Illustration: Amrei Binzer-Panchal

Basic Biostatistics and Bioinformatics

Session 3: PCA

Swedish University of Agricultural Sciences, Alnarp

11 December 2023

Basic Biostatistics and Bioinformatics

A seminar series on fundamentals

Organised by *SLUBI* and *Statistics at SLU*

Presentation of background and a practical exercise

Upcoming topics

- 27 November. Linux Basics
- **11 December. PCA**
- 15 January. Introduction to Markdown
- 29 January. Population Structure

Topic suggestions are welcome

SLUBI

- SLU bioinformatics center
- Weekly online drop-in (Wednesdays at 13.00)
- slubi@slu.se, https://www.slubi.se
- Alnarp: Lizel Potgieter (Dept. of Plant Breeding)

Statistics at SLU

- SLU statistics center
- Free consultations for all SLU staff
- statistics@slu.se
- Alnarp: Jan-Eric Englund and Adam Flöhr (Dept. of Biosystems and Technology)

☰

Today's Presentation

Principal Component Analysis

Some background and justification

Interpretation of results

Implementation in R

Exercise session

`PCAtools` in Bioconductor

- https://bioconductor.org/packages/devel/bioc/vignettes/PCAtools/inst/doc/PCAtools.html

The nature of multivariate data

Multiple measurements of the same unit

n units and d measured variables

Examples

- Phenological measures on the same plant
- Expressions of genes on the same biological sample
- Chemical compound measurements on the same soil sample

Example data

Palmer Archipelago (Antarctica) penguin data

Bill, flipper, and body mass measurements for 344 individuals from 3 species

```r
library(palmerpenguins)
penguins <- penguins %>% drop_na()

penguins
```

```
# A tibble: 333 × 8
   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
 1 Adelie  Torgersen           39.1          18.7               181        3750
 2 Adelie  Torgersen           39.5          17.4               186        3800
 3 Adelie  Torgersen           40.3          18                 195        3250
 4 Adelie  Torgersen           36.7          19.3               193        3450
 5 Adelie  Torgersen           39.3          20.6               190        3650
 6 Adelie  Torgersen           38.9          17.8               181        3625
 7 Adelie  Torgersen           39.2          19.6               195        4675
 8 Adelie  Torgersen           41.1          17.6               182        3200
 9 Adelie  Torgersen           38.6          21.2               191        3800
10 Adelie  Torgersen           34.6          21.1               198        4400
# ℹ 323 more rows
# ℹ 2 more variables: sex <fct>, year <int>
```

$n = 333, d = 4$

Linear combinations and variance

A *linear combination* of variables is a weighted sum

Say we have a set of variables $x_1, x_2, \ldots, x_d$

We can construct linear combinations

$$z_1 = l_1 \cdot x_1 + l_2 \cdot x_2 + \ldots + l_d \cdot x_d$$

Common to use some restriction on the coefficients l

For PCA purposes the relevant restriction is that squared ls equals one

Variance of sums

The variance of a sum is the sum of the variances plus twice the correlation between each pair

☰

Penguin example

The penguin data contains columns for bill length (x_1) and flipper length (x_2)

We can combine these, for example $z = \sqrt{\frac{1}{5}} \cdot x_1 + \sqrt{\frac{4}{5}} \cdot x_2$

and get

$$Var(z) = \frac{1}{5}Var(x_1) + \frac{4}{5}Var(x_2) + 2\sqrt{\frac{1}{5}}\sqrt{\frac{4}{5}}Cor(x_1, x_2)$$

Variance and correlation is given by

```
var(penguins[c(3,5)])
```

```
                  bill_length_mm flipper_length_mm
bill_length_mm          29.90633          50.05819
flipper_length_mm       50.05819         196.44168
```

and the variance of the sum becomes

```
var(1/sqrt(5) * penguins$bill_length_mm + sqrt(4) / sqrt(5) * penguins$flipper_length_mm)
```

```
[1] 203.1812
```

The linear combination has higher variance than either original variable

Dimension reduction

Original data has d dimensions

Want to reduce the number of dimensions but keep as much information as possible

PCA

Two highly correlated variables contain (some of) the same information

Merging correlated variables gives combinations which capture more of the variation

We can order these linear combinations by variance explained

Combinations with little variance explained can be dropped

Principal Component Analysis (PCA)

PCA forms a new set of variables (principal components) as linear combinations of the original variables

The first PC contains the most of the original variance

Results from a PCA

Three primary outputs

- *Variance decomposition:* shows the proportion of variance in each component
- *Scores:* Principal components for the observations
- *Loadings:* weight parameters of the original variables

PCA does not rely on any formal assumptions

Works best on continuous data with somewhat even distributions

Tests of components may have assumptions, such as requiring normal distribution

PCA, penguin example

We can run a PCA using `prcomp()` from base-R

```r
mod <- prcomp(penguins[,3:6], scale. = T)
summary(mod)
```

```
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.6569 0.8821 0.60716 0.32846
Proportion of Variance 0.6863 0.1945 0.09216 0.02697
Cumulative Proportion  0.6863 0.8809 0.97303 1.00000
```

The principal components are ordered by importance

If the later components explain little of the total variance they may be removed without a great loss

Here we lose 12 percent of the total variance if we drop the two final components

Penguin examples. Loadings and scores

The components are given by multipling original variables with *loadings* and summing

Loadings are contained in the object as `rotation`

```r
mod$rotation
```

```
                         PC1         PC2        PC3        PC4
bill_length_mm     0.4537532 -0.60019490 -0.6424951  0.1451695
bill_depth_mm     -0.3990472 -0.79616951  0.4258004 -0.1599044
flipper_length_mm  0.5768250 -0.00578817  0.2360952 -0.7819837
body_mass_g        0.5496747 -0.07646366  0.5917374  0.5846861
```

A *score* can be calculated for each observation and component

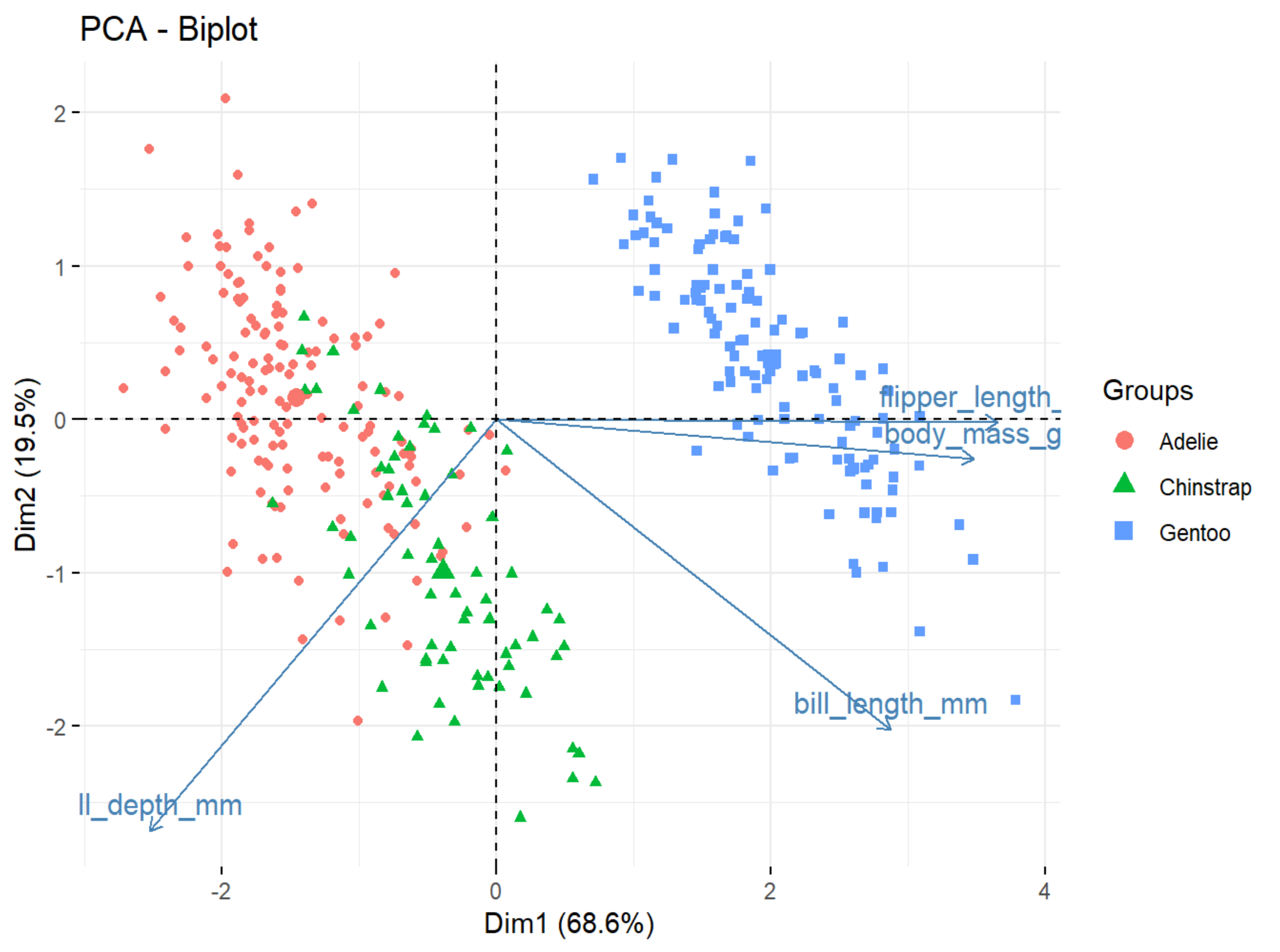
```r
mod$x[1:5,] # Scores of the first five observations
```

```
           PC1         PC2        PC3        PC4
[1,] -1.850808 -0.03202119  0.2345487  0.5276026
[2,] -1.314276  0.44286031  0.0274288  0.4011230
[3,] -1.374537  0.16098821 -0.1894042 -0.5278675
[4,] -1.882455  0.01233268  0.6279277 -0.4721826
[5,] -1.917096 -0.81636958  0.6999980 -0.1961213
```

PCA, biplot

PCA results are often visualised in a *biplot*

```r
library(factoextra)
fviz_pca_biplot(mod, geom = "point",
                habillage = penguins$species)
```



The biplot summarises similarity between individuals (points) and variables (arrows)

- Close points correspond to more similar individuals
- Loadings (arrows) with similar angles are correlated
- Longer loadings are more important in the corresponding component
- Points in the direction of a loading indicate individuals with high values in that variable

Alternatives and complements to PCA

Factor analysis

Factor analysis re-combines the components (by rotation)

Clarifies the PCA by strengthening the connection between components and original variables

nMDS (non-metric Multi-dimensional Scaling)

Replicates multivariate distances in a smaller number of dimensions

Generalises the PCA by allowing the use of any type of distance measure

Regression-type methods

A large number of methods for situations with two or more multidimensional datasets

Want to explain one multivariate response using some multivariate explanatory set

Includes PLS (Partial Least Squares), RDA (Redundancy Analysis) and CCA (Canonical Correspondence Analysis)

Illustration: Amrei Binzer-Panchal

The End. Stick around for practical exercise