Principal Component Analysis (PCA) Lab Instructions

Multivariate Data Analysis Course

Objective

The objective of this lab is to:

- Program the steps involved in performing centered PCA and normalized PCA (centered and reduced),
- Assess the representativeness of PCA by reducing dimensionality with i) selecting dimension $m < p$, and ii) evaluating the quality of individual projections (projection from R^p into R^n) as well as the quality of the projections of variables X_j,
- Visualize the quality of dimensionality reduction based on the intrinsic properties of the sample,

Part I: PCA in the Variable Space R^p

1. Initial Script Setup

Develop a script that allows you to:

- Visualize the matrix $\mathbf{X}$ of dimension (n, p) where n represents individuals and p represents variables,
- Construct basic statistical indicators: variance, covariance, and standard deviation.

2. Implement PCA for R^p Space

Create a script to perform PCA by:

- Centering the data points in R^p to find the hyperplanes that maximize projected inertia. These hyperplanes correspond to the p directions of the space, where each factorial axis is represented by an eigenvector $\mathbf{u}$ (or matrix $\mathbf{U}$) and an eigenvalue λ (or matrix $\mathbf{\Lambda}$) obtained through diagonalizing the covariance matrix.

Part II: Quality Assessment of PCA

To evaluate the quality of dimensionality reduction, you should:

- Assess the quality of the data cloud reduction,
- Visualize the projection of individuals,
- Calculate the contribution and quality of the projected individuals.
- Additionally, consider the case of normalized PCA (centered and reduced data).

3. Inertia Explained by Principal Components

For representing the explained inertia λ_j for each of the j principal components obtained (with j ranging from 1 to p):

- Use bar plots to display both the explained inertia for each component and the cumulative sum of inertia from components 1 to p.
- Experiment with different criteria for selecting the number of principal components to retain based on λ_j for each component.

4. New Coordinates for Each Individual

Compute the new coordinates C_{ij} of each individual i along each of the components j, which helps in defining the quality of the individual's projection Q_{ik}, given by:

$$Q_{ik} = \frac{\sum_{j=1}^{k}(C_{ij})^2}{\sum_{j=1}^{p}(C_{ij})^2}$$

5. Contribution to Factorial Axes

Define the contribution of each individual i to the inertia of factorial axis j as:

$$\gamma_{ij} = \frac{1}{n}\frac{(C_{ij})^2}{\lambda_j}$$

This allows calculating each individual's contribution and quality within the new subspace.

6. Validation of the First Factorial Plane

Verify the accuracy of your first factorial plane by comparing it with PCA results from functions such as ‘dudi.pca()‘ in R (or equivalent in Python). Functions like ‘plot‘, ‘eigen‘, and ‘plot3d‘ are useful for visualization.

7. Graphical Representation of Individuals

Represent the individuals in the new subspace according to the selected first and second factorial planes (e.g., ‘plot(CP1, CP2)‘ or ‘plot(CP1, CP3)‘). Use the function ‘dudi.pca()‘ in R or equivalent functions in Python.

Part III: Studying Cloud Shape and Dimensionality Reduction

8. Isotropic Cloud

Generate data for three variables in a nearly spherical point cloud. Possible methods include generating a sample of size n for the three variables X, Y, and Z as independent Gaussian vectors following $N(0, 1)$ and obtaining vector $\mathbf{V}$ with three components normalized by $\|\mathbf{V}\|$.

8.1 Script for Data Generation

Write a script to generate such data and test PCA on it.

8.2 Observing Covariance and Quality of Projection

Observe the evolution of the covariance or correlation matrix, eigenvalues, and projection quality as the sample size changes. Provide an interpretation of your findings.

9. Anisotropic Cloud

To observe stronger correlations among variables, generate data for three variables X, Y, and Z where X and Y have a linear relationship, and similarly between Y and Z. Adding noise can adjust the correlation level.

9.1 Testing PCA with Variable Sample Sizes

Test PCA on these data and observe changes in the covariance or correlation matrix, eigenvalue distribution, and quality of projection as the sample size changes. Interpret your observations.

10. Extreme Points

For a non-isotropic point cloud, add a few extreme points in one or two variables, stretching the cloud in that direction, or remove extreme points to compress the cloud along specific variables.

10.1 Experiment with Centered and Normalized PCA

Propose experiments with centered and normalized PCA on this dataset. Discuss your observations and conclusions regarding centering and normalizing the data.