People's Democratic Republic of Algeria
University Med Khider of Biskra
Faculty of SNVSTU

Principal Component Analysis (PCA) and Factorial Correspondence Analysis (FCA)

Dr. Ben Gherbal Hanane
Biostatistics – Level: M1 Biology
Email: hanane.bengherbal@univ-biskra.dz

1 Introduction

In biological studies, researchers often collect a large amount of data on individuals, samples, or different species. These data may be:

- **quantitative**, such as height, weight, biomass, or protein content,
- or **qualitative**, such as diet type, variety, health status, or ecological environment.

When the number of variables becomes large, it is difficult to interpret the relationships among them directly. For this reason, some multivariate statistical methods are used to summarize the information and reveal the general structure of the data. Among the most important of these methods are:

- **PCA**: Principal Component Analysis,
- **FCA**: Factorial Correspondence Analysis.

2 Principal Component Analysis (PCA)

2.1 Definition

PCA is a statistical method used when the variables are **quantitative**. Its purpose is to reduce the number of original variables to a smaller number of new axes called **principal components**, while preserving as much information as possible.

2.2 Objective of PCA

PCA is used in order to:

- reduce the number of variables,
- study the relationships among quantitative variables,
- represent individuals and variables graphically,
- detect similar groups of individuals.

2.3 Principle

PCA constructs new axes:

- the first axis explains the largest proportion of the variance,
- the second axis explains the largest possible proportion of the remaining variance,
- and so on.

Thus, the data can be represented in two or three dimensions instead of a large number of variables.

2.4 Biological Example

Suppose we study a group of plants, and for each plant we measure:

- stem height,
- number of leaves,
- root length,
- dry mass.

These are quantitative variables. By using PCA, we may observe, for example, that:

- taller plants generally have a greater number of leaves,
- plants with longer roots have a greater dry mass,
- the individuals are divided into two groups: weakly growing plants and strongly growing plants.

In this case, the first axis may represent the **overall growth of the plant**.

2.5 When is PCA used?

PCA is used when:

- the variables are quantitative,
- we want to study the correlations among them,
- we seek to summarize the structure of the data.

3 Factorial Correspondence Analysis (FCA)

3.1 Definition

FCA is a statistical method mainly used when the data are presented in the form of a **contingency table** containing frequencies or proportions, that is, when we study the relationship between **two qualitative variables**.

3.2 Objective of FCA

FCA is used in order to:

- study the relationship between the rows and columns of a contingency table,
- represent qualitative categories graphically,
- highlight the proximity or distance between categories,
- facilitate the interpretation of large tables.

3.3 Principle

FCA is based on a correspondence table, for example:

- animal species $\times$ food types,
- plant species $\times$ soil types,
- bacterial strains $\times$ culture media.

Then, both rows and columns are represented on factorial axes that make their relationships easier to visualize.

3.4 Biological Example

Suppose we study the relationship between **insect type** and **preferred plant**, and we obtain a frequency table showing the number of times each type of insect is found on each type of plant.

By using FCA, we may discover, for example, that:

- a certain insect type is more strongly associated with certain plants,
- some plants attract the same types of insects,
- there are groups of plant or insect types that are close to one another.

3.5 Another Example

FCA may also be used to study:

- **bacterial species** $\times$ **antibiotic response level**,
- **bird species** $\times$ **habitats**,
- **plant varieties** $\times$ **soil nature**.

In all these cases, FCA helps us understand the relationships between qualitative categories.

3.6 When is FCA used?

FCA is used when:

- the data are qualitative,
- the data are organized in a contingency table,
- we want to study the relationship between the row categories and the column categories.

4 Importance of These Methods in Biology

PCA and FCA are of great importance in biology because they make it possible to:

- summarize a large amount of data,
- interpret the relationships among variables or categories,
- detect similar groups,
- present results in a clear graphical form,
- facilitate the biological understanding of complex data.

These methods are used in many fields such as:

- botany,
- zoology,
- ecology,
- microbiology,
- genetics,
- nutrition,
- laboratory analysis.

5 Conclusion

Both **Principal Component Analysis (PCA)** and **Factorial Correspondence Analysis (FCA)** are among the most important tools of multivariate statistics in biology.

- **PCA** is used when the data are quantitative,
- **FCA** is used when the data are qualitative and represented in a contingency table.

Their importance lies in the fact that they help the researcher move from complex and detailed tables to a concise and clear representation from which biological conclusions can be drawn more easily.

6 Exercise: PCA in Biology

A biologist studies the growth of **6 plants**. For each plant, the following variables are measured:

- stem height (cm),
- root length (cm),
- number of leaves,
- dry mass (g).

The collected data are summarized below:

cccccc

Plant	Stem height	Number of leaves	Root length	Dry mass
P1	15,00	6,00	8,00	1,50
P2	17,00	7,00	9,00	1,80
P3	16,00	6,00	10,00	1,70
P4	30,00	13,00	13,00	4,10
P5	32,00	14,00	15,00	4,60
P6	31,00	12,00	14,00	4,20

Questions

1. What is the nature of the variables in this dataset?
2. Which multivariate method is the most appropriate?
3. Why is this method appropriate?
4. Give a possible biological interpretation of the first principal axis.

Solution

1. The variables are **quantitative**, since they are numerical measurements: stem height, root length, number of leaves, and dry mass.
2. The most appropriate method is **PCA (Principal Component Analysis)**.
3. PCA is appropriate because:
   - the dataset contains **quantitative variables**,
   - we want to study the relationships among these variables,
   - we want to summarize the information and detect possible groups of similar plants.
4. The data suggest that plants with greater stem height also tend to have longer roots, more leaves, and greater dry mass. Therefore, the first principal axis may be interpreted as an **overall plant growth axis**.

   From a biological point of view, PCA would likely distinguish:

   - a group of weakly developed plants: P1, P2, P3,
   - a group of strongly developed plants: P4, P5, P6.

   Thus, PCA helps summarize the general growth pattern of the plants.

6.1 Solution with SPSS

6.1.1 Data Entry in SPSS

In **Variable View**, define the variables as follows:

- `Plant`: String
- `Height`: Numeric
- `Leaves`: Numeric
- `Root`: Numeric
- `DryMass`: Numeric

Then enter the data in **Data View**.

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
Plant	String	8	0		None	None	8	Left	Nominal	Input
Height	Numeric	8	2		None	None	8	Right	Unknown	Input
Leaves	Numeric	8	2		None	None	8	Right	Unknown	Input
Root	Numeric	8	2		None	None	8	Right	Unknown	Input
Drymass	Numeric	8	2		None	None	8	Right	Unknown	Input

Figure 1: Variable View and Data View

Plant	Height	Leaves	Root	Drymass
P1	15,00	8,00	6,00	1,50
P2	17,00	9,00	7,00	1,80
P3	16,00	8,00	6,00	1,60
P4	30,00	14,00	13,00	4,20
P5	32,00	15,00	14,00	4,50
P6	31,00	14,00	13,00	4,30

Figure 2: Variable View and Data View

6.1.2 Steps for Performing PCA in SPSS

6.1.3 Step 1: Open the procedure

Go to:

`Analyze` → `Dimension Reduction` → `Factor`

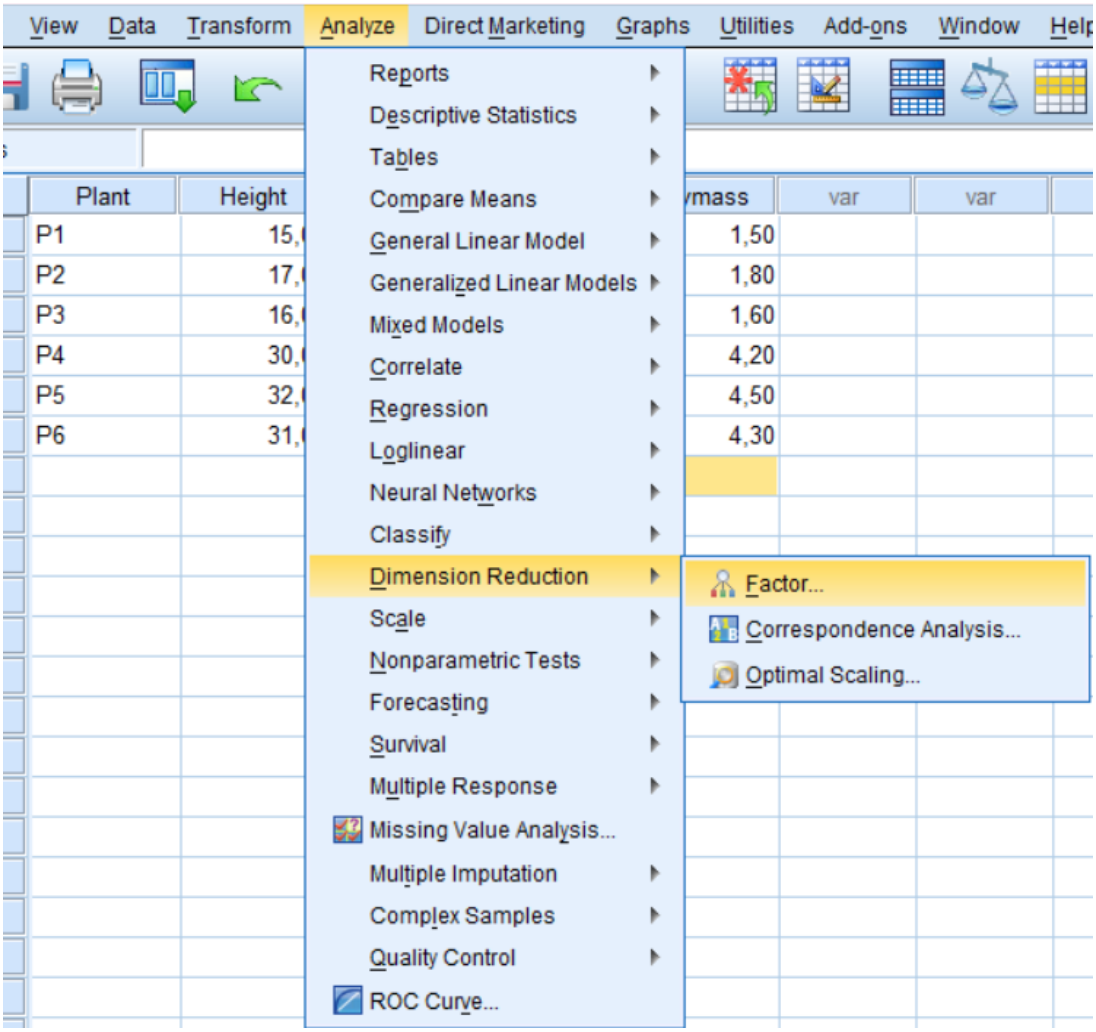

Figure 3: Open the procedure

6.1.4 Step 2: Select the variables

Move the following quantitative variables into the analysis box:

- Height
- Leaves
- Root
- DryMass

Do not include `Plant`, since it is only an identifier.

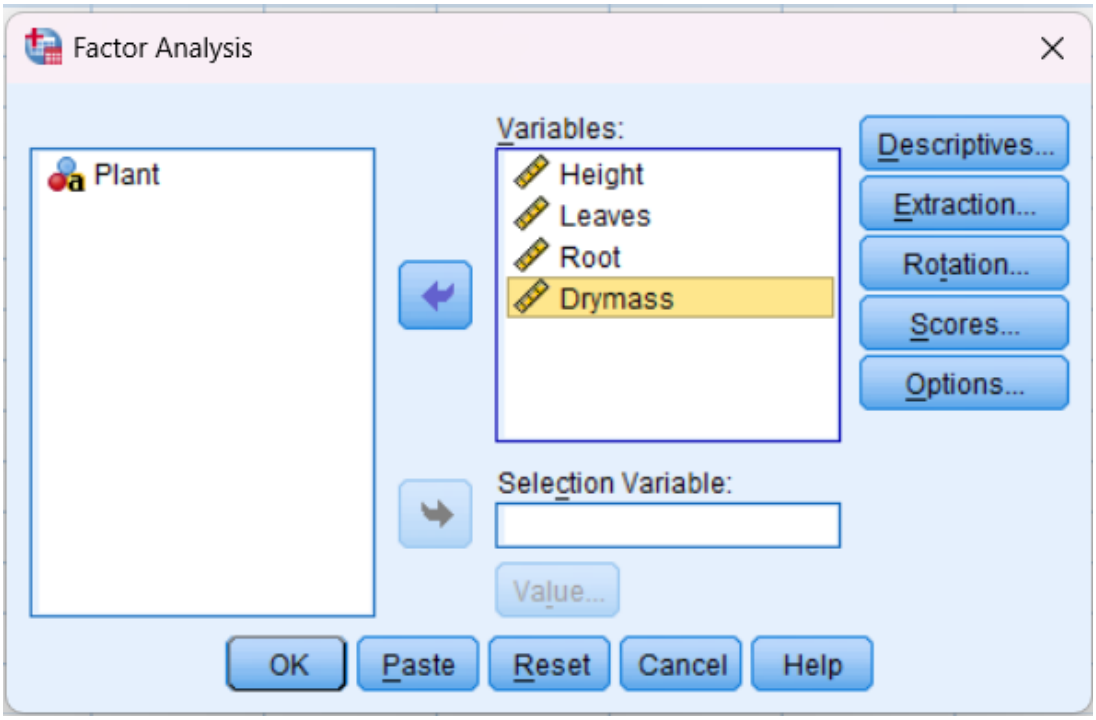

Figure 4: Select the variables

6.1.5 Step 3: Descriptive options

Click on **Descriptives**, then select:

- Correlation matrix
- KMO and Bartlett's test of sphericity
- Anti-image

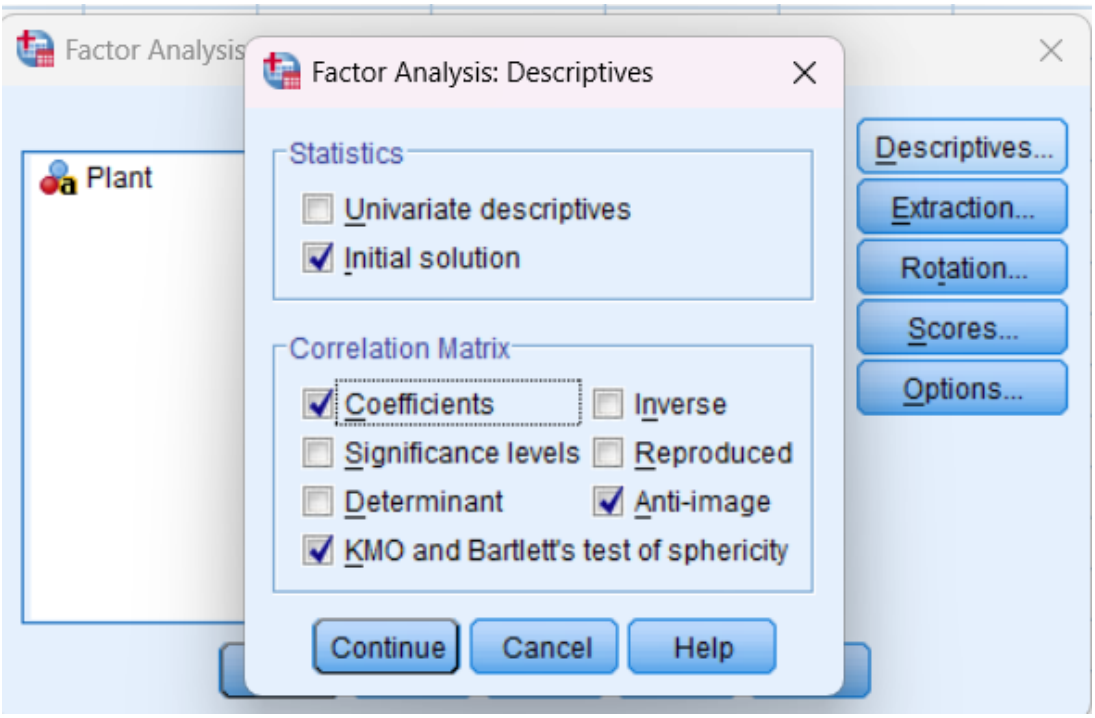

Figure 5: Descriptive options

6.1.6 Step 4: Extraction options

Click on **Extraction**, then choose:

- Method: **Principal Components**
- **Scree Plot**
- **Eigenvalues greater than 1**

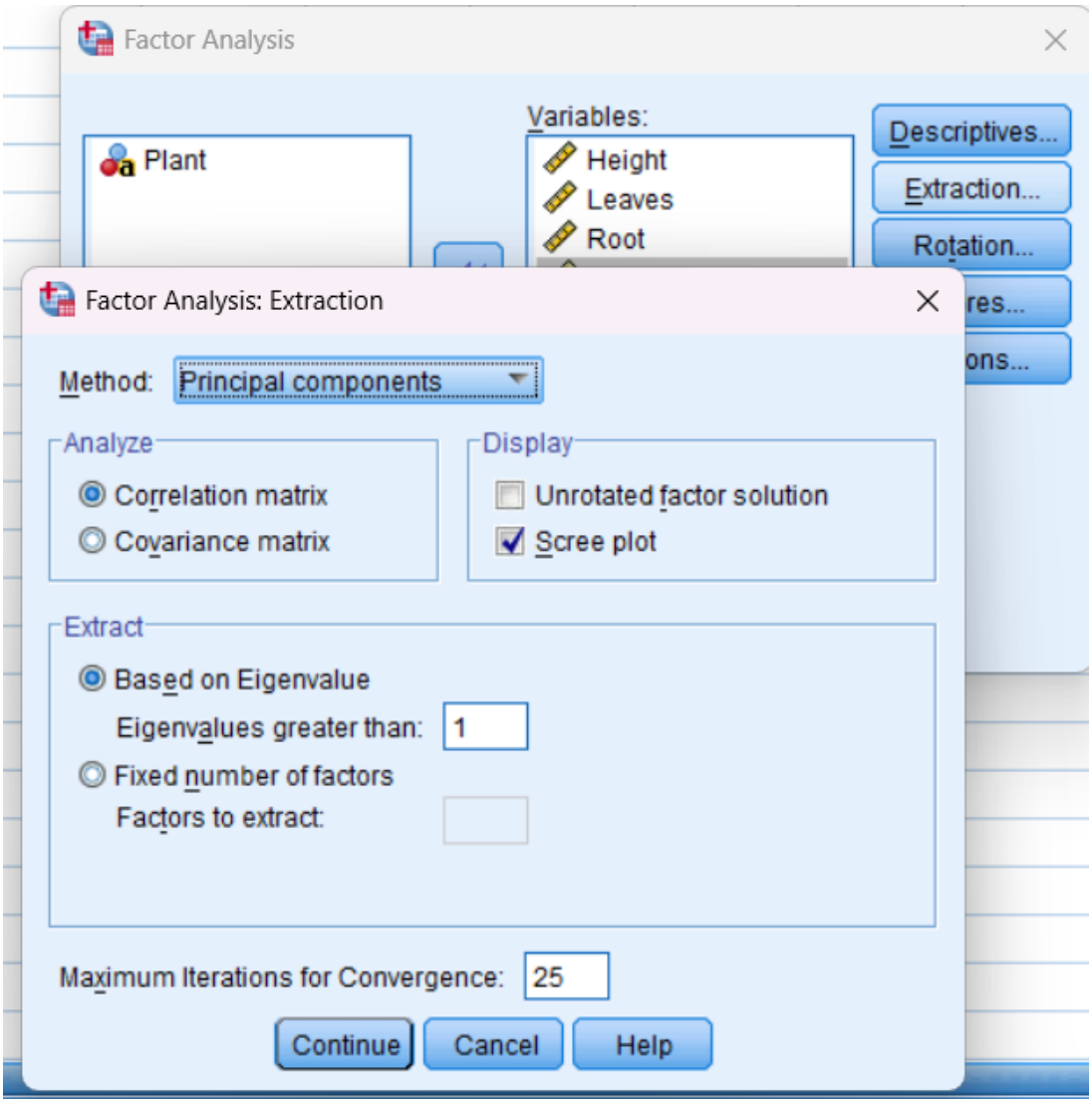

Figure 6: Extraction options

6.1.7 Step 5: Rotation

Click on **Rotation**, then choose:

- **Varimax**

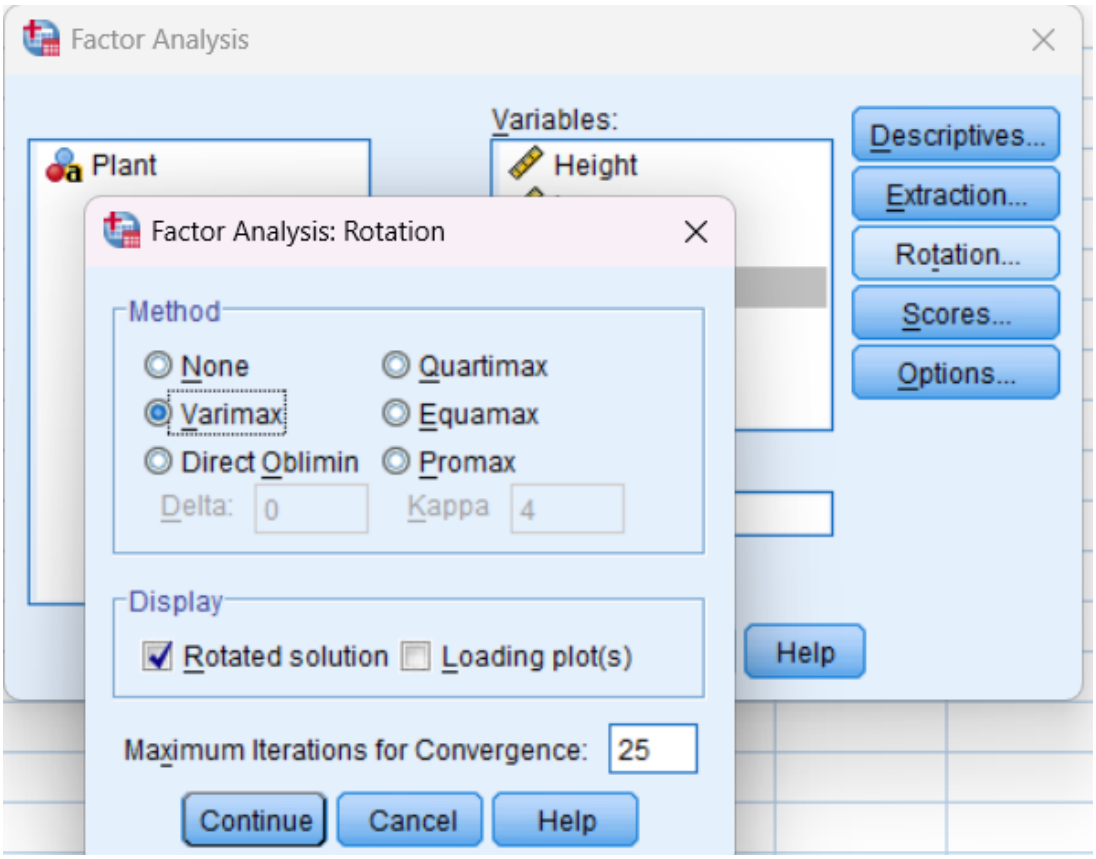

Figure 7: Rotation

6.1.8 Step 6: Save component scores

Click on **Scores**, then select:

- **Save as variables**
- **Regression**

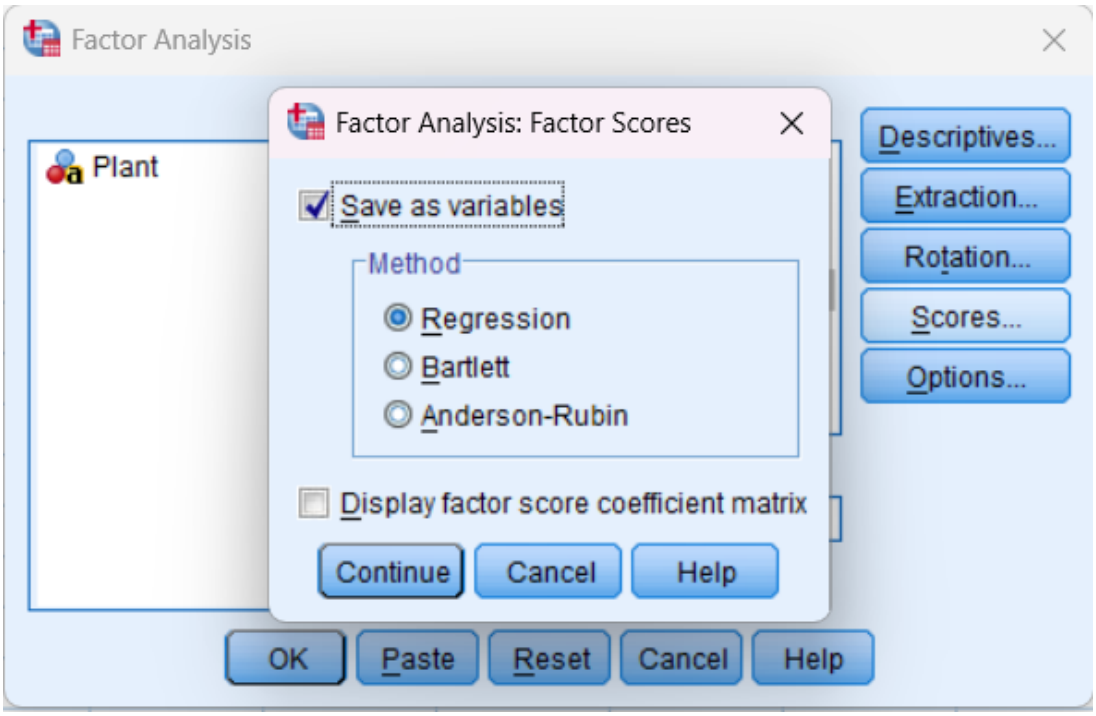

Figure 8: Save component scores

6.1.9 Step 7: Run the analysis

Click **OK** to obtain the results.

6.2 SPSS Interpretation of PCA Results

KMO and Bartlett's Test. The KMO value is 0.580, which indicates an acceptable but weak adequacy of the data for PCA. Bartlett's test is significant ($p = 0.000 < 0.05$), so the variables are sufficiently correlated. Therefore, PCA can be applied.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,580
Bartlett's Test of Sphericity	Approx. Chi-Square	40,620
	df	6
	Sig.	,000

Figure 9: KMO and Bartlett's Test

Correlation Matrix. The correlation matrix shows very strong positive relationships among Height, Leaves, Root, and Drymass. This means that plants with greater height also tend to have more leaves, longer roots, and higher dry mass. Therefore, these variables do not describe separate biological phenomena, but rather a common pattern related to overall plant growth.

Correlation Matrix

		Height	Leaves	Root	Drymass
Correlation	Height	1,000	,988	,970	,999
	Leaves	,988	1,000	,951	,992
	Root	,970	,951	1,000	,976
	Drymass	,999	,992	,976	1,000

Figure 10: Correlation Matrix

Scree Plot. The scree plot shows that the first component has a much larger eigenvalue than the others. After the first component, the eigenvalues drop sharply and become very small. This means that the first component explains most of the variability in the data, while the remaining components contribute very little. Therefore, only one principal component should be retained.

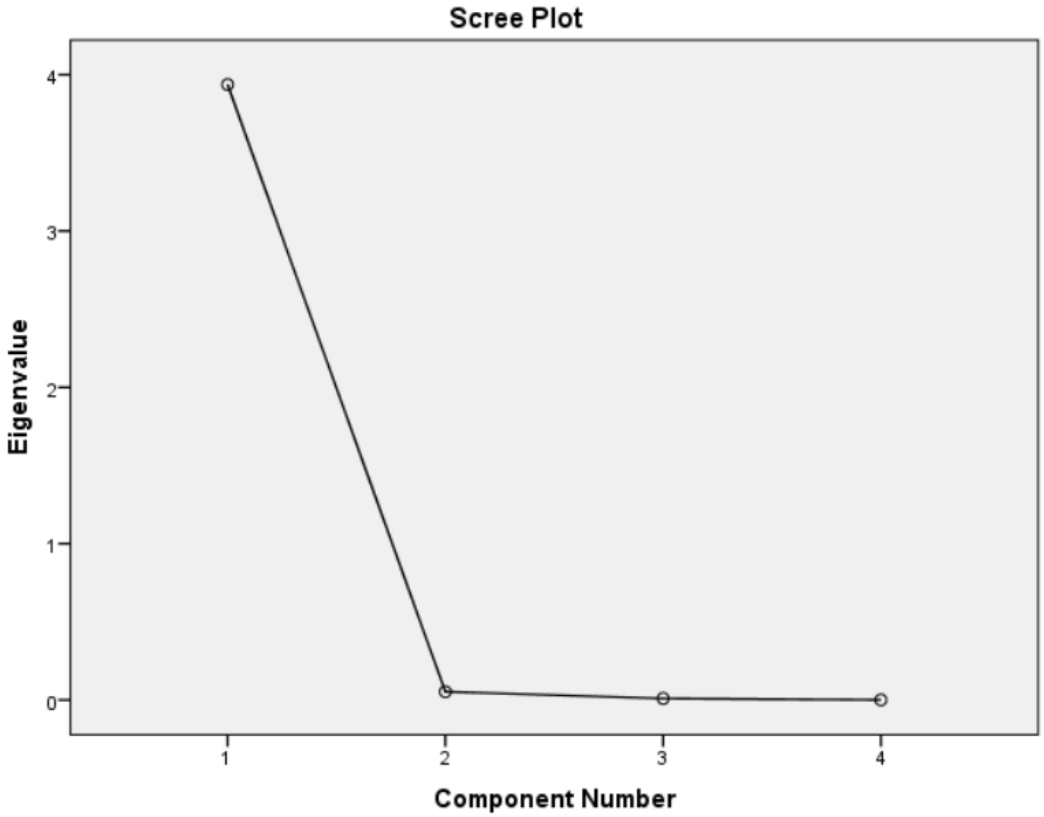

Figure 11: Scree Plot

Anti-image Matrix. The anti-image matrix provides the MSA value for each variable. These values are used to assess whether each variable is suitable for inclusion in PCA. Since all MSA values are greater than 0.5, all variables are considered acceptable and can be retained in the analysis. However, because the values are only slightly above 0.5, the adequacy of the variables remains moderate.

Anti-image Matrices

		Height	Leaves	Root	Drymass
Anti-image Covariance	Height	,001	,002	,003	-,001
	Leaves	,002	,005	,007	-,001
	Root	,003	,007	,015	-,002
	Drymass	-,001	-,001	-,002	,000
Anti-image Correlation	Height	,594[a]	,734	,732	-,960
	Leaves	,734	,593[a]	,807	-,883
	Root	,732	,807	,592[a]	-,859
	Drymass	-,960	-,883	-,859	,546[a]

a. Measures of Sampling Adequacy(MSA)

Figure 12: Anti-image Matrix

Biological Conclusion. The four variables do not describe separate biological traits. Instead, they vary together and reflect a single common characteristic of the plants. Since taller plants also tend to have more leaves, longer roots, and greater dry mass, this common dimension can be interpreted as the overall plant growth axis