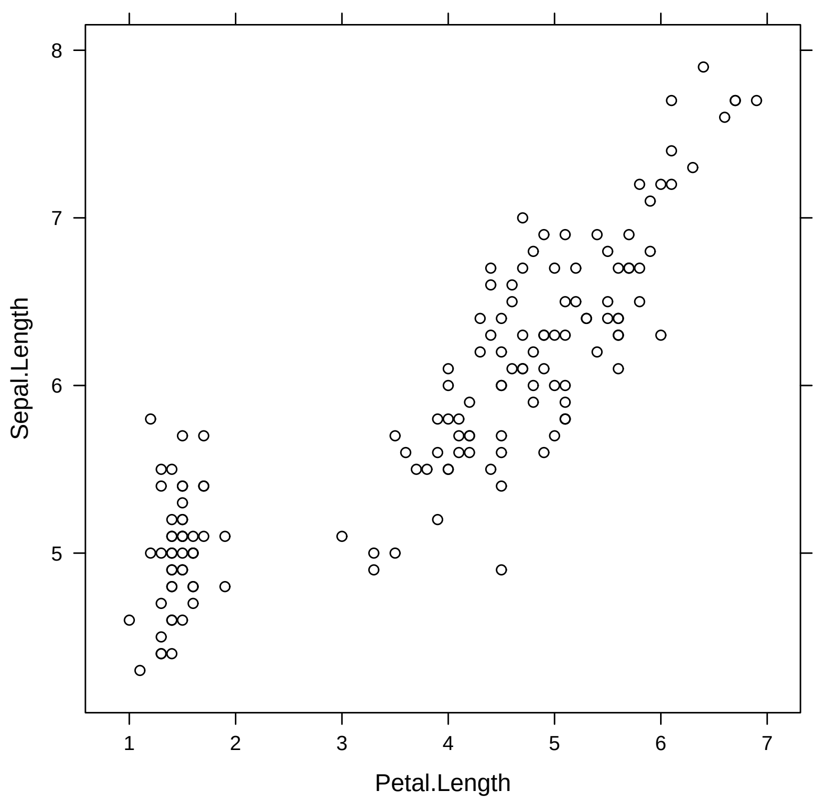
Multivariate statistics

Multivariate statistics focus less upon independent→dependent relationships and more on relationships among all variables. Multivariate statistics are used in:

- classification
- hypothesis generation
- dimension-reduction

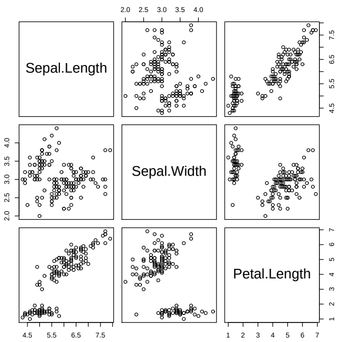
Bivariate Plots



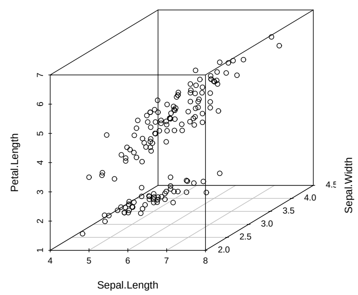
Multivariate plots

Multivariate data are difficult to present visually. As you might expect in that case, there are several approaches that are used.

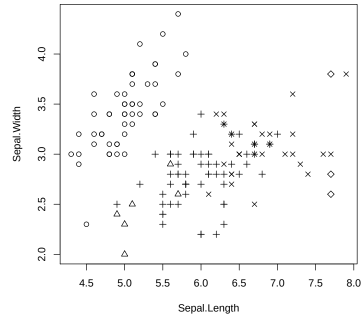
Pairs



Clouds



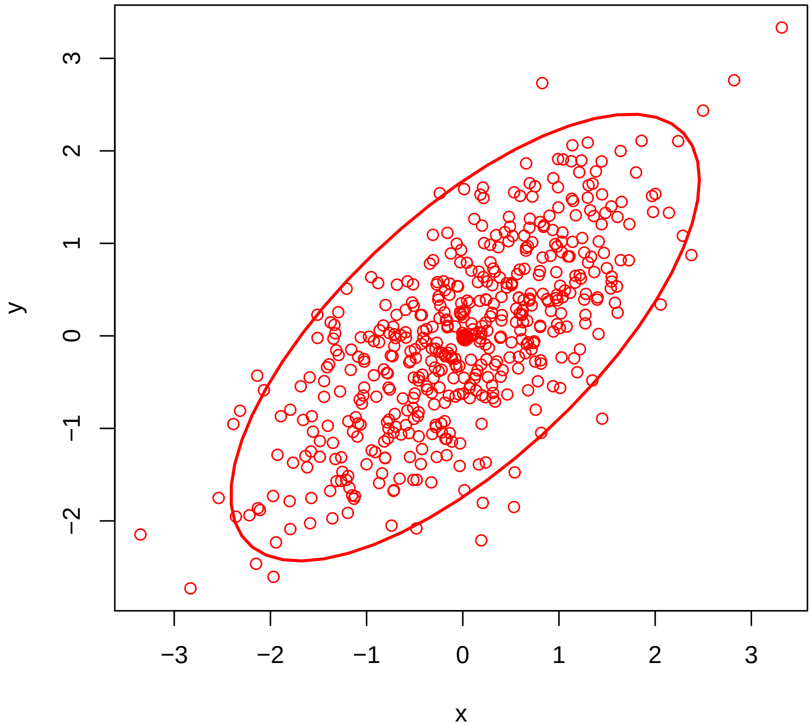
Symbols

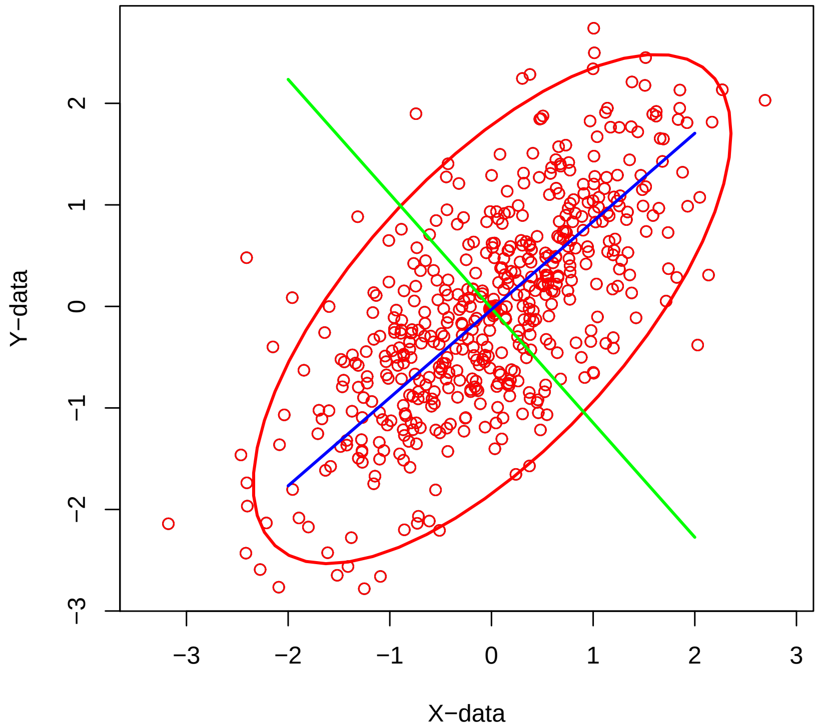


Ordination

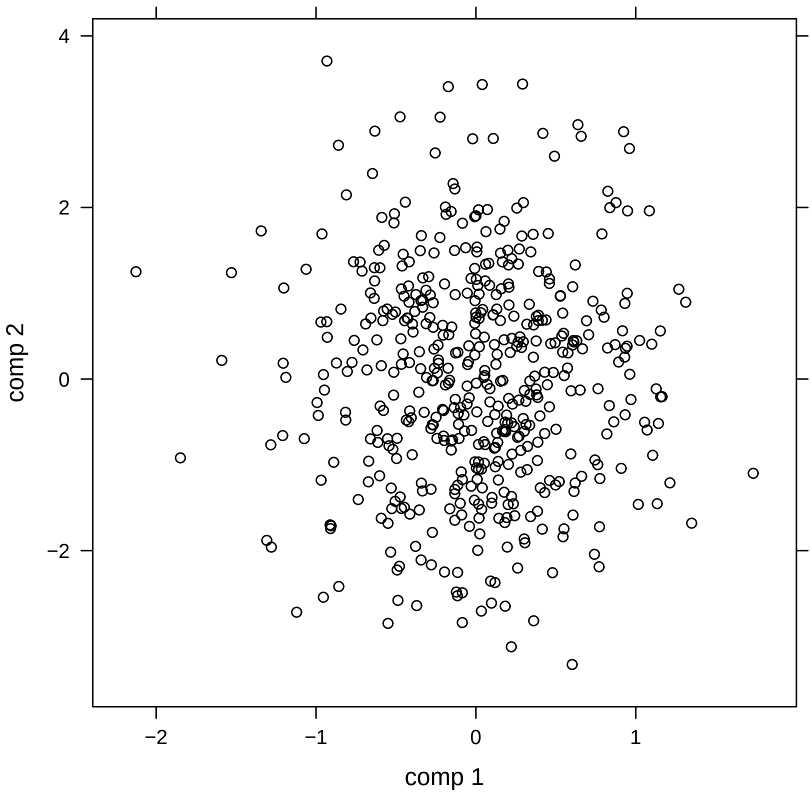
These methods include principle components analysis, metric and non-metric multidimensional scaling.
In general, they seek to develop new variables that try to explain the variation in the data *not* due to covariance.

Data ellipse





Rotated data



Eigenvectors

The lines on the previous slide represent *Eigenvectors* calculated from the covariance matrix between variables in the dataframe. They have useful properties:

- They are *orthogonal*
- They are associated with *eigenvalues* which can be used to estimate the proportion of variation explained by each vector.
- They are relatively easy to compute (not by hand!)

Calculating eigensystems in R

```
> cz <- cov(Z)
> cz

          [,1]      [,2]
[1,] 0.9158870 0.6485146
[2,] 0.6485146 0.9872295
```

```
> eig <- eigen(cz)
> eig$values

[1] 1.6010532 0.3020633

> eig$vectors

          [,1]       [,2]
[1,] 0.6874149 -0.7262650
[2,] 0.7262650  0.6874149

> eig

$values
[1] 1.6010532 0.3020633

$vectors
          [,1]       [,2]
[1,] 0.6874149 -0.7262650
[2,] 0.7262650  0.6874149
```

How much variation is explained?

The eigenvalues can be used to determine how much variance is explained by each eigenvector

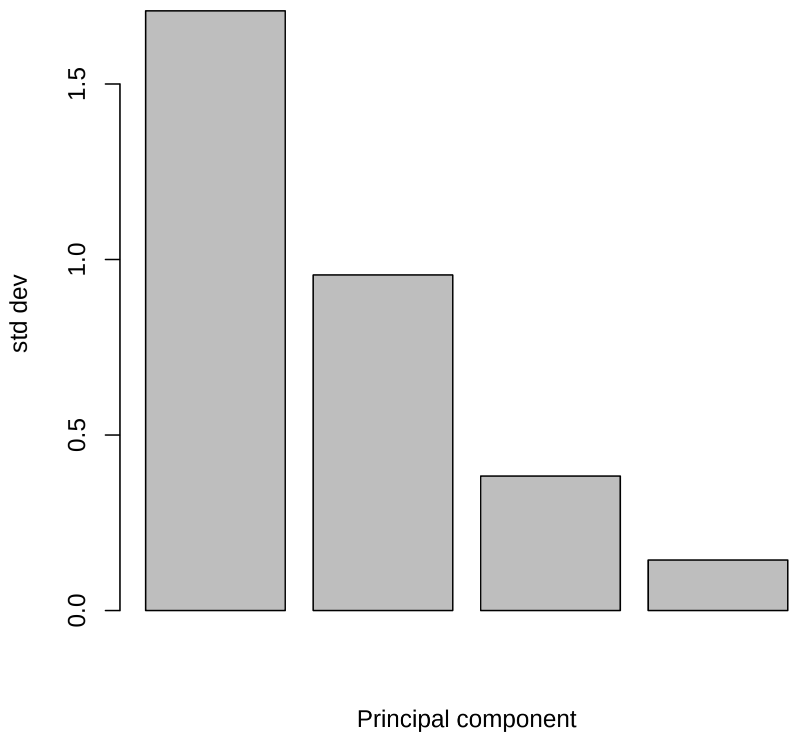
```
> ev <- eig$values
> 100 * (ev/sum(ev))

[1] 84.12797 15.87203
```

R functions

`prcomp()`, `princomp()`

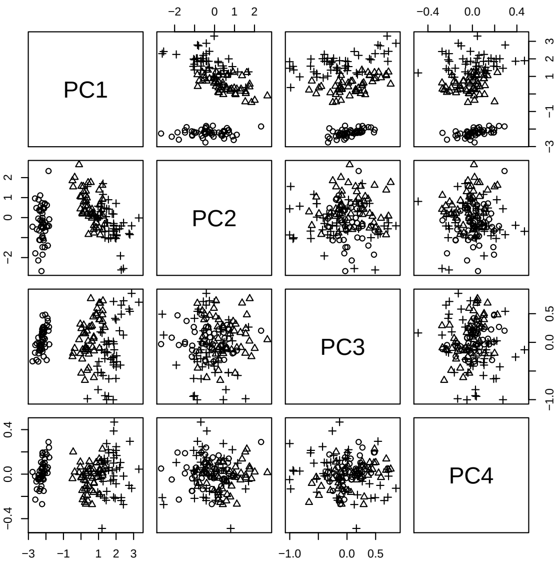
```
> data(iris)
> iris.pc <- prcomp(iris[, 1:4], scale = T)
> summary(iris.pc)

Importance of components:
                        PC1   PC2    PC3     PC4
Standard deviation     1.71 0.956 0.3831 0.14393
Proportion of Variance 0.73 0.229 0.0367 0.00518
Cumulative Proportion  0.73 0.958 0.9948 1.00000
```

Screeplots

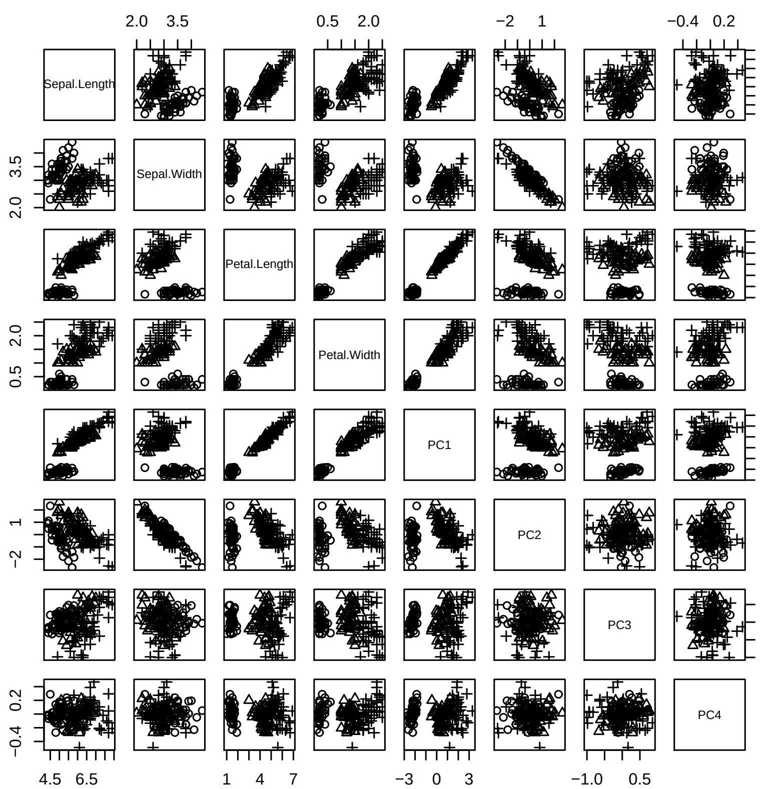


PCA scores

The eigenvalues and vectors can be used to *Transform* the coordinates of each variable into the orthogonal space defined by the eigenvectors

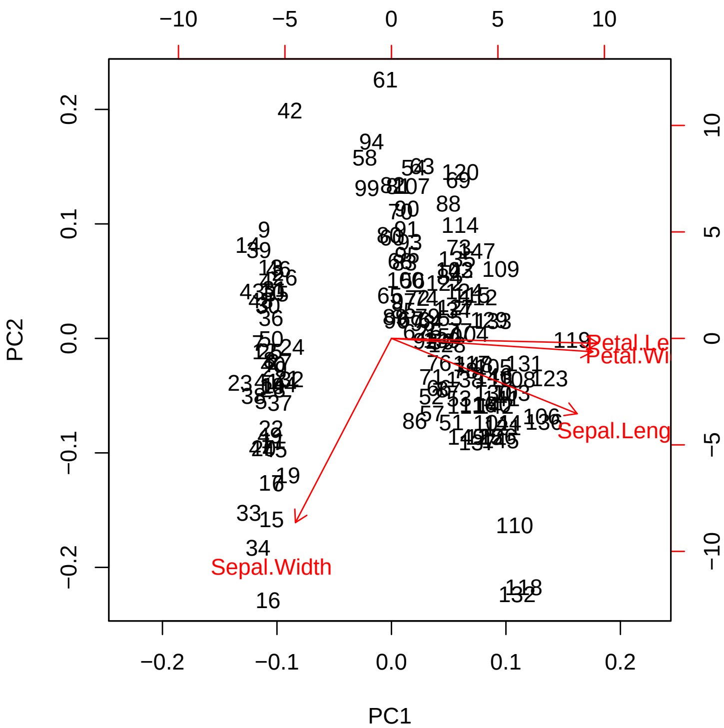


What variables contribute?

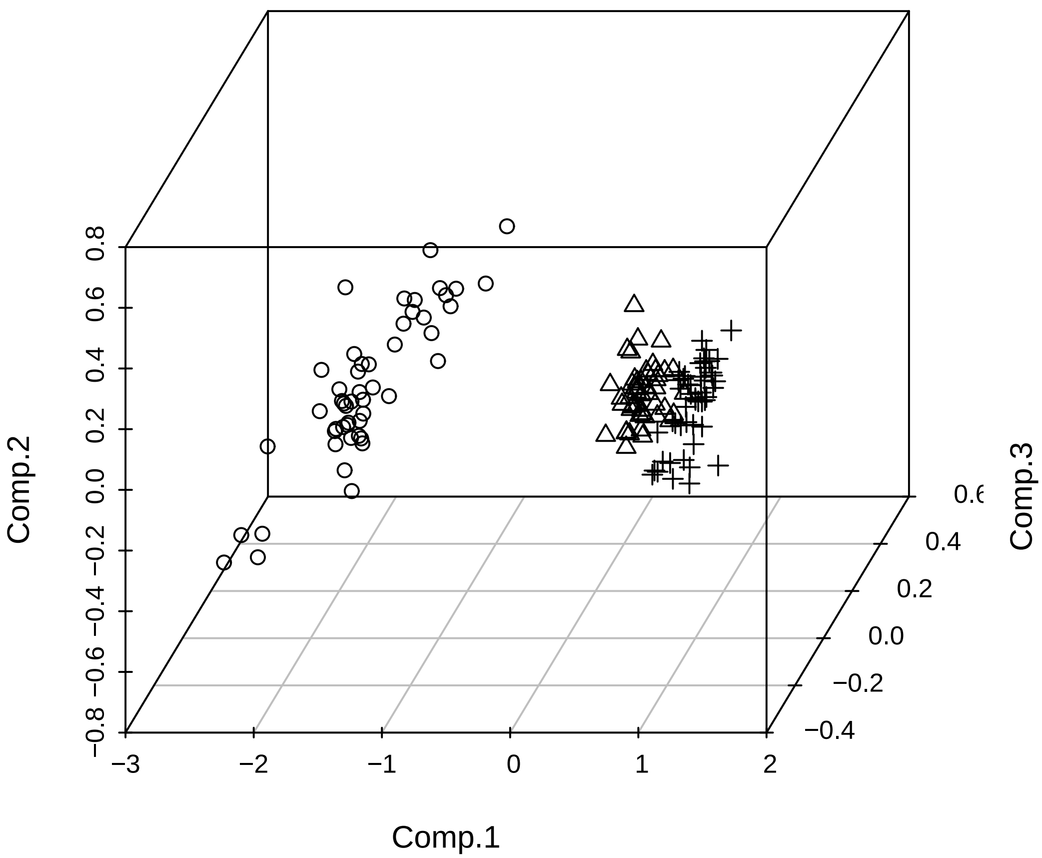


Biplots

```
> biplot(iris.pc)
```



Do the PCA scores help?



Conditions/Assumptions

- Variables all on a comparable scale with comparable variance.
  - If not, variables can be scaled using `'scale()'` or scale option in principle components functions
- multivariate normality
  - each variable has to be normally distributed
  - does not mean that they are multivariate normal though.
- PCA performs better with multivariate normal data, but is still robust to deviations from this assumption.

Uses of PCA

- Eliminate multicollinearity
- Reduce variables
- Examine patterns