- $\sum_{i=1}^{p} Var(X_i) = \sum_i h_i^2 + \sum_i \psi_i = \sum_i \sum_j \ell_{ij}^2 + \sum_i \psi_i$ (c.f. PCA) $\sum_i Var(X_i) = \sum_{i=1}^{p} Var(PC_i) = \lambda_1 + \cdots + \lambda_p$
- $\text{Cov}(X_i, X_k) = \ell_{i1}\ell_{k1} + \cdots + \ell_{im}\ell_{km}$ ← inner product of the rows of L
  - The covariance are not dependent on the specific factors in any way. The common factors account for the relationship (covariance/correlation) between the observed variables $X_i$'s ← $\Sigma$ is seperated into two parts - diagonal common & specific - off-diagonal common factors

➢ $$(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}' = (\mathbf{LF} + \boldsymbol{\varepsilon})\mathbf{F}' = \mathbf{LF}\mathbf{F}' + \boldsymbol{\varepsilon}\mathbf{F}'$$

so
$$\text{Cov}(\mathbf{X}, \mathbf{F}) = E(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}' = \mathbf{L}\underbrace{E(\mathbf{FF}')}_{=\mathbf{I}} + \underbrace{E(\boldsymbol{\varepsilon}\mathbf{F}')}_{=\mathbf{0}} = \mathbf{L}.$$

That is, $\text{Cov}(X_i, F_j) = \ell_{ij}$
$\ell_{ij} = cor(X_i, F_j)$ if $\Sigma$ replaced by $\rho$

$\Sigma =$ # of parameters in $\Sigma$

➢ The factor model assumes that the $p + p(p-1)/2 = p(p+1)/2$ variances and covariances for $\mathbf{X}$ can be reproduced from the $pm$ factor loadings $\ell_{ij}$ and the $p$ specific variances $\psi_i$. When $m = p$, any covariance matrix $\boldsymbol{\Sigma}$ can be reproduced exactly as $\mathbf{LL}'$, so $\boldsymbol{\Psi}$ can be the zero matrix. However, it is when $m$ is small relative to $p$ that factor analysis is most useful. In this case, the factor model provides a "simple" explanation of the covariation in $\mathbf{X}$ with fewer parameters than the $p(p+1)/2$ parameters in $\boldsymbol{\Sigma}$.

- Unfortunately for the factor analyst, most covariance matrices cannot be factored as $\mathbf{LL}' + \boldsymbol{\Psi}$, where the number of factors $m$ is much less than $p$.

maximum # of common factors ← # of parameters in $LL' + \Psi$ $= pm + p = p(m+1)$
e.g. $p=3$, $m=1$
$p(p+1)/2 = 6$
$p(m+1) = 6$



➢ rotational indeterminacy ($\mathbf{L}$ is not unique) → Note: columns of T form a new coordinate axes.

even though $p(m+1) \ll p(p+1)/2$

- let $\mathbf{T}$ be any $m \times m$ orthogonal matrix, so that $\mathbf{TT}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$.

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon} = \mathbf{LTT}'\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{L}^*\mathbf{F}^* + \boldsymbol{\varepsilon}$$

where
$$\mathbf{L}^* = \mathbf{LT} \quad \text{and} \quad \mathbf{F}^* = \mathbf{T}'\mathbf{F}$$

- Notice that $E(\mathbf{F}^*) = \mathbf{T}'E(\mathbf{F}) = \mathbf{0}$ and $\text{Cov}(\mathbf{F}^*) = \mathbf{T}'\text{Cov}(\mathbf{F})\mathbf{T} = \mathbf{T}'\mathbf{T} = \underset{(m \times m)}{\mathbf{I}}$
- $$\boldsymbol{\Sigma} = \mathbf{LL}' + \boldsymbol{\Psi} = \mathbf{LTT}'\mathbf{L}' + \boldsymbol{\Psi} = (\mathbf{L}^*)(\mathbf{L}^*)' + \boldsymbol{\Psi}$$
- it is impossible, on the basis of observations on $\mathbf{X}$, to distinguish the loadings $\mathbf{L}$ from the loadings $\mathbf{L}^*$. That is, the factors $\mathbf{F}$ and $\mathbf{F}^* = \mathbf{T}'\mathbf{F}$ have the same statistical properties, and even though the loadings $\mathbf{L}^*$ are, in general, different from the loadings $\mathbf{L}$, they both generate the same covariance matrix $\boldsymbol{\Sigma}$.
  - this "ambiguity" provides the rational for "factor rotations."
- Factor loadings $\mathbf{L}$ are determined only up to an orthogonal matrix $\mathbf{T}$. Thus, the loadings
$$\mathbf{L}^* = \mathbf{LT} \quad \text{and} \quad \mathbf{L}$$
both give the same representation. The communalities $= diag(\Sigma - \Psi)$, given by the diagonal elements of $\mathbf{LL}' = (\mathbf{L}^*)(\mathbf{L}^*)'$ are also unaffected by the choice of $\mathbf{T}$.
  - The analysis of the factor model proceeds by imposing conditions that allow one to uniquely estimate $\mathbf{L}$ and $\boldsymbol{\Psi}$.
  - The loading matrix is then rotated, where the rotation is determined by some "ease-of-interpretation" criterion

❖ **Reading**: Textbook, 9.1, 9.2

- Estimation of the factor model
  - Q: what to estimate? $m, L, \Psi$
  - For estimation purpose, we will use the sample covariance matrix $\mathbf{S}$ as our estimate of the population covariance matrix $\boldsymbol{\Sigma}$.
  - The estimation problem in factor analysis is essentially that of finding $\hat{\mathbf{L}}$ and $\hat{\boldsymbol{\Psi}}$ for which

$$\mathbf{S} \approx \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}}. \qquad \text{(c.f. PCA)}$$

$$S = \hat{P}\hat{\Lambda}\hat{P}' \quad \hat{P} = [\hat{e}_1 | \cdots | \hat{e}_p] \quad \hat{\Lambda} = \begin{bmatrix} \hat{\lambda}_1 & & \\ & \ddots & \\ & & \hat{\lambda}_p \end{bmatrix}$$

(If the $x_i$s are standardized, then $\mathbf{S}$ is replaced by $\mathbf{R}$.)

  - Principal Component Approach
    - Recall: spectral decomposition

Let $\boldsymbol{\Sigma}$ have eigenvalue–eigenvector pairs $(\lambda_i, \mathbf{e}_i)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Note: $L = P\Lambda^{1/2}$ the columns of $L$ are orthogonal.

Q: how many constraints have been imposed? if the # of columns of $L$ is $m$ $\binom{m}{2} = \frac{m(m-1)}{2}$

$$\boldsymbol{\Sigma} = \lambda_1\mathbf{e}_1\mathbf{e}_1' + \lambda_2\mathbf{e}_2\mathbf{e}_2' + \cdots + \lambda_p\mathbf{e}_p\mathbf{e}_p'$$

$$= \underbrace{\left[\sqrt{\lambda_1}\,\mathbf{e}_1 \,\vdots\, \sqrt{\lambda_2}\,\mathbf{e}_2 \,\vdots\, \cdots \,\vdots\, \sqrt{\lambda_p}\,\mathbf{e}_p\right]}_{=L=P\Lambda^{1/2}} \begin{bmatrix} \sqrt{\lambda_1}\,\mathbf{e}_1' \\ \sqrt{\lambda_2}\,\mathbf{e}_2' \\ \vdots \\ \sqrt{\lambda_p}\,\mathbf{e}_p' \end{bmatrix}$$

$$= \underset{(p\times p)}{\mathbf{L}}\;\underset{(p\times p)}{\mathbf{L}'} + \underset{(p\times p)}{\mathbf{0}} = \mathbf{LL}'$$

$$\Sigma = P\Lambda P', \quad P = [e_1 | \cdots | e_p]$$
$$= P\Lambda^{1/2}\Lambda^{1/2}P'$$
$$= (P\Lambda^{1/2})(P\Lambda^{1/2})'$$



- when the last $p - m$ eigenvalues are small, neglect the contribution of $\lambda_{m+1}\mathbf{e}_{m+1}\mathbf{e}_{m+1}' + \cdots + \lambda_p\mathbf{e}_p\mathbf{e}_p'$ to $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} \doteq \left[\sqrt{\lambda_1}\,\mathbf{e}_1 \,\vdots\, \sqrt{\lambda_2}\,\mathbf{e}_2 \,\vdots\, \cdots \,\vdots\, \sqrt{\lambda_m}\,\mathbf{e}_m\right] \begin{bmatrix} \sqrt{\lambda_1}\,\mathbf{e}_1' \\ \sqrt{\lambda_2}\,\mathbf{e}_2' \\ \vdots \\ \sqrt{\lambda_m}\,\mathbf{e}_m' \end{bmatrix} = \underset{(p\times m)}{\mathbf{L}}\;\underset{(m\times p)}{\mathbf{L}'}$$

Note: it may not be diagonal matrix

- specific variances may be taken to be the diagonal elements of $\boldsymbol{\Sigma} - \mathbf{LL}'$
- The principal component factor analysis of the sample covariance matrix $\mathbf{S}$ is specified in terms of its eigenvalue–eigenvector pairs $(\hat{\lambda}_1, \hat{\mathbf{e}}_1)$, $(\hat{\lambda}_2, \hat{\mathbf{e}}_2), \ldots,$ $(\hat{\lambda}_p, \hat{\mathbf{e}}_p)$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. Let $m < p$ be the number of common factors. Then the matrix of estimated factor loadings $\{\tilde{\ell}_{ij}\}$ is given by

$$\tilde{\mathbf{L}} = \left[\sqrt{\hat{\lambda}_1}\,\hat{\mathbf{e}}_1 \,\vdots\, \sqrt{\hat{\lambda}_2}\,\hat{\mathbf{e}}_2 \,\vdots\, \cdots \,\vdots\, \sqrt{\hat{\lambda}_m}\,\hat{\mathbf{e}}_m\right] \qquad (9\text{-}15)$$

The estimated specific variances are provided by the diagonal elements of the matrix $\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}'$, so

$$\tilde{\boldsymbol{\Psi}} = \begin{bmatrix} \tilde{\psi}_1 & 0 & \cdots & 0 \\ 0 & \tilde{\psi}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\psi}_p \end{bmatrix} \quad \text{with} \quad \tilde{\psi}_i = s_{ii} - \sum_{j=1}^{m} \tilde{\ell}_{ij}^2 \qquad (9\text{-}16)$$

Communalities are estimated as

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

The principal component factor analysis of the sample correlation matrix is obtained by starting with $\mathbf{R}$ in place of $\mathbf{S}$.

- For this approach, the estimated loadings for a given factor $\boxed{\text{do not change}}$ as the number of factors increases from $m$ to $m+1$
- By the definition of $\tilde{\psi}_i$, the diagonal elements of $\mathbf{S}$ are equal to the diagonal elements of $\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\mathbf{\Psi}}$. However, the off-diagonal elements of $\mathbf{S}$ are not usually reproduced by $\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\mathbf{\Psi}}$.
- Q: how to select the number of common factors, $m$?

  If the number of common factors is not determined by a priori considerations, such as by theory or the work of other researchers, the choice of $m$ can be based on the estimated eigenvalues in much the same manner as with principal components.

- Consider the *residual matrix* $\boxed{\hat{P}(\hat{\Lambda}-\hat{\Lambda}_k)\hat{P}'}$

$$\mathbf{S} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\mathbf{\Psi}}) = \begin{bmatrix} 0 & & \\ & \ddots & \\ \neq 0 & & 0 \end{bmatrix}$$

$tr[\hat{P}(\hat{\Lambda}-\hat{\Lambda}_k)\hat{P}'\hat{P}(\hat{\Lambda}-\hat{\Lambda}_k)\hat{P}']$ $\quad \hat{P}\hat{\Lambda}\hat{P}'$, $\hat{\Lambda} = \begin{bmatrix} \hat\lambda_1 & & 0 \\ & \ddots & \\ 0 & & \hat\lambda_p \end{bmatrix}$ $\quad \hat{P}\hat{\Lambda}_k\hat{P}'$, $\hat{\Lambda}_k = \begin{bmatrix} \hat\lambda_1 & & & \\ & \ddots & & \\ & & \hat\lambda_m & \\ & & & 0 \end{bmatrix}$

Then,

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

variance not explained by L.

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 approximation.



- The contribution to the *total* variance,

  $tr(\Sigma) = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^p Var(X_i)$ $= \sum_i \sum_j \ell_{ij}^2 + \sum_i \psi_i = \sum_j\left(\sum_i \ell_{ij}^2\right) + \sum_i \psi_i$

  from the $i$th common factor is

  $\ell_{1i}^2 + \ell_{2i}^2 + \cdots + \ell_{pi}^2$ ← length$^2$ of the ith ~~factor~~ columns of L

  - For the principal component solution, say the 1st common factor

$$\tilde{\ell}_{11}^2 + \tilde{\ell}_{21}^2 + \cdots + \tilde{\ell}_{p1}^2 = (\sqrt{\hat{\lambda}_1}\,\hat{\mathbf{e}}_1)'(\sqrt{\hat{\lambda}_1}\,\hat{\mathbf{e}}_1) = \hat{\lambda}_1$$

  and

$$\begin{pmatrix} \text{Proportion of total} \\ \text{sample variance} \\ \text{due to } j\text{th common factor} \end{pmatrix} = \begin{cases} \dfrac{\hat{\lambda}_j}{s_{11} + s_{22} + \cdots + s_{pp}} & \text{for a factor analysis of } \mathbf{S} \\[2ex] \dfrac{\hat{\lambda}_j}{p} & \text{for a factor analysis of } \mathbf{R} \end{cases}$$

  is frequently used as a heuristic device for determining the appropriate number of common factor

- **Example 9.3 (Factor analysis of consumer-preference data)** In a consumer-preference study, a random sample of customers were asked to rate several attributes of a new product. The responses, on a 7-point semantic differential scale, were tabulated and the attribute correlation matrix constructed. The correlation matrix is presented next:

$P=5$

$P(P+1)/2 = 15$

$m=2$

$P(m+1) = \cancel{10}\,15$

$m=3$

$P(m+1) = \cancel{15}\,20$

| *Attribute (Variable)* | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Taste | 1 | 1.00 | .02 | .96 | .42 | .01 |
| Good buy for money | 2 | .02 | 1.00 | .13 | .71 | .85 |
| Flavor | 3 | .96 | .13 | 1.00 | .50 | .11 |
| Suitable for snack | 4 | .42 | .71 | .50 | 1.00 | .79 |
| Provides lots of energy | 5 | .01 | .85 | .11 | .79 | 1.00 |

| Variable | Estimated factor loadings orthogonal $\tilde{\ell}_{ij} = \sqrt{\hat{\lambda}_i}\,\hat{e}_{ij}$ PC1 $F_1$ | PC2 $F_2$ | Communalities $\tilde{h}_i^2$ | Specific variances $\tilde{\psi}_i = 1 - \tilde{h}_i^2$ |
|---|---|---|---|---|
| 1. Taste ($e_{11}$) | .56 | .82 $-\Sigma\bigcirc^2\rightarrow$ | .98 + | .02 $=Var(X)=1$ |
| 2. Good buy for money ($e_{12}$) | .78 | −.53 | .88 | .12 |
| 3. Flavor ($e_{13}$) | .65 $=\sqrt{\hat\lambda_1}\hat{e}_1$ | .75 | .98 | .02 |
| 4. Suitable for snack ($e_{14}$) | .94 | −.10 | .89 | .11 |
| 5. Provides lots of energy ($e_{15}$) | .80 $-\Sigma\bigcirc^2$ | −.54 | .93 | .07 |
| Eigenvalues | 2.85 $=\hat{\lambda}_1$ | + 1.81 = | $\Sigma\bigcirc^2 = \sum_i\sum_j \ell_{ij}^2$ | |
| Cumulative proportion of total (standardized) sample variance | .571 | .932 | | |

$\hat{e}_1$

$PC1 = \Sigma e_{1j} X_j$

$X_j \leftarrow \ell_{ij} F_1$

Group 1: {Taste, Flavor}

Group 2: {GBFM, SFS, PLOE}

- ◆ comparison of PCA and FA — ① $\ell_{ij}$ in FA may not be the final loading. ② meaning of $\ell_{ij}$ & $e_{ij}$. ③ compare PCA score & FA score



➢ Modified Principal Component Approach --- Principal Factor Solution

- Idea: the common factor should account for the *off-diagonal* elements of Σ, as well as the *communality portion* of the diagonal elements

$$h_i^2 = \sigma_{ii} - \psi_i^2$$

- algorithm (Note: $\mathbf{\Sigma} = \mathbf{LL}' + \mathbf{\Psi}$) — perform spectral decomposition on $S$ (PC solution); perform spectral decomposition on $S-\hat{\Psi}$ (PF solution)

  1. Guess $\tilde{\Psi}$
  2. Set $\tilde{L}$ = the largest $m$ eigenvectors of the eigendecomposition of $S - \tilde{\Psi}$
  3. Set $\tilde{\Psi} = \text{diag}(S - \tilde{L}\tilde{L}')$

  Repeat steps 2 and 3 until convergence

  - ◆ A popular choice for initial estimates of specific variances is the inverse of the diagonal elements of $\mathbf{S}^{-1}$ (for factoring the correlation matrix $\mathbf{R}$, one minus the initial estimates is equal to the square of the multiple correlation coefficient between $X_i$ and other $p$–1 variables) — Is this still a cov matrix? $cov^2 \le Var(X_1)Var(X_2)$

- some of the eigenvalues of $S - \tilde{\Psi}$ may be negative
- In the iteration, some communality estimates may exceeds Var($X_i$), which results in a negative estimate of the specific variance. This is referred to as *Heywood case*.

➢ Maximum Likelihood Method

- Assume the common factors ($\underset{\sim}{F}$) and the specific factors ($\underset{\sim}{\varepsilon}$) are jointly normally distributed