MULTIVARIATE PROBABILITY DISTRIBUTIONS IN R

Reading multivariate data

Surajit Ray
Senior Lecturer, University of Glasgow

Course topics

- Read and analyze multivariate data
- Explore plotting techniques
- Use common statistical distributions
    - Gaussian and T distribution
- Techniques for high-dimensional data
    - Principal component analysis (PCA)

Structure of multivariate data

- Rectangular in shape - organized by rows and columns
  - Rows represent observations
  - Columns represent variables
- May or may not include:
  - Row names or numbers
  - Column headers
- Possible missing data

Multivariate data examples

Iris Data from Cambridge University website

```
5.1  3.5  1.4  0.2  1
4.9  3.0  1.4  0.2  1
4.7  3.2  1.3  0.2  1
```

Birth Weight data (CSV with column header)

```
"","case","bwt","gestation","parity","age","height","weight","smoke"
"1",1,120,284,0,27,62,100,0
"2",2,113,282,0,33,64,135,0
```

Reading data

From a URL

```
iris_url <- "http://mlg.eng.cam.ac.uk/teaching/3f3/1011/iris.data"
iris_raw <- read.table(iris_url,  sep ="", header = FALSE)
```

Locally

```
iris_raw <- read.table("iris.txt", sep = "", header = FALSE)
```

Viewing the dataset

```
head(iris_raw, n = 4)

     V1  V2  V3  V4 V5
1   5.1 3.5 1.4 0.2  1
2   4.9 3.0 1.4 0.2  1
3   4.7 3.2 1.3 0.2  1
4   4.6 3.1 1.5 0.2  1
```

Assigning column names

```
colnames(iris_raw) <- c("Sepal.Length", "Sepal.Width", "Petal.Length",
                        "Petal.Width",  "Species" )
```

```
head(iris_raw)

  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2       1
2          4.9         3.0          1.4         0.2       1
3          4.7         3.2          1.3         0.2       1
4          4.6         3.1          1.5         0.2       1
5          5.0         3.6          1.4         0.2       1
6          5.4         3.9          1.7         0.4       1
```

Accessing specific columns

Check current names of columns

```
names(iris_raw)

"Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

Accessing Sepal length and Sepal width columns

```
iris_raw[, 1:2]
iris[, c('Sepal.Length', 'Sepal.Width')]
```

Changing data types

Change the last variable `Species` to a factor

```
iris_raw$species <- as.factor(iris_raw$species)
```

```
str(iris_raw)

'data.frame':    150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
```

Assigning factor labels

Recode the species labels from `1`, `2` and `3` to `setosa`, `versicolor` and `virginica`

- Assign factor labels
- Change first variable to a factor

```
library(car)
iris_raw$Species <- recode(iris_raw$Species,
                           " 1 ='setosa'; 2 = 'versicolor'; 3 = 'virginica'")
```

```
str(iris_raw)

'data.frame':    150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1
```

Reading csv data with named columns

Birth Weight data (CSV with column header)

```
"","case","bwt","gestation","parity","age","height","weight","smoke"
"1",1,120,284,0,27,62,100,0
"2",2,113,282,0,33,64,135,0
"3",3,128,279,0,28,64,115,1
"4",4,123,NA,0,36,69,190,0
```

Reading Birth Weight data

```
bwt <- read.csv("birthweight.csv", row.names = 1)
head(bwt, n = 3)
```

```
  case bwt gestation parity age height weight smoke
1    1 120       284      0  27     62    100     0
2    2 113       282      0  33     64    135     0
3    3 128       279      0  28     64    115     1
```

MULTIVARIATE PROBABILITY DISTRIBUTIONS IN R

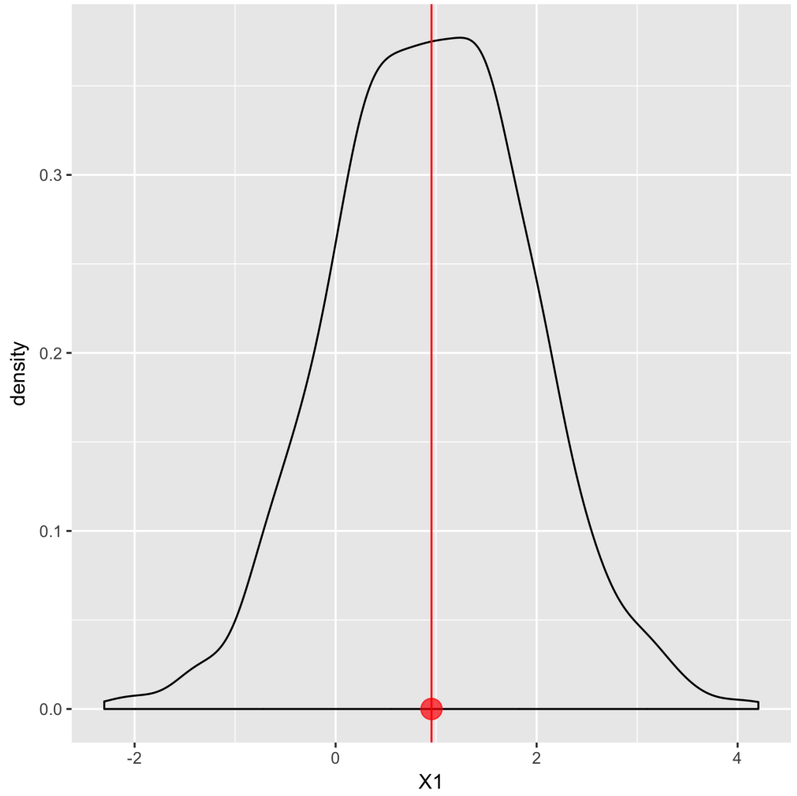
Let's read some multivariate data!

MULTIVARIATE PROBABILITY DISTRIBUTIONS IN R

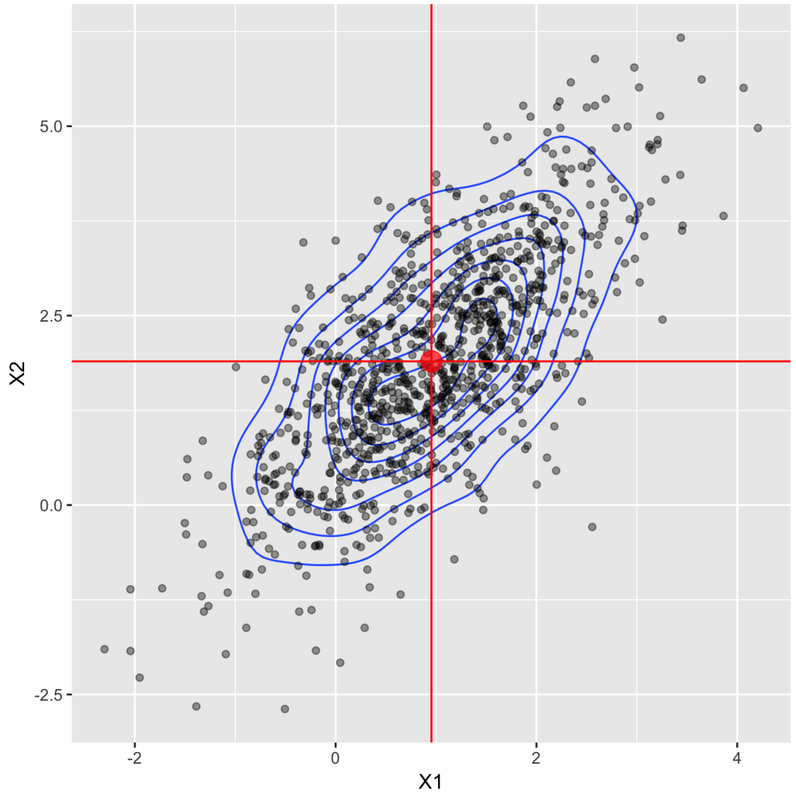
Mean vector and variance-covariance matrix

Surajit Ray
Senior Lecturer, University of Glasgow

Mean represents the location of the distribution

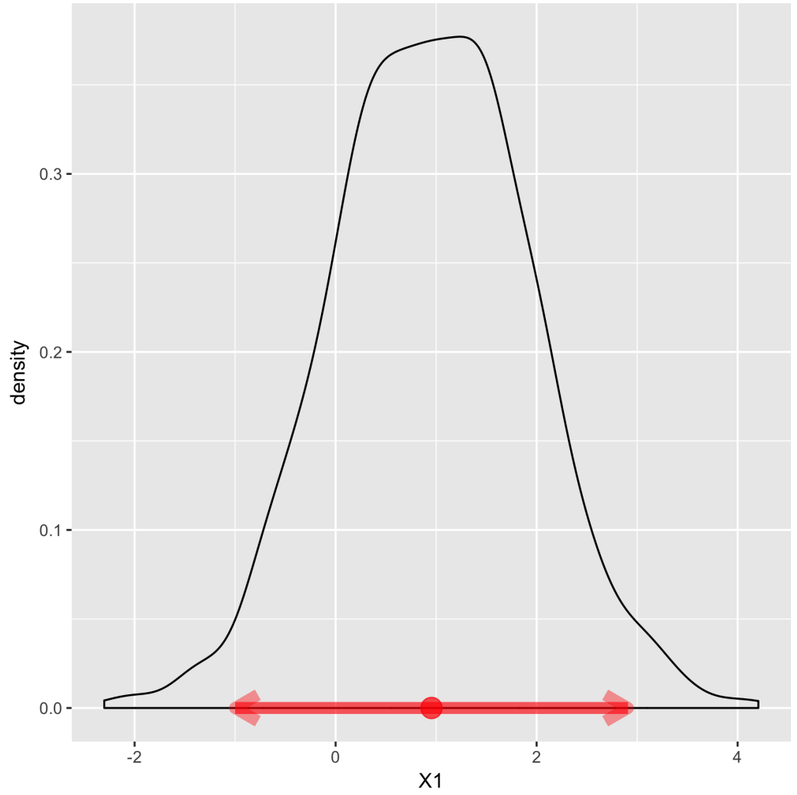


Mean 0.95

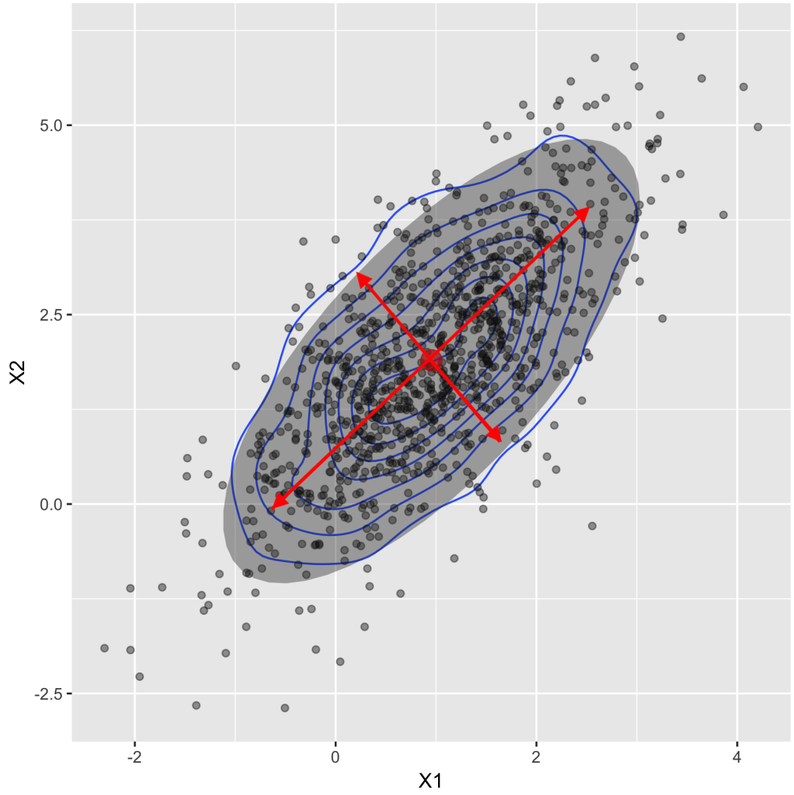


Mean vector (0.95 1.89)

Variance-covariance matrix represents the spread



Variance 1.02



Variance-covariance $\begin{pmatrix} 1.02 & 0.97 \\ 0.97 & 2 \end{pmatrix}$

Calculating the mean

```
colMeans(iris_raw[, 1:4])

Sepal.Length  Sepal.Width Petal.Length  Petal.Width
        5.84         3.05         3.76         1.20
```

Functions that calculate means by subgroups

- `by()`
- `aggregate()`

Calculating the group mean using by

```
by(data = iris[,1:4], INDICES = iris$Species, FUN = colMeans)
```

```
iris_raw$Species: setosa
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       5.006        3.418        1.464        0.244
------------------------------------------------------
iris_raw$Species: versicolor
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       5.936        2.770        4.260        1.326
------------------------------------------------------
iris_raw$Species: virginica
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       6.588        2.974        5.552        2.026
```

Calculating the group mean using aggregate

```
aggregate(. ~ Species, iris_raw, mean)
```

```
     Species Sepal.Length Sepal.Width Petal.Length Petal.Width
1     setosa         5.01        3.42         1.46       0.244
2 versicolor         5.94        2.77         4.26       1.326
3  virginica         6.59        2.97         5.55       2.026
```

Calculating the variance-covariance and correlation matrices

Variance

```
var(iris_raw[, 1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length       0.6857     -0.0393        1.274       0.517
Sepal.Width       -0.0393      0.1880       -0.322      -0.118
Petal.Length       1.2737     -0.3217        3.113       1.296
Petal.Width        0.5169     -0.1180        1.296       0.582
```

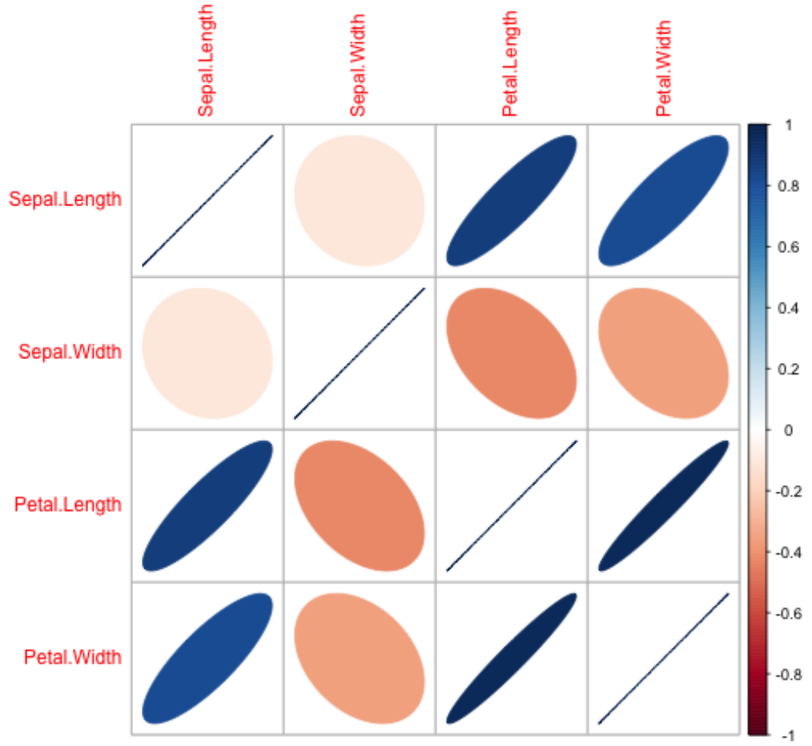
Correlation

```
cor(iris_raw[, 1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length        1.000      -0.109        0.872       0.818
Sepal.Width        -0.109       1.000       -0.421      -0.357
Petal.Length        0.872      -0.421        1.000       0.963
Petal.Width         0.818      -0.357        0.963       1.000
```

Visualization of correlation matrix

`corrplot` function to visualize correlation plot

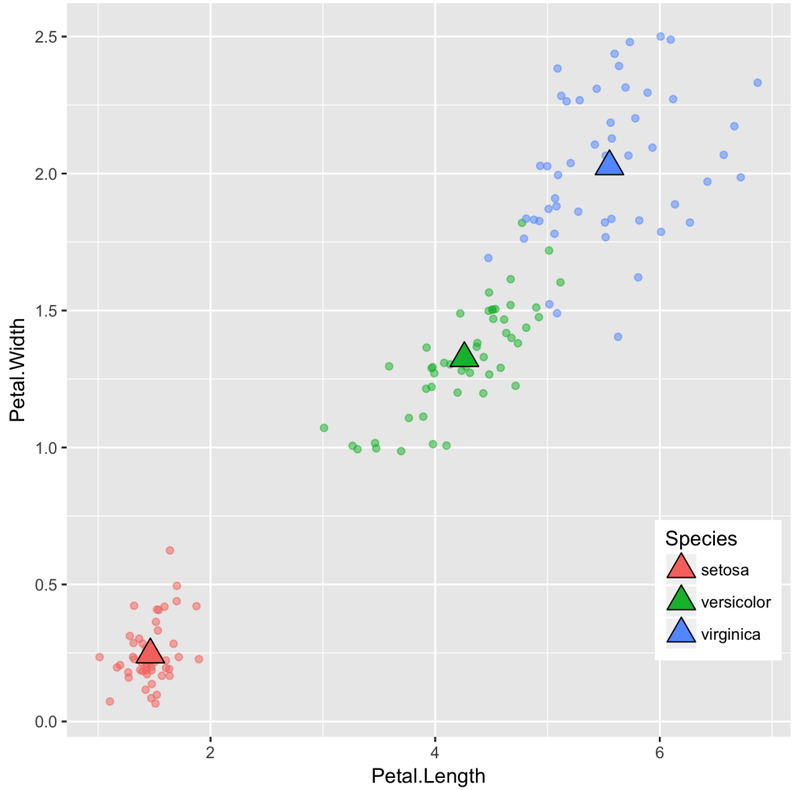
```
corrplot(cor(iris_raw[, 1:4]), method = "ellipse")
```



Interpretation of means

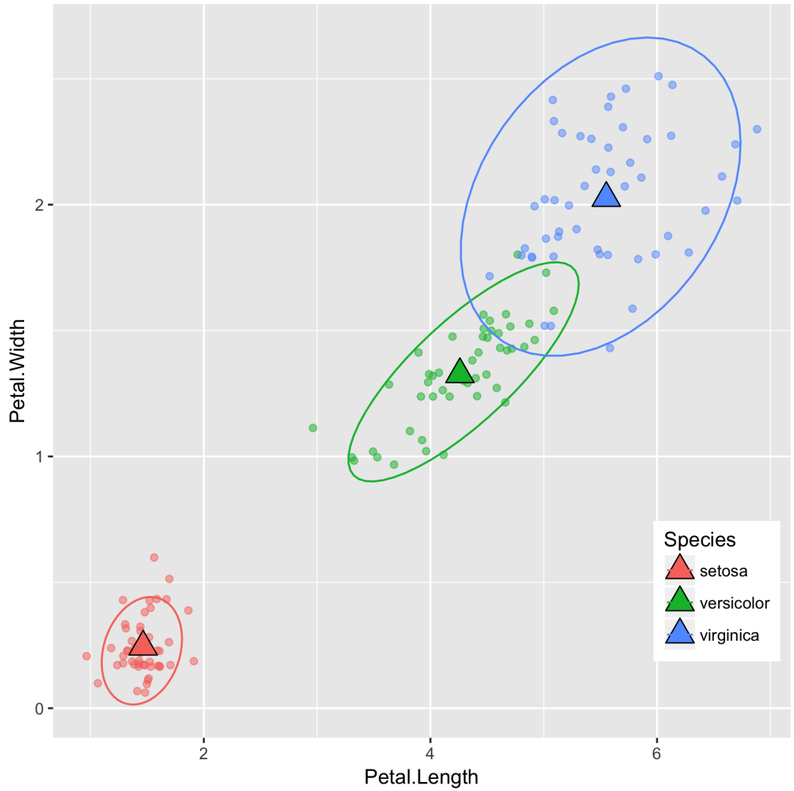
Means

Species	Petal.Length	Petal.Width
setosa	1.46	0.244
versicolor	4.26	1.326
virginica	5.55	2.026



Interpretation of variances

setosa	Petal.Length	Petal.Width
Petal.Length	0.030	0.006
Petal.Width	0.006	0.011
versicolor	Petal.Length	Petal.Width
Petal.Length	0.221	0.073
Petal.Width	0.073	0.039
virginica	Petal.Length	Petal.Width
Petal.Length	0.305	0.049
Petal.Width	0.049	0.075



MULTIVARIATE PROBABILITY DISTRIBUTIONS IN R

Let's practice!

MULTIVARIATE PROBABILITY DISTRIBUTIONS IN R

Plotting multivariate data

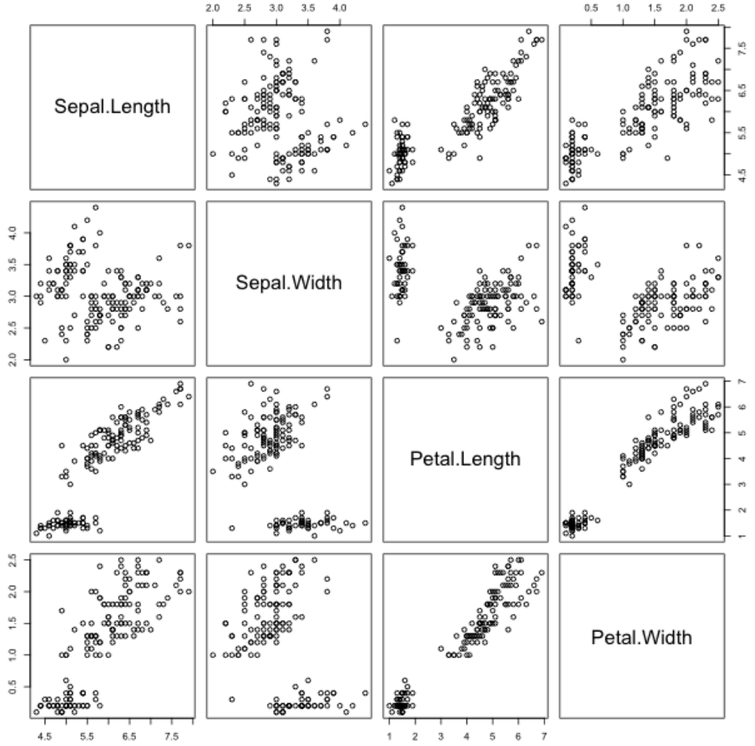
Surajit Ray
Senior Lecturer, University of Glasgow

Various plotting options

- Basic `R` plot
- `lattice` library
- `ggplot`
- 3D plotting options

Basic R plot for multivariate data

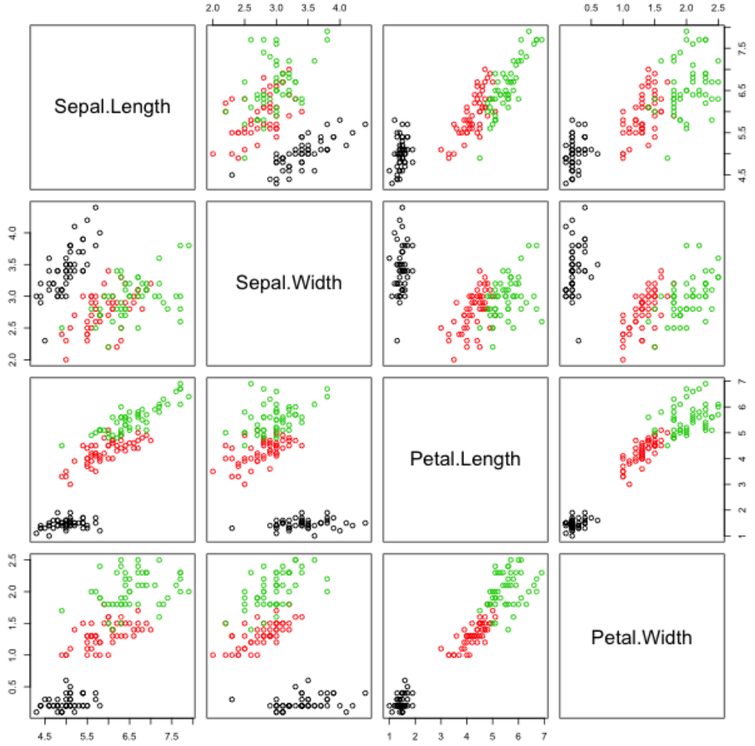
```
pairs(iris_raw[, 1:4])
```



- Plot not as useful with many variables

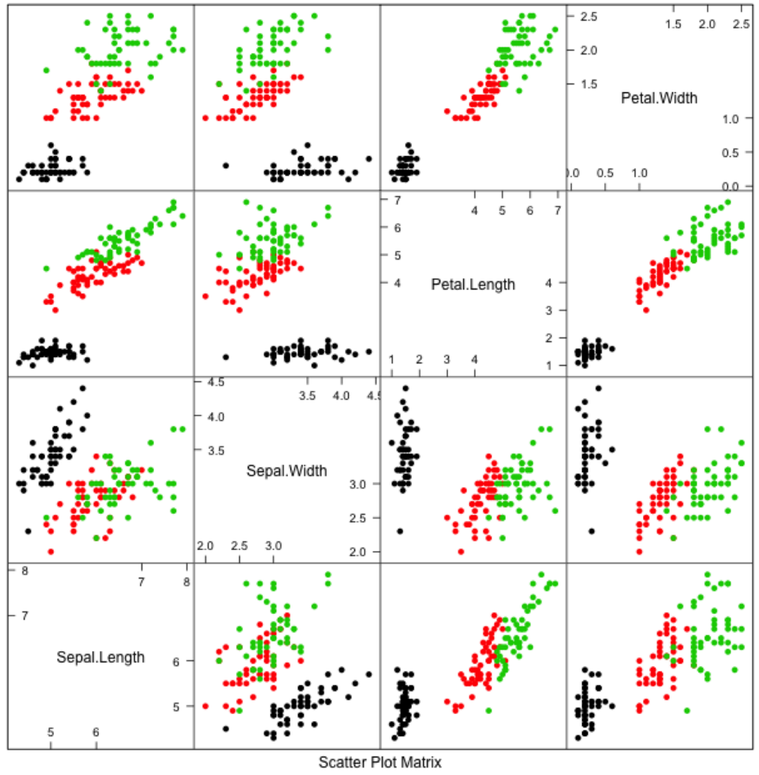
Pairs plot by color

```
pairs(iris_raw[, 1:4], col = iris_raw$Species)
```



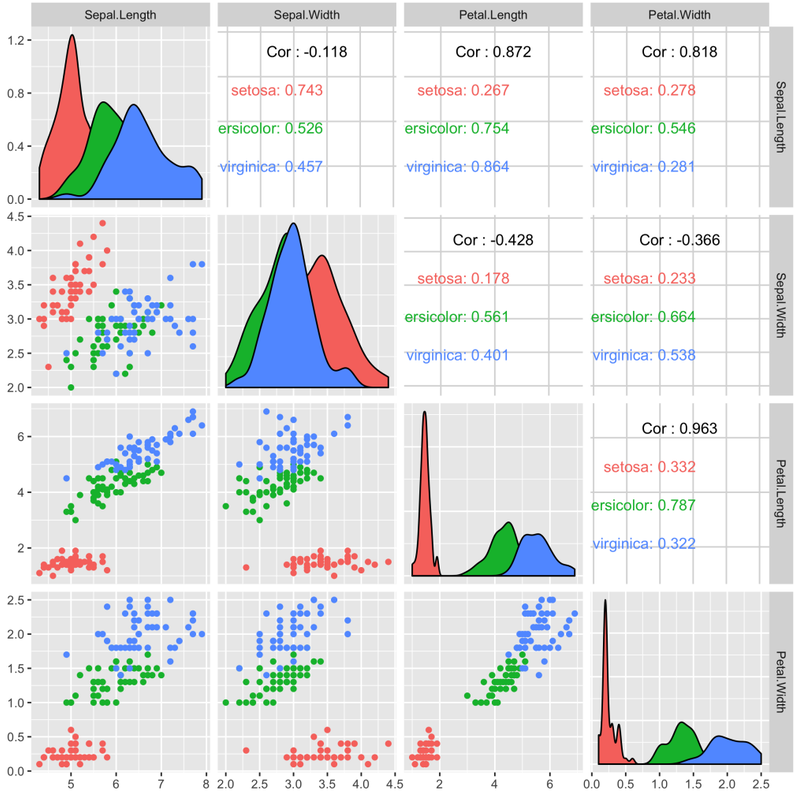
Lattice

```
library(lattice)
splom(~iris_raw[, 1:4], col = iris_raw$Species, pch = 16)
```



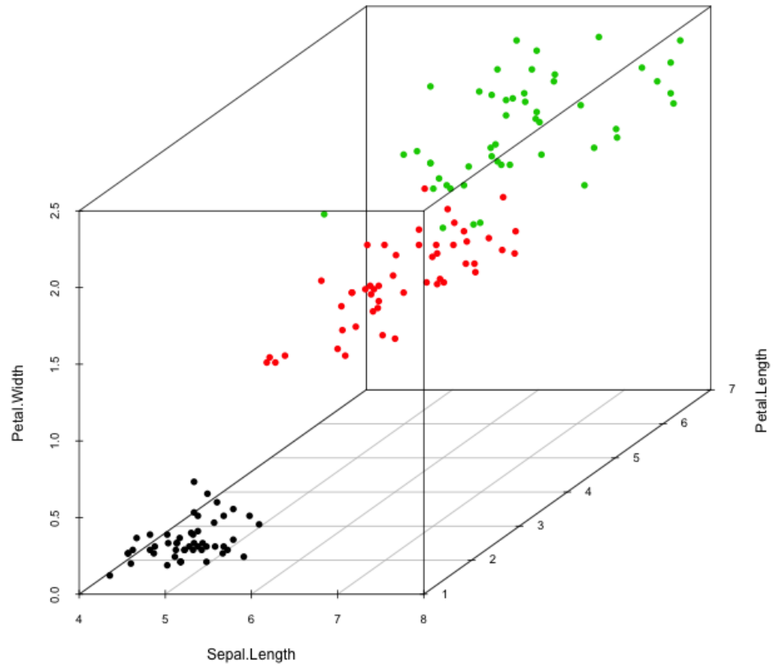
Using ggplot

```
library(ggplot2)
library(GGally)
ggpairs(data = iris_raw, columns = 1:4, mapping = aes(color = Species))
```



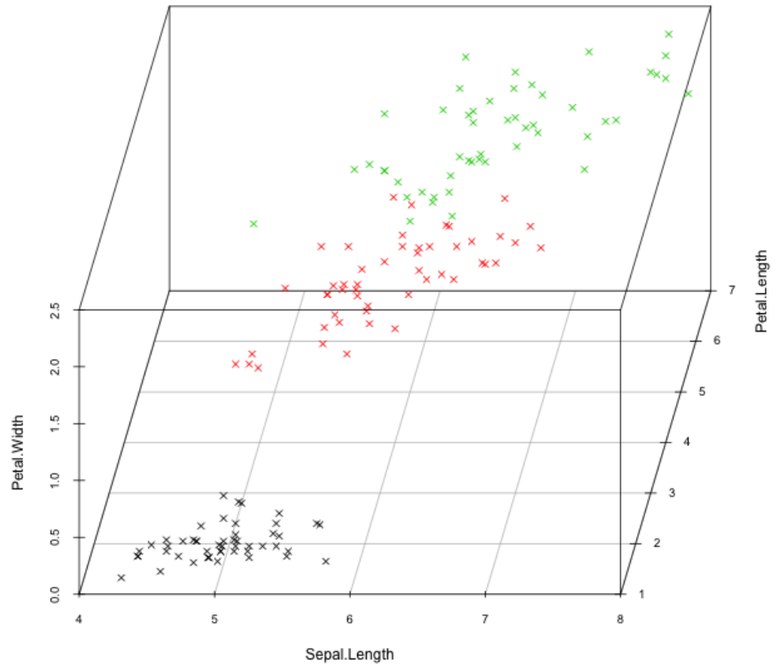
3D plots

```
library(scatterplot3d)
scatterplot3d(iris_raw[, c(1, 3, 4)], color = as.numeric(iris_raw$Species))
```



3D plots

```
scatterplot3d(iris_raw[, c(1, 3, 4)], color = as.numeric(iris_raw$Species),
              pch = 4, angle = 80)
```



MULTIVARIATE PROBABILITY DISTRIBUTIONS IN R

Let's practice some plotting with the wine data!