Lab06: Unsupervised Learning – Identifying Structural Patterns and Development Paths

Course: Programming for Data Science
Professor: Dr. Carlos J. Costa, PhD.

March 15, 2026

Objective

The objective of this laboratory is to explore how unsupervised learning techniques can be used to identify structural patterns in multivariate datasets.

Unlike supervised learning methods, which focus on predicting a specific outcome variable, unsupervised learning seeks to uncover hidden structures and similarities within the data itself.

In this lab, students will apply two important techniques:

- Factor Analysis, used to identify latent dimensions underlying a set of correlated indicators.
- Cluster Analysis, used to group observations according to similarities across multiple indicators.

Using the dataset provided, the goal is to identify structural profiles among countries and to analyze how these profiles evolve over time.

Learning Objectives

After completing this lab, students should be able to:

- Understand the difference between supervised and unsupervised learning.
- Prepare real-world datasets for multivariate analysis.
- Apply factor analysis to identify latent dimensions in multivariate data.
- Use clustering methods to classify observations into homogeneous groups.
- Interpret clusters based on the underlying indicators.
- Analyze how countries may transition between structural profiles over time.

Dataset

The dataset used in this lab is available at:

`https://github.com/masterfloss/data/raw/refs/heads/main/WBESG6123.xlsx`

The dataset contains several indicators for many countries over multiple years. The structure follows a long (tidy) format where each row corresponds to a single observation of one indicator for one country in one year.

The main variables in the dataset are:

- **INDICATOR_LABEL** – name of the indicator
- **FREQ_LABEL** – frequency of the observation (annual)
- **TIME_PERIOD** – year of the observation
- **REF_AREA_LABEL** – country or region
- **OBS_VALUE** – numerical value of the indicator
- **UNIT_TYPE_LABEL** – unit of measurement

Because of this structure, each row represents one indicator for one country in one year.
The dataset contains approximately 450,000 observations.

Data Structure and Preparation

The dataset is stored in long format, which is common in statistical databases and panel data repositories.

However, techniques such as factor analysis and clustering require a different structure in which each observation is described by multiple variables.

Therefore, before performing the analysis, the dataset must be transformed into a wide format.
In the wide format:

- Each row corresponds to a country–year observation.
- Each column corresponds to a specific indicator.
- The cell values correspond to the numerical value of the indicator.

After reshaping the dataset, each country–year observation will be described by several indicators simultaneously, making it suitable for multivariate analysis.

Students should also check for missing values and consider appropriate strategies for handling them.

Tasks

Students should complete the following steps during the laboratory session.

Task 1: Data Exploration and Preparation

Begin by loading and exploring the dataset.

- Load the dataset from the provided link.
- Inspect its structure and variables.
- Identify the key variables describing indicators, countries, and years.
- Transform the dataset from long format into wide format so that each observation corresponds to a country–year combination and each column corresponds to an indicator.
- Examine descriptive statistics for the resulting variables.
- Analyze correlations between indicators.

Discuss whether the indicators appear to be correlated and whether dimensionality reduction techniques may be useful.

Task 2: Factor Analysis

Apply factor analysis to identify latent dimensions in the dataset.

- Select the indicators to include in the analysis.
- Standardize the variables if necessary.
- Determine an appropriate number of factors using common criteria (for example eigenvalues or scree plots).
- Estimate the factor model.
- Examine the factor loadings.

Interpret the factors by identifying which variables load strongly on each factor and discuss what structural dimensions these factors may represent.

Task 3: Factor Scores

After estimating the factor model, compute factor scores for each observation.

- Calculate the factor scores.
- Examine how countries are distributed across the identified dimensions.
- Visualize the observations in the factor space.

Discuss what the factor scores reveal about similarities and differences between countries.

Task 4: Cluster Analysis

Use clustering techniques to group observations with similar characteristics.

- Choose an appropriate clustering method.
- Determine an appropriate number of clusters.
- Assign each observation to a cluster.
- Examine the characteristics of each cluster by analyzing the average values of the indicators.

Interpret each cluster as a structural profile of countries.

Task 5: Cluster Visualization

Use the factor dimensions to visualize the clusters.

- Plot observations using the factor scores.
- Identify how clusters are distributed in the factor space.
- Evaluate whether clusters appear clearly separated.

Discuss what this visualization reveals about similarities between countries.

Task 6: Cluster Evolution Over Time

Because the dataset contains observations for multiple years, it is possible to study how cluster membership changes over time.

- Track the cluster assignment of each country across years.
- Identify countries that remain in the same cluster.
- Identify countries that move between clusters.

Discuss whether these transitions suggest structural changes.

Task 7: Development Paths

Using the cluster transitions identified in the previous step, analyze possible development trajectories.

- Identify countries that move from one cluster to another over time.
- Determine whether some transitions correspond to improvements in key indicators.
- Identify clusters that appear stable and clusters that change frequently.

Discuss possible explanations for these trajectories.

Discussion Questions

1. What are the main latent dimensions identified by factor analysis?
2. How many clusters best represent the data?
3. What are the key characteristics of each cluster?
4. Which variables contribute most to the separation between clusters?
5. Do countries tend to remain in the same cluster over time?
6. Can you identify examples of countries moving between clusters?
7. What factors might explain these transitions?

Concluding Remarks

Unsupervised learning techniques provide powerful tools for exploring complex datasets where the objective is to understand structure rather than prediction.

In this lab, factor analysis and clustering are used to reduce the dimensionality of the data, identify structural similarities across countries, classify countries into distinct profiles, and analyze how these profiles evolve over time.

These methods are widely used in empirical research to analyze heterogeneity and structural trajectories across countries.