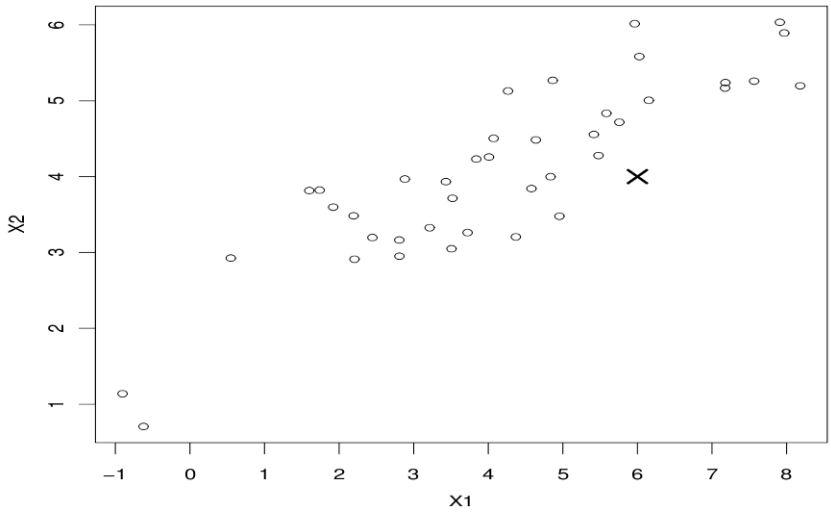
Solution Sheet 3

The full R code will be made available in a separate file on the course homepage.

1. **Warning: The original figure on the exercise sheet is distorted. Distances on the X1 and X2 axes are not measured in the same unit of length. Below you can find a properly scaled version of the figure.**



   a) As in the example in the lecture.

   b) For normal univariate data, roughly 2/3 of the date lie within ± 1 standard deviation of the mean. In the direction of the first PC (from bottom left to top right), there are roughly 25 observations within the central 5 units range, the standard deviation of this PC is therefore equal to 2.5. The spread orthogonal to the first principal axis is roughly 1/3 this large, i.e., a standard deviation of 0.8 may be a good estimate.
   The eigenvalues of the covariance matrix are equal to the variances of the principal components; squaring the above estimated standard deviations, we obtain the eigenvalues 6.25 and 0.64.
   Remark: The 2 outlying observations at the bottom left indicate that the data are probably not normal.

   c) Starting from the center of the point cloud, we get to the point × by moving 2 units along the first principal axis and −1 unit along the second principal axis, hence, we obtain the following coordinates in the rotated coordinate system: $[2, -1]^T$.

2. **a)** See Fig. 1.

   b) Square root transformed data: Fig. 2. The transformation does not improve the result in this case. But using `scale=FALSE` does (see Fig. 3), since differences in rare species, that will often be random, are then weighted less than abundant ones.

   c) `summary(r.pca)` yields the proportion of the variance that is explained by the principal components.

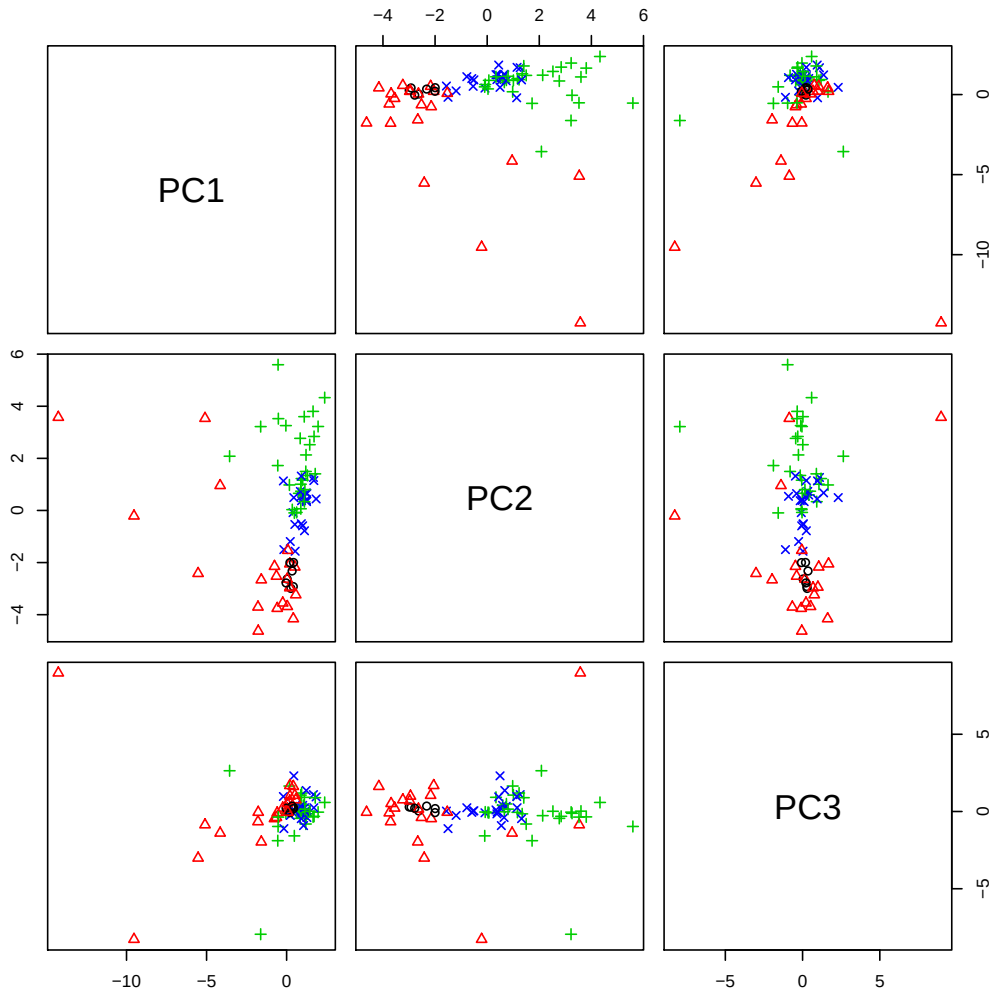


Figure 1: The first 3 principal components for the vegetation data. The symbols mark the vegetation types.

```
Importance of components:
                          PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8  PC9 PC10 PC11
Standard deviation        2.4 2.28 1.84 1.72 1.63 1.58 1.50 1.45 1.44 1.36 1.33
Proportion of Variance    0.1 0.08 0.05 0.05 0.04 0.04 0.04 0.03 0.03 0.03 0.03
Cumulative Proportion     0.1 0.17 0.23 0.28 0.32 0.36 0.40 0.43 0.46 0.49 0.52
...
```

With 3 components, 23% are explained, and 90% are achieved by 33 components. For the transformed data, the numbers are 26% and 32 components, and for the unscaled (transformed) data, 54% and 19. (For many datasets, the proportion increases much faster!)

3. **a)** If all the variables (species) are used, the picture looks messy. The function `g.biplot` allows for coloring the points, using the argument `xcol`, or specifying symbols by the argument `xlabs`.

b) Using only those species that occur more than once in the mean over the parcels, the arrows can now be distinguished, see Fig. 4.

c) The ellipse shown in the display allows for judging how well the variables are represented in the graph: The proportion of the length of the arrow to the distance to the ellipse along the same

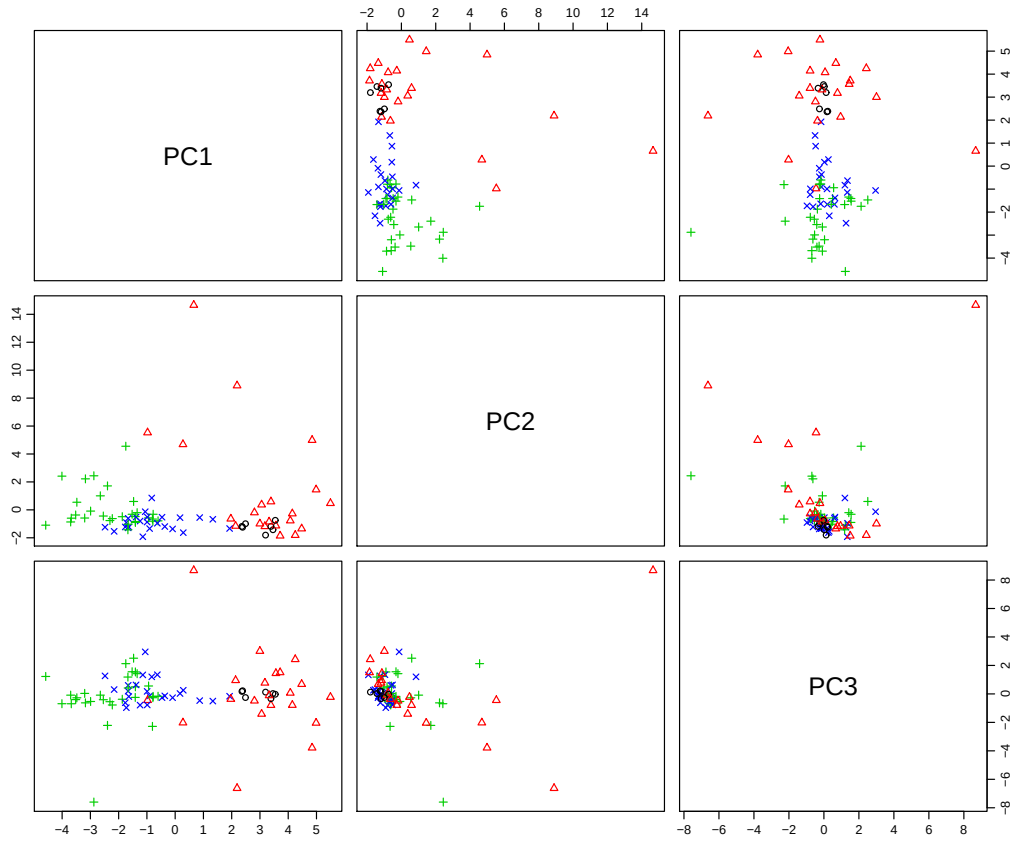


Figure 2: As Fig. 1, after transformation

direction represents the proportion of the variance of the respective variable that is represented in the graph. Here, the species `Nardstri` appears to be best represented.

d) The (signed) distance from the origin to the projection on the arrow approximately equals the centered and standardized square root transformed count of the species `Nardstri`, so we expect observation 34 to have the highest count, and observation 30 to have the lowest count. This is only an approximation since we have more than 2 variables, and comparing with the measured counts, we see that observation 46 actually has a lower count than observation 30.

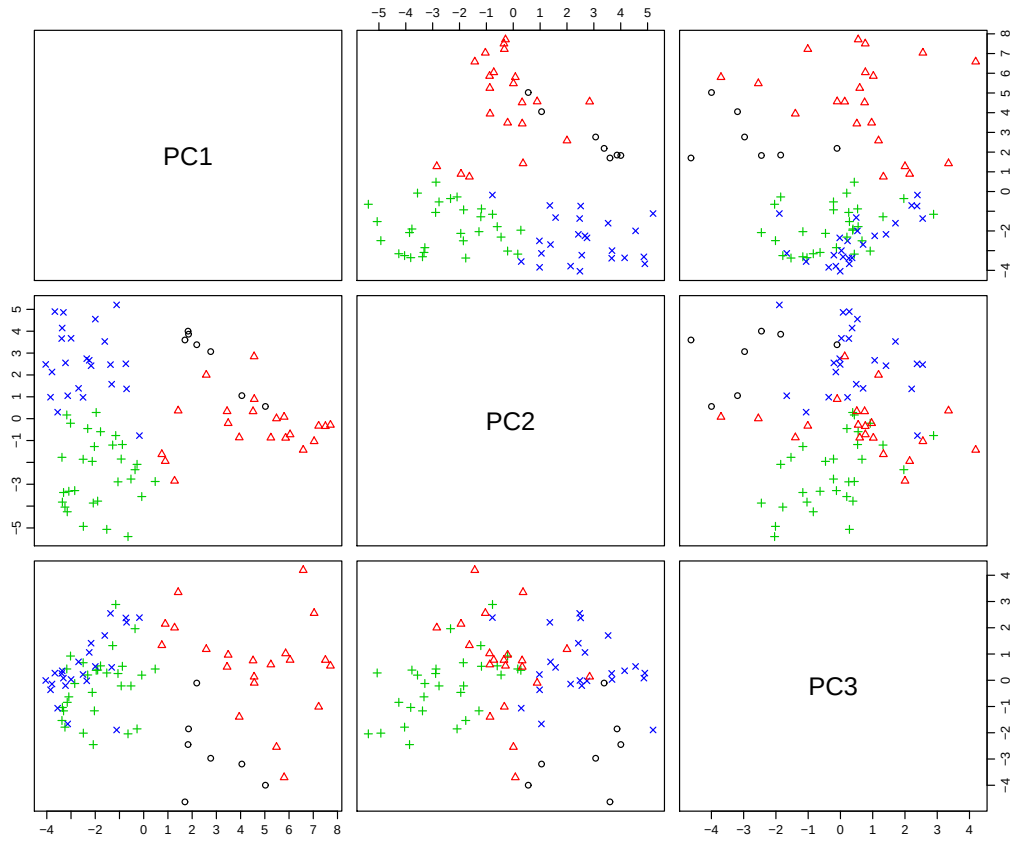


Figure 3: As Fig. 1, after transformation, no scaling

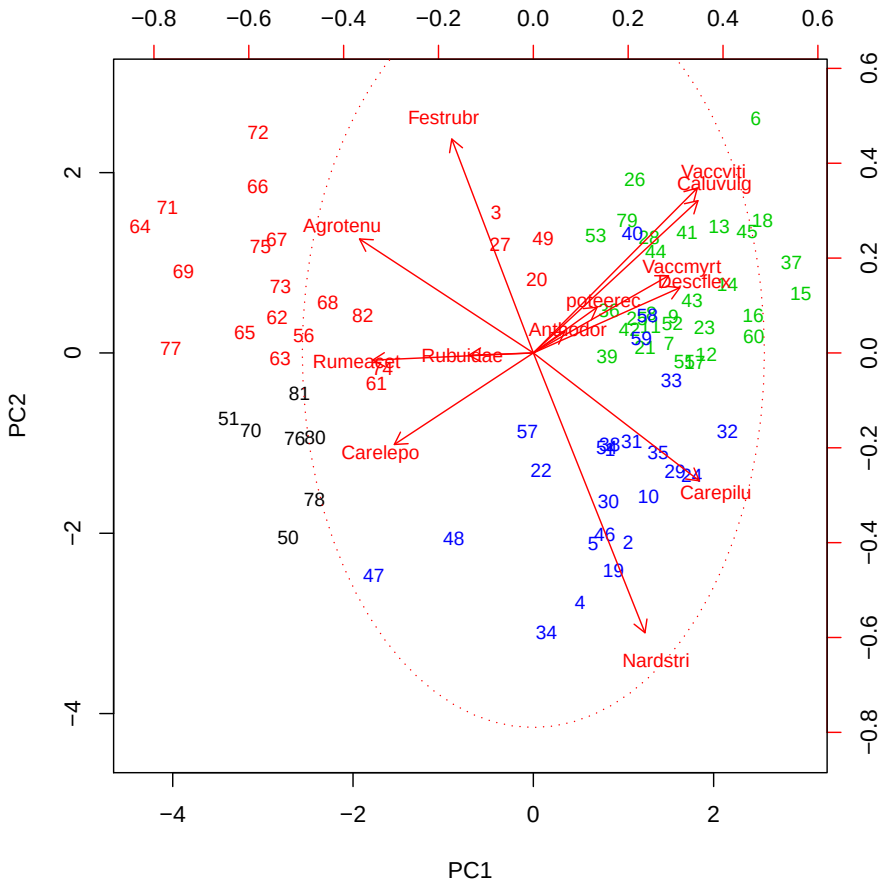


Figure 4: PCA-Biplot for the more abundant species.