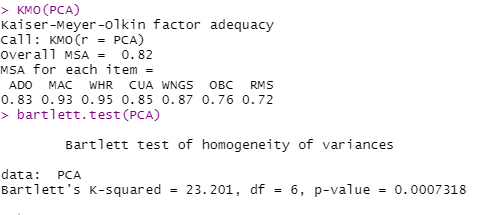
HW5\_PCA

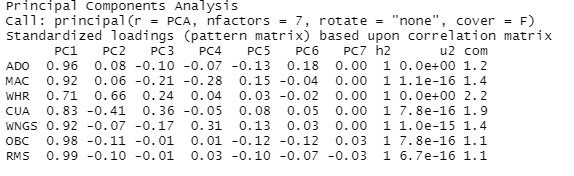
0859605陳冠景



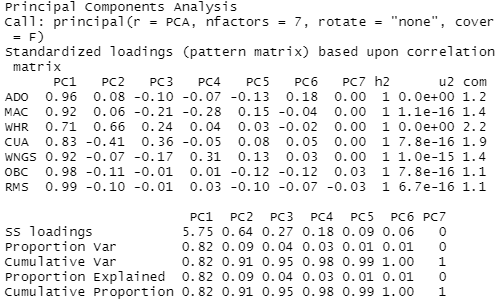
Five variables ADO(average daily occupancy),MAC (monthly average number of check ins),WHR(number of hour per week service desk in operation),CUA(total common use area),WNGS(number of building wings) are factor well. The other two variables(OBC,RMC) are only mediocre degree of common variance, they’re KMO index are lower than .8, but higher than .6.

The Bartlett’s test of Sphericity is significant (Bartlett’s K- squared = 23.201, *p*<0.001) and the *p*-value is small enough to reject the null hypothesis.

The strength of the relationship among the variables are strong and appropriate for factor analysis or principal component analysis.



The community of 7 variables are all 1, which means these variables can be totally explained by these 7 principal components.



According to the principal component analysis, the SS loadings in the result represent the sum of square loading, which is the eigenvalues of each variables.

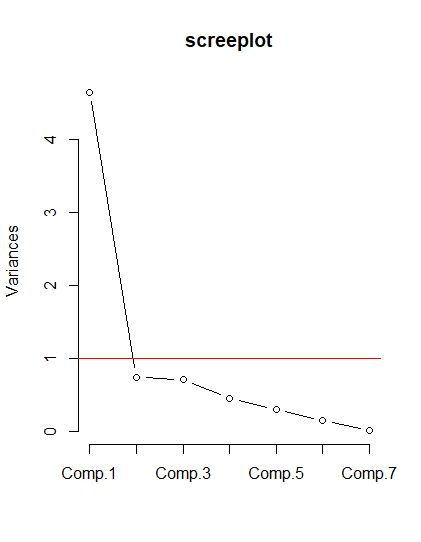
With 7 factors in PCA , only eigenvalues of first principal component is larger than

1, and account for 82% of variance explained among variables. The second

principal component accounts for 9% of variance explained. The cumulative 91%

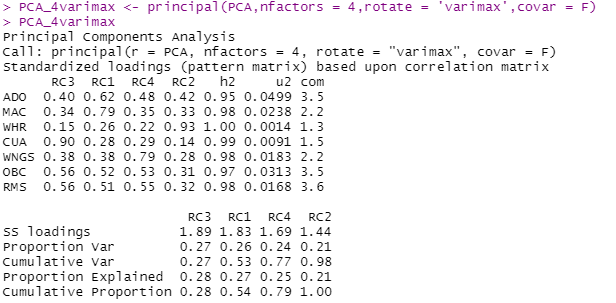
of variance explained among first and second principal component. The total

cumulative variance can be explained by six principal components.



According to the screeplot, only 2 component values are higher than 1, the other eigenvalues of the other 5 components are close enough to zero that they can be ignored.

As a result, the number of principal components among these variables is 2.



Since the second principal component accounts little variance. In this case, the varimax orthogonal rotation have been applied to find out the true dimension that well explain the variance among variables.

The result of varimax rotation with 4 factors shows that the first principal component can account for 27% of variance and the second is 26%, the third is 24% , the forth is 21%. The cumulative variance of four principal components is almost 98 %, and each component accounts for one fourth of total variance among 7 variables.

As a result, the principal component analysis with 4 factors and varimax rotation can better represent the underlying latent variables among these 7 variables. Four principal components can be well explained variance among the original variables.

1. Name the principal component

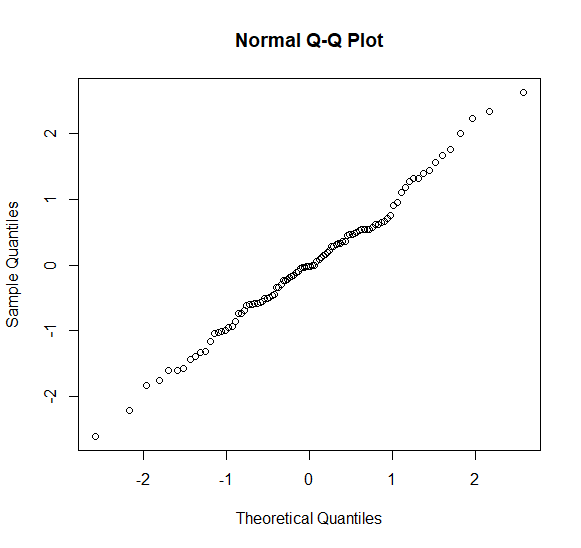
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table. 1 dominate cross loading in each principal component | | | | |
| Dominate cross loading | PC1 (RC3) | PC2(RC1) | PC3(RC4) | PC4(RC2) |
| ADO |  | .62 |  |  |
| MAC |  | .79 |  |  |
| WHR |  |  |  | .93 |
| CUA | .9 |  |  |  |
| WNGS |  |  | .79 |  |
| OBC | .56 |  |  |  |
| RMS | .56 |  |  |  |

As a result, the first principal component is dominated by 3 variables ,total common use area in square feet (CUA), operational building capacity (OBC), total numbers of room (RMS). These three variables are about the scale of building. As a result, the first principal component is named ‘the scale of building’.

The second principal component is dominated by average daily occupancy (ADO)

and monthly average number of check-ins (MAC), which are about number of occupancy. The second principal component is named ‘average number of occipancy’. The third principal component is dominated by number of building wings(WNGS). Therefore, the third principal component is named ‘number of building wings’. The last principal component is dominated by the variable,the number of hours per week service desk is in operation. Than this component can be named ‘the open hour of counter’.

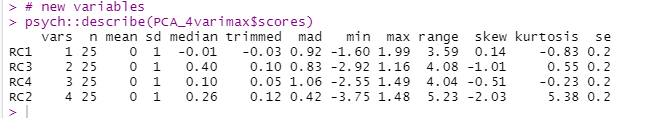
6.



Use the normal Q-Q plot to obtain the normal distribution of principal component

Scores. The line in the plot show almost linear, so the the principal components (new

variables ) have the multivariate normal distribution.



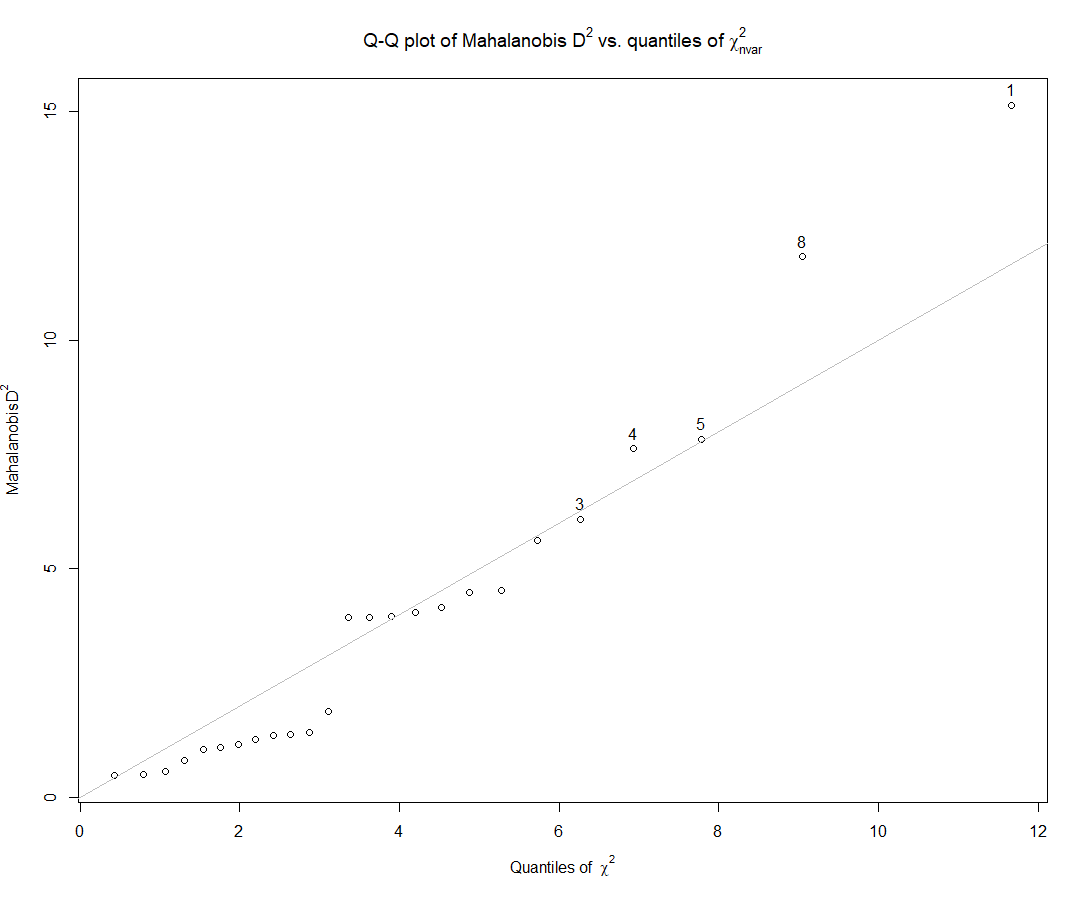
The skew and kurtosis for four principal component between -3~3. Thus, the four principal component variables are normal distribution.

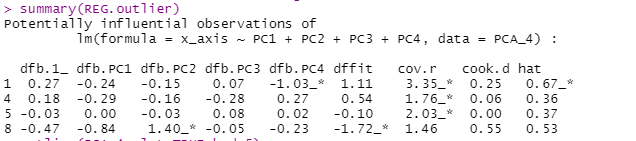
7. possible outlier

#1

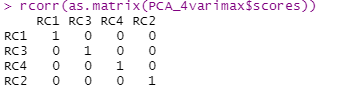
The outlier of PCA result as the picture below.

The sample unit 1.4.5.8 may be the outlier of the regression model of PCA.





8.

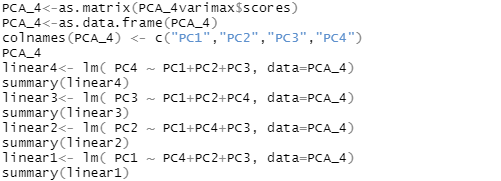


As the correlation matrix for PCA with four components above.

Since the principal component axes is orthogonal, the principal component score for

the measured variables on different axes are not correlated and there’re not

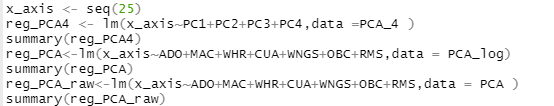
multicollinearities among the principal components.



The *p*- value of four linear regression model are all 1, so four variables are not

correlated.

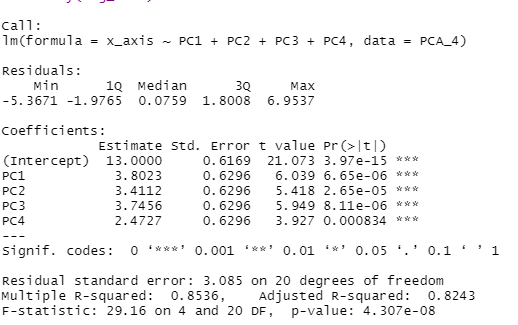
9. Compare PCA result and original variables and log transform ones.



The regression model of four principal components from PCA shows significant *p*-

values. As a result, PC1~PC4 have significantly correlated with sample units(n=25).

The total R squared is 82.4% and these four principal components can significantly explained 82.4% of the sample units’ variance(F(4,20)=29.16,*p*<0.001).

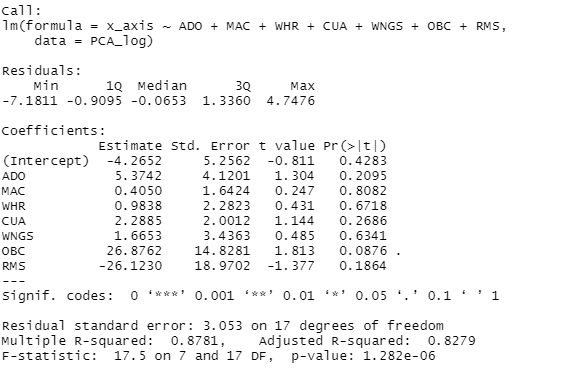


The regression model of original variables on sample unit shows only OBC have

significantly correlated with sample units(n=25)(OBC:*p*-value=.08).

The total R squared is 82.8% and these seven original variables can significantly

explained 82.8% of the sample units’variance(F(7,17)=17.5,*p*<0.001).



The regression model with data without log transform shows that the regression

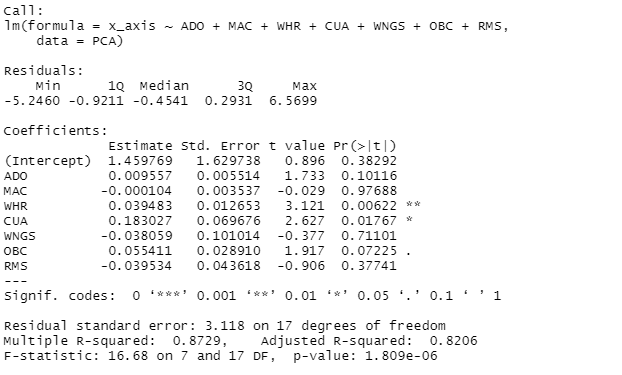
coefficient of WHR,CUA are significant(WHR: *p* =.006;CUA: *p* =.017) , which means

these 2 variables have significantly correlated with sample units(n=25).

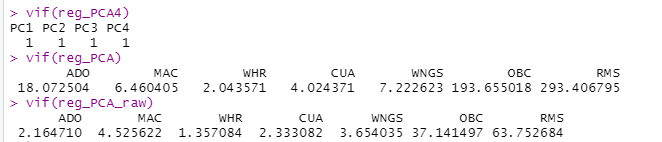
The total R squared is 82% and these seven original variables without log transform

can significantly explained 82% of variance of the sample units

(F(7,17)=16.68,*p*<0.001).



Result of VIF of these 3 model :



Only independent variables in PCA regression model aren’t collinear, since the vif of

four principal components are smaller than 10.

Original variables with log and without log, OBC and RMS in regression model have

collinear effect.

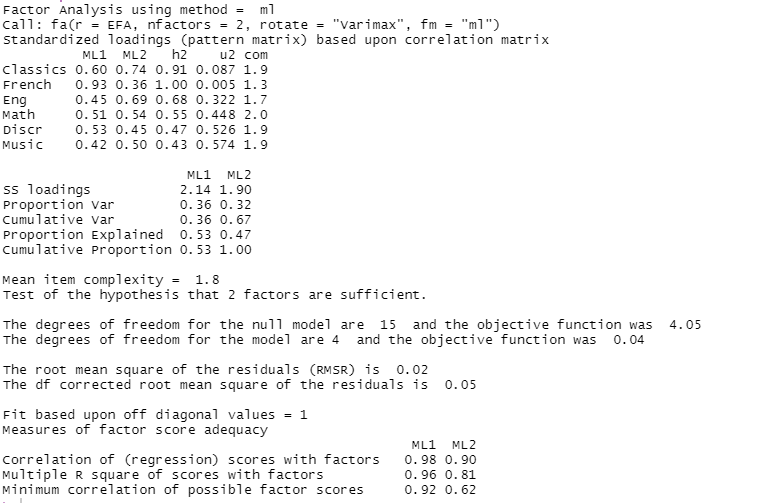
As a result, PCA can identify the meaningful underly latent variables among these 7

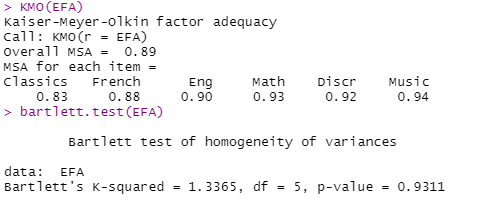
original variables and help to get a better understanding of the correlation structure

among the responses.

\_EFA

1. EFA with ‘Maximum Likelihood’, factor rotation: ‘Varimax’ with 2 factors.

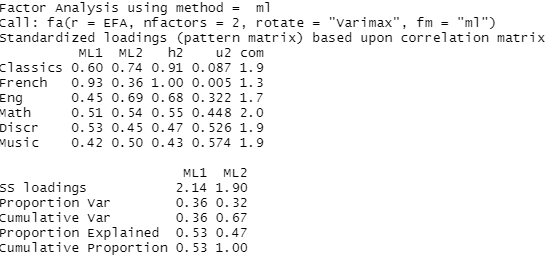




KMO of six variables are all higher 0.8 and they are factor well. It has common variances among these variables.

The Bartlett’s test of Sphericity isn’t significant( Bartlett’s K- squared = 1.33, *p*=.93).

The relationship among the variables isn’t strong enough to identify their underlying factors.



Total variance explained : 67 %

Community:



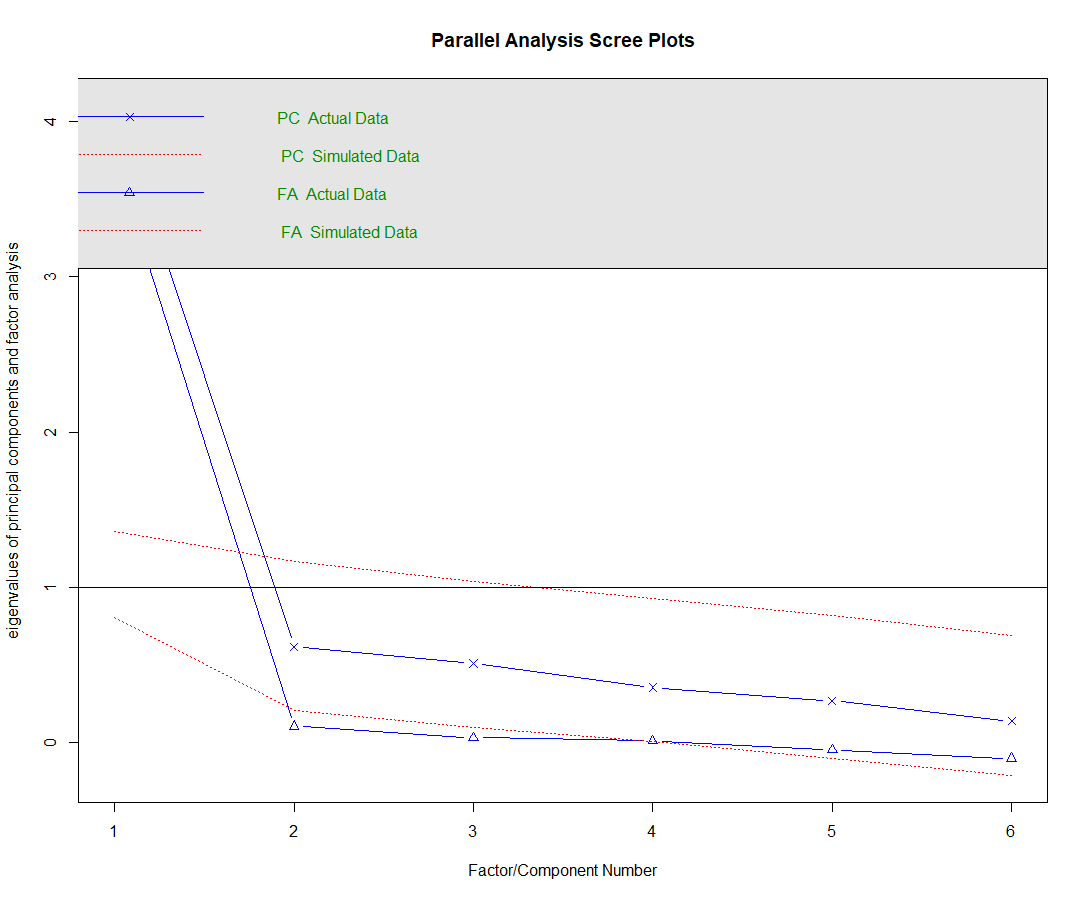
Community represent sum of squared factor loadings for that item(response variables).

Eigenvalue:

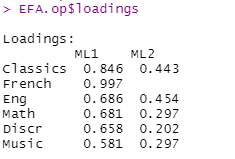


Eigenvalues of 6 response variables from common factors.

Scree plot:



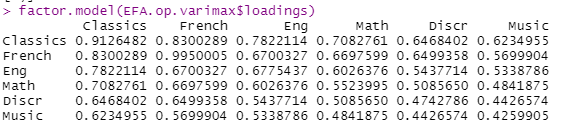
Factor matrix: This table contains the unrotated factor loadings, which are the correlations between the variable and the factor.



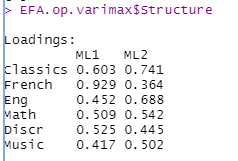
Goodness- of – fit



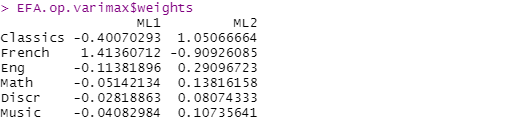
Reproduced correlation matrix : (the communality is the diagonal of the reproduced correlation matrix )



Rotated factor matrix:This table contains the rotated factor loadings, which represent both how the variables are weighted for each factor but also the correlation between the variables and the factor.



Factor score coefficient matrix : This is the factor weight matrix and is used to compute the factor scores.

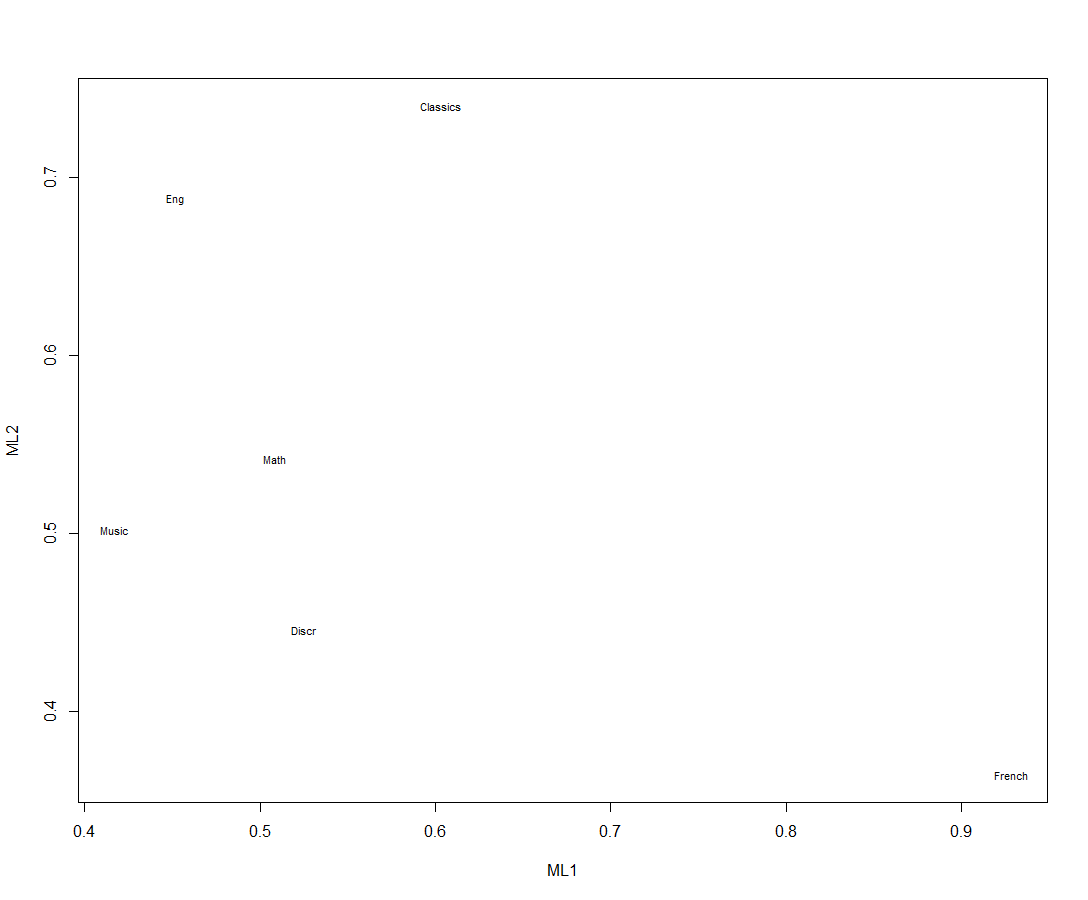


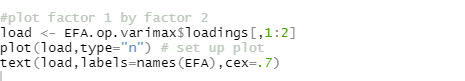
Factor score covariance matrix: The correlations of the factor score estimates using the specified model.



Because we used an orthogonal rotation, this should be a diagonal matrix.

Factor plot in rotated factor space:





3.

Because only six variables in the data, it’s possible that the factor analysis could only create a new set of uncorrelated variables, called underlying factors to better understanding the data.

4.

|  |  |  |
| --- | --- | --- |
| Variables | Community | Uniquenss |
| X1 | 24.3% | 75.7% |
| X2 | 22.8% | 77.2% |
| X3 | 21.3% | 78.7% |
| X4 | 19.7% | 80.3% |
| X5 | 18.2% | 81.8% |
| X6 | 16.7% | 83.3% |

