2 Exploratory Factor Analysis (EFA)

Reference:

• Tabachnick & Fidell (2013). Chapter 13.

2.1. Introduction

• Factor analysis is a ***multivariate*** statistical technique for identifying a relatively ***small*** set of common ***underlying*** dimensions (k), known as ***factors***, that explain the relationships (correlation, covariance) among a set of p ($> k$) ***interrelated*** variables.

• Use ***exploratory factor analysis*** (EFA) for

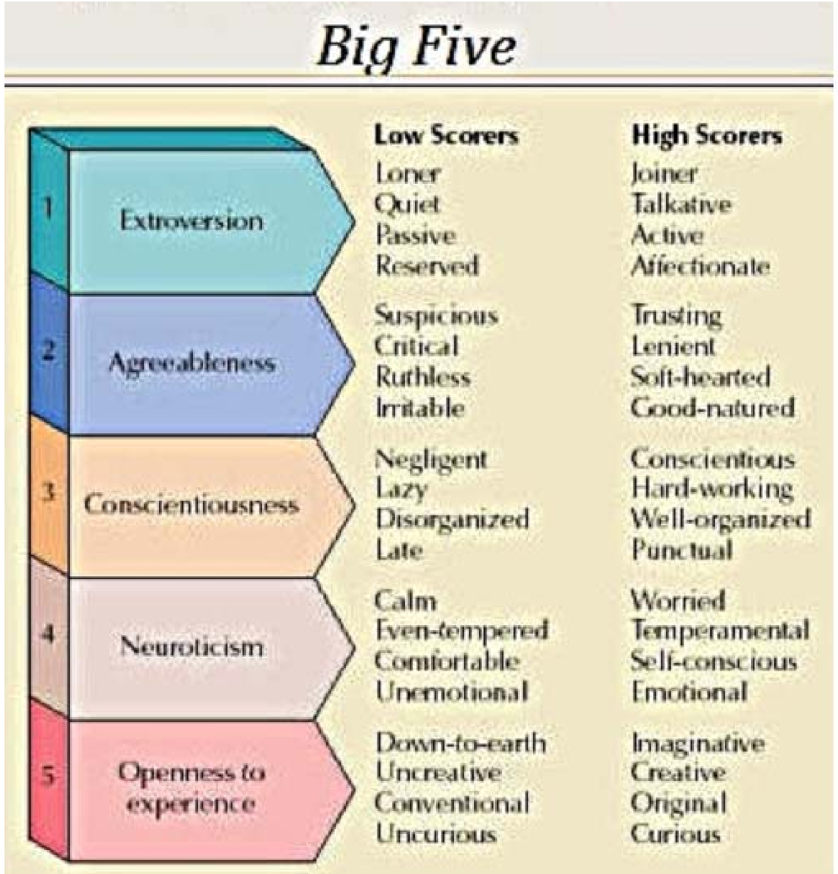
- theory development/construct validation

- data simplification/dimension reduction (principal component analysis)

• Example 1: Personality Descriptors

Loner	Joiner
Quiet	Talkative
Passive	Active
Reserved	Affectionate
Suspicious	Trusting
Critical	Lenient
Ruthless	Soft-hearted
Irritable	Good-natured
Negligent	Conscientious
Lazy	Hard-working
Disorganized	Well-organized
Late	Punctual
Calm	Worried
Even-tempered	Temperamental
Comfortable	Self-conscious
Unemotional	Emotional
Down-to-earth	Imaginative
Uncreative	Creative
Conventional	Original
Uncurious	Curious

• Relevant theory is the Five-Factor Model of Personality (Costa & McCrae, 1992)



• Example 2: Study of Intelligence

Dimension/Scale	Subtests (WAIS-IV)
Verbal Comprehension	Similarities[a]
	Vocabulary[a]
	Information[a]
	Comprehension[b]
Perceptual Reasoning	Block Design[a]
	Matrix Reasoning[a]
	Visual Puzzles[a]
	Picture Completion[b]
	Figure Weights[b]
Working Memory	Digit Span[a]
	Arithmetic[a]
	Letter-Number Sequencing[b]
Processing Speed	Symbol Search[a]
	Coding[a]
	Cancellation[b]

a Core subtest.
b Supplemental subtest.

- Typical questions in EFA:
  - How many factors?
  - How to interpret the factors?
  - Are the factors interrelated?
- Use ***confirmatory factor analysis*** (CFA) for
  - theory/model testing
  - model comparison

2.2. Example 3. Junior Executive Attitude Survey (JEAS)

V1 My job pays me well. (+)
V2 I have my career well planned out. (+)
V3 I would do anything to win my boss' approval. (+)
V4 This is the best job I have ever had. (+)
V5 I find my work tedious. (-)
V6 My job provides me with a sense of achievement. (+)
V7 I perform well in competitive situations. (+)
V8 I think its unfair to promote a person simply because he's more senior. (+)
V9 I am happy with my job. (+)
V10 I hate to be in a responsible position with several people reporting to me. (-)
V11 I am quite content with what I have achieved with my job. (+)
V12 I would leave my job for another offer that pays better. (-)

• 70 junior executives responded to the 12 statements on a 5-point scale (1=strongly disagree to 5=strongly agree)

• filename: *jeas.dat*

• Purpose is to identify and understand the key underlying dimensions about people's work attitudes

```
# import data
mydata <- read.table("jeas.dat")
```

```
> describe(mydata)

    vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
V1     1 70 2.93 1.55    3.0    2.91 2.97   1   5     4  0.09    -1.50 0.19
V2     2 70 3.16 1.46    3.0    3.20 1.48   1   5     4 -0.10    -1.43 0.17
V3     3 70 3.17 1.46    3.0    3.21 1.48   1   5     4 -0.13    -1.44 0.18
V4     4 70 3.29 1.46    3.5    3.36 2.22   1   5     4 -0.25    -1.37 0.17
V5     5 70 2.76 1.53    2.0    2.70 1.48   1   5     4  0.24    -1.52 0.18
V6     6 70 3.26 1.46    3.5    3.32 2.22   1   5     4 -0.17    -1.46 0.17
V7     7 70 3.17 1.44    4.0    3.21 1.48   1   5     4 -0.32    -1.31 0.17
V8     8 70 3.40 1.44    4.0    3.50 1.48   1   5     4 -0.39    -1.19 0.17
V9     9 70 3.34 1.55    3.5    3.43 2.22   1   5     4 -0.18    -1.60 0.19
V10   10 70 2.91 1.45    3.0    2.89 1.48   1   5     4 -0.05    -1.44 0.17
V11   11 70 3.11 1.51    3.0    3.14 1.48   1   5     4 -0.12    -1.47 0.18
V12   12 70 2.70 1.60    2.0    2.62 1.48   1   5     4  0.36    -1.49 0.19
```

```
> mycor <- cor(mydata)
> mycor

             V1          V2          V3         V4          V5          V6          V7          V8          V9
V1   1.00000000  0.15821543 -0.09644228 -0.0548983  0.09641207 -0.02370147  0.02490447  0.13613251 -0.09798478
V2   0.15821543  1.00000000  0.90181699 -0.3007463  0.02385264 -0.24998789  0.84557891  0.85920083 -0.18413966
V3  -0.09644228  0.90181699  1.00000000 -0.1660543 -0.06537628 -0.10893983  0.79451470  0.75129679 -0.09651299
V4  -0.05489830 -0.30074626 -0.16605429  1.0000000 -0.87443420  0.88441935 -0.11321495 -0.17297166  0.91920819
V5   0.09641207  0.02385264 -0.06537628 -0.8744342  1.00000000 -0.86786709 -0.16485114 -0.06069938 -0.87059873
V6  -0.02370147 -0.24998789 -0.10893983  0.8844194 -0.86786709  1.00000000 -0.06238172 -0.17371199  0.88181558
V7   0.02490447  0.84557891  0.79451470 -0.1132150 -0.16485114 -0.06238172  1.00000000  0.92906278 -0.02663450
V8   0.13613251  0.85920083  0.75129679 -0.1729717 -0.06069938 -0.17371199  0.92906278  1.00000000 -0.06239536
V9  -0.09798478 -0.18413966 -0.09651299  0.9192082 -0.87059873  0.88181558 -0.02663450 -0.06239536  1.00000000
V10 -0.06697632 -0.85435640 -0.82456640  0.2859392  0.03623131  0.23590733 -0.82903448 -0.85058201  0.25793778
V11 -0.21897369 -0.62628204 -0.65840748 -0.3185293  0.51546989 -0.44072740 -0.63424183 -0.62227659 -0.45697956
V12  0.10781026  0.20642710  0.09648489 -0.9143198  0.93060463 -0.90226987  0.03511355  0.11583384 -0.89863477
            V10        V11         V12
V1  -0.06697632 -0.2189737  0.10781026
V2  -0.85435640 -0.6262820  0.20642710
V3  -0.82456640 -0.6584075  0.09648489
V4   0.28593919 -0.3185293 -0.91431975
V5   0.03623131  0.5154699  0.93060463
V6   0.23590733 -0.4407274 -0.90226987
V7  -0.82903448 -0.6342418  0.03511355
V8  -0.85058201 -0.6222766  0.11583384
V9   0.25793778 -0.4569796 -0.89863477
V10  1.00000000  0.5138811 -0.16091305
V11  0.51388110  1.0000000  0.41055324
V12 -0.16091305  0.4105532  1.00000000
```

2.3. The Basic Factor Analysis Model

• Model equations:

$$
\begin{aligned}
Y_1 &= \mu_1 + a_{11}F_1 + a_{12}F_2 + \ldots + a_{1k}F_k + e_1 \\
Y_2 &= \mu_2 + a_{21}F_1 + a_{22}F_2 + \ldots + a_{2k}F_k + e_2 \\
&\vdots \\
Y_p &= \mu_p + a_{p1}F_1 + a_{p2}F_2 + \ldots + a_{pk}F_k + e_p
\end{aligned}
$$

where

$Y_1, \ldots, Y_p$	observed/measured/indicator variables
$\mu_1, \ldots, \mu_p$	intercepts
$F_1, \ldots, F_k$	unobserved/latent/common factors
a_{ij}	factor loading (regression coefficient) of variable i on factor j
$e_1, \ldots, e_p$	measurement errors/unique factors

• Each measured variable can be expressed as a linear combination of common factors plus error.

- Using matrix notation,

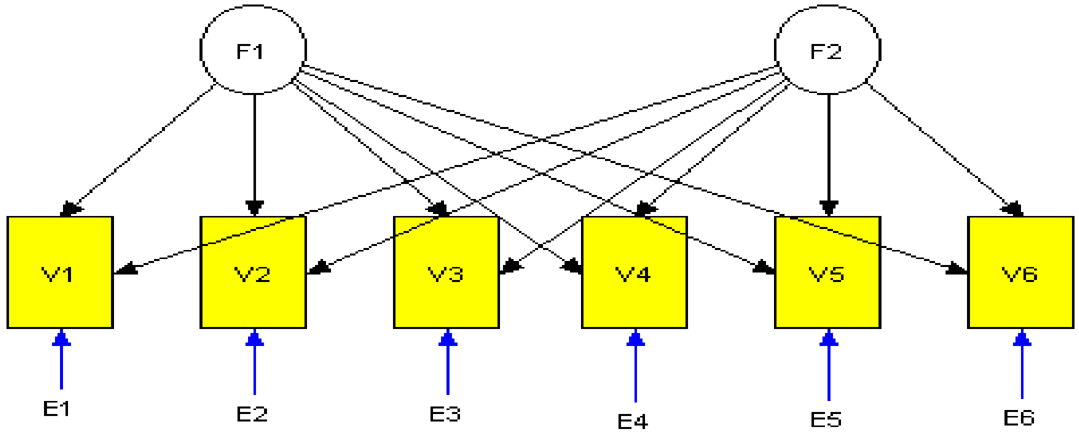
$$\mathrm{y} - \mu = \mathrm{A}\ \mathrm{F} + \mathrm{e} \tag{1}$$

where

$$\mathrm{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}, \qquad \mathrm{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pk} \end{pmatrix},$$

$$\mathrm{F} = \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_k \end{pmatrix}, \qquad \mathrm{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{pmatrix}.$$

• ***Path diagram***:



2.4. Technical Details

• Standard assumptions in factor analysis:

1. Common factors and errors are uncorrelated

$$\operatorname{cov}(F, e) = 0$$

2. Errors are uncorrelated of each other

var (e) $= \Psi$, where Ψ is the $p \times p$ diagonal error variance matrix (uniqueness)

3. Means of F and e are zero

E (F) $= 0$ and E (e) $= 0$

4. Factors are independent of each other (orthogonal)

var (F) $=$ I, where I is the $k \times k$ identity matrix

- Under these assumptions, the $p \times p$ population covariance matrix of y is Σ, where

$$\Sigma = AA' + \Psi, \tag{2}$$

and A is the $p \times k$ factor loading (pattern) matrix.

2.5. Interpreting the Factor Solutions

2.5.1. Factor loadings, $\widehat{a}_{ij}$ (pattern matrix)

```
> fit_pc <- principal(mycor, n.obs=70, nfactors=3, residuals=TRUE, rotate="none")
> fit_pc

Principal Components Analysis
Call: principal(r = mycor, nfactors = 3, residuals = TRUE, rotate = "none", n.obs = 70)
Standardized loadings (pattern matrix) based upon correlation matrix
      PC1   PC2   PC3   h2    u2 com
V1   0.13  0.02  0.99 0.99 0.011 1.0
V2   0.82  0.51  0.04 0.93 0.071 1.7
V3   0.71  0.57 -0.21 0.87 0.126 2.1
V4  -0.74  0.61  0.01 0.92 0.076 1.9
V5   0.52 -0.80  0.06 0.92 0.075 1.7
V6  -0.70  0.65  0.06 0.91 0.087 2.0
V7   0.69  0.64 -0.08 0.88 0.117 2.0
V8   0.74  0.57  0.04 0.88 0.122 1.9
V9  -0.67  0.69 -0.01 0.92 0.079 2.0
V10 -0.79 -0.48  0.07 0.86 0.139 1.7
V11 -0.22 -0.87 -0.21 0.84 0.155 1.3
V12  0.68 -0.69  0.04 0.94 0.061 2.0

                       PC1  PC2  PC3
SS loadings           5.11 4.69 1.08
Proportion Var        0.43 0.39 0.09
Cumulative Var        0.43 0.82 0.91
Proportion Explained  0.47 0.43 0.10
Cumulative Proportion 0.47 0.90 1.00
```

• $\widehat{a}_{ij}$ indicates the effect of F_j on Y_i, with the influence of other factors partialled out (regression coefficient)

• If variables are standardized, which is usually the case in EFA, $\widehat{a}_{ij}$ can be intrepreted as the estimated correlation between the variable (Z_i) and the factor (F_j)

• Reproduced equations:

$$
\begin{aligned}
\widehat{Z}_1 &= .13F_1 + .02F_2 + .99F_3 \\
\widehat{Z}_2 &= .82F_1 + .51F_2 + .04F_3 \\
\vdots & \qquad \vdots \\
\widehat{Z}_{12} &= .68F_1 - .69F_2 + .04F_3
\end{aligned}
$$

2.5.2. Communality, $\widehat{h}_i^2$

• $\widehat{h}_i^2$ estimates the amount (proportion) of variance of variable *i* that is accounted for by the common factors

• It is equal to the sum of squared loadings over the factors on each variable *i*:

$$\hat{h}_i^2 = \sum_{j=1}^{k} \widehat{a}_{ij}^2$$

• $\hat{h}_3^2 = 0.87$, that means 87% variance of Z_3 is accounted for by the 3 factors

2.5.3. Uniqueness, $\widehat{\psi}_{ii}$

- $\widehat{\psi}_{ii}$ measures the amount (proportion) of unexplained variance of variable *i* (variance not accounted for by the common factors)

$$\hat{\psi}_{ii} = r_{ii} - \hat{h}_i^2$$

- $\hat{\psi}_{33} = 1.00 - .874 = .126$, that means 12.6% variance of Z_3 is not explained by the factors

2.5.4. Eigenvalue, $\hat{\lambda}_j$

- $\hat{\lambda}_j$ estimates the amount of variance that is accounted for by factor *j*

- It is equal to the sum of squared loadings of the variables on a particular factor *j*:

$$\hat{\lambda}_j = \sum_{i=1}^{p} \widehat{a}_{ij}^2$$

- Percentage of variance accounted for by factor $j = \frac{\hat{\lambda}_j}{\text{total variance}}$

- With standardized variables, total variance $= p$

```
> # Initial factor solutions by principal component extraction
> fit0 <- principal(mycor, n.obs=70, nfactors=12, rotate="none")
> fit0
Principal Components Analysis
Call: principal(r = mycor, nfactors = 12, rotate = "none", n.obs = 70)
Standardized loadings (pattern matrix) based upon correlation matrix
      PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10  PC11  PC12 h2      u2 com
V1   0.13  0.02  0.99  0.04 -0.07  0.05  0.04 -0.01  0.00  0.00  0.02  0.01  1 5.6e-16 1.1
V2   0.82  0.51  0.04 -0.11 -0.02  0.16  0.00 -0.16  0.06  0.03 -0.04 -0.03  1 4.0e-15 1.9
V3   0.71  0.57 -0.21 -0.29 -0.11  0.11  0.11  0.06 -0.02  0.01  0.03  0.04  1 2.2e-15 2.7
V4  -0.74  0.61  0.01  0.08  0.00  0.18  0.10  0.14 -0.07 -0.02 -0.03 -0.03  1 3.3e-15 2.3
V5   0.52 -0.80  0.06 -0.13  0.15  0.00  0.04  0.11  0.03  0.13  0.02 -0.03  1 2.9e-15 2.0
V6  -0.70  0.65  0.06 -0.08 -0.05 -0.19  0.16  0.01  0.12  0.03 -0.03  0.00  1 2.9e-15 2.4
V7   0.69  0.64 -0.08  0.21  0.14 -0.10  0.16 -0.08 -0.08  0.00  0.04 -0.01  1 1.9e-15 2.6
V8   0.74  0.57  0.04  0.25  0.21  0.01 -0.06  0.09  0.05  0.04 -0.05  0.03  1 1.7e-15 2.4
V9  -0.67  0.69 -0.01 -0.02  0.21  0.11 -0.07 -0.01  0.10 -0.03  0.07  0.00  1 1.3e-15 2.3
V10 -0.79 -0.48  0.07 -0.17  0.29  0.05  0.08 -0.10 -0.06  0.02 -0.03  0.03  1 1.9e-15 2.2
V11 -0.22 -0.87 -0.21  0.31 -0.11  0.15  0.13 -0.05  0.08  0.03  0.01  0.02  1 1.1e-16 1.7
V12  0.68 -0.69  0.04 -0.07  0.12  0.00  0.09  0.05  0.07 -0.16 -0.01 -0.01  1 2.2e-15 2.3

                       PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8  PC9 PC10 PC11 PC12
SS loadings           5.11 4.69 1.08 0.36 0.27 0.15 0.12 0.09 0.06 0.05 0.02 0.01
Proportion Var        0.43 0.39 0.09 0.03 0.02 0.01 0.01 0.01 0.00 0.00 0.00 0.00
Cumulative Var        0.43 0.82 0.91 0.94 0.96 0.97 0.98 0.99 0.99 1.00 1.00 1.00
Proportion Explained  0.43 0.39 0.09 0.03 0.02 0.01 0.01 0.01 0.00 0.00 0.00 0.00
Cumulative Proportion 0.43 0.82 0.91 0.94 0.96 0.97 0.98 0.99 0.99 1.00 1.00 1.00
```

2.5.5. Reproduced (predicted) correlations, $\widehat{r}_{ij}$

• $\widehat{r}_{ij}$ is the predicted correlation between Y_i and Y_i based on the EFA factor solutions

$$\hat{r}_{12} = (.13)(.82)+(.02)(.51)+(.99)(.04) = .16$$
$$\hat{r}_{13} = (.13)(.71)+(.02)(.57)+(.99)(-.21) = -.10$$
$$\vdots$$
$$\hat{r}_{11,12} = (-.22)(.68)+(-.87)(-.69)+(-.21)(.04) = .44$$

• General formula: $\hat{r}_{ij} = \sum_{s=1}^{k} \widehat{a}_{is}\widehat{a}_{js}$

2.5.6. Residuals, $\widehat{e}_{ij}$

• $\widehat{e}_{ij}$ is the difference between the observed and reproduced correlations:

$$\hat{e}_{ij} = r_{ij} - \hat{r}_{ij}$$

• Useful for assessing the ***goodness of fit*** of the factor model.

- A good fit of the model is indicated when all the residuals are very small, say, < .05

$$\hat{e}_{12} = r_{12} - \hat{r}_{12} = .158 - .155 = .003$$
$$\hat{e}_{13} = r_{13} - \hat{r}_{13} = \text{-}.096 - (\text{-}.102) = .006$$
$$\vdots$$
$$\hat{e}_{11,12} = r_{11,12} - \hat{r}_{11,12} = .411 - .437 = \text{-}.026$$

```
> error <- as.data.frame(fit_pc["residual"])
> round(error,3)
    residual.V1 residual.V2 residual.V3 residual.V4 residual.V5 residual.V6 residual.V7 residual.V8 residual.V9
V1        0.011       0.003       0.006       0.014      -0.015      -0.003       0.002      -0.007      -0.013
V2        0.003       0.071       0.039      -0.004      -0.001      -0.012      -0.036      -0.039       0.020
V3        0.006       0.039       0.126       0.014       0.034       0.025      -0.074      -0.094      -0.012
V4        0.014      -0.004       0.014       0.076       0.005      -0.032       0.009       0.024       0.002
V5       -0.015      -0.001       0.034       0.005       0.075       0.018      -0.009       0.010       0.033
V6       -0.003      -0.012       0.025      -0.032       0.018       0.087       0.007      -0.034      -0.032
V7        0.002      -0.036      -0.074       0.009      -0.009       0.007       0.117       0.059      -0.001
V8       -0.007      -0.039      -0.094       0.024       0.010      -0.034       0.059       0.122       0.043
V9       -0.013       0.020      -0.012       0.002       0.033      -0.032      -0.001       0.043       0.079
V10      -0.022       0.034       0.026      -0.006       0.059      -0.004       0.026       0.008       0.058
V11       0.034       0.003      -0.049       0.051      -0.052      -0.018       0.053       0.047      -0.015
V12      -0.009      -0.004       0.015       0.011       0.018       0.018       0.009       0.005       0.030
    residual.V10 residual.V11 residual.V12
V1        -0.022        0.034       -0.009
V2         0.034        0.003       -0.004
V3         0.026       -0.049        0.015
V4        -0.006        0.051        0.011
V5         0.059       -0.052        0.018
V6        -0.004       -0.018        0.018
V7         0.026        0.053        0.009
V8         0.008        0.047        0.005
V9         0.058       -0.015        0.030
V10        0.139       -0.065        0.042
V11       -0.065        0.155       -0.026
V12        0.042       -0.026        0.061
```

2.6. Four Steps in EFA

1. Data preparation and inspection

2. Factor extraction

3. Factor rotation

4. Factor scores computation

2.7. Step 1: Data Preparation and Inspection

• To prepare the input data and determine whether factor analysis is appropriate

• In EFA, we can use the raw data, the sample correlation matrix (R), or the sample covariance matrix (S) as input

• Interval or ratio scale

• Ordinal data are acceptable if no. of categories is large

• Subject/variable ratio $\simeq 10$

• Examine the intercorrelations: if ALL are small, then no need to run factor analysis

$$\begin{pmatrix} 1.0 & & & & \\ 0.1 & 1.0 & & & \\ 0.07 & 0.11 & 1.0 & & \\ 0.02 & 0.03 & 0.13 & 1.0 & \\ 0.12 & 0.09 & 0.14 & 0.12 & 1.0 \end{pmatrix} \text{Vs} \begin{pmatrix} 1.0 & & & & \\ 0.5 & 1.0 & & & \\ 0.57 & 0.41 & 1.0 & & \\ 0.02 & 0.03 & 0.13 & 1.0 & \\ 0.12 & 0.09 & 0.14 & 0.62 & 1.0 \end{pmatrix}$$

2.7.1. Bartlett's test of sphericity

• To test whether the variables are all independent (a matrix is an identity matrix)

• Assumption: Data are multivariate normal

• H_o: All variables are independent

$$\mathrm{X}^2 = -\,(n\text{-}1\text{-}\tfrac{2p+5}{6})\ln|\mathrm{R}| \sim \chi^2(\tfrac{1}{2}p(p\text{-}1))$$

• Reject $H_o \Rightarrow$

```
> # upload package "psych"
> library(psych)
> # Bartlett's test
> bartlett <- cortest.bartlett(mycor, n=70)
> bartlett
$chisq
[1] 1304.504

$p.value
[1] 2.474003e-229

$df
[1] 66
```

2.7.2. Measure of sampling adequacy (Kaiser, 1970)

- *KMO* is a measure of the homogeneity of variables:

$$KMO = \frac{\sum_{i \neq}\sum_{j} r_{ij}^2}{\sum_{i \neq}\sum_{j} r_{ij}^2 + \sum_{i \neq}\sum_{j} r_{ij(p)}^2} \in [0,1]$$

- Rule of Thumb:

> .90	Very good
.80 - .90	Good
.70 - .80	OK
.60 - .70	Acceptable
< .50	Unacceptable

```
> KMO measure of sampling adequacy
> kmo <- KMO(mycor)
> kmo
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = mycor)
Overall MSA =  0.59
MSA for each item =
  V1   V2   V3   V4   V5   V6   V7   V8   V9  V10  V11  V12
0.05 0.58 0.47 0.56 0.60 0.84 0.69 0.55 0.73 0.57 0.57 0.85
```

2.8. Step 2: Factor Extraction

• To determine the number of factors and factor loadings

2.8.1. How many factors (*k*)?

• $k < p$

Rule #1 : Examine the percentage of variance explained by each factor. Ignore any additional factor if it can only explain a small percentage

a. Kaiser's (1960) criterion

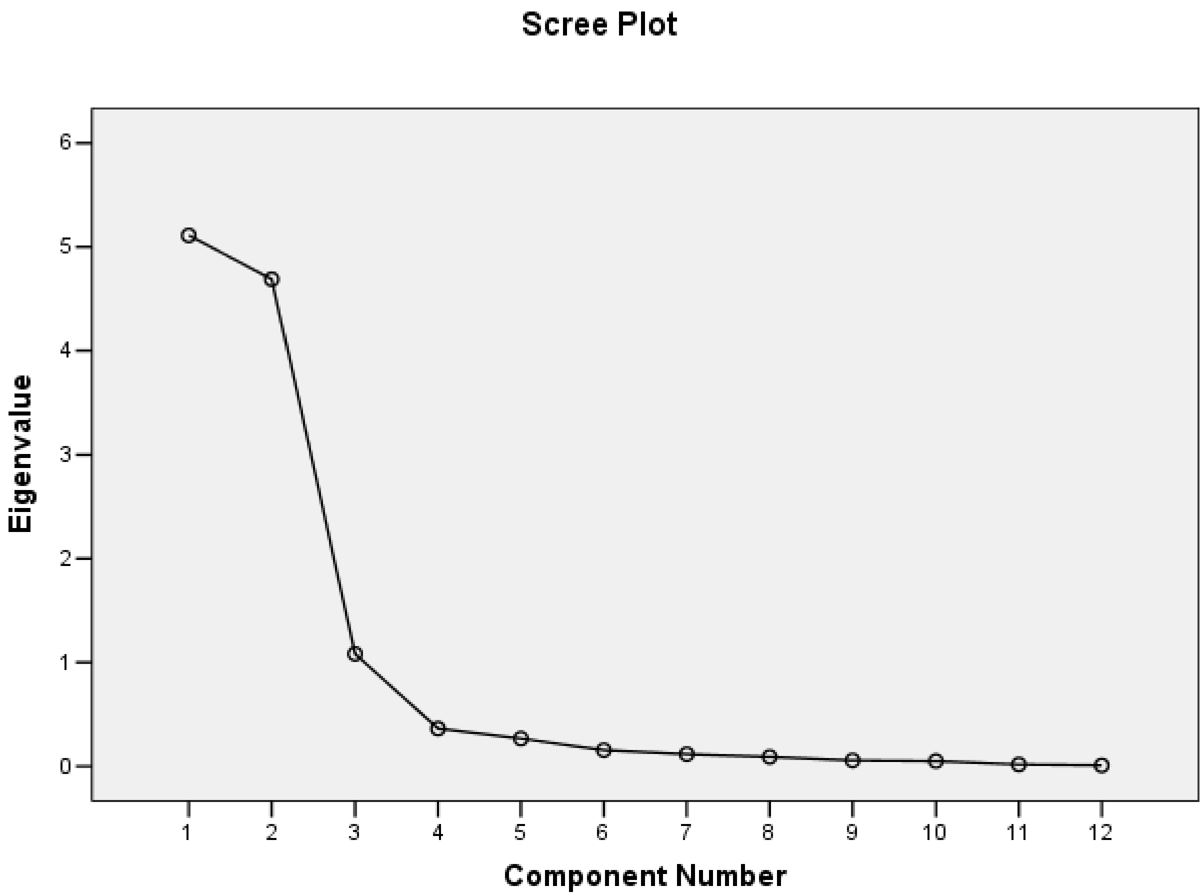
• Set $k =$ no. of factors that have eigenvalues > 1.0

• Over-extraction may occur when we have a large no. of variables (> 40) and low communalities (< 0.4)

• Reliable when $p < 30$ and $h_i^2 > .70$

b. Cattell's (1966) scree test

• A graphical method in which the eigenvalue of each successive factor is plotted. Keep all factors in the steep slope before the first one which starts to level off



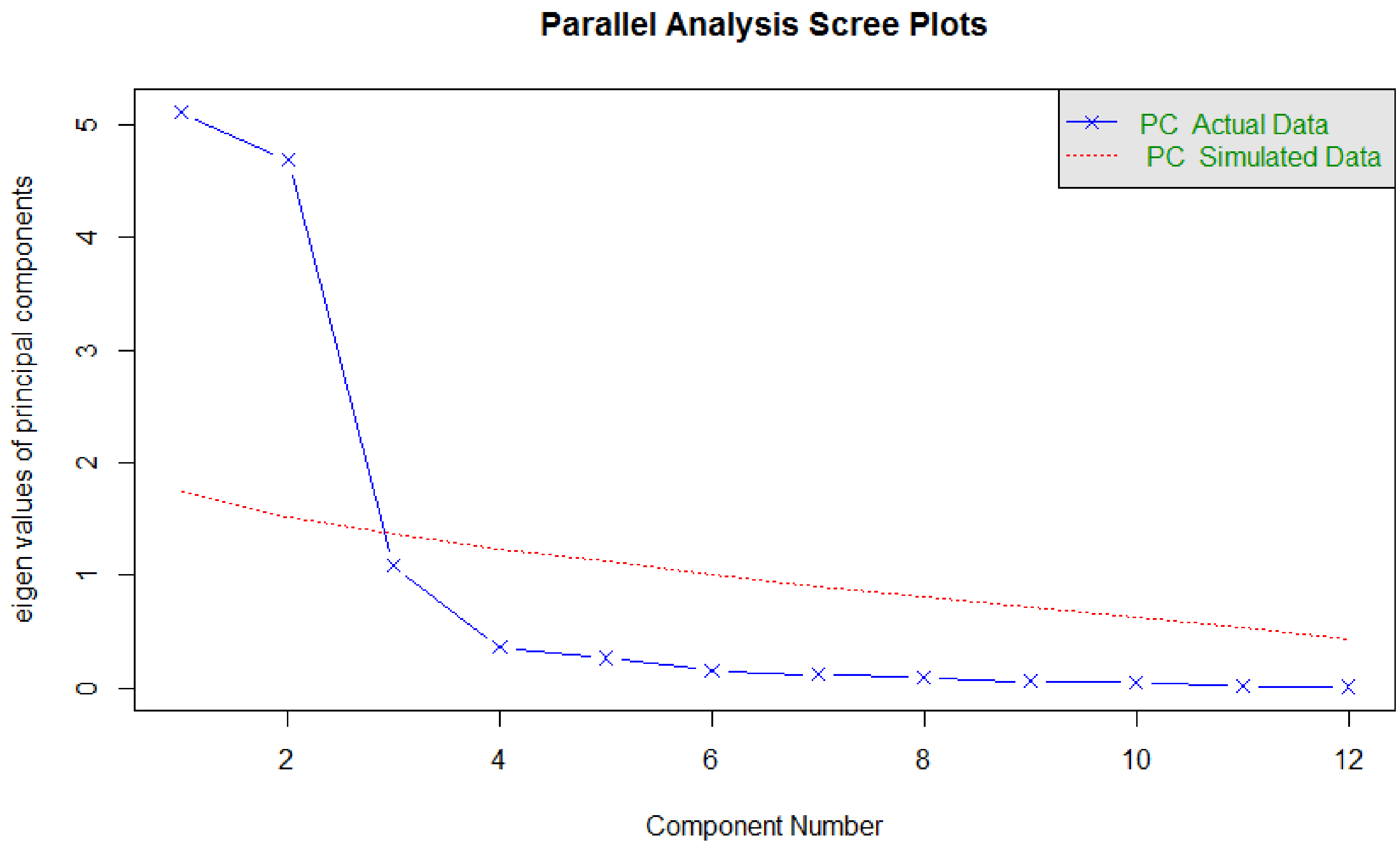
c. Parallel Analysis (Horn, 1965)

1. Generate random datasets (e.g., 100) with the same numbers of observations and variables as the original data

2. Compute the average eigenvalues from these random data

3. Select factors with eigenvalues greater than that from PA

```
> # Scree plot and parallet analysis
> fa.parallel(mycor, n.obs=70, n.iter=100, fa="pc", nfactors=12)
```



Rule #2 : Examine the communalities of the variables. Make sure they are high enough. The presence of low communalities suggests more factors should be extracted.

1-factor solution:

	PC1	h2	u2
V1	0.13	0.016	0.98
V2	0.82	0.671	0.33
V3	0.71	0.506	0.49
V4	-0.74	0.549	0.45
V5	0.52	0.274	0.73
V6	-0.70	0.484	0.52
V7	0.69	0.473	0.53
V8	0.74	0.548	0.45
V9	-0.67	0.451	0.55
V10	-0.79	0.625	0.38
V11	-0.22	0.049	0.95
V12	0.68	0.462	0.54

2-factor solution:

	PC1	PC2	h2	u2
V1	0.13	0.02	0.017	0.983
V2	0.82	0.51	0.927	0.073
V3	0.71	0.57	0.831	0.169
V4	-0.74	0.61	0.924	0.076
V5	0.52	-0.80	0.921	0.079
V6	-0.70	0.65	0.910	0.090
V7	0.69	0.64	0.877	0.123
V8	0.74	0.57	0.876	0.124
V9	-0.67	0.69	0.920	0.080
V10	-0.79	-0.48	0.857	0.143
V11	-0.22	-0.87	0.801	0.199
V12	0.68	-0.69	0.937	0.063

3-factor solution:

	PC1	PC2	PC3	h2	u2
V1	0.13	0.02	0.99	0.99	0.011
V2	0.82	0.51	0.04	0.93	0.071
V3	0.71	0.57	-0.21	0.87	0.126
V4	-0.74	0.61	0.01	0.92	0.076
V5	0.52	-0.80	0.06	0.92	0.075
V6	-0.70	0.65	0.06	0.91	0.087
V7	0.69	0.64	-0.08	0.88	0.117
V8	0.74	0.57	0.04	0.88	0.122
V9	-0.67	0.69	-0.01	0.92	0.079
V10	-0.79	-0.48	0.07	0.86	0.139
V11	-0.22	-0.87	-0.21	0.84	0.155
V12	0.68	-0.69	0.04	0.94	0.061

Rule #3 : The extracted factors should be interpretable (most important)

“each extracted factor contains only a group of numbers (loadings), we must literally ***label*** the factor in order to achieve greatest psychological meaning”

2.8.2. Factor loading estimation

1. ***principal component (PC)*** - based on PCA of the correlation matrix (default)

```
# PC solutions with 3 factors extracted
fit_pc <- principal(mycor, n.obs=70, nfactors=3, residuals=TRUE, rotate="none")
```

2. ***principal axis (PAF)*** - using squared multiple correlations (SMC) as the initial estimates of the communalities and proceed as PC

```
# PAF solutions with 3 factors extracted
fit_paf <- fa(mycor, n.obs=70, nfactors=3, fm="pa", rotate="none")
```

3. ***generalized least squares (GLS)*** - loadings that minimize the squared residuals

4. ***maximum likelihood (ML)*** - loadings that are most likely to produce the observed covariances if data are multivariate normal

- PC is the easiest but it extracts the total variances instead of the common variances. So it tends to overestimate the factor loadings, esp. when correlations are small

- PAF is a modified approach of PC and it overcomes some of the drawbacks of PC

• Comparing PC and PAF soluitions:

A 3-factor solution based on PC (fit_pc):

	PC1	PC2	PC3	h2	u2
V1	0.13	0.02	0.99	0.99	0.011
V2	0.82	0.51	0.04	0.93	0.071
V3	0.71	0.57	-0.21	0.87	0.126
V4	-0.74	0.61	0.01	0.92	0.076
V5	0.52	-0.80	0.06	0.92	0.075
V6	-0.70	0.65	0.06	0.91	0.087
V7	0.69	0.64	-0.08	0.88	0.117
V8	0.74	0.57	0.04	0.88	0.122
V9	-0.67	0.69	-0.01	0.92	0.079
V10	-0.79	-0.48	0.07	0.86	0.139
V11	-0.22	-0.87	-0.21	0.84	0.155
V12	0.68	-0.69	0.04	0.94	0.061

A 3-factor solution based on PAF (fit_paf):

	PA1	PA2	PA3	h2	u2
V1	0.12	0.03	0.90	0.82	0.182
V2	0.79	0.55	0.04	0.93	0.071
V3	0.67	0.60	-0.19	0.84	0.160
V4	-0.77	0.56	0.02	0.91	0.091
V5	0.56	-0.77	0.06	0.91	0.089
V6	-0.72	0.60	0.06	0.89	0.110
V7	0.64	0.66	-0.07	0.86	0.143
V8	0.70	0.60	0.04	0.85	0.149
V9	-0.70	0.64	-0.01	0.90	0.098
V10	-0.75	-0.52	0.06	0.83	0.175
V11	-0.17	-0.85	-0.19	0.79	0.211
V12	0.71	-0.65	0.04	0.93	0.070

• GLS and ML are more complicated procedures, usually assume multivariate normality and give goodness of fit test for the factor model

2.9. Step 3: Factor Rotation

• To transform the initial pattern matrix into ***simple structure*** for easier interpretation

Before rotation: initial solutions

	PC1	PC2	PC3	h2	u2
V1	0.13	0.02	0.99	0.99	0.011
V2	0.82	0.51	0.04	0.93	0.071
V3	0.71	0.57	-0.21	0.87	0.126
V4	-0.74	0.61	0.01	0.92	0.076
V5	0.52	-0.80	0.06	0.92	0.075
V6	-0.70	0.65	0.06	0.91	0.087
V7	0.69	0.64	-0.08	0.88	0.117
V8	0.74	0.57	0.04	0.88	0.122
V9	-0.67	0.69	-0.01	0.92	0.079
V10	-0.79	-0.48	0.07	0.86	0.139
V11	-0.22	-0.87	-0.21	0.84	0.155
V12	0.68	-0.69	0.04	0.94	0.061

	PC1	PC2	PC3
SS loadings	5.11	4.69	1.08
Proportion Var	0.43	0.39	0.09
Cumulative Var	0.43	0.82	0.91
Proportion Explained	0.47	0.43	0.10
Cumulative Proportion	0.47	0.90	1.00

After rotation: Varimax solutions

	RC1	RC2	RC3	h2	u2
V1	0.04	-0.06	0.99	0.99	0.011
V2	0.95	-0.16	0.10	0.93	0.071
V3	0.92	-0.04	-0.15	0.87	0.126
V4	-0.16	0.95	0.00	0.92	0.076
V5	-0.14	-0.95	0.06	0.92	0.075
V6	-0.10	0.95	0.04	0.91	0.087
V7	0.94	0.03	-0.02	0.88	0.117
V8	0.93	-0.05	0.10	0.88	0.122
V9	-0.05	0.96	-0.03	0.92	0.079
V10	-0.91	0.16	0.00	0.86	0.139
V11	-0.72	-0.51	-0.25	0.84	0.155
V12	0.06	-0.97	0.06	0.94	0.061

	RC1	RC2	RC3
SS loadings	4.91	4.87	1.10
Proportion Var	0.41	0.41	0.09
Cumulative Var	0.41	0.82	0.91
Proportion Explained	0.45	0.45	0.10
Cumulative Proportion	0.45	0.90	1.00

Example: Junior Executive Attitude Survey

Factor 1

V2 I have my career well planned out. (.95)
V3 I would do anything to win my boss' approval. (.92)
V7 I perform well in competitive situations. (.94)
V8 I think its unfair to promote a person simply because he's more senior. (.93)
V10 I hate to be in a responsible position with several people reporting to me. (-.91)
V11 I am quite content with what I have achieved with my job. (-.72)

Factor 2

V4 This is the best job I have ever had. (.95)
V5 I find my work tedious. (-.95)
V6 My job provides me with a sense of achievement. (.95)
V9 I am happy with my job. (.96)
~~V11 I am quite content with what I have achieved with my job. (-.51)~~
V12 I would leave my job for another offer that pays better. (-.965)

Factor 3

V1 My job pays me well. (.99)

- Simple structure is achieved when (Thurstone, 1947)
    - each variable is only related to “a few” factors, preferably one
    - each factor is only related to “a few” variables
- Factor indeterminacy due to rotation

2.9.1. Orthogonal rotations

• Factors are uncorrelated after rotation (relative position of factor axes won't change)

• Communality of each variable will not change

• Percentage of variance accounted for by ***each*** factor will change

• But ***total*** percentage over the k factors will not change (rotation redistributes the explained variance among the factors)

• ***Varimax rotation -*** each factor will tend to have high loadings on a small number of variables and low loadings on the others (Kaiser, 1960)

```
> fit_varimax=principal(mydata, nfactors=3, rotate="varimax", scores=TRUE)
```

2.9.2. Oblique rotations

• Factors become correlated after rotation $\Rightarrow$ more realistic?

• The pattern matrix (A) contains factor loadings (regression coefficients)

• The structure matrix (T) gives correlations between the variables and the factors:

$$T = \text{cov}(y, F) = \text{cov}(\mu + AF + e, F) = A\,\text{cov}(F,F) = A\Phi$$

• The component correlation matrix (Φ) tells us the relationships among the factors

• Communality remains unchanged and is equal to the sum of product of pattern and structural loadings

$$\hat{h}_i^2 = \sum_{j=1}^{k} \widehat{a}_{ij}\widehat{t}_{ij}$$

• Percentage of variance accounted for by each factor will change

• Total percentage over *k* factors will not change

• *Oblimin rotation*

```
> fit_oblimin=principal(mydata, nfactors=3, rotate="oblimin", scores=TRUE)
Loading required namespace: GPArotation
> fit_oblimin
Principal Components Analysis

Standardized loadings (pattern matrix) based upon correlation matrix
      TC1   TC2   TC3   h2    u2 com
V1  -0.03 -0.06  1.00 0.99 0.011 1.0
V2   0.94 -0.14  0.08 0.93 0.071 1.1
V3   0.94 -0.02 -0.18 0.87 0.126 1.1
V4  -0.14  0.95  0.00 0.92 0.076 1.0
V5  -0.16 -0.95  0.07 0.92 0.075 1.1
V6  -0.09  0.95  0.04 0.91 0.087 1.0
V7   0.94  0.05 -0.05 0.88 0.117 1.0
V8   0.92 -0.03  0.07 0.88 0.122 1.0
V9  -0.04  0.96 -0.02 0.92 0.079 1.0
V10 -0.92  0.14  0.03 0.86 0.139 1.0
V11 -0.71 -0.52 -0.23 0.84 0.155 2.1
V12  0.04 -0.96  0.05 0.94 0.061 1.0

                       TC1  TC2  TC3
SS loadings           4.91 4.87 1.10
Proportion Var        0.41 0.41 0.09
Cumulative Var        0.41 0.81 0.91
Proportion Explained  0.45 0.45 0.10
Cumulative Proportion 0.45 0.90 1.00

 With component correlations of
      TC1   TC2  TC3
TC1  1.00 -0.04 0.11
TC2 -0.04  1.00 0.00
TC3  0.11  0.00 1.00
```

2.9.3. Summary of effect of rotations

	initial (a)	orthogonal (b)	oblique (c)	effect
communality ($\widehat{h}_i^2$)	$\sum_{j=1}^{k} \widehat{a}_{ij}^2$	$\sum_{j=1}^{k} \widehat{a}_{ij}^2$	$\sum_{j=1}^{k} \widehat{a}_{ij}\widehat{t}_{ij}$	a=b=c
eigenvalue ($\widehat{\lambda}_j$)	$\sum_{i=1}^{p} \widehat{a}_{ij}^2$	$\sum_{i=1}^{p} \widehat{a}_{ij}^2$	$\sum_{i=1}^{p} \widehat{t}_{ij}^2$	a $\neq$ b $\neq$ c
% of variance	λ_j/p	λ_j/p	λ_j/p	a $\neq$ b $\neq$ c
total % of var	$\sum_{j=1}^{k} \lambda_j/p$	$\sum_{j=1}^{k} \lambda_j/p$	??	a=b=c

2.10. Factor Scores Computation

• A composite score indicates the value of each common factor for each individual

2.10.1. Regression method

• If covariance matrix is used:

$$\hat{F}_i = \hat{T}' S^{-1}(y_i - \bar{y})$$

• If correlation matrix is used

$$\hat{F}_i = \hat{T}' R^{-1} z_i = W' z_i$$

- W is the factor (component) score coefficient matrix:

```
> fit_varimax["weights"]
$weights
             RC1          RC2          RC3
V1  -0.0386492057 -0.005020744  0.910526043
V2   0.1888798510 -0.023670883  0.050067134
V3   0.1971556159 -0.001464922 -0.179163289
V4  -0.0241048137  0.193873081  0.009809348
V5  -0.0392761235 -0.196145028  0.055898939
V6  -0.0145185424  0.194840909  0.049381138
V7   0.1949751158  0.012973685 -0.060677576
V8   0.1867722360 -0.002834116  0.047446935
V9  -0.0021272915  0.196397150 -0.013617758
V10 -0.1877613035  0.024810306  0.049205428
V11 -0.1411752640 -0.112516770 -0.203095830
V12  0.0008614333 -0.197792846  0.042612419
```

```
> fit_oblimin["weights"]
$weights
            TC1          TC2          TC3
V1  -0.010563219 -0.005102940  0.905096002
V2   0.190797611 -0.026416849  0.064018729
V3   0.191548350 -0.004171467 -0.163942967
V4  -0.028001079  0.194192431  0.008077287
V5  -0.033268638 -0.195599686  0.052714031
V6  -0.017227935  0.194994724  0.048254858
V7   0.192692561  0.010213957 -0.045943737
V8   0.188158106 -0.005550124  0.061258702
V9  -0.006813096  0.196416936 -0.013642200
V10 -0.186657355  0.027470534  0.035060785
V11 -0.144863689 -0.110334800 -0.213125142
V12  0.006468493 -0.197814617  0.042460719
```

```
> # output factor scores under varimax rotation
> fs_varimax=as.data.frame(fit_varimax["scores"])

> # output factor scores under oblimin rotation
> fs_oblimin=as.data.frame(fit_oblimin["scores"])
```

2.10.2. Factor-based scales

• Select important items for each factor (e.g., $\widehat{a}_{ij} > .40$)

• Scores on factor-based scales are obtained by summing or averaging the scores on those important items (zero or unit weight)

$$\hat{F}_1 = [\mathrm{V2} + \mathrm{V3} + \mathrm{V7} + \mathrm{V8} + (6 - \mathrm{V10}) + (6 - \mathrm{V11})]/6$$
$$\hat{F}_2 = [\mathrm{V4} + (6 - \mathrm{V5}) + \mathrm{V6} + \mathrm{V9} + (6 - \mathrm{V12})]/5$$
$$\hat{F}_3 = \mathrm{V1}$$

```
> # Factor-based scales
> attach(mydata)
> fb1 <- (V2+V3+V7+V8+(6-V10)+(6-V11))/6
> fb2 <- (V4+(6-V5)+V6+V9+(6-V12))/5
> fb3 <- V1
```

```
> # compute summary statistics and correlation matrix of factor scores
> fs <- data.frame(fs_varimax, fs_oblimin, fb1, fb2, fb3)
> mean_fs <- describe(fs)
> mean_fs

           vars  n mean   sd median trimmed  mad   min  max range  skew kurtosis   se
scores.RC1    1 70 0.00 1.00   0.06    0.02 1.50 -1.39 1.23  2.62 -0.20    -1.62 0.12
scores.RC2    2 70 0.00 1.00   0.23    0.04 1.24 -1.54 1.11  2.65 -0.23    -1.69 0.12
scores.RC3    3 70 0.00 1.00  -0.12   -0.02 1.40 -1.42 1.51  2.93  0.20    -1.47 0.12
scores.TC1    4 70 0.00 1.00   0.06    0.02 1.52 -1.38 1.22  2.60 -0.17    -1.63 0.12
scores.TC2    5 70 0.00 1.00   0.23    0.04 1.21 -1.55 1.13  2.68 -0.24    -1.68 0.12
scores.TC3    6 70 0.00 1.00  -0.15   -0.02 1.37 -1.35 1.53  2.88  0.22    -1.47 0.12
fb1           7 70 3.15 1.31   3.33    3.17 1.85  1.33 4.83  3.50 -0.19    -1.60 0.16
fb2           8 70 3.29 1.45   3.70    3.34 1.63  1.20 5.00  3.80 -0.28    -1.65 0.17
fb3           9 70 2.93 1.55   3.00    2.91 2.97  1.00 5.00  4.00  0.09    -1.50 0.19

> cor_fs <- cor(fs)
> round(cor_fs,3)

           scores.RC1 scores.RC2 scores.RC3 scores.TC1 scores.TC2 scores.TC3    fb1    fb2    fb3
scores.RC1      1.000      0.000      0.000      0.999     -0.014      0.075  0.998 -0.046  0.044
scores.RC2      0.000      1.000      0.000     -0.022      1.000      0.000  0.027  0.997 -0.060
scores.RC3      0.000      0.000      1.000      0.031     -0.001      0.997  0.055 -0.023  0.991
scores.TC1      0.999     -0.022      0.031      1.000     -0.036      0.105  0.998 -0.068  0.076
scores.TC2     -0.014      1.000     -0.001     -0.036      1.000     -0.001  0.013  0.998 -0.062
scores.TC3      0.075      0.000      0.997      0.105     -0.001      1.000  0.129 -0.026  0.992
fb1             0.998      0.027      0.055      0.998      0.013      0.129  1.000 -0.021  0.095
fb2            -0.046      0.997     -0.023     -0.068      0.998     -0.026 -0.021  1.000 -0.081
fb3             0.044     -0.060      0.991      0.076     -0.062      0.992  0.095 -0.081  1.000
```