Supervision Meeting Demo

J.P. Meagher

8 November 2017

Here is a summary and extension of the results I presented to Kate at the CBER on Monday 6th November 2017.

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      7    Ves    Anpa <dbl [104 × 50]>
```

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.

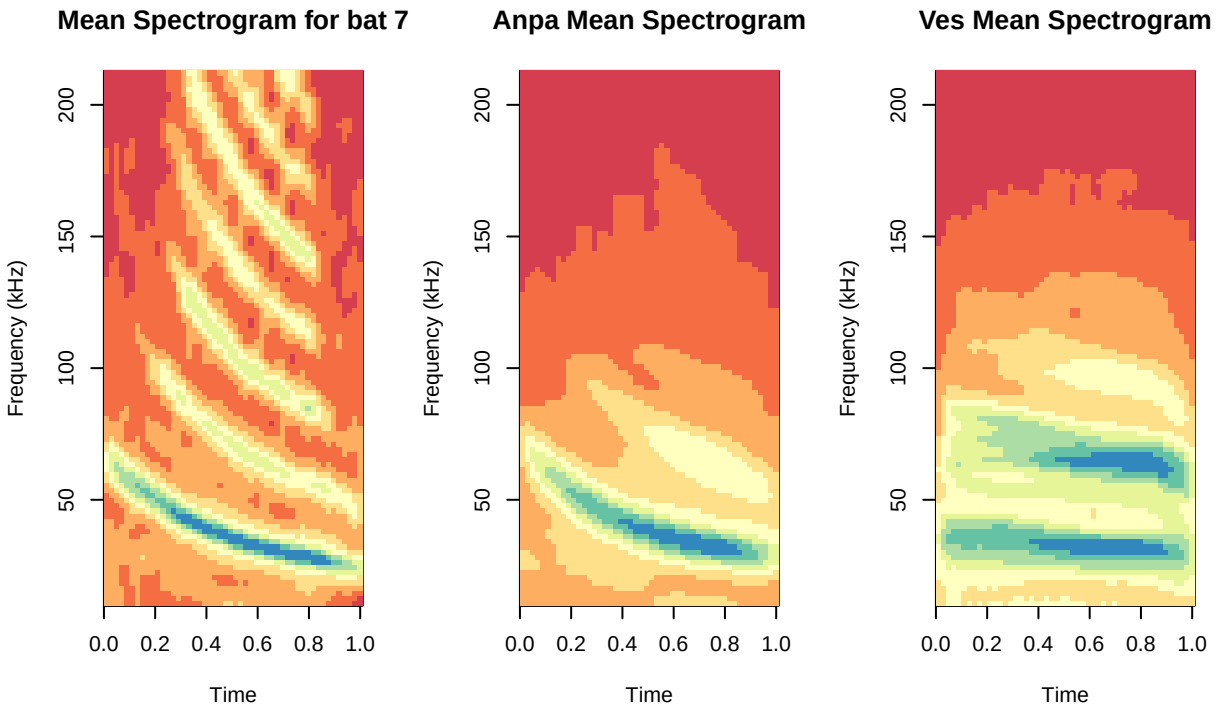

Figure 1: Plotted above the mean spectrogram for a randomly selected individual bat along with the species and family level mean spectrograms for that individual. These plots serve as a sense check on the data.

PCA

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.

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?

ICA

An Independent Components Analysis offers an alternative suite of evolutionary features, however, this method forces the analyst to choose the number of components desired and this choice will affect the structure of Independent Components selected. If the PCA revealed a definite number of modes of variation then selecting the number of independent components is simplified, but in this example this is not the case. Looking at various numbers of independent components and simply picking the 'best set' that is problematic, what would constitute the best set?

Despite these misgivings, 15 independent components will be identified and the proportion of variance in the sample explained by these components explored,

It seems that the sum of the variance of the orthogonal Independent Component scores is less that that for the principal components. I believe this indicates that the independent components capture less of the samples variation than the principal components. It is worth noting however that no independent component

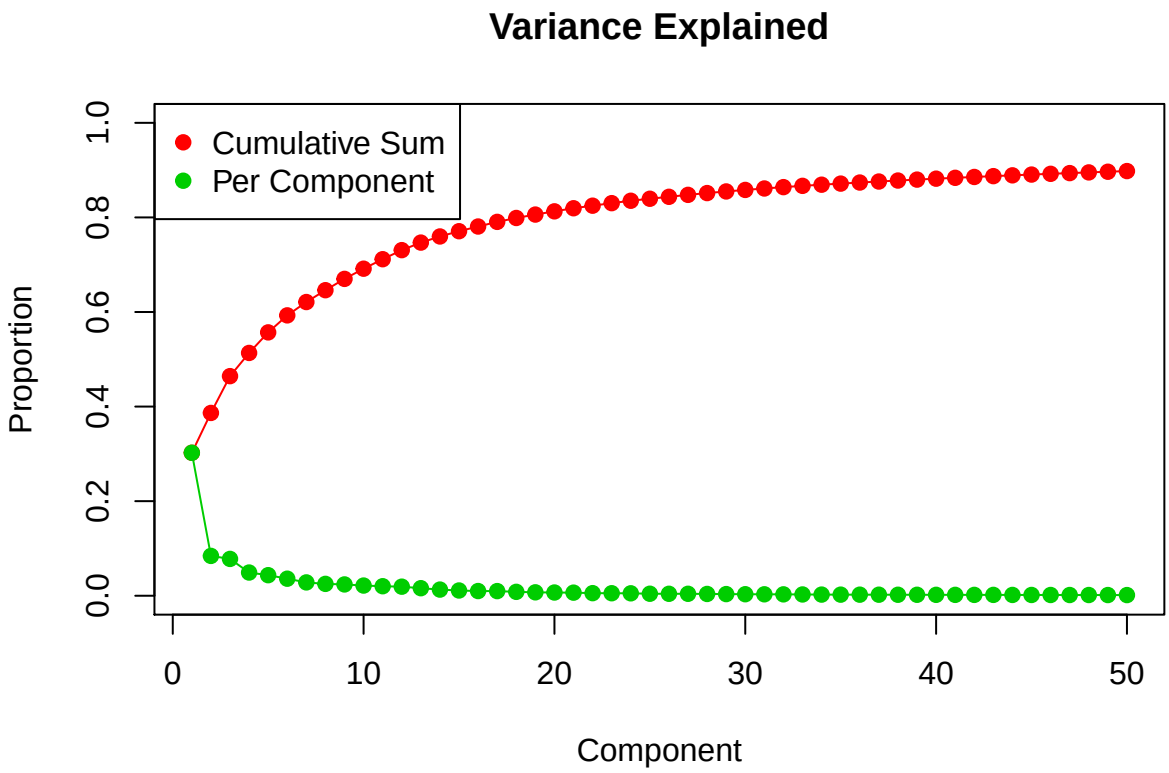

Figure 2: The scores associated with each component provide insight into the proportion of the sample variance captured by the corresponding principal component. The scores above suggest that the dataset of echolocation call spectrograms is a very high dimensional, requiring 3 principal components to 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%

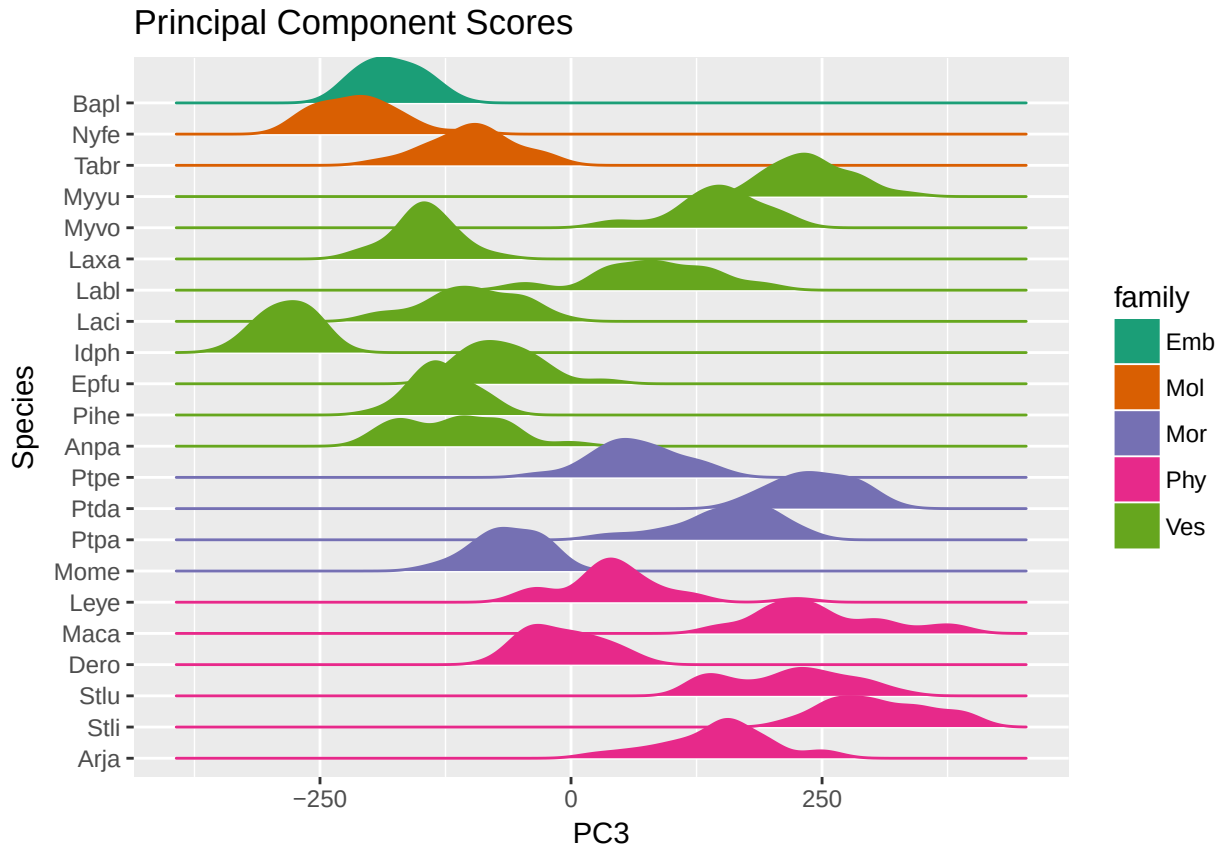

Figure 3: Distribution of component scores for a selected Principal component loading, separated out by species.

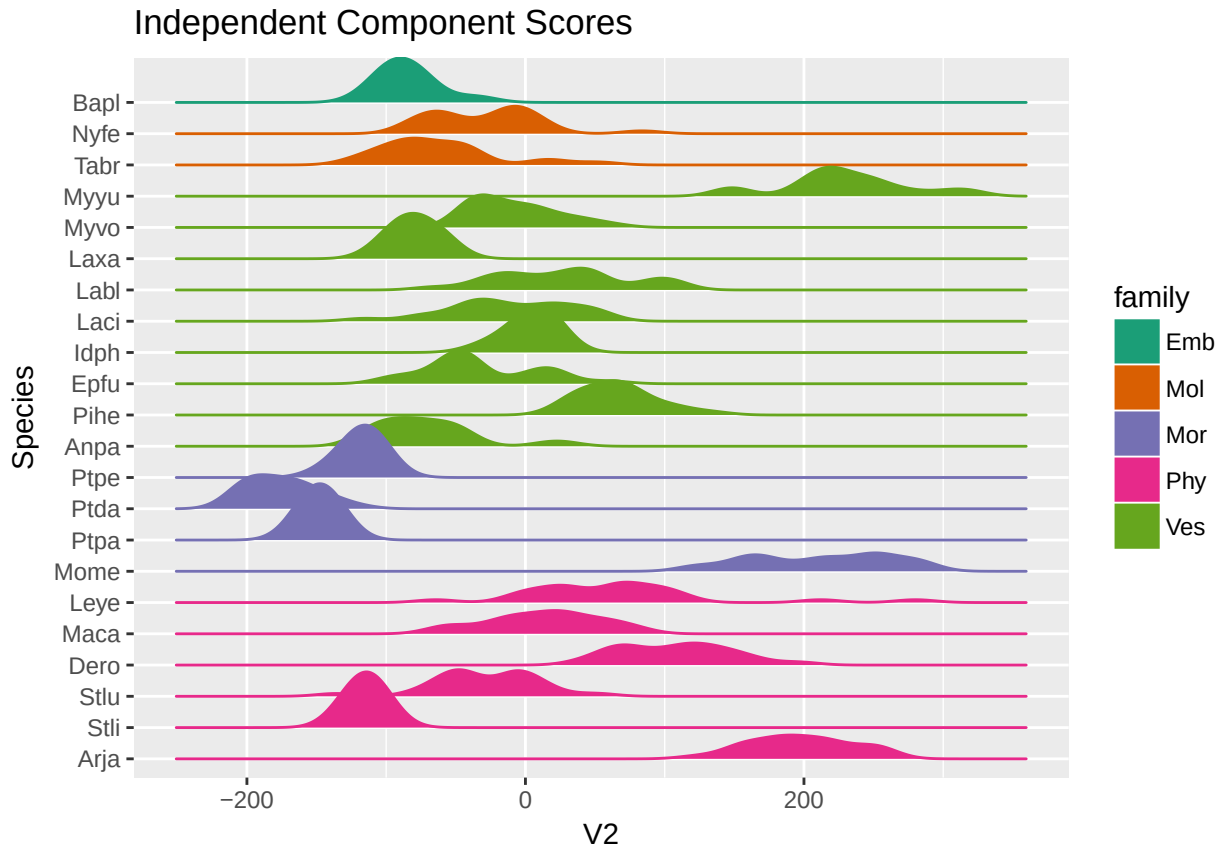

Figure 4: Distribution of component scores for a selected Independent component loading, separated out by species.

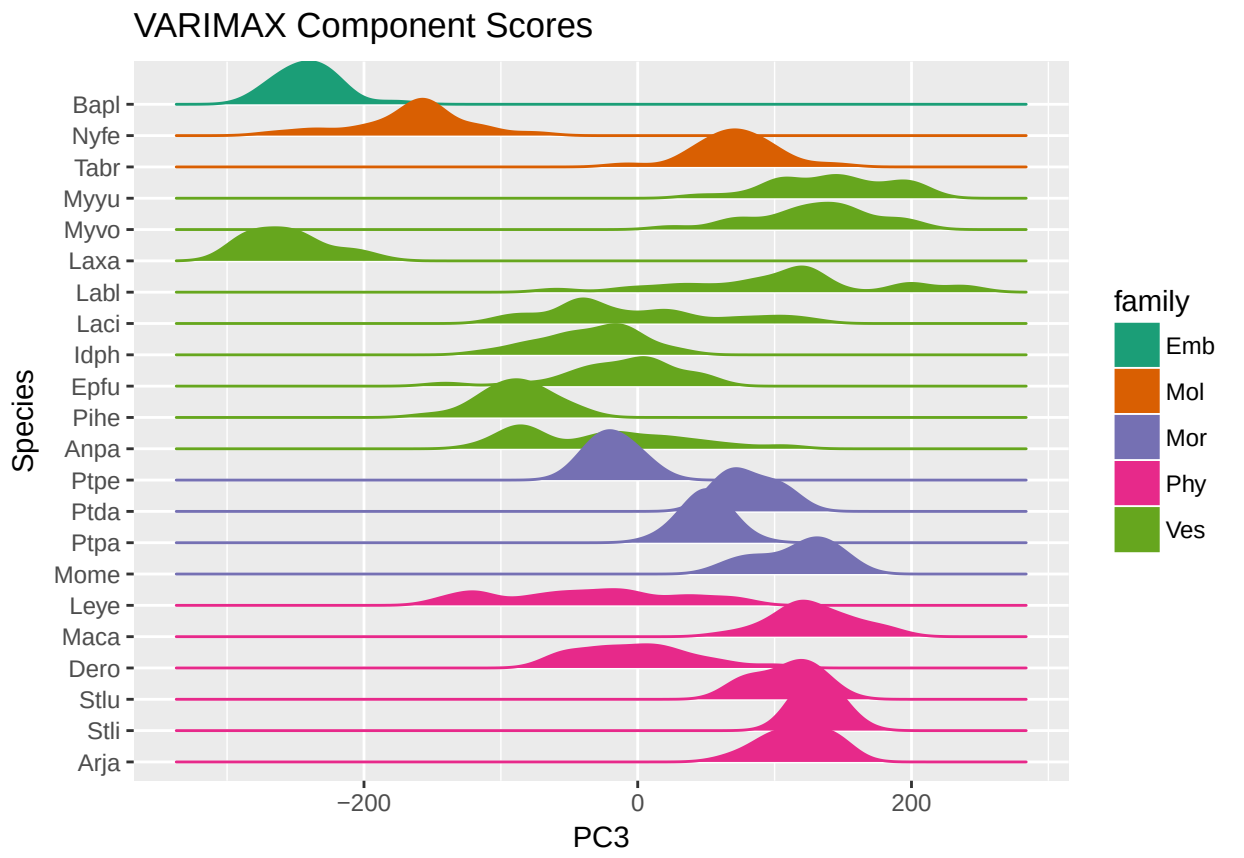

Figure 5: Distribution of component scores for a selected VARIMAX component loading, separated out by species.

is especially dominant in terms of the proportion of variation it captures. This may be a desirable property of evolutionary features.

VARIMAX

Another set of alternative components can be obtained by a VARIMAX rotation of the principal components. A varimax rotation rotates components such that the sum of the variances of the squared loadings is maximised. This results in loadings where weights at a particular point are encouraged to be either very large or very small.

The VARIMAX scores do a much better job at preserving the proportion of variation explained than the independent components, However the scores associated with VARIMAX components are not uncorrelated. I an not fully sure if the scores associated with Independent Components are or are not uncorrelated.

From a theoretical perspective the PCA components are sufficient, however, it may be that when it comes to applying these methods, ICA or VARIMAX offers a better solution.

Estimating Hyperparameters

Type II Maximum Likelihood Estimation can be performed on the Component Scores to estimate hyperparameters of the phylogeneetic Gaussian process for evolution. The Principal component scores can be treated as independent Phylogenetic Gaussian processes.

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

```
##         PC1   PC2   PC3    PC4    PC5    PC6    PC7    PC8    PC9   PC10
## s_p    0.80  0.98  1.00   0.96   0.97   0.95   0.93   0.93   0.84   0.93
## l     34.60 18.17 44.15   9.34  23.26   0.59   4.31   0.41   2.01   0.37
## s_n    0.57  0.25  0.26   0.31   0.30   0.31   0.35   0.36   0.54   0.37
## logl 512.68 75.92 98.40 190.36 184.25 204.62 257.26 277.15 488.17 288.68
##        PC11   PC12   PC13   PC14   PC15
## s_p    0.91   0.88   0.87   0.79   0.71
## l      5.17  11.06  20.13   0.43   5.48
## s_n    0.41   0.48   0.50   0.61   0.69
## logl 344.96 423.44 444.85 545.98 613.51
```

```
## [1] "IC MLE HYP"
```

```
##         V1     V2     V3     V4     V5     V6     V7     V8     V9    V10
## s_p   0.97   0.95   1.08   0.92   0.93   0.91   0.96   1.45   0.91   0.99
## l    26.04  12.59  49.04   0.37   8.10  12.32  22.76 164.13   0.33  33.24
## s_n   0.25   0.32   0.29   0.39   0.37   0.35   0.29   0.29   0.42   0.32
## logl 87.60 208.00 162.65 312.53 284.92 261.20 167.53 155.70 353.60 219.43
##         V11    V12    V13    V14    V15
## s_p    0.96   0.96   0.87   0.87   0.93
## l     19.00  28.64   4.67   8.45   6.88
## s_n    0.30   0.43   0.48   0.47   0.34
## logl 175.62 371.53 423.65 415.14 251.65
```

```
## [1] "VARIMAX MLE HYP"
```

```
##         PC1   PC2    PC3    PC4    PC5    PC6    PC7    PC8    PC9   PC10
## s_p    0.78  0.98   0.94   1.00   0.91   0.95   0.96   1.03   0.99   0.93
## l     16.44 20.47   6.63  40.31   0.09  17.04  18.27  38.45  83.54   0.43
## s_n    0.64  0.23   0.32   0.32   0.42   0.32   0.31   0.35   0.42   0.35
## logl 577.88 41.04 215.94 215.32 354.52 206.63 197.61 258.11 351.28 263.69
##        PC11   PC12   PC13   PC14   PC15
## s_p    0.95   0.79   0.78   0.93   0.86
## l     26.59  34.11  23.36  12.31  33.81
## s_n    0.44   0.56   0.63   0.31   0.36
## logl 382.39 499.31 566.17 200.20 266.27
```