Report of Factor Analysis

June 27 2019

1 Introduction of Factor Analysis

Factor Analysis (FA) is a statistical method to describe the variability among observed. Therefore, it is similar with Principal Components Analysis (PCA) in some way. The ideology of FA is that there may be two or three unobserved variables that can represent the variance of ten variables. FA can be regarded as an update of PCA, as FA is clearly designed with the objective to identify certain unobservable factors from the observed variables, whereas PCA does not directly address this objective (from wiki).

2 Notation

Symbol	Meaning	Note
p	Dimension of the data	$-$
k	Dimension of the common Factor	$k < p$
n	Number of Observations	$-$
x_i^t	Random variable i of observation t	R
x^t	Observation t	R^p
μ_i	Mean of each random variable	R^p
F	Common factor	R^k
L	Loading Matrix	$R_{p\times k}$

The method of implementing FA is the following:
We assume that the value of x can be presented by the following formula:

$$x - \mu = LF + \epsilon$$

Or elementwisely:

$$x_i - \mu_i = \sum_j^k l_{ij} F_j + \epsilon_i$$

Here, ϵ is a white noise, and the following assumptions are satisfied:

- F and ϵ are independent;
- $E(F) = 0$;
- $cov(F) = I$;

3 Test the Validity of Implementing FA

3.1 Bartlett's Test of Sphericity

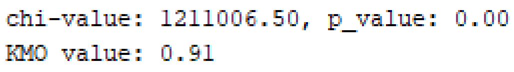
Bartlett's test of sphericity tests whether the correlation matrix is an identity matrix. That is, it shows that whether our data is uncorrelated and suitable for structure detection. If we get significant values, for example less than 0.05, it means that our data is suitable for implementing FA.

3.2 Kaiser-Meyer-Olkin Measure of Sampling Adequacy

The short version is the KMO test. It is a measure that shows approximately what proportion of the variance in our data might be caused by unobserved factors. If the value of the measurement is greater than 0.5, then FA can provide useful information.

3.3 Implementing Two Tests to Our Data

The related package of factor analysis in python is "factor_analyzer", and two functions called "calculate_bartlett_sphericity" and "calculate_kmo" are for two tests. I implemented the two tests to the data and get the following result:



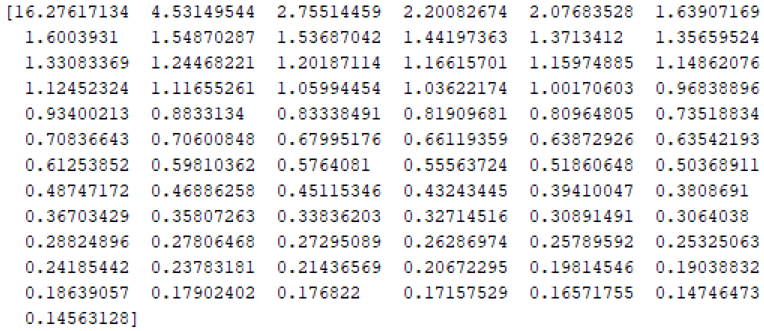
```
chi-value: 1211006.50, p_value: 0.00
KMO value: 0.91
```

Figure 1: Result of BS and KMO Tests

The p value of BS test is 0.00, which means that the categories of our data are uncorrelated. The value of KMO test is 0.91, which means that 0.91 of the variance of our data can be interpret by some underlying factors. Both of the tests show that FA is suitable for our data.

4 Determine the Number of Factors

To determine how many factors we should decompose to, first we try to decompose to 73 factors (maximum factors) and see the eigenvalues of those factors, which represent the variance. The eigenvalues are shown as following:



```
[16.27617134  4.53149544  2.75514459  2.20082674  2.07683528  1.63907169
  1.6003931   1.54870287  1.53687042  1.44197363  1.3713412   1.35659524
  1.33083369  1.24468221  1.20187114  1.16615701  1.15974885  1.14862076
  1.12452324  1.11655261  1.05994454  1.03622174  1.00170603  0.96838896
  0.93400213  0.8833134   0.83338491  0.81909681  0.80964805  0.73518834
  0.70836643  0.70600848  0.67995176  0.66119359  0.63872926  0.63542193
  0.61253852  0.59810362  0.5764081   0.55563724  0.51860648  0.50368911
  0.48747172  0.46886258  0.45115346  0.43243445  0.39410047  0.3808691
  0.36703429  0.35807263  0.33836203  0.32714516  0.30891491  0.3064038
  0.28824896  0.27806468  0.27295089  0.26286974  0.25789592  0.25325063
  0.24185442  0.23783181  0.21436569  0.20672295  0.19814546  0.19038832
  0.18639057  0.17902402  0.176822    0.17157529  0.16571755  0.14746473
  0.14563128]
```

Figure 2: Eigenvalues of 73 Factors

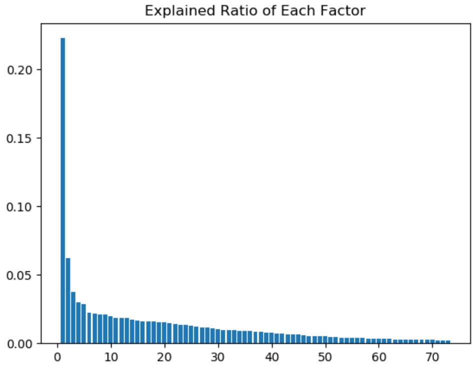


Figure 3: Explained Ratio of Each Factor

Originally, according to the Kaiser Criterion, we only should preserve number of components with eigenvalues greater than 1. However, Kaiser Criterion is for data that are standardized. As in our case, it is not meaningful to standardize data, so to reduce components, we can increase the threshold. For example, I make the threshold as $\lambda = 2$, then only the first 5 factors are taken. (If the threshold is 1.25, then the first 12 factors are taken, which is also possible, but it takes more time to evaluate and interpret) Here I will take 5 as example.

5 Performing FA

First we performed FA when the number of factor is 5, and the set the rotation parameter as "varimax". The result is attached to the appendix 1. Note that, I pre-process the data by when the answer of question t_{38}, t_{39} is greater than 6, I set it to 6. If the response is negative number -1, I set it to NaN and regard it as invalid data. Therefore, if the pre-process of data is different, then the result may be a little bit different. However, we see that there are many coefficients in the loading matrix are small, but we want to interpret each factor, then we need to abandon those small coefficients. When we set all the coefficient that are smaller than 0.25 to 0, we get the matrix attached to appendix 2. However, I still save some coefficients for interpretability.

6 Interpretation of the Result of FA

6.1 Interpretation Based on Factors

When we conclude the variables that are dominating each factors, we get the following table:

Factor	Dominating Questions
Factor1	$a1, a2, a3, a4, a5, a6, b7, b8, b9, c11, c14, de17, de18, de19, de20,$ $f21, f22, f23, f24, f25, f26, g27, g28, g29, g30, g31, h32, h33$
Factor2	$c10, c11, c12, c13, h32, h33, h34, h35, h36, h37, t38, t39,$ $ij40_6, ij41_6, ij42_6, ij43_6, ij47_6, ij48_6, ij55_6, ij56_6,$ $ij40_e, ij41_e, ij42_e, ij43_e, ij45_e, ij47_e, ij48_e, ij55_e, ij56_e$
Factor3	$c11, c13, c14, c15, c16, de17, de18, de19, de20, f21, f22, f25, f26,$ $ij43_6, ij44_6, ij45_6, ij46_6, ij52_6,$ $ij43_e, ij44_e, ij45_e, ij46_e, ij52_e$
Factor4	$a6, f23, f24,$ $ij45_6, ij46_6, ij49_6, ij50_6, ij51_6,$ $ij45_e, ij46_e, ij49_e, ij50_e, ij51_e$
Factor5	$h34, h37,$ ij$47_6, ij48_6, ij53_6, ij54_6, ij55_6, ij56_6,$ $ij47_e, ij48_e, ij53_e, ij54_e, ij55_e, ij56_e$

Table 1: Classification of Questions

The distribution of factor shows an interesting pattern to interpret.

- For Factor 1, it contains ALL questions from **a,b,de,f,g** categories, and it also contains some questions from category **c,h**. They represent the performance of antisocial tendencies, weak parental supervision, impulsive risk taking, delinquent beliefs or neutralization of guilt for offending, and negative peer influence. **This factor mainly measures the background and personality of the child.**

- For Factor 2, it contains questions from **c,h,t** and some question from **ij-6 month, ij-ever**. Questions in **c,h,t** represents the influence of critical events, friends, and parents behaviors (peer delinquency and family gang membership). For questions in **ij** sections, $ij40, ij41, ij42$ are question about whether the child use cigarettes, alcohol, and drugs; $ij43$ is about skipping classes; $ij47, ij48$ are about illegally holding weapons and painting; $ij55, ij56$ are about gang fight and marijuana. Those questions are related with friends and surroundings. **This factor mainly measures the effect of critical event and children's surroundings.**

- For Factor 3, it mostly contains questions from **c,de,f**, which is measuring the impact of critical life events relating with impulsive risk taking factor and delinquent beliefs. It also contains some question from **ij-6 month, ij-ever**. For those **ij** problems, some of them are measuring risk behavior (such as $ij43$,$ij46$,$ij52$), and some of them are measuring delinquent beliefs (such as $ij44$,$ij45$). **Therefore, the factor mainly measures the critical events influence together with people's personality.**

- For Factor 4, it contains three questions in section **a,f** sections. However, viewing the three questions, they are all about **stealing behaviors**. Also $ij45, ij49, ij50, ij51$ are all about **stealing behaviors**, while $ij46$ is about destroying items that do not belong to you (with a lower weight). **Therefore, the factor mainly measures the stealing behavior of children.**

- For Factor 5, the two questions other from **ij** sections are whether your friends belong to a gang or have attacked someone. Also, for questions in **ij** sections, some questions are about gang related behaviors (such as $ij48$, $ij56$), the others are about violence behaviors (such as $ij47$,$ij53$,$ij54$,$ij55$). **Therefore, the factor mainly measures the gang and violence behavior related with friends and gangs.**

6.2 Interpretation Based on Questions

It is worth mentioning that all the questions are classified to some clusters. That is, 5 factors are enough to cover all the questions, that is, the whole questionnaire.

Then we can analyze the number of appearance in the 5 factors for each question:

Time	Question
3 Times	$c11, ij43_6, ij43_e$
2 Times	$a6, c13, c14, de17, de18, de19, de20, f21, f22, f23, f24, f25, f26,$ $h32, h33, h34, h37, ij45_6, ij45_e, ij46_6, ij46_e,$ $ij47_6, ij47_e, ij48_6, ij48_e, ij55_6, ij55_e, ij56_6, ij56_e$

Table 2: Questions that Appears Multiple Times

From the table, we can observe that all questions in section **de,f** appears at least twice, which in some sense indicates that Impulsive Risk Taking and Neutralization may play important roles in the questionnaire. Also, the table also indicates the questions that appear at least twice in section **ij**. It is noticeable that question $ij43$ appears three times, and it is about "Have you ever skipped classes". This question is somehow indicated to be important.

Also, another finding from the factor analysis is that: **It verifies the validity of considering questions with "in 6 month" and "ever" as a whole.** Check in Table 1, we can observe that in each factor, whenever $ijxx_6$ exists, its corresponding question $ijxx_e$ also exists in that factor. It holds all the time for 5 factors, which indicates the strong correlation between $ijxx_6$ and $ijxx_e$.

Appendix 1

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
0	0.557	0.034	0.142	0.058	0.056
1	0.471	0.041	0.225	-0.026	0.065
2	0.590	0.130	0.128	0.033	0.030
3	0.483	0.039	0.094	0.102	0.099
4	0.538	0.073	0.217	0.090	0.058
5	0.481	0.081	0.084	0.334	0.083
6	0.563	0.197	0.011	0.092	0.083
7	0.598	0.227	0.030	0.108	0.072
8	0.587	0.233	0.053	0.104	0.066
9	0.226	0.247	0.213	0.010	-0.018
10	0.257	0.261	0.277	0.058	0.097
11	0.168	0.249	0.221	0.002	-0.000
12	0.165	0.304	0.252	0.008	-0.003
13	0.291	0.034	0.336	0.010	0.028
14	0.131	0.007	0.177	-0.005	-0.001
15	0.095	0.035	0.263	-0.001	0.104
16	0.628	0.148	0.260	0.063	0.071
17	0.683	0.144	0.275	0.057	0.055
18	0.596	0.094	0.299	0.026	0.032
19	0.635	0.155	0.302	0.026	0.048
20	0.579	0.122	0.244	0.121	0.033
21	0.611	0.085	0.264	0.126	0.018
22	0.490	0.129	0.063	0.336	0.094
23	0.496	0.156	0.083	0.368	0.081
24	0.531	0.155	0.377	0.034	0.055
25	0.500	0.090	0.338	0.016	0.046
26	0.396	0.102	-0.040	0.072	0.081
27	0.384	0.150	-0.066	0.089	0.119
28	0.520	0.120	0.138	0.069	-0.025
29	0.564	0.153	0.159	0.076	-0.010
30	0.451	0.255	0.020	0.100	0.065
31	0.321	0.454	0.246	0.061	0.071
32	0.316	0.397	0.260	0.199	0.151
33	0.156	0.347	0.143	0.082	0.348
34	0.181	0.627	0.100	0.068	0.208
35	0.240	0.699	0.187	0.084	0.114
36	0.149	0.479	0.238	0.028	0.272

Table 3: Loading Matrix when # of Factor is 5 (First Half)

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
37	0.204	0.390	0.196	0.036	0.226
38	0.112	0.250	0.279	0.013	0.158
39	0.153	0.635	0.052	0.148	0.029
40	0.140	0.701	0.011	0.113	0.057
41	0.078	0.236	0.007	0.087	0.129
42	0.325	0.386	0.329	0.122	-0.026
43	0.097	0.075	0.432	0.235	0.087
44	0.170	0.223	0.391	0.315	0.008
45	0.226	0.088	0.390	0.252	0.181
46	0.068	0.260	0.182	0.081	0.435
47	0.115	0.377	0.131	0.122	0.277
48	0.199	0.131	0.177	0.601	0.060
49	0.101	0.166	0.013	0.532	0.237
50	0.090	0.048	0.132	0.503	0.125
51	0.229	0.044	0.562	0.113	0.138
52	0.058	0.032	0.101	0.063	0.531
53	0.049	0.038	-0.016	0.139	0.409
54	0.049	0.287	0.165	-0.001	0.393
55	0.028	0.376	-0.036	0.105	0.297
56	0.142	0.675	0.0656	0.144	-0.016
57	0.125	0.748	0.006	0.105	0.025
58	0.081	0.282	0.014	0.098	0.132
59	0.307	0.436	0.332	0.115	-0.041
60	0.087	0.108	0.442	0.242	0.083
61	0.159	0.266	0.402	0.318	-0.010
62	0.222	0.135	0.399	0.264	0.165
63	0.061	0.319	0.199	0.094	0.430
64	0.107	0.442	0.135	0.129	0.268
65	0.183	0.192	0.193	0.587	0.039
66	0.091	0.237	0.021	0.540	0.243
67	0.081	0.103	0.146	0.526	0.139
68	0.218	0.100	0.563	0.118	0.117
69	0.052	0.102	0.106	0.089	0.568
70	0.051	0.081	-0.015	0.167	0.438
71	0.044	0.377	0.176	0.001	0.417
72	0.017	0.460	-0.034	0.117	0.308

Table 4: Loading Matrix when # of Factor is 5 (Second Half)

Appendix 2

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
0	0.557	0.000	0.000	0.000	0.000
1	0.471	0.000	0.000	0.000	0.000
2	0.590	0.000	0.000	0.000	0.000
3	0.483	0.000	0.000	0.000	0.000
4	0.538	0.000	0.000	0.000	0.000
5	0.481	0.000	0.000	0.334	0.000
6	0.563	0.000	0.000	0.000	0.000
7	0.598	0.000	0.000	0.000	0.000
8	0.587	0.000	0.000	0.000	0.000
9	0.000	0.247	0.000	0.000	0.000
10	0.254	0.261	0.277	0.000	0.000
11	0.000	0.249	0.000	0.000	0.000
12	0.000	0.304	0.252	0.000	0.000
13	0.291	0.000	0.336	0.000	0.000
14	0.000	0.000	0.177	0.000	0.000
15	0.000	0.000	0.263	0.000	0.000
16	0.628	0.000	0.260	0.000	0.000
17	0.683	0.000	0.275	0.000	0.000
18	0.596	0.000	0.299	0.000	0.000
19	0.635	0.000	0.302	0.000	0.000
20	0.579	0.000	0.244	0.000	0.000
21	0.611	0.000	0.264	0.000	0.000
22	0.490	0.000	0.000	0.336	0.000
23	0.496	0.000	0.000	0.368	0.000
24	0.531	0.000	0.377	0.000	0.000
25	0.500	0.000	0.338	0.000	0.000
26	0.396	0.000	0.000	0.000	0.000
27	0.384	0.000	0.000	0.000	0.000
28	0.520	0.000	0.000	0.000	0.000
29	0.564	0.000	0.000	0.000	0.000
30	0.451	0.000	0.000	0.000	0.000
31	0.321	0.454	0.000	0.000	0.000
32	0.316	0.397	0.000	0.000	0.000
33	0.000	0.347	0.000	0.000	0.348
34	0.000	0.627	0.000	0.000	0.000
35	0.000	0.699	0.000	0.000	0.000
36	0.000	0.479	0.000	0.000	0.272

Table 5: Sparse Loading Matrix when # of Factor is 5 (First Half)

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
37	0.000	0.390	0.000	0.000	0.000
38	0.000	0.250	0.000	0.000	0.000
39	0.000	0.635	0.000	0.000	0.000
40	0.000	0.701	0.000	0.000	0.000
41	0.000	0.236	0.000	0.000	0.000
42	0.325	0.386	0.329	0.000	0.000
43	0.000	0.000	0.432	0.000	0.000
44	0.000	0.000	0.391	0.315	0.000
45	0.000	0.000	0.390	0.252	0.000
46	0.000	0.260	0.000	0.000	0.435
47	0.000	0.377	0.000	0.000	0.277
48	0.000	0.000	0.000	0.601	0.000
49	0.000	0.000	0.000	0.532	0.000
50	0.000	0.000	0.000	0.503	0.000
51	0.000	0.000	0.562	0.000	0.000
52	0.000	0.000	0.000	0.000	0.531
53	0.000	0.000	0.000	0.000	0.409
54	0.000	0.287	0.000	0.000	0.393
55	0.000	0.376	0.000	0.000	0.297
56	0.000	0.675	0.000	0.000	0.000
57	0.000	0.748	0.000	0.000	0.000
58	0.000	0.282	0.000	0.000	0.000
59	0.307	0.436	0.332	0.000	0.000
60	0.000	0.000	0.442	0.000	0.000
61	0.000	0.266	0.402	0.318	0.000
62	0.000	0.000	0.399	0.264	0.000
63	0.000	0.319	0.000	0.000	0.430
64	0.000	0.442	0.000	0.000	0.268
65	0.000	0.000	0.000	0.587	0.000
66	0.000	0.000	0.000	0.540	0.000
67	0.000	0.000	0.000	0.526	0.000
68	0.000	0.000	0.563	0.000	0.000
69	0.000	0.000	0.000	0.000	0.568
70	0.000	0.000	0.000	0.000	0.438
71	0.000	0.377	0.000	0.000	0.417
72	0.000	0.460	0.000	0.000	0.308

Table 6: Sparse Loading Matrix when # of Factor is 5 (Second Half)