# Factor Analysis Audio data set

Factor Analysis is a method of modeling the variable and their co variance structure in terms of common factors. It is also used to reduce the dimension of the data in principal component analysis sometime one variable contributes in more than one principal component. Ideally, we want each variable to be in different principal component. Factor Analysis is used to solve this.

Our objective is to describe relationship among the variables. Before doing factor analysis we need to determine number of factors should be included in the model

For 8 variable, co variance matrix will have 8*(8+1)/2=36 unique values

For factor analysis no of factor =m, Number of parameters in factor model should be = m (8+1)

If we have m=4 then no of parameter will be same as the original model so there would be any reduction in dimension.

So, we will take m=3 that Number of parameters in the model=27

Calculated Eigen value and Eigen vector from correlation matrix

```
> eigenvalues
[1] 3.929005 1.618322 0.975325 0.466782 0.340090 0.315891 0.200111 0.154474

Eigen value is >1 for first and second factor
> (eigenvalues)>1
[1]  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE

> sum(eigenvalues[1:m])/8
[1] 0.693
 Proportion of total variance lie to factor 1 and factor 2 is 69.3

> eigenvectors=eigen(rho)$vectors
> eigenvectors
          [,1]      [,2]       [,3]      [,4]       [,5]       [,6]       [,7]       [,8]
[1,] -0.401095  0.316964 -0.1581569 -0.327758 -0.0231364  0.4459041 -0.3292553  0.5462999
[2,] -0.420991  0.225464  0.0519613 -0.481631  0.3792268 -0.0674582  0.0331212 -0.6227389
[3,] -0.366375 -0.238593  0.4702930 -0.282429 -0.4392466 -0.0637999  0.5255167  0.1863469
[4,] -0.280856 -0.474154 -0.4295025 -0.161081 -0.3503196 -0.4169270 -0.4269440 -0.0839347
[5,] -0.343251  0.386020 -0.2593193  0.487600 -0.4975031  0.1947771  0.1593507 -0.3425302
[6,] -0.411421  0.231773  0.0288540  0.372316  0.3513176 -0.6136377  0.0836779  0.3613654
[7,] -0.311548 -0.317059  0.5629330  0.391417  0.1107857  0.2650301 -0.4778158 -0.1465875
[8,] -0.254221 -0.513512 -0.4262229  0.159098  0.3959590  0.3660466  0.4139353  0.0508206
```

Factor loading are nothing but correlation between the variable and Common factors.

| | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| L500 | -0.79504 | 0.40322 | -0.15619 |

| | | | |
|---|---|---|---|
| L1000 | -0.83448 | 0.28682 | 0.051316 |
| L2000 | -0.72622 | -0.30352 | 0.464454 |
| L4000 | -0.5567 | -0.60319 | -0.42417 |
| R500 | -0.68038 | 0.491068 | -0.2561 |
| R1000 | -0.81551 | 0.294845 | 0.028496 |
| R2000 | -0.61754 | -0.40334 | 0.555944 |
| R4000 | -0.50391 | -0.65326 | -0.42093 |

1) Factor 1 is correlated most strongly with L1000 (-0.834) and correlated with L500, L2000, R500, R2000 and R1000. We can say that the first factor is primarily a measure of these variables.

2) Factor 2 is primarily related to R4000 and L4000 Here we can see that Factor 2 is associated with hearing issue with the high frequency. This distinguishes person with hearing loss for high frequency values to the persons with hearing loss for lower frequency sound

3) Factor 3 is primarily a measure of R2000 and is also Positively related. Rest all values are less than .5

Common variance: it is variance in common with common factor for all 8 variables

```
> common
  L500      L1000    L2000   L4000     R500     R1000    R2000    R4000
[1] 0.819070 0.781249 0.835236 0.853676 0.769656 0.752797 0.853117 0.857851
```

We can say that these values as multiple $R2$ values for regression models predicting the variables from the 3 factors. The communality for a given variable can be interpreted as the proportion of variation in that variable explained by t he three factors. In other words, if we perform multiple regression of L500 against the three common factors, we obt ain an $R2$ = 0.81, indicating that about 81% of the variation in L500 is explained by the factor model. The results sug gest that the factor analysis does the best job of explaining variation L500, L1000, L2000,L4000, R500, R1000 ,R20 00,R4000.

Total Communality is 6.42 and proportion of total variance explained by # factors is 81%

Unique Variance: This Variance is unique to each variable.

```
> unique

[1] 0.180930 0.218751 0.164764 0.146324 0.230344 0.247203 0.146883 0.142149
```

Specific variance for L500 is .180

Factor Analysis model provides an approximation to the correlation matrix. We can check the model by recreating the correlation matrix.

```
> phi=diag(8)*unique
> recreate=L%*%t(L)+phi
> recreate
```

```
        [,1]     [,2]     [,3]     [,4]      [,5]     [,6]      [,7]     [,8]
[1,] 1.000000 0.771077 0.382440 0.265637 0.7789402 0.762796 0.2415001 0.202969
[2,] 0.771077 1.000000 0.542789 0.269784 0.6954691 0.766551 0.4281668 0.211533
[3,] 0.382440 0.542789 1.000000 0.390362 0.2261090 0.515978 0.8291041 0.368723
[4,] 0.265637 0.269784 0.390362 1.000000 0.1911960 0.264062 0.3512638 0.853111
[5,] 0.778940 0.695469 0.226109 0.191196 1.0000000 0.692348 0.0797197 0.129859
[6,] 0.762796 0.766551 0.515978 0.264062 0.6923481 1.000000 0.4005287 0.206338
[7,] 0.241500 0.428167 0.829104 0.351264 0.0797197 0.400529 1.0000000 0.340656
[8,] 0.202969 0.211533 0.368723 0.853111 0.1298591 0.206338 0.3406561 1.000000
```

Below is the original correlation matrix

```
> rho
          L500    L1000    L2000    L4000     R500    R1000    R2000    R4000
L500  1.000000 0.777542 0.401220 0.255358 0.696287 0.641617 0.237188 0.204089
L1000 0.777542 1.000000 0.536550 0.274945 0.551541 0.707026 0.359745 0.216887
L2000 0.401220 0.536550 1.000000 0.425018 0.239118 0.445983 0.701144 0.326214
L4000 0.255358 0.274945 0.425018 1.000000 0.178980 0.263196 0.316452 0.709740
R500  0.696287 0.551541 0.239118 0.178980 1.000000 0.663439 0.158890 0.132108
R1000 0.641617 0.707026 0.445983 0.263196 0.663439 1.000000 0.414232 0.220109
R2000 0.237188 0.359745 0.701144 0.316452 0.158890 0.414232 1.000000 0.374559
R4000 0.204089 0.216887 0.326214 0.709740 0.132108 0.220109 0.374559 1.000000
```

We can assess the model's appropriateness with the residuals

```
> options(digits=3)
> residual=rho-recreate
> residual
          L500    L1000    L2000     L4000     R500     R1000    R2000    R4000
L500   0.00000  0.00646  0.01878 -0.010280 -0.08265 -0.121179 -0.00431  0.00112
L1000  0.00646  0.00000 -0.00624  0.005161 -0.14393 -0.059524 -0.06842  0.00535
L2000  0.01878 -0.00624  0.00000  0.034656  0.01301 -0.069996 -0.12796 -0.04251
L4000 -0.01028  0.00516  0.03466  0.000000 -0.01222 -0.000867 -0.03481 -0.14337
R500  -0.08265 -0.14393  0.01301 -0.012216  0.00000 -0.028909  0.07917  0.00225
R1000 -0.12118 -0.05952 -0.07000 -0.000867 -0.02891  0.000000  0.01370  0.01377
R2000 -0.00431 -0.06842 -0.12796 -0.034812  0.07917  0.013703  0.00000  0.03390
R4000  0.00112  0.00535 -0.04251 -0.143371  0.00225  0.013772  0.03390  0.00000
```

Residual value is almost zero. This indicates that how well factor model fits with the data.

Let’s consider m=2

Loading:

```
> L
       [,1]   [,2]
[1,] -0.795  0.403
[2,] -0.834  0.287
[3,] -0.726 -0.304
[4,] -0.557 -0.603
[5,] -0.680  0.491
[6,] -0.816  0.295
[7,] -0.618 -0.403
[8,] -0.504 -0.653
```

Communality variance and unique variance dependent on the number of factors

```
> common
[1] 0.795 0.779 0.620 0.674 0.704 0.752 0.544 0.681
```

Unique Variance

```
> unique
[1] 0.205 0.221 0.380 0.326 0.296 0.248 0.456 0.319
```

We are able to almost recreate the co relation matrix with the factor model where factor equals to 2

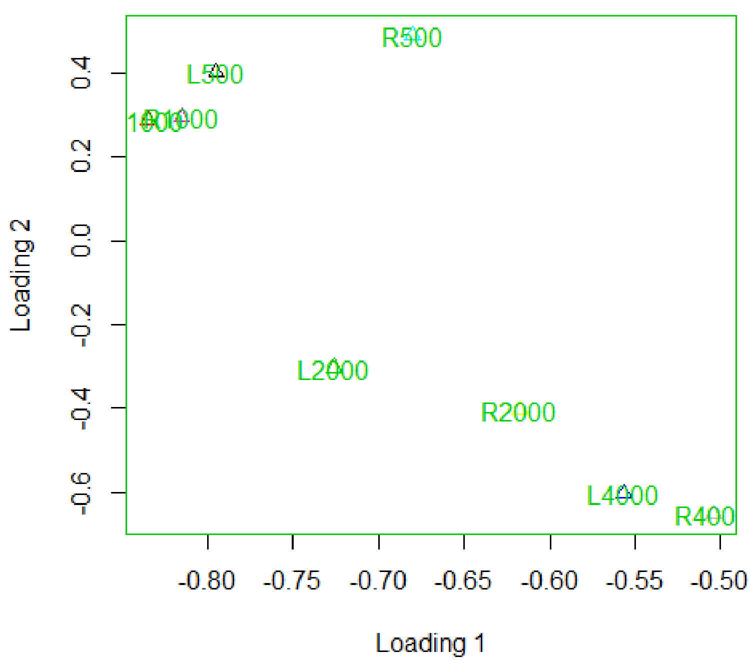
```
> recreate
      [,1]  [,2]  [,3]   [,4]   [,5]  [,6]  [,7]   [,8]
[1,] 1.000 0.779 0.455 0.1994 0.7389 0.767 0.328 0.1372
[2,] 0.779 1.000 0.519 0.2916 0.7086 0.765 0.400 0.2331
[3,] 0.455 0.519 1.000 0.5874 0.3451 0.503 0.571 0.5642
[4,] 0.199 0.292 0.587 1.0000 0.0826 0.276 0.587 0.6746
[5,] 0.739 0.709 0.345 0.0826 1.0000 0.700 0.222 0.0221
[6,] 0.767 0.765 0.503 0.2761 0.6996 1.000 0.385 0.2183
[7,] 0.328 0.400 0.571 0.5871 0.2221 0.385 1.000 0.5747
[8,] 0.137 0.233 0.564 0.6746 0.0221 0.218 0.575 1.0000
> rho
       L500 L1000 L2000 L4000  R500 R1000 R2000 R4000
L500   1.000 0.778 0.401 0.255 0.696 0.642 0.237 0.204
L1000  0.778 1.000 0.537 0.275 0.552 0.707 0.360 0.217
L2000  0.401 0.537 1.000 0.425 0.239 0.446 0.701 0.326
L4000  0.255 0.275 0.425 1.000 0.179 0.263 0.316 0.710
R500   0.696 0.552 0.239 0.179 1.000 0.663 0.159 0.132
```

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| R1000 | 0.642 | 0.707 | 0.446 | 0.263 | 0.663 | 1.000 | 0.414 | 0.220 |
| R2000 | 0.237 | 0.360 | 0.701 | 0.316 | 0.159 | 0.414 | 1.000 | 0.375 |
| R4000 | 0.204 | 0.217 | 0.326 | 0.710 | 0.132 | 0.220 | 0.375 | 1.000 |

Lets check the residual

| | L500 | L1000 | L2000 | L4000 | R500 | R1000 | R2000 | R4000 |
|---|---|---|---|---|---|---|---|---|
| L500 | 0.00000 | -0.00155 | -0.0538 | 0.0560 | -0.0427 | -0.12563 | -0.0911 | 0.06687 |
| L1000 | -0.00155 | 0.00000 | 0.0176 | -0.0166 | -0.1571 | -0.05806 | -0.0399 | -0.01625 |
| L2000 | -0.05377 | 0.01760 | 0.0000 | -0.1624 | -0.1059 | -0.05676 | 0.1303 | -0.23801 |
| L4000 | 0.05597 | -0.01661 | -0.1624 | 0.0000 | 0.0964 | -0.01295 | -0.2706 | 0.03518 |
| R500 | -0.04265 | -0.15707 | -0.1059 | 0.0964 | 0.0000 | -0.03621 | -0.0632 | 0.11005 |
| R1000 | -0.12563 | -0.05806 | -0.0568 | -0.0130 | -0.0362 | 0.00000 | 0.0295 | 0.00178 |
| R2000 | -0.09115 | -0.03989 | 0.1303 | -0.2706 | -0.0632 | 0.02955 | 0.0000 | -0.20011 |
| R4000 | 0.06687 | -0.01625 | -0.2380 | 0.0352 | 0.1100 | 0.00178 | -0.2001 | 0.00000 |

In above residual matrix off diagonal values are quite significant. We can say that when we increase number of factors residual also increase.




Loading 1 lower value plots the Variable L2000, R2000,L4000,R4000 Hence Factor 1 define the hearing loss at high frequency sounds

Loading 2 High value plots the variable L500, R500, L1000,R1000 ,Factor 2 defines the hearing loss at lower frequency, Confirming output with `Principal Components Analysis`

```
Call: principal(r = data, nfactors = 2, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
       PC1   PC2   h2   u2 com
L500  0.80 -0.40 0.79 0.21 1.5
L1000 0.83 -0.29 0.78 0.22 1.2
L2000 0.73  0.30 0.62 0.38 1.3
L4000 0.56  0.60 0.67 0.33 2.0
R500  0.68 -0.49 0.70 0.30 1.8
R1000 0.82 -0.29 0.75 0.25 1.3
R2000 0.62  0.40 0.54 0.46 1.7
R4000 0.50  0.65 0.68 0.32 1.9

                       PC1  PC2
SS loadings           3.93 1.62
Proportion Var        0.49 0.20
Cumulative Var        0.49 0.69
Proportion Explained  0.71 0.29
Cumulative Proportion 0.71 1.00

Mean item complexity =  1.6
Test of the hypothesis that 2 components are enough.

The root mean square of the residuals (RMSR) is  0.11
with the empirical chi square  64.7  with prob <  7.4e-09
```

RMSR is very small hence our model is good.