Covariance structure and factor models

PSYC/EPSY/SOC/STAT 588 Spring 2011

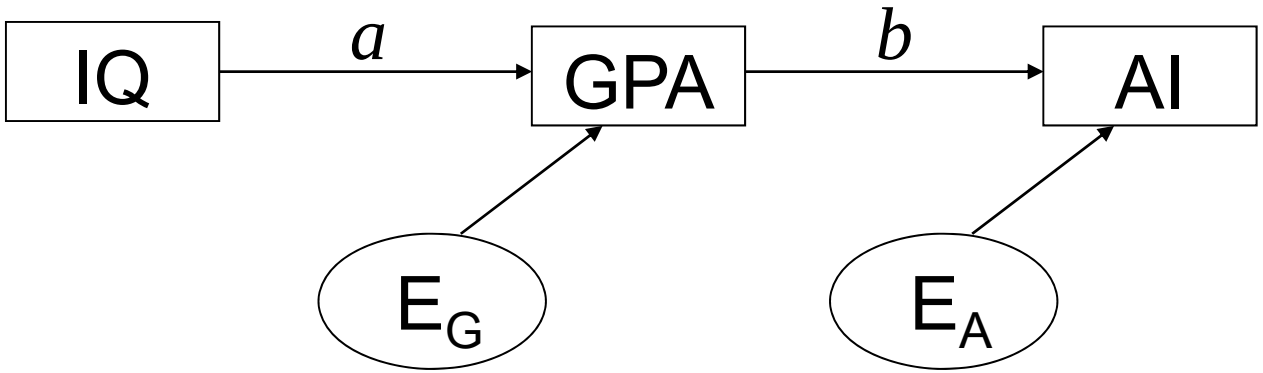
Instructor: Sungjin Hong
Office: Psych 429 (244-8296, hongsj AT illinois DOT edu)
Class meetings: TR, 3:00-4:50, room 29 (219A/289 to access AMOS; 289 also reserved for Wed 3-5pm)
Office hours: W, 2:00-3:00 or by appointment
Course website: https://compass.illinois.edu

Historical backgrounds

1. Path analysis
2. Factor analysis (measurement model in SEM terms)
3. General estimation procedure

1. Path analysis

- To parsimoniously describe relationships between a set of observed variables with directed paths by which we wish to represent **causal** influences of one variable to another



$$\mathrm{GPA} = a\,\mathrm{IQ} + \mathrm{E_G}, \qquad \mathrm{AI} = b\,\mathrm{GPA} + \mathrm{E_A}$$

$$\mathrm{cov}\left(\mathrm{GPA}, \mathrm{AI}\right) = a^2 b\ \mathrm{var}\left(\mathrm{IQ}\right) + b\ \mathrm{var}\left(\mathrm{E_G}\right)$$

2. Factor analysis

- Scientifically useful variables are often not directly observable (measurable) --- e.g., IQ

- The FA model defines the unobservable (latent) variables such that they **linearly** relate to (influence or determine) a set of correlated, yet distinctive, observable variables (indicators)

$$X_1 = a_1 + b_1\,\mathrm{IQ} + E_1$$
$$\vdots \qquad \vdots \qquad \vdots$$
$$X_q = a_q + b_q\,\mathrm{IQ} + E_q$$

$cf.$ $\mathrm{IQ} = c_1 X_1 + c_2 X_2 + \cdots + c_q X_q + E_{\mathrm{IQ}}$

3. General estimation procedure

- Relationships among observed and unobserved variables are quantitatively defined --- “parameterized”, “modeled”
- A number is derived to indicate discrepancy between the modeled structure and the data with some side conditions (assumptions) --- model fit

$$F = f(\text{data, parameters, combining rule})$$

$$F = \text{tr}\left[\left(\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta})\right)^2\right]: \quad \text{ULS}$$

- Distributional characteristics of F (as a random variable) are established according to a set of parametric assumptions (e.g., multivariate normal data), which allows for inference on statistical reliability of some hypotheses, e.g., $F = 0$ or $F_1 - F_2 = 0$

SEM related myths

- Confirmatory vs. exploratory:
  - a priori model specified
  - statistical test available (parametric or nonparametric)
  - what software used (e.g., AMOS or SPSS)
- An SEM analysis proves something:

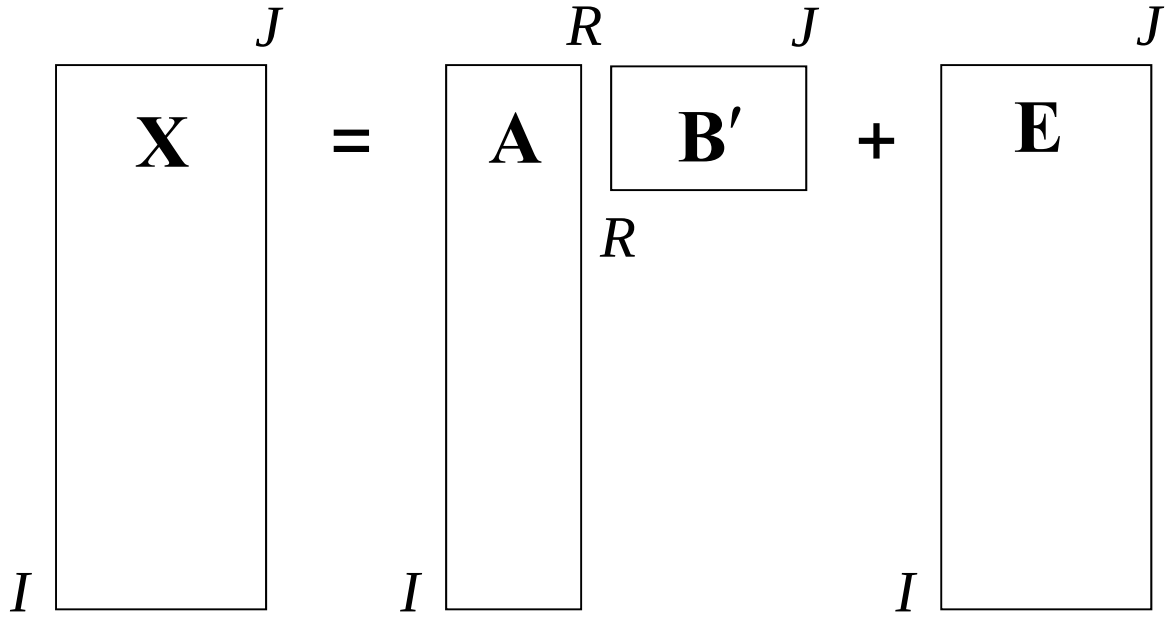
  “I got a dataset from somewhere and model fit from an AMOS output is acceptable. Now, I have proved my hypothesis (the model) and so I am entitled a Ph.D.”
- Hopefully, we will unveil many more in this course

Notational conventions

- Data matrix tall or wide?

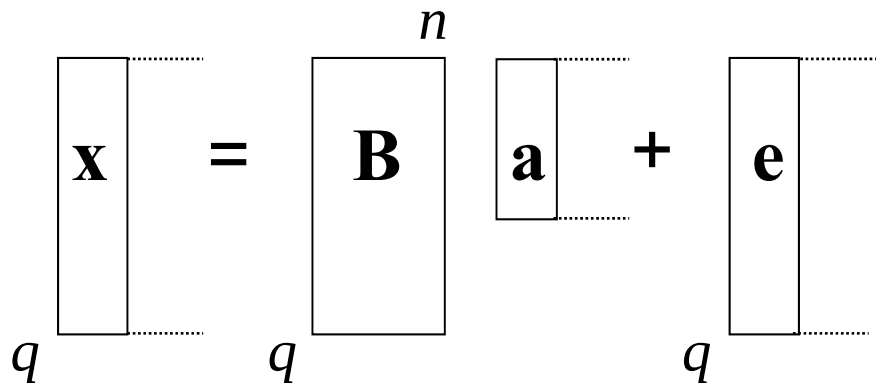
  The principal component model written for tall data matrix --- "mathematical" notation

$$\mathbf{X} = \mathbf{A}\mathbf{B}' + \mathbf{E}$$



The principal component model written for a set of random variables (wide data matrix) --- "statistical" notation

$$\mathbf{x} = \mathbf{B}\mathbf{a} + \mathbf{e}$$



- Variables in the data vector $\mathbf{x}$ are considered as random quantities; not interested in their realized values themselves but their distributions --- equivalent to the data matrix represented horizontally
- The "LISREL" notation follows this convention as is most common for SEM; and we'll do so in this course

Sample covariance matrix

- Given data matrix $\mathbf{X}\ (N \times q)$, deviation matrix (column mean centered) is written as:

$$\mathbf{Z} = \mathbf{LX}, \quad \mathbf{L} = \mathbf{I} - \tfrac{1}{N}\mathbf{1}\mathbf{1}' \text{ : idempotent centering matrix}$$

And unbiased sample covariance matrix:

$$\mathbf{S} = \tfrac{1}{N-1}\mathbf{Z}'\mathbf{Z} \quad \text{or} \quad \tfrac{1}{N-1}\mathbf{z}\mathbf{z}' \quad \text{with the horizontal notation}$$

$$\begin{bmatrix} c_{11}, c_{12} \cdots \\ c_{21}, c_{22} \cdots \\ \vdots \quad \vdots \quad \ddots \end{bmatrix} = \frac{1}{N-1} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \end{bmatrix} \begin{bmatrix} z_1, z_2, \cdots \end{bmatrix}$$