Multivariate Data Analysis using R

Samarendra Das, A. R. Rao
Email: samarendra.das@icar.gov.in
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-12

Background

Multivariate statistical techniques simultaneously analyze measurements on multiple variables for each individual under investigation and are widely used in plant breeding programs. The main purpose of multivariate data analysis is to study the relationships among the (multiple) variables and perform several analyses on the collected samples. The multivariate techniques are efficient compared to the univariate counterparts due to their ability to consider inter-variable relationships. In this tutorial, we will focus on important multivariate data analytical techniques including: (1) Principal Component Analysis; (2) Factor Analysis; (3) Cluster Analysis; (4) Discriminant Analysis with real data examples. These four types of multivariate analytical techniques are extensively used in Agricultural experimental data analysis.

1. Principal Components Analysis and Factor Analysis

Principal Components Analysis (PCA) and Factor Analysis (FA) are usually viewed as attempts to approximate the relationships among a set of (*i.e.*, multiple) variables. PCA is concerned with explaining the variance-covariance structure through a few *linear* combinations of the original variables. Whereas FA is concerned with explaining covariance relationships among original variables in terms of a few underlying, but unobservable, random quantities called factors. Factors which are generated are thought to be representative of the underlying processes that have created the correlations among variables. FA is considered as an extension of PCA and its model is also considered to be more elaborative than PCA model. Many a time, PCA and FA together called as Common Factor Analysis (CFA).

The key underlying base to Common Factor Analysis (PCA and FA) is that the chosen variables can be transformed into linear combinations of factors. Factors may either be associated with 2 or more of the original variables (common factors) or associated with an individual variable (unique factors). Loadings relate the specific association between factors and original variables. Therefore, it is necessary to find the loadings, then solve for the factors, which will approximate the relationship between the original variables and underlying factors. The loadings are derived from the magnitude of eigenvalues associated to individual variables. The difference between PCA and FA is that for the purposes of matrix computations PCA assumes that all variance is common, with all unique factors set equal to zero; while FA assumes that there is some unique variance. The level of unique variance is dictated by the FA model which is chosen. Accordingly, PCA is a model of a closed system, while FA is a model of an open system. Rotation in CFA attempts to put the factors in a simpler position with respect to the original variables, which aids in the interpretation of factors. Rotation places the factors into positions that only the variables, which are distinctly related to a factor, will be associated. Varimax, quartimax, and equimax are all orthogonal rotations, while oblique rotations are non-orthogonal. The varimax rotation maximizes the variance of the loadings, and is also the most commonly used rotation method. To analyze data with either PCA or FA, three key decisions must be made. They are (i) the factor extraction method (ii) the number of factors to extract and (iii) the transformation method to be used.

For example: Foresters measure data on several characters (*e.g.*, variables) of tree species, such as, growth, volume, yield, forest potential, height, collar diameter, diameter at breast height, crown diameter, *etc*. The example data is shown below. Here, the main idea is to illustrate CFA approach on this example data using R. The main purpose of this lecture is to focus more on CFA approach using public statistical R software than dealing with its theory.

Principal Components Analysis

Example 1: The following data pertains to variables, such as Height, Collar diameter, Diameter Breast Height (DBH), and Crown diameter of 36 trees of a particular species. We perform PCA on this data using R and provide hand-on to interpret the obtained results.

Obs.	**Height**	**Collar diameter**	**DBH**	**Crown diameter**
1	4.00	10.50	6.90	15.13
2	3.80	7.00	4.30	2.63
3	4.90	10.30	7.30	21.71
4	3.00	9.10	5.80	5.24
5	3.80	9.80	6.40	7.57
6	4.00	10.90	6.50	8.67
7	5.30	11.10	6.90	13.09
8	4.50	10.30	6.50	10.55
9	4.40	10.30	6.30	11.53
10	4.70	13.70	9.10	20.66
11	5.20	14.90	10.50	23.19
12	5.30	14.90	9.40	18.59
13	3.60	9.30	6.30	10.21
14	3.30	6.40	3.70	4.67
15	5.00	9.70	6.20	12.34
16	3.70	8.10	5.10	5.89
17	3.80	9.20	5.60	5.36
18	4.00	10.30	7.80	7.54
19	4.90	12.10	8.00	12.93
20	5.50	12.70	8.70	17.79
21	5.30	13.60	9.00	12.76
22	4.80	14.90	10.00	25.62
23	4.30	13.20	9.10	15.57
24	5.10	14.50	10.10	22.56
25	1.50	2.50	1.20	0.17
26	1.90	3.80	2.50	1.56
27	2.40	3.90	1.70	0.58
28	3.80	7.50	5.50	4.47
29	3.60	9.30	6.50	5.94
30	3.30	7.00	3.90	4.63
31	5.60	13.70	9.20	15.26
32	4.50	9.90	4.70	12.11
33	5.20	11.20	7.70	12.57
34	4.60	11.70	9.70	16.21
35	5.00	18.40	10.80	21.15
36	4.60	12.80	8.60	14.44

R-codes:

Step 1: Create and set the working directory.

```
setwd("../../file location")
```

Step 2: Save the data in a text file and data reading. For example name the data file as: “data.txt”

```
dat <- read.table(file="data.txt", header=T, row.names = 1, sep="\t")
```

Step 3: PCA

```
dat.pca <- prcomp(dat, center = TRUE,scale. = TRUE)
```

Step 4: Results summary(dat.pca)

```
Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     1.9008 0.45588 0.37934 0.18753
Proportion of Variance 0.9033 0.05196 0.03597 0.00879
Cumulative Proportion  0.9033 0.95523 0.99121 1.00000

Step 5: Representation through a lower ortho-dimensional space.
Rotation (n x k) = (4 x 4):
                   PC1        PC2        PC3          PC4
Height       0.4832289  0.8509266  0.1995658 -0.050865416
Collar diam. 0.5127805 -0.1463091 -0.4324909  0.727049786
DBH          0.5104606 -0.2202849 -0.4712747 -0.684693116
Crown diam.  0.4929258 -0.4538637  0.7423108  0.002579604
```

Interpretation:

The proportion of total variation accounted for by the first principal component is 0.903 and the first two components account for a proportion of .9552. Hence, in further analysis, the first or first two principal components PCA1 and PCA2 could replace four variables by sacrificing negligible information about the total variation in the system. The scores of principal components can be obtained by substituting the values of variables in equations of PCAs. For the data under study, the first two principal component scores for first observation i.e. for tree1 can be worked out as

$$\text{PC1 score} = .483 \times 4.00 + .512 \times 10.50 + .510 \times 6.90 + .49 \times 15.13$$

$$\text{PC2 score} = .851 \times 4.00 + -.146 \times 10.50 + -.220 \times 6.90 - .4533 \times 15.13$$

Similarly for all other trees the first two principal components scores can be worked out. Thus the whole data with four variables can be converted to a new data set with two principal components.

Factor Analysis

We have demonstrated the FA with the following data example in R.

Example 2: Consider a hypothetical data on six characters with 15 observations as below:

Obs.	X1	X2	X3	X4	X5	X6
1	609.40	164.99	61.11	15.77	449.89	318.38
2	1960.90	4.30	54.74	33.47	37.14	1.43
3	1846.20	72.92	64.28	36.09	927.87	79.43
4	1002.70	211.76	49.15	42.60	1198.60	280.20
5	2801.10	59.43	82.32	4.40	329.55	108.05
6	1060.00	156.00	69.97	14.07	318.33	229.67
7	512.80	642.81	68.59	8.74	497.83	865.83
8	919.40	18.50	77.13	7.63	403.25	142.80
9	450.40	13.90	54.46	3.48	124.42	25.42
10	1449.90	129.93	67.04	20.04	530.53	210.44
11	2153.40	96.49	90.92	12.72	881.04	84.66
12	1237.85	147.97	64.38	37.43	643.96	165.81

13	744.90	95.75	77.27	25.03	551.82	154.47
14	1320.90	29.11	68.87	28.54	344.87	63.15
15	1846.20	21.40	63.31	33.87	261.71	12.48

R code: Repeat the Steps 1 and 2.

```
dat.fa <- factanal(dat, factors = n)  ###choose ‘n’
```

Results:

```
Uniquenesses:
   X1    X2    X3    X4    X5    X6 
0.775 0.030 0.741 0.005 0.738 0.005 

Loadings:
   Factor1 Factor2
X1 -0.474         
X2  0.980         
X3         -0.509 
X4 -0.249   0.966 
X5  0.200   0.471 
X6  0.997         

               Factor1 Factor2
SS loadings      2.282   1.424
Proportion Var   0.380   0.237
Cumulative Var   0.380   0.618

Null hypothesis: 2 factors are sufficient
Chi square statistic: 6.62 (4 degrees of freedom)
p-value: 0.158
```

Interpretation: Before we interpret the results of the FA, recall the basic idea behind it. FA creates linear combinations of factors to abstract the variable’s underlying communality. To the extent that the variables have an underlying communality, fewer factors capture most of the variance in the dataset. This allows us to aggregate a large number of observable variables in a model to represent an underlying concept, making it easier to understand the data. The variability in our data, is given by Σ, and its estimate $\hat{\Sigma}$ is composed of the variability explained by the factors (linear combination of the factors (communality)) and part of the variability cannot be explained by a linear combination of the factors (uniqueness).

From the above FA it is evident that two factors are sufficient as the test is not significant. Variables X1, X2, and X6 define *factor 1* (high loadings on factor 1, small or negligible loadings on factor 2), variables X3, X4, and X5 define *factor 2* (high loadings on factor 2, small or negligible loadings on factor 1).

Cluster Analysis and Discriminant Analysis

Rudimentary, exploratory procedures are often quite helpful in understanding the complex nature of multivariate relationship. Searching the data for a structure of "natural" grouping is an important exploratory technique. The most important techniques for data classification are: Cluster analysis and Discriminant analysis.

Although both cluster and discriminant analyses classify objects into different categories, discriminant analysis requires one to know group membership for the cases (*i.e*., prior class information) used to decide the classification rule whereas in cluster analysis group membership for all cases is unknown. In addition to membership, the number of groups is also generally unknown. In cluster analysis the units within cluster are similar but different between clusters. The grouping is done on the basis of some criterion like similarities measures etc. Thus in the case of cluster analysis the inputs are similarity measures or the data from which these can be computed.

Cluster Analysis

Cluster analysis is a technique used for combining observations into groups such that:

(a) Each group is homogeneous or compact with respect to certain characteristics i.e., observations in each group are similar to each other.

(b) Each group should be different from other groups with respect to the characteristics i.e., observations of one group should be different from the observations of other groups.

There are various mathematical methods which help to sort objects in to a group of similar objects called a Cluster. Cluster analysis is used in diversified research fields. In biology, cluster analysis is used to identify diseases and their stages. For example by examining patients who are diagnosed as depressed, one finds that there are several distinct sub-groups of patients with different types of depression. In marketing cluster analysis is used to identify persons with similar buying habits. By examining their characteristics it becomes possible to plan future strategies more efficiently.

Example 3: We will use the data given in Example 2 for cluster analysis using R.

Steps 1, 2: Follow the Steps 1 and 2 mentioned in PCA.

Step 3: Normalize the data (Sometimes normalization is essential for cluster analysis)

```
means <- apply(dat1,2,mean)
sds <- apply(dat1,2,sd)
nor <- scale(dat1,center=means,scale=sds)
```

Step 4: Calculate the distance matrix

```
distance = dist(nor)
```

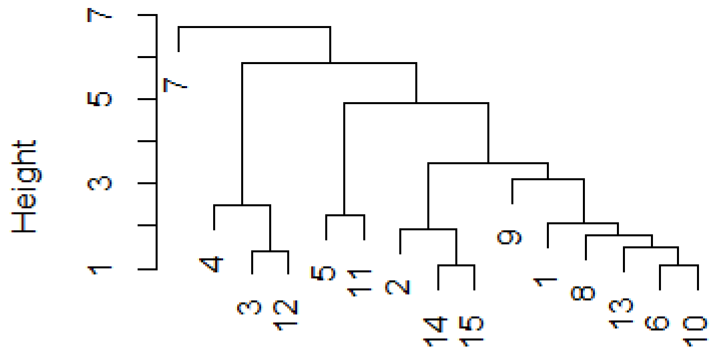
Step 5: Select the clustering method

a. Default method (Hierarchical agglomerative clustering)

```
mydata.hclust = hclust(distance)
plot(mydata.hclust)
```

```
Cluster membership:
member = cutree(mydata.hclust, n) ###Select n (number of clusters, say
n=3)
[1] 1 1 2 2 1 1 3 1 1 1 1 2 1 1 1
Table: member
 1  2  3
 11  3  1
```

Dendrogram:

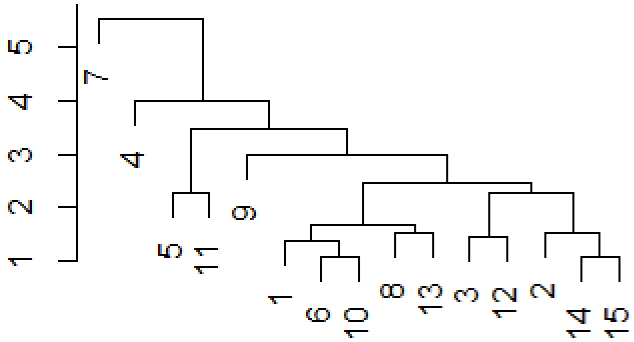


b. **Average linkage method**

```
mydata.hclust = hclust(distance, method = “average”)
plot(mydata.hclust)
```

```
Cluster membership:
member = cutree(mydata.hclust, n) ###Select n (number of clusters,
say n=3)
[1] 1 1 1 2 1 1 3 1 1 1 1 1 1 1 1
Table: member
   1  2  3
   11  3  1
```

Dendrogram:



c. **K-means clustering**

```
set.seed(123)  ##set seed for results reproducibility
kc<-kmeans(nor,3)
print(kc)
```

```
K-means clustering with 3 clusters of sizes 5, 3, 7
Cluster members:
 [1] 3 1 2 2 1 3 3 3 3 3 1 2 3 1 1
Cluster means:
           x1         x2          x3          x4         x5         x6
1  1.01492019 -0.5231903  0.40335875  0.07683345 -0.4159255 -0.6089339
2  0.05085677  0.1264368 -0.75013641  1.30454261  1.3631827 -0.0362353
3 -0.74673875  0.3195201  0.03337364 -0.61397073 -0.2871315  0.4504822
```

```
Within cluster sum of squares by cluster:
[1] 17.946535  4.653991 29.628320 (between_SS / total_SS =  37.8 %)
```

REFERENCES

Chatfield, C. and Collins, A.J. (1990). Introduction to multivariate analysis. *Chapman and Hall publications*.

Cheng (1997). Applications of GIS and Multivariate Statistical Analysis in Planning Water Conservation Protected Forest -- An Example of the Experimental Forest of National Taiwan University. *Taiwan Journal of Forestry*.

Johnson, R.A. and Wichern, D.W. (1996). Applied multivariate statistical analysis. *Prentice-Hall of India Private Limited*.

Salam and Naguchi. (1998). Factors influencing the loss of forest cover in bangladesh: An analysis from socioeconomic and demographic perspectives, *Journal of Forest Research*, **3**, 145-150.

Sharma, S. (1996). Applied Multivariate Techniques, *John Wiley & Sons*, New York.