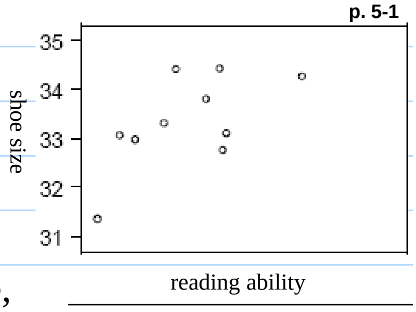
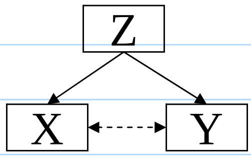
# Factor Analysis




- A motivating example: for children in elementary school
  - ➢ observed variables: shoe size and reading ability
    - ▪ there exists strong correlation between them
  - ➢ latent (lurking) variable: age
  - ➢ **Q**: can we extract information about the latent variable, called *factor*, from the observed variables? If yes, what information?




- Purpose of factor analysis
  - ➢ reduce high dimensional data down to just a few representative variables (similar to PCA)
  - ➢ describe the relationship between many variables (as captured by the covariance/correlation matrix) by a few underlying, but unobservable (latent) variables
    - ▪ *Suppose variables can be grouped by their correlation*. Then, it is conceivable that each group of variables represents a single underlying factor. For example,
      - ◆ 1st group of variables: test scores in classics, French, English, mathematics, and music ⇒ suggest an underlying "intelligence" factor
      - ◆ 2nd group of variables representing physical-fitness scores ⇒ might correspond to another factor





- Modeling (orthogonal factor model)
  - ➢ The observable random vector $\mathbf{X}$, with $p$ components, has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
  - ➢ $\mathbf{X}$ is linearly dependent upon a few unobservable random variables $F_1, F_2, \ldots, F_m$, called *common factors*, and $p$ additional sources of variation $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$, called *errors* or, sometimes, *specific factors*.

$$
\begin{aligned}
X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \varepsilon_1 \\
X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \varepsilon_2 \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \varepsilon_p
\end{aligned}
$$

or, in matrix notation,

$$
\underset{(p\times 1)}{\mathbf{X} - \boldsymbol{\mu}} = \underset{(p\times m)}{\mathbf{L}} \underset{(m\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\varepsilon}}
$$

    - ▪ The coefficient $\ell_{ij}$ is called the *loading* of the $i$th variable on the $j$th factor
    - ▪ Note: $m < p$. (otherwise, why bother?) We want $m$ as small as possible.
    - ▪ Note: $F_1, F_2, \ldots, F_m, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ are unobservable. This distinguishes the factor model from the regression model
  - ➢ assume that

$$
E(\mathbf{F}) = \underset{(m\times 1)}{\mathbf{0}}, \qquad \text{Cov}(\mathbf{F}) = E[\mathbf{F}\mathbf{F}'] = \underset{(m\times m)}{\mathbf{I}}
$$

➢ assume that

$$E(\boldsymbol{\varepsilon}) = \underset{(p\times 1)}{\mathbf{0}}, \qquad \text{Cov}(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \underset{(p\times p)}{\boldsymbol{\Psi}} = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$$

➢ assume that $\mathbf{F}$ and $\boldsymbol{\varepsilon}$ are independent, so

$$\text{Cov}(\boldsymbol{\varepsilon}, \mathbf{F}) = E(\boldsymbol{\varepsilon}\mathbf{F}') = \underset{(p\times m)}{\mathbf{0}}$$

- Some results of the orthogonal factor model

➢
$$\begin{aligned}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' &= (\mathbf{LF} + \boldsymbol{\varepsilon})(\mathbf{LF} + \boldsymbol{\varepsilon})' \\ &= (\mathbf{LF} + \boldsymbol{\varepsilon})((\mathbf{LF})' + \boldsymbol{\varepsilon}') \\ &= \mathbf{LF}(\mathbf{LF})' + \boldsymbol{\varepsilon}(\mathbf{LF})' + \mathbf{LF}\boldsymbol{\varepsilon}' + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\end{aligned}$$

so that

$$\begin{aligned}\boldsymbol{\Sigma} = \text{Cov}(\mathbf{X}) &= E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' \\ &= \mathbf{L}E(\mathbf{FF}')\mathbf{L}' + E(\boldsymbol{\varepsilon}\mathbf{F}')\mathbf{L}' + \mathbf{L}E(\mathbf{F}\boldsymbol{\varepsilon}') + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\ &= \mathbf{LL}' + \boldsymbol{\Psi}\end{aligned}$$

▪ $$\underbrace{\sigma_{ii}}_{\text{Var}(X_i)} = \underbrace{\ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2}_{\text{communality}} + \underbrace{\psi_i}_{\text{specific variance}} = h_i^2 + \psi_i$$

- ◆ $h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2$ is called the $i$th *communality*, which represents the variance shared with the other observed variables via the common factor
- ◆ $\psi_i$ is called the $i$th *specific variance*, which relates to the variability in $X_i$ not shared with other variables





- ◆ $\sum_{i=1}^{p} Var(X_i) =$ (c.f. PCA)

▪ $\text{Cov}(X_i, X_k) = \ell_{i1}\ell_{k1} + \cdots + \ell_{im}\ell_{km}$

- ◆ 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

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

so

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

That is, $\text{Cov}(X_i, F_j) = \ell_{ij}$

➢ 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 $p$.

- rotational indeterminacy (**L** is not unique)
  - 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, 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?
  - For estimation purpose, we will use the sample covariance matrix **S** as our estimate of the population covariance matrix **Σ**.
  - The estimation problem in factor analysis is essentially that of finding $\hat{\Lambda}$ and $\hat{\Psi}$ for which
    $$\mathbf{S} \approx \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}. \qquad \text{(c.f. PCA)}$$
    (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 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

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

$$= \left[\sqrt{\lambda_1}\,\mathbf{e}_1 \,\vdots\, \sqrt{\lambda_2}\,\mathbf{e}_2 \,\vdots\, \cdots \,\vdots\, \sqrt{\lambda_p}\,\mathbf{e}_p\right] \begin{bmatrix} \sqrt{\lambda_1}\,\mathbf{e}_1' \\ \hdashline \sqrt{\lambda_2}\,\mathbf{e}_2' \\ \hdashline \vdots \\ \hdashline \sqrt{\lambda_p}\,\mathbf{e}_p' \end{bmatrix}$$

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

- 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 $\mathbf{\Sigma}$

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

- specific variances may be taken to be the diagonal elements of $\mathbf{\Sigma} - \mathbf{L}\mathbf{L}'$

- 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

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





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 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*

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

Then,

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

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 approximation.