DATA MINING

SAI TEJA BANDARU

OBJECTIVE:

This project demonstrates the practical and theoretical aspects of a Data Mining course, highlighting how various tools and techniques can be applied to analyze and extract insights from data. It aims to provide a clearer understanding of how these methods can make data more accessible and manageable.

WHAT IS THE NEED FOR THE DATA MINING AND DATA COLLECTING ?

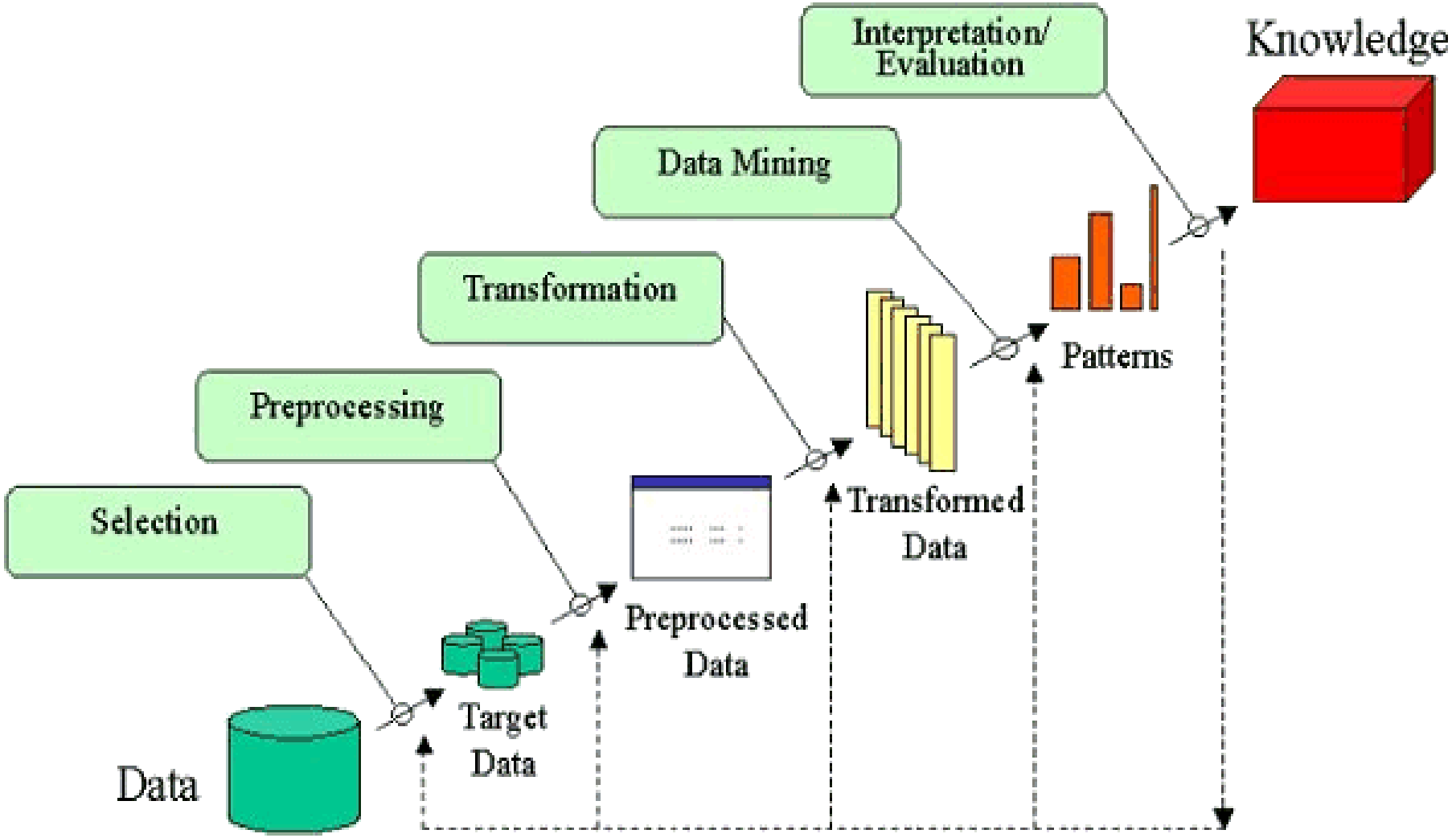
- Data Collection and Data Mining enable informed decision-making, improve efficiency, enhance customer experience, provide a competitive edge, and support effective risk management.

1. Web Data(ex: facebook has billions of active users)
2. Retail Market Insights
3. Bank/Credit card transactions
4. Bank Fraud Detection

KDD

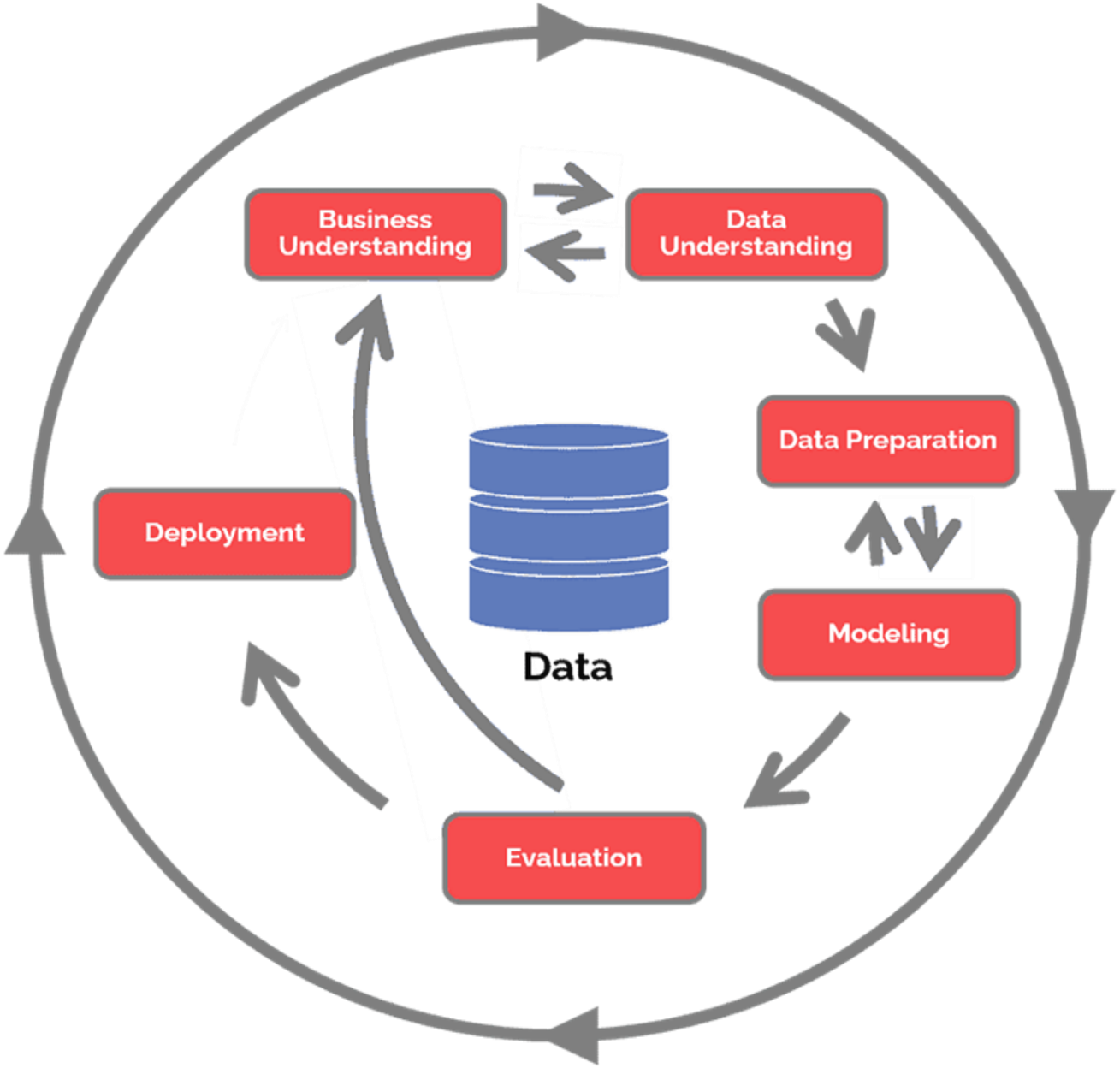
- The full form of KDD is "Knowledge discovery in databases", and it refers to the process of finding useful knowledge or the patterns from the large volumes of the data.
- Data mining is the integral part of the KDD process.
- The objective of KDD is to convert raw data into actionable insights through various techniques and methodologies.

KDD PROCESS STEPS



KEY STEPS IN THE KDD PROCESS:

- **Data Selection:** Collecting relevant data from various sources.
- **Data Preprocessing:** Cleaning and organizing data to handle missing values, noise, and inconsistencies.
- **Data Transformation:** Converting data into formats suitable for mining, such as normalization or aggregation.
- **Data Mining:** Applying algorithms to discover patterns, correlations, and insights.
- **Evaluation:** Assessing the mined patterns to ensure they are valid and useful.
- **Knowledge Representation:** Presenting discovered knowledge in a user-friendly format for decision-making



DATA

Data is a collection of facts, figures, and details.

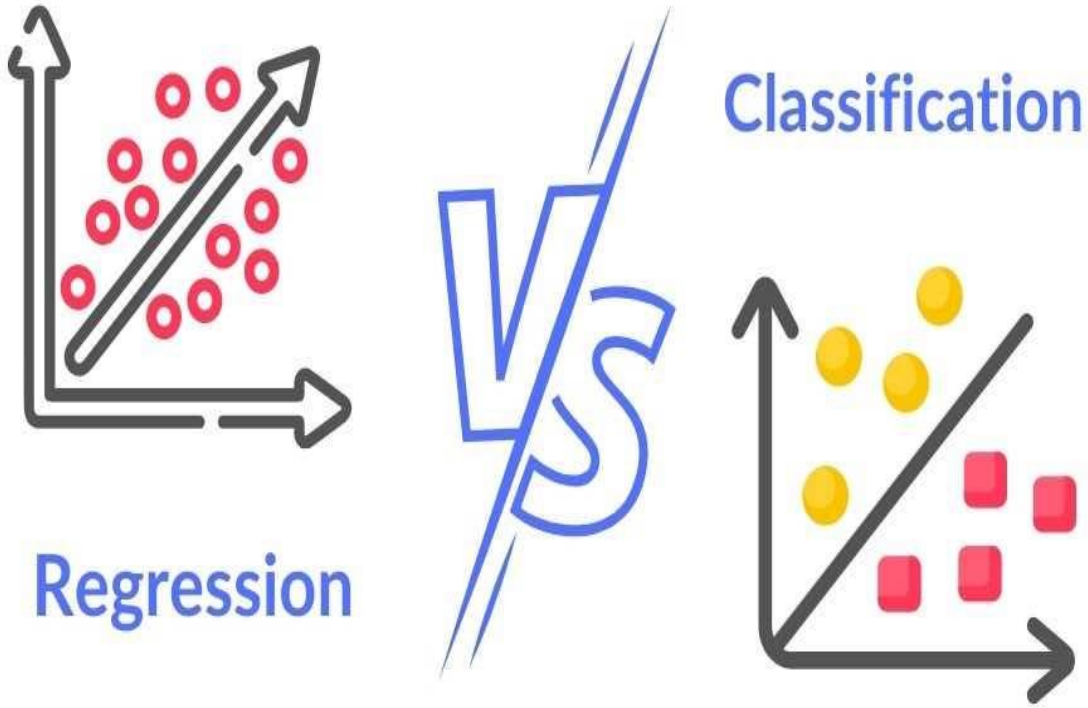
A **pattern** is a recognizable structure or relationship within this data, often expressed in a specific language or model that applies to a subset of the data.

MAIN GOALS OF DATA MINING

PREDICTION: Uses variables/fields to predict unknown or future values of interest.

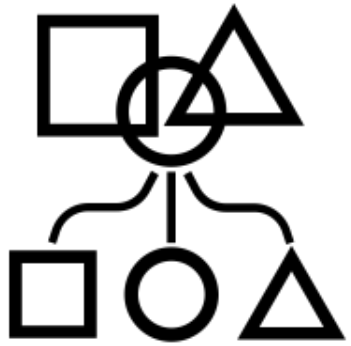
DESCRIPTION: Focuses on finding interpretable patterns that describe the data.

COMMON TECHNIQUES

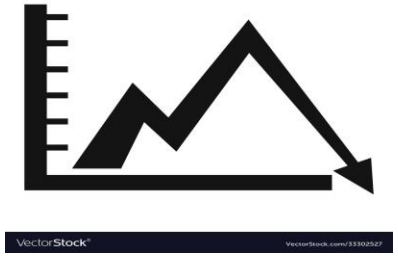


- **Classification:** Defines a function/rule to classify an element into predefined classes.
- **Regression:** Defines a function to attribute a value to an element based on a predictive variable with real values.

CLASSIFICATION EXAMPLES



- Classifying credit card transactions as legitimate or fraudulent
- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant



REGRESSION:

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively studied in statistics, neural network fields.

Examples:

- Predicting sales amounts of new product based on advetising expenditure.
- Time series prediction of stock market indices.

ATTRIBUTES

Attribute values are numbers or symbols (categories, names, …) assigned to an attribute for a particular object

Difference between attributes and attribute values

Same attribute can be mapped to different attribute values
Example: height can be measured in feet or meters

Different attributes can be mapped to the same set of values
Example: Attribute values for ID and age are integers
Discrete attribute values

TYPES OF ATTRIBUTES

1.Nominal

Examples: ID numbers, eye color, zip codes

2.Ordinal

Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}

3.Interval

Examples: calendar dates, temperatures in Celsius or Fahrenheit.

4.Ratio

Examples: temperature in Kelvin, length, counts

Discrete Attribute

- Has only a finite or countably infinite set of values

Examples: zip codes, counts, or the set of words in a collection of documents

- Often represented as integer variables.
- Note: binary attributes are a special case of discrete attributes

Continuous Attribute

- Has real numbers as attribute values

Examples: temperature, height, or weight.

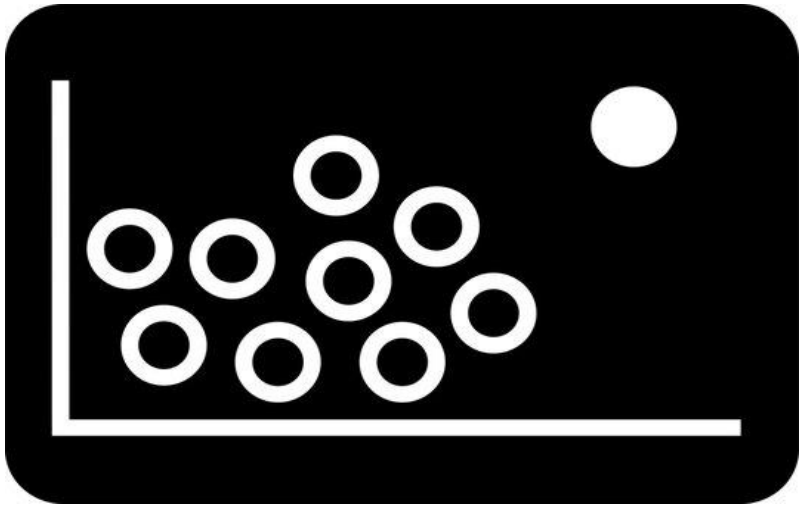
- Practically, real values can only be measured and represented using a finite number of digits.
- Continuous attributes are typically represented as floating-point variables.

NOISE DEFINITION

- For objects, noise is an extraneous object
- For attributes, noise refers to modification of original values

Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen

OUTLIERS



Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set

EUCLIDEAN DISTANCE

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

where n is the number of dimensions (attributes) and xk and yk are, respectively, the kth attributes (components) or data objects x and y.

MINKOWSKI DISTANCE

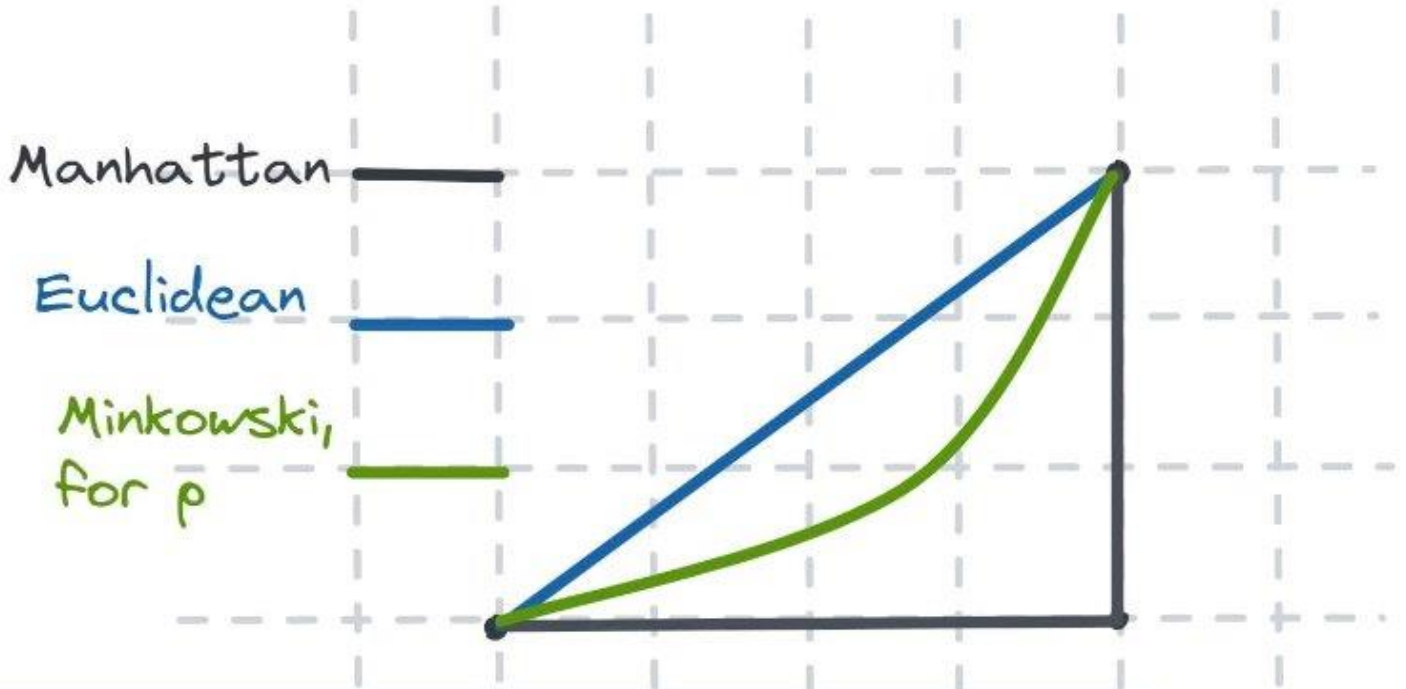
Minkowski Distance is a generalization of the Euclidean distance.

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Where r = parameter

n = no of dimensions(attributes)

xk or yk = kth attributes(components) or data objects x and y.



Minkowski when p = 1 --> Manhattan
Minkowski when p = 2 --> Euclidean

SAMPLING

- Sampling is the main technique employed for data reduction.
- It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians often sample because obtaining the entire set of data of interest is too expensive or time consuming.
- Sampling is typically used in data mining because processing the entire set of data of interest is too expensive or time consuming.

TYPES OF SAMPLING

Simple Random Sampling

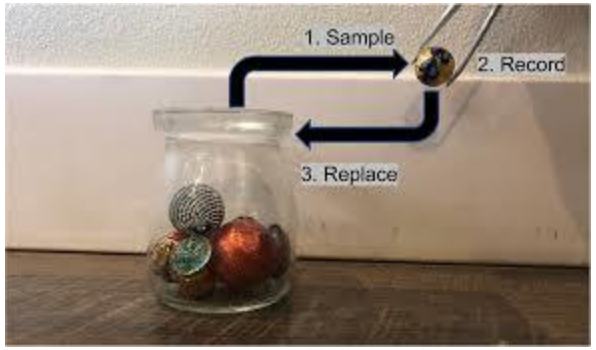
- There is an equal probability of selecting any object (individual) Sampling without replacement
- As each object is selected, it is removed from the population

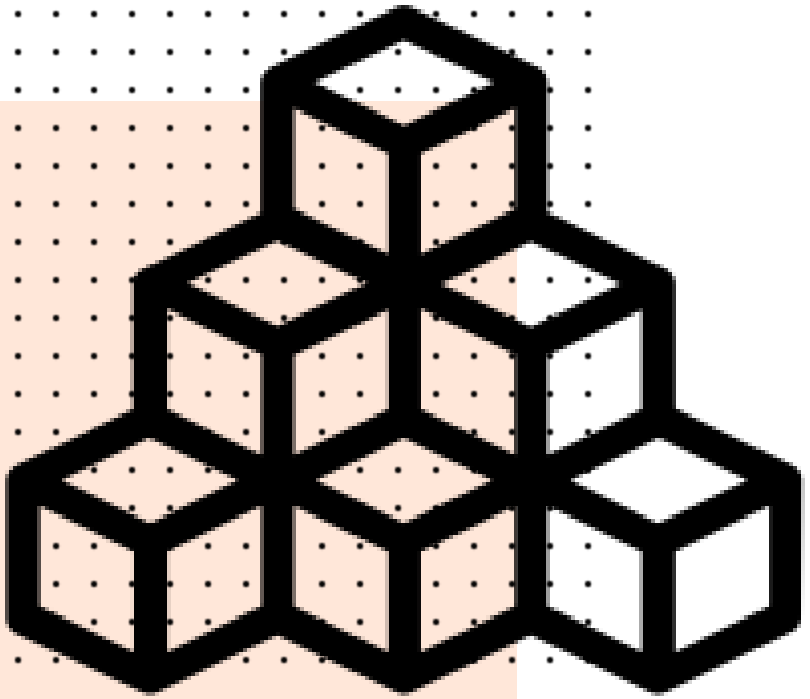
Sampling with replacement

- Objects are not removed from the population as they are selected for the sample.
- In sampling with replacement, the same object can be picked up more than once

Stratified sampling

- Split the data into several partitions; then draw random samples from each partition

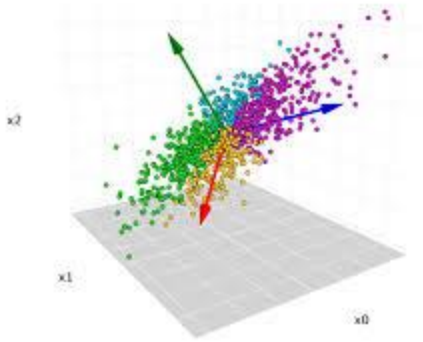




MULTIDIMENSIONAL DATA ANALYSIS

Set of descriptive methods having for objective to summarize, visualize and explore relevant information contained in large data tables.

PRINCIPAL COMPONENT ANALYSIS

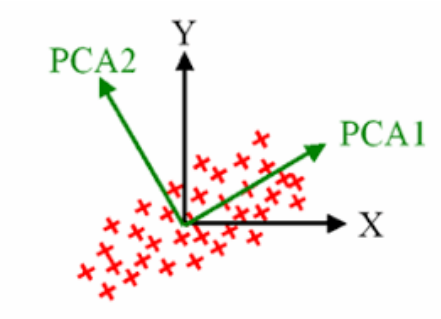


- **Principal Component Analysis (PCA)** is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional representation, capturing the most important information while minimizing the loss of variance. PCA is widely applied in various fields, including data analysis, image processing, and feature extraction.

Objective: PCA aims to identify the principal components in the data, which are linear combinations of the original variables. These principal components capture the maximum variance in the dataset.

Use case: data reduction and noise reduction

PCA PROCESSING

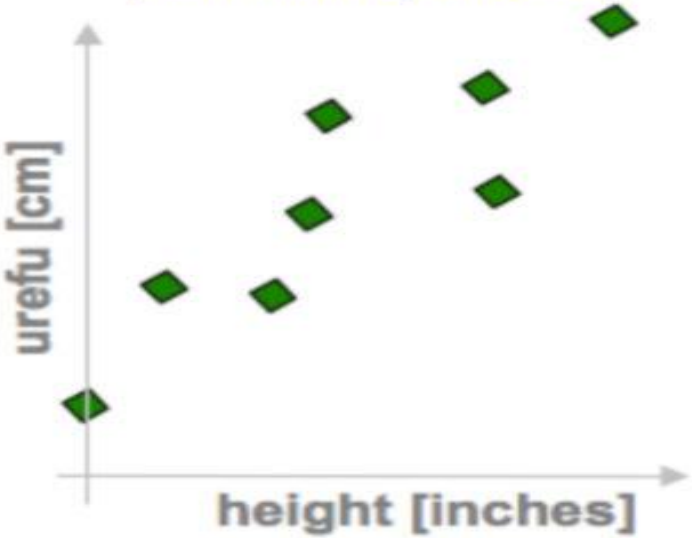


Steps in PCA:

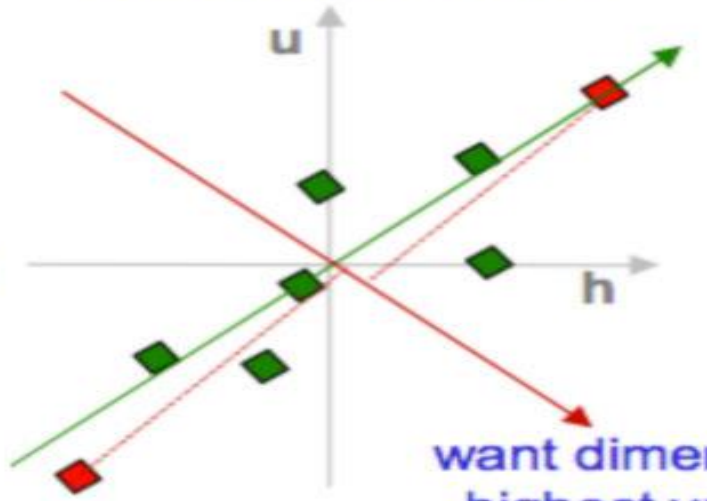
- Standardization: Standardize the variables to have zero mean and unit variance.
- Covariance Matrix: Compute the covariance matrix of the standardized data.
- Eigendecomposition: Find the eigenvectors and eigenvalues of the covariance matrix.
- Principal Components: Sort the eigenvalues in descending order and select the corresponding eigenvectors as principal components.
- Projection: Project the original data onto the new lower-dimensional space defined by the principal components.

PCA in a nutshell

1. correlated hi-d data
("urefu" means "height" in Swahili)



2. center the points



want dimension of highest variance

3. compute covariance matrix

	h	u
h	2.0	0.8
u	0.8	0.6

$$\text{cov}(h,u) = \frac{1}{n}\sum_{i=1}^{n} h_i u_i$$

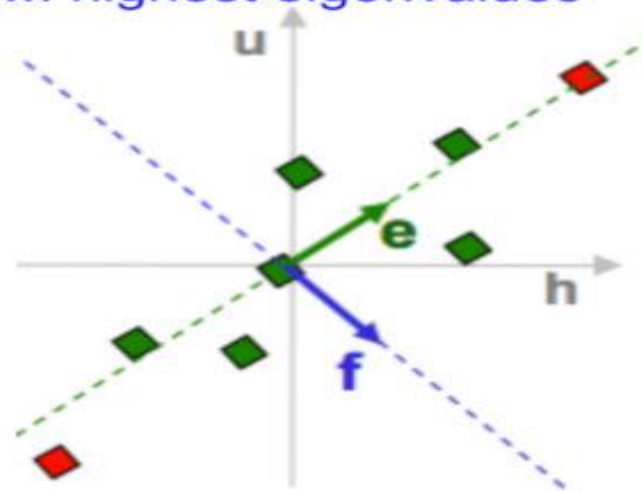
4. eigenvectors + eigenvalues

$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} e_h \\ e_u \end{pmatrix} = \lambda_e \begin{pmatrix} e_h \\ e_u \end{pmatrix}$$

$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} f_h \\ f_u \end{pmatrix} = \lambda_f \begin{pmatrix} f_h \\ f_u \end{pmatrix}$$

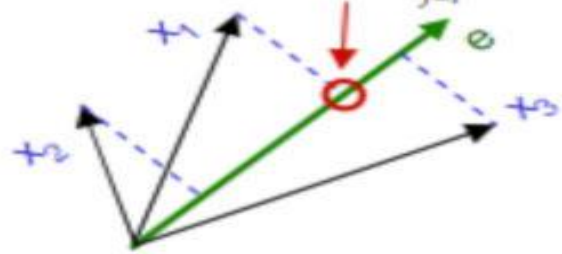
```
eig(cov(data))
```

5. pick m<d eigenvectors
w. highest eigenvalues

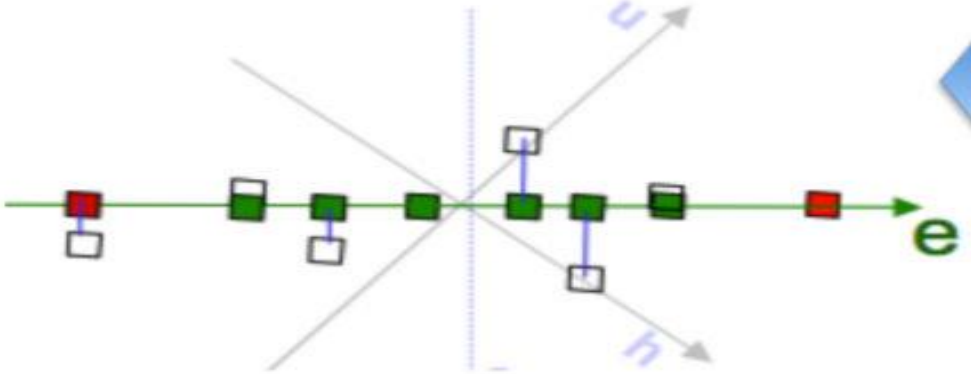


6. project data points to those eigenvectors

$$x_e' = x^T e = \sum_{j=1}^{a} x_{ij} e_j$$



7. uncorrelated low-d data



SINGULAR VALUES IN PCA

- **Definition:** **Singular values measure the amount of variance captured by each principal component.**
- **Computation:** **Using SVD, X = U Σ V^T.**
- **Importance:** **Higher singular values indicate more significant principal components.**

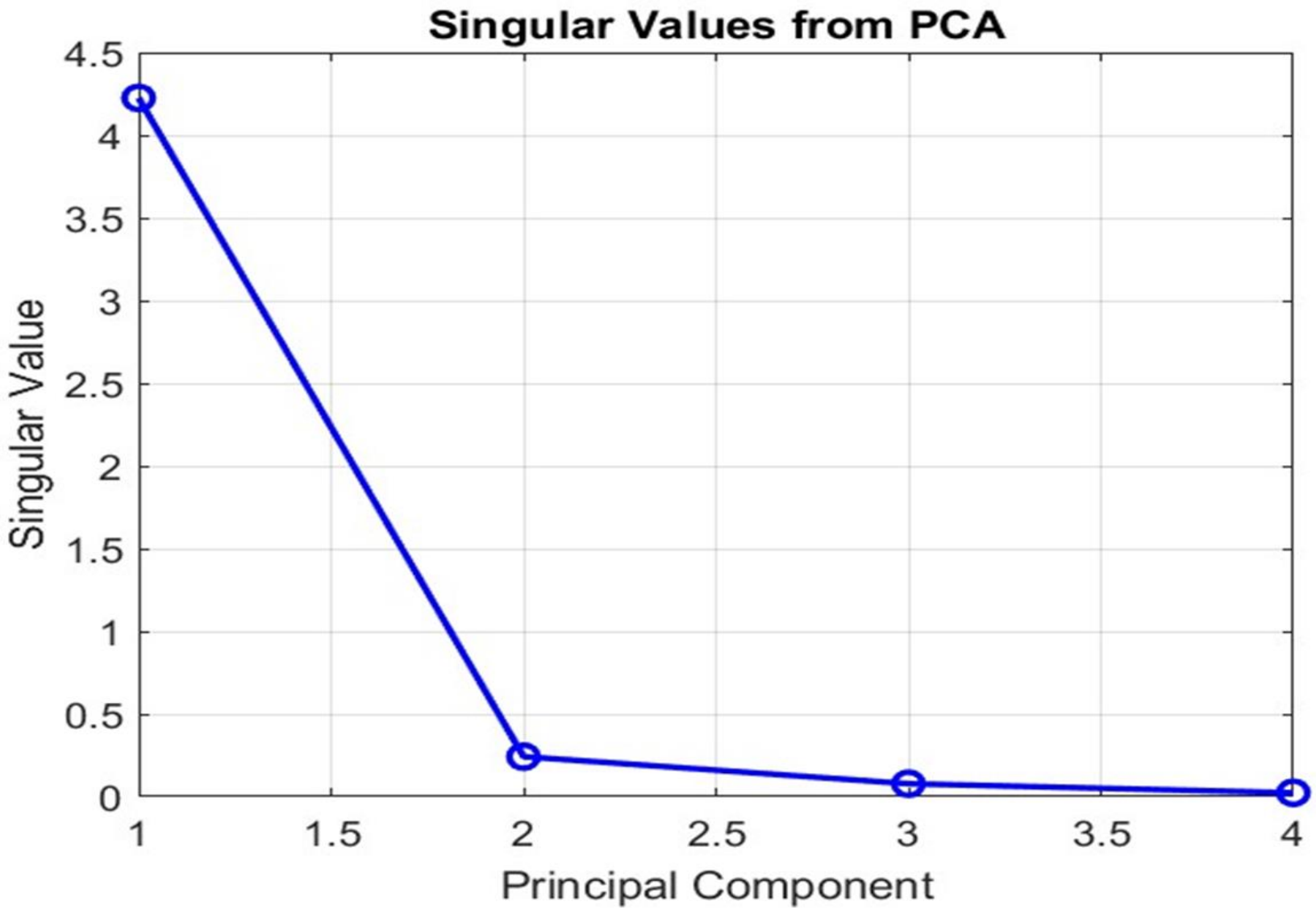
EXAMPLE ON PCA

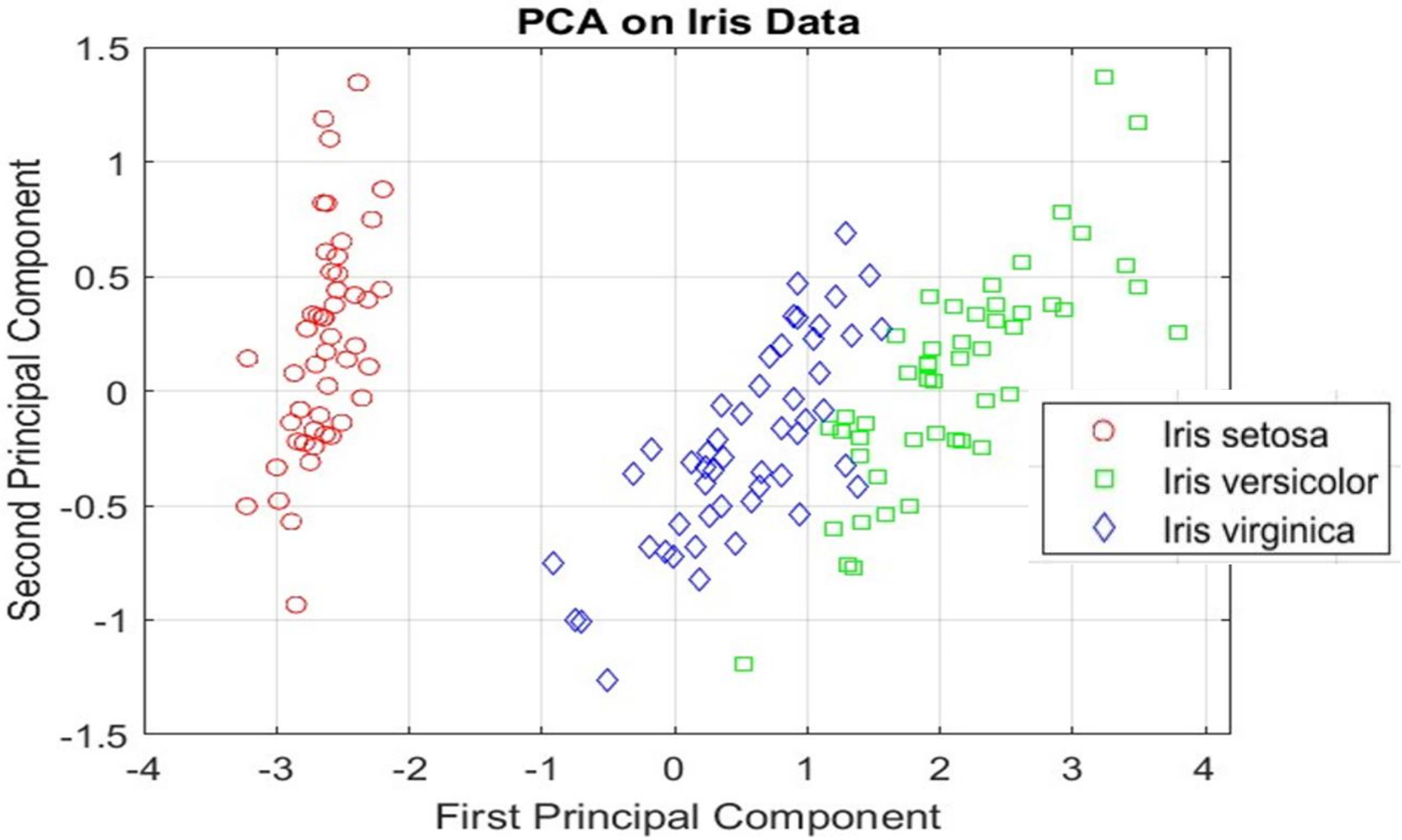
DATA SET: IRIS

Specie 'virginica'

Specie 'setosa'

Specie 'versicolor'


Singular Values from PCA
Singular Value
Principal Component
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
1
1.5
2
2.5
3
3.5
4


PCA on Iris Data
Second Principal Component
First Principal Component
1.5
1
0.5
0
-0.5
-1
-1.5
-4
-3
-2
-1
0
1
2
3
4
Iris setosa
Iris versicolor
Iris virginica

CORRESPONDENCE ANALYSIS

Correspondence Analysis (CA) is a multivariate statistical technique used to analyze and visualize the relationships between two categorical variables. It transforms data from a large contingency table into a lower-dimensional space, helping to identify patterns and associations between categories.

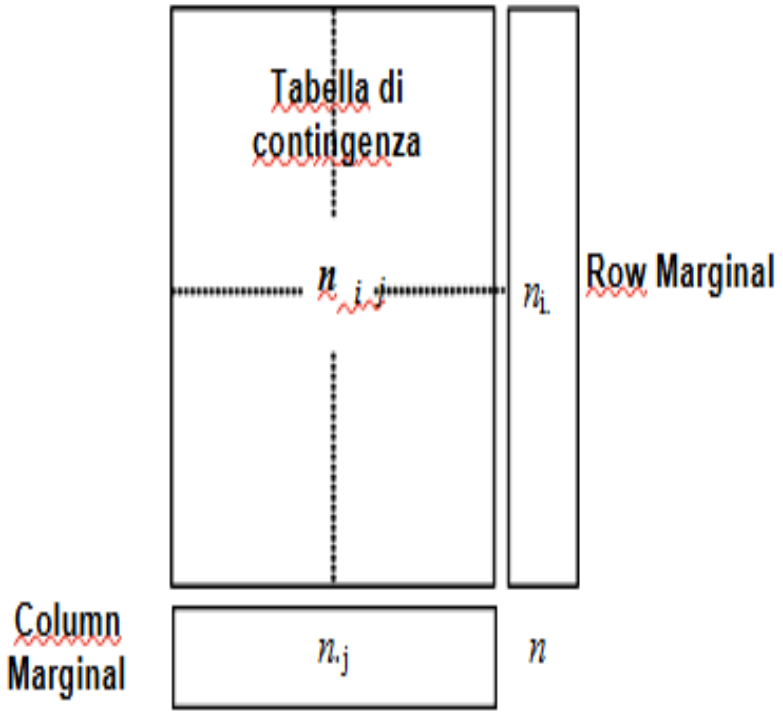
WHEN TO USE CA?

- When you have a contingency table (cross-tabulation) of two categorical variables.
- To explore associations between categories.
- To visualize the data in a low-dimensional space for easier interpretation.

STEPS OF CORRESPONDENCE ANALYSIS

1. Construct the Contingency Table:

Contingency Table: A matrix where rows represent one categorical variable and columns represent another, with each cell containing the frequency count of occurrences for each combination of Categories.



2. CALCULATE ROW AND COLUMN PROFILES

Row Profile: The proportion of each cell within a row relative to the row sum.

$$f_{ij} = \frac{n_{ij}}{n_{i.}}$$

- n_{ij}: Frequency count for row i and column j.
- $n_{i.}$: Sum of frequencies for row i (row marginal total).

Column Profile: The proportion of each cell within a column relative to the column sum.

$$f_{ij} = \frac{n_{ij}}{n_{.j}}$$

- n_{ij}: Frequency count for row i and column j.
- $n_{.j}$: Sum of frequencies for column j (column marginal total).

3. COMPUTE MARGINAL TOTALS AND EXPECTED FREQUENCIES

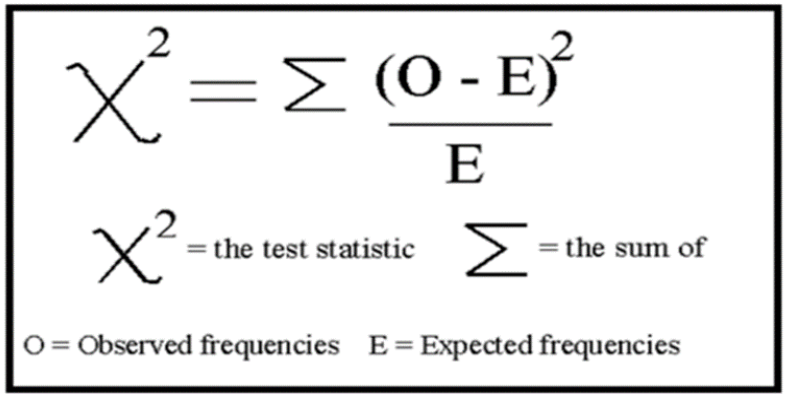
Marginal Totals: The sums of rows and columns in the contingency table.
Expected Frequencies: The expected count in each cell if rows and columns were independent.

$$e_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$$

- $n_{i\cdot}$: Row marginal total for row i.
- $n_{\cdot j}$: Column marginal total for column j.
- n: Grand total of all frequencies in the table.

CHI-SQUARE STATISTIC FOR CA

- The chi-square statistic is a measure that assesses the difference between the observed and expected frequencies in a contingency table. In Correspondence Analysis, a higher chi-square value indicates a stronger association between the categorical variables.
- According to chi-squared distance, rows that have similar profiles can be aggregated without affecting the geometry of the columns, and vice-versa.



4. CALCULATE CHI-SQUARE DISTANCES

- Measures the distance between row (or column) profiles and the average profile (barycenter).
- **Chi-Square Distance for Rows and Columns:**

$$d^2(R_i, R_j) = \sum_k \frac{(f_{ik} - f_{jk})^2}{f_{.k}}$$

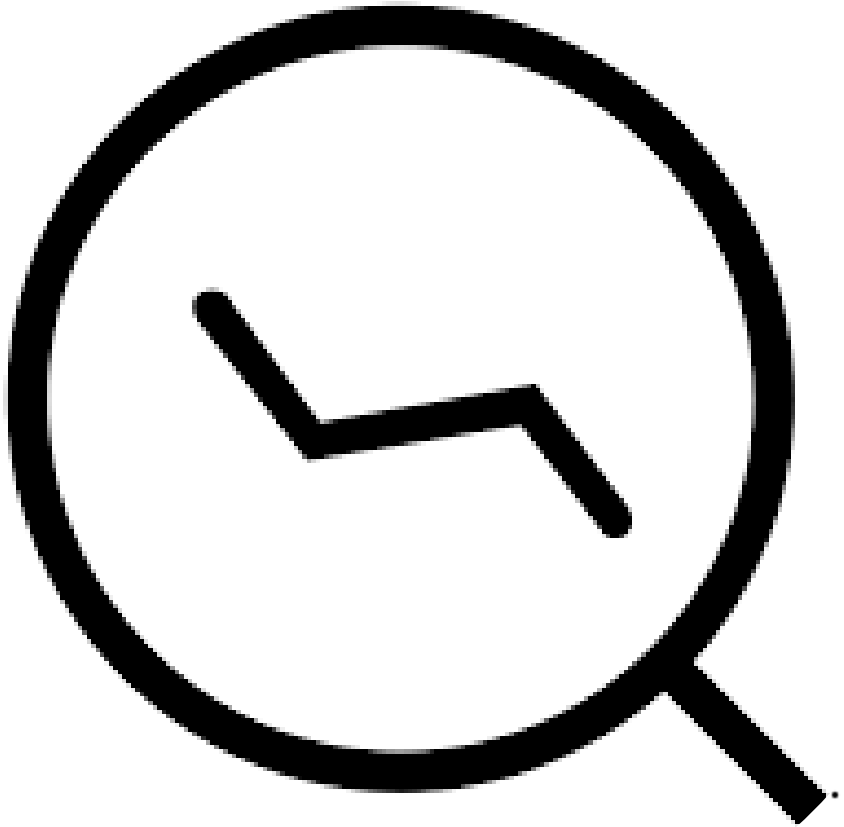
- $d^2(R_i, R_j)$: Chi-square distance between row profiles i and j.
- f_{ik}: Proportion of row i in column k.
- f_{jk}: Proportion of row j in column k.
- $f_{.k}$: Proportion of total in column k.

$$d^2(C_i, C_j) = \sum_k \frac{(f_{ki} - f_{kj})^2}{f_{k.}}$$

- $d^2(C_i, C_j)$: Chi-square distance between column profiles i and j.
- f_{ki}: Proportion of column i in row k.
- f_{kj}: Proportion of column j in row k.
- $f_{k.}$: Proportion of total in row k.

STANDARDIZED RESIDUALS

Standardized residuals indicate how far the observed counts in a contingency table deviate from what would be expected if the variables were independent. Positive values signify over-representation, and negative values signify under-representation.



5. DECOMPOSES THE STANDARDIZED RESIDUALS MATRIX Z TO IDENTIFY PRINCIPAL COMPONENTS

$$Z = D_r^{-1/2}(P - E)D_c^{-1/2} = U\Sigma V^T$$

- Z: Matrix of standardized residuals.
- P: Matrix of relative frequencies.
- E: Matrix of expected frequencies.
- D_r: Diagonal matrix of row marginal totals.
- D_c: Diagonal matrix of column marginal totals.
- U: Left singular vectors.
- Σ: Diagonal matrix of singular values.
- V^T: Transpose of the right singular vectors.

6. COMPUTE PRINCIPAL COORDINATES AND STANDARD COORDINATES

Principal Coordinates for Rows:

$$F = D_r^{-1/2} U \Sigma$$

- F: Matrix of principal coordinates for rows.
- $D_r^{-1/2}$: Inverse square root of row marginal totals.
- U: Left singular vectors.
- Σ: Diagonal matrix of singular values.

Principal Coordinates for Columns:

$$G = D_c^{-1/2} V \Sigma$$

- G: Matrix of principal coordinates for columns.
- $D_c^{-1/2}$: Inverse square root of column marginal totals.
- V: Right singular vectors.
- Σ: Diagonal matrix of singular values.

Standard Coordinates for Rows:

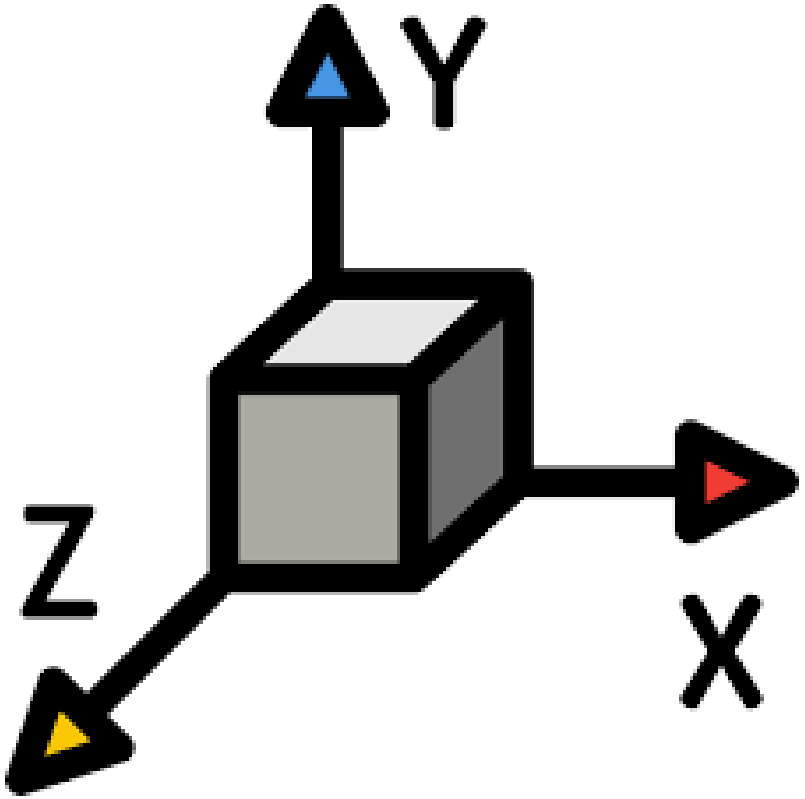
$$\Phi = D_r^{-1/2} U$$

- Φ: Matrix of standard coordinates for rows.
- $D_r^{-1/2}$: Inverse square root of row marginal totals.
- U: Left singular vectors.

Standard Coordinates for Columns:

$$\Psi = D_c^{-1/2} V$$

- Ψ: Matrix of standard coordinates for columns.
- $D_c^{-1/2}$: Inverse square root of column marginal totals.
- V: Right singular vectors.



PRINCIPAL COORDINATES

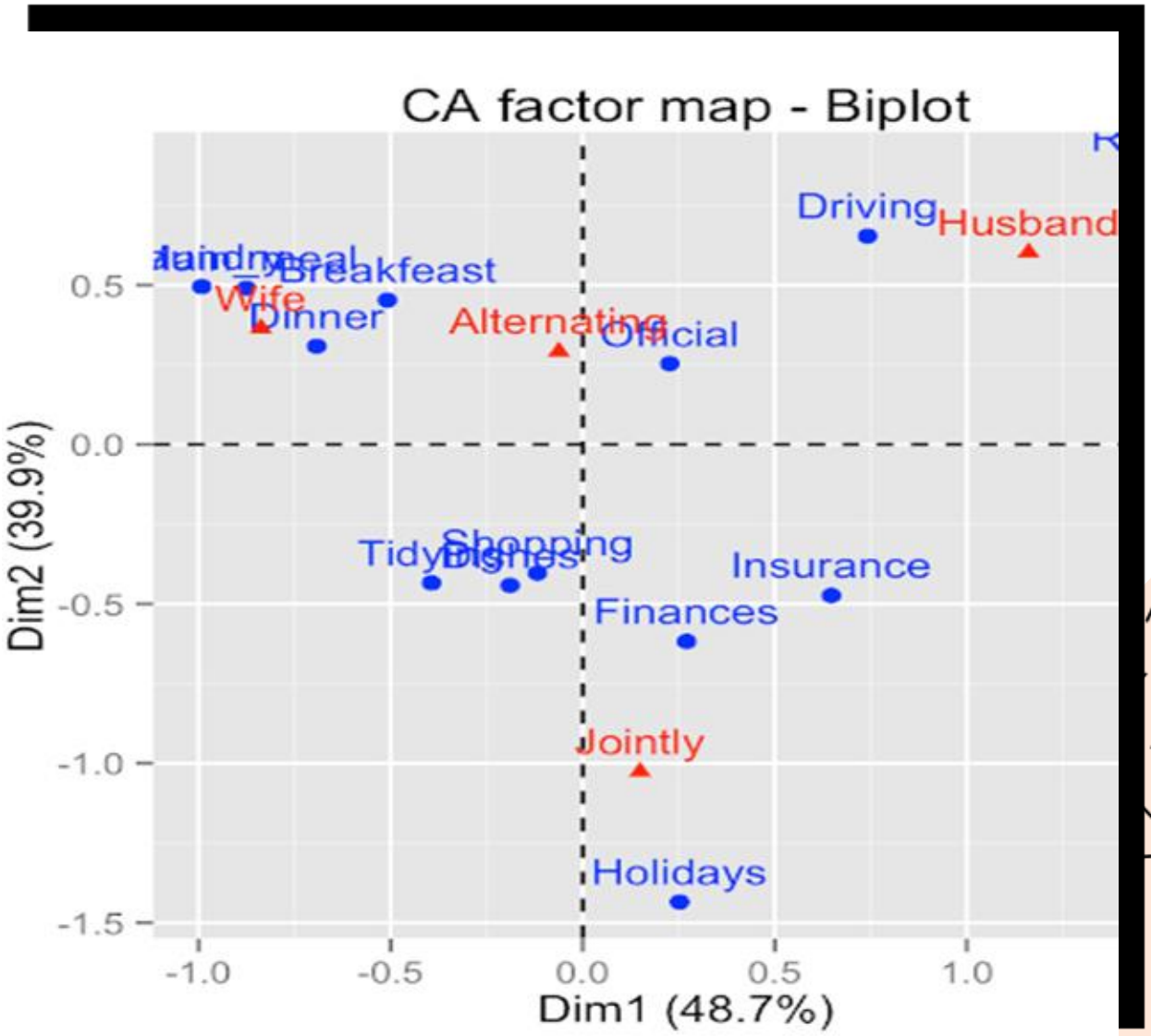
Principal coordinates represent the positions of categories in reduced-dimensional space. These coordinates capture the essential patterns in the data, allowing for a simplified and interpretable representation.

BIPLOT VISUALIZATION

A biplot is a graphical representation that combines information from both rows and columns of a data matrix, providing a way to visualize relationships between variables and observations. Biplots are commonly used in multivariate analysis, including techniques like Correspondence Analysis (CA), Principal Component Analysis (PCA), and more

CORRESPONDENCE ANALYSIS VISUALIZATION

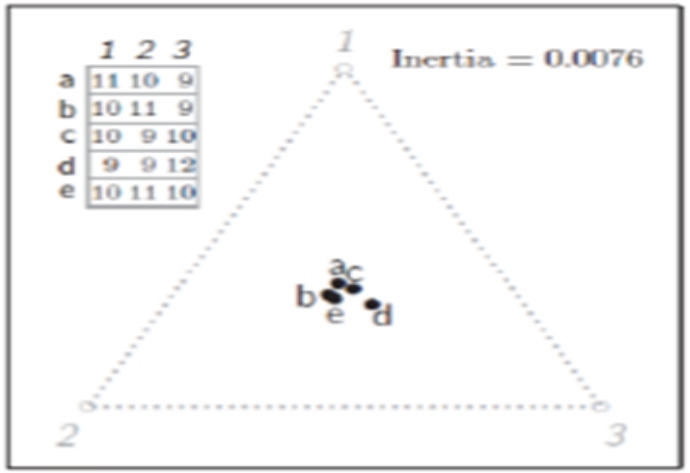
A biplot is a graphical representation that visually captures relationships between rows and columns in a contingency table. It simplifies complex relationships and allows for the interpretation of patterns in the data.

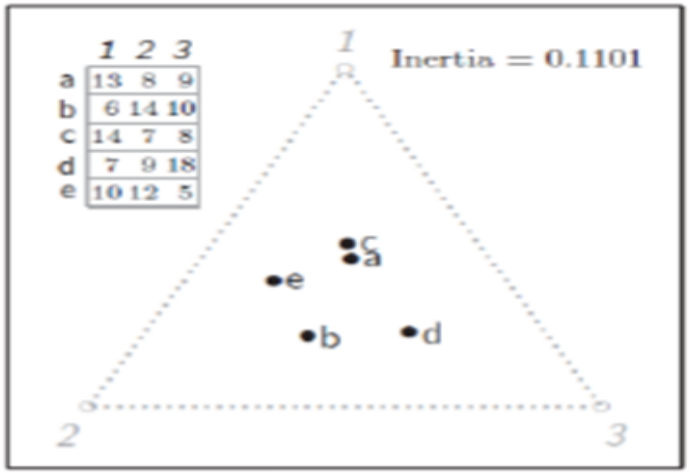


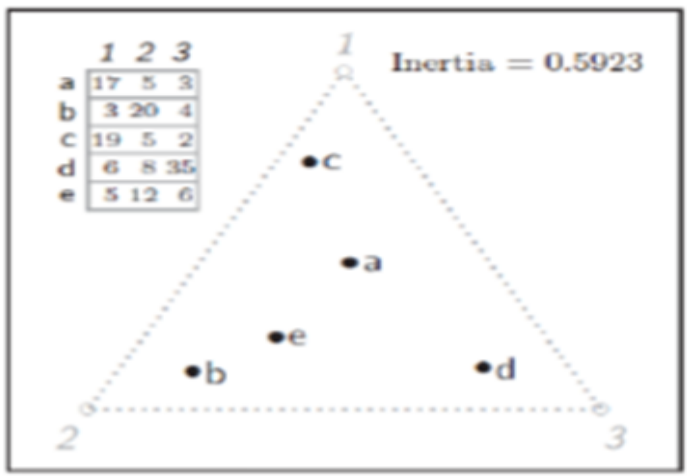
A series of data tables with increasing total **inertia. The higher** the total inertia, the greater is the association between the rows and **columns, displayed** by the higher dispersion of the profile points in the **profile space**

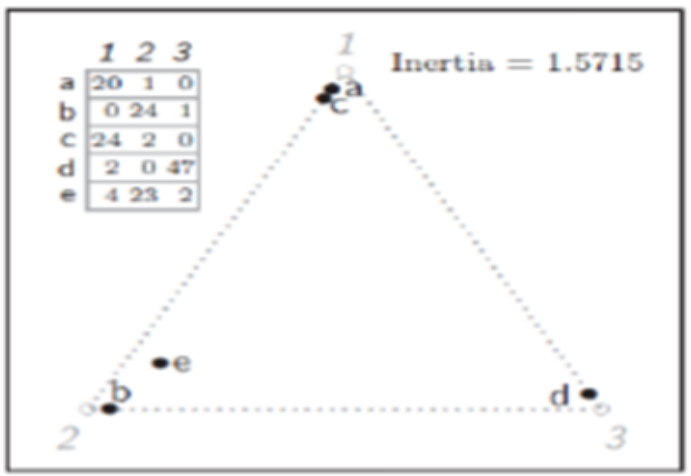
$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(f_{ij}-f_{i.}f_{.j}\right)^2}{f_{i.}\,f_{.j}}$$

Chi-square association index









- When the inertia is low, the row profiles are not dispersed very much and lie close to their average profile. In this case we say that there is low association, or correlation, between the rows and columns.
- **The higher the inertia, the more the row profiles lie closer to the column vertices, i.e., the higher is the row–column association.**

Multiple Correspondence Analysis (MCA)

- MCA is one of the most important methods for the analysis of qualitative or mixed variables;
- MCA is an extension of two-way analysis;
- MCA is particularly suitable for the analysis of survey data and the description of large tables;
- MCA is performed on a table R where:
  - the rows are generally individuals or observations;
  - the columns are the modalities of *nominal variables* (or classes of values of *continuous variables*) and often represent the *modalities of response* relating to the questions of a questionnaire

1. Construct the Condensed Coding Matrix

- **Condensed Coding Matrix**: This matrix R represents the original survey data where rows correspond to individuals and columns to the categories of the variables (e.g., gender, age, occupation).

2. Create the Complete Disjunctive Matrix Z

- **Complete Disjunctive Matrix**: This is a binary matrix where each column corresponds to a category of a variable, and each row represents an individual. An entry is 1 if the individual belongs to the category, and 0 otherwise.

$$Z = [Z_1|Z_2|\ldots|Z_Q]$$

  - Z: Complete disjunctive matrix.
  - Z_q: Disjunctive coding matrix for variable q.
  - Q: Total number of variables.
  - p: Total number of categories (sum of all categories of the variables).

3. Construct Burt's Matrix B

- **Burt's Matrix**: This is the product of the complete disjunctive matrix and its transpose.

$$B = Z'Z$$

  - B: Burt's matrix, which is symmetric and consists of all two-way cross-tabulations of the variables.

4. Perform Correspondence Analysis on Matrix Z

- **Perform CA on Z**: This involves standard steps of Correspondence Analysis but applied to the complete disjunctive matrix Z.

5. Calculate Characteristic Equation in $\mathbb{R}^p$

- **Characteristic Equation**: The characteristic equation is derived from the CA on Z, which involves calculating eigenvalues and eigenvectors.

$$Z'D_r^{-1}ZD_c^{-1}Z' = Z'D_r^{-1}ZD_c^{-1}$$

6. Determine Coordinates and Inertia

- **Row Coordinates**: These are obtained by multiplying the eigenvectors by the square root of the eigenvalues.

$$F = D_r^{-1/2} U \Sigma$$

  - F: Matrix of row coordinates.
  - D_r: Diagonal matrix of row sums.
  - U: Matrix of left singular vectors.
  - Σ: Diagonal matrix of singular values.

- **Column Coordinates**: Similarly, column coordinates are calculated.

$$G = D_c^{-1/2} V \Sigma$$

  - G: Matrix of column coordinates.
  - D_c: Diagonal matrix of column sums.
  - V: Matrix of right singular vectors.
  - Σ: Diagonal matrix of singular values.

7. Interpret Results

- **Inertia**: The total inertia is a measure of the total variance explained by the dimensions. For MCA, the total inertia depends only on the number of variables and their categories, not on the data itself.

$$\text{Total Inertia} = \frac{p - Q}{Q}$$

  - p: Total number of categories.
  - Q: Total number of variables.

1. DATA TRANSFORMATION

- This involves converting each categorical variable into dummy variables. While not a specific formula, the basic principle is creating a 0/1 matrix where:
- Rows represent observations.
- Columns represent categories for each variable.
- Entries are 1 if the observation belongs to that category, otherwise 0.

2.CORRESPONDENCE MATRIX

- This uses the dummy variable matrix to calculate the "chi-square distance" between categories, reflecting their association. Essentially, you calculate:
- Chi-square statistic for each pair of categories within each variable.
- Normalize these values by the total chi-square for the variable.
- The result is a "correspondence matrix" containing association strengths between categories.

3. EIGENVALUE DECOMPOSITION

- This step involves applying linear algebra to the correspondence matrix:
- Decompose the matrix into eigenvalues (representing variance explained) and eigenvectors (defining category and observation coordinates).
- Usually, only the top few dimensions with the highest eigenvalues are retained for analysis.
- This identifies the most important dimensions capturing the data's variance.

4.COORDINATES CALCULATION

Coordinates of the rows (individuals) and columns (categories of the Q variables) of **F**

Row coordinates

$$\hat{\Psi}_\alpha = \frac{1}{Q\sqrt{\mu_\alpha}}\mathbf{Z}\Phi_\alpha \Leftrightarrow \hat{\Psi}_\alpha = \frac{1}{Q\sqrt{\mu_\alpha}}\mathbf{F}\mathbf{D}_p^{-1}u_\alpha$$

Φ_α

Column coordinates

$$\hat{\Phi}_\alpha = \frac{1}{\sqrt{\mu_\alpha}}\mathbf{Z}'\Psi_\alpha \Leftrightarrow \hat{\Phi}_\alpha = \frac{1}{\sqrt{\mu_\alpha}}\mathbf{F}'\mathbf{D}_n^{-1}v_\alpha$$

Ψ_α

5.INTERPRETATION

- This final step involves analyzing the positions of categories and observations in the low-dimensional space:
- Examine distances between points to understand category associations.
- Interpret the meaning of each dimension based on the variable contributions.

Summary:

Correspondence Analysis explores associations in categorical data.

Expected frequencies and chi-square statistics quantify associations.

Principal coordinates and total inertia provide a reduced-dimensional view.

Biplots visually represent relationships for interpretation.

SUMMARY OF PCA, CA, MCA

Principal Component Analysis (PCA)

- Reduce the dimensionality of continuous data while retaining most of the variability present in the dataset.
- Identify the directions (principal components) that maximize the variance in the data.

Correspondence Analysis (CA)

- Explore the relationships between two categorical variables in a contingency table.
- Visualize associations between row and column categories.

Multiple Correspondence Analysis (MCA)

- Extend Correspondence Analysis to more than two categorical variables.
- Analyze and visualize the relationships among multiple categorical variables.

THANK YOU