35459 Multivariate Statistics

Week 8 Seminar - Factor Analysis I

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$ be an observable random vector with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)^T$ and covariance matrix $\boldsymbol{\Sigma}$ and $F_1, F_2, \ldots, F_m$ be unobserved random variables called common factors. The factor analysis model is given by

$$\begin{aligned}
X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \ldots + \ell_{1m}F_m + \epsilon_1 \\
X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \ldots + \ell_{2m}F_m + \epsilon_2 \\
&\vdots \\
X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \ldots + \ell_{pm}F_m + \epsilon_p
\end{aligned}$$

which can be written as $\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{LF} + \boldsymbol{\epsilon}$, where

1. $F \sim N_m(\mathbf{0}, \boldsymbol{I}_m)$
2. $\boldsymbol{\epsilon} \sim N_p(\mathbf{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = diag(\psi_1, \psi_2, \ldots, \psi_p)$
3. $\boldsymbol{F}$ and $\boldsymbol{\epsilon}$ are independent.

$\boldsymbol{\epsilon}$ is a vector of unobserved random errors and ϵ_i are called specific factors or unique factors. The factor loading, ℓ_{ij}, is the covariance between the i^{th} response variable (X_i) and the j^{th} common factor (F_j). $\boldsymbol{L}$ is the matrix of factor loadings which is NOT unique.

Properties:

1. $E(\boldsymbol{X}) = \boldsymbol{\mu}$,
2. $Var(\boldsymbol{X}) = \boldsymbol{LL}^T + \boldsymbol{\Psi}$, and
3. $\sigma_{ii} = h_i^2 + \psi_i$, where $h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + \ldots + \ell_{im}^2$ is called the i^{th} communality which is the sum of squares of the factor loadings of the i^{th} variable on the m common factors and measures the proportion of variation of the i^{th} variable explained by the m factors. ψ_i is called the specific variance which measures the proportion of variation of the i^{th} variable NOT explained by the m factors.

Estimating $\boldsymbol{L}$ and $\boldsymbol{\Psi}$

- **Principal components method**: Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the eigenvalues of $Var(\boldsymbol{X})$ in decreasing order, with associated eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$ Then the columns of $\boldsymbol{L}$ are given by $\sqrt{\lambda_i}\boldsymbol{e}_i$ for $i = 1, \ldots, m$, where $m < p$. Then $\boldsymbol{\Psi}$ is approcimated by the diagonal elements of $\boldsymbol{S} - \boldsymbol{LL}^T$.

- **Maximum likelihood method**: Use the likelihood function for estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for a multivriate normal sample and estimate $\boldsymbol{L}$ and $\boldsymbol{\Psi}$ through $\widehat{\boldsymbol{\sigma}}$ subject to the constraint that $\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L}$ be diagonal.

Boston Data

Harrison and Rubenfeld (1978) collected data to determine whether clean air had any influence on house prices. The following variables were collected.

- CRIM: per capita crime rate by town
- ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS: proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX: nitric oxides concentration (parts per 10 million)
- RM: average number of rooms per dwelling
- AGE: proportion of owner-occupied units built prior to 1940
- DIS: weighted distances to five Boston employment centres
- RAD: index of accessibility to radial highways
- TAX: full-value property-tax rate per $\$10,000$
- PTRATIO: pupil-teacher ratio by town
- B: $1000(Bk - 0.63)^2$ where Bk is the proportion of African Americans by town
- LSTAT: % lower status of the population
- MEDV: Median value of owner-occupied homes in $1000's

We consider the variables after Box–Cox transfomations (see Seminar 1). We do not analyse the fourth variable (as it is binary).

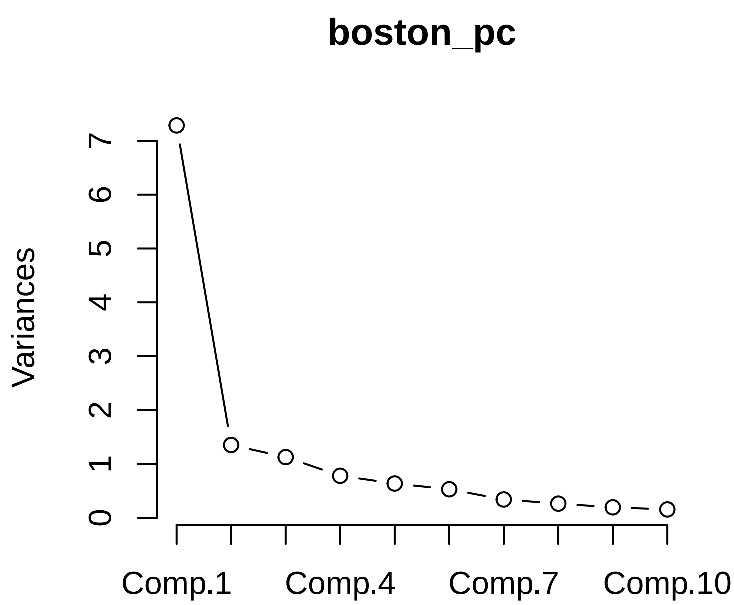
```
library(psych)
```

```
boston <- read.csv("C:/Documents/boston.csv") boston<-
boston[,-4]
boston$CRIM<-log(boston$CRIM)
boston$ZN<-boston$ZN/10
boston$INDUS<-log(boston$INDUS)
boston$NOX<-log(boston$NOX)
boston$RM<-log(boston$RM)
boston$AGE<-(boston$AGE)^2.5/10000
boston$DIS<-log(boston$DIS)
boston$RAD<-log(boston$RAD)
boston$TAX<-log(boston$TAX)
boston$PTRATIO<-exp(0.4*boston$PTRATIO)/1000
boston$B<-boston$B/100
boston$LSTAT<-sqrt(boston$LSTAT)
boston$MEDV<-log(boston$MEDV)
```

```r
round(cor(boston),3)
```

```
##            CRIM     ZN  INDUS    NOX     RM    AGE    DIS    RAD    TAX PTRATIO      B  LSTAT   MEDV
## CRIM      1.000 -0.517  0.740  0.807 -0.324  0.697 -0.744  0.839  0.810   0.454 -0.479  0.622 -0.567
## ZN       -0.517  1.000 -0.656 -0.569  0.309 -0.526  0.591 -0.351 -0.306  -0.350  0.176 -0.452  0.363
## INDUS     0.740 -0.656  1.000  0.750 -0.430  0.658 -0.730  0.581  0.659   0.455 -0.331  0.621 -0.554
## NOX       0.807 -0.569  0.750  1.000 -0.318  0.783 -0.860  0.613  0.668   0.344 -0.379  0.609 -0.515
## RM       -0.324  0.309 -0.430 -0.318  1.000 -0.277  0.281 -0.213 -0.306  -0.321  0.130 -0.639  0.610
## AGE       0.697 -0.526  0.658  0.783 -0.277  1.000 -0.796  0.469  0.541   0.378 -0.286  0.637 -0.482
## DIS      -0.744  0.591 -0.730 -0.860  0.281 -0.796  1.000 -0.542 -0.600  -0.322  0.325 -0.556  0.406
## RAD       0.839 -0.351  0.581  0.613 -0.213  0.469 -0.542  1.000  0.820   0.398 -0.411  0.461 -0.435
## TAX       0.810 -0.306  0.659  0.668 -0.306  0.541 -0.600  0.820  1.000   0.476 -0.428  0.534 -0.557
## PTRATIO  0.454 -0.350  0.455  0.344 -0.321  0.378 -0.322  0.398  0.476   1.000 -0.205  0.434 -0.508
## B        -0.479  0.176 -0.331 -0.379  0.130 -0.286  0.325 -0.411 -0.428  -0.205  1.000 -0.361  0.402
## LSTAT     0.622 -0.452  0.621  0.609 -0.639  0.637 -0.556  0.461  0.534   0.434 -0.361  1.000 -0.825
## MEDV     -0.567  0.363 -0.554 -0.515  0.610 -0.482  0.406 -0.435 -0.557  -0.508  0.402 -0.825  1.000
```

```r
boston_pc<-princomp(boston,cor=TRUE)
screeplot(boston_pc, type = "lines")
```



Maximum likelihood method

```
sapply(1:8, function(nf) factanal(boston, factors = nf, method = "mle")$PVAL)

##  objective  objective  objective  objective  objective  objective  objective  objective
## 2.084e-306 5.322e-182  5.873e-42  8.987e-24  3.954e-13  8.476e-08  8.063e-02  7.662e-01

factanal(boston,factors=6)

##
## Call:
## factanal(x = boston, factors = 6)
##
## Uniquenesses:
##    CRIM      ZN   INDUS     NOX      RM     AGE     DIS     RAD     TAX PTRATIO       B   LSTAT
##   0.045   0.312   0.208   0.126   0.469   0.005   0.132   0.174   0.005   0.595   0.712   0.137
##    MEDV
##   0.159
##
## Loadings:
##         Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
## CRIM     0.779   0.253   0.403   0.334
## ZN      -0.121  -0.228  -0.296  -0.725
## INDUS    0.408   0.352   0.382   0.564   0.100   0.166
## NOX      0.483   0.252   0.631   0.386  -0.134   0.113
## RM              -0.691          -0.200
## AGE      0.245   0.245   0.879   0.221   0.222
## DIS     -0.385  -0.164  -0.682  -0.439   0.118  -0.145
## RAD      0.848   0.138   0.175   0.201   0.131
## TAX      0.811   0.263   0.261           0.233   0.371
## PTRATIO  0.291   0.369   0.104   0.234   0.342
## B       -0.444  -0.248  -0.142
## LSTAT    0.273   0.785   0.371   0.170
## MEDV    -0.318  -0.830  -0.181
##
##                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
## SS loadings      3.068   2.459   2.335   1.529   0.305   0.224
## Proportion Var   0.236   0.189   0.180   0.118   0.023   0.017
## Cumulative Var   0.236   0.425   0.605   0.722   0.746   0.763
##
## Test of the hypothesis that 6 factors are sufficient.
## The chi square statistic is 62.74 on 15 degrees of freedom.
## The p-value is 8.48e-08
```

Principal component method

```r
principal(boston,nfactors=9,rotate="none")

## Principal Components Analysis
## Call: principal(r = boston, nfactors = 9, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##           PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9   h2     u2
## CRIM     0.91  0.22  0.15  0.04  0.09 -0.06  0.05 -0.10  0.01 0.92 0.0777
## ZN      -0.64 -0.03  0.51 -0.09  0.35  0.38 -0.19  0.10  0.04 0.99 0.0143
## INDUS    0.86  0.04 -0.18  0.07 -0.04 -0.18 -0.10  0.37  0.18 0.99 0.0128
## NOX      0.87  0.24 -0.18 -0.12  0.08  0.09 -0.08  0.03 -0.23 0.94 0.0581
## RM      -0.51  0.70  0.09  0.08 -0.23  0.23  0.31  0.15 -0.02 1.00 0.0040
## AGE      0.80  0.16 -0.29 -0.10  0.02  0.38  0.02 -0.12  0.22 0.97 0.0327
## DIS     -0.83 -0.29  0.30  0.11 -0.04 -0.11  0.17  0.02  0.13 0.93 0.0702
## RAD      0.75  0.29  0.38  0.20  0.20 -0.21  0.13 -0.17  0.06 0.96 0.0365
## TAX      0.81  0.16  0.37  0.17  0.21 -0.03 -0.04  0.13 -0.02 0.92 0.0849
## PTRATIO  0.57 -0.27  0.15  0.60 -0.37  0.23 -0.14 -0.05 -0.02 1.00 0.0034
## B       -0.49 -0.10 -0.52  0.50  0.46  0.06  0.13  0.03 -0.04 1.00 0.0006
## LSTAT    0.80 -0.43 -0.03 -0.19  0.08  0.13  0.21 -0.03  0.09 0.93 0.0675
## MEDV    -0.74  0.52 -0.17  0.10  0.01 -0.11 -0.25 -0.14  0.16 0.97 0.0325
##
##                        PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8  PC9
## SS loadings           7.29 1.35 1.13 0.78 0.64 0.53 0.34 0.26 0.19
## Proportion Var        0.56 0.10 0.09 0.06 0.05 0.04 0.03 0.02 0.01
## Cumulative Var        0.56 0.66 0.75 0.81 0.86 0.90 0.93 0.95 0.96
## Proportion Explained  0.58 0.11 0.09 0.06 0.05 0.04 0.03 0.02 0.02
## Cumulative Proportion 0.58 0.69 0.78 0.84 0.89 0.94 0.96 0.98 1.00
##
## Test of the hypothesis that 9 components are sufficient.
##
## The degrees of freedom for the null model are  78  and the objective function was  11.43
## The degrees of freedom for the model are -3  and the objective function was  1.66
## The total number of observations was  506  with MLE Chi Square =  820  with prob <  NA
##
## Fit based upon off diagonal values = 1
```

Drug use

(From Everitt and Hothorn)

There is a large amount of literature on the patterns of drug abuse. A study collected drug usage rates for 1634 high school students in Los Angeles, where each respondent was asked to state the number of times that they had used a particular substance. The substances studied are:

- cigarettes
- beer
- wine
- other alcohol
- cocaine
- tranquilisers
- other pharmaceuticals

In this case, we are provided with the covariance matrix, rather than the data itself.

```
d <-c(
1.000,0.447,0.422,0.435,0.114,0.203,0.091,0.082,0.513,0.304,0.245,0.101,0.245,
0.447,1.000,0.619,0.604,0.068,0.146,0.103,0.063,0.445,0.318,0.203,0.088,0.199,
0.422,0.619,1.000,0.583,0.053,0.139,0.110,0.066,0.365,0.240,0.183,0.074,0.184,
0.435,0.604,0.583,1.000,0.115,0.258,0.122,0.097,0.482,0.368,0.255,0.139,0.293,
0.114,0.068,0.053,0.115,1.000,0.349,0.209,0.321,0.186,0.303,0.272,0.279,0.278,
0.203,0.146,0.139,0.258,0.349,1.000,0.221,0.355,0.315,0.377,0.323,0.367,0.545,
0.091,0.103,0.110,0.122,0.209,0.221,1.000,0.201,0.150,0.163,0.310,0.232,0.232,
0.082,0.063,0.066,0.097,0.321,0.355,0.201,1.000,0.154,0.219,0.288,0.320,0.314,
0.513,0.445,0.365,0.482,0.186,0.315,0.150,0.154,1.000,0.534,0.301,0.204,0.394,
0.304,0.318,0.240,0.368,0.303,0.377,0.163,0.219,0.534,1.000,0.302,0.368,0.467,
0.245,0.203,0.183,0.255,0.272,0.323,0.310,0.288,0.301,0.302,1.000,0.340,0.392,
0.101,0.088,0.074,0.139,0.279,0.367,0.232,0.320,0.204,0.368,0.340,1.000,0.511,
0.245,0.199,0.184,0.293,0.278,0.545,0.232,0.314,0.394,0.467,0.392,0.511,1.000)
druguse <- matrix(d,nrow=13)
colnames(druguse) <- c("cigarettes", "beer", "wine", "liquor", "cocaine",
        "tranquillizers", "drug store medication", "heroin","marijuana",
        "hashish", "inhalants", "hallucinogenics", "amphetamine")
rownames(druguse) <-colnames(druguse)
```

Maximum likelihood method

```
sapply(1:6, function(nf) factanal(covmat = druguse, factors = nf,method = "mle", n.obs = 1634)$PVAL)

## objective objective objective objective objective objective
## 0.000e+00 9.786e-70 7.364e-28 1.795e-11 3.892e-06 9.753e-02

factanal(covmat = druguse, factors = 6, method = "mle",n.obs = 1634)

##
## Call:
## factanal(factors = 6, covmat = druguse, n.obs = 1634, method = "mle")
##
## Uniquenesses:
##           cigarettes                  beer                  wine                liquor
##                0.563                 0.368                 0.374                 0.412
##              cocaine       tranquillizers drug store medication                heroin
##                0.681                 0.522                 0.785                 0.669
##            marijuana               hashish             inhalants       hallucinogenics
##                0.318                 0.005                 0.541                 0.620
##          amphetamine
##                0.005
##
## Loadings:
##                       Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
## cigarettes             0.494                           0.407   0.110
## beer                   0.776                           0.112
## wine                   0.786
## liquor                 0.720   0.121   0.103   0.115   0.160
## cocaine                        0.519           0.132           0.158
## tranquillizers         0.130   0.564   0.321   0.105   0.143
## drug store medication          0.255                           0.372
## heroin                         0.532   0.101                   0.190
## marijuana              0.429   0.158   0.152   0.259   0.609   0.110
## hashish                0.244   0.276   0.186   0.881   0.194   0.100
## inhalants              0.166   0.308   0.150           0.140   0.537
## hallucinogenics                0.387   0.335   0.186           0.288
## amphetamine            0.151   0.336   0.886   0.145   0.137   0.187
##
##                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
## SS loadings      2.301   1.415   1.116   0.964   0.676   0.666
## Proportion Var   0.177   0.109   0.086   0.074   0.052   0.051
## Cumulative Var   0.177   0.286   0.372   0.446   0.498   0.549
##
## Test of the hypothesis that 6 factors are sufficient.
## The chi square statistic is 22.41 on 15 degrees of freedom.
## The p-value is 0.0975
```

Principal component method

```
principal(druguse,nfactors=6,rotate="none")

## Principal Components Analysis
## Call: principal(r = druguse, nfactors = 6, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                       PC1   PC2   PC3   PC4   PC5   PC6   h2   u2
## cigarettes            0.58 -0.40 -0.06  0.01  0.28 -0.38 0.73 0.27
## beer                  0.60 -0.57  0.13  0.09 -0.15  0.12 0.74 0.26
## wine                  0.55 -0.56  0.21  0.13 -0.27  0.13 0.77 0.23
## liquor                0.67 -0.46  0.05  0.06 -0.16  0.14 0.71 0.29
## cocaine               0.44  0.41  0.05  0.53  0.38  0.30 0.88 0.12
## tranquillizers        0.61  0.37 -0.17  0.08 -0.11  0.05 0.56 0.44
## drug store medication 0.37  0.27  0.71 -0.30  0.22  0.22 0.90 0.10
## heroin                0.42  0.45  0.14  0.48 -0.28 -0.32 0.82 0.18
## marijuana             0.71 -0.23 -0.23 -0.10  0.31 -0.12 0.73 0.27
## hashish               0.69  0.07 -0.35 -0.11  0.22  0.20 0.70 0.30
## inhalants             0.58  0.24  0.31 -0.18  0.07 -0.42 0.70 0.30
## hallucinogenics       0.52  0.47 -0.11 -0.26 -0.31  0.12 0.68 0.32
## amphetamine           0.69  0.33 -0.23 -0.24 -0.18  0.02 0.73 0.27
##
##                        PC1  PC2  PC3  PC4  PC5  PC6
## SS loadings           4.38 2.05 0.95 0.82 0.77 0.69
## Proportion Var        0.34 0.16 0.07 0.06 0.06 0.05
## Cumulative Var        0.34 0.49 0.57 0.63 0.69 0.74
## Proportion Explained  0.45 0.21 0.10 0.08 0.08 0.07
## Cumulative Proportion 0.45 0.67 0.76 0.85 0.93 1.00
##
## Test of the hypothesis that 6 components are sufficient.
##
## The degrees of freedom for the null model are  78  and the objective function was  4.05
## The degrees of freedom for the model are 15  and the objective function was  1.84
##
## Fit based upon off diagonal values = 0.95
```

In Class Exercises

Swiss Bank Notes

The data in Notes.csv contain various characteristics of 100 genuine and 100 counterfeit Swiss bank notes. The characteristics include:

- Length of the bank note
- Height of the bank note, measured on the left
- Height of the bank note, measured on the right
- Distance of inner frame to the lower border
- Distance of inner frame to the upper border
- Length of the diagonal

Observations 1-100 are the genuine bank notes and the other 100 observations are the counterfeit bank notes. Determine whether these measures are different between the two types of note and produce confidence intervals for the difference between the notes. Perform a factor analysis on the six continuous variables of this data set. Is it possible to interpret these factors?

Activities

The dataset Activities.csv contains data from 28 individuals, measuring the amount of time (in hours) spent on 10 activities over 100 days, as well as some demographic information. Perform a principal components analysis on the 10 variables related to time spent on activities.

- prof: professional activity
- tran: transportation linked to professional activity
- hous: household occupation
- kids: occupation linked to children
- shop: shopping
- pers: time spent for personal care
- eat: eating
- slee: sleeping
- tele: watching television
- leis: other leisure activities

Perform a factor analysis on this data set. Is it possible to interpret these factors?

Exercises

Johnson and Wichern Exercises 9.10 (Redo the factor analysis in R), 9.18a-b, 9.19a-d, 9.20,