BHARATHIDASAN UNIVERSITY

Tiruchirappalli- 620024

Tamil Nadu, India.

Programme: M.Sc. Statistics

Course Title: Multivariate Analysis

Course Code: 23ST12CC

Unit-V

Factor Analysis and Canonical Correlations

Dr. T. Jai Sankar
Associate Professor and Head
Department of Statistics

Ms. I. Angel Agnes Mary
Guest Faculty
Department of Statistics

UNIT – V

FACTOR ANALYSIS AND CANONICAL CORRELATIONS

Principal components analysis (PCA)

- Principal components analysis is one of a family of techniques for taking high-dimensional data, and using the dependencies between the variables to represent it in a more tractable, lower-dimensional form, without losing too much information.
- PCA is one of the simplest and most robust ways of doing such dimensionality reduction. It is also one of the oldest, and has been rediscovered many times in many fields, so it is also known as the Karhunen-Loève transformation, the Hotelling transformation, the method of empirical orthogonal functions, and singular value decomposition.

Objectives of Principal components

- To discover or to reduce the dimensionality of the data set.
- To identify new meaningful underlying variables.

Advantages of Principal Component Analysis

- **Dimensionality Reduction:** Principal Component Analysis is a popular technique used for dimensionality reduction, which is the process of reducing the number of variables in a dataset. By reducing the number of variables, PCA simplifies data analysis, improves performance, and makes it easier to visualize data.
- **Feature Selection:** PCA can be used for feature selection, which is the process of selecting the most important variables in a dataset. This is useful in machine learning, where the number of variables can be very large, and it is difficult to identify the most important variables.
- **Data Visualization:** PCA can be used for data visualization. By reducing the number of variables, PCA can plot high-dimensional data in two or three dimensions, making it easier to interpret.
- **Multicollinearity:** PCA can be used to deal with multicollinearity, which is a common problem in a regression analysis where two or more independent variables

are highly correlated. PCA can help identify the underlying structure in the data and create new, uncorrelated variables that can be used in the regression model.

- **Noise Reduction:** Principal Component Analysis can be used to reduce the noise in data. By removing the principal components with low variance, which are assumed to represent noise, Principal Component Analysis can improve the signal-to-noise ratio and make it easier to identify the underlying structure in the data.

- **Data Compression:** Principal Component Analysis can be used for data compression. By representing the data using a smaller number of principal components, which capture most of the variation in the data, PCA can reduce the storage requirements and speed up processing.

- **Outlier Detection:** Principal Component Analysis can be used for outlier detection. Outliers are data points that are significantly different from the other data points in the dataset. Principal Component Analysis can identify these outliers by looking for data points that are far from the other points in the principal component space.

Disadvantages of Principal Component Analysis

- **Interpretation of Principal Components:** The principal components created by Principal Component Analysis are linear combinations of the original variables, and it is often difficult to interpret them in terms of the original variables. This can make it difficult to explain the results of PCA to others.

- **Data Scaling:** Principal Component Analysis is sensitive to the scale of the data. If the data is not properly scaled, then PCA may not work well. Therefore, it is important to scale the data before applying Principal Component Analysis.

- **Information Loss:** Principal Component Analysis can result in information loss. While Principal Component Analysis reduces the number of variables, it can also lead to loss of information. The degree of information loss depends on the number of principal components selected. Therefore, it is important to carefully select the number of principal components to retain.

- **Non-linear Relationships:** Principal Component Analysis assumes that the relationships between variables are linear. However, if there are non-linear relationships between variables, Principal Component Analysis may not work well.

- **Computational Complexity:** Computing Principal Component Analysis can be computationally expensive for large datasets. This is especially true if the number of variables in the dataset is large.

- **Over fitting:** Principal Component Analysis can sometimes result in overfitting, which is when the model fits the training data too well and performs poorly on new data. This can happen if too many principal components are used or if the model is trained on a small dataset.

Extraction of principal components

Principal component analysis (PCA) is a technique for extracting information from a high-dimensional space by reducing it to a lower-dimensional subspace. The process of extracting principal components involves the following steps:

- **Center the data:** Subtract the mean from each dimension of the data to create a centered data matrix with a mean of zero.

- **Calculate the covariance matrix:** Calculate the covariance matrix of the centered data.

- **Calculate the eigenvectors and eigenvalues:** Calculate the eigenvectors and eigenvalues of the covariance matrix.

- **Order the eigenvectors:** Order the eigenvectors from highest to lowest eigenvalues.

- **Select the principal components:** The eigenvectors with the highest eigenvalues are the principal components of the data set.

- **Recast the data:** Recast the data along the principal components axes.

PCA is a feature extraction technique that produces new variables that are linear combinations of the original variables. These new variables, or principal components, are uncorrelated and contain most of the information in the original variables.

PCA can be used to identify the most important variables in a data set. It can also be used to shorten a measurement scale by removing items that are superfluous.

Factor analysis

- Factor analysis is a dimension reduction technique. It is used when in analysis a large number of variables and it is not possible to deal with all the variables simultaneously.
- The factor analysis is classified into two types:
  - ***Exploratory Factor Analysis (EFA):*** The EFA is used when the structure of underlying factors is unknown and is to be determined.
  - ***Confirmatory Factor Analysis (CFA):*** The CFA is used when the structure of underlying factors is already known and it is required to check whether the data collected confirm that structure or not.

Objectives of Factor analysis

- To identifying the underlying dimensions or factors that explain the variation (or correlations) among the set of variables.
- To obtain a new smaller set of uncorrelated variables to replace the original set of correlated variables in subsequent analysis.
- To obtain a smaller set of salient variables from a large set for use in subsequent analysis.

Factoring Methods

There are several different types of factor analysis

- Principal component method
- Principal axes method
- Summation method
- Centroid method

Steps in Exploratory Factor Analysis:

1. Collect data: choose relevant variables.
2. Extract initial factors (via principal component).
3. Choose number of factors to retain.

4. Choose estimation method, estimate model.

5. Rotate and interpret.

6. (a) Decide on changes need to be made (e.g. drop items include items)

   (b) Repeat (4), (5).

7. Construct scales and use on further analysis.

Principal component method

In this method, factors are selected one at a time such that each factor best fits the data. The first fraction is created such that it represents the most highly correlated set of variables. Each subsequent selected factor explains less variance than its predecessor. This procedure is continued till all the factors are selected. All the factors selected explain the largest amount of residual variance in the entire set of standardized response scores.

Centroid Method

1) Obtain the correlation matrix.

2) Obtain grand matrix sum, row sum, column sum.

3) Calculate $N = \frac{1}{Grand\,Total}$

4) Multiply each column sum with N, which gives the first factor loading.

5) To find the second factor loading, find the cross product matrix of factor 1 by testing first factor loading horizontally and vertically and then multiplying corresponding rows and columns.

6) Find the first factor residual matrix given as,

Residuals = total variations - explained variations

$= r_{ij} - l_{ij}$

7) Reflection:

Reflection means that each test vector retains its length but extends in opposite directions. The major purpose of reflection is to get a reflected cost matrix having the highest possible grand total. This step is taken due to the reason that some of the factors loading is with grand total. The method of reflection is by trial and error. This can be done by changing the signs of the variables from positive to negative and negative to positive column wise and row wise. The outcome we get is a reflective residual correlation matrix.

8) Repeat steps from (3) to (7).

Methods for finding number of factors to be extracted

1) **Thumb Rule:** All the interrelated factors must explain at least as much as variances as an average variable. Check, if a variable is under a factor then the percentage of variable explaining variance should be less than the percentage of factor explaining.

2) **Eigen Value Index:** When the Eigen value of a factor is less than 1, it explains less variance then the variables included in the factor itself such a factor should not be considered.

3) **Fruckter Formula:**

$$\text{Number of factors} = \frac{(2n-1)-\sqrt{8n+1}}{2}$$

Where n is the number of variables included in the study.

4) **Residual correlation matrix method:** In this method, the residual correlation is observed and if it is soon that most of the correlation coefficient in this matrix are zero, and then the extraction of the factor can be determinate.

5) **Scree Plot test:** This method is to decide about the number of factors to be retained from the extracted factors. The test determines which of the extracted factors are actually contributing variance and does not measure random errors. The number of factors is plotted against the proportion of variance. It extracts in the order of the extracted factors.

Standardization of Responses

$$\hat{X}_i = \frac{X_i - \overline{X}}{\sigma}$$

where X_i is a value corresponding to a response and σ is the variance.

Factor Loading

Factor loadings play a crucial role in factor analysis, representing the correlation between the variable and the factor. A factor loading of 0.7 or higher typically indicates that the factor sufficiently captures the variance of that variable. These loadings help in determining the importance and contribution of each variable to a factor. All variables load on all factors but they load highly on some specific factors. Range is ±1.

Eigen Values

It is the measurement of the amount of variants explained by a factor. A factor eigen value is the sum of the square of its factor loading.

Communality

It indicates the proportion of variance in the responses to the statement which is explained by the identified factors.

$$Var(X_i) = Var\left(\sum_{j=1}^{p} \lambda_{ij} F_j + e_i\right)$$

$$= \sum_{j=1}^{p} \lambda_{ij}^2 Var(F_j) + Var(e_i)$$

$$= \sum_{j=1}^{p} \lambda_{ij}^2 + \omega_i$$

where, $\sum_{j=1}^{p} \lambda_{ij}^2$ is the proportion of variance in variable i, that comes from the common factors. It is called the communality of variable i. The communality cannot exceed one.

Percentage of Variance

$$= \frac{Eigen\, value\, of\ the\ factor}{Sum\, of\ all\ the\, eigen\, values} \times 100$$

If any variable is not combined in any of the groups, then it can be left or can be considered as another factor. To remove variable which is totally different use rotation i.e., we are basically changing its direction.

Rotation of Factors

For the purpose of simplifying the interpretation of obtained factors and to increase the number of high and low positive loadings in the columns of a factor, factor rotation is used. There are two methods for this:

1) Orthogonal Rotation/ Variance Rotation: Here factors are rotated such that the original factors as well as rotated factors are orthogonal. The line between the factors axis remains 90^o.

2) Promax Rotation: The factors are rotated such that the line between original and rotated factors is more than or less than 90^o.

Advantages of Factor Analysis:

- Both objective and subjective attributes can be used.
- It can be used to identify the hidden dimensions or constraints which may or may not be apparent from direct analysis.
- It is not extremely difficult to do and at the same time its inexpensive and gives accurate results.
- There is flexibility in naming and using dimensions.

Disadvantages of Factor Analysis

- The usefulness depends on the researcher's ability to develop a complete and accurate set of product attributes. If important attributes are missed the value of procedure is reduced accordingly.
- Naming of the factors can be difficult multiple attributes can be highly correlated with no apparent reasons.
- If the observed variables are completely unrelated the factor analysis is unable to produce meaningful pattern.

Confirmatory Factor Analysis

Confirmatory Factor Analysis (CFA) is a sophisticated statistical technique used to verify the factor structure of a set of observed variables. It allows researchers to test the hypothesis that a relationship between observed variables and their underlying latent constructs exists. CFA is distinct from Exploratory Factor Analysis (EFA), where the structure of the data is not predefined and is instead determined through the analysis.

Purpose and Procedure of CFA

The primary goal of CFA is to confirm whether the data fits a hypothesized measurement model based on theory or prior research. This involves several critical steps:

- **Defining Constructs:** The process begins by clearly defining the theoretical constructs. This stage often involves a pretest to evaluate the construct's items and ensure they are well-defined and represent the concept accurately.
- **Developing the Measurement Model:** In CFA, it is essential to establish the concept of unidimensionality, where each factor or construct is represented by multiple observed variables that are presumed to measure only that specific construct. Typically, a good practice involves having at least three items per construct.
- **Specifying the Model:** Researchers must specify the number of factors and the pattern of loadings (which variables load on which factors). This specification is based on theoretical expectations or results from previous studies.
- **Assessing Model Fit:** The validity of the measurement model is assessed by comparing the theoretical model with the actual data. This includes examining factor loadings (with a standard threshold of 0.7 or higher for adequate loadings), and fit indices such as Chi-square, Root Mean Square Error of Approximation (RMSEA), Goodness of Fit Index (GFI), and Comparative Fit Index (CFI).

Assumptions in CFA

- **Multivariate Normality:** The data should follow a multivariate normal distribution.
- **Sample Size:** Adequate sample size is crucial, generally n > 200, to ensure reliable results.
- **Model Specification:** The model should be correctly specified a priori based on theoretical or empirical justification.

- **Random Sampling:** Data must be collected from a random sample to generalize findings.

Concepts in CFA

- **Theory and Model:** A theory is a systematic set of causal relationships explaining a phenomenon, while a model is a specified set of dependent relationships within that theory used for testing.

- **Path Analysis and Diagram:** Path analysis is utilized to test structural equation models, with path diagrams visually representing the cause-effect relationships.

- **Endogenous and Exogenous Variables:** Endogenous variables are outcomes within the model, influenced by other variables, while exogenous variables are predictors not influenced by other variables within the model.

- **Confirmatory Analysis and Cronbach's Alpha:** Confirmatory analysis tests pre-specified relationships, and Cronbach's Alpha assesses the reliability of construct indicators.

- **Identification:** This refers to the ability of the data to provide sufficient information to estimate the model. Models can be under-identified, exactly identified, or over-identified.

- **Goodness of Fit:** This measures how well the model fits the observed data. Fit indices help in evaluating whether the model is acceptable.

Cluster Analysis

- The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering.

- A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters.

- A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression.

Applications of Cluster Analysis

- Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing.
- In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns.
- In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations.
- Clustering may also help in the identification of areas of similar land use in an earth observation database and in the identification of groups of houses in a city according to house type, value and geographic location, as well as the identification of groups of automobile insurance policy holders with a high average claim cost.
- Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity.
- Clustering can also be used for outlier detection, Applications of outlier detection include the detection of credit card fraud and the monitoring of criminal activities in electronic commerce.

Methods of Major Clustering

- ❖ Partitioning Methods
- ❖ Hierarchical Methods
- ❖ Density-Based Methods
- ❖ Grid-Based Methods
- ❖ Model-Based Methods

Partitioning Methods

A partitioning method constructs k partitions of the data, where each partition represents a cluster and $k \leq n$. That is, it classifies the data into k groups, which together satisfy the following requirements:

- Each group must contain at least one object, and
- Each object must belong to exactly one group.

A partitioning method creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another.

Hierarchical Methods

A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed.

- The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one or until a termination condition holds.

- The divisive approach, also called the top-down approach, starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds.

Density-based methods

- Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shapes.

- Other clustering methods have been developed based on the notion of density. Their general idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold; that is, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise (outliers) and discover clusters of arbitrary shape.

- DBSCAN and its extension, OPTICS, are typical density-based methods that grow clusters according to a density-based connectivity analysis. DENCLUE is a method that clusters objects based on the analysis of the value distributions of density functions.

Grid-Based Methods

- Grid-based methods quantize the object space into a finite number of cells that form a grid structure.
- All of the clustering operations are performed on the grid structure i.e., on the quantized space. The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space.
- STING is a typical example of a grid-based method. Wave Cluster applies wavelet transformation for clustering analysis and is both grid-based and density-based.

Model-Based Methods

- Model-based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model.
- A model-based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points.
- It also leads to a way of automatically determining the number of clusters based on standard statistics, taking noise or outliers into account and thus yielding robust clustering methods.

Classical Partitioning Methods

The most well-known and commonly used partitioning methods are

- The k-Means Method
- k-Medoids Method

Centroid-Based Technique

The K-Means Method

The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid or center of gravity. The k-means algorithm proceeds as follows.

- First, it randomly selects k of the objects, each of which initially represents a cluster mean or center.
- For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean.
- It then computes the new mean for each cluster.
- This process iterates until the criterion function converges.
- Typically, the square-error criterion is used, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2$$

where E is the sum of the square error for all objects in the data set, p is the point in space representing a given object and m_i is the mean of cluster C_i.

The k-means partitioning algorithm

The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input

- k is the number of clusters
- D is a data set containing n objects.

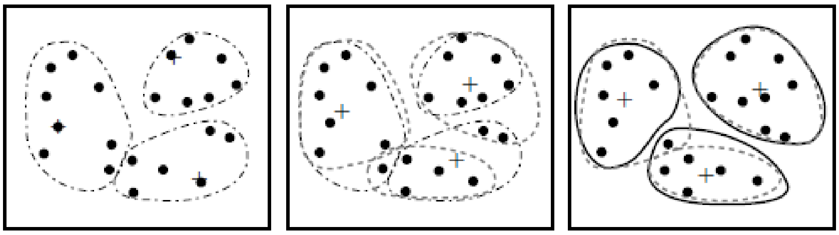
Output

- A set of k clusters.

Method

- arbitrarily choose k objects from D as the initial cluster centers
- repeat
- re-assign each object to the cluster to which the object is the most similar based on the mean value of objects in the cluster
- update the cluster means, i.e., calculate the mean value of the objects for each cluster

- Until no changes.



Clustering of a set of objects based on the k-means method

Hierarchical Clustering Methods

- A hierarchical clustering method works by grouping data objects into a tree of clusters.
- The quality of a pure hierarchical clustering method suffers from its inability to perform adjustment once a merge or split decision has been executed. That is, if a particular merge or split decision later turns out to have been a poor choice, the method cannot backtrack and correct it.

Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up or top-down fashion.

Agglomerative hierarchical clustering

- This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied.
- Most hierarchical clustering methods belong to this category. They differ only in their definition of intercluster similarity.

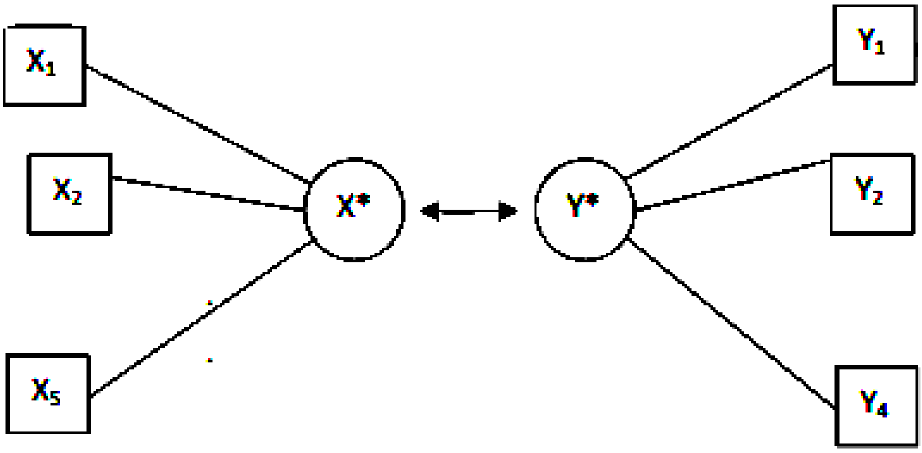
Divisive hierarchical clustering

- This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster.
- It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions, such as a desired number of clusters is obtained or the diameter of each cluster is within a certain threshold.

Canonical correlation

Canonical correlation analysis is a multivariate statistical model used to study the interrelationships among sets of multiple dependent variables and multiple independent variables. This technique is distinct from the multiple regression model in the sense that multiple regression predicts a single dependent variable from a set of multiple independent variables whereas canonical correlation simultaneously predicts multiple dependent variables from multiple independent variables.

For example, as a student of economics, may like to know the association between economic inequality X and political instability Y. The economic inequality can be measured by five variables i.e. (i) the division of farmland (X_1), (ii) the gini coefficient (X_2), (iii) the percentage of tenant farmers (X_3), (iv) the gross national product (X_4) and (v) the percentage of farmers (X_5). Similarly the political instability can be measured by four variables (indicators) i.e. (i) the instability of leadership (Y_1), (ii) the level of internal group violence (Y_2), (iii) the occurrence of internal war (Y_3) and (iv) stability of democracy (Y_4). These two theoretical concepts X and Y can be called two sets of variables or canonical variables. These can be shown in the following figure:



Canonical variables and canonical correlations

The purpose of canonical correlation analysis is to find the correlation between a linear combination of the variables in one set and a linear combination of the variables in another set. The idea behind this approach is first to determine the pair of linear combinations having the largest correlation. Next, we determine the pair of linear combinations having the largest correlation among all pairs uncorrelated with the initially selected pair, and so on. This pairs of linear combinations are called the canonical variables and their correlations are called canonical correlations.

Types of canonical correlations

There are three types of correlations

- Correlation between X variables, the correlation matrix is R_{xx}.
- Correlation between Y variables, the correlation matrix is R_{yy}.
- Correlation between X and Y variables, the correlation matrix is $R_{xy} = R'_{xy}$.

Assumptions of Canonical Correlation

- The correlation coefficient between any two variables is based on linear relationship.
- The canonical correlation is the linear relationship between the variates.
- The distribution of variables is normal.
- Hetro-scedasticity, to the extent it decreases the correlation between variables.

Derivation of canonical correlation coefficients

In 1935-36, Hotelling proposed a method, known as Canonical Correlation Analysis to investigate "linear" relationship between the two sets of variates.

James Press (2005) has expressed the whole idea of this Canonical Correlation Analysis in the following words:

"The Canonical correlation" model selects weighted sums of variables from each of the two sets to form new variables in each of the sets, so that the correlation between the new variables in "different sets" is maximized while the new variables within each set are constrained to be uncorrelated with mean zero and unit variance.

Let

$\alpha : p_1 \times 1$ and $\gamma : p_2 \times 1$ be two unknown vectors to be determined such that the correlation between $\alpha'Y$ and $\gamma'Z$ be as large as possible.

So, let

$U_1 = \alpha'Y$, $V_1 = \gamma'Z$ and $\rho(U_1, V_1)$ = Correlation coefficient between U_1 and V_1.

The problem of correlation coefficient now amounts to the following:

Maximize $\rho(U_1, V_1)$

Subject to

$$Var(U_1) = Var(V_1) = 1$$

and $E(U_1) = E(V_1) = 0.$

Hotelling solved this problem using the celebrated method of Lagrange Multipliers. However, we shall just state the final results. Hotelling showed that this maximization problem is equivalent to following algorithm:

Step 1: Solve for λ the equation

$$\begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix} = 0 \qquad \text{----------- (A)}$$

Here, $\Sigma_{11} = Var(Y)$

$\Sigma_{22} = Var(Z)$

$\Sigma_{12} = \Sigma_{21} = Var(Y, Z)$

Let λ_1 be the largest positive root of the above equation.

Step 2: Now, solve the system of equations

$$\begin{pmatrix} -\lambda_1\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda_1\Sigma_{22} \end{pmatrix}\begin{pmatrix} \alpha \\ \gamma \end{pmatrix} = 0$$

for α and γ.

Mathematically, it is not so simple to solve the above system of linear equations. However, there is an equivalent formulation. To compute α and γ, we solve the pair of equations:

$$\left(\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \lambda_1^2\Sigma_{11}\right)\alpha = 0$$

$$\left(\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \lambda_1^2\Sigma_{22}\right)\gamma = 0$$

Step 3: We now call these α and γ as α_1 and γ_1 respectively and

$U_1 = \alpha_1' Y$ and $V_1 = \gamma_1' Z$ as First Canonical Variates.

It will turn out that λ_1 will be the correlation coefficient between U_1 and V_1. Therefore, we write $\lambda_1 = \rho(U_1, V_1)$ and call this as First Canonical Correlation.

Step 4: We now proceed to the next iteration. We now define

$U_2 = \alpha_2' Y$ and $V_2 = \gamma_2' Z$

where α_2 and γ_2 are to be determined.

Maximize $\rho(U_2, V_2)$

Subject to

$$Var(U_2) = Var(V_2) = 1$$

and $E(U_2) = E(V_2) = 0.$

We repeat the above procedure to compute α_2 and γ_2 as solution of

$$\begin{pmatrix} -\lambda_2 \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda_2 \Sigma_{22} \end{pmatrix} \begin{pmatrix} \alpha_2 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

where λ_2 is the second largest positive root of the equation (A).

Step 5: We continue this procedure until the smallest positive root.

Step 6: The result is now to be collected in a vector format

$$U = \begin{pmatrix} U_1 \\ \vdots \\ U_{p_1} \end{pmatrix} = \begin{pmatrix} U_1 & \cdots & U_{p_1} \end{pmatrix}'$$

and

$$V = \begin{pmatrix} V_1 \\ \vdots \\ V_{p_1} \end{pmatrix} = \begin{pmatrix} V_1 & \cdots & V_{p_1} \end{pmatrix}'$$

The elements of U and V are called Canonical Variates and λ_i is corresponding canonical correlation.

Concepts of multidimensional scaling

Multidimensional Scaling (MDS) is a statistical technique that visualizes the similarity or dissimilarity among a set of objects or entities by translating high-dimensional data into a more comprehensible two- or three-dimensional space. MDS is particularly used in fields such as psychology, sociology, marketing, geography, and biology, where understanding complex structures is crucial for decision-making and strategic planning. There are three classes of MDS

- **Classical MDS** is chosen when the distance data are Euclidean and accurate preservation of these distances is crucial.
- **Metric MDS** is suitable when distances are non-Euclidean or when the scale of measurement levels varies.
- **Non-metric MDS** is beneficial for qualitative data or when only the order of distances (not the actual distances) matters.

1. Classical Multidimensional Scaling

Classical Multidimensional Scaling is a technique that takes an input matrix representing dissimilarities between pairs of items and produces a coordinate matrix that minimizes the strain.

Mathematically, strain is defined as:

$$Strain_D(x_1,x_2,\ldots,x_n)=\left(\frac{\sum_{i,j}(b_{ij}-x_i^T x_j)^2}{\sum_{i,j} b_{ij}^2}\right)^{1/2}$$

where,

x_i denotes vectors in an N-dimensional space.

$x_i^T x_j$ denotes the scalar product between x_i and x_j.

b_{ij} are the elements of the matrix B.

2. Metric Multidimensional Scaling

Metric Multidimensional Scaling generalizes the optimization procedure to various loss functions and input matrices with known distances and weights. It minimizes a cost function called "stress," often minimized using a procedure called stress majorization.

Stress is defined as a residual sum of squares:

$$Stress_D(x_1,x_2,\ldots,x_n) = \sqrt{\sum_{i\neq j=1,\cdots,n}\left(d_{ij}-\left\|x_i-x_j\right\|\right)^2}$$

3. Non-metric Multidimensional Scaling

Non-metric Multidimensional Scaling finds a non-parametric monotonic relationship between dissimilarities and Euclidean distances between items, along with the location of each item in the low-dimensional space. It defines a "stress" function to optimize, considering a monotonically increasing function f.

$$S(x_1,\cdots,x_n;f)=\frac{\sum_{i<j}(f(d_{ij})-\hat{d}_{ij})^2}{\sum_{i<j}\hat{d}_{ij}^2}$$

where

- d_{ij} are the observed dissimilarities between pairs of items i and j.
- $\hat{d}_{ij}$ are the distances between items i and j in the lower-dimensional space.
- $f(d_{ij})$ is a monotonic transformation of the observed dissimilarities d_{ij} to best approximate the distances $\hat{d}_{ij}$ in the reduced space.
- The summation $\Sigma_{i<j}$ is taken over all pairs of items.

Comparison with Other Dimensionality Reduction Techniques

Dimensionality Reduction Technique	**Objective**	**Visualization**	**Applicability**	**Interpretation**
Multidimensional Scaling (MDS)	Preserves original pair-wise distances or dissimilarities	Provides intuitive visualizations of similarities/dissimilarities	Suitable for data with known dissimilarities or similarities, applicable across various domains	Emphasizes the preservation of relationships, facilitating qualitative interpretation
Principal Component Analysis (PCA)	Maximizes variance along orthogonal axes	Efficient for capturing global structure but may not preserve pair-wise distances	Suitable for linear data transformations, often used for feature extraction	Focuses on capturing variance, useful for dimensionality reduction in high-dimensional data

Correspondence analysis

Correspondence analysis is a statistical methodology utilized for identifying and visualizing the hidden patterns and associations between categorical variables in multivariate data where variables have discrete categories rather than numerical values.

Similar to Principal Component Analysis (PCA) for continuous data, Correspondence Analysis aims to uncover patterns and associations in categorical data by reducing the dimensionality of the data while preserving the essential relationships between variables· It effectively condenses large volumes of data into a simplified visual representation, demonstrating the relationships and correspondences between different categories of variables.

CA starts with a contingency table, which displays the frequencies of occurrences for each combination of categories between two categorical variables· This table serves as the basis for analyzing associations between the variables.

Concepts of Correspondence Analysis

- **Contingency Table:** A contingency table is a matrix that displays the frequency distribution of the variables. Each cell in the table represents the count or frequency of occurrences for the combination of row and column categories.
- **Singular Value Decomposition (SVD):** Singular Value Decomposition is a mathematical technique used in Correspondence Analysis to decompose the contingency table into its principal components. This decomposition helps in reducing the dimensionality of the data while preserving as much variability as possible.
- **Chi-Square Distance:** Chi-square distance is used to measure the association between categories in the contingency table. It helps in quantifying the differences between observed and expected frequencies, which is crucial for identifying significant relationships.

Applications of Correspondence Analysis

Correspondence Analysis (CA) is a useful technique in various scenarios where categorical data analysis is involved.

- **Exploring Relationships in Categorical Data:** When you have categorical variables and want to understand how they are related or associated with each other, CA can provide insights into the structure of these relationships.

- **Large Contingency Tables:** When dealing with large contingency tables with many rows and columns, it can be challenging to interpret the relationships between categories. CA helps by reducing the dimensionality of the data and visualizing the associations in a more manageable way.

- **Visualization of Multivariate Data:** CA provides graphical representations, such as biplots, that allow for the visualization of multivariate categorical data. These visualizations can reveal patterns, clusters, or trends in the data that may not be apparent from the raw contingency table.

- **Market Research:** In market research, CA can be used to analyze consumer preferences, brand associations, and market segmentation based on categorical survey data. It helps identify relationships between consumer demographics, product features, and purchasing behavior.

- **Social Science Research:** CA is valuable in social science research for analyzing survey responses, studying relationships between demographic variables, and exploring patterns in categorical data collected from social surveys.

- **Text Analysis:** In text analysis, CA can be used to explore relationships between words in documents, such as identifying word co-occurrences, document clustering, and thematic analysis.

- **Ecology and Biology:** CA is also applied in ecology and biology for analyzing species abundance data, community composition, and environmental variables.

Procedure for Correspondence Analysis

In correspondence analysis it transforms the contingency table into a lower-dimensional space, typically two dimensions, in order to visualize and interpret the relationships between the rows and columns of the table.

Step 1: Data Collection

Find or collect data relevant to the question trying to answer. This data should consist of categorical variables, like types of customers, movie genres, or answer choices on a survey.

Step 2: Preparing the Data for Analysis

Before starting the analysis, it is essential to structure the data to make it suitable for this type of statistical comparison. Pre-processing can involve converting categorical data into numbers, combining categories into broader groups if there are too many within a variable, and reviewing data to ensure it is complete and high-quality for reliable results. This is like cleaning and organizing workspace before a project.

- If the data isn't already in categories, might need to group it (e.g., income brackets instead of individual income values).
- Double-check for any missing information or errors and fix them if possible.
- Organize data into a contingency table. Imagine a grid where rows represent categories from one variable and columns represent categories from another. Each cell in the table shows how many times a specific combination of categories occurs (e.g., how many young adults prefer comedies).

Step 3: Contingency Table Construction

Create a contingency table (also known as a cross-tabulation table) that displays the frequency counts of each combination of categories for the variables being analyzed. Each cell in the table represents the count of observations falling into a particular combination of categories.

- **Normalization:** Normalize the contingency table to adjust for differences in marginal totals (row and column totals). This step ensures that the analysis focuses on the relative frequencies of observations rather than absolute counts.
- **Expected Values Calculation:** Calculate the expected values for each cell of the contingency table under the assumption of independence between the variables. These expected values represent what would be expected if there were no association between the variables.
- **Residual Computation:** Compute residuals by subtracting the expected values from the observed values in each cell of the contingency table. Residuals quantify the deviations from independence and indicate the strength and direction of the associations between categories.

- **Eigenvalue Decomposition:** Apply eigenvalue decomposition to the matrix of residuals. This mathematical technique decomposes the matrix into eigenvectors and eigenvalues, revealing the underlying structure in the data.
- **Dimensionality Reduction:** Retain a subset of the most significant eigenvectors (usually two or three) to represent the data in a lower-dimensional space. This reduces the complexity of the data while preserving the most important relationships between categories.

Step 4: Visualize and Interpret the Results

Visualize the results of Correspondence Analysis using biplots, which display both the rows and columns of the contingency table in the reduced-dimensional space. Biplots allow for the interpretation of relationships between categories and the identification of clusters or patterns in the data. It is important to remember that CA reveals associations, not necessarily causation. Just because two categories are close on the biplot doesn't mean one causes the other.

The key to interpreting the biplot lies in the distances between the points. The closer two points are, the stronger the association between the corresponding categories. For example, if comedies and the "young adult" category are close on the biplot, it suggests a strong connection – young adults tend to favor comedies in this data set.

SCL PROBLEMS

1. Agglomerative Cluster

Given the dataset {a, b, c, d and e} and the following distance matrix, construct a dendrogram by complete-linkage hierarchical clustering using the agglomerative method.

	a	b	c	d	E
a	0	9	3	6	11
b	9	0	7	5	10
c	3	7	0	9	2
d	6	5	9	0	8
e	11	10	2	8	0

Procedure

- Assignment of each data item to its own cluster, so that we have (N = 5) cluster, each containing just one item.

Data set = {a, b, c, d, e}

Initial cluster set c_1 = {a}, {b}, {c}, {d}, {e}

- Find the closed pair of clusters and merge them into a single cluster. Now we have one less cluster.

Calculation

Distance Table of c_1

	a	b	c	d	e
a	0	9	3	6	11
b	9	0	7	5	10
c	3	7	0	9	2
d	6	5	9	0	8
e	11	10	2	8	0

Minimum distance between {c} and {e} is,

d({c}, {e}) = 2

New set of clusters c_2: {a}, {b}, {d}, {c, e}

Compute the distance between the new cluster and each of the old clusters.

d({c, e}, {a}) = Max(d{c, a}, d{e, a})

= Max (3, 11) = 11

d({c, e}, {b}) = Max(d{c, b}, d{e, b})

= Max (7, 10) = 10

d({c, e}, {d}) = Max(d{c, d}, d{e, d})

= Max (9, 8) = 9

Distance Table of c$_2$

	a	b	d	{c, e}
a	0	9	6	11
b	9	0	5	10
d	6	5	0	9
{c, e}	11	10	9	0

Repeat previous steps until all items are clustered into a single cluster of size N.

Minimum distance between {b} and {d} is,

d({b}, {d}) = 5

Now, new set of clusters c$_3$: {a}, {b, d}, {c, e}

Compute the distance between the new cluster and each of the old clusters.

d({b, d}, {a}) = Max(d{b, a}, d{d, a})

= Max (9, 6) = 9

d({b, d}, {c, e}) = Max(d{b, c}, d{b, e}, d{d, c}, d{d, e})

= Max (7, 10, 9, 8) = 10

Distance Table of c$_3$

	a	{b, d}	{c, e}
a	0	9	11
{b, d}	9	0	10
{c, e}	11	10	0

Minimum distance between {a} and {b, d} is,

d({a}, {b, d}) = 9

Now, new set of clusters c$_4$: {a, b, d}, {c, e} d({a, b, d}, {c, e}) = Max(d{a, c}, d{a, e}, d{b, c}, d{b, e}, d{d, c}, d{d, e})

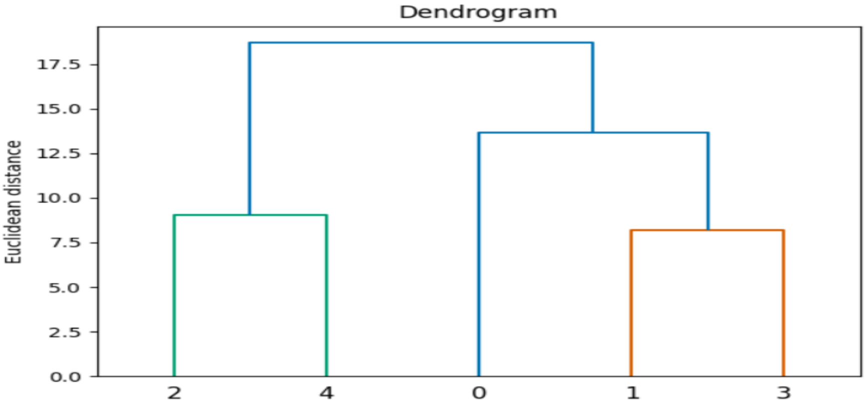
= Max (3, 11, 7, 10, 9, 8) = 11

Distance Table of c$_4$

	{a, b, d}	{c, e}
{a, b, d}	0	11
{c, e}	11	0

Conclusion

New set of clusters c$_5$: {a, b, c, d, e}



2. K-Means Cluster Analysis

Use k-means clustering algorithm to divide the following data into two clusters:

X$_1$	1	2	2	3	4	5
X$_2$	1	1	3	2	3	5

Procedure

k = 2, choose randomly 2 cluster centers, say v$_1$ = (2, 1) (2, 3)

Euclidean Distance: $(x_1, x_2)(y_1, y_2) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$

Calculation

Data points:	a$_1$ (1, 1)	a$_2$ (2, 1)	a$_3$ (2, 3)	a$_4$ (3, 2)
	a$_5$ (4, 3)	a$_6$ (5, 5)		

$$a_1(1,1)\ \ v_1(2,1) = \sqrt{(1-2)^2 + (1-1)^2} = 1$$

$$a_1(1,1)\ \ v_2(2,3) = \sqrt{(1-2)^2 + (1-3)^2} = 2.24$$

$$a_2(2,1)\ \ v_1(2,1) = \sqrt{(2-2)^2 + (1-1)^2} = 0$$

$$a_2(2,1)\ \ v_2(2,3) = \sqrt{(2-2)^2 + (1-3)^2} = 2$$

$$a_3(2,3)\ \ v_1(2,1) = \sqrt{(2-2)^2 + (3-1)^2} = 2$$

$$a_3(2,3)\ \ v_2(2,3) = \sqrt{(2-2)^2 + (3-3)^2} = 0$$

$$a_4(3,2)\ \ v_1(2,1) = \sqrt{(3-2)^2 + (2-1)^2} = 1.41$$

$$a_4(3,2)\ \ v_2(2,3) = \sqrt{(3-2)^2 + (2-3)^2} = 1.41$$

$$a_5(4,3)\ \ v_1(2,1) = \sqrt{(4-2)^2 + (3-1)^2} = 2.83$$

$$a_5(4,3)\ \ v_2(2,3) = \sqrt{(4-2)^2 + (3-3)^2} = 2$$

$$a_6(5,5)\ \ v_1(2,1) = \sqrt{(5-2)^2 + (5-1)^2} = 5$$

$$a_6(5,5)\ \ v_2(2,3) = \sqrt{(5-2)^2 + (5-3)^2} = 3.61$$

Distance table -1

Data points	Distance from v_1 (2, 1)	Distance from v_2 (2, 3)	Assigned Center
a_1 (1, 1)	1	2.24	v_1
a_2 (2, 1)	0	2	v_1
a_3 (2, 3)	2	0	v_2
a_4 (3, 2)	1.41	1.41	v_1
a_5 (4, 3)	2.83	2	v_2
a_6 (5, 5)	5	3.61	v_2

K-means Clustering

Cluster 1 and v_1 = { a_1, a_2, a_4}

Cluster 2 and v_2 = { a_3, a_5, a_6}

v_1 = (1/3) [(1, 1) + (2, 1) + (3, 2)]

= (1/3) (6, 4)

v_1 = (2, 1.33)

v_2 = (1/3) [(2, 3) + (4, 3) + (5, 5)]

= (1/3) (11, 11)

v_2 = (3.67, 3.67)

Data points: a_1 (1, 1) a_2 (2, 1) a_3 (2, 3) a_4 (3, 2)
a_5 (4, 3) a_6 (5, 5)

v_1 = (2, 1.33) V_2 = (3.67, 3.67)

$$a_1(1,1)\ \ v_1(2,1.33) = \sqrt{(1-2)^2 + (1-1.33)^2} = 1.05$$

$$a_1(1,1)\ \ v_2(3.67,3.67) = \sqrt{(1-3.67)^2 + (1-3.67)^2} = 3.78$$

$$a_2(2,1)\ \ v_1(2,1.33) = \sqrt{(2-2)^2 + (1-1.33)^2} = 0.33$$

$$a_2(2,1)\ \ v_2(3.67,3.67) = \sqrt{(2-3.67)^2 + (1-3.67)^2} = 3.15$$

$$a_3(2,3)\ \ v_1(2,1.33) = \sqrt{(2-2)^2 + (3-1.33)^2} = 1.67$$

$$a_3(2,3)\ \ v_2(3.67,3.67) = \sqrt{(2-3.67)^2 + (3-3.67)^2} = 1.8$$

$$a_4(3,2)\ \ v_1(2,1.33) = \sqrt{(3-2)^2 + (2-1.33)^2} = 1.204$$

$$a_4(3,2)\ \ v_2(3.67,3.67) = \sqrt{(3-3.67)^2 + (2-3.67)^2} = 1.8$$

$$a_5(4,3)\ \ v_1(2,1.33) = \sqrt{(4-2)^2 + (3-1.33)^2} = 2.605$$

$$a_5(4,3)\ \ v_2(3.67,3.67) = \sqrt{(4-3.67)^2 + (3-3.67)^2} = 0.75$$

$$a_6(5,5)\ \ v_1(2,1.33) = \sqrt{(5-2)^2 + (5-1.33)^2} = 4.74$$

$$a_6(5,5)\ \ v_2(3.67,3.67) = \sqrt{(5-3.67)^2 + (5-3.67)^2} = 1.88$$

Distance table - 2

Data points	Distance from v_1 (2, 1.33)	Distance from v_2 (3.67, 3.67)	Assigned Center
a_1 (1, 1)	1.05	3.78	v_1
a_2 (2, 1)	0.33	3.15	v_1
a_3 (2, 3)	1.67	1.8	v_1
a_4 (3, 2)	1.204	1.8	v_1
a_5 (4, 3)	2.605	0.75	v_2
a_6 (5, 5)	4.74	1.88	v_2

K-means Clustering

Cluster 1 and v_1 = { a_1, a_2, a_3, a_4}

Cluster 2 and v_2 = { a_5, a_6}

Recalculating the cluster centers,

v_1 = (1/4) [(1, 1) + (2, 1) + (2, 3) + (3, 2)]

= (1/4) (8, 7)

v_1 = (2, 1.75)

v_2 = (1/2) [(4, 3) + (5, 5)]

= (1/2) (9, 8)

V_2 = (4.5, 4)

Data points:	a_1 (1, 1)	a_2 (2, 1)	a_3 (2, 3)	a_4 (3, 2)
	a_5 (4, 3)	a_6 (5, 5)		

v_1 = (2, 1.75) V_2 = (4.5, 4)

$$a_1(1,1) \quad v_1(2,1.75) = \sqrt{(1-2)^2 + (1-1.75)^2} = 1.25$$

$$a_1(1,1) \quad v_2(4.5,4) = \sqrt{(1-4.5)^2 + (1-4)^2} = 4.61$$

$$a_2(2,1) \quad v_1(2,1.75) = \sqrt{(2-2)^2 + (1-1.75)^2} = 0.75$$

$$a_2(2,1) \quad v_2(4.5,4) = \sqrt{(2-4.5)^2 + (1-4)^2} = 3.905$$

$$a_3(2,3) \quad v_1(2,1.75) = \sqrt{(2-2)^2 + (3-1.75)^2} = 1.25$$

$$a_3(2,3) \quad v_2(4.5,4) = \sqrt{(2-4.5)^2 + (3-4)^2} = 2.69$$

$$a_4(3,2)\quad v_1(2,1.75)=\sqrt{(3-2)^2+(2-1.75)^2}=1.03$$

$$a_4(3,2)\quad v_2(4.5,4)=\sqrt{(3-4.5)^2+(2-4)^2}=2.5$$

$$a_5(4,3)\quad v_1(2,1.75)=\sqrt{(4-2)^2+(3-1.75)^2}=2.36$$

$$a_5(4,3)\quad v_2(4.5,4)=\sqrt{(4-4.5)^2+(3-4)^2}=1.12$$

$$a_6(5,5)\quad v_1(2,1.75)=\sqrt{(5-2)^2+(5-1.75)^2}=4.42$$

$$a_6(5,5)\quad v_2(4.5,4)=\sqrt{(5-4.5)^2+(5-4)^2}=1.12$$

Distance table - 3

Data points	Distance from v_1 (2, 1.75)	Distance from v_2 (4.5, 4)	Assigned Center
a_1 (1, 1)	1.25	4.61	v_1
a_2 (2, 1)	0.75	3.905	v_1
a_3 (2, 3)	1.25	2.69	v_1
a_4 (3, 2)	1.03	2.5	v_1
a_5 (4, 3)	2.36	1.12	v_2
a_6 (5, 5)	4.42	1.12	v_2

Cluster 1 of v_1 = a_1(1, 1), a_2(2,1), a_3(2, 3), a_4(3, 2)

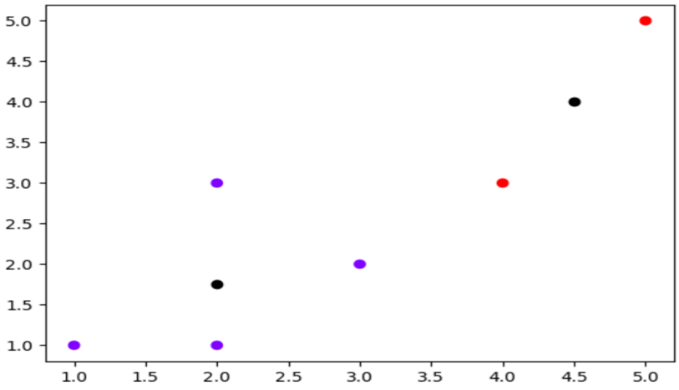
Cluster 2 of v_2 = a_5(4, 3), a_6(5, 5)

Result

Cluster elements are same in the previous iteration.

Cluster 1 of v_1 = a_1(1, 1), a_2(2,1), a_3(2, 3), a_4(3, 2)

Cluster 2 of v_2 = a_5(4, 3), a_6(5, 5)


5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0