Structural Equation Models: The General Case

STA431: Spring 2015

An Extension of Multiple Regression

- More than one regression-like equation
- Includes latent variables
- Variables can be explanatory in one equation and response in another
- Modest changes in notation
- Vocabulary
- Path diagrams
- No intercepts, all expected values zero
- Serious modeling (compared to ordinary statistical models)
- Parameter identifiability

Variables can be response in one equation and explanatory in another

- Variables (IQ = Intelligence Quotient):
  - X_1 = Mother's adult IQ
  - X_2 = Father's adult IQ
  - Y_1 = Person's adult IQ
  - Y_2 = Child's IQ in Grade 8

$$
\begin{aligned}
Y_1 &= \alpha_1 + \gamma_1 X_1 + \gamma_2 X_2 + \epsilon_1 \\
Y_2 &= \alpha_2 + \beta Y_1 + \epsilon_2
\end{aligned}
$$

- Of course all these variables are measured with error.
- We will lose the intercepts very soon.

Modest changes in notation

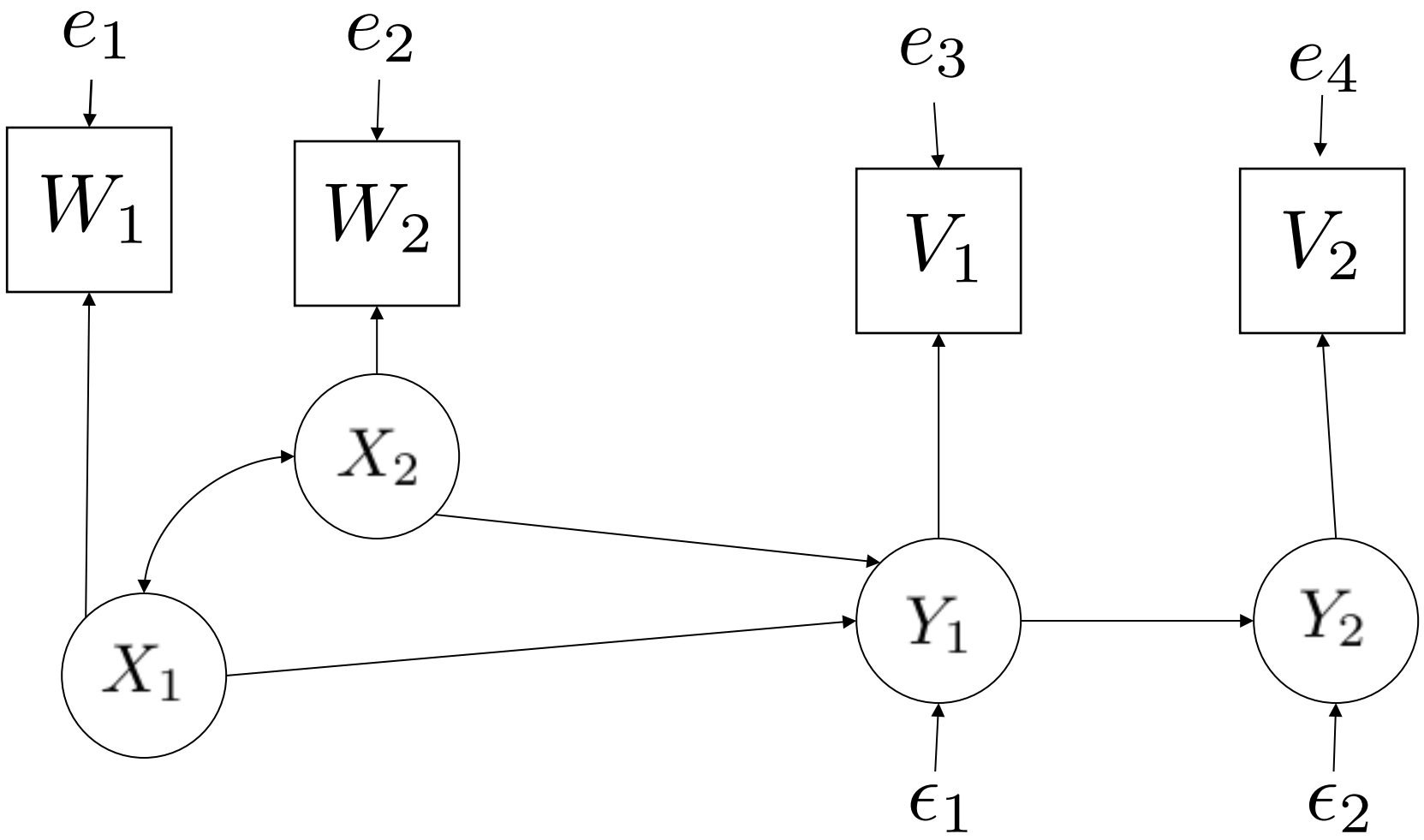
- Regression coefficients are now called gamma instead of beta
- Betas are used for links between Y variables
- Intercepts are alphas but they will soon disappear.
- We feel free to drop the subscript *i*; implicitly, everything is independent and identically distributed for *i = 1, …, n*.

$$
\begin{aligned}
Y_1 &= \alpha_1 + \gamma_1 X_1 + \gamma_2 X_2 + \epsilon_1 \\
Y_2 &= \alpha_2 + \beta Y_1 + \epsilon_2
\end{aligned}
$$

Vocabulary

- Variables can be Latent or Manifest.
  - Manifest means observable
  - All error terms are latent
- Variables can be Exogenous or Endogenous
  - **Ex**ogenous variables appear only on the right side of the = sign.
    - Think "X" for explanatory variable.
    - All error terms are exogenous
  - **End**ogenous variables appear on the left of at least one = sign.
    - Think "end" of an arrow pointing from exogenous to endogenous
    - Betas link endogenous variables to other endogenous variables.

Path diagrams



Path Diagram Rules

- Latent variables are enclosed by ovals.
- Observable (manifest) variables are enclosed by rectangles.
- Error terms are not enclosed
  - Sometimes the arrows from the error terms seem to come from nowhere. The symbol for the error term does not appear in the path diagram.
  - Sometimes there are no arrows for the error terms at all. It is just assumed that such an arrow points to each endogenous variable.
- Straight, single-headed arrows point from each variable on the right side of an equation to the endogenous variable on the left side.
  - Sometimes the coefficient is written on the arrow, but sometimes it is not.
- A curved, double-headed arrow between two variables (always exogenous variables) means they have a non-zero covariance.
  - Sometimes the symbol for the covariance is written on the curved arrow, but sometimes it is not.

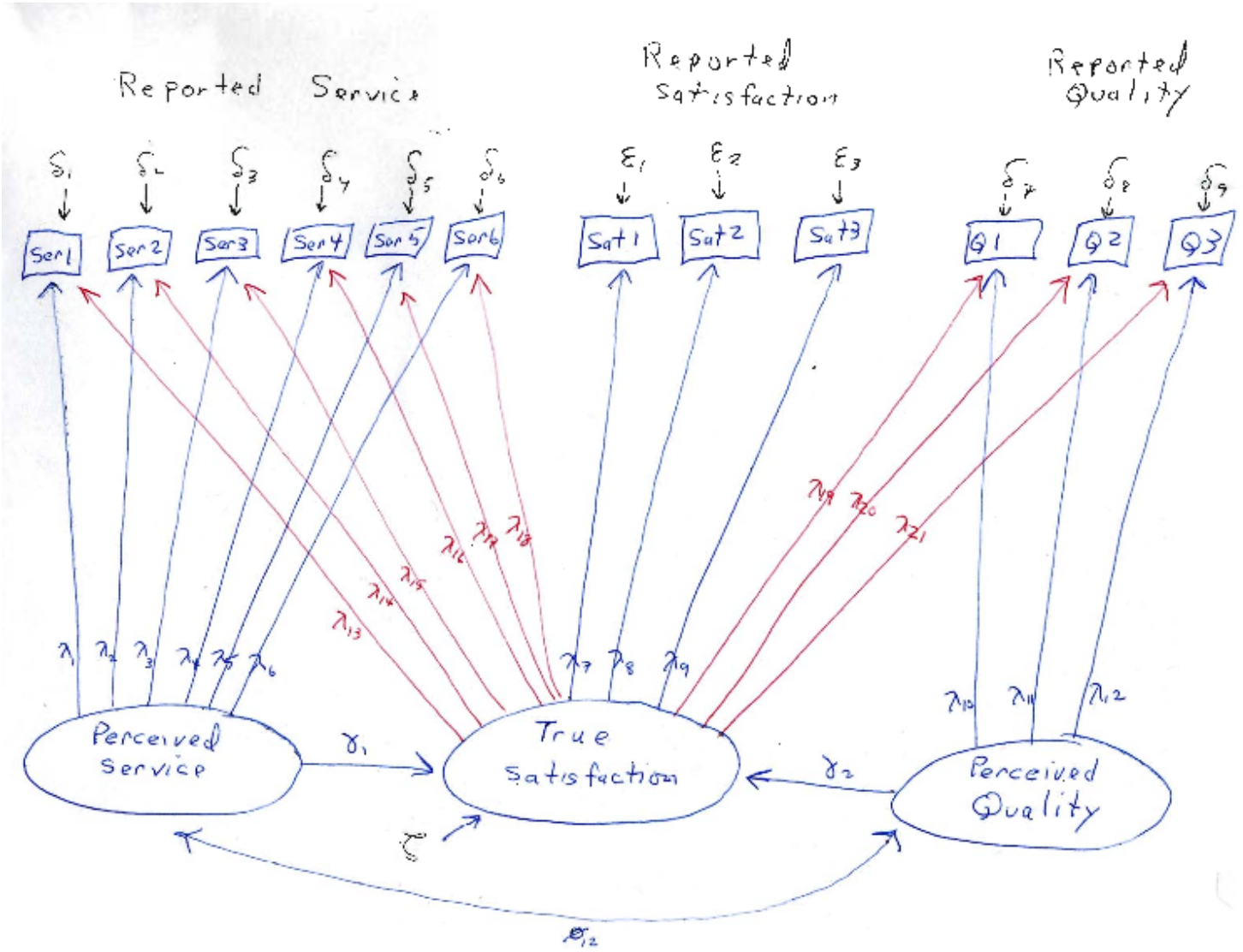
Causal Modeling (cause and effect)

- The arrows deliberately imply that if A → B, we are saying A *contributes* to B, or partly *causes* it.
- There may be other contributing variables. All the ones that are unknown are lumped together in the error term.
- Are these unknown variables are independent of the variables in the model? Probably not.
- Sometimes we can get around the problem with instrumental variables.

Serious Modeling

- Once you accept that model equations are statements about what contributes to what, you realize that structural equation models represent a rough *theory* of the data, with some parts (the parameter values) unknown.
- They are somewhere between ordinary statistical models, which are like one-size-fits-all clothing, and true scientific models, which are like tailor made clothing.
- So they are very flexible and potentially valuable. It is *good* to combine what the data can tell you with what you already know.
- But structural equation models can require a lot of input and careful thought to construct. In this course, we will get by mostly on common sense.
- In general, the parameters of the most reasonable model need not be identifiable. It depends upon the form of the data as well as on the model. Identifiability needs to be checked. Frequently, this can be done by inspection.

Example: Halo Effects in Real Estate



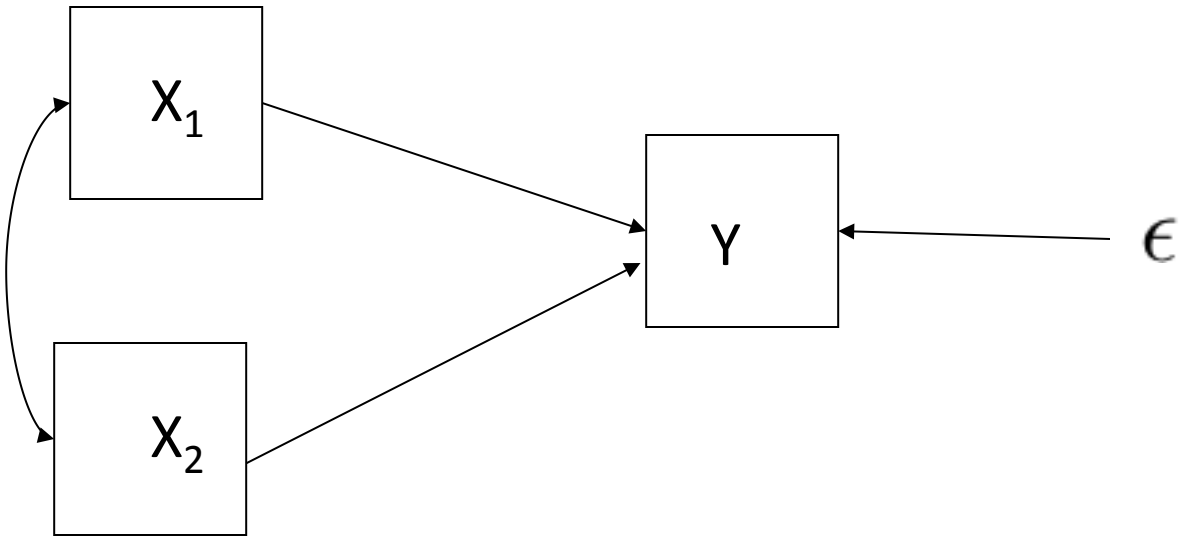
Losing the intercepts and expected values

- Mostly, the intercepts and expected values are not identifiable anyway, as in multiple regression with measurement error.
- We have a chance to identify a *function* of the parameter vector – the parameters that appear in the covariance matrix $\boldsymbol{\Sigma} = V(\mathbf{D})$.
- Re-parameterize. The new parameter vector is the set of parameters in $\boldsymbol{\Sigma}$, and also $\boldsymbol{\mu} = E(\mathbf{D})$. Estimate $\boldsymbol{\mu}$ with x-bar, forget it, and concentrate on inference for the parameters in $\boldsymbol{\Sigma}$.
- To make calculation of the covariance matrix easier, write the model equations in centered form. The little letters $_c$ over the variables are invisible.

From this point on the models have no means and no intercepts.

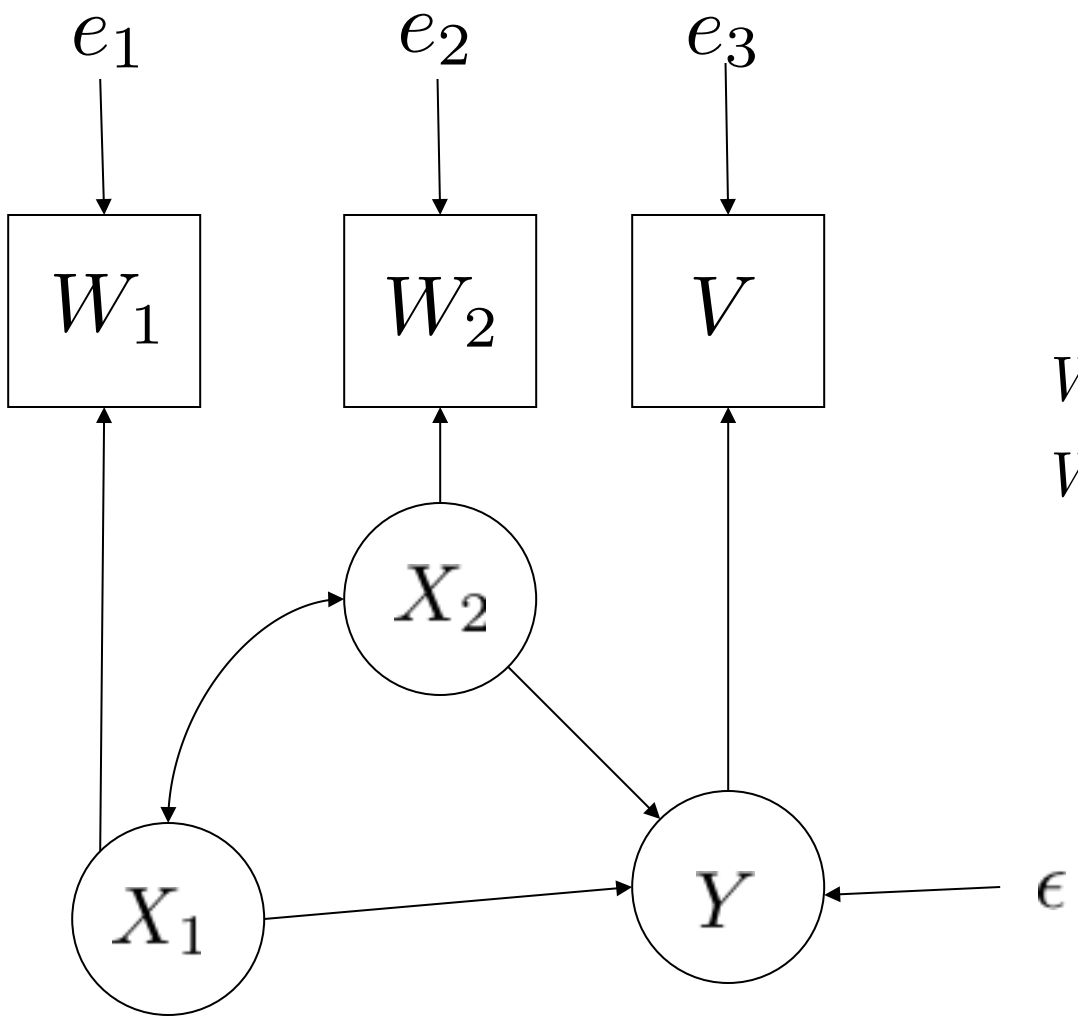
Now more examples

Multiple Regression



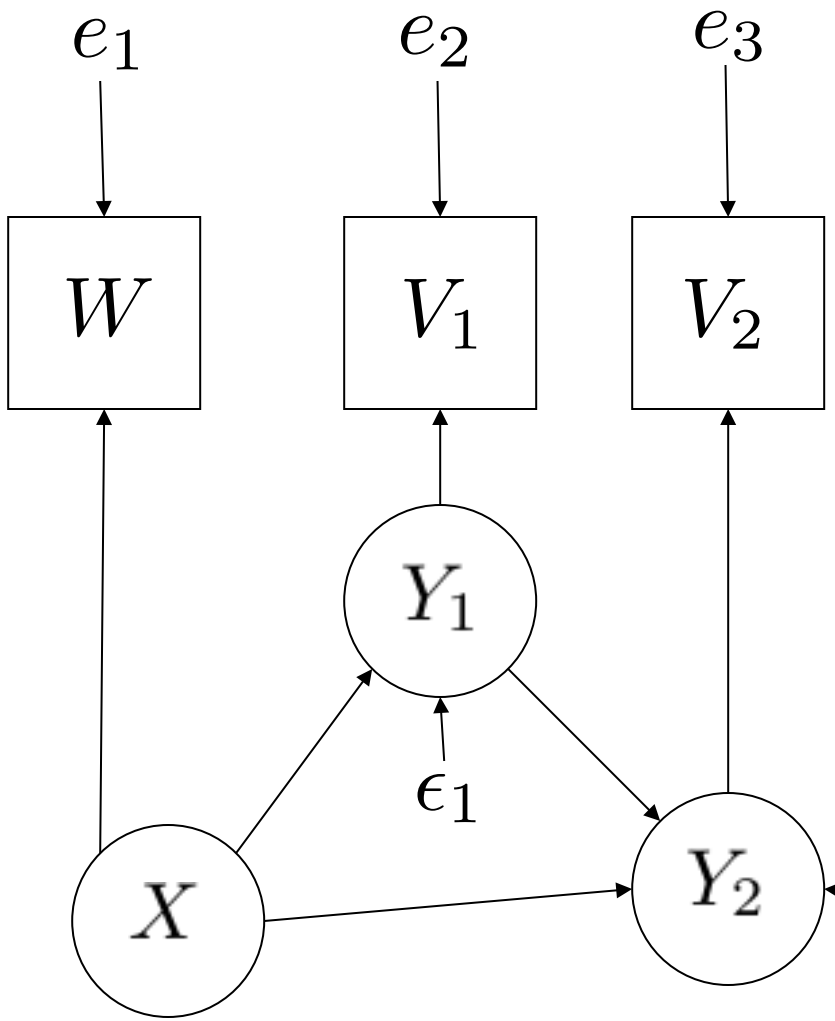
$$Y = \gamma_1 X_1 + \gamma_2 X_2 + \epsilon$$

Regression with measurement error



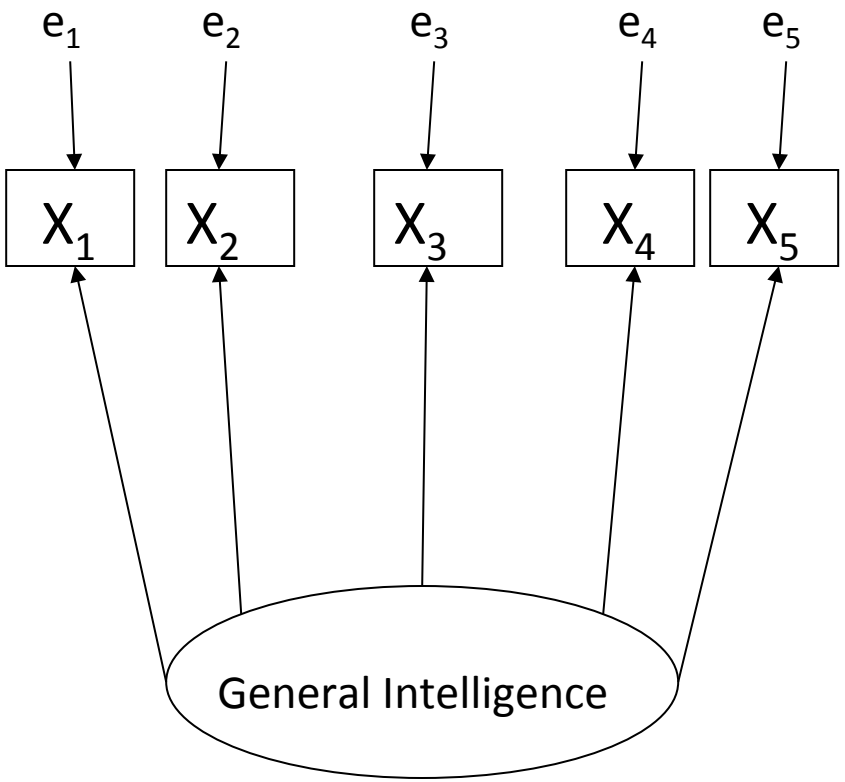
$$
\begin{aligned}
Y &= \gamma_1 X_1 + \gamma_2 X_2 + \epsilon \\
W_1 &= X_1 + e_1 \\
W_2 &= X_2 + e_2 \\
V &= Y + e_3
\end{aligned}
$$

A Path Model with Measurement Error



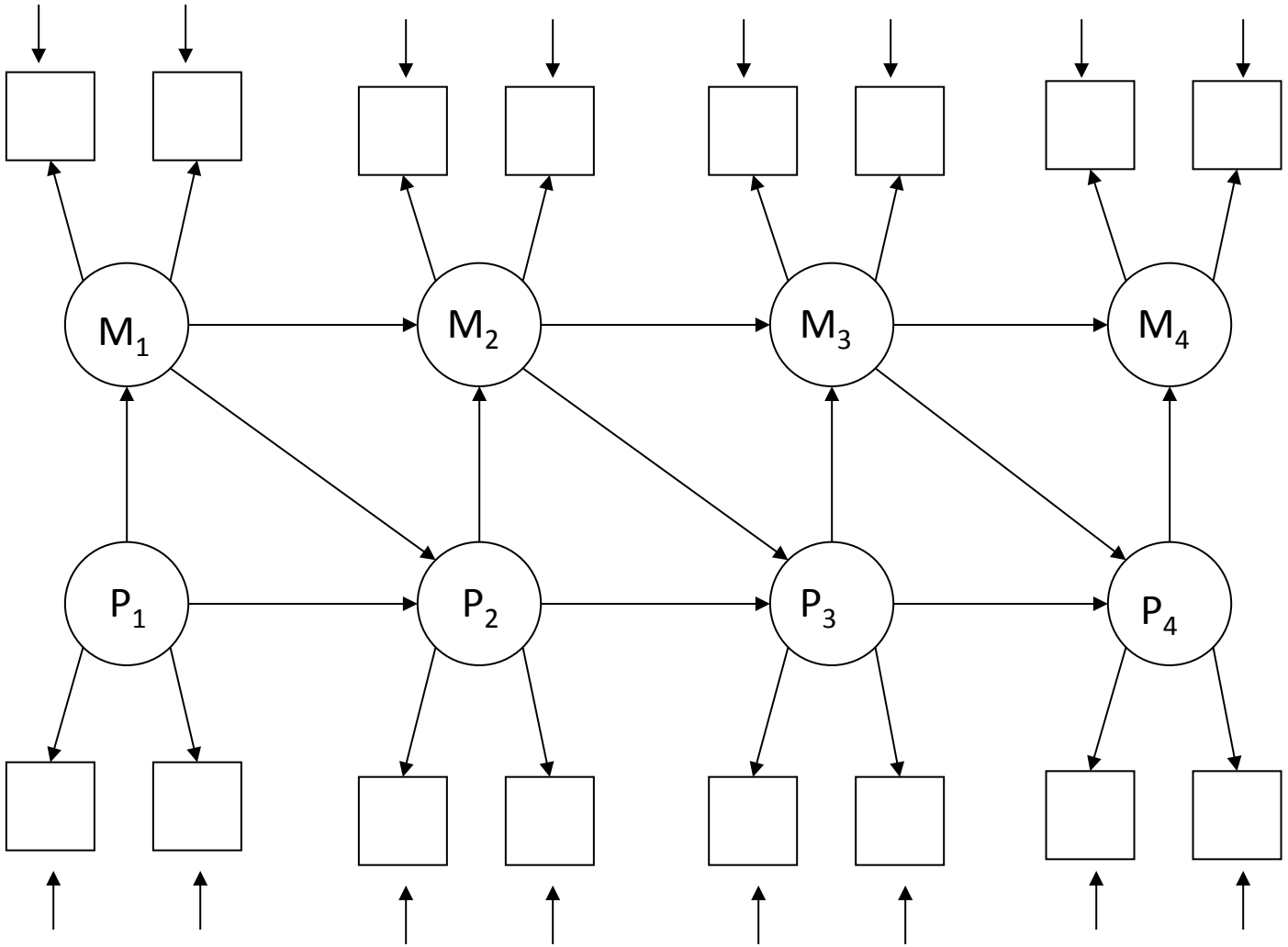
$$
\begin{aligned}
Y_1 &= \gamma_1 X + \epsilon_1 \\
Y_2 &= \beta Y_1 + \gamma_2 X + \epsilon_2 \\
W &= X + e_1 \\
V_1 &= Y_1 + e_2 \\
V_2 &= Y_2 + e_3
\end{aligned}
$$

A Factor Analysis Model

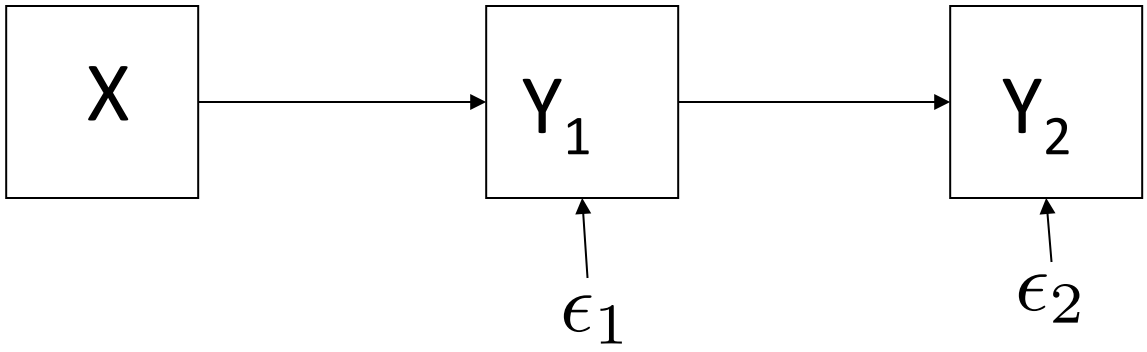


$$X_1 = \lambda_1 F + e_1$$
$$X_2 = \lambda_2 F + e_2$$
$$X_3 = \lambda_3 F + e_3$$
$$X_4 = \lambda_4 F + e_4$$
$$X_5 = \lambda_5 F + e_5$$

A Longitudinal Model



Estimation and Testing as Before



$$
\begin{aligned}
Y_1 &= \gamma X + \epsilon_1 \\
Y_2 &= \beta Y_1 + \epsilon_2
\end{aligned}
$$

All expected values equal zero.

$V(X) = \phi$, $V(\epsilon_1) = \psi_1$, $V(\epsilon_2) = \psi_2$,

$X, \epsilon_1, \epsilon_2$ are all independent.

Everything is normal.

Distribution of the data

$$\begin{pmatrix} X_1 \\ Y_{1,1} \\ Y_{1,2} \end{pmatrix} \dots \begin{pmatrix} X_n \\ Y_{n,1} \\ Y_{n,2} \end{pmatrix}$$ are independent normal with mean zero.

and covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \phi & \gamma\phi & \beta\gamma\phi \\ \gamma\phi & \gamma^2\phi + \psi_1 & \beta(\gamma^2\phi + \psi_1) \\ \beta\gamma\phi & \beta(\gamma^2\phi + \psi_1) & \beta^2(\gamma^2\phi + \psi_1) + \psi_2 \end{pmatrix}$$

$$\boldsymbol{\theta} = (\gamma, \beta, \phi, \psi_1, \psi_2)$$

A General Two-Stage Model

$$
\begin{aligned}
\mathbf{Y}_i &= \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i \\
\mathbf{F}_i &= \begin{pmatrix} \mathbf{X}_i \\ \mathbf{Y}_i \end{pmatrix} \\
\mathbf{D}_i &= \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i
\end{aligned}
$$

- $\mathbf{D}_i$ (the data) are observable. All other variables are latent.
- $\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i$ is called the *Latent Variable Model*
- The latent vectors $\mathbf{X}_i$ and $\mathbf{Y}_i$ are collected into a "factor" $\mathbf{F}_i$. This is *not* a categorical independent variable, the usual meaning of factor in experimental design.
- $\mathbf{D}_i = \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i$ is called the *Measurement Model*.

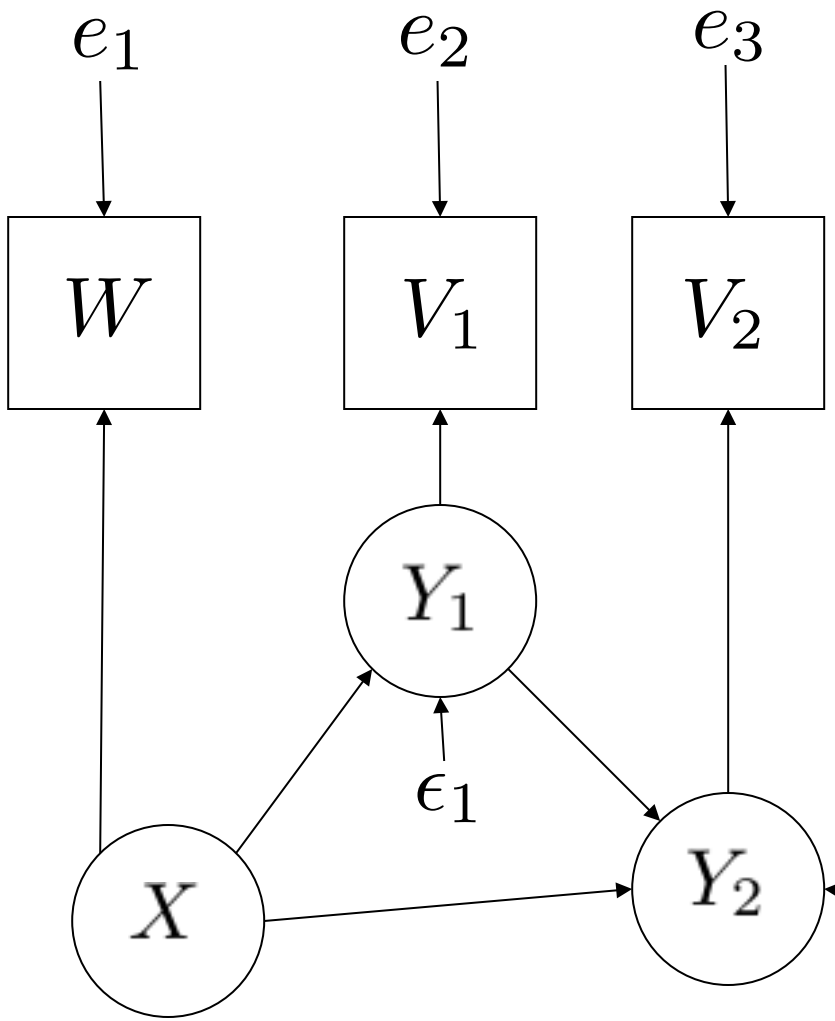
More Details

- $\mathbf{Y}_i$ is a $q \times 1$ random vector.
- $\boldsymbol{\beta}$ is a $q \times q$ matrix of constants with zeros on the main diagonal.
- $\boldsymbol{\Gamma}$ is a $q \times p$ matrix of constants.
- $\mathbf{X}_i$ is a $p \times 1$ random vector.
- $\boldsymbol{\epsilon}_i$ is a $q \times 1$ random vector.
- $\mathbf{F}_i$ (F for Factor) is just $\mathbf{X}_i$ stacked on top of $\mathbf{Y}_i$. It is a $(p+q) \times 1$ random vector.
- $\mathbf{D}_i$ is a $k \times 1$ random vector. Sometimes, $\mathbf{D}_i = \begin{pmatrix} \mathbf{W}_i \\ \mathbf{V}_i \end{pmatrix}$
- $\boldsymbol{\Lambda}$ is a $k \times (p+q)$ matrix of constants.
- $\mathbf{D}_i$ is a $k \times 1$ random vector.
- $\mathbf{e}_i$ is a $k \times 1$ random vector.
- $\mathbf{X}_i$, $\boldsymbol{\epsilon}_i$ and $\mathbf{e}_i$ are independent.

$$
\begin{aligned}
\mathbf{Y}_i &= \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i \\
\mathbf{F}_i &= \begin{pmatrix} \mathbf{X}_i \\ \mathbf{Y}_i \end{pmatrix} \\
\mathbf{D}_i &= \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i
\end{aligned}
$$

- $V(\mathbf{X}_i) = \boldsymbol{\Phi}_x$
- $V(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$
- $V(\mathbf{F}_i) = V\begin{pmatrix} \mathbf{X}_i \\ \mathbf{Y}_i \end{pmatrix} = \begin{pmatrix} V(\mathbf{X}_i) & C(\mathbf{X}_i, \mathbf{Y}_i) \\ C(\mathbf{Y}_i, \mathbf{X}_i) & V(\mathbf{Y}_i) \end{pmatrix} = \boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12} \\ \boldsymbol{\Phi}'_{12} & \boldsymbol{\Phi}_{22} \end{pmatrix}$
- $V(\mathbf{e}_i) = \boldsymbol{\Omega}$
- $V(\mathbf{D}_i) = \boldsymbol{\Sigma}$

Recall the example



$$
\begin{aligned}
Y_1 &= \gamma_1 X + \epsilon_1 \\
Y_2 &= \beta Y_1 + \gamma_2 X + \epsilon_2 \\
W &= X + e_1 \\
V_1 &= Y_1 + e_2 \\
V_2 &= Y_2 + e_3
\end{aligned}
$$

$$\begin{aligned}\mathbf{Y} &= \boldsymbol{\beta}\mathbf{Y} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\ \mathbf{D} &= \boldsymbol{\Lambda}\mathbf{F} + \mathbf{e}\end{aligned}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \beta & 0 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} + \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} X + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

$$\begin{pmatrix} W \\ V_1 \\ V_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y_1 \\ Y_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}$$

- $V(\mathbf{X}) = \boldsymbol{\Phi}_x = \phi$
- $V(\boldsymbol{\epsilon}) = \boldsymbol{\Psi} = \begin{pmatrix} \psi_1 & 0 \\ 0 & \psi_2 \end{pmatrix}$
- $V(\mathbf{e}) = \boldsymbol{\Omega} = \begin{pmatrix} \omega_1 & 0 & 0 \\ 0 & \omega_2 & 0 \\ 0 & 0 & \omega_3 \end{pmatrix}$
- $V(\mathbf{D}) = \boldsymbol{\Sigma}$

Observable variables in the latent variable model (fairly common)

- These present no problem
- Let $P(e_j=0) = 1$, so $Var(e_j) = 0$
- And $Cov(e_i,e_j)=0$ because if $P(e_j=0) = 1$

$$\begin{aligned} Cov(e_i, e_j) &= E(e_i e_j) - E(e_i)E(e_j) \\ &= E(e_i \cdot 0) - E(e_i) \cdot 0 \\ &= 0 - 0 = 0 \end{aligned}$$

- So in the covariance matrix $\mathbf{\Omega}$=V(**e**), just set $\omega_{ij} = \omega_{ji} = 0$, i=1,…,k

What should you be able to do?

- Given a path diagram, write the model equations and say which exogenous variables are correlated with each other.
- Given the model equations and information about which exogenous variables are correlated with each other, draw the path diagram.
- Given either piece of information, write the model in matrix form and say what all the matrices are.
- Calculate model covariance matrices
- Check identifiability

Recall the notation

$$
\begin{aligned}
\mathbf{Y}_i &= \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i \\
\mathbf{F}_i &= \begin{pmatrix} \mathbf{X}_i \\ \mathbf{Y}_i \end{pmatrix} \\
\mathbf{D}_i &= \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i
\end{aligned}
$$

- $V(\mathbf{X}_i) = \boldsymbol{\Phi}_x$
- $V(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$
- $V(\mathbf{F}_i) = V\begin{pmatrix} \mathbf{X}_i \\ \mathbf{Y}_i \end{pmatrix} = \begin{pmatrix} V(\mathbf{X}_i) & C(\mathbf{X}_i, \mathbf{Y}_i) \\ C(\mathbf{Y}_i, \mathbf{X}_i) & V(\mathbf{Y}_i) \end{pmatrix} = \boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12} \\ \boldsymbol{\Phi}'_{12} & \boldsymbol{\Phi}_{22} \end{pmatrix}$
- $V(\mathbf{e}_i) = \boldsymbol{\Omega}$
- $V(\mathbf{D}_i) = \boldsymbol{\Sigma}$

For the latent variable model, calculate $\mathbf{\Phi} = V(\mathbf{F})$

$$
\begin{aligned}
& \mathbf{Y} = \boldsymbol{\beta}\mathbf{Y} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\
\Rightarrow \quad & \mathbf{Y} - \boldsymbol{\beta}\mathbf{Y} = \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\
\Rightarrow \quad & \mathbf{I}\mathbf{Y} - \boldsymbol{\beta}\mathbf{Y} = \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\
\Rightarrow \quad & (\mathbf{I} - \boldsymbol{\beta})\mathbf{Y} = \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\
\Rightarrow \quad & (\mathbf{I} - \boldsymbol{\beta})^{-1}(\mathbf{I} - \boldsymbol{\beta})\mathbf{Y} = (\mathbf{I} - \boldsymbol{\beta})^{-1}(\boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon}) \\
\Rightarrow \quad & \mathbf{Y} = (\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\mathbf{X} + (\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\epsilon}
\end{aligned}
$$

So,

$$
\begin{aligned}
V(\mathbf{Y}) &= V\left((\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\mathbf{X} + (\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\epsilon}\right) \\
&= V\left((\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\mathbf{X}\right) + V\left((\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\epsilon}\right) \\
&= (\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\, V(\mathbf{X})\left((\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\right)^{\top} + (\mathbf{I} - \boldsymbol{\beta})^{-1} V(\boldsymbol{\epsilon})(\mathbf{I} - \boldsymbol{\beta})^{-1\top} \\
&= (\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\, \boldsymbol{\Phi}_{11}\boldsymbol{\Gamma}^{\top}(\mathbf{I} - \boldsymbol{\beta})^{-1\top} + (\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Psi}(\mathbf{I} - \boldsymbol{\beta})^{-1\top}
\end{aligned}
$$

For the measurement model, calculate **Σ** = V(**D**)

$$
\begin{aligned}
\mathbf{D} &= \mathbf{\Lambda F} + \mathbf{e} \\
\Rightarrow V(\mathbf{D}) &= V(\mathbf{\Lambda F} + \mathbf{e}) \\
&= V(\mathbf{\Lambda F}) + V(\mathbf{e}) \\
&= \mathbf{\Lambda} V(\mathbf{F}) \mathbf{\Lambda}^\top + V(\mathbf{e}) \\
&= \mathbf{\Lambda \Phi \Lambda}^\top + \mathbf{\Omega} \\
&= \mathbf{\Sigma}
\end{aligned}
$$

Two-stage Proofs of Identifiability

- Show the parameters of the measurement model ($\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Omega}$) can be recovered from $\boldsymbol{\Sigma} = V(\mathbf{D})$.
- Show the parameters of the latent variable model ($\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Phi}_{11}$, $\boldsymbol{\Psi}$) can be recovered from $\boldsymbol{\Phi} = V(\mathbf{F})$.
- This means *all* the parameters can be recovered from $\boldsymbol{\Sigma}$.
- Break a big problem into two smaller ones.
- Develop *rules* for checking identifiability at each stage.

Copyright Information