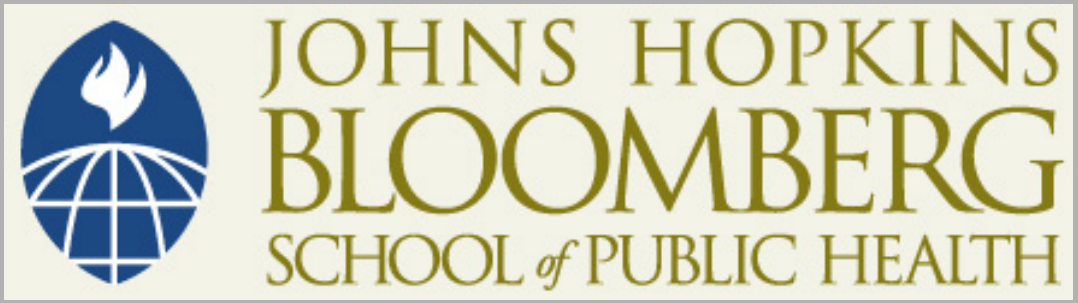
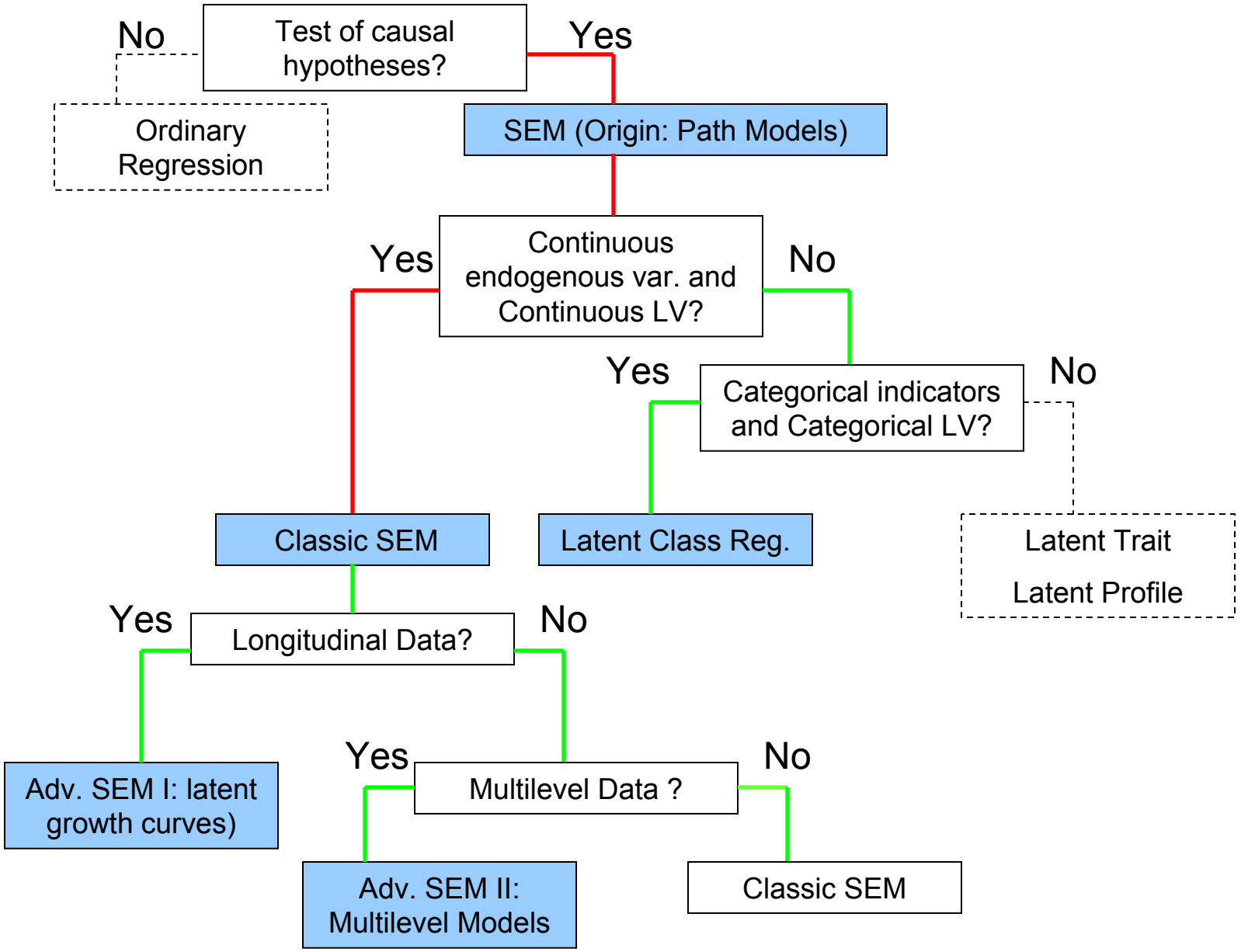


Introduction to Structural Equations with Latent Variables

Statistics for Psychosocial Research II:
Structural Models

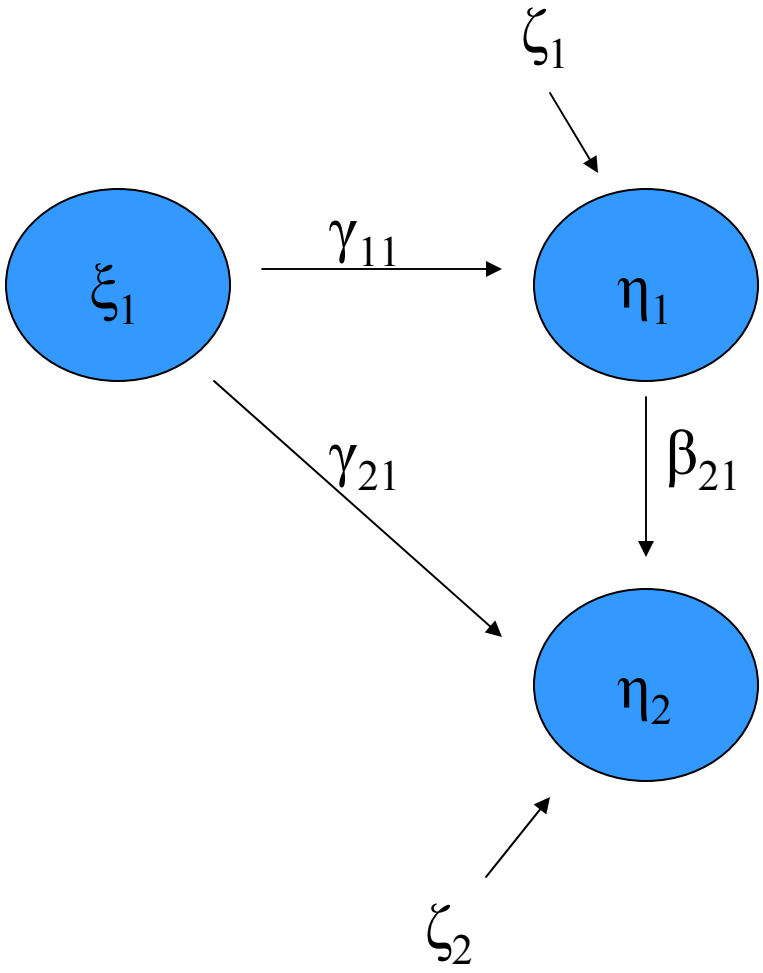
Qian-Li Xue


Test of causal hypotheses?
No
Ordinary Regression
Yes
SEM (Origin: Path Models)
Continuous endogenous var. and Continuous LV?
Yes
No
Categorical indicators and Categorical LV?
Yes
No
Classic SEM
Latent Class Reg.
Latent Trait
Latent Profile
Longitudinal Data?
Yes
No
Adv. SEM I: latent growth curves)
Multilevel Data ?
Yes
No
Adv. SEM II: Multilevel Models
Classic SEM

Adding Latent Variables to the Model

- So far…we've only included **observed** variables in our "path models"
- Extension to latent variables:
  - need to add in a "measurement piece"
  - how do we "define" the latent variable (think factor analysis)
  - more complicated to look at, but same general principles apply

Path Diagram



Notation for Latent Variable Model

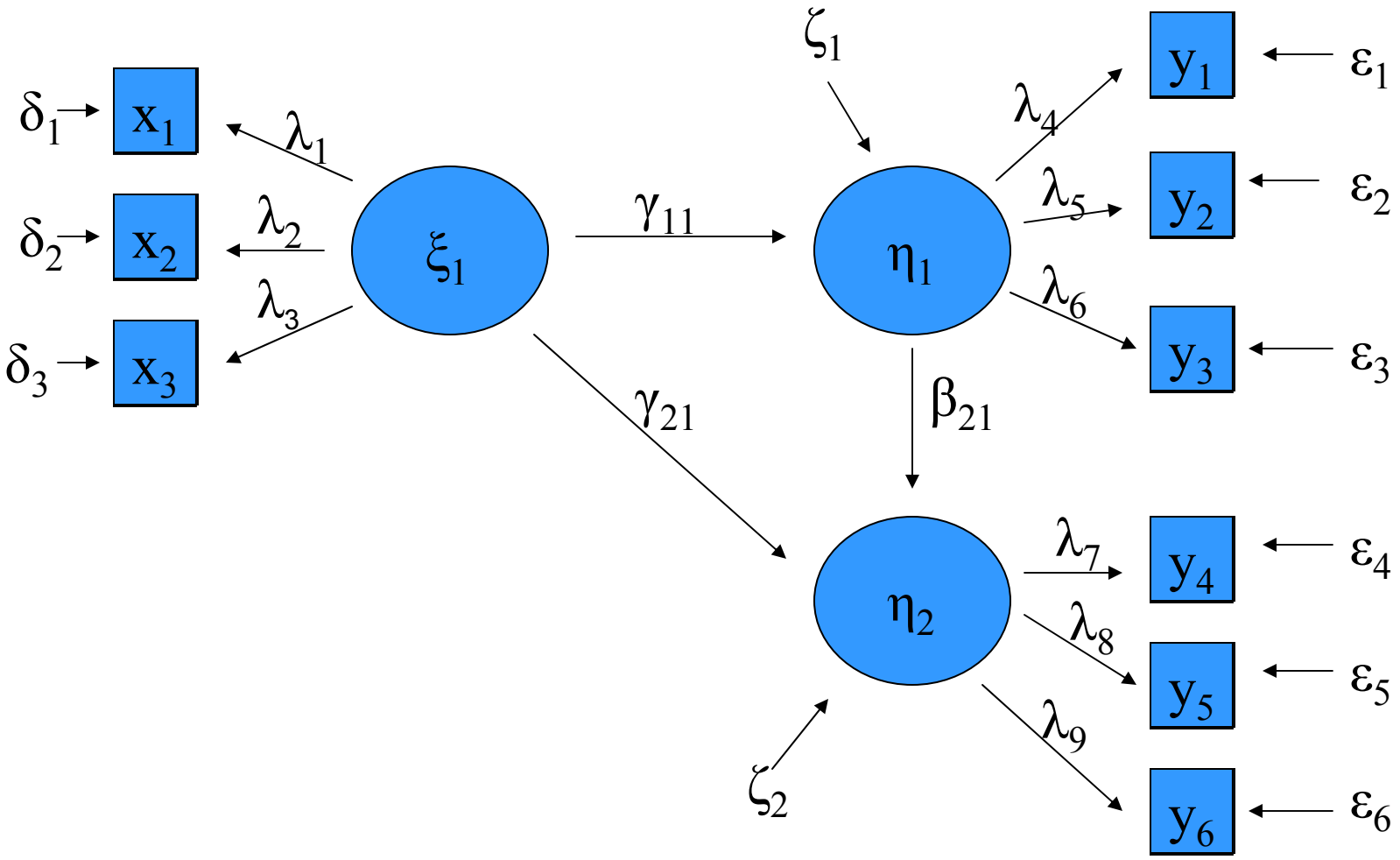
- η = latent endogenous variable (eta)
- ξ = latent exogenous variable (ksi, pronounced “kah-see”)
- ζ = latent error (zeta)
- β = coefficient on latent endogenous variable (beta)
- γ = coefficient on latent exogenous variable (gamma)

$$\eta_1 = \gamma_{11}\xi_1 + \zeta_1$$
$$\eta_2 = \beta_{21}\eta_1 + \gamma_{21}\xi_1 + \zeta_2$$

e.g. ξ_1 is childhood home environment, η_1 is social support, and η_2 is cancer coping ability

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta, \mathrm{Cov}(\xi) = \Phi, \mathrm{Cov}(\zeta) = \Psi$$

Path Diagram



Notation for Measurement Model

y = observed indicator of η

x = observed indicator of ξ

ε = measurement error on y

δ = measurement error on x

λ_x = coefficient relating x to ξ

λ_y = coefficient relating y to η

$$x_1 = \lambda_1 \xi_1 + \delta_1 \qquad y_1 = \lambda_4 \eta_1 + \varepsilon_1 \qquad y_4 = \lambda_7 \eta_2 + \varepsilon_4$$

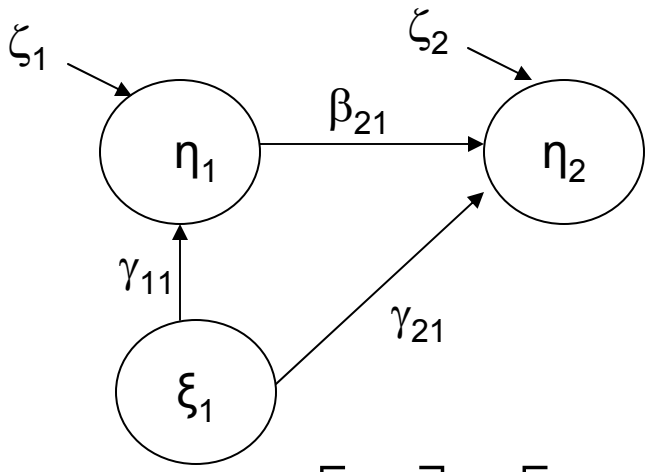
$$x_2 = \lambda_2 \xi_1 + \delta_2 \qquad y_2 = \lambda_5 \eta_1 + \varepsilon_2 \qquad y_5 = \lambda_8 \eta_2 + \varepsilon_5$$

$$x_3 = \lambda_3 \xi_1 + \delta_3 \qquad y_3 = \lambda_6 \eta_1 + \varepsilon_3 \qquad y_6 = \lambda_9 \eta_2 + \varepsilon_6$$

$$y = \Lambda_y \eta + \varepsilon, \mathrm{Cov}(\varepsilon) = \Theta_\varepsilon$$

$$x = \Lambda_x \xi + \delta, \mathrm{Cov}(\delta) = \Theta_\delta$$

Example: Model Specification

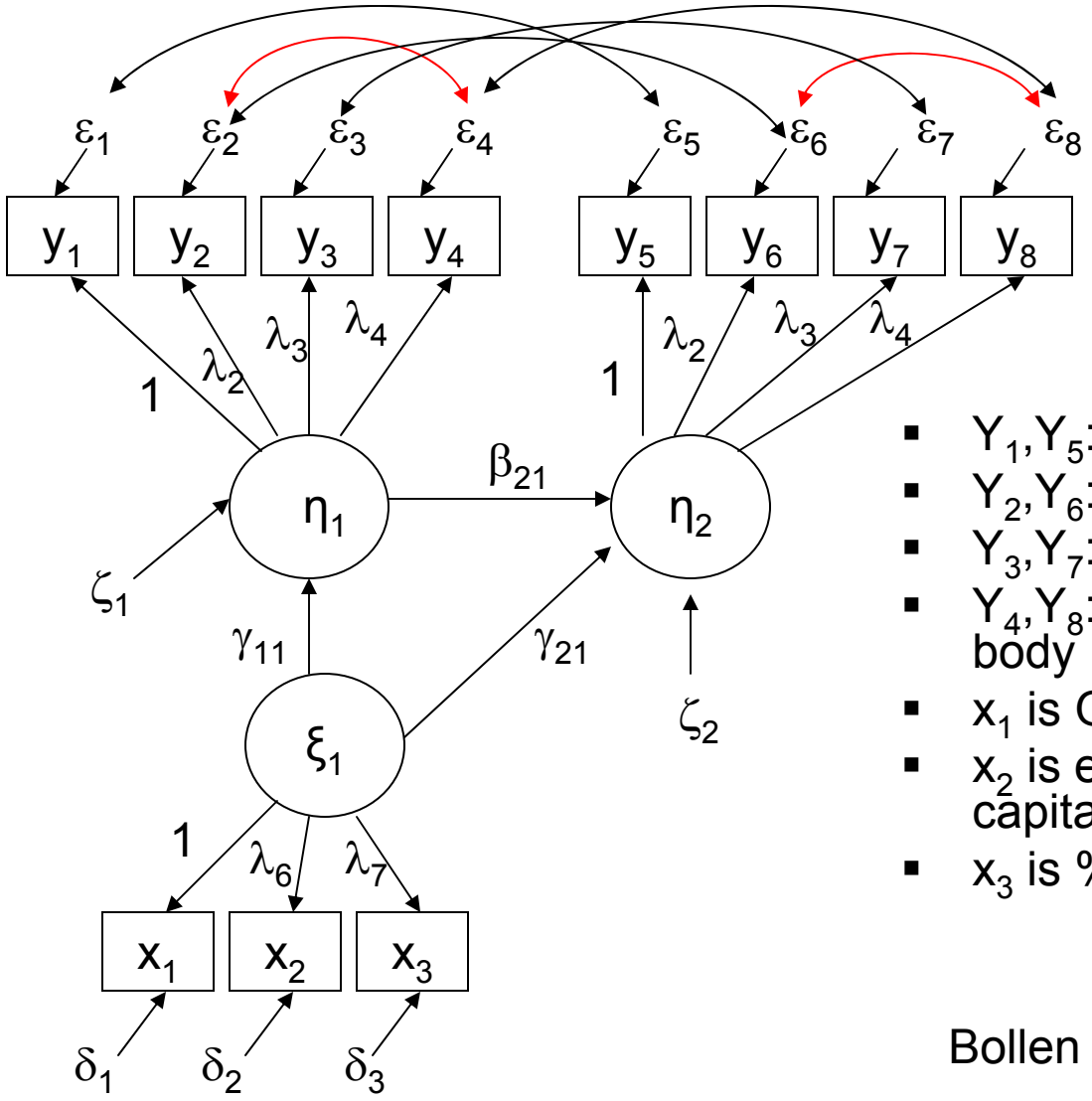


- η_2 is democracy in 1965
- η_1 is democracy in 1960
- ξ_1 is industrialization in 1960

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \gamma_{11} \\ \gamma_{21} \end{bmatrix} [\xi_1] + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}$$

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta, \mathrm{Cov}(\xi) = \Phi, \mathrm{Cov}(\zeta) = \Psi$$

Example: Model Specification



- Y_1, Y_5: freedom of press
- Y_2, Y_6: freedom of group opposition
- Y_3, Y_7: fairness of election
- Y_4, Y_8: effectiveness of legislative body
- x_1 is GNP per capita
- x_2 is energy consumption per capita
- x_3 is % labor force

Bollen pp322-323

Model Estimation

- In multiple regression, estimation is based on individual cases via LS, i.e. minimization of

$$\sum_n \left(\hat{Y} - Y\right)^2$$

- In SEM, estimation is based on covariances

  If Model is correct and θ is known

$$\Sigma = \Sigma(\theta)$$

  Where Σ is the population covariance matrix of observed variables and $\Sigma(\theta)$ is the covariance matrix written as a function of θ

Model Estimation

- Regression analysis, confirmatory factor analysis are special cases
- E.g. $y=\gamma x+\zeta$, $\zeta \perp \gamma$, $E(\zeta)=0$

$$\begin{bmatrix} VAR(y) & \\ COV(x,y) & VAR(x) \end{bmatrix} = \begin{bmatrix} \gamma^2 VAR(x)+VAR(\zeta) & \\ \gamma VAR(x) & VAR(x) \end{bmatrix}$$

$$COV(x,y)=\gamma VAR(x) \Rightarrow \gamma = \frac{COV(x,y)}{VAR(x)}$$

- Does γ look familiar?

  Remember in SLR, $\beta=(x'x)^{-1}x'y$

Model Estimation

- In reality, neither population covariances Σ nor the parameters θ are known
- What we have is sample estimate of Σ: S
- Goal: estimate θ based on S by choosing the estimates of θ such that $\hat{\Sigma}$ is as close to S as possible
- But how close is "close"?
- Define and minimize objective functions or called "fitting functions": $F(S, \hat{\Sigma})$

  e.g. $F(S,\hat{\Sigma}) = S - \hat{\Sigma}$

Common Fitting Functions

- **Maximum Likelihood (ML)**

  To minimize

  $F_{ML}=(1/2)tr(\{[S-\Sigma(\theta)]\ \Sigma^{-1}(\theta)\}^2)$

  or

  $F_{ML}=\log|\Sigma(\theta)|+tr(S\Sigma^{-1}(\theta))-\log|S|-(p+q)$

  - Explicit solutions of θ may not exist
  - Iterative numeric procedure is needed
  - Asymptotic properties of ML estimators:
    - Consistent, i.e. $\hat{\theta}\rightarrow\theta$ as $n\rightarrow\infty$
    - Efficient, i.e. smallest asymptotic variance
    - Asymptotic normality

Common Fitting Functions

- **Maximum Likelihood (ML)**

  Advantages
  - Scale invariant
    - $F(S,\Sigma(\theta))$ if scale invariance if $F(S,\Sigma(\theta)) = F(DSD,D\Sigma(\theta)D)$, where D is a diagonal matrix with positive elements
    - E.g. D consists of inverses of standard deviation of observed variables, DSD becomes a correlation matrix
    - More general, the value of F is the same for any change of scale (e.g. dollars to cents)
  - Scale free

    Knowing D, we can calculate $\hat{\theta}^*$ (based on transformed data) from $\hat{\theta}$ (based on non-transformed data) without actually rerunning the model
  - Test of overall model fit for overidentified model based on the fact: $(N-1)F_{ML}$ is a χ^2 distribution with $\frac{1}{2}(p+q)(p+q+1)-t$

  Disadvantage
  - Assumption of multivariate normality

Common Fitting Functions

- **Unweighted Least Squares (ULS)**
  To minimize
  $F_{ULS}=(1/2)tr[(S-\Sigma(\theta))^2]$
  - Analogous to OLS, minimize the sum of squares of each element in the residual matrix $(S-\Sigma(\theta))$
  - Give greater weights to off covariance terms than variance terms
  - Explicit solutions of θ may not exist
  - Iterative numeric procedure is needed
  - Advantages of ULS estimators:
    - Intuitive
    - Consistent, i.e. $\hat{\theta} \rightarrow \theta$ as $n \rightarrow \infty$
    - No distributional assumptions
  - Disadvantages
    - Not most efficient
    - Not scale invariant, not scale free

Common Fitting Functions

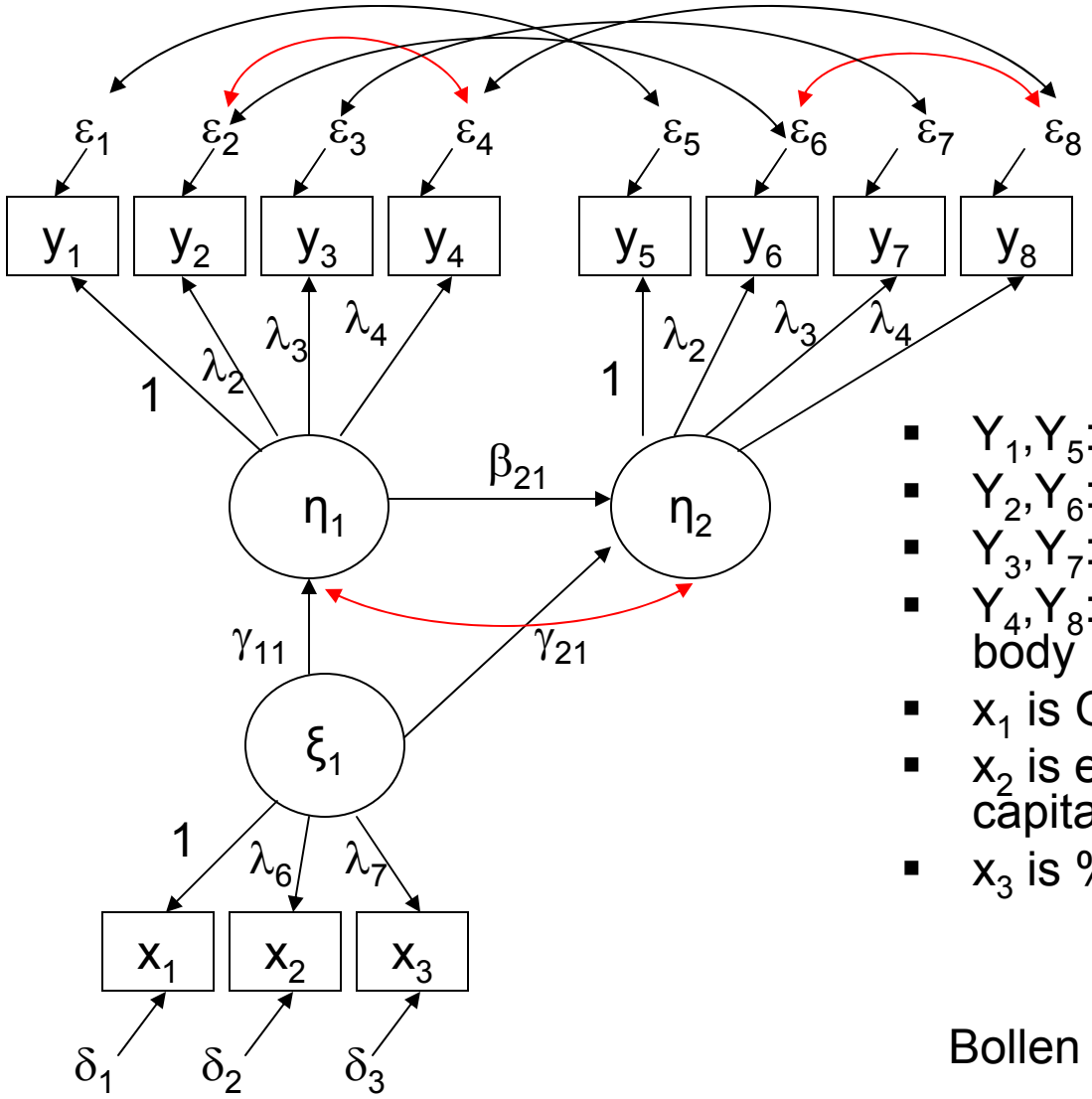
- **Generalized Least Squares (GLS)**

  To minimize

  $F_{GLS}=(1/2)tr(\{[S-\Sigma(\theta)]S^{-1}\}^2)$

  - Weights the elements of $(S- \Sigma(\theta))$ according to variances and covariances
  - F_{ULS} is a special case of F_{GLS} with $S^{-1}=I$
  - Advantages of ULS estimators:
    - Intuitive
    - Consistent, i.e. $\hat{\theta} \rightarrow \theta$ as $n\rightarrow\infty$
    - Asymptotic normality (availability of significance test)
    - Asymptotically efficient
    - Scale invariant and scale free
    - Test of overall model fit for overidentified model based on the fact: $(N-1)F_{GLS}$ is a χ^2 distribution with $\frac{1}{2}(p+q)(p+q+1)-t$
  - Disadvantages
    - Sensitive to “fat” or “thin” tails

Example: Model Specification



- Y_1, Y_5: freedom of press
- Y_2, Y_6: freedom of group opposition
- Y_3, Y_7: fairness of election
- Y_4, Y_8: effectiveness of legislative body
- x_1 is GNP per capita
- x_2 is energy consumption per capita
- x_3 is % labor force

Bollen pp322-323

Simple Case of SEM with Latent Variables: Confirmatory Factor Analysis

Recap of Basic Characteristics of Exploratory Factor Analysis (EFA)

- Most EFA extract orthogonal factors, which is “boring” to SEM users
- Distinction between common and unique variances
- EFA is underidentified (i.e. no unique solution)
  - Remember rotation? Equally good fit with different rotations!
- All measures are related to each factor

Confirmatory Factor Analysis (CFA)

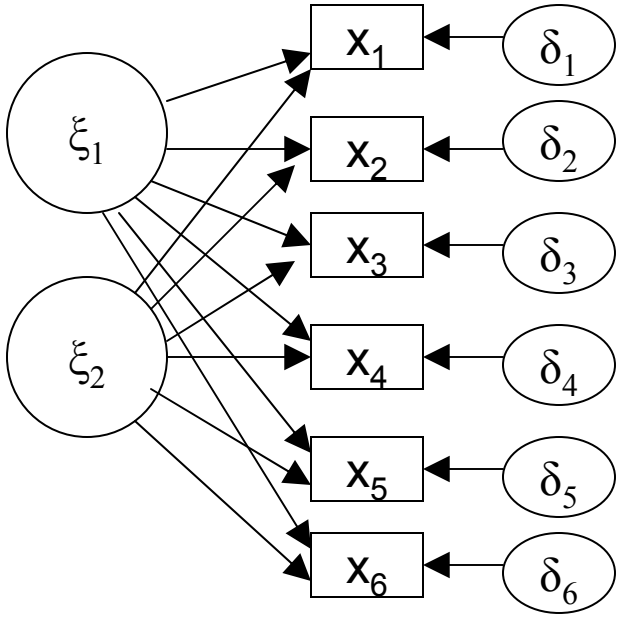
- Takes factor analysis a step further.
- We can “test” or “confirm” or “implement” a “highly constrained a priori structure that meets conditions of model identification”
- But be careful, a model can never be confirmed!!
- CFA model is constructed in advance
- number of latent variables (“factors”) is pre-set by analyst (not part of the modeling usually)
- Whether latent variable influences observed is specified
- Measurement errors may correlate
- Difference between CFA and the usual SEM:
  - SEM assumes causally interrelated latent variables
  - CFA assumes interrelated latent variables (i.e. exogenous)

Exploratory Factor Analysis

Two factor model:

$$x = \Lambda \xi + \delta$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ \lambda_{41} & \lambda_{42} \\ \lambda_{51} & \lambda_{52} \\ \lambda_{61} & \lambda_{62} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{bmatrix}$$

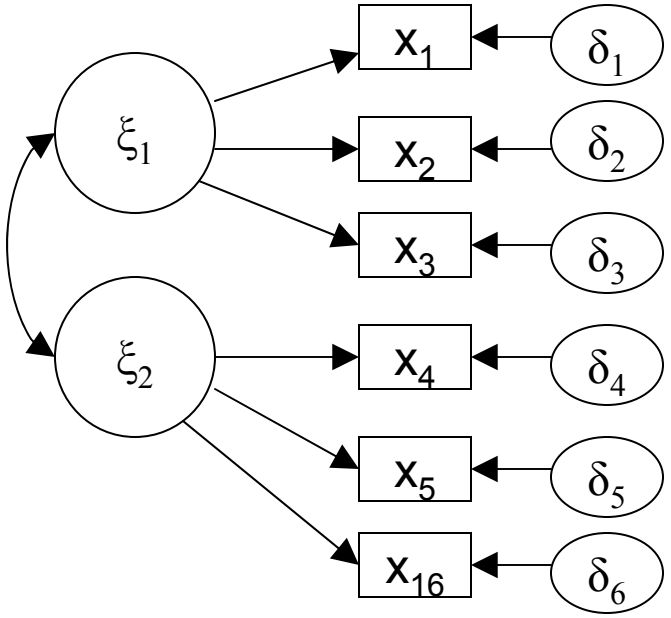


CFA Notation

Two factor model:

$$x = \Lambda \xi + \delta$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{bmatrix}$$



For the "matrix-challenged"

CFA

$$x_1 = \lambda_{11}\xi_1 + \delta_1$$
$$x_2 = \lambda_{21}\xi_1 + \delta_2$$
$$x_3 = \lambda_{31}\xi_1 + \delta_3$$
$$x_4 = \lambda_{42}\xi_2 + \delta_4$$
$$x_5 = \lambda_{52}\xi_2 + \delta_5$$
$$x_6 = \lambda_{62}\xi_2 + \delta_6$$
$$\text{cov}(\xi_1, \xi_2) = \varphi_{12}$$

EFA

$$x_1 = \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \delta_1$$
$$x_2 = \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \delta_2$$
$$x_3 = \lambda_{31}\xi_1 + \lambda_{32}\xi_2 + \delta_3$$
$$x_4 = \lambda_{41}\xi_1 + \lambda_{42}\xi_2 + \delta_4$$
$$x_5 = \lambda_{51}\xi_1 + \lambda_{52}\xi_2 + \delta_5$$
$$x_6 = \lambda_{61}\xi_1 + \lambda_{62}\xi_2 + \delta_6$$
$$\text{cov}(\xi_1, \xi_2) = 0$$

More Important Notation

- Φ (capital of φ): covariance matrix of factors

$$\Phi = \text{var}(\xi) = \begin{bmatrix} \varphi_{11} & \varphi_{12} \\ \varphi_{12} & \varphi_{22} \end{bmatrix} \qquad \text{var}(\xi_1) = \varphi_{11}$$
$$\text{cov}(\xi_1, \xi_2) = \varphi_{12}$$

- Ψ(capital of ψ): covariance matrix of errors

$$\Psi = \text{var}(\Delta) = \begin{bmatrix} \psi_{11} & & & & & \\ \psi_{21} & \psi_{22} & & & & \\ \psi_{31} & \psi_{32} & \psi_{33} & & & \\ \psi_{41} & \psi_{42} & \psi_{43} & \psi_{44} & & \\ \psi_{51} & \psi_{52} & \psi_{53} & \psi_{54} & \psi_{55} & \\ \psi_{61} & \psi_{62} & \psi_{63} & \psi_{64} & \psi_{65} & \psi_{66} \end{bmatrix}$$

$$\text{var}(\delta_1) = \psi_{11}$$
$$\text{cov}(\delta_1, \delta_2) = \psi_{12}$$

NOTE: usually, $\psi_{ij} = 0 \text{ if } i \neq j$

Model Estimation

- In our previous CFA example, we have six equations, and many more unknowns
- In this form, not enough information to uniquely solve for λ and the factor correlation
- What if we multiple both sides of $x = \Lambda \xi + \delta$ by **x'**

$$xx' = (\lambda\xi + \delta)(\lambda\xi + \delta)'$$

$$= (\lambda\xi)(\lambda\xi)' + (\lambda\xi)\delta' + \delta(\lambda\xi)' + \delta\delta'$$

because ξ and δ are independent

$$xx' = (\lambda\xi)(\lambda\xi)' + \delta\delta'$$

$$= \lambda\xi\xi'\lambda' + \delta\delta'$$

$$\Sigma = \lambda\Phi\lambda + \Psi = \Sigma(\Theta)$$

where Φ is the covariance matrix of factors ξ, and Ψ is error covariance matrix

Model Constraints

- Hallmark of CFA
- Purposes for setting constraints:
  - Test a priori theory
  - Ensure identifiability
  - Test reliability of measures

Model Constraints: Identifiability

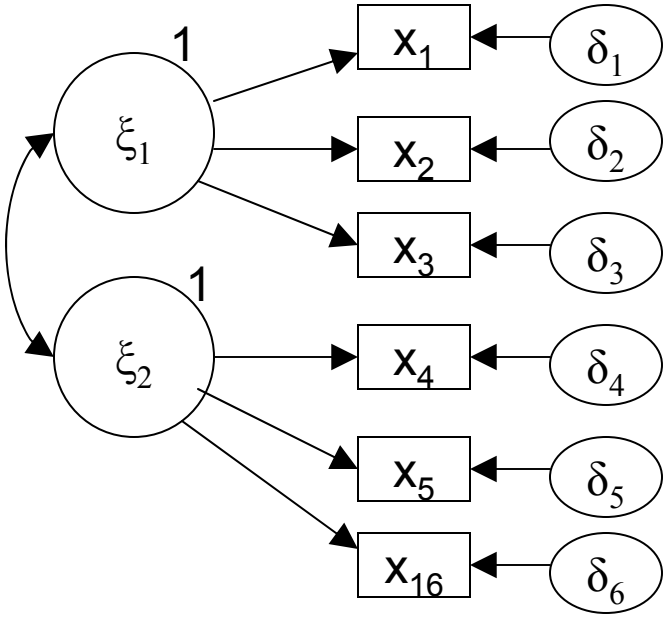
- Latent variables (LVs) need some constraints
- Because factors are unmeasured, their variances can take different values
- Recall EFA where we constrained factors:

$$F \sim N(0,1)$$

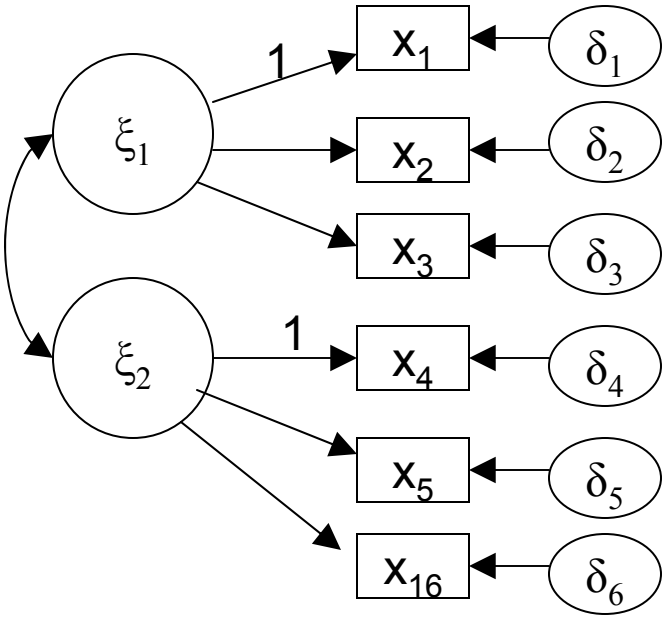
- Otherwise, model is not identifiable.
- Here we have two options:
  - Fix variance of latent variables (LV) to be 1 (or another constant)
  - Fix one path between LV and indicator

Necessary Constraints

Fix variances:



Fix path:



Model Parametrization

Fix variances:

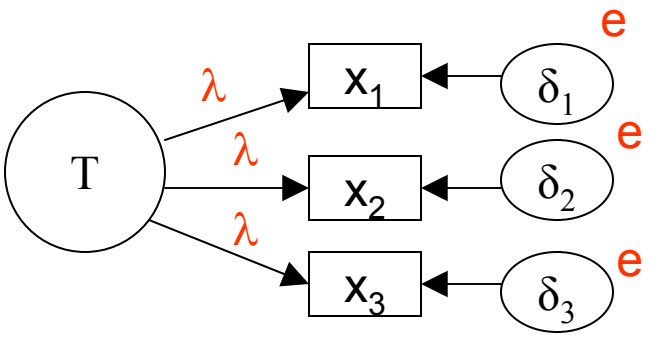
$$x_1 = \lambda_{11}\xi_1 + \delta_1$$
$$x_2 = \lambda_{21}\xi_1 + \delta_2$$
$$x_3 = \lambda_{31}\xi_1 + \delta_3$$
$$x_4 = \lambda_{42}\xi_2 + \delta_4$$
$$x_5 = \lambda_{52}\xi_2 + \delta_5$$
$$x_6 = \lambda_{62}\xi_2 + \delta_6$$
$$\mathrm{cov}(\xi_1, \xi_2) = \varphi_{12}$$
$$\mathrm{var}(\xi_1) = 1$$
$$\mathrm{var}(\xi_2) = 1$$

Fix path:

$$x_1 = \xi_1 + \delta_1$$
$$x_2 = \lambda_{21}\xi_1 + \delta_2$$
$$x_3 = \lambda_{31}\xi_1 + \delta_3$$
$$x_4 = \xi_2 + \delta_4$$
$$x_5 = \lambda_{52}\xi_2 + \delta_5$$
$$x_6 = \lambda_{62}\xi_2 + \delta_6$$
$$\mathrm{cov}(\xi_1, \xi_2) = \varphi_{12}$$
$$\mathrm{var}(\xi_1) = \varphi_{11}$$
$$\mathrm{var}(\xi_2) = \varphi_{22}$$

Model Constraints: Reliability Assessment

- Parallel measures
  - $T_{x1} = T_{x2}$ [= E(x)] (True scores are equal)
  - T affects x1 and x2 equally
  - Cov(δ_1, δ_2) = 0 (Errors not correlated)
  - Var (δ_1) = Var (δ_2) (Equal error variances)



Model Constraints: Reliability Assessment

- **Tau equivalent measures**
  - $T_{x1} = T_{x2}$
  - T affects x1 and x2 equally
  - Var (δ_1) ≠ Var (δ_2)
  - Note: for standardized measures, it makes no sense to constrain the loadings without also constraining the residuals, since Var(x) =1.0