1. Introduction to Multivariate Analysis

Isabel M. Rodrigues

1.1 Overview of multivariate methods and main objectives.

WHY MULTIVARIATE ANALYSIS?

- Multivariate statistical analysis is concerned with analysing and understanding data in high dimensions.
- Consists of a collection of methods that can be used when several measurements are made on each individual or object in one or more samples.
- We will refer to the measurements as variables and to the individuals or objects as units (research units, sampling units, or experimental units) or observations.

1.1 Overview of multivariate methods and main objectives.

- Ordinarily the variables are measured simultaneously on each sampling unit.

Examples of Multivariate Data

Units	Variables
Students	Several exam scores in a single course
Students	Grades in mathematics, history, music, art, physics
People	Height, weight, percentage of body fat
Birds	Lengths of various bones

- Typically, these variables are correlated. We need to untangle the overlapping information provided by correlated variables to see the underlying structure. We seek to express what is going on in terms of a reduced set of dimensions. Thus the goal of many multivariate approaches is simplification.

1.1 Overview of multivariate methods and main objectives.

Data types:

1. A single sample with several variables measured on each sampling unit (subject or object);

2. A single sample with two sets of variables measured on each unit;

3. Two samples with several variables measured on each unit;

4. Three or more samples with several variables measured on each unit.

Types of Measurement:

1. **Nominal**: Categorical variables with no meaningful order (Examples: Gender, Hair color);
2. **Ordinal**: Categorical variables where a meaningful order exists (Examples: Social class, Rating of instructor);
   - Nominal and Ordinal are Non-Metric data $\Rightarrow$ this data can not be manipulated mathematically.
3. **Interval**: Numerical variables where taking differences is meaningful, but there is no fixed zero position (Examples: Temperature using Celsius/Fahrenheit);
4. **Ratio**: Numerical variables where taking ratios is meaningful since there is a fixed zero (Examples: Age, Height, Weight).
   - Interval and Ratio are Metric data $\Rightarrow$ this data can be manipulated mathematically.

1.1 Overview of multivariate methods and main objectives.

Types of Multivariate Techniques:

1. **Dependence techniques:** a variable or set of variables is identified as the dependent variable to be predicted or explained by other variables known as independent variables.

2. **Interdependence techniques:** involve the simultaneous analysis of all variables in the set, without distinction between dependent variables and independent variables.

Multivariate methods in this course

1. Dependence methods

   Investigation of the relation among variables
   - Regression (Cap. 2)
   - Analysis of variance - ANOVA (Cap. 3)

2. Interdependence methods

   Data reduction and simplification
   - Principal components analysis (Cap. 4)

3. Sorting and grouping
   - Cluster analysis (Cap. 5)

1.2 Some definitions and notation.

Data Organization

p: number of numeric response variables (or characters) of being measured;
n: number of items (individuals, or experiment units) on which variables are being measured;
x_{jk}: measurement of the kth variable on the jth item.

Data Matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2k} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3k} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

$\mathbf{X}$: n rows and p columns.

1.2 Some definitions and notation.

Data Vector

$$\mathbf{x}_j^{\mathrm{T}} = (x_{j1}, x_{j2}, \ldots, x_{jp}) \Rightarrow \mathbf{x} = \begin{bmatrix} x_{j1} \\ x_{j2} \\ x_{j3} \\ \vdots \\ x_{jk} \\ \vdots \\ x_{jp} \end{bmatrix} \Rightarrow \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \mathbf{x}_3^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_j^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{bmatrix}$$

All arrays (matrices and vectors) will be symbolized by boldfaced font.

Descriptive Statistics Review - Sample quantities

- Sample Mean
- Sample Variance
- Sample Covariance
- Sample Correlation

1.2 Some definitions and notation.

Sample Mean: For the kth variable, the sample mean is:

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}$$

$$\text{sample mean vector - } \bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

Sample Variance: For the kth variable, the Sample Variance is:

$$s_k^2 = s_{kk} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2 = \frac{1}{n-1}\left(\sum_{j=1}^{n} x_{jk}^2 - n\bar{x}_k^2\right)$$

1.2 Some definitions and notation.

For a pair of variables, i and k, the Sample Covariance is:

$$s_{ik} = s_{ki} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) = \frac{1}{n-1}\left(\sum_{j=1}^{n} x_{ji}x_{jk} - n\bar{x}_i\bar{x}_k\right)$$

Sample Covariance matrix - symmetric matrix

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

The Variable Space

- Each row of $\mathbf{X}$ is a point in "$p-$space" or variable space.
- The variables (columns of $\mathbf{X}$) define the axes.
- Consider the n points in the p dimensional space.

The center of the point "cloud" is $\bar{\mathbf{x}}$.

The variability and covariability is measured by $\mathbf{S}$.

1.2 Some definitions and notation.

- Sample covariances are dependent upon the scale of the variables under study.
- For this reason, the Sample Correlation is often used to describe the linear association between two variables.

1.2 Some definitions and notation.

- For a pair of variables, i and k, the Sample Correlation is obtain by dividing the sample covariance by the product of the standard deviation of the variables:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$$

$r_{ik} \in [-1, 1]$ and $r_{ii} = 1$

Sample Correlation matrix - symmetric matrix

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

Measures of multivariate scatter:

- With just one variable, we usually only need just one statistics (the sample variance) to describe the variability in the data.
- With p variable we need p variances and $p(p-1)/2$ covariances.
- It might be nice to have a single statistic that summarizes the information in $\mathbf{S}$. This statistic should reflect all the variances and covariance.

Generalized Sample Variance

- Generalized Sample Variance (GSV) is computed by $|\mathbf{S}| =$ product of eigenvalues of $\mathbf{S}$.

1.2 Some definitions and notation.

- The equation $(\mathbf{x} - \bar{\mathbf{x}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) = c^2$ describes an equation where all points are equal distant from the mean $\bar{\mathbf{x}}$ (i.e., it will form an hyper-ellipsoid).

- The GSV is proportional to the volume of this ellipsoid. The ellipsoid represents statistical distances (Sample Mahananobis distance) of observations from the vector of means.

Sample Statistical Distance

- To obtain a useful distance measure in a multivariate setting, we must consider not only the sample variances of the variables but also their sample covariances or sample correlations.
- For a statistical distance, we standardize by inserting the inverse of the sample covariance matrix:

$$d^2 = (\mathbf{x}_1 - \mathbf{x}_2)^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x}_1 - \mathbf{x}_2)$$

or a distance to the sample mean vector;

$$D^2 = (\mathbf{x} - \bar{\mathbf{x}})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}})$$

- These (squared) distances between two vectors were first proposed by Mahalanobis (1936) and are often referred to as Mahalanobis distances.

Total Sample Variance

- Another way to characterize the sample variance is with the Total Sample Variance (TSV).
- $TSV = \sum_{i=1}^{p} s_{ii} = trace(\mathbf{S}) =$ sum of eigenvalues of $\mathbf{S}$.
- Describes the variability of the data without taking in to account the covariances.

Some useful results of matrix theory

- **Norm of a Vector:** $\|\mathbf{x}\| = \sqrt{\mathbf{x}^{\mathrm{T}}\mathbf{x}}$
- **Trace of a Matrix:** If $\mathbf{A}$ is a squared matrix $p \times p$ then $\mathrm{tr}(\mathbf{A}) = \sum_{\mathrm{i}=1}^{\mathrm{p}} \mathrm{a}_{\mathrm{ii}}$.
- **Properties:**
  1. $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$.
  2. $\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$.
  3. **Determinant:** Let $\mathbf{A}$ be a 2×2 matrix, then
     $$|\mathbf{A}| = \det(\mathbf{A}) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$
     Let $\mathbf{A}$ be a 3×3 matrix, then
     $$|\mathbf{A}| = \det(\mathbf{A}) = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} +$$
     $$a_{12}a_{23}a_{31} - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}.$$

Some useful results of matrix theory

- **Determinant Properties:** Let $\mathbf{A}$ be a $n \times n$ matrix.
  1. $|\mathbf{A}^{\mathrm{T}}| = |\mathbf{A}|$.
  2. $|k\mathbf{A}| = k^n|\mathbf{A}|$, $k \in \mathbb{R}$.
  3. $|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}$.
  4. $|\mathbf{AB}| = |\mathbf{BA}|$.
  5. $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$.
- **Singular matrix:** The matrix $\mathbf{A}$ is singular if $|\mathbf{A}| = 0$. The matrix $\mathbf{A}$ is non-singular if $|\mathbf{A}| \neq 0$.
- **Inverse of a matrix:** The inverse of a squared matrix, $\mathbf{A}$, is the matrix $\mathbf{A}^{-1}$ which satisfies: $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. The inverse, $\mathbf{A}^{-1}$, only exists if $\mathbf{A}$ is non-singular.
  Inverse Properties:
  1. $(k\mathbf{A})^{-1} = k^{-1}\mathbf{A}^{-1}$, $k \in \mathbb{R}$.
  2. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

Some useful results of matrix theory

- **Rank of a matrix:** $rank(\mathbf{A})$ is the maximum number of linearly independent rows or columns of the matrix. If $\mathbf{A}$ is a $p \times q$ matrix then
$$rank(\mathbf{A}) \leq \min(p, q).$$

- **Orthogonal vectors:** Two vectors are orthogonal if and only if $\mathbf{x}^{\mathrm{T}}\mathbf{y} = 0$ (which implies that the angle between the two vectors is 90^o).

- **Orthogonal matrix:** A squared matrix is orthogonal if and only if $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{A}\mathbf{A}^{\mathrm{T}} = \mathbf{I} \Rightarrow \mathbf{A}^{-1} = \mathbf{A}^{\mathrm{T}}$.

Some useful results of matrix theory

- **Eigenvalues and eigenvectors:** Eigenvalues and eigenvectors are defined for squared matrices, and play an important key role in multivariate data analysis. We introduce them by considering the following equation (where $\mathbf{A}$ is a known matrix):

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \tag{1}$$

The scaler quantities, λ_j, that solve this equation are called eigenvalues, and the corresponding vectors, $\mathbf{x}_j$, are the eigenvectors. Equation (1) can be rewrite as:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}. \tag{2}$$

Equation (2) have a nontrivial solution (i.e. $\mathbf{x} \neq \mathbf{0}$) if $|\mathbf{A}-\lambda\mathbf{I}| = 0$, and the solution of this equation gives us the eigenvalues of $\mathbf{A}$.

1.3 Some definitions and notation.

- We will organized the eigenvalues such that:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p.$$

Properties:

1. $|\mathbf{A}| = \prod_{i=1}^{p} \lambda_i$.
2. $\mathrm{tr}(\mathbf{A}) = \sum_{\mathrm{i}=1}^{\mathrm{p}} \lambda_{\mathrm{i}}$.

- **Spectral decomposition:** A symmetric $p \times p$ matrix, $\mathbf{A}$, can be written in the following form:

$$\mathbf{A} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^{\mathrm{T}} = \sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i^{\mathrm{T}},$$

where $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues of $\mathbf{A}$, i.e. $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_{\mathrm{p}}\}$ and $\mathbf{\Gamma}$ is a matrix formed by the eigenvectors, γ_i, of $\mathbf{A}$, i.e. $\mathbf{\Gamma} = [\gamma_1, \ldots, \gamma_p]$.

1.2 Some definitions and notation.

Note that:

- $\mathbf{\Gamma}$ is an orthogonal matrix i.e. $\mathbf{\Gamma\Gamma}^{\mathrm{T}} = \mathbf{\Gamma}^{\mathrm{T}}\mathbf{\Gamma} = \mathbf{I}$, i.e. the eigenvectors are chosen to be mutually orthogonal (perpendicular) and with to have length$= 1$.
- If the eigenvalues of $\mathbf{A}$ are λ_i the eigenvalues of $\mathbf{A}^{-1}$ are λ_i^{-1}.
- The eigenvectors of $\mathbf{A}^{-1}$ are the same as the eigenvectors of $\mathbf{A}$.
- $\mathbf{A}^{-1} = \mathbf{\Gamma\Lambda}^{-1}\mathbf{\Gamma}^{\mathrm{T}}$, where $\mathbf{\Lambda}^{-1} = \mathrm{diag}\{\lambda_1^{-1}, \ldots, \lambda_{\mathrm{p}}^{-1}\}$, since:

$$\mathbf{AA}^{-1} = (\mathbf{\Gamma\Lambda\Gamma}^{\mathrm{T}})(\mathbf{\Gamma\Lambda}^{-1}\mathbf{\Gamma}^{\mathrm{T}}) = (\mathbf{\Gamma\Lambda\Lambda}^{-1}\mathbf{\Gamma}^{\mathrm{T}}) = \mathbf{\Gamma\Gamma}^{\mathrm{T}} = \mathbf{I}$$

.

- $\mathbf{A}^{1/2} = \mathbf{\Gamma\Lambda}^{1/2}\mathbf{\Gamma}^{\mathrm{T}}$, where $\mathbf{\Lambda}^{1/2} = \mathrm{diag}\{\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_{\mathrm{p}}}\}$.
- $\mathbf{A}^{-1/2} = \mathbf{\Gamma\Lambda}^{-1/2}\mathbf{\Gamma}^{\mathrm{T}}$, where $\mathbf{\Lambda}^{-1/2} = \mathrm{diag}\{\sqrt{\lambda_1^{-1}}, \ldots, \sqrt{\lambda_{\mathrm{p}}^{-1}}\}$.

Mean Vector, Covariance and Correlation Matrices of a Random Vector $\mathbf{X}^{\mathrm{T}} = (X_1, \ldots, X_p)$ (Population Parameters)

- **Mean Vector:** $E(\mathbf{X}) = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$

- **Covariance Matrix:** $\mathrm{COV}(\mathbf{X}) = \boldsymbol{\Sigma} = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}}$

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

- with $\mathrm{VAR}(X_i) = \sigma_i^2 = \sigma_{ii}$, for $i = 1, \ldots, p$.

1.2 Some definitions and notation.

- $\boldsymbol{\Sigma}$ is a symmetric positive definite (s.p.d.) matrix.
- **Correlation Matrix (Covariance of standardized variables):**

$$\mathrm{COR}\,(\mathbf{X}) = \boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}, \text{ is a s.p.d. matrix.}$$

- Considering $\mathbf{V}^{1/2} = \mathrm{diag}\{\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{\mathrm{pp}}}\}$, we have that:

$$\boldsymbol{\Sigma} = \mathbf{V}^{1/2}\boldsymbol{\rho}\mathbf{V}^{1/2} \quad \text{and} \quad \boldsymbol{\rho} = (\mathbf{V}^{1/2})^{-1}\boldsymbol{\Sigma}(\mathbf{V}^{1/2})^{-1}.$$

- $|\boldsymbol{\Sigma}| = (\sigma_{11} \times \sigma_{22} \ldots \sigma_{pp})|\boldsymbol{\rho}|$

Generalized Variance

- $GV = |\mathbf{\Sigma}|$ = product of eigenvalues of $\mathbf{\Sigma}$ or
  $GV = |\boldsymbol{\rho}|$ = product of eigenvalues of $\boldsymbol{\rho}$ (considering standard variables).

Total Variance

- $TV = \sum_{i=1}^{p} \sigma_{ii} = trace(\mathbf{\Sigma})$ = sum of eigenvalues of $\mathbf{\Sigma}$ or
  $TV = p$ = sum of eigenvalues of $\boldsymbol{\rho}$ (considering standard variables).

Mahalanobis Distance

- $\Delta^2 = (\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})$

Some properties of the covariance:

- Let $\mathbf{X}$ be a random vector, with dimension p and covariance matrix $\mathbf{\Sigma}$.

  1. $\text{VAR}(\mathbf{a}^{\text{T}}\mathbf{X}) = \mathbf{a}^{\text{T}}\mathbf{\Sigma}\mathbf{a}$, $\mathbf{a} \in \mathbb{R}^p$.
  2. $\text{COV}(\mathbf{a}_1^{\text{T}}\mathbf{X}, \mathbf{a}_2^{\text{T}}\mathbf{X}) = \mathbf{a}_1^{\text{T}}\mathbf{\Sigma}\mathbf{a}_2$, $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^p$.
  3. $\text{COV}(\mathbf{AX}, \mathbf{BX}) = \mathbf{A\Sigma B}^{\text{T}}$, where $\mathbf{A}$ and $\mathbf{B}$ are $q \times p$ matrices.

Using the R software: descriptive methods

- Edgar Anderson's Iris Data: This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris, setosa, versicolor, and virginica. iris3 gives the same data arranged as a 3-dimensional array of size 50 by 4 by 3. The first dimension gives the case number within the species subsample, the second the measurements with names Sepal L., Sepal W., Petal L., and Petal W., and the third the species. **iris3[,,1]** is the R command for the specie setosa.

1.3 Exploratory analysis: descriptive methods and graphical multivariate data display.

Data setosa

```
setosa=iris3[,,1]
dim(setosa)
```

```
50 4
```

summary statistics

```
summary(setosa)
```

```
       Sepal L.       Sepal W.       Petal L.       Petal W.
Min.   :4.300   Min.   :2.300   Min.   :1.000   Min.   :0.100
1st Qu.:4.800   1st Qu.:3.200   1st Qu.:1.400   1st Qu.:0.200
Median :5.000   Median :3.400   Median :1.500   Median :0.200
Mean   :5.006   Mean   :3.428   Mean   :1.462   Mean   :0.246
3rd Qu.:5.200   3rd Qu.:3.675   3rd Qu.:1.575   3rd Qu.:0.300
Max.   :5.800   Max.   :4.400   Max.   :1.900   Max.   :0.600
```

1.3 Exploratory analysis: descriptive methods and graphical multivariate data display.

Sample mean vector

```
apply(setosa,2,mean)
```

```
Sepal L. Sepal W. Petal L. Petal W.
5.006      3.428     1.462      0.246
```

Sample covariance matrix

```
round (cov(setosa),digits=5)
```

```
             Sepal L. Sepal W. Petal L. Petal W.
Sepal L.  0.12425  0.09922  0.01636  0.01033
Sepal W.  0.09922  0.14369  0.01170  0.00930
Petal L.  0.01636  0.01170  0.03016  0.00607
Petal W.  0.01033  0.00930  0.00607  0.01111
```

1.3 Exploratory analysis: descriptive methods and graphical multivariate data display.

Sample correlation matrix

```
round (cor(setosa),digits=5)
```

```
          Sepal L. Sepal W. Petal L. Petal W.
Sepal L.  1.00000  0.74255  0.26718  0.27810
Sepal W.  0.74255  1.00000  0.17770  0.23275
Petal L.  0.26718  0.17770  1.00000  0.33163
Petal W.  0.27810  0.23275  0.33163  1.00000
```

Alternative: Covariance of the standard variables

```
round(cov(scale(setosa)), digits=5)
```

```
          Sepal L. Sepal W. Petal L. Petal W.
Sepal L.  1.00000  0.74255  0.26718  0.27810
Sepal W.  0.74255  1.00000  0.17770  0.23275
Petal L.  0.26718  0.17770  1.00000  0.33163
Petal W.  0.27810  0.23275  0.33163  1.00000
```

eigenvalues and eigenvector of S

```
eigen(cov(setosa))
$values

0.236455690 0.036918732 0.026796399 0.009033261

$vectors


-0.66907840 0.5978840 0.4399628 -0.03607712
-0.73414783 -0.6206734 -0.2746075 -0.01955027
-0.09654390 0.4900556 -0.8324495 -0.23990129
-0.06356359 0.1309379 -0.1950675 0.96992969

Svalues=round(eigen(cov(setosa))$values, digits=4)
Svalues
0.2365 0.0369 0.0268 0.0090
```

1.3 Exploratory analysis: descriptive methods and graphical multivariate data display.

eigenvalues and eigenvector of R

```
eigen(cor(setosa))
$values

2.0585402 1.0221782 0.6678202 0.2514613

$vectors


-0.6044164 0.3349908 -0.0673598261 0.71966982
-0.5756194 0.4408461 -0.0007138239 -0.68870645
-0.3754348 -0.6269717 -0.6770628102 -0.08683986
-0.4029788 -0.5480350 0.7328356536 -0.01475204

Rvalues=round(eigen(cor(setosa))$values, digits=4)
Svalues
2.0585 1.0222 0.6678 0.2515
```

1.3 Exploratory analysis: descriptive methods and graphical multivariate data display.

Generalized and Total sample variances

```
prod(Svalues)
```

2.104916e-06

```
prod(Rvalues)
```

0.3534037

```
sum(Svalues)
```

0.3092

```
sum(Rvalues)
```

4

$$|\mathbf{S}| = (s_{11} \times s_{22} \times \ldots s_{pp})|\mathbf{R}|$$

```
diagS=round(diag(cov(setosa)), digits=4)
diagS
```

Sepal L. Sepal W. Petal L. Petal W.

0.1242 0.1437 0.0302 0.0111

```
prod(diagS)*prod(Rvalues)
```

2.114362e-06

1.3 Exploratory analysis: descriptive methods and graphical multivariate data display.

Mahalanobis sample distances

```
meanSet=apply(setosa,2,mean)
round(mahalanobis(setosa,meanSet,cov(setosa)), digits=3)
```

0.449 2.081 1.284 1.706 0.762 3.713 3.424 0.343 2.996 3.200

1.891 2.015 2.947 7.040 10.222 7.654 5.742 0.636 5.186 1.612

5.349 2.722 11.044 7.230 9.748 3.771 2.526 0.829 1.323 2.174

1.995 4.889 7.699 5.248 1.267 3.302 5.721 3.086 3.270 0.589

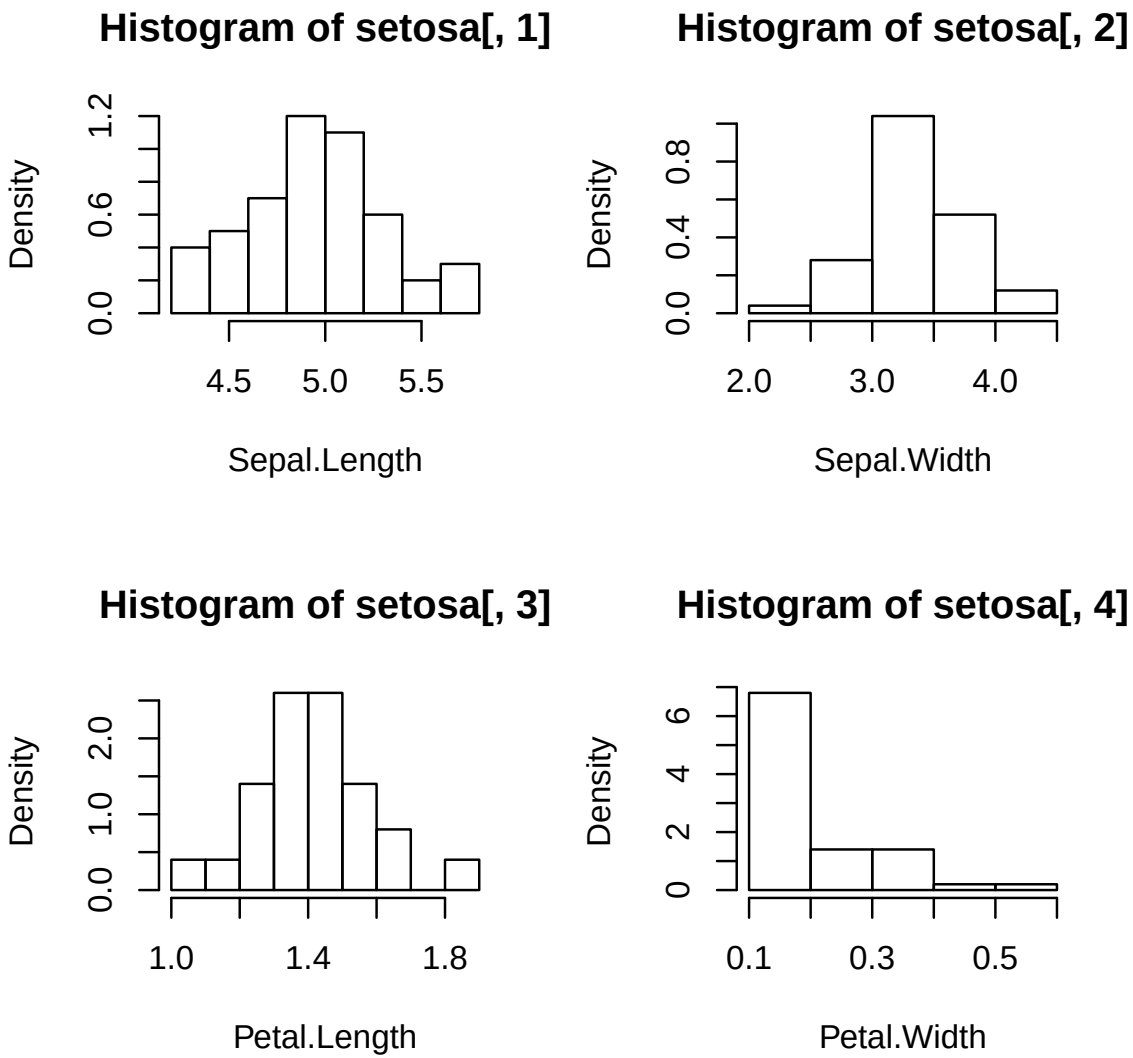
1.685 12.328 4.201 12.310 8.601 2.195 2.756 1.489 1.253 0.495

Using the R software: graphical multivariate data display

Histograms

```
par(mfrow=c(2,2))
hist(setosa[,1],prob=TRUE,xlab="Sepal.Length")
hist(setosa[,2],prob=TRUE,xlab="Sepal.Width")
hist(setosa[,3],prob=TRUE,xlab="Petal.Length")
hist(setosa[,4],prob=TRUE,xlab="Petal.Width")
```

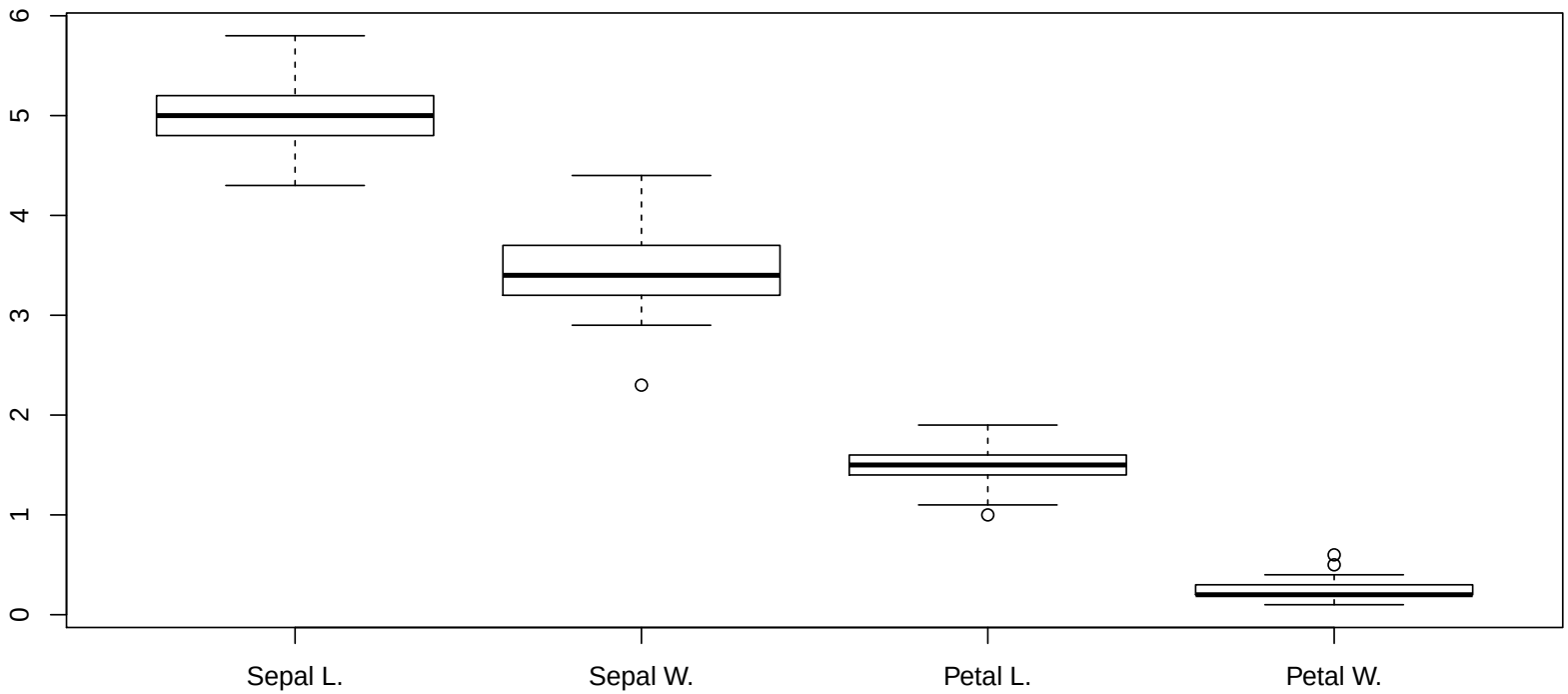
1.3 Exploratory analysis: descriptive methods and graphical multivariate data display.



1.3 Exploratory analysis: descriptive methods and graphical multivariate data display.

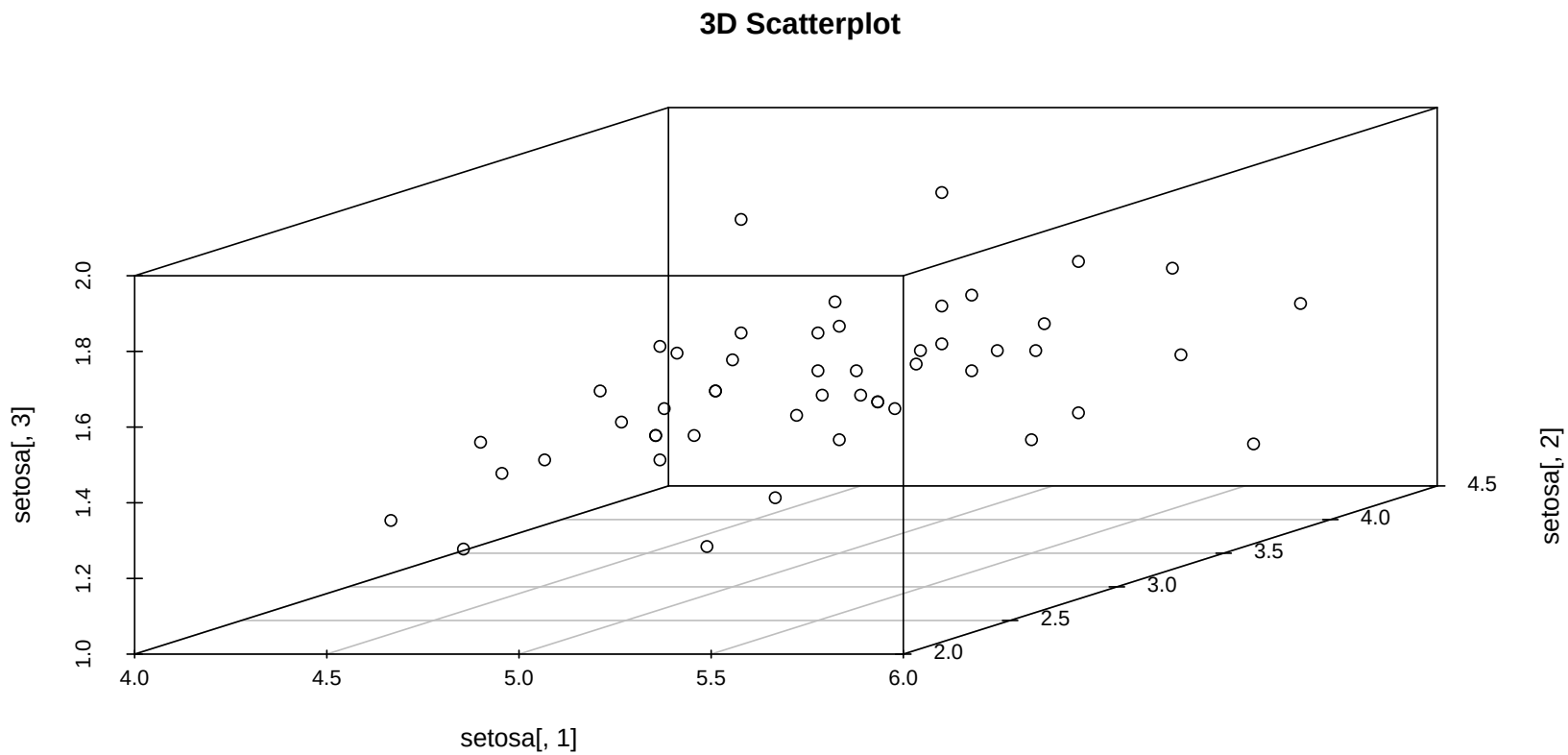
Boxplot

```
boxplot(setosa,prob=TRUE,xlab="")
```



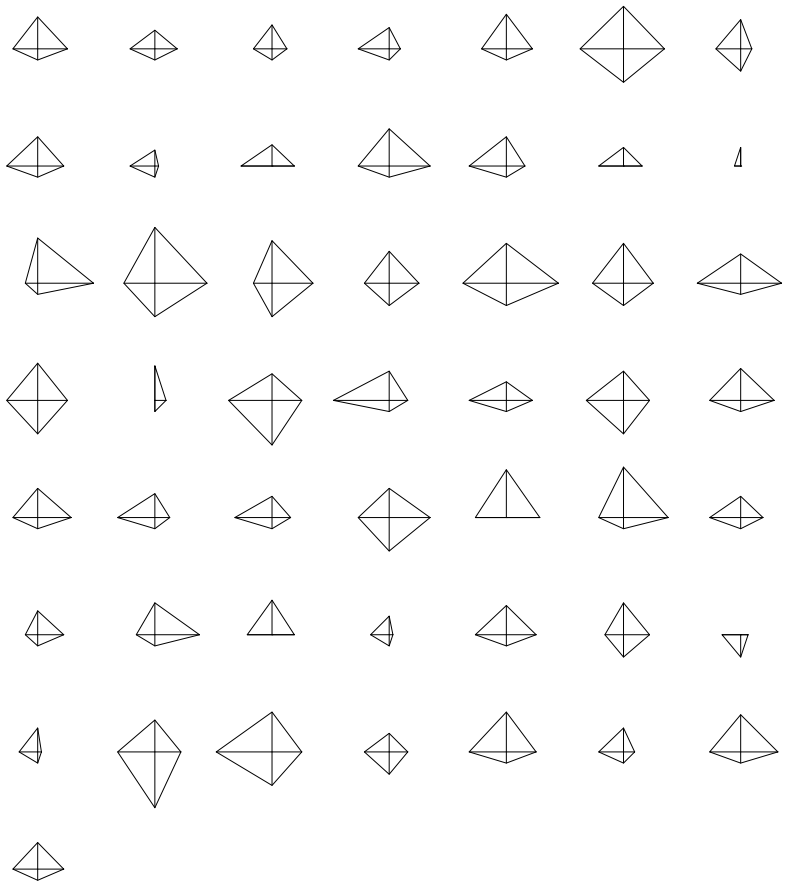
3D Scatterplot

```
library(scatterplot3d)
scatterplot3d(setosa[,1],setosa[,2],setosa[,3],main="3D Scatterplot")
```



1.3 Exploratory analysis: descriptive methods and graphical multivariate data display.

```
library(graphics)
stars(setosa)
```



1.3 Exploratory analysis: descriptive methods and graphical multivariate data display.

Chernoff Faces

```
library(aplpack)
faces(setosa)
```