Student name: Syed Noor Razi Ali

SrudentID: 2070326

DSC 324/424 Assignment 3
Due: Tuesday, April 26th at 11:59pm CST

1) Principal Component Analysis

The data given in the file 'Employment.txt' is the percentage employed in different industries in Europe countries during 1979. Techniques such as Principal Component Analysis (PCA) can be used to examine which countries have similar employment patterns. There are 26 countries in the file and 10 variables as follows:

Variable Names:

1. Country: Name of country
2. Agr: Percentage employed in agriculture
3. Min: Percentage employed in mining
4. Man: Percentage employed in manufacturing
5. PS: Percentage employed in power supply industries
6. Con: Percentage employed in construction
7. SI: Percentage employed in service industries
8. Fin: Percentage employed in finance
9. SPS: Percentage employed in social and personal services
10. TC: Percentage employed in transport and communications.

Perform a principal component analysis using the covariance matrix:

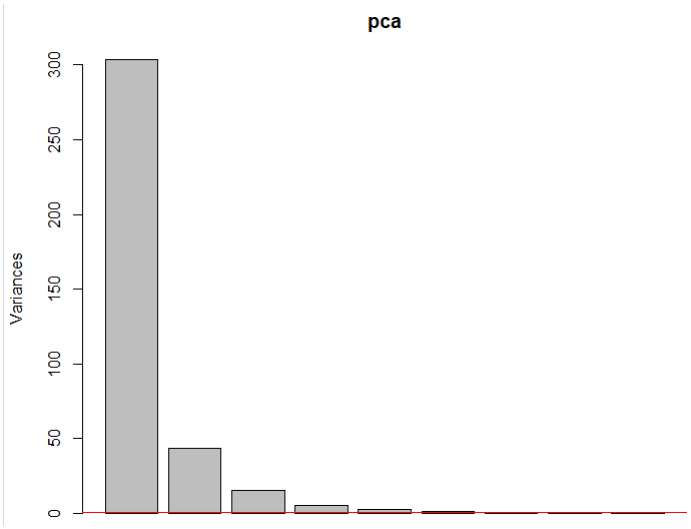
a. **How many principal components are required to explain 90% of the total variation for this data?**

Performing a principal component analysis on the employment dataset will allow us to see how many components explain a certain percentage of the variance in our data. Running the following commands produces the below output.

Looking at the output from our prcomp() command we see moving left to right that by the second component we already have a cumulative proportion of 93.33% the variability.

```
> pca = prcomp(emp)
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     17.4200 6.6107 3.89966 2.37473 1.56314 1.02276 0.64873 0.25481 0.04373
Proportion of Variance  0.8158 0.1175 0.04088 0.01516 0.00657 0.00281 0.00113 0.00017 0.00001
Cumulative Proportion   0.8158 0.9333 0.97415 0.98931 0.99588 0.99869 0.99982 0.99999 1.00000
```

To effectively analyze the right components to consider we want to make sure that there is not something being missed. Another way to decide on which components to use, is to plot a scree plot and look for a knee or leveling off of the graph. Using plot(pca) will produce the following:



From the above scree plot we can see that there may be some disagreement on the number of components to include. As the leveling off doesn't seem to occur until after 4 factors.

2 principle components explains 90% of the variance

b. For each of component from part a, give the formula for each component and a brief interpretation. Think about what kind of label you might give to each component.

Using print(pca) we can find the information for the equations for PC1 & PC2. We can see that:
PC1 = 0.8918*Agr + 0.0019*Min -0.2713*Man – 0.0084*Ps - 0.0496*Con – 0.1918*Si - 0.0311*Fin – 0.2980*Sps - 0454*Tc

PC2 = -0.0068*Agr + 0.0923*Min +0.7703*Man + 0.0120*Ps + 0.0690*Con - 0.2344*SI - 0.1301*Fin - 0.5668*Sps - 0.010*Tc

Looking at PC1 we see strong positive contributions from AGR, relatively flat MIN, and negative contributions from the rest. PC2 has strong positive contributions from MAN and strong negative contributions from SI and SPS. At this point in the analysis it is difficult to discern what the grouping or meaning of these different components are. The values show no strong patterns or relationships and seem to fluctuate a lot.

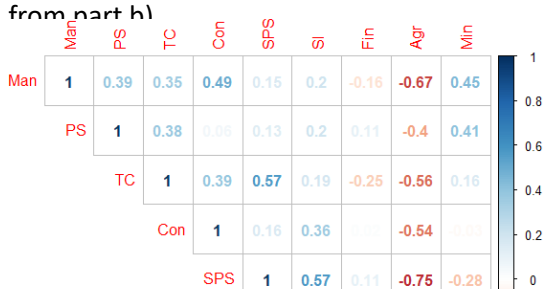
c. What countries have the highest and lowest values for each principal component (only include the number of components specified in part a). For each of those countries, give their principal component scores (again only for the number of components specified in part a).

Looking at the country level scores we can look at the top two positively contributing countries and top negatively contributing countries for each factor. In RC1 we see, Yugoslavia and Spain (-2.35 & -1.67 respectively) have the lowest contributing components and Norway and Denmark have the highest with scores of(1.76 & 1.30). Looking at the countries components for RC4 we get large negative contributions from Turkey and Yugoslavia again (-3.0 & -1.91) with similar positive contribution amounts from Spain and Switzerland (1.67 & 1.62). RC3 contains negative contributions from USSR & Poland (-2.06 1.41) with positive contributions from Yugoslavia and Netherlands (1.70 & 1.17). Finally RC2 has negative contributions from Turkey and
Italy (-1.57 & -1.40) and positive contributions from Hungary and E. Germany (2.37 & 1.70).

```
> scores = pfa$scores
> temp = cbind(employ[1], scores)
> temp
          Country         RC1         RC4         RC3         RC2
1         Belgium  0.92591453  0.07542789  0.85956704 -0.04092151
2         Denmark  1.30009650 -0.43551230  0.58946314 -1.09479554
3          France -0.02352915  0.43181322  0.97863849 -0.19674089
4       W. Germany  0.04017680  0.31719331  0.45082695  0.44795140
5         Ireland  0.32243811 -0.82393361  0.34997919  0.36647804
6           Italy -0.07982299  1.19034531 -0.07359559 -1.40413475
7      Luxembourg -0.36665929  1.04759193  0.18464058  0.58916378
8     Netherlands  0.90292228  0.24373341  1.17400734 -0.59271606
9  United Kingdom  0.84699227 -0.58752844  0.94904837  1.23421402
10        Austria -0.06371437  0.45409828  0.68367652  0.83050432
11        Finland  0.91735745 -0.73468043  0.61607651  0.47323270
12         Greece -0.56376872 -0.42036655 -0.55997229 -1.11492172
13         Norway  1.76758382 -0.30735511 -0.09081679 -0.66891853
14       Portugal -0.39608311  0.17171666 -0.06651434 -1.11253909
15          Spain -1.67414935  1.67037267  0.92730964 -0.64219876
16         Sweden  1.28629848 -0.67302949  0.60572134 -0.37875681
17    Switzerland -0.67544790  1.62334279  0.98399132 -0.39161800
18         Turkey -1.29459489 -3.00346779 -0.84773126 -1.56826256
19       Bulgaria -0.26112514  0.34246633 -1.38516227 -0.25395597
20 Czechoslovakia -0.30445415  0.62949955 -1.18931043  1.18139744
21     E. Germany  0.61389684  0.44261806 -1.18847064  1.69753674
22        Hungary  0.11804574 -0.41581898 -1.07268825  2.36620274
23         Poland -0.42476189 -0.05879183 -1.41297674  0.26259356
24        Rumania -1.55229776  0.66980722 -1.10583823 -0.29501981
25           USSR  0.98821691  0.05733526 -2.05669202 -0.86633094
26     Yugoslavia -2.34953105 -1.90687736  1.69682241  1.17255619
```

d. Analyze the entries in the correlation matrix for fields that are highly correlated or completely uncorrelated with the other fields. If there are highly uncorrelated fields, try removing them from the model. Does this help your interpretation from part b)

after plotting the correlation plot



We can look at the upper triangle of the correlation matrix at a confidence level of .9 and look for fields that a re highly correlated with the other fields.

We can see that AGR stands out right away with negative correlations at 40% with all but one of the other fields. MIN also seams to have correlations with the other variables but a less severe level. We can rerun analysis with AGR taken out to see if we can get improved interpretability of the other variables.

If we want to look at correlations that are above our 90% confidence level we can run a corr.test() and look for p- values < .10. Looking at highly uncorrelated variables it looks like FIN shows statistically insignificant correlations with 6 of the 9 variables. We can rerun our analysis without FIN to also see if it helps with interpretability.

```
> corrTest = corr.test(emp, adjust="none")
> round(corrTest$p, 2)
     Agr  Min  Man   PS  Con   SI  Fin  SPS   TC
Agr 0.00 0.86 0.00 0.04 0.00 0.00 0.28 0.00 0.00
Min 0.86 0.00 0.02 0.04 0.90 0.04 0.02 0.16 0.44
Man 0.00 0.02 0.00 0.05 0.01 0.32 0.45 0.45 0.08
PS  0.04 0.04 0.05 0.00 0.77 0.32 0.59 0.52 0.06
Con 0.00 0.90 0.01 0.77 0.00 0.07 0.94 0.44 0.05
SI  0.00 0.04 0.32 0.32 0.07 0.00 0.07 0.00 0.36
Fin 0.28 0.02 0.45 0.59 0.94 0.07 0.00 0.60 0.23
SPS 0.00 0.16 0.45 0.52 0.44 0.00 0.60 0.00 0.00
TC  0.00 0.44 0.08 0.06 0.05 0.36 0.23 0.00 0.00
```

Rerunning the analysis without the AGR or FIN variable, we see no clear benefit in interpretability of the output. With AGR we get less components on RC4 and without FIN we get MIN now negatively contributing to RC1 instead of TC contributing when included.

2) Principal Component Analysis

Begin with the “census2.csv” datafile, which contains census data on various tracts in a district. The

fields in the data are:

- Total Population (thousands)
- Professional degree (percent)
- Employed age over 16 (percent)
- Government employed (percent)
- Median home value (dollars)

a. Conduct a principal component analysis using the covariance matrix. How much of the variance is accounted for in the first component? Why is this happening?

Performing a principal component analysis on the “census2.csv” data set using the covariance matrix, we can use prcomp(census), as covariance is the standard matrix

used. At first the output may seem strange:

```
> cen = read.csv("census2.csv")
> pca = prcomp(cen)
> pca
Standard deviations (1, .., p=5):
[1] 56446.885008    10.206857     6.218887     2.246707     1.559823

Rotation (n x k) = (5 x 5):
                      PC1           PC2           PC3           PC4           PC5
Population     8.537905e-07 -4.108282e-02 -7.059713e-02  4.826860e-01  8.719762e-01
Professional   3.775797e-05  7.080539e-02 -7.460074e-02 -8.714029e-01  4.796648e-01
Employed      -1.367095e-06 -5.126328e-01 -8.542663e-01 -1.524163e-02 -8.487872e-02
Government     3.004471e-05  8.546967e-01 -5.095880e-01  8.624903e-02 -4.873218e-02
MedianHomeVal  1.000000e+00 -2.901832e-05  1.701961e-05  2.987813e-05 -1.750755e-05
> summary(pca)
Importance of components:
                          PC1   PC2   PC3   PC4  PC5
Standard deviation     56447 10.21 6.219 2.247 1.56
Proportion of Variance     1  0.00 0.000 0.000 0.00
Cumulative Proportion      1  1.00 1.000 1.000 1.00
```

The first principal component accounts for 100% of the variance. PC2, PC3, PC4, and PC5 contain none of the variance. Because PC1 contained 100% of the variance this indicates that all the variables can be written as a linear transformation of just one of the variables. The other factor likely contributing to the large proportion of variance in PC1 is the scaling factor between MedianHomeValue and all the other variables, as this variable is in $100,000's and the rest are in integers.

b. Try dividing the MedianHomeValue field by 100,000 so that the median home value in the dataset is measured in $100,000's rather than in dollars. How does this change the analysis?

Taking the MedianHomeVal variable and dividing all observations by 100,000 greatly improves the analysis. We can see that after scaling the variable we get much more interpretable variables.

```
> centran = cen %>% mutate(MedianHomeVal = MedianHomeVal/100000)
> ###Rerun PCA
> pcaTran = prcomp(centran)
> summary(pcaTran)
Importance of components:
                            PC1    PC2     PC3     PC4        PC5
Standard deviation     10.3444 6.2979 2.87105 1.68799 3.977e-06
Proportion of Variance  0.6783 0.2514 0.05225 0.01806 0.000e+00
Cumulative Proportion   0.6783 0.9297 0.98194 1.00000 1.000e+00
```

We can see that there variance captured by the different PC's is much more spread out. After PC2 we already have ~93% of the variance.

c. Compute the PCA with the correlation matrix instead. How does this change the result and how does your answer compare with your answer in b)?

Using the correlation matrix instead of the covariance matrix requires additional principal components to be included in order to encompass > 90% of the variance. The correlation matrix forces us to use 4 principal components to get to ~95% of the variance.

```
> cenCor = prcomp(centran, scale. = TRUE)
> summary(cenCor)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5
Standard deviation     1.4114 1.1694 0.9296 0.7315 0.49126
Proportion of Variance 0.3984 0.2735 0.1728 0.1070 0.04827
Cumulative Proportion  0.3984 0.6719 0.8447 0.9517 1.00000
> print(cenCor)
Standard deviations (1, .., p=5):
[1] 1.4113534 1.1694129 0.9296006 0.7314787 0.4912604

Rotation (n x k) = (5 x 5):
                    PC1         PC2         PC3         PC4         PC5
Population     0.2625829 -0.4629936  0.78390268 -0.2169291  0.2347882
Professional  -0.5933541 -0.3256442 -0.16407255  0.1446471  0.7028828
Employed       0.3256978 -0.6051419 -0.22487455  0.6628689 -0.1943206
Government    -0.4792022  0.2524850  0.55070086  0.5716730 -0.2766497
MedianHomeVal -0.4932213 -0.4996473 -0.06882436 -0.4072024 -0.5801162
```

On the other hand, using the correlation matrix greatly improves the interpretability of PC1. Using the covariance matrix, we get very small and wildly varying components as shown below.

```
> print(pcaTran)
Standard deviations (1, .., p=5):
[1] 1.034439e+01 6.297903e+00 2.871047e+00 1.687995e+00 3.976986e-06

Rotation (n x k) = (5 x 5):
                      PC1           PC2           PC3           PC4           PC5
Population     3.889563e-02 -7.114843e-02  1.943086e-01  9.775834e-01 -5.436147e-07
Professional  -1.052217e-01 -1.293106e-01 -9.680732e-01  1.871936e-01 -1.370029e-06
Employed       4.924464e-01 -8.644066e-01  4.428274e-02 -9.130649e-02  4.545578e-08
Government    -8.630829e-01 -4.806432e-01  1.520444e-01 -3.086221e-02  6.308581e-08
MedianHomeVal -9.094877e-08 -1.462226e-07 -1.232264e-06  7.939867e-07  1.000000e+00
> print(cenCor)
Standard deviations (1, .., p=5):
[1] 1.4113534 1.1694129 0.9296006 0.7314787 0.4912604
```

From the scaled PCA analysis we can clearly see negative contributions from professional, government and medianHomeVal, with smaller positive contributions from population and employed.

d. Discuss what using the correlation matrix does and why it may or may not be appropriate in this case.

When performing PCA analysis with the correlation matrix, you are effectively normalizing each of the variables. In virtually all circumstances, scaling the variables helps smooth out discrepancies in how the variables were assessed and is strongly recommended. Although the correlation matrix standardizes all variables rather than just the one we standardized to use the covariance matrix, it may have been appropriate to use the covariance matrix if the variables were all on similar scales, as we had done. The correlation matrix should be used for datasets with big variables, as the covariance matrix would allow the large variables to dominate the study and generate unnecessary bias.

3. Common Factor Analysis

For this problem, you will analyze partial from intelligence tests given to children in the 'wiscsem.csv' dataset. Each child was given 11 tests on which they were rated. These were:

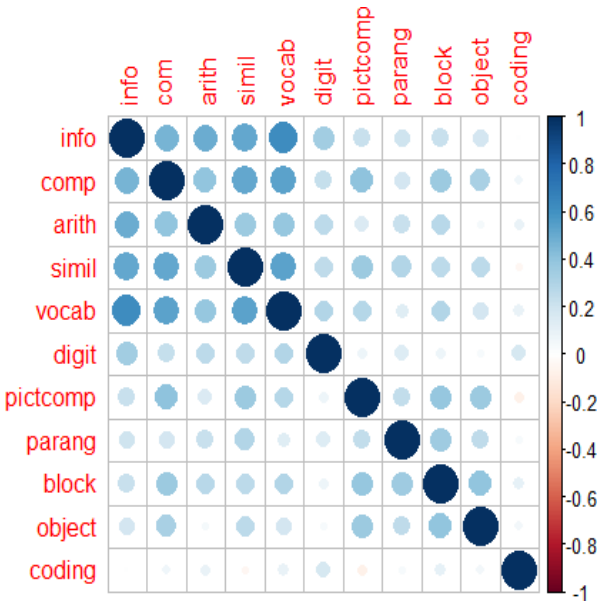
info = 'Information'
comp = 'Comprehension'
arith = 'Arithmetic'
simil = 'Similarities'
vocab = 'Vocabulary'
digit = 'Digit Span'
pictcomp = 'Picture Completion'
parang = 'Paragraph Arrangement'
block = 'Block Design'
object = 'Object Assembly'
coding = 'Coding';

a. Should the data be scaled or not for running PCA? Explain why/why not in detail.

This data does not need to be scaled when running PCA. No scaling is needed because all the variables measure the same thing and we can assume that the grading method was not changed from test to test as this would make little practical since when comparing performance across the factors.

b. Run an initial corrplot and an initial PCA. Use the corrplot and a scree plot to determine the appropriate number of factors to extract.

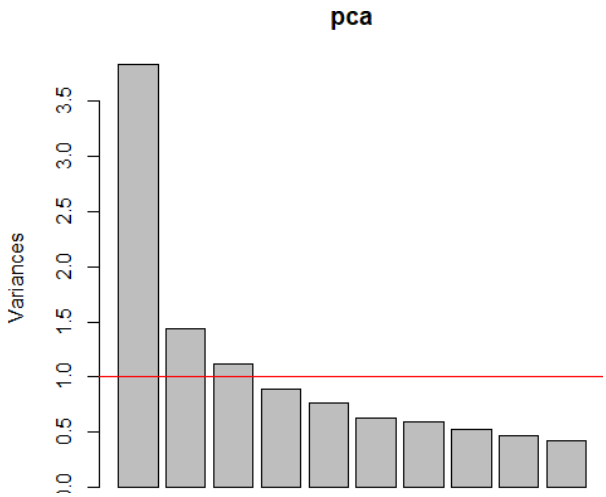
```
> pca = prcomp(int)
> summary(pca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     1.9567 1.2008 1.0565 0.94345 0.87656 0.79547 0.77155 0.72274
Proportion of Variance 0.3481 0.1311 0.1015 0.08092 0.06985 0.05753 0.05412 0.04749
Cumulative Proportion  0.3481 0.4792 0.5806 0.66153 0.73138 0.78891 0.84303 0.89051
                          PC9    PC10    PC11
Standard deviation     0.68621 0.64701 0.56111
Proportion of Variance 0.04281 0.03806 0.02862
Cumulative Proportion  0.93332 0.97138 1.00000
```



From the corrplot on the left we can see two different groups start to form in squares one, in the upper left corner and one in the bottom right. This could indicate two groups that have meaningful relationships and provide insightful information. We can then run PCA to see how much variance is captured by the components and get a better idea of how many we need to include moving forward.

Looking at the above output we can see that PC1 contains ~35% of the variance and going out to PC4 we get ~66% of the variance. We don't get 90% until all the way at the 9th

The next thing we can do to decide how many factors is right is plot the components and look for which components are above the variance = 1 line. On the right we can see that there are three components clearly above the line and then a steady leveling off below.



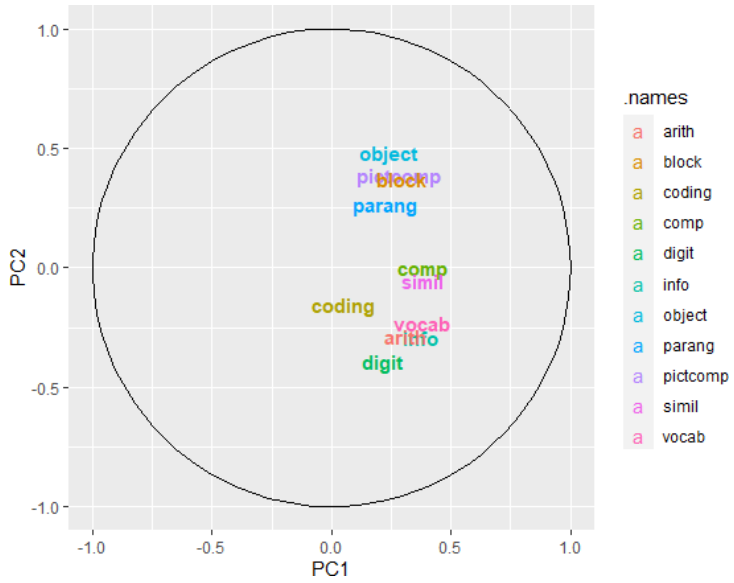
Looking at these results in tandem it seems like including more than 6 PC's would be force the components into separate groups. Because of the grouping in the corrplot and plot of the components on the right it seems like 3 is the right number of components to include.

c. Are there any variables that will likely be single-variable factors? Explain.

```
> pca = prcomp(int)
> pca
Standard deviations (1, .., p=11):
 [1] 1.9566923 1.2008306 1.0564764 0.9434468 0.8765614 0.7954743 0.7715489 0.7227389
 [9] 0.6862114 0.6470055 0.5611119

Rotation (n x k) = (11 x 11):
                PC1         PC2         PC3         PC4         PC5         PC6
info      0.3759142 -0.292206046  0.16171349 -0.01123982  0.04408530 -0.26962707
comp      0.3854910  0.004458532  0.08878963  0.27930558  0.06351471  0.12680963
arith     0.3089402 -0.287998864 -0.04146540 -0.27722242  0.52012156  0.20016393
simil     0.3804344 -0.051138453  0.22776197 -0.04104799 -0.15148512 -0.23906435
vocab     0.3802284 -0.231987278  0.10988177  0.26675798  0.06244542 -0.17797878
digit     0.2151143 -0.391218257 -0.26032409 -0.17010947 -0.68341696  0.32299764
pictcomp  0.2856439  0.389194610  0.16350032  0.12833053 -0.17175699  0.61634646
parang    0.2295444  0.266792609 -0.23872977 -0.74426179 -0.07037295 -0.22049472
block     0.2935647  0.373593094 -0.25052810 -0.02884532  0.34404849  0.21441716
object    0.2365222  0.485568453 -0.13035366  0.24719591 -0.26018301 -0.44839856
coding    0.0515469 -0.155461173 -0.81795291  0.33300989  0.09525039 -0.04962257
                PC7         PC8         PC9         PC10        PC11
info       0.14255703 -0.12694786  0.44572266  0.26535809  0.60569171
comp      -0.04500436  0.24738709 -0.65644077  0.48418207  0.12959272
arith      0.22586174  0.51326745  0.15399000 -0.15854725 -0.26294820
simil     -0.39068265  0.10120144 -0.23356465 -0.68784280  0.16864097
vocab     -0.13386030 -0.47121444  0.10518427  0.10819733 -0.64812540
digit      0.33443402 -0.06246306 -0.09582146 -0.06939902 -0.05325721
pictcomp  -0.32725162  0.09109001  0.44332862  0.04898130  0.00449372
parang    -0.30504694 -0.05796978 -0.04693002  0.32006047 -0.10289803
block      0.35927776 -0.52459464 -0.18028703 -0.27006872  0.19127717
object     0.41812098  0.35533445  0.16174921 -0.03470215 -0.18158030
coding    -0.37441691  0.09966942  0.12040379 -0.03722570  0.12526898
```

Looking at the variable contributions to component one we can see that there appears to be two different levels of contributions. One group with variables like info, comp, arith, simil, and vocab si contributing at a ~0.3 level, with the other variables (except for coding) contributing at a ~0.2 level. To better assess where these groups might be forming, we can use PCA_Plot to visualize these groupings.



With the plot above we can see that because everything is positively contributing to PC1, it's hard to clearly see any real groupings. Yet, these two groupings are much more apparent in PC2 when they have opposite contributions. We can see the more visual and spatial variables (object, block, parang, pictcomp, etc.) with positive contributions to PC2, indicating the significance of spatial intelligence to this component. On the other hand, the other group, the more written critical thinking skills, contribute negatively to PC2.

d. Run a Principal Factor Analysis with VARIMAX rotation and report the loadings with a cutoff of 0.4. How clean are the variable separations? Give a name to each factor.

Running a principal factor analysis with a cutoff of .4 and rotating the data with varimax makes it even clearer to interpret the groupings of distinct factors. On the right we can see the critical thinking group contributing to RC1, while RC2 has contributions from the more spatial and visual factors.

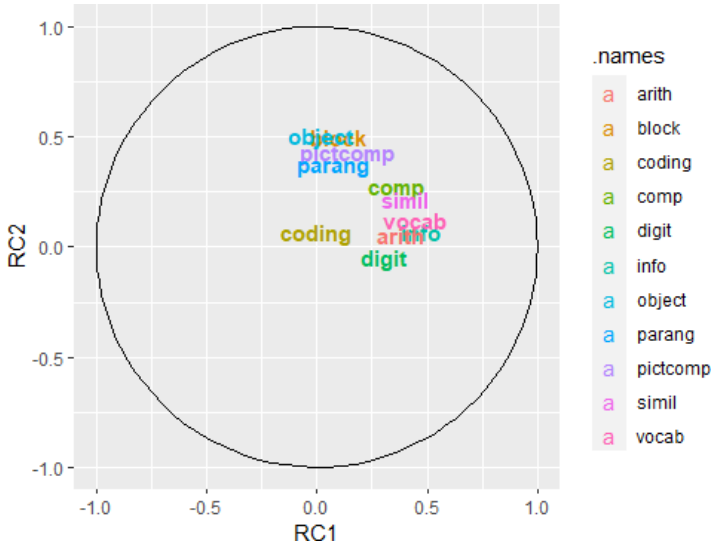
RC3 doesn't have a lot of highly contributing factors but the fact that coding was not correlated with any other variables this could indicate it as its own group and that there is a relationship between it and the digit variable. Performing the rotation also helped spotlight the interesting aspect that comprehension seems to contribute positively to both RC1 and RC2 indicating it could potentially be a sign of

```
> pfa = principal(int, nfactors = 3, rot = "varimax")
> print(pfa$loadings, cutoff = .4)

Loadings:
        RC1    RC2    RC3
info     0.826
comp     0.634  0.416
arith    0.669
simil    0.694
vocab    0.782
digit    0.535         0.428
pictcomp        0.649
parang          0.567
block           0.743
object          0.756
coding                 0.883

                RC1   RC2   RC3
SS loadings   3.022 2.211 1.154
Proportion Var 0.275 0.201 0.105
Cumulative Var 0.275 0.476 0.581
> PCA_Plot_Psyc(pfa)
```

intelligence in both groups. Using the PCA_Plopt_Psych function we can visualize these groupings and see that the two groups both positively contribute to their respective factors at roughly the same level.



e. Run a Common Factor Analysis (exploratory) and compare the loadings to those of the principal factor analysis. Note any significant differences and explain how they affect the factors practically.

PFA

```
> pfa = principal(int, nfactors =3, rot = "varimax")
> print(pfa$loadings, cutoff = .4, sort = TRUE)

Loadings:
         RC1    RC2    RC3
info      0.826
comp      0.634  0.416
arith     0.669
simil     0.694
vocab     0.782
digit     0.535         0.428
pictcomp         0.649
parang           0.567
block            0.743
object           0.756
coding                  0.883

                 RC1   RC2   RC3
SS loadings    3.022 2.211 1.154
Proportion Var 0.275 0.201 0.105
Cumulative Var 0.275 0.476 0.581
```

CFA

```
> fact = factanal(int,3)
> print(fact$loadings, cutoff = .2, sort = TRUE)

Loadings:
         Factor1 Factor2 Factor3
info      0.779
comp      0.551   0.449
arith     0.556           0.269
simil     0.620   0.366
vocab     0.721   0.252
pictcomp  0.202   0.605
block             0.714   0.380
object            0.573
digit     0.431
parang            0.392
coding                    0.290

               Factor1 Factor2 Factor3
SS loadings      2.399   1.801   0.410
Proportion Var   0.218   0.164   0.037
Cumulative Var   0.218   0.382   0.419
```

Running a common factor analysis (above right) we can see the similarity between the output. While overall there is less variance captured by the common factor analysis the

contributions to the factors is very similar. We can see that the loadings on Factor 1, 2 & 3 are generally slightly lower than in the principal factor analysis. This optimized method for computing the factors seems to be in agreement with the principal factor method except for on factor 3. While both factors have contributions from coding indicating it could be a separate group that is important to include, the models do not agree with what other variables share a relationship with it. The other thing it could be telling us is that it is not in fact important and possibly better to only include two factors one for overall performance and the other for differences in spatial intelligence.