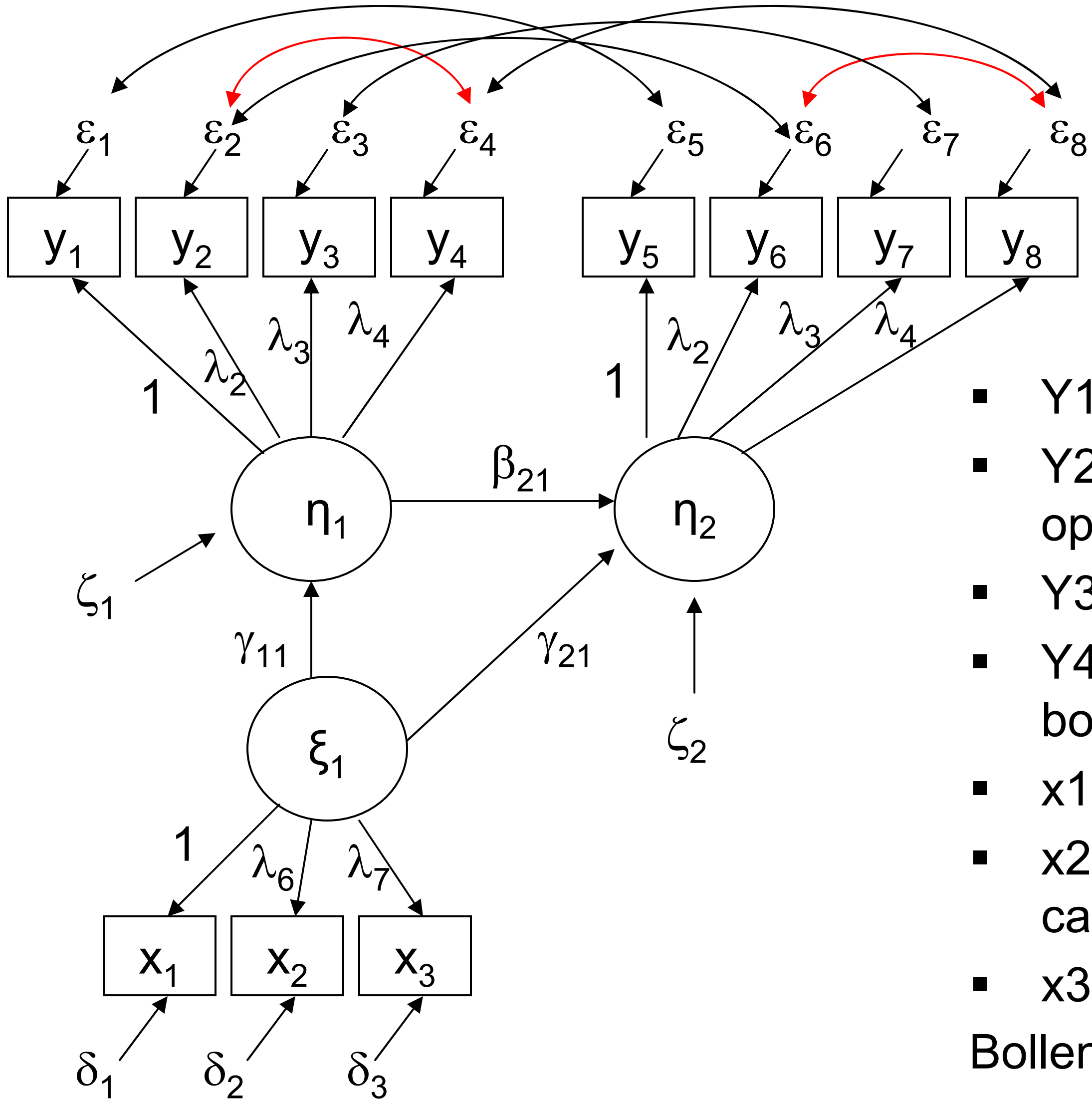
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Inference using structural equations with latent variables

Statistics for Psychosocial Research II:
Structural Models

Qian-Li Xue

Example: Industrialization and Political Policy



- Y1,Y5: freedom of press
- Y2,Y6: freedom of group opposition
- Y3,Y7: fairness of election
- Y4,Y8: effectiveness of legislative body
- x1 is GNP per capita
- x2 is energy consumption per capita
- x3 is % labor force

Bollen pp322-323

Key Components of SEM

- Formulation of Theory
- Model specification
  - Model identification √
- Model estimation
- Model evaluation √

Identifiability

- Let θ be a $t\times 1$ vector containing all unknown and unconstrained parameters in a model. The parameters θ are identified if $\Sigma(\theta_1)= \Sigma(\theta_2) \Leftrightarrow \theta_1=\theta_2$
- Estimability ≠ Identifiability !!
- Identifiability – attribute of the model
- Estimability – attribute of the data

Identifiability Rules for CFA

(1) T-rule (revisited)

- necessary, but not sufficient
- "t" "things" to estimate
- "n" observed variables

$$t \leq \frac{1}{2} n(n+1)$$

- More than one way of assessing the t-rule
- Bollen:
  - ❖ $n(n+1)/2 \geq$ number of exogenous variances + number of error variances + number of direct effects + number of double-headed arrows.
- Maruyama:
  - ❖ $n(n-1)/2 \geq$ number of direct effects + number of double-headed arrows.

- Both give the same result in terms of number of degrees of freedom of the model.

Identifiability Rules for CFA

(2) Two-indicator rule *(<u>sufficient, not necessary</u>)*

1) At least two factors
2) At least two indicators per factor
3) Exactly one non-zero element per row of Λ

   (translation: each x only is pointed at by one LV)
4) Non-correlated errors (Θ is diagonal)

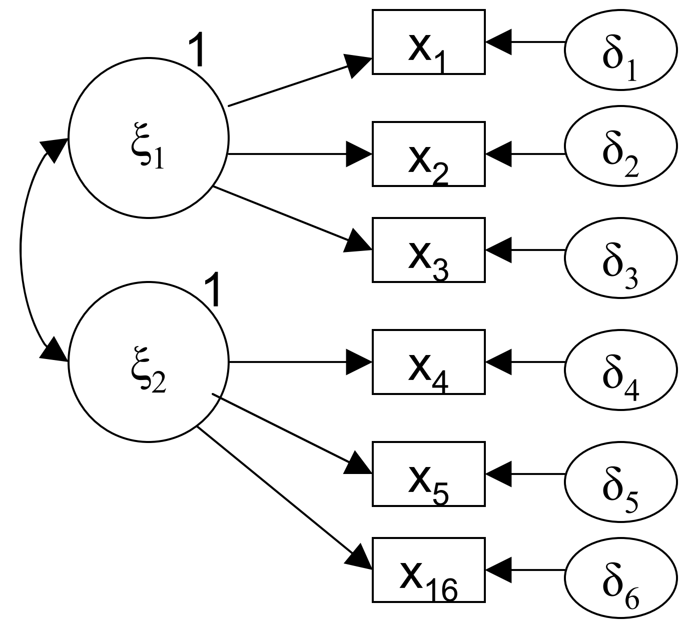
   (translation: no double-header arrows between the δ's)
5) Factors are correlated (Φ has no zero elements)*

   (translation: there are double-header arrows between all of the LVs)

* Alternative less strict criteria: each factor is correlated with **at least** one other factor.

(see page 247 on Bollen)

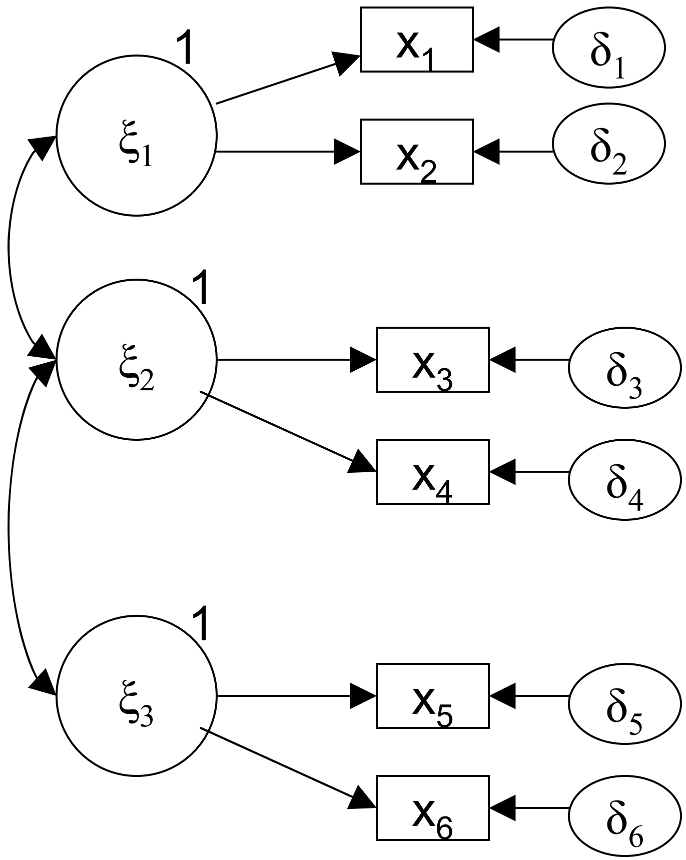
$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2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



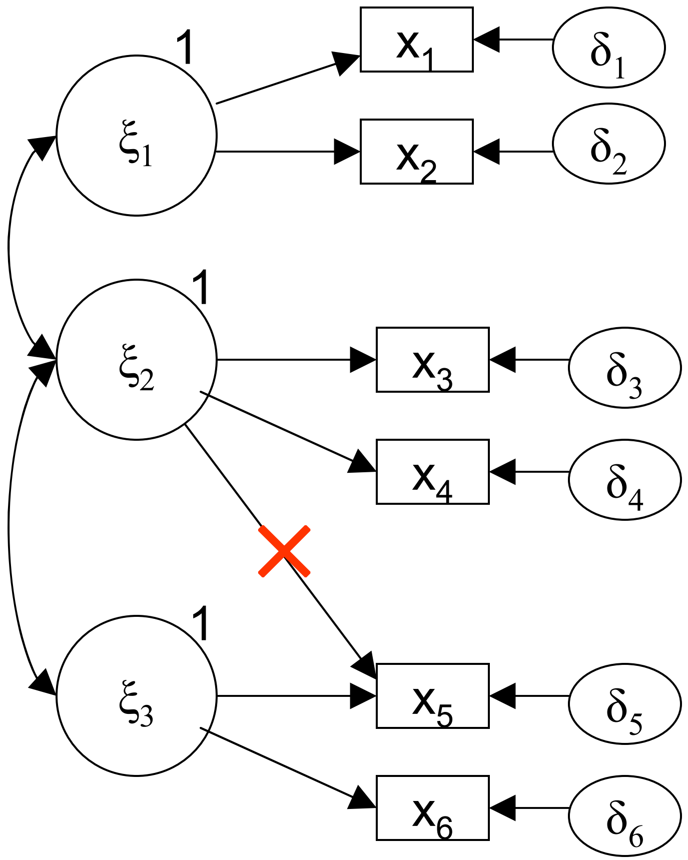
$$\Theta = \text{var}(\delta) = \begin{bmatrix} \theta_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & \theta_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & \theta_{66} \end{bmatrix}$$

$$\Phi = \text{var}(\xi) = \begin{bmatrix} 1 & \varphi_{12} \\ \varphi_{12} & 1 \end{bmatrix}$$

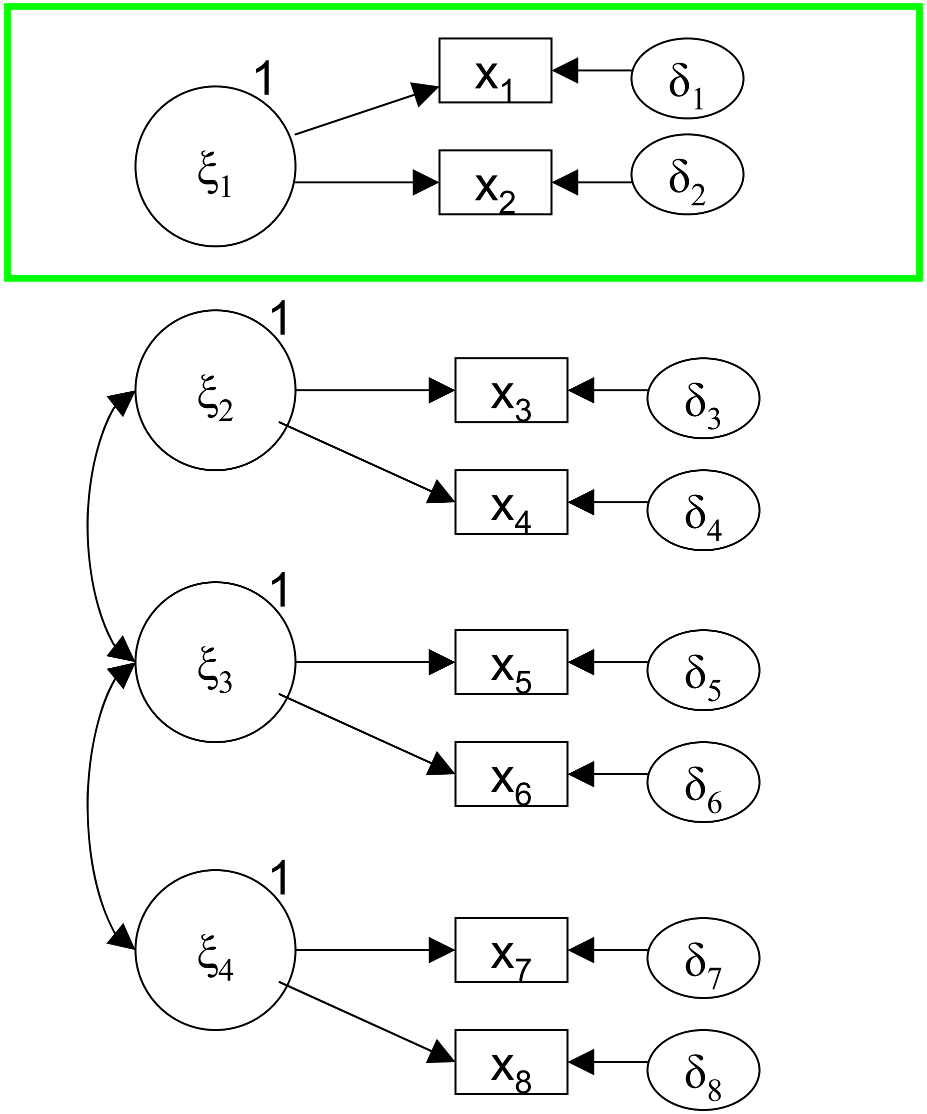
Example: Two-Indicator Rule



Example: Two-Indicator Rule



Example: Two-Indicator Rule



Identifiability Rules for CFA

(3) *Three-indicator rule* (*<u>sufficient, not necessary</u>*)

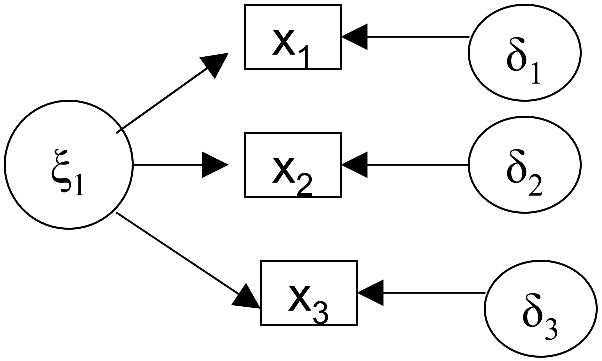
1) at least one factor
2) at least three indicators per factor
3) one non-zero element per row of Λ

   (translation: each x only is pointed at by one LV)

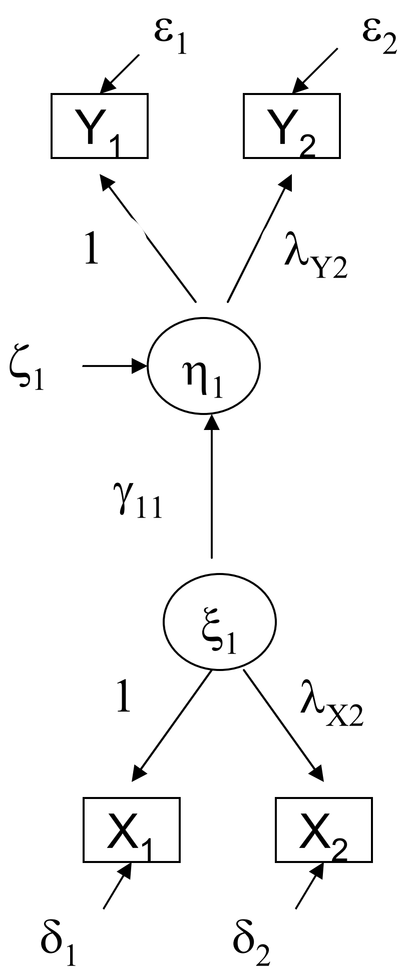
4) non-correlated errors (Φ is diagonal)

   (translation: no double-headed arrows between the δ's)

[Note: no condition about correlation of factors (no restrictions on Φ).]



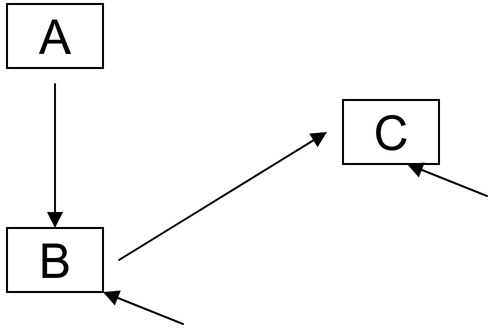
Identifiability: Latent Variable and Measurement Models Combined



- Apply the t-rule
  - # equations = n(n+1)/2=10
  - # unknows = 9:
    - Direct effects: λ_{Y2}, λ_{X2}, γ_{11}
    - Variance: var(η_1)=ϕ_{11}
    - Residual Variances: var(ε_1), var(ε_2), var(δ_1), var(δ_2), var(ζ_1)=ψ_{11}
  - # equations > # unknows
  - T-rule is met, identification is possible!

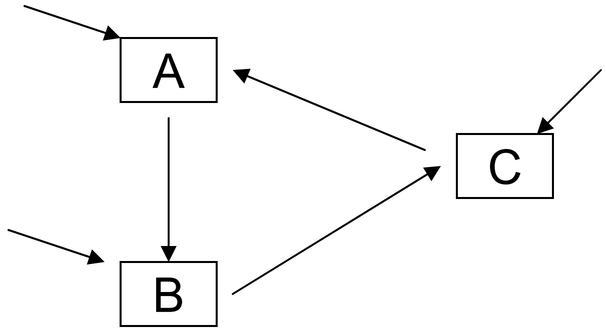
Two more Identification Rules for Path Models

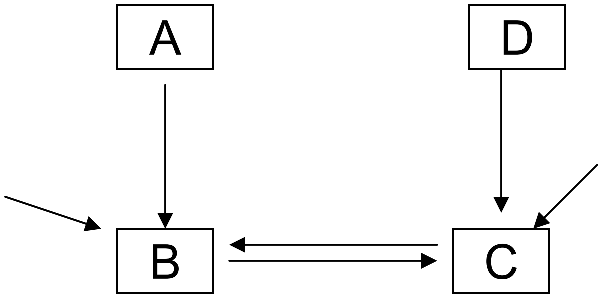
- "Null B rule"
  - a sufficient, but not necessary condition for identification
  - when no endogenous variable affects any other endogenous variable then the model is identified.



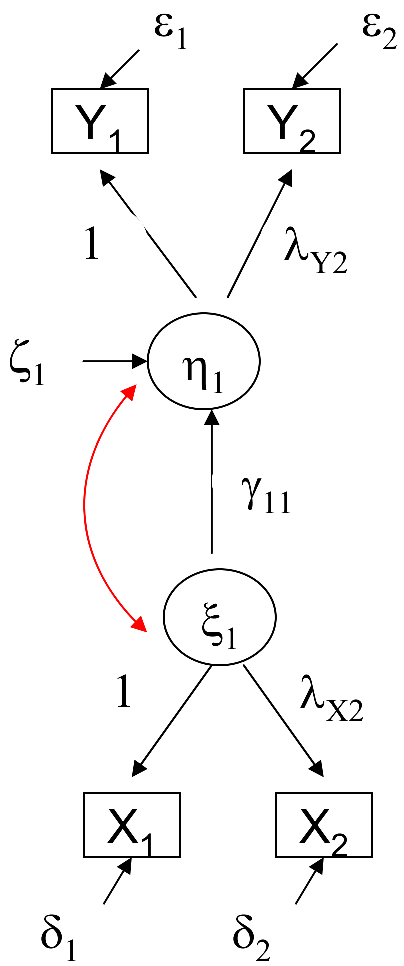
Two more Identification Rules for Path Models

- The "recursive rule"
  - a sufficient, but not necessary condition for identification
  - recursive models are identified
  - Definition: a model is recursive if there are no loops



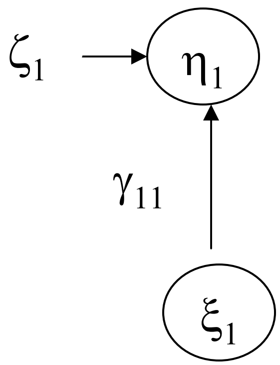


Identifiability: Latent Variable and Measurement Models Combined



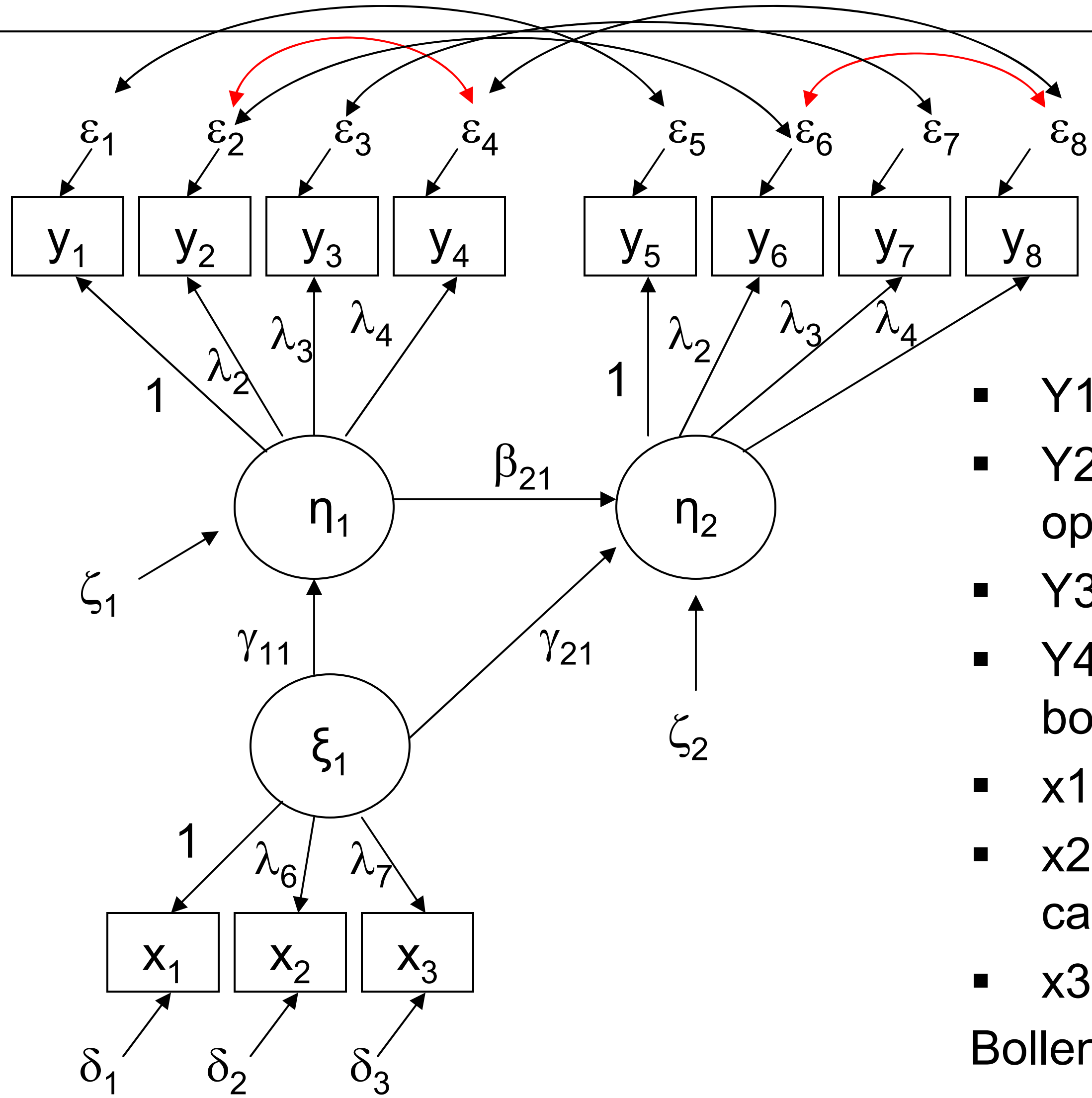
- Two-step rule
  - Step 1:
    - reformulate the model as a confirmatory factor analysis
    - assess identifiability
    - if identified, move to Step 2
  - Example
    - 2 factors, 2 indicators for each factor, each indicator is associated with only one factor, independent errors, correlated factors
    - The 2-indicator rule for CFA applies
    - Therefore, the CFA model is identified

Identifiability: Latent Variable and Measurement Models Combined



- Two-step rule
  - Step 2
    - Examines the latent variable equation of the model
    - Treat it as if it were a structural eqn. for observed variables
    - Assess identifiability using the rules for path models
    - If the LV eqn. is identified, it is sufficient to identify the whole model
  - Example:
    - The LV model is identified based on the Null-B rule and the recursive rule
    - The whole model is identified

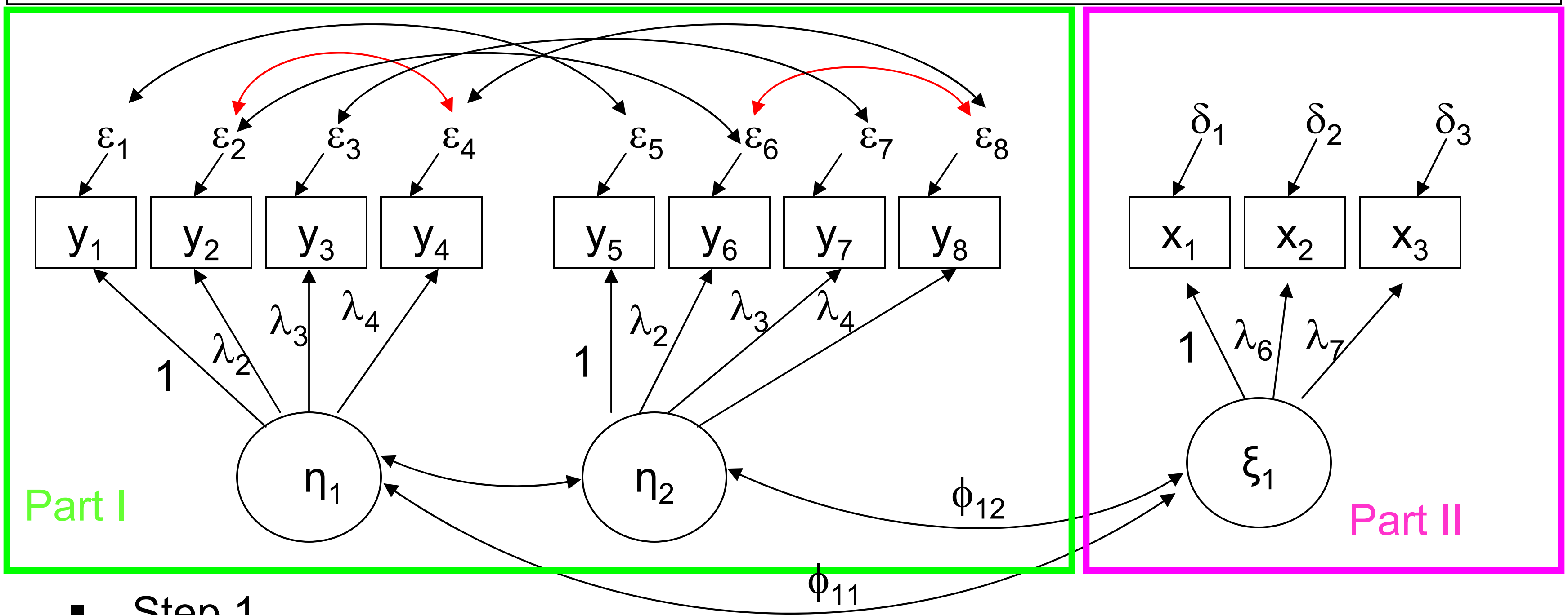
Identifiability: Latent Variable and Measurement Models Combined



- Y1,Y5: freedom of press
- Y2,Y6: freedom of group opposition
- Y3,Y7: fairness of election
- Y4,Y8: effectiveness of legislative body
- x1 is GNP per capita
- x2 is energy consumption per capita
- x3 is % labor force

Bollen pp322-323

Identifiability: Latent Variable and Measurement Models Combined



- Step 1
  - # equations=n(n+1)/2=11*12/2=66; # unknowns=28
  - The CFA model may be identified by the T-rule
  - Part I is identified by covariance algebra (Bollen, pp. 251-254)
  - Part II is identified by the 3-indicator rule
  - ϕ_{11} and ϕ_{12} are identified based on covariance algebra: cov(x_1,y_1) and cov(x_1,y_5)
  - The entire CFA model is identified

Empirical Tests of Identification

- Unfortunately, the rules introduced so far do not cover all models (e.g. correlated errors, non-recursive models)
- Neither are they necessary <u>nor</u> sufficient
- Empirical (data-based) tests do exist in routine SEM programs for "local" identification
  - Wald's rank rule (Wald, 1950)
  - Invertible information matrix (Rothenberg, 1971)
- But what do you mean by "local"?

Global vs. Local Identifiability

- θ is globally identified if no θ_1 and θ_2 exist such that $\Sigma(\theta_1) = \Sigma(\theta_2)$ unless $\theta_1 = \theta_2$
- θ is locally identified at a point θ_1, if in the neighborhood of θ_1 there is no θ_2 for which $\Sigma(\theta_1) = \Sigma(\theta_2)$ unless $\theta_1 = \theta_2$
- Global identifiability $\Rightarrow$ Local identifiability
- Local identifiability $\not\Rightarrow$ Global identifiability

See Figure 2.3 on page 39 of Loehlin's Latent Variable Models

Checking Identifiability

1) Apply appropriate identifiability rules
2) If the model meets a necessary but not sufficient condition, if possible, solve parameters in terms of the elements of the covariance matrix
3) Run empirical test on the information matrix, or
4) Estimate model with different starting values, or
5) Run empirical tests for random subsamples

Bollen, 1989; p251

Parameter Evaluation

- Estimate/SE = Z-statistic
- Standard interpretation:
  - if |Z| > 2, then "significant"
- Consider both statistical and scientific value of including a variable in the model
- Notes for parameter testing in CFA:
  - Not usually interesting in determining if loadings are equal to zero
  - Might be interested in testing whether or not covariance between factors is zero.

Model Evaluation

- Two basic types of model evaluation statistics
  - Global tests of Goodness-of-Fit
  - Comparison of models

Model Evaluation:
Global Tests of Goodness of Fit

<u>Based on predicted vs. observed covariances:</u>

1. **χ^2 tests**
   - d.f.=(# non-redundant components in S) – (# unknown parameter in the model)
   - Null hypothesis: lack of significant difference between $\Sigma(\theta)$ and S
   - Sensitive to sample size
   - Sensitive to the assumption of multivariate normality
   - χ^2 tests for difference between <u>**NESTED**</u> models

2. **Root Mean Square Error of Approximation** (RMSEA)
   - A population index, <u>insensitive to sample size</u>
   - No specification of baseline model is needed
   - Test a null hypothesis of poor fit
   - Availability of confidence interval
   - <0.10 “good”, <0.05 “very good” (Steiger, 1989, p.81)

3. **Standardized Root Mean Residual** (SRMR)
   - Squared root of the mean of the squared standardized residuals
   - SRMR = 0 indicates “perfect” fit, < .05 “good” fit, < .08 adequate fit

Model Evaluation: Global Tests of Goodness of Fit

Comparing the given model with the “null” model

1. **Comparative Fit Index** (CFI; Bentler 1989)
   - compares the existing model fit with a null model which assumes uncorrelated variables in the model (i.e. the "independence model")
   - Interpretation: % of the covariation in the data can be explained by the given model
   - Insensitive to sample size
   - CFI ranges from 0 to 1, with 1 indicating a very good fit; acceptable fit if CFI>0.9

2. **The Tucker-Lewis Index** (TLI) or **Non-Normed Fit Index** (NNFI)
   - Relatively independent of sample size (Marsh et al. 1988, 1996)
   - NNFI >= .95 indicates a good model fit, <0.9 poor fit

Model Evaluation: Comparing Models

- You may compare the overall goodness-of-fit statistics (e.g. χ^2, RMSEA)
- Prefer those that adjust for degrees of freedom (e.g. RMSEA, TLI)
- But, they're still descriptive, no tests of statistical significance
- Formal tests are available
  - Choice of fit statistics depends on types of models in comparison
    - Nested models
    - Non-nested models

Model Evaluation: Comparing Models

For comparison of <u>nested</u> models:

1. Likelihood Ratio Test (LRT)

$$LR = -2\left[\log L\left(\hat{\theta}_r\right) - \log L\left(\hat{\theta}_u\right)\right] = -2\log\left(\frac{L\left(\hat{\theta}_r\right)}{L\left(\hat{\theta}_u\right)}\right),$$

where $\hat{\theta}_r$ and $\hat{\theta}_u$ are the ML estimators for the restricted and nested model and the unrestricted model, respectively.

- **With d.f. = difference in the d.f. for the two models**
- **Valid if the ML estimators are valid**
- **Actual model fitting is needed for the two models**

Model Evaluation: Comparing Models

For comparison of <u>nested</u> models:

2. Lagrangian Multiplier Test (LMT)

- Only requires fitting of the restricted model
- Based on the score $s(\theta_u)=\partial \log L(\theta_u)/\partial\theta_u$, where $L(\theta_u)$ is the <u>unrestricted</u> likelihood function
- $s(\theta_u)=0$ when evaluated at the MLE of θ_u
- The Idea: substitute the MLE of <u>θ_r</u>, assess departure from 0

$$LM = \left[s\left(\hat{\theta}_r\right)\right]' I^{-1}\left(\hat{\theta}_r\right)\left[s\left(\hat{\theta}_r\right)\right],$$

where $I^{-1}\left(\hat{\theta}_r\right)$ is the variance of $\hat{\theta}_r$.

- LM ~ χ^2 with d.f.=difference in the d.f. of the two nested models

Model Evaluation: Comparing Models

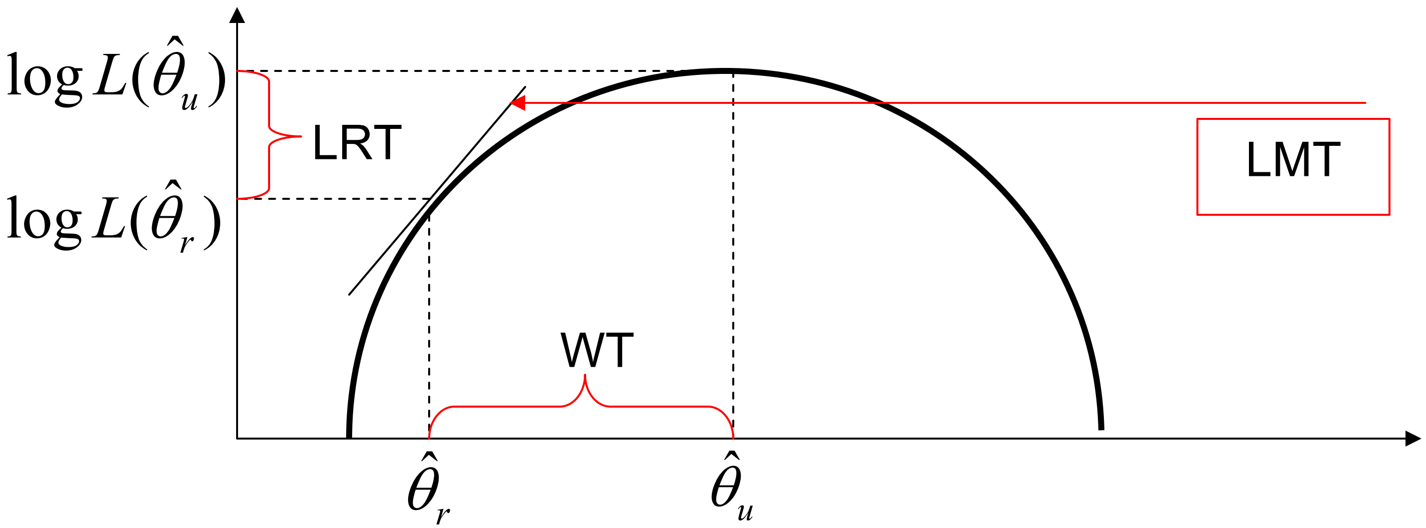
For comparison of <u>nested</u> models:

2. Wald Test (LMT)

- Let r(θ) be a vector of model constraints
- We know that r($\hat{\theta}_r$) = 0
- Fit the unrestricted model, if the restricted model is false, r($\hat{\theta}_u$) >> 0
- W ~ χ with d.f. = # of constraints in r(θ)
- E.g. r(θ)= θ_1=0

$$W = \frac{\hat{\theta}_1^{\,2}}{\mathrm{var}(\hat{\theta}_1)}, \mathrm{d.f.} = 1$$

Comparisons of LRT, LMT, and WT



	Restrictions		Estimates Required	
	Imposed	Removed	θ_u $\hat{\theta}_r$	θ_r
LRT	√	√	√	√
LMT		√		√
WT	√		√	

Bollen, pp. 295-296

Model Evaluation: Comparing Models

For comparison of <u>non-nested</u> models:

4. Information Criteria

- Akaike:

  $$AIC = -2LL + 2*s$$

- Schwarz:

  $$BIC = -2LL + s*\log(N)$$

- “consistent” AIC:

  $$CAIC = -2LL + s*(\log(N) + 1)$$

- s is number of free parameters
- the smaller the better