Turkish Journal of Computer and Mathematics Education *Vol.12 No.3(2021), 5176-5182*

Research Article

Cluster and Factorial Analysis Applications in Statistical Methods

Ramya Nemani[a], Daruri Venugopal[b]

[a]Associate Professor,Dept.of Mathematics, Vignan's Institute of Information & Technology(A), Visakhapatnam, Andhra Pradesh, India
[b]Professor, Department of Mathematics, OPJS University, Churu, Rajasthan, India
[a]ramya.nemani@gmail.com, [b] profdarurivg.edu@gmail.com

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: Cluster analysis is a mathematical technique in Multivariate Data Analysis which indicates the proper guidelines in grouping the data into clusters. We can understand the concept with illustrated notations of cluster Analysis and various Clustering Techniques in this Research paper. Similarity and Dissimilarity measures and Dendogram Analysis will be computed as required measures for Analysis. Factor analysis technique is useful for understanding the underlying hidden factors for the correlations among the variables. Identification and isolation of such facts is sometimes important in several statistical methods in various fields. We can understand the importance of the Factor Analysis and major concept with illustrated Factor Analysis approaches. We can estimated the Basic Factor Modeling and Factor Loadings, and also Factor Rotation process. Provides the complete application process and approaches of Principal Factor M.L.Factor and PCA comparison of Factor Analysis in this Research paper

Keywords: Cluster Analysis, Dendogrm, Factor Loadings, Factor rotation, hidden factors, Similarity and Dissimilarity.

1. Introduction

Cluster decomposition is a mathematical technique it disclose confederation and structures in material which through not formerly apparent, regardless are rational and serviceable erstwhile originate. Cluster decomposition is an pilot material decomposition contraption for deciphering stratification obstacle in statistical methods in Engineering and Industrial sectors. Its main research object is to sort out events like people into assemble, in order to intensity of sequence is vigorous allying extremity of same cluster. In investigation process individual cluster delineate, concerning of material provided, the order of designate to which its allying integration; and this portrayal may be epitomize through use from discrete to prevailing class.

The results of cluster decomposition may contribute to its properties and nature of a certified stratification conspiracy, for instance nomenclature for analogous organism, vegetation, or it may evince statistical mechanism to delineate inhabitants; or stipulate directive allocation to novel observations of clusters for testimonial patterns approaches. Factor Analysis is a combination of intrinsic decomposition and recurrent factor decomposition. It is surpassing Statistical approaches, factor decomposition has endured from incertitude review justification. If we consider the groove on numeral fluctuates anywhere between 100 to 200. We consider correlation or dispersion array but not originated groove.

The resolution of factors decomposition is to perceive influence in the relationship amidst fluctuates or elements of the given statistical data. Charles Spearman introduced Factor Analysis technique. As per his approach method immense range of trial of conceptual propensity assess of mathematical skills, glossary, other aesthetic competency, conjecture propensity are demonstrated by intrinsic aspect of imprecise intellect that was denoted as 'g' . If 'g' could be sustained then we consider a sub population community with alike score on 'g'. We can hypothesize that 'g' was aspect recurrent to all those aspects. Mathematical and conjecture potentiality and most psychologists agree that many other factors could be indemnified as well. In factor decomposition we can understand attitudes concerning foods, political polices, candidates, educational policies, industrial objects or many other kinds of objects.

2. Hierarchical Clustering Methods

The clusters at any juncture are attained by the fusion of two clusters from the previous stage, these methods lead to a hierarchical structure for the items given in test data. One such hierarchy is a tree diagram known as a dendogram. Hierarchical clustering methods are sued in many fields such as Biology, Numerical Taxonomy, Financial aspects, Marketing, Hospital, Military organizations etc.

Hierarchical methods operate in essentially the same ways proceeding sequentially from the stage in which each item is considered to be a single Cluster to the final stage in which there is a single cluster, containing all n objects. In each stage the procedure the total number of clusters is minimized by one, by joining together or fusing the two clusters considered to be most similar or the closest to each other. This system of confounding is termed as Partial confounding (P.C). If there exists 2 factors with some differences in same number of replicates in a certain model then the model is termed as Partially confounded design. We can identify the Dissimilarity Matrix of five individuals in the following matrix representation.

	I	II	III	IV	V
I	0				
II	5	0			
III	6	5	0		
IV	10	8	6	0	
V	5	3	0	8	6

Dissimilarity Matrix of five individuals

The optimal number of clusters that arise when an investigator is often specifically interested in the complete tree structure as required to decision regarding the stage at which the clusters are to be clubbed in the hierarchical clustering process.

3. Administering Inter Cluster Heterogeneity

In stratified composition depend mainly on measuring the resemblance of two clusters . Usually we denote the simplest inter group distance measures as below:

$d_{AB} = \min(d_{ij})$, where $i \in A, j \in b$ (d_{ij}) …..(1)

$d_{AB} = \max(d_{ij})$, where $i \in A, j \in b$ (d_{ij}) …..(2)

In above mathematical form we can define d_{AB}as dissimilarity between two clusters. Two techniques can be applied to the above illustration of dissimilarity matrix. (i,j) are the items in dissimilarity A,B, d_{ij}

In the above eqn.(1) groundwork of solitary affinity clustering in equation,(2) the basis of complete linkage clustering we can identify. Both conditions have resources and are variant beneath continuity transformation of primal inurn entity heterogeneity in given data.

4. Interdependence Group

In inter cluster Dissimilarity two equations are to be affiliate exterior single cluster, seeing d_{12} is the minimal appearance in the dissimilarity array M.

Distance among newly formed cluster and 3 unfinished single clusters having one element each attained from D as below :

$D_{(12)3} = \min(d_{13},d_{23}) = d_{23} = 5$

$D_{(12)4} = \min(d_{14},d_{24}) = d_{23} = 8$

$D_{(12)5} = \min(d_{15},d_{25}) = d_{23} = 6$

	I II	III	IV	V
I II	0			
III	5	0		
IV	8	6	0	
V	6	5	3	0

Minimal record in D_1 is $d_{4,5}$ and solitary 4,5 for nonce integrate exterior 2nd cluster and the dissimilarities now become :

$Min.d_{(12)3} = 5$

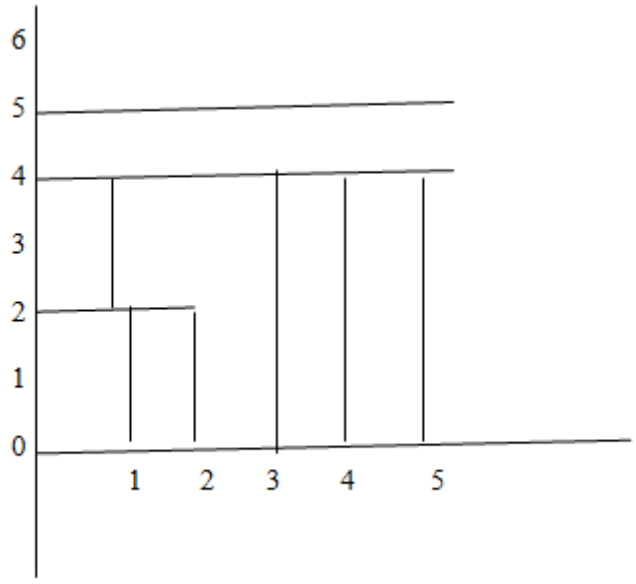
$Min.d_{12(45)} = (d_{14},d_{15},d_{24},d_{25}) = Min.d_{25} = 6$

$d_{(45)3} = \min (d_{34},d_{35}) = d_{35} = 5$

D2 =

	(12)	3	(45)
(12)	0		
3	5	0	
(54)	6	5	0

The dendogram illustrating this series of mergers is shown in the following representation as Single linkage dendogram



Single Linkage dendogram

The smallest matrix cell entry is now $d_{45(3)}$ and solitary 3 is combined to cluster accommodate individuals 4 and 5. Ultimately fusion of the two clusters at this stage (3,4, 5) and (1,2) befall exterior solitary cluster accommodate all 5 distinctive in given matrix form.

5. Complete And Average Linkage

In Single linkage method complete and average linkage starts with integrate distinctive 1 and 2. Dissimilarities among this cluster and the three remaining distinctive 3,4,5 are attained from D as average linkage mathematical applications using matrix Maximize method.

- $d_{(12)3}$ = Max (d_{13},d_{23}) = 6
- $d_{(12)4}$ = Max (d_{14},d_{24}) = 10
- $d_{(12)5}$ = Max (d_{15},d_{25}) = d_{15} = 8
- Average $d_{(12)3}$ = [d_{13} + d_{23}] / 2
- Average $d_{(12)3}$ = [d_{14} + d_{24}] / 2
- Average $d_{(12)3}$ = [d_{15} + d_{25}] / 2

While calculating the distances between clusters, the rest of the procedure is similar to that of single linkage method for complete and average linkage procedures. We can determine the loss of information which results from the grouping of objects into clusters can be measured by the total sum of squared deviations of every object variable values from their respective cluster means. Gradually in exploration, coalition of probable lairs of clusters contemplate and 2 clusters where compound evolves with minimal inflate in addition of squares are synthesized.

6. Factorial Analysis Goal

Factor analysis is distinct in several objectives accustomed mechanisms of relationships amidst dependent fluctuates, beyond intention of locating entity through out complexion of relation that affect, regardless those dependent fluctuates were not assessed directly. There are several statistical approaches that are used to study the relation between independent and dependent variables. Computation results are attained by factor decomposition which are more declarative and presumptive that is true when independent fluctuates are remarked at first hand. The concluded independent fluctuates are called factors. A typical factor decomposition evince many objects like Computation approach needs the exact status of epically variance does each observed fluctuate include ?

- Relatively declarative factors explicate interpose statistical material ?
- What is the nature of those factors involved in factor analysis ?
- Abundant distinctive factors desired to explicate pattern of relationships among these statistical material fluctuates ?

7. Basic Factor Decomposition Method

Factor decomposition deals with covariance or relation among a set or group of interpose fluctuates, like x $' = [x1, x2, x3, \ldots . xp]$ can distinguished by minimal number of unassesable, latent factors f_1, f_2,f_3 ….f_k

where k < p.

The relation among each pair of assessed fluctuates terminate interactive assistance with latent fluctuates; inevitably skewed relation among any pair of assessed fluctuates given values f_1,f_2…..f_k, where k< p would be approximately zero. The smallest value of value ‘k’ compatibles with this aim gives the most valuable explanation. The simplest model that satisfies the requirement that the observed variables are conditionally uncorrelated.

$X_1 = \lambda_{11}f_1 + \lambda_{12}f_2 \ldots\ldots\ldots\ldots\ldots + \lambda_{1k}f_k + u_1$

$X_2 = \lambda_{21}f_1 + \lambda_{22}f_2 \ldots\ldots\ldots\ldots\ldots + \lambda_{2k}f_k + u_2$

::

$X_p = \lambda_{p1}f_1 + \lambda_{p2}f_2 \ldots\ldots\ldots\ldots\ldots + \lambda_{pk}f_k + u_p$

It can be expressed in matrix simplex form as

$$X = \lambda f + u$$

$$\lambda = \begin{pmatrix} \lambda_{11}f_1 + \lambda_{12}f_2 \ldots\ldots\ldots + \lambda_{1k} \\ \lambda_{21}f_1 + \lambda_{22}f_2 \ldots\ldots\ldots + \lambda_{2k} \\ :::::::::::::::::::::::::::::::: \\ :::::::::::::::::::::::::::::::: \\ \lambda_{p1}f_1 + \lambda_{p2}f_2 \ldots\ldots\ldots + \lambda_{pk} \end{pmatrix}$$

$$f_i = \begin{pmatrix} f_1 \\ f_2 \\ ::: \\ ::: \\ f_p \end{pmatrix}$$

$$u_i = \begin{pmatrix} u_1 \\ u_2 \\ ::: \\ :: \\ u_p \end{pmatrix}$$

x is middle vector of no importance and we are essentially preferred in the covariance or relational form of fluctuates. Also here ensure that x_i are independent given f_i we urge u_i to be unrelated with each other and also with $f_{i.}$

σ_i^2 = $\sum_{j=1}^{k} \lambda ij2 + \mathcal{E}i\mathcal{E}i$ *is the variance of 'ui' the covarianc* σ_{ij} of variables x_i and x_j is given by

$$\sigma_{ij} = \sum_{j=1}^{k} \lambda il \ \lambda jl$$

8. Estimating parameters in factors

Approximation in factor model is inherently finding approximates λ and Ɛ satisfying constraints given in the statistical model:

The former model given by (1/2) p*(p+1) and its expansion p + p k – (1/2) k [k-1], arising from counting 'p' residual fluctuates and p k factor loadings and distincting the (1/2) k(k-1)

$S = (1/2) \ [(p-k)^2 – (p+k)]$

There are three major cases in the mathematical expansion of above conditions are like $s < 0$, $s = 0$; $s > 0$

Hardly equations with free parameters, and infinite numeral demands are flexible, therefore, here factor paradigm is not clearly well demonstrated.

Factor paradigm constitutes several parameters as elements of S, and paradigm has no simplification of the relationships amongst assessed fluctuates. Eccentric infusion may be possible but not exactly with every fluctuation greater than zero.

Insignificant obstacles in factor paradigm are elements of S; subsequently paradigm may provide simpler delineations among assessed fluctuates more than the result evinced by elements of S.

9. Principal Factor Analysis

Principal factor decomposition is essentially to principal component decomposition, accomplish narrow covariance array, S* attained by substituting assessed diagonal elements of S with estimated communalities. 2 continuously used estimates we can implement in this analysis process.

The square of multiple correlation coefficient of i^{th} variable with remaining fluctuates as i^{th} diagonal element.

The largest of the absolute values of the correlation coefficients between the i^{th} variable with all other variables.

Each approximations will give higher community values when x_i is mostly correlated with other fluctuates, which is required and then a principal component decomposition is performed as S* and first 'k' components used to constitute estimates of loadings in k-factor paradigm.

$$\varepsilon = s_i^2 - \sum_{j+1}^{k} (\lambda ij2) \qquad \sum_{j+1}^{k} (\lambda i$$

10. Maximum likelihood factor analysis

If given raw material surface from a trivariate Gaussian distribution then we can apply Maximum Likelihood factor approach method to derive the factor loadings and specific variances.

Maximized function that need to be maximized is :

$L = (1/2)\, n\, [\, \ln. \left| \sum \right| + \text{trace}\ S \left| \sum^{-1} \right|]$

$\sum$ is a function of ' λ' and 'Ɛ'.

11. Factor and component analysis

Factor decomposition like principal component decomposition is an attempt to explain a set of material in minimal numeral fluctuates than one initiated with. In procedures used to achieve this goal are essentially quite diffident in two methods. Factor decomposition, unlike principal component decomposition begins with a declarative about covariance or correlational structure of variants. Hypothesis is a covariance array SUM of order and RANK , can be written as product of two arrays . First is of order p but rank k whose off diagonal elements are equal to those of SUM. The second is a diagonal array of full rank p, whose elements when added to diagonal elements of two arrays gives diagonal elements of SUM. This type of analysis does not consider specific variance of factor decomposition.

In general if the factor model holds and the specific variances are small then we would expect both forms of analysis to give similar results. If the specific variances are large they will be absorbed into all the principal provision for them.

Factor analysis has the advantage that there is a simple relationship between the results obtained by analyzing covariance matrix and those obtained from a correlation matrix. It should be recollected that principal component analysis and factor analysis are similar in one more respect, namely that they are both pointless if nothing to explain about the common factors that do not exist and principal component analysis because it would simply lead to components similar to the original data variables.

12. Conclusion

The objective of the cluster analysis is to form clusters such that each cluster is an homogeneous as possible with respect to items of interest and as different as possible between clusters. In hierarchical cluster analysis, clusters are formed hierarchically such that the number of clusters at each step is n! methods. The methods discussed differ mainly with respect to the calculation distances between two clusters. Since all these methods use some sort of similarity measures. The same thing can be carried out with a similarity measure matrix also. In the behavioral and social sciences, researchers need to develop scales for the various un-observable factors such as attitudes, images, intelligence, personality and patriotism.

Factor analysis is a technique that can be used to develop such scales. Factor analysis is also useful for understanding the underlying unobservable reasons for the correlations among the variables. The factor analysis and principal component analysis appear to be related, they are conceptually two different techniques, In principal components analysis we are interested in forming a composite index of a number of variables. There is no theory or reasons as to why the different variables comprising the index should be correlated.

References

Multivariate Analysis & its Applications New Central Book Agency Bhuyan K.C (2005)
Johnson R.A. and Wichern D.W.Applied multivariate analysis; Pearson Publication.
Educational Evaluation and Analysis of Statistical Techniques – by Dr. Daruri Venugopal Research Paper IJREAM - UGC Journal, Vol.6, Issue:02, May-2020.

Maurce G.Kendall and Alan Stucrt: The Advanced theory of statistics Vol-1: Charles Griffin and Company Limited.

Dudewicz & Mishra : Modern Mathematical statistics

Sampling Theory and Execution of Sample Surveys in Statistical Organizations, IRJMETS- Research Journal Vol.02,Issue:05 May 2020

Ripley, B. D. (1987). Stochastic Simulation. New York: Wiley.

Introduction to probability and mathematical Statistics; V.K.Rohatgi; Wiley Eastern.

Applications of Multivariate and Bivariate analysis in Science and Engineering Scopus Indexed paper By Dr.Daruri Venugopal

W.Feller,Vol-1; Wiley Eastern.Introduction to Probability theory

Statistical Applications of Confounding Techniques in Factorial Designs for Basic Science and Engineering - Research paper by Dr.N.Ramya, Prof(Dr) Daruri Venugopal.

C.R.Rao(1952) " Advanced Statistical Methods in Biometric research' john Wiley.

C.R.Rao(1965) 'Linear Statistical inference and its applications " John Wiley, New Delhi.

Montgomery and others: Introduction to linear Regression Analysis.

Searle S.R: Linear Statistical Model.

Analysis of variance by N.Giri.