A COMPARATIVE STUDY ON GROUPING OF VARIABLES USING MULTIVARIATE TECHNIQUES

LATHA V, DR. P RAJALAKSHMI

Abstract: Most of the data sets available for researchers have several characteristics which are recorded on each unit or individual with a multivariate data structure. It is of interest to examine the relationship among these variables. Multivariate analysis consists of techniques which can be used to explore and describe such data. Some of the techniques available in multivariate analysis for grouping variables are classification, principal component analysis, factor analysis, cluster analysis and structural equation modelling. In this paper, an attempt has been made to study the above techniques by applying them to a psychometric data set relating to 8 variables, which are subscales of a widely used personality measure from a sample of 500 psychotherapy outpatients: anxiety , hostility , depression , self- consciousness , warmth , gregariousness , assertiveness and positive emotions .A report of the comparative study of these techniques is discussed.

Key words: Cluster analysis, Factor analysis, Principal Component Analysis

Introduction: The need to understand the relationships between variables makes multivariate analysis an inherently interesting area of learning. Multivariate analysis consists of a collection of methods that can be used when several measurements are made on each individual or object in one or more samples Multivariate data analysis methods are used for making inferences about the structure of the mean and covariance of several variables, for exploring the pattern of data that may exist in one or more dimensions and for modelling relationships among variables.

Certain aspects in the social and behavioural sciences are not well defined and some terms like social class, extrovert or introvert personality, public opinion etc. are variables which cannot be directly observed but are hypothetical constructs which help the researcher to understand an area of interest which has no method for direct measurement. Such variables are called latent variables. Factor analysis is an interdependence technique used to understand the underlying structure among the variables and examine the relationship between variables by determining whether the information available can be condensed by using a smaller set of factors. Factors represent the constructs that summarize the observed variables. If factor analysis is used as a data reduction method by searching for a structure among a group of variables then it is an exploratory tool. But if a researcher has some idea about the actual structure of the data, then the problem reduces to assessing the degree to which the data meets the expected structure. In this case factor analysis is a confirmatory tool. Factor analysis determines the number of latent variables or factors and the nature of these variables which account for variation and covariation among a set of observed variables called indicators. A latent factor is an unobservable variable that influences more than one observed variable and accounts for the correlation between these observed measures. It studies the covariation among a set of variables.

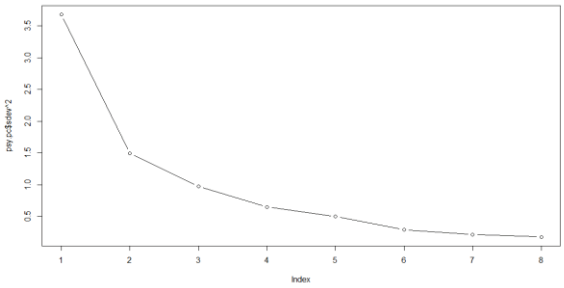
Let $X=(x_1,x_2,...x_p)'$ be the set of observed variables which are assumed to be linked to unobservable latent variables $f_1,f_2,...f_k$, $k<p$ and the residuals $U=(u_1,u_2,...u_p)'$ by a model of the form

$X=\Lambda F+U$ where $\Lambda=\begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1k} \\ \vdots & \ddots & \vdots \\ \lambda_{p1} & \cdots & \lambda_{pk} \end{pmatrix} F=\begin{pmatrix} f_1 \\ \vdots \\ f_p \end{pmatrix}$. It is assumed that the residual terms $u_1,u_2,...u_p$ are uncorrelated with each other and they are uncorrelated with the factors $f_1,f_2,...f_k$. Further it is also assumed that the latent factors occur in the standard form with mean zero and variance 1 and are uncorrelated with one another. The elements in Λ are called factor loadings. Factor analysis model partitions the variance of each observed variable into two parts- common variance and unique variance. Common variance is the variance shared with the other variables through the common factor. The second part is the specific or unique variance which is the variability in a specific variable not shared by other variables.

Confirmatory factor analysis is used when the researcher starts with a hypothesis about the common factors and their relationship with the observed variables. This field developed with the work of Joreskog(1966) .Confirmatory factor analysis is used when there is a need to test a hypothesis about how a set of latent variables can be studied using another set of observed variables which are called indicators. An attempt is made to assess and explain the structure in a set of correlated variables which can be observed in terms of a small number of latent variables.

Data Analysis: The data set on Neuroticism and extraversion from Thomas A Brown(2015) has been

used for data analysis. The hypothesized model is drawn from the vast literature on the five-factor model of personality. In this example, a researcher has collected eight measures (subscales of a widely used personality measure) from a sample of 500 psychotherapy outpatients: anxiety (N1), hostility(N2), depression (N3), self-consciousness(N4), warmth (E1), gregariousness (E2),assertiveness (E3), and positive emotions (E4). A two-factor model is used in which the observed measures of anxiety, hostility, depression, and self-consciousness are conjectured to load on a latent dimension of Neuroticism, and the observed measures of warmth, gregariousness, assertiveness, and positive emotions are predicted to load on to a distinct factor of Extraversion. The Latent factors are neuroticism and extraversion. Neuroticism is a higher-order personality trait in the study of psychology characterized by anxiety, fear, moodiness, worry, envy, frustration, jealousy, and loneliness. Extraversion is defined as a behavior where someone enjoys being around people more than being alone. An example of extraversion is when someone always likes to be around people and enjoys being the center of attention. The correlation matrix and the scree plot have been obtained using the stats package of R software.



Correlation Matrix

	A	B	C	D	E	F	G	H
A	1.0000	0.7793	0.4116	0.3663	0.2800	0.2266	0.4174	-0.1178
B	0.7793	1.0000	0.4628	0.3112	0.2726	0.1928	0.4522	-0.1222
C	0.4116	0.4628	1.0000	0.2855	0.2688	0.2179	0.3654	-0.1462
D	0.3663	0.3112	0.2855	1.0000	0.7843	0.6958	0.5386	-0.0824
E	0.2800	0.2726	0.2688	0.7843	1.0000	0.7478	0.6053	-0.0734
F	0.2266	0.1928	0.2179	0.6958	0.7478	1.0000	0.4559	-0.1192
G	0.4174	0.4522	0.3654	0.5386	0.6053	0.4559	1.0000	-0.0964
H	-0.1178	-0.1222	-0.1462	-0.0824	-0.0734	-0.1192	-0.0964	1.0000

As the elbow of the scree plot is between 2 and 3, factor analysis was conducted with two and three factors and the results were obtained. Factor analysis conducted with two factors showed a cumulative variation of 56% while that with three factors showed a high uniqueness value for factors C and H and the cumulative variation indicated that only 59.6% of the variation was being accounted for in this case. The results of factor analysis conducted with two factors and three factors are given below.

For three factors Uniquenesses:

A	B	C	D	E	F	G	H
0.187	0.205	0.706	0.252	0.130	0.337	0.439	0.978

Loadings:

	Factor1	Factor2	Factor3
A	0.156	0.880	-0.115
B	0.880	0.102	
C	0.176	0.474	0.196
D	0.823	0.267	
E	0.895	0.185	0.189
F	0.805	0.122	
G	0.513	0.424	0.345
H	-0.131		

	Factor1	Factor2	Factor3
SS loadings	2.457	2.092	0.218
Proportion Var	0.307	0.262	0.027
Cumulative Var	0.307	0.569	0.596

For 2 factors Uniquenesses:

A	B	C	D	E	F	G	H
0.299	0.131	0.730	0.275	0.134	0.348	0.513	0.978

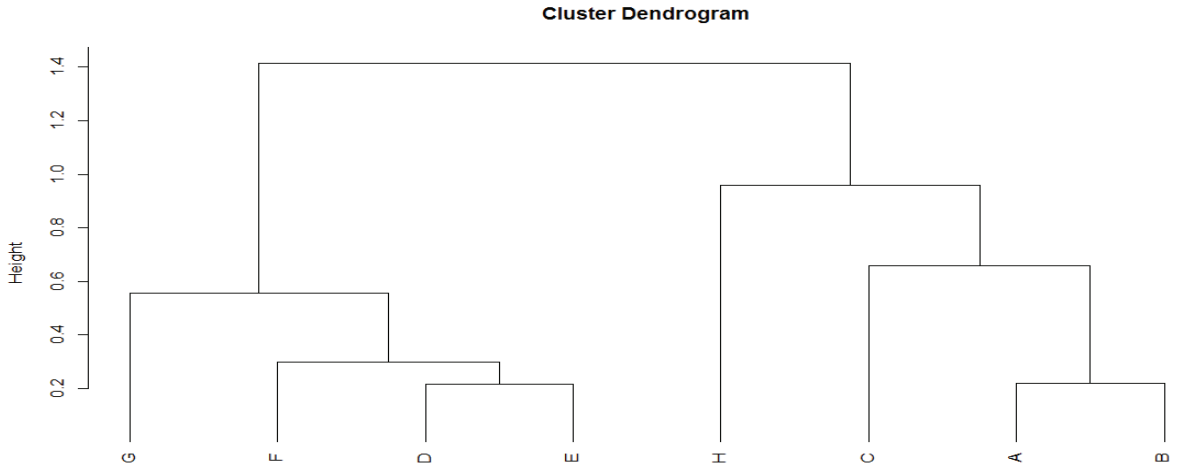
Loadings:

	Factor1	Factor2
A	0.147	0.824
B		0.927
C	0.201	0.479
D	0.810	0.261
E	0.910	0.193
F	0.798	0.125
G	0.552	0.427
H		-0.130

	Factor1	Factor2
SS loadings	2.502	2.088
Proportion Var	0.313	0.261
Cumulative Var	0.313	0.574

For data analysis and management it is vital to classify or group data into categories or clusters. In order to understand a new object or phenomenon, one tries to look at the features that describe it and also compare it with other objects or phenomena based on similarity according to certain rules or standards. Classification systems are either supervised or unsupervised depending on whether they assign new inputs to one of a finite number of discrete supervised classes or unsupervised categories respectively. In supervised classification, the various groups or categories are labeled and an object is classified as belonging to one of these groups. In unsupervised classification or clustering, no labeled

data is available. The aim of clustering is to separate a finite unlabeled data set into a finite and discrete set of objects which exhibit similar characteristics so that there is maximum dissimilarity between clusters. Clustering algorithms partition data into a number of clusters such that internal homogeneity is maximum and external separation is also maximum. Clustering of variables is a data reduction technique which helps in obtaining maximum information from the minimum number of variables.



Using the ClustofVar package in R, a dendrogram for the data set with eight variables has been obtained using hierarchical clustering, which shows a grouping of the variables A, B, C and H in one group and the variables D, E, F and G in another group.

The model for confirmatory factor analysis is the same as that of the exploratory factor analysis model with the assumption that the latent factors are allowed to be correlated which is not so in the exploratory case. Using confirmatory factor analysis on the same data set with two latent factors as suggested by Brown(2015)by taking the first four variables under one latent factor and the next four under the second, it is observed that the value of the chi square statistic is very high and the model is not a good fit for the data.

The following information was also obtained

Model Chi square = 567.4533 Df = 19 Pr(>Chisq) = 3.675228e-108

AIC = 601.4533

BIC = 449.3757

Normalized Residuals

Min. 1st Qu. Median Mean 3rd Qu. Max.

-2.787 -1.398 0.000 1.037 1.105 13.360

R-square for Endogenous Variables

A B C D E F G H

0.7627 0.7642 0.2619 0.1945 0.8916 0.6145 0.4141 0.0097

The analysis was then conducted grouping the variables A, B ,C and H as the first factor and the remaining variables as the second factor and it was found that statistically it was a better fit as compared to the previous case where the grouping was done based on the theory in psychology. The present grouping was as per the cluster analysis grouping of variables using hierarchical clustering.

Model Chi square = 124.7415 Df = 19 Pr(>Chisq) = 1.431708e-17

AIC = 158.7415

BIC = 6.663983

Normalized Residuals

Min. 1st Qu. Median Mean 3rd Qu. Max.

-1.988000 -0.732800 0.000002 0.317000 0.209700 5.232000

R-square for Endogenous Variables

A B C H D E F G

0.7418 0.8131 0.2575 0.0211 0.7318 0.8514 0.6403 0.4147

When we observe the correlation matrix also, it indicates that the variables H and C do not contribute any extra information as the value of the correlation coefficient is low.

When the variables H and C were considered as one group and the analysis was done with 3 factors, the result indicated that the model was under- identified.

Principal component analysis indicated the first four components contributed 92% of the total variation and since each component is a linear combination of the observed variables , the loadings indicate the contribution of each variable to the particular component.

Standard deviations:

```
 Comp.1   Comp.2   Comp.3    Comp.4   Comp.5   Comp.6   Comp.7   Comp.8
25.373614 15.632768 11.681802  8.365667  6.730084  4.862945  2.998532  2.791578
 Importance of components:
                            Comp.1       Comp.2       Comp.3       Comp.4       Comp.5
Standard deviation       25.3736141 15.6327679   11.6818024    8.36566748    6.73008425
Proportion of Variance    0.5454352    0.2070381    0.1156108    0.05928976    0.03837245
Cumulative Proportion     0.5454352    0.7524733    0.8680841    0.92737386    0.96574631
                             Comp.6       Comp.7       Comp.8
Standard deviation       4.86294548   2.998531628   2.791577923
Proportion of Variance   0.02003445   0.007617208   0.006602038
Cumulative Proportion    0.98578075   0.993397962   1.000000000
Loadings:
  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
A -0.740 -0.281  0.600
B -0.553 -0.217 -0.735 -0.314
C                -0.208 -0.973
D -0.115  0.220         0.112  0.302 -0.891  0.197
E -0.177  0.444         0.314  0.704  0.398
F -0.233  0.761  0.127 -0.476 -0.344
G -0.193  0.204 -0.267  0.744 -0.539
H         -0.991
```

Conclusion: The correlation values obtained from the correlation matrix indicated that correlation was significant for variables A and B, and moderate for the variables D, E and F. C and H have low correlation indicating that they might not contribute significant information to the group.

- Clustering of variables by hierarchical clustering indicated that the variables A, B, C and H formed one group and the remaining variables D, E , F and G formed the second group.
- Factor analysis was performed with two and three factors which indicated groupings which supported the cluster analysis results. Further increase in the number of factors was not possible since the Hessian could not be computed. The grouping done by clustering was tested for model fit using confirmatory factor analysis which indicated that the two factor grouping by cluster analysis gave a better fit as compared to the factor grouping proposed by T.A. Brown(2015)from a psychological perspective. Brown's grouping with variables A,B,C and D as the first factor and with variables E,F, G and H as the second factor has a high chi square value indicating that the model is not a good fit from the statistical point of view.
- The analysis of the data set using principal component indicated that the variables A and B contribute significantly to the first principal component , variables E and F to the second, H appears separately in the third, E and G in the sixth, E and D in the seventh and the variable C in the eighth component.
- Comparison of the observations from these methods indicate that the variables C and H do not contribute significantly to any statistical grouping pattern and can be considered as atypical or noise variables.

References:

1. Everitt B.S and G Dunne (2001), Applied multivariate data analysis. Second Edition, John Wiley and sons
2. Everitt B.S, S Landau, M Leese and D Stahl (2011), Cluster analysis, Fifth Edition, John Wiley and Sons
3. Chavent .M(2011),ClustofVar, an R package
4. Fox.J(2006), Package 'sem' in R
5. Hair J, Black W, Babin B and anderson R, Multivariate data analysis, Seventh Edition, 2010, Pearson
6. Timothy A Brown (2015), Confirmatory Factor Analysis for Applied Research. Second Edition, The Guilford Press, New York
7. www.guilford.com/brown3-materials

**

Latha V , Department of statisitcs,Jyoti Nivas College, Autonomous, Bangalore
Dr. P Rajalakshmi, Department of Statistics, Bangalore University, Bangalore