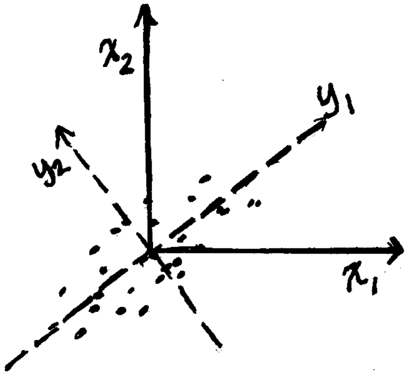
IV. Analysis of Covariance Structure

IV.A. Principal Component Analysis

- Seek to maximize the variance of a linear combination of the variables
- Purpose: explain the covariance structure of a set of variables (one-sample technique)
- Often useful as inputs to another analysis
  - Principal Component Regression involves converting p collinear predictors into $k < p$ independent components
    - Often yields better estimates of regression coefficients

A Geometric Motivation:



- Suppose we have an ellipsoidal swarm of points where $x_1, \ldots, x_p$ are correlated.
- Ellipsoidal swarm of points not parallel to any of the $x_1, \ldots, x_p$ axes
- Wish to find "natural axes" for the points, that is, the axes of the ellipsoid

- Assuming $\mathbf{x}_i, i = 1, \ldots, n$, are centered at $\mathbf{0}$, we rotate the axes by multiplying $\mathbf{x}_i$ by an orthogonal matrix $\mathbf{A}$:

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i$$

  - Distance from origin is same for both old and new data

$$\mathbf{y}_i'\mathbf{y}_i = (\mathbf{A}\mathbf{x}_i)'(\mathbf{A}\mathbf{x}_i) = \mathbf{x}_i' \underbrace{\mathbf{A}'\mathbf{A}}_{\substack{\mathbf{A}\text{ is}\\ \text{orthogonal}}} \mathbf{x}_i = \mathbf{x}_i'\mathbf{x}_i$$

  - If $\mathbf{A}$ were chosen to rotate axes in harmony with ellipsoid, then

$$\text{var}\{\mathbf{y}\} = \mathbf{S}_y = \mathbf{A}\mathbf{S}\mathbf{A}' = \begin{bmatrix} s_{y_1}^2 & & & \mathbf{0} \\ & s_{y_2}^2 & & \\ & & \ddots & \\ \mathbf{0} & & & s_{y_p}^2 \end{bmatrix}$$

- But, by definition, the matrix $\mathbf{A}$ which diagonalizes $\mathbf{S}$ is a matrix containing the normalized eigenvectors of $\mathbf{S}$.

- $y_{i1} = \mathbf{a}_1'\mathbf{x}_i, y_{i2} = \mathbf{a}_2'\mathbf{x}_i$, etc. are the “principal component scores” for subject i, where $\text{var}\{y_{ij}\} = s^2_{y_j} = \lambda_j$ (the j^{th} ordered eigenvalue of $\mathbf{S}$).
  - Since $\mathbf{a}_1$ specifies the single dimension of greatest separation among the data, it is expected that $\text{var}\{y_1\}$ should have largest variance among $(s^2_{y_j})$ and $\text{var}\{y_p\}$ should have smallest variance.
- Number of “significant” eigenvalues indicates number of “essential dimensions” in data.

$$\begin{array}{c}\text{Proportion}\\ \text{of variance}\\ \text{explained}\\ (\text{using } \lambda_1, \ldots, \lambda_k)\end{array} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\sum_{i=1}^{p} s_{ii}}$$

  - We then represent the p-dimensional data $(x_{i1}, \ldots, x_{ip})$ where $k < p$

An Algebraic Motivation

We want a linear combination of $\mathbf{x}$ with maximal variance $\widehat{\text{var}}\{\mathbf{a}'\mathbf{x}\} = \mathbf{a}'\mathbf{S}\mathbf{a}$

Since $\mathbf{a}'\mathbf{S}\mathbf{a}$ has no max for arbitrary $\mathbf{a}$, we wish to maximize (for $\mathbf{a} \neq \mathbf{0}$)

$$\lambda = \frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{\mathbf{a}'\mathbf{a}} \tag{1}$$

$$\Rightarrow \mathbf{a}'\mathbf{S}\mathbf{a} = \lambda\mathbf{a}'\mathbf{a}$$

$$\Rightarrow \mathbf{a}'(\mathbf{S}\mathbf{a} - \lambda\mathbf{a}) = 0$$

$$\Rightarrow \mathbf{S}\mathbf{a} - \lambda\mathbf{a} = 0 \qquad \text{since } \mathbf{a} \text{ in } * \neq \mathbf{0}$$

$$\Rightarrow (\mathbf{S} - \lambda\mathbf{I})\mathbf{a} = 0$$

So $\mathbf{a}_1$ (the normalized eigenvector associated with largest eigenvalue) maximizes λ in (1) at the value of:

$$\lambda_1 = \frac{\mathbf{a}_1'\mathbf{S}\mathbf{a}_1}{\mathbf{a}_1'\mathbf{a}_1} = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1 = \text{var}\{y_1\}$$

- k^{th} p.c. $(k = 2, \ldots, p)$ is defined by $\mathbf{a}_k$ (the eigenvector corresponding to the k^{th} largest eigenvalue) since

$$\max_{\mathbf{a} \perp \mathbf{a}_1, \ldots, \mathbf{a}_{k-1}} \frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{\mathbf{a}'\mathbf{a}} = \lambda_k$$

  with the maximum occurring at $\mathbf{a} = \mathbf{a}_k$

Notes:

1. P.C.'s $y_i = \mathbf{a}_i'\mathbf{x}$ and $y_j = \mathbf{a}_j'\mathbf{x}$ are orthogonal (and uncorrelated) for $i \neq j$

2. "Component scores" can be calculated for each individual:

$$\begin{aligned} y_{i1} &= \mathbf{a}_1'\mathbf{x}_i \\ y_{i2} &= \mathbf{a}_2'\mathbf{x}_i \\ &\vdots \\ y_{ip} &= \mathbf{a}_p'\mathbf{x}_i \end{aligned}$$

or

$$\underset{n\times p}{\mathbf{Y}} = \begin{bmatrix} y_{\cdot 1} & \cdots & y_{\cdot p} \end{bmatrix} = \underset{n\times p}{\mathbf{X}}\underset{p\times p}{\mathbf{A}}$$

where $\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_p \end{bmatrix}$.

Plotting the first 2 or more component scores against each other can be useful for checking:

- non-linearity (indication of non-normality)
- outliers
- clusters or groupings

3. "Component loadings" (elements of $\mathbf{a}_i$) are useful for gaining insight about which variables are important in the principal component

   $*$ Interpretation of standardized loadings is not the panacea that it is in discriminant analysis or canonical correlation analysis . . .

4. Principal components are NOT scale invariant ☹

   - Orientation of P.C.'s changes if we convert from: inches to centimeters, pounds to kg, original scale to standardized scale, etc.

$\boxed{\text{ex}}$

$$\underset{\sim}{S} = \begin{bmatrix} 1 & 2 \\ 2 & 100 \end{bmatrix} \xrightarrow[\text{matrix}]{\text{convert to corr'n}} \underset{\sim}{R} = \begin{bmatrix} 1 & .2 \\ .2 & 1 \end{bmatrix}$$

$\lambda_1 = 100.04$

$\lambda_2 = 0.96$

$\underset{\sim}{a}_1 = [.0202 \quad .9998]$

$\underset{\sim}{a}_2 = [.9998 \quad -.0202]$

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = 99.04\%$$

$\lambda_1^{(R)} = 1.2$

$\lambda_2^{(R)} = 0.8$

$\underset{\sim}{a}_1^{(R)} = [.7071 \quad .7071]$

$\underset{\sim}{a}_2^{(R)} = [.7071 \quad -.7071]$

$$\frac{\lambda_1^{(R)}}{\lambda_1^{(R)} + \lambda_2^{(R)}} = 60\%$$

- No simple relationships exist between:
  - λ_i and $\lambda_i^{(R)}$
  - $\frac{\lambda_i}{\Sigma\lambda_i}$ and $\frac{\lambda_i^{(R)}}{\Sigma\lambda_i^{(R)}}$
  - $\mathbf{a}_1$ and $\mathbf{a}_1^{(R)}$

- P.C. #1 (based on **S**):

$$y_1 = .0202x_1 + .9998x_2$$

- P.C. #1 (based on **R**):

$$\begin{aligned} y_1 &= .7071\left(\frac{x_1-\bar{x}_1}{\sqrt{s_{11}}}\right) + .7071\left(\frac{x_2-\bar{x}_2}{\sqrt{s_{22}}}\right) \\ &= .7071\,(x_1-\bar{x}_1) + .0707\,(x_2-\bar{x}_2) \end{aligned}$$

  - Even expressing the p.c.'s from **R** in terms of the original (unstandardized) variables gives a different component

5. Components for a given **R** are not unique to that matrix. E.g.,

$$\mathbf{R} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

always has eigenvalues $\lambda_1 = 1 + r$ and $\lambda_2 = 1 - r$, regardless of the value of r, and the corresponding principal components are:

$$y_1 = .7071 \left(\frac{x_1 - \bar{x}_1}{\sqrt{s_{11}}} \right) + .7071 \left(\frac{x_2 - \bar{x}_2}{\sqrt{s_{22}}} \right)$$

and

$$y_2 = .7071 \left(\frac{x_1 - \bar{x}_1}{\sqrt{s_{11}}} \right) - .7071 \left(\frac{x_2 - \bar{x}_2}{\sqrt{s_{22}}} \right)$$

Question: Is it reasonable to assign a meaningful interpretation to the p.c.'s from the correlation matrix?

6. P.C.A. can still be carried out with a singular **S** matrix

7. All variables uncorrelated ⇒ the variables <u>are</u> the P.C.'s
   - Characteristic equation

     $$0 = |\mathbf{S} - \lambda \mathbf{I}| = \prod_{i=1}^{p} (s_{ii} - \lambda)$$

     has solutions $\lambda_i = s_{ii}$ with eigenvectors

     $$\mathbf{a}_i = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}$$

     i^{th} P.C.:

     $$y_i = \mathbf{a}_i' \mathbf{x} = x_i$$

   - So, 1^{st} P.C. is x_i with largest variance
     2^{nd} P.C. is x_i with second largest variance, etc.

8. When one or more variables has dramatically larger variance than others, those variables will dominate the 1^{st} P.C.

9. All correlations/covariances are positive ⇒ all elements of $\mathbf{a}_1 > 0$
   - So, $\mathbf{a}_1$ will be similar to an average or sum and can be interpreted as an overall measure of "size"
   - Since $\mathbf{a}_2, \ldots, \mathbf{a}_p$ are orthogonal to $\mathbf{a}_1$ $(\mathbf{a}_1'\mathbf{a}_2 = \mathbf{a}_1'\mathbf{a}_3 = \cdots = \mathbf{a}_1'\mathbf{a}_p = 0)$, $\quad \mathbf{a}_2, \ldots, \mathbf{a}_p$ will each contain positive and negative coefficients. These weighted differences can be interpreted as aspects of "shape"
     – These patterns often still hold true when only most of covariances are positive

Interpretation of P.C.'s

- Examine coefficients (although PC's from **S** and **R** require different interpretations)
- Correlation between i^{th} variable x_i and j^{th} P.C. y_j can be calculated to identify which variables are most important
  - Not useful in a multivariate context (Rencher, 1992)

    $$(\text{corr}\{x_i, y_1\})^2 + \cdots + (\text{corr}\{x_i, y_k\})^2 = R^2_{x_i|y_1,\ldots,y_k}$$

    Squared correlations between x_i and the P.C.'s reflect the univariate relationship between x_i and P.C.'s

    ∗ <u>NO</u> info about importance of x_i in the presence of $x_1, \ldots, x_{i-1}, x_{i+1}, \cdots, x_p$

$\boxed{\text{ex}}$ Football helmet design

3 groups of subjects:

1 = high school football players
2 = college football players
3 = non-football players

6 variables:

6 variables :

wdim = head width at widest dimension

circum = head circumference

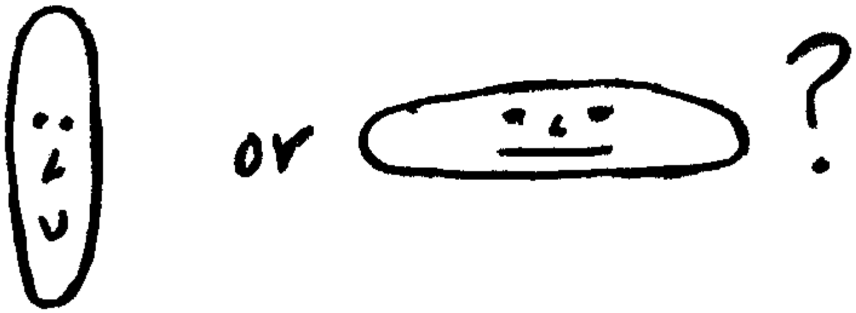
fbeye = front-to-back measure at eye level

eyehd = eye-to-top-of-head measure

earhd = ear-to-top-of-head measure

jaw = jaw width

Which is the high school player?



Number of Principal Components

How many do we retain to sufficiently capture and summarize the covariance/correlation structure?

Approaches:

(1) Retain enough components to account for $100k\%$ of the total variability (where k is .80 or .90 or $\ldots$)

(2) Retain components whose variance (eigenvalue) is greater than the average eigenvalue ($\Sigma\lambda_i/p$)

[For PCA on correlation matrix, this is equivalent to retaining y_i when $\lambda_i > 1 \Longrightarrow$ sometimes called the "rule of 1"]

(3) Use scree plot to find natural break between "large/important" components and "small/unimportant" components [scree = "rocky debris at base of a cliff"]

$*$ Approaches (1), (2), and (3) lack theoretical justification

$\boxed{\text{ex}}$ football helmet data

(4) Test that the last k population eigenvalues (denoted $\lambda_{p-k+1}^{\text{pop}}, \ldots, \lambda_p^{\text{pop}}$) are equal (and hence, usually small):

$$H_0 : \lambda_{p-k+1}^{\text{pop}} = \cdots = \lambda_p^{\text{pop}}$$

- Calculate $\bar{\lambda} = \frac{1}{k} \sum_{i=p-k+1}^{p} \lambda_i$
- $u = \left(n - \frac{2p+11}{6}\right) \left(k \ln \bar{\lambda} - \sum_{i=p-k+1}^{p} \ln \lambda_i\right) \overset{\text{approx}}{\sim} \chi^2_{\frac{1}{2}(k-1)(k+2)}$

IV.B. Factor Analysis and S.E.M.

ref. Bollen (1989), Fuller (1987)

Brief Roadmap for Latent Variable Models:

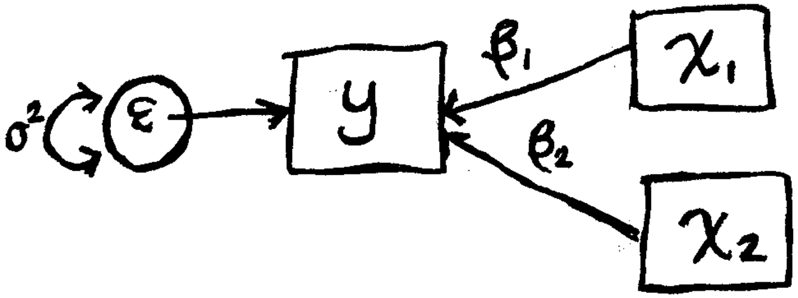
- Use observable variables in order to understand nature of, and relationship between latent (unobservable) variables.
- Traditionally, latent variables have been concepts or constructs which are impossible to measure directly
  - attitudes
  - aptitudes/abilities
  - intelligence
  - experience
  - quality
  - etc.

- Path diagrams show relationship between

observables (manifest variables and latent variables

Ex. Regression model

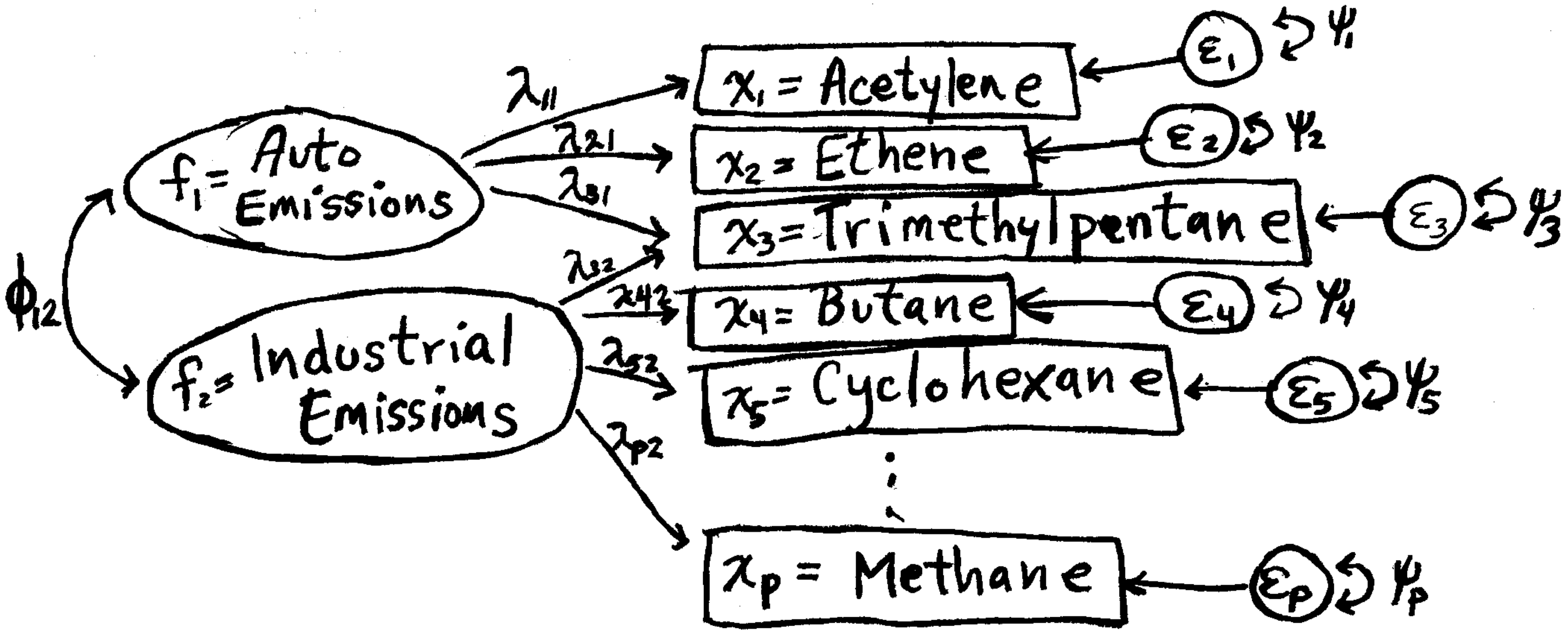
$$y = \beta_1 x + \beta_2 x_2 + \varepsilon \quad , \quad \text{var}\{\varepsilon\} = \sigma^2$$



∗ Factor Analysis Models:

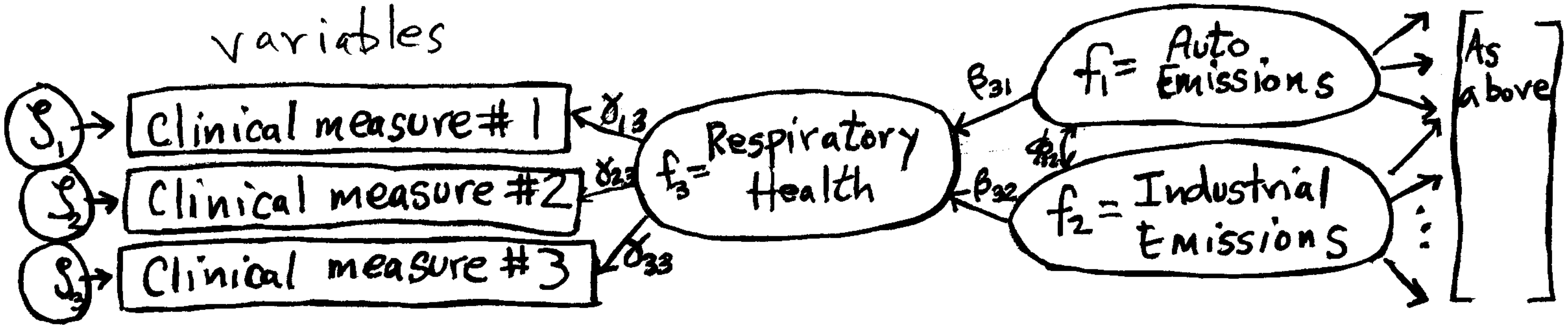
- Explore and describe structure in the p observable variables using model with $k < p$ factors
- Represent/validate subject matter theory

$\boxed{\text{ex}}$ Path diagram



∗ Structural Equation Models:

- Objectives of F.A.
- Model causal relationships <u>between</u> latent variables



Factor Analysis Model:

$$\left\{ \underset{p\times 1}{\mathbf{x}} = \underset{p\times 1}{\boldsymbol{\mu}} + \underset{p\times k}{\boldsymbol{\Lambda}} \underset{k\times 1}{\mathbf{f}} + \underset{p\times 1}{\mathbf{e}} \right\} \qquad \text{(IV – 1)}$$

or

$$\left\{ \begin{array}{ll} x_1 & = \mu_1 + \lambda_{11} f_1 + \lambda_{12} f_2 + e_1 \\ x_2 & = \mu_2 + \lambda_{21} f_1 + \lambda_{22} f_2 + e_2 \\ \ \vdots & \\ x_p & = \mu_p + \lambda_{p1} f_1 + \lambda_{p2} f_2 + e_p \end{array} \right\}$$

λ_{ij}: "factor loading" of i^{th} variable on j^{th} factor (not an eigenvalue)

μ_i = mean of i^{th} variable

f_j = j^{th} common factor

e_i = i^{th} error ("specific factor")

$$\left\{ \underset{p\times 1}{\mathbf{x}} = \underset{p\times 1}{\boldsymbol{\mu}} + \underset{p\times k}{\boldsymbol{\Lambda}}\underset{k\times 1}{\mathbf{f}} + \underset{p\times 1}{\mathbf{e}} \right\} \qquad (\mathrm{IV}-1)$$

- $\mathbf{f}$ and $\mathbf{e}$ are independent
- $E\{\mathbf{e}\} = \mathbf{0}$ and $\mathrm{var}\{\mathbf{e}\} = \underset{p\times p}{\boldsymbol{\Psi}} = \begin{bmatrix} \psi_1 & & 0 \\ & \ddots & \\ 0 & & \psi_p \end{bmatrix}$
- $\mathrm{var}\{\mathbf{f}\} = \underset{k\times k}{\boldsymbol{\Phi}} = \begin{bmatrix} \phi_{11} & \phi_{21} & \cdots & \phi_{k1} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2k} \\ \vdots & \vdots & \ddots & \\ \phi_{k1} & \phi_{k2} & \cdots & \phi_{kk} \end{bmatrix}$ and $E\{\mathbf{f}\} = \mathbf{0}$
  - So many unobservable quantities that the factor model cannot be verified from the data ... need further assumptions.

IV.B.i. Exploratory Factor Analysis (via the Orthogonal Factor Model)

Key Assumption:

Facilitate verification of the factor analysis model (IV-1) by assuming

$$\mathrm{var}\{\underset{k\times 1}{\mathbf{f}}\} = \underset{k\times k}{\Phi} = \mathbf{I}_k \qquad \text{(IV-2)}$$

Thus, all factors f_j are independent with zero mean and unit variance.

→ Underlying "sources" affecting the observed variables are non-redundant

Under model (IV-1) and orthogonal factor assumption (IV-2):

$$
\begin{aligned}
\mathrm{var}\{\mathbf{x}\} &= \boldsymbol{\Lambda}\mathrm{var}\{\mathbf{f}\}\boldsymbol{\Lambda}' + \mathrm{var}\{\mathbf{e}\} \\
&= \boldsymbol{\Lambda}\mathbf{I}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \\
&= \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{cov}\{\mathbf{x}, \mathbf{f}\} &= \{\boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \mathbf{e}, \mathbf{f}\} \\
&= \mathrm{cov}\{\boldsymbol{\Lambda}\mathbf{f}, \mathbf{f}\} \\
&= \boldsymbol{\Lambda}
\end{aligned}
$$

$$
\begin{array}{ccl}
\mathrm{var}\{x_i\} & = & [\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{ik}][\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{ik}]' + \psi_i \\
\underbrace{\sigma_{ii}}_{\substack{\uparrow \\ \text{variance for} \\ \text{variable } i}} & = & \underbrace{\lambda_{i1}^2 + \lambda_{i2}^2 + \ldots + \lambda_{ik}^2}_{\substack{\uparrow \\ h_i^2 \,=\, \text{communality} \\ \text{for variable } i}} + \underbrace{\psi_i}_{\substack{\uparrow \\ \text{specific variance} \\ \text{for variable } i}}
\end{array}
$$

$$\boxed{\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}}$$

Notes:

1. Refer to var$\{\mathbf{x}\}$ as $\boldsymbol{\Sigma}[\boldsymbol{\theta}]$ where the argument $\boldsymbol{\theta}$ is a vector containing all model parameters affecting var$\{\mathbf{x}\}$.

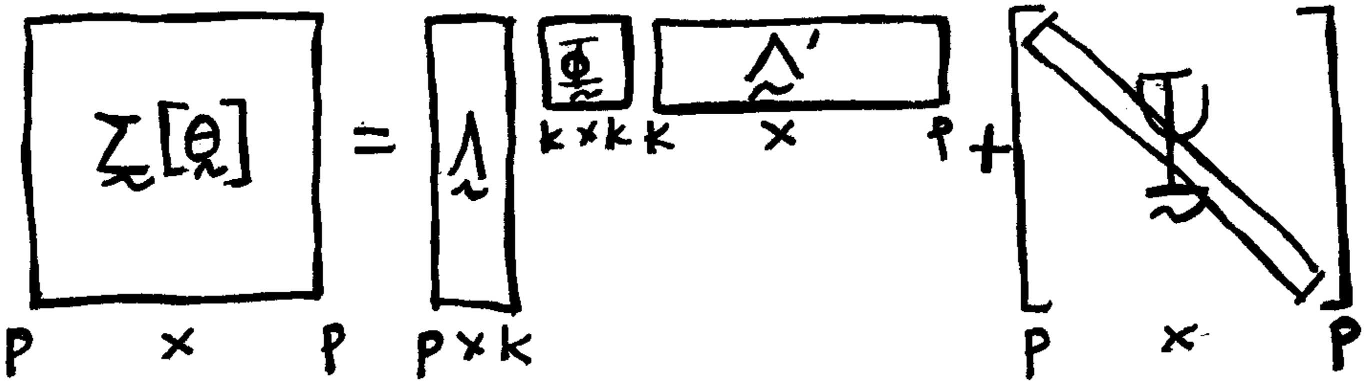
   - In general,

$$\boldsymbol{\theta} = (\underset{\substack{\uparrow \\ \text{vector} \\ \text{containing} \\ \text{all unique} \\ \text{factor} \\ \text{loading} \\ \text{parameters}}}{\boldsymbol{\lambda}'} \quad , \quad \underset{\substack{\uparrow \\ \text{vector} \\ \text{containing} \\ \text{all unique} \\ \text{params in} \\ \boldsymbol{\Phi} = \text{var}\{\mathbf{f}\}}}{\boldsymbol{\phi}'} \quad , \quad \underset{\substack{\uparrow \\ \text{contains} \\ \psi_1, \ldots, \psi_p}}{\boldsymbol{\psi}'})'$$

   - For factor model (IV-1) and orthogonality assumptions (IV-2) :

$$\underset{(pk+p)\times 1}{\boldsymbol{\theta}} = \left(\underset{1\times pk}{\boldsymbol{\lambda}'}, \underset{1\times p}{\boldsymbol{\psi}'}\right)'$$

2. Structure of $\mathbf{\Sigma}[\boldsymbol{\theta}] = \text{var}\{\mathbf{x}\}$ can be visualized as



when the model holds. Factor analysis evaluates whether covariance matrix estimate (i.e., **S**) fits the above structure

3. Non-uniqueness of $\mathbf{\Lambda}$

   Loadings in $\mathbf{x} = \boldsymbol{\mu} + \mathbf{\Lambda f} + \mathbf{e}$ can be multiplied by an orthogonal matrix without affecting the model fit.

$$\begin{aligned} \mathbf{x} &= \boldsymbol{\mu} + \mathbf{\Lambda f} + \mathbf{e} \\ &= \boldsymbol{\mu} + \underbrace{\mathbf{\Lambda T}}_{\mathbf{\Lambda}^*}\underbrace{\mathbf{T}'\mathbf{f}}_{\mathbf{f}^*} + \mathbf{e}, \qquad \text{where} \underset{k\times k}{\mathbf{T}}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}_k \end{aligned}$$

   Note that…

$$
\begin{aligned}
E\{\mathbf{f}^*\} &= E\{\mathbf{T}'\mathbf{f}\} = \mathbf{T}'\mathbf{0} = \mathbf{0} \\
\mathrm{var}\{\mathbf{f}^*\} &= \mathbf{T}'\mathrm{var}\{\mathbf{f}\}\mathbf{T} \\
&= \mathbf{T}'\mathbf{I}_k\mathbf{T} = \mathbf{I}_k \\
\mathrm{var}\{\mathbf{x}\} &= \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*\prime} + \boldsymbol{\Psi} \\
&= (\boldsymbol{\Lambda}\mathbf{T})(\boldsymbol{\Lambda}\mathbf{T})' + \boldsymbol{\Psi} \\
&= \boldsymbol{\Lambda}\mathbf{T}\mathbf{T}'\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \\
&= \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \\
\mathrm{var}\{x_i\} &= \underbrace{\boldsymbol{\lambda}_i^{*\prime}\boldsymbol{\lambda}_i^*}_{\substack{\text{communality}\\ \text{after transform}}} + \psi_i \\
&= \boldsymbol{\lambda}_i'\mathbf{T}'\mathbf{T}\boldsymbol{\lambda}_i + \psi_i \\
&= \underbrace{\boldsymbol{\lambda}_i'\boldsymbol{\lambda}_i}_{\substack{\text{communality}\\ \text{before transform}}} + \psi_i
\end{aligned}
$$

… <u>SO</u> any orthogonal transformation of $\boldsymbol{\Lambda}$ is <u>equally valid</u>.

4. Principal Components Analysis vs. Factor Analysis

- P.C.A. and E.F.A. are purely exploratory where confirmatory F.A. and S.E.M. can be used for modeling and inference
- P.C.'s are linear combinations of the p observed variables

  $$\text{PC1} = y_1 = \mathbf{a}_1'\mathbf{x}$$

  In F.A., the p observed variables are expressed as linear combinations of the factors

  $$x_1 = \mu_1 + \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{ik} f_k + e_1$$

  (Factor scores <u>are</u> linear combinations of the p observed variables)
- Removing one or more P.C.'s does not affect the nature of the other P.C.'s

  Removing one or more factors will change the factor loadings and factor scores associated with other factors when using any estimation method other than "principal component method"

<u>Estimation Methods</u>

$*$ *Principal Component Method*

Let $d_1, \ldots, d_p$ be the p ordered eigenvalues of $\mathbf{S}$ (or $\mathbf{R}$) and let $\mathbf{a}_1, \ldots, \mathbf{a}_p$ be the corresponding eigenvectors. Let $\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_p \end{bmatrix}$ and $\mathbf{D} = \begin{bmatrix} d_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & d_p \end{bmatrix}$

We neglect $\hat{\mathbf{\Psi}}$ in

$$\mathbf{S} \cong \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}$$

and say

$$\begin{aligned} \mathbf{S} &= \underset{p\times p}{\mathbf{A}}\,\underset{p\times p}{\mathbf{D}}\,\underset{p\times p}{\mathbf{A}'} \\ &\cong \underset{p\times k}{\mathbf{A}_1}\,\underset{k\times k}{\mathbf{D}_1}\,\underset{k\times p}{\mathbf{A}_1'} \qquad \left(\begin{array}{c}\text{retaining only } k < p \text{ factors,} \\ \text{where } \mathbf{A}_1 = [\mathbf{a}_1 \ldots \mathbf{a}_k] \\ \text{and } \mathbf{D}_1 = \text{diag}(d_1 \ldots d_k)\end{array}\right) \\ &= (\mathbf{A}_1\mathbf{D}_1^{1/2})(\mathbf{A}_1\mathbf{D}_1^{1/2})' \\ &= \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' \end{aligned}$$

$$\Rightarrow \hat{\mathbf{\Lambda}} = \mathbf{A}_1\mathbf{D}_1^{1/2}$$

and

$$\hat{\mathbf{\Psi}} = \mathbf{S} - \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' \qquad (\hat{\psi}_i = s_{ii} - \underbrace{\sum_{j=1}^{k} \hat{\lambda}_{ij}^2}_{\substack{h_i^2 \text{ or} \\ \text{“communality”}}})$$

or, if using $\mathbf{R}$

$$\hat{\mathbf{\Psi}} = \mathbf{R} - \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' \qquad (\hat{\psi}_i = 1 - \underbrace{\sum_{j=1}^{k} \hat{\lambda}_{ij}^2}_{h_i^2})$$

- Proportion of variance due to j^{th} factor is:

$$\frac{d_j}{\mathrm{tr}\{\mathbf{S}\}} \quad (\text{where } d_j \text{ also equal to } \textstyle\sum_{i=1}^{p} \lambda_{ij}^2)$$

Notes about principal component method:

1. Before "rotation," loadings on j^{th} factor $(\lambda_{1j}, \ldots, \lambda_{pj})$ are proportional to coefficients on j^{th} P.C.

2. Before "rotation," loadings on a retained factor are unchanged by adding/deleting other factors.

3. "Principal Factor Method" extends the "principal component method" by taking an initial error variance estimate $\mathbf{\Psi}^{\circ}$ and defining
$$\hat{\mathbf{\Lambda}} = \mathbf{A}_1^* \mathbf{D}_1^{*\frac{1}{2}}$$
where $\mathbf{A}^* \mathbf{D}^* \mathbf{A}^{*\prime}$ is the spectral decomposition of $\mathbf{S} - \mathbf{\Psi}^{\circ}$ (or $\mathbf{R} - \mathbf{\Psi}^{\circ}$). Process can be iterated since $\hat{\mathbf{\Psi}}$ is updated after each estimation of $\mathbf{\Lambda}$. For the $(i+1)$th iteration, let $(\hat{\psi}_1, \ldots, \hat{\psi}_p)$ be equal to diag$\{\mathbf{S} - \hat{\mathbf{\Lambda}}^{(i)} \hat{\mathbf{\Lambda}}^{\prime(i)}\}$.

- An initial estimate of h_i^2 (when using $\mathbf{R} - \mathbf{\Psi}^\circ$) is

$$\hat{h}_i^2 = 1 - \frac{1}{r^{ii}} = R^2_{x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p}$$

where r^{ii} is the i^{th} diagonal element of $\mathbf{R}^{-1}$. So,

$$\hat{\psi}_i^\circ = 1 - \left(1 - \frac{1}{r^{ii}}\right) = \frac{1}{r^{ii}}$$

or, if using $\mathbf{S} - \mathbf{\Psi}^\circ$

$$\hat{\psi}_i^\circ = s_{ii} - \left(s_{ii} - \frac{1}{s^{ii}}\right) = \frac{1}{s^{ii}}$$

where s_{ii} and s^{ii} are i^{th} diagonal elements of $\mathbf{S}$ and $\mathbf{S}^{-1}$, respectively.

4. Choosing the number of factors

   (a) Choose k equal to number of factors necessary to account for 80%, 85%, or 90% of the total variance $\text{tr}\{\mathbf{S}\}$ (or $\text{tr}\{\mathbf{R}\}$).

   (b) Choose k equal to: number of eigenvalues of $\mathbf{S}$ greater than $\frac{\text{tr}\{\mathbf{S}\}}{p}$ (or number of eigenvalues of $\mathbf{R}$ greater than 1).

   (c) Use scree plot to determine which eigenvalues are associated with the cliff (retain) and which eigenvalues are the "scree" (remove).

$*$ *Maximum Likelihood Method*

Assuming $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we can find $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Psi}}$ such that $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is maximized subject to a uniqueness condition

$$\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} = \underset{\substack{\uparrow \\ \text{a diagonal} \\ \text{matrix}}}{\boldsymbol{\Delta}}$$

As before, communalities are

$$\hat{h}_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2$$

and proportion of variance due to j^{th} factor

$$\frac{\left(\sum_{i=1}^{p} \lambda_{ij}^2\right)}{\text{tr}\{\mathbf{S}\}}$$

$\rightarrow$ Choosing # of factors

- Can use the 3 methods suggested with the principal component/principal factor methods.
- When using ML method, more often we test $H_0 : \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$ using Bartlett's modified likelihood ratio statistic

$$\left(n - \frac{2p + 4k + 11}{6}\right) \ln \left(\frac{|\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}|}{|\mathbf{S}|} \right) \sim \chi^2_{\frac{1}{2}[(p-k)^2 - (p+k)]}$$

  - Degrees of freedom for test of $H_0 : \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$ is equal to the number of independent parameters in the basic factor analysis model (IV-1) subtracted from number of unique sample statistics in $\mathbf{S}$ (and $\bar{\mathbf{x}}$).

# of sample statistics =

$+\frac{1}{2}p(p+1)$ ← variances and covariances in $\underset{\sim}{S}$

$+\ p$ ← means in $\underset{\sim}{\bar{X}}$

of <u>independent</u> parameters =

$+\ pk$ ← elements of $\underset{\sim}{\Lambda}$

$+\ \frac{1}{2}k(k+1)$ ← unique elements in $\underset{k\times k}{\underset{\sim}{\Phi}} = \text{var}\{\underset{k\times 1}{\underset{\sim}{f}}\}$

$+\ p$ ← error variances $\psi_1, \dots, \psi_p$

$+\ p$ ← means in $\underset{\sim}{\mu}$

$-\ k^2$ ← since parameterization

$$\underset{\sim}{x} = \underset{\sim}{\mu} + \underset{\sim}{\Lambda}\underset{\sim}{f} + \underset{\sim}{e}$$

is unique up to an orthogonal transformation

$$\underset{\sim}{x} = \underset{\sim}{\mu} + \underbrace{\underset{\sim}{\Lambda}\underset{k\times k}{\underset{\sim}{T}}}_{\underset{\sim}{\Lambda}^*}\underbrace{\underset{\sim}{T}'\underset{\sim}{f}}_{\underset{\sim}{f}^*} + \underset{\sim}{e}$$

elements in arbitrary transformation matrices

So,

$$\begin{aligned} df &= \left(\frac{1}{2}p(p-1)+p\right) \\ &\quad - \left(pk + \frac{1}{2}k(k+1) + p + p - k^2\right) \\ &= \frac{1}{2}\left[(p-k)^2 - (p+k)\right] \end{aligned}$$

In fact, maximum likelihood estimation yields valid solutions only if $df \geq 0$ and test of H_0 valid only if $df > 0$.

Rule-of-thumb: Identifiability of a factor analysis model:

$$(p-k)^2 \text{ must be } \geq (p+k)$$

or

$$k \leq \frac{1}{2}\left(2p+1-\sqrt{8p+1}\right)$$

p	k	$(p-k)^2$	$(p+k)$	identified model?
2	1	1	3	No
3	1	4	4	Yes $(\underset{\sim}{S} \equiv \underset{\sim}{\hat{\Lambda}}\,\underset{\sim}{\hat{\Lambda}}' + \underset{\sim}{\hat{\Psi}})$
3	2	1	5	No
4	1	9	5	Yes
4	2	4	6	No
5	2	9	7	Yes
5	3	4	8	No

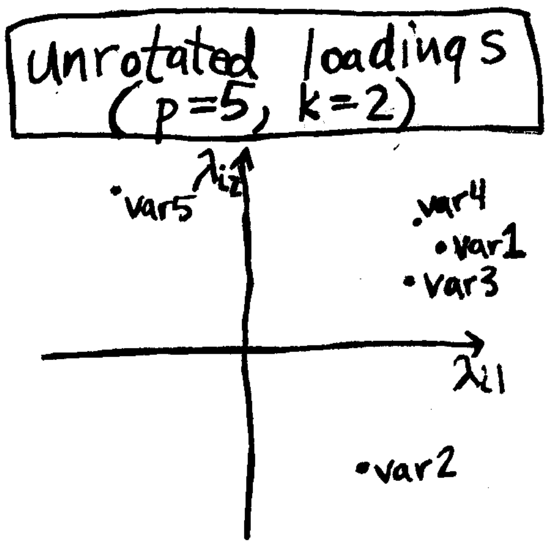
Factor Rotation

As previously noted, factor loadings are unique only up to multiplication by an orthogonal matrix that rotates the loadings.

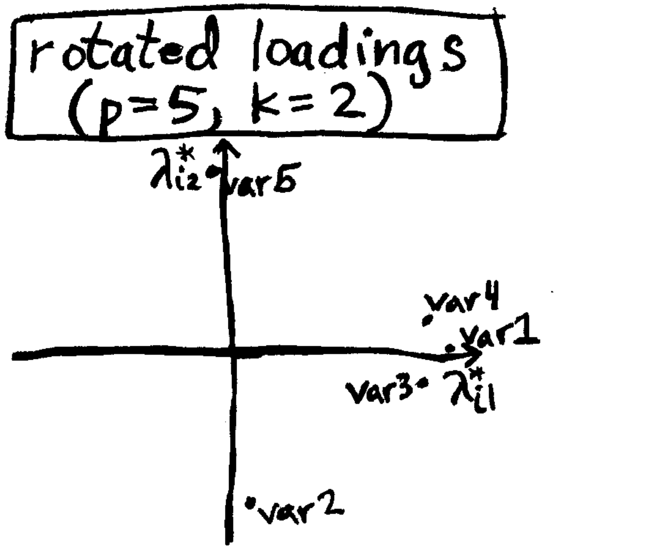
- Rotated loadings (and factors) reproduce the covariance matrix

$$\begin{aligned}\mathbf{S} &\cong \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}} \\ &= \underbrace{\hat{\mathbf{\Lambda}}\mathbf{T}}\underbrace{\mathbf{T}'\hat{\mathbf{\Lambda}}} + \hat{\mathbf{\Psi}} \\ &= \hat{\mathbf{\Lambda}}^{*}\hat{\mathbf{\Lambda}}^{*'} + \hat{\mathbf{\Psi}}\end{aligned}$$

- Rotate $\hat{\mathbf{\Lambda}}$ to obtain more interpretable structure— “simple structure” is where each variable loads only on one factor is the ideal.



All var's are linear combinations of both factors



Vars 1, 3, and 4 depend on factor 1 and vars 2 and 5 depend on factor 2.

- "Complexity" of variable is number of factors on which a variable loads highly—we want low complexity for variables.

- Orthogonal Rotation
  - Communalities stay the same

$$h_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2 = \sum_{j=1}^{k} {\lambda_{ij}^*}^2 = {h_i^*}^2$$

    but variance accounted for by factor j changes with rotation

$$\sum_{i=1}^{p} \lambda_{ij}^2 \neq \sum_{i=1}^{p} {\lambda_{ij}^*}^2$$

  - "Varimax Rotation"

    Choose $\mathbf{T}$ in $\mathbf{\Lambda}^* = \mathbf{\Lambda T} = \left(\lambda_{ij}^*\right)$ such that

$$\frac{1}{p} \sum_{j=1}^{k} \left[\sum_{i=1}^{p} \left(\frac{\hat{\lambda}_{ij}^*}{\hat{h}_i} \right)^4 - \frac{1}{p} \left(\sum_{i=1}^{p} \frac{\hat{\lambda}_{ij}^{*2}}{\hat{h}_i^2} \right)^2 \right]$$

    is maximized.

Essentially, varimax "spreads out" the loadings for each factor

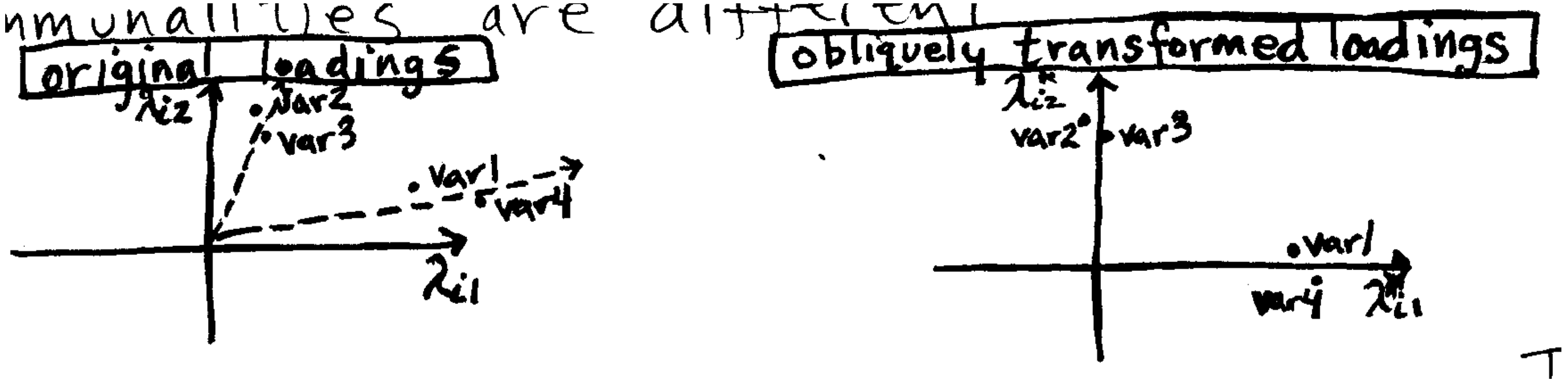
$\boxed{\text{ex}}$

$$(\lambda_{1j}^*, \lambda_{2j}^*, \lambda_{3j}^*, \ldots, \lambda_{pj}^*) = (\underbrace{.95, -.04, \ldots, .12}_{\text{high variability}})$$

has simpler structure (more desirable) than

$$(\lambda_{1j}^*, \lambda_{2j}^*, \lambda_{3j}^*, \ldots, \lambda_{pj}^*) = (\underbrace{.85, .62, \ldots, .71}_{\text{low variability}})$$

- Oblique "Rotation" Transformation
  - $\mathbf{f}^* = \mathbf{Q}'\mathbf{f}$ where $\text{var}\{\mathbf{f}^*\} = \mathbf{Q}'\mathbf{I}\mathbf{Q} = \mathbf{Q}'\mathbf{Q} \neq \mathbf{I}$
  - Axes are no longer perpendicular (factors are correlated)
  - Communalities are different



Factor Score Estimation

After we have estimates of $\boldsymbol{\mu}, \boldsymbol{\Lambda}$, and $\boldsymbol{\Psi}$, we can use the model for the t^{th} individual

$$\underset{p\times 1}{\mathbf{x}_t} - \bar{\mathbf{x}} = \hat{\boldsymbol{\Lambda}} \underset{k\times 1}{\mathbf{f}_t} + \boldsymbol{\varepsilon}, \qquad \widehat{\text{var}}\{\boldsymbol{\varepsilon}\} = \hat{\boldsymbol{\Psi}}$$

to obtain estimates of the unknown quantities $\mathbf{f}_t$.

- Weighted Least Squares Method

$$\underset{k\times 1}{\hat{\mathbf{f}}_t} = (\underset{k\times p}{\hat{\boldsymbol{\Lambda}}'} \underset{p\times p}{\hat{\boldsymbol{\Psi}}^{-1}} \underset{p\times k}{\hat{\boldsymbol{\Lambda}}})^{-1} \underset{k\times p}{\hat{\boldsymbol{\Lambda}}'} \underset{p\times p}{\hat{\boldsymbol{\Psi}}^{-1}} (\underset{p\times 1}{\mathbf{x}_t} - \underset{p\times 1}{\bar{\mathbf{x}}})$$

- Regression Method
  - Treats factors and observations as jointly normally distributed

$$\hat{\mathbf{f}}_t = \hat{\boldsymbol{\Lambda}}'(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}})^{-1}(\mathbf{x}_t - \bar{\mathbf{x}})$$

or

$$\hat{\mathbf{f}}_t = \hat{\boldsymbol{\Lambda}}'\mathbf{S}^{-1}(\mathbf{x}_t - \bar{\mathbf{x}})$$

$\boxed{\text{ex}}$ Stat 2301 Grades

- Principal Component Method
  - Varimax (orthogonal) rotation
  - Promax (oblique) rotation
- Maximum Likelihood
  - Varimax rotation

IV.B.ii Confirmatory Factor Analysis (CFA)

Because of the indeterminacy in our model

$$\begin{aligned}\mathbf{x} &= \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \mathbf{e} \qquad\qquad \text{(IV-1)} \\ &= \boldsymbol{\mu} + \underbrace{\boldsymbol{\Lambda}\mathbf{T}}_{\boldsymbol{\Lambda}^*}\underbrace{\mathbf{T}'\mathbf{f}}_{\mathbf{f}^*} + \mathbf{e}\end{aligned}$$

we cannot assign any practical physical meaning to $\boldsymbol{\Lambda}$ or $\mathbf{f}$

The approach of CFA is to choose a sensible parameterization to make parameter estimates unique.

A widely used parameterization is the "errors-in-variables" or "measurement error" model:

$$\underbrace{\begin{bmatrix} \underset{(p-k)\times 1}{\mathbf{x}_1} \\ \underset{k\times 1}{\mathbf{x}_2} \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} \underset{(p-k)\times 1}{\boldsymbol{\beta}_0} \\ \underset{k\times 1}{\mathbf{0}} \end{bmatrix}}_{\boldsymbol{\mu}} + \underbrace{\begin{bmatrix} \underset{(p-k)\times k}{\mathbf{B}} \\ \mathbf{I}_k \end{bmatrix}}_{\boldsymbol{\Lambda}} \mathbf{f} + \mathbf{e}$$

- Defines k of the variables to be equal to one of the factors plus error.
- Allows meaningful physical interpretation and statistical inference associated with model parameters (e.g., λ_{ij}).

$\boxed{\text{ex}}$ Stat 2301 Grades

As an expert in statistical education, I want to test my theory that course assignments (excluding group assignments) are functions of 2 factors associated with daily effort and knowledge mastery.

Specifically, I hypothesize:

$$\begin{array}{rllll}
\text{Labmean} & = \beta_{01} & & +\lambda_{12} f_2 & +e_1 \\
\text{PQmean} & = \beta_{02} & +\lambda_{21} f_1 & +\lambda_{22} f_2 & +e_2 \\
\text{Exam1} & = \beta_{03} & +\lambda_{31} f_1 & & +e_3 \\
\text{Exam2} & = \beta_{04} & +\lambda_{41} f_1 & & +e_4 \\
\text{ExamFin} & = & \phantom{+\lambda_{41}} f_1 & & +e_5 \\
\text{HWmean} & = & & \phantom{+\lambda_{22}} f_2 & +e_6
\end{array}$$

- Like an errors-in-variables regression in which predictors (f_1 and f_2) are only observed in their error-contaminated states ($f_1 + e_5$ and $f_2 + e_6$)

- f_1 now constrained to have same mean as Examfin with $\text{var}\{f_1\} \leq \text{var}\{\text{ExamFin}\}$. In fact, we can consider f_1 to be an error-free version of ExamFin. (although f_1 is in fact estimated using all the variables).

- Parameters are uniquely determined so interpretation and inference are feasible. For example,

  H_0: PQmean is not affected by knowledge mastery

  is equivalent to

  $$H_0 : \lambda_{21} = 0$$

  and can be tested with

  $$t = \frac{\hat{\lambda}_{21}}{\text{s.e.}(\hat{\lambda}_{21})}$$

- C.F.A. model allows for standard statistical model-building and model assessment using goodness-of-fit (χ^2) test.

- Note that the errors-in-variables parameterization removes the factor indeterminacy by replacing a $k \times k$ portion of

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1k} \\ \vdots & & \vdots \\ \lambda_{p1} & \cdots & \lambda_{pk} \end{bmatrix}$$

with the full rank matrix $\mathbf{I}_k$ to obtain

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1k} \\ \vdots & & \\ \lambda_{(p-k),1} & \cdots & \lambda_{(p-k),k} \\ 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$$

BUT, any full rank $k \times k$ matrix of constants could be placed in k of the rows of Λ to yield an identified model.

- When a full-rank $k \times k$ matrix of constants replaces k rows of $\boldsymbol{\Lambda}$, we have as parameters:
  - $\star$ p means ($p - k$ in β_0, plus $E\{f_1\}, \ldots, E\{f_x\}$)
  - $\star$ $(p-k)k$ loadings in $\boldsymbol{\Lambda}$
  - $\star$ $\frac{1}{2}k(k+1)$ unique elements in Φ
  - $\star$ p specific variances $\psi_1, \ldots, \psi_p$
- Because the p mean parameters can be (optionally) estimated with the p sample means, the difference between [# of statistics in $\mathbf{S}$ and $\bar{\mathbf{x}}$] and [number of parameters] is

$$\begin{aligned} df &= \frac{1}{2}p(p+1) - \{(p-k)k + \frac{1}{2}k(k+1) + p\} \\ &= \frac{1}{2}[(p-k)^2 - (p+k)] \qquad \leftarrow \boxed{df \text{ for a } \chi^2 \text{ GOF test}} \end{aligned}$$

When $\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Psi}$ have additional constants, the df is larger:

$$df = \frac{1}{2}p(p+1) - \text{ \# of parameters in } \boldsymbol{\theta} = (\boldsymbol{\lambda}', (\text{vech } \boldsymbol{\Phi})', \boldsymbol{\psi}').$$

- Basic estimation approach: Estimator $\hat{\boldsymbol{\theta}}$ chosen to minimize the difference between $\mathbf{S}$ and $\boldsymbol{\Sigma}[\boldsymbol{\theta}]$ over $\boldsymbol{\theta} \in$ parameter space

Estimation and Inference for C.F.A. Model Parameters

Develop estimators assuming

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}[\boldsymbol{\theta}], \boldsymbol{\Sigma}[\boldsymbol{\theta}])$$

$$\Rightarrow (n-1)\mathbf{S} \sim W_p(\boldsymbol{\Sigma}[\boldsymbol{\theta}], n-1)$$

We can then argue that these estimators (and associated inferential methods) are valid even when $\mathbf{x}$ is NOT normal (see Anderson and Amemiya, 1988; Amemiya and Anderson, 1990)

$*$ Maximum Likelihood Estimator

Likelihood (under Wishart):

$$L(\mathbf{\Sigma}[\boldsymbol{\theta}], \mathbf{S}) = \text{constant} \times |\mathbf{\Sigma}[\boldsymbol{\theta}]|^{-n/2} \exp\left\{-\frac{1}{2}\text{tr}\left\{n\mathbf{S}\left(\mathbf{\Sigma}[\boldsymbol{\theta}]\right)^{-1}\right\}\right\}$$

Define $-2\log(\text{likelihood ratio})$ as :

$$\ell(\boldsymbol{\theta}, \mathbf{S}) = n\left(\log|\mathbf{\Sigma}[\boldsymbol{\theta}]| + \text{tr}\left\{\mathbf{S}\left(\mathbf{\Sigma}[\boldsymbol{\theta}]\right)^{-1}\right\} - \log|\mathbf{S}| - p\right)$$

- $\hat{\boldsymbol{\theta}}_{ML}$ minimizes $\ell(\boldsymbol{\theta}; \mathbf{S})$ over parameter space Θ
- $\ell(\hat{\boldsymbol{\theta}}_{ML}; \mathbf{S}) \to \chi^2_{\frac{1}{2}p(p+1)-q}$

  is a goodness-of-fit statistic where $\boldsymbol{\theta}$ is $q \times 1$ and $\underbrace{\frac{1}{2}p(p+1)}_{=p^*}$ is # of statistics in $\mathbf{S}$.

  – Reject H_0 : "$\mathbf{\Sigma}[\boldsymbol{\theta}]$ model holds" if $\ell(\hat{\boldsymbol{\theta}}_{ML}, \mathbf{S}) > \chi^2_{p^*-q,\alpha}$

Some definitions:

- Let $\mathbf{A} = \begin{bmatrix} \boxed{\mathbf{a}_1} & \boxed{\mathbf{a}_2} & \cdots & \boxed{\mathbf{a}_p} \end{bmatrix}$ be $p \times p$ and symmetric. Then,
  - vec $\mathbf{A} = \underset{p^2 \times 1}{\begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_p \end{bmatrix}}$

- $$\text{vech } \mathbf{A} = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \\ a_{22} \\ a_{32} \\ \vdots \\ a_{p2} \\ a_{33} \\ a_{43} \\ \vdots \\ a_{pp} \end{bmatrix}_{\underbrace{\left(\frac{1}{2}p(p+1)\right)}_{p^*} \times 1}$$

– For any $p \times p$ symmetric $\mathbf{A}$, there is a $p^2 \times \overbrace{\left(\frac{1}{2}p(p+1)\right)}^{p^*}$ matrix $\mathbf{H}_p$ such that vec $\mathbf{A} = \mathbf{H}_p$vech $\mathbf{A}$

$\boxed{\text{ex}}$ $p = 2$

$$\underbrace{\begin{bmatrix} a_{11} \\ a_{21} \\ a_{12} \\ a_{22} \end{bmatrix}}_{\substack{\text{vec } \mathbf{A} \\ 4\times 1}} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\substack{\mathbf{H}_2 \\ 4\times 3}} \underbrace{\begin{bmatrix} a_{11} \\ a_{21} \\ a_{22} \end{bmatrix}}_{\substack{\text{vech } \mathbf{A} \\ 3\times 1}}$$

– $\underset{p^* \times p^2}{\mathbf{H}_p^+} = (\mathbf{H}_p'\mathbf{H}_p)^{-1}\mathbf{H}_p'$ (Moore-Penrose generalized inverse)

So for any $p \times p$ symmetric matrix $\mathbf{A}$

$$\begin{aligned}\underset{p^* \times p^2}{\mathbf{H}_p^+} \underbrace{\text{vec } \mathbf{A}}_{p^2 \times 1} &= \underset{p^* \times p^2}{\mathbf{H}_p^+} \underset{p^2 \times p^*}{\mathbf{H}_p} \underbrace{\text{vech } \mathbf{A}}_{p^* \times 1} \\ &= \text{vech } \mathbf{A}\end{aligned}$$

$\boxed{\text{ex}}$ $p = 2$

$$\mathbf{H}_2^+ = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Lemma:

If $(n-1)\mathbf{S} \sim W_p(\boldsymbol{\Sigma}[\boldsymbol{\theta}], n-1)$

- $E\{\text{vech } \mathbf{S}\} = \text{vech } \boldsymbol{\Sigma}[\boldsymbol{\theta}]$
- $\text{var}\{\text{vech } \mathbf{S}\} = \frac{2}{n-1} \underbrace{\mathbf{H}_p^+(\boldsymbol{\Sigma}[\boldsymbol{\theta}] \otimes \boldsymbol{\Sigma}[\boldsymbol{\theta}])\mathbf{H}_p^{+\prime}}_{p^* \times p^*}$
  - * Note that because $\mathbf{X} \sim N_p$, the 4th moments of $\mathbf{X}$ are functions of 2nd moments

Further, if $\boldsymbol{\Sigma}[\boldsymbol{\theta}]$ is positive definite

- $(\text{var}\{\text{vech } \mathbf{S}\})^{-1} = \frac{n-1}{2}\mathbf{H}_p' \left((\boldsymbol{\Sigma}[\boldsymbol{\theta}])^{-1} \otimes (\boldsymbol{\Sigma}[\boldsymbol{\theta}])^{-1}\right) \mathbf{H}_p$

Proof: see Fuller, 1987, p. 386

Define

$$\mathbf{V}[\boldsymbol{\theta}] = \frac{2}{n-1}\mathbf{H}_p^+(\boldsymbol{\Sigma}[\boldsymbol{\theta}] \otimes \boldsymbol{\Sigma}[\boldsymbol{\theta}])\mathbf{H}_p^{+\prime}$$

and

$$\hat{\mathbf{V}} = \frac{2}{n-1}\mathbf{H}_p^+(\mathbf{S} \otimes \mathbf{S})\mathbf{H}_p^{+\prime}$$

Least Squares Estimator

Let

$$q(\boldsymbol{\theta}; \mathbf{S}) = (\text{vech } \mathbf{S} - \text{vech } \boldsymbol{\Sigma}[\boldsymbol{\theta}])' \hat{\mathbf{V}}^{-1} (\text{vech } \mathbf{S} - \text{vech } \boldsymbol{\Sigma}[\boldsymbol{\theta}])$$

- $\hat{\boldsymbol{\theta}}_{LS}$ minimizes $q(\boldsymbol{\theta}; \mathbf{S})$ over parameter space Θ
- $q(\hat{\boldsymbol{\theta}}_{LS}; \mathbf{S}) \rightarrow \chi^2_{p^*-q,\alpha}$ $\qquad (p^* = \frac{1}{2}p(p+1))$

Iteratively Reweighted Least Squares Estimator

$\hat{\boldsymbol{\theta}}^{(i)}_{IRLS}$ minimizes

$$w(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{i-1}_{IRLS}, \mathbf{S}) = (\text{vech } \mathbf{S} - \text{vech } \boldsymbol{\Sigma}[\boldsymbol{\theta}])' \left(\mathbf{V}[\hat{\boldsymbol{\theta}}^{(i-1)}_{IRLS}] \right)^{-1} (\text{vech } \mathbf{S} - \text{vech } \boldsymbol{\Sigma}[\boldsymbol{\theta}])$$

- Iterative procedure converges to $\hat{\theta}_{ML}$ (approximately...there are numerical/computational issues)
- $w(\hat{\boldsymbol{\theta}}_{IRLS}; \hat{\boldsymbol{\theta}}_{IRLS}, \mathbf{S}) \rightarrow \chi^2_{p^*-q,\alpha}$

Goodness-of-fit measures for testing H_0 vs. H_1

H_0 : the hypothesized model is correct.

H_1 : the hypothesized model is incorrect (more factors are needed).

- $\ell(\hat{\boldsymbol{\theta}}_{ML}; \mathbf{S}) \rightarrow \chi^2_{\frac{1}{2}p(p+1)-q}$
- Bentler's CFI: scaled reduction of lack of fit when using the specified model instead of a baseline model such as the independence model:

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

$$\text{CFI} = 1 - \frac{\max(\chi^2_M - df_M, 0)}{\max(\chi^2_B - df_B, \chi^2_M - df_M, 0)}$$

Hu and Bentler (1999) recommend that CFI values greater than 0.95 indicate a good fit, although other authors argue that values greater than 0.90 are acceptable.

- RMSEA: excess lack of fit associated with the hypothesized model

$$\text{RMSEA} = \sqrt{\frac{1}{n-1}\frac{\max(\chi^2_M - df_M, 0)}{df_M}}$$

Hu and Bentler (1999) recommend that an RMSEA value less than 0.06 indicates a good fit.

- SRMR: quantifies the difference between the sample covariance matrix for the data ($\mathbf{S}$) and the model-constructed estimates of the covariance matrix $[\mathbf{\Sigma}(\hat{\boldsymbol{\theta}})]$

$$\text{SRMR} = \sqrt{\frac{2}{p(p+1)}\sum_{i=1}^{p}\sum_{j=1}^{i}\frac{(s_{ij}-\sigma_{ij})^2}{s_{ii}s_{jj}}}$$

where s_{ij} and σ_{ij} are the (i,j) elements of $\mathbf{S}$ and $\mathbf{\Sigma}(\hat{\boldsymbol{\theta}})$, respectively. Hu and Bentler (1999) recommend that SRMR values less than 0.08 indicate a good fit, although other authors argue that values less than 0.10 imply an adequate fit of the model.

Inference for $\boldsymbol{\theta}$

Regardless of distribution of $\mathbf{x}$, define $\underset{p^* \times p^*}{\mathbf{\Gamma}} = \mathrm{var}\{\mathrm{vech}\ \mathbf{S}\}$

- Note

$$\mathrm{vech}\ \mathbf{S} = \frac{1}{n-1} \sum_{t=1}^{n} \underbrace{\mathrm{vech}\ \ ((\mathbf{x}_t - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})')}_{\underset{p^* \times 1}{\mathbf{z}_t}}$$

$$= \frac{1}{n-1} \sum_{t=1}^{n} \underset{p^* \times 1}{\mathbf{z}_t}$$

and $\hat{\mathbf{\Gamma}} = \mathrm{var}\{\frac{1}{n-1} \sum_{t=1}^{n} \mathbf{z}_t\} = \frac{1}{(n-1)^2} \sum_{t=1}^{n} (\mathbf{z}_t - \bar{\mathbf{z}})(\mathbf{z}_t - \bar{\mathbf{z}})'$

When $\hat{\boldsymbol{\theta}}$ is $\hat{\boldsymbol{\theta}}_{LS}$ or $\hat{\boldsymbol{\theta}}_{ML}$,

$$\widehat{\mathrm{var}}\{\hat{\boldsymbol{\theta}}\} = (\hat{\mathbf{F}}'(\mathbf{V}[\hat{\boldsymbol{\theta}}])^{-1}\hat{\mathbf{F}})^{-1}\hat{\mathbf{F}}'(\mathbf{V}[\hat{\boldsymbol{\theta}}])^{-1}\hat{\mathbf{\Gamma}}(\mathbf{V}[\hat{\boldsymbol{\theta}}])^{-1}\hat{\mathbf{F}}(\hat{\mathbf{F}}'(\mathbf{V}[\hat{\boldsymbol{\theta}}])^{-1}\hat{\mathbf{F}})^{-1}$$

When the null model holds and $\mathbf{x}_t \sim N_p$,

$$\hat{\boldsymbol{\Gamma}} = \mathbf{V}[\hat{\boldsymbol{\theta}}]$$

and

$$\widehat{\text{var}}\{\hat{\boldsymbol{\theta}}\} = (\hat{\mathbf{F}}'(\mathbf{V}[\hat{\boldsymbol{\theta}}])^{-1}\hat{\mathbf{F}})^{-1}$$

where $\underset{p^* \times q}{\hat{\mathbf{F}}} = \frac{\partial \text{vech } \boldsymbol{\Sigma}[\boldsymbol{\theta}]}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ and q is the number of parameters in $\boldsymbol{\theta}$

Note: As previously mentioned, A & A (1988, 1990) show that as $n \to \infty$ (and innocuous identification conditions holding) and <u>any</u> distribution for $\mathbf{x}$

- $\hat{\boldsymbol{\lambda}}_{LS}$ and $\hat{\boldsymbol{\lambda}}_{ML} \to \boldsymbol{\lambda}$ and $\widehat{\text{var}}\{\hat{\boldsymbol{\lambda}}_{LS}\}$ and $\widehat{\text{var}}\{\hat{\boldsymbol{\lambda}}_{ML}\}$ still unbiased estimates
- $\ell(\hat{\boldsymbol{\theta}}_{ML}; \mathbf{S}) \to \chi^2_{p^*-q}$
  $q(\hat{\boldsymbol{\theta}}_{LS}; \mathbf{S}) \to \chi^2_{p^*-q}$
  $w(\hat{\boldsymbol{\theta}}_{IRLS}; \hat{\boldsymbol{\theta}}_{IRLS}, \mathbf{S}) \to \chi^2_{p^*-q}$

If additionally, $\mathbf{e}$ (errors) are N_p,

- $\hat{\boldsymbol{\psi}}_{LS}$ and $\hat{\boldsymbol{\psi}}_{ML} \to \boldsymbol{\psi}$
  and $\widehat{\text{var}}\{\hat{\boldsymbol{\psi}}_{LS}\}$ and $\widehat{\text{var}}\{\hat{\boldsymbol{\psi}}_{ML}\}$ still unbiased estimates

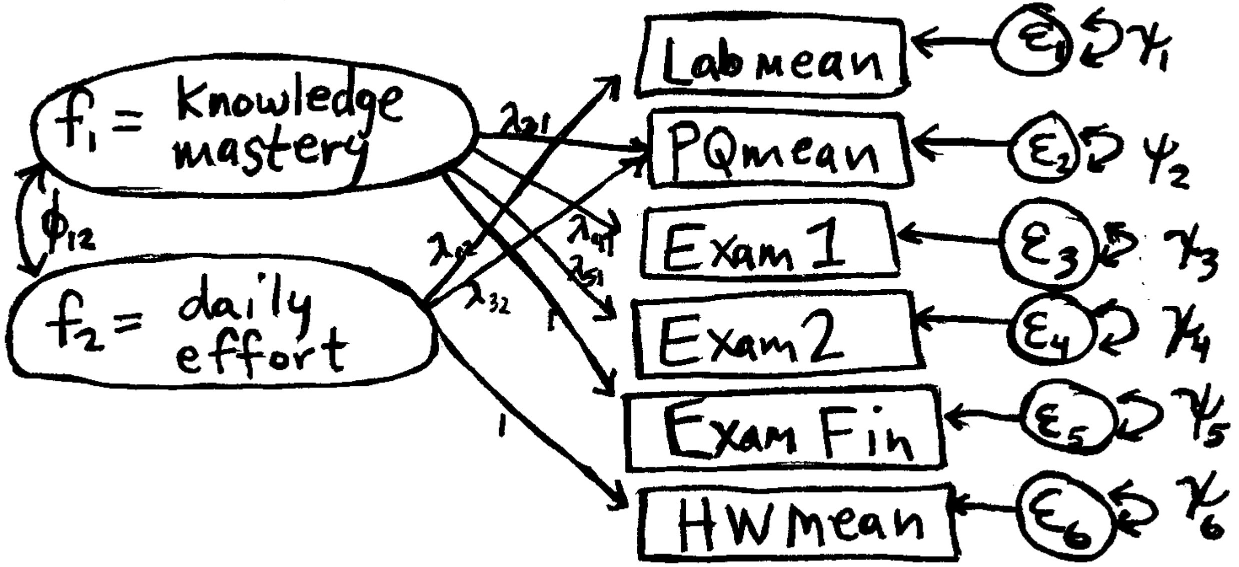
<u>Proc Calis</u>

(see SectIVB.sas)

- Use method = ml (to get $\hat{\boldsymbol{\theta}}_{ML}$)
  or method = gls (to get $\hat{\boldsymbol{\theta}}_{IRLS} \cong \hat{\theta}_{ML}$)
- Use LINEQS and write all equations
- Use STD to define variances of factors and errors
- Use COV to define covariances of factors (generally errors uncorrelated)
- Use BOUNDS to constrain estimates (such as var$\{e_1\} \geq 0$)

$\boxed{\text{ex}}$ Stat 2301 Grades

I think we have two factors (daily effort and knowledge mastery). Specifically, I want to test my theory:

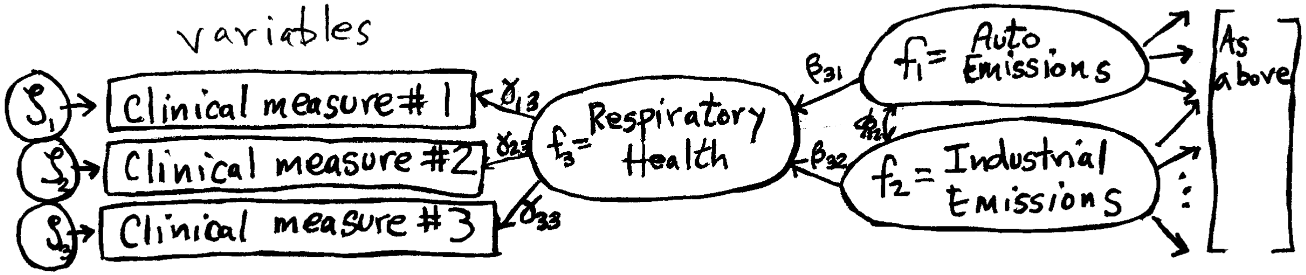


- Follow-up (pre-cursor to SEM): Does the Effort factor influence the Knowledge Mastery factor?
- Follow-up (pre-cursor to SEM): Is there a common factor that influences both Effort and Knowledge Mastery?

IV.C Structural Equation Modeling (SEM)

[Sources: *Latent Variable Models and Factor Analysis: A Unified Approach* by Bartholomew, Knott, and Moustaki; *Principles and Practice of Structural Equation Modeling* by Kline; *Structural Equations with Latent Variables* by Bollen]

- As noted earlier, a latent variable model is generally called a *structural equation model* (SEM) when relationships between latent factors are being considered.



- Common software: Mplus (standalone, expensive), AMOS (part of SPSS system), SAS (Proc Calis), R (lavaan package is good, but comparitively limited)

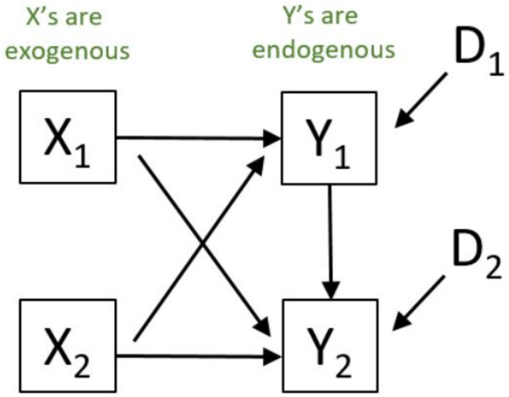
- Much of the terminology, motivation, and software used for CFA extends to SEM

Concepts Important to SEM

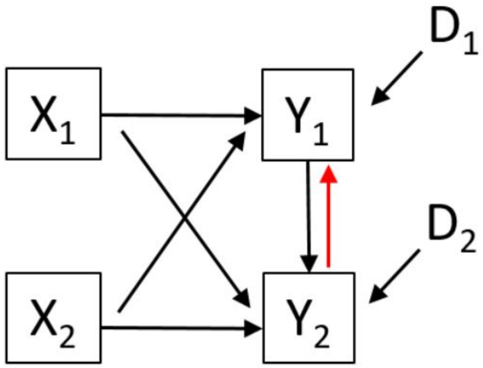
- Sample Size Rule of Thumb
  - $n/q > 20$: ideal
  - $n/q > 10$: livable
- Variable types
  - exogenous: predictor variable; “causes” of exogenous variable are unknown (“of external origin”)
  - endogenous: the “effect” of another variable (“of internal origin”); can also be the “cause” of another variable

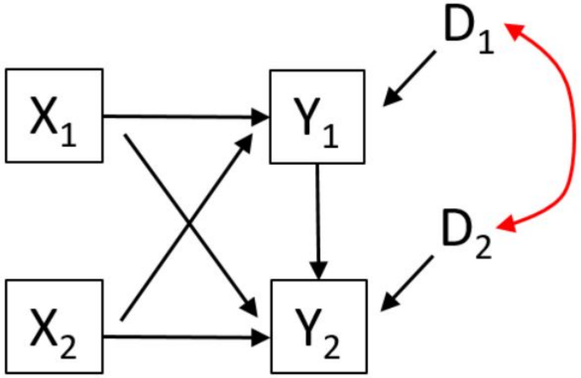
- Causality (the Holy Grail)
  - If we want to argue for *causality* in $X \rightarrow Y$, all the following must be met:
    1. X precedes Y in time
    2. the presumed unidirectionality $(X \rightarrow Y)$ is plausible, while the alternatives $(X \leftarrow Y$ and $X \rightleftharpoons Y)$ are not
    3. the relation does not disappear when external variables such as common causes of both X and Y are held constant

- Recursive vs. Nonrecursive Models
  - Recursive models: unidirectional effects and uncorrelated disturbances/errors; could be fit with regression techniques



  - Nonrecursive models: bidirectional effects, feedback loops, and/or correlated disturbances/errors; cannot be fit with regression techniques





Figures adapted from Figure 5.1 of Kline (1998).

- Direct and Indirect Effects
  - In the path diagram below, the variable Achievement acts as both a predictor (“cause”) and a response (“effect”); such variables are called *mediator* variables

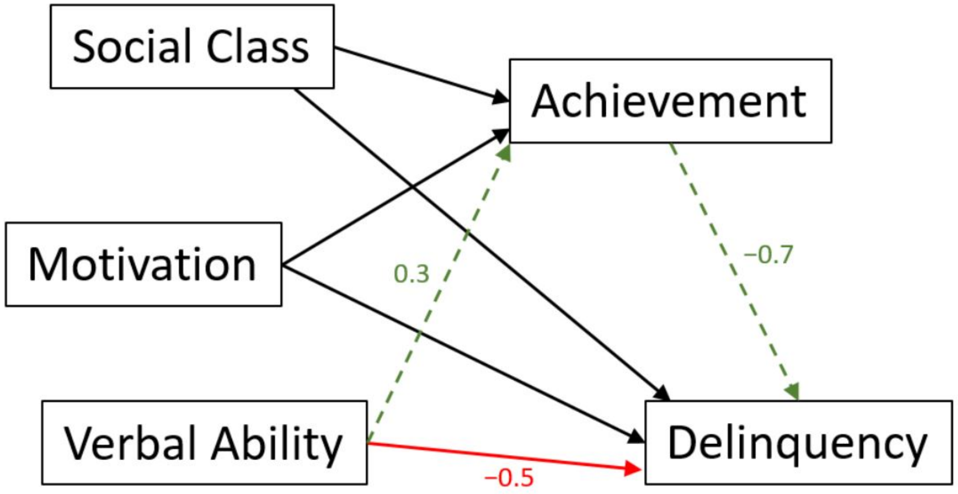
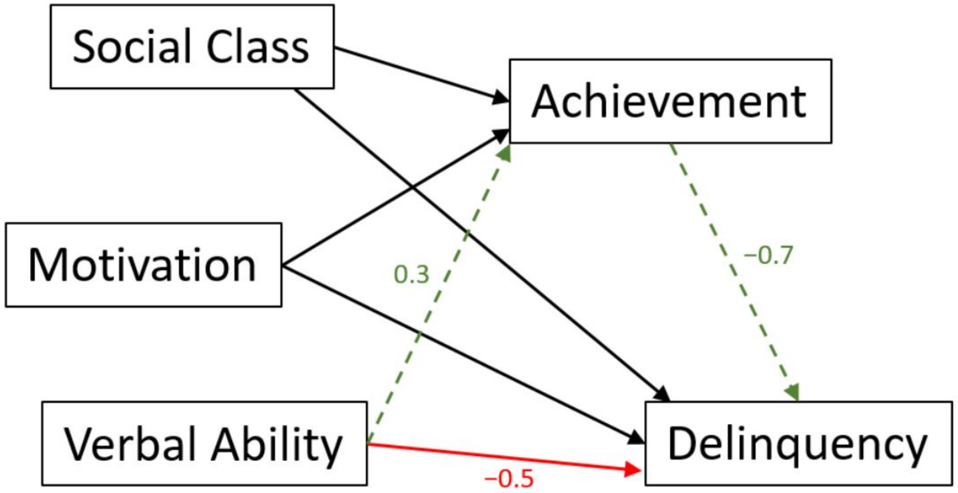


Figure adapted from Figure 3.1 of Kline (1998).

  - When mediator variables are included in an SEM, we are generally interested in both *direct* and *indirect* effects.



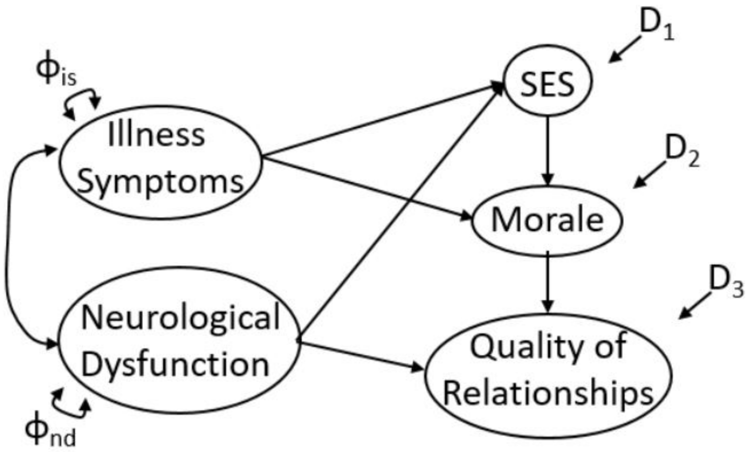
– If the estimated coefficients given above are standardized coefficients, then:

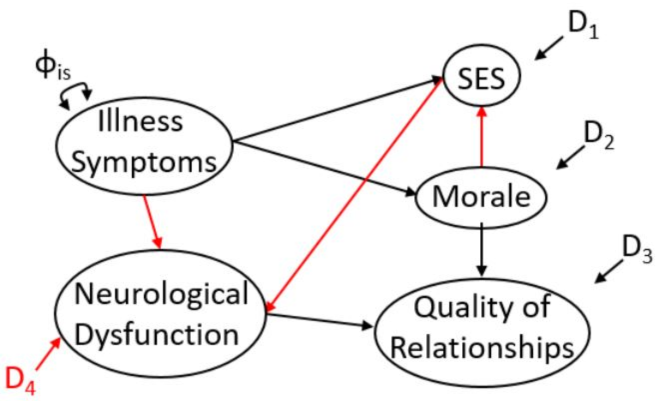
  ∗ Direct effect of Verbal Ability on Delinquency is -0.5 (i.e., a 1 standard deviation increase in Verbal Ability has a direct effect of a 0.5 standard deviation DECREASE in Delinquency)

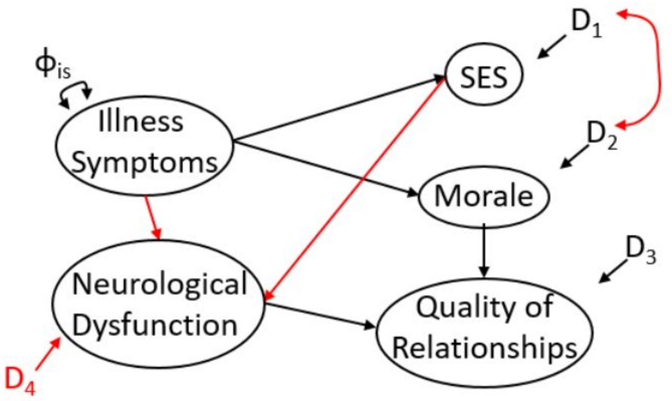
  ∗ Indirect effect of Verbal Ability on Delinquency is (0.3)(-0.7) = -0.21

  ∗ Total effect of Verbal Ability on Delinquency is -0.5 + (-0.21) = -0.71 (i.e., a 1 sd increase in Verbal Ability has a total effect of a 0.71 sd DECREASE in Delinquency)

- Equivalent Models
  - After fitting a “final” model, it is important to consider potentially equivalent models before making strong assertions about your hypothesis
  - Three models for cardiac surgery patients that are equivalent in terms of fit:

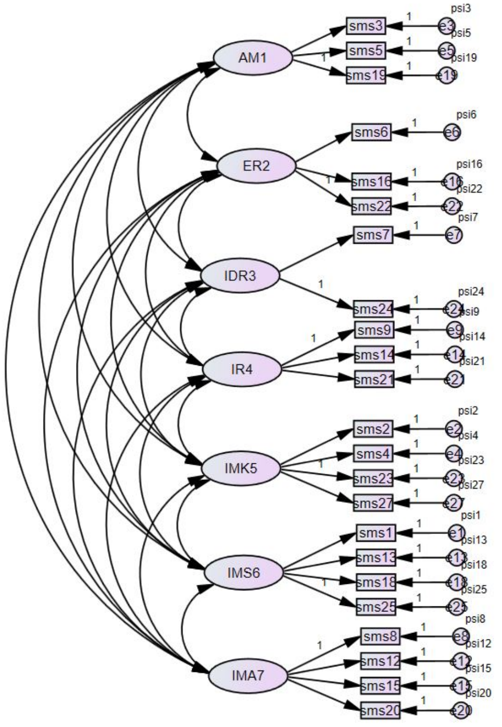






Figures adapted from Figure 5.5 of Kline (1998). (Manifest variables omitted from diagram)

Example: Motivation for P.E. class in middle school (AM=amotivation, ER=external motivation,…,IMA=intrinsic motivation to accomplish)



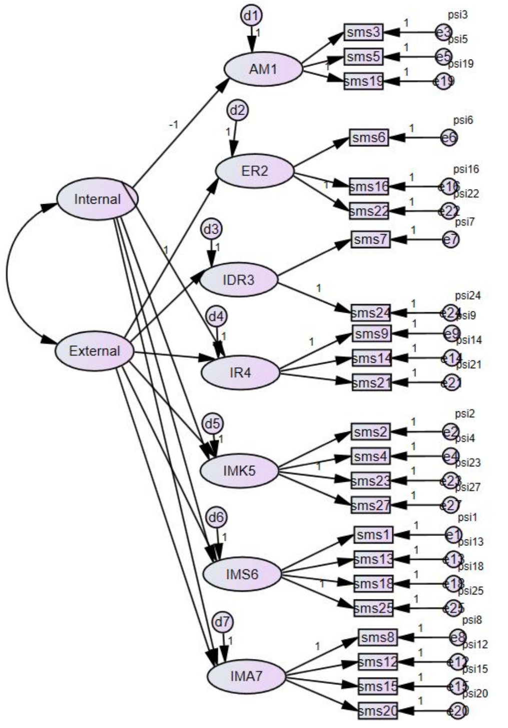
First-order CFA

$\chi^2 = 619.796$ ($df = 209$)

$\chi^2/df = 2.966$ (RoT: < 3)

CFI $= 0.958$

RMSEA $= 0.039$



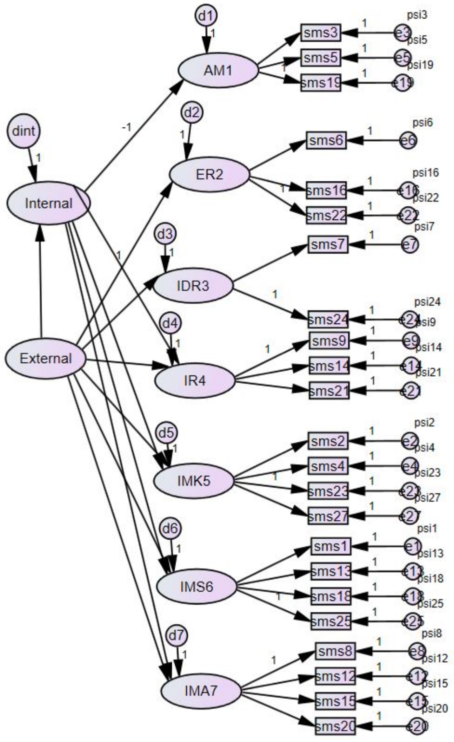
Second-order CFA #1*

$\chi^2 = 629.284$ ($df = 218$)

$\chi^2/df = 2.887$ (RoT: < 3)

CFI $= 0.958$

RMSEA $= 0.038$



Second-order CFA #2*

$\chi^2 = 629.284$ ($df = 218$)

$\chi^2/df = 2.887$ (RoT: < 3)

CFI $= 0.958$

RMSEA $= 0.038$