SEM 2: Structural Equation Modeling

Week 1 - Causal modeling and SEM

Sacha Epskamp

07-05-2019

Causal modeling

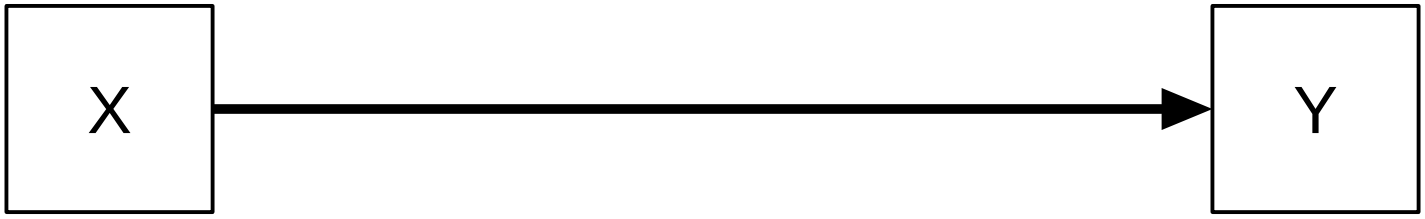
- This course will introduce structural equation modeling (SEM)
- In SEM, we will discuss modeling complex causal hypotheses
- Again, all variables are assumed normally distributed and all associations are assumed linear
- Causal hypotheses can be specified between observed and latent variables
- CFA is a special case of SEM

Causal models

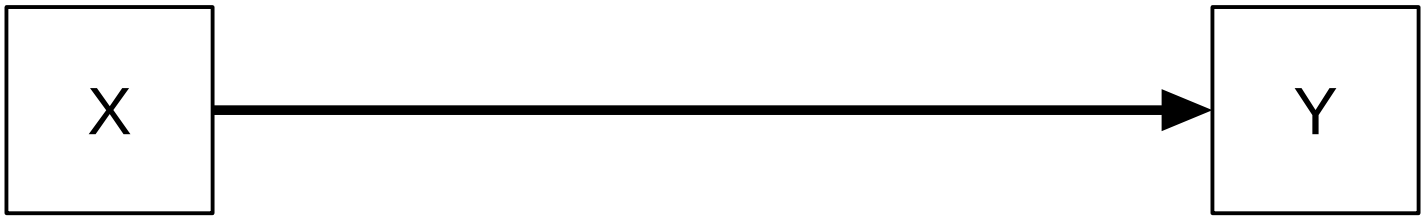
X causes Y

Endogenous and exogenous

- **Exogenous** (independent) variables are variables of which the causal origin are not modeled
  - Exogenous variables have a variance (sometimes not drawn)
  - Exogenous variables, except residuals, are allowed to covary (sometimes not drawn)
  - Latents: ξ (xi); observed: x (x is also used for indicators of latent exogenous variables)
  - Residuals are exogenous
- **Endogenous** (dependent) variables are variables of which the causal origin are modeled
  - Simply stated: endogenous variables have incoming arrows
  - Endogenous variables do not have a variance by themselves
  - Latents: η (eta); observed: y
  - The causal equation for endogenous variables can be derived from the path diagram by summing all incoming edges

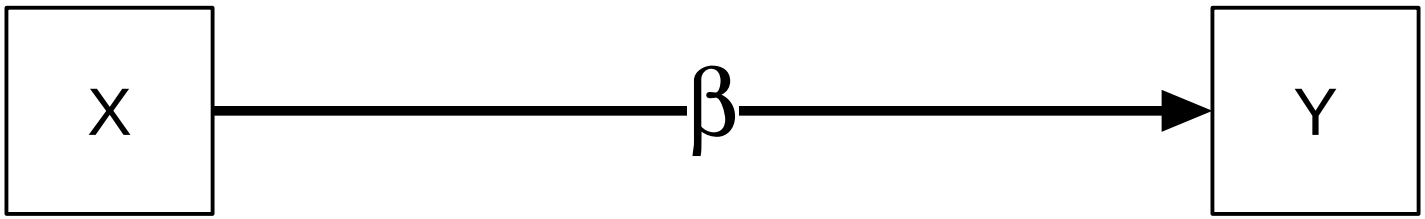


X is exogenous, Y is endogenous

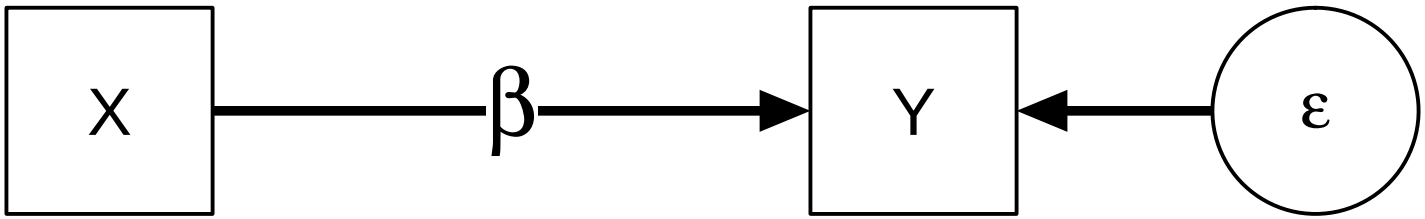


$$y_i = x_i$$

Causal effect goes from right hand side to left hand side.
Experimentally changing x will change y, experimentally changing y will not change x

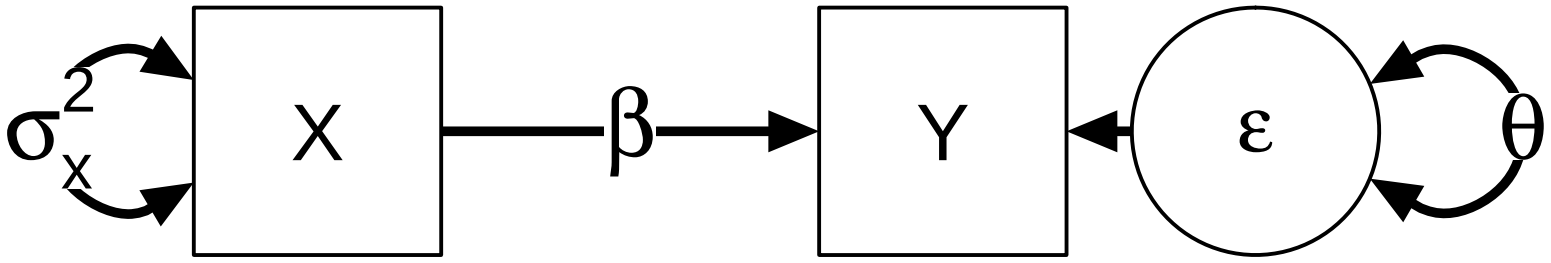


$$y_i = \beta x_i$$

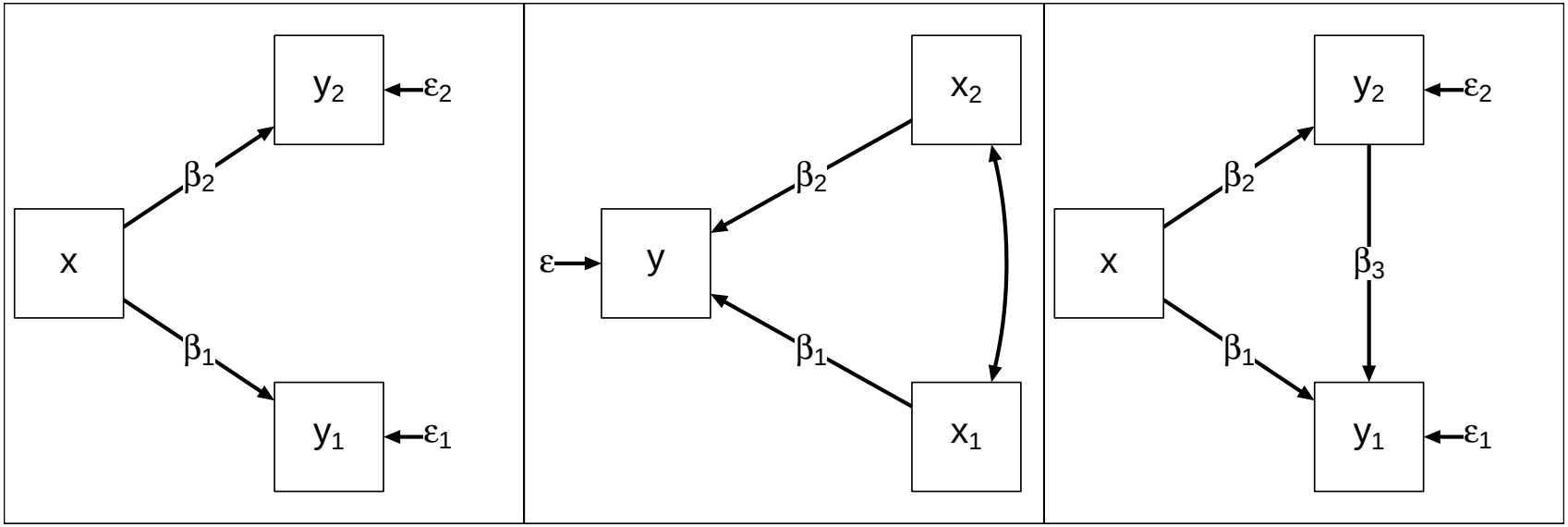


$$y_i = \beta x_i + \varepsilon_i$$

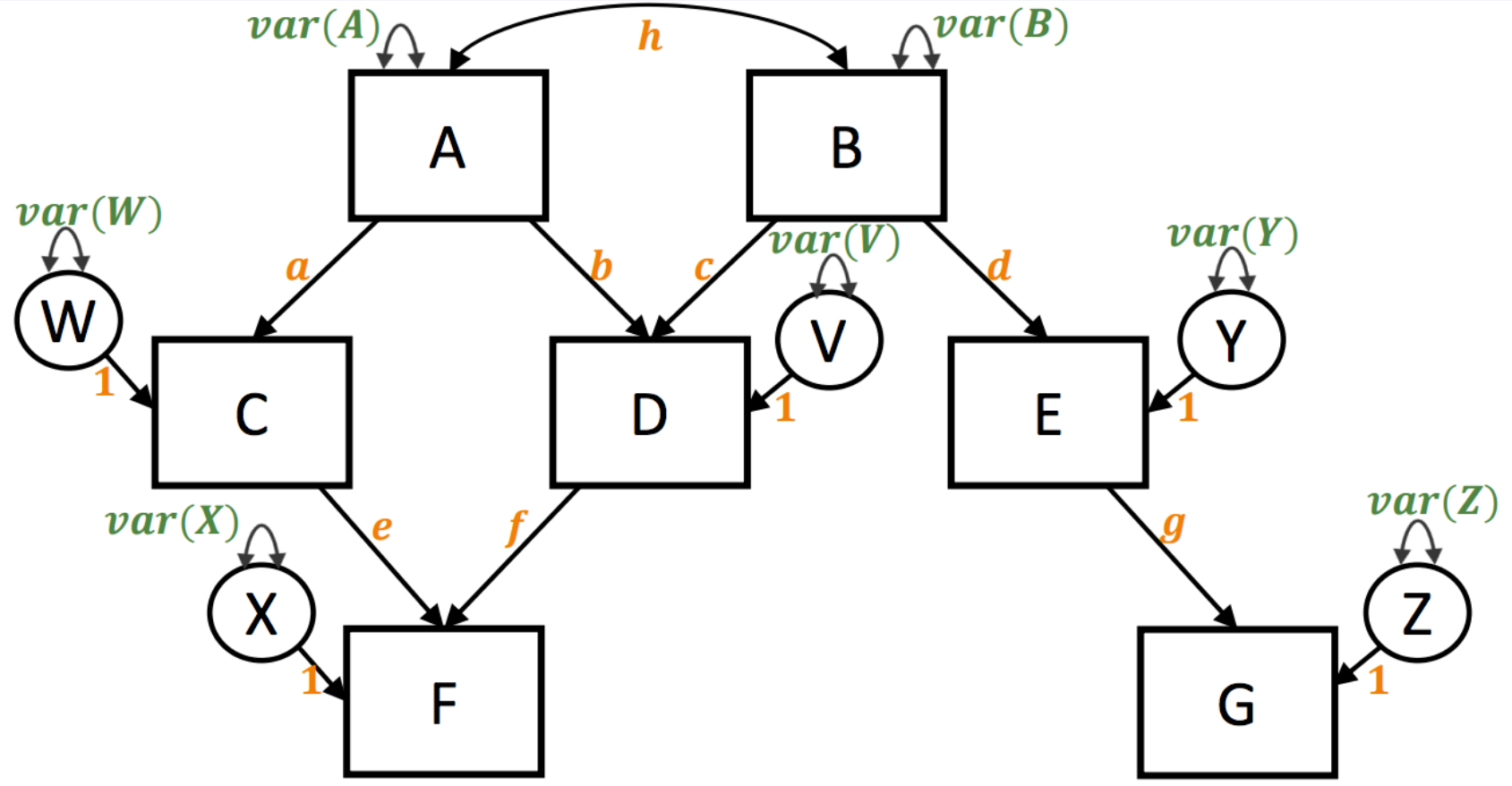
Exogenous variables have a variance (often not drawn)



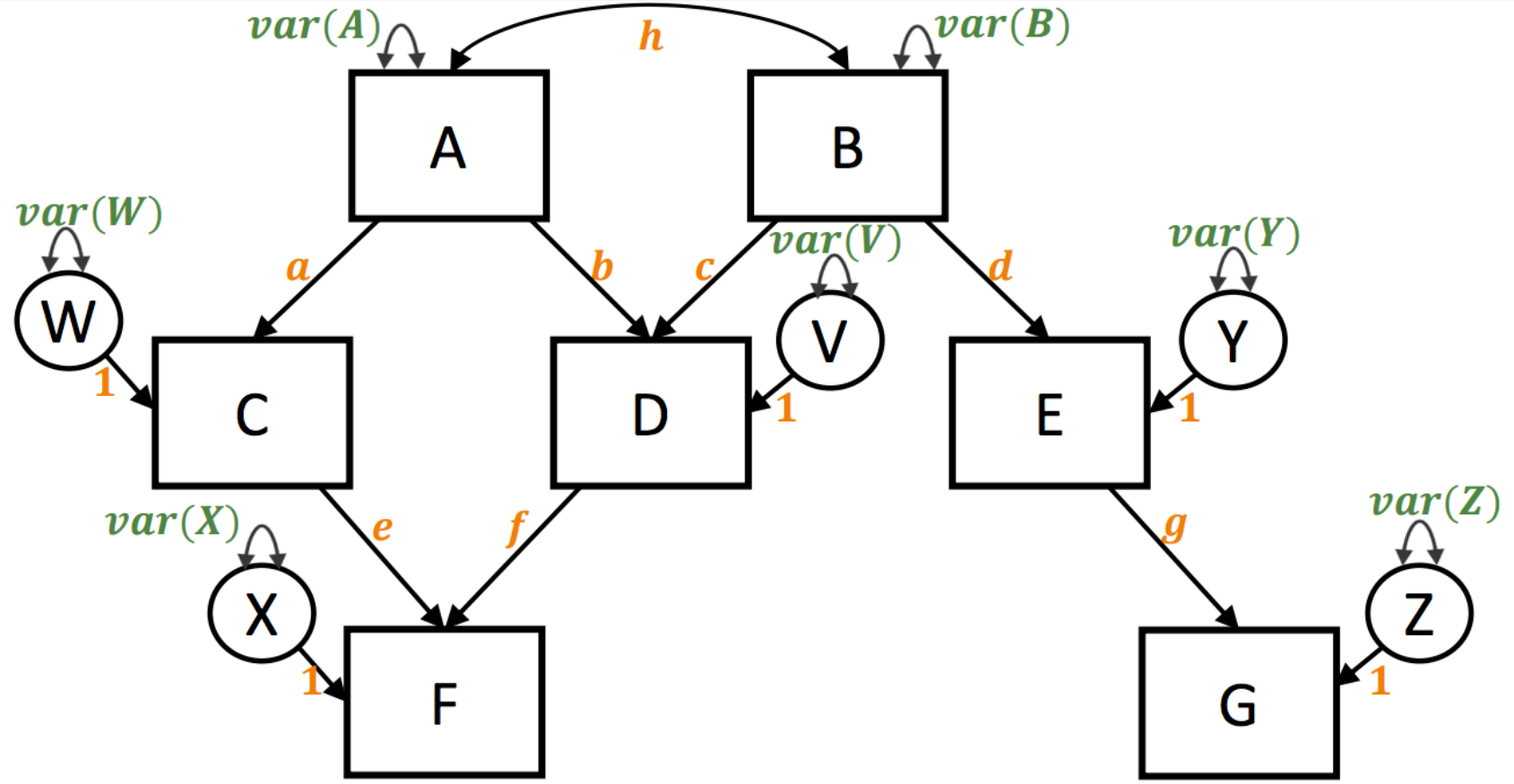
$$y_i = \beta x_i + \varepsilon_i$$
$$x \sim N(\mu_x, \sigma_x)$$
$$\varepsilon \sim N(0, \theta)$$



Goal: derive a model to explain (co)variances related to observed endogenous variables (including covariances between observed endogenous and exogenous variables), using parameters and (co)variances of observed and unobserved exogenous variables.



How many degrees of freedom?



How many degrees of freedom?

- $7 \times 8/2 = 28$ observed variances and covariances
- 7 regressions +7 variances +1 covariance = 15 parameters
- 13 degrees of freedom

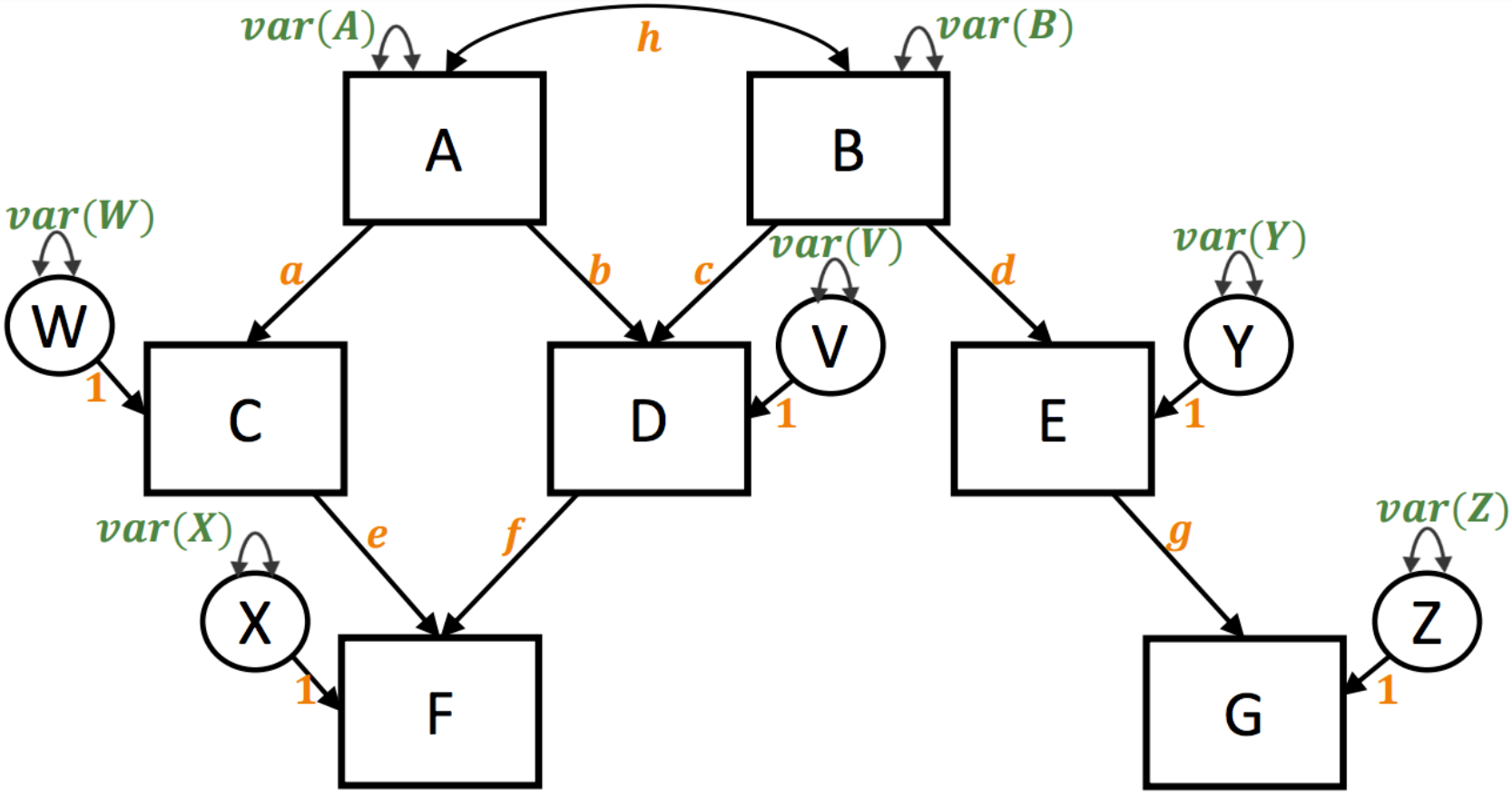
Wright's path tracing Rules

The covariance between any two variables can be expressed as the sum of the compound paths connecting them.
To obtain a compound path:

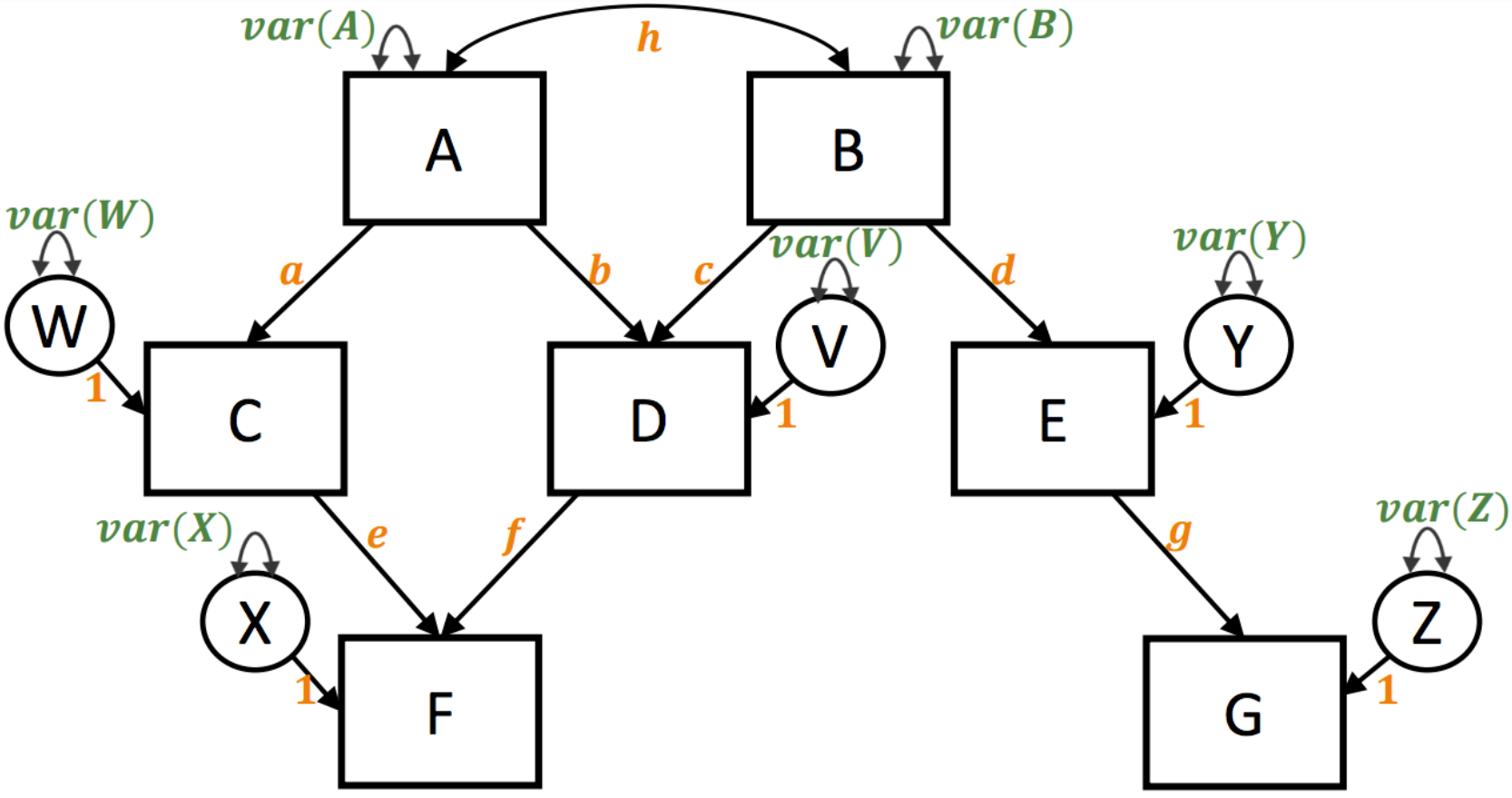
- Trace backwards, change direction at a two-headed arrow, then trace forwards
- Do not go forward and then backward
  - You can never pass out of one arrow head and into another arrowhead: heads-tails, or tails-heads, not heads-heads
- Must contain one, and only one, variance or covariance (bidirectional edge)

Then to obtain the implied (co)variance:

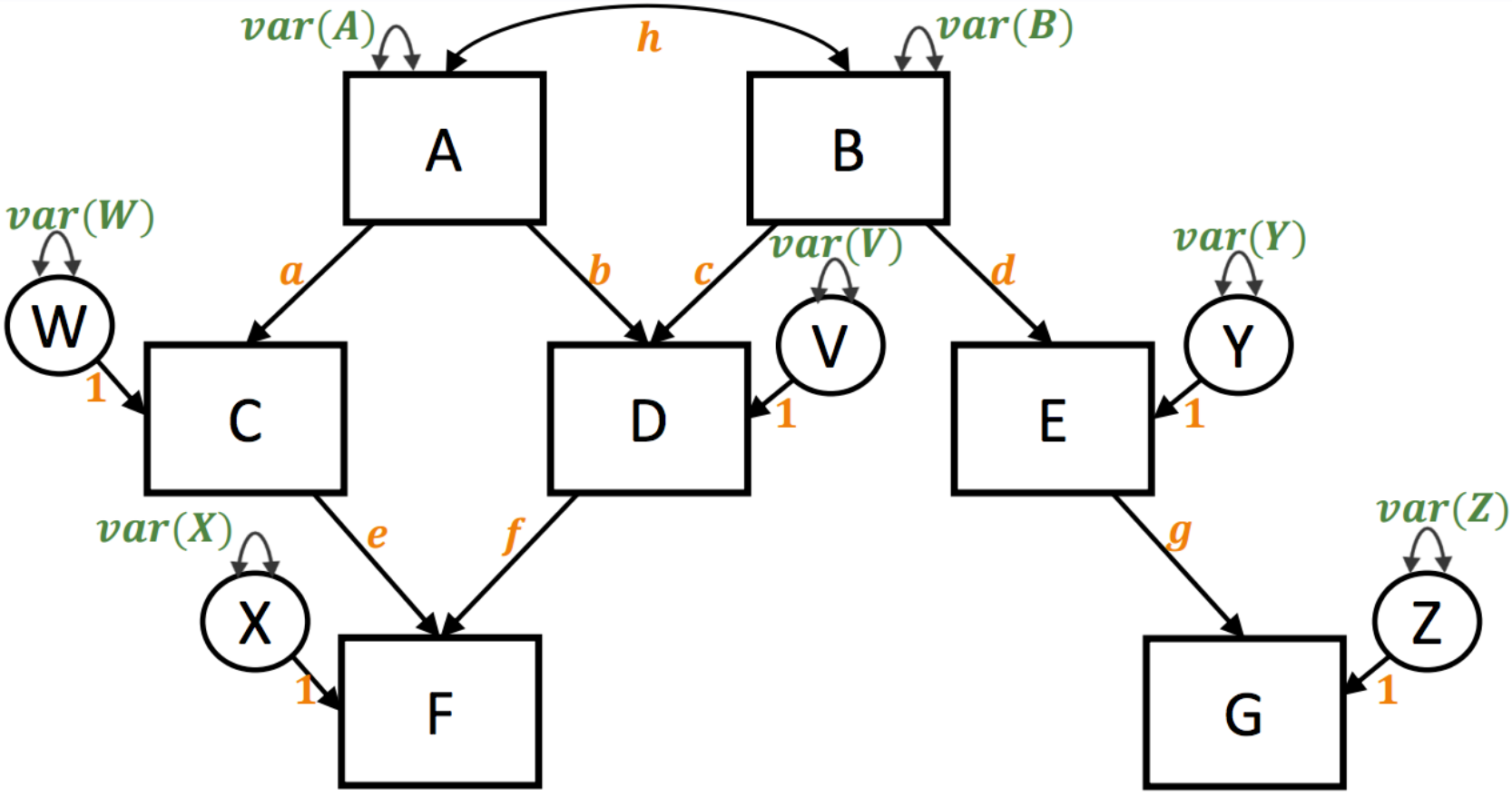
- Compute the product of coefficients in each route between the variables of interest
- Sum over all distinct routes, where pathways are considered distinct if they contain different coefficients, or encounter those coefficients in a different order



Compound paths between A and B:

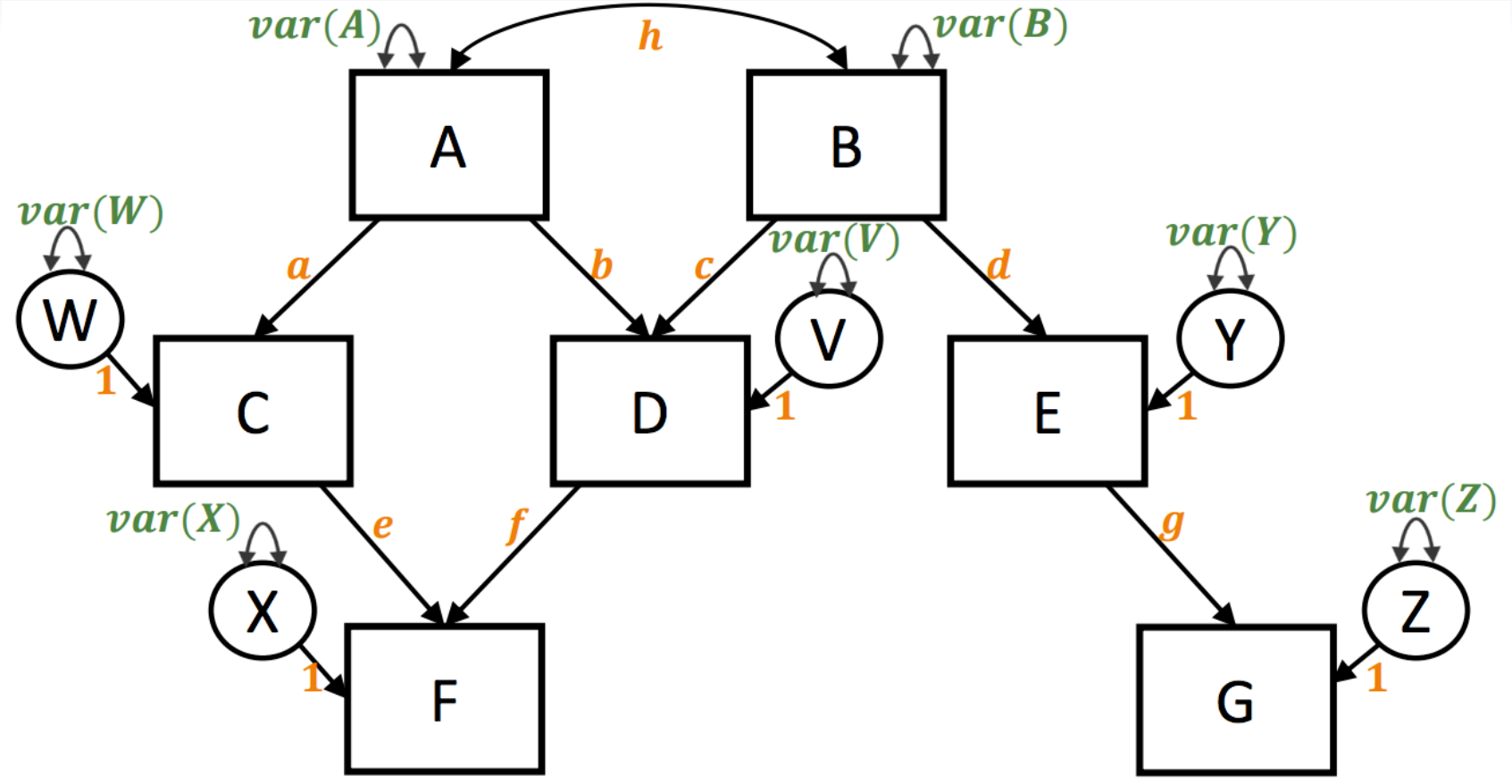


Compound paths between A and B: $A \leftrightarrow B$

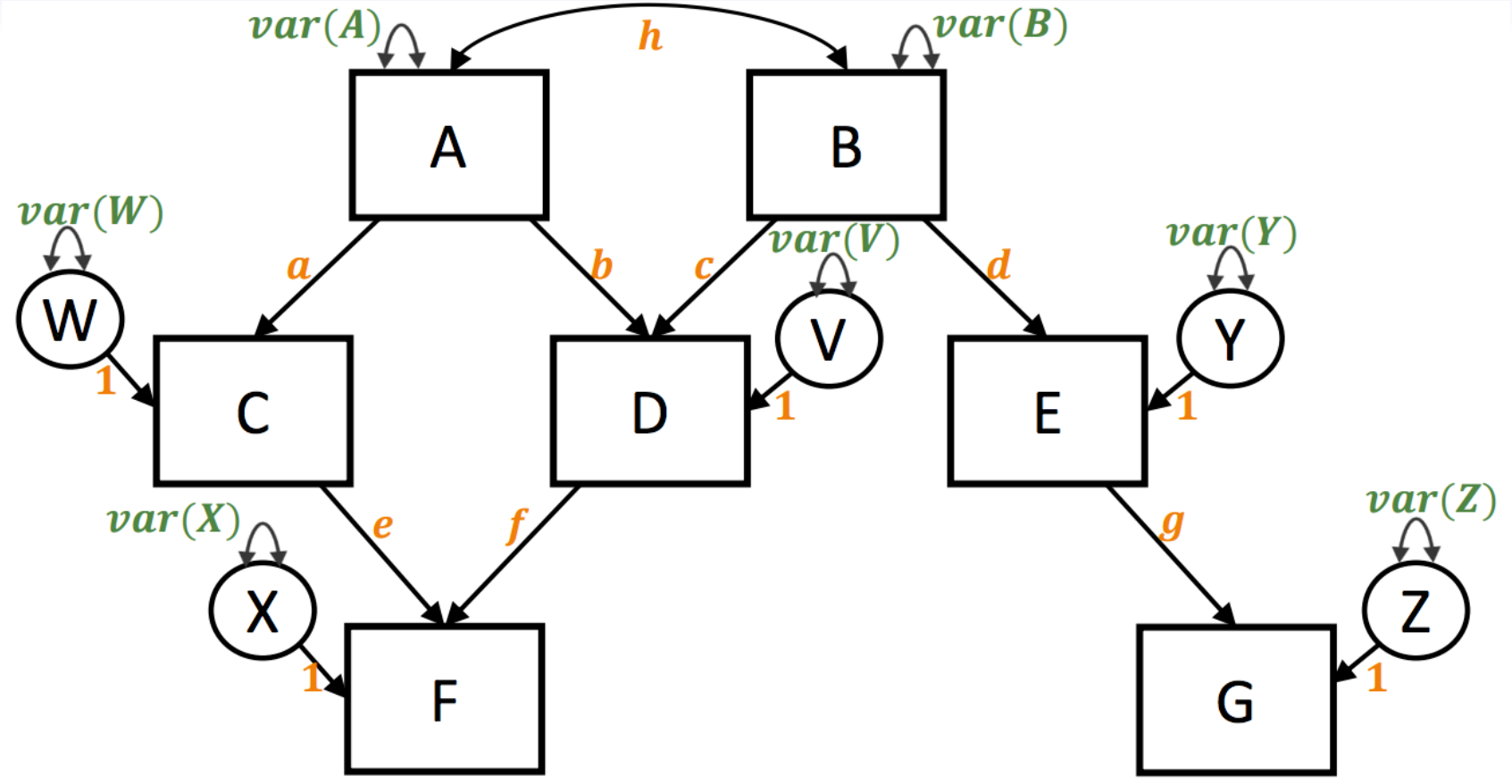


Compound paths between A and B: $A \leftrightarrow B$

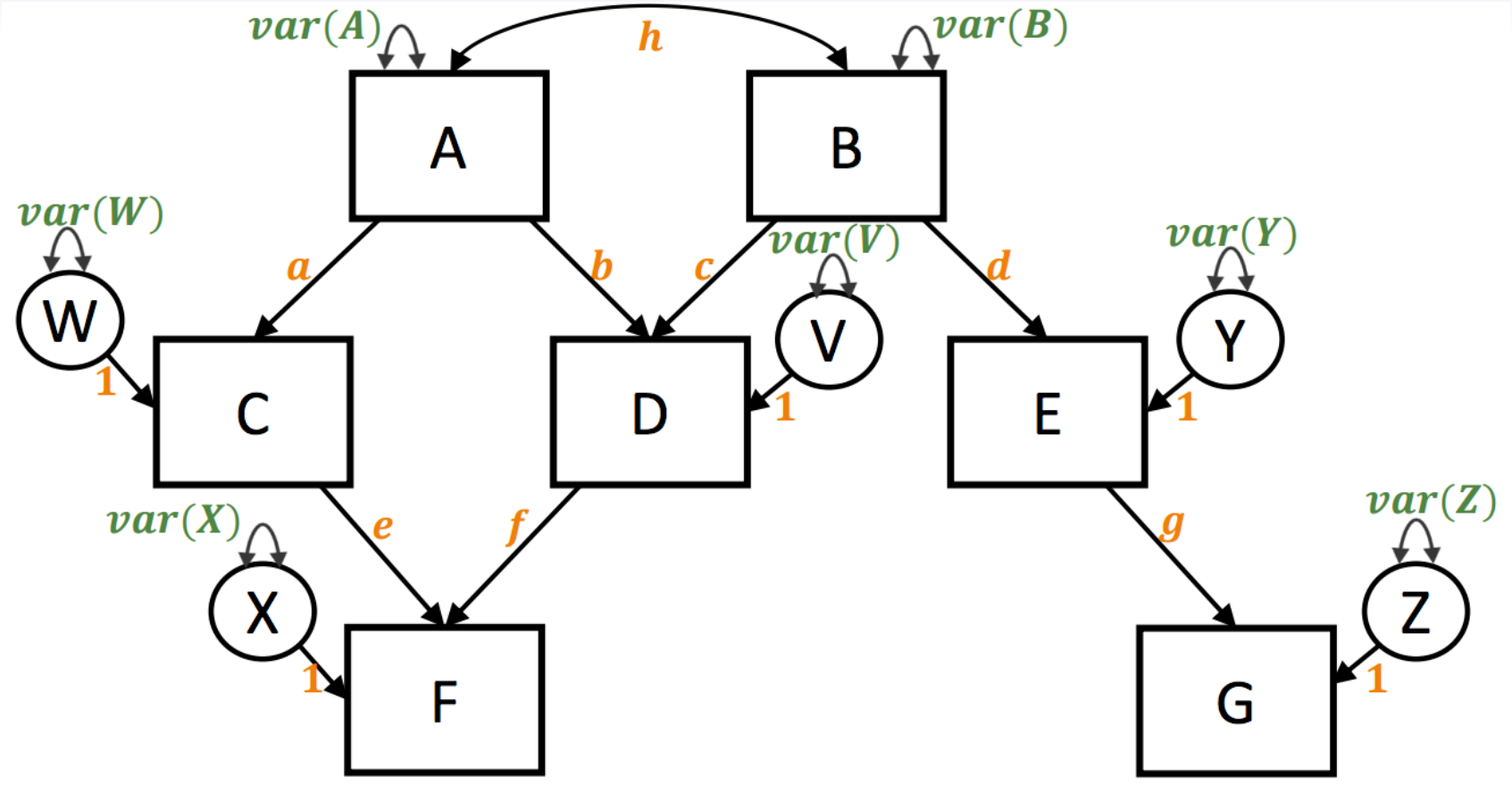
$$\mathrm{Cov}\,(A, B) = h$$



Compound paths between C and D:

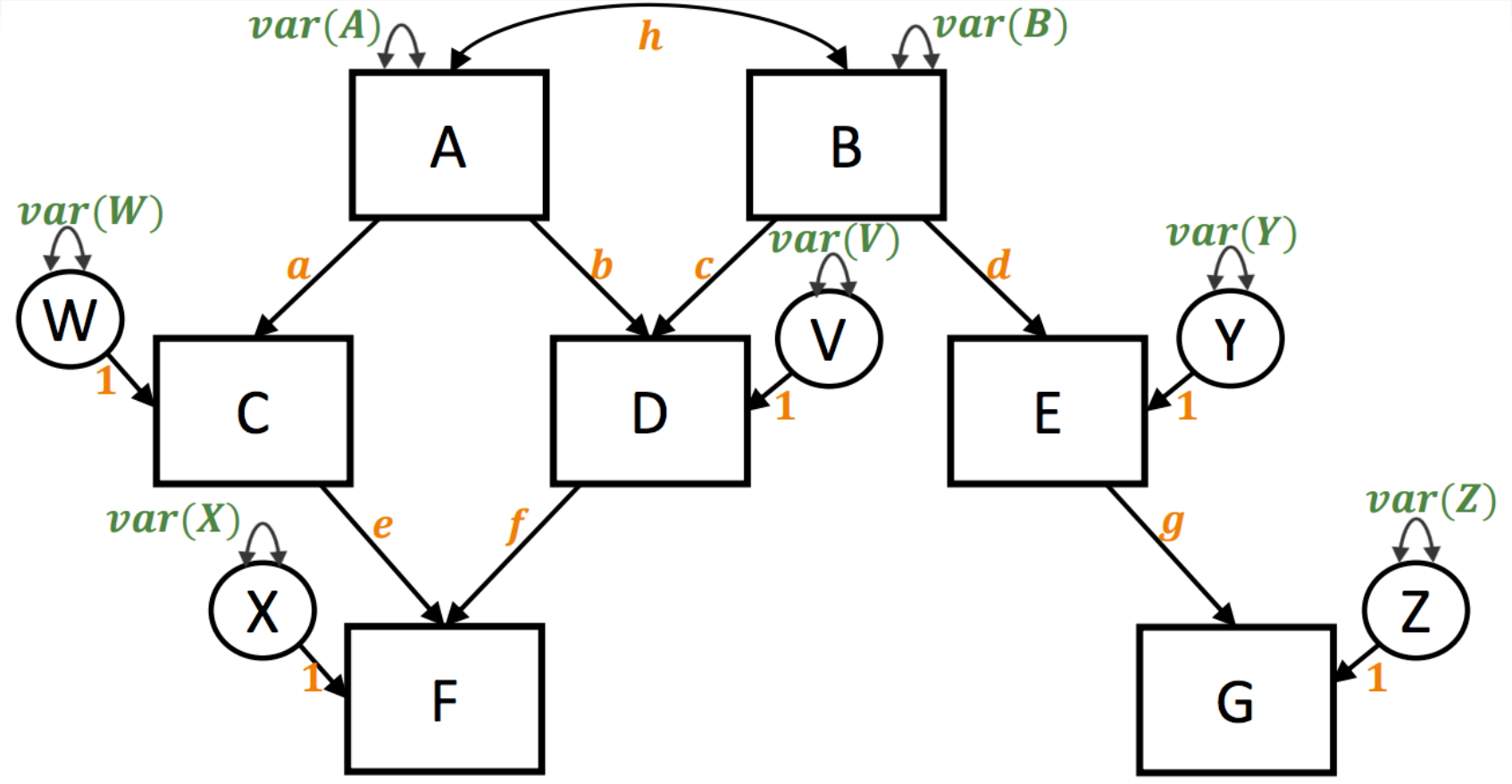


Compound paths between C and D: $C \leftarrow A \leftrightarrow A \rightarrow D$ and $C \leftarrow A \leftrightarrow B \rightarrow D$

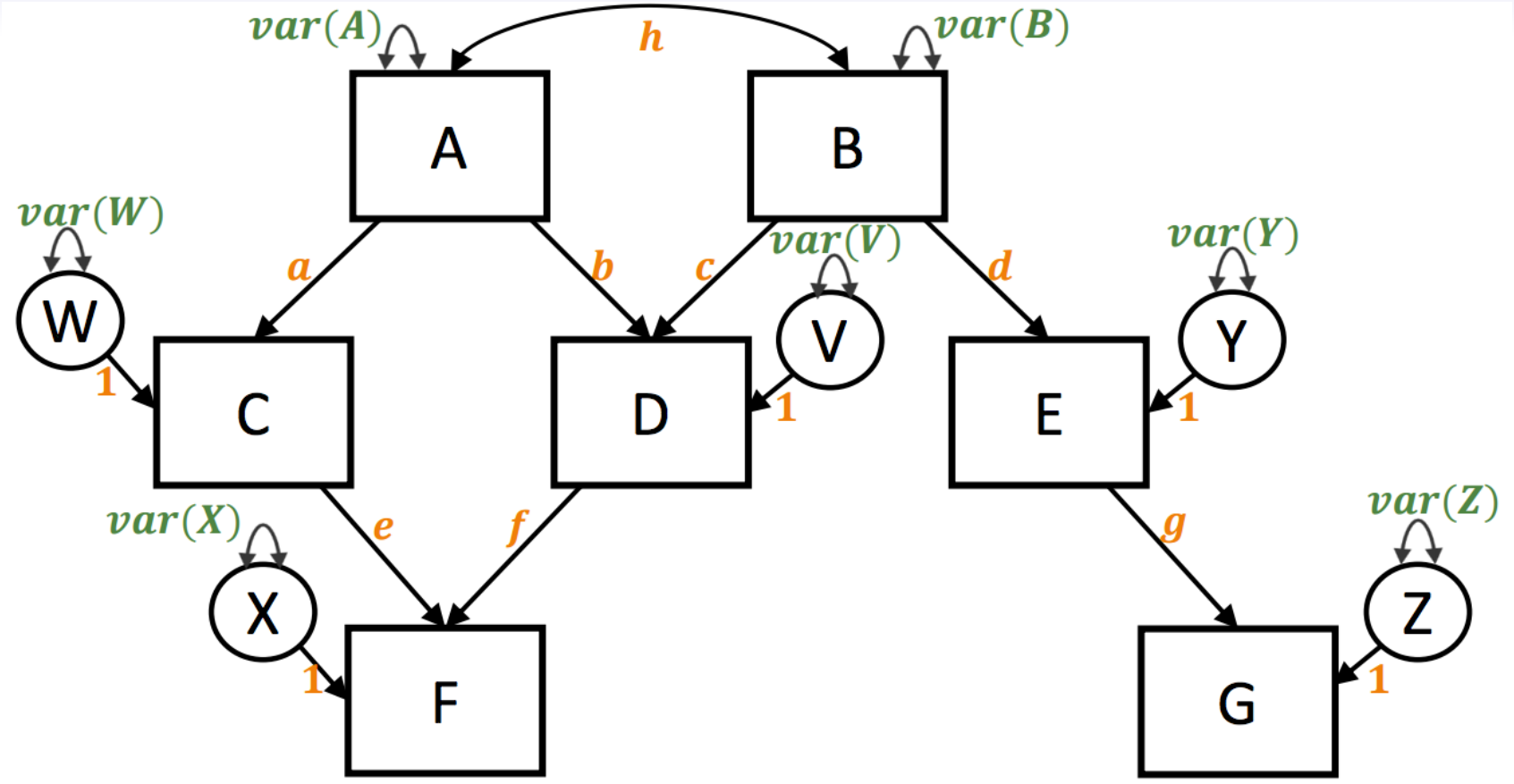


Compound paths between C and D: $C \leftarrow A \leftrightarrow A \rightarrow D$ and $C \leftarrow A \leftrightarrow B \rightarrow D$

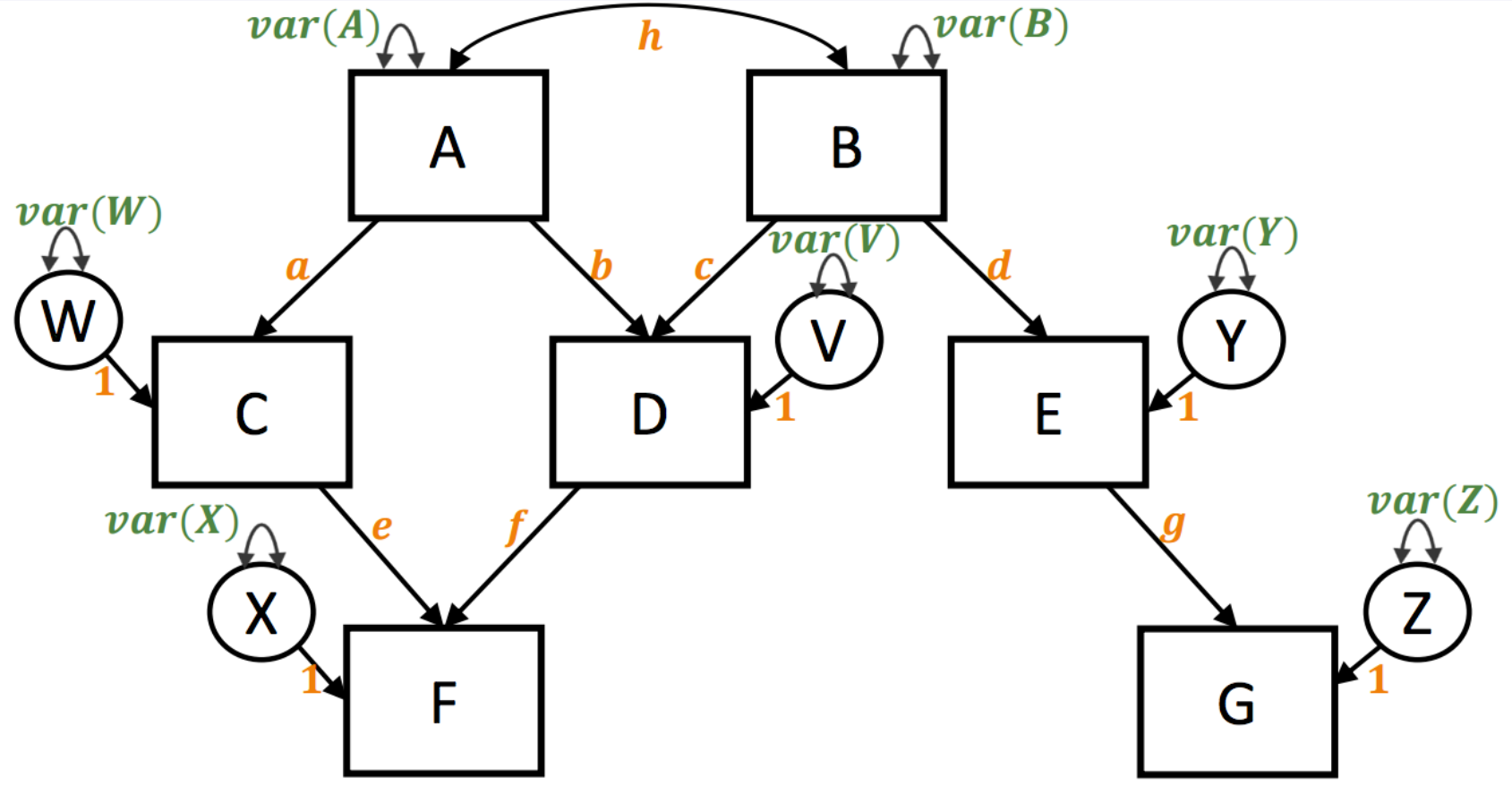
$$\mathrm{Cov}\,(C, D) = a(\mathrm{var}(A))b + ahc$$



Compound paths between C and itself:

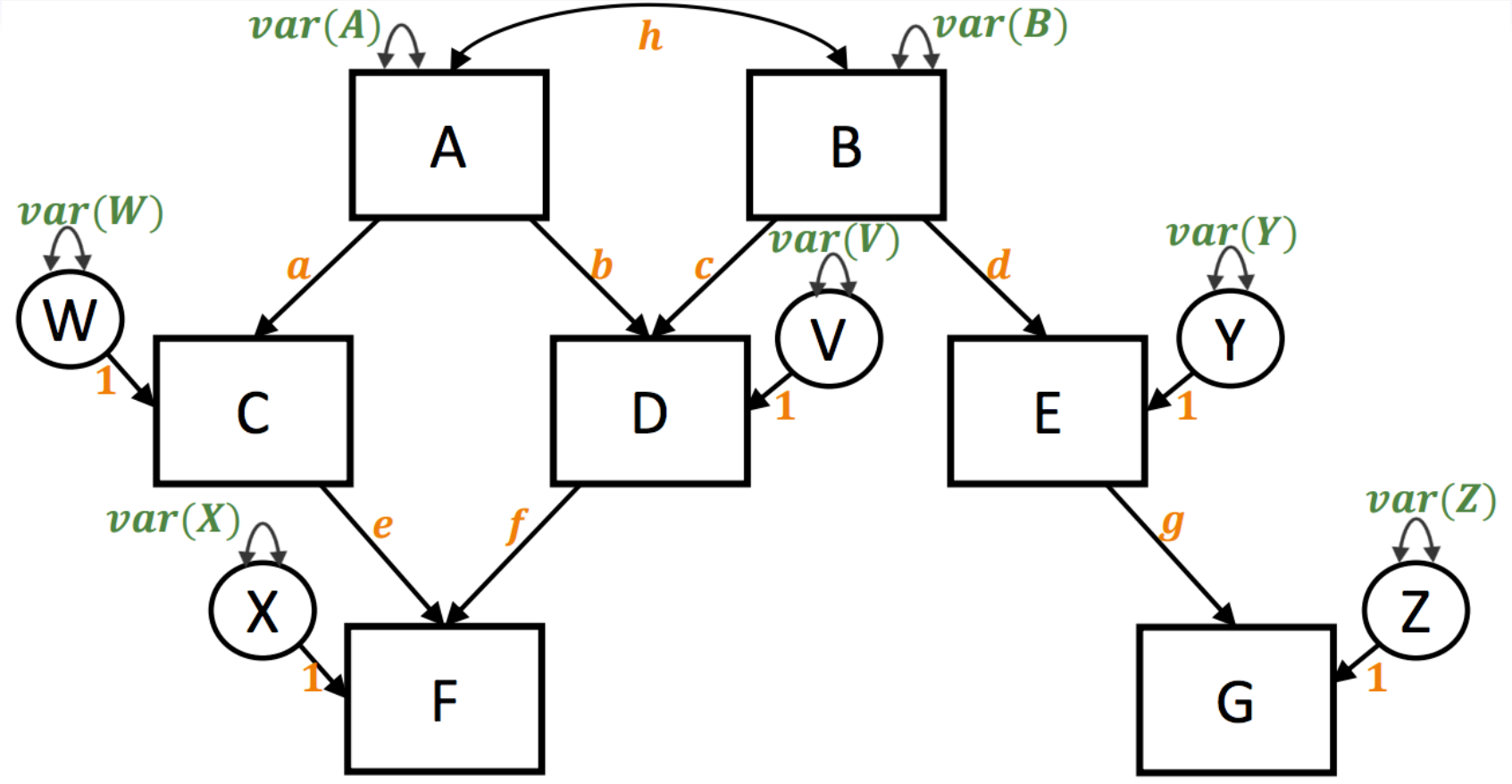


Compound paths between C and itself: $C \leftarrow W \leftrightarrow W \rightarrow C$ and $C \leftarrow A \leftrightarrow A \rightarrow C$

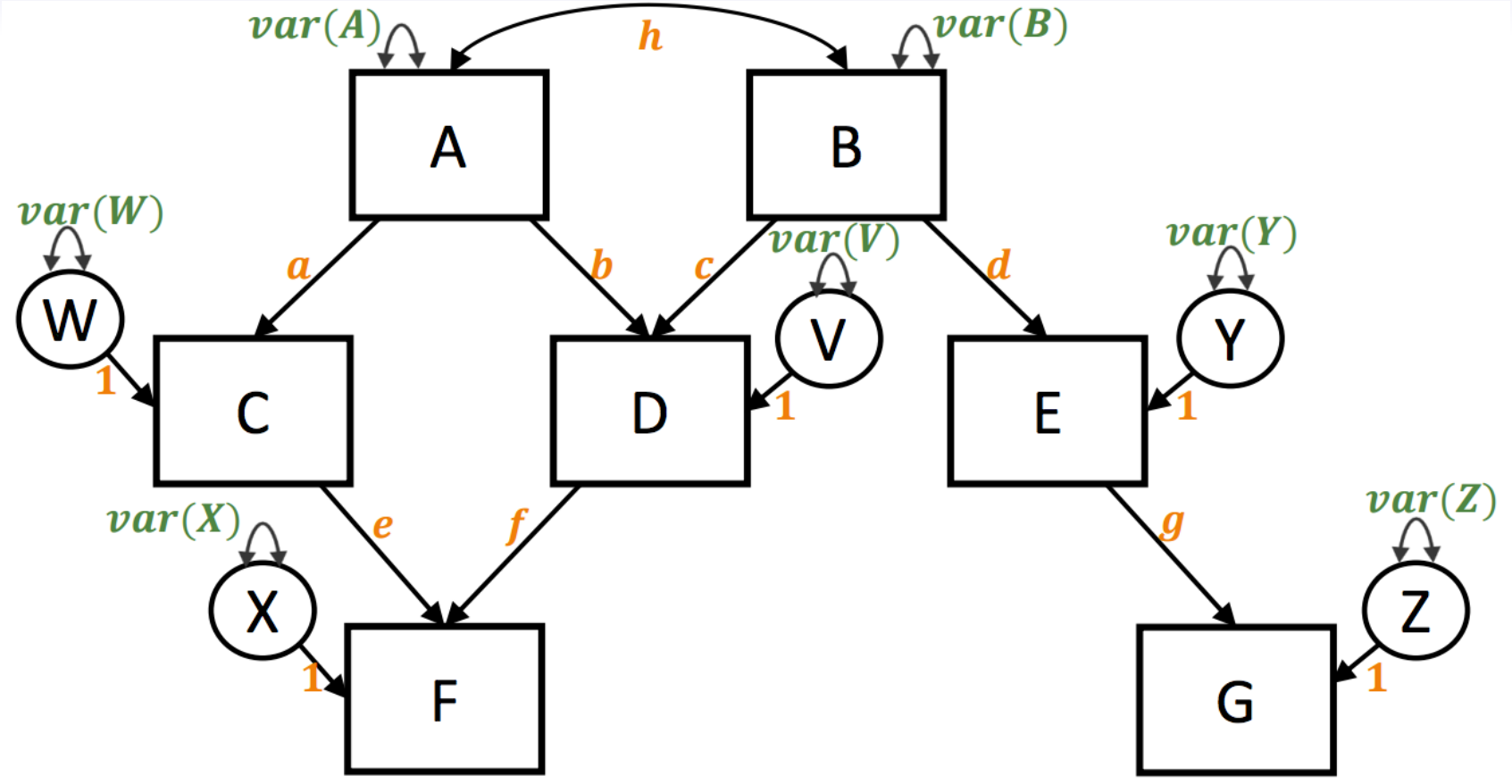


Compound paths between C and itself: $C \leftarrow W \leftrightarrow W \rightarrow C$ and $C \leftarrow A \leftrightarrow A \rightarrow C$

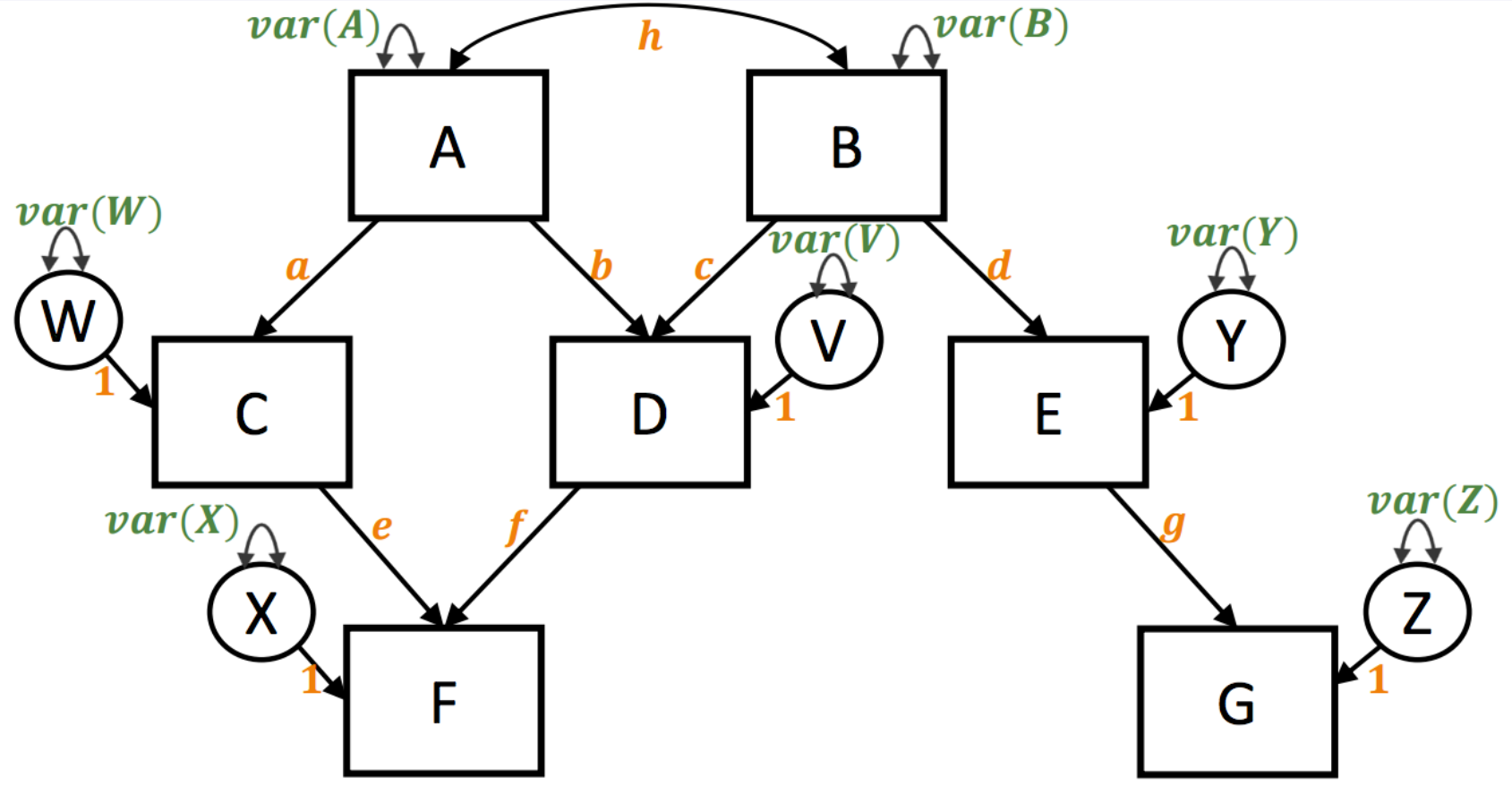
$$\mathrm{Var}\,(C) = a^2(\mathrm{var}(A)) + \mathrm{Var}(W)$$



Compound paths between F and G:



Compound paths between F and G: $F \leftarrow D \leftarrow B \leftrightarrow B \rightarrow E \rightarrow G$, $F \leftarrow C \leftarrow A \leftrightarrow B \rightarrow E \rightarrow G$, and $F \leftarrow D \leftarrow A \leftrightarrow B \rightarrow E \rightarrow G$.

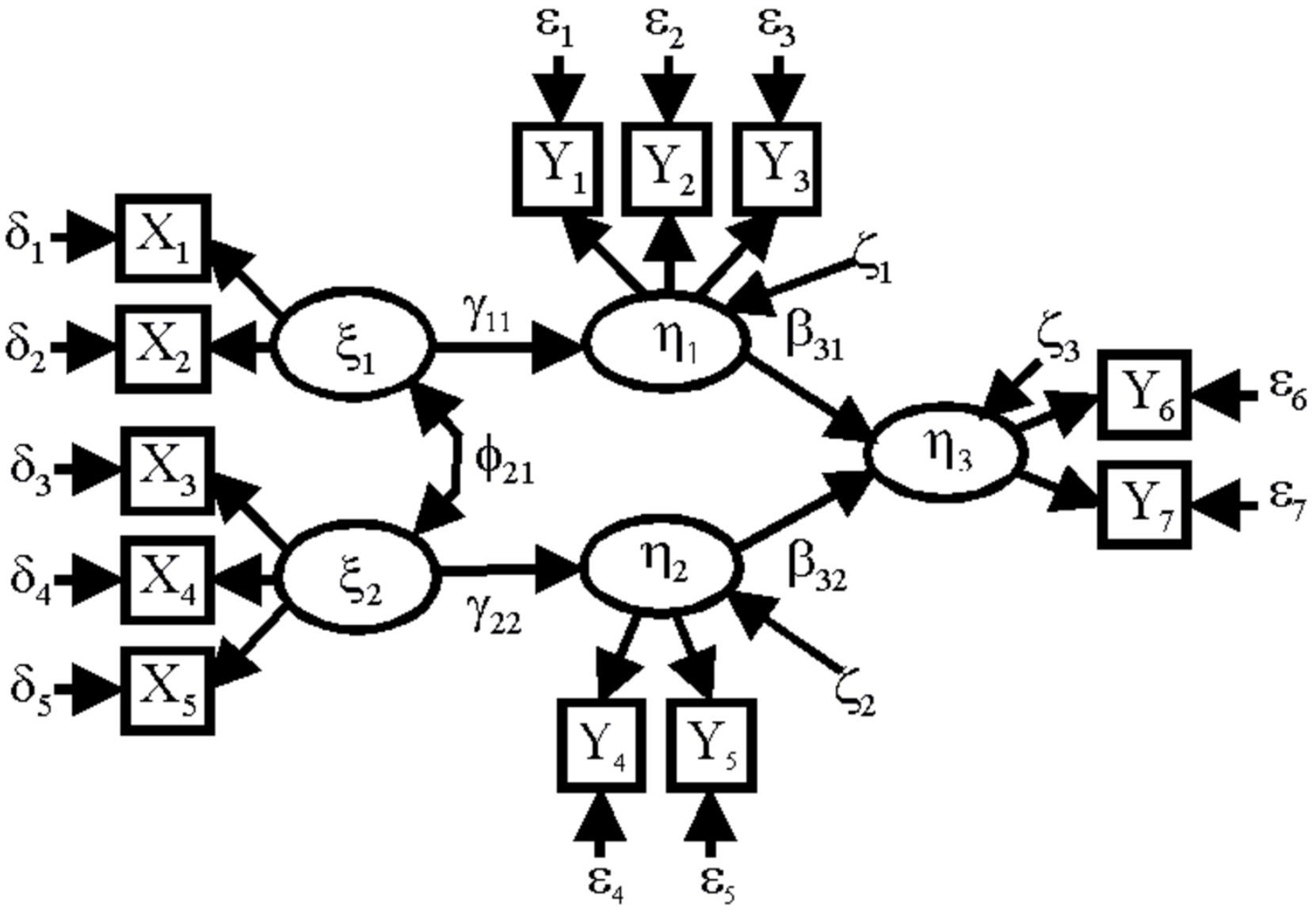


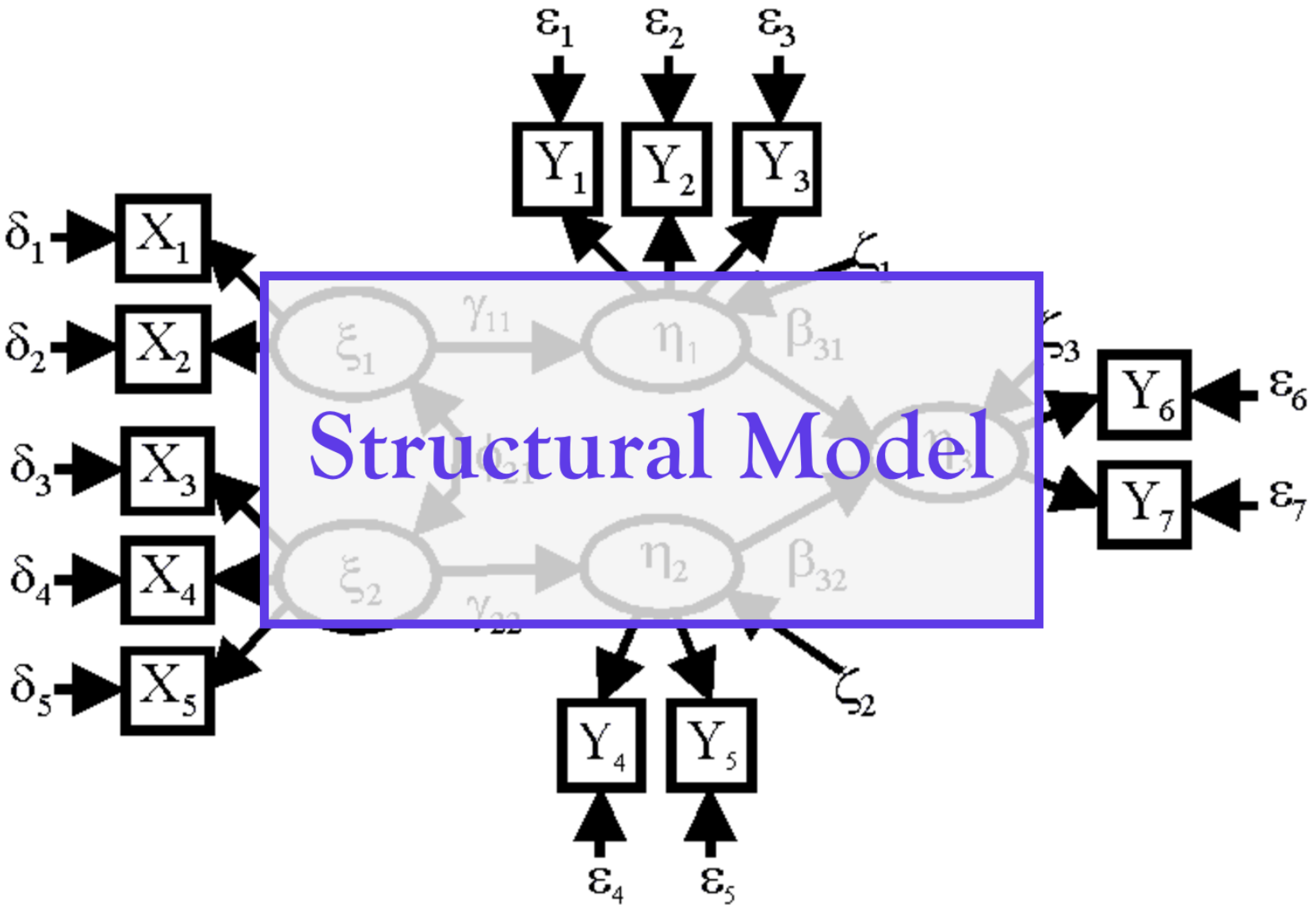
Compound paths between F and G: $F \leftarrow D \leftarrow B \leftrightarrow B \rightarrow E \rightarrow G$, $F \leftarrow C \leftarrow A \leftrightarrow B \rightarrow E \rightarrow G$, and $F \leftarrow D \leftarrow A \leftrightarrow B \rightarrow E \rightarrow G$.

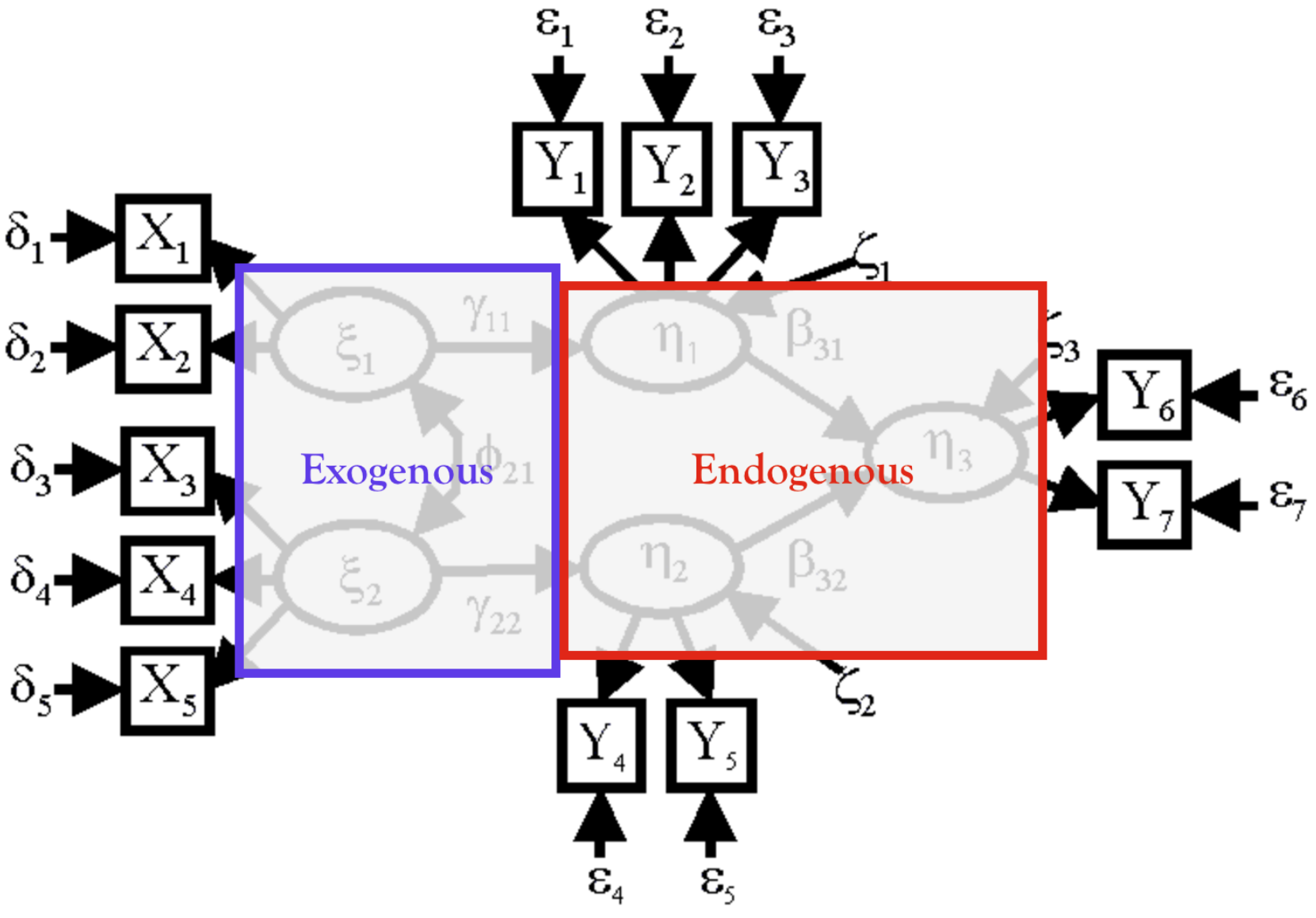
$$\mathrm{Cov}\,(F, G) = fc(\mathrm{Var}(B))dg + fbhdg + eahgd$$

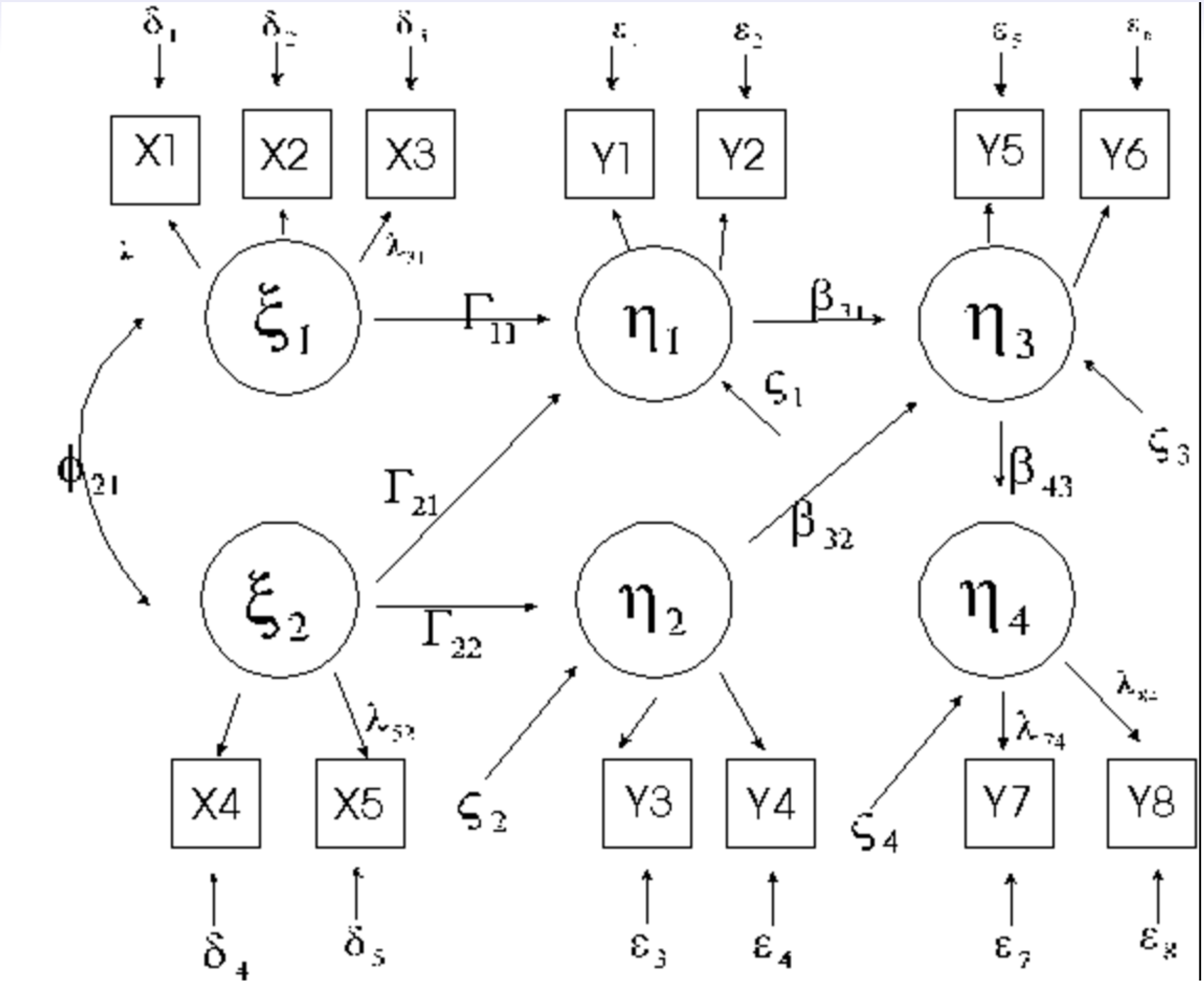
Structural Equation Modeling

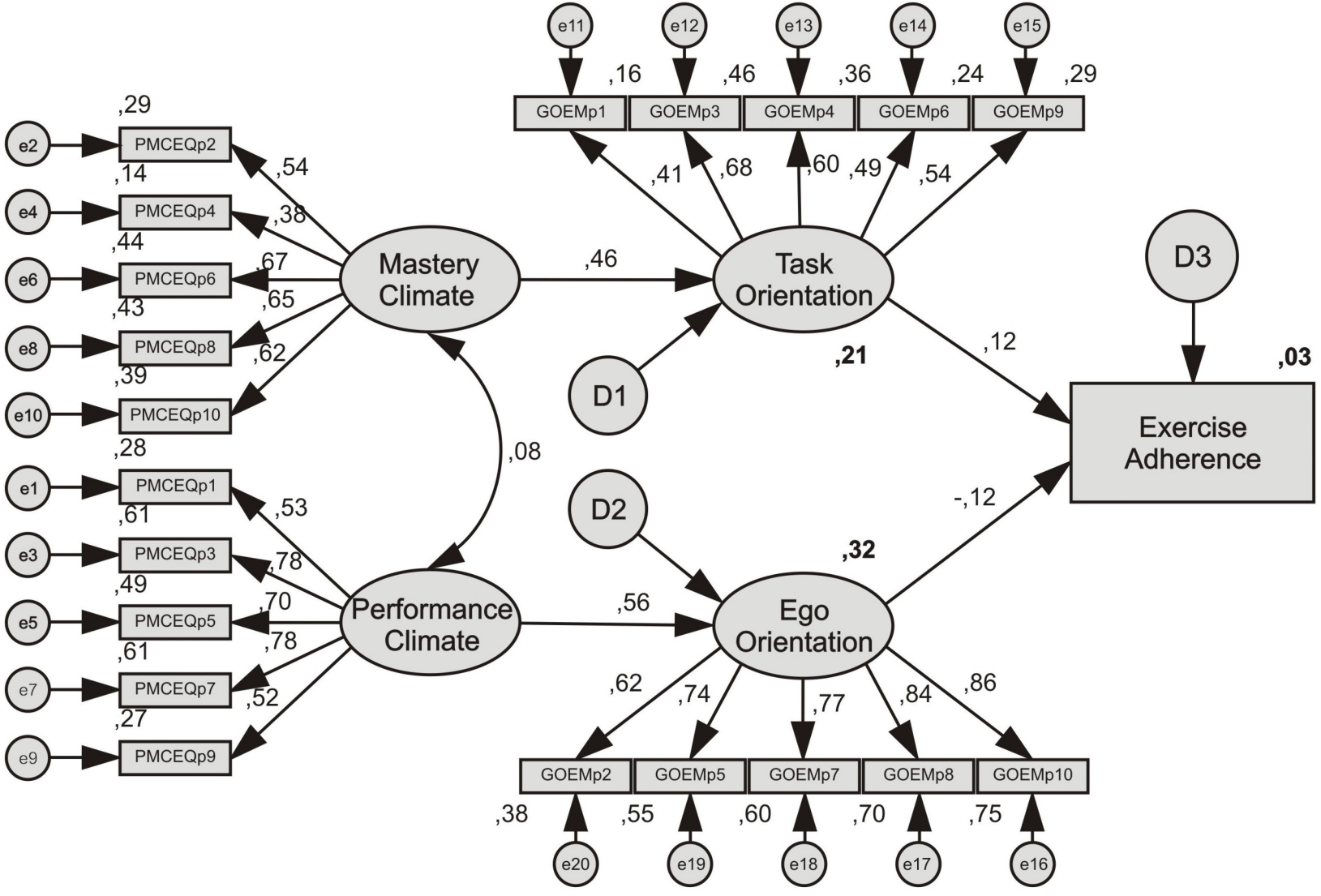
- Structural equation modeling (SEM) extends confirmatory factor analysis (CFA) by modeling the variance–covariance matrix of latent variables with a path model
- Allows one to test causal hypotheses on the latent variables
- Includes path analysis for observed variables:
  - Define one latent per observed variable
  - Set factor loading to 1
  - Set residual variance to 0
- In lavaan: use `sem()` function and define structural relationships using the ~ operator (same as used in regression analysis)
- In psychonetrics: use the `beta` argument in `lvm()` to model direct effects

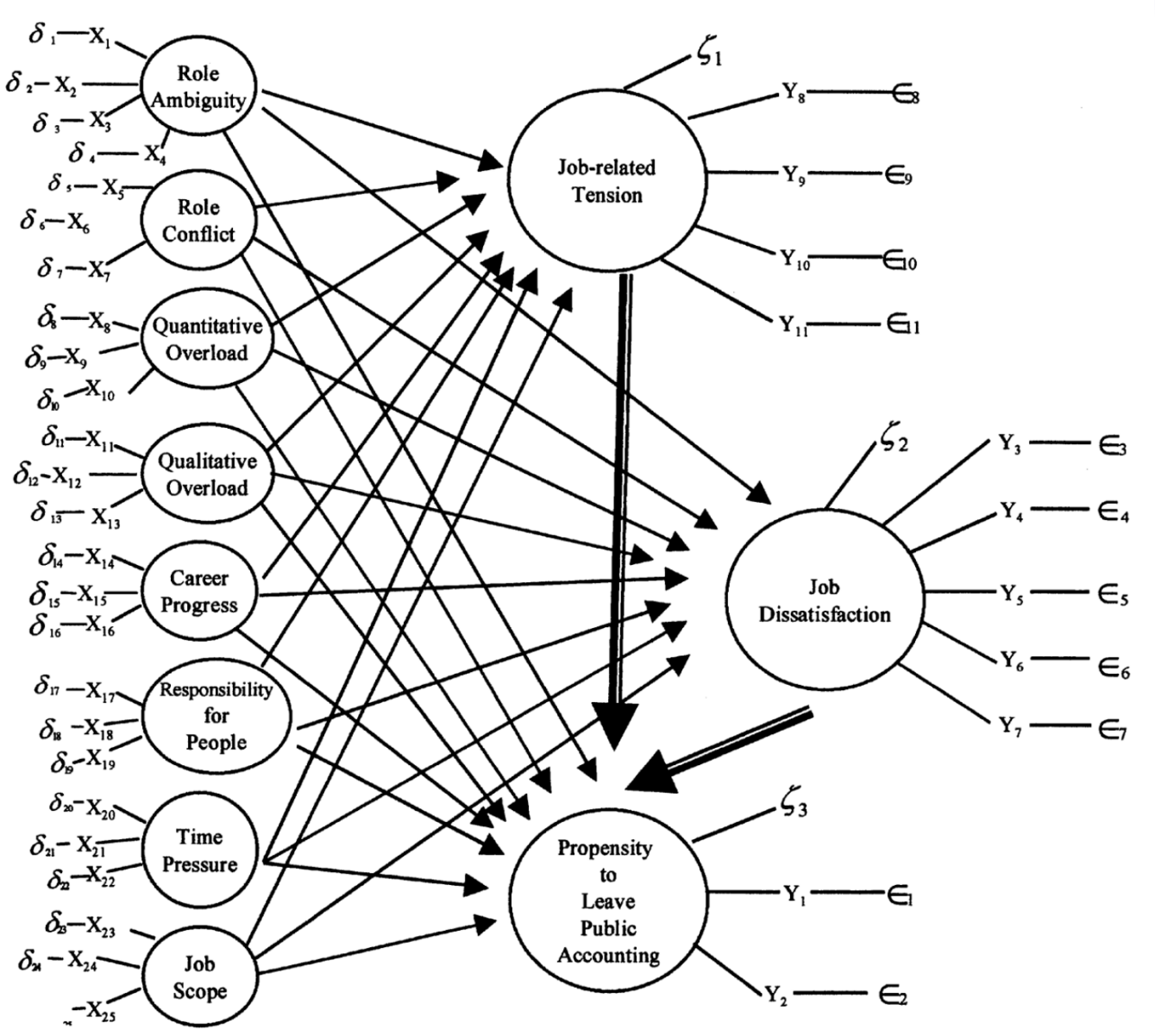

ε1
ε2
ε3
Y1
Y2
Y3
δ1
X1
δ2
X2
ξ1
γ11
η1
ζ1
β31
ζ3
Y6
ε6
η3
ϕ21
δ3
X3
Y7
ε7
δ4
X4
ξ2
η2
β32
γ22
δ5
X5
ζ2
Y4
Y5
ε4
ε5

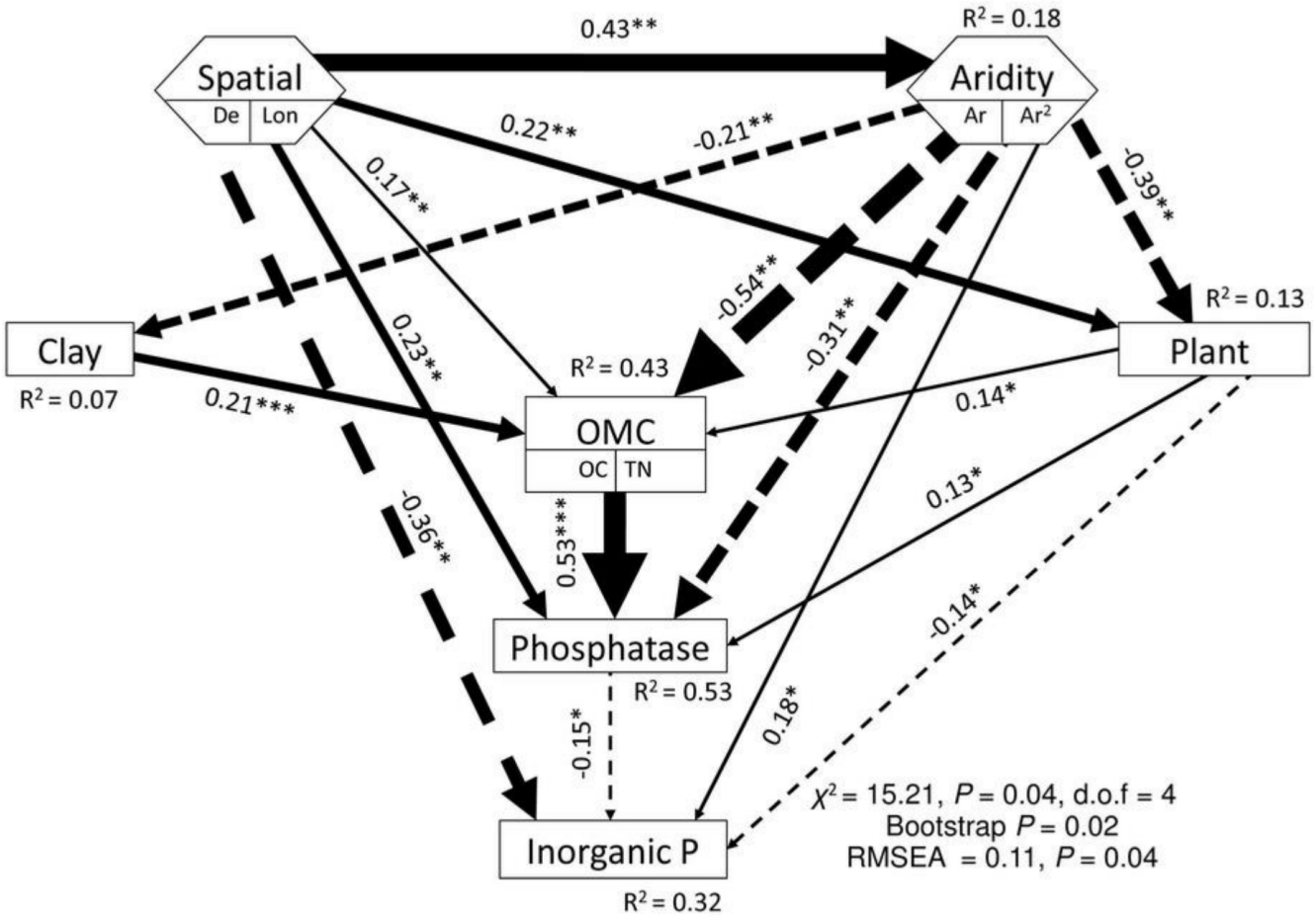

Structural Model

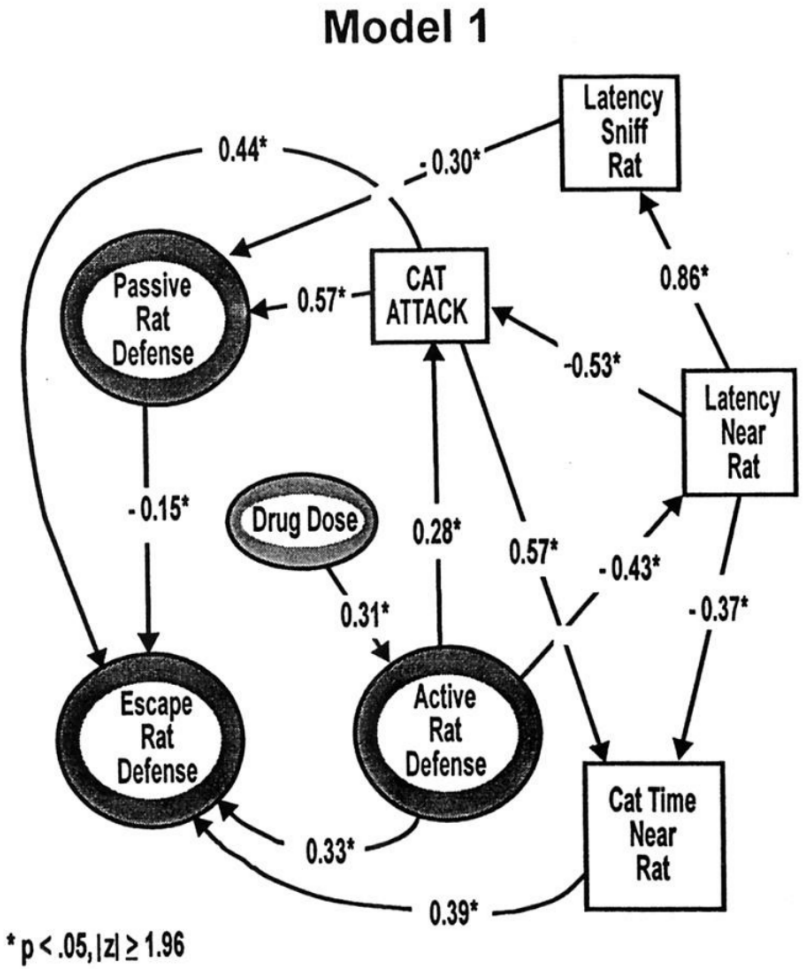

ε1
ε2
ε3
Y1
Y2
Y3
δ1
δ2
δ3
δ4
δ5
X1
X2
X3
X4
X5
ξ1
ξ2
γ11
γ22
ϕ21
Exogenous
η1
η2
η3
β31
β32
ζ1
ζ2
ζ3
Endogenous
Y6
Y7
ε6
ε7
Y4
Y5
ε4
ε5

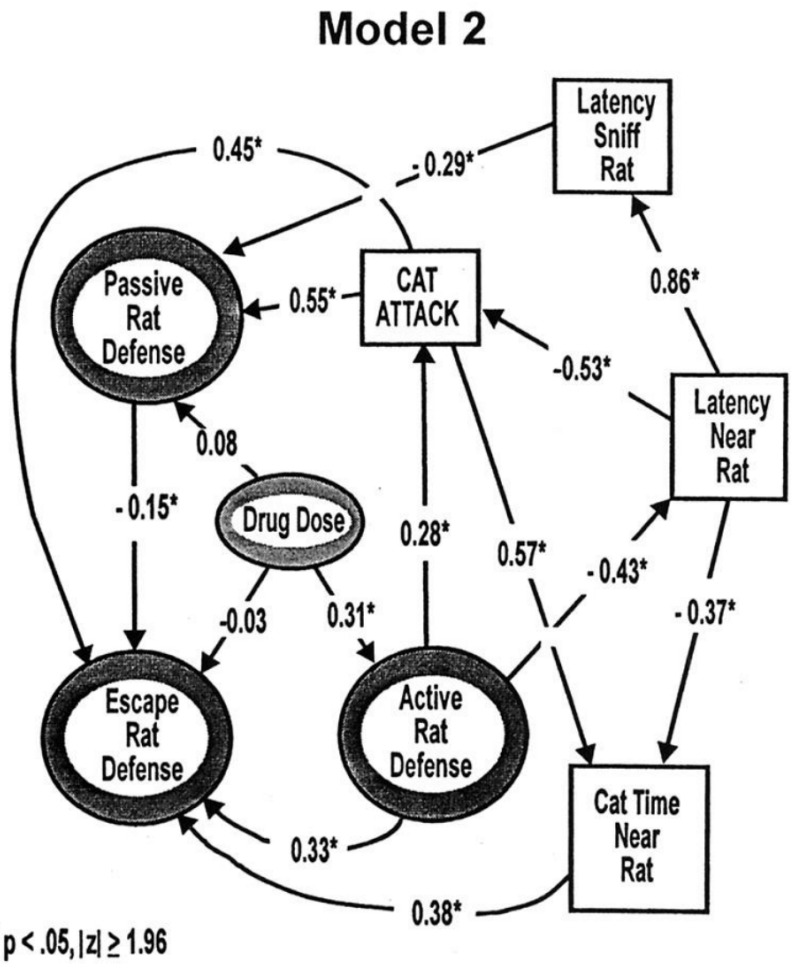

X1
X2
X3
Y1
Y2
Y5
Y6
X4
X5
Y3
Y4
Y7
Y8
ξ_1
ξ_2
η_1
η_2
η_3
η_4
Γ_{11}
Γ_{21}
Γ_{22}
β_{31}
β_{32}
β_{43}
ϕ_{21}
ζ_1
ζ_2
ζ_3
ζ_4
δ_4
δ_5
ε_3
ε_4
ε_7
ε_8


e11
e12
e13
e14
e15
,16
,46
,36
,24
,29
GOEMp1
GOEMp3
GOEMp4
GOEMp6
GOEMp9
,41
,68
,60
,49
,54
,29
e2
PMCEQp2
,54
,14
e4
PMCEQp4
,38
,44
e6
PMCEQp6
,67
,43
e8
PMCEQp8
,65
,62
,39
e10
PMCEQp10
,28
Mastery Climate
,46
Task Orientation
D3
,03
,21
,12
D1
Exercise Adherence
,08
e1
PMCEQp1
,53
,61
D2
-,12
,32
e3
PMCEQp3
,78
,49
e5
PMCEQp5
,70
,78
,61
Performance Climate
,56
Ego Orientation
e7
PMCEQp7
,52
,27
e9
PMCEQp9
,62
,74
,77
,84
,86
GOEMp2
GOEMp5
GOEMp7
GOEMp8
GOEMp10
,38
,55
,60
,70
,75
e20
e19
e18
e17
e16


Role Ambiguity
Role Conflict
Quantitative Overload
Qualitative Overload
Career Progress
Responsibility for People
Time Pressure
Job Scope
Job-related Tension
Job Dissatisfaction
Propensity to Leave Public Accounting


0.43**
R² = 0.18
Spatial
De
Lon
Aridity
Ar
Ar²
0.22**
-0.21**
0.17**
-0.39**
-0.54**
R² = 0.13
Clay
R² = 0.07
0.23**
-0.31**
R² = 0.43
Plant
0.21***
0.14*
OMC
OC
TN
0.13*
-0.36**
0.53***
-0.14*
Phosphatase
R² = 0.53
-0.15*
0.18*
X² = 15.21, P = 0.04, d.o.f = 4
Bootstrap P = 0.02
RMSEA = 0.11, P = 0.04
Inorganic P
R² = 0.32


Model 1
Latency Sniff Rat
Passive Rat Defense
CAT ATTACK
Latency Near Rat
Drug Dose
Escape Rat Defense
Active Rat Defense
Cat Time Near Rat
0.44*
- 0.30*
0.57*
0.86*
-0.53*
- 0.15*
0.28*
0.57*
- 0.43*
- 0.37*
0.31*
0.33*
0.39*
* p < .05, |z| ≥ 1.96


Model 2
Latency Sniff Rat
Passive Rat Defense
CAT ATTACK
Latency Near Rat
Drug Dose
Escape Rat Defense
Active Rat Defense
Cat Time Near Rat
0.45*
- 0.29*
0.55*
0.86*
-0.53*
0.08
- 0.15*
0.28*
0.57*
- 0.43*
- 0.37*
-0.03
0.31*
0.33*
0.38*
* p < .05, |z| ≥ 1.96

Exogenous measurement model (using full LISREL notation, it gets easier!):

$$\begin{aligned}
\boldsymbol{x}_i &= \boldsymbol{\Lambda}_x \boldsymbol{\xi}_i + \boldsymbol{\delta}_i \\
\boldsymbol{x} &\sim N(\mathbf{0}, \boldsymbol{\Sigma}_x) \\
\boldsymbol{\xi} &\sim N(\mathbf{0}, \boldsymbol{\Phi}) \\
\boldsymbol{\delta} &\sim N(\mathbf{0}, \boldsymbol{\Theta}_\delta),
\end{aligned}$$

Allows you to derive the model-implied variance–covariance matrix:

$$\begin{aligned}
\boldsymbol{\Sigma}_x = \mathrm{Var}(\boldsymbol{x}) &= \mathrm{Var}(\boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}) \\
&= \boldsymbol{\Lambda}_x \mathrm{Var}(\boldsymbol{\xi}) \boldsymbol{\Lambda}_x^\top + \mathrm{Var}(\boldsymbol{\delta}) \\
&= \boldsymbol{\Lambda}_x \boldsymbol{\Phi} \boldsymbol{\Lambda}_x^\top + \boldsymbol{\Theta}_\delta
\end{aligned}$$

Endogenous model:

$$\begin{aligned}
\boldsymbol{y}_i &= \boldsymbol{\Lambda}_y \boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \\
\boldsymbol{\eta}_i &= \boldsymbol{\Gamma} \boldsymbol{\xi}_i + \boldsymbol{B} \boldsymbol{\eta}_i + \boldsymbol{\zeta}_i \\
\boldsymbol{y} &\sim N(\mathbf{0}, \boldsymbol{\Sigma}_y) \\
\boldsymbol{\xi} &\sim N(\mathbf{0}, \boldsymbol{\Phi}) \\
\boldsymbol{\zeta} &\sim N(\mathbf{0}, \boldsymbol{\Psi}), \text{Note that } \boldsymbol{\Psi} \text{ is now diagonal!} \\
\boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \boldsymbol{\Theta}_\varepsilon)
\end{aligned}$$

Only different from CFA model in the added regression parameters $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$. Note that $\boldsymbol{\eta}_i$ appears twice in the structural model, so let's first solve that:

$$\begin{aligned}
\boldsymbol{\eta}_i &= \boldsymbol{\Gamma} \boldsymbol{\xi}_i + \boldsymbol{B} \boldsymbol{\eta}_i + \boldsymbol{\zeta}_i \\
\boldsymbol{\eta}_i - \boldsymbol{B} \boldsymbol{\eta}_i &= \boldsymbol{\Gamma} \boldsymbol{\xi}_i + \boldsymbol{\zeta}_i \\
(\boldsymbol{I} - \boldsymbol{B}) \boldsymbol{\eta}_i &= \boldsymbol{\Gamma} \boldsymbol{\xi}_i + \boldsymbol{\zeta}_i \\
\boldsymbol{\eta}_i &= (\boldsymbol{I} - \boldsymbol{B})^{-1} \boldsymbol{\Gamma} \boldsymbol{\xi}_i + (\boldsymbol{I} - \boldsymbol{B})^{-1} \boldsymbol{\zeta}_i
\end{aligned}$$

Now:

$$\begin{aligned}\mathrm{Var}(\boldsymbol{\eta}) &= \mathrm{Var}\left((\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\xi}_i + (\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\zeta}_i\right) \\ &= \mathrm{Var}\left((\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\xi}_i\right) + \mathrm{Var}\left((\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\zeta}_i\right) \\ &= (\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi}\left((\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\Gamma}\right)^{\top} + (\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\Psi}(\boldsymbol{I}-\boldsymbol{B})^{-1\top}\end{aligned}$$

Which can be used in:

$$\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{\Lambda}_y \mathrm{Var}(\boldsymbol{\eta})\boldsymbol{\Lambda}_y^{\top} + \boldsymbol{\Theta}_{\varepsilon}$$

And $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})$ can similarly be derived. Way too complicated...

All-y notation

Much easier, just treat exogenous variables as endogenous variables. All latents are then contained in $\boldsymbol{\eta}$ and all indicators in $\boldsymbol{y}$. Only important to note is that $\boldsymbol{\Psi}$ then contains both exogenous variances and covariances (all freely estimated) as well as latent residual variances (usually without covariances).

All-y model:

$$
\begin{aligned}
\boldsymbol{y}_i &= \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \\
\boldsymbol{\eta}_i &= \boldsymbol{B}\boldsymbol{\eta}_i + \boldsymbol{\zeta}_i \\
&= (\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\zeta}_i \\
\boldsymbol{y} &\sim N(\boldsymbol{0}, \boldsymbol{\Sigma}) \\
\boldsymbol{\zeta} &\sim N(\boldsymbol{0}, \boldsymbol{\Psi}) \\
\boldsymbol{\varepsilon} &\sim N(\boldsymbol{0}, \boldsymbol{\Theta})
\end{aligned}
$$

Results in:

$$
\begin{aligned}
\boldsymbol{\Sigma} = \operatorname{Var}(\boldsymbol{y}) &= \operatorname{Var}\left(\boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}\right) \\
&= \operatorname{Var}\left(\boldsymbol{\Lambda}\boldsymbol{\eta}\right) + \operatorname{Var}\left(\boldsymbol{\varepsilon}\right) \\
&= \boldsymbol{\Lambda}\operatorname{Var}\left(\boldsymbol{\eta}\right)\boldsymbol{\Lambda}^\top + \boldsymbol{\Theta} \\
&= \boldsymbol{\Lambda}\operatorname{Var}\left((\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\zeta}\right)\boldsymbol{\Lambda}^\top + \boldsymbol{\Theta} \\
&= \boldsymbol{\Lambda}(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Psi}(\boldsymbol{I} - \boldsymbol{B})^{-1\top}\boldsymbol{\Lambda}^\top + \boldsymbol{\Theta}
\end{aligned}
$$

SEM model:

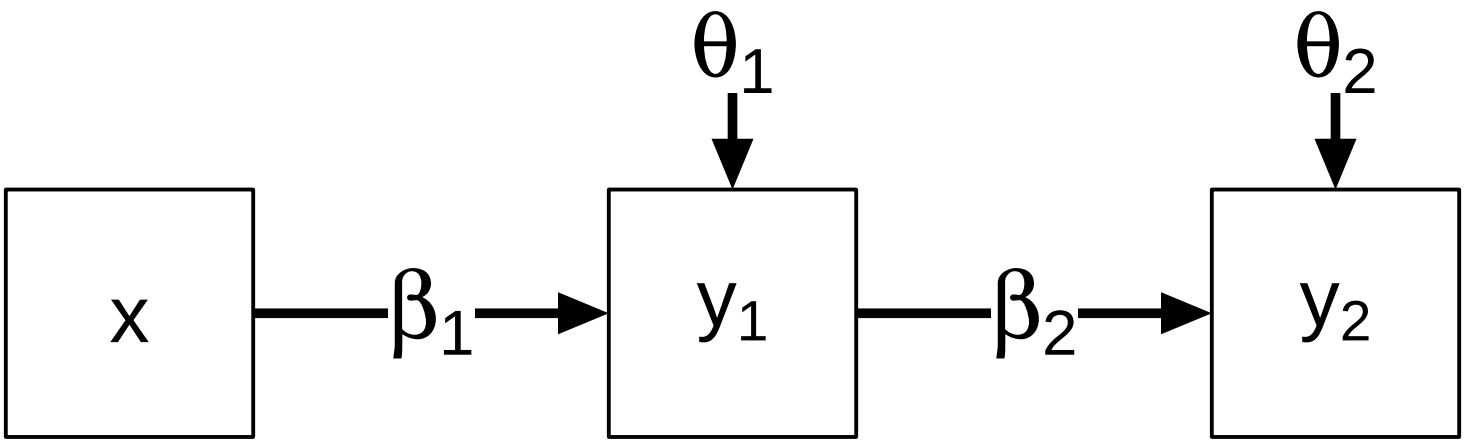
$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Psi}(\boldsymbol{I} - \boldsymbol{B})^{-1\top}\boldsymbol{\Lambda}^{\top} + \boldsymbol{\Theta}$$

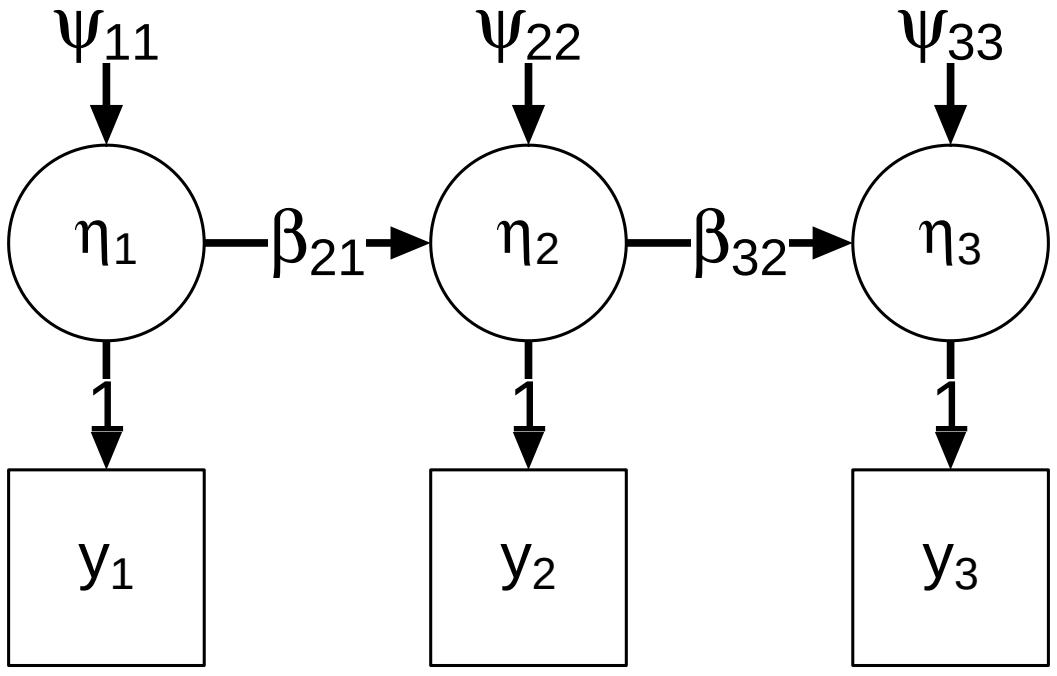
Simply the CFA model with one extra matrix: $\boldsymbol{B}$ encoding regression parameters. Element β_{ij} encodes the effect from variable j to variable i (note, this is opposite of how normally a directed network is encoded).

The same identification rules as in CFA apply:

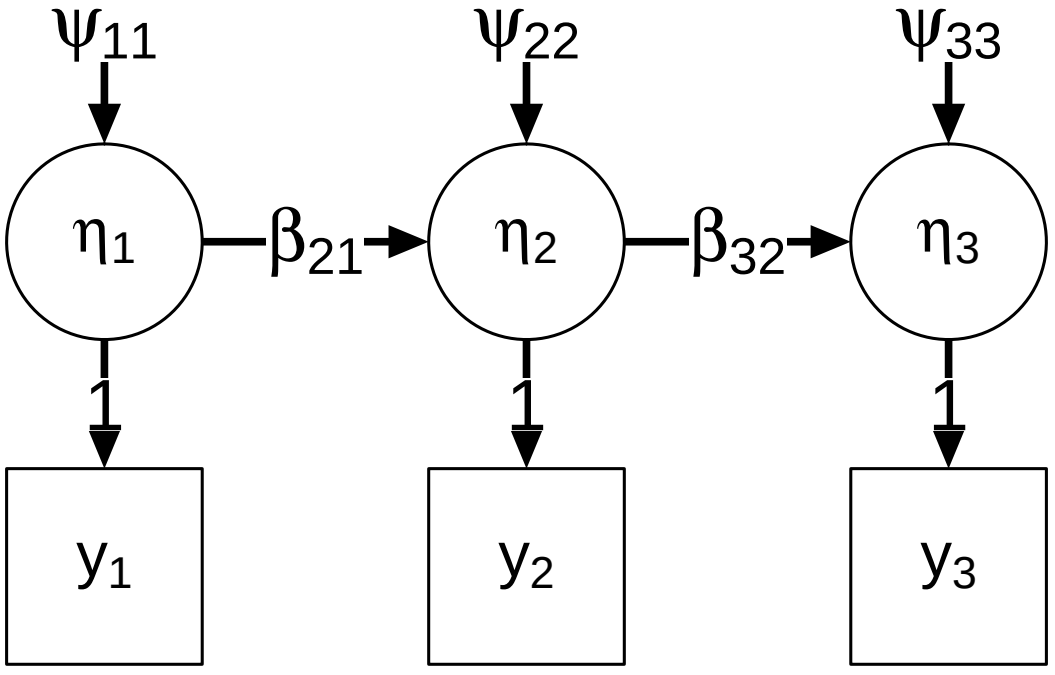
- Latent variables must be scaled by setting one factor loading or (residual) variance to 1
- Model must have at least 0 degrees of freedom

Next week we will discuss equivalent models.


θ1
θ2
x
β1
y1
β2
y2

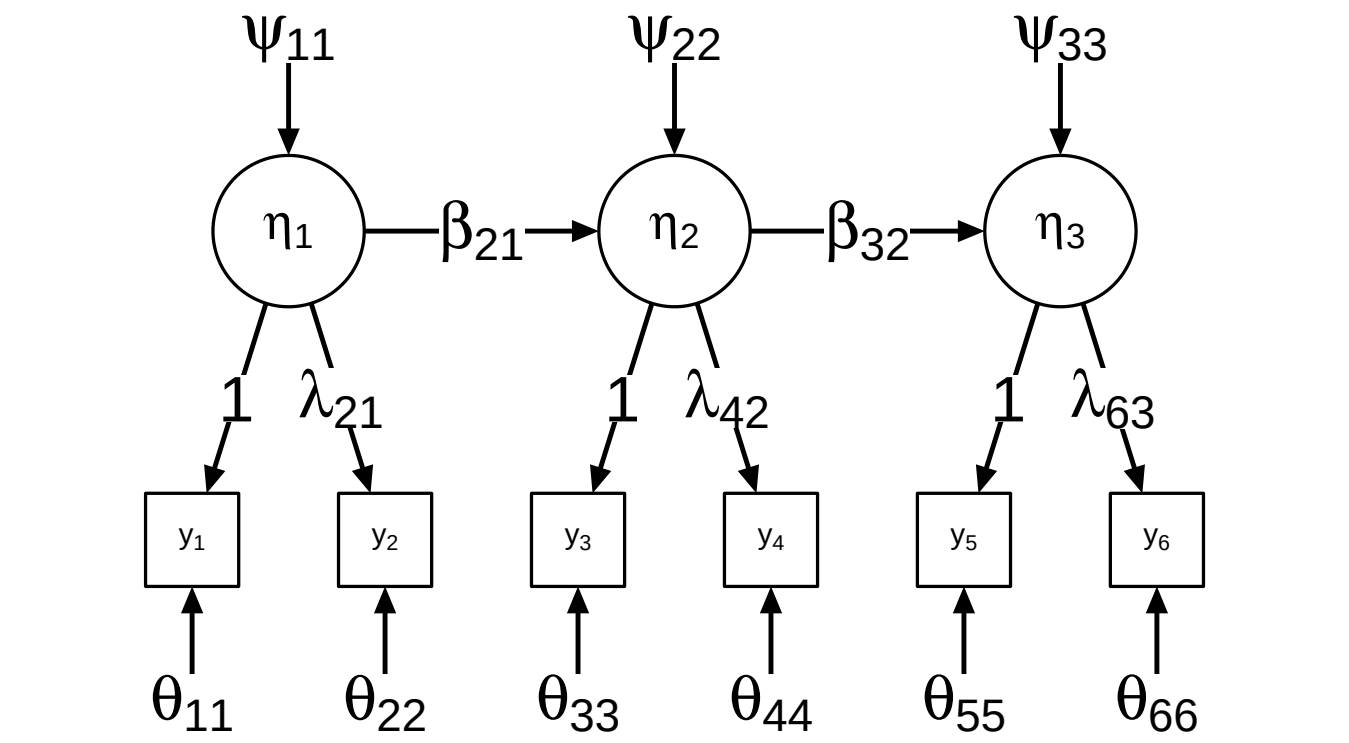


$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{\Psi} = \begin{bmatrix} \psi_{11} & 0 & 0 \\ 0 & \psi_{22} & 0 \\ 0 & 0 & \psi_{33} \end{bmatrix}, \mathbf{\Theta} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ 0 & \beta_{32} & 0 \end{bmatrix}$$



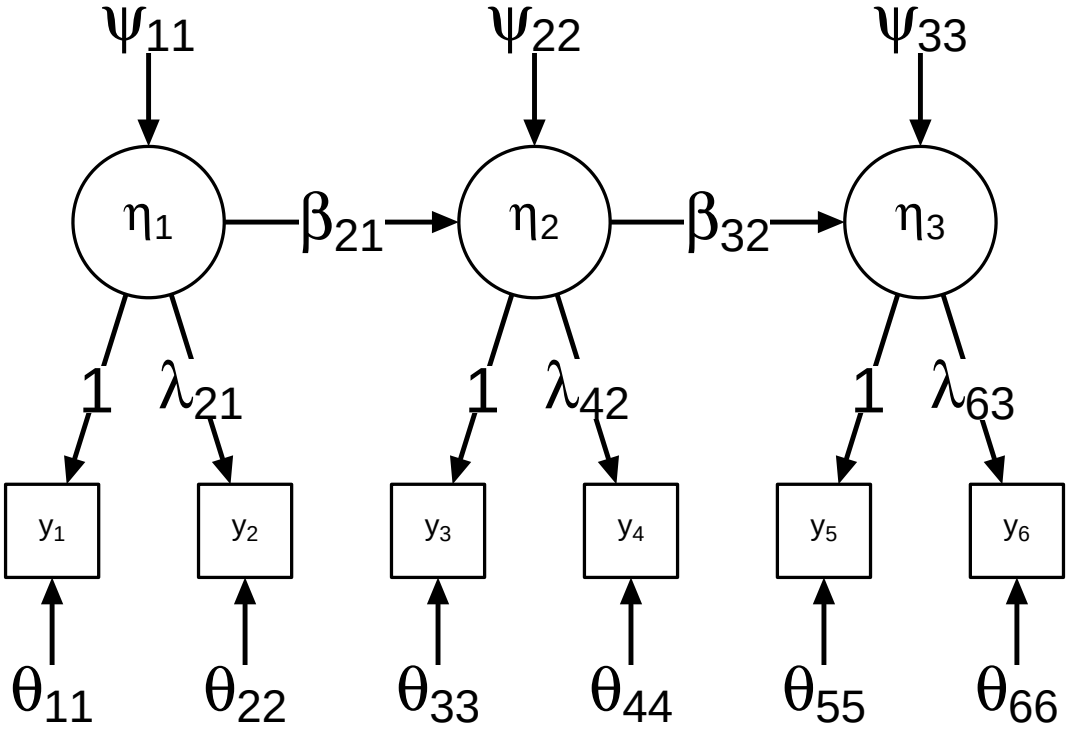
Lavaan automatically adds the latent dummy variables for you!
The model is just:

```
y2 ~ y1
y3 ~ y2
```



$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_{63} \end{bmatrix}, \mathbf{\Psi} = \begin{bmatrix} \psi_{11} & 0 & 0 \\ 0 & \psi_{22} & 0 \\ 0 & 0 & \psi_{33} \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ 0 & \beta_{32} & 0 \end{bmatrix}$$

$\mathbf{\Theta}$ diagonal as usual.



Lavaan model (using `sem()`):

```
eta1 =~ y1 + y2
eta2 =~ y3 + y4
eta3 =~ y5 + y6
eta2 ~ eta1
eta3 ~ eta2
```

Structural Equation Modeling (SEM)

- SEM is a more general modeling framework than CFA
- Allows for modeling complex causal hypotheses
- Requires fitting CFA model (as it is nested in CFA model)
- Fitting SEM models using lavaan:
  - Use `sem()` instead of `cfa()`
  - Specify regressions using the `~` operator
- Same identification rules as in CFA apply
- Models can be assessed and compared in the same way as CFA models