Multivariate analysis

DAAG Chapter 12

Learning objectives

In this section, we will learn some basic approaches to multivariate analysis.

- Principal components analysis
  - What is principal components analysis?
  - What does principal components analysis do?
  - How can principal components analysis be used?
- Multi-dimensional scaling (MDS)
  - What is a distance measure?
  - What are Euclidean, Manhattan, Canberra distances?
  - What does MDS do?
  - How can MDS be used?

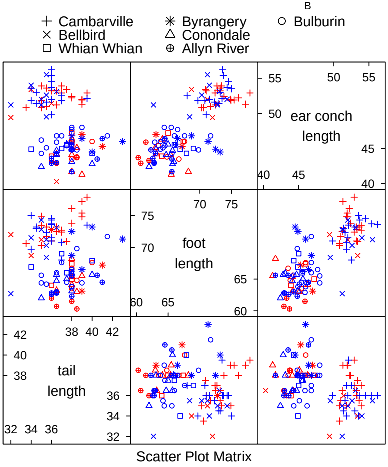
Multivariate analysis: Motivating problem

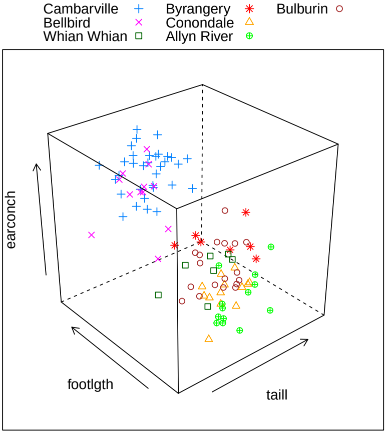
Possum morphology data. 104 possums trapped at seven sites in Australia.

- sex
- age
- head length
- skull width
- total length
- tail length
- foot length
- ear conch length
- eye measurement
- chest girth
- belly girth

How can we analyze these data to uncover the patterns that exist?

Plots of possum data





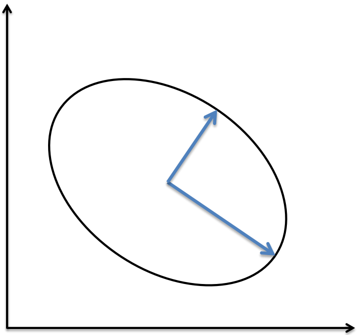
Principal components analysis

For the possum data, we have 9 morphological measurements.

- This is a lot to visualize.
- Also, there is no "response" variable
- How can we uncover structure in these data?

Principal components analysis creates new variables (components) using linear combinations of the existing variables.

- The first component is chosen to explain as much variation as possible
- Subsequent components are chosen in the same way
- Components are orthogonal



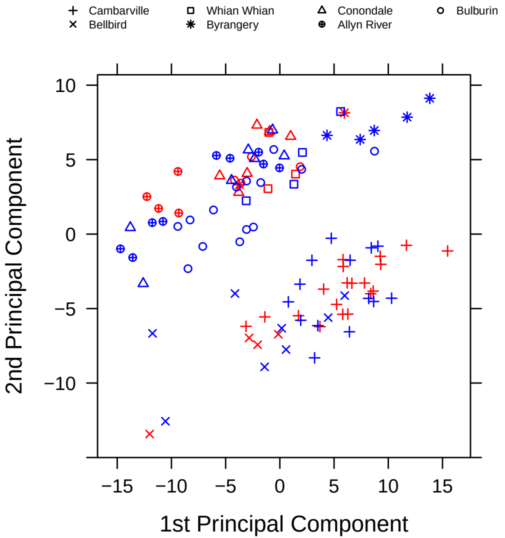
Principal components on possums

```
Importance of components:
                          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
Standard deviation         6.800  5.033 2.6993 2.1601 1.7372
Proportion of Variance     0.498  0.273 0.0785 0.0503 0.0325
Cumulative Proportion      0.498  0.771 0.8495 0.8998 0.9323
                          Comp.6 Comp.7 Comp.8  Comp.9
Standard deviation        1.5989 1.2860 1.1111 0.91696
Proportion of Variance    0.0275 0.0178 0.0133 0.00906
Cumulative Proportion     0.9598 0.9776 0.9909 1.00000
```

Principal components on possums

```
Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
hdlngth   0.413  0.282  0.339 -0.185  0.695
skullw    0.296  0.269  0.540 -0.338 -0.519
totlngth  0.518  0.315 -0.648 -0.156
taill            0.251 -0.350        -0.194
footlgth  0.514 -0.468               -0.336
earconch  0.309 -0.650                0.249
eye
chest     0.219         0.175  0.174 -0.177
belly     0.246  0.178  0.134  0.891
         Comp.6 Comp.7 Comp.8 Comp.9
hdlngth   0.277        -0.184
skullw   -0.276  0.259  0.112
totlngth -0.226 -0.145  0.336
taill            0.437 -0.753  0.106
footlgth  0.633
earconch -0.584  0.208 -0.172
eye              0.195  0.242  0.942
chest    -0.189 -0.763 -0.404  0.267
belly    -0.102  0.239  0.144
```

Principal components on possums



Uses of principal components

- Description of patterns in high-dimensional data
  - Direct interpretation of components
  - Graphical display using components
  - Grouping/clustering
- Transformation for subsequent statistical analysis
  - Use components as explanatory variables in regression
    - Good for summarizing the effects of many covariates
    - Avoid problems with multicollinearity
  - Use first component as response variable in regression

Multidimensional scaling

We have seen how to use principal components analysis to display multivariate information in fewer dimensions.

- Principal components analysis is a specific version of a more general class of methods called multidimensional scaling (MDS)
- In MDS, we take multivariate data and display them in fewer dimensions, doing our *best* to maintain the distance between points
- Classical MDS with Euclidean distance is equivalent to the principal components representation
- However, we can extend the lower-dimensional representation in two ways:
  1. Use a different distance (or *dissimilarity*) metric.
  2. Use a different criteria for *ordination* (display of objects).

Distance or dissimilarity metrics

- Euclidean distance

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}$$

- Manhattan distance

$$d_{ij} = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|$$

- Canberra distance

$$d_{ij} = \frac{|x_{i1} - x_{j1}|}{|x_{i1} + x_{j1}|} + \frac{|x_{i2} - x_{j2}|}{|x_{i2} + x_{j2}|} + \ldots + \frac{|x_{ip} - x_{jp}|}{|x_{ip} + x_{jp}|}$$

where all $x_{\cdot\cdot} \geq 0$.

Ordination methods

- Classical MDS
  - Distances are treated as Euclidean.
  - Find the lower-dimensional representation that best preserves distances.
- Sammon method
  - Similar to classical MDS.
  - Minimize weighted sum of squared differences between dissimilarities and representation distances.
  - Weights are proportional to dissimilarities (more dissimilar $=$ more weight).
- Kruskal's non-metric MDS
  - Dissimilarities are allowed a monotonic transformation
    - Only the ranks of the dissimilarities matter
  - Minimize *stress* $S = \sqrt{\frac{\sum_i (d_i - r_i)^2}{\sum d_i^2}}$ where
    - d_i are the input dissimilarities (transformed)
    - r_i are the output representation (Euclidean) distances

MDS example

Data are for 47 swiss provinces circa 1888 (undergoing demographic transition). Variables are proportion of population (agricultural, education, religion, infant mortality,...).