Multivariate analysis for species differentiation

- Examine the species (striped dolphins and sperm whales) differentiation in the multivariate space of click characteristics.
- Study how each click characteristic contributes to the species differentiation.
- Choose the significant parameters/features to be used in classification.

Where we're at...

- Our labeled dataset consists of 347 striped dolphin (SD) and 184 sperm whale (SW) audio files. The duration of each audio file is approximately 4 – 6 seconds.
- Clicks are detected for each audio file in our dataset, using the Click Detector PAMGuard module. The output of the Click Detector module, i.e. the detected clicks per audio file, is stored in .pgdf binary files.
- The information in the .pgdf files is loaded in R using the PAMpal package. For each audio file, i.e. event, in our dataset, a number of clicks, i.e. audio segments, is loaded.
- These click segments are parts of the signal that have their own time domain characteristics (e.g. duration) and spectral characteristics (e.g. peak frequency), which are easily calculated using the PAMpal functions.

Multivariate analysis objectives (1/2)

- We want to devise a classification algorithm that predicts the label of a given audio file recording, using the click detections produced as PAMGuard output.
- The classification parameters are going to be fit on our dataset. Thus, the classification success depends on the degree in which the click detections differ between striped dolphin (SD) and sperm whale (SW) recordings.
- Before proceeding to the classification, we need to **examine if and how the per event/file click characteristics differ for the two species that are present in our dataset** (SDs and SWs).

Multivariate analysis objectives (2/2)

- In order to examine the species differentiation in the multivariate space of click characteristics:
  1. The PAMGuard detected clicks are loaded in R. The click characteristics of each detected click signal/waveform are calculated using PAMpal functions. The detected clicks are grouped in events, according to the dataset audio file that they correspond to.
  2. The mean value of each click characteristic is calculated in each event. The per species distribution of each click characteristic is visualized using boxplots.
  3. A series of statistical tests that examine group/species differentiation, as well as feature dimensionality reduction via principal component analysis (PCA), are performed.

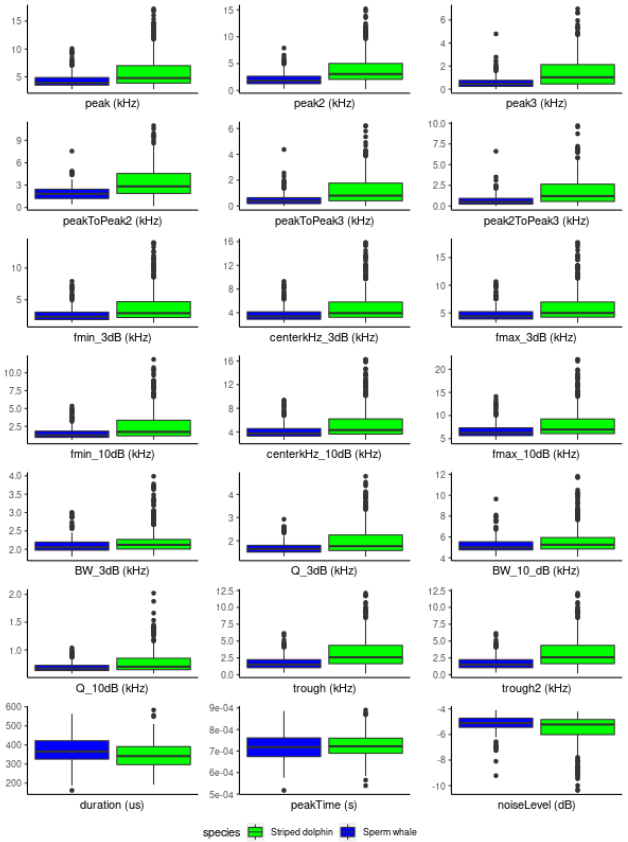
Per species distribution of click characteristics (1/3)

- The distribution of click characteristics is naturally affected by the quantity and quality of the detected clicks.
- The Click Detector PAMGuard module uses a set of parameters to detect the high amplitude samples of the signal as clicks.
- Adjusting the detection parameters (e.g. the threshold over which detection is applied) affects the detection strictness.
- A strict detection detects less clicks of higher quality, while a lax detection detects more clicks but of lower quality (e.g. more samples are detected as clicks even if they’re not).

Per species distribution of click characteristics (2/3)

- The click characteristics are aggregated on an event basis, e.g. their mean value is calculated in each event.
- Statistical inferences are more reliable on a big high quality sample, but increasing the detection quantity reduces the detection quality.
- This tradeoff between detection quantity and quality was explored via the exploration of the detection parameters in PAMGuard.
- Increased values for the detection threshold (default 10dB) and the noise weight factor α_N (default 1E-5) were tested, in order to make the detection more strict.

Per species distribution of click characteristics (3/3)

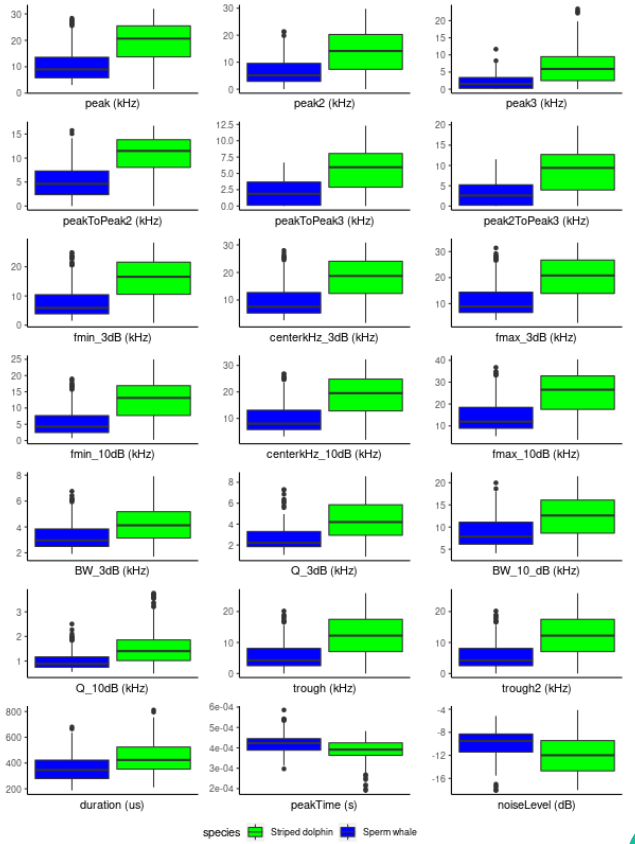


Left: default click detection

Right: optimal click detection (α_N = 1E-2)

The default click detection is lax. Many clicks are detected. The per species distribution of the click characteristics is not reliable, many outliers are detected and there exists overlap between the two species distributions.

The optimal click detection is more strict (increased value for α_N). The number of outliers is reduced and the distribution of the click characteristics differs between species. The spectral characteristics have higher values for striped dolphins, as expected.



Statistical testing for species differentiation (1/2)

- The boxplot visualization of the per species distribution for each click characteristic highlights the individual click characteristics for which the two species differ.
- For instance, the peak frequency distribution is centered around 10kHz for sperm whales and around 20kHz for striped dolphins.
- However, this method demonstrates the species differentiation for each click characteristic separately. It cannot explain species (dis)similarity in the multivariate space that includes all click characteristics.

Statistical testing for species differentiation (2/2)

- In order to examine the species differentiation in the multivariate space of click characteristics, two statistical tests are performed:
  - ✔ A PerMANOVA test that examines if the centroids and dispersions in multivariate space are equivalent for the two species/groups.
  - ✔ A PERMDISP test that examines the homogeneity of dispersion in multivariate space between the two species/groups.

PerMANOVA (1/4)

- Permutational multivariate analysis of variance (PerMANOVA) is a non-parametric multivariate statistical permutation test.
- It tests the null hypothesis that the centroids and dispersions in the multivariate space of dataset features are equivalent for all dataset groups.
- A dissimilarity matrix is calculated in feature space between all pairs of dataset samples. For instance, a matrix of the euclidean distances between the dataset samples, as measured with respect to the dataset features.
- The squares of the dissimilarity matrix elements that correspond to samples in same groups are summed to calculate the within-group sum-of-squares SS_W, while the squares of the dissimilarity matrix elements that correspond to samples in different groups are summed to calculate the between-group sum-of-squares SS_A.

PerMANOVA (2/4)

- Then, the pseudo F-statistic is calculated as a ratio that expresses the relationship between the between-groups distance and the within-groups dispersion.
  - ✔ SS_A: between-group sum-of-squares
  - ✔ SS_W: within-group sum-of-squares
  - ✔ a: number of groups
  - ✔ N: number of samples

$$F = \frac{\frac{SS_A}{a-1}}{\frac{SS_W}{N-a}}$$

- High values for the pseudo F-statistic mean that the between-groups distance is more significant than the within-groups dispersion (i.e. the groups differ in feature space).
- On the other hand, an F-statistic that approximates 1 means that the between-groups distance is as significant as the within-groups dispersion (i.e. the groups don't differ).

PerMANOVA (3/4)

- In order to ensure that the result is not random and that the group is in fact a significant factor that affects the value of this F-statistic, the previously described process is repeated for a fixed number of group permutations, i.e. random assignments of samples to groups.
- The F-statistic is calculated for each permutation and compared to the one acquired for the original dataset, where each sample was assigned to its true group.
- Finally, the probability p that a random permutation corresponds to a value for the F-statistic that is greater than the one corresponding to the original dataset is calculated.
- Adequately small values for p (smaller that 0.05) indicate a statistically significant result, i.e. that the group differentiation is significantly more prominent in the original dataset.

PerMANOVA (4/4)

- The PerMANOVA test was executed on our z-score transformed dataset for 999 permutations. The species labels were used to group the data. The features that have been calculated for both -10dB and -3dB are only considered in their -3dB version.
- The features included in the default test (top-left) are marked with * in Table 1.
- A test that also considered the peak time feature was performed (bottom-right).

```
          Df SumOfSqs     R2      F Pr(>F)
species    1   1728.2 0.2329 160.61  0.001 ***
Residual 529   5691.8 0.7671
Total    530   7420.0 1.0000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.
```

```
          Df SumOfSqs      R2     F Pr(>F)
species    1   1776.2 0.22342 152.2  0.001 ***
Residual 529   6173.8 0.77658
Total    530   7950.0 1.00000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

PERMDISP (1/2)

- PerMANOVA assumes that the dataset samples are exchangeable under the null hypothesis, which further implies that:
  - ✔ The dataset samples are independent (e.g. no duplicates are present).
  - ✔ The dataset samples within each group have similar multivariate dispersion (i.e. each group has a similar degree of multivariate scatter).
- In order to test the second assumption, a permutation test of multivariate homogeneity of dispersions (PERMDISP) is executed.
- PERMDISP tests the null hypothesis that the groups within a dataset present similar dispersions by calculating the F-statistic and probability p for a number of permutations, in a manner similar to PerMANOVA.

PERMDISP (2/2)

- The PERMDISP test was executed on our z-score transformed dataset for 999 permutations.
- Again, two tests were performed, one for the default features (top-left) and one also including the peak time feature (bottom-right).

```
           Df  Sum Sq Mean Sq     F N.Perm Pr(>F)
Groups      1  121.66 121.665 63.08    999  0.001 ***
Residuals 529 1020.30   1.929
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
           Df  Sum Sq Mean Sq      F N.Perm Pr(>F)
Groups      1   96.38  96.384 48.817    999  0.001 ***
Residuals 529 1044.45   1.974
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Principal Component Analysis (1/4)

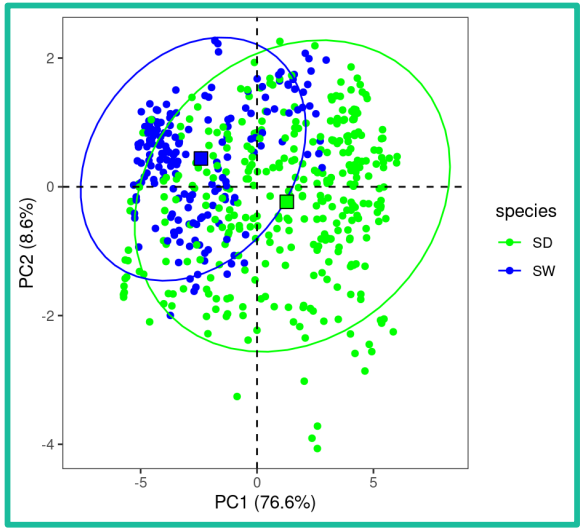
- The executed PerMANOVA test suggests that the two species/groups in our dataset differ significantly in the multivariate space of click characteristics.
- However, the between-groups dispersion homogeneity, a significant assumption for the validity of the PerMANOVA test, is disproved by the executed PERMDISP test.
- At this point, it is expected that:
  - ✔ The distance between the centroids of the two species in the multivariate feature space is considerable.
  - ✔ One of the two species is more spread out in the multivariate feature space.
- We can’t know for sure that our dataset is separable with respect to species.

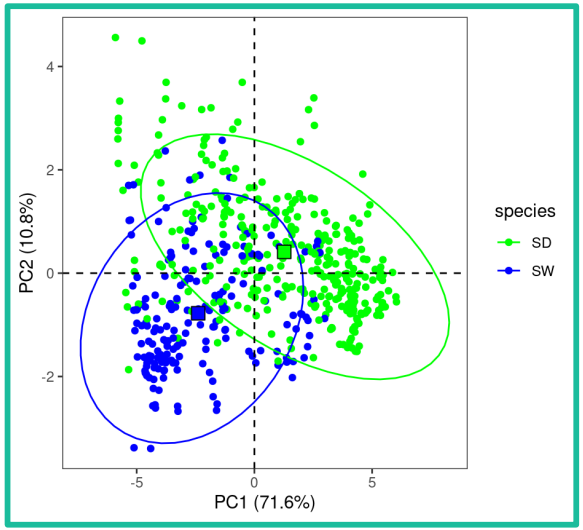
Principal Component Analysis (2/4)

- In order to plot the samples in 2D and visually examine the species differentiation, we need to reduce the dimensionality of the multivariate feature space.
- The ordination technique that is used for this purpose is principal component analysis.
- PCA is a technique that rotates the feature space, so as to change its base to a number of independent principal components, each one accounting for a percentage of the total samples variation.
- The variation explained by each principal component is decreased as the component number is increased, i.e. the first component accounts for the most sample variation while the last component accounts for the least sample variation.

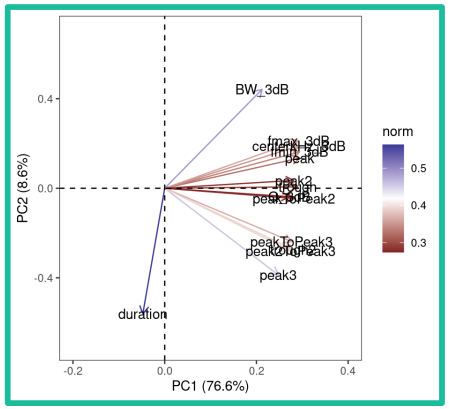
Principal Component Analysis (3/4)

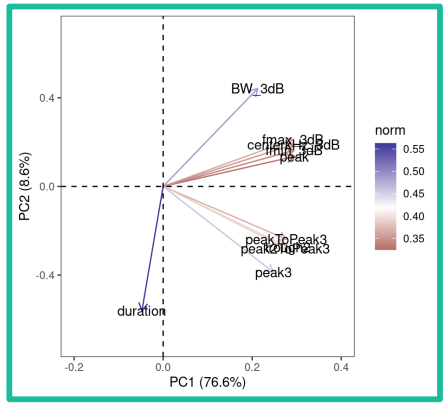
- PCA was performed on our dataset to reduce the dimensionality of both the default feature space (left) and the feature space that includes the peak time feature (right).
- The two first principal components account for approximately 80% of the variation, which means that it's safe to use only them to visualize the samples.





Principal Component Analysis (4/4)





Left: scree plots for the default feature space

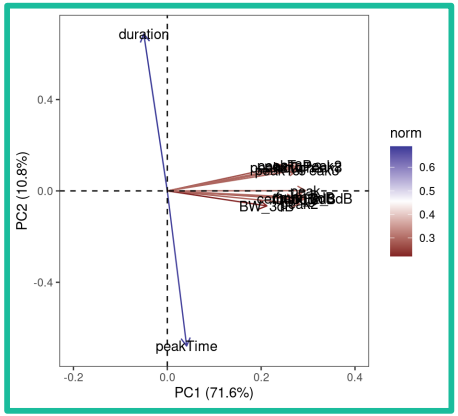
Right: scree plots for the feature space that also includes the peak time characteristic

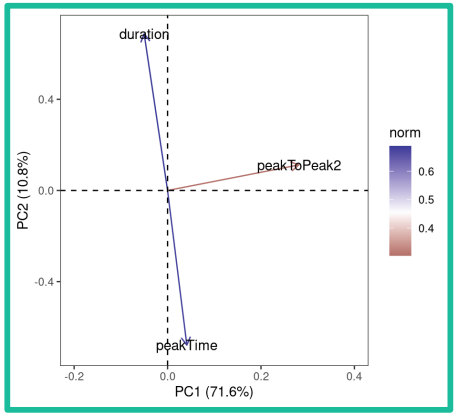
The scree plots demonstrate how each feature of the original multivariate feature space contributes to the two first principal components of the PCA.

The greater the norm of the feature vector the greater its contribution to the sample distribution in the 2D space of the two primary PCs.

In the bottom plots only the features with a norm greater than 0.3 are included, in order to make the scree plots more readable.

For instance, in the left scree plots the BW_3dB feature moves the samples to increased values for both PCs, while in the right scree plots the peakTime feature affects primary the second PC.





Questions – Next steps

- Should we revisit the click detection in PAMGuard to readjust the detection parameters?
- How are the PerMANOVA and PERMDISP test results interpreted?
  - ✔ Does the fact that our dataset is unbalanced affect the results?
  - ✔ PerMANOVA assumes dispersion homogeneity between groups, but PERMDISP disproves this.
- Is our dataset separable in the click characteristics space (with respect to species)?
- Which click characteristics are significant in the species differentiation and can thus be beneficial for species classification?
- How do we utilize the statistical testing and PCA results?
- What other statistical tests and/or ordination techniques can we employ?