Basilico Eleonora, Caporali Francesco, Malgieri Luigi Maria
05 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)
```

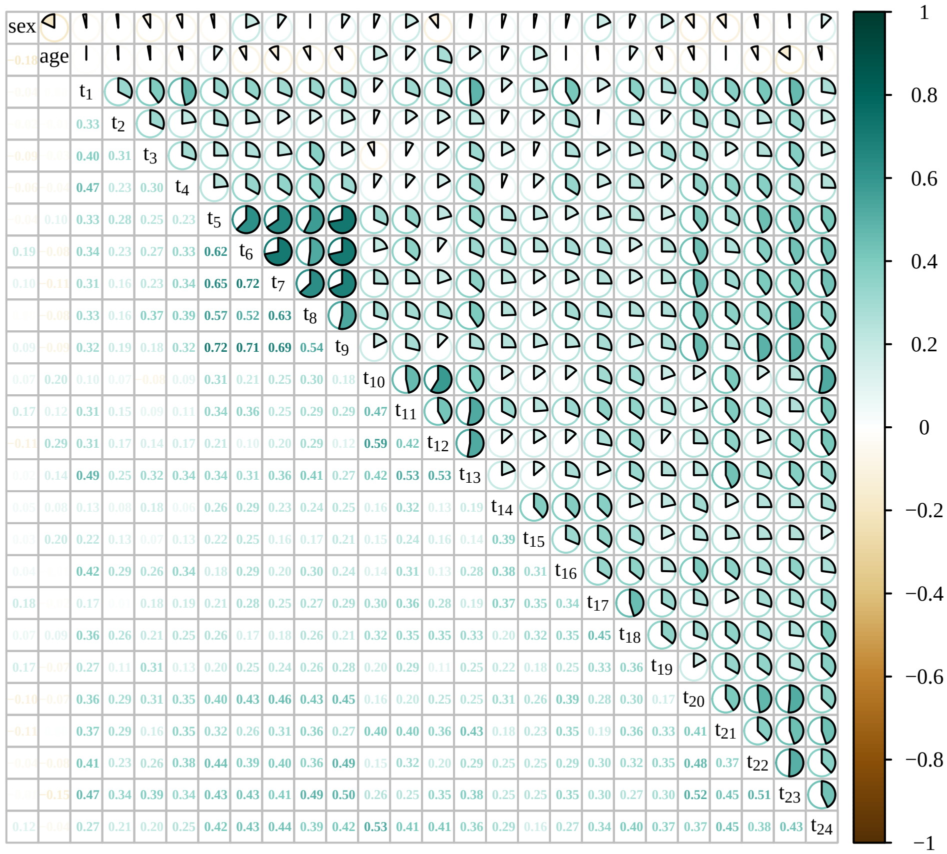
1.1

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)
```



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 $\boldsymbol{F} = (F_1, \ldots, F_m)$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)$ are normally distributed then

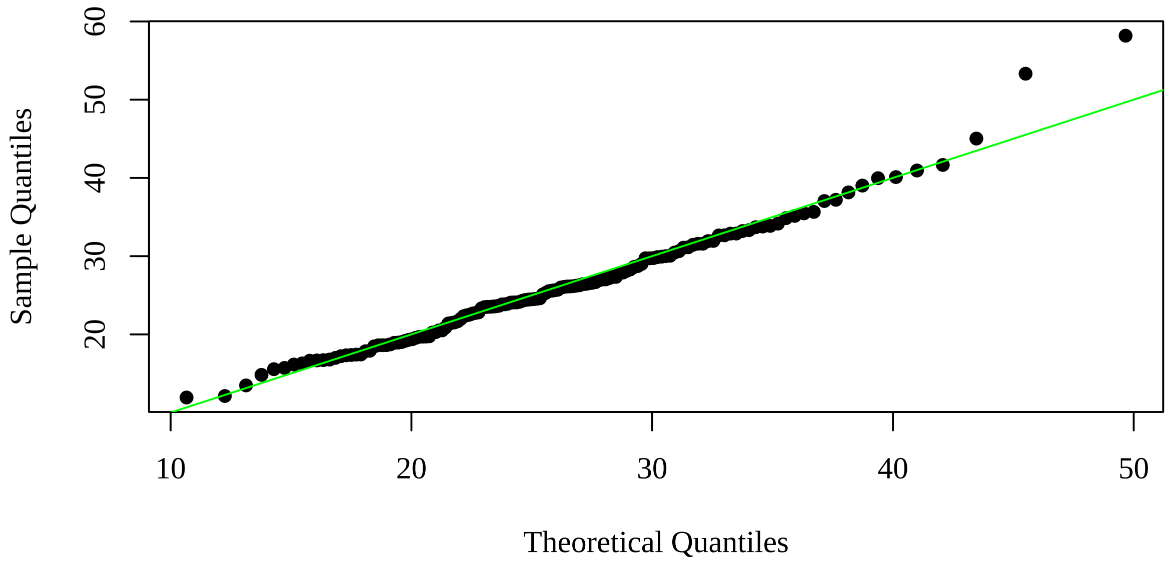
$$\boldsymbol{X} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon} \sim \mathcal{N}(\mu, \Sigma), \text{ with } L \in \mathbb{R}^{p \times m}$$

a first check that can be made is that our input data $\boldsymbol{x} \in \mathbb{R}^{2+24}$, appropriately rescaled using the command `scale()`, actually comes from a $\boldsymbol{X} \sim \mathcal{N}(0, I)$.

```
gws = scale(gw)
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_{2+24}.

```
par(family = "serif", mar = c(4, 4, 1, 1))
d = mahalanobis(gws, center = colMeans(gws), cov = cov(gws))
plot(qchisq(ppoints(d), df = ncol(gws)), sort(d), pch = 16,
    xlab = "Theoretical Quantiles", ylab = "Sample Quantiles")
abline(0, 1, col = "green")
```



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
sex	0.0657	-0.0564	-0.1206	0.4143	-0.1095
age	0.0070	0.3803	-0.1446	-0.2894	0.5550
t_1	0.5429	0.0317	0.4381	-0.2115	-0.0160
t_2	0.3437	-0.0008	0.2826	-0.1686	-0.0028
t_3	0.3724	-0.1074	0.4089	-0.1633	-0.0189
t_4	0.4600	-0.0722	0.3104	-0.1582	-0.1046
t_5	0.7483	-0.2092	-0.2492	-0.1689	0.1616
t_6	0.7354	-0.3391	-0.1479	0.0700	0.0392
t_7	0.7435	-0.3081	-0.2196	-0.0493	-0.0801
t_8	0.6983	-0.1180	-0.0315	-0.0981	-0.0996
t_9	0.7506	-0.3863	-0.1782	0.0206	0.0625
t_10	0.4907	0.6061	-0.3551	0.0612	-0.1304
t_11	0.5389	0.3451	-0.0412	0.1451	0.0693
t_12	0.4578	0.6001	-0.0294	-0.2123	-0.0467
t_13	0.5784	0.3220	0.1060	-0.2265	-0.0709
t_14	0.3957	0.0501	0.0985	0.3024	0.3043
t_15	0.3528	0.1084	0.1603	0.2167	0.4215
t_16	0.4501	0.0702	0.4206	0.1912	0.1233
t_17	0.4552	0.1746	0.0985	0.4406	0.1178
t_18	0.4679	0.3049	0.2414	0.1645	0.0928
t_19	0.4161	0.0656	0.1771	0.2654	-0.0656
t_20	0.6040	-0.1014	0.2247	-0.0157	0.0116
t_21	0.5591	0.2520	0.1783	-0.0526	-0.1354
t_22	0.6027	-0.0964	0.2131	0.0243	0.0023
t_23	0.6669	-0.0268	0.2465	-0.0453	-0.1442
t_24	0.6550	0.2163	-0.0795	0.1679	-0.1832

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

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
sex	0.0676	-0.0558	-0.1812	0.3802	-0.3144	0.3257
age	0.0144	0.3574	-0.1107	-0.1982	0.5429	0.1707
t_1	0.5518	0.0388	0.4574	-0.1513	-0.0359	0.1191
t_2	0.3460	-0.0040	0.2888	-0.1002	0.0277	-0.0408
t_3	0.3754	-0.1030	0.4260	-0.1208	-0.0361	0.0475
t_4	0.4612	-0.0729	0.3207	-0.1236	-0.0880	-0.0451
t_5	0.7396	-0.2330	-0.2276	-0.1670	0.2039	-0.0074
t_6	0.7317	-0.3648	-0.1591	0.0444	-0.0298	0.1606
t_7	0.7365	-0.3264	-0.2100	-0.1002	-0.0842	0.0082
t_8	0.6949	-0.1299	-0.0223	-0.1155	-0.0676	-0.0636
t_9	0.7376	-0.4096	-0.1738	0.0026	0.0578	-0.0013
t_10	0.5014	0.5852	-0.3856	0.0401	-0.0669	-0.1403
t_11	0.5568	0.3508	-0.0680	0.1413	-0.0147	0.2864
t_12	0.4729	0.5871	-0.0182	-0.1793	0.0356	-0.0956
t_13	0.6117	0.3663	0.1453	-0.2846	-0.1413	0.2941
t_14	0.3938	0.0260	0.0639	0.3334	0.2486	0.0956
t_15	0.3498	0.0746	0.1290	0.3055	0.3929	0.0723
t_16	0.4517	0.0559	0.3850	0.2630	0.0750	0.0619
t_17	0.4510	0.1413	0.0364	0.4653	0.0618	0.0172
t_18	0.4700	0.2823	0.2050	0.2343	0.0845	-0.0167
t_19	0.4161	0.0564	0.1382	0.2734	-0.1372	0.0578
t_20	0.6014	-0.1322	0.2282	0.0484	0.1027	-0.2664
t_21	0.5663	0.2437	0.1718	-0.0174	-0.0614	-0.1788
t_22	0.5989	-0.1136	0.2082	0.0717	0.0433	-0.1344
t_23	0.6669	-0.0428	0.2469	-0.0049	-0.0780	-0.2230
t_24	0.6551	0.1880	-0.1167	0.1681	-0.1534	-0.1643

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

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

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 $\mathrm{trace}\,(\boldsymbol{S}) = \mathtt{size}((\boldsymbol{S}) = 26$ (it is indeed a sample correlation matrix).

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

	Factor1	Factor2	Factor3	Factor4	Factor5
ptsv_5	0.2901983	0.069617	0.0531564	0.0402975	0.0305235

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

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
ptsv_6	0.291827	0.0696332	0.0530655	0.0419415	0.0302067	0.023172

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 loadings      7.545    1.81   1.382   1.048   0.794
## Proportion Var   0.290    0.07   0.053   0.040   0.031
## Cumulative Var   0.290    0.36   0.413   0.453   0.484
```

```
faml_6
```

```
##
## Call:
## factanal(x = gws, factors = 6, rotation = "none")
##
## Uniquenesses:
##   sex   age   t_1   t_2   t_3   t_4   t_5   t_6   t_7   t_8   t_9  t_10  t_11
## 0.610 0.497 0.446 0.784 0.649 0.654 0.277 0.278 0.290 0.478 0.255 0.232 0.460
##  t_12  t_13  t_14  t_15  t_16  t_17  t_18  t_19  t_20  t_21  t_22  t_23  t_24
## 0.389 0.283 0.658 0.602 0.566 0.555 0.595 0.708 0.485 0.554 0.560 0.437 0.443
##
## Loadings:
##      Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
## sex                  -0.181   0.380  -0.314   0.326
## age           0.357  -0.111  -0.198   0.543   0.171
## t_1   0.552           0.457  -0.151           0.119
## t_2   0.346           0.289  -0.100
## t_3   0.375  -0.103   0.426  -0.121
## t_4   0.461           0.321  -0.124
## t_5   0.740  -0.233  -0.228  -0.167   0.204
## t_6   0.732  -0.365  -0.159                   0.161
## t_7   0.737  -0.326  -0.210  -0.100
## t_8   0.695  -0.130          -0.115
## t_9   0.738  -0.410  -0.174
## t_10  0.501   0.585  -0.386                  -0.140
## t_11  0.557   0.351           0.141           0.286
## t_12  0.473   0.587          -0.179
## t_13  0.612   0.366   0.145  -0.285  -0.141   0.294
## t_14  0.394                   0.333   0.249
## t_15  0.350           0.129   0.306   0.393
## t_16  0.452           0.385   0.263
## t_17  0.451   0.141           0.465
## t_18  0.470   0.282   0.205   0.234
## t_19  0.416           0.138   0.273  -0.137
## t_20  0.601  -0.132   0.228           0.103  -0.266
## t_21  0.566   0.244   0.172                  -0.179
## t_22  0.599  -0.114   0.208                  -0.134
## t_23  0.667           0.247                  -0.223
## t_24  0.655   0.188  -0.117   0.168  -0.153  -0.164
##
```

```
##                  Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
## SS loadings        7.588   1.810   1.380   1.090   0.785   0.602
## Proportion Var     0.292   0.070   0.053   0.042   0.030   0.023
## Cumulative Var     0.292   0.361   0.415   0.456   0.487   0.510
##
## Test of the hypothesis that 6 factors are sufficient.
## The chi square statistic is 190.46 on 184 degrees of freedom.
## The p-value is 0.357
```


```
##                  Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
## SS loadings        7.588   1.810   1.380   1.090   0.785   0.602
## Proportion Var     0.292   0.070   0.053   0.042   0.030   0.023
## Cumulative Var     0.292   0.361   0.415   0.456   0.487   0.510
```


Both models seem to fit very poorly. In both cases ($m = 5, 6$), they explain about 50% 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 $(\psi_j)_{j=1}^{2+24}$ 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
```

sex	age	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_10	t_11
0.7944	0.4426	0.4674	0.7736	0.6555	0.6508	0.2796	0.3159	0.2952	0.4779	0.2513	0.2451	0.5629

t_12	t_13	t_14	t_15	t_16	t_17	t_18	t_19	t_20	t_21	t_22	t_23	t_24
0.3822	0.4942	0.6472	0.6134	0.5638	0.5445	0.5941	0.7164	0.574	0.571	0.5814	0.4709	0.4561

```
psi_6 = faml_6$uniquenesses
```

sex	age	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_10	t_11
0.61	0.4967	0.4464	0.7844	0.6488	0.654	0.2773	0.2776	0.2897	0.4778	0.2546	0.2316	0.46

t_12	t_13	t_14	t_15	t_16	t_17	t_18	t_19	t_20	t_21	t_22	t_23	t_24
0.3888	0.283	0.6581	0.6025	0.566	0.5546	0.5951	0.7077	0.485	0.5544	0.56	0.4366	0.4431

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. $\sum_{i=m+1}^{\mathtt{size}(\boldsymbol{S})} \lambda_i^2$.

```
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	1.007146	7.001307

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

```
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	0.7251166	5.952033

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

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

and in our case:

$m = 5:\ 1.0071465 \leq 7.0013073;$

$m = 6:\ 0.7251166 \leq 5.9520329.$

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 very inaccurate, but in some sense acceptable, and although the improvement given by the choice of $m = 6$ is not particularly significant, we tend to prefer it as it is closer to the sufficiency criteria.

1.2

1.3

1.4

1.5

Exercise 2

```
# code
```

2.1

2.2

2.3

2.4

2.5 (optional)