Modelling complex data using R

Dr Eleni Matechou

Schedule for the day

- 9:30-10.30: Principal Components Analysis
- 10:30-11.00: Principal Components Analysis - Practical
- 11:00-11.30: Break
- 11:30-12.30: Factor Analysis
- 12:30-13.00: Factor Analysis - Practical
- 13:00-14.00: Break
- 14:00-15.00: Discriminant Analysis
- 15:00-15.30: Discriminant Analysis - Practical

Outline

Complex data

For this course, the complexity of the data results from the fact that we have a potentially large number of measurements for each row (individual or subject or observation).

These data are also referred to as **multivariate data** and typically include a large number of columns (variables).

Visualizing, **interpreting** or **finding patterns** in multivariate data is challenging.

We will learn how to **reduce the dimensions** of the data using Principal Components Analysis, to **estimate the relationship between observable and unobservable variables** using Factor Analysis and to **obtain rules to classify observations** using Discriminant Analysis.

Principle Component Analysis (PCA)

Principal Component Analysis (PCA) attempts to summarize the information in the data in a **new, smaller set of variables** whilst retaining **most** of the information in the original set.

PCA essentially aims to **reduce the dimensionality** of a multivariate data set, while accounting for as much of the variation as possible from the original data.

This aim is achieved by creating a new set of **uncorrelated** variables, called the **principal components** (PC).

PC are **linear combinations** of the original variables and are **ordered** so that the first **few**, eg. 2 or 3, of them account for most of the variation in **all** of the original variables.

This new set of variables can provide the basis for further analyses as, for example, in regression modelling with many and/or highly correlated explanatory variables. In this case, the first few PC can be used as covariates.

The original (potentially correlated) variables are $\mathbf{X} = (\underline{x}_1, \ldots, \underline{x}_p)^T$.

The first PC, $\underline{z}_1$, is the linear combination

$$\underline{z}_1 = a_{11}\underline{x}_1 + a_{12}\underline{x}_2 + \ldots + a_{1p}\underline{x}_p$$

and it has **variance that is greatest among all such linear combinations**.

The second PC, $\underline{z}_2$, is the linear combination

$$\underline{z}_2 = a_{21}\underline{x}_1 + a_{22}\underline{x}_2 + \ldots + a_{2p}\underline{x}_p$$

and it accounts for a maximal proportion of the **remaining** variance, **subject to being uncorrelated with the first PC,** $\underline{z}_1$.

etc until the pth PC.

Generally, the jth PC, $\underline{z}_j$, is constructed as

$$\underline{z}_j = \sum_{k=1}^{p} a_{jk}\underline{x}_k,$$

with

$$\sum_{k=1}^{p} a_{jk}^2 = 1$$

Construction of PC

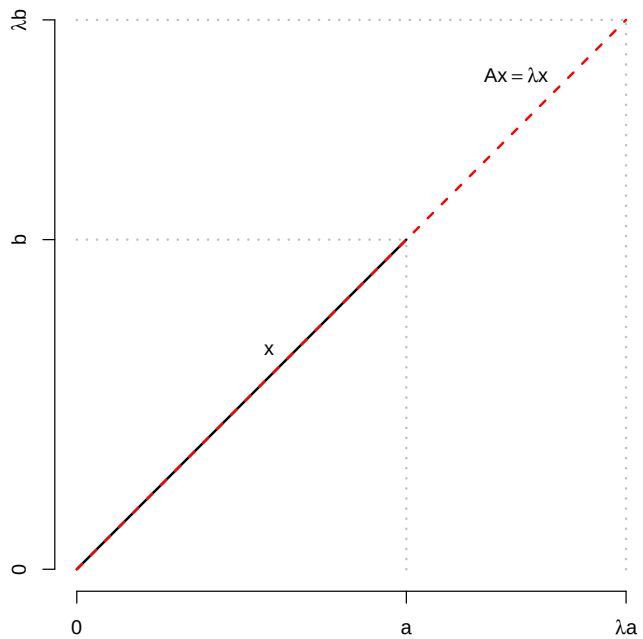
It can be shown that $\underline{a}_1 = (a_{11}, a_{12}, \ldots, a_{1p})$ is the **eigenvector** related to the **largest eigenvalue** λ_1 of the **sample covariance matrix**, S_x, or equivalently of the **sample correlation matrix**, R_x.

Similarly, $\underline{a}_2$ is the eigenvector related to the 2nd largest eigenvalue λ_2 etc.

In general, we refer to $\underline{a}_j$ as the **loading** of the j-th PC, which is the eigenvector related to the j-th largest eigenvalue λ_j.

Eigenvalues and eigenvectors

If x is an eigenvector of matrix A with λ the corresponding eigenvalue, then multiplying x by A stretches x by λ, but does not change its direction.



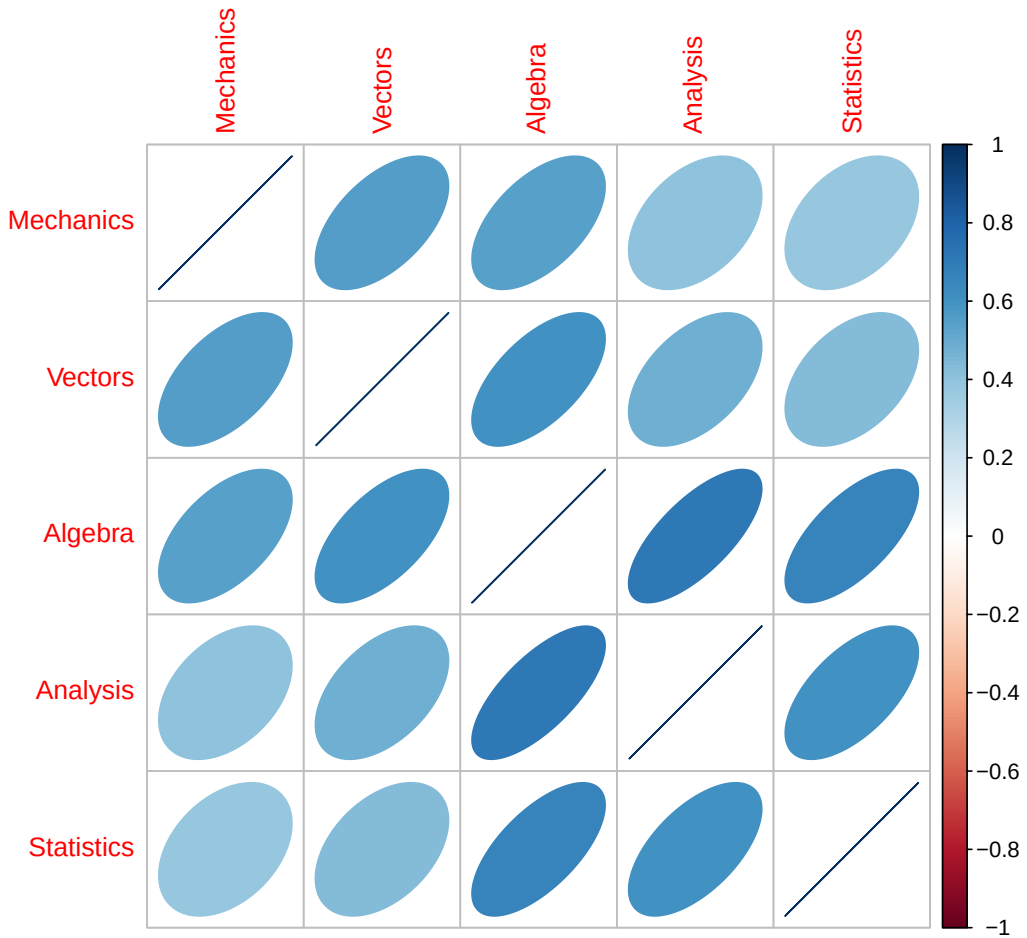
Example: Open/close book exam marks data

Data extract:

Mechanics (C)	Vectors (C)	Algebra (O)	Analysis (O)	Statistics (O)
77	82	67	67	81
63	78	80	70	81
75	73	71	66	81
55	72	63	70	68
63	63	65	70	63
53	61	72	64	73
51	67	65	65	68
59	70	68	62	56
62	60	58	62	70
64	72	60	62	45

Correlation matrix

The exam results are highly correlated with each other, so PCA will work well.



PCA of the exam scores data

e-values	e-vectors					
λ_1 : 686.990	$\underline{a}_1^T$	-0.505	-0.368	-0.346	-0.451	-0.535
λ_2 : 202.111	$\underline{a}_2^T$	-0.749	-0.207	0.076	0.301	0.548
λ_3 : 103.747	$\underline{a}_3^T$	0.300	-0.416	-0.145	-0.597	0.600
λ_4 : 84.630	$\underline{a}_4^T$	-0.296	0.782	0.003	-0.519	0.176
λ_5 : 32.153	$\underline{a}_5^T$	-0.079	-0.189	0.924	-0.285	-0.152

Note: the loadings are **not unique**, multiplying all elements of the eigenvectors by -1 is an alternative solution.

The first PC is

$$-0.505 \times \text{Mechanics} - 0.368 \times \text{Vectors} - 0.346 \times \text{Algebra}$$
$$-0.451 \times \text{Analysis} - 0.535 \times \text{Statistics}$$

This PC has negative loadings on all variables: we could change all signs to + and interpret it as an “average” grade.

The second PC is

$$-0.749 \times \text{Mechanics} - 0.207 \times \text{Vectors} + 0.076 \times \text{Algebra}$$
$$+ 0.301 \times \text{Analysis} + 0.548 \times \text{Statistics}$$

This PC discriminates between "close-book" and "open-book" examinations.

The third PC is

$$0.300 \times \text{Mechanics} - 0.416 \times \text{Vectors} - 0.145 \times \text{Algebra}$$
$$- 0.597 \times \text{Analysis} + 0.600 \times \text{Statistics}$$

This PC is a contrast between "applied maths" (Mechanics and Statistics) and "pure maths" (Vectors, Algebra and Analysis).

Interpretations of PC

These are some guidelines for interpreting PC:

- The interpretation of PC will often require some knowledge of the problem.
- Some PC can be interpreted as a weighted average (where all weights have the same sign).
- Small loadings can often be ignored.
- Some PC can be interpreted as a contrast between two groups of variables (ones with positive loadings and ones with negative loadings).

Properties of PC

- The variance of the j-th PC is λ_j.
- The total variance of the p PC is equal to the total variance of the original variables so that

$$\sum_{j=1}^{p} \lambda_j = s_1^2 + s_2^2 + \ldots s_p^2,$$

  where s_k^2 is the sample variance of $\underline{x}_k$.
- The relative importance of PC $\underline{z}_j$ (proportion of the total variation of $\mathbf{X}$ explained by $\underline{z}_j$) is

$$\frac{\lambda_j}{\lambda_1 + \cdots + \lambda_p}.$$

- The proportion of variation explained by the first k PC is

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p}.$$

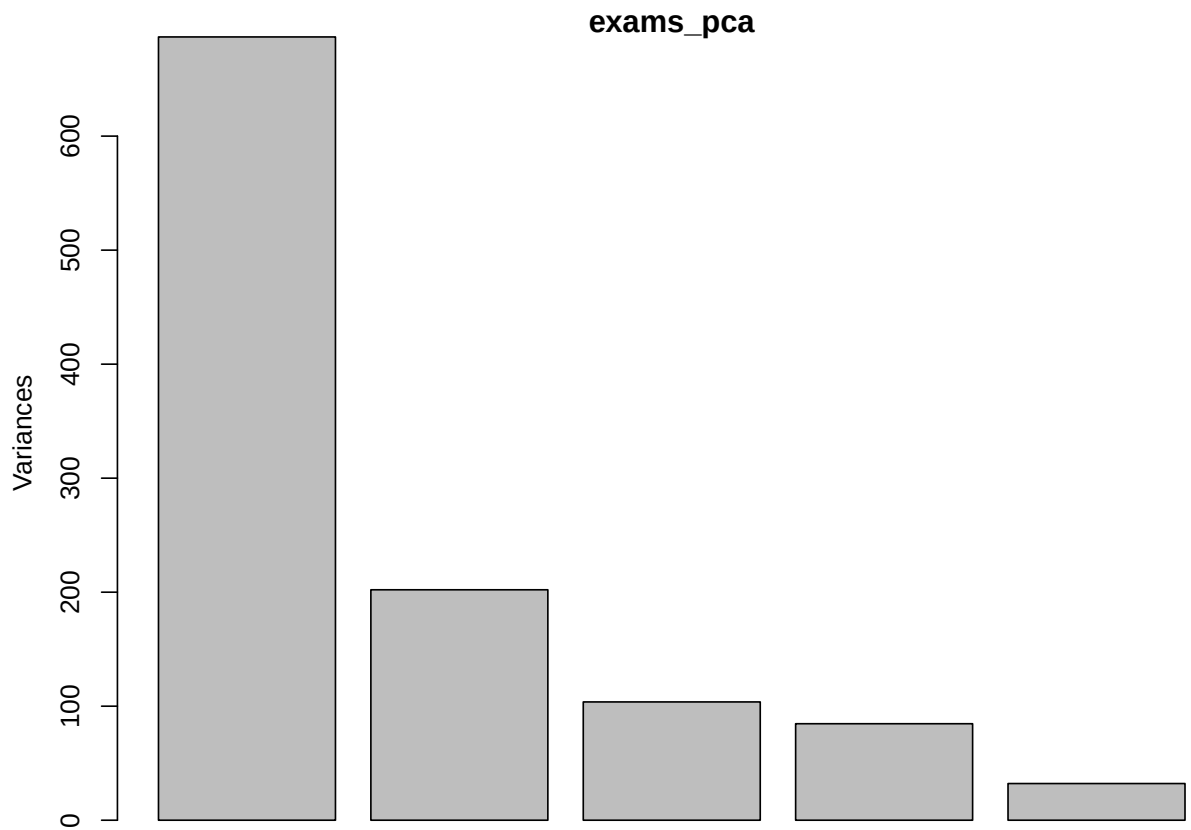
Choosing the number of principal components (dimension reduction)

It is often the case that a lot of the variation in $\mathbf{X}$ is contained in the first few PC. We can therefore ignore the other PC whilst maintaining a good representation of the data.

There are several methods for choosing the number of PC.

- Include just enough PC to explain a fixed proportion of the total variation (eg. 80% or 90%).
- Use scree plots (see next slide) and wait for it to "level off" (elbow method).
- Exclude PC with eigenvalues smaller than the average eigenvalue (note: the average eigenvalue is also the average variance of the original variables). In this case, the average eigenvalue is 219.4, so we would only keep the first PC.

Scree plot for the open/close book exam marks data



Variation explained by the 5 PC:

$$61.91\%, \quad 18.21\%, \quad 9.35\%, \quad 7.63\%, \quad 2.90\%$$

Variation explained by the first two PC:

$$61.91\% + 18.21\% = 80.13\%$$

Variation explained by the first three PC:

$$61.91\% + 18.21\% + 9.35\% = 89.47\%$$

The scores

The scores refer to **individual rows** in the original data (remember the original variables are in the columns of that data set).

If we decide to keep k PC, then the k scores for individual i with $p \times 1$ vector of variable values $\underline{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ are

$$z_{i1} = a_{11}x_{i1} + a_{12}x_{i2} + \ldots + a_{1p}x_{ip}$$

$$z_{i2} = a_{21}x_{i1} + a_{22}x_{i2} + \ldots + a_{2p}x_{ip}$$

$$\vdots$$

$$z_{ik} = a_{k1}x_{i1} + a_{k2}x_{i2} + \ldots + a_{kp}x_{ip}$$

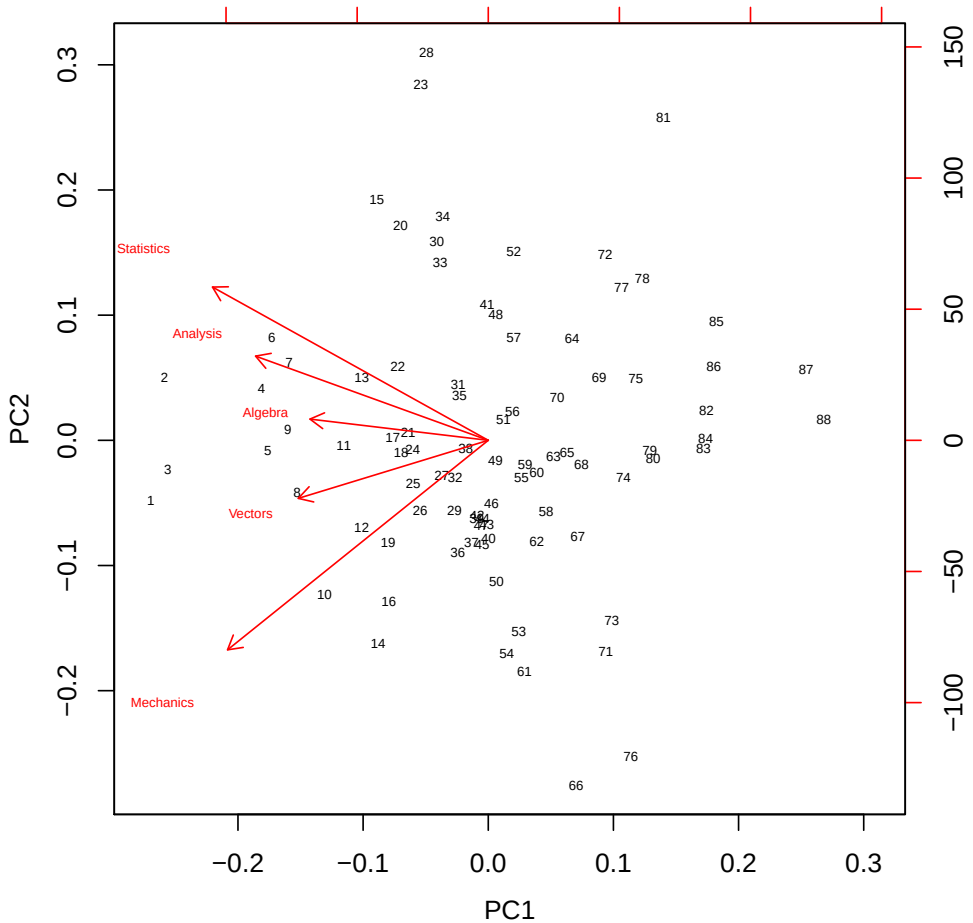
Uses of scores

The scores can be useful for identifying outliers or identifying nonlinear relationships in the data.

They can also sometimes be used instead of the original variables in regression analysis (Principal Component Regression) to avoid problems of “collinearity”. Potential problem of losing information that is important in the regression.

Biplot

A **biplot** displays both the loadings (shown here in red) and the scores (shown here in black) and helps with interpreting the results.

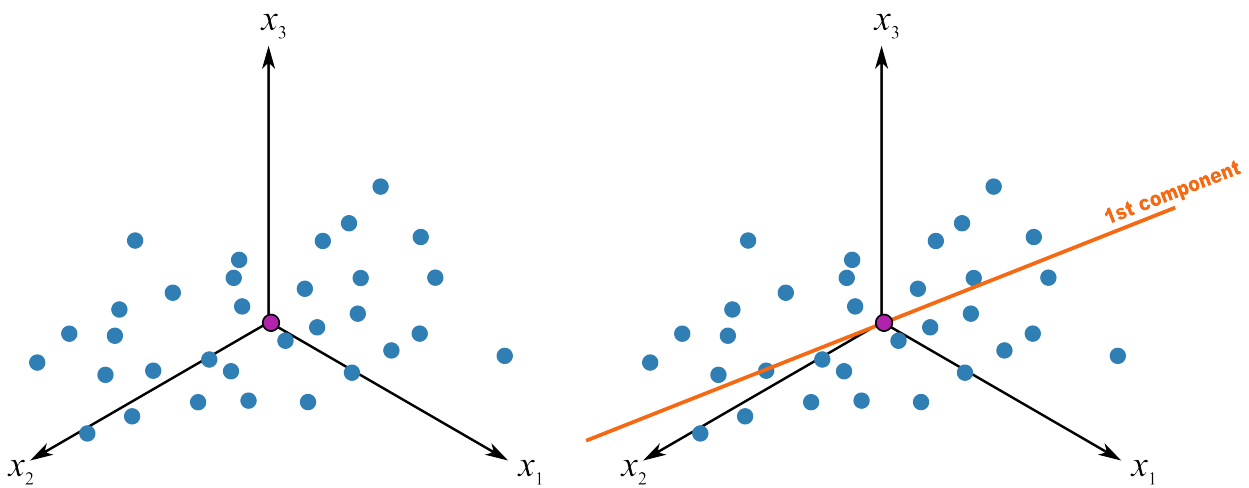


Geometry of PC

Suppose that the original data set includes three variables, with data centred to have mean 0.

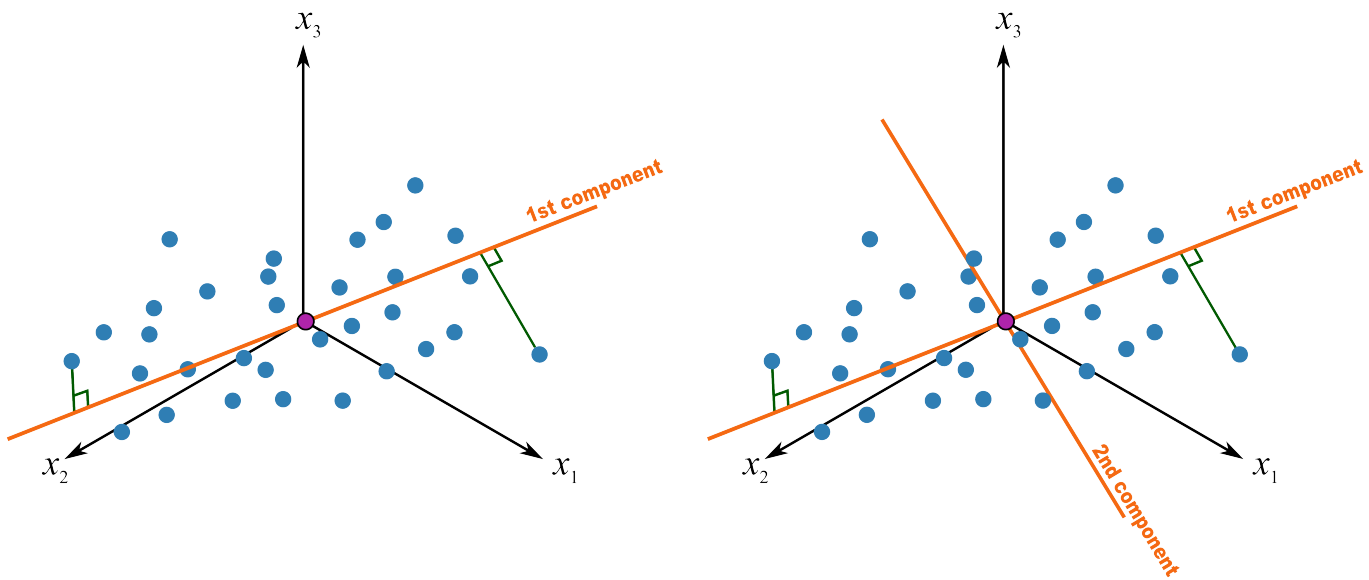
The first PC is the line that goes through the origin and in the direction of maximum variance of the projections of the observations onto the line.

The distance from the origin to the projection of each point along the line is the score for that observation.



Geometry of PC

The second PC is perpendicular to the first, also starts at the origin, and its direction is chosen so that it gives the greatest variance in the score values when projected on this new direction vector.



Geometry of PC

Therefore, the loadings are directions of the new coordinate system while the scores are projections along these directions, that is locations of the observations in the new coordinate system.

Covariance or correlation matrix?

In general, PC are **not invariant** to changes of scale.

For example, suppose that a multivariate data set consists of three variables: weight, measured in pounds, height, measured in feet, and age, measured in years but we want the PC expressed in ounces, inches and decades.[1]

We could

- Multiply the variables by 16, 12 and 1/10, respectively, and then carry out PCA on the covariance matrix of the transformed variables.
- Carry out PCA on the original variables and then multiply the elements of the relevant component by 16, 12 and 1/10.

Generally, the results of these two procedures will be different.

[1] 1 pound = 16 ounces, 1 foot = 12 inches

Implications:

- PC can be sensitive to the relative scales of the data. PC obtained using the covariance matrix will depend on the (arbitrary) units of measurement and if there are large differences between the variances of the original variables then the variables with the largest variances will tend to dominate the first few components.
- Since it is typically unlikely that the original variables are measured on the same scale, in practice PCA is performed on the **correlation matrix**. This is equivalent to calculating the PC on the original variables, after standardising them to have unit variance.
- PC derived from S_x (sample covariance matrix) and those from R_x (sample correlation matrix) are in general **different**.

Factor Analysis

Often, it is not possible to measure the concepts of interest, such as intelligence, social class etc, directly. Concepts such as these that cannot be measured directly are called **latent variables.**

In these cases, researchers collect data on variables that are observable, called **manifest variables**, and are indicators of the latent variables, for example, examination scores, education background, income.

Factor analysis (FA) helps **uncover the relationships** between latent and manifest variables.

Introduction

The purpose of FA is to

- explain the **correlation** structure of the manifest variables in terms of a small number of underlying factors
- provide a linear model to describe the interrelationship of variables
- provide factors that are easier to interpret than PC
- have factors that are invariant to changes of scales
- provide criteria for deciding the number of factors

Rationale

The FA model assumes that the observed relationships between manifest variables (as measured by their covariances or correlations) are a result of the relationships of these variables to the latent variables.

Factor model

For $\underline{X} = (X_1, X_2, \ldots, X_p)^T$, a q-factor model assumes that

$$X_i = \lambda_{i1} Y_1 + \lambda_{i2} Y_2 + \cdots + \lambda_{iq} Y_q + e_i$$

where

- $Y_1, \ldots, Y_q$ are uncorrelated random variables, called **factors**, with $\mathsf{E}(Y_j) = 0$, $\mathsf{Var}(Y_j) = 1$ and $\mathsf{Corr}(Y_j, Y_k) = 0,\ j \neq k = 1, \cdots, q$
- λ_{ij} for $i = 1, \cdots, p$, $j = 1, \cdots, q$ are fixed scalars, called **factor loadings**.
- e_i is the i-th specific-factor variate, which contributes only to X_i: $\mathsf{E}(e_i) = 0$, $\mathsf{Corr}(e_i, e_k) = 0,\ i \neq k = 1, \cdots, q$, $\mathsf{Corr}(e_i, Y_j) = 0,\ i = 1, \cdots, p,\ j = 1, \cdots, q$.

Matrix form

The factor model can be written as

$$\underline{X} = \Lambda \, \underline{Y} + \underline{e}$$

where

- $\underline{Y}$ is a $q \times 1$-dimensional vector of (unobservable) common factors
- Λ is a $(p \times q)$-dimensional matrix of loadings with (i, j)-th element λ_{ij};
- $\underline{e}$ is a vector of specific factors.

$\mathsf{E}(\underline{Y}) = \underline{0}$, $\mathsf{Cov}(\underline{Y}) = \mathsf{I}_q$, $\mathsf{Cov}(\underline{Y}, \underline{e}) = \mathbf{0}_{q \times p}$ (a $(q \times p)$-dim. matrix of 0's), $\mathsf{Cov}(\underline{e}) = \mathsf{diag}(\psi_{11}, \ldots, \psi_{pp}) \stackrel{def}{=} \Psi$

Therefore, $\Sigma = \text{Cov}(\underline{X}) = \Lambda\ \Lambda^T + \Psi$

Definition: we say a q-factor model holds for $\underline{X}$ if and only if

$$\text{Cov}(\underline{X}) = \Sigma = \Lambda\ \Lambda^T + \Psi$$

for some $(p \times q)$-dimensional matrix Λ and a diagonal matrix Ψ with nonnegative elements.

Comments and properties

- The model has a "linear regression" structure where $Y_1, \cdots, Y_q$ can be regarded as uncorrelated explanatory variables (which are unobserved) and $e_1, \ldots, e_p$ are observation errors.
- $\mathsf{E}[X_i] = 0,\ i = 1, \cdots, p$, and

$$\begin{aligned}\mathsf{Cov}(X_i, X_j) &= \mathsf{Cov}\left(\sum_{k=1}^{q} \lambda_{ik} Y_k + e_i, \sum_{m=1}^{q} \lambda_{jm} Y_m + e_j\right) \\ &= \sum_{k=1}^{q} \lambda_{ik}\lambda_{jk}\end{aligned}$$

Implications for covariance structure

- The i-th **specific** term e_i contributes only to the variance of X_i.

$$\text{Var}(X_i) = \text{Var}\left(\sum_{k=1}^{q} \lambda_{ik} Y_k + e_i\right) = \sum_{k=1}^{q} \lambda_{ik}^2 + \text{Var}(e_i)$$

- $h_i^2 = \sum_{k=1}^{q} \lambda_{ik}^2$ is the **communality** of X_i; the part of the variance of X_i shared with the other variables via the common factors.

- $\psi_{ii} = \text{Var}(e_i)$ is the **specific variance**.

Factor scores

Suppose that $\underline{X} = (X_1, X_2, X_3)^T$ are scores in three IQ tests.

The 1-factor model is

$$X_i = \lambda_i Y + e_i, \qquad i = 1, 2, 3$$

where Y could be interpreted as a measure of overall intelligence.

Given the IQ test scores for an individual, we can obtain a score for their intelligence.

Example: Open/close book exam marks data

```
Call:
factanal(x = exam, factors = 1)

Uniquenesses:
 Mechanics    Vectors    Algebra   Analysis Statistics
     0.641      0.555      0.158      0.403      0.476

Loadings:
           Factor1
Mechanics  0.599
Vectors    0.667
Algebra    0.917
Analysis   0.772
Statistics 0.724

               Factor1
SS loadings      2.766
Proportion Var   0.553
```

Results

- `uniquenesses` are the $\psi_{11}, \ldots, \psi_{pp}$.
- `loadings` are the entries of Λ.
- `SS loadings` is the sum of the squared loadings.
- `Proportion Var` is the proportion of the variation explained by one factor.

Example: Open/close book exam marks data

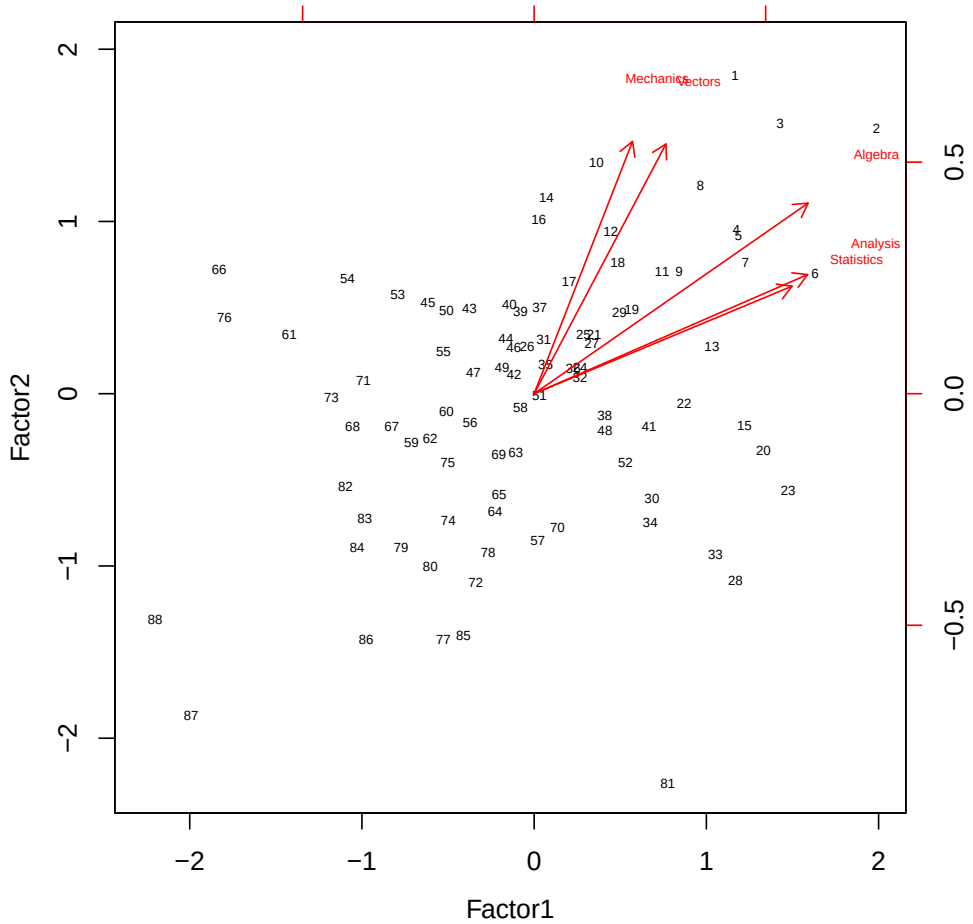
```
Call:
factanal(x = exam, factors = 2)

Uniquenesses:
 Mechanics    Vectors    Algebra   Analysis Statistics
     0.466      0.419      0.189      0.352      0.431

Loadings:
           Factor1 Factor2
Mechanics  0.265   0.681
Vectors    0.356   0.674
Algebra    0.740   0.514
Analysis   0.738   0.322
Statistics 0.696   0.290

                 Factor1 Factor2
SS loadings        1.774   1.370
Proportion Var     0.355   0.274
Cumulative Var     0.355   0.629
```

Plotting the factor loadings



Scaling, Rotation of Common Factors

Factor analysis is invariant to change of scale.

If a q-factor model holds for $\underline{X}$, i.e.

$$\Sigma = \text{Cov}(\underline{X}) = \Lambda\ \Lambda^T + \Psi,$$

for some $(p \times q)$-dimensional matrix Λ and a diagonal matrix Ψ, then a q-factor model also holds for any re-scaling of $\underline{X}$.

Consequently, the FA model can be applied to either the covariance or the correlation matrix and the results will essentially be equivalent.

Common factors are not unique for $q > 1$

Suppose that a q-factor model holds for $\underline{X}$ and let Γ be an $(q \times q)$-dimensional orthogonal matrix ($\Gamma\Gamma^T = I_q$), then we can also write

$$\underline{X} = \Lambda(\Gamma\ \Gamma^T)\underline{Y} + \underline{e} = (\Lambda\Gamma)(\Gamma^T\underline{Y}) + \underline{e} = \Lambda^\star\underline{Y}^\star + \underline{e}$$

where $\Lambda^\star = \Lambda\Gamma$ and $\underline{Y}^\star = \Gamma^T Y$.

Similarly,

$$\Sigma = \Lambda\Gamma\ \Gamma^T\Lambda^T + \Psi = \Lambda\Lambda^T + \Psi.$$

Therefore, a new q-factor model can be defined by rotating Y to $\Gamma^T Y$ and defining new loadings $\Lambda\Gamma$.

For fixed Ψ, this rotation is the only indeterminacy in the decomposition of Σ.

Choosing a rotation

This implies that there are, generally, an infinite number of Λ, which are all rotations of each other.

By default, R uses the Varimax criterion.

The Varimax criterion

Let the (i, j)-th element of Δ be δ_{ij}.

The **varimax rotation** maximizes the sum of the variances of the squared loadings within each column of Δ, given by

$$\phi(\Delta) = \sum_{j=1}^{q} \sum_{i=1}^{p} \left(\delta_{ij}^2 - p^{-1} \sum_{i=1}^{p} \delta_{ij}^2 \right)^2 .$$

Other rotations

Other rotations are available, see *e.g.* the `GPArotation` package in R.

These use alternative criteria to define the loadings. Our main aim is to find a rotation that leads to a suitable interpretation of the factors and so “simple” factors (which have only a few variables away from zero) are often preferred.

Choosing the number of factors q

The decision of how many factors to include in the model is typically critical.

Solutions differing by one factor can give very different factor loadings **for all** factors.

In general, a small q is preferred (*i.e.* factor models should offer a simplification).

The adequacy of a q-factor solution can be assessed by testing the goodness of fit of a q-factor model and comparing the fit with different values of q.

Example: Open/close book exam marks data

R provides these statistics

```
Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 8.65 on 5 degrees of freedom.
The p-value is 0.124
```

For the one factor model, the test statistic is 8.65 and $r = 5$. The null hypothesis cannot be rejected (p-value = 0.124). Therefore, a one factor model **adequately** describes the data.

```
Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 0.07 on 1 degree of freedom.
The p-value is 0.785
```

The two factor model is included for completeness. Again, the null hypothesis cannot be rejected and the two factor model **adequately** describes the data. We prefer the one factor model as it is simpler.

Discriminant Analysis

There are many problems where we can divide people/objects into groups. We want a rule for allocating new people/objects to groups using observed characteristics.

For example,

- Credit scoring – The two groups are those who will re-pay a loan and those who will not. Useful characteristics might be: size of loan, history of bad debts, salary, monthly outgoings, etc.
- Disease severity – Patients could be divided into groups with mild, moderate or severe forms of a disease. Useful characteristics might be: blood test results, age, gender, measure of co-morbidities.

Example 1: Iris Data

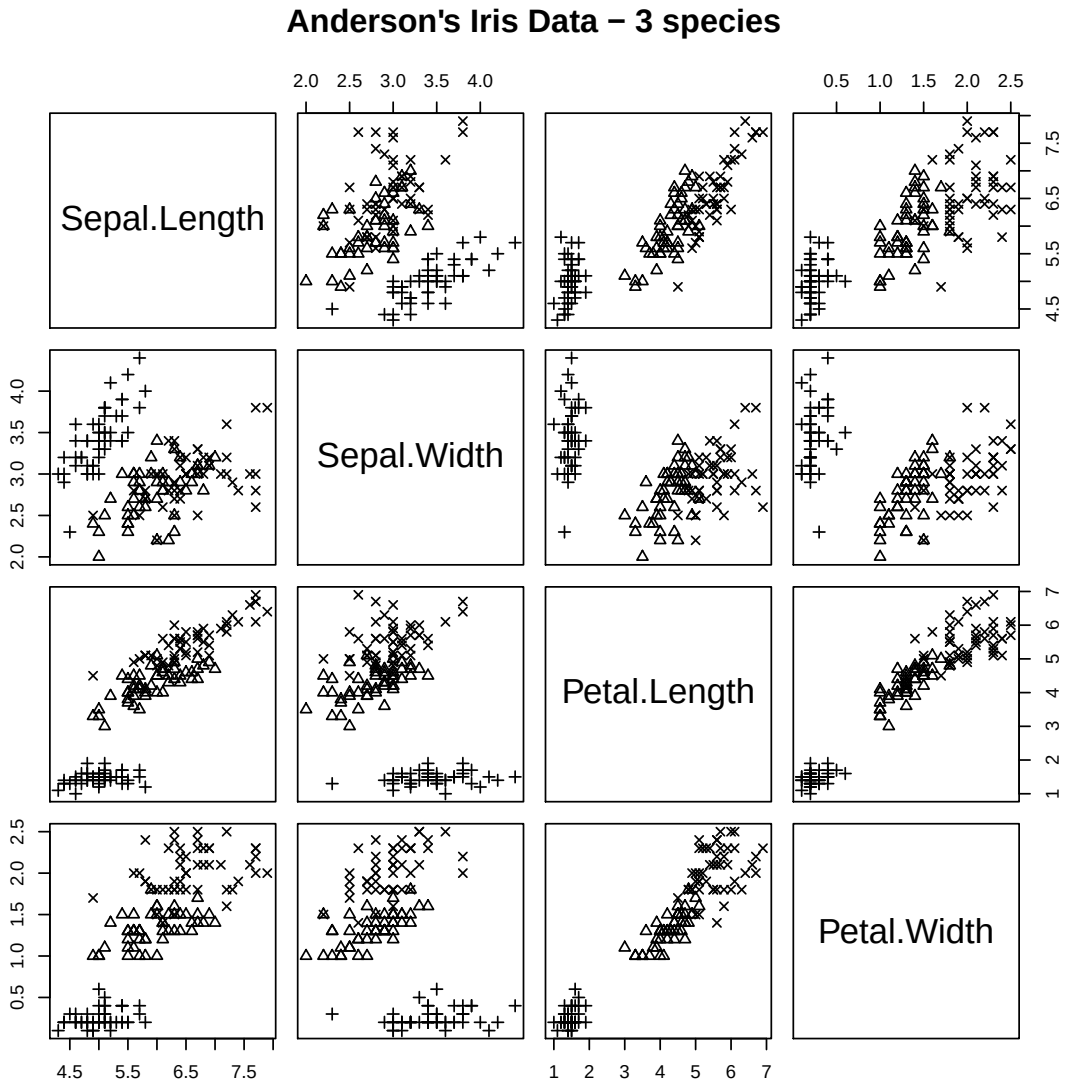
This is the famous (Fisher's or Anderson's) iris data set, which gives the measurements in centimetres on 50 flowers of the variables:

- Sepal length
- Sepal width
- Petal length
- Petal width

The flowers were divided into the species:

- Setosa (circles)
- Versicolor (triangles)
- Virginica (crosses)

Example 1: Iris Data



Suppose that we observe the sepal length and width, and petal length and width for a new iris. To which species does it belong?

Discriminant rule

Consider G populations or groups that are considered "distinctive" or "different".

Let $\underline{X}_{new} \in \mathbb{R}^p$ be the characteristics of a new object to be allocated to one of our G groups.

A **discriminant rule** is constructed by

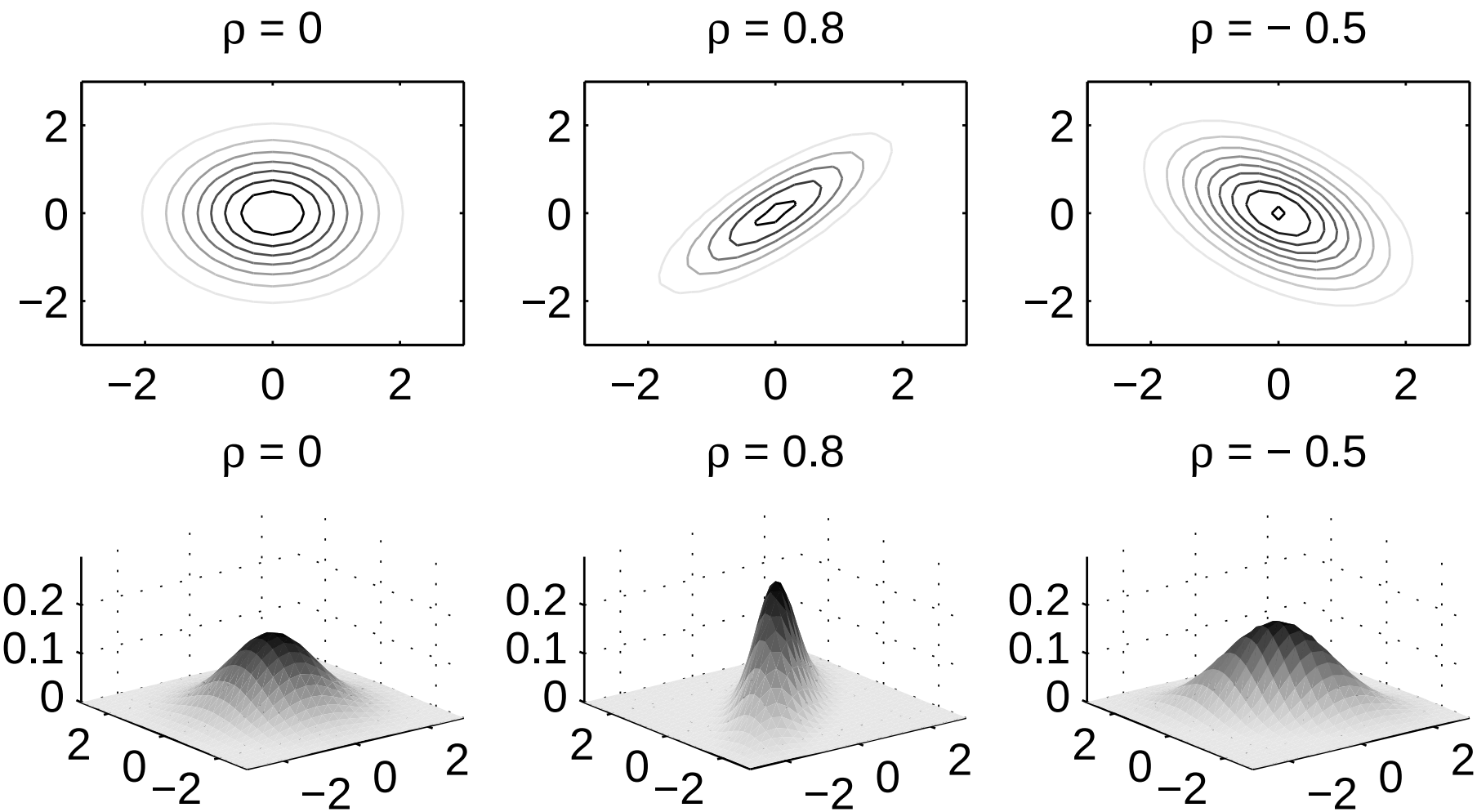
- partitioning $\mathbb{R}^p$ into disjoint regions $R_1, \ldots, R_G$ for which $\cup_{j=1}^{G} R_j = \mathbb{R}^p$.
- allocating $\underline{X}_{new}$ to the j-th group if $\underline{X}_{new} \in R_j, \quad j = 1, \ldots, G$.

Linear discriminant analysis

Let $\underline{X} = (X_1, X_2, \ldots, X_p)^T$ be the characteristics of a randomly chosen object. We assume $\underline{X} \overset{i.i.d.}{\sim} MN(\underline{\mu}_j, \Sigma)$ for an object from the j-th group. By MN we mean the multivariate normal distribution, which assumes that each of the p characteristics is a normally distributed random variable with some correlation structure between each pair of characteristics.

We want to assign $\underline{X}_{new}$ to one of the G groups.

MN distribution
Contour plots



Linear discriminant analysis

LDA approximates the so called Bayes classifier, which classifies an observation to the class for which the posterior probability of allocation is largest.

Bayes classifier takes its name from Bayes theorem, which states that for two events A and B with corresponding probabilities $Pr(A)$ and $Pr(B)$, the *conditional probability* of A given B, $Pr(A|B)$, is

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}$$

We refer to $Pr(A|B)$ as the posterior probability and to $Pr(A)$ as the prior probability of allocation.

LDA and Bayes classifier

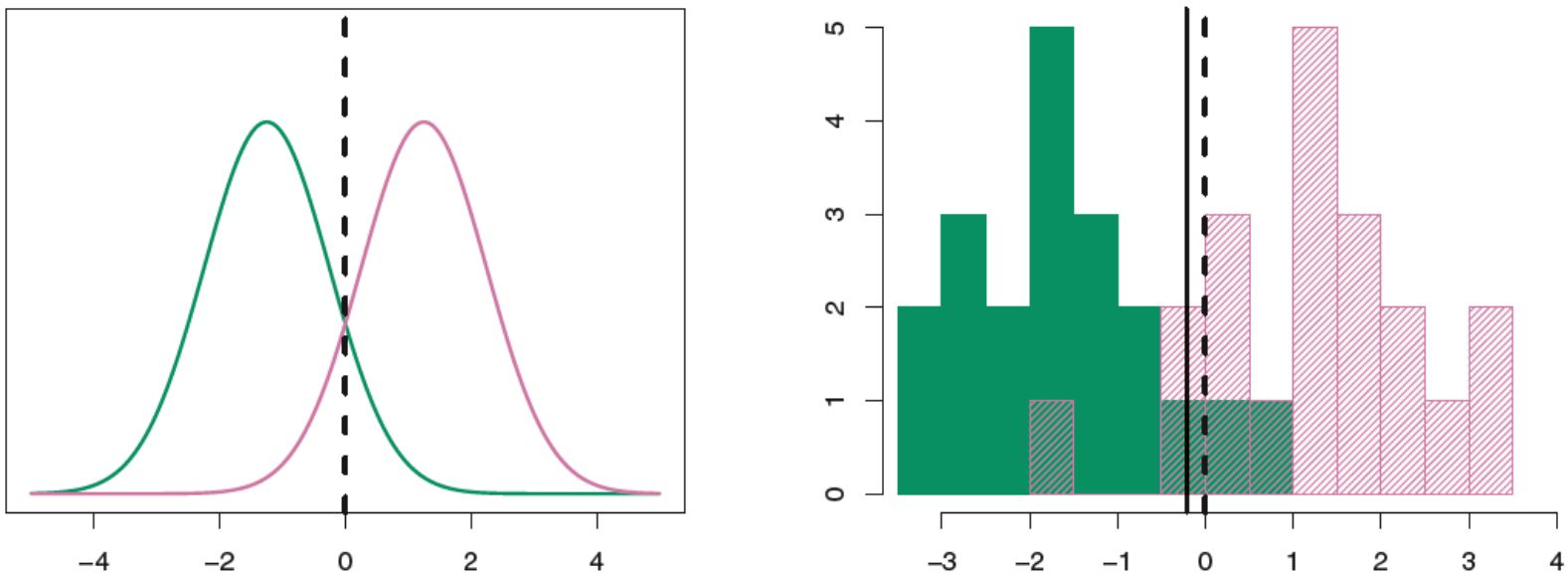


FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.*

Example: Fisher's Iris data

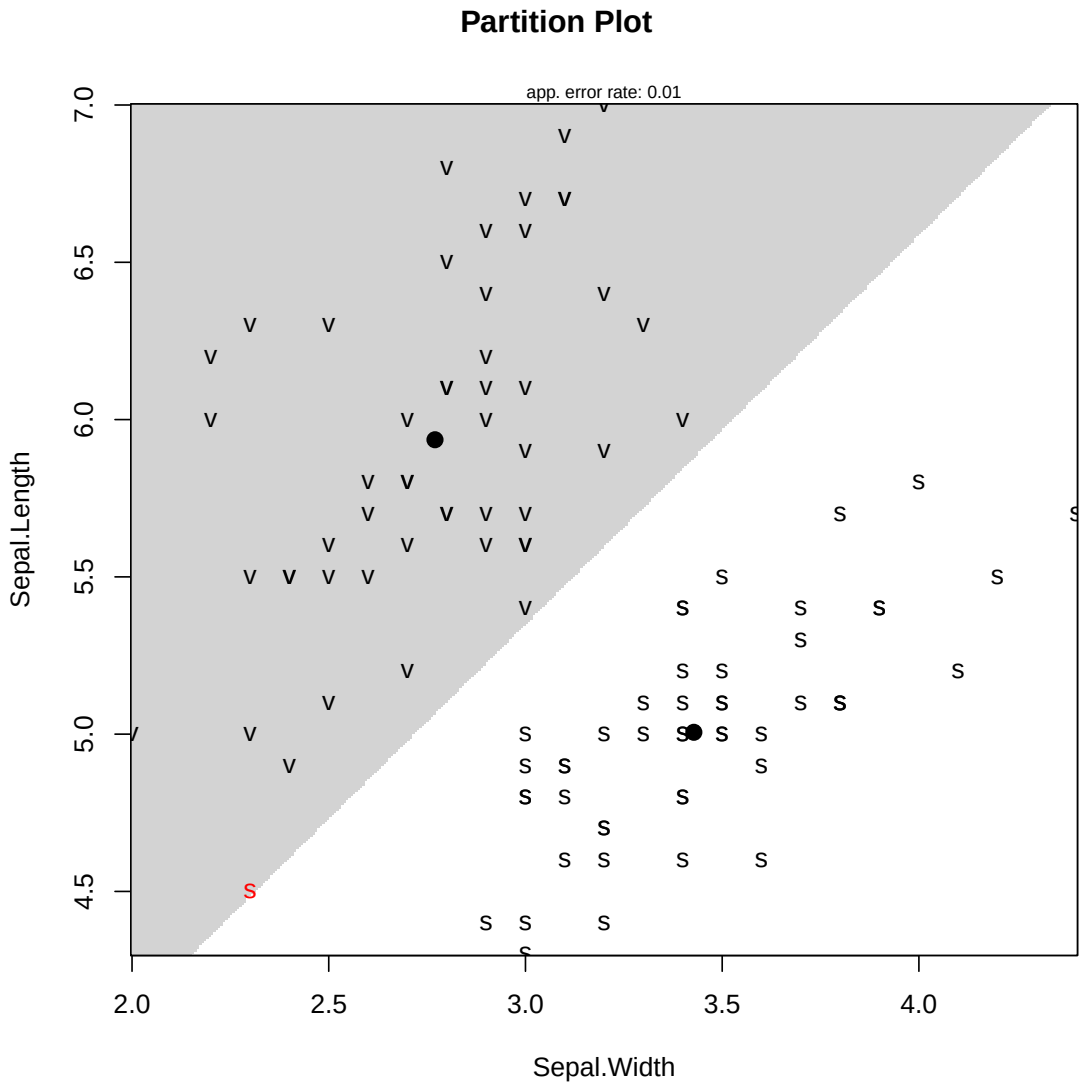
We want to discriminate between Setosa (Group 1) and Versicolor (Group 2) using the first two variables (Sepal length and Sepal width).

```
Prior probabilities of groups:
    setosa versicolor
       0.5        0.5

Group means:
           iris_subdata[, 1] iris_subdata[, 2]
setosa                 5.006             3.428
versicolor             5.936             2.770

Coefficients of linear discriminants:
                         LD1
iris_subdata[, 1]  2.560968
iris_subdata[, 2] -3.167079
```

The linear discriminant function is



Confusion matrix

A confusion matrix compares the linear discriminant analysis predictions to the true class assignments.

Elements on the diagonal of the matrix represent cases that were correctly classified, while off-diagonal elements represent cases that were misclassified.
For this example, the confusion matrix is

	setosa	versicolor
setosa	49	1
versicolor	0	50

Example: Fisher's Iris data
Three groups

Suppose that we consider all three species and assume

- F_{Setosa} is MN($\underline{\mu}_{Setosa}, \Sigma$)
- $F_{Versicolor}$ is MN($\underline{\mu}_{Versicolor}, \Sigma$)
- $F_{Virginica}$ is MN($\underline{\mu}_{Virginica}, \Sigma$)

with unknown parameters $\underline{\mu}_{Setosa}$, $\underline{\mu}_{Versicolor}$, $\underline{\mu}_{Virginica}$ and Σ.

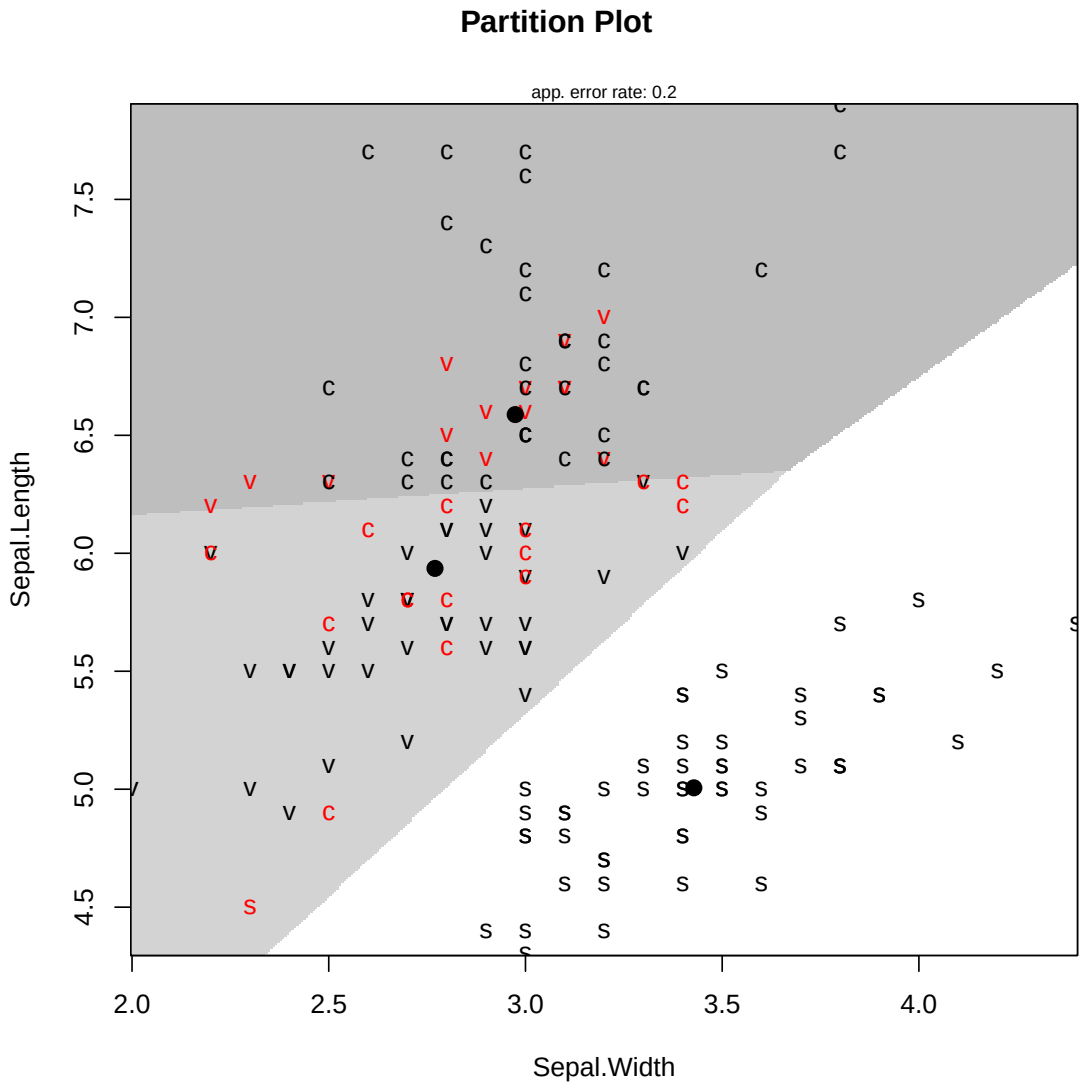
```
Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
           iris[, 1] iris[, 2]
setosa         5.006     3.428
versicolor     5.936     2.770
virginica      6.588     2.974

Coefficients of linear discriminants:
                 LD1        LD2
iris[, 1] -2.141178 -0.8152721
iris[, 2]  2.768109 -2.0960764

Proportion of trace:
   LD1    LD2
0.9628 0.0372
```

All three groups



The confusion matrix is given below:

	setosa	versicolor	virginica
setosa	49	1	0
versicolor	0	36	14
virginica	0	15	35

We can almost perfectly classify cases from the species “setosa”, but, our linear discriminant analysis does not perform as well in distinguishing between the other two species, at least when we consider the length and width of petals.

Improving the discriminant rule

Some ways to improve the discriminant rules are

- relax assumptions *e.g.*
  - not assuming common covariance matrix, leading to the so-called quadratic discriminant function.
  - use a more flexible distribution for F_i or a nonparametric estimate.

- use more variables if these are available.

Quadratic discriminant analysis

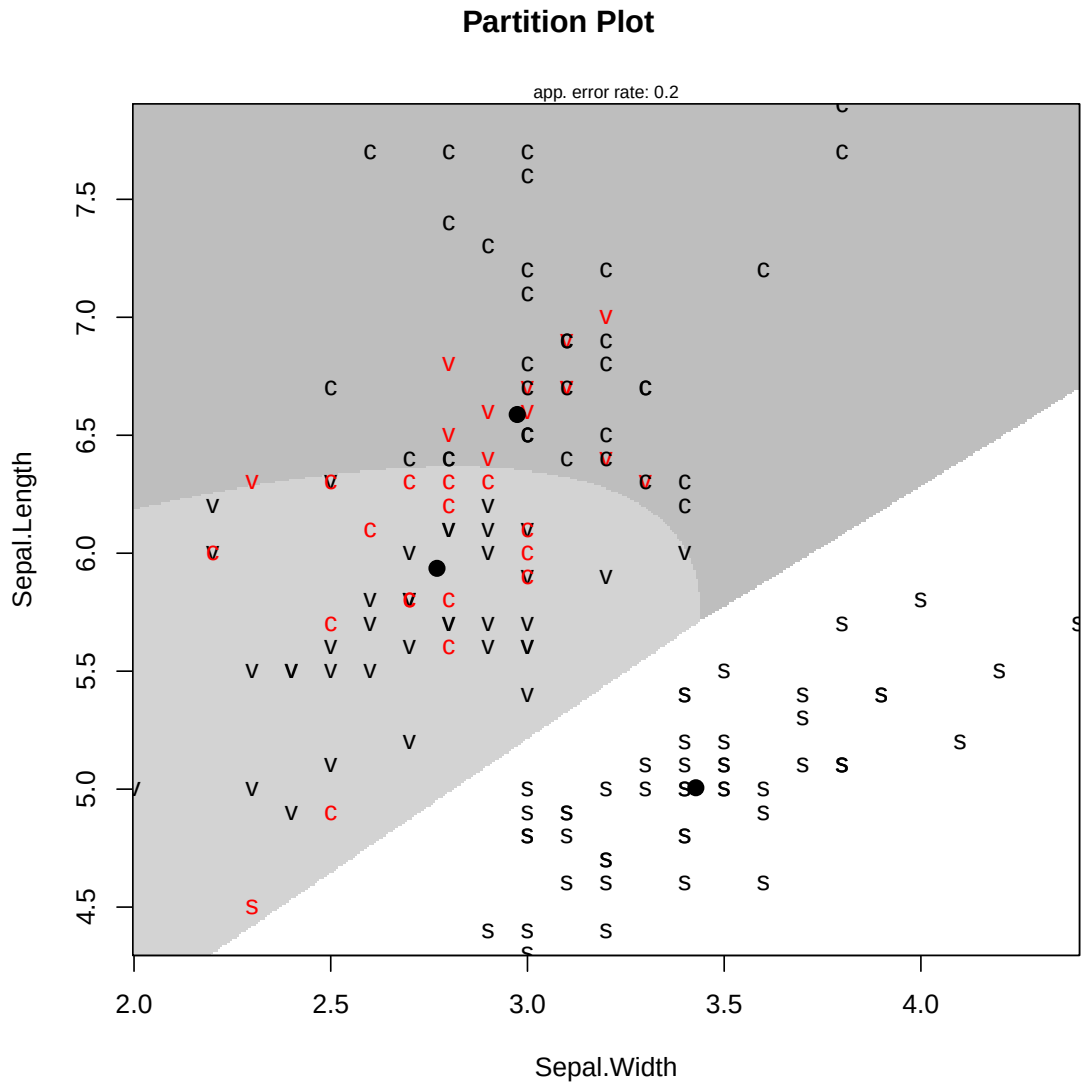
In this case we assume that each class has its own covariance matrix, instead of assuming that they all share a common covariance matrix.

The function takes its name from the fact that $\underline{X}_{new}$ appears as a quadratic in the discriminant function.

When there are p predictors, then estimating a covariance matrix requires estimating $p(p+1)/2$ parameters. QDA estimates a separate covariance matrix for each class, for a total of $Kp(p+1)/2$ parameters.

LDA on the other hand only estimates one covariance matrix, so it is a much less flexible classifier than QDA and as a result could be biased, but will also have smaller variance.

So when choosing between LDA and QDA we have to consider this bias-variance trade-off.


Partition Plot
app. error rate: 0.2
Sepal.Length
Sepal.Width
4.5
5.0
5.5
6.0
6.5
7.0
7.5
2.0
2.5
3.0
3.5
4.0

Want to learn more?

- An Introduction to Statistical Learning by G. James, D. Witten, T. Hastie and R. Tibshirani
  `http://www-bcf.usc.edu/~gareth/ISL/index.html`
- An introduction to Applied Multivariate Analysis with R by B. Everitt and T. Hothorn (library e-book)
- Introduction to Statistical Learning (MA7529 MSc in Statistical Data Science-SMSAS)