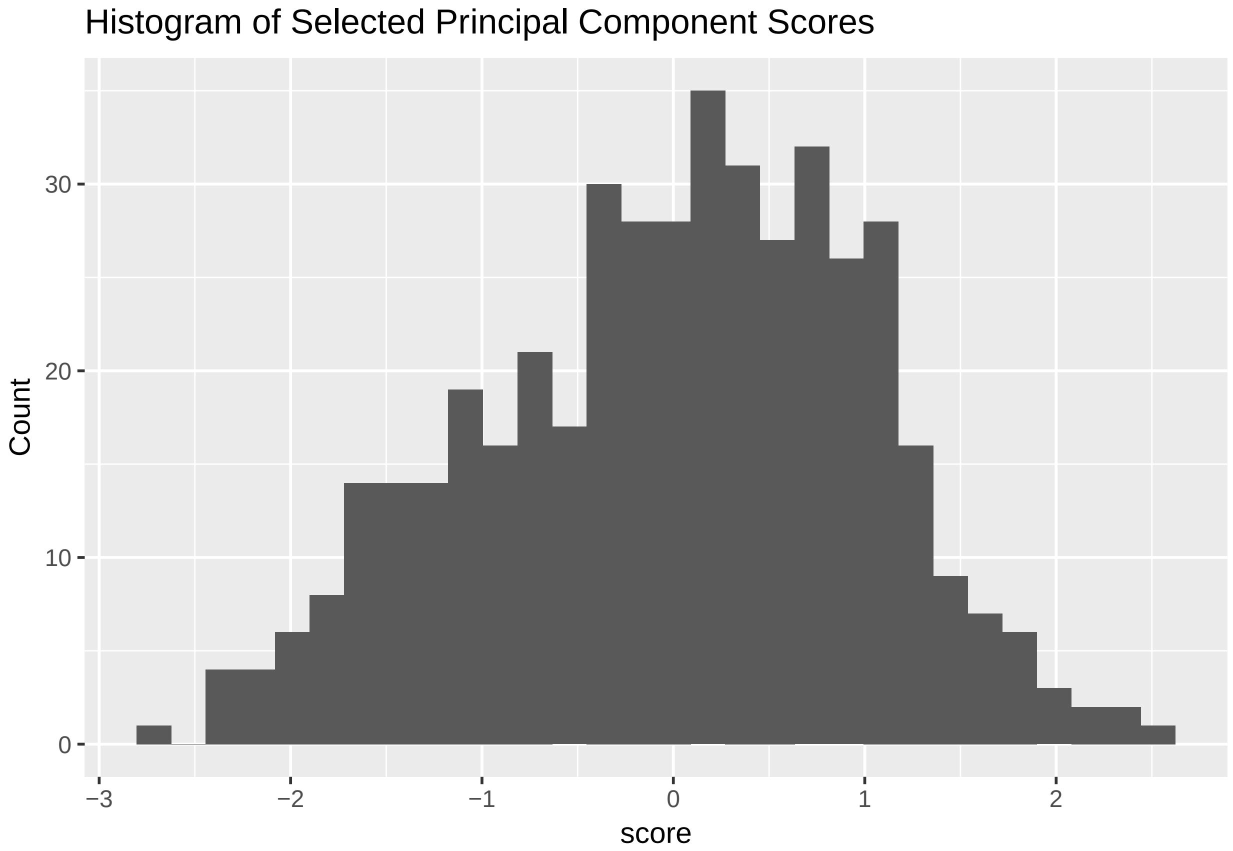


Figure 7: Distribution of component scores for a selected Principal component loading.

Component Scores for each Bat

These components can be used to find scores for every echolocation call. In this case I will consider the mean call of each individual bat as an observation and find the associated scores for the first 15 principal components.

These scores can then by explored through various plots.

Visual inspections of the score distributions do not seem to indicate particularly strong phylogenetic relationships within the Principal Component Scores of each individual bats mean spectrogram.

Bootstrapped mean estimates

When performing this analysis for spectral density curves, bootstrapped estimates of the mean species call were required. This approach will be attempted here also.

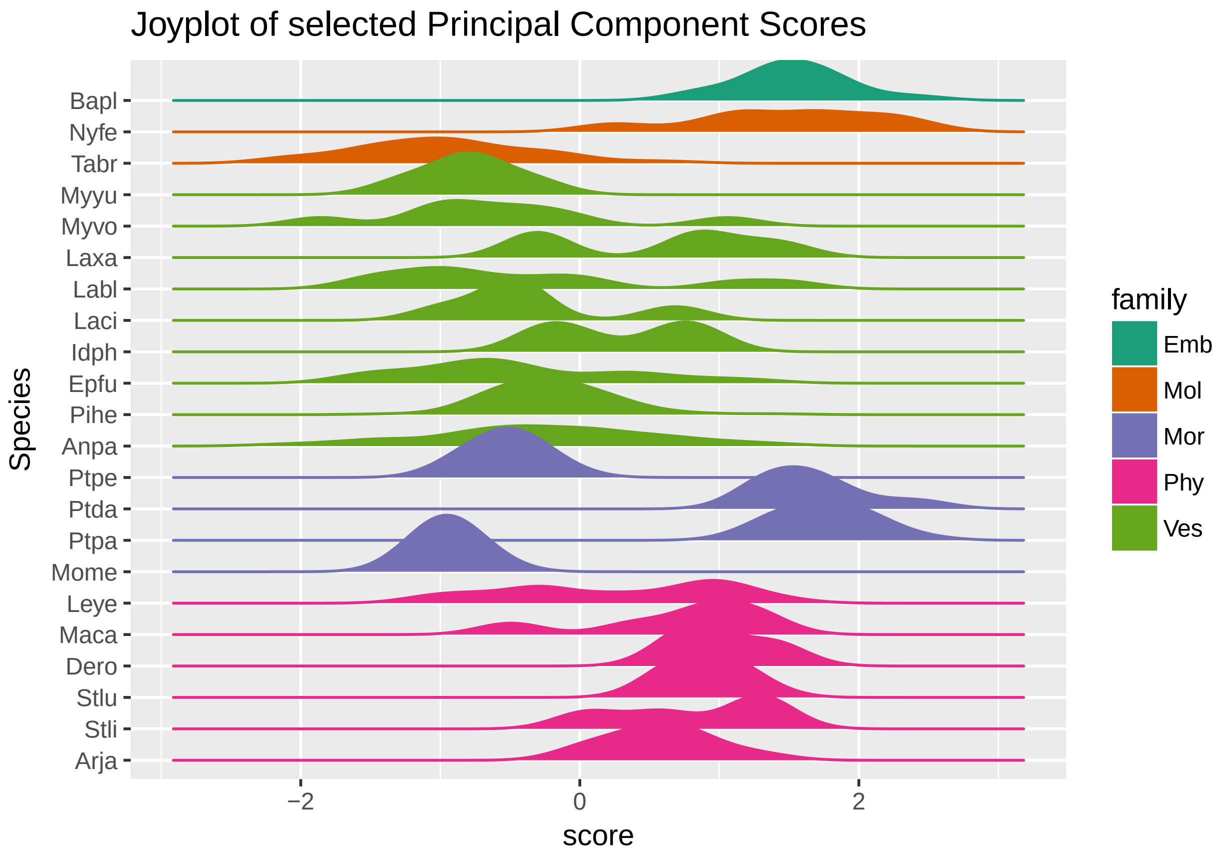
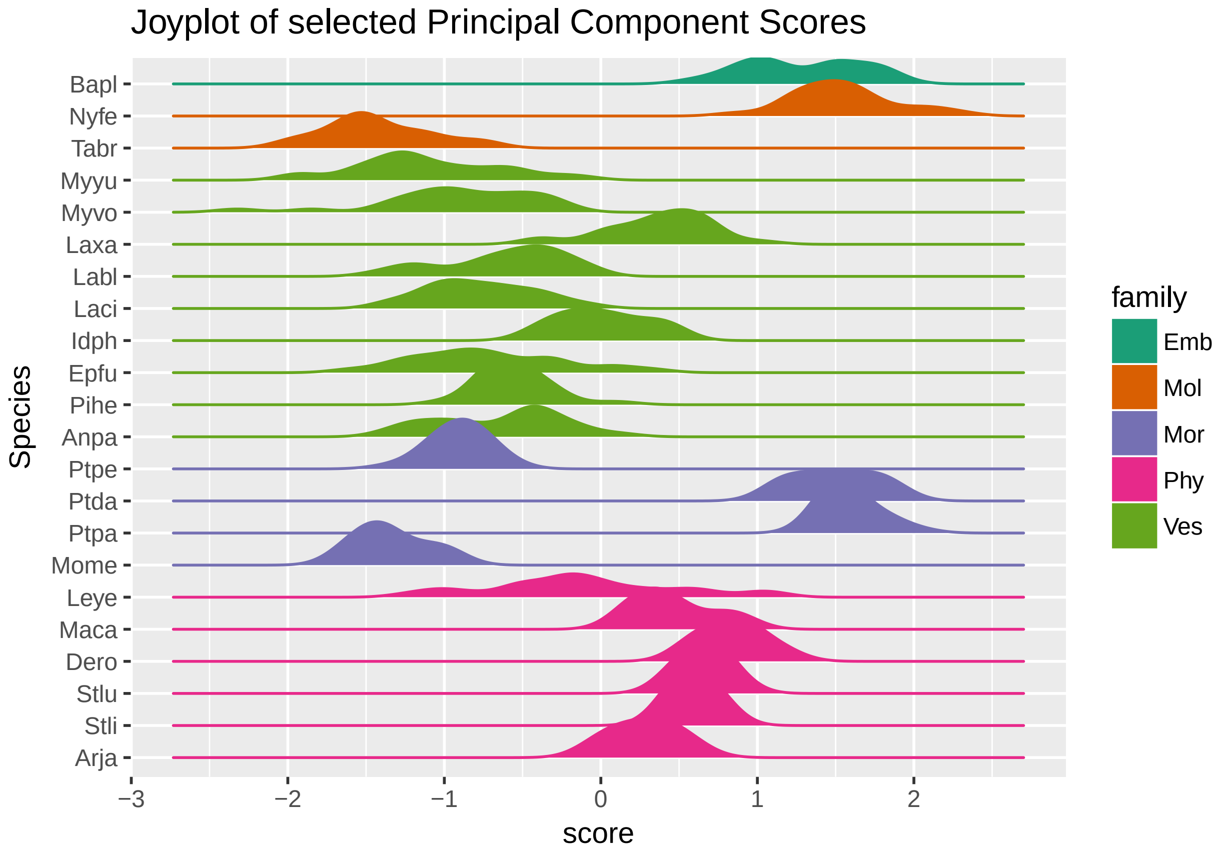
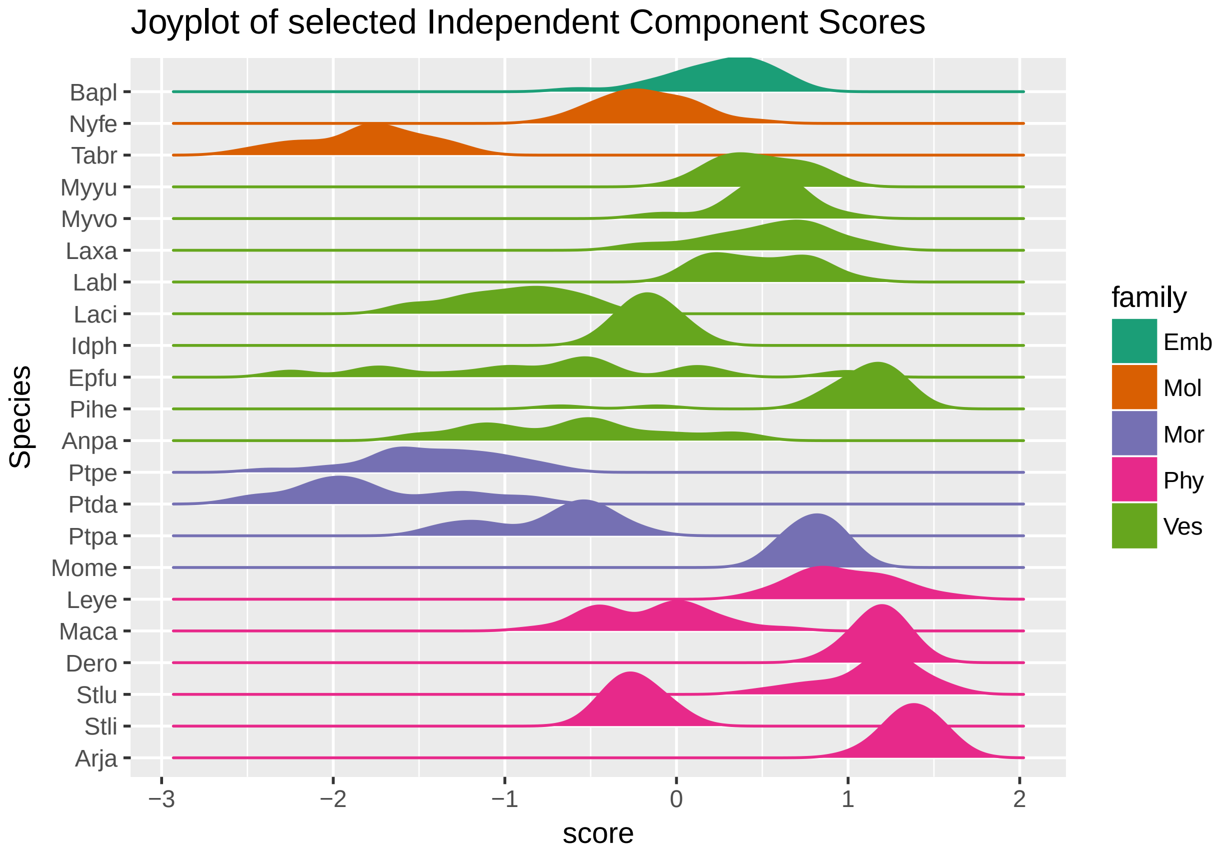


Figure 8: Distribution of component scores for a selected Principal component loading, separated out by species.



Independent Components

Another set of evolutionary features can be estimated by an Independent Components Analysis of the data, this may provide more interpretable features with clearer phylogenetic relationships.



Evolutionary Inference

Ancestral reconstruction can be performed by finding the posterior predictive distribution of the Phylogenetic Gaussian Process. In order to do this we must first identify appropriate hyperparameters for the PGP. This can be performed by type II maximum likelihood estimation of the hyperparameters. This can be obtained by optimisation, or with a MCMC chain, however in this case it may be more informative to simply do a grid search over the space of the hyperparameters. This will allow a better understanding of the hyperparameter space rather than simply reporting some values that produce a local optimum.

References

Collen, AL. 2012. “The Evolution of Echolocation in Bats: A Comparative Approach.” PhD thesis, UCL (University College London).

Damoulas, Theodoros, Samuel Henry, Andrew Farnsworth, Michael Lanzone, and Carla Gomes. 2010. “Bayesian Classification of Flight Calls with a Novel Dynamic Time Warping Kernel.” In *Machine Learning and Applications (Icmla), 2010 Ninth International Conference on*, 424–29. IEEE.

Garcia, Damien. 2010. “Robust Smoothing of Gridded Data in One and Higher Dimensions with Missing Values.” *Computational Statistics & Data Analysis* 54 (4). Elsevier: 1167–78.

———. 2011. “A Fast All-in-One Method for Automated Post-Processing of Piv Data.” *Experiments in Fluids* 50 (5). Springer: 1247–59.

Stathopoulos, Vassilios, Veronica Zamora-Gutierrez, Kate E Jones, and Mark Girolami. 2017. “Bat

Echolocation Call Identification for Biodiversity Monitoring: A Probabilistic Approach.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. Wiley Online Library.

Tang, Rong, and Hans-Georg Müller. 2008. “Pairwise Curve Synchronization for Functional Data.” *Biometrika* 95 (4). Oxford University Press: 875–89.