Honours Multivariate Analysis

Lecture 1 - Introduction, Examples of Multivariate Data, Visualization and Summary Statistics

Stefan S. Britz

Department of Statistical Sciences
University of Cape Town

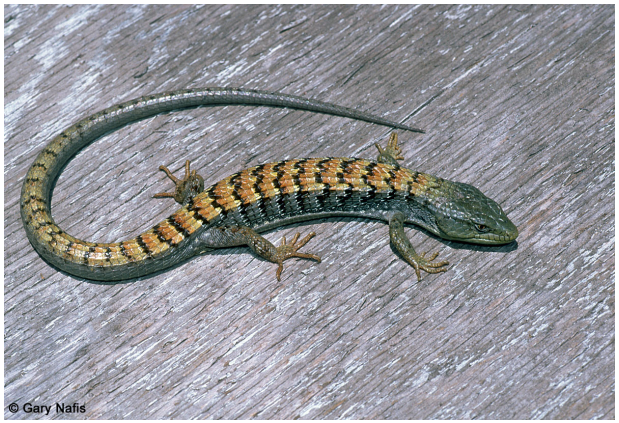
Course Outline

1. Introduction, Examples of Multivariate Data †
2. TOOLS
   1. Visualization and Summary Statistics †
   2. Singular Value Decomposition, Eigenvalue Decomposition and Spectral Decomposition revisited †
   3. The Multivariate Normal Distribution †
   4. Multivariate Maximum Likelihood Estimation †
   5. Multivariate Inference †
3. EXPLORATORY ANALYSIS
   1. Principal Component Analysis §
   2. Factor Analysis §
   3. Correspondence Analysis §
4. CONFIRMATORY ANALYSIS
   1. For grouped Multivariate Data:
      1. Manova †
      2. Discriminant Analysis §
   2. Regression
      1. Multivariate Regression §
      2. Canonical Correlation Analysis §

†Mr Stefan Britz

§Mr Miguel Rodo

Introduction



Multivariate Analysis is concerned with the analysis of multiple response variables. For example,

1. Mass (in grams), snout-vent length (in mm) and hind limb span (in mm) measurements for lizards.
2. Air pollution data consisting of 42 measurements of 7 air-pollution variables recorded on different days.
3. National Track records for men and women.

Examples of datasets

```
head(lizards)
```

```
     Mass   SVL    HLS
1   5.526  59.0  113.5
2  10.401  75.0  142.0
3   9.213  69.0  124.0
4   8.953  67.5  125.0
5   7.063  62.0  129.5
6   6.610  62.0  123.0
```

```
head(AirPollution)
```

```
  Wind Radiation CO NO NO2 O3 HC
1    8        98  7  2  12  8  2
2    7       107  4  3   9  5  3
3    7       103  4  3   5  6  3
4   10        88  5  2   8 15  4
5    6        91  4  2   8 10  3
6    8        90  5  2  12 12  4
```

Examples of datasets

```
head(Track)
```

```
  Country  m100  m200  m400 m800 m1500 m3000 Marathon Gender
1     ARG 11.57 22.94 52.50 2.05  4.25  9.19   150.32 Female
2     AUS 11.12 22.23 48.63 1.98  4.02  8.63   143.51 Female
3     AUT 11.15 22.70 50.62 1.94  4.05  8.78   154.35 Female
4     BEL 11.14 22.48 51.45 1.97  4.08  8.82   143.05 Female
5     BER 11.46 23.05 53.30 2.07  4.29  9.81   174.18 Female
6     BRA 11.17 22.60 50.62 1.97  4.17  9.04   147.41 Female
```

Matrix representation of data

Data can be represented in the form of a matrix where the rows reflect the observations and the columns reflect the variables. Typically

$$n = \text{Number of observations}$$

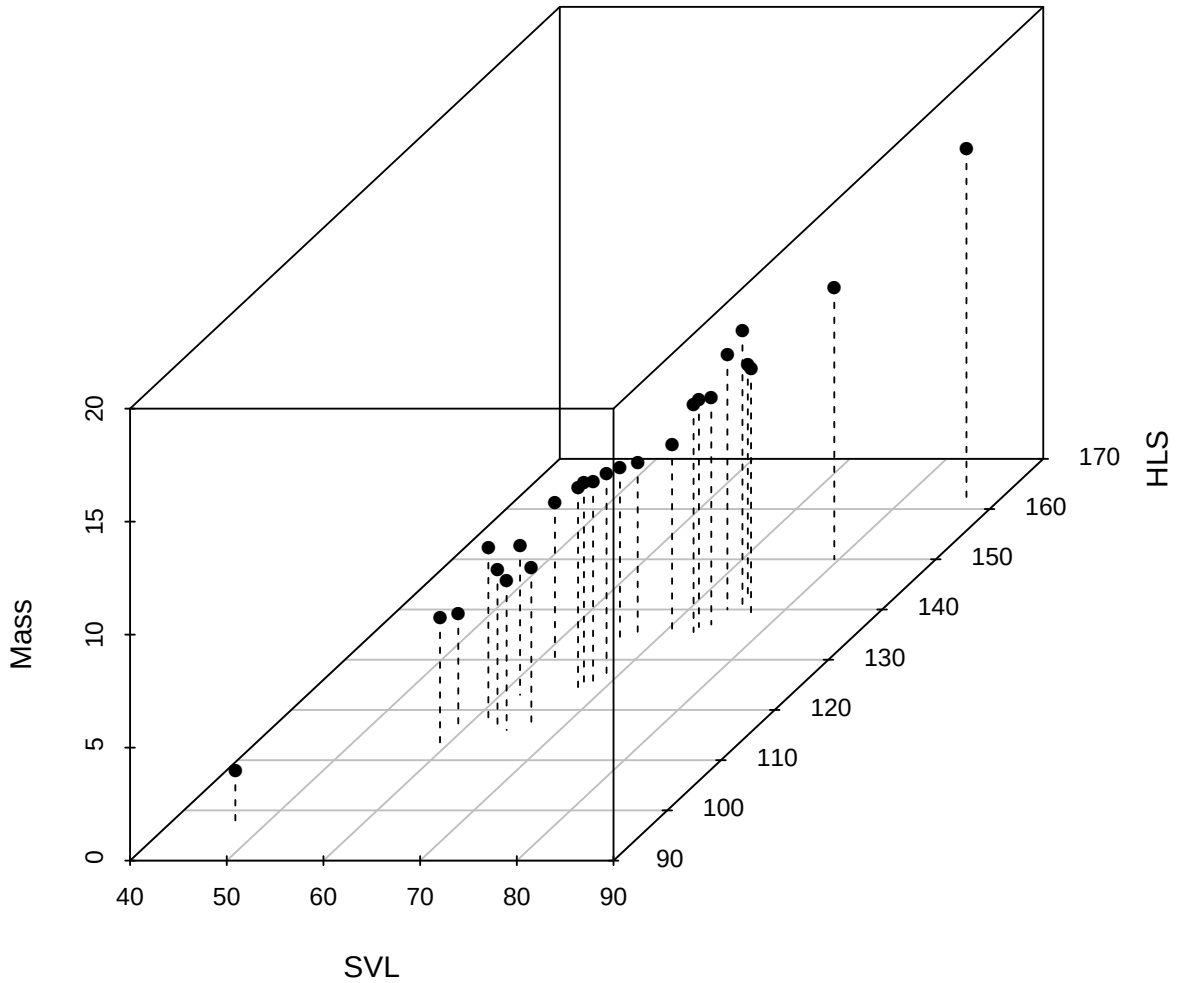
$$p = \text{Number of variables}$$

$$\boldsymbol{X}_{n\times p} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}$$

So for the lizard data,

$$\boldsymbol{X}_{25\times 3} = \begin{bmatrix} 5.526 & 59.0 & 113.5 \\ 10.401 & 75.0 & 142.0 \\ \vdots & \vdots & \vdots \\ 6.890 & 63.0 & 117.0 \end{bmatrix}$$

The data can be viewed as n points in $p-$dimensional space



Summary Statistics: Mean Vector

Similar to the univariate case, we can define a sample **mean** and variance, as well as covariances and correlations.

$$\bar{\boldsymbol{x}}_{p\times 1} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ is the sample mean of the j^{th} variable (j^{th} column of $\boldsymbol{X}$)

Can be written as

$$\bar{\boldsymbol{x}} = \frac{1}{n}\boldsymbol{X}'\mathbf{1}$$

where $\mathbf{1}_{n\times 1}$ is a vector of ones

Summary Statistics: Sample Mean Vector

Example:

If

$$\boldsymbol{X}_{3\times 2} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

then

$$\frac{1}{n} \times \boldsymbol{X}'\mathbf{1} = \frac{1}{3}\begin{bmatrix} 4 & -1 & 3 \\ 1 & 3 & 5 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 6/3 = 2 \\ 9/3 = 3 \end{bmatrix}$$

Summary Statistics: Sample Covariance Matrix

Similar to the univariate case, we can define a sample mean and **variance**, as well as **covariances** and correlations.

$\boldsymbol{S}_{p\times p}$ is a symmetric matrix with

$$s_{jj} = \mathrm{Var}(X_j) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

$$s_{jk} = \mathrm{Cov}(X_j, X_k) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

Summary Statistics: Sample Covariance Matrix

In matrix notation, the **sample covariance matrix** can be expressed in terms of:

$\mathbf{1}\bar{\boldsymbol{x}}' = \frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{X} \rightarrow$ an $n \times p$ matrix of means, and

$\boldsymbol{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{X} \rightarrow$ an $n \times p$ matrix of deviations such that:

Summary Statistics: Sample Covariance Matrix

In matrix notation, the **sample covariance matrix** can be expressed in terms of:

$\mathbf{1}\bar{\boldsymbol{x}}' = \frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{X} \rightarrow$ an $n \times p$ matrix of means, and

$\boldsymbol{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{X} \rightarrow$ an $n \times p$ matrix of deviations such that:

$$\boldsymbol{S} = \frac{1}{n-1}\left[\left(\boldsymbol{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{X}\right)'\left(\boldsymbol{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{X}\right)\right]$$

$$\begin{aligned}
(n-1)\boldsymbol{S} &= \left(\boldsymbol{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{X}\right)'\left(\boldsymbol{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{X}\right) \\
&= \boldsymbol{X}'\left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)'\left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\boldsymbol{X} \\
&= \boldsymbol{X}'\left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\boldsymbol{X} \\
&= \boldsymbol{X}'\boldsymbol{X} - \frac{1}{n}\boldsymbol{X}'\mathbf{1}\mathbf{1}'\boldsymbol{X} \\
&= \boldsymbol{X}'\boldsymbol{X} - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}'
\end{aligned}$$

Summary Statistics: Sample Correlation Matrix

Similar to the univariate case, we can define a sample mean and variance, as well as covariances and **correlations**.

$$\boldsymbol{R}_{p\times p} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

where

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}}$$

Can be written as

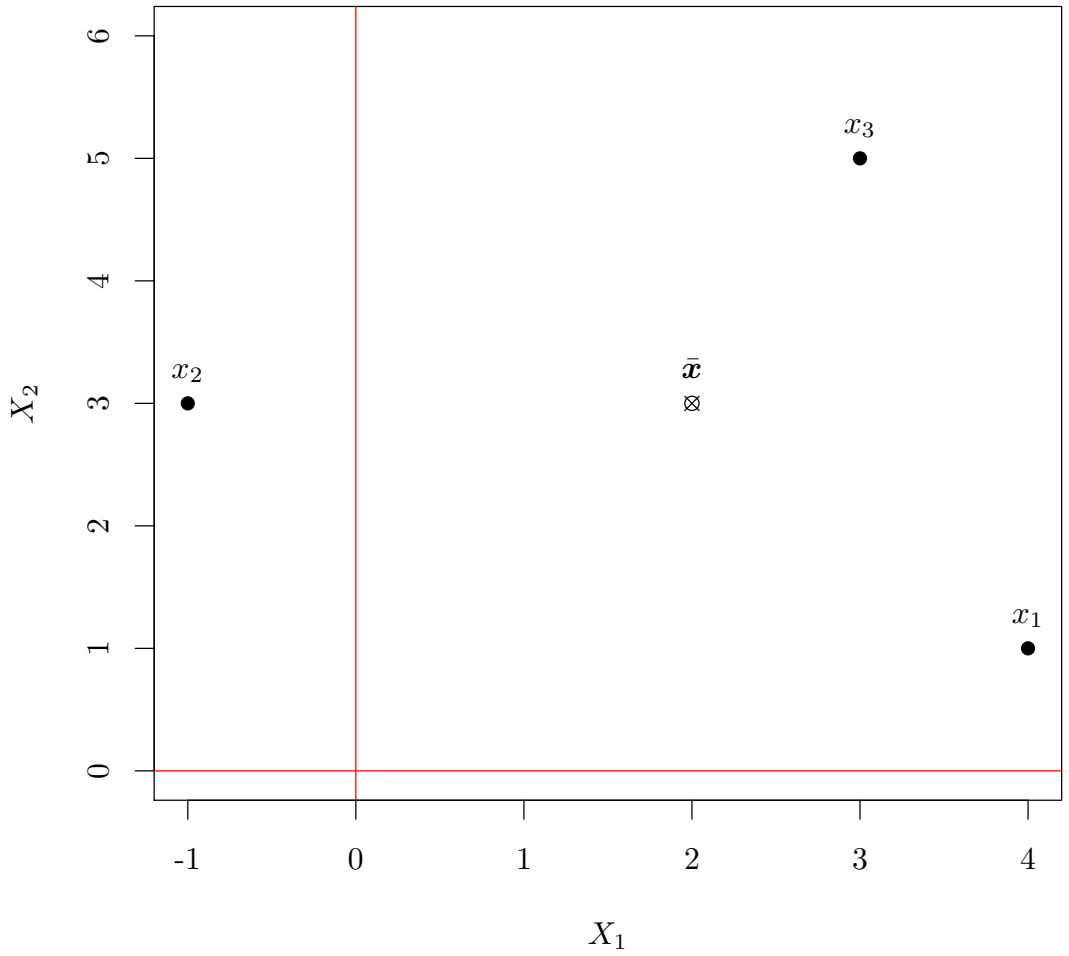
$$\boldsymbol{R} = [diag(\boldsymbol{S})]^{-\frac{1}{2}}\boldsymbol{S}[diag(\boldsymbol{S})]^{-\frac{1}{2}} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{D}^{-\frac{1}{2}}$$

Summary Statistics: Applications in R

See `Lecture1_1.R`

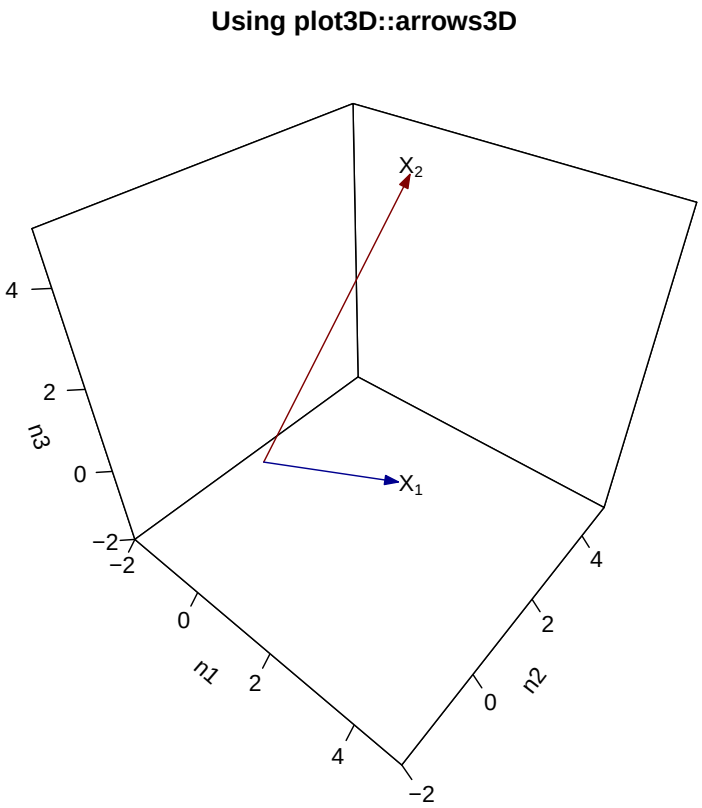
Graphical interpretation of a matrix

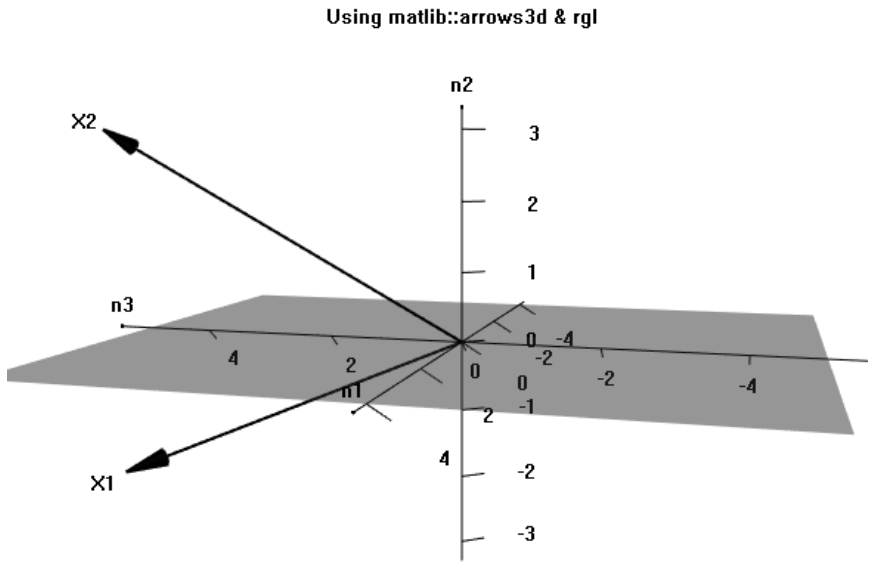
An $n \times p$ matrix $\boldsymbol{X}$ can be viewed as n observations in $p-$dimensional space...



Graphical interpretation of a matrix

...or as p vectors in $n-$dimensional space.





Deviation Vectors

We note that the expressions for the covariance matrix:

$$\boldsymbol{S} = \frac{1}{n-1}\left[\left(\boldsymbol{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{X}\right)'\left(\boldsymbol{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{X}\right)\right]$$

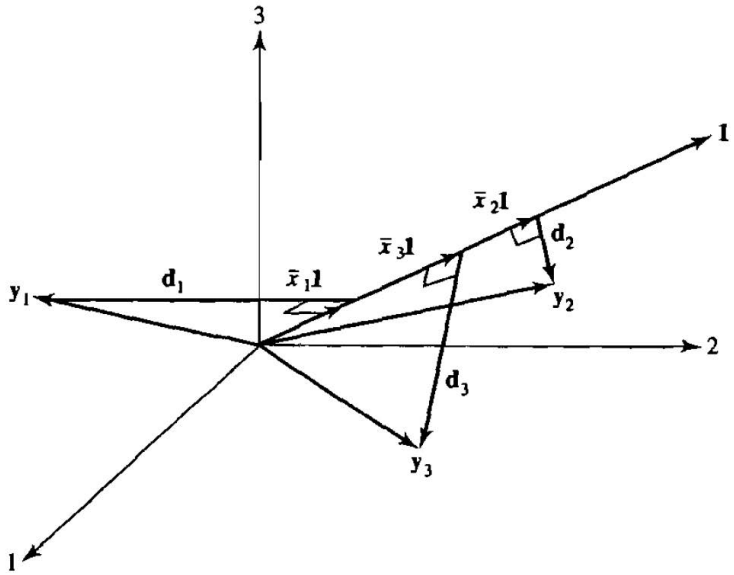
depend on the deviation of the observed values from the variable means. We can define and calculate so-called "deviation vectors" and then calculate the elements of the covariance matrix using these deviation vectors. The graphical display and interpretation of the deviation vectors is also informative.

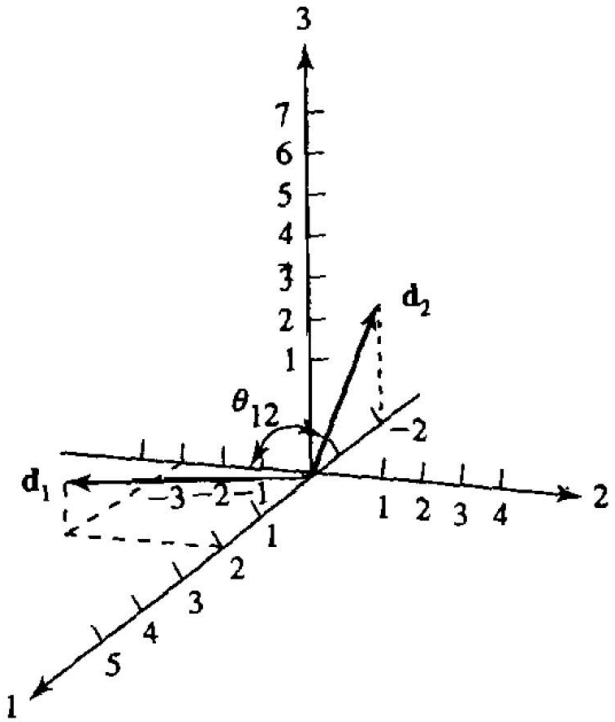
For an illustration, continuing with the previous example, see again `Lecture1_1.R`.

Geometry of deviation vectors

1. The projection of a column $\boldsymbol{x}_i$ of the data matrix $\boldsymbol{X}$ into the equal angular vector $\mathbf{1}$ is the vector $\bar{x}_i\mathbf{1}$ with length $\sqrt{n}|\bar{x}_i|$, i.e., the i^{th} sample mean is related to the projection of $\boldsymbol{x}_i$ on $\mathbf{1}$.
2. The information comprising $\boldsymbol{S}$ is obtained from the deviation vectors $\boldsymbol{d}_i = \boldsymbol{x}_i - \bar{x}_i\mathbf{1}$. The square of the length of $\boldsymbol{d}_i$ is $(n_i - 1)s_{ii}$ and the inner product between $\boldsymbol{d}_i$ and $\boldsymbol{d}_k$ is $(n-1)s_{ik}$.
3. The sample correlation coefficient r_{ik} is the cosine of the angle between $\boldsymbol{d}_i$ and $\boldsymbol{d}_k$.

Geometry of deviation vectors





Homework exercise 1

Johnson & Wichern exercise 3.1.

Given the data matrix

$$\mathbf{X} = \begin{bmatrix} 9 & 1 \\ 5 & 3 \\ 1 & 2 \end{bmatrix}$$

1. Graph the scatterplot in $p = 2$ dimensions. Locate the sample mean on your diagram.
2. Sketch the $n = 3$ dimensional representation of the data, and plot the two deviation vectors.
3. Sketch the deviation vectors in (2) emanating from the origin. Calculate the lengths of these vectors and the cosine of the angle between them. Relate these quantities to $\boldsymbol{S}$ and $\boldsymbol{R}$.

Random Vectors

Let $\boldsymbol{X}_{p\times 1}$ be a random vector whose components are random variables

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

where each random variable X_j has its own marginal probability distribution $f_{X_j}(x)$.

Expected value of X

$$\boldsymbol{\mu}_{p\times 1} = \mathrm{E}(\boldsymbol{X}) = \begin{bmatrix} \mathrm{E}(X_1) \\ \mathrm{E}(X_2) \\ \vdots \\ \mathrm{E}(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

where

$$\mu_j = \mathrm{E}(X_j) = \begin{cases} \displaystyle\int_{-\infty}^{\infty} x f_{X_j}(x)\, dx \\ \displaystyle\sum_x x \Pr(X_j = x) \end{cases}$$

Covariance & Correlation Matrix of $\boldsymbol{X}$

The covariance matrix $\boldsymbol{\Sigma}_{p\times p}$ is symmetric, positive semi-definite, and contains the pairwise covariances:

$$\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{X}) = \mathrm{E}\left[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})'\right]$$

with

$$\sigma_{jj} = \mathrm{Var}(X_j) = \mathrm{E}[(X_j-\mu_j)^2]$$

and

$$\sigma_{jk} = \mathrm{Cov}(X_j, X_k) = \mathrm{E}[(X_j-\mu_j)(X_k-\mu_k)]$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

Covariance & Correlation Matrix of $\boldsymbol{X}$

Similarly, the correlation matrix $\boldsymbol{P}_{p\times p}$ contains the pairwise correlations:

$$\boldsymbol{P} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

which can be written as

$$\boldsymbol{P} = [diag(\boldsymbol{\Sigma})]^{-\frac{1}{2}} \boldsymbol{\Sigma} [diag(\boldsymbol{\Sigma})]^{-\frac{1}{2}}$$

Moments of Sample Mean and Covariance Matrix

We consider the rows of the $\boldsymbol{X}$ matrix to represent a random sample from the joint distribution of the p variables with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Therefore, $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ are random variables! We'll look at their sampling distributions in detail later, for now we'll focus on their first central moments:

$$\mathrm{E}(\bar{\boldsymbol{X}}) = \boldsymbol{\mu} \text{ and } \mathrm{Cov}(\bar{\boldsymbol{X}}) = \frac{1}{n}\boldsymbol{\Sigma}$$

$$\mathrm{E}(\boldsymbol{S}) = \boldsymbol{\Sigma}$$

Therefore,

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'$$

is an unbiased estimator of $\boldsymbol{\Sigma}$.

Linear Combinations of Variables

Let $\boldsymbol{X}_{p\times 1}$ be a random vector with $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Now consider the following two linear combinations of the variables contained in $\boldsymbol{X}$: $\boldsymbol{b'X} = b_1X_1 + b_2X_2 + \ldots + b_pX_p$ and $\boldsymbol{c'X} = c_1X_1 + c_2X_2 + \ldots + c_pX_p$.

Then

1. the mean of $\boldsymbol{b'X}$ is $\mathrm{E}(\boldsymbol{b'X}) = \boldsymbol{b'\mu}$. Likewise $\mathrm{E}(\boldsymbol{c'X}) = \boldsymbol{c'\mu}$.
2. the variance of $\boldsymbol{b'X}$ is $\mathrm{Var}(\boldsymbol{b'X}) = \boldsymbol{b'\Sigma b}$. Likewise $\mathrm{Var}(\boldsymbol{c'X}) = \boldsymbol{c'\Sigma c}$.
3. the covariance of $\boldsymbol{b'X}$ and $\boldsymbol{c'X}$ is $\mathrm{Cov}(\boldsymbol{b'X}, \boldsymbol{c'X}) = \boldsymbol{b'\Sigma c}$.

Linear Combinations of Variables

Let $\boldsymbol{X}_{p\times 1}$ be a random vector with $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Now consider the following two linear combinations of the variables contained in $\boldsymbol{X}$: $\boldsymbol{b'X} = b_1X_1 + b_2X_2 + \ldots + b_pX_p$ and $\boldsymbol{c'X} = c_1X_1 + c_2X_2 + \ldots + c_pX_p$.

Then

1. the mean of $\boldsymbol{b'X}$ is $\mathrm{E}(\boldsymbol{b'X}) = \boldsymbol{b'\mu}$. Likewise $\mathrm{E}(\boldsymbol{c'X}) = \boldsymbol{c'\mu}$.
2. the variance of $\boldsymbol{b'X}$ is $\mathrm{Var}(\boldsymbol{b'X}) = \boldsymbol{b'\Sigma b}$. Likewise $\mathrm{Var}(\boldsymbol{c'X}) = \boldsymbol{c'\Sigma c}$.
3. the covariance of $\boldsymbol{b'X}$ and $\boldsymbol{c'X}$ is $\mathrm{Cov}(\boldsymbol{b'X}, \boldsymbol{c'X}) = \boldsymbol{b'\Sigma c}$.

Homework exercise: Johnson & Wichern exercise 3.18

Energy consumption by state from the major sources, x_1 =petroleum, x_2 =natural gas, x_3 = hydroelectric power and x_4 = nuclear electric power is recorded in quadrillions (10^{15}).

Homework exercise 2

The resulting mean and covariance matrix are

$$\bar{\mathbf{x}} = \begin{bmatrix} 0.766 \\ 0.508 \\ 0.438 \\ 0.161 \end{bmatrix}$$

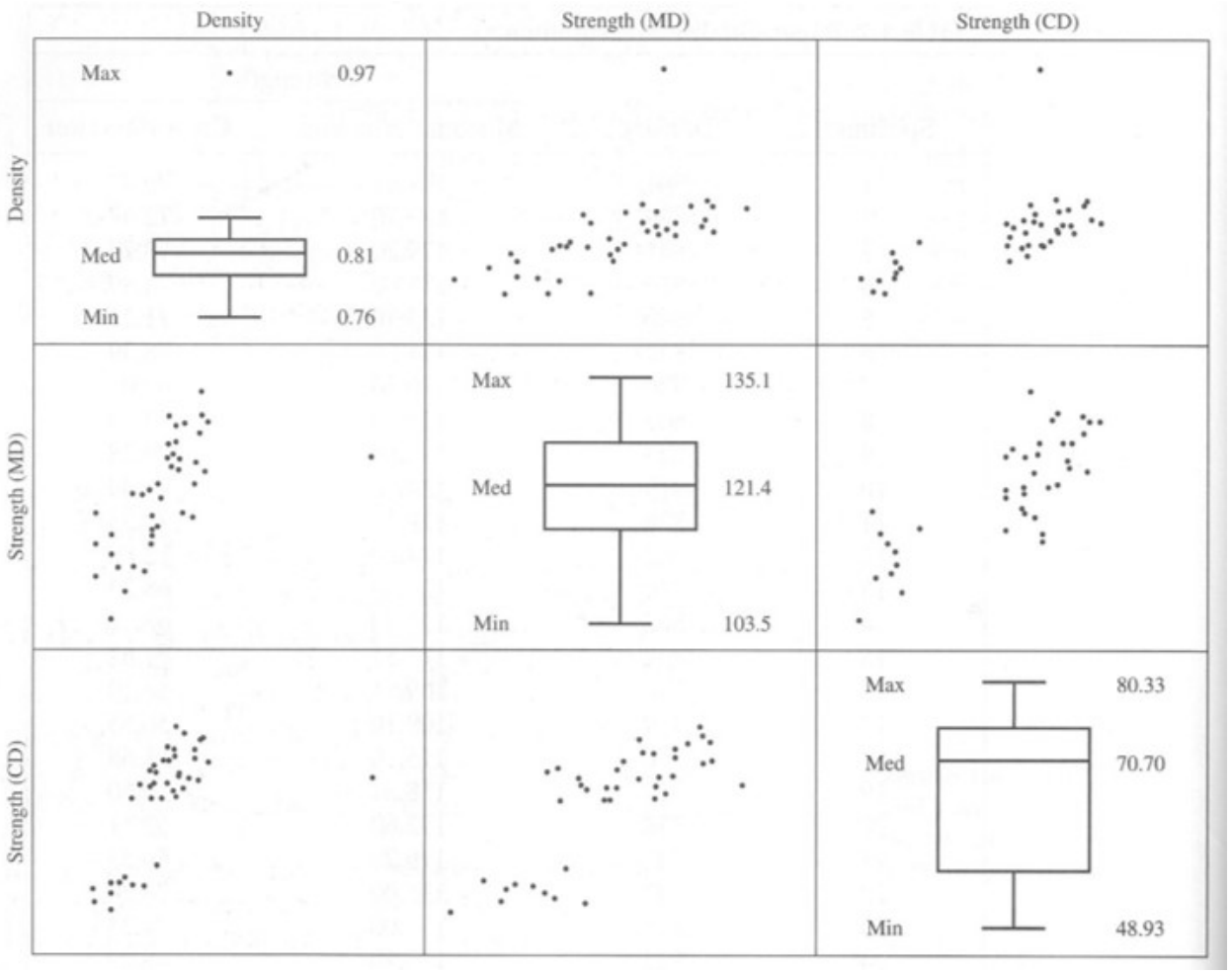
and

$$\mathbf{S} = \begin{bmatrix} 0.856 & 0.635 & 0.173 & 0.096 \\ 0.635 & 0.568 & 0.128 & 0.067 \\ 0.173 & 0.128 & 0.171 & 0.039 \\ 0.096 & 0.067 & 0.039 & 0.043 \end{bmatrix}$$

1. Determine the sample mean and variance of a state's total energy consumption for these major sources.
2. Determine the sample mean and variance of the excess of petroleum consumption over natural gas consumption. Also find the sample covariance of this variable with the total variable in part 1.

Graphical Summaries and Exploratory Plots in **R**

Function: pairs(...) #See `Lecture1_2.R` for examples on usage

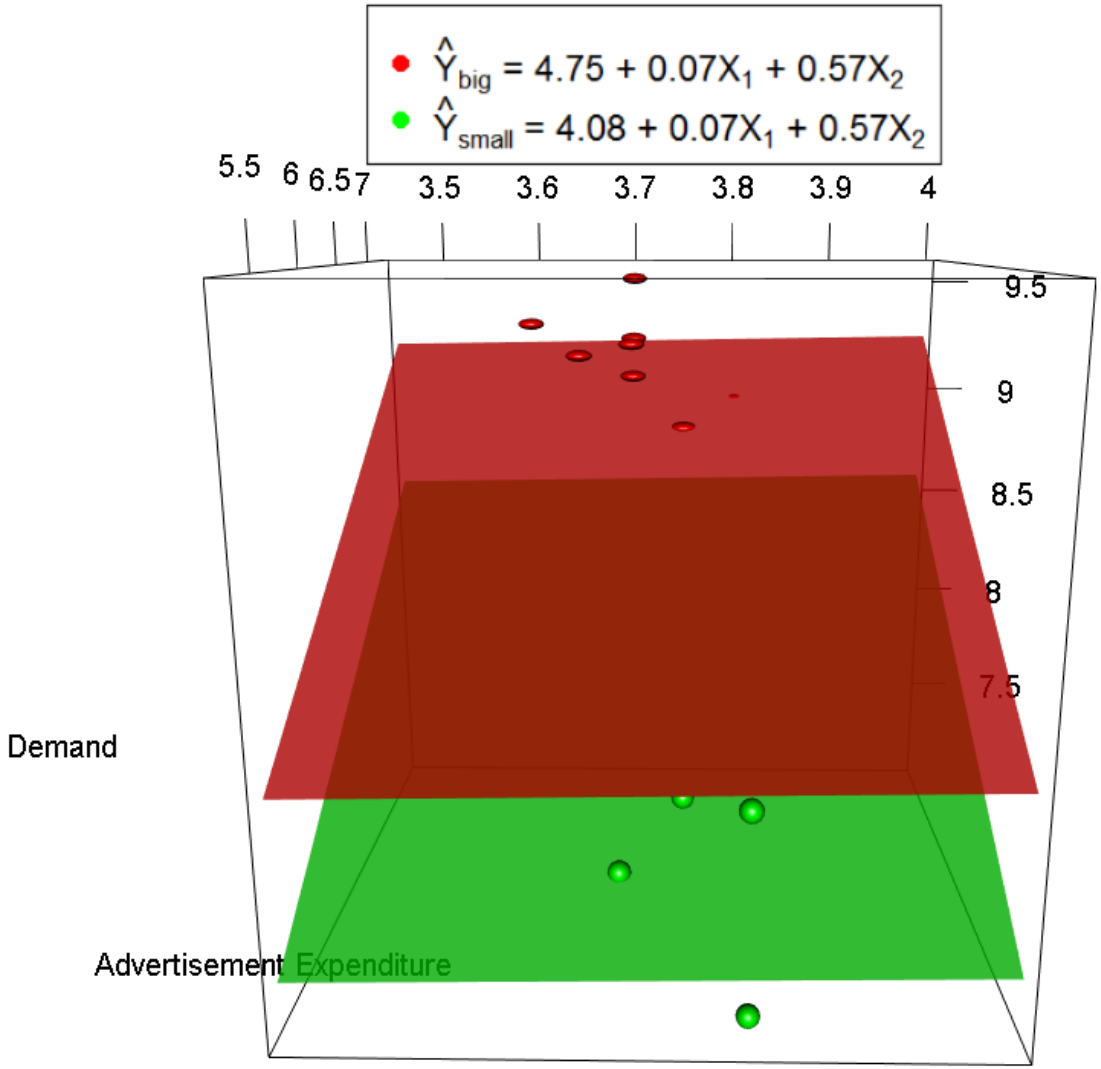


Graphical Displays in **R**

Package: rgl

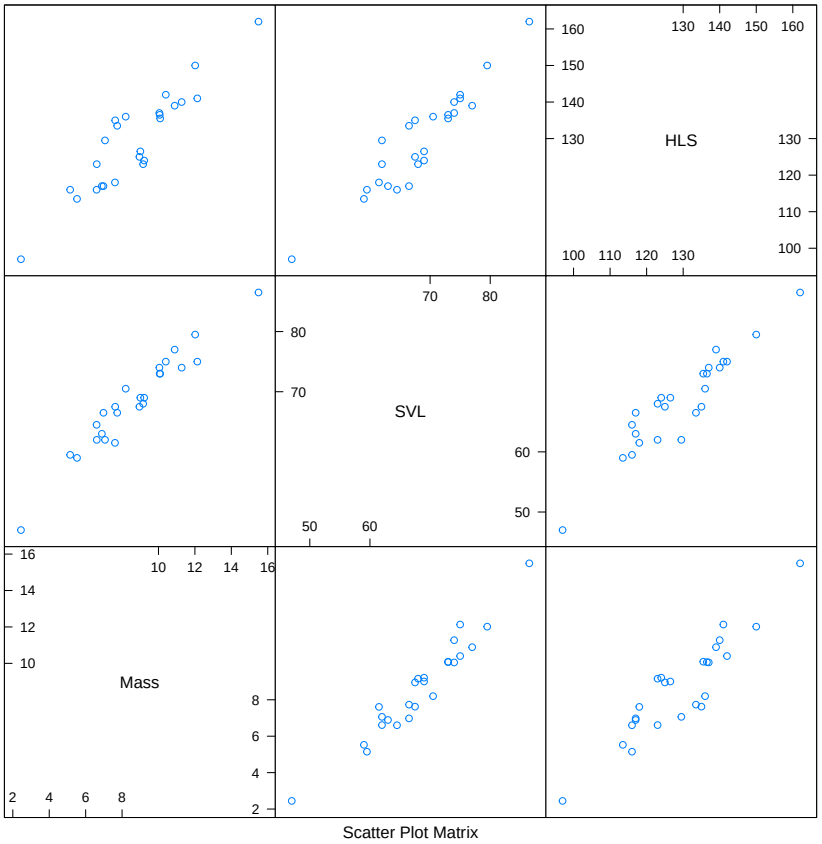
Functions: plot3d
spheres3d
points3d
planes3d

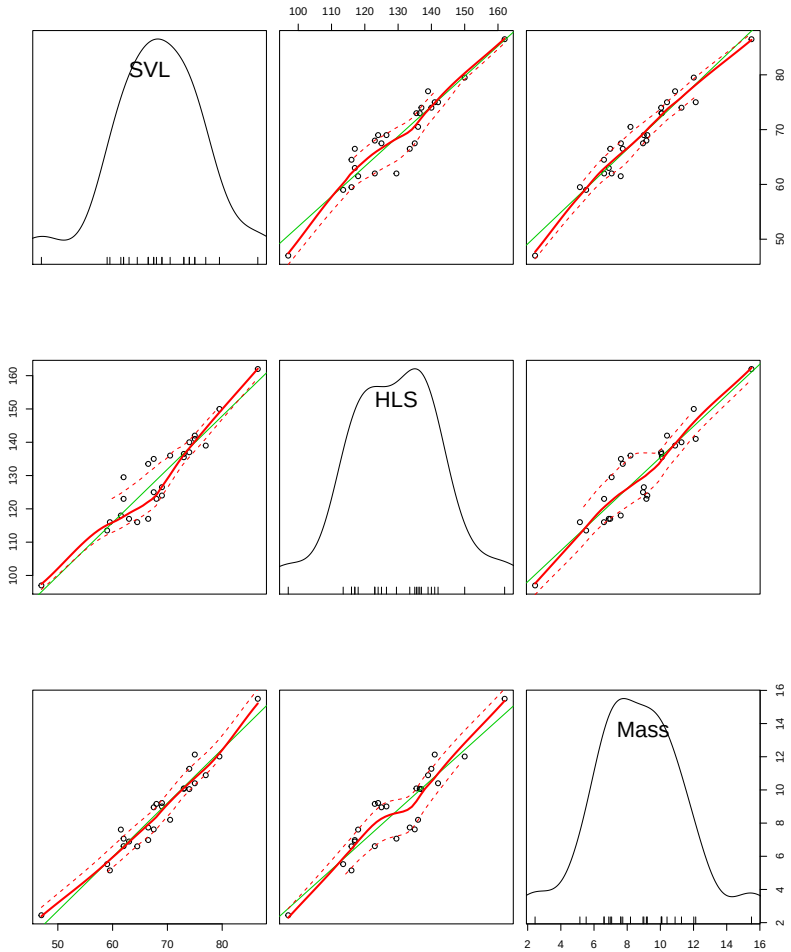
Note that there are many other **R** libraries to produce plots, this is just my favourite!



Exploratory plots for multivariate data

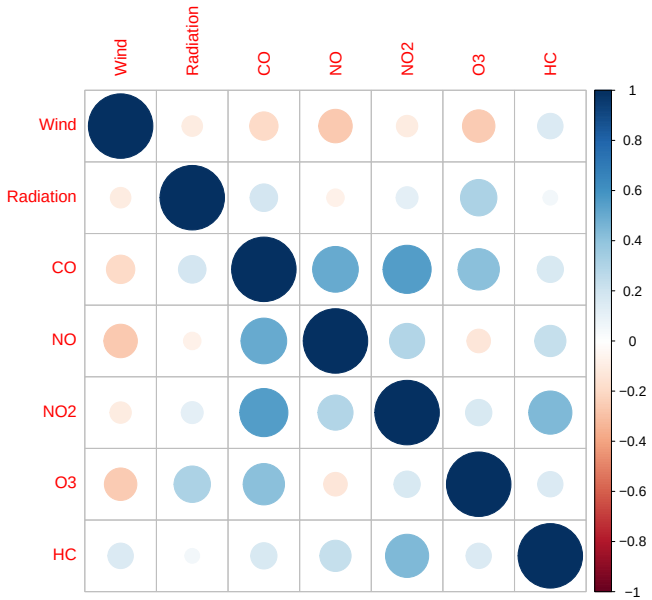
See `Lecture1_2.R` for scatterplot examples.

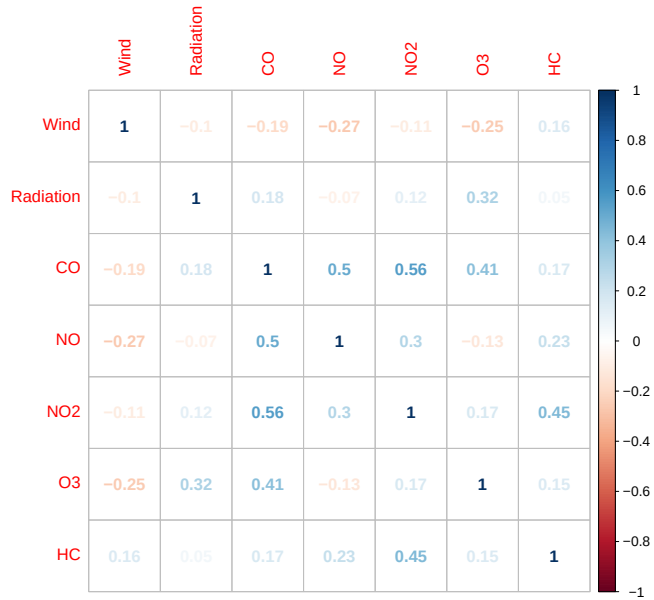




Exploratory plots for multivariate data

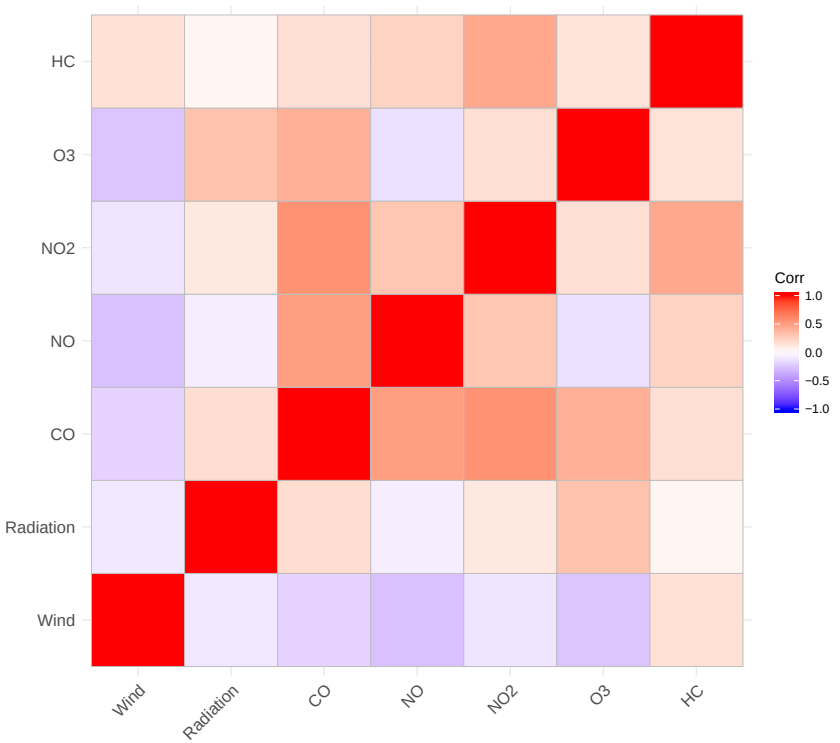
See `Lecture1_2.R` for covariance matrix heatmap examples.

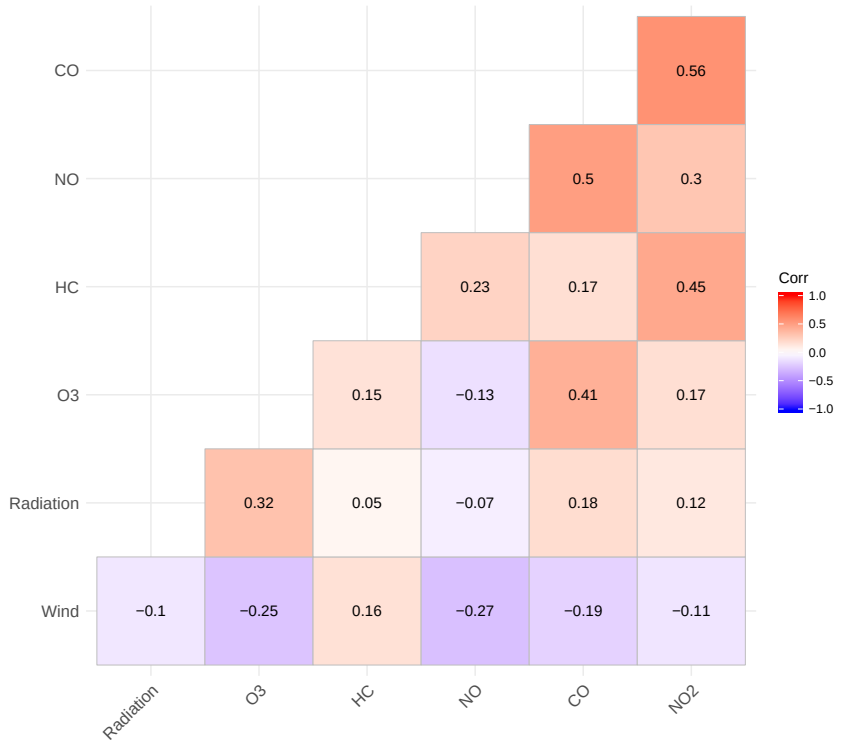




Exploratory plots for multivariate data

See `Lecture1_2.R` for covariance matrix heatmap examples.





Homework exercise 3

There are many different R functions for data visualization. Two R libraries that are particularly useful for multivariate data are the `lattice` library and the `ggplot2` library.

A good book on the use of lattice is Seepayan Sarkar, "Lattice. Multivariate Visualization with R", Springer, 2008. There are good blogs on the web that give you the ggplot2 functions that are equivalent to the lattice functions, for example, https://learnr.wordpress.com/2009/06/28/ggplot2-version-of-figures-in-lattice-multivariate-data-visualization-with-r-part-1/.

Use R to visualize the data on National Track records for Men in a creative manner. Prepare a presentation of your visualization that shows both the graphs and the R-code that generated the graphs, as well as an interpretation of what the graphs are telling you about the variables and the associations in the data.

Looking forward: Dimension reduction

So far we have concentrated on pairwise associations of the many variables or on at most 3-dimensional displays. But how do we visualize multivariate data with many more variables, or equivalently, much higher dimensions?

This is going to be a topic of much discussion later in this course.