Basilico Eleonora, Caporali Francesco, Malgieri Luigi Maria
10 May 2023

Multivariate Statistical Analysis
Problem Set 2

Exercise 1

Consider the data set `psych`, which contains 24 psychological tests ($t_i, \forall i \in \{1, \ldots, 24\}$) administered to 301 students, with ages ranging from 11 to 16, in a suburb of Chicago:

- 1st group of 156 students (74 boys, 82 girls) from the *Pasteur School*;
- 2nd group of 145 students (72 boys, 73 girls) from the *Grant-White School*.

```
psych_0 = read.table("data/psych.txt", header = T)
dim_p = dim(psych_0)
colnames(psych_0) = c(c("case", "sex", "age"), paste0("t_", 1:(dim_p[2] - 4)), "group")
psych_0[2] = tolower(unlist(psych_0[2]))
psych_0[28] = tolower(unlist(psych_0[28]))
```

case	sex	age	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_10	t_11	t_12	t_13
1	m	13.1	20	31	12	3	40	7	23	22	9	78	74	115	229
2	f	13.6	32	21	12	17	34	5	12	22	9	87	84	125	285
3	f	13.1	27	21	12	15	20	3	7	12	3	75	49	78	159
4	m	13.2	32	31	16	24	42	8	18	21	17	69	65	106	175
5	f	12.2	29	19	12	7	37	8	16	25	18	85	63	126	213
6	f	14.1	32	20	11	18	31	3	12	25	6	100	92	133	270

t_14	t_15	t_16	t_17	t_18	t_19	t_20	t_21	t_22	t_23	t_24	group
170	86	96	6	9	16	3	14	34	5	24	pasteur
184	85	100	12	12	10	-3	13	21	1	12	pasteur
170	85	95	1	5	6	-3	9	18	7	20	pasteur
181	80	91	5	3	10	-2	10	22	6	19	pasteur
187	99	104	15	14	14	29	15	19	4	20	pasteur
164	84	104	6	6	14	9	2	16	10	22	pasteur

The 24 tests correspons to the following subjects:

	t
t_1	visual perception
t_2	cubes
t_3	paper form board
t_4	flags
t_5	general information
t_6	paragraph comprehension
t_7	sentence completion
t_8	word classification
t_9	word meaning

	t
t_10	addition
t_11	code
t_12	counting dots
t_13	straight-curved capitals
t_14	word recognition
t_15	number recognition
t_16	figure recognition
t_17	object-number
t_18	number-figure
t_19	figure-word
t_20	deduction
t_21	numerical puzzles
t_22	problem reasoning
t_23	series completion
t_24	arithmetic problems

We can observe that part of our data is not numerical, in particular the variable `sex`. Since this variable has only two levels, we can proceed by transforming it into boolean.
We assign the values as reported:

$$\begin{cases} 0 & \text{if } \texttt{sex} = \text{M} \\ 1 & \text{if } \texttt{sex} = \text{F} \end{cases}.$$

```
psych_1 = psych_0
psych_1[2] = as.integer(psych_1[2] == "f")
```

Another important observation concerns the fact that the variable `case` is not relevant as it only corresponds to an enumeration of the students who were tested in sequential order (containing some gaps probably due to the absence of data for some students).

```
psych_2 = subset(psych_1, select = -case)
```

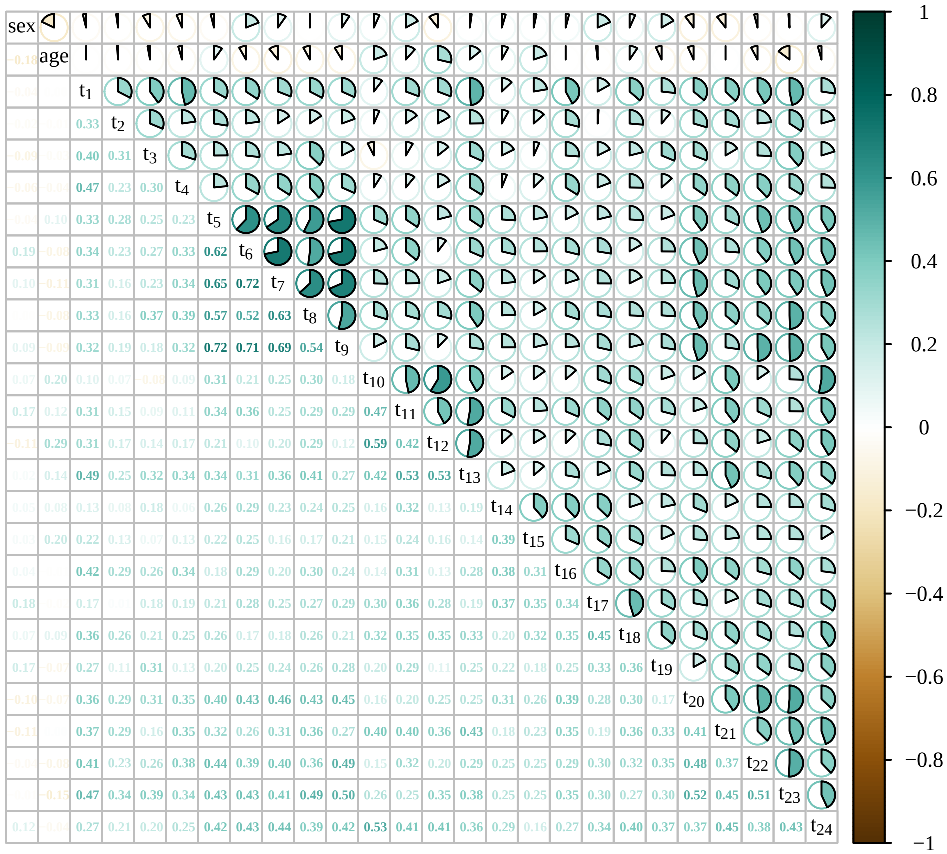
1.1

We are asked to use only the Grant-White students data, so we subset our data frame in accordance with the request.

```
gw = subset(psych_2, group == "grant", select = -group)
```

At this point we look at the correlation matrix of our data, a central object in the execution of the *Factor Analysis*.
Since we have a very large number of variables, we choose not to display the values directly of the matrix entries, but rather to display them via a plot.

```
cor_gw = cor(gw)
colnames(cor_gw) = c(c("sex", "age"), paste0("$t[", 1:(dim_p[2] - 4), "]"))
rownames(cor_gw) = colnames(cor_gw)
par(family = "serif")
corrplot.mixed(cor_gw, upper = "pie",
    upper.col = COL2("BrBG"), lower.col = COL2("BrBG"),
    number.cex = 0.4, tl.col = "black", tl.cex = 0.7, cl.cex = 0.7)
```



Looking at the `corrplot()` we just performed, we immediately realise that the variables `sex` and `age` are scarcely correlated with the 24 tests. For this reason it is reasonable to expect that in a Factor Analysis, including them would entirely characterise the factors in which they appear and have negligible loadings in the others. We will therefore initially avoid considering these first two variables and then comment on how the analysis would change by including them.
Another, more substantial, reason why we discard them is that we do not expect there to be a common factor on which they can depend, as they are in a sense primitive factors themselves.

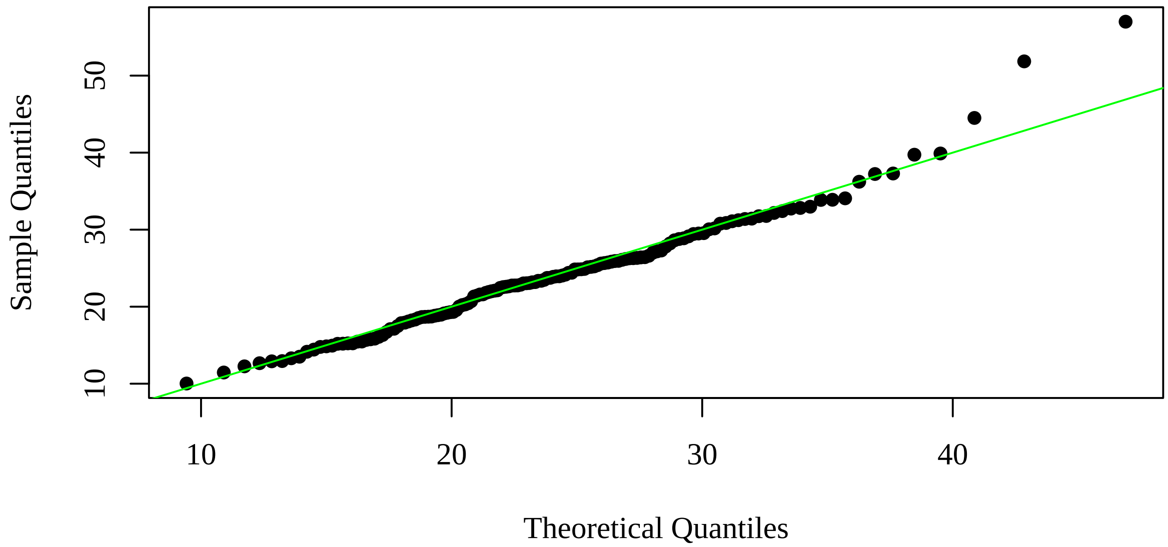
To obtain the maximum likelihood solution for $m = 5$ and $m = 6$ factors in R we can use the built-in function `factanal()`.
Before proceeding with the direct computation performed via software, we would like to recall that the *maximum likelihood* method, unlike the *principal component method*, relies on the necessary assumption of normality of the *common factors* ($\boldsymbol{F}$) and the *specific error terms* ($\boldsymbol{\varepsilon}$). Recalling also that if $\boldsymbol{F} = (F_1, \dots, F_m)$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_p)$ are normally distributed then

$$\boldsymbol{X} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon} \sim \mathcal{N}(\mu, \Sigma), \text{ with } L \in \mathbb{R}^{p \times m}$$

a first check that can be made is that our input data $\boldsymbol{x} \in \mathbb{R}^{24}$, appropriately rescaled using the command `scale()`, actually comes from a $\boldsymbol{X} \sim \mathcal{N}(0, I)$.

```
gws = scale(gw[, 3:dim(gw)[2]])
cor_gws = cor(gws)
dim_gws = dim(gws)
```

For this purpose we look at the Q-Q plot of the squared Mahalanobis distances *vs* a χ^2_{24}.



The plot shows that the variables jointly seem to follow Gaussian behaviour.

We now proceed with the computation of the maximum likelihood solution, first in the case of $m = 5$ factors, then with $m = 6$ (without any rotation):

```
faml_5 = factanal(gws, factors = 5, rotation = "none")
load_5 = faml_5$loadings[, ]
```

	Factor1	Factor2	Factor3	Factor4	Factor5
t__1	0.5549	-0.0032	0.4659	-0.1495	0.0015
t__2	0.3444	-0.0287	0.2917	-0.0563	0.1250
t__3	0.3734	-0.1422	0.4267	-0.1045	0.0418
t__4	0.4634	-0.1044	0.3032	-0.1128	0.1482
t__5	0.7226	-0.2536	-0.2249	-0.0756	-0.0044
t__6	0.7208	-0.3742	-0.1685	-0.0139	-0.1453
t__7	0.7278	-0.3355	-0.2323	-0.1317	0.0131
t__8	0.6917	-0.1442	-0.0421	-0.1066	0.0801
t__9	0.7232	-0.4245	-0.1967	0.0169	-0.0214
t__10	0.5182	0.6034	-0.3795	0.0411	0.1158
t__11	0.5701	0.3495	-0.0240	0.0649	-0.3670
t__12	0.4872	0.5444	0.0052	-0.1179	0.1277
t__13	0.6305	0.3467	0.2011	-0.3833	-0.2058
t__14	0.3929	-0.0013	0.0648	0.3688	-0.2378
t__15	0.3456	0.0268	0.1282	0.3678	-0.1281
t__16	0.4559	0.0247	0.3781	0.2755	-0.0855
t__17	0.4530	0.1283	0.0333	0.4382	-0.1130
t__18	0.4749	0.2521	0.2182	0.2588	0.0177
t__19	0.4179	0.0511	0.1376	0.1964	-0.0669
t__20	0.5961	-0.1672	0.1806	0.1546	0.2271
t__21	0.5741	0.2267	0.1539	0.0252	0.1590
t__22	0.5946	-0.1395	0.1803	0.1287	0.0982
t__23	0.6650	-0.0636	0.2131	0.0332	0.2445
t__24	0.6571	0.1864	-0.1262	0.1451	0.1292

```r
faml_6 = factanal(gws, factors = 6, rotation = "none")
load_6 = faml_6$loadings[, ]
```

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
t_1	0.5486	0.0039	0.4562	-0.1968	-0.0599	0.0333
t_2	0.3388	-0.0273	0.3009	-0.1585	0.0715	0.2322
t_3	0.3725	-0.1392	0.4443	-0.1107	0.0336	-0.2323
t_4	0.4600	-0.1066	0.3043	-0.1332	0.1257	-0.0871
t_5	0.7243	-0.2605	-0.2170	-0.0734	-0.0261	0.0920
t_6	0.7240	-0.3674	-0.1557	0.0278	-0.1458	0.0009
t_7	0.7329	-0.3539	-0.2340	-0.0919	0.0229	-0.1481
t_8	0.6953	-0.1550	-0.0401	-0.1049	0.0908	-0.2071
t_9	0.7277	-0.4211	-0.1804	0.0539	-0.0332	0.0922
t_10	0.5131	0.5871	-0.3853	-0.0239	0.1601	0.0291
t_11	0.5786	0.3898	-0.0434	0.0797	-0.4217	0.1269
t_12	0.4816	0.5361	-0.0146	-0.1655	0.1279	-0.1027
t_13	0.6175	0.3280	0.1533	-0.3573	-0.2278	-0.1395
t_14	0.3978	0.0305	0.0803	0.3532	-0.1307	0.0058
t_15	0.3494	0.0578	0.1457	0.3323	-0.0393	0.0969
t_16	0.4568	0.0562	0.3879	0.2097	-0.0402	0.0753
t_17	0.4744	0.1802	0.0697	0.5696	0.0082	-0.2565
t_18	0.4783	0.2777	0.2330	0.2208	0.0730	0.0072
t_19	0.4218	0.0713	0.1544	0.1842	-0.0259	-0.0171
t_20	0.5961	-0.1556	0.2009	0.0750	0.2310	0.0915
t_21	0.5706	0.2318	0.1513	-0.0958	0.1371	0.2158
t_22	0.5970	-0.1208	0.1977	0.0858	0.0702	0.1688
t_23	0.6616	-0.0583	0.2287	-0.0376	0.2257	0.0688
t_24	0.6561	0.1904	-0.1127	0.0757	0.1584	0.0672

Then we proceed with the computatio of the proportion of total sample variance due to each factor.
We recall that, according to the theory the operator ptsv that compute the proportion of total sample variance due to a factor is defined as

$$\mathrm{ptsv}(k) = \frac{\sum_{j=1}^{m} \hat{l}_{j,k}^2}{\mathrm{trace}\left(\boldsymbol{S}\right)},$$

with $\left(\hat{l}_{j,k}\right)_{j,k=1}^{m}$ loadings and $\boldsymbol{S}$ sample covariance matrix.
Due to the scaling performed at the beginning of the computation in our case $\mathrm{trace}\left(\boldsymbol{S}\right) = \mathtt{size}(\boldsymbol{S}) = 24$ (it is indeed a sample correlation matrix).

```r
ptsv_5 = colSums(load_5^2) / dim_gws[2]
```

	Factor1	Factor2	Factor3	Factor4	Factor5
ptsv_5	0.3159	0.0698	0.0548	0.04	0.0223

```r
ptsv_6 = colSums(load_6^2) / dim_gws[2]
```

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
ptsv_6	0.3168	0.0711	0.0563	0.0417	0.0212	0.0175

Actually this computation is also performed as a part of the output of the command `factanal()`, together with the sum of the squares of the loadings and the cumulative proportion of sample variance:

```
faml_5
```

	Factor1	Factor2	Factor3	Factor4	Factor5
ss_load_5	7.5813	1.6743	1.3161	0.9589	0.5351
ptsv_5	0.3159	0.0698	0.0548	0.0400	0.0223
ctsv_5	0.3159	0.3856	0.4405	0.4804	0.5027

```
faml_6
```

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
ss_load_6	7.6024	1.7068	1.3515	1.0000	0.5086	0.4192
ptsv_6	0.3168	0.0711	0.0563	0.0417	0.0212	0.0175
ctsv_6	0.3168	0.3879	0.4442	0.4859	0.5071	0.5245

Both models seem to fit very poorly. In both cases ($m = 5, 6$), they explain about 50% (respectively 50.27% and 52.45%) of the total variance collectively.
Recall that a general criterion, valid for both factor extraction methods seen, is to take m factors with m such that

$$m = \#\text{ factors necessary to account for 80\% of the total variance.}$$

Next, as requested, the specific variances $(\psi_j)_{j=1}^{24}$ are reported below, again for both $m = 5$ and 6. In this case we directly exploit the output of `factanal()` in order not to have to recalculate the values of the specific variances of the factors by hand. We report the results of the computation below:

```
psi_5 = faml_5$uniquenesses
```

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_10	t_11	t_12
0.4526	0.7766	0.6456	0.6477	0.3573	0.2907	0.2863	0.4812	0.2573	0.2082	0.4134	0.4361
t_13	t_14	t_15	t_16	t_17	t_18	t_19	t_20	t_21	t_22	t_23	t_24
0.2525	0.6489	0.7118	0.5654	0.5724	0.596	0.7607	0.5086	0.5695	0.5683	0.4474	0.4799

```
psi_6 = faml_6$uniquenesses
```

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_10	t_11	t_12
0.4474	0.7098	0.5771	0.6433	0.346	0.2946	0.2519	0.4288	0.2481	0.2164	0.3111	0.4262
t_13	t_14	t_15	t_16	t_17	t_18	t_19	t_20	t_21	t_22	t_23	t_24
0.2885	0.6925	0.732	0.5864	0.3473	0.5857	0.7583	0.5128	0.5233	0.5491	0.4494	0.4853

Finally, it is required to assess the accuracy of the given approximation of the correlation matrix. For this purpose, we analyse the residual given by the difference of the starting correlation matrix, $\boldsymbol{R}$, and the correlation matrix given by the approximation performed by the previous procedure, i.e. $\boldsymbol{S} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}}$.
Then we compare its squared Frobenius norm with the sum of the squares of the neglected eigenvalues, i.e. $\sum_{i=m+1}^{\texttt{size}(\boldsymbol{S})} \lambda_i^2$.

```r
eig = eigen(cor_gws)$values
residual_5 = cor_gws - (load_5 %*% t(load_5) + diag(psi_5))
eig_negl_5 = eig[(5 + 1):dim_gws[2]]
comparison_5 = c(sum(residual_5^2), sum(eig_negl_5^2))
```

	ss_residual_5	ss_eig_negl_5
comparison_5	0.7335	5.7823

Then we repeat exactly the same computation for $m = 6$:

```r
residual_6 = cor_gws - (load_6 %*% t(load_6) + diag(psi_6))
eig_negl_6 = eig[(6 + 1):dim_gws[2]]
comparison_6 = c(sum(residual_6^2), sum(eig_negl_6^2))
```

	ss_residual_6	ss_eig_negl_6
comparison_6	0.602	4.9392

Clearly the theoretical required condition is fulfilled. Indeed it should be valid

$$\left\| \boldsymbol{R} - \left(\hat{\boldsymbol{L}} \hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}} \right) \right\|_{\mathrm{F}}^2 \leq \sum_{i=m+1}^{\texttt{size}(\boldsymbol{S})} \lambda_i^2$$

and in our case:

$m = 5: \; 0.7335059 \leq 5.7822848;$

$m = 6: \; 0.6020222 \leq 4.9391922.$

But it is also evident that in both cases the approximation error of the correlation matrix is not negligible.

We can therefore conclude that both choices are acceptable, but in some sense inaccurate, and observing that the improvement given by the choice of $m = 6$ is not particularly significant, we tend to prefer $m = 5$. Indeed the last factor obtained with $m = 6$ accounts only for the 1.75% of the total sample variance.

The same computation including the variables `sex` and `age` leads to a very similiar result:

```
gws_2 = scale(gw)
cor_gws_2 = cor(gws_2)
dim_gws_2 = dim(gws_2)
eig_2 = eigen(cor_gws_2)$values

faml_5_2 = factanal(gws_2, factors = 5, rotation = "none")
load_5_2 = faml_5_2$loadings[, ]
psi_5_2 = faml_5_2$uniquenesses
residual_5_2 = cor_gws_2 - (load_5_2 %*% t(load_5_2) + diag(psi_5_2))
eig_negl_5_2 = eig_2[(5 + 1):dim_gws_2[2]]
comparison_5_2 = c(sum(residual_5_2^2), sum(eig_negl_5_2^2))
ctsv_5_2 = cumsum(colSums(load_5_2^2) / dim_gws_2[2])

faml_6_2 = factanal(gws_2, factors = 6, rotation = "none")
load_6_2 = faml_6_2$loadings[, ]
psi_6_2 = faml_6_2$uniquenesses
residual_6_2 = cor_gws_2 - (load_6_2 %*% t(load_6_2) + diag(psi_6_2))
eig_negl_6_2 = eig_2[(6 + 1):dim_gws_2[2]]
comparison_6_2 = c(sum(residual_6_2^2), sum(eig_negl_6_2^2))
ctsv_6_2 = cumsum(colSums(load_6_2^2) / dim_gws_2[2])
```

	Factor1	Factor2	Factor3	Factor4	Factor5
ctsv_5_2	0.2902	0.3598	0.413	0.4533	0.4838

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
ctsv_6_2	0.2918	0.3615	0.4145	0.4565	0.4867	0.5098

	ss_residual_5_2	ss_eig_negl_5_2
comparison_5_2	1.0071	7.0013

	ss_residual_6_2	ss_eig_negl_6_2
comparison_6_2	0.7251	5.952

1.2

We now have to give an interpretation to the common factors in the $m = 5$ solution. Without any rotation the loadings are pretty difficult to interpret. Indeed, as we noticed in the previous point, when $m = 5$ almost all variables load on the first factor higher than on the other four factors. Therefore, a rotation may help in the interpretation process. As requested, we perform the Varimax rotation.

```
faml_5_var = factanal(gws, factors = 5, rotation = "varimax")
load_5_var = faml_5_var$loadings[,]
```

	Factor1	Factor2	Factor3	Factor4	Factor5
t_1	0.1654	0.6549	0.1250	0.1810	0.2066
t_2	0.1079	0.4416	0.0871	0.0954	0.0024
t_3	0.1341	0.5595	-0.0473	0.1115	0.0934
t_4	0.2305	0.5333	0.0895	0.0811	0.0124

	Factor1	Factor2	Factor3	Factor4	Factor5
t_5	0.7383	0.1893	0.1916	0.1486	0.0547
t_6	0.7724	0.1867	0.0318	0.2477	0.1243
t_7	0.7983	0.2140	0.1427	0.0883	0.0502
t_8	0.5710	0.3429	0.2391	0.1275	0.0423
t_9	0.8079	0.2024	0.0332	0.2188	-0.0072
t_10	0.1807	-0.1082	0.8451	0.1803	0.0264
t_11	0.1952	0.0661	0.4233	0.4365	0.4177
t_12	0.0297	0.2322	0.6944	0.1022	0.1285
t_13	0.1863	0.4329	0.4793	0.0775	0.5382
t_14	0.1846	0.0614	0.0443	0.5522	0.0797
t_15	0.1043	0.1223	0.0586	0.5089	-0.0028
t_16	0.0698	0.4061	0.0559	0.5087	0.0540
t_17	0.1543	0.0716	0.2104	0.5947	-0.0269
t_18	0.0323	0.2999	0.3219	0.4576	0.0043
t_19	0.1563	0.2209	0.1440	0.3785	0.0451
t_20	0.3728	0.4614	0.1265	0.2930	-0.1939
t_21	0.1717	0.3980	0.4312	0.2382	-0.0004
t_22	0.3637	0.4232	0.1139	0.3204	-0.0689
t_23	0.3615	0.5421	0.2482	0.2307	-0.1147
t_24	0.3680	0.1786	0.4952	0.3208	-0.0683

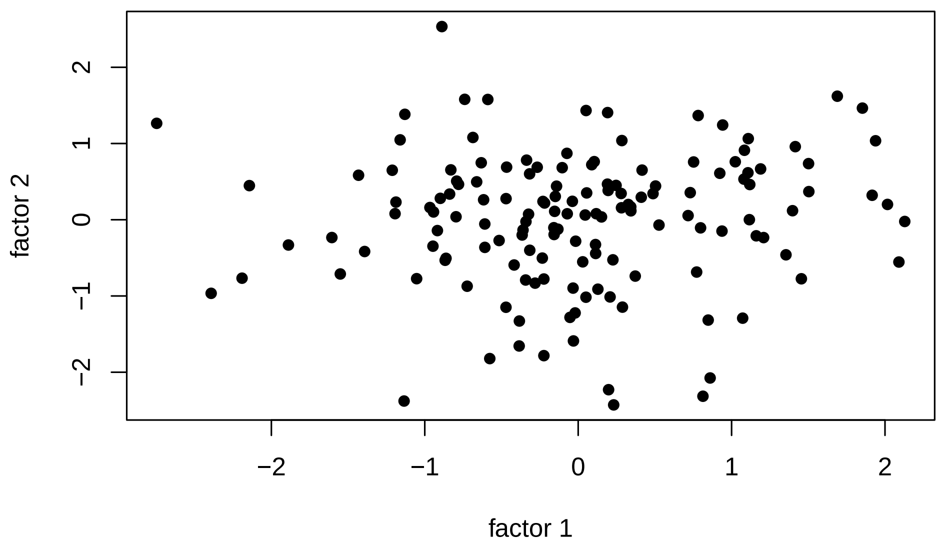
We choose not to visualize the results in a plot since there are too many factors and variables, so it woud have not been of any help.

After the rotation, things become a little better: as expected, the loadings are in general smaller or larger than the previous ones, and this facilitates the interpretation of the factors. In particular:

- The variables t_5, t_6, t_7, t_8 and t_9 load highly on the first common factor. The psychological tests associated to these variables primarily assess the language-related capacities of an individual, including reading comprehension, vocabulary knowledge, word associations, sentence construction and general knowledge. Hence, we can interpret the first factor as "verbal ability";
- The variables from the 14th to the 19th determine the fourth common factor. The tests associated to these variables measure an individual's capacity of recognising numbers, words and figures and of making associations between them. Hence, the fourth factor can be interpreted as "recognition and association ability";
- The fifth factor is solely determined by the variable t_{13}...
- The variables t_{10} and t_{12} load highly on the third factor, which is also determined by the variables t_{21} and t_{24}. These variables refer to psychological tests that assess cognitive capacities related to numerical processing, mathematical reasoning, problem-solving and arithmetic skills. We refer to the fourth factor as "numerical/mathematical ability";
- Finally, the second factor is determined by the variables from t_1 to t_4 togheter with t_{20}, t_{22} and t_{23}. The first four tests measure the visual ability of an individual, while the last three tests assess the logical ability of an individual. Hence, we choose to interpret the second factor as "logical and visual ability";
- Note also that the variable t_{11} loads similarly on the last three common factors....

1.3

We report below the scatterplot of the first two factor scores for the $m = 5$ solution obtained by the regression method, as requested.



It seems there is no particular correlation between the two factors. In fact, if we compute it we obtain

```
## [1] 0.07425218
```

1.4

1.5

Exercise 2

```
# code
```

2.1

2.2

2.3

2.4

2.5 (optional)