

Multivariate Statistics (I)

3. Factor Analysis (FA)

Contents

3.1 Comprehension of FA

3.2 Concept of common factor

3.3 Factor model

3.4 Estimation of factor model

3.5 Factor rotation and factor loadings plot

3.6 Application of factor scores

3.7 Visualizations of FA

3.8 R for FA : Practice Time

3.1 Comprehension of FA

- **Definition**

FA : technique for describing the covariance relationship among many variables in terms of a few *factors* which are underlying, but unobservable random quantities.

History :

- **K. Pearson and Charles Spearman** provided beginnings of FA in the early 20th century.
- Charles Spearman is known for being the one who coined the term factor analysis and actually used it to measure children's cognitive performance.
- **Spearman(1904)**. General intelligence objectively determined and measured, *American Journal of Psychology*, 15, 201–293.

Charles Spearman

From Wikipedia, the free encyclopedia

Charles Edward Spearman, **FRS** (10 September 1863 - 17 September 1945) was an English psychologist known for work in statistics, as a pioneer of factor analysis, and for Spearman's rank correlation coefficient. He also did seminal work on models for human intelligence, including his theory that disparate cognitive test scores reflect a single general factor and coining the term *g* factor.



3.1 Introduction of FA - Process Steps for FA

- [STEP 1] Prepare a multivariate data matrix X .
- [STEP 2] Obtain a covariance matrix S (or a correlation matrix R).
- [STEP 3] PCFA is performed to obtain the first $m(\leq p)$ common factors of 70% or more of the goodness-of-fit, and the common factors are can be interpreted after the orthogonal transformation.
- [STEP 4] MLFA is performed and the common factors are obtained through the test and the common factors are interpreted after the orthogonal transformation.
- [STEP 5] Compare the tendency of the common factors and the factor scores obtained from [STEP3] and [STEP 4].
- [STEP 6] Repeat [Step 3] - [Step 5] while changing the number of common factors.
- [STEP 7] Consider factor scores as a new multivariate data reducing dimensionally .

3.2 Concept of common factor

◆ Spearman's study of 1904 : $n = 33$ children of private elementary school

Classic French English Mathematics Discrimination of pitch Music

	고전	프랑스어	영어	수학	음감	음악
고 전	1.00					
프랑스어	0.83	1.00				
영 어	0.78	0.67	1.00			
수 학	0.70	0.67	0.64	1.00		
음 감	0.66	0.65	0.54	0.45	1.00	
음 악	0.63	0.57	0.51	0.51	0.40	1.00

Factor loadings

General ability(intelligence) factor

$$\text{Model : } x_j = \lambda_j f + e_j, \quad j = 1, \dots, 6.$$

Specific factor

PCFA

$$\lambda_1 = -0.94, \lambda_2 = -0.89, \lambda_3 = -0.84, \lambda_4 = -0.80, \lambda_5 = -0.74, \lambda_6 = -0.72,$$

3.3 Factor model

Vector of specific factors

❖ Model with m common factors

$$\mathbf{x} = (x_1, \dots, x_p)^t \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} > 0$$

$$x_j - \mu_j = \lambda_{j1}f_1 + \dots + \lambda_{jm}f_m + e_j, \quad j = 1, \dots, p \quad \rightarrow \quad \mathbf{x} - \boldsymbol{\mu} = \boldsymbol{\Lambda} \mathbf{f} + \mathbf{e}$$

Matrix of factor loadings

❖ Assumptions

$\mathbf{f}$ and $\boldsymbol{\epsilon}$: independent $\Leftrightarrow \text{Cov}(\boldsymbol{\epsilon}, \mathbf{f}) = E(\boldsymbol{\epsilon} \mathbf{f}') = \mathbf{0}$

$$E(\mathbf{f}) = \mathbf{0}, \text{Cov}(\mathbf{f}) = E(\mathbf{f} \mathbf{f}') = \mathbf{I} \quad \Leftrightarrow \mathbf{f} \sim (\mathbf{0}, \mathbf{I})$$

$$E(\mathbf{e}) = \mathbf{0}, \text{Cov}(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon} \boldsymbol{\epsilon}') = \boldsymbol{\Psi} = \text{diag}(\psi_1, \dots, \psi_p) \quad \Leftrightarrow \mathbf{e} \sim (\mathbf{0}, \boldsymbol{\Psi})$$

❖ Properties : Covariance Structure

1) $\boldsymbol{\Sigma} = \boldsymbol{\Lambda} \boldsymbol{\Lambda}^t + \boldsymbol{\Psi}$: **Common factors decomposition**

$$- \text{var}(x_j) = \sigma_{jj} = \lambda_{j1}^2 + \dots + \lambda_{jm}^2 + \psi_j = h_j^2 + \psi_j$$

$$- \text{cov}(x_j, x_k) = \sigma_{jk} = \lambda_{j1}\lambda_{k1} + \dots + \lambda_{jm}\lambda_{km}$$

$$- h_j^2 = \lambda_{j1}^2 + \dots + \lambda_{jm}^2 = \sum_{k=1}^m \lambda_{jk}^2 : j\text{th communality}$$

(sum of squared loading of the x_j)

2) $\text{Cov}(\mathbf{x}, \mathbf{f}) = \boldsymbol{\Lambda}$

$$- \text{cov}(x_j, f_k) = \lambda_{jk} : \text{loadings of the } j\text{th variable } x_j \text{ on the } k\text{th factor}$$

3.4 Estimation of factor model - PCFA

[Table 3.4.1] PCFA of S based on the PC method

[step 1] Data matrix : $X = [\mathbf{x}_1, \dots, \mathbf{x}_n]^t$, $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^t$, $i = 1, \dots, n$.

[step 2] Centred data matrix : $Y = HX$, $H = I - n^{-1} \mathbf{1}_n \mathbf{1}_n^t$.

[step 3] Spectral decomposition :

$$Y^t Y / (n - 1) = S = V D V^t = \sum_{k=1}^p l_k \mathbf{v}_k \mathbf{v}_k^t$$

- $V = (\mathbf{v}_1, \dots, \mathbf{v}_p)$: Orthogonal matrix satisfying $V^t V = V V^t = I$
- $D = \text{diag}(l_1, \dots, l_p)$: A matrix of eigenvalues satisfying $l_1 \geq \dots \geq l_p > 0$

[step 4] Proportion of total sample variance due to j th factor : $\frac{l_k}{\sum_{j=1}^p s_{jj}} \times 100$, $k = 1, \dots, m$

[step 5] Estimation of factor loading matrix: $\hat{\Lambda} = \hat{\Lambda}_y = [\sqrt{l_1} \mathbf{v}_1, \dots, \sqrt{l_m} \mathbf{v}_m]$, $m < p$.

[step 6] Estimation of specific variance : $= V_{(m)} D_{(m)}^{1/2}$

$$S = \hat{\Lambda}_y \hat{\Lambda}_y^t + \hat{\Psi}_y \Rightarrow \hat{\Psi} = \hat{\Psi}_y = \text{diag}(\hat{\psi}_{y1}, \dots, \hat{\psi}_{yp}), \hat{\psi}_{yj} = s_{jj} - \sum_{k=1}^m \hat{\lambda}_{yjk}^2$$

: **Sample common factors decomposition**

[step 7] Residual matrix : $R_e = S - (\hat{\Lambda}_y \hat{\Lambda}_y^t + \hat{\Psi}_y)$

3.4 Estimation of factor model : PCFA

- How do we select the number of factors m in PCM?

- ✓ Set m equal to the number of eigenvalues of R greater than 1 or the number of positive eigenvalues of S (Rule of thumb= Kaiser(1960)'s rule)

- ✓ Percentage of variation accounted for by the first m eigenvalues are more than equal to about 70%, i.e.

$$1) \frac{\sum_{k=1}^m l_k}{\sum_{j=1}^p s_{jj}} \times 100 \geq 70\% \text{ for } S$$

$$2) \frac{\sum_{k=1}^m l_k}{p} \times 100 \geq 70\% \text{ for } R$$

- ✓ Residual matrix: $R_e = S - (\hat{\Lambda}_y \hat{\Lambda}_y^t + \hat{\Psi}_y)$ vs. $R_e = R - (\hat{\Lambda}_z \hat{\Lambda}_z^t + \hat{\Psi}_z)$



The diagonal elements are zero and the other elements are small:
 m factors model is appropriate !

3.4 Estimation of factor model : PCFA

[Example 3.4.1] PCFA of KLPGA Data (klpgaa.txt)

- **[STEP 1]** Prepare a multivariate data matrix X form [Data 1.3.2]
- **[STEP 2]** Obtain a covariance matrix S (or a correlation matrix R).

	평균퍼팅수	그린적중률	파세이브율	파브레이크율	평균타수	상금률
$R =$ 평균퍼팅수	1.000	0.128	-0.376	-0.440	0.444	-0.407
그린적중률	0.128	1.000	0.759	0.731	-0.800	0.641
파세이브율	-0.376	0.759	1.000	0.717	-0.937	0.736
파브레이크율	-0.440	0.731	0.717	1.000	-0.897	0.829
평균타수	0.444	-0.800	-0.937	-0.897	1.000	-0.829
상금률	0.407	0.641	0.736	0.829	-0.829	1.000

Putting average
Green in regulation %
Par save %
Par break %
Scoring average
Prize rate

- **[STEP 3]** Spectral decomposition - $R = VDV^t : V = (v_1, \dots, v_p), D = \text{diag}(l_1, \dots, l_p), l_1 \geq \dots \geq l_p > 0$.

Eigenvector : $V = (v_1, v_2, v_3, v_4, v_5, v_6)$ eigenvalue : $(l_1, \dots, l_6) = (4.31, 1.12, 0.33, 0.20, 0.03, 0.01)$

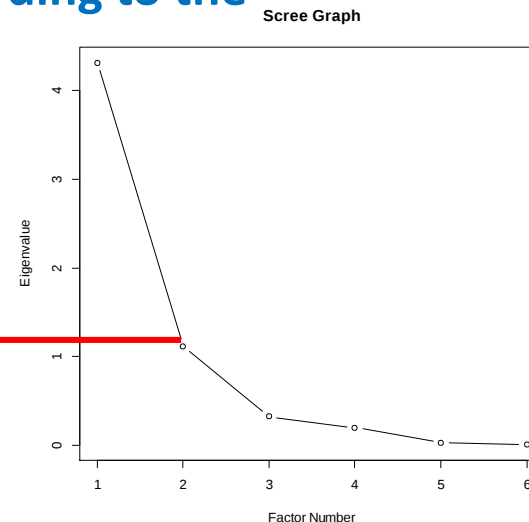
v_1	v_2	v_3	v_4	v_5	v_6
-0.21	0.84	-0.14	-0.16	0.45	0.04
0.39	0.53	0.06	0.23	-0.71	-0.06
0.44	0.04	0.66	-0.27	0.27	-0.47
0.45	-0.05	-0.47	0.56	0.38	-0.35
-0.48	0.00	-0.20	-0.12	-0.25	-0.81
0.43	-0.07	-0.53	-0.72	-0.09	0.01

3.4 Estimation of factor model : PCFA

- **[STEP 4]** Proportion of total sample variance according to the number of factors

$$l_1 = 4.31, l_2 = 1.12 \quad \longrightarrow \quad \frac{4.31 + 1.12}{6} \times 100 = 90.48\%$$

$$m = 2$$



- **[STEP 5]** Estimation of factor loading matrix

$$V_{(m)} D_{(m)}^{1/2} = \hat{\Lambda} = \hat{\Lambda}_z = [\sqrt{l_1} \mathbf{v}_1, \dots, \sqrt{l_m} \mathbf{v}_m], \quad m < p,$$

$$= V_{(m)} D_{(m)}^{1/2}$$

$$V_{(2)} D_{(2)}^{1/2} = \hat{\Lambda} = \hat{\Lambda}_z = [\sqrt{l_1} \mathbf{v}_1, \sqrt{l_2} \mathbf{v}_2] = [\sqrt{4.31} \mathbf{v}_1, \sqrt{1.12} \mathbf{v}_2]$$

- **[STEP 6]** Estimation of specific variance

$$\hat{\Psi}_z = R - \hat{\Lambda}_z \hat{\Lambda}_z^t \Rightarrow \hat{\Psi} = \hat{\Psi}_z = \text{diag}(\hat{\psi}_{z1}, \dots, \hat{\psi}_{zp}), \quad \hat{\psi}_{zj} = 1 - \sum_{k=1}^m \hat{\lambda}_{zjk}^2$$

3.4 Estimation of factor model : PCFA

- **[STEP 7] Residual matrix** $R_e = R - (\hat{\Lambda}_z \hat{\Lambda}_z^t + \hat{\Psi}_z)$

	평균퍼팅수	그린적중률	파세이브율	파브레이크율	평균타수	상금률
평균퍼팅수	0.000	-0.017	-0.015	0.014	0.010	0.048
그린적중률	-0.017	0.000	-0.004	0.004	0.007	-0.040
파세이브율	-0.015	-0.004	0.000	-0.134	-0.027	-0.076
파브레이크율	0.014	0.004	-0.134	0.000	0.034	-0.009
평균타수	0.010	0.007	-0.027	0.034	0.000	0.061
상금률	0.048	-0.040	-0.076	-0.009	0.061	0.000

Results of PCFA

	Factor loadings		Communalities	Specific variances
variables	$\lambda_{jk} = \sqrt{l_k} v_{jk}$		$h_j^2 = \lambda_{j1}^2 + \lambda_{j2}^2$	$\psi_j = 1 - h_j^2$
	f_1	f_2		
Putting average	-0.436	0.888	0.979	0.021
Green in regulation %	0.810	0.561	0.970	0.030
Par save %	0.913	0.042	0.836	0.164
Par break %	0.934	-0.053	0.876	0.124
Scoring average	-0.997	0.000	0.993	0.007
Prize rate	0.893	-0.074	0.802	0.198
eigenvalue	4.31	1.12		
contribution rate	71.83	18.65	total contribution rate = 90.48%	

3.4 Estimation of factor model - MLFA

MLFA based on the Maximum Likelihood Method

$$f(x) = \frac{1}{(\sqrt{2\pi})^p |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu)^t \Sigma^{-1}(x - \mu)\right\}$$

[step 1] Given $x = (x_1, \dots, x_p)^t \sim N_p(\mu, \Sigma)$, $\Sigma > 0$, consider the likelihood function

$$l(\mu, \Sigma) = \log L(\mu, \Sigma) = -\frac{n}{2} \log |2\pi \Sigma| - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^t \Sigma^{-1} (x_i - \mu)$$

$$S_n = \frac{n-1}{n} S$$

$$\Rightarrow \text{Max}_{\mu} l(\mu, \Sigma) \Rightarrow \hat{\mu} = \bar{x}$$

$$= -\frac{n}{2} \log |2\pi \Sigma| - \frac{n}{2} \text{tr}(\Sigma^{-1} S_n) - \frac{n}{2} (\bar{x} - \mu)^t \Sigma^{-1} (\bar{x} - \mu)$$

[step 2] $\Sigma = \Lambda \Lambda^t + \Psi \rightarrow \rho = D_{\hat{\sigma}}^{-1/2} \Sigma D_{\hat{\sigma}}^{-1/2}$

$$l(\hat{\mu}, \Lambda, \Psi) = -\frac{n}{2} \left[\log |2\pi (\Lambda \Lambda^t + \Psi)| + \text{tr}((\Lambda \Lambda^t + \Psi)^{-1} S_n) \right]$$

[step 3] $(\hat{\Psi}^{-1/2} S_n \hat{\Psi}^{-1/2}) \hat{\Psi}^{-1/2} \hat{\Lambda} = \hat{\Psi}^{-1/2} \hat{\Lambda} (I + \hat{\Lambda}^t \hat{\Psi}^{-1} \hat{\Lambda})$

$\text{Max}_{\Lambda, \Psi} l$

$$\hat{\Psi} = \text{diag}(S_n - \hat{\Lambda} \hat{\Lambda}^t)$$

$\hat{\Lambda}^t \hat{\Psi}^{-1} \hat{\Lambda} : \text{diagonal matrix}$

$$\hat{\Lambda} = \hat{\Lambda}_{\hat{\sigma}}$$

$$\hat{\Psi}_{\hat{\sigma}}$$

$$\hat{\Lambda}_z = D_{\hat{\sigma}}^{-1/2} \hat{\Lambda}_{\hat{\sigma}}$$

$$\hat{\Psi}_z = D_{\hat{\sigma}}^{-1} \hat{\Psi}_{\hat{\sigma}}$$

3.4 Estimation of factor model: MLFA

- How do we select the number of factors m in MLM?

✓ Goodness-of fit: $\frac{\sum_{k=1}^m l_k}{\sum_{j=1}^p s_{jj}} \times 100$ for S $\frac{\sum_{k=1}^m l_k}{p} \times 100$ for R

✓ Test the hypotheses $H_0 : \Sigma = \Lambda \Lambda^t + \Psi$, with an appropriate m vs. $H_1 : \Sigma \neq \Lambda \Lambda^t + \Psi$
 where $\mathbf{x} = (x_1, \dots, x_p)^t \sim N_p(\boldsymbol{\mu}, \Sigma)$, $\Sigma > 0$

$$\left(n - \frac{2p + 4m + 11}{6} \right) \ln \left(\frac{|\hat{\Lambda} \hat{\Lambda}^t + \hat{\Psi}|}{|(n-1)S/n|} \right) \simeq \chi^2_{([(p-m)^2 - p - m] / 2)}$$

: Bartlett's test statistic based on the chi-square approximation when $n \wedge n-p$ are large

✓ ← Likelihood Ratio Test : $-2 \log \Lambda \sim \chi^2_{df}$

✓ Residual matrix :

$$\hat{R}_e = S - (\hat{\Lambda}_y \hat{\Lambda}_y^t + \hat{\Psi}_y) \rightarrow \hat{R}_e = R - (\hat{\Lambda}_z \hat{\Lambda}_z^t + \hat{\Psi}_z)$$

The diagonal elements are zero and the other elements are small: m factor model is appropriate !

3.4 Estimation of factor model: MLFA

❖ [Example 3.4.3] MLFA of KLPGA

Comparison of MLFA and PCFA						
	MLFA			PCFA		
variable	$\hat{\lambda}_{zjk}$		$\hat{\psi}_{zj} = 1 - \hat{h}_{zj}^2$	$\lambda_{jk} = \sqrt{l_k} v_{jk}$		$\psi_j = 1 - h_j^2$
	f_1	f_2		f_1	f_2	
Putting average	-0.811	0.581	0.005	-0.436	0.888	0.021
Green in regulation %	0.455	0.855	0.062	0.810	0.561	0.030
Par save %	0.805	0.476	0.124	0.913	0.042	0.164
Par break %	0.814	0.382	0.191	0.934	-0.053	0.124
Scoring average	-0.882	-0.467	0.005	-0.997	0.000	0.007
Prize rate	0.754	0.352	0.307	0.893	-0.074	0.198
Contribution rate	58.70	29.70	total contribution rate = 88.40%	71.83	18.65	total contribution rate = 90.48%

	평균퍼팅수	그린적중률	파세이브율	파브레이크율	평균타수	상금률
$\hat{R}_e =$ 평균퍼팅수	0.000	0.000	0.001	-0.001	0.000	0.000
그린적중률	0.000	0.000	-0.015	0.034	0.000	-0.003
파세이브율	0.001	-0.015	0.000	-0.121	-0.005	-0.039
파브레이크율	-0.001	0.034	-0.121	0.000	-0.001	0.080
평균타수	0.000	0.000	-0.005	-0.001	0.000	0.000
상금률	0.000	-0.003	-0.039	0.080	0.000	0.000

Residual matrix of PCFA

MLFA < PCFA

3.5 Factor rotation and factor loadings plot

• Definition

- An orthogonal transformation of the factor loadings : $T = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$ $T = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$
Clockwise Counterclockwise

• Concepts

- $\hat{\Lambda}$ the $p \times m$ matrix of estimated factor loadings obtained by **PCFA** or **MLFA**

- $\hat{\Lambda}^* = \hat{\Lambda} T$: matrix of rotated loadings where $TT^t = T^t T = I$.

=> The estimated **factor common decomposition** remains **unchanged**:

Conclusion

$$\Sigma \doteq \hat{\Lambda} \hat{\Lambda}^t + \hat{\Psi} = \hat{\Lambda} (TT^t) \hat{\Lambda}^t + \hat{\Psi} = \hat{\Lambda}^* \hat{\Lambda}^{*t} + \hat{\Psi}$$

From a mathematical viewpoint, it is immaterial whether $\hat{\Lambda}$ or $\hat{\Lambda}^* = \hat{\Lambda} T$ is obtained.

Since the **original loadings may not be readily interpretable**, it is usual practice to rotate them until a simple structure is achieved.

Question: How can you choose an orthogonal matrix T? **Varimax**

$$\text{Max}_T \nu = \frac{1}{p} \sum_{k=1}^m \sum_{j=1}^p (d_{jk}^2 - \bar{d}_k)^2 = \frac{1}{p} \sum_{k=1}^m \left[\sum_{j=1}^p d_{jk}^4 - p \bar{d}_k^2 \right]$$

$d_{jk} = \hat{\lambda}_{jk}^* / \hat{h}_j \rightarrow \bar{d}_k = p^{-1} \sum_{j=1}^p d_{jk}^2$
↗ Standardizing
 $\hat{h}_j^2 = \hat{\lambda}_{j1}^2 + \dots + \hat{\lambda}_{jm}^2$

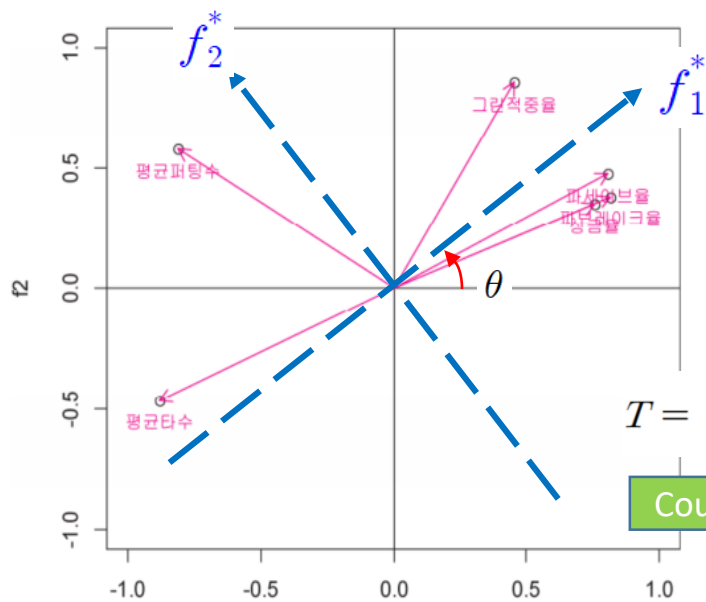
$\hat{\Lambda}^* = \hat{\Lambda} T = (\hat{\lambda}_{jk}^*), \quad j = 1, \dots, p, \quad k = 1, \dots, m$

3.5 Factor rotation and factor loadings plot

[Example 3.5.1] Comparison of before and after the rotation in MLFA

variable	Before rotation		After rotation	
	$\hat{\Lambda}_z$		$\hat{\Lambda}_z^*$	
	f_1	f_2	f_1	f_2
Putting average	-0.811	0.581	-0.168	0.983
Green in regulation %	0.455	0.855	0.925	0.288
Par save %	0.805	0.476	0.908	-0.228
Par break %	0.814	0.382	0.848	-0.301
Scoring average	-0.882	-0.467	-0.955	0.288
Prize rate	0.754	0.352	0.784	-0.280

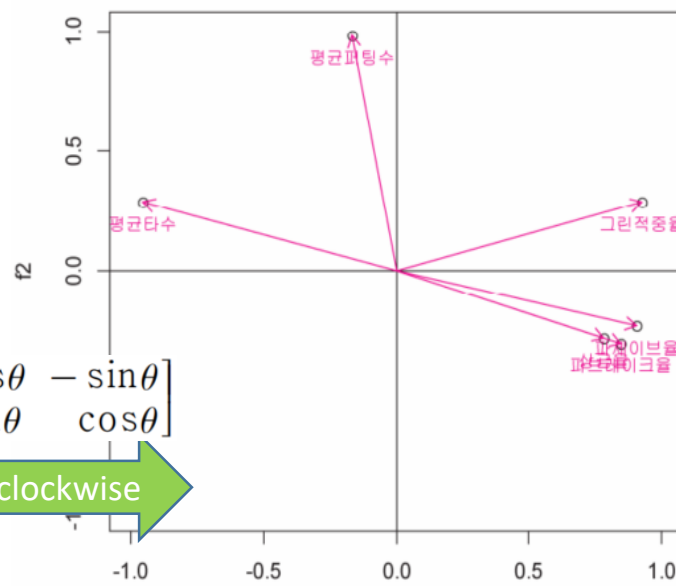
Plot of Factor Loadings : None



$$T = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

Counterclockwise

Plot of Factor Loadings : Varimax



Before

After

3.6 Application of factor scores

Bartlett(1973) :

$$x_i - \mu = \Lambda f_i + e_i, \quad i = 1, \dots, n$$



Min
 f_i

Sum of the weighted squares of the errors

$$e_i^t \Psi^{-1} e_i = (x_i - \mu - \Lambda f_i)^t \Psi^{-1} (x_i - \mu - \Lambda f_i)$$

with $\hat{\Lambda}, \hat{\Psi}, \hat{\mu} = \bar{x}$

Estimations of Factor Score

- MLFA-WLSM: $\hat{f}_i^{ls} = (\hat{\Lambda}^t \hat{\Psi}^{-1} \hat{\Lambda})^{-1} \hat{\Lambda}^t \hat{\Psi}^{-1} (x_i - \bar{x}), \quad i = 1, \dots, n$

- MLFA-REGM: $\hat{f}_i^{re} = \hat{\Lambda}^t S^{-1} (x_i - \bar{x}), \quad i = 1, \dots, n$

$$\hat{f}_i^{re} = \hat{\Lambda}_z^t R^{-1} z_i, \quad i = 1, \dots, n$$

- PCFA-LSM: $\hat{f}_i^{pc} = (\hat{\Lambda}^t \hat{\Lambda})^{-1} \hat{\Lambda}^t (x_i - \bar{x}), \quad i = 1, \dots, n$

$f \sim N_m(0, I)$ Thomson(1951) :

$$x - \mu \sim N_p(0, \Sigma)$$



$$f|x \sim N_m(\Lambda^t \Sigma^{-1} (x - \mu), I_m - \Lambda^t \Sigma^{-1} \Lambda)$$

$$E(f|x) = \Lambda^t \Sigma^{-1} (x - \mu)$$

R: Factor Scores

library(psych)

pcfa=principal()



fpc = pcfa\$scores

method="regression" : default

mlfa=factanal()



fml = mlfa\$scores

score="regression"/"Bartlett"

3.6 Application of factor scores

- ❖ [Example 3.6.1] PCFA and MLFA of air-pollution data [Data 2.8.2] in LA
- ❖ Comparison of varimax rotations of PCFA and MLFA : [R-code 3.6.1]

	PCFA				MLFA			
variable	f_1	f_2	f_3	Specific distribution	f_1	f_2	f_3	Specific distribution
Wind	-0.059	-0.172	<u>0.839</u>	0.263	<u>0.000</u>	-0.210	-0.334	0.840
Solar	0.040	<u>0.736</u>	-0.017	0.456	0.000	0.318	0.000	0.891
CO	<u>0.718</u>	0.278	-0.364	0.275	<u>0.487</u>	0.318	0.507	0.405
NO	<u>0.665</u>	-0.380	-0.456	0.205	0.238	-0.269	<u>0.931</u>	0.005
NO2	<u>0.810</u>	0.156	0.034	0.319	<u>0.989</u>	0.000	0.000	0.005
O3	0.167	<u>0.820</u>	-0.148	0.278	0.000	<u>0.987</u>	0.124	0.005
HC	<u>0.687</u>	0.076	0.495	0.278	<u>0.427</u>	0.103	0.172	0.778
Contribution rate	33.38	19.80	17.20	total contribution rate : 70.38%	30.00	21.00	19.00	total contribution rate : 70%

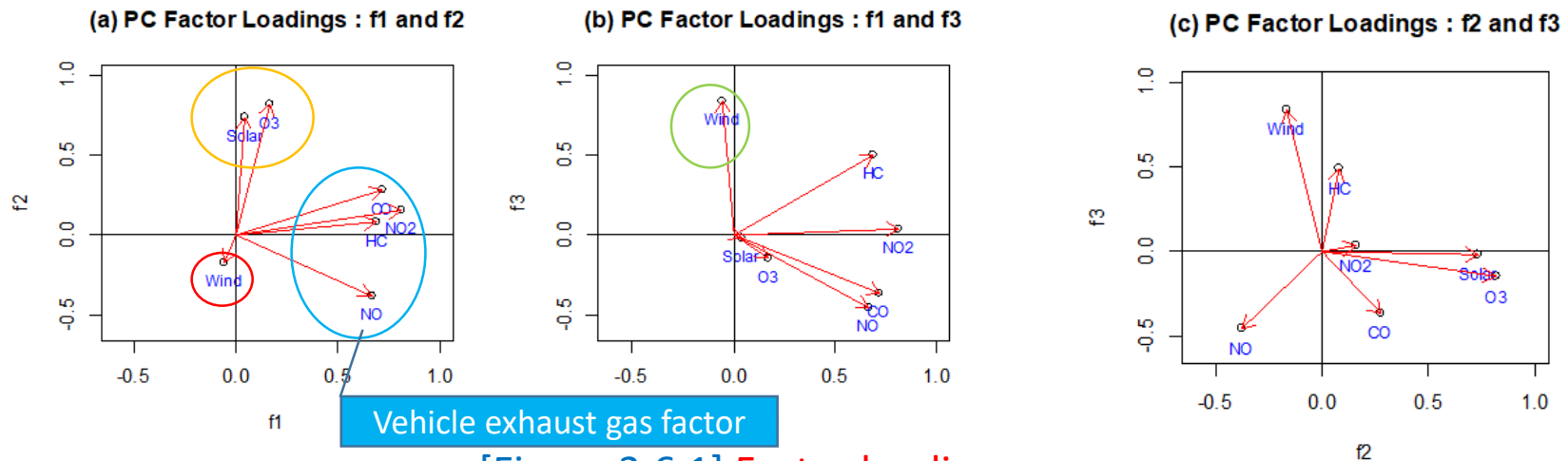
Vehicle exhaust gas factor

Residual Matrices of PCFA and MLFA:

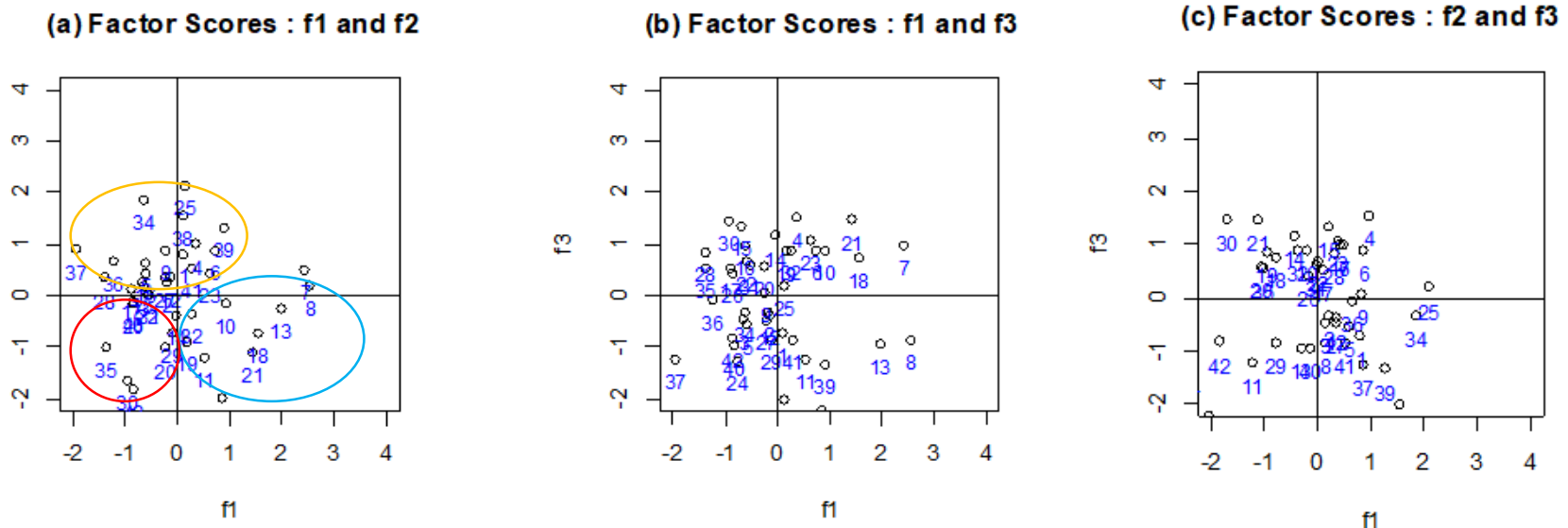
$\hat{R}_e$

	PCFA							MLFA						
	Wind	Solar	CO	NO	NO2	O3	HC	Wind	Solar	CO	NO	NO2	O3	HC
Wind	0.000	0.042	0.202	0.087	-0.064	0.021	-0.205	0.000	-0.032	0.072	-0.001	-0.001	-0.001	0.261
Solar	0.042	0.000	-0.056	0.172	-0.031	-0.294	-0.023	-0.032	0.000	0.043	-0.001	0.000	0.000	-0.018
CO	0.202	-0.056	0.000	-0.036	-0.056	0.010	-0.168	0.072	0.043	0.000	0.000	0.001	0.000	-0.162
NO	0.087	0.172	-0.036	0.000	-0.166	0.000	0.033	-0.001	-0.001	0.000	0.000	0.000	0.000	0.002
NO2	-0.064	-0.031	-0.056	-0.166	0.000	-0.092	-0.137	-0.001	0.000	0.001	0.000	0.000	0.000	0.001
O3	0.021	-0.294	0.010	0.000	-0.092	0.000	0.050	-0.001	0.000	0.000	0.000	0.000	0.000	0.001
HC	-0.205	-0.023	-0.168	0.033	-0.137	0.050	0.000	0.261	-0.018	-0.162	0.002	0.001	0.001	0.000

3.6 Application of factor scores



[Figure 3.6.1] Factor loadings



[Figure 3.6.2] PCFA's factor scores

3.6 Application of factor scores

- 42 days LA air pollution

42 measurements on air-pollution variables recorded at 12:00 noon in the LA area on different days.

<i>X</i> : Law								<i>Z</i> : Standardized							<i>F</i> : Factor scores		
Wind	Solar	CO	NO	NO2	O3	HC		Wind	Solar	CO	NO	NO2	O3	HC	f_1	f_2	f_3
1	8	98	7	2	12	8	2	0.316	1.393	1.988	-0.175	0.579	-0.252	-1.583	0.116	0.800	-0.737
2	7	107	4	3	9	5	3	-0.316	1.912	-0.444	0.744	-0.311	-0.791	-0.138	-0.175	0.230	-0.341
3	7	103	4	3	5	6	3	-0.316	1.681	-0.444	0.744	-1.497	-0.612	-0.138	-0.629	0.167	-0.492
4	10	88	5	2	8	15	4	1.581	0.816	0.367	-0.175	-0.607	1.005	1.308	0.367	0.978	1.521
5	6	91	4	2	8	10	3	-0.949	0.989	-0.444	-0.175	-0.607	0.107	-0.138	-0.587	0.603	-0.563
6	8	90	5	2	12	12	4	0.316	0.931	0.367	-0.175	0.579	0.466	1.308	0.749	0.860	0.867
7	9	84	7	4	12	15	5	0.949	0.585	1.988	1.664	0.579	1.005	2.754	2.429	0.492	0.975
8	5	72	6	4	21	14	4	-1.581	-0.107	1.177	1.664	3.249	0.826	1.308	2.558	0.172	-0.877
9	7	82	5	1	11	11	3	-0.316	0.470	0.367	-1.095	0.283	0.287	-0.138	-0.230	0.845	0.050
10	8	64	5	2	13	9	4	0.316	-0.569	0.367	-0.175	0.876	-0.073	1.308	0.940	-0.181	0.865
.....																	
38	5	86	7	2	13	18	2	-1.581	0.700	1.988	-0.175	0.876	1.544	-1.583	0.137	1.548	-2.038
39	7	79	7	4	9	25	3	-0.316	0.297	1.988	1.664	-0.311	2.802	-0.138	0.909	1.293	-1.362
40	7	79	5	2	8	6	2	-0.316	0.297	0.367	-0.175	-0.607	-0.612	-1.583	-0.817	-0.126	-0.976
41	6	68	6	2	11	14	3	-0.949	-0.338	1.177	-0.175	0.283	0.826	-0.138	0.306	0.524	-0.877
42	8	40	4	3	6	5	2	0.316	-1.953	-0.444	0.744	-1.201	-0.791	-1.583	-0.845	-1.830	-0.849

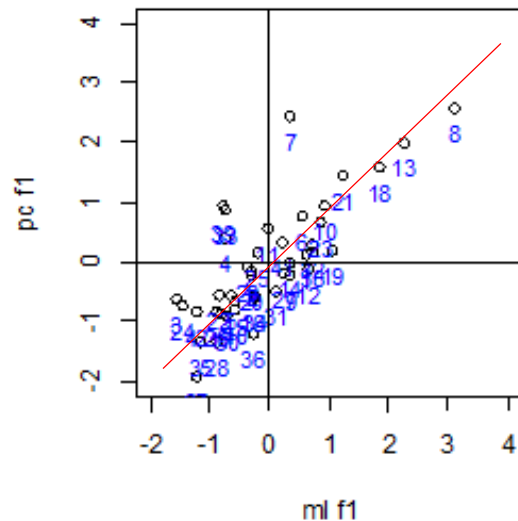
3.6 Application of factor scores

- ✓ Comparison of PCFA and MLFA

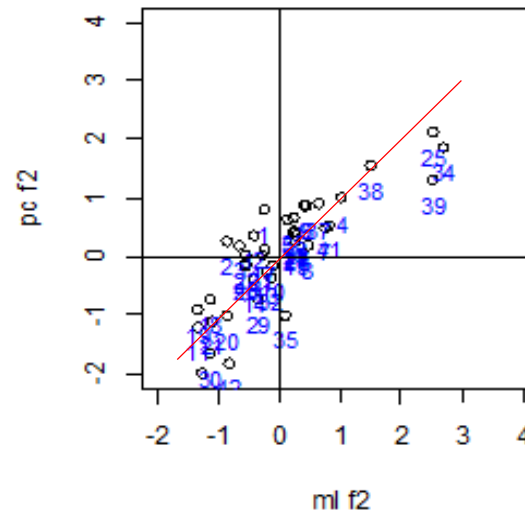
Similar Pattern fort Factor Scores of PCFA and MLFA



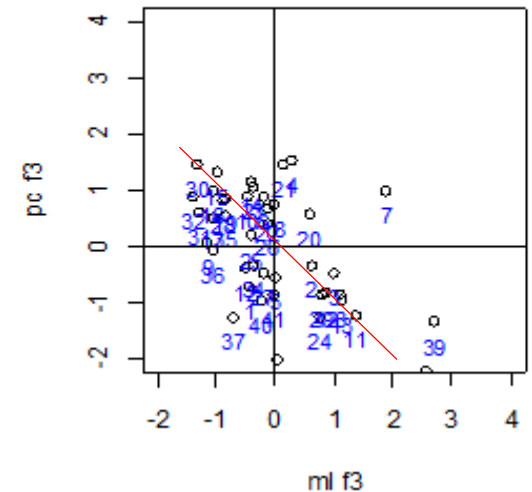
(a) Factor Scores : ml f1 and pc f1



(b) Factor Scores : ml f2 and pc f2



(c) Factor Scores : ml f3 and pc f3



[Figure 3.6.3] Plots of Pairs of Factor Scores PCFA and MLFA after rotation

3.7 Visualizations of FA

❖ [Example 3.7.1] Air pollution- FA Biplot

$$Y = U\Lambda V^t = \sum_{k=1}^p \lambda_k \mathbf{u}_k \mathbf{v}_k^t$$

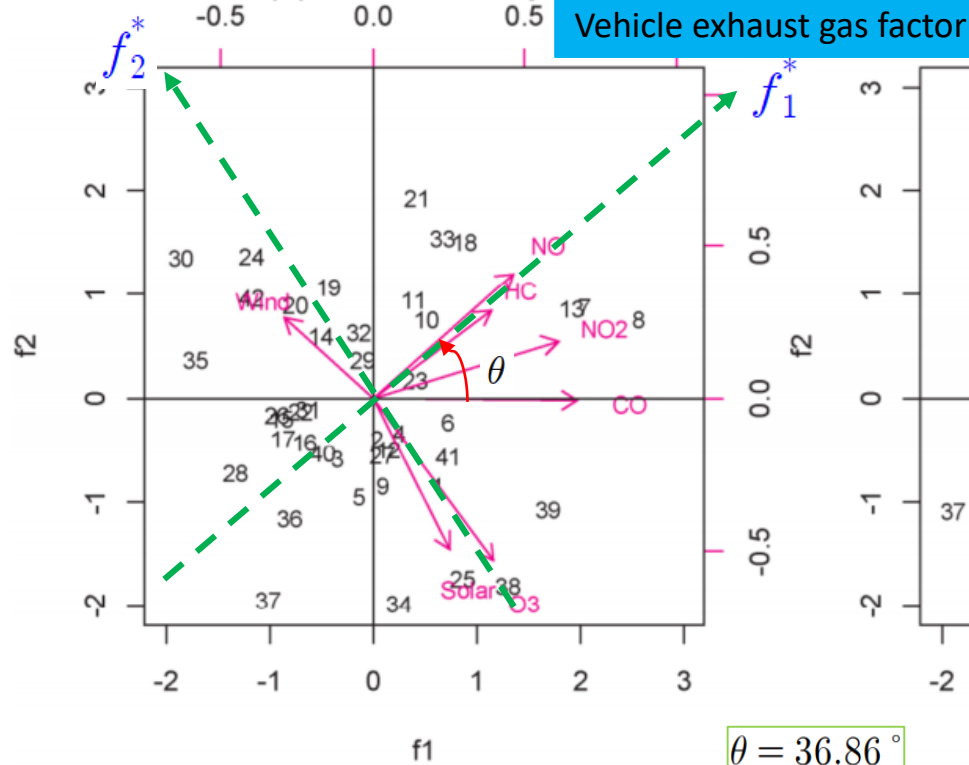
$$\hat{\Lambda}_{(m)} = (n-1)^{-1/2} V_{(m)} \Lambda_{(m)} = (n-1)^{-1/2} [\lambda_1 \mathbf{v}_1, \dots, \lambda_m \mathbf{v}_m]$$

$$F_{(m)} = (n-1)^{1/2} U_{(m)} = (n-1)^{1/2} (\mathbf{u}_1, \dots, \mathbf{u}_m)$$

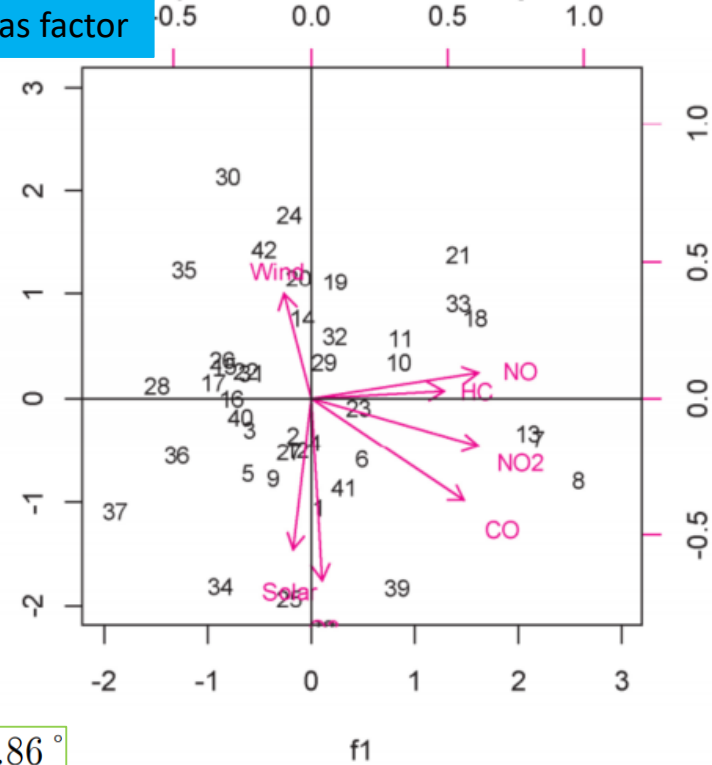
$$\hat{\Lambda}^* = \hat{\Lambda} T$$

$$F^* = F T$$

(a) Unrotated Biplot



(b) Varimax Rotated Biplot



$$T = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} 0.837 & -0.547 \\ 0.547 & 0.837 \end{bmatrix}$$

3.8 R for FA : Practice Time

R-Code: FA(PCFA , MLFA) and FA Biplot

<code>library(psych), principal()</code>	PCFA
<code>library(psych), factanal()</code>	MLFA
<code>principal(, rotation="varimax")</code> <code>factanal(, rotation="varimax")</code>	Varimax Rotation
FA	Klpga-PCFAsteps-scee.R Klpga-MLFAfactanal.R Klpga-MLFAvarimax.R airpollution-PCMLFAvarimax-scores.R
FA Biplot	airpollution-PCFABiplot.R

3.8 R for FA : Practice Time

R-code list of Chapter 3 Factor Analysis

spearman-PCFA.R	[R-코드 3.2.1]	스피어만의 여섯 과목 성적의 PCFA
klpga-PCFAsteps-scrree.R	[R-코드 3.4.1]	KLPGA 선수 성적의 PCFA
5subjects-PCFAsteps.R	[R-코드 3.4.2]	두 가지 시험성적의 PCFA
klpga-MLFAfactanal.R	[R-코드 3.4.3]	KLPGA 선수 성적의 MLFA
5subjects-MLFAfactanal.R	[R-코드 3.4.4]	두 가지 시험성적의 MLFA
klpga-MLFAvarimax.R	[R-코드 3.5.1]	KLPGA 선수 성적에 대한 MLFA의 Varimax 회전 후 인자적재그림
5subjects-MLFAvarimax.R	[R-코드 3.5.2]	두 가지 시험성적에 대한 MLFA의 Varimax 회전 전과 후의 인자적재그림
airpollution-PCMLFAvarimax-scores.R	[R-코드 3.6.1]	LA시 대기오염 자료의 PCMA와 MLFA의 실행과 비교
airpollution-PCFAbiplot.R	[R-코드 3.7.1]	LA시 대기오염 자료의 PCFA의 회전 전과 후의 인자적재와 인자점수 행렬도

3.8 R for FA : Practice Time

[R-code 3.4.1] klpga-PCFAsteps-scee.R in [Example 3.4.1]

```
# PCFA Steps for KLPGA
#[Step 1] Data Matrix X
Data1.3.2<-read.table("klpga.txt", header=T)
X=Data1.3.2
rownames<-rownames(X)
p=ncol(X)
#[Step 2] Covariance Matrix S(or Correlation Matix R)
R=round(cor(X),3)
R

#[Step 3] Spectral Decomposition
eigen.R=eigen(R)
round(eigen.R$values, 2) # Eigenvalues
V=round(eigen.R$vectors, 2) # Eigenvectors

#[Step 4] Number of factors : m
gof=eigen.R$values/p*100 # Goodness-of fit
round(gof, 3)
plot(eigen.R$values, type="b", main="Scree Graph",
      xlab="Factor Number", ylab="Eigenvalue")
```

```
#[Step 5] Factor Loadings and Communality
V2=V[,1:2]
L=V2%*%diag(sqrt(eigen.R$values[1:2]))
round(L, 3)
round(diag(L%*%t(L)), 3)

#[Step 6] Specific Variance : Psi
Psi=diag(R-L%*%t(L))
round(Psi, 3)

#[Step 7] Residual Matrix
Rm = R-(L%*%t(L) + diag(Psi))
round(Rm, 3)

# PCFA using the principal()
library(psych)
pcfa<-principal(R, nfactors=2, rotate="none")
pcfa
round(pcfa$values, 2)
gof=pcfa$values/p*100 # Goodness-of fit
round(gof, 3)
round(pcfa$residual, 2)
```

3.8 R for FA : Practice Time

[R-code 3.4.3] klpga-MLFAfactanal.R in [Example 3.4.3]

```
# MLFA Steps for KLPGA
# Data Matrix X
Data1.3.2<-read.table("klpga.txt", header=T)
X=Data1.3.2
rownames<-rownames(X)
p=ncol(X)
Z<-scale(X, scale=T)

# Covariance Matrix S(or Correlation Matix R)
R=round(cor(X),3)
R

# ML Estimation using the factanal( )
library(psych)
mlfa<-factanal(Z, factors = 2, rotation="none")
mlfa

# Residual Matrix
L=mlfa$loading[, 1:2]
Psi=mlfa$uniquenesses
Rm = R-(L%*%t(L) + diag(Psi))
round(Rm, 3)
```

```
Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 61.74 on 4 degrees of freedom.
The p-value is 1.25e-12
```

3.8 R for FA : Practice Time

[R-code 3.5.1] klpga-MLFAvarimax.R in [Example 3.5.1]

```
# MLFA : None and Varimax Rotation for KLPGA
# Data Matrix X
```

```
Data1.3.2<-read.table("klpga.txt", header=T)
```

```
X=Data1.3.2
```

```
rownames<-rownames(X)
```

```
p=ncol(X)
```

```
# Covariance Matrix S(or Correlation Matix R)
```

```
R=round(cor(X),3)
```

```
R
```

```
# ML Estimation using the factanal( ): None
```

```
library(psych)
```

```
mlfa<-factanal(covmat=R, factors = 2, rotation="none" )
```

```
mlfa
```

```
# Residual Matrix
```

```
L=mlfa$loading[, 1:2]
```

```
Psi=mlfa$uniquenesses
```

```
Rm = R-(L%*%t(L) + diag(Psi))
```

```
round(Rm, 3)
```

```
par(mfrow=c(1,2))
```

```
# Factor Loadings Plot : None
```

```
lim<-range(pretty(L))
```

```
plot(L[,1], L[,2],main="Plot of Factor Loadings : None", xlab="f1", ylab="f2",
      xlim=lim, ylim=lim)
```

```
text(L[,1], L[, 2], labels=rownames(L), cex=0.8, col="blue", pos=1)
```

```
abline(v=0, h=0)
```

```
arrows(0,0, L[,1], L[, 2], col=2, code=2, length=0.1)
```

```
# ML Estimation using the factanal( ): Varimax
```

```
library(psych)
```

```
mlfa<-factanal(covmat=R, factors = 2, rotation="varimax" ) # rotation="none"
mlfa
```

```
# Residual Matrix
```

```
L=mlfa$loading[, 1:2]
```



$$\hat{A}^* = \hat{A} T$$

```
L
```

```
Psi=mlfa$uniquenesses
```

```
Rm = R-(L%*%t(L) + diag(Psi))
```

```
round(Rm, 3)
```

```
# Factor Loadings Plot : Varimax
```

```
lim<-range(pretty(L))
```

```
plot(L[,1], L[,2],main="Plot of Factor Loadings : Varimax ", xlab="f1", ylab="f2",
      xlim=lim, ylim=lim)
```

```
text(L[,1], L[, 2], labels=rownames(L), cex=0.8, col="blue", pos=1)
```

```
abline(v=0, h=0)
```

```
arrows(0,0, L[,1], L[, 2], col=2, code=2, length=0.1)
```

3.8 R for FA : Practice Time

[R-code 3.6.1] airpollution-PCMLFAvarimax-scores.R in [Example 3.6.1]

```
Data2.8.2<-read.table("airpollution.txt", header=T)
X=Data2.8.2
rownames<-rownames(X)
p=ncol(X)
n=nrow(X)
Z<-scale(X, scale=T) # 표준화자료행렬

# Covariance Matrix S(or Correlation Matix R)
R=round(cor(X),3)
R
```

```
##### PCFA using the principal( ): 주성분인자분석 #####
library(psych)
pcfa<-principal(Z, nfactors=3, rotate="varimax")

# Factor Scores : Regression Method
fpc=pcfa$scores
round(fpc, 3)
```

method="regression" : default

```
##### MLFA using the factanal( ) : 최대우도인자분석####
library(psych)
mlfa<-factanal(Z, factors = 3, rotation="varimax", score="regression")

# Factor Scores : Regression Method
fml=mlfa$scores
round(fml, 3)
```

"Bartlett"

3.8 R for FA : Practice Time

[R-code 3.7.1] airpollution-PCFAbiplot.R in [Example 3.7.1]

```
Data2.8.2<-read.table("airpollution.txt", header=T)
```

```
X=Data2.8.2
```

```
rownames(X)
```

```
colnames(X)
```

```
p=ncol(X)
```

```
n=nrow(X)
```

```
Z<-scale(X, scale=T)
```

```
# Biplot based on the Singular Value Decomposition
```

```
svd.Z <- svd(Z)
```

```
U <- svd.Z$u
```

```
V <- svd.Z$v
```

```
D <- diag(svd.Z$d)
```

```
F <- (sqrt(n-1)*U)[,1:2] # Factor Scores Matrix : F
```

```
L <- (sqrt(1/(n-1))*V%*%D)[,1:2] # Factor Loadings Matrix : Lambda
```

```
C<- rbind(F, L)
```

```
rownames(F)<-rownames(X)
```

```
rownames(L)<-colnames(X)
```

```
# Godness-of-fit
```

```
eig <- (svd.Z$d)^2
```

```
per <- eig/sum(eig)*100
```

```
gof <- sum(per[1:2])
```

```
per
```

```
gof
```

```
# Biplot: Joint Plot of Factor Loadings and Scores
```

```
par(mfrow=c(1,2))
```

```
par(pty="s")
```

```
lim1 <- range(pretty(L))
```

```
lim2 <- range(pretty(F))
```

```
biplot(F,L, xlab="f1",ylab="f2", main=" (a)Unrotated Biplot", xlim=lim2, ylim=lim2,  
cex=0.8,pch=16)
```

```
abline(v=0,h=0)
```

```
# Varimax Rotated Biplot: Joint Plot of Rotated Factor Loadings and Scores
```

```
varimax<-varimax(L)
```

```
Lt = varimax$loadings
```

```
T=varimax$rotmat
```

```
T
```

```
Ft= F%*%T
```

```
biplot(Ft,Lt, xlab="f1",ylab="f2", main=" Varimax Rotated Biplot",  
xlim=lim2,ylim=lim2,cex=0.8,pch=16)
```

```
abline(v=0,h=0)
```