UNIVERSIDADE FEDERAL FLUMINENSE

Programa de Mestrado e Doutorado em Engenharia de Produção

Multivariate Data Analysis

Exploratory Factor Analysis

Professor: Valdecy Pereira, D. Sc.

email: valdecy.pereira@gmail.com

Outline

1. Definition
2. EFA vs PCA
3. Example
4. Bibliography

MVDA - *EFA*

- It is performed to understand the interdependencies (correlations) between variables.
- Each dimension (Factors, Latent Variables, Unobserved Variables or Constructs) is formed by highly correlated variables.
- The idea is to come up with a simple structure, easy to be explained (Law of Parsimony).
- The correlation between a variable and a factor is called Factor Loadings.
- The Commonality (h^2), measures how the variance of each variable can be explained by factors.

MVDA - *EFA*

Exploratory Factor Analysis - OBJECTIVE

- Identify the data structure: it can be used to discover and explore the basic dataset structure.

Principal Component Analysis - OBJECTIVE

- Reduce the volume of data: it can be used to reduce the mass of variables into a manageable amount.

Both methods differ in purpose so we can affirm that:

EFA is not PCA!

MVDA - *EFA*

Exploratory Factor Analysis - TOTAL VARIANCE

- The initial variance is estimated, and the total variance is composed by:

Total Variance = Common Variance + Specific Variance + Error

Principal Component Analysis - TOTAL VARIANCE

- The initial variance is equal 1, and the total variance is composed by:

Total Variance = Common Variance + Unique Variance

Unique Variance = Specific Variance + Error

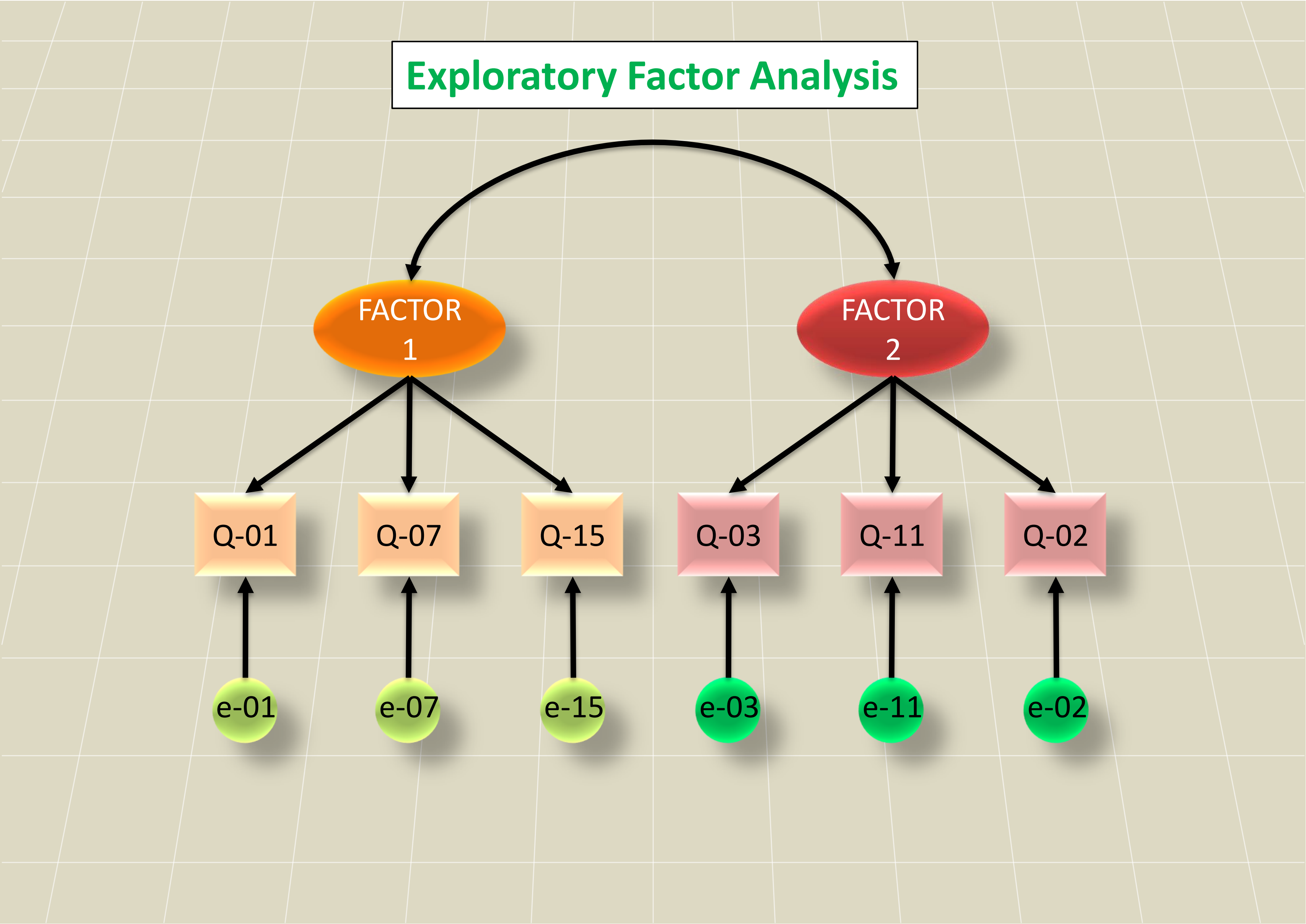
MVDA - *EFA*

Factors Extraction

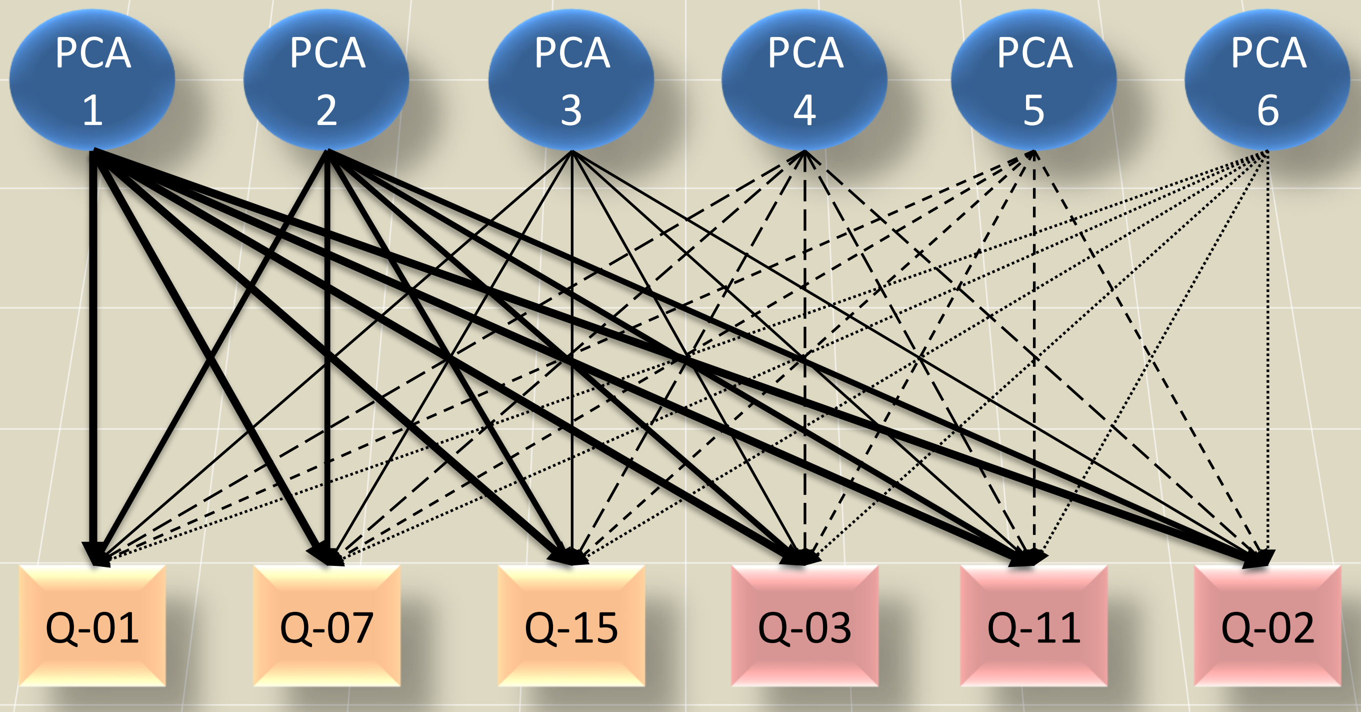
- Exploratory Factor Analysis: It is necessary to explore the structure to find the optimal number of Factors.

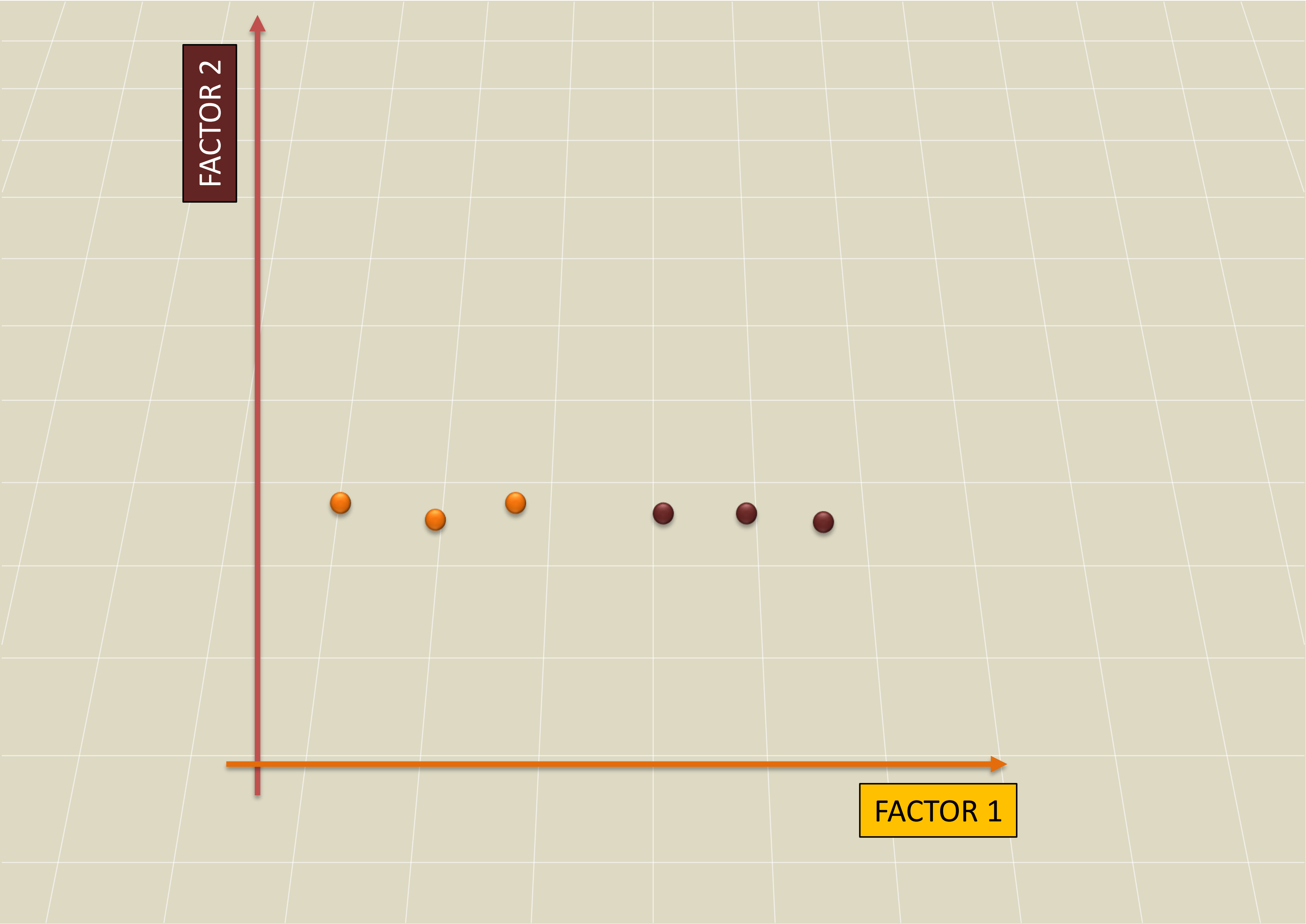
- Principal Component Analysis: Generates a Factor for each variable, but the variance is concentrated in the first ones.

FACTORS / QUESTIONS	Factor 1	Factor 2
Question 01	●	
Question 07	●	
Question 15	●	
Question 03		◆
Question 11		◆
Question 02		◆

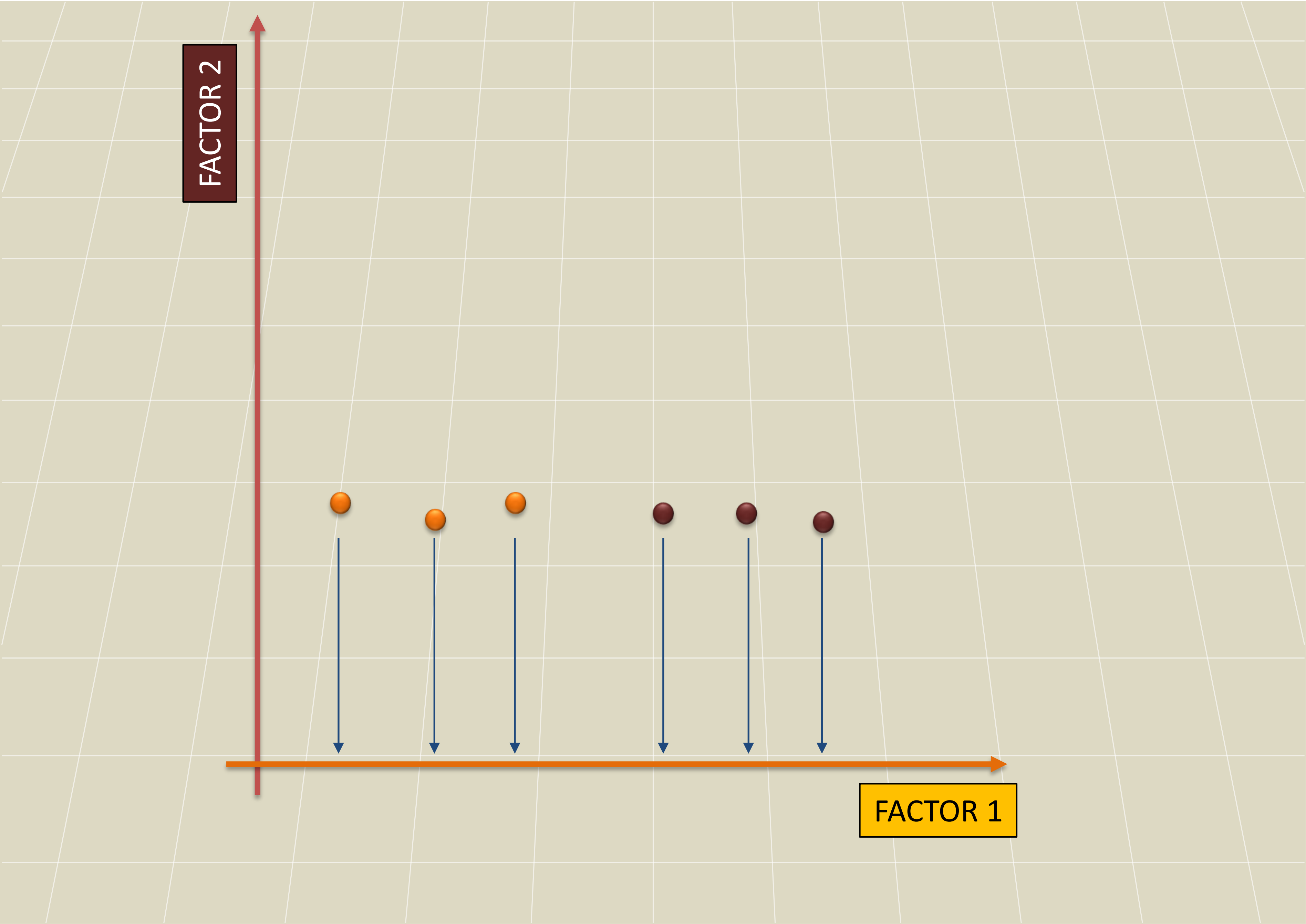

Exploratory Factor Analysis
FACTOR 1
FACTOR 2
Q-01
Q-07
Q-15
Q-03
Q-11
Q-02
e-01
e-07
e-15
e-03
e-11
e-02

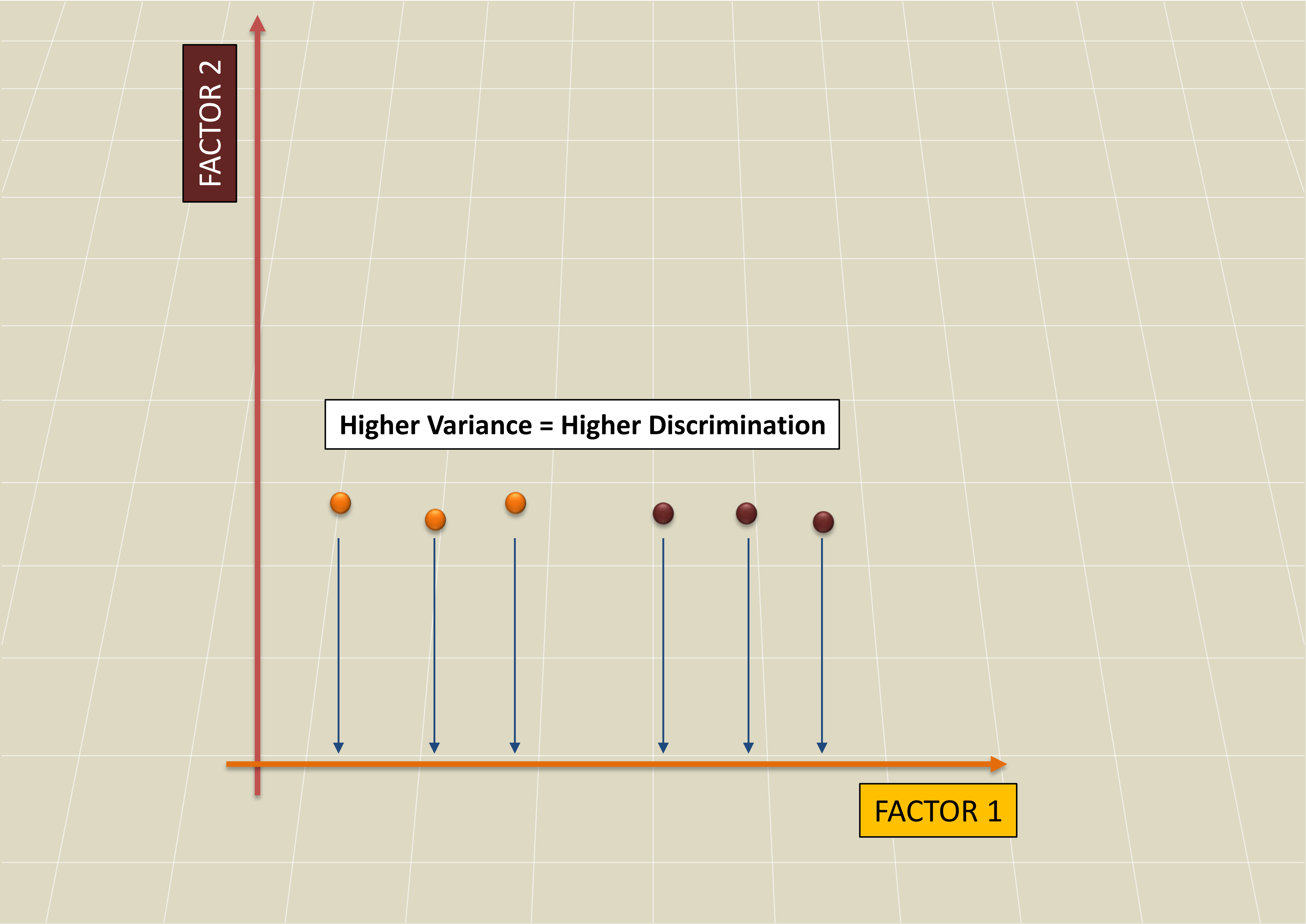
Principal Component Analysis



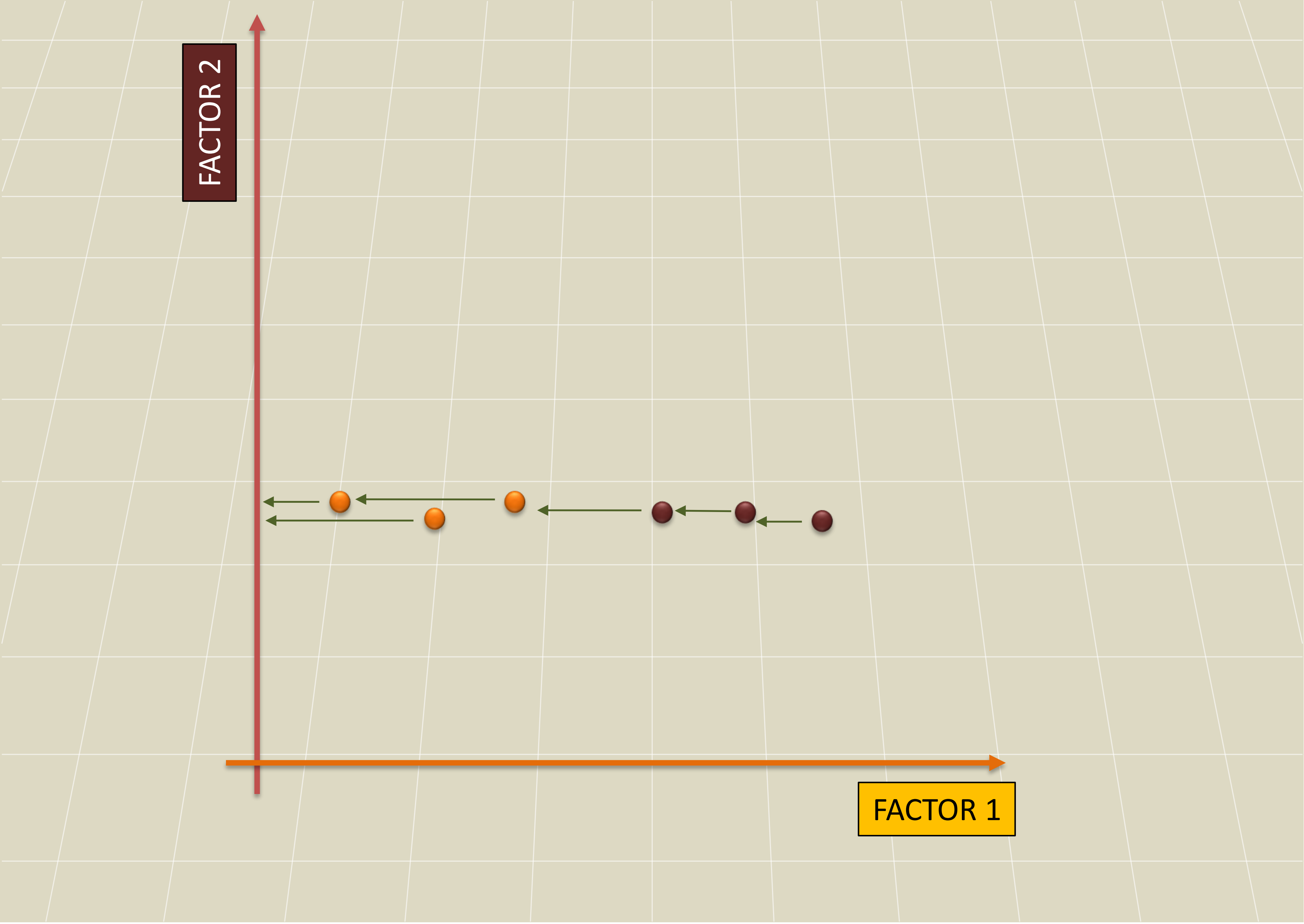


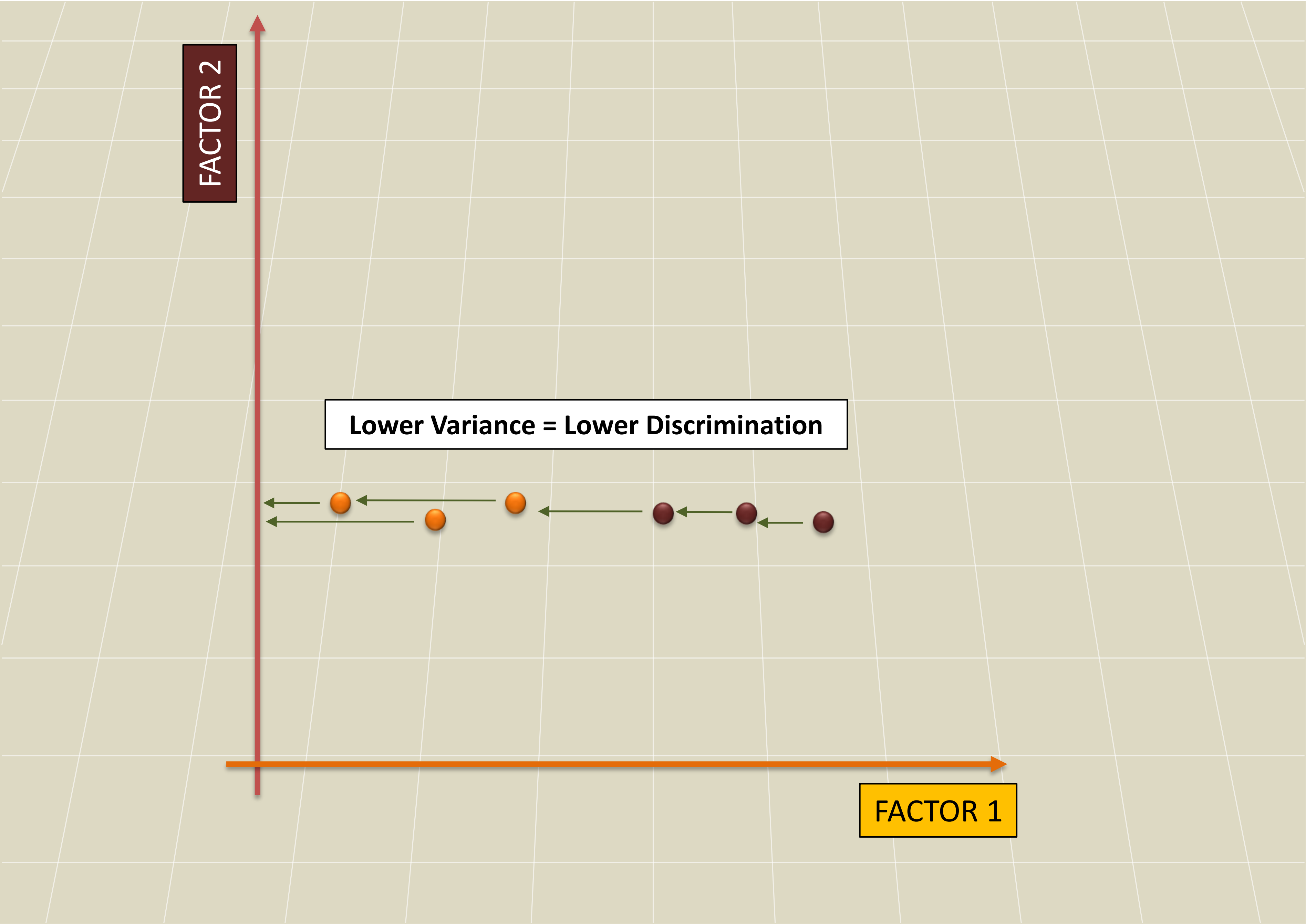
FACTOR 2
FACTOR 1

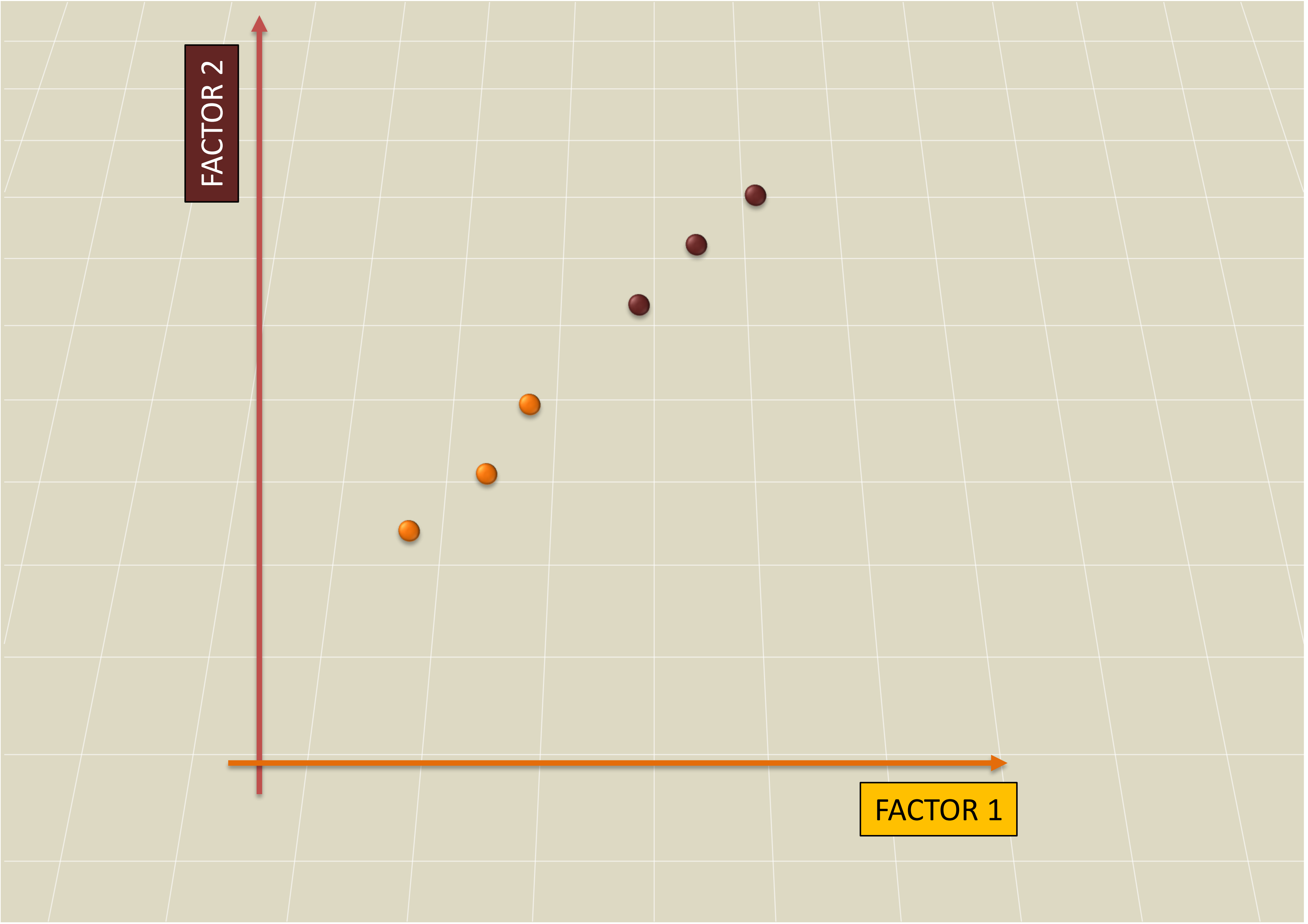

FACTOR 2
FACTOR 1

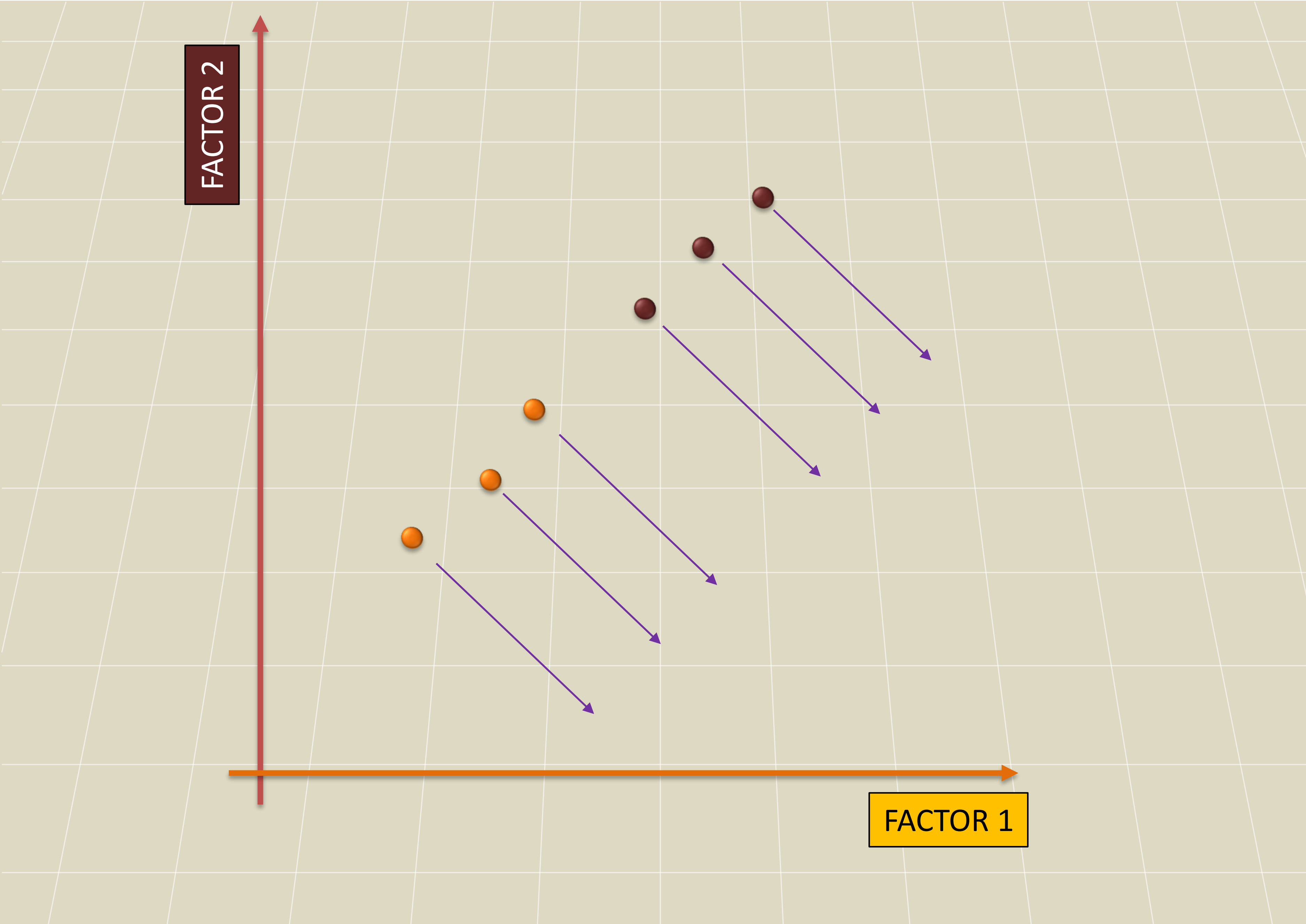


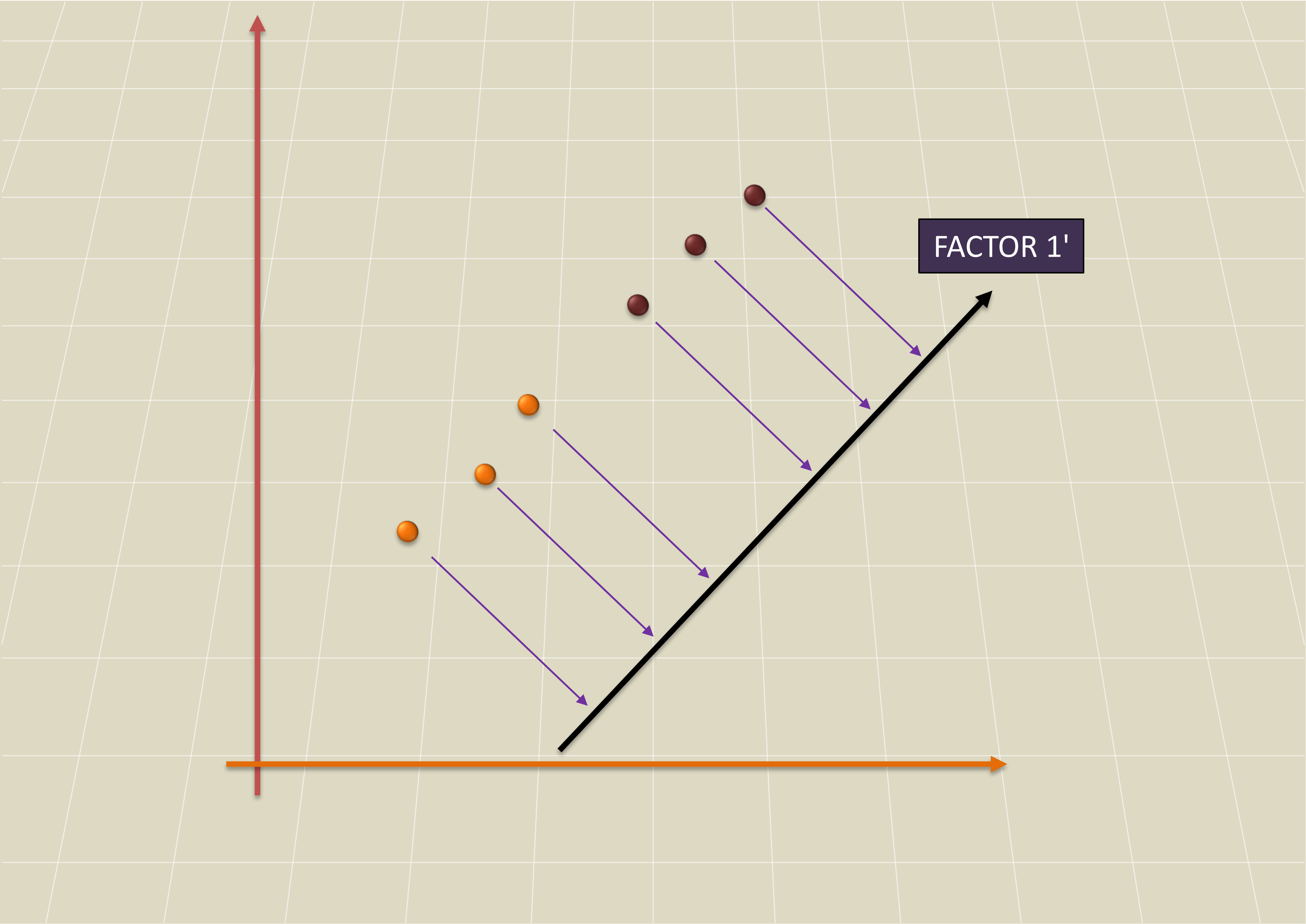
FACTOR 2
Higher Variance = Higher Discrimination
FACTOR 1

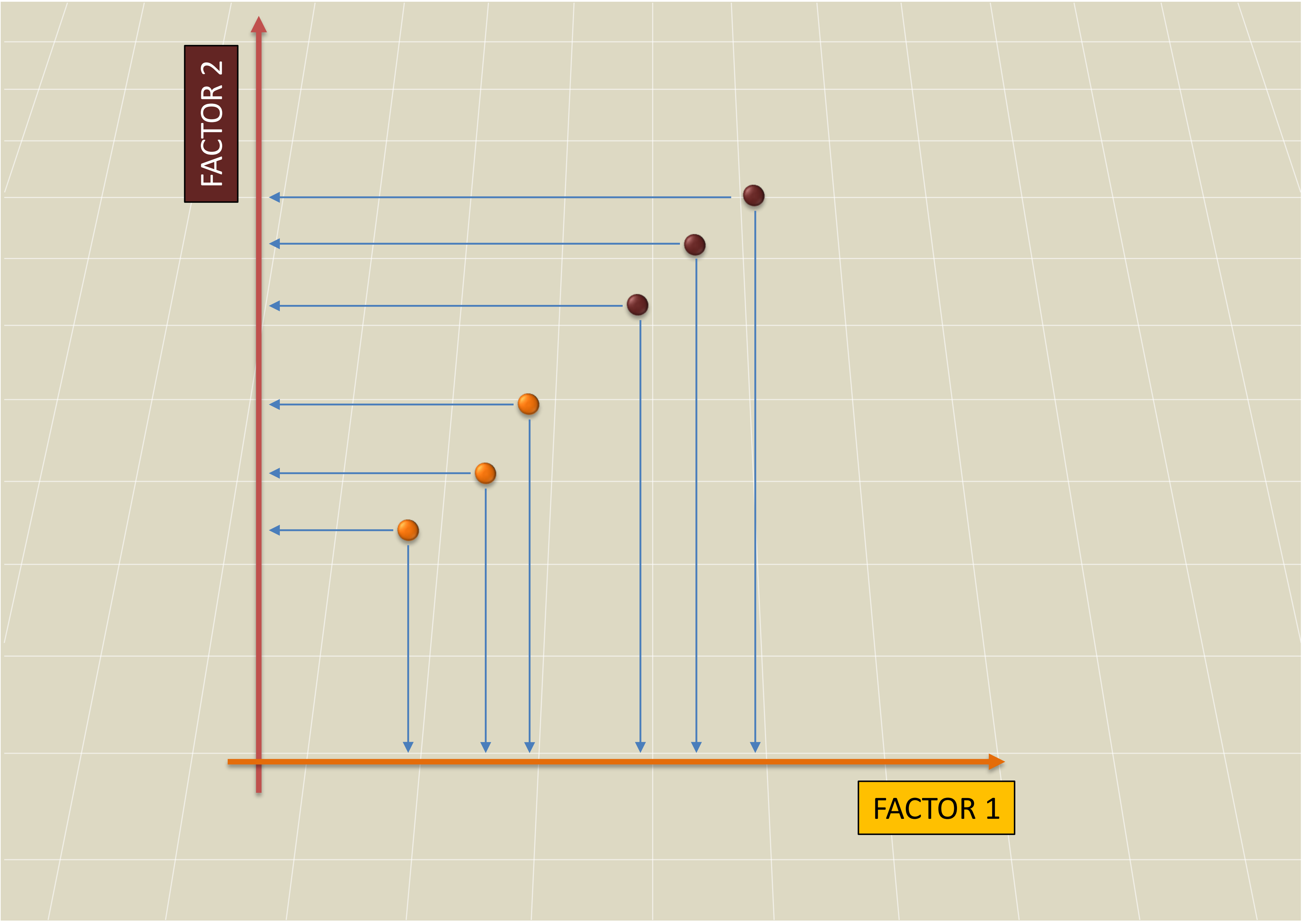

FACTOR 2
FACTOR 1


FACTOR 2
Lower Variance = Lower Discrimination
FACTOR 1


FACTOR 2
FACTOR 1


FACTOR 2
FACTOR 1


FACTOR 1'


FACTOR 2
FACTOR 1

MVDA - *EFA*

In order to explain an **EFA** approach, the following dataset will be used: The data of 1536 students (cases) of different levels of education were collected in 1976. The dataset has eight variables:

1. Moed (Mother's Education) – 1: $\leq 8th$ grade; 2: partial high school; 3: high school; 4: part college; 5: college; 6: post-graduate degree;
2. Faed (Father's Education) – 1: $\leq 8th$ grade; 2: partial high school; 3: high school; 4: part college; 5: college; 6: post-graduate degree;
3. Faminc (Family Income) – 1: $< \$5000$; 2: $\$5000 - \$7499$; 3: $\$7500 - \$9999$; 4: $\$10000 - \$14999$; 5: $\$15000 - \$19999$; 6: $\$20000 - \$24999$; 7: $\geq \$25000$;
4. English (English Test) – Metric;
5. Math (Math Test) – Metric;
6. SocSci (Social Science Test) – Metric;
7. NatSci (Natural Science Test) – Metric;
8. Vocab (Vocabulary Test) – Metric.

Assumptions

MVDA - *EFA*

ASSUMPTIONS

- Multivariate Normality: Statistics are improved if the dataset has Multivariate Normal Distribution. Relaxation – Most of the variables in the dataset are Normally Distributed (Univariate Normality).
- Linear Relationships Between Variables (Correlation): Statistics are improved if most variables have a linear relationship.
- Sample Size: It is recommend at least 200 cases (the more the better). Relaxation – 100 cases.
- Relationship Between Cases and Variables: $\frac{m}{p} \geq 20 \rightarrow$ at least 20 cases (m) by variable (p). Relaxation – $\frac{m}{p} \geq 5$

Assumptions - Multivariate Normality

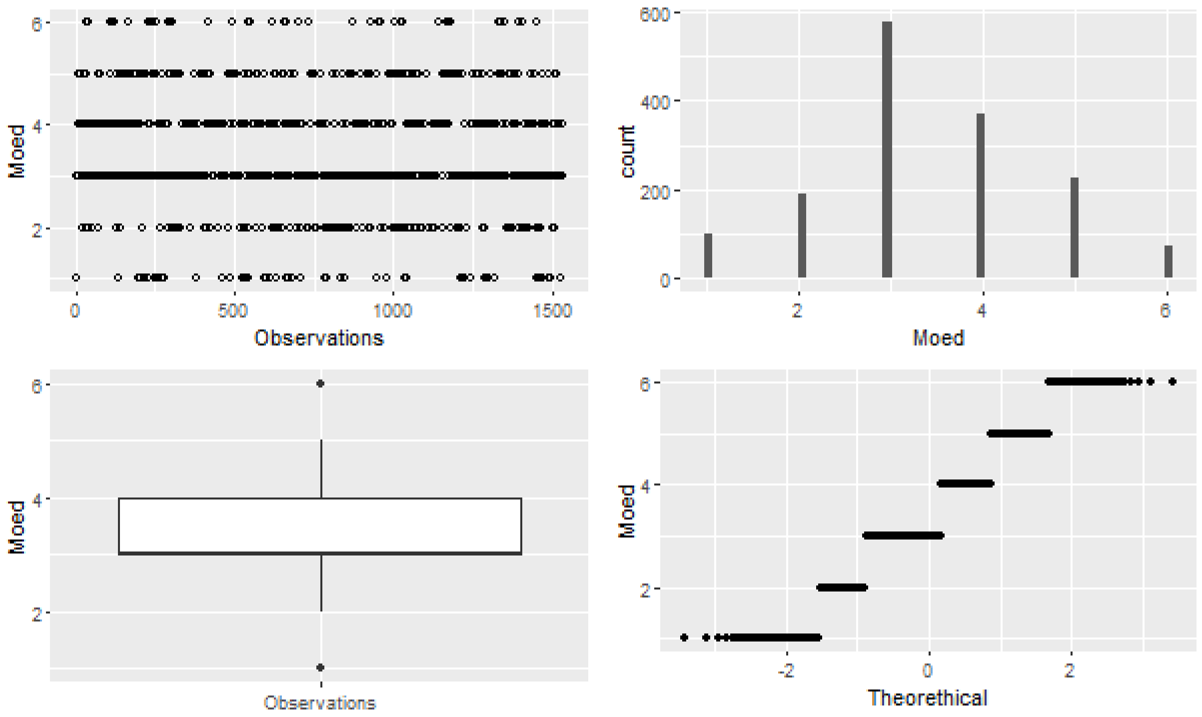
MVDA - *EFA*

```
# Normality Analysis
library(ggplot2)
Observations <- 1:1536
ggplot(data = my_data, aes(x = Observations, y = Moed)) + geom_point(shape = 1)
ggplot(data = my_data, aes(x = "", y = Moed)) + geom_boxplot() + theme(axis.title.x = element_blank())
ggplot(my_data, aes(Moed)) + geom_histogram(bins = 50)
ggplot(data = my_data, aes( sample = my_data[ ,1])) + stat_qq()+ xlab("Theorethical") + ylab("Moed")
ggplot(data = my_data, aes(x = Observations, y = Faed)) + geom_point(shape = 1)
ggplot(data = my_data, aes(x = "", y = Faed)) + geom_boxplot() + theme(axis.title.x = element_blank())
ggplot(my_data, aes(Faed)) + geom_histogram(bins = 50)
ggplot(data = my_data, aes( sample = my_data[ ,2])) + stat_qq()+ xlab("Theorethical") + ylab("Faed")
ggplot(data = my_data, aes(x = Observations, y = Famin)) + geom_point(shape = 1)
ggplot(data = my_data, aes(x = "", y = Famin)) + geom_boxplot() + theme(axis.title.x = element_blank())
ggplot(my_data, aes(Famin)) + geom_histogram(bins = 50)
ggplot(data = my_data, aes( sample = my_data[ ,3])) + stat_qq()+ xlab("Theorethical") + ylab("Famin")
ggplot(data = my_data, aes(x = Observations, y = Eng)) + geom_point(shape = 1)
ggplot(data = my_data, aes(x = "", y = Eng)) + geom_boxplot() + theme(axis.title.x = element_blank())
ggplot(my_data, aes(Eng)) + geom_histogram(bins = 50)
ggplot(data = my_data, aes( sample = my_data[ ,4])) + stat_qq()+ xlab("Theorethical") + ylab("Eng")
```

MVDA - *EFA*

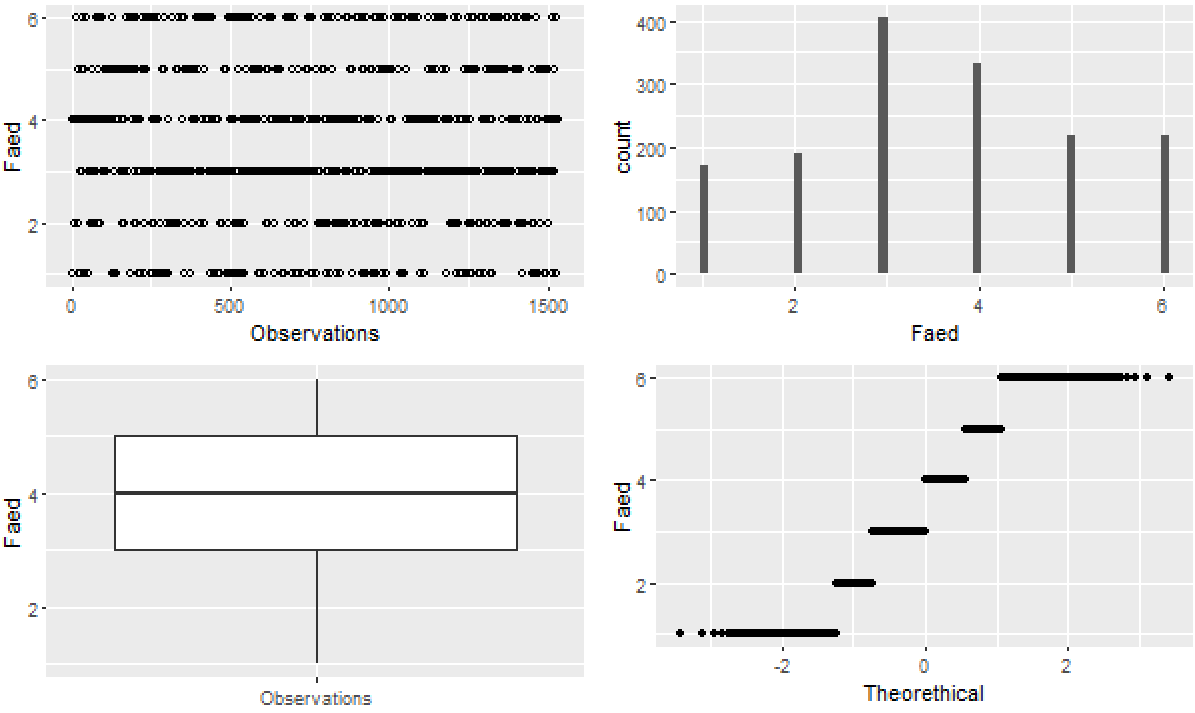
```
ggplot(data = my_data, aes(x = Observations, y = Math)) + geom_point(shape = 1)
ggplot(data = my_data, aes(x = "", y = Math)) + geom_boxplot() + theme(axis.title.x = element_blank())
ggplot(my_data, aes(Math)) + geom_histogram(bins = 50)
ggplot(data = my_data, aes( sample = my_data[ ,5])) + stat_qq()+ xlab("Theorethical") + ylab("Math")
ggplot(data = my_data, aes(x = Observations, y = Soc)) + geom_point(shape = 1)
ggplot(data = my_data, aes(x = "", y = Soc)) + geom_boxplot() + theme(axis.title.x = element_blank())
ggplot(my_data, aes(Soc)) + geom_histogram(bins = 50)
ggplot(data = my_data, aes( sample = my_data[ ,6])) + stat_qq()+ xlab("Theorethical") + ylab("Soc")
ggplot(data = my_data, aes(x = Observations, y = Nat)) + geom_point(shape = 1)
ggplot(data = my_data, aes(x = "", y = Nat)) + geom_boxplot() + theme(axis.title.x = element_blank())
ggplot(my_data, aes(Nat)) + geom_histogram(bins = 50)
ggplot(data = my_data, aes( sample = my_data[ ,7])) + stat_qq()+ xlab("Theorethical") + ylab("Nat")
ggplot(data = my_data, aes(x = Observations, y = Vocab)) + geom_point(shape = 1)
ggplot(data = my_data, aes(x = "", y = Vocab)) + geom_boxplot() + theme(axis.title.x = element_blank())
ggplot(my_data, aes(Vocab)) + geom_histogram(bins = 50)
ggplot(data = my_data, aes( sample = my_data[ ,8])) + stat_qq()+ xlab("Theorethical") + ylab("Vocab")
```

MVDA - *EFA*



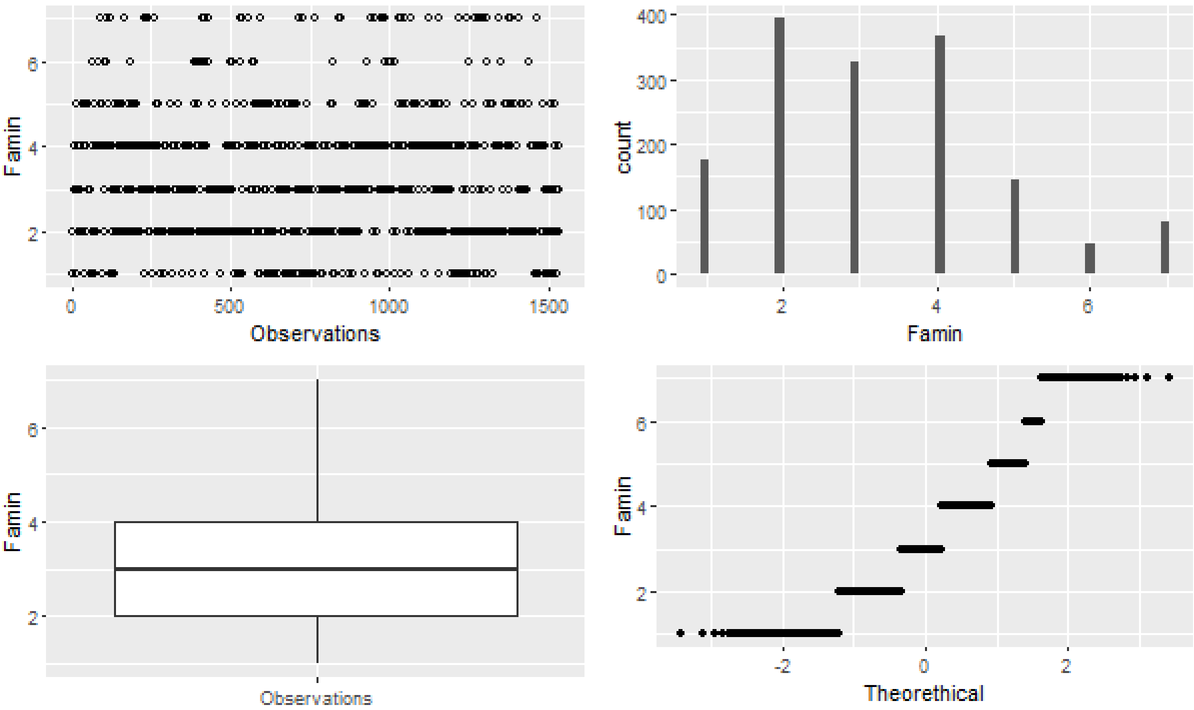
Moed
Observations
0
500
1000
1500
2
4
6
count
0
200
400
600
Moed
Observations
Theorethical
-2
0
2

MVDA - *EFA*

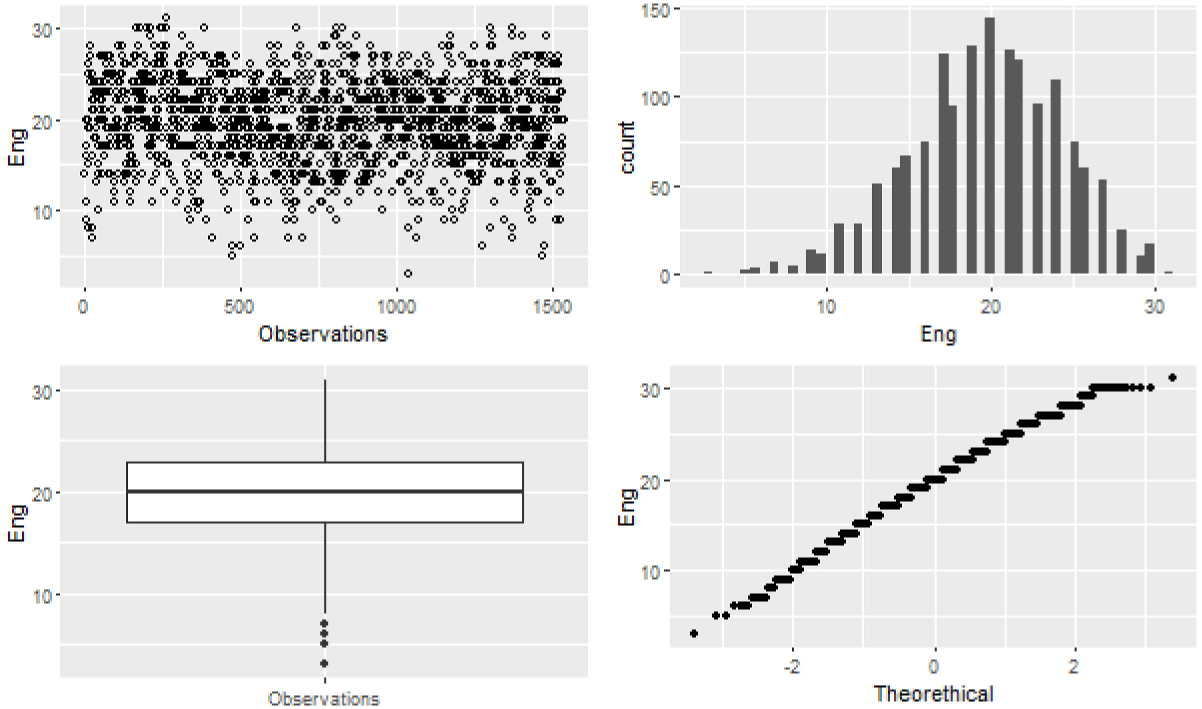


Faed
6
4
2
0
500
1000
1500
Observations
count
400
300
200
100
0
2
4
6
Faed
Faed
6
4
2
Observations
Faed
6
4
2
-2
0
2
Theorethical

MVDA - *EFA*

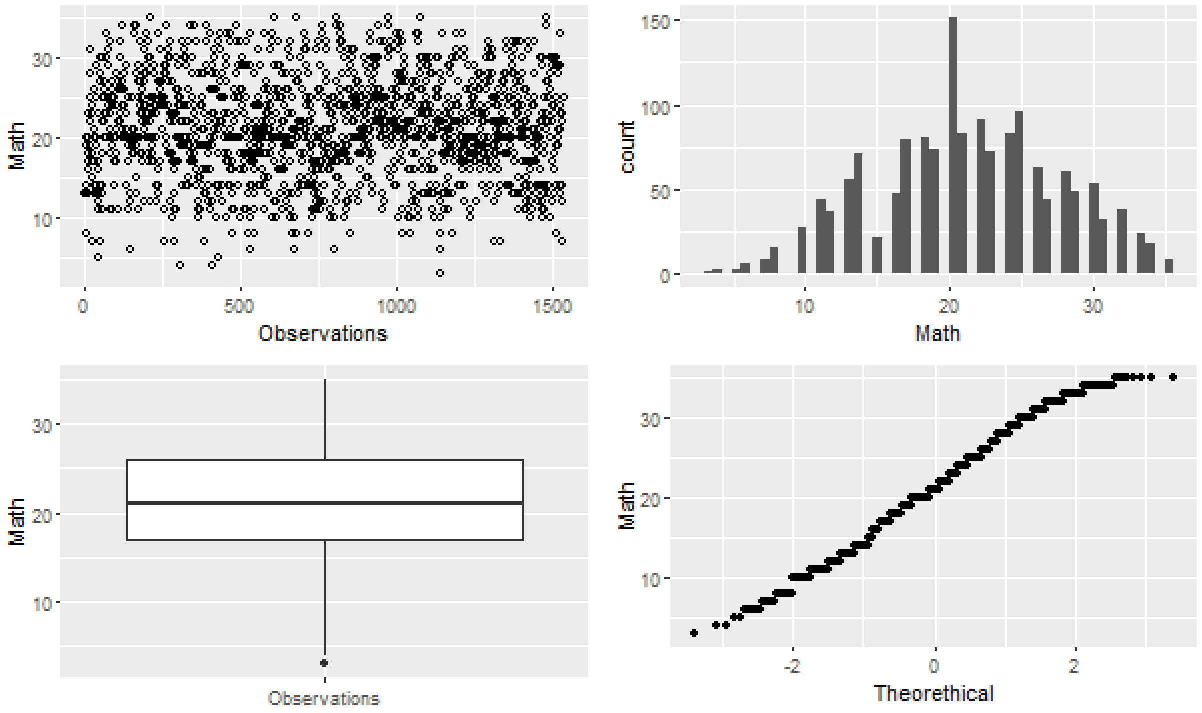

Famin
Observations
0
500
1000
1500
count
400
300
200
100
Theorethical
-2
2
4
6

MVDA - *EFA*

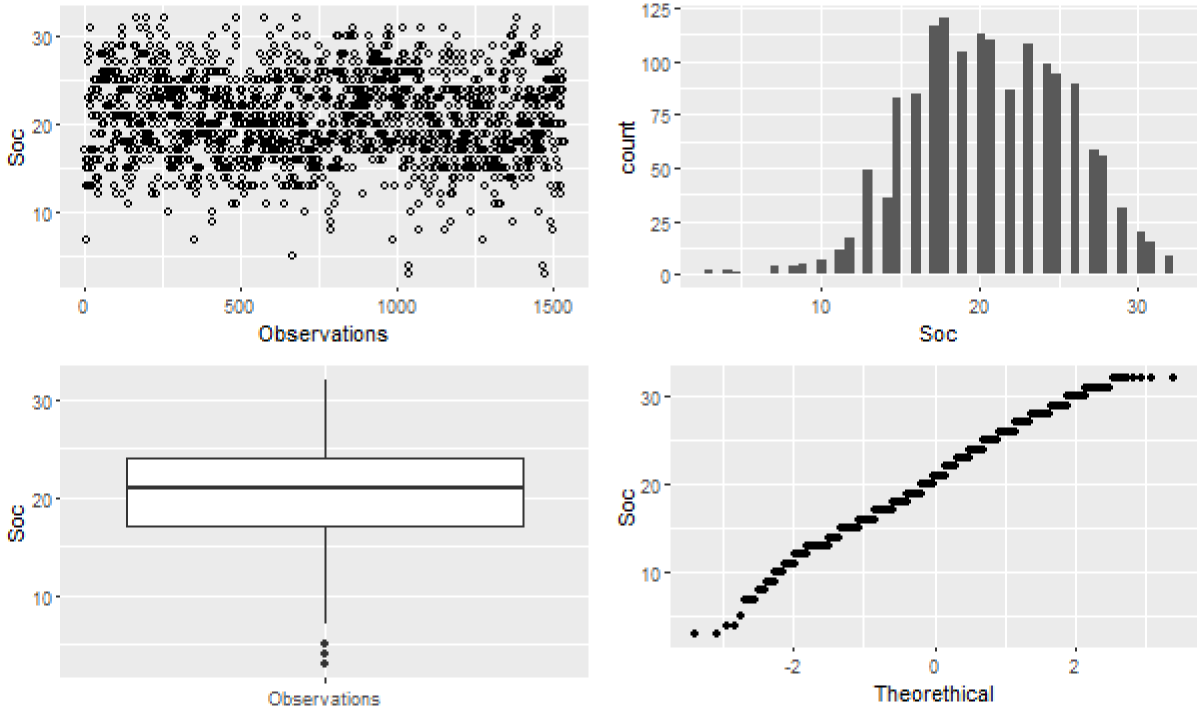


30
20
10
Eng
0
500
1000
1500
Observations
150
100
50
0
count
10
20
30
Eng
30
20
10
Eng
Observations
30
20
10
Eng
-2
0
2
Theorethical

MVDA - *EFA*

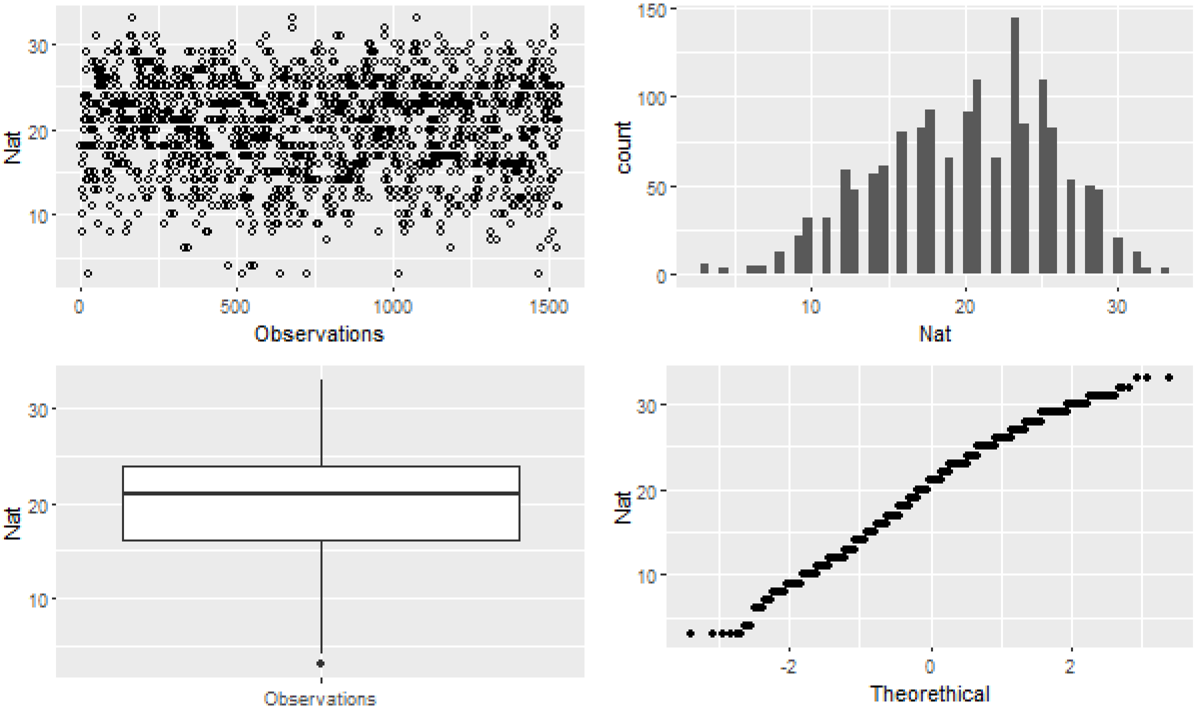

Math
Observations
count
Math
Math
Observations
Math
Theorethical

MVDA - *EFA*



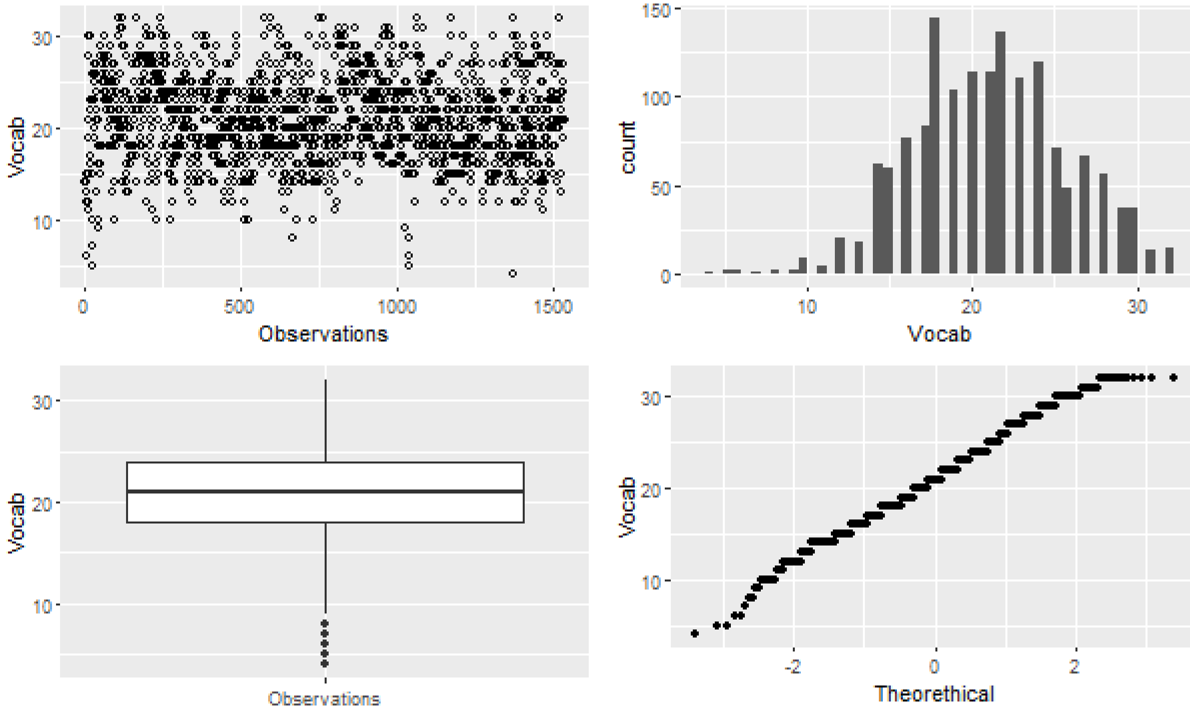
Soc
Observations
0
500
1000
1500
10
20
30
count
125
100
75
50
25
0
Soc
Observations
Theorethical
-2
0
2

MVDA - *EFA*



Nat
30
20
10
0
500
1000
1500
Observations
count
150
100
50
0
10
20
30
Nat
Nat
30
20
10
Observations
Nat
30
20
10
-2
0
2
Theorethical

MVDA - *EFA*



Vocab
30
20
10
0
500
1000
1500
Observations
count
150
100
50
0
10
20
30
Vocab
Vocab
30
20
10
Observations
Vocab
30
20
10
-2
0
2
Theorethical

MVDA - *EFA*

```
# Univariate Normality
shapiro.test(my_data$Moed)
p-value < 2.2e-16

shapiro.test(my_data$Faed)
p-value < 2.2e-16

shapiro.test(my_data$Famin)
p-value < 2.2e-16

shapiro.test(my_data$Eng)
p-value = 4.563e-09

shapiro.test(my_data$Math)
p-value = 8.405e-09

shapiro.test(my_data$Soc)
p-value = 6.692e-09

shapiro.test(my_data$Nat)
p-value = 1.485e-12

shapiro.test(my_data$Vocab)
p-value = 4.172e-08
```

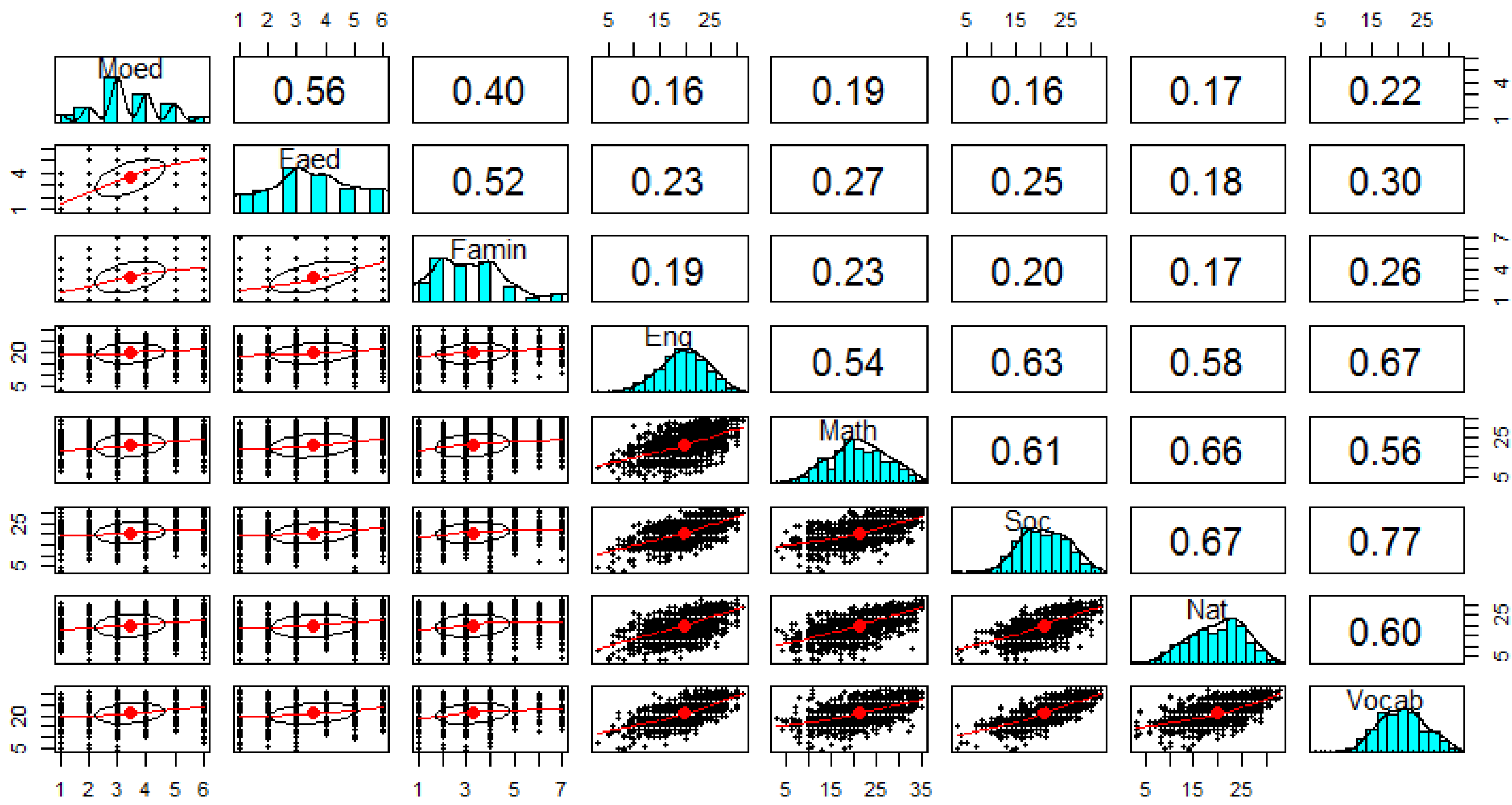
MVDA - *EFA*

```
# Multivariate Normality
library(MVN)
mardiaTest(my_data, qqplot = FALSE)

 Mardia's Multivariate Normality Test
---------------------------------------
   data : my_data

   g1p            : 2.297712
   chi.skew       : 588.2143
   p.value.skew   : 7.370124e-63

   g2p            : 80.91948
   z.kurtosis     : 1.424451
   p.value.kurt   : 0.1543158

   chi.small.skew : 589.6189
   p.value.small  : 4.200743e-63

   Result         : Data are not multivariate normal.
---------------------------------------
```

Assumptions - Linearity

MVDA - *EFA*

```
# Linearity
library (psych)
pairs.panels(my_data)
```



Assumptions - Sample Size and Ratio

MVDA - *EFA*

```
# Sample Size
nrow(my_data)
[1] 1536

# Ratio
nrow(my_data)/ncol(my_data)
[1] 192
```

Measures of Sampling Adequacy

MVDA - *EFA*

- The determinant of the correlation matrix indicates if the array can be rotated, thus allowing the application of the technique. The value of 0 indicates that the matrix can not be rotated.
- The diagonal of the Anti-Image Correlation Matrix, shows values between 0 and 1, that indicate the degree of adjustment of each variable to factor analysis. Variable with values below 0.5 should be excluded because they do not have a significant amount of common variance.
- The Kaiser-Meyer-Olkin (KMO) test is an indicator that varies between 0 (no adequacy) and 1 (perfect adequacy), and it shows the proportion of variance shared by all variables due to common factors. A test with a value below 0.5 indicates that the data set is not suitable for the EFA analysis.
- The Bartlett's sphericity is based on a chi-squared statistical distribution and it tests the following hypothesis:

$$\mathrm{H_0}: \text{ The correlation matrix is an identity matrix (no correlation among the variables).}$$
$$\mathrm{H_1}: \text{ The correlation matrix is not an identity matrix (there is correlation among the variables).}$$

MVDA - *EFA*

KMO	Interpretation
$0.9 < kmo \leq 1.0$	Excellent
$0.8 < kmo \leq 0.9$	Very Good
$0.7 < kmo \leq 0.8$	Good
$0.6 < kmo \leq 0.7$	Fair
$0.5 < kmo \leq 0.6$	Poor
$kmo \leq 0.5$	Inacceptable

Measures of Sampling Adequacy – Determinant

MVDA - *EFA*

```
# Correlation matrix
c_mat <- cor(my_data)
```

	Moed	Faed	Famin	Eng	Math	Soc	Nat	Vocab
Moed	1.0000000	0.5582127	0.4010453	0.1611124	0.1885132	0.1619355	0.1713547	0.2222116
Faed	0.5582127	1.0000000	0.5230169	0.2299847	0.2672347	0.2470533	0.1848430	0.2999410
Famin	0.4010453	0.5230169	1.0000000	0.1903187	0.2335749	0.1957009	0.1652116	0.2569234
Eng	0.1611124	0.2299847	0.1903187	1.0000000	0.5412699	0.6339606	0.5783527	0.6692687
Math	0.1885132	0.2672347	0.2335749	0.5412699	1.0000000	0.6086261	0.6583328	0.5556737
Soc	0.1619355	0.2470533	0.1957009	0.6339606	0.6086261	1.0000000	0.6749258	0.7719540
Nat	0.1713547	0.1848430	0.1652116	0.5783527	0.6583328	0.6749258	1.0000000	0.6004781
Vocab	0.2222116	0.2999410	0.2569234	0.6692687	0.5556737	0.7719540	0.6004781	1.0000000

```
# Determinant of the correlation matrix
det(c_mat)
[1] 0.0223
```

Measures of Sampling Adequacy – Others

MVDA - *EFA*

```
# KMO and Diagonal
library(psych)
KMO(cor(my_data))

Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor(my_data))
Overall MSA =  0.84
MSA for each item =
Moed  Faed  Famin  Eng  Math   Soc  Nat  Vocab
  0.74    0.73     0.80   0.91   0.89  0.84  0.86     0.84

# Barlett Sphericity Test
library(psych)
cortest.bartlett(cor(my_data), n = nrow(my_data))
$chisq
[1] 5823.107

$p.value
[1] 0

$df
[1] 28
```

Number of Factors

MVDA - *EFA*

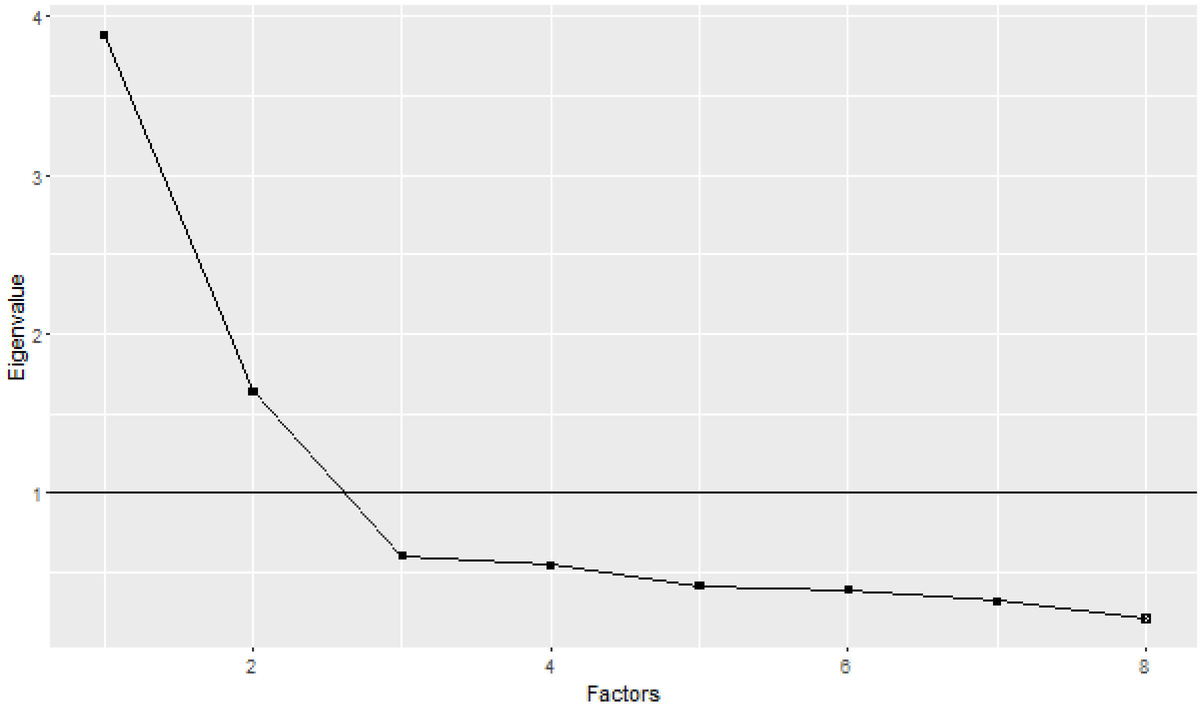
Scree Plot – A visual method that helps decide the amount of factors to be extracted. that employs the plot of the eigenvalues against the order factor extraction. The obtained curve indicates the number of factors to be extracted, using the following criterion:

- Kaiser Criterion: Extracts only the factors that have eigenvalues above 1. Not recommended.

- Elbown Criterion: Draw a straight line fitting the smallest eigenvalues. The point where a departure from this line occurs indicates the number of factors to be retained.

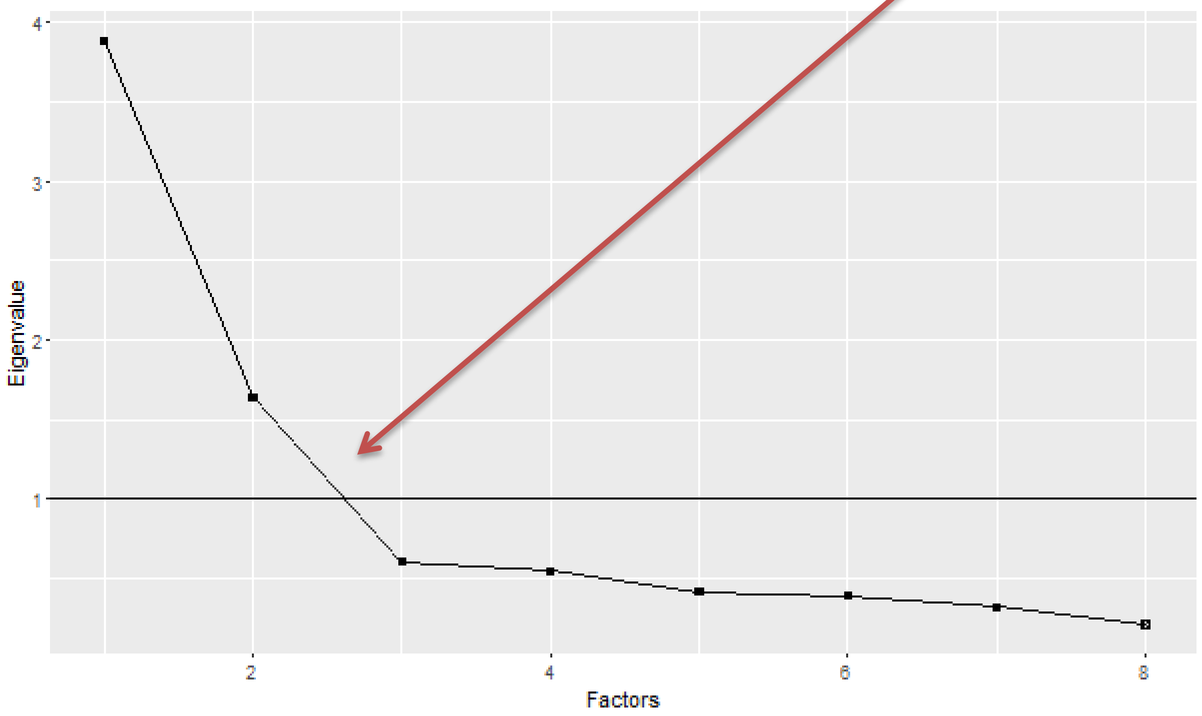
```
# Scree Plot
scree <- as.data.frame(eigen(cor(my_data))$values)
ggplot(data = scree, aes( x = 1:ncol(my_data), y = scree[,1])) + geom_point(shape = 7, size = 1.2) + geom_line() +
geom_hline(yintercept = 1) + xlab("Factors") + ylab("Eigenvalue")
```

MVDA - *EFA*

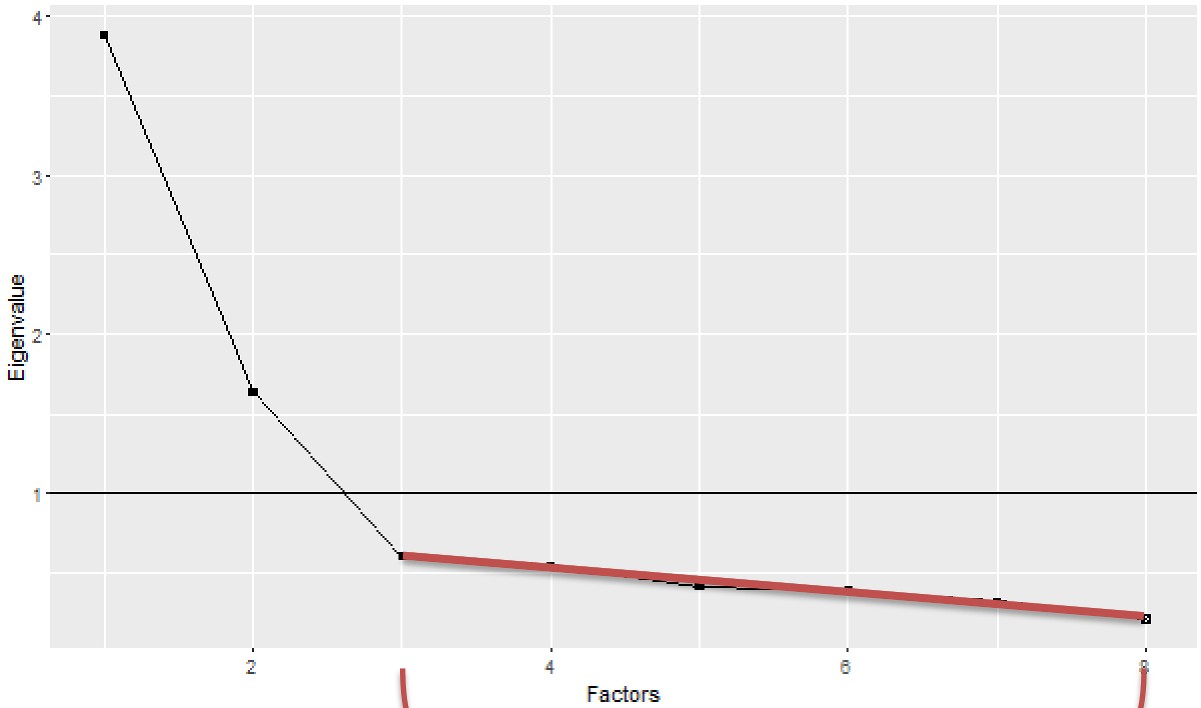


MVDA - *EFA*

Kaiser Criterion = 2 Factors



MVDA - *EFA*



Elbow Criterion = 2 Factors

Extraction

MVDA - *EFA*

- Weighted Least Squares: This is one of the extraction method that minimizes the sum of squared differences between the data matrix and reproduced correlation matrix, ignoring the diagonals.
- Generalized Least Squares: Same as above, but in this case the correlation is weighed by the inverse of their singularities (perfect correlation between two variables) and variables with high singularities are made with less weight than those with lower singularities.
- Maximum Likelihood: This method estimate parameters that will most likely to reproduce the original correlation matrix. The sample must have a multivariate normal distribution.
- Principal Axis Factoring: In the principal axis factoring method, the initial estimate of common variance assumes that the communality of each variable is equal to the square multiple regression coefficient of that variable with respect to the other variables. The principal axis factoring method is implemented by replacing the main diagonal of the correlation matrix by these initial estimates of the communalities. An algorithm is repeated until a predefined maximum number of iterations are performed or the communalities converge.

Rotation

MVDA - *EFA*

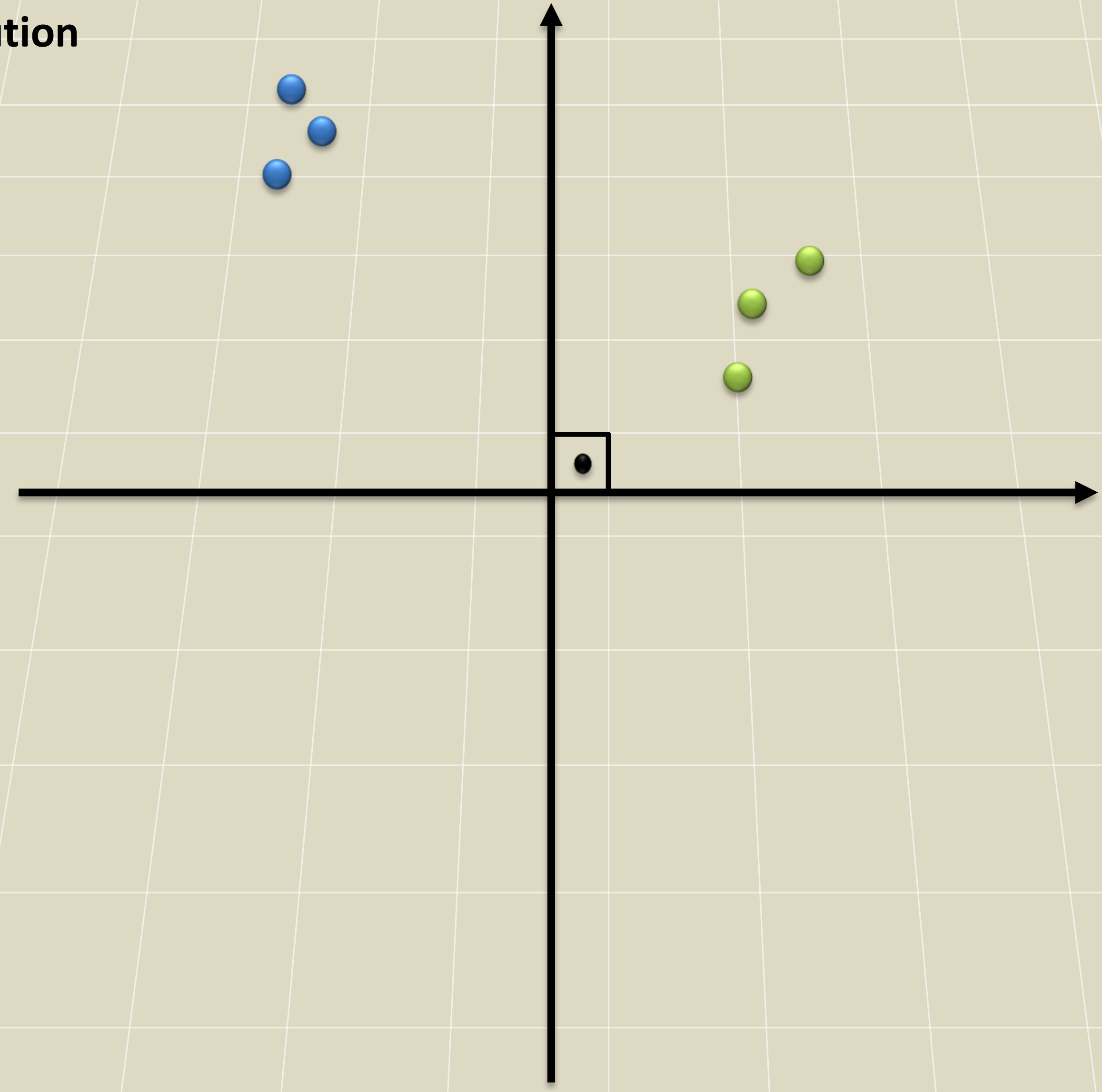
The interpretability of factors can be improved by rotation methods and two major approaches are available:

- **Orthogonal**: the factors are maintained uncorrelated;

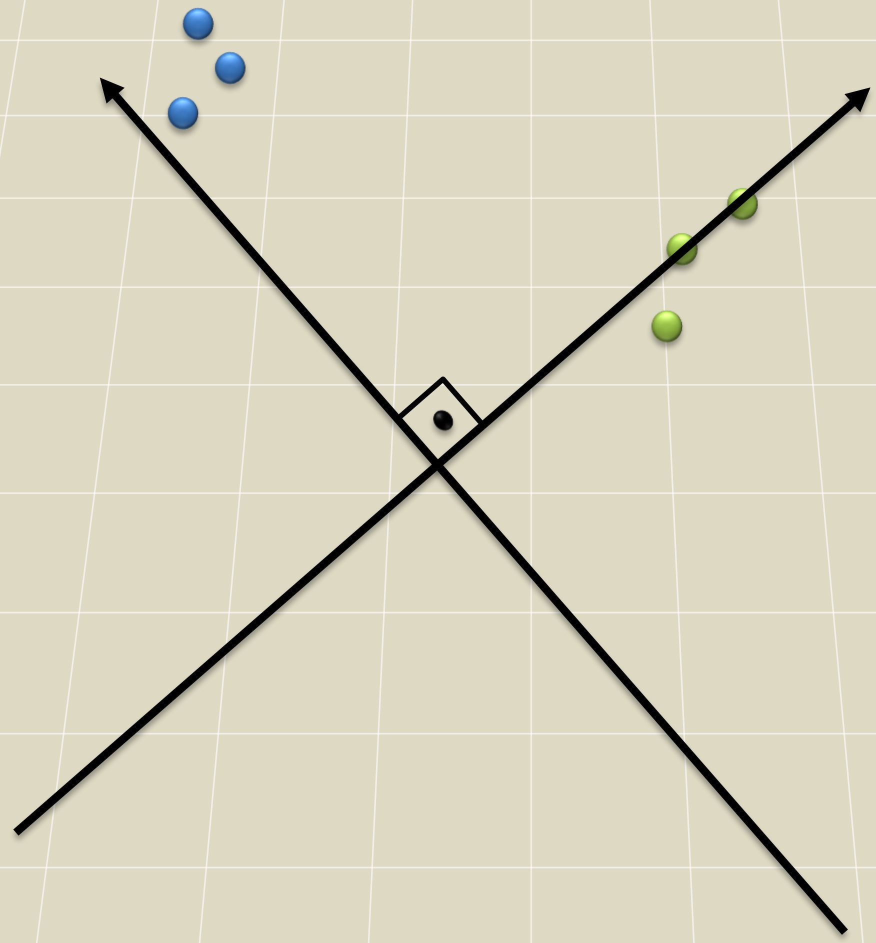
- **Oblique**: the factors can be correlated.

The most used orthogonal rotation is the varimax method, and the most used oblique rotation is the direct oblimin method.

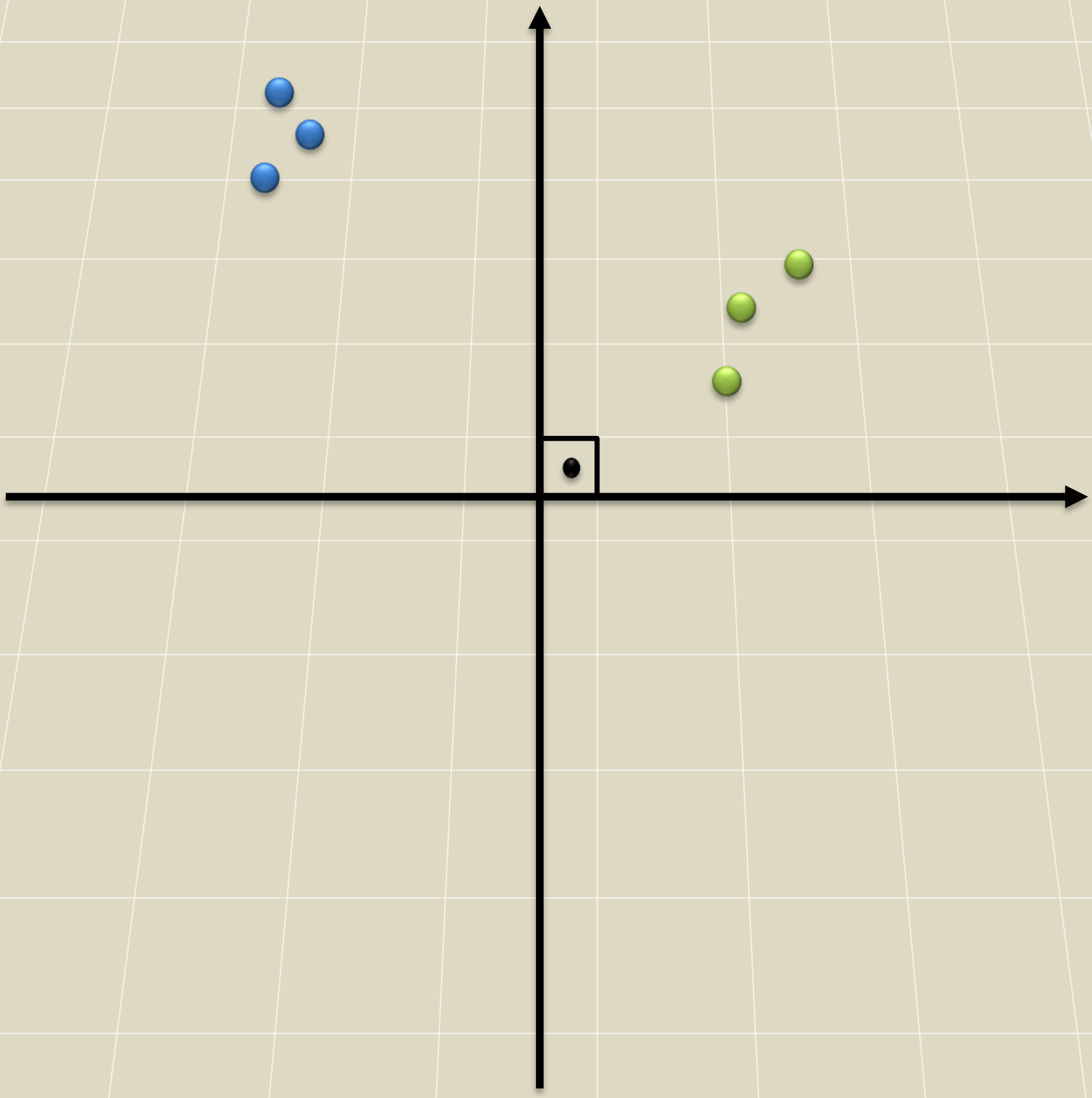
Unrotated Solution



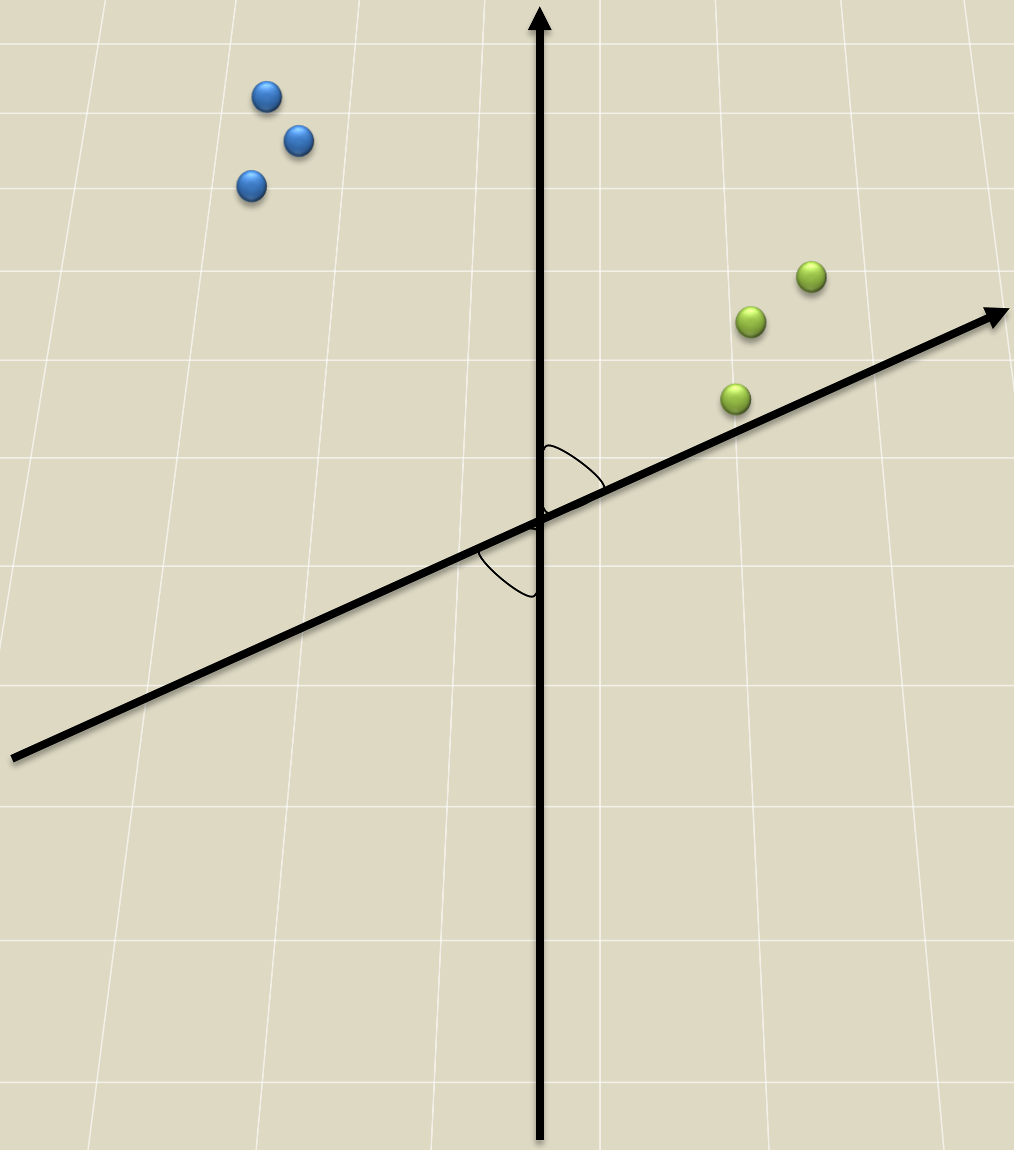
Orthogonal Rotation



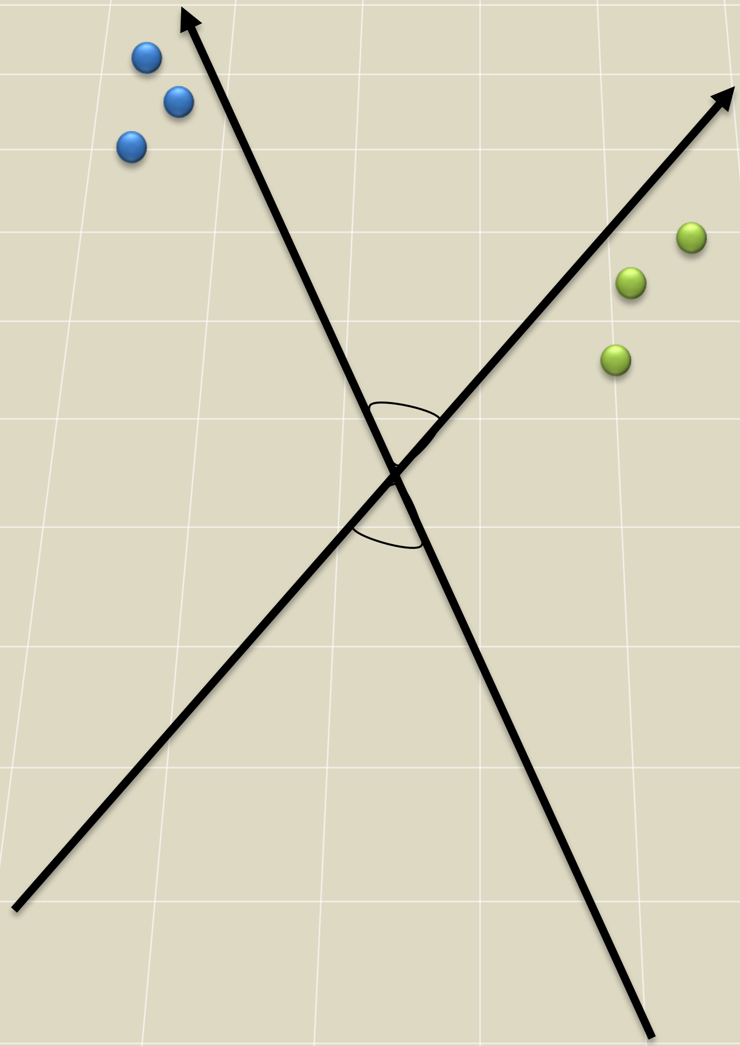
Unrotated



Oblique Rotation



Oblique Rotation



MVDA - *EFA*

- **Varimax**: It is an orthogonal rotation method that minimizes the number of variables for each factor. It simplifies the interpretation of the factors.
- **Quartimax**: It is an orthogonal method that minimizes the number of factors needed to explain each variable. It simplifies the interpretation of the variables.
- **Equamax**: It is an orthogonal method that seeks a combination of both varimax and quartimax methods. The variables will have factor loadings with higher values and the number of factors is minimized.
- **Oblimin**: It is an oblique method of rotation that increases the correlation between factors.
- **Promax**: It is an oblique method of rotation with an algorithm that finds rotated solutions faster than the oblimin rotation.

Factors

MVDA - *EFA*

```
# Extraction and Rotation
library(psych)
fa_obl <- fa(my_data, nfactors = 2, rotate = "oblimin", fm = "pa")
fa_var <- fa(my_data, nfactors = 2, rotate = "varimax", fm = "pa")

# Factor Loadings (Correlation between a variable and a factor, values greater than |0.3| should be interpretated)
fa_obl$loadings
```

	PA1	PA2
Moed		0.660
Faed		0.849
Famin		0.604
Eng	0.753	
Math	0.709	
Soc	0.883	
Nat	0.805	
Vocab	0.807	

	PA1	PA2
SS loadings	3.151	1.533
Proportion Var	0.394	0.192
Cumulative Var	0.394	0.585

MVDA - *EFA*

Factor Loadings
fa_var$loadings

	PA1	PA2
Moed	0.111	0.645
Faed	0.159	0.833
Famin	0.154	0.599
Eng	0.741	0.132
Math	0.709	0.182
Soc	0.862	0.119
Nat	0.783	
Vocab	0.807	0.208

	PA1	PA2
SS loadings	3.151	1.585
Proportion Var	0.390	0.198
Cumulative Var	0.390	0.588

MVDA - *EFA*

Factor Loadings
fa_var$loadings

	PA1	PA2
Moed	0.111	0.645
Faed	0.159	0.833
Famin	0.154	0.599
Eng	0.741	0.132
Math	0.709	0.182
Soc	0.862	0.119
Nat	0.783	
Vocab	0.807	0.208

	PA1	PA2
SS loadings	3.151	1.585
Proportion Var	0.390	0.198
Cumulative Var	0.390	0.588

$$Comunality \rightarrow h(Moed)^2 = (0{,}111)^2 + (0{,}645)^2 = \ 0{,}428$$

MVDA - *EFA*

Communality (Proportion of variance for each variable that can be explained by the extracted factors)

fa_obl$communality

Moed	Faed	Famin	Eng	Math	Soc	Nat	Vocab
0.4287644	0.7189107	0.3831450	0.5667775	0.5352446	0.7573750	0.6214141	0.6944233

fa_var$communality

Moed	Faed	Famin	Eng	Math	Soc	Nat	Vocab
0.4287644	0.7189107	0.3831450	0.5667775	0.5352446	0.7573750	0.6214141	0.6944233

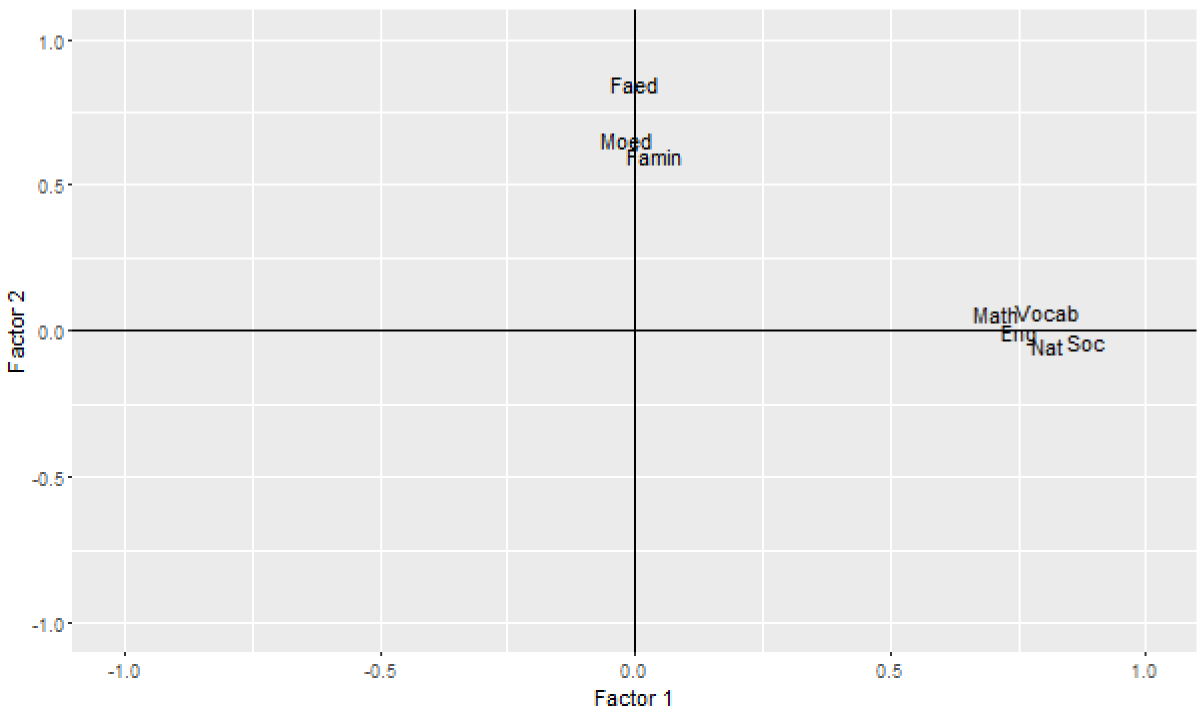
Factor Correlation (Correlations greater than 0.3 are significant)

fa_obl$Phi

	PA1	PA2
PA1	1.0000000	0.3603697
PA2	0.3603697	1.0000000

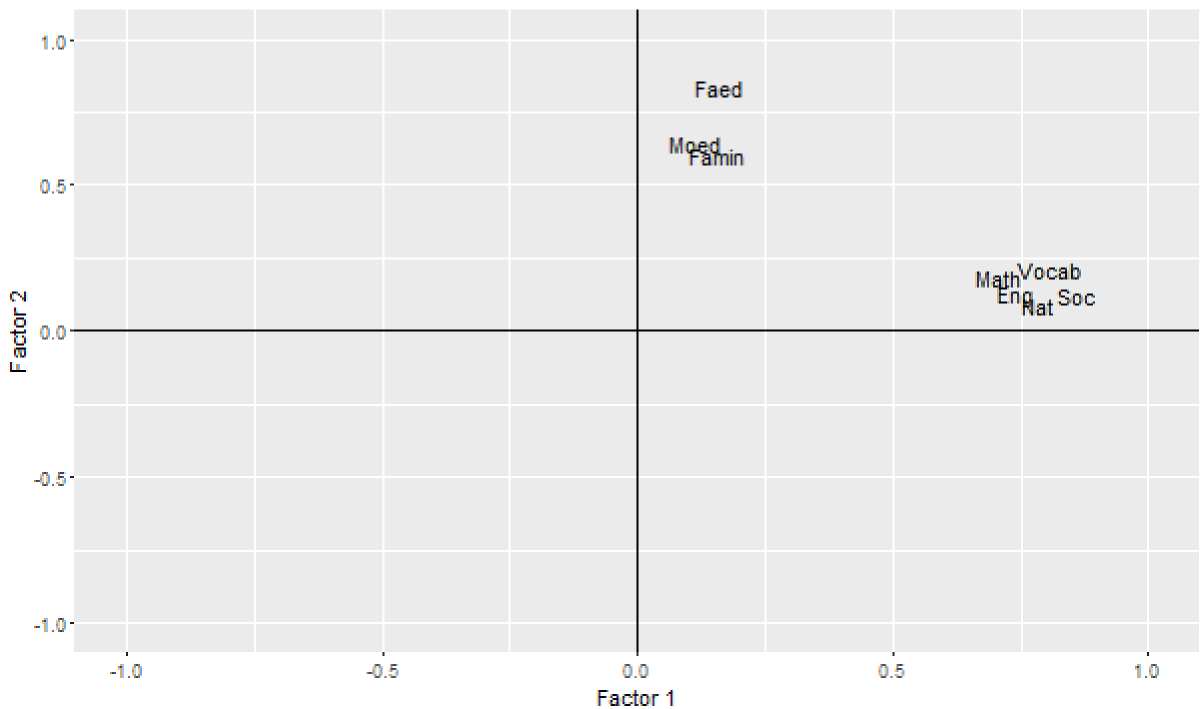
MVDA - *EFA*

```
# Factor Loadings Plot
L1_obl <- fa_obl$loadings[1:ncol(my_data)]
L2_obl <- fa_obl$loadings[(ncol(my_data) + 1):(2*ncol(my_data))]
FL_obl <- as.data.frame(cbind(L1_obl,L2_obl))
ggplot(data = FL_obl, aes( x = L1_obl, y = L2_obl, label = colnames(my_data))) + geom_text() + xlim(c(-1,1)) +
ylim(c(-1,1)) + geom_vline(xintercept = 0) + geom_hline(yintercept = 0) + xlab("Factor 1") + ylab("Factor 2")
```



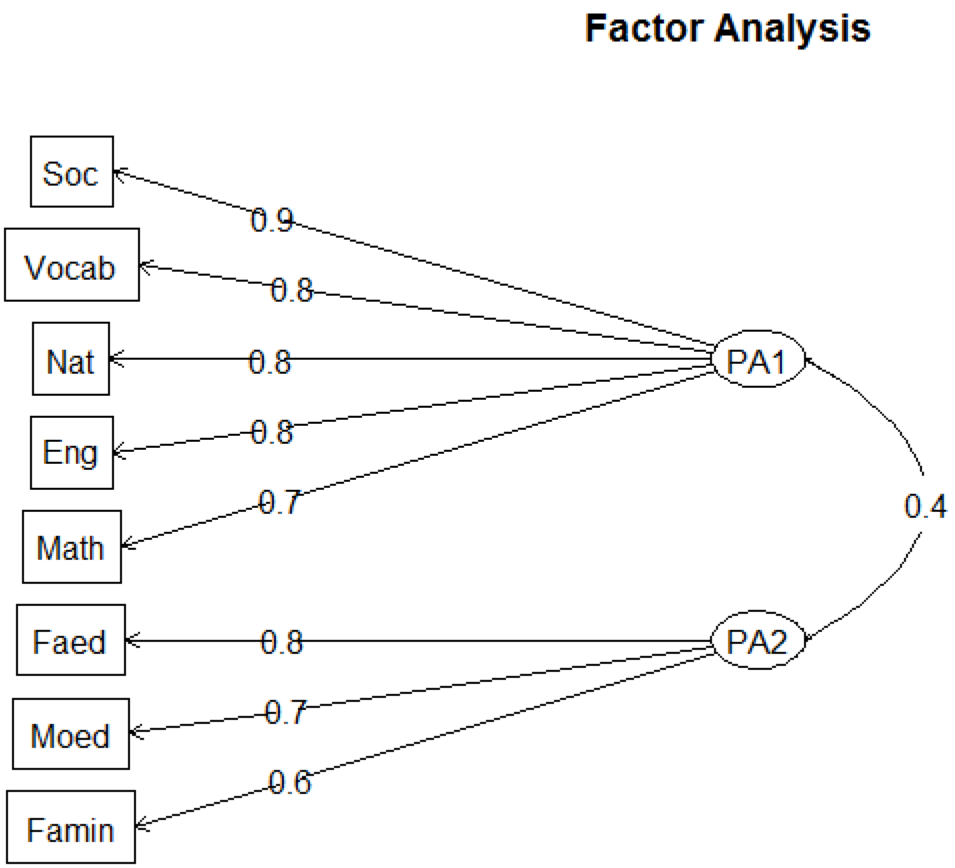
MVDA - *EFA*

```
# Factor Loadings Plot
L1_var <- fa_var$loadings[1:ncol(my_data)]
L2_var <- fa_var$loadings[(ncol(my_data) + 1):(2*ncol(my_data))]
FL_var <- as.data.frame(cbind(L1_var,L2_var))
ggplot(data = FL_var, aes( x = L1_var, y = L2_var, label = colnames(my_data))) + geom_text() + xlim(c(-1,1)) +
ylim(c(-1,1)) + geom_vline(xintercept = 0) + geom_hline(yintercept = 0) + xlab("Factor 1") + ylab("Factor 2")
```



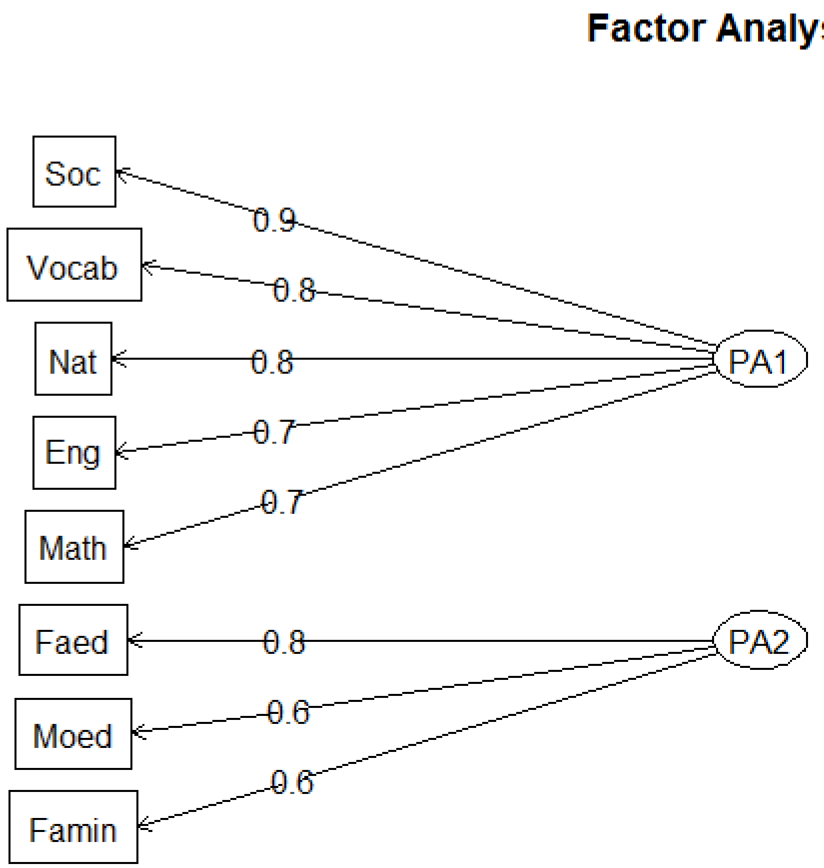
MVDA - *EFA*

```
# Factor Plot
fa.diagram(fa_obl)
```

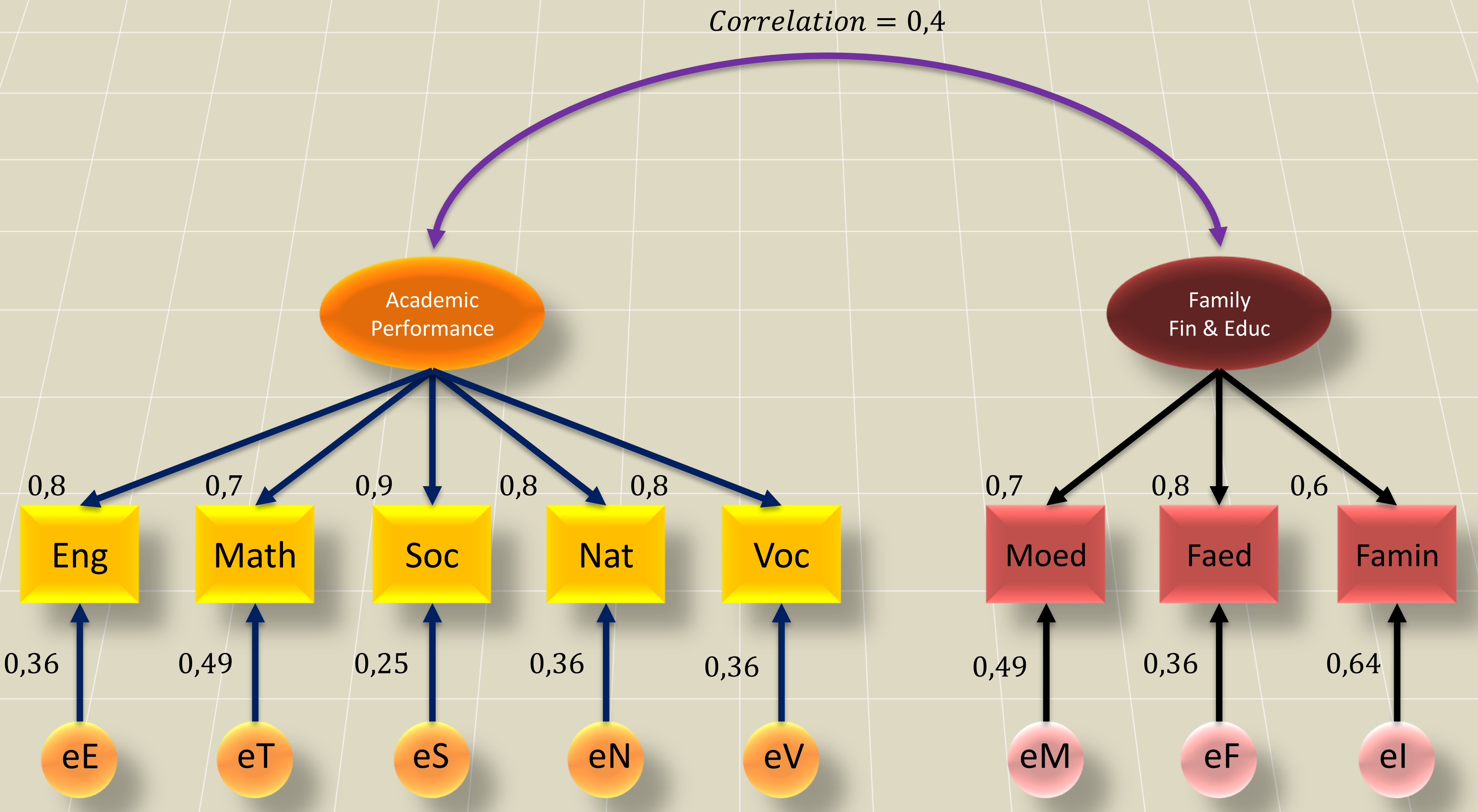


MVDA - *EFA*

```
# Factor Plot
fa.diagram(fa_var)
```



Interpretation



$$Unique\ Variance_{Eng\ eE} = 1 - (0{,}8)^2 = 0{,}36$$

MVDA

https://github.com/Valdecy/Multivariate_Data_Analysis

##
##

Created by: Prof. Valdecy Pereira, D.Sc.
UFF - Universidade Federal Fluminense (Brazil)
email: valdecypereira@yahoo.com.br
Course: Multivariate Data Analysis
Lesson: Exploratory Factor Analysis

Citation:
PEREIRA, V. (2016). Project: Multivariate Data Analysis, File: R-MVDA-03-EFA.pdf, GitHub repository: <https://github.com/Valdecy/Multivariate_Data_Analysis>

##
##

Bibliography

CORRAR, L.J.; PAULO, E.; DIAS FILHO, J. M. **Análise Multivariada para Cursos de Administração, Ciências Contábeis e Economia**. ATLAS, 2009.

FÁVERO, L. P.; BELFIORE, P.; SILVA, F. L.; CHAN, B. **Análise de Dados: Modelagem Multivariada para Tomada de Decisões**. CAMPUS, 2009.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise Multivariada de Dados**. BOOKMAN, 2009.

LATTIN, J.; CARROLL, J. D.; GREEN, P. E. **Análise de Dados Multivariados**. CENGAGE Learning, 2011.

LEVINE, D. M.; STEPHAN, D. F.; KREHBIEL, T. C.; BERENSON, M. L. **Estatística - Teoria e Aplicações - Usando Microsoft Excel**. LTC, 2012.