FACTOR ANALYSIS USING R

1 Introduction

Factor analysis is a technique that is used to reduce a large number of variables into fewer numbers of factors. This technique extracts maximum common variance from all variables and puts them into a common score. As an index of all variables, we can use this score for further analysis.
Factor analysis can be considered an extension of principal component analysis. Both can be viewed as attempts to approximate the covariance matrix. However the approximation based on the factor analysis model is more elaborate.

2 Quick Review

Suppose $\mathbf{X}=\left(X_1, X_2, ..., X_m\right)^T$ be a vector of m variables having the dispersion matrix Σ

Let $\mathbf{X}=\mu + \mathbf{L}\mathrm{F} + \epsilon$

Where
$\mu=\left(\mu_1, \mu_2, ..., \mu_m\right)^T$ is the vector of means
$\mathrm{F}=(F_1, F_2, ..., F_p)^T$ is the vector of p factors
$\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_m)^T$ is the vector of specific factors
$\mathbf{L}=((l_{ij}))$is an m×p matrix of loadings

Assumptions:

- $\mathbf{E}(\epsilon)=0$
- $\mathbf{E}(\mathrm{F})=0$
- $\mathbf{D}(\epsilon)=\Psi \rightarrow$ Diagonal matrix
- $\mathbf{D}(\mathrm{F})=\mathbf{I}_p$
- cov(F,ϵ)=0

This is known as the orthogonal factor model

Thus we have

$\mathbf{E}(\mathrm{X})=\mu$ and $\mathbf{D}(\mathrm{X})=\mathrm{LL}^T + \Psi = \Sigma$

Note that:

$$Var(X_i) = \underbrace{{l_{i1}}^2 + {l_{i2}}^2 + ... + {l_{ip}}^2}_{Communality} + \underbrace{\Psi_i}_{Uniqueness}$$

For a good Factor model we want communality to be large and uniqueness to be small
Also p $\ll$ m, that is the number of factors are required to be much smaller than the number of variables

- Contribution of the jth factor to the total variability of the data is

  $\frac{\lambda_j}{\sum_{j=1}^{m} Var(X_j)}$, where λ_j is the jth largest eigen value of the matrix Σ

- Cumulative contribution of the first j factors to the total variability is

  $\sum_{k=1}^{j} \left(\frac{\lambda_k}{\sum_{j=1}^{m} Var(X_j)} \right)$

Example

Consider the data on the times taken by some athletes to complete different category races like 100m race, 200m race etc

100m	200m	400m	800m	1500m	5K	10K	Marathon
10.39	20.81	46.84	1.81	3.7	14.04	29.36	137.72
10.31	20.06	44.84	1.74	3.57	13.28	27.66	128.3
10.44	20.81	46.82	1.79	3.6	13.26	27.72	135.9
10.34	20.68	45.04	1.73	3.6	13.22	27.45	129.95
10.28	20.58	45.91	1.8	3.75	14.68	30.55	146.62
10.22	20.43	45.21	1.73	3.66	13.62	28.62	133.13
10.64	21.52	48.3	1.8	3.85	14.45	30.28	139.95
10.17	20.22	45.68	1.76	3.63	13.55	28.09	130.15
10.34	20.8	46.2	1.79	3.71	13.61	29.3	134.03
10.51	21.04	47.3	1.81	3.73	13.9	29.13	133.53
10.43	21.05	46.1	1.82	3.74	13.49	27.88	131.35
12.18	23.2	52.94	2.02	4.24	16.7	35.38	164.7
10.94	21.9	48.66	1.87	3.84	14.03	28.81	136.58
...	...	...	...	...	...	...	...
10.82	21.86	49	2.02	4.24	16.28	34.71	161.83

Notice that the time in the first 3 categories of races (first three columns) are given in seconds whereas the other times are given in minutes

We are first going to see how many factors are required to best explain the data. We perform eigen analysis of the sample correlation matrix from the first principles. We find the cumulative proportion of variability explained by factors

```
> R=cor(runners)
> lambda=eigen(R)$values
> lambda
[1] 6.62214613 0.87761829 0.15932114 0.12404939 0.07988027 0.06796515
    0.04641953 0.02260010
> v=eigen(R)$vectors

> P=cumsum(lambda)/sum(diag(R))
> P
[1] 0.8277683 0.9374706 0.9573857 0.9728919 0.9828769 0.9913725
    0.9971750 1.0000000
```

Thus we see that $\approx 94\%$ of the total variability is explained by 2 factors

First we are going to look at the principal component method of estimation.

Principal Component Method of Estimation:

Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ..... \geq \hat{\lambda}_m$ be the eigen values of the matrix $\hat{\Sigma}$, the sample estimate of the dispersion matrix Σ with $\hat{p}_1, \hat{p}_2, ..., \hat{p}_m$ as the corresponding estimated normalized eigen vectors. If p$\ll$m be the number of factors which can best explain the data, then the estimate of the matrix L, by the principal component method of estimation is given by

$$\hat{\mathbf{L}}=\left[\sqrt{\hat{\lambda}_1}\hat{\mathrm{p}}_1 \ \sqrt{\hat{\lambda}_2}\hat{\mathrm{p}}_2 \cdots \sqrt{\hat{\lambda}_p}\hat{\mathrm{p}}_p\right]$$

Also the matrix Ψ is estimated as

$$\hat{\Psi} = \mathrm{Diag}\left(\hat{\Sigma} - \hat{\mathrm{L}}\hat{\mathrm{L}}^{\mathrm{T}}\right)$$

We see that 2 factors explain $\approx 94\%$ of the total data variability. We estimate the matrix of loadings L and matrix of uniqueness Ψ from the first principles.

```
> L=cbind(sqrt(lambda[1])*v[,1],sqrt(lambda[2])*v[,2])
> rownames(L)=colnames(runners)
> colnames(L)=c("Factor1","Factor2")
> L                    ## The Matrix of loadings
```

```
            Factor1    Factor2
X100m       -0.817     -0.531
X200m       -0.867     -0.432
X400m       -0.915     -0.232
X800m       -0.949     -0.012
X1500m      -0.959      0.131
X5K         -0.938      0.292
X10K        -0.944      0.287
Marathon    -0.880      0.411
```

```
#The residual matrix is given by

> Resid=R-(L%*%t(L))

> Shi=diag(Resid)
> Shi                ## Uniqueness
```

```
X100m       X200m       X400m       X800m      X1500m         X5K      X10K     Marathon
0.050       0.061       0.108       0.099       0.062       0.035    0.026       0.057
```

To Find the Factor scores:

For the Principal Component Method:

The estimated factor scores corresponding to the ith individual is:

$$\hat{f}_i = \left(\hat{\mathrm{L}}^{\mathrm{T}}\hat{\mathrm{L}}\right)^{-1}\hat{\mathrm{L}}^{\mathrm{T}}\left(X_i - \bar{X}\right)$$

```
> for(i in 1:nrow(runners))
+ {
+   runner[i,]=runners[i,]-apply(runners,MARGIN=2,FUN=mean)
+ }
> f=matrix(0,nrow=nrow(runners),ncol=2)
> A=as.matrix(solve(t(L)%*%L)%*%t(L))
> for(i in 1:nrow(runners))
+ {
+    f[i,1]=A[1,]%*%t(runner[i,])    #1st Factor Scores
+    f[i,2]=A[2,]%*%t(runner[i,])    #2nd Factor Scores

+ }


> colnames(f)=c("Factor 1","Factor 2")
```

Let us assume that Factor 1="Endurance" and Factor 2="Strength"

```
> f
        Factor 1      Factor 2
[1,] -0.25702652  0.70673628
[2,]  1.75788702 -3.58763316
[3,]  0.34293545 -0.98266780
[4,]  1.46163011 -3.27532407
[5,] -1.53340311  5.91393352
[6,]  0.83106172 -1.10909353
[7,] -1.08840895  1.32401144
        . . . .       . . . .
[54,]  1.55509360 -2.36132934
[55,] -5.13765520 13.23024453
```

Data Corresponding to 4^{th} individual

X100m	X200m	X400m	X800m	X1500m	X5K	X10K	Marathon
10.34	20.68	45.04	1.73	3.6	13.22	27.45	129.95

Data Corresponding to 6^{th} individual

X100m	X200m	X400m	X800m	X1500m	X5K	X10K	Marathon
10.22	20.43	45.21	1.73	3.66	13.62	28.62	133.13

Interpretation:

- The score corresponding to "Endurance" (Factor 1) is more for the 4^{th} individual as compared to the 6^{th} individual. Looking at the data corresponding to the two individuals we see that for the 4^{th} individual time taken to complete a longer distance race i.e 5k, 10k or Marathon is less as compared to the 6^{th} individual.
- Likewise the score corresponding to "Strenth" (Factor 2) is more for the 6^{th} individual as compared to the 4^{th} individual. Looking at the data corresponding to the two individuals we see that for the 6^{th} individual time taken to complete a shorter distance race i.e 100m, 200m, etc is less as compared to the 4^{th} individual.

Maximum Likelihood Method of Estimation

The built in R function for ML estimation of a Factor model is **factanal()**

```
>fit_ML=factanal(runners,factors=2,rotation="varimax")
```

From the output fit_ML we see

Test of the hypothesis that 2 factors are sufficient. The chi square statistic is 16.36 on 13 degrees of freedom. The p-value is 0.23

This implies that the model fit is appropriate
Here fit_ML is the output from the factanal() function. We extract the loadings and uniqueness from the output

```
> L.ML=fit_ML$loadings
> L.ML            #Getting the matrix of loadings
Loadings:
        Factor1 Factor2
X100m    0.291   0.914
X200m    0.382   0.882
X400m    0.543   0.744
X800m    0.691   0.622
X1500m   0.799   0.530
X5K      0.901   0.394
X10K     0.907   0.399
Marathon 0.915   0.278

                Factor1  Factor2
SS loadings      4.112   3.225
Proportion Var   0.514   0.403
Cumulative Var   0.514   0.917


> Shi.MLE=fit_ML$uniqueness
> Shi.MLE        #Getting the uniquesness

X100m       0.081
X200m       0.076
X400m       0.151
X800m       0.135
X1500m      0.082
X5K         0.034
X10K        0.018
Marathon    0.086
```

To Find the Factor scores:

For the Maximum Likelihood Method:

The estimated factor scores corresponding to the ith individual is:

$$\hat{f}_i = \left(\hat{\mathrm{L}}^{\mathrm{T}}\hat{\Psi}^{-1}\hat{\mathrm{L}}\right)^{-1}\hat{\mathrm{L}}^{\mathrm{T}}\hat{\Psi}^{-1}\left(X_i - \bar{X}\right)$$

In R we can get the factor scores from the R output by extracting the scores

```
> fit_ML_scores=factanal(runners,factors=2,scores="regression",rotation="varimax")$scores
> fit_ML_scores
```

	Factor1	Factor2
1	0.33633782	-0.265151192
2	-0.49395787	-0.812133498
3	-0.74199914	0.176415083
4	-0.79602754	-0.238852529
5	1.46541593	-1.170446573
6	0.07780163	-0.887129076
	. . . . .	. . . . .
54	-0.32473110	-1.223659041
55	3.30259721	0.370069289

The interpretation of factor scores remain the same as in that obtained by the Principal Component Method of estimation

Conclusion:

Originally we had 8 variables. With the help of Factor Analysis we can consider a 2 factor model, without losing much information about the data variability and simultaneously achieving reduction in dimension

Example where Factor Analysis is not worthwhile

Suppose that we have $\mathbf{X}=(X_1, X_2, X_3, X_4, X_5, X_6)^T$ with correlation matrix $\mathbf{R}$

$$\mathbf{R} = \begin{bmatrix} 1.0000 & 0.4919 & 0.2636 & 0.4653 & -0.2277 & 0.0652 \\ 0.4919 & 1.0000 & 0.3127 & 0.3506 & -0.1917 & 0.2045 \\ 0.2636 & 0.3127 & 1.0000 & 0.4108 & 0.0647 & 0.2493 \\ 0.4653 & 0.3506 & 0.4108 & 1.0000 & -0.2249 & 0.2293 \\ -0.2277 & -0.1917 & 0.0647 & -0.2249 & 1.0000 & -0.2144 \\ 0.0652 & 0.2045 & 0.2493 & 0.2293 & -0.2144 & 1.0000 \end{bmatrix}$$

From the first principals:

We find the eigen values of the matrix R and the cumulative proportion of variability explained

```
eigen.values=eigen(R)$values
eigen.values
[1] 2.3549437 1.0718555 0.9842359 0.6643850 0.5003684 0.4242116

cumsum(eigen.values)/sum(eigen.values)
[1] 0.3924906 0.5711332 0.7351725 0.8459033 0.9292981 1.0000000
```

Cumulative Proportion of total variability

j	1	2	3	4	5	6
P_j	0.3924906	0.5711332	0.7351725	0.8459033	0.9292981	1

Note that:

There are 6 variables and 5 factors are required to explain ≈93% of the total variability

So there is not much reduction in the number of variables by performing Factor Analysis

SUMMARY

- In R we can perform Factor Analysis using inbuilt R functions
- If the data is available in raw form we can go for either estimation using Principal Component method of Maximum Likelihood method and also estimate the factor scores
- If the data is not available in raw form, but we have the correlation or dispersion matrix, we can only go for estimation using Principal Component method and estimate the factor scores from first principles
- If the variables are highly correlated amongst themselves we achieve Dimension Reduction using Factor Analysis, however such dimension reduction is not significant if the correlations are low