

# Multivariate Statistics in Ecology and Quantitative Genetics

## **Combining PCA and GLMs**

Dirk Metzler & Paul Staab

[http://evol.bio.lmu.de/\\_statgen](http://evol.bio.lmu.de/_statgen)

Summer semester 2015

- Lets again try to predict the species richness on sandy beaches
- Data from the dutch National Institute for Coastal and Marine Management (RIKZ: Rijksinstituut voor Kust en Zee)
- see also
  - 📖 Zuur, Ieno, Smith (2007) *Analysing Ecological Data*. Springer

```
rikz_data <- read.delim("RIKZGroups.txt")  
dim(rikz_data)
```

```
## [1] 45 17
```

```
colnames(rikz_data)
```

```
## [1] "Sample"           "Polychaeta"       "Crustacea"  
## [4] "Mollusca"         "Insecta"          "week"  
## [7] "angle1"           "angle2"           "exposure"  
## [10] "salinity"         "temperature"      "NAP"  
## [13] "penetrability"   "grainsize"        "humus"  
## [16] "chalk"            "sorting1"
```

# Meaning of the Variables

**index i** index of sampling station

**richness** Number of species that were found in a plot.

**angle1** angle of the station

**angle2** slope of the beach at the plot

**exposure** index composed of wave action etc.

**NAP** altitude of the plot compared to the mean sea level.

**grainsize** average diameter of sand grains

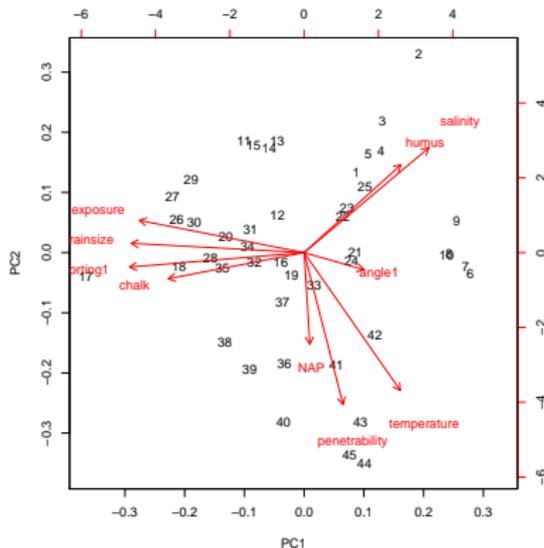
**humus** fraction of organic material

**week** in which of 4 weeks was this plot probed.

(many more variables in original data set)

# Explanatory Variables

```
explanatory <- c(7, 9:17)
pca <- prcomp(rikz_data[ , explanatory], scale = TRUE)
biplot(pca, scale = 1)
```



# Explanatory Variables

- Several of the explanatory variables are highly correlated. This can be problematic when doing a linear regression.

# Explanatory Variables

- Several of the explanatory variables are highly correlated. This can be problematic when doing a linear regression.
- Most likely, not more than one of each group will have a significant influence.

# Explanatory Variables

- Several of the explanatory variables are highly correlated. This can be problematic when doing a linear regression.
- Most likely, not more than one of each group will have a significant influence.
- We can use the biplot to manually choose only one of each group of highly correlated variables.

# Explanatory Variables

- Several of the explanatory variables are highly correlated. This can be problematic when doing a linear regression.
- Most likely, not more than one of each group will have a significant influence.
- We can use the biplot to manually choose only one of each group of highly correlated variables.
- Or we can use the PCA components as explanatory variables.

# Model

We use the model

$$\text{Crustacea} = \beta_0 + \beta_1 Z_1 + \dots + \beta_n Z_n + \varepsilon$$

where the  $Z_i$  are the values of the data's  $i$ -th PC.

# Model

```
rikz_data_pca <- data.frame(Crustacea = rikz_data$Crustacea, pca$x)
fitted_model <- glm(Crustacea ~ ., data = rikz_data_pca)
summary(fitted_model)
```

```
##
## Call:
## glm(formula = Crustacea ~ ., data = rikz_data_pca)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -71.660   -8.869   -0.340    7.445   124.395
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.4444    5.2207   3.916 0.000412 ***
## PC1          9.4165    2.7633   3.408 0.001701 **
## PC2         3.3396    3.9664   0.842 0.405690
## PC3         6.1369    4.8754   1.259 0.216693
## PC4        -0.9854    5.1151  -0.193 0.848379
## PC5         1.9342    5.5169   0.351 0.728062
## PC6         1.1733    6.9085   0.170 0.866146
## PC7        12.2134    7.7256   1.581 0.123161
## PC8        19.9986   13.5043   1.481 0.147839
## PC9        32.8808   15.2320   2.159 0.038025 *
## PC10       -7.6220   16.7327  -0.456 0.651635
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1226.502)
##
```

# Model

After dropping non-significant variables, only PC1 and PC9 remain:

```
fitted_model_2 <- glm(Crustacea ~ PC1 + PC9, data = rikz_data_pca)
summary(fitted_model_2)
```

```
##
## Call:
## glm(formula = Crustacea ~ PC1 + PC9, data = rikz_data_pca)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -50.155  -19.186   -5.232   11.129  141.578
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.444      5.182   3.945 0.000297 ***
## PC1           9.417      2.743   3.433 0.001354 **
## PC9          32.881     15.120   2.175 0.035333 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1208.457)
##
##      Null deviance: 70713  on 44  degrees of freedom
## Residual deviance: 50755  on 42  degrees of freedom
## AIC: 451.97
##
## Number of Fisher Scoring iterations: 2
```



# Interpretation

Now, how can we interpret these results?

# Interpretation

Now, how can we interpret these results?

Because the principal components are linear combinations of the original variables, we translate the results back.

# Interpretation

Now, how can we interpret these results?

Because the principal components are linear combinations of the original variables, we translate the results back.

To do so, we multiply the coefficients with the loadings of the corresponding principal component.

# Interpretation

```
pca$rotation[ , c(1,9)] %*%  
  fitted_model_2$coefficients[-1]
```

```
##           [,1]  
## angle1      4.2215458  
## exposure   -22.0660880  
## salinity    8.2313874  
## temperature -16.0715593  
## NAP         6.6778369  
## penetrability 11.3565887  
## grainsize   5.5566109  
## humus        0.4565255  
## chalk        7.0286887  
## sorting1    -9.2184292
```

# Interpretation

- These are the factors for the corresponding normalized variables (plus Intercept of 20.44).

# Interpretation

- These are the factors for the corresponding normalized variables (plus Intercept of 20.44).
- As the variables were normalized, we can directly compare the factors. The larger absolute value, the larger is influence of the variable.

# Interpretation

- These are the factors for the corresponding normalized variables (plus Intercept of 20.44).
- As the variables were normalized, we can directly compare the factors. The larger absolute value, the larger is influence of the variable.
- We can use them to predict the species richness of new observations. We however need to apply the same normalization transformation, i.e. subtract mean and divide by the standard distribution of the training data.

# Interpretation

- These are the factors for the corresponding normalized variables (plus Intercept of 20.44).
- As the variables were normalized, we can directly compare the factors. The larger absolute value, the larger is influence of the variable.
- We can use them to predict the species richness of new observations. We however need to apply the same normalization transformation, i.e. subtract mean and divide by the standard distribution of the training data.
- Maybe better suited: Linear Discriminant Analysis instead of PCA (not covered here).