

Lecture 7: Independent Component Analysis & Factor Models

1 Independent Components

Reference: *Independent Component Analysis*, by Aapo Hyvärinen, Juha Karhunen, and Erkki Oja (2002), John Wiley.

Independent component analysis (ICA) can be regarded as an extension of the principal component analysis (PCA).

Let $\mathbf{X} = (X_1, \dots, X_p)'$ be a p -dimensional random vector. Let $\mathbf{S} = (S_1, \dots, S_p)'$ be another p -dimensional random vector. Typically, \mathbf{X} denotes the observed variable and \mathbf{S} the *latent* variable, which is not directly observed. Assume that $\mathbf{X} = \mathbf{A}\mathbf{S}$, where $\mathbf{A} = [a_{ij}]$ is a $p \times p$ matrix.

Assumptions:

1. The components of \mathbf{S} are statistically independent, i.e.

$$f(\mathbf{s}) = \prod_{i=1}^p f_i(s_i),$$

where $f_i(s_i)$ is the marginal density function of S_i and $f(\mathbf{s})$ is the joint density of \mathbf{S} .

2. The components S_i are not Gaussian. [Otherwise, PCA is sufficient.]
3. The matrix $\mathbf{A}_{p \times p}$ is non-singular.

ICA is to recover \mathbf{S} from the observed \mathbf{X} .

Remark. In this introduction, we assume that there is no observational noise and the number of independent components is the same as the dimension of the observed variable. These two assumptions can be relaxed.

Ambiguities of ICA, i.e. lack of identifiability

1. The variances of the independent components can not be determined. To simplify the problem, we assume that $\text{Cov}(\mathbf{S}) = \mathbf{I}$.
2. The sign of S_i cannot be determined.
3. The ordering of \mathbf{S} cannot be determined, because for any non-singular permutation matrix \mathbf{P} , we have $\mathbf{P}\mathbf{P}' = \mathbf{I}$ so that

$$\mathbf{X} = \mathbf{A}\mathbf{S} = \mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{S},$$

and $\text{Cov}(\mathbf{P}\mathbf{S}) = \mathbf{P}\text{Cov}(\mathbf{S})\mathbf{P}' = \mathbf{P}\mathbf{P}' = \mathbf{I}$.

1.1 Illustration

Suppose that $p = 2$ and S_i s are uniform on $[-\sqrt{3}, \sqrt{3}]$, i.e.

$$f_i(s_i) = \begin{cases} \frac{1}{2\sqrt{3}}, & \text{if } |s_i| < \sqrt{3} \\ 0 & \text{o.w.} \end{cases}$$

Let

$$\mathbf{X} = \mathbf{A}\mathbf{S} = \begin{bmatrix} 5 & 10 \\ 10 & 2 \end{bmatrix}.$$

Suppose that we observe $\mathbf{x}_1, \dots, \mathbf{x}_{3000}$. What are the results of PCA? What are the results of ICA?

R demonstration: fastICA package.

1.2 Estimation

Several methods are available. A simple method is to maximize nongaussianity of the data, i.e. find the linear combinations that make the data as far away to normality as possible. For instance, maximize the absolute value of the kurtosis. [Recall the PCA is to maximize the component variances.]

Basic idea: Any non-zero linear combination of random variables is closer to normality than the individual variables are. Thus, to recover independent components, one seeks linear combinations that undo the effects of the Central Limit Theorem.

Properties of excess kurtosis relevant to ICA: $K(y) = E(y^4) - 3(E(y^2))^2$, where $E(y) = 0$.

1. If y_1 and y_2 are independent, then $K(y_1 + y_2) = K(y_1) + K(y_2)$.
2. For a constant c , $K(cy_1) = c^4 K(y_1)$.

Let $y = b_1 X_1 + b_2 X_2$ be a linear combination of \mathbf{X} , i.e. $y = \mathbf{b}'\mathbf{X}$. Then, $y = \mathbf{b}'\mathbf{A}\mathbf{S} = (q_1, q_2)\mathbf{S}$. By properties of excess kurtosis,

$$K(y) = K(q_1 S_1) + K(q_2 S_2) = q_1^4 K(S_1) + q_2^4 K(S_2).$$

Since we assume $\text{Var}(y) = 1$, we have $1 = E(y^2) = E(q_1^2 S_1^2 + q_2^2 S_2^2) = q_1^2 + q_2^2$. This means $\mathbf{q} = (q_1, q_2)'$ is on the unit circle.

The question then becomes: what are the maxima of $|K(y)| = |q_1^4 K(S_1) + q_2^4 K(S_2)|$. If we further assume $K(S_1) = K(S_2) = 1$, the problem becomes maximizing

$$F(\mathbf{q}) = q_1^4 + q_2^4, \quad \text{subject to } q_1^2 + q_2^2 = 1.$$

We can plot the contours of $F(\mathbf{q})$ and the unit circle. It is easy to see that the maxima are at $(1, 0)$, $(0, 1)$, $(-1, 0)$ or $(0, -1)$. Thus, \mathbf{q} identifies the components of \mathbf{S} .

Remark. In practice, one often applies PCA first to the observed data and normalizes the PC's so that they have unit variance. This normalized data set is then used to perform ICA.

1.3 Possible applications

Multivariate volatility modeling and dimension reduction.

2 Factor Models

2.1 The orthogonal factor model

Let $\mathbf{X} = (X_1, \dots, X_p)'$ be a p -dimensional random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. \mathbf{X} follows an orthogonal factor model if

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{L}\mathbf{F} + \boldsymbol{\epsilon} \quad (1)$$

where $\mathbf{F} = (F_1, \dots, F_m)'$ is the vector of common factors with $m \leq p$, $\mathbf{L} = [\ell_{ij}]$ is a $p \times m$ matrix of factor loadings, \mathbf{F} and $\boldsymbol{\epsilon}$ are independent, $E(\mathbf{F}) = \mathbf{0}$, $\text{Cov}(\mathbf{F}) = \mathbf{I}_m$, $E(\boldsymbol{\epsilon}) = \mathbf{0}$, and $\text{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi} = \text{diag}\{\psi_1, \dots, \psi_p\}$.

Note that ℓ_{ij} is the *loading* of the i th variable on the j th common factor. From the definition,

$$\begin{aligned} \text{Cov}(\mathbf{X}) &= E(\mathbf{L}\mathbf{F} + \boldsymbol{\epsilon})(\mathbf{L}\mathbf{F} + \boldsymbol{\epsilon})' \\ &= \mathbf{L}E(\mathbf{F}\mathbf{F}')\mathbf{L}' + E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')\mathbf{L}' + \mathbf{L}E(\mathbf{F}\boldsymbol{\epsilon}') + E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') \\ &= \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}. \end{aligned}$$

Thus,

$$\begin{aligned} \text{Var}(X_i) &= \ell_{i1}^2 + \dots + \ell_{im}^2 + \psi_i \\ \text{Cov}(X_i, X_j) &= \ell_{i1}\ell_{j1} + \dots + \ell_{im}\ell_{jm}. \end{aligned}$$

The first identity says that

$$\sigma_{ii} = h_i^2 + \psi_i, \quad \text{with} \quad h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + \dots + \ell_{im}^2,$$

where h_i^2 is called *communality* and ψ_i specific variance.

Also,

$$\text{Cov}(\mathbf{X}, \mathbf{F}) = E(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}' = \mathbf{L}.$$

Thus, $\text{Cov}(X_i, F_j) = \ell_{ij}$.

Ambiguity of factor models: Lack of identifiability

For $m > 1$, let \mathbf{T} be any $m \times m$ orthogonal matrix so that $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$. Then,

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\epsilon} = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{F} + \boldsymbol{\epsilon} = \mathbf{L}^*\mathbf{F}^* + \boldsymbol{\epsilon},$$

where $\mathbf{L}^* = \mathbf{L}\mathbf{T}$ and $\mathbf{F}^* = \mathbf{T}'\mathbf{F}$. It is easy to verify that $\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}^*\mathbf{F}^* + \boldsymbol{\epsilon}$ is also an orthogonal factor model.

Remark: Not all covariance matrix $\boldsymbol{\Sigma}$ gives rise to a proper factor model. See Example 9.2 of the textbook.

2.2 Estimation

Two methods are available to estimate the orthogonal factor model. They are the *principal component method* and the *maximum likelihood method*. The PC method is easy to carry out, but it is an approximation method. The ML method needs some further identification constraint, but may not exist for a given number of PCs.

PC method: Use the spectral decomposition of the covariance matrix Σ .

From spectral decomposition,

$$\Sigma = \lambda_1 \mathbf{e}_1 \mathbf{e}_1' + \cdots + \lambda_p \mathbf{e}_p \mathbf{e}_p',$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are the eigenvalues of Σ and \mathbf{e}_i is the eigenvector associated with eigenvalue λ_i such that $\mathbf{e}_i' \mathbf{e}_i = 1$. This decomposition can be written as

$$\Sigma = [\sqrt{\lambda_1} \mathbf{e}_1, \sqrt{\lambda_2} \mathbf{e}_2, \dots, \sqrt{\lambda_p} \mathbf{e}_p] \begin{bmatrix} \sqrt{\lambda_1} \mathbf{e}_1' \\ \sqrt{\lambda_2} \mathbf{e}_2' \\ \vdots \\ \sqrt{\lambda_p} \mathbf{e}_p' \end{bmatrix}.$$

This is a special case of the orthogonal factor model with $m = p$ and special variance $\psi_i = 0$. In matrix notation, this is

$$\Sigma = \mathbf{L}\mathbf{L}' + \mathbf{0} = \mathbf{L}\mathbf{L}'.$$

This factor analysis is exact, but is not particularly useful. It employs as many common factors as the number of variables.

A truncated version is often used. Let $m < p$ and

$$\mathbf{L} = [\sqrt{\lambda_1} \mathbf{e}_1, \sqrt{\lambda_2} \mathbf{e}_2, \dots, \sqrt{\lambda_m} \mathbf{e}_m].$$

In addition, define $\Psi = \text{diag}\{\psi_1, \psi_2, \dots, \psi_p\}$, where $\psi_i = \sigma_{ii} - \sum_{j=1}^m \ell_{ij}^2$ for $i = 1, \dots, p$. Then

$$\Sigma \approx \mathbf{L}\mathbf{L}' + \Psi.$$

The communalities of the model are estimated as

$$h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2.$$

In application, Σ is estimated by the sample covariance matrix of the data, denoted by \mathbf{S} . Denote the eigenvalues and eigenvectors of \mathbf{S} as $(\hat{\lambda}_i, \hat{\mathbf{e}}_i)$, $i = 1, \dots, p$ with the eigenvalues in decreasing order. The truncated version of factor model is then as follows. Let $m < p$ and

$$\tilde{\mathbf{L}} = [\sqrt{\hat{\lambda}_1} \hat{\mathbf{e}}_1, \sqrt{\hat{\lambda}_2} \hat{\mathbf{e}}_2, \dots, \sqrt{\hat{\lambda}_m} \hat{\mathbf{e}}_m].$$

In addition, define $\tilde{\Psi} = \text{diag}\{\tilde{\psi}_1, \tilde{\psi}_2, \dots, \tilde{\psi}_p\}$, where $\tilde{\psi}_i = s_{ii} - \sum_{j=1}^m \tilde{\ell}_{ij}^2$ for $i = 1, \dots, p$. Then

$$\Sigma \approx \tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\Psi}.$$

The communalities of the model are estimated as

$$\tilde{h}_i^2 = \tilde{\ell}_{i1}^2 + \tilde{\ell}_{i2}^2 + \cdots + \tilde{\ell}_{im}^2.$$

Discussion:

1. Under the PC method, the estimated loadings for a given factor do not change as the number of factors is increased.
2. The *residual matrix* is defined as

$$\mathbf{S} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\mathbf{\Psi}}).$$

By definition, the diagonal elements of this residual matrix are zero. The off-diagonal elements can be shown to satisfy

$$\text{Sum of squared entries of } (\mathbf{S} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\mathbf{\Psi}})) \leq \hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_p^2.$$

Consequently, a small value for the sum of the squares of the neglected eigenvalues implies a small value for the sum of the squared errors of the approximation. This property can be used to select the number of factors m .

3. The proportion of total sample variance due to the j th factor is

$$\frac{\hat{\lambda}_j}{tr(\mathbf{S})} = \frac{\hat{\lambda}_j}{\sum_{i=1}^p s_{ii}}.$$

When the sample correlation matrix \mathbf{R} is used, instead of \mathbf{S} , we have $tr(\mathbf{R}) = p$.

See Examples 9.3 and 9.4 for illustration.

```
> x=read.table("T8-4.DAT")
> dim(x)
[1] 103 5
> colnames(x) <- c("JPM", "C", "WFC", "RDS", "XOM")
> mean(x)
      JPM          C          WFC          RDS          XOM
0.0010627806 0.0006554204 0.0016260816 0.0040491252 0.0040386417
> var(x)
      JPM          C          WFC          RDS          XOM
JPM 4.332695e-04 0.0002756679 1.590265e-04 6.411929e-05 8.896616e-05
C   2.756679e-04 0.0004387172 1.799737e-04 1.814512e-04 1.232623e-04
WFC 1.590265e-04 0.0001799737 2.239722e-04 7.341348e-05 6.054612e-05
RDS 6.411929e-05 0.0001814512 7.341348e-05 7.224964e-04 5.082772e-04
XOM 8.896616e-05 0.0001232623 6.054612e-05 5.082772e-04 7.656742e-04
```

```

> m1=princomp(x) % Use covariance matrix
> summary(m1)
Importance of components:
              Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
Standard deviation  0.03680217 0.02635056 0.01585365 0.01188352 0.01085046
Proportion of Variance 0.52926066 0.27133298 0.09821584 0.05518400 0.04600652
Cumulative Proportion 0.52926066 0.80059364 0.89880948 0.95399348 1.00000000
> names(m1)
[1] "sdev"      "loadings" "center"   "scale"    "n.obs"    "scores"   "call"
> m1$loadings

```

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
JPM	0.223	0.625	0.326	0.663	0.118
C	0.307	0.570	-0.250	-0.414	-0.589
WFC	0.155	0.345		-0.497	0.780
RDS	0.639	-0.248	-0.642	0.309	0.148
XOM	0.651	-0.322	0.646	-0.216	

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
SS loadings	1.0	1.0	1.0	1.0	1.0
Proportion Var	0.2	0.2	0.2	0.2	0.2
Cumulative Var	0.2	0.4	0.6	0.8	1.0

```

> L=cbind(0.0368*m1$loading[,1],0.02635*m1$loading[,2])
> L

```

	[,1]	[,2]
JPM	0.008199880	0.016474706
C	0.011308271	0.015029777
WFC	0.005697019	0.009077705
RDS	0.023514022	-0.006533417
XOM	0.023953282	-0.008480689

```

> LLT = L%*%t(L)
> LLT

```

	JPM	C	WFC	RDS	XOM
JPM	3.386540e-04	0.0003403376	1.962674e-04	8.517603e-05	5.669718e-05
C	3.403376e-04	0.0003537712	2.008593e-04	1.677071e-04	1.434073e-04
WFC	1.962674e-04	0.0002008593	1.148608e-04	7.465140e-05	5.947711e-05
RDS	8.517603e-05	0.0001677071	7.465140e-05	5.955948e-04	6.186459e-04
XOM	5.669718e-05	0.0001434073	5.947711e-05	6.186459e-04	6.456818e-04

```

> Psi=diag(var(x))-diag(LLT)

```

```

> Psi
          JPM          C          WFC          RDS          XOM
9.461549e-05 8.494600e-05 1.091114e-04 1.269016e-04 1.199924e-04
>
> v=var(x) % Use correlation matrix.
> v1=diag(v)
> v1
          JPM          C          WFC          RDS          XOM
0.0004332695 0.0004387172 0.0002239722 0.0007224964 0.0007656742
> v1=sqrt(v1)
> v2=diag(v1)
> v2
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.02081513 0.00000000 0.00000000 0.00000000 0.00000000
[2,] 0.00000000 0.02094558 0.00000000 0.00000000 0.00000000
[3,] 0.00000000 0.00000000 0.0149657 0.00000000 0.00000000
[4,] 0.00000000 0.00000000 0.00000000 0.02687929 0.00000000
[5,] 0.00000000 0.00000000 0.00000000 0.00000000 0.02767082

> v2inv=solve(v2)
> rtn=as.matrix(x)%*%v2inv
> var(rtn)
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 1.0000000 0.6322878 0.5104973 0.1146019 0.1544628
[2,] 0.6322878 1.0000000 0.5741424 0.3222921 0.2126747
[3,] 0.5104973 0.5741424 1.0000000 0.1824992 0.1462067
[4,] 0.1146019 0.3222921 0.1824992 1.0000000 0.6833777
[5,] 0.1544628 0.2126747 0.1462067 0.6833777 1.0000000
> m2=princomp(rtn)
> summary(m2)
Importance of components:
          Comp.1      Comp.2      Comp.3      Comp.4      Comp.5
Standard deviation  1.5535798 1.1804035 0.7040266 0.62940272 0.50268529
Proportion of Variance 0.4874546 0.2814025 0.1001025 0.08000632 0.05103398
Cumulative Proportion 0.4874546 0.7688572 0.8689597 0.94896602 1.00000000
> m2$loading

Loadings:
          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
[1,] -0.469 0.368 0.604 0.363 0.384
[2,] -0.532 0.236 0.136 -0.629 -0.496
[3,] -0.465 0.315 -0.772 0.289

```

```

[4,] -0.387 -0.585          -0.381  0.595
[5,] -0.361 -0.606  0.109  0.493 -0.498

          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
SS loadings      1.0   1.0   1.0   1.0   1.0
Proportion Var   0.2   0.2   0.2   0.2   0.2
Cumulative Var   0.2   0.4   0.6   0.8   1.0
> L=cbind(1.554*m2$loading[,1],1.1804*m2$loading[,2])
> L
          [,1]      [,2]
[1,] -0.7289553  0.4343954
[2,] -0.8273581  0.2791202
[3,] -0.7228638  0.3720379
[4,] -0.6019356 -0.6905780
[5,] -0.5604999 -0.7151410
> LLT=L%*%t(L)
> LLT
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.72007521 0.7243556 0.6885470 0.1388002 0.09792543
[2,] 0.72435562 0.7624295 0.7019105 0.3052621 0.26412390
[3,] 0.68854696 0.7019105 0.6609443 0.1781963 0.13910558
[4,] 0.13880023 0.3052621 0.1781963 0.8392244 0.83124544
[5,] 0.09792543 0.2641239 0.1391056 0.8312454 0.82558674
> Psi=rep(1,5)-diag(LLT)
> Psi
[1] 0.2799248 0.2375705 0.3390557 0.1607756 0.1744133
>

```

The maximum likelihood method: If the common factors \mathbf{F}_j and the specific factors $\boldsymbol{\epsilon}_j$ are normally distributed, then $\mathbf{X}_j - \boldsymbol{\mu} = \mathbf{L}\mathbf{F}_j + \boldsymbol{\epsilon}_j$ are normally distributed. The likelihood function is

$$\begin{aligned}
L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= (2\pi)^{-(np/2)} |\boldsymbol{\Sigma}|^{-n/2} \exp\left[-\frac{1}{2} \text{tr} \left[\boldsymbol{\Sigma}^{-1} \left(\sum_{j=1}^n (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})' + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})' \right) \right]\right] \\
&= (2\pi)^{-(n-1)p/2} |\boldsymbol{\Sigma}|^{-(n-1)/2} \exp\left[-\frac{1}{2} \text{tr} \left[\boldsymbol{\Sigma}^{-1} \left(\sum_{j=1}^n (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})' \right) \right]\right] \\
&\times (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2} (\bar{\mathbf{x}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})\right], \tag{2}
\end{aligned}$$

which depends on \mathbf{L} and $\boldsymbol{\Psi}$ through $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$. This model is not well-defined because \mathbf{L} is not uniquely determined. A commonlu used constraint is

$$\mathbf{L}'\boldsymbol{\Psi}^{-1}\mathbf{L} = \boldsymbol{\Delta},$$

which is a diagonal matrix.

Result 9.1. Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$ is the covariance matrix for the m common (orthogonal) factor model. The MLE $\hat{\mathbf{L}}, \hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ maximize the likelihood function (2) subject to $\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}}$ being diagonal. The maximum likelihood estimates of the communalities are

$$\hat{h}_i^2 = \hat{\ell}_{i1}^2 + \hat{\ell}_{i2}^2 + \dots + \hat{\ell}_{im}^2, \quad i = 1, \dots, p.$$

The proportion of total sample variance due to the j th common factor is

$$\frac{\hat{\ell}_{1j}^2 + \hat{\ell}_{2j}^2 + \dots + \hat{\ell}_{pj}^2}{s_{11} + s_{22} + \dots + s_{pp}}.$$

Remark: In most packages, the sample correlation matrix \mathbf{R} is inserted for $[(n-1)/n]\mathbf{S}$ in the likelihood function (2) and the maximum likelihood estimates $\hat{\mathbf{L}}_z$ and $\hat{\boldsymbol{\Psi}}_z$ are obtained, where the subscript “z” is used to denote standardized data. It turns out that this is equivalent to obtaining the MLE $\hat{\mathbf{L}}$ and $\hat{\boldsymbol{\Psi}}$ based on the sample covariance matrix \mathbf{S} , setting $\hat{\mathbf{L}}_z = \hat{\mathbf{V}}^{-1/2}\hat{\mathbf{L}}$, $\hat{\boldsymbol{\Psi}}_z = \hat{\mathbf{V}}^{-1/2}\hat{\boldsymbol{\Psi}}\hat{\mathbf{V}}^{-1/2}$, where $\hat{\mathbf{V}}^{-1/2}$ is the diagonal matrix with the reciprocal of the sample standard deviations (computed with the divisor \sqrt{n}) on the main diagonal. Similarly, given $\hat{\mathbf{L}}_z$ and specific variances $\hat{\boldsymbol{\Psi}}_z$ obtained from \mathbf{R} , we find that the resulting maximum likelihood estimates for a factor analysis of the covariance matrix $[(n-1)/n]\mathbf{S}$ are

$$\hat{\mathbf{L}} = \hat{\mathbf{V}}^{1/2}, \quad \hat{\boldsymbol{\Psi}} = \hat{\mathbf{V}}^{1/2}\hat{\boldsymbol{\Psi}}_z\hat{\mathbf{V}}^{1/2},$$

or

$$\hat{\ell}_{ij} = \hat{\ell}_{z,ij}\sqrt{\hat{\sigma}_{ii}}, \quad \hat{\psi}_{z,i} = \hat{\psi}_{z,i}\hat{\sigma}_{ii}.$$

See Appendix 9A of the textbook.

```
> m3=factanal(rtn,2) % varimax rotation is used.
> m3
```

Call:

```
factanal(x = rtn, factors = 2)
```

Uniquenesses:

```
[1] 0.417 0.275 0.542 0.005 0.530
```

Loadings:

```
Factor1 Factor2
[1,] 0.763
```

```
[2,] 0.819 0.232
[3,] 0.668 0.108
[4,] 0.113 0.991
[5,] 0.108 0.677
```

```
                Factor1 Factor2
SS loadings      1.725  1.507
Proportion Var   0.345  0.301
Cumulative Var   0.345  0.646
```

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 1.97 on 1 degree of freedom.
The p-value is 0.16
> m4=factanal(rtn,2,rotation="none")
> m4

Call:
factanal(x = rtn, factors = 2, rotation = "none")

Uniquenesses:
[1] 0.417 0.275 0.542 0.005 0.530

Loadings:
 Factor1 Factor2
[1,] 0.121 0.754
[2,] 0.328 0.786
[3,] 0.188 0.650
[4,] 0.997
[5,] 0.685

```
                Factor1 Factor2
SS loadings      1.622  1.610
Proportion Var   0.324  0.322
Cumulative Var   0.324  0.646
```

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 1.97 on 1 degree of freedom.
The p-value is 0.16
>

3 A large sample test for the number of common factors

Assume normality. Consider the test problem

$$H_o : \Sigma = \mathbf{L}_{p \times m} \mathbf{L}'_{m \times p} + \Psi \quad vs \quad H_a : \Sigma \text{ is positive definite.}$$

Under H_a , Σ does not have any other constraint so that $\hat{\Sigma} = ((n-1)/n)\mathbf{S} = \mathbf{S}_n$ and the maximum of the likelihood function is proportional to $|\mathbf{S}_n|^{-n/2} e^{-np/2}$. Under H_o , $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$, and $\hat{\Sigma} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi}$, and the maximum of the likelihood function is proportional to

$$|\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi}|^{-n/2} \exp\left(-\frac{1}{2}n \times \text{tr}[(\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi})^{-1}\mathbf{S}_n]\right).$$

The likelihood ratio statistic is

$$-2 \ln(\Lambda) = -2 \ln\left(\frac{|\hat{\Sigma}|}{|\mathbf{S}_n|}\right)^{-n/2} + n[\text{tr}(\hat{\Sigma}^{-1}\mathbf{S}_n) - p]$$

with degrees of freedom

$$v - v_o = \frac{1}{2}p(p+1) - [p(m+1) - 0.5m(m-1)] = \frac{1}{2}[(p-m)^2 - p - m].$$

It can be shown that $\text{tr}(\hat{\Sigma}^{-1}\mathbf{S}_n) - p = 0$ provided that $\hat{\Sigma} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi}$ is the MLE of $\Sigma = \mathbf{L}\mathbf{L}' + \Psi$. Thus, we have

$$-2 \ln \Lambda = n \ln\left(\frac{|\hat{\Sigma}|}{|\mathbf{S}_n|}\right).$$

Under Bartlett's correction, we reject H_o at the α level of significance if

$$(n - 10(2p + 4m + 5)/6) \ln\left(\frac{|\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi}|}{|\mathbf{S}_n|}\right) > \chi^2_{[(p-m)^2 - p - m]/2}(\alpha)$$

provided that n and $n-p$ are large. Since the number of degrees of freedom must be positive, it follows that

$$m < \frac{1}{2}(2p + 1 - \sqrt{8p + 1})$$

in order to apply the test.

See R demonstration: In R, the test statistic is "STATISTIC" and the associated p-value is "PVAL". All under the command **factanal**.

4 Factor Rotation

Suppose that $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{\Psi}}$ are the estimates of factor model so that $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{L}}\widehat{\mathbf{L}}' + \widehat{\mathbf{\Psi}}$. Let \mathbf{T} be any rotation matrix such that $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$. Then,

$$\widehat{\mathbf{L}}\widehat{\mathbf{L}}' + \widehat{\mathbf{\Psi}} = \widehat{\mathbf{L}}\mathbf{T}\mathbf{T}'\widehat{\mathbf{L}} + \widehat{\mathbf{\Psi}} = \widehat{\mathbf{L}}_*\widehat{\mathbf{L}}_*' + \widehat{\mathbf{\Psi}},$$

where $\widehat{\mathbf{L}}_* = \widehat{\mathbf{L}}\mathbf{T}$ is a “rotated” loading matrix. Thus, rotated loading matrix also provides a valid model with the same specific variances and residual matrix.

One can then use rotation to seek a factor model that has a *simpler structure* such as the factors are readily interpretable. In practice, different criteria lead to different types of rotation. Kaiser (1958, Psychometrika) suggested an analytical measure of simple structure known as the *varimax* criterion. Let $\tilde{\ell}_{ij}^* = \ell_{ij}^*/\hat{h}_i$ be the rotated coefficients scaled by the square root of the communalities. The varimax procedure selects the orthogonal transformation \mathbf{T} that makes

$$V = \frac{1}{p} \sum_{j=1}^m \left[\sum_{i=1}^p (\tilde{\ell}_{ij}^*)^4 - \left(\sum_{i=1}^p (\tilde{\ell}_{ij}^*)^2 \right)^2 / p \right]$$

as large as possible.

The function V is complicated, but has a simple interpretation.

$$V \propto \sum_{j=1}^m \left(\begin{array}{c} \text{variance of the squares of (scaled) loadings} \\ \text{for } j\text{th factor} \end{array} \right).$$

Effectively, maximizing V corresponds to *spreading out* the squares of the loadings on each factor as much as possible. Therefore, we seek to find groups of the large and negligible coefficients in any column of the rotated loading matrix $\widehat{\mathbf{L}}_*$.

Remark: Varimax is the default rotation used in R command **factanal**.

R demonstration: Stock return example continued.

```
> m1=factanal(x,2)
> m1
```

Call:

```
factanal(x = x, factors = 2)
```

Uniquenesses:

V1	V2	V3	V4	V5
0.417	0.275	0.542	0.005	0.530

Loadings:

	Factor1	Factor2
V1	0.763	

```
V2 0.819  0.232
V3 0.668  0.108
V4 0.113  0.991
V5 0.108  0.677
```

```
                Factor1 Factor2
SS loadings      1.725  1.507
Proportion Var   0.345  0.301
Cumulative Var   0.345  0.646
```

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 1.97 on 1 degree of freedom.
The p-value is 0.16

```
> m2=factanal(x,2,rotation="none")
> m2
```

Call:

```
factanal(x = x, factors = 2, rotation = "none")
```

Uniquenesses:

```
    V1    V2    V3    V4    V5
0.417 0.275 0.542 0.005 0.530
```

Loadings:

```
    Factor1 Factor2
V1  0.121   0.754
V2  0.328   0.786
V3  0.188   0.650
V4  0.997
V5  0.685
```

```
                Factor1 Factor2
SS loadings      1.622  1.610
Proportion Var   0.324  0.322
Cumulative Var   0.324  0.646
```

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 1.97 on 1 degree of freedom.
The p-value is 0.16

5 Factor scores

The estimated values of the common factors are called *factor scores*. These quantities can be used as inputs to a subsequent analysis or as statistics in model checking. Factor scores are not unknown parameters in the usual sense. They are estimates of values for the unobserved random factor vectors $\mathbf{F}_i = (F_{i1}, \dots, F_{im})'$ for $i = 1, \dots, n$.

Notice that the unobserved quantities $\hat{\mathbf{f}}_i$ and ϵ_i outnumber the observed data \mathbf{x}_i . To overcome this difficulty, some heuristic, but reasoned, approaches have been proposed in the literature. Following the textbook, we discuss two such approaches. These approaches typically make the following assumptions:

1. Treat the estimated loadings $\hat{\ell}_{ij}$ and specific variances $\hat{\psi}_i$ as if they were the true values.
2. They involve linear transformation of the original data, perhaps centered or standardized.

5.1 The weighted least squares method

Assume that $\boldsymbol{\mu}$, \mathbf{L} and $\boldsymbol{\Psi}$ are known for the factor model

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\epsilon}.$$

Treating $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_p)'$ as errors, we have a regression model with unequal variances ψ_i , $i = 1, \dots, p$. Bartlett (1937) suggested that weighted least squares be used to estimate the common factor values. The sum of squares of the errors, weighted by the reciprocal of their variances, is

$$\sum_{i=1}^p \frac{\epsilon_i^2}{\psi_i} = \boldsymbol{\epsilon}'\boldsymbol{\Psi}^{-1}\boldsymbol{\epsilon} = (\mathbf{X} - \boldsymbol{\mu} - \mathbf{L}\mathbf{F})'\boldsymbol{\Psi}^{-1}(\mathbf{X} - \boldsymbol{\mu} - \mathbf{L}\mathbf{F}).$$

For a given observation \mathbf{x} , Bartlett proposed choosing the estimates $\hat{\mathbf{f}}$ of \mathbf{F} to minimize the prior equation. The solution is

$$\hat{\mathbf{f}} = (\mathbf{L}'\boldsymbol{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}'\boldsymbol{\Psi}^{-1}(\mathbf{x} - \boldsymbol{\mu}).$$

Consequently, for the j th observation, we have

$$\hat{\mathbf{f}}_j = (\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}})^{-1}\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}(\mathbf{x} - \bar{\mathbf{x}}).$$

For MLE $\hat{\mathbf{L}}$ and $\hat{\boldsymbol{\Psi}}$, they satisfy $\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}} = \hat{\boldsymbol{\Delta}}$, a diagonal matrix. Consequently, the factor scores obtained by weighted least squares from the MLEs are

$$\begin{aligned} \hat{\mathbf{f}}_j &= (\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}})^{-1}\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \\ &= \hat{\boldsymbol{\Delta}}^{-1}\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}) \end{aligned}$$

for $j = 1, 2, \dots, n$. If the correlation matrix is factored, then

$$\begin{aligned}\hat{\mathbf{f}}_j &= (\hat{\mathbf{L}}_z' \hat{\mathbf{\Psi}}_z^{-1} \hat{\mathbf{L}}_z)^{-1} \hat{\mathbf{L}}_z' \hat{\mathbf{\Psi}}_z^{-1} \mathbf{z}_j \\ &= \hat{\mathbf{\Delta}}_z^{-1} \hat{\mathbf{L}}_z' \hat{\mathbf{\Psi}}_z^{-1} \mathbf{z}_j\end{aligned}$$

where $\mathbf{z}_j = \mathbf{D}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}})$ and $\hat{\boldsymbol{\rho}} = \hat{\mathbf{L}}_z \hat{\mathbf{L}}_z' + \hat{\mathbf{\Psi}}_z$.

Note that the factor scores generated above have sample mean vector zero and zero sample covariances.

5.2 The regression method

Under the orthogonal factor model and the joint normality of \mathbf{F} and $\boldsymbol{\epsilon}$, $\mathbf{X} - \boldsymbol{\mu} + \mathbf{L}\mathbf{F} + \boldsymbol{\epsilon}$ is distributed as $N_p(\mathbf{0}, \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi})$. Furthermore, $\text{Cov}(\mathbf{X}, \mathbf{F}) = E[(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}'] = \mathbf{L}$. Consequently, we have

$$\begin{bmatrix} \mathbf{X} - \boldsymbol{\mu} \\ \mathbf{F} \end{bmatrix} \sim N_{p+m} \left(\mathbf{0}, \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{L} \\ \mathbf{L}' & \mathbf{I} \end{bmatrix} \right),$$

where $\mathbf{0}$ is an $(p + m)$ -dimensional vector of zeros and \mathbf{I} is the $m \times m$ identity matrix.

Using the property of multivariate normal distribution, $\mathbf{F}|\mathbf{X} = \mathbf{x}$ is normally distributed with mean

$$E(\mathbf{F}|\mathbf{X} = \mathbf{x}) = \mathbf{L}'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{L}'(\mathbf{L}\mathbf{L}' + \boldsymbol{\Psi})^{-1}(\mathbf{x} - \boldsymbol{\mu}),$$

and covariance matrix

$$\text{Cov}(\mathbf{F}|\mathbf{X} = \mathbf{x}) = \mathbf{I} - \mathbf{L}'\boldsymbol{\Sigma}^{-1}\mathbf{L} = \mathbf{I} - \mathbf{L}'(\mathbf{L}\mathbf{L}' + \boldsymbol{\Psi})^{-1}\mathbf{L}.$$

The conditional mean equation is essentially the coefficients in multivariate linear regression of factors on the variables. Consequently, using the MLE $\hat{\mathbf{L}}$ and $\hat{\mathbf{\Psi}}$, we have

$$\hat{\mathbf{f}}_j = \hat{\mathbf{L}}' \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}), \quad j = 1, 2, \dots, n.$$

In practice, the covariance matrix $\hat{\boldsymbol{\Sigma}}$ is often replaced by \mathbf{S} , the original sample covariance matrix. Consequently, the factor scores obtained by the regression are as follows:

1. $\hat{\mathbf{f}}_j = \hat{\mathbf{L}}' \mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})$, $i = 1, 2, \dots, n$,

2. If correlation matrix is used,

$$\hat{\mathbf{f}}_j = \hat{\mathbf{L}}_z' \mathbf{R}^{-1} \mathbf{z}_j, \quad j = 1, 2, \dots, n,$$

where $\mathbf{z}_j = \mathbf{D}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}})$ and $\hat{\boldsymbol{\rho}} = \hat{\mathbf{L}}_z \hat{\mathbf{L}}_z' + \hat{\mathbf{\Psi}}_z$.

Remark: Let $\hat{\mathbf{f}}_j^{LS}$ and $\hat{\mathbf{f}}_j^W$ be the estimates of factor scores by the regression and weighted LS method, respectively. It can be shown that

$$\hat{\mathbf{f}}_j^{LS} = (\mathbf{I} + (\hat{\mathbf{L}}' \hat{\mathbf{\Psi}}^{-1} \hat{\mathbf{L}})^{-1}) \hat{\mathbf{f}}_j^W.$$

R demonstration: Stock return example continued.

```
> m5=factanal(x,2,scores=c("Bartlett"))
> names(m5)
 [1] "converged"      "loadings"      "uniquenesses" "correlation"  "criteria"
 [6] "factors"       "dof"           "method"        "scores"       "STATISTIC"
[11] "PVAL"          "n.obs"         "call"
> dim(m5$scores)
[1] 103  2
> m5$scores[1,]
  Factor1    Factor2
0.2508994 -1.8536707
> m6=factanal(x,2,scores=c("regression"))
> m6$scores[1,]
  Factor1    Factor2
0.1653586 -1.8342740
```