DA410_Project7_MattGraham

Exploratory Factor Analysis

```
library(nnspat)  # used for dist2full()
library("dplyr")  # used to select numeric datatypes
library("ggplot2")
library(reshape)  # used for melting matricies
library(klaR)
library(ggvis)
library(class)
library(gmodels)
library(MASS)
library(readxl)
library(psych)
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.2.2
```

Get data

```
subject <- read.csv("subject.csv", header=TRUE)
subject
```

BIO	GEO	CHEM	ALG	CALC	STAT
<int>	<int>	<int>	<int>	<int>	<int>
1	1	1	1	1	1
4	4	3	4	4	4
2	1	3	4	1	1
2	3	2	4	4	3
3	1	2	2	3	4
1	1	1	4	4	4
3	3	3	2	3	1
4	3	4	2	3	2
2	1	3	3	4	3
2	3	3	2	3	4

1-10 of 300 rows Previous **1** 2 3 4 5 6 ... 30 Next

Assumption check:

Variables used should be metric. Dummy variables can also be considered, but only in special cases. -> check

Sample size: Sample size should be more than 200. -> check

Homogeneous sample: A sample should be homogenous. Violation of this assumption increases the sample size as the number of variables increases. Reliability analysis is conducted to check the homogeneity between variables.

Correlation: At least 0.30 correlations are required between the research variables.

```
subject.cor <- cor(subject)
as.data.frame(subject.cor)
```

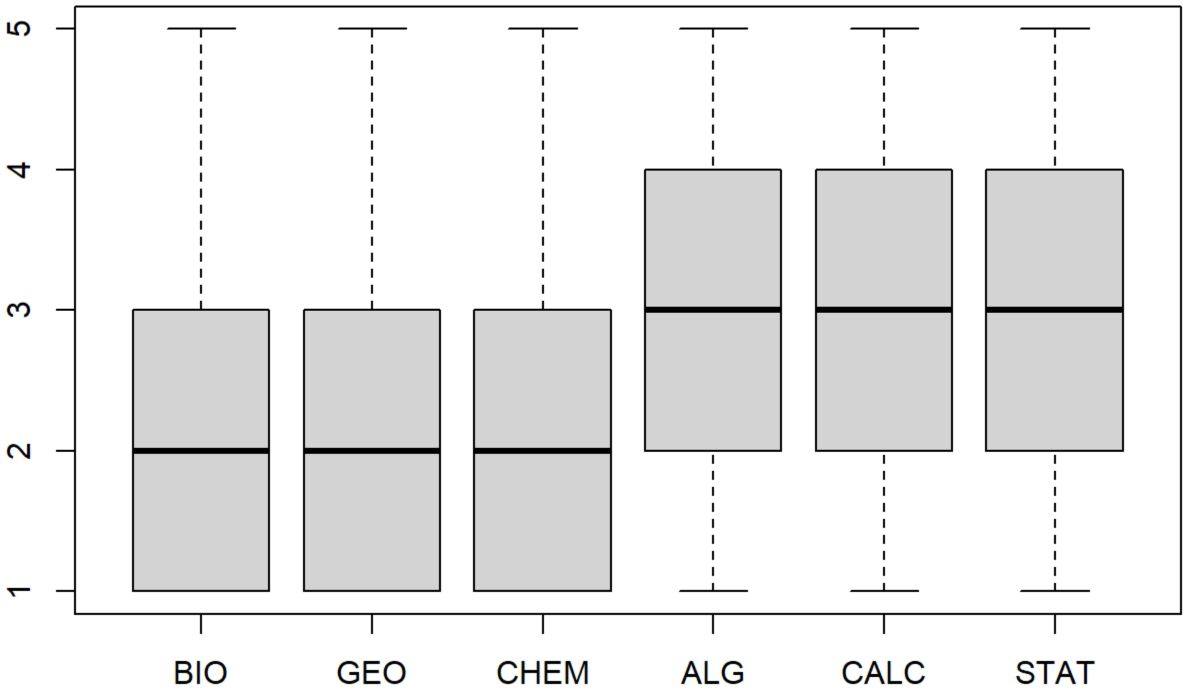
	BIO	GEO	CHEM	ALG	CALC	STAT
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
BIO	1.0000000	0.6822208	0.7470278	0.1153204	0.2134271	0.2028315
GEO	0.6822208	1.0000000	0.6814857	0.1353557	0.2045215	0.2316288
CHEM	0.7470278	0.6814857	1.0000000	0.0838225	0.1364251	0.1659747
ALG	0.1153204	0.1353557	0.0838225	1.0000000	0.7709303	0.4094324
CALC	0.2134271	0.2045215	0.1364251	0.7709303	1.0000000	0.5073147
STAT	0.2028315	0.2316288	0.1659747	0.4094324	0.5073147	1.0000000

6 rows

We have a few correlations that are unable to be compared, and will be noted through analysis

There should be no outliers in the data.

```
boxplot(subject)
```



There are no outliers

Running fa

```
solution <- fa(r = subject.cor, nfactors = 2, rotate = "oblimin", fm="pa")
```

```
## Loading required namespace: GPArotation
```

```
## Warning in fac(r = r, nfactors = nfactors, n.obs = n.obs, rotate = rotate, : I
## am sorry, to do these rotations requires the GPArotation package to be installed
```

```
solution
```

```
## Factor Analysis using method =  pa
## Call: fa(r = subject.cor, nfactors = 2, rotate = "oblimin", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##       PA1   PA2   h2    u2 com
## BIO  0.76 -0.42 0.75 0.255 1.6
## GEO  0.71 -0.36 0.63 0.369 1.5
## CHEM 0.72 -0.47 0.75 0.253 1.7
## ALG  0.51  0.62 0.65 0.354 1.9
## CALC 0.65  0.70 0.92 0.081 2.0
## STAT 0.45  0.30 0.29 0.709 1.8
##
##                        PA1  PA2
## SS loadings           2.48 1.50
## Proportion Var        0.41 0.25
## Cumulative Var        0.41 0.66
## Proportion Explained  0.62 0.38
## Cumulative Proportion 0.62 1.00
##
## Mean item complexity =  1.7
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  15  and the objective function was  2.87
## The degrees of freedom for the model are 4  and the objective function was  0.01
##
## The root mean square of the residuals (RMSR) is  0.01
## The df corrected root mean square of the residuals is  0.02
##
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                    PA1  PA2
## Correlation of (regression) scores with factors   0.96 0.94
## Multiple R square of scores with factors          0.91 0.89
## Minimum correlation of possible factor scores     0.83 0.79
```

Similar to maximum likelihood. WLS = weighted least squared.

Overall, our model does a great job explaining ~96% of variation when using 2 factors. Our most-ideal values to model from would be Calculus and either Biology or Geography. We can also see in our console output that hypothesis tests with 2 factors are sufficient.