Agricultural Statistics in Practice

Multivariate Analysis

Principal Component Analysis – Scree Plot

Dr. Amandeep Singh

Imagineering Laboratory
Indian Institute of Technology Kanpur

Principal Component Analysis

- PCA is a statistical method that employs an orthogonal transformation to convert a collection of observations with potentially correlated variables into a set of values known as principal components.

- These components are linearly uncorrelated, providing a new representation of the data.

How does PCA work?

Graphical

- Calculate a matrix summarizing the relationships among variables.
- Break down the matrix into direction and magnitude components to understand the data's directions and their importance.
- The Principal Component (PC) matrix has the same dimensions as the original data matrix, but reduction of PCs is often necessary. To determine this reduction, we examine the variance explained by each PC vector.
- Approaches for determining the number of PCs:

1.   Eigenvalue criterion

2.   Proportion of variance explained criterion

3.   Scree plot criterion   (graphical)

Scree Plot

- Scree plots are a graphical representation used to determine the number of retained principal components (PCs).

- They display eigenvalues on the y-axis and the number of factors on the x-axis.

- Scree plots typically exhibit a downward curve, starting high on the left and falling quickly.

- The plot flattens out at some point, indicating diminishing returns in explaining variability.

- The first component usually explains a significant portion of the variability.

Charactceristics roots

Scree Plot

A

elbow

Eigenvalue

Component Number
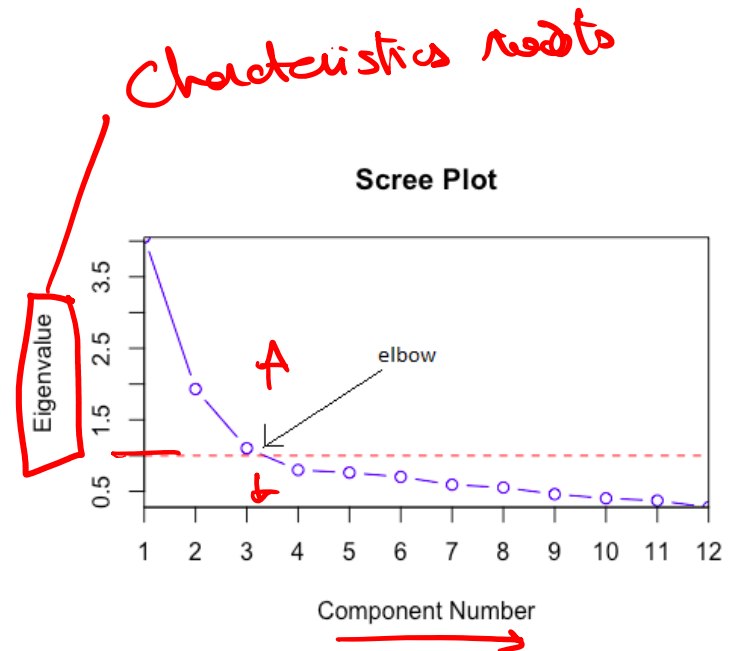
Scree Plot

Scree Plot

- The next few components explain a moderate amount, while the latter components contribute less.

- The scree plot criterion identifies the "elbow" in the curve as the cutoff for retaining components.

- Components before the elbow are selected as they capture the most significant variability.

- The term "Scree Plot" derives from its resemblance to a slope of fallen rocks on a mountain.

- Scree plots are commonly used in PCA to guide the selection of principal components.
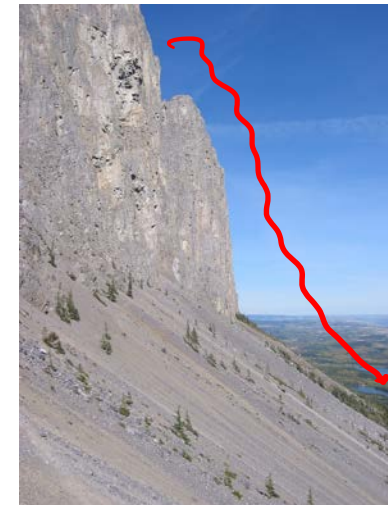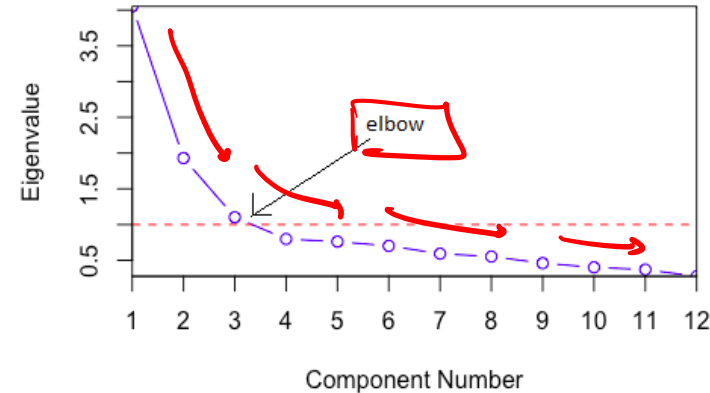
Scree Plot

- To address an irregular scree plot curve, there are a few approaches:
  - Kaiser rule: Select principal components (PCs) with eigenvalues of at least 1.
  - Proportion of variance plot: Choose PCs that collectively account for at least 80% of the variance
- If the number of principal components exceeds 3, alternative dimension reduction techniques like t-SNE and MDS may be more suitable for visualizing the data.

Scree Plot

- However, the scree plot test has faced criticism for its subjectivity. Multiple "elbows" in the plot can make it challenging to determine the appropriate number of factors or components, leading to unreliable results.

- Moreover, the lack of standard scaling for the x and y axes means that different statistical programs may generate different plots from the same data.

- To address these concerns, a more objective variation of the scree test, known as the Cattell-Nelson-Gorsuch scree test (CNG scree test), can be used.

Thank You