Training Fiche Template

Title	LINEAR DISCRIMINANT ANALYSIS
Keywords (meta tags)	discriminant analysis, classification, R, Bayesian analysis
Language	English
Objectives / Goals / Learning outcomes	**the objective of this module is to introduce and explain the basics of Linear Discriminant Analysis (LDA).** **At the end of this module you will be able to:** - **Identify situations on which LDA can be useful** - **Calculate LDA functions** - **Interpret the results produced by descriptive and predictive LDA**
Training course:	
Data Science Literacy	
Data Visualisation and Visual Analytics Module	X
Introduction to Data science for Human & Social Sciences	
Data Science for good	
Data Journalism and Storytelling	
Description	In this training module you will be introduced to the use of Linear Discriminant Analysis (LDA). LDA is as a method for finding linear combinations of variables that best separates observations into groups or classes, and it was originally developed by Fisher (1936). This method maximizes the ratio of between-class variance to the within-class variance in any particular data set. By doing this, the between-groups variability is maximized, which results in maximal separability. LDA can be used with purely classification purposes, but also with predictive objectives.

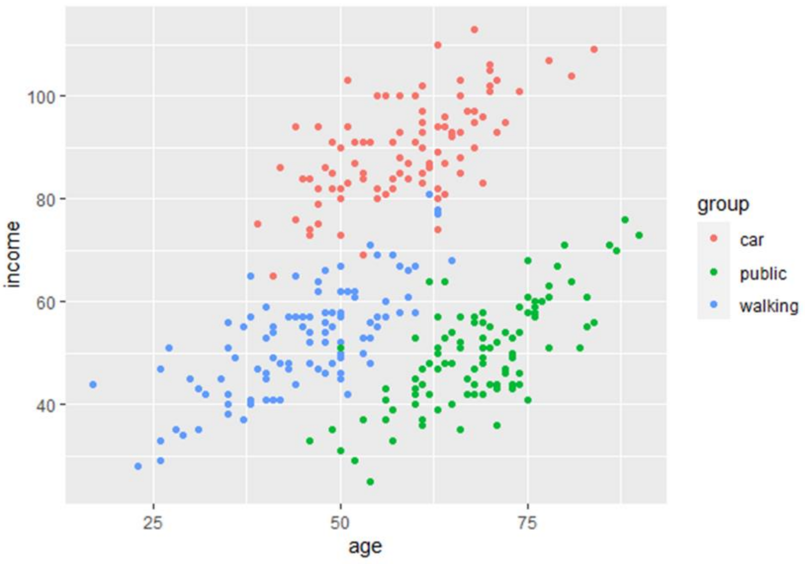
Co-funded by the
Erasmus+ Programme
of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Contents arranged in 3 levels

1. INTRODUCTION: MOTIVATION BY AN ILLUSTRATIVE EXAMPLE

Suppose that we have a sample of individuals, and we observe the transportation mode (by car, by public transport or by walking) they usually take to move within a city. We know that the choice of the transportation mode is partially influenced by their economic status, and we observe data on their age in years and their household annual income, together with the chosen mean of transportation:



We want to know how these two covariates help to classify (i.e., discriminate) the individuals by assigning them to a specific category of transportation mode. We can see that there is not perfect classification: individuals with high income tend to use cars more frequently, but there is great overlap of “walking” and “public transport” categories for those with lower incomes. And there is a larger overlapping among categories regarding their distribution by age: older individuals do not walk, but at younger values age is not a good predictor of transportation mode. This is the typical problem that LDA addresses.

2. LDA for classification

2.1. Formulation

LDA functions can be recovered to help with the classification of the data based on a matrix of covariates $\mathbf{X}$. Similar to Principal Component Analysis (PCA), LDA functions aim at finding a linear combination of the original data as:

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

$$\mathrm{LDA} = \mathbf{u}^{\mathbf{T}}\mathbf{X}$$

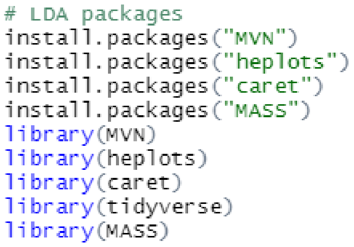
where the between-class variance ($\mathbf{B}$) is maximized relative to the within-class variance ($\mathbf{W}$), which can be approached as a generalized eigenvalue problem:

$$\mathrm{u} = \arg\max_{\mathbf{u}} \frac{\mathbf{u}^{\mathbf{T}}\mathbf{B}\mathbf{u}}{\mathbf{u}^{\mathbf{T}}\mathbf{W}\mathbf{u}}$$

Discriminant coordinates are obtained from the eigenvectors of $\mathbf{W}$^(−$\mathbf{1}$) $\mathbf{B}$.

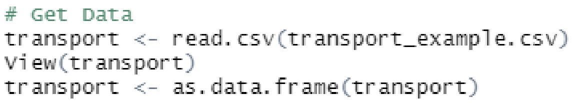
2.2. An example

As an illustrative example, we solve the classification problem of transportation mode basing on age and income by LDA in R. This can be easily done by the "lda" function within the "mass" library. For all the analysis presented here, we will need to install and load the following R pakages:



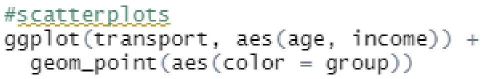
```
# LDA packages
install.packages("MVN")
install.packages("heplots")
install.packages("caret")
install.packages("MASS")
library(MVN)
library(heplots)
library(caret)
library(tidyverse)
library(MASS)
```

The data studied comes in a csv file (called "trasnpor_example"), which can be easily imported to R by runing this piece of code:



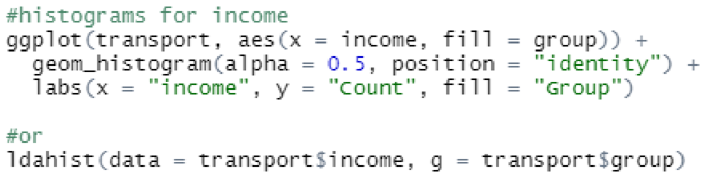
```
# Get Data
transport <- read.csv(transport_example.csv)
View(transport)
transport <- as.data.frame(transport)
```

In ordser to have a first impresion of the data, we can plot the sample in the form of a scatter plot as:



```
#scatterplots
ggplot(transport, aes(age, income)) +
  geom_point(aes(color = group))
```

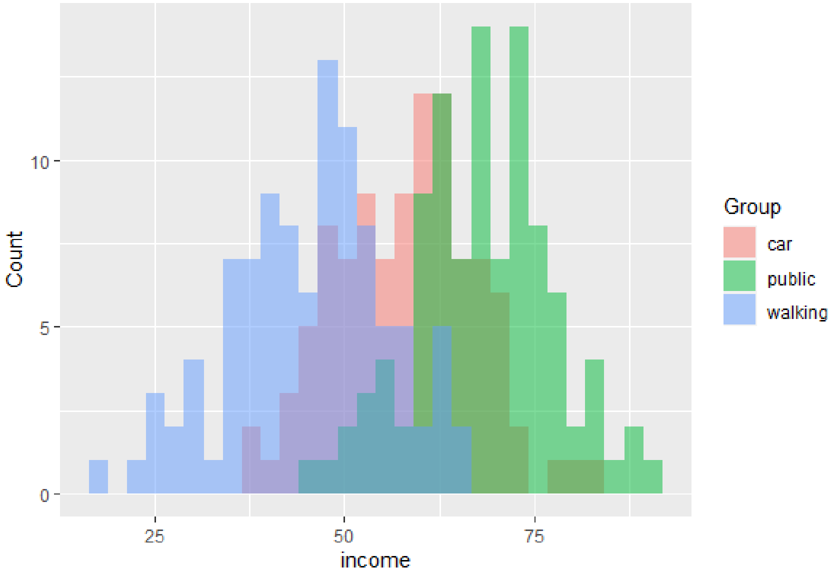
The code lines above produces the scatterplot shown in the introductory section of thid document. Alternatively, we could plot the data as a series of histograms as:



```
#histograms for income
ggplot(transport, aes(x = income, fill = group)) +
  geom_histogram(alpha = 0.5, position = "identity") +
  labs(x = "income", y = "Count", fill = "Group")

#or
ldahist(data = transport$income, g = transport$group)
```

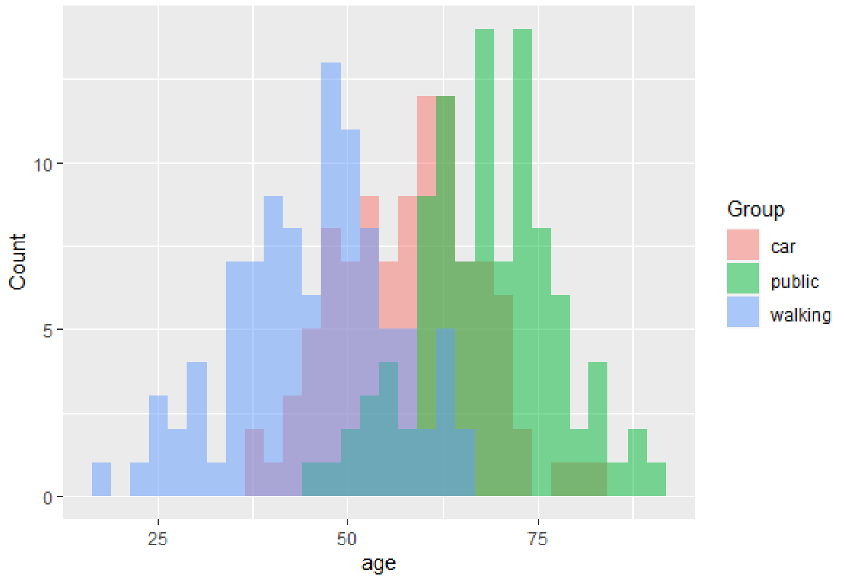
By running any of these two lines, we can have in a glimpse an idea on how transportation mode distribute across values og age and income. For example:



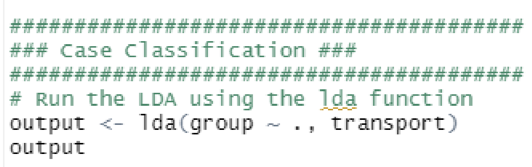
Or:

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

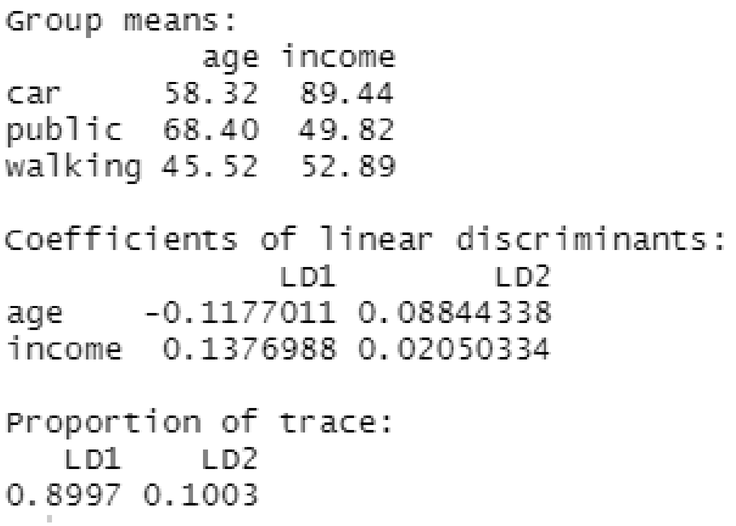


LDA is conducted by simply running:



```
##########################################
### Case Classification ###
##########################################
# Run the LDA using the lda function
output <- lda(group ~ ., transport)
output
```

The typical output shows the initial means by group, the coefficients in the LD projections and the proportion of the between variance (trace) that each LD coordinate explains:



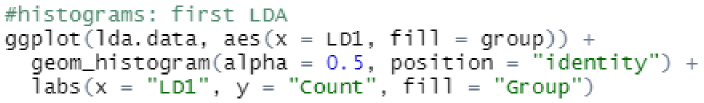
```
Group means:
             age  income
car        58.32   89.44
public     68.40   49.82
walking  45.52   52.89

Coefficients of linear discriminants:
                 LD1           LD2
age      -0.1177011  0.08844338
income    0.1376988  0.02050334

Proportion of trace:
   LD1     LD2
0.8997  0.1003
```

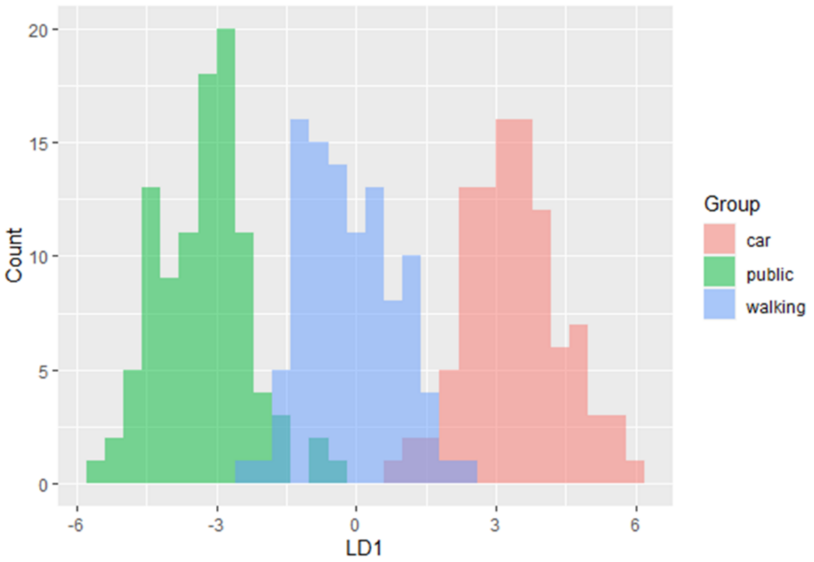
In our example, the first LD coordinate is positively correlated with income and negatively with age, and contains almost 90% of the inter-class variability. The second LD function shows positive but weaker correlation with both variables, and only accounts for approximately 10% of the between variability.

The new coordinates are produced projecting the original data points with the LDA coefficients by the expression $\mathbf{u}^{\mathsf{T}}\mathbf{X}$. In these new coordinates, observations are more clearly separated across groups. In our example, we have two LD coordinates for each individual, given their age and income. The coordinates corresponding to the first LD function have the larger discriminant power. We can easily see this discriminant power by plotting in R an histogram, now putting the first LD coordinates in the horizontal axis:



```
#histograms: first LDA
ggplot(lda.data, aes(x = LD1, fill = group)) +
  geom_histogram(alpha = 0.5, position = "identity") +
  labs(x = "LD1", y = "Count", fill = "Group")
```

Obtaining:



This plot shows how the amount of overlapping diminishes considerably. In other words, the first LD coordinate (remember that it is a "composite" that correlates negatively with age and positively with income) adequately discriminates among the transportation categories.

3. Predictive LDA

3.1 The procedure

LDA can be used not only for (descriptive) classification purposes, but also with the objective of predicting class membership. For example, suppose that we have data of the age and annual household income for an (in-sample or out-of-sample) individual, and we'd like to predict the transportation mode that this person is most likely to use. LDA can be helpful in providing us with a prediction, in a similar fashion to multinominal logit or probit models.

For this predictive purpose, some assumptions are required:

- groups are multivariate normal
- equal variances-covariances across groups

The formulation of predicitive LDA is related to the formulation of Bayes tehorem for updating probabilities: Let g be the number of groups and q_i the prior probability (usually observed relative frequencies) for group i. Let $\mathbf{x}$ be a vector of observations of covariates for one individual.The (posterior) probability of belonging to group G_i conditional on $\mathbf{x}$, P(G_i |$\mathbf{x}$), can be expressed as:

$$\mathrm{P}(G_i|\mathbf{x}) = \frac{q_i \mathrm{P}(\mathbf{x}|G_i)}{\sum_{j=1}^{g} q_j \, \mathrm{P}(\mathbf{x}|G_j)}$$

This is a Bayesian approach that updates the prior probabilities q_i basing on the conditional probabilities P($\mathbf{x}$|G_i). Under the normality assumptions:

$$\mathrm{P}(\mathbf{x}|G_i) = (2\pi)^{(-p/2)}|\mathbf{W}|^{(-1/2)}e^{(-D_i^2/2)}$$

where |$\mathbf{W}$| is the determinant of the within-class variance matrix and D_i is $D_i \ = (\mathbf{x} - \bar{\mathbf{x}}_\mathbf{i})^T\mathbf{W}^{-1} \ (\mathbf{x} - \bar{\mathbf{x}}_\mathbf{i})$. Plugging the expression of $\mathrm{P}(\mathbf{x}|G_i)$ in the formula for $\mathrm{P}(G_i|\mathbf{x})$, we have:

Co-funded by the
Erasmus+ Programme
of the European Union

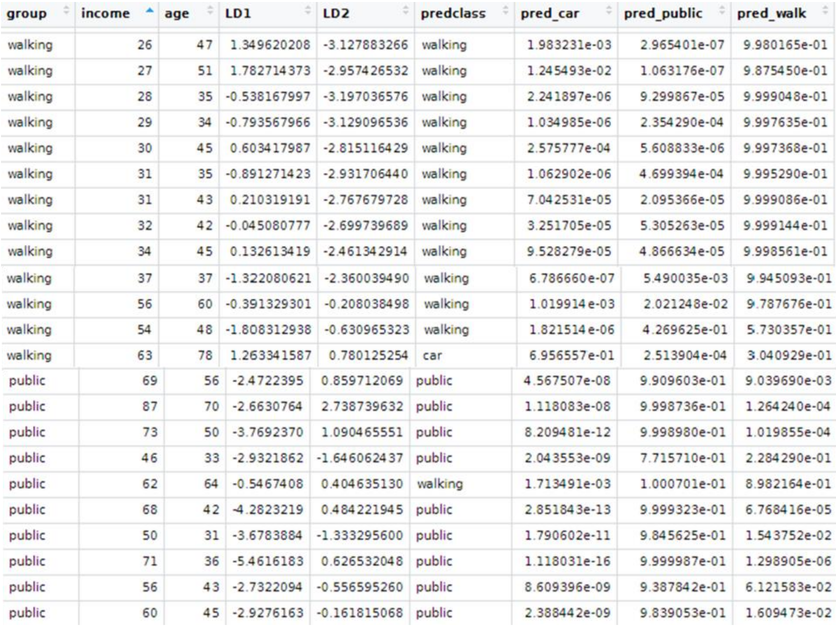
With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

$$\mathrm{P}(G_i|\mathbf{x}) = \frac{q_i e^{\left(-D_i^2/2\right)}}{\sum_{j=1}^{g} q_j\, e^{\left(-D_j^2/2\right)}}$$

3.2. An example with R

The LDA routine in R can produce posterior probabilities basing on the assumptions and the formulation detailed before. The LDA functions allows for predicting the most likely class membership for any individual, given a vector of covariates (age and household income in the example).

As an illustration, the table displayed below shows the predicted probabilities for each group for a subset of individuals in the sample. The priors q_i are assumed to be identical for each one of the three transportation modes ($q_i = 1/3$).

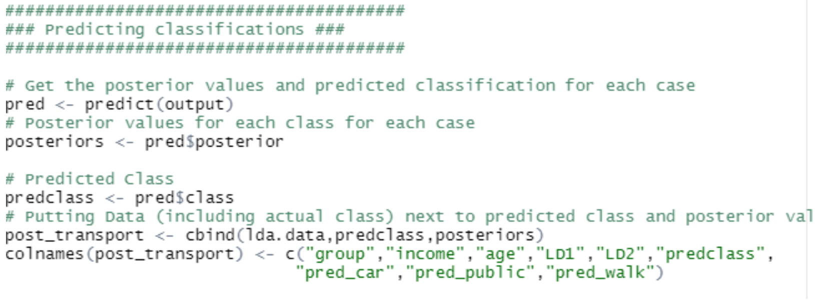


group	income	age	LD1	LD2	predclass	pred_car	pred_public	pred_walk
walking	26	47	1.349620208	-3.127883266	walking	1.983231e-03	2.965401e-07	9.980165e-01
walking	27	51	1.782714373	-2.957426532	walking	1.245493e-02	1.063176e-07	9.875450e-01
walking	28	35	-0.538167997	-3.197036576	walking	2.241897e-06	9.299867e-05	9.999048e-01
walking	29	34	-0.793567966	-3.129096536	walking	1.034985e-06	2.354290e-04	9.997635e-01
walking	30	45	0.603417987	-2.815116429	walking	2.575777e-04	5.608833e-06	9.997368e-01
walking	31	35	-0.891271423	-2.931706440	walking	1.062902e-06	4.699394e-04	9.995290e-01
walking	31	43	0.210319191	-2.767679728	walking	7.042531e-05	2.095366e-05	9.999086e-01
walking	32	42	-0.045080777	-2.699739689	walking	3.251705e-05	5.305263e-05	9.999144e-01
walking	34	45	0.132613419	-2.461342914	walking	9.528279e-05	4.866634e-05	9.998561e-01
walking	37	37	-1.322080621	-2.360039490	walking	6.786660e-07	5.490035e-03	9.945093e-01
walking	56	60	-0.391329301	-0.208038498	walking	1.019914e-03	2.021248e-02	9.787676e-01
walking	54	48	-1.808312938	-0.630965323	walking	1.821514e-06	4.269625e-01	5.730357e-01
walking	63	78	1.263341587	0.780125254	car	6.956557e-01	2.513904e-04	3.040929e-01
public	69	56	-2.4722395	0.859712069	public	4.567507e-08	9.909603e-01	9.039690e-03
public	87	70	-2.6630764	2.738739632	public	1.118083e-08	9.998736e-01	1.264240e-04
public	73	50	-3.7692370	1.090465551	public	8.209481e-12	9.998980e-01	1.019855e-04
public	46	33	-2.9321862	-1.646062437	public	2.043553e-09	7.715710e-01	2.284290e-01
public	62	64	-0.5467408	0.404635130	walking	1.713491e-03	1.000701e-01	8.982164e-01
public	68	42	-4.2823219	0.484221945	public	2.851843e-13	9.999323e-01	6.768416e-05
public	50	31	-3.6783884	-1.333295600	public	1.790602e-11	9.845625e-01	1.543752e-02
public	71	36	-5.4616183	0.626532048	public	1.118031e-16	9.999987e-01	1.298905e-06
public	56	43	-2.7322094	-0.556595260	public	8.609396e-09	9.387842e-01	6.121583e-02
public	60	45	-2.9276163	-0.161815068	public	2.388442e-09	9.839053e-01	1.609473e-02

The predicted class corresponds to the highest $\mathrm{P}(G_i|\mathbf{x})$ for each individual. They are calculated by applying the following routine in Rstudio:

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



```
########################################
### Predicting classifications ###
########################################

# Get the posterior values and predicted classification for each case
pred <- predict(output)
# Posterior values for each class for each case
posteriors <- pred$posterior

# Predicted Class
predclass <- pred$class
# Putting Data (including actual class) next to predicted class and posterior values
post_transport <- cbind(lda.data,predclass,posteriors)
colnames(post_transport) <- c("group","income","age","LD1","LD2","predclass",
                             "pred_car","pred_public","pred_walk")
```

In most of cases, LDA correctly predicts the group to which each individual belongs. There are some cases, however, for which LDA does not predict correctly. These cases correspond to the overlapping observations that still remain in the LDA classification

Self-assessment (multiple choice queries and answers)	LDA is a statistical technique that allows for: a) Classification of data into groups b) Predicting class membership c) Both answers are true Assumptions required to apply LDA for predictive purposes are: a) Multivariate normality across groups b) Equal variances-covariances across groups c) Both answers are true LDA bases on maximizing the ratio: a) Between-groups versus within-class variability b) Within-groups versus total variability c) Total versus Within-groups variability
Resources (videos, reference link)	
Related material	
Related PPT	
Bibliography	Boedeker, P., & Kearns, N. T. (2019). Linear discriminant analysis for prediction of group membership: A user-friendly primer. Advances in Methods and Practices in Psychological Science, 2, 250-263.
Provided by	[Uniovi]

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.