ST437/537 – HW #04- Solution

Due date: February 14, 2019

Instructions

Please follow the instructions below when you prepare and submit your assignment.

- **Include a cover-page** with your homework. It should contain

i. Full name,
ii. Course#: ST 437/537 and
iii. HW-#
iv. Submission date

- Assignments should be submitted in class on the date specified ("due date").
- Neatly typed or hand-written solution on standard letter-size papers (stapled on the top-left corner) should be submitted. **All R code/output should be well commented, with relevant outputs highlighted.**
- **Always staple (upper left corner) your homework <u>before coming to class.</u> Ten percent points will be deducted otherwise.**
- When you solve a particular problem, do not only give the final answer. Instead **show all your work** and the steps you used (with proper explanation) to arrive at your answer to get full credit.

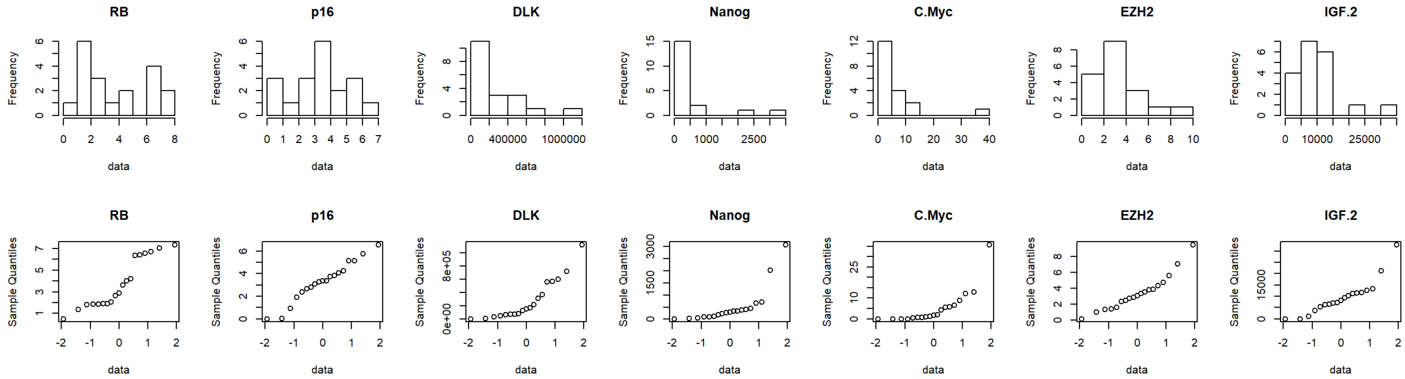
Problems

Soleve the following problems. You may use `R` for these problems unless I specifically instruct otherwise.

1. (15 points) Consider the [hemangioma data] discussed in class.

a. Examine the marginal distributions of genetic markers in the hemangioma data. Which of these appear to be normally distributed? Identify both large and small outliers.

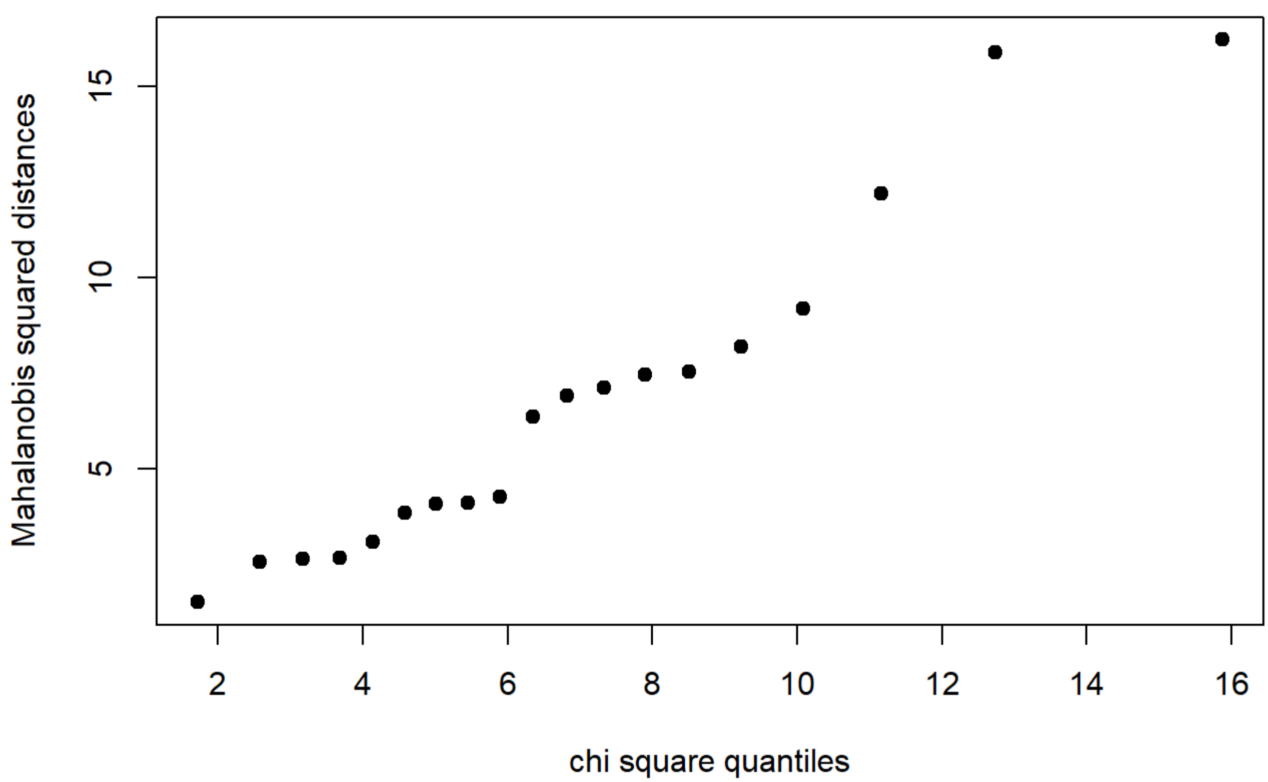
```
data<-read.table("https://www.stat.ncsu.edu/people/maity/courses/st537-S2019/data/hemang
ioma.txt",header = TRUE)
data=data[,2:8]
par(mfrow=c(2,7))
for(i in 1:7){
  hist(data[,i],main=colnames(data)[i],xlab = "data")
}
for(i in 1:7){
  qqnorm(data[,i],main=colnames(data)[i],xlab = "data")
}
```



From the histograms, only p16 and EZH2 seem to be approximately normally distributed.

Next, we will identify outliers using MahaLanobis distance.

```
s=cov(data)
x.cen=scale(data,center = T,scale = F)
d2=diag(x.cen%*%solve(s)%*%t(x.cen))
sortd=sort(d2)
p=ncol(data)
n=nrow(data)
qchi=qchisq((1:n-0.5)/n,df=p)
plot(qchi,sortd,xlab="chi square quantiles", ylab="Mahalanobis squared distances", pch=1
9)
```



```
###Potential outliers
ind = order(d2,decreasing=TRUE)
round( cbind(data[ind,], d2[ind])[1:5,], 2 )
```

```
##      RB  p16        DLK   Nanog C.Myc EZH2    IGF.2 d2[ind]
## 12 6.70 2.67  126015.95 3072.50  0.00 4.35 11762.76   16.23
## 11 1.81 5.15  164881.06 2012.48 35.65 9.45 32721.81   15.89
## 6  2.87 5.76 1119257.50  176.75  8.76 3.51  9342.13   12.22
## 19 4.18 4.24  560208.30  339.93  5.43 1.36 21173.78    9.20
## 14 7.33 0.92   43438.04  697.57  1.77 3.32 11517.89    8.20
```

It seems that individual 11 and 12 are potentials outliers.

b. Perform an EFA after removing the outliers. Do you obtain same or different conclussions?

Before removing the outliers:

```
### Without removing outliers
factanal(data, factors = 3)
```

```
##
## Call:
## factanal(x = data, factors = 3)
##
## Uniquenesses:
##    RB   p16   DLK Nanog C.Myc  EZH2 IGF.2
## 0.050 0.293 0.005 0.609 0.005 0.490 0.249
##
## Loadings:
##       Factor1 Factor2 Factor3
## RB     0.141  -0.144   0.954
## p16    0.366   0.757
## DLK   -0.163   0.961  -0.211
## Nanog  0.559           0.275
## C.Myc  0.841   0.295  -0.448
## EZH2   0.682           0.193
## IGF.2  0.780   0.377
##
##                Factor1 Factor2 Factor3
## SS loadings      2.274   1.757   1.269
## Proportion Var   0.325   0.251   0.181
## Cumulative Var   0.325   0.576   0.757
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 1.86 on 3 degrees of freedom.
## The p-value is 0.603
```

After removing the top two individuals:

```
### Removing outliers
newdata = data[-c(11, 12), ]
factanal(newdata, factors = 3)
```

```
##
## Call:
## factanal(x = newdata, factors = 3)
##
## Uniquenesses:
##    RB   p16   DLK Nanog C.Myc  EZH2 IGF.2
## 0.005 0.005 0.277 0.360 0.392 0.586 0.032
##
## Loadings:
##       Factor1 Factor2 Factor3
## RB    -0.195   0.954   0.219
## p16    0.979   0.168
## DLK    0.766  -0.288   0.231
## Nanog          0.722   0.344
## C.Myc  0.724  -0.229   0.178
## EZH2           0.616  -0.177
## IGF.2  0.434   0.174   0.865
##
##                Factor1 Factor2 Factor3
## SS loadings      2.299   2.004   1.040
## Proportion Var   0.328   0.286   0.149
## Cumulative Var   0.328   0.615   0.763
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 1.69 on 3 degrees of freedom.
## The p-value is 0.638
```

The 3 factor model is still sufficient. However, the contributions of the measured variables to the factors/ the grouping of the variables into the factors are different.

2. (15 points) The correlation matrix given below arises from the scores of 220 boys in six school subjects: (1) French, (2) English, (3) History, (4) Arithmetic, (5) Algebra, and (6) Geometry. We wish to perform an EFA on this data.

```
### Load data
library(sem)
```

```
## Warning: package 'sem' was built under R version 3.5.2
```

```
lt <- readMoments("https://www.stat.ncsu.edu/people/maity/courses/st537-S2019/data/Everi
ttEx5.5.txt", diag = T)
R <- (lt + t(lt)) - diag(1, 6)
colnames(R) <- c("French", "English", "History", "Arithmetic", "Algebra", "Geometry")
rownames(R) <- c("French", "English", "History", "Arithmetic", "Algebra", "Geometry")
R
```

```
##            French English History Arithmetic Algebra Geometry
## French       1.00    0.44    0.41       0.29    0.33     0.25
## English      0.44    1.00    0.35       0.35    0.32     0.33
## History      0.41    0.35    1.00       0.16    0.19     0.18
## Arithmetic   0.29    0.35    0.16       1.00    0.59     0.47
## Algebra      0.33    0.32    0.19       0.59    1.00     0.46
## Geometry     0.25    0.33    0.18       0.47    0.46     1.00
```

a. Find the two-factor solution from a maximum likelihood factor analysis with no rotation applied. Interpret the factors as best as you can.

```
library("psych")
n=220
fa.out.none<-fa(r=R,nfactors=2,n.obs=n,fm="ml",rotate="none")
fa.out.none$loadings
```

```
##
## Loadings:
##            ML1    ML2
## French      0.558  0.425
## English     0.569  0.286
## History     0.392  0.450
## Arithmetic  0.738 -0.279
## Algebra     0.718 -0.209
## Geometry    0.595 -0.133
##
##                  ML1   ML2
## SS loadings    2.204 0.603
## Proportion Var 0.367 0.101
## Cumulative Var 0.367 0.468
```

From the above loading, it appears that factor 1 influences all the variables almost equally (this might be a `general ability` factor) and factor 2 seems to a capture the difference between arts courses and non-art (science) courses.

b. Find the two-factor solution from a maximum likelihood factor analysis with varimax rotation applied. Interpret the factors as best as you can.

```
library("psych")
n=220
fa.out.varimax<-fa(r=R,nfactors=2,n.obs=n,fm="ml",rotate="varimax")
fa.out.varimax$loadings
```

```
##
## Loadings:
##            ML1   ML2
## French     0.233 0.661
## English    0.319 0.551
## History          0.591
## Arithmetic 0.770 0.172
## Algebra    0.715 0.220
## Geometry   0.570 0.215
##
##                  ML1   ML2
## SS loadings    1.593 1.215
## Proportion Var 0.265 0.202
## Cumulative Var 0.265 0.468
```

We can find that factor 1 controls the science subjects and factor 2 controls the arts subjects.

c. Do you think a two factor model is sufficient? Explain your answwer.

```
factanal(covmat=R,factors=2,n.obs=n)
```

```
##
## Call:
## factanal(factors = 2, covmat = R, n.obs = n)
##
## Uniquenesses:
##     French    English    History Arithmetic    Algebra   Geometry
##      0.508      0.595      0.644      0.377      0.440      0.628
##
## Loadings:
##            Factor1 Factor2
## French     0.233   0.661
## English    0.319   0.551
## History            0.591
## Arithmetic 0.770   0.172
## Algebra    0.715   0.220
## Geometry   0.570   0.215
##
##                Factor1 Factor2
## SS loadings      1.593   1.215
## Proportion Var   0.265   0.202
## Cumulative Var   0.265   0.468
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 2.18 on 4 degrees of freedom.
## The p-value is 0.703
```

From the hypothesis testing (the output at the bottom) that p-value is larger than 0.05, and thus we conclude that a two factor model is sufficient

(20 points) The [matrix below] shows the correlations between ratings on nine statements about pain made by 123 people suffering from extreme pain. Each statement was scored on a scale from 1 to 6, ranging from agreement to disagreement. The nine pain statements were as follows:

```
lt <- readMoments("https://www.stat.ncsu.edu/people/maity/courses/st537-S2019/data/Everi
ttEx7.1.txt", diag = T)
R <- (lt + t(lt)) - diag(1, 9)
R2 <- R[-9, -9]
R2
```

```
##       X1    X2    X3    X4    X5    X6    X7    X8
## X1  1.00 -0.04  0.61  0.45  0.03 -0.29 -0.30  0.45
## X2 -0.04  1.00 -0.07 -0.12  0.49  0.43  0.30 -0.31
## X3  0.61 -0.07  1.00  0.59  0.03 -0.13 -0.24  0.59
## X4  0.45 -0.12  0.59  1.00 -0.08 -0.21 -0.19  0.63
## X5  0.03  0.49  0.03 -0.08  1.00  0.47  0.41 -0.14
## X6 -0.29  0.43 -0.13 -0.21  0.47  1.00  0.63 -0.13
## X7 -0.30  0.30 -0.24 -0.19  0.41  0.63  1.00 -0.26
## X8  0.45 -0.31  0.59  0.63 -0.14 -0.13 -0.26  1.00
```

a. Fit a correlated two-factor model in which questions 1, 3, 4, and 8 are assumed to be indicators of the latent variable Doctor`s Responsibility and questions 2, 5, 6, and 7 are assumed to be indicators of the latent variable Patient’s Responsibility.

```
####Loading specific model form
library(sem)
n=123
pain_model<-specifyModel(file="model.txt")
```

```
## NOTE: it is generally simpler to use specifyEquations() or cfa()
##       see ?specifyEquations
```

```
pain_model
```

```
##    Path                  Parameter StartValue
## 1  Doctor -> X1          lambda11
## 2  Doctor -> X3          lambda31
## 3  Doctor -> X4          lambda41
## 4  Doctor -> X8          lambda81
## 5  Patient -> X2         lambda22
## 6  Patient -> X5         lambda52
## 7  Patient -> X6         lambda62
## 8  Patient -> X7         lambda72
## 9  X1        <-> X1      psi1
## 10 X2        <-> X2      psi2
## 11 X3        <-> X3      psi3
## 12 X4        <-> X4      psi4
## 13 X5         <-> X5     psi5
## 14 X6         <-> X6     psi6
## 15 X7         <-> X7     psi7
## 16 X8         <-> X8     psi8
## 17 Patient   <-> Patient <fixed>   1
## 18 Doctor <-> Doctor     <fixed>   1
## 19 Doctor    <-> Patient rho
```

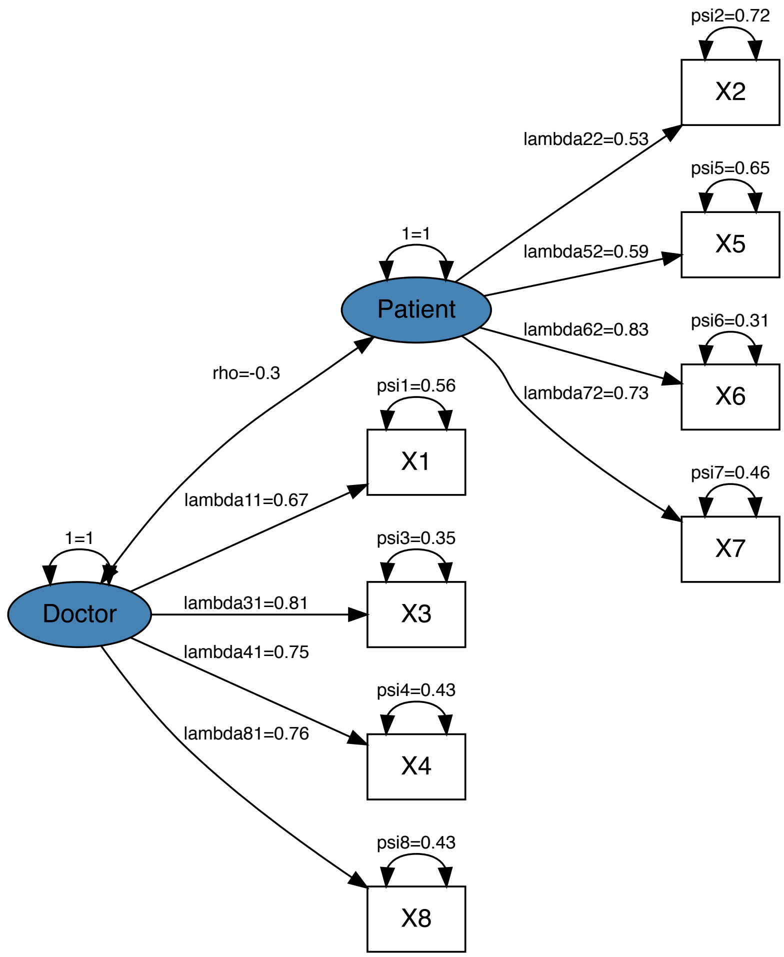
Fitting sem and visualizing.

```
pain_sem<-sem::sem(model=pain_model,S=R2,N=n)
pain_sem
```

```
##
##  Model Chisquare =  63.2304   Df =  19
##
##   lambda11   lambda31   lambda41   lambda81   lambda22   lambda52
##  0.6670173  0.8063408  0.7546241  0.7562965  0.5295387  0.5911277
##   lambda62   lambda72       psi1       psi2       psi3       psi4
##  0.8323020  0.7314687  0.5550877  0.7195887  0.3498144  0.4305423
##       psi5       psi6       psi7       psi8        rho
##  0.6505678  0.3072733  0.4649534  0.4280155 -0.3049759
##
##  Iterations =  17
```

```
pathDiagram(pain_sem,     ## output from `sem()` fit
            ignore.double = FALSE,   ## whether to suppress the variances
            edge.labels = "both",   ## Put both the name and estimated value of edge lab
els
            file = "ability_seb_fitted",  ## Output file name
            output.type = "dot",   ## Output file extension
            node.colors = c("steelblue", "transparent"))  ## Node colors

# Load the DiagrammeR library
library(DiagrammeR)
# Plot the estimated graph
grViz("ability_seb_fitted.dot")
```



b. Find a 95% confidence interval for the correlation between the two latent variables.

```
summary(pain_sem)
```

```
##
##  Model Chisquare =  63.2304   Df =  19 Pr(>Chisq) = 1.180358e-06
##  AIC =  97.2304
##  BIC =  -28.20111
##
##  Normalized Residuals
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.0597 -0.3085 -0.0200  0.0122  0.6249  1.9168
##
##  R-square for Endogenous Variables
##     X1     X3     X4     X8     X2     X5     X6     X7
## 0.4449 0.6502 0.5695 0.5720 0.2804 0.3494 0.6927 0.5350
##
##  Parameter Estimates
##          Estimate   Std Error  z value   Pr(>|z|)
## lambda11  0.6670173 0.08657199  7.704771 1.310781e-14 X1 <--- Doctor
## lambda31  0.8063408 0.08164213  9.876528 5.262474e-23 X3 <--- Doctor
## lambda41  0.7546241 0.08341040  9.047122 1.467856e-19 X4 <--- Doctor
## lambda81  0.7562965 0.08335235  9.073487 1.152677e-19 X8 <--- Doctor
## lambda22  0.5295387 0.09314848  5.684888 1.308985e-08 X2 <--- Patient
## lambda52  0.5911277 0.09149425  6.460818 1.041384e-10 X5 <--- Patient
## lambda62  0.8323020 0.08700206  9.566463 1.106251e-21 X6 <--- Patient
## lambda72  0.7314687 0.08859275  8.256531 1.499678e-16 X7 <--- Patient
## psi1      0.5550877 0.08394571  6.612461 3.779834e-11 X1 <--> X1
## psi2      0.7195887 0.10156760  7.084826 1.392193e-12 X2 <--> X2
## psi3      0.3498144 0.07038309  4.970148 6.690186e-07 X3 <--> X3
## psi4      0.4305423 0.07437535  5.788777 7.090077e-09 X4 <--> X4
## psi5      0.6505678 0.09583263  6.788584 1.132394e-11 X5 <--> X5
## psi6      0.3072733 0.08761380  3.507133 4.529624e-04 X6 <--> X6
## psi7      0.4649534 0.08655697  5.371646 7.802124e-08 X7 <--> X7
## psi8      0.4280155 0.07422086  5.766782 8.079941e-09 X8 <--> X8
## rho      -0.3049759 0.10136386 -3.008725 2.623469e-03 Patient <--> Doctor
##
##  Iterations =  17
```

From the last row in the `Parameter estimate` section, we see that the estimated correlation is -0.30497 with standard error 0.10136. Thus The 95% CI is (-0.30497 $\pm$ 1.96*0.10136) = (-0.5036,-0.1063).

4. (20 points) Consider the partial output from an exploratory factor analysis:

```
##
## Loadings:
##    Factor1 Factor2 Factor3
## X1  0.665  -0.354   0.167
## X2          0.205   0.664
## X3  0.798  -0.127
## X4  0.717          -0.121
## X5          0.318   0.609
## X6          0.831   0.367
## X7 -0.218   0.594   0.314
## X8  0.810          -0.366
##
##                Factor1 Factor2 Factor3
## SS loadings      2.315   1.344   1.229
## Proportion Var   0.289   0.168   0.154
## Cumulative Var   0.289   0.457   0.611
```

```
Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 12.34 on 7 degrees of freedom.
The p-value is 0.0898
```

Assume that each manifest variable has been standardized (so that they each have variance 1). Assume that the model fitted above is a orthogonal factor model. Answer the following questions.

a. How many manifest variables are observed? How many common factors are extracted?

A total of 8 manifest variables are observed and 3 common factors are extracted.

b. Create a table showing the communalities and uniquenesses for each manifest variable.

We can compute Communality = Row sum of squares of loadings, and uniqueness = 1 - communality.

```
##           cm        un
## X1 0.5955475 0.4044525
## X2 0.4920209 0.5079791
## X3 0.6603966 0.3396034
## X4 0.5375153 0.4624847
## X5 0.4749429 0.5250571
## X6 0.8320485 0.1679515
## X7 0.4991360 0.5008640
## X8 0.7968027 0.2031973
```

c. Is the fitted model sufficient for the data? Explain your answer.

Yes, the fitted model is sufficient for the data because the p-value is 0.0898 which is larger than 0.05. Therefore, the hypothesis that 3 factors are sufficient cannot be rejected.

d. What proportion of $var(X_1)$ is captured/explained by `Factor 1`? By `Factor 1` and `Factor 2` together?

Proportion of $var(X_1)$ captured by Factor 1 is $\lambda_{11}^2 = 0.665^2 = 0.442225.$

Proportion of $var(X_1)$ captured by `Factor 1` and `Factor 2` together is
$\lambda_{11}^2 + \lambda_{12}^2 = 0.665^2 + (-0.354)^2 = 0.567541.$

e. Which factor has the highest correlation with X7? Justify your answer.

Factor 2 has the highest correlation with X7 since Factor 2 has the highest loading for X7.