Introduction to Structural Equation Modelling (SEM)

David Evans[1,2,3]

1 Institute for Molecular Bioscience, University of Queensland

2 University of Queensland Diamantina Institute

3 MRC Integrative Epidemiology Unit, University of Bristol

What is SEM?

- A statistical method for analyzing the relationship between observed and latent variables
- Used mostly in social and behavioural sciences and also genetic epidemiology
- Causal and correlational relationships between variables are modelled explicitly
- Involves constructing a statistical (structural) model, seeing how well this model fits some data, and obtaining estimates of parameters
- Also known as “Confirmatory Factor Analysis” / “Analysis of covariance structure” / “Path analysis”

Why SEM?

- Flexibility- almost any linear model can be written as a SEM
- SEM makes it easy to create new models/methods
- Super useful for deriving expected variances/covariances in genetics
- SEM means that you can think about a problem multiple ways
- Advantages for modelling human genetic data:
  - Latent variables
  - Multivariate phenotypes
  - Feedback loops
  - Assortative mating
  - Vertical transmission
  - Gene-environment covariance
  - Non-linear constraints

SEM and Genetics

Galton (1865-ish)
Correlation
Family Resemblance
Twins
Ancestral Heredity

Mendel (1865)
Particulate Inheritance
Genes: single in gamete
double in zygote
Segregation ratios

Darwin (1858,1871)
Natural Selection
Sexual Selection
Evolution

Fisher (1918)
Correlation & Mendel
Maximum Likelihood
ANOVA: partition of variance

Spearman (1904)
Common Factor Analysis

Wright (1921)
Path Analysis

Mather (1949) & **Jinks** (1971)
Biometrical Genetics
Model Fitting (plants)

Thurstone (1930's)
Multiple Factor Analysis

Watson & Crick (1953)

Joreskog (1960)
Covariance
Structure Analysis
LISREL

Jinks & Fulker (1970)
Model Fitting applied to humans

Morton (1974)
Path Analysis &
Family Resemblance

Population Genetics

Elston etc (19..)
Segregation
Linkage

Rao, Rice, Reich, Cloninger (1970's)
Assortment
Cultural Inheritance

Martin & Eaves (1977)
Genetic Analysis of
Covariance Structure

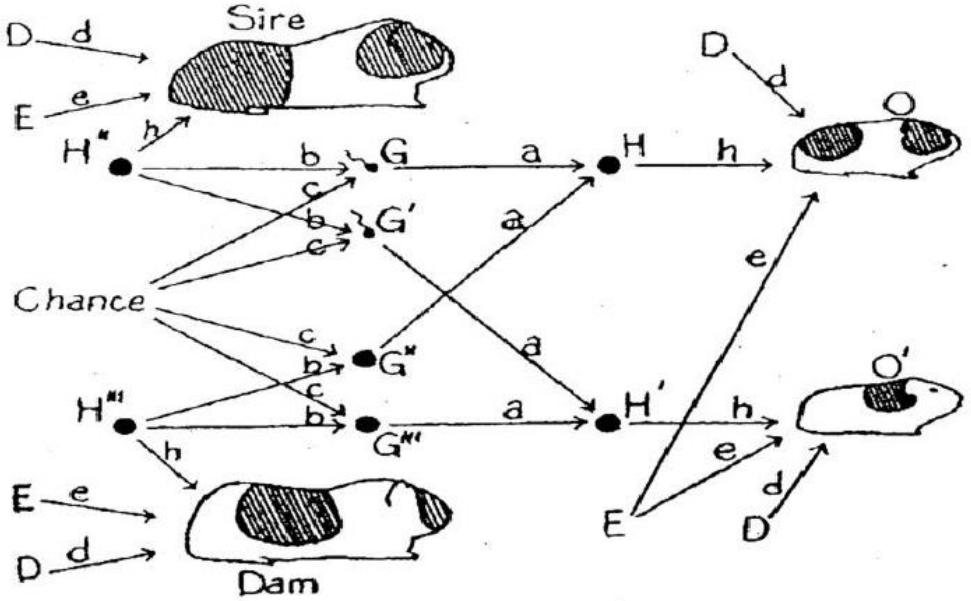
2000

Neale (1990) Mx

Molecular Genetics

Neale & Cardon (1992)

Sewall Wright



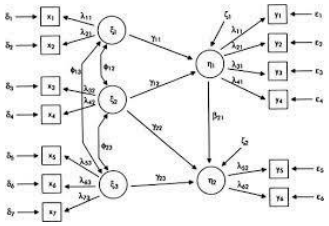
How does SEM Work?

- (1) Start of with a theory
- (2) Express this theory as a model using a series of structural equations or as a path diagram (i.e. a “Structural Equation Model”)
- (3) Collect the data
- (4) Fit the model to the data. Obtain parameter estimates and a measure of how well the model fits the data.
- (5) Revise the theory/model
- (6) Repeat

How does SEM Work?

THEORY

BUILD MODEL



Implies covariance structure

$\boldsymbol{\Sigma(\theta)}$

$\boldsymbol{\Sigma}$

FIT TO DATA

Yes

OBTAIN PARAMETER ESTIMATES AND CIs

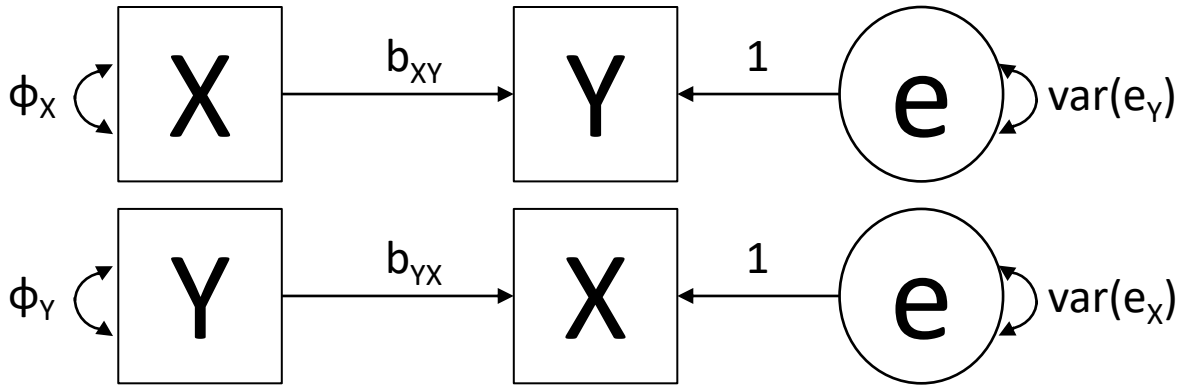
No

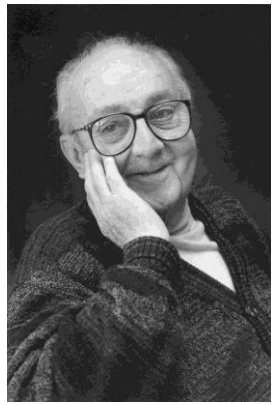
REVISE MODEL/ THEORY

“All Models and Wrong – Some Models are Useful”

- This adage is true for all models, not just SEMs!
- Sometimes different models give exactly the same fit
- In genetic epidemiology, our SEMs are constructed based on biometrical genetics principles increasing their validity
- SEM and model falsification
- SEM and parameter estimation and confidence intervals

Which model is “correct”?



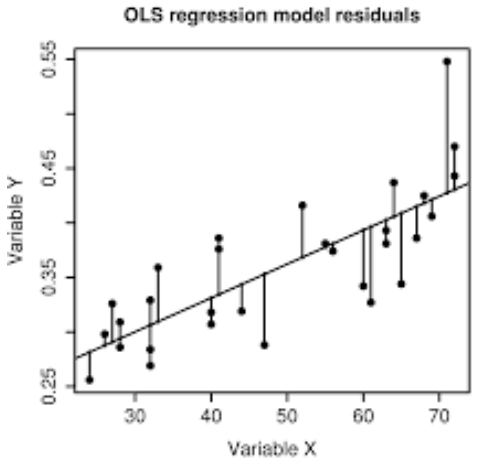


George Box

Fitting Functions and Estimators

- We fit the model to our data using a fitting function. This provides us with a measure of the goodness of fit of our model and estimates of population parameters of interest

- Most students will be familiar with ordinary least squares (OLS) which we typically use in linear regression. OLS minimizes the sum of the squared residuals between the estimated regression line and the observed (y) values of our data



- In SEM we typically use another fitting function based on the method of maximum likelihood. The estimates of the population parameters are called “maximum likelihood estimates” (MLEs)

(Full Information) Maximum Likelihood

Maximum Likelihood Properties

- Consistent
- Asymptotically unbiased
- Efficient
- Scale Invariant
- Sampling distribution of estimates is asymptotically normal
- Asymptotically, twice the difference in log-likelihood between nested models is distributed as chi-square (e.g. Consider $\boldsymbol{\theta}_F$ = (a, b, c); $\boldsymbol{\theta}_R$ = (a, b, c=0)- twice the difference in log-likelihoods between the models would be distributed as χ^2_1)

Likelihood

Imagine tossing a single coin. What is the probability of throwing a head?

If the coin is fair:

P(H) = 0.5

P(T) = 0.5

If the coin isn't fair:

P(H) = p

P(T) = 1 - p

Now imagine tossing the same coin ten times and obtaining the following sequence:
HEAD, HEAD, HEAD, TAIL, HEAD, HEAD, HEAD, HEAD, TAIL, TAIL

What is the probability of observing this particular sequence?

P(HHHTHHHTT) = p^7*(1-p)^3

Recall that because the events are independent, the probabilities of each coin toss are multiplied together.

Likelihood

Probability is concerned with estimating the chances of observing a particular event given a model for the data (in this case that the coin is fair p = 0.5)

Probability of the **Data** given the **Model** P(**D**|**M**) P(HHHTHHHTT|p) = p^7*(1-p)^3

Likelihood flips this relationship around. Now we are interested in what is the likely value for the parameter *p* given we have observed the data.

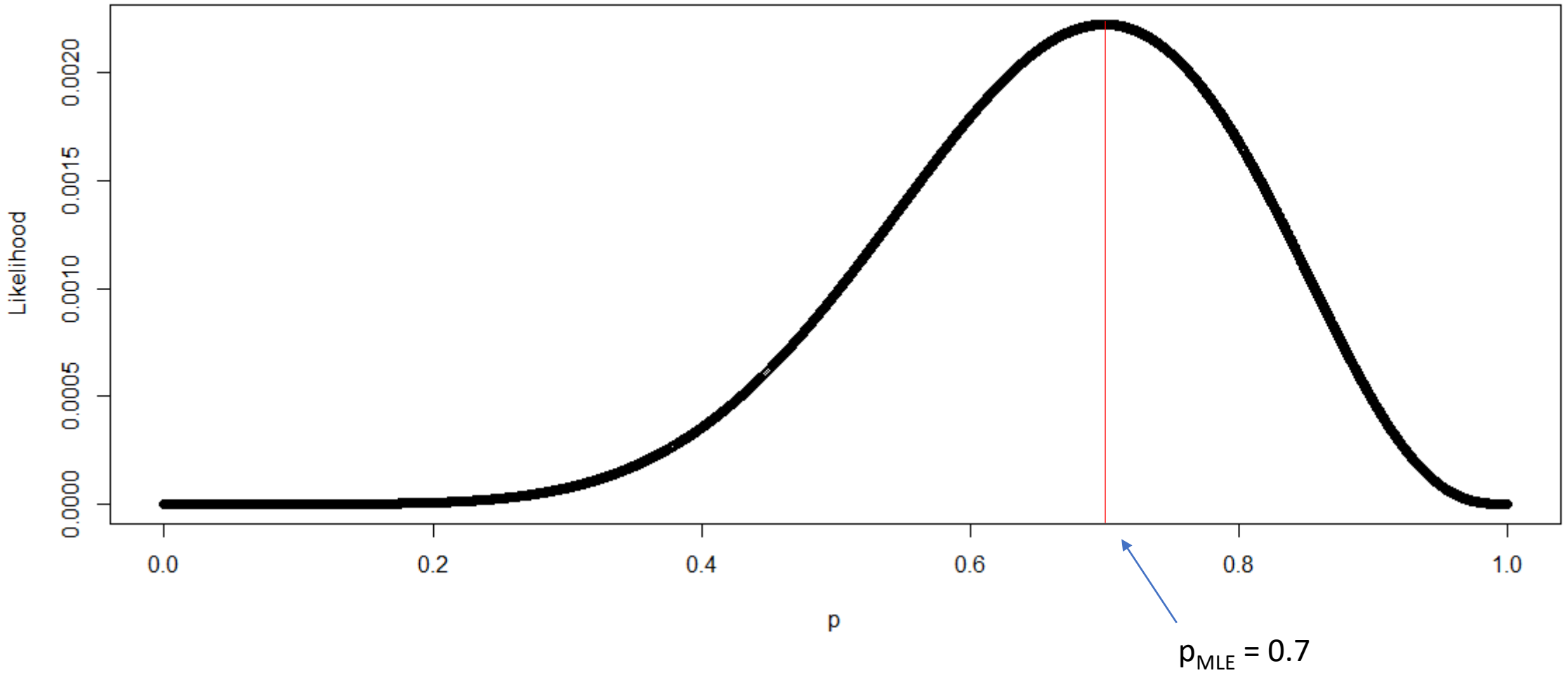
Likelihood of the **Model** given the **Data** L(**M**|**D**) L(p|HHHTHHHTT) = p^7*(1-p)^3

In other words, is our coin fair given we have observed seven heads and three tails?

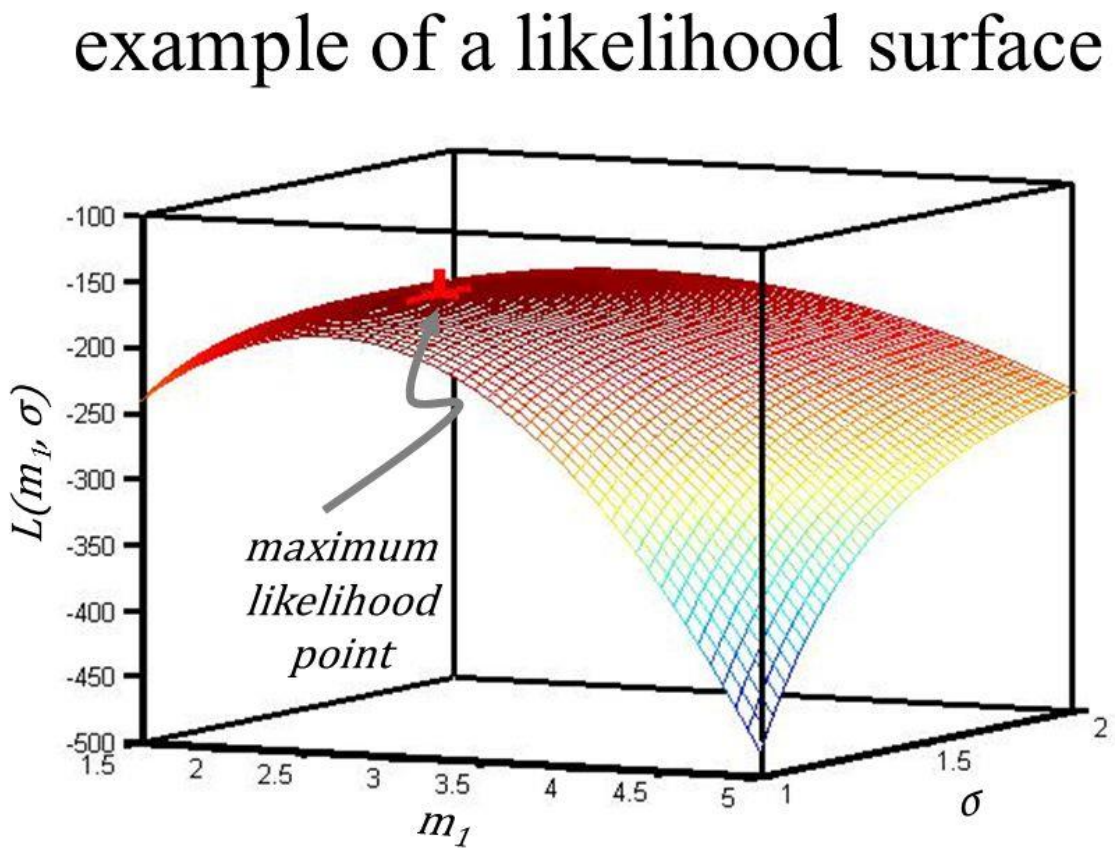
Likelihood obeys many of the same rules as probability but not all of them!:
-e.g. The likelihood of independent events are multiplied together (just like probabilities)

We are interested in finding the value of the parameter (here “p”) that maximizes the likelihood of the data.
Q. What is the MLE for p in this particular example?

Likelihood


Likelihood
0.0000
0.0005
0.0010
0.0015
0.0020
0.0
0.2
0.4
0.6
0.8
1.0
p
p_{MLE} = 0.7

Likelihood >1 parameter



Likelihood- Quantitative Traits

What about quantitative traits?

Imagine we have measured the height of one hundred unrelated individuals.

We assume that height is normally distributed in the population:

$$P(D) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

This probability density is specified by:
-A mean μ (let's say 170cm)
-A variance σ^2 (let's say 25cm^2)

The probability density associated with a single individual who has height 180 is therefore:

$$P(D) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{180-\mu}{\sigma})^2}$$

where μ = 170 and σ^2 = 25cm^2

Since individuals are independent, the likelihood of observing the hundred individuals is given by:
$L(\mu, \sigma^2 | Data) = P(D_1)P(D_2)...P(D_N)$

We wish to find the values of μ and σ^2 that maximise the likelihood of the data

Likelihood- Multivariate Quantitative Data

What about multivariate quantitative data (e.g. *k* traits measured on the same individual; related individuals etc)?

Imagine we have measured the height and BMI (k = 2) of 100 unrelated individuals.

Here we typically assume that the joint distribution of the quantitative traits is multivariate normal:

$$P(D) = (\frac{1}{\sqrt{2\pi}})^{k/2} |\boldsymbol{\Sigma}|^{-1/2} e^{-\frac{1}{2}(x-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (x-\boldsymbol{\mu})}$$

This probability density is specified by:
-A (*k* x 1) vector of means **μ**
-A (*k* x *k*) variance-covariance matrix **Σ**

Since individuals are independent, the likelihood of observing the hundred individuals is given by:
$L(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \text{Data}) = P(D_1)P(D_2)...P(D_N)$

We wish to find the values of **μ** and **Σ** that maximise the likelihood of the data

Likelihood- “Modelling the Means”

So far our modelling has not been particularly interesting, only finding MLEs of Mean vectors and Covariance matrices!

However, we can go a step further than this and model **μ** and **Σ** in terms of other parameters

For example, we can model the effect of covariates like sex or genotype on our variables e.g.:

$$\boldsymbol{\mu} = \alpha + \beta_{sex} \times \text{sex} + \beta_{SNP} \times \text{SNP}$$

We call this modelling fixed effects or “modelling the means”

Here our likelihood is maximized wrt α, β_{sex}, β_{SNP}, **Σ** :
$L(\alpha, \beta_{sex}, \beta_{SNP}, \boldsymbol{\Sigma} \mid \text{Data}) = P(D_1)P(D_2)\ldots P(D_N)$

If we were interested in testing for evidence of genetic association, we would typically (but not always!) do it here in the model for the means

Unlike our outcome variables, there is no requirement for these covariates to be normally distributed

Likelihood- “Modelling the Covariances”

Likewise, we can also (simultaneously) model the covariance structure of our data

This is what SEM is all about!

For example, if we had a set of sibling pairs we might be able to model the variance of our data and the covariance between relatives in terms of genetic and environmental variance components

$$\sigma^2 = \sigma_G^2 + \sigma_E^2$$
$$\sigma_{1,2} = \frac{1}{2}\sigma_G^2$$

We refer to this as “modelling the variances/covariances”

Here our likelihood is maximized wrt σ_G^2, σ_E^2 and the mean vector:
$L(\boldsymbol{\mu}, \sigma_G^2, \sigma_E^2 \mid \text{Data}) = P(D_1)P(D_2)\ldots P(D_N)$

Most of the rest of our course concerns how we can informatively parameterize our covariance structure to answer interesting questions in genetic epidemiology

Likelihood- Modelling Both Means and Covariances Simultaneously

In SEM we can fit our model to individual level data in which case we maximize the multivariate normal likelihood according to the model for the means and the model for the covariances we have just discussed.

This is particularly useful if:

-We have missing data
-We wish to remove the effect of covariates from the variance/covariance structure
-We wish to perform specific hypothesis tests/estimate certain parameters in the model for the means

Alternatively, particularly if we do not have missing data, we can fit the model to summary results level data (i.e. to covariance matrices)

This can be thought of heuristically as minimizing the difference between the covariance matrix implied by the SEM (the "expected covariance matrix") and the observed covariance matrix

$$\Sigma - \Sigma(\boldsymbol{\theta})$$

Observed Covariance Matrix — Expected Covariance Matrix

Understanding SEM

$$\Sigma - \Sigma(\boldsymbol{\theta})$$

Observed Covariance Matrix — Expected Covariance Matrix

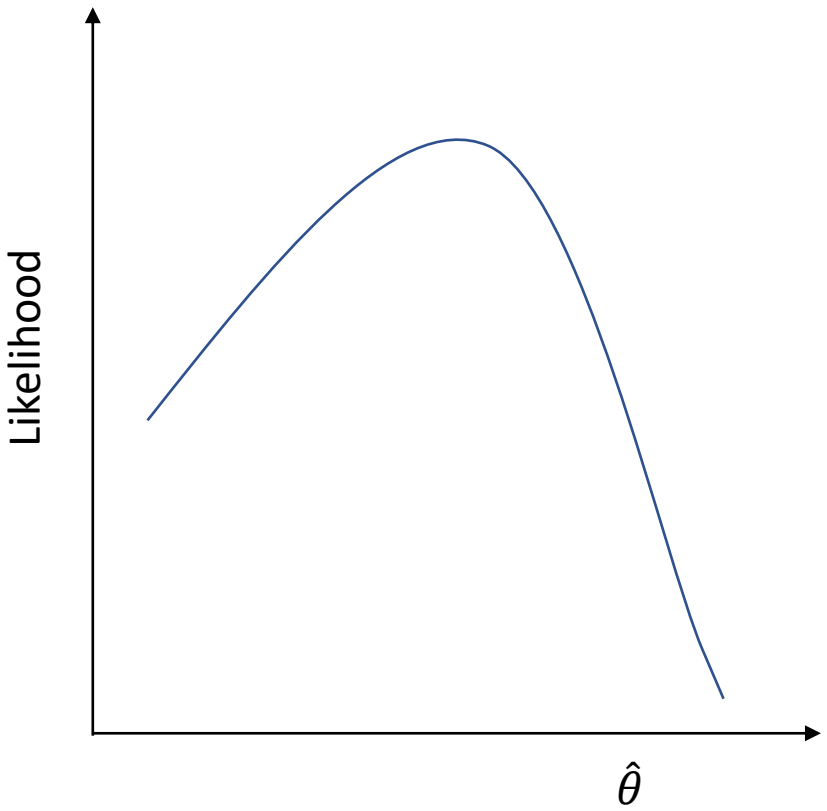
- Expected covariance matrix is a function of model parameters
- These expectations depend on the particular structural equation model
- Parameters chosen to minimize the difference between observed and expected covariance matrices (MLEs)

SEM- Assumptions

- Linearity
- Multivariate normality
  - Exogenous variables exempt
  - Binary/ordinal variables can be modelled assuming an underlying normal distribution of liability
  - Methods exist for combining binary and continuous variables
- “All models are wrong- some models are useful”

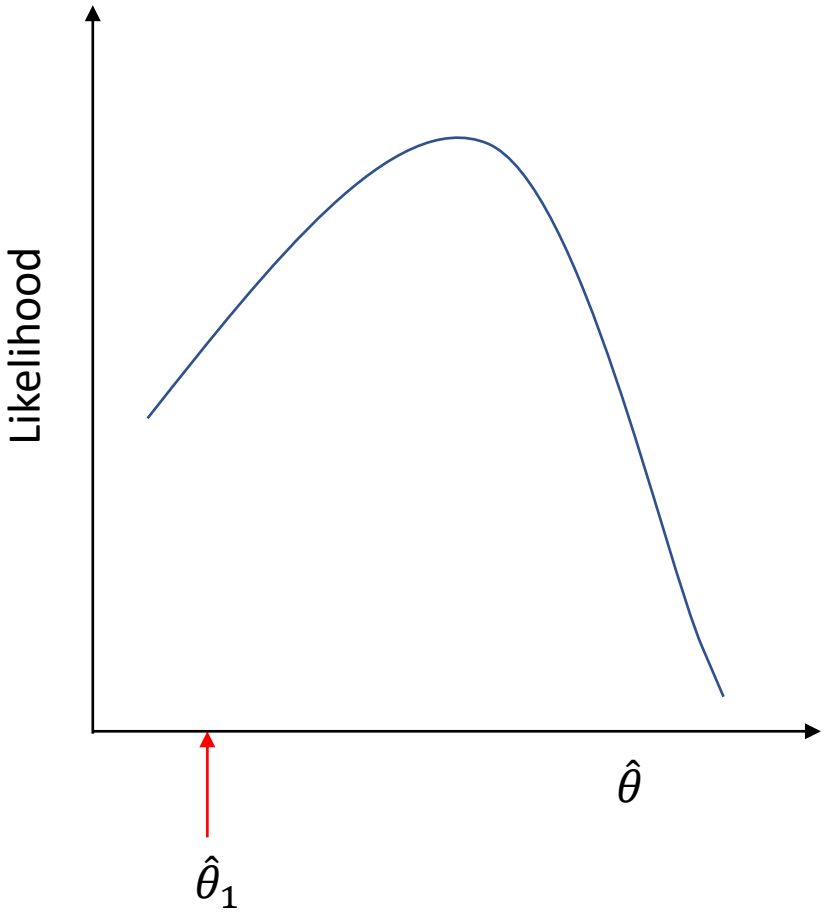
Optimization

Optimization



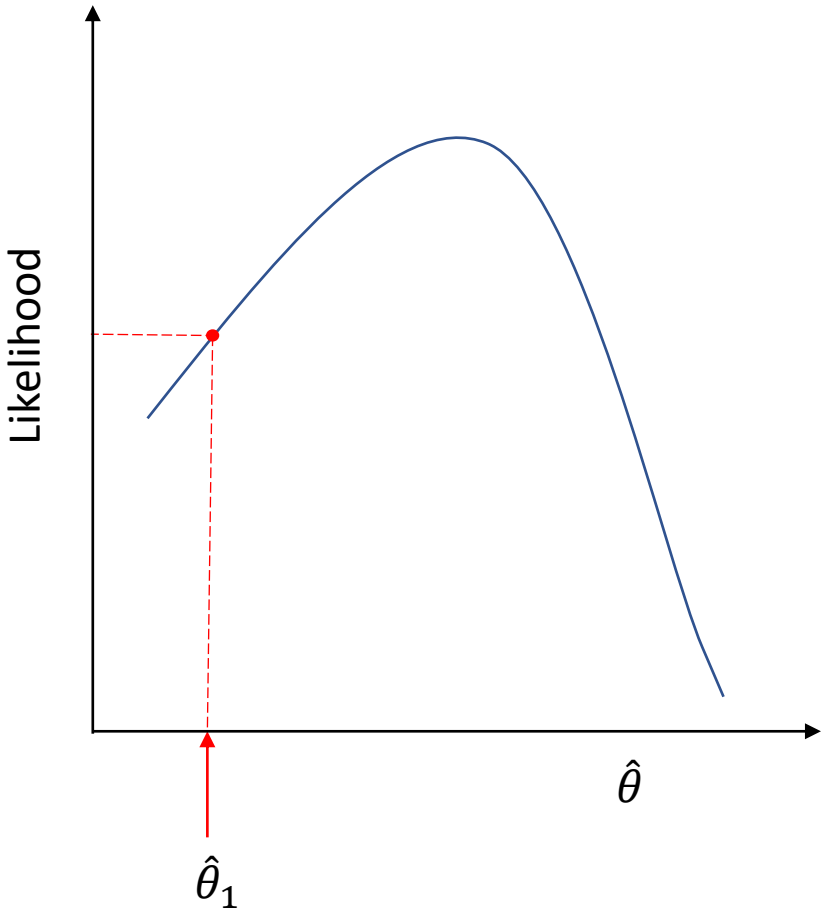
- Maximum likelihood solutions can rarely be solved in closed form- rather iterative optimization procedures are commonly needed
- Choose starting values for parameters
- Calculate likelihood of these parameter estimates, as well as the first and second derivative of the likelihood with respect to the parameters
- Adjust parameter values, and repeat process until stopping criterion is reached

Optimization



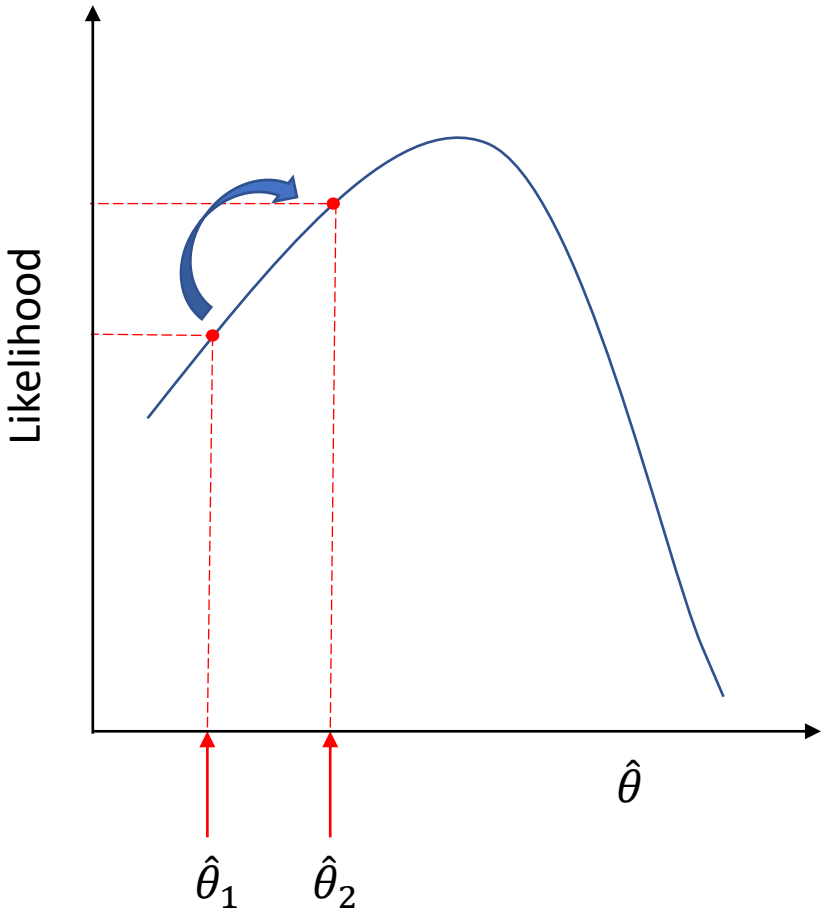
- Maximum likelihood solutions can rarely be solved in closed form- rather iterative optimization procedures are commonly needed
- Choose starting values for parameters
- Calculate likelihood of these parameter estimates, as well as the first and second derivative of the likelihood with respect to the parameters
- Adjust parameter values, and repeat process until stopping criterion is reached

Optimization



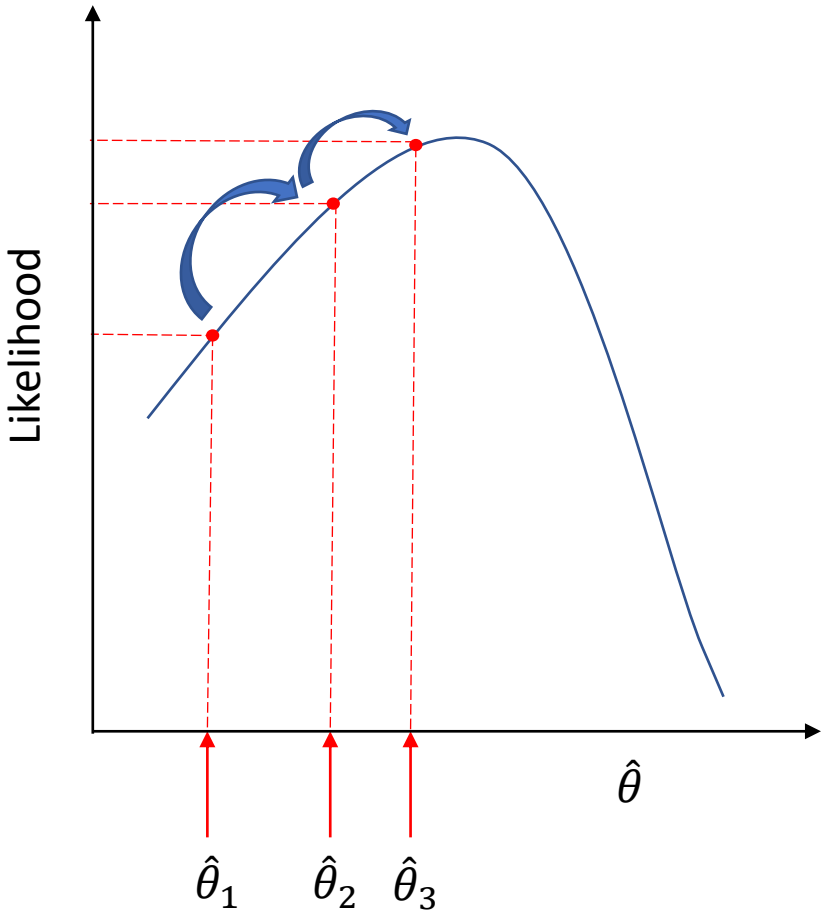
- Maximum likelihood solutions can rarely be solved in closed form- rather iterative optimization procedures are commonly needed
- Choose starting values for parameters
- Calculate likelihood of these parameter estimates, as well as the first and second derivative of the likelihood with respect to the parameters
- Adjust parameter values, and repeat process until stopping criterion is reached

Optimization



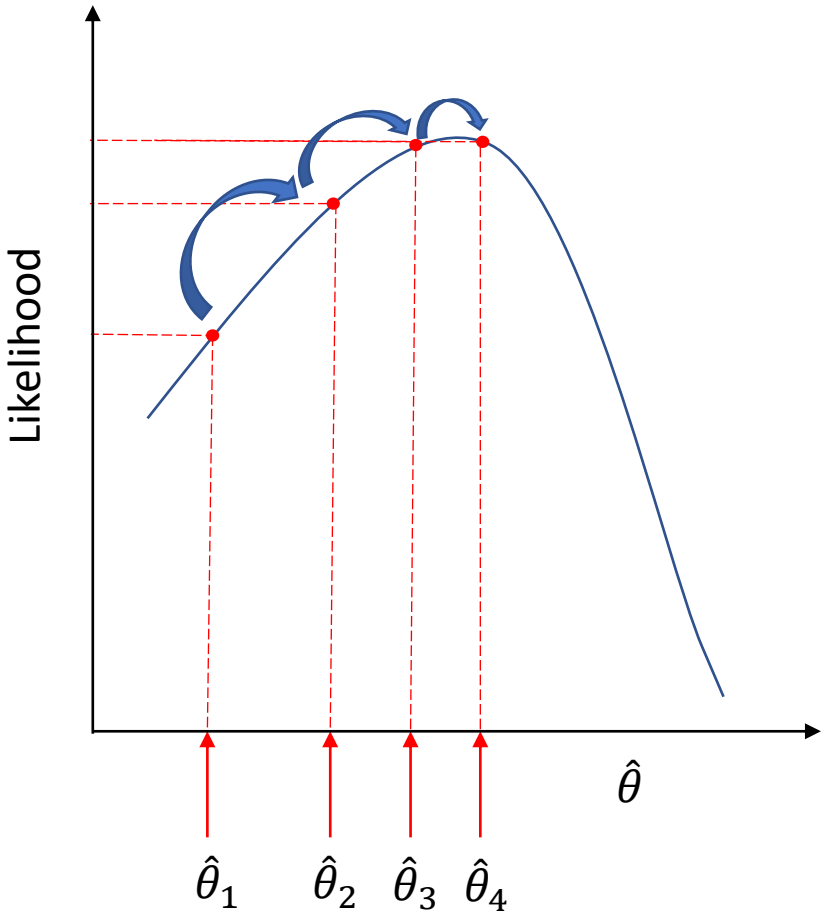
- Maximum likelihood solutions can rarely be solved in closed form- rather iterative optimization procedures are commonly needed
- Choose starting values for parameters
- Calculate likelihood of these parameter estimates, as well as the first and second derivative of the likelihood with respect to the parameters
- Adjust parameter values, and repeat process until stopping criterion is reached

Optimization



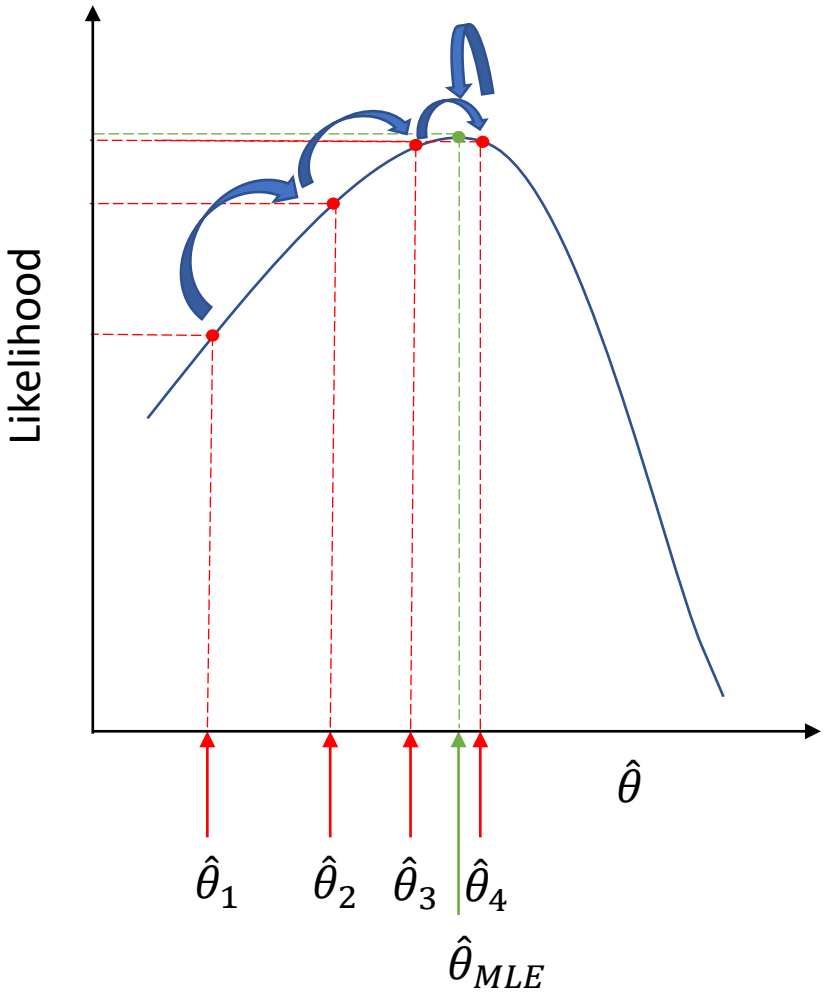
- Maximum likelihood solutions can rarely be solved in closed form- rather iterative optimization procedures are commonly needed
- Choose starting values for parameters
- Calculate likelihood of these parameter estimates, as well as the first and second derivative of the likelihood with respect to the parameters
- Adjust parameter values, and repeat process until stopping criterion is reached

Optimization



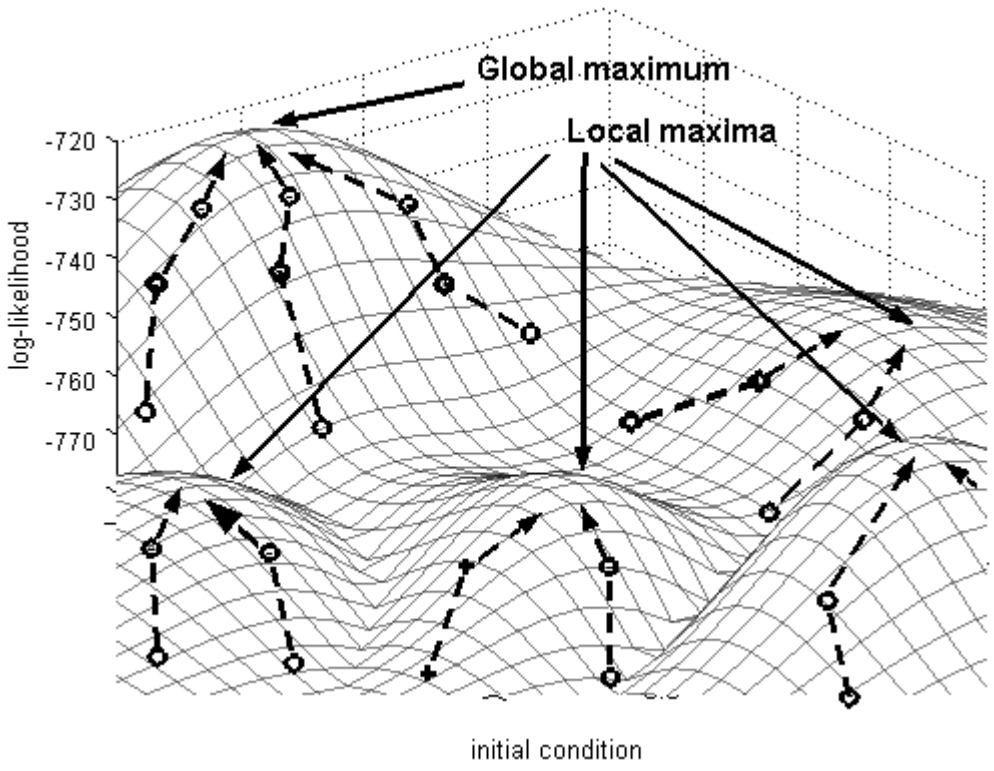
- Maximum likelihood solutions can rarely be solved in closed form- rather iterative optimization procedures are commonly needed
- Choose starting values for parameters
- Calculate likelihood of these parameter estimates, as well as the first and second derivative of the likelihood with respect to the parameters
- Adjust parameter values, and repeat process until stopping criterion is reached

Optimization



- Maximum likelihood solutions can rarely be solved in closed form- rather iterative optimization procedures are commonly needed
- Choose starting values for parameters
- Calculate likelihood of these parameter estimates, as well as the first and second derivative of the likelihood with respect to the parameters
- Adjust parameter values, and repeat process until stopping criterion is reached

Optimization



- Typically we maximize the log-likelihood because computers find it easier to add rather than multiply
- The likelihood surface may be complicated with one or more local maxima
- Choosing different starting values can increase confidence in a global solution
- In general it is good practice to choose starting values as close as possible to the global solution

Identification

Identification

- Means that all parameters in a model can be estimated uniquely given the data
- A necessary (but not sufficient condition) for identifiability is that you have the same (or more) observed statistics than parameters you want to estimate
- If all parameters in a model are identified, then the model as a whole is identified
- Even though the model as a whole may be unidentified some parameters may be identified

Identified or Not?

(1) $\theta_1 + \theta_2 = 10$

(2) $\theta_1 + \theta_2 = 10$

$\theta_1 - \theta_2 = 0$

(3) $\theta_1 + \theta_2 = 10$

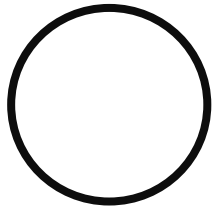
$2\theta_1 + 2\theta_2 = 20$

Building Models With Path Diagrams

Path Diagrams

- Path diagrams pictorially represent causal models. They aid in deriving the variances and covariances implied by the model.

Observed Variables

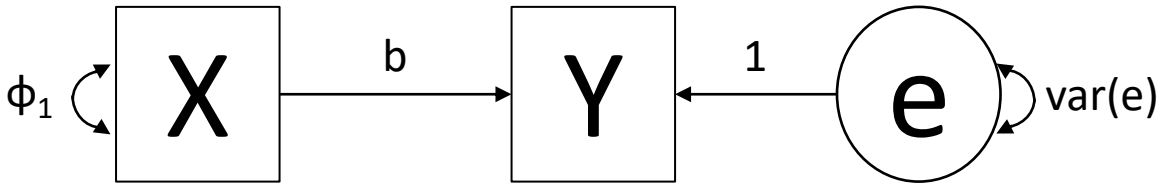


Latent Variables

Causal Paths

(Co)variance Paths

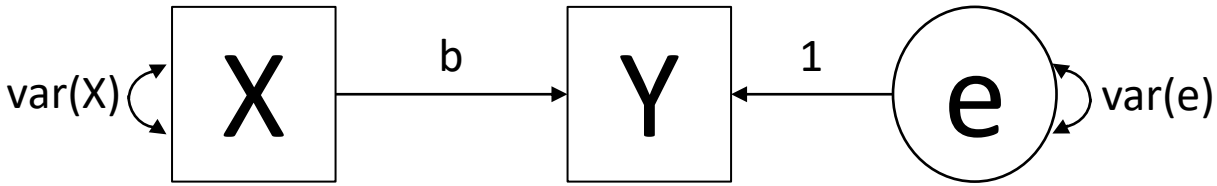
Path Diagrams



$$Y = bX + e$$

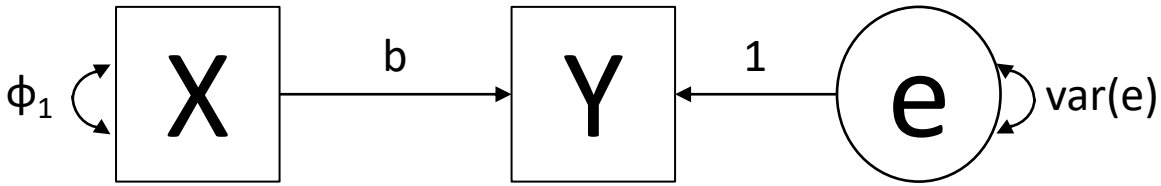
- Assume variables measured in deviation form
- “b” is a path coefficient
- It quantifies the expected change in Y for every unit change in X is “b”

Path Diagrams- Univariate Regression



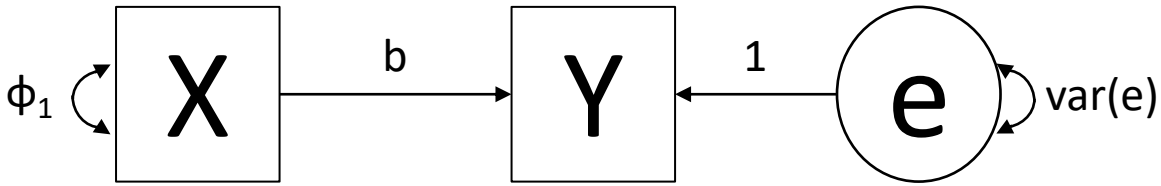
- Y = BX + e (explicit)
- Measurement error in Y (explicit)
- No measurement error in X (explicit)
- No covariance between X and epsilon (explicit)
- Covariance between X and Y is b*var(X) (explicit)
- Linear relationships between the variables (implicit)
- Multivariate normality (implicit)

Path Diagrams- Univariate Regression



Structural Equation: $Y = bX + e$

Path Diagrams- Univariate Regression

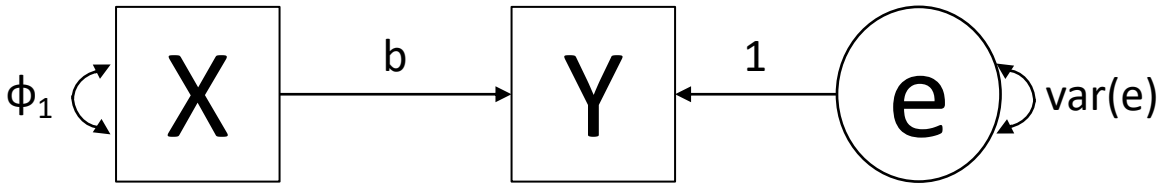


Structural Equation: $Y = bX + e$

Observed Covariance Matrix:

$$\Sigma = \begin{matrix} \text{VAR(X)} & \text{COV(X,Y)} \\ \text{COV(X,Y)} & \text{VAR(Y)} \end{matrix}$$

Path Diagrams- Univariate Regression



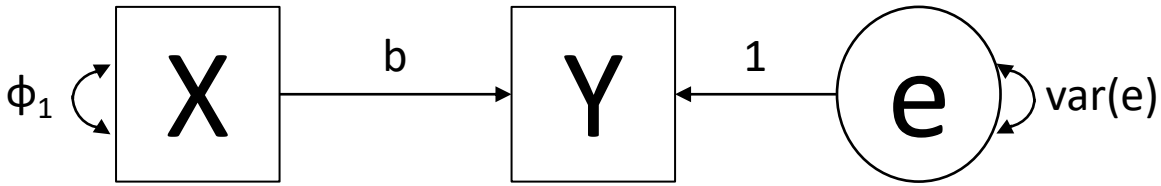
Structural Equation: $Y = bX + e$

Observed Covariance Matrix:

$$\Sigma = \begin{matrix} \text{VAR(X)} & \text{COV(X,Y)} \\ \text{COV(X,Y)} & \text{VAR(Y)} \end{matrix}$$

Number of observed statistics: 3

Path Diagrams- Univariate Regression



Structural Equation: $Y = bX + e$

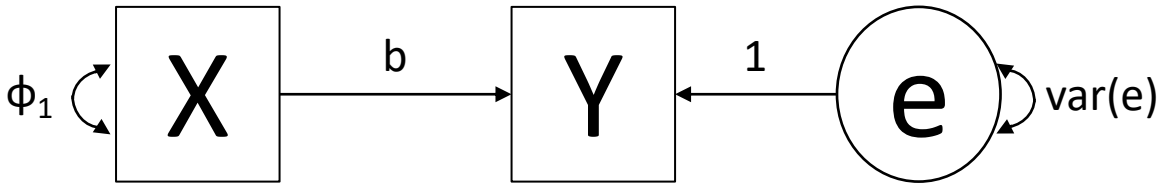
Observed Covariance Matrix:

$$\Sigma = \begin{matrix} \text{VAR(X)} & \text{COV(X,Y)} \\ \text{COV(X,Y)} & \text{VAR(Y)} \end{matrix}$$

Number of estimated parameters: 3 (ϕ_1, b, var(e))

Number of observed statistics: 3

Path Diagrams- Univariate Regression



Structural Equation: $Y = bX + e$

Observed Covariance Matrix:

$$\Sigma = \begin{matrix} \text{VAR(X)} & \text{COV(X,Y)} \\ \text{COV(X,Y)} & \text{VAR(Y)} \end{matrix}$$

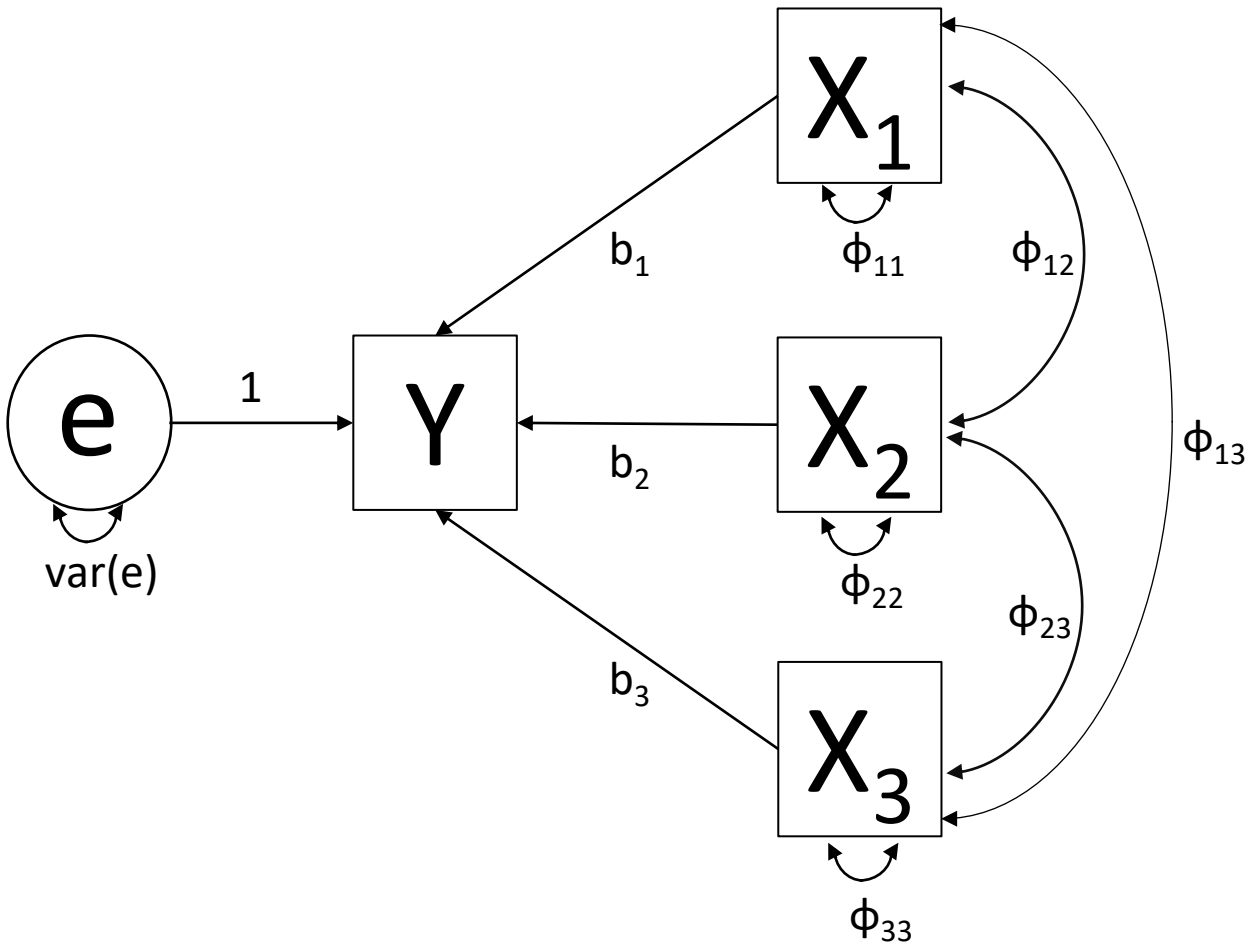
Number of observed statistics: 3

Number of estimated parameters: 3 (ϕ_1, b, var(e))

Expected/Implied Covariance Matrix:

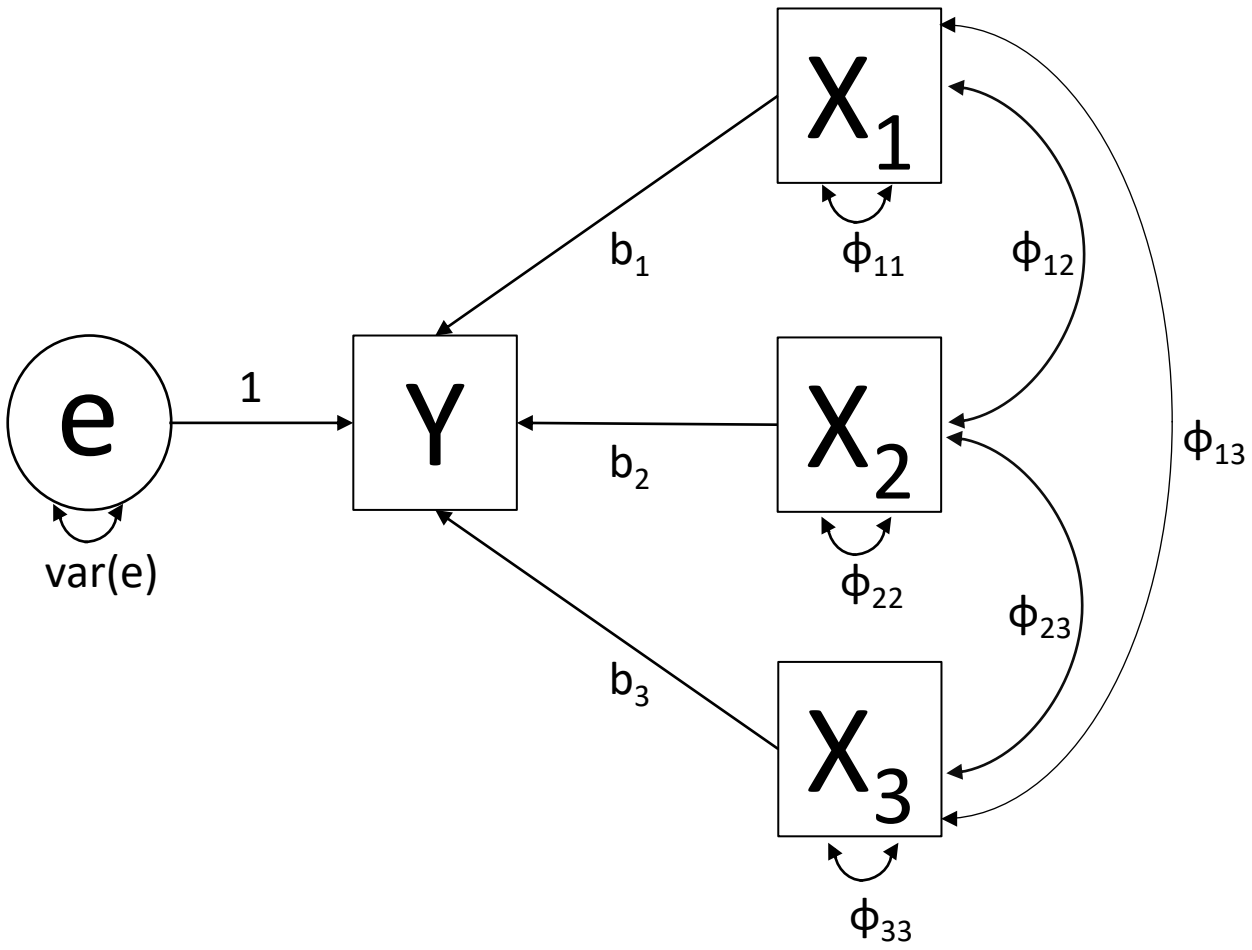
$$\Sigma(\theta) = \begin{matrix} \phi_1 & b\phi_1 \\ b\phi_1 & b^2\phi_1+\text{var(e)} \end{matrix}$$

Path Diagrams

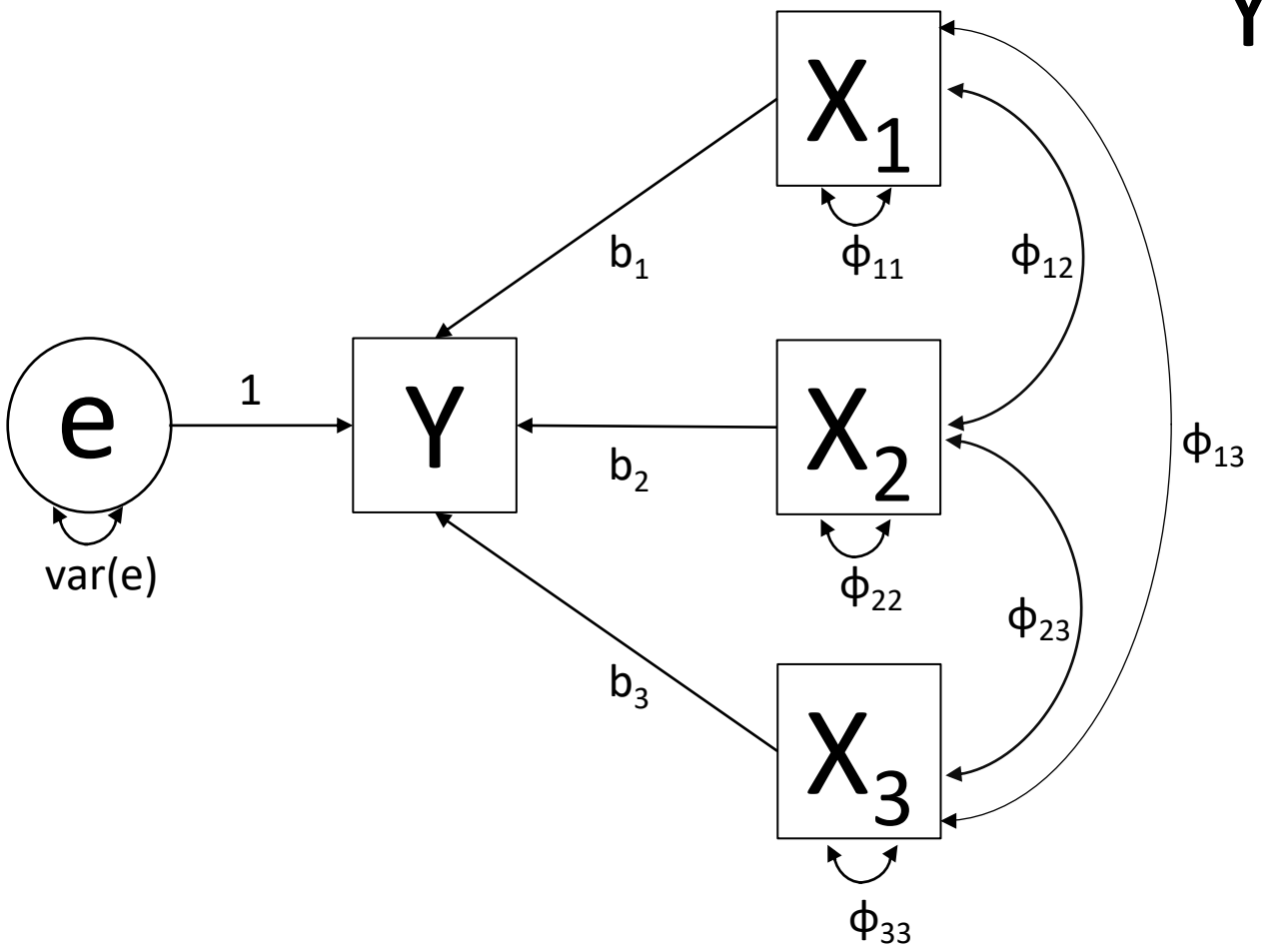


Path Diagrams- Multivariable Regression

Structural Equation:



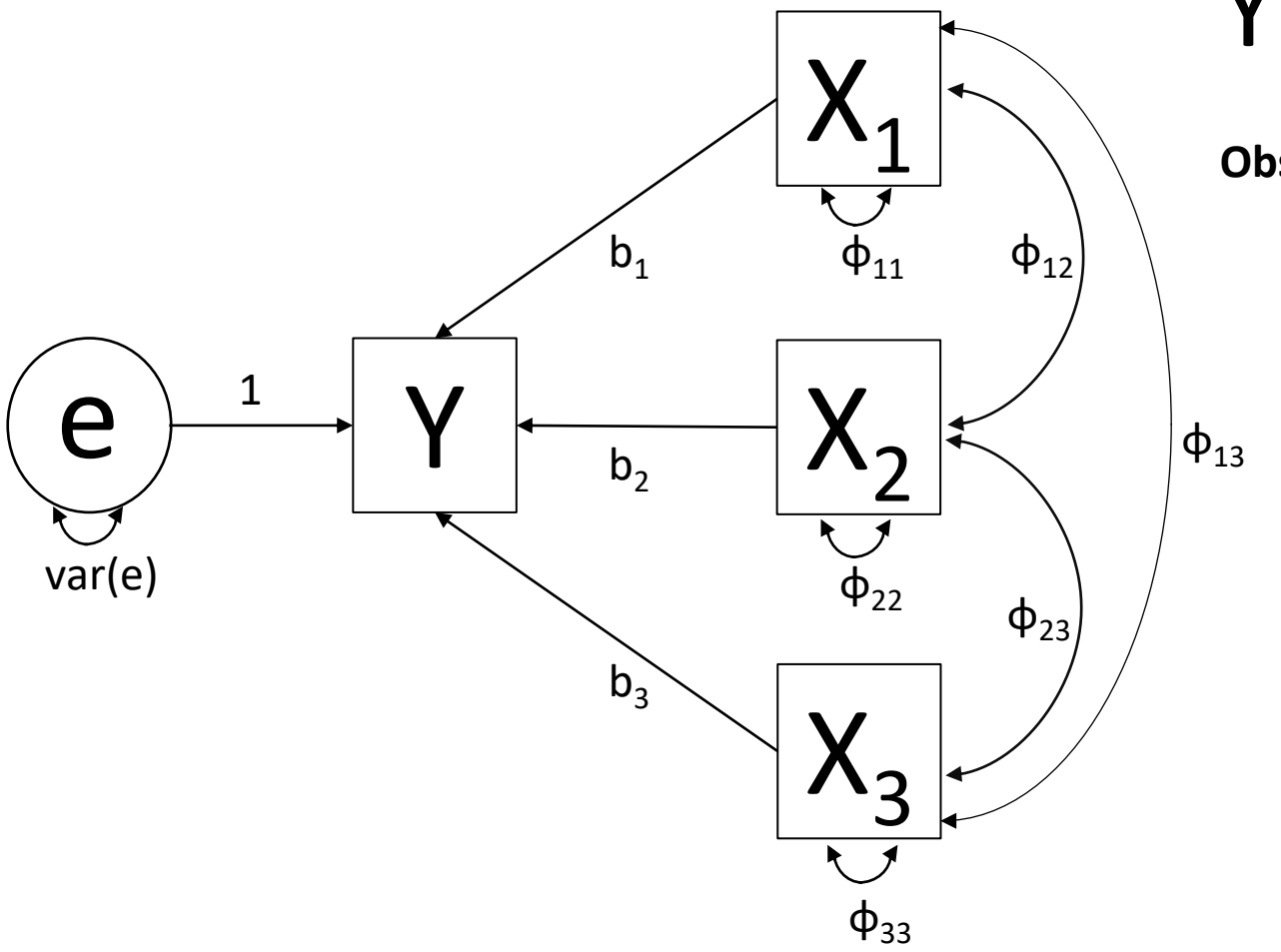
Path Diagrams- Multivariable Regression



Structural Equation:

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + e$$

Path Diagrams- Multivariable Regression

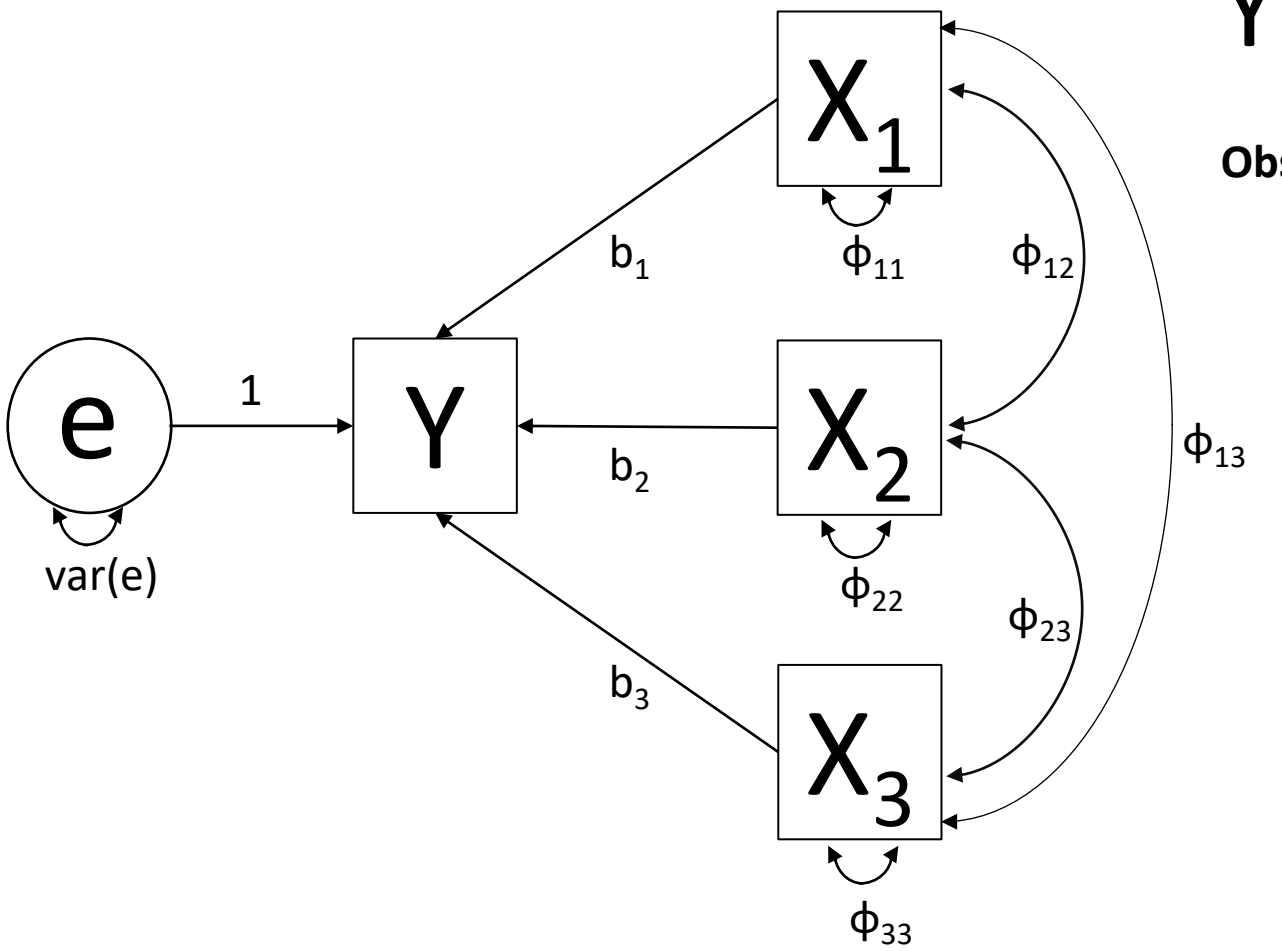


Structural Equation:

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + e$$

Observed Covariance Matrix:

Path Diagrams- Multivariable Regression



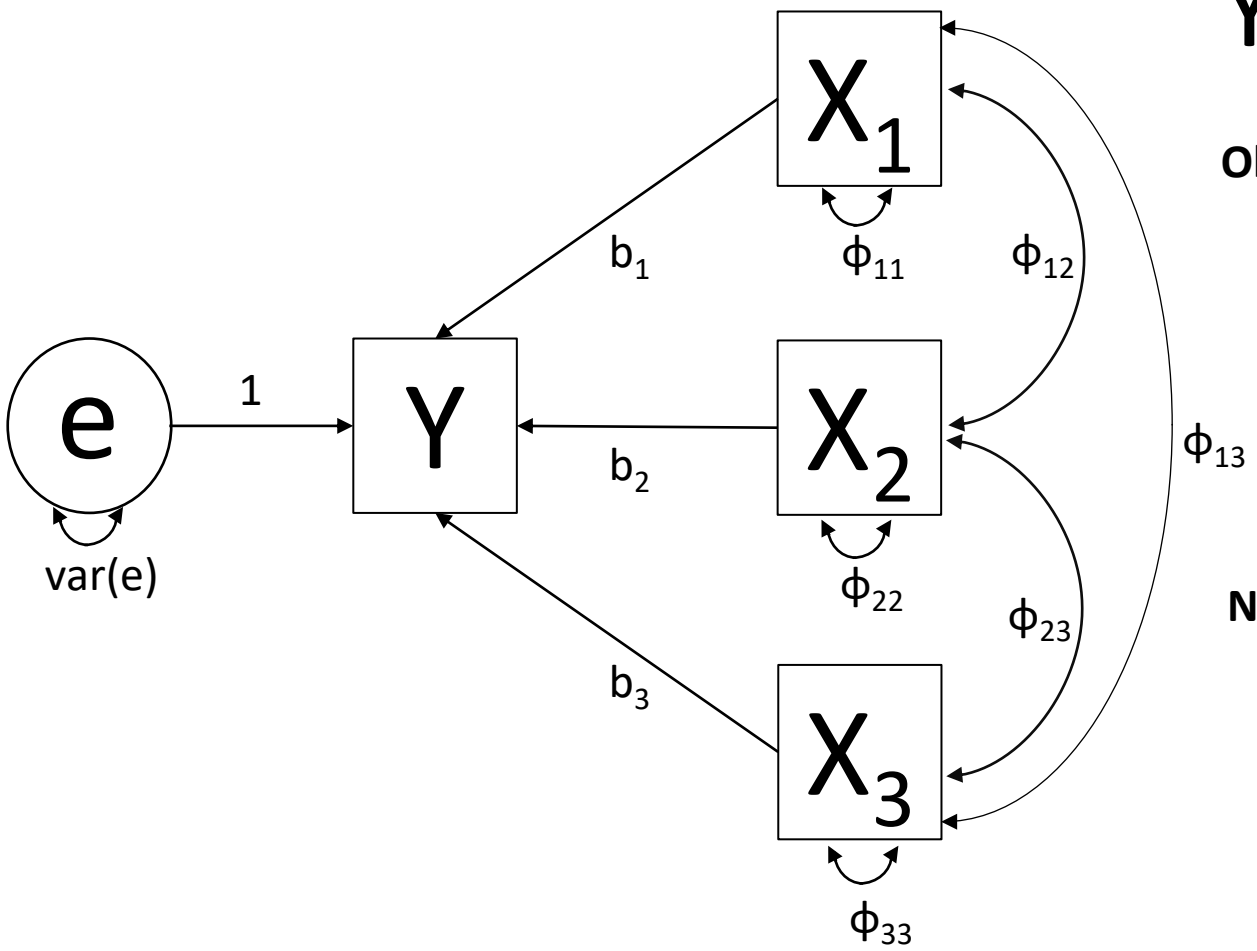
Structural Equation:

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + e$$

Observed Covariance Matrix:

$$\Sigma = \begin{bmatrix} VAR(X_1) & COV(X_1,X_2) & COV(X_1,X_3) & COV(X_1,Y) \\ COV(X_2,X_1) & VAR(X_2) & COV(X_2,X_3) & COV(X_2,Y) \\ COV(X_3,X_1) & COV(X_3,X_2) & VAR(X_3) & COV(X_3,Y) \\ COV(Y,X_1) & COV(Y,X_2) & COV(Y,X_3) & VAR(Y) \end{bmatrix}$$

Path Diagrams- Multivariable Regression



Structural Equation:

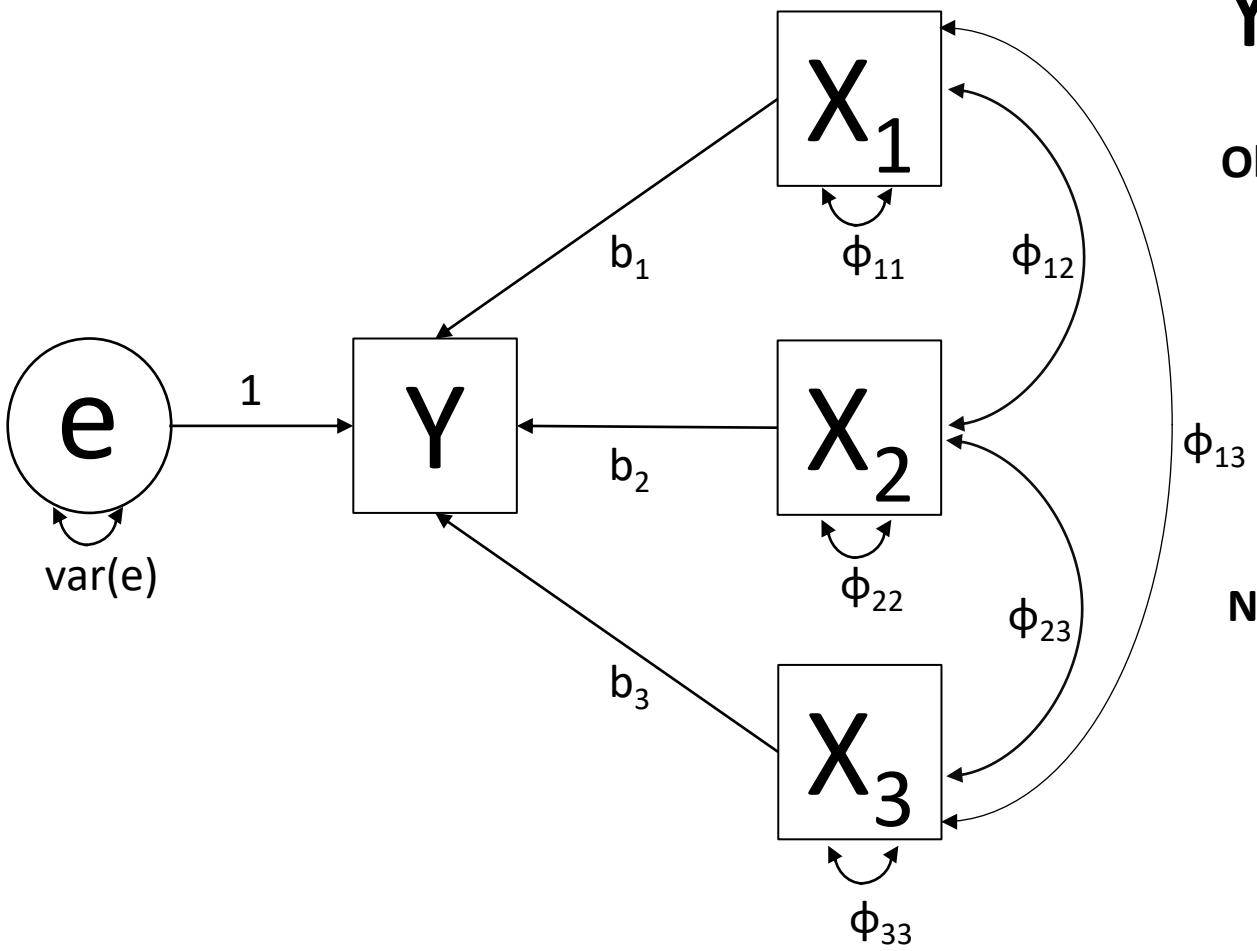
$$Y = b_1X_1 + b_2X_2 + b_3X_3 + e$$

Observed Covariance Matrix:

$$\Sigma = \begin{bmatrix} VAR(X_1) & COV(X_1,X_2) & COV(X_1,X_3) & COV(X_1,Y) \\ COV(X_2,X_1) & VAR(X_2) & COV(X_2,X_3) & COV(X_2,Y) \\ COV(X_3,X_1) & COV(X_3,X_2) & VAR(X_3) & COV(X_3,Y) \\ COV(Y,X_1) & COV(Y,X_2) & COV(Y,X_3) & VAR(Y) \end{bmatrix}$$

Number of observed statistics:

Path Diagrams- Multivariable Regression



Structural Equation:

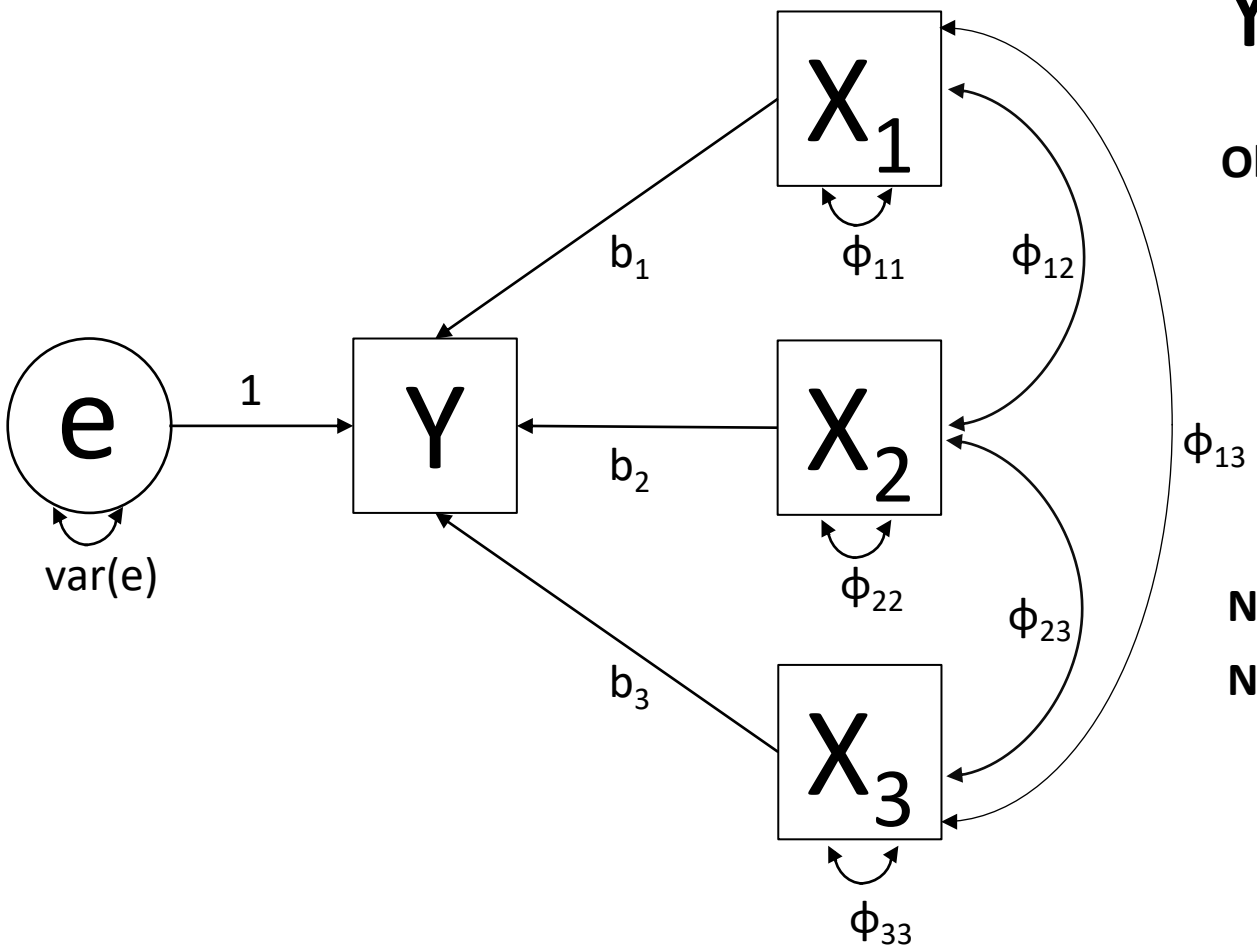
$$Y = b_1X_1 + b_2X_2 + b_3X_3 + e$$

Observed Covariance Matrix:

$$\Sigma = \begin{bmatrix} VAR(X_1) & COV(X_1,X_2) & COV(X_1,X_3) & COV(X_1,Y) \\ COV(X_2,X_1) & VAR(X_2) & COV(X_2,X_3) & COV(X_2,Y) \\ COV(X_3,X_1) & COV(X_3,X_2) & VAR(X_3) & COV(X_3,Y) \\ COV(Y,X_1) & COV(Y,X_2) & COV(Y,X_3) & VAR(Y) \end{bmatrix}$$

Number of observed statistics: 10

Path Diagrams- Multivariable Regression



Structural Equation:

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + e$$

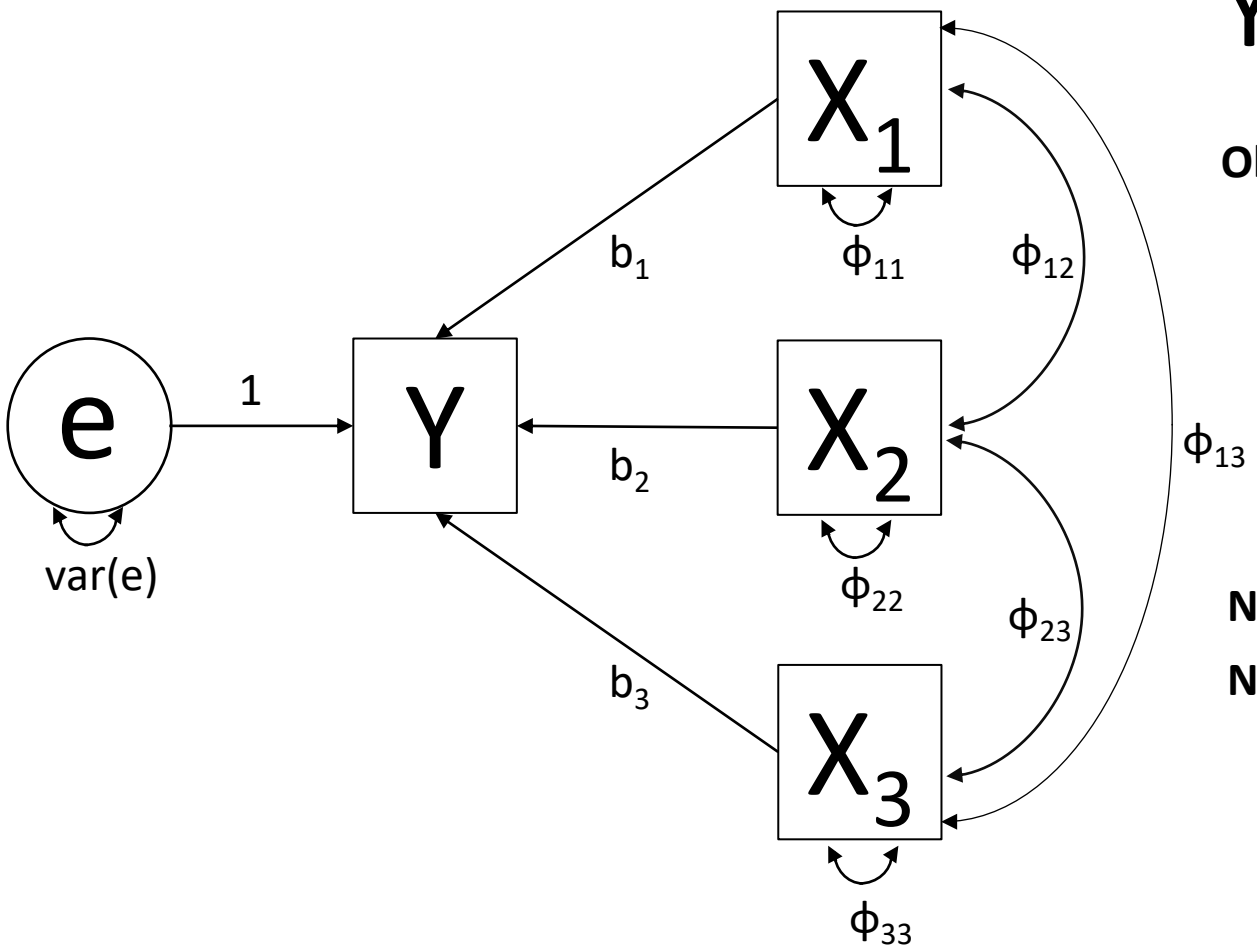
Observed Covariance Matrix:

$$\Sigma = \begin{bmatrix} VAR(X_1) & COV(X_1,X_2) & COV(X_1,X_3) & COV(X_1,Y) \\ COV(X_2,X_1) & VAR(X_2) & COV(X_2,X_3) & COV(X_2,Y) \\ COV(X_3,X_1) & COV(X_3,X_2) & VAR(X_3) & COV(X_3,Y) \\ COV(Y,X_1) & COV(Y,X_2) & COV(Y,X_3) & VAR(Y) \end{bmatrix}$$

Number of observed statistics: 10

Number of estimated parameters:

Path Diagrams- Multivariable Regression



Structural Equation:

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + e$$

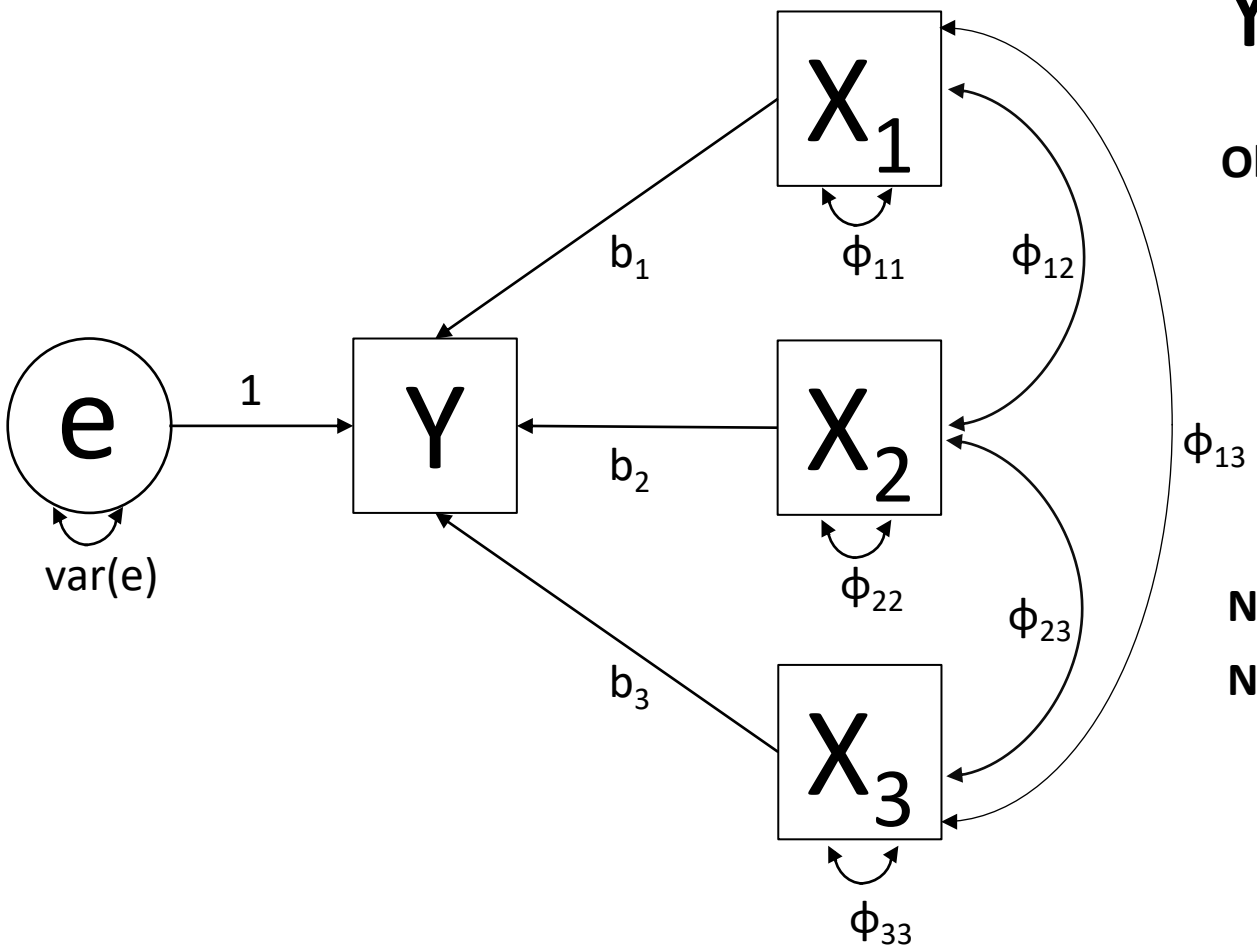
Observed Covariance Matrix:

$$\Sigma = \begin{bmatrix} VAR(X_1) & COV(X_1,X_2) & COV(X_1,X_3) & COV(X_1,Y) \\ COV(X_2,X_1) & VAR(X_2) & COV(X_2,X_3) & COV(X_2,Y) \\ COV(X_3,X_1) & COV(X_3,X_2) & VAR(X_3) & COV(X_3,Y) \\ COV(Y,X_1) & COV(Y,X_2) & COV(Y,X_3) & VAR(Y) \end{bmatrix}$$

Number of observed statistics: 10

Number of estimated parameters: 10

Path Diagrams- Multivariable Regression



Structural Equation:

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + e$$

Observed Covariance Matrix:

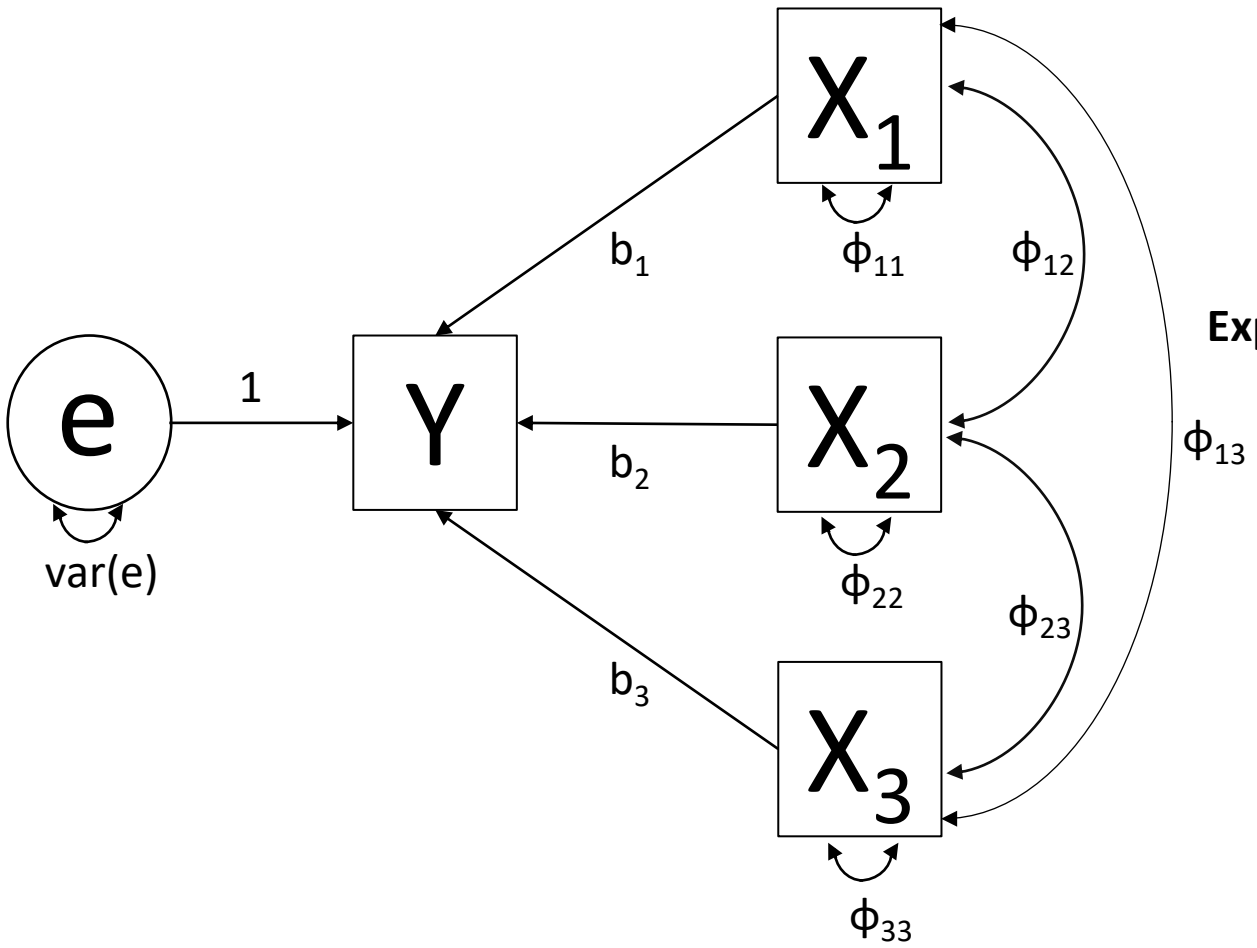
$$\Sigma = \begin{bmatrix} VAR(X_1) & COV(X_1,X_2) & COV(X_1,X_3) & COV(X_1,Y) \\ COV(X_2,X_1) & VAR(X_2) & COV(X_2,X_3) & COV(X_2,Y) \\ COV(X_3,X_1) & COV(X_3,X_2) & VAR(X_3) & COV(X_3,Y) \\ COV(Y,X_1) & COV(Y,X_2) & COV(Y,X_3) & VAR(Y) \end{bmatrix}$$

Number of observed statistics: 10

Number of estimated parameters: 10

$(b_1,b_2,b_3,\phi_{11},\phi_{12},\phi_{13},\phi_{22},\phi_{23},\phi_{33},var(e))$

Path Diagrams- Multivariable Regression



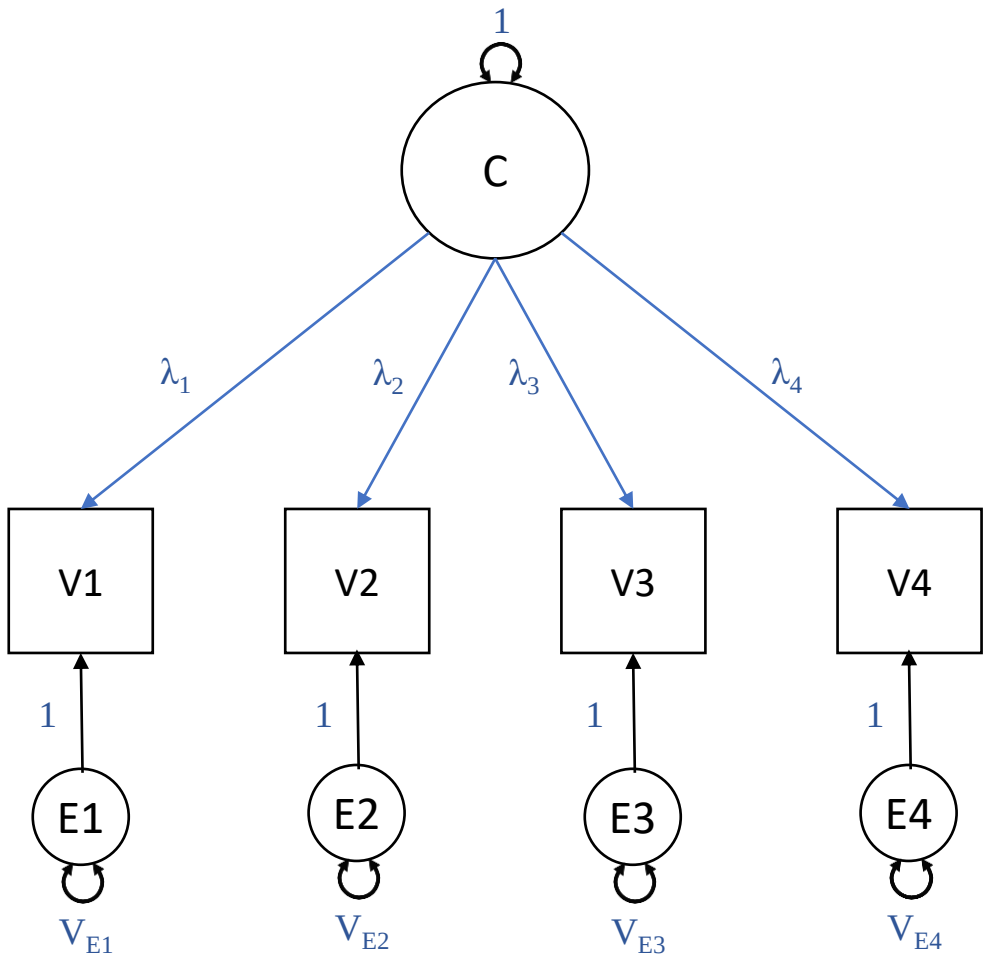
Observed Covariance Matrix:

$$\Sigma = \begin{bmatrix} VAR(X_1) & COV(X_1,X_2) & COV(X_1,X_3) & COV(X_1,Y) \\ COV(X_2,X_1) & VAR(X_2) & COV(X_2,X_3) & COV(X_2,Y) \\ COV(X_3,X_1) & COV(X_3,X_2) & VAR(X_3) & COV(X_3,Y) \\ COV(Y,X_1) & COV(Y,X_2) & COV(Y,X_3) & VAR(Y) \end{bmatrix}$$

Expected Covariance Matrix:

$$\Sigma(\theta) = \begin{bmatrix} \phi_{11} & \phi_{12} & \phi_{13} & b_1\phi_{11}+b_2\phi_{12}+b_3\phi_{13} \\ \phi_{12} & \phi_{22} & \phi_{23} & b_2\phi_{22}+b_1\phi_{12}+b_3\phi_{23} \\ \phi_{13} & \phi_{23} & \phi_{33} & b_3\phi_{33}+b_1\phi_{13}+b_2\phi_{23} \\ b_1\phi_{11}+b_2\phi_{12}+b_3\phi_{13} & b_2\phi_{22}+b_1\phi_{12}+b_3\phi_{23} & b_3\phi_{33}+b_1\phi_{13}+b_2\phi_{23} & b_1^2\phi_{11}+b_2^2\phi_{22}+b_3^2\phi_{33}+2b_1b_2\phi_{12}+2b_1b_3\phi_{13}+2b_2b_3\phi_{23}+var(e) \end{bmatrix}$$

Path Diagrams



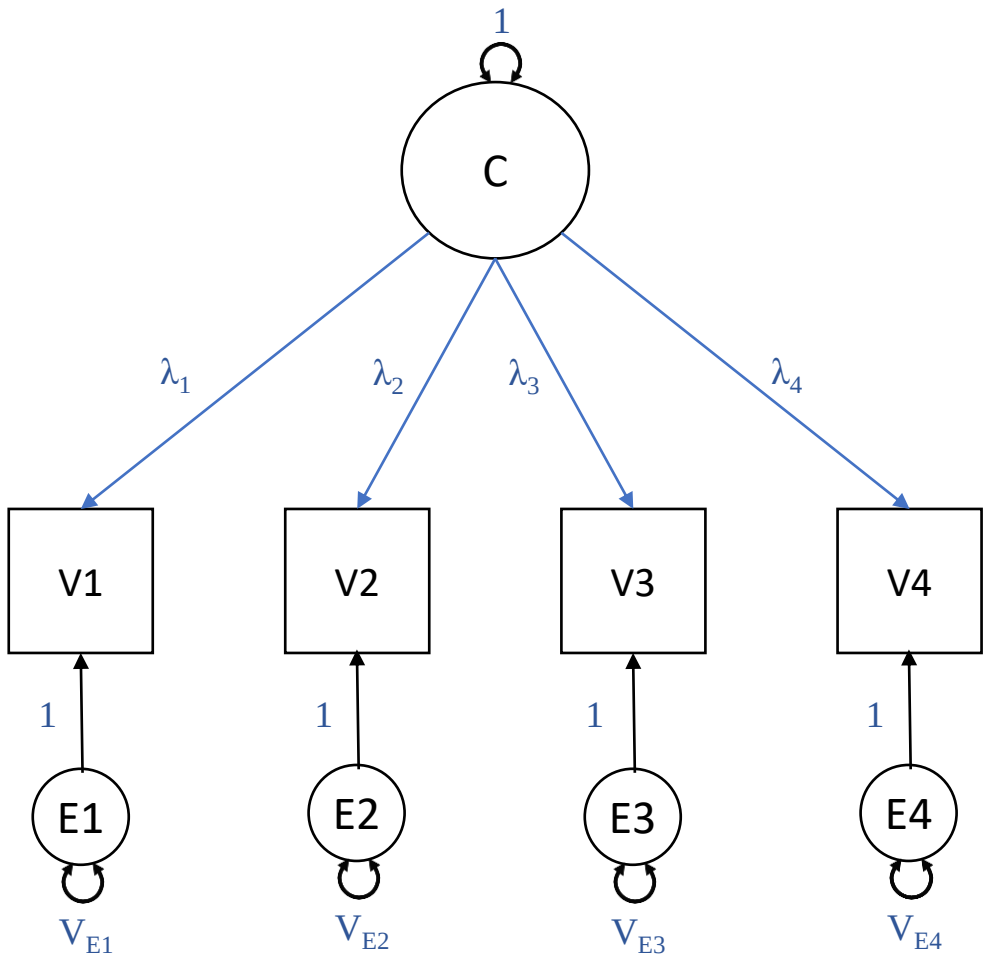
Structural Equations:

Observed Covariance Matrix:

Number of observed statistics:

Number of estimated parameters:

Path Diagrams- Common Factor Model



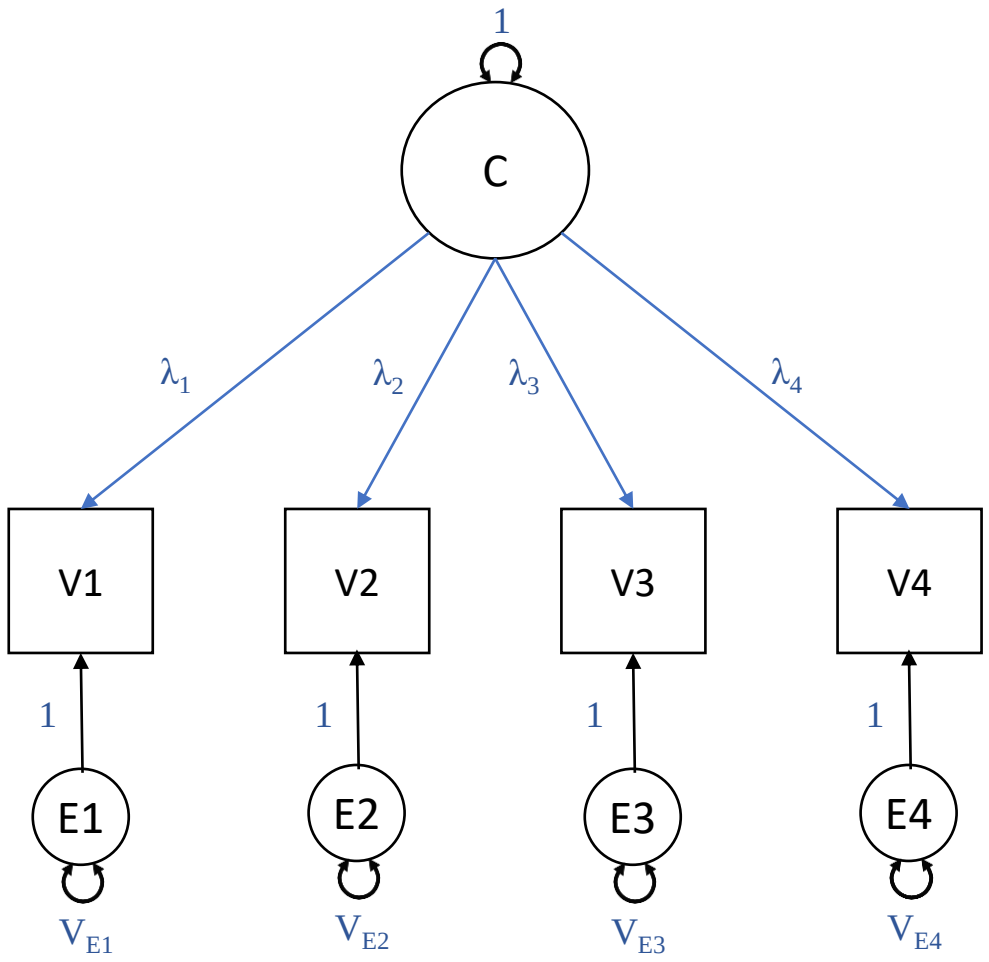
Structural Equations:

Observed Covariance Matrix:

Number of observed statistics:

Number of estimated parameters:

Path Diagrams- Common Factor Model



Structural Equations:

$V_1 = \lambda_1 C + E_1$

$V_2 = \lambda_2 C + E_2$

$V_3 = \lambda_3 C + E_3$

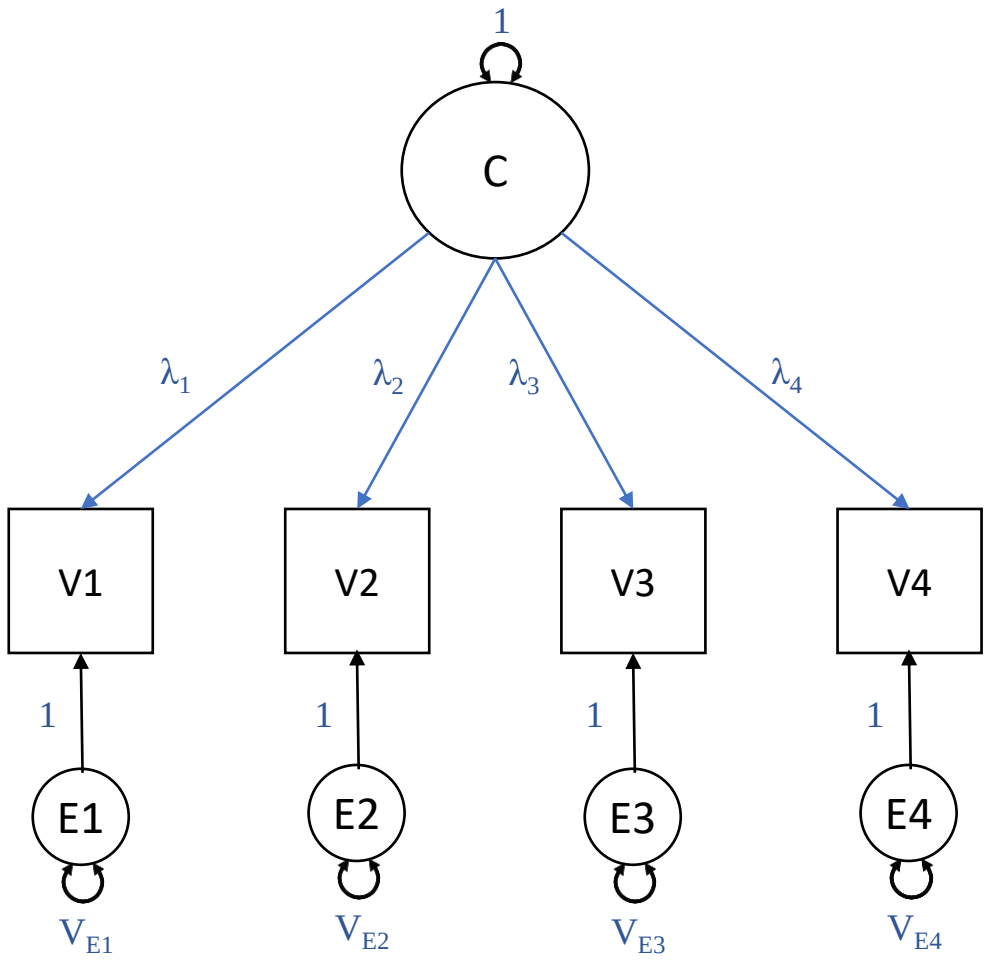
$V_4 = \lambda_4 C + E_4$

Observed Covariance Matrix:

Number of observed statistics:

Number of estimated parameters:

Path Diagrams- Common Factor Model



Structural Equations:

$V_1 = \lambda_1 C + E_1$

$V_2 = \lambda_2 C + E_2$

$V_3 = \lambda_3 C + E_3$

$V_4 = \lambda_4 C + E_4$

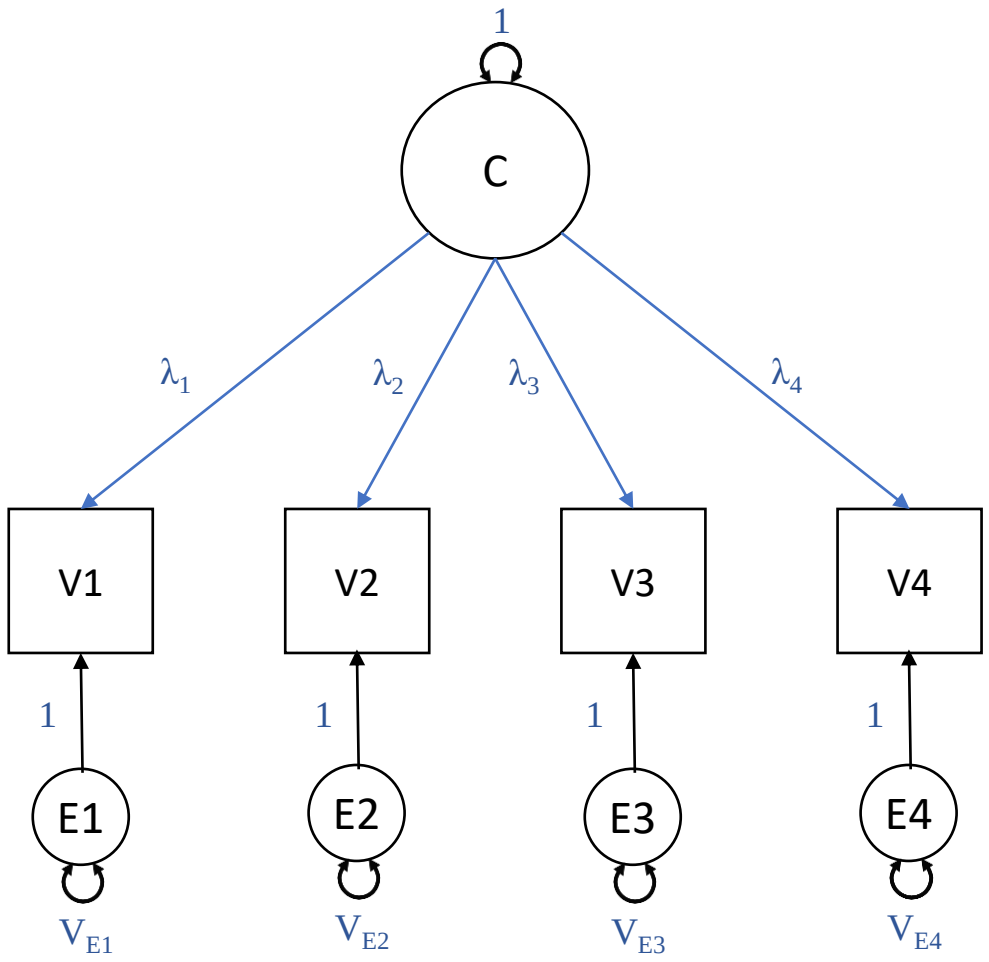
Observed Covariance Matrix:

$$\Sigma = \begin{bmatrix} VAR(V_1) & COV(V_1,V_2) & COV(V_1,V_3) & COV(V_1,V_4) \\ COV(V_2,V_1) & VAR(V_2) & COV(V_2,V_3) & COV(V_2,V_4) \\ COV(V_3,V_1) & COV(V_3,V_2) & VAR(V_3) & COV(V_3,V_4) \\ COV(V_4,V_1) & COV(V_4,V_2) & COV(V_4,V_3) & VAR(V_4) \end{bmatrix}$$

Number of observed statistics:

Number of estimated parameters:

Path Diagrams- Common Factor Model



Structural Equations:

$V_1 = \lambda_1 C + E_1$

$V_2 = \lambda_2 C + E_2$

$V_3 = \lambda_3 C + E_3$

$V_4 = \lambda_4 C + E_4$

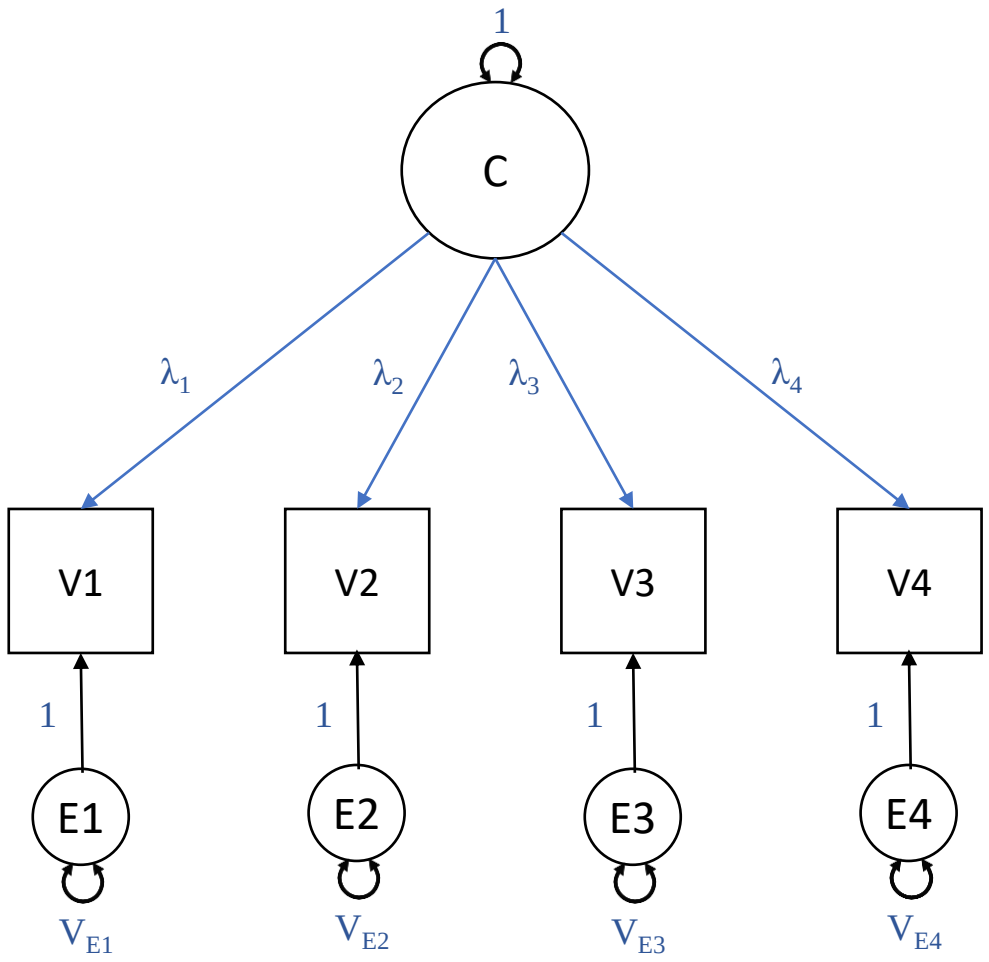
Observed Covariance Matrix:

$$\Sigma = \begin{bmatrix} VAR(V_1) & COV(V_1,V_2) & COV(V_1,V_3) & COV(V_1,V_4) \\ COV(V_2,V_1) & VAR(V_2) & COV(V_2,V_3) & COV(V_2,V_4) \\ COV(V_3,V_1) & COV(V_3,V_2) & VAR(V_3) & COV(V_3,V_4) \\ COV(V_4,V_1) & COV(V_4,V_2) & COV(V_4,V_3) & VAR(V_4) \end{bmatrix}$$

Number of observed statistics: 10

Number of estimated parameters:

Path Diagrams- Common Factor Model



Structural Equations:

$V_1 = \lambda_1 C + E_1$

$V_2 = \lambda_2 C + E_2$

$V_3 = \lambda_3 C + E_3$

$V_4 = \lambda_4 C + E_4$

Observed Covariance Matrix:

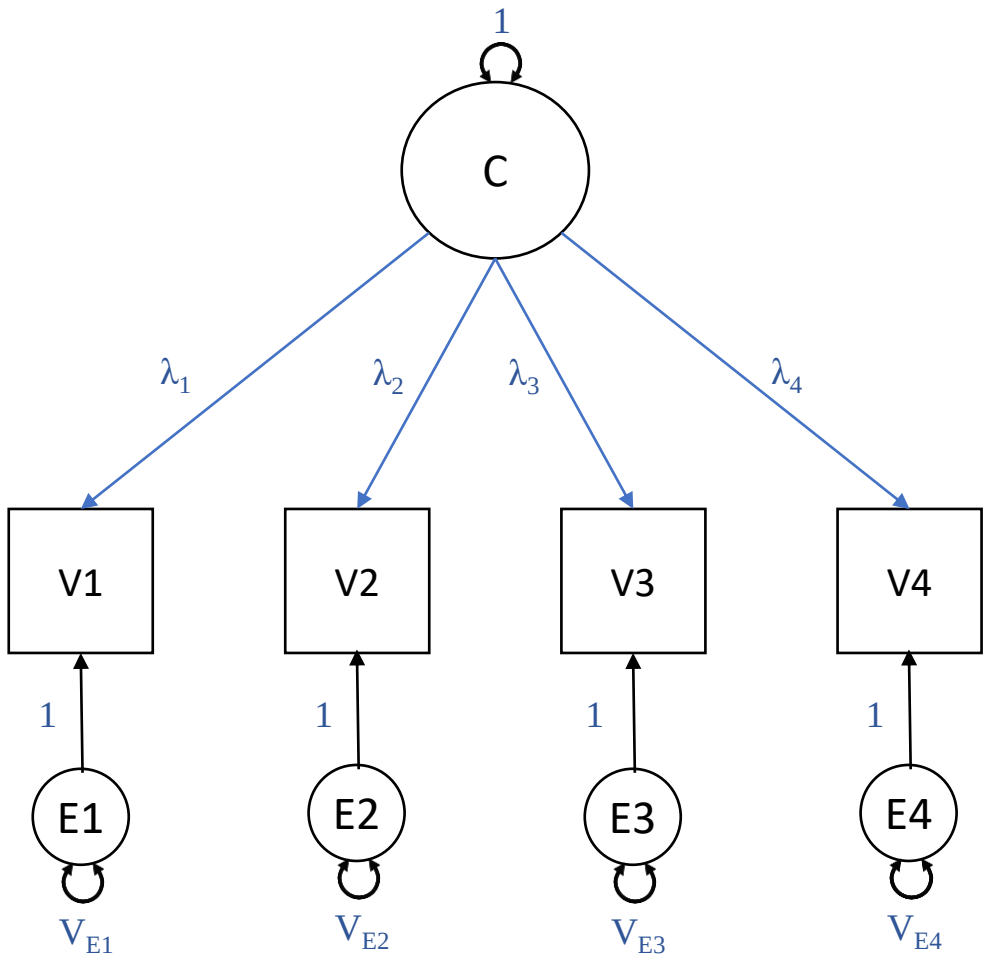
$$\Sigma = \begin{bmatrix} VAR(V_1) & COV(V_1,V_2) & COV(V_1,V_3) & COV(V_1,V_4) \\ COV(V_2,V_1) & VAR(V_2) & COV(V_2,V_3) & COV(V_2,V_4) \\ COV(V_3,V_1) & COV(V_3,V_2) & VAR(V_3) & COV(V_3,V_4) \\ COV(V_4,V_1) & COV(V_4,V_2) & COV(V_4,V_3) & VAR(V_4) \end{bmatrix}$$

Number of observed statistics: 10

Number of estimated parameters: 8

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4, V_{E1}, V_{E2}, V_{E3}, V_{E4})$

Path Diagrams- Common Factor Model



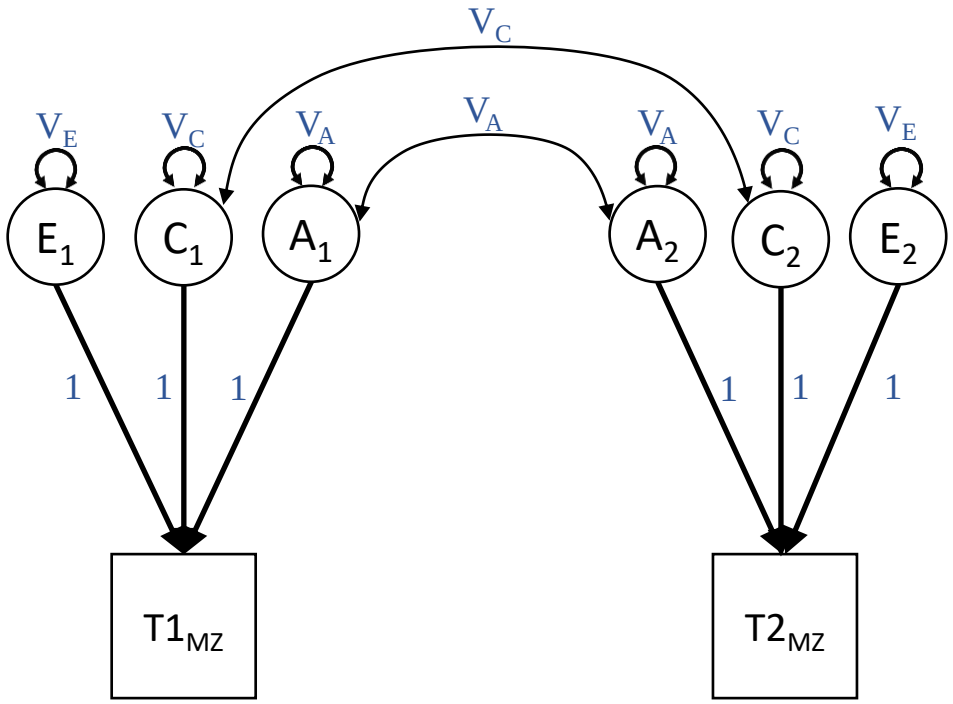
Observed Covariance Matrix:

$$\Sigma = \begin{bmatrix} VAR(V_1) & COV(V_1,V_2) & COV(V_1,V_3) & COV(V_1,V_4) \\ COV(V_2,V_1) & VAR(V_2) & COV(V_2,V_3) & COV(V_2,V_4) \\ COV(V_3,V_1) & COV(V_3,V_2) & VAR(V_3) & COV(V_3,V_4) \\ COV(V_4,V_1) & COV(V_4,V_2) & COV(V_4,V_3) & VAR(V_4) \end{bmatrix}$$

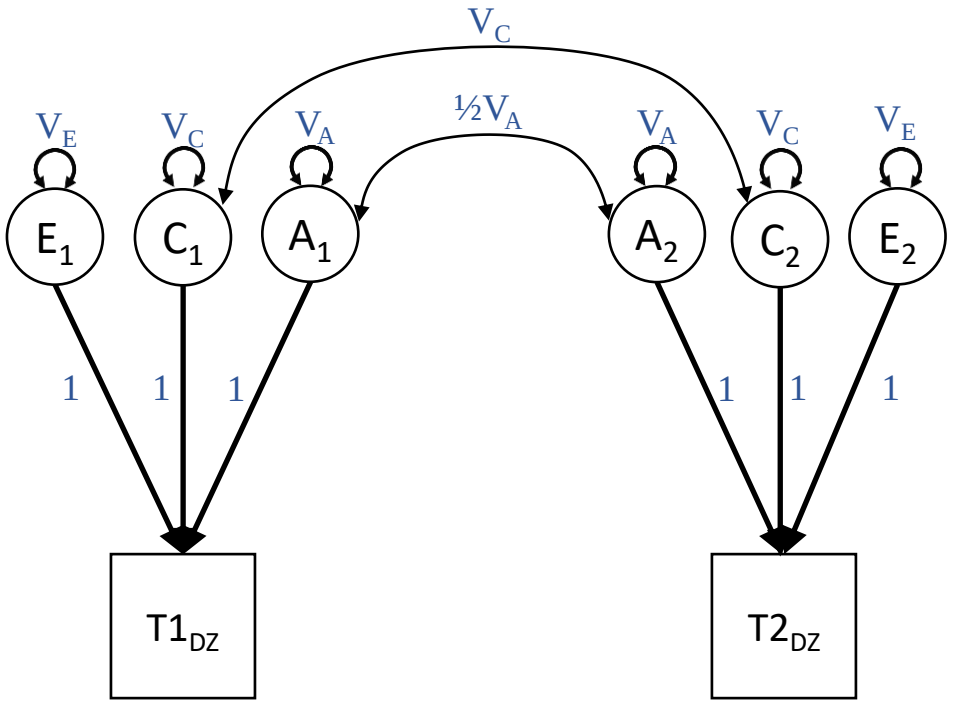
Expected Covariance Matrix:

$$\Sigma(\theta) = \begin{bmatrix} \lambda_1^2+V_{E1} & \lambda_1\lambda_2 & \lambda_1\lambda_3 & \lambda_1\lambda_4 \\ \lambda_2\lambda_1 & \lambda_2^2+V_{E2} & \lambda_2\lambda_3 & \lambda_2\lambda_4 \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3^2+V_{E3} & \lambda_3\lambda_4 \\ \lambda_4\lambda_1 & \lambda_4\lambda_2 & \lambda_4\lambda_3 & \lambda_4^2+V_{E4} \end{bmatrix}$$

Path Diagrams

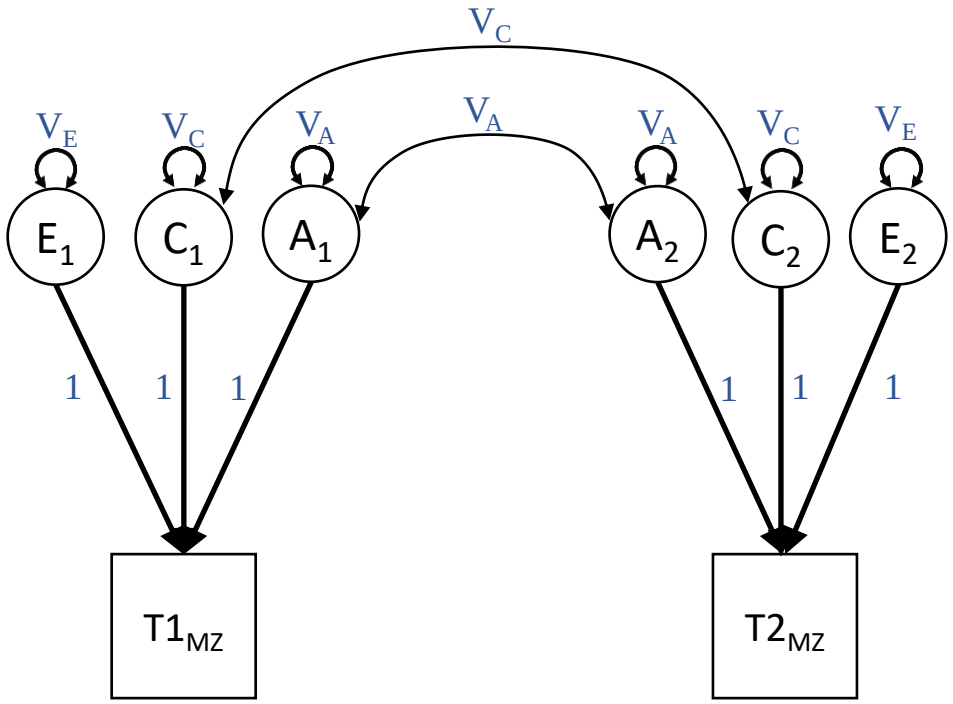


Monozygotic Twins

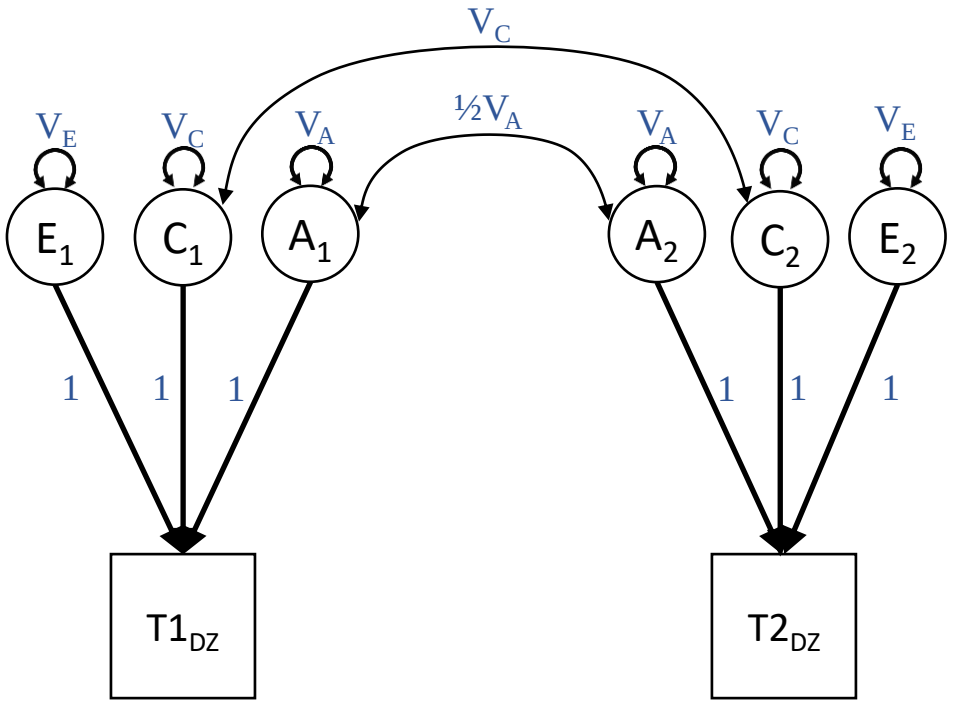


Dizygotic Twins

Path Diagrams- Classical Twin Design



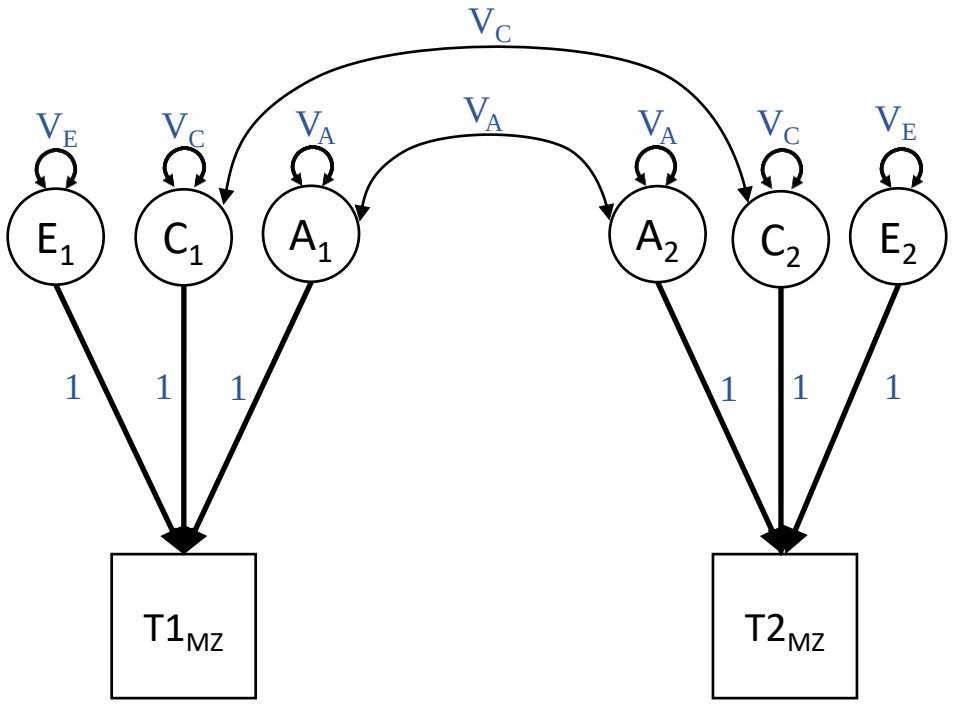
Monozygotic Twins



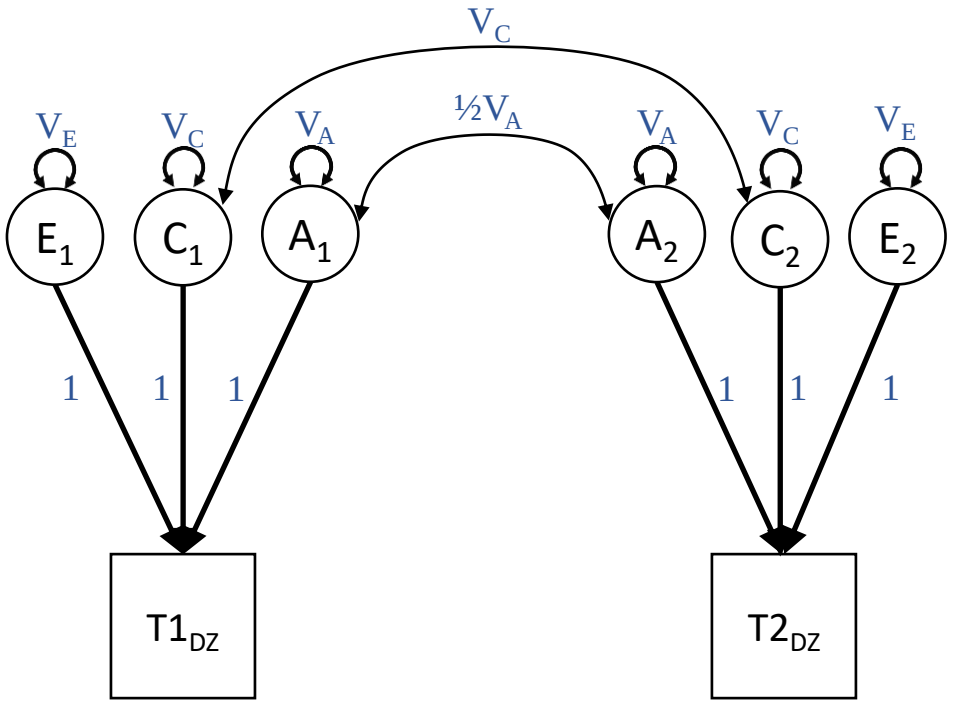
Dizygotic Twins

Structural Equations:

Path Diagrams- Classical Twin Design



Monozygotic Twins



Dizygotic Twins

Structural Equations:

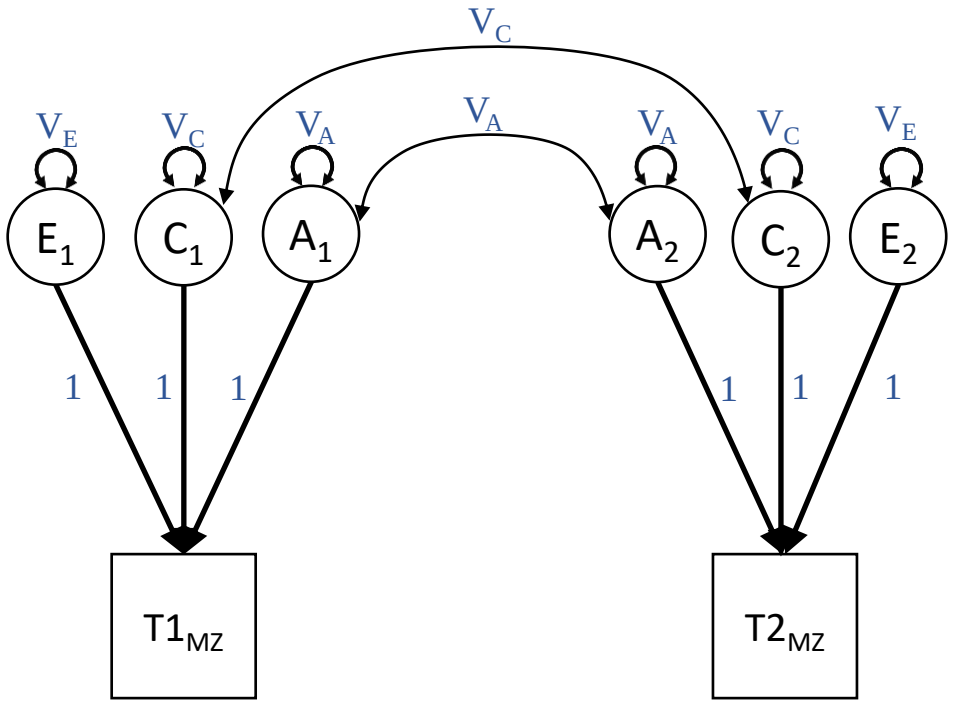
$T1_{MZ} = A_1 + C_1 + E_1$

$T2_{MZ} = A_2 + C_2 + E_2$

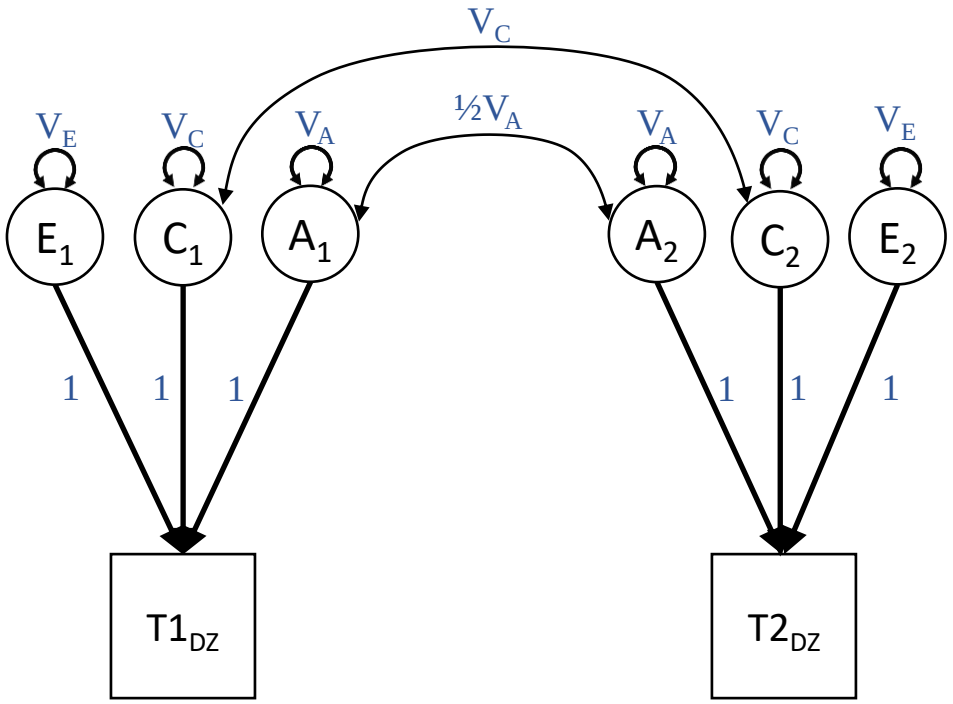
$T1_{DZ} = A_1 + C_1 + E_1$

$T2_{DZ} = A_2 + C_2 + E_2$

Path Diagrams- Classical Twin Design



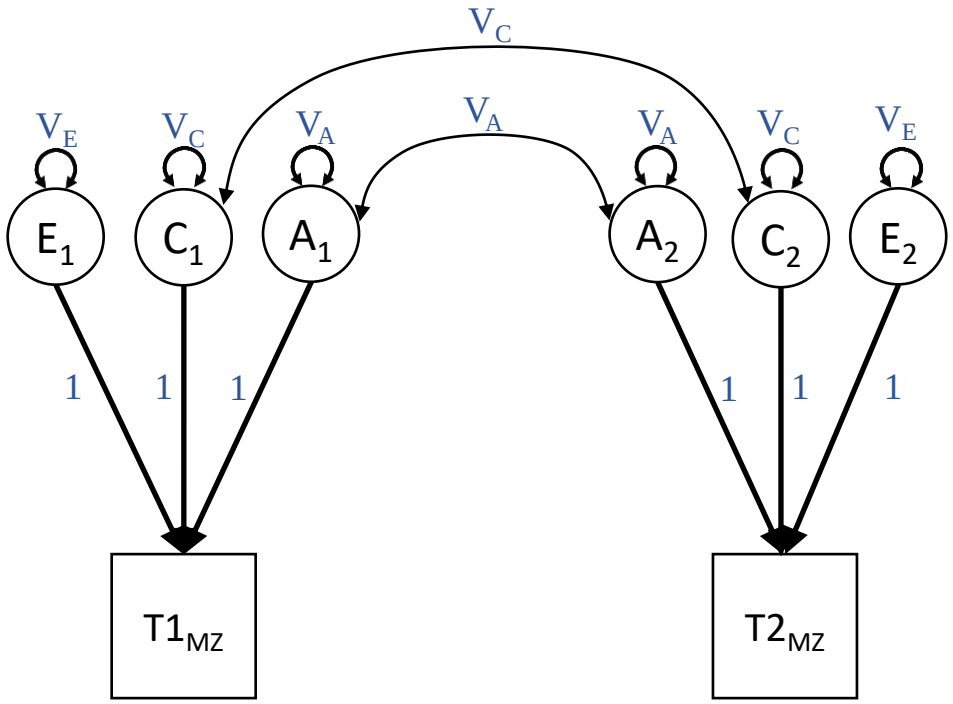
Monozygotic Twins



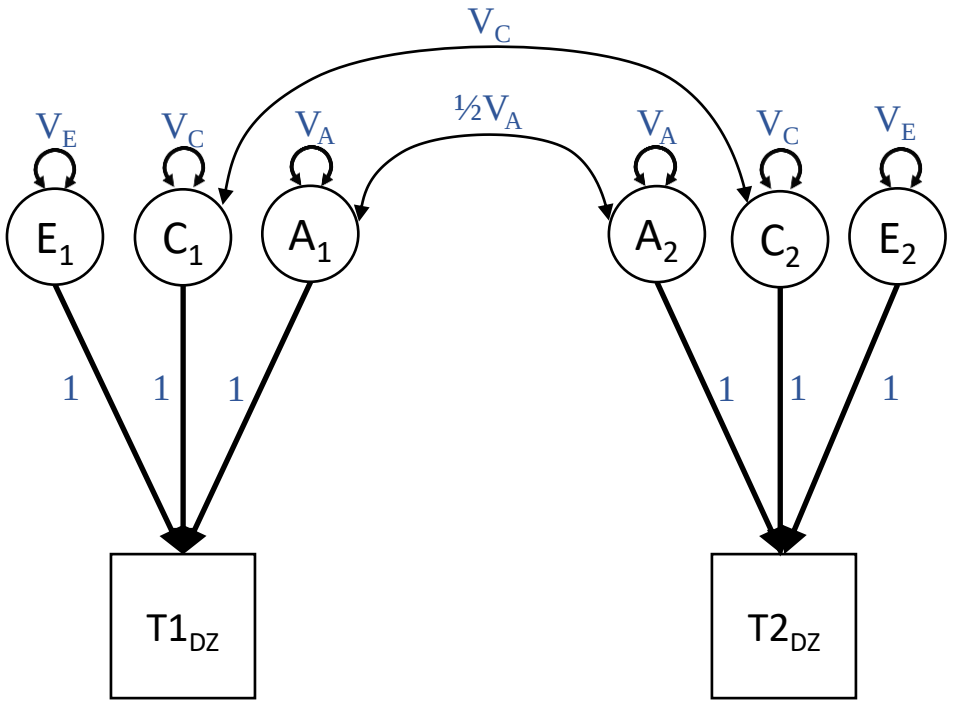
Dizygotic Twins

Number of Observed Variables:

Path Diagrams- Classical Twin Design



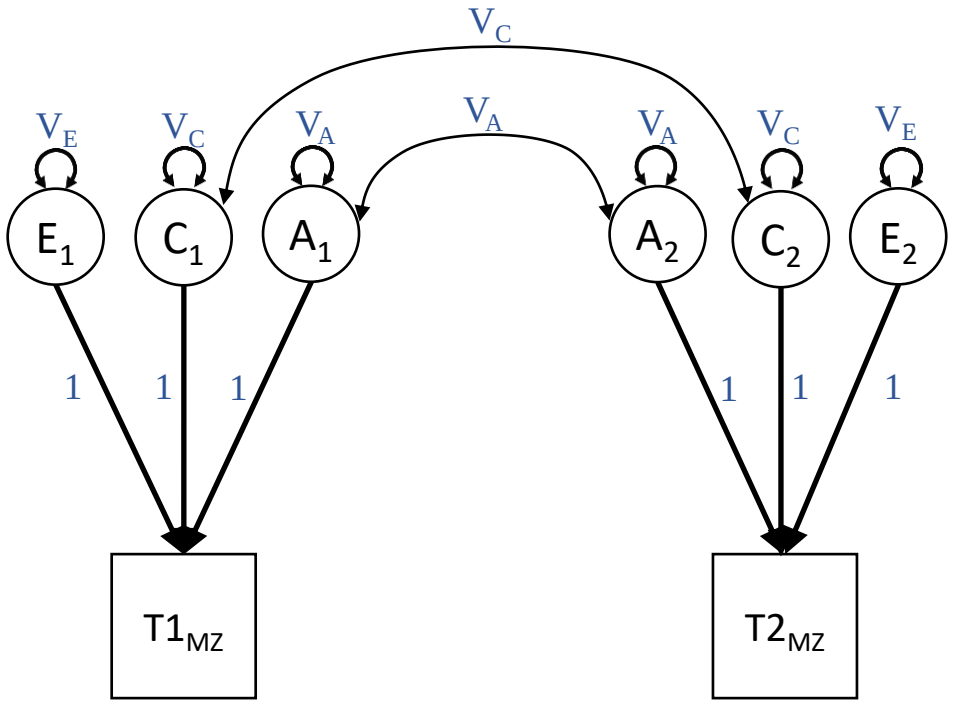
Monozygotic Twins



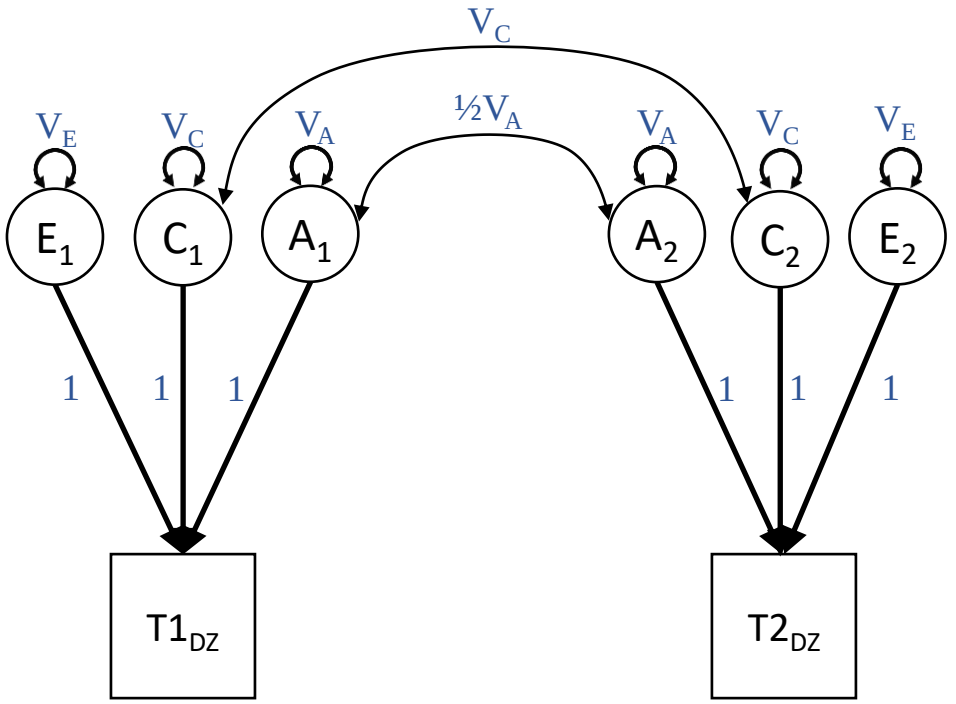
Dizygotic Twins

Number of Observed Variables: 3

Path Diagrams- Classical Twin Design



Monozygotic Twins



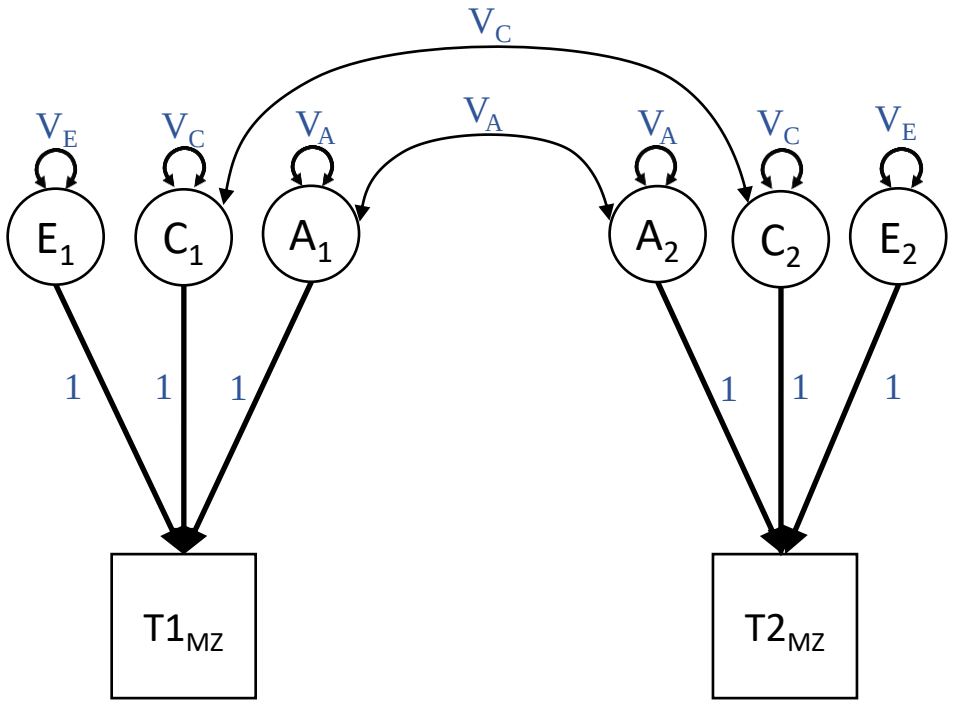
Dizygotic Twins

Observed Covariance Matrices:

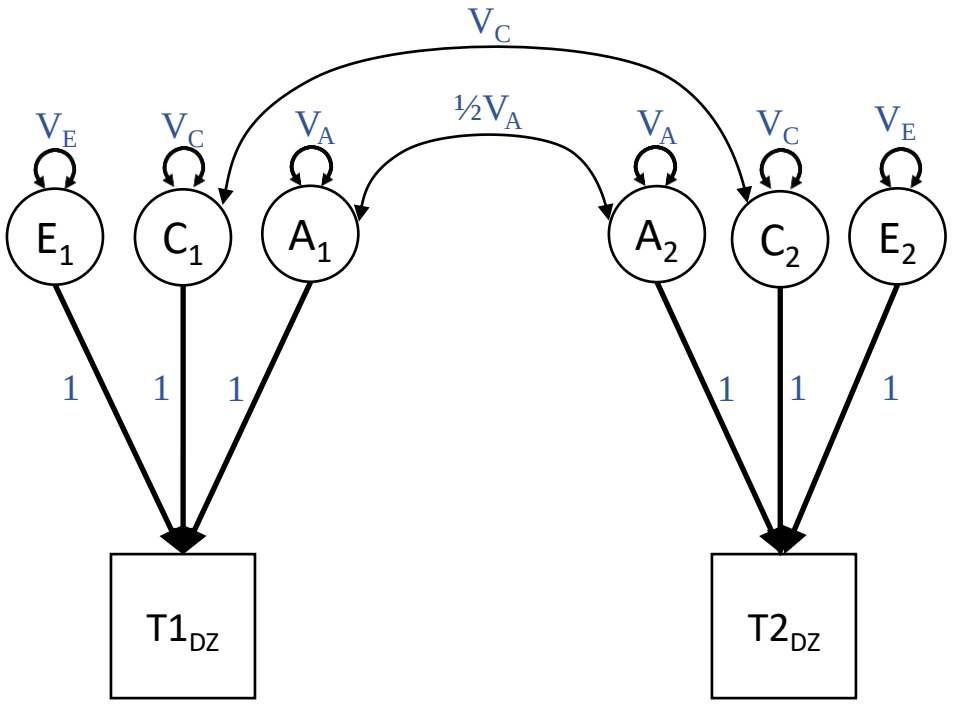
$$\Sigma_{MZ} = \begin{matrix} VAR(T1_{MZ}) & COV(T1_{MZ},T2_{MZ}) \\ COV(T1_{MZ},T2_{MZ}) & VAR(T2_{MZ}) \end{matrix}$$

$$\Sigma_{DZ} = \begin{matrix} VAR(T1_{DZ}) & COV(T1_{DZ},T2_{DZ}) \\ COV(T1_{DZ},T2_{DZ}) & VAR(T2_{DZ}) \end{matrix}$$

Path Diagrams- Classical Twin Design



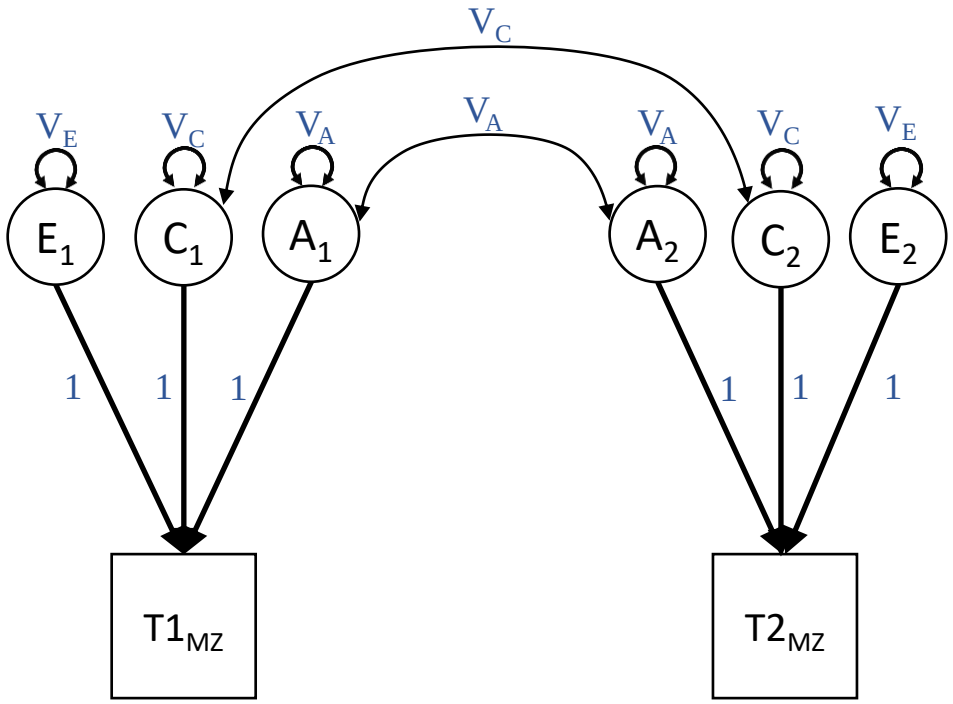
Monozygotic Twins



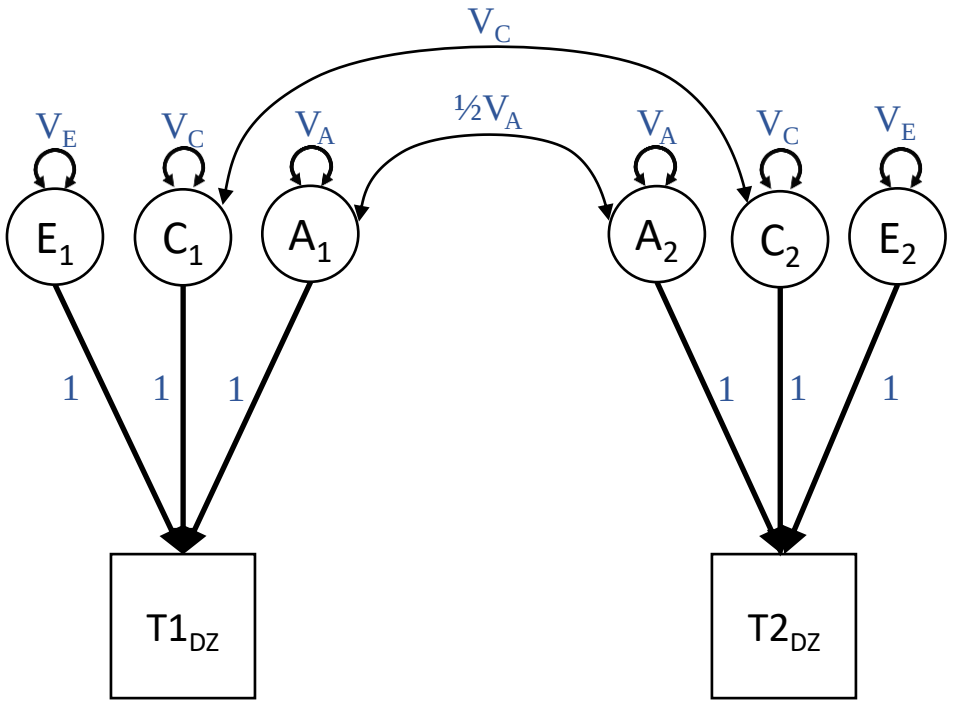
Dizygotic Twins

Number of parameters: 3; **θ** = (V_A, V_C, V_E)

Path Diagrams- Classical Twin Design



Monozygotic Twins



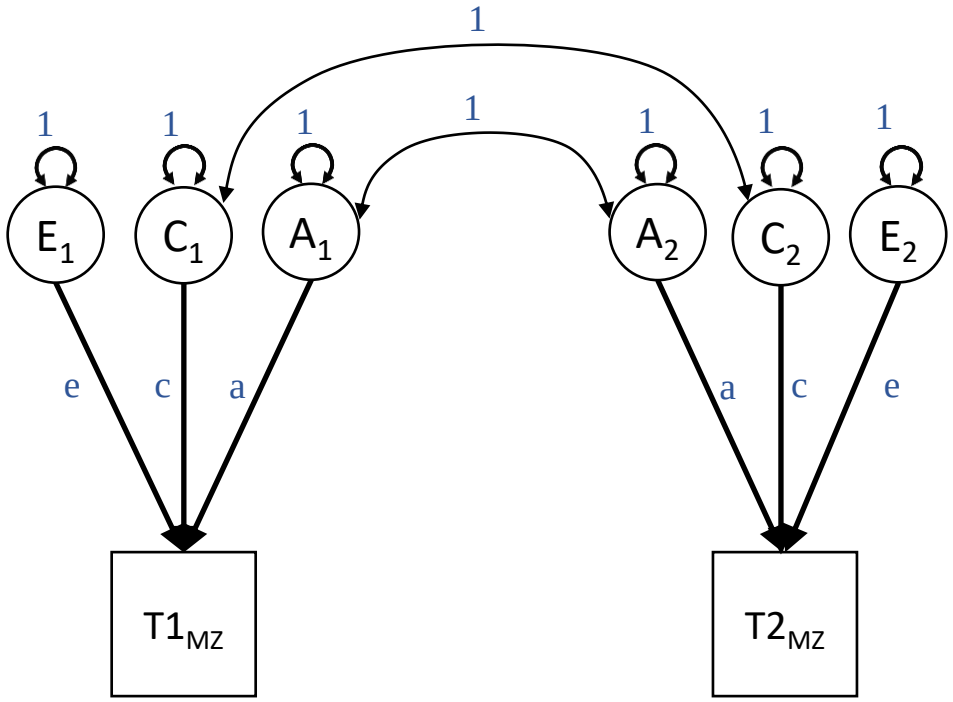
Dizygotic Twins

Expected Covariance Matrices:

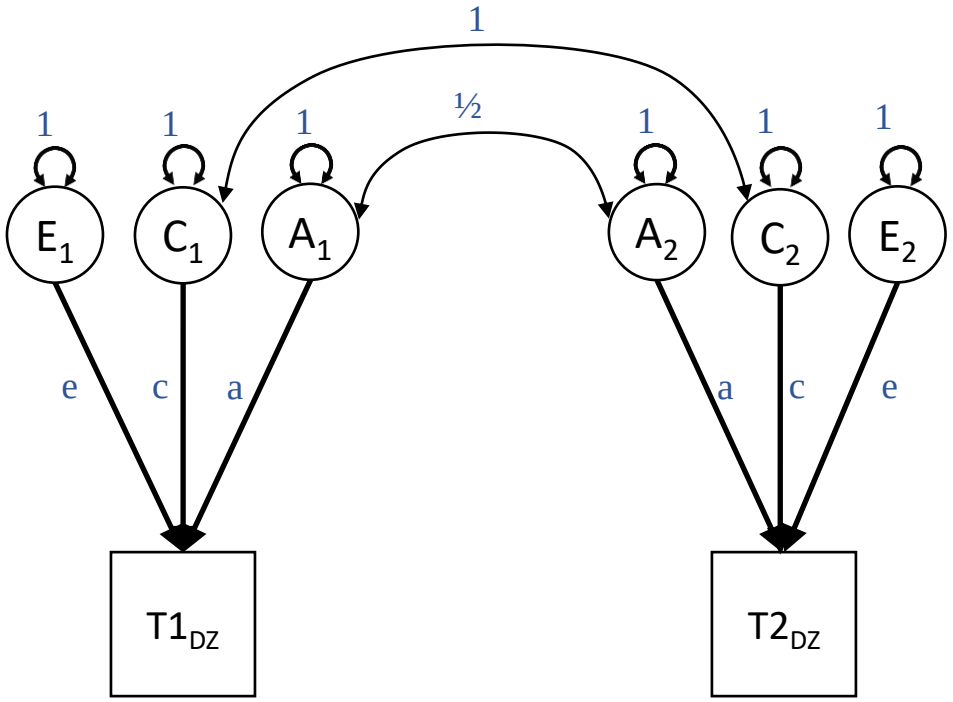
$$\Sigma_{MZ} = \begin{matrix} V_A+V_C+V_E & V_A+V_C \\ V_A+V_C & V_A+V_C+V_E \end{matrix}$$

$$\Sigma_{DZ} = \begin{matrix} V_A+V_C+V_E & ½V_A+V_C \\ ½V_A+V_C & V_A+V_C+V_E \end{matrix}$$

Path Diagrams- Classical Twin Design



Monozygotic Twins



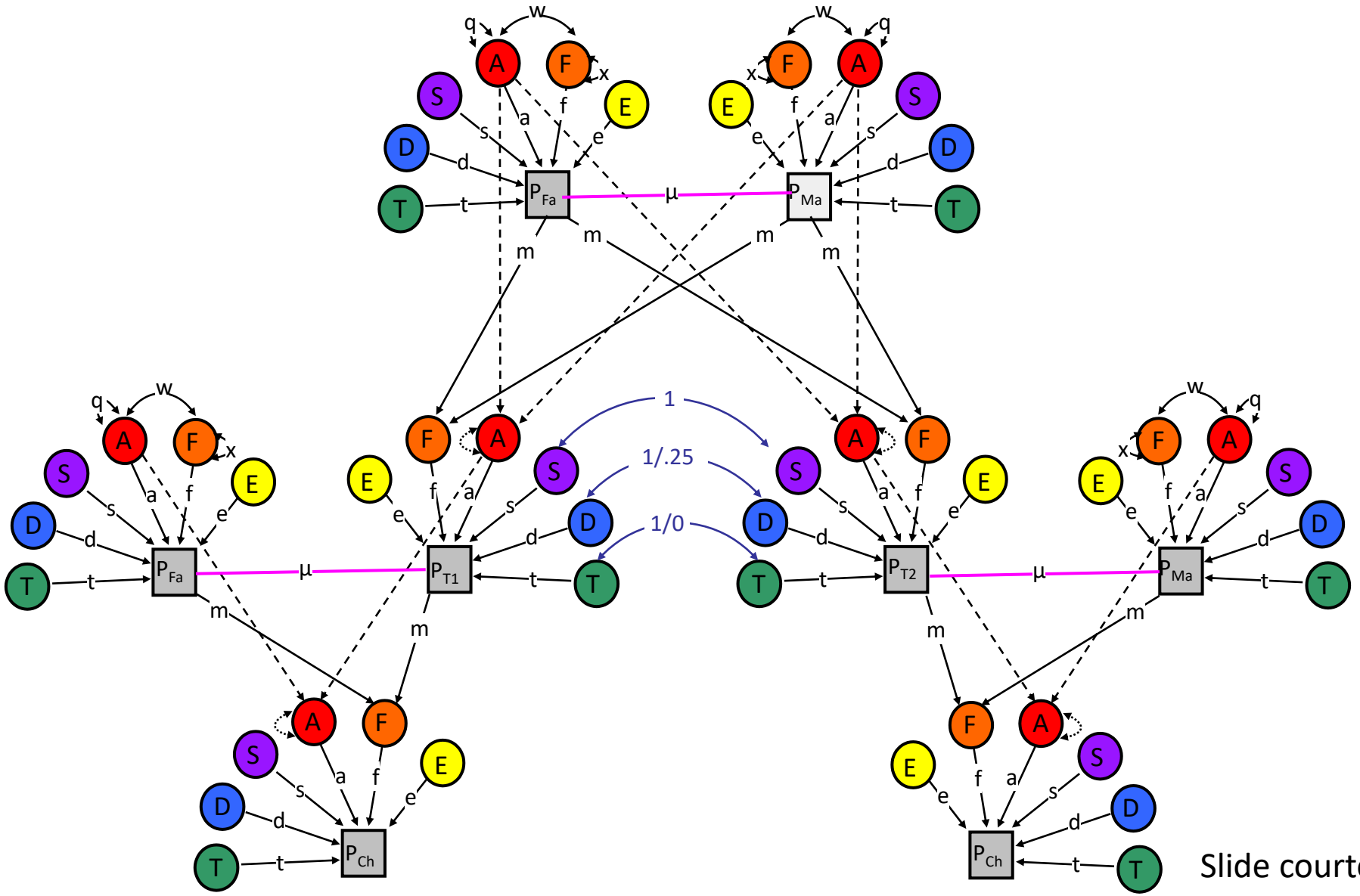
Dizygotic Twins

Expected Covariance Matrices:

$$\Sigma_{MZ} = \begin{matrix} a^2+c^2+e^2 & a^2+c^2 \\ a^2+c^2 & a^2+c^2+e^2 \end{matrix}$$

$$\Sigma_{DZ} = \begin{matrix} a^2+c^2+e^2 & \frac{1}{2}a^2+c^2 \\ \frac{1}{2}a^2+c^2 & a^2+c^2+e^2 \end{matrix}$$

Path Diagrams- Extended Twin Design



Slide courtesy of Matt Keller

Other Path Models

- Mendelian randomization models
- G-REML models
- Multivariate models
- Models involving feedback loops
- Many, many others…

Deriving Expected Variances and Covariances Using Path Tracing Rules

Deriving variances & covariances

1. Identify all legitimate chains (a series of paths) that connect one variable to another (covariances) or connect a variable back to itself (variances)
2. The expected value of a chain is the product of all coefficients associated with each path making up that chain
3. The final expected variance or covariance equals the sum of the values of all legitimate chains

Path Tracing Rules. Legitimate chains:

1. All chains begin by travelling backwards against the direction of a (single or double-headed) arrow, head to tail.

Path Tracing Rules. Legitimate chains:

1. All chains begin by travelling backwards against the direction of a (single or double-headed) arrow, head to tail.
2. Once a double headed arrow has been traversed, the direction reverses such that the chain travels forward

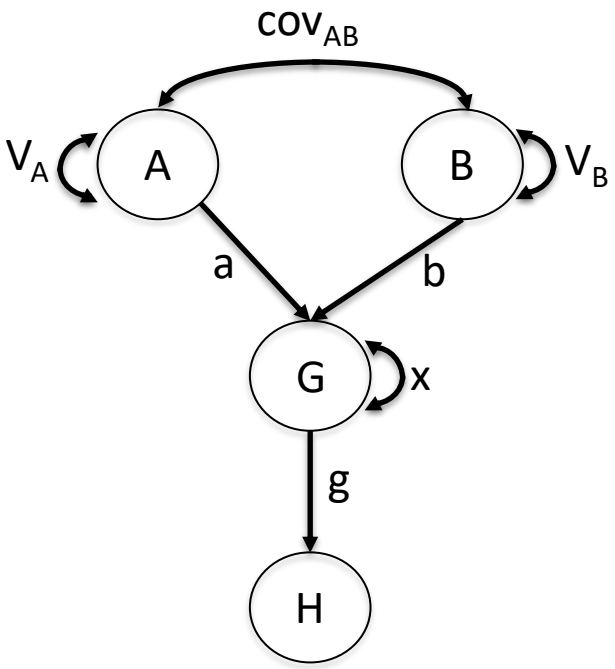
Path Tracing Rules. Legitimate chains:

1. All chains begin by travelling backwards against the direction of a (single or double-headed) arrow, head to tail.
2. Once a double headed arrow has been traversed, the direction reverses such that the chain travels forward
3. All chains must include exactly one double-headed arrow. This implies a chain must change directions exactly once.

Path Tracing Rules. Legitimate chains:

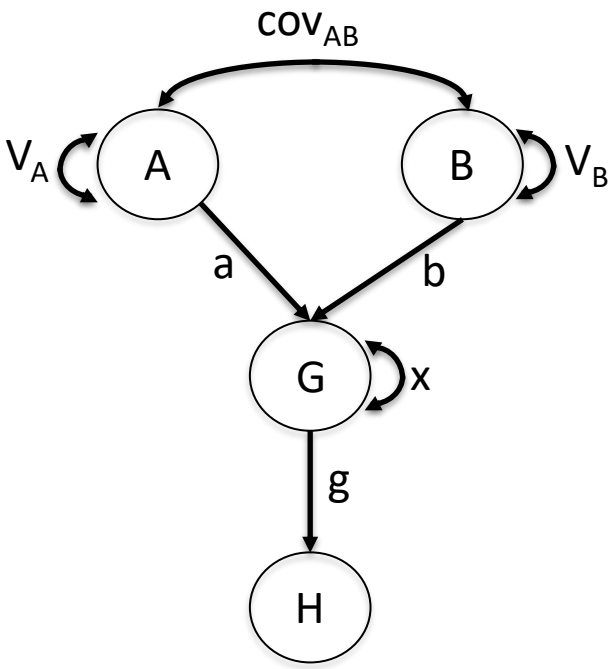
1. All chains begin by travelling backwards against the direction of a (single or double-headed) arrow, head to tail.
2. Once a double headed arrow has been traversed, the direction reverses such that the chain travels forward
3. All chains must include exactly one double-headed arrow. This implies a chain must change directions exactly once.
4. All chains must be counted exactly once and each must be unique. However, order matters: *abc* is a distinct chain from *cba*.

Path Tracing Example



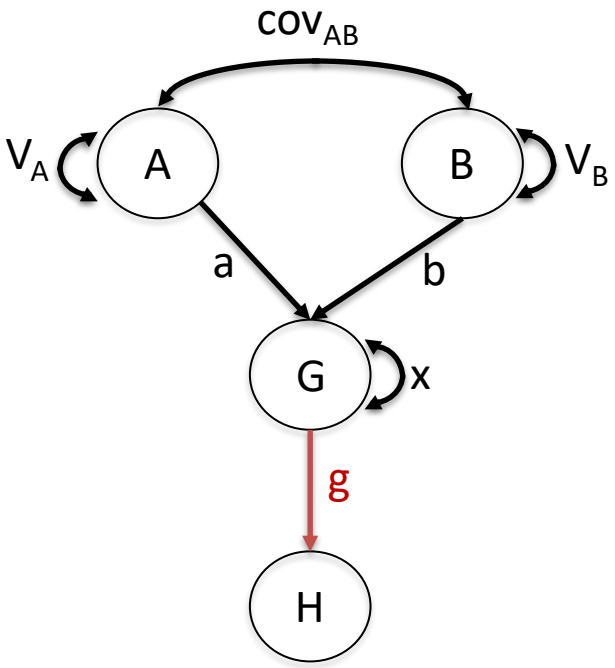
Path Tracing Example

COV(H,A) =



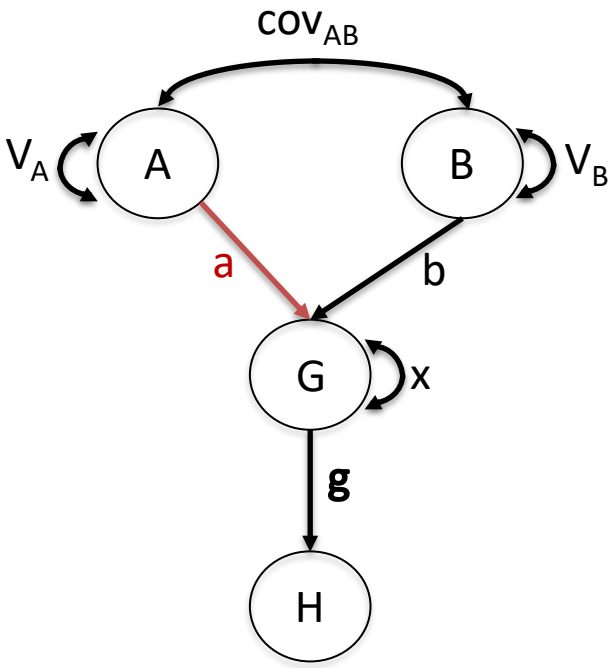
Path Tracing Example

COV(H,A) = **g**



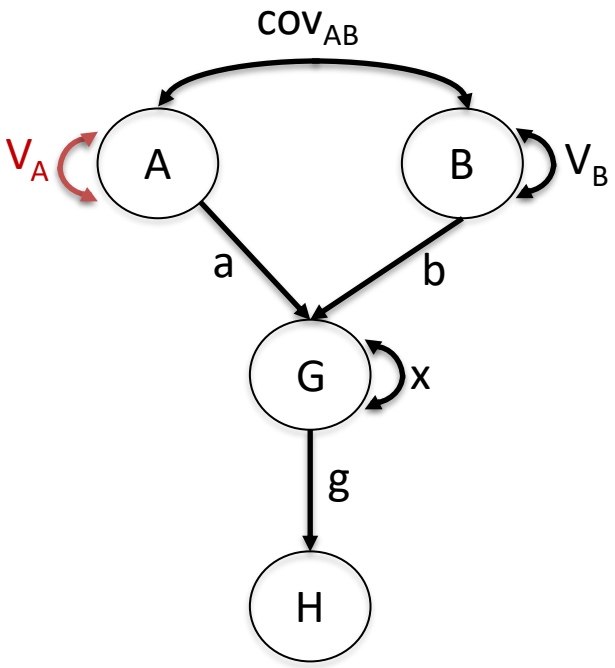
Path Tracing Example

COV(H,A) = **g** * **a**

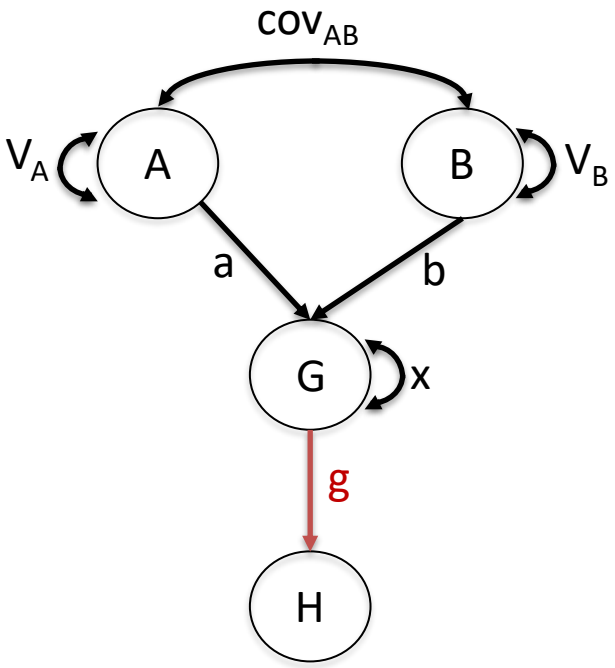


Path Tracing Example

COV(H,A) = **g * a *** V_A

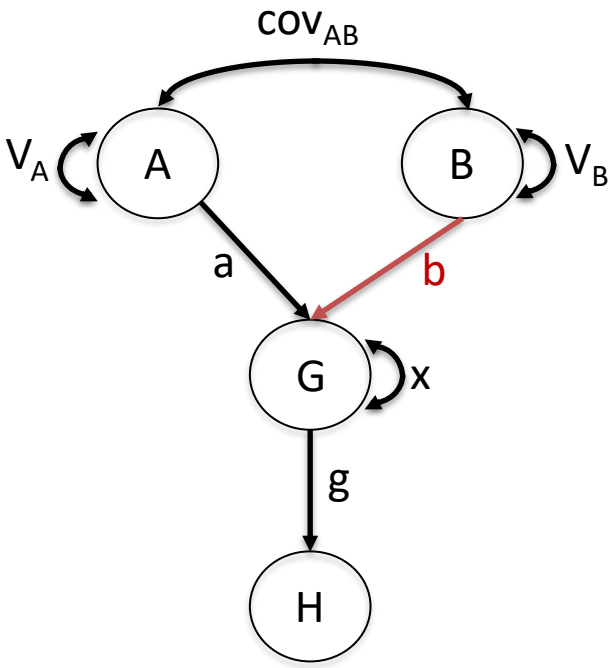


Path Tracing Example



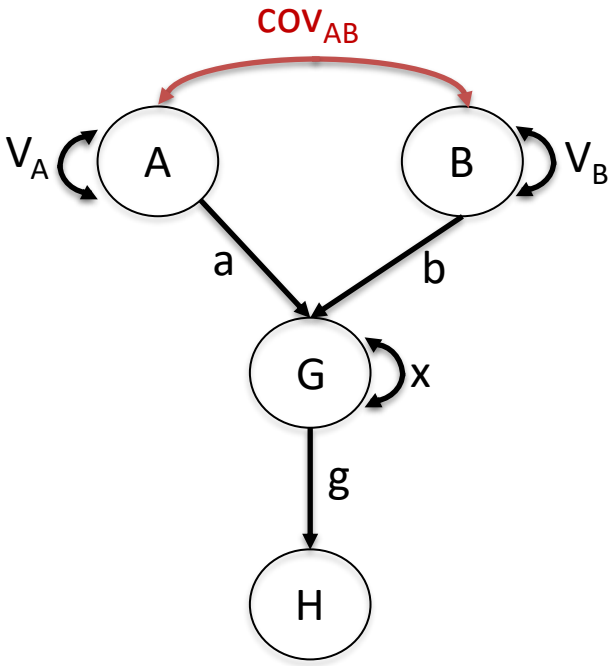
COV(H,A) = **g * a * V_A + g**

Path Tracing Example



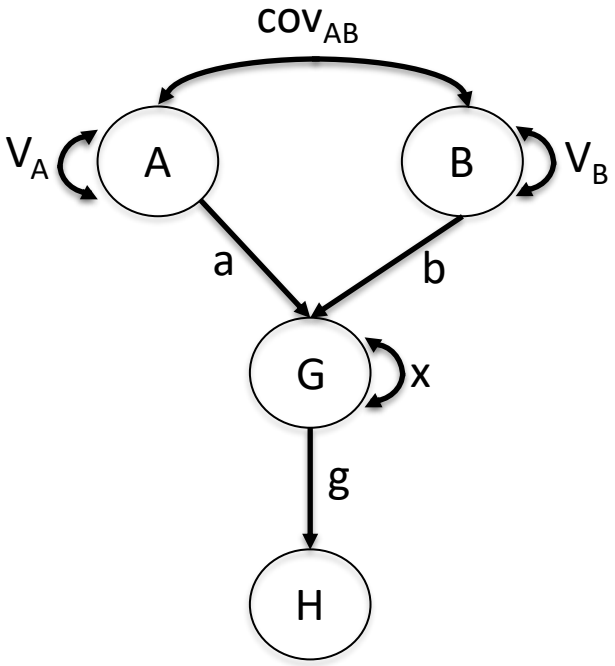
COV(H,A) = **g * a * V_A + g * b**

Path Tracing Example



COV(H,A) = **g * a * V_A + g * b * COV_{AB}**

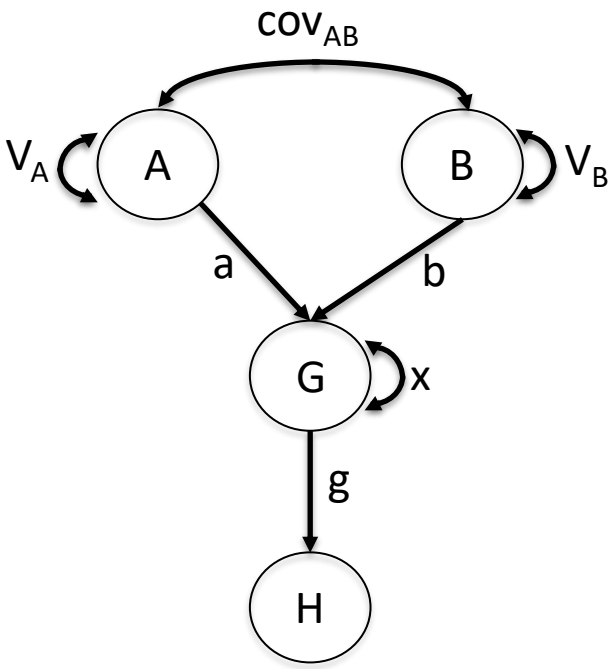
Path Tracing Example



COV(H,A) = **g * a * V_A + g * b * COV_{AB}**

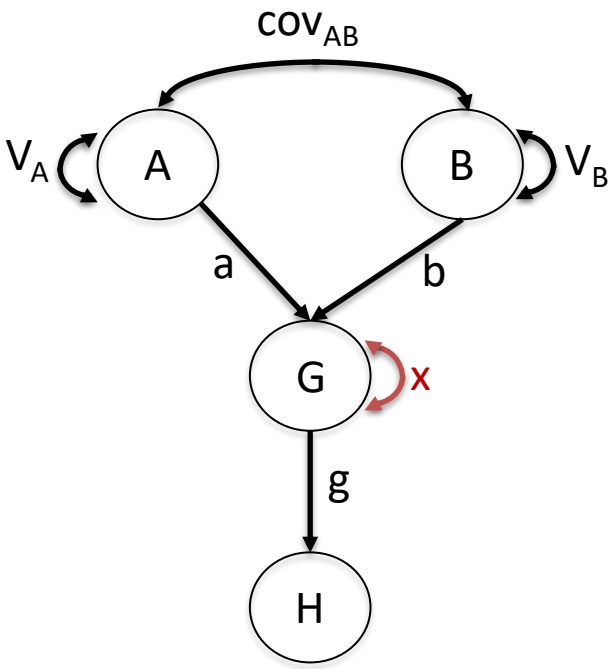
Path Tracing Example

VAR(G) =



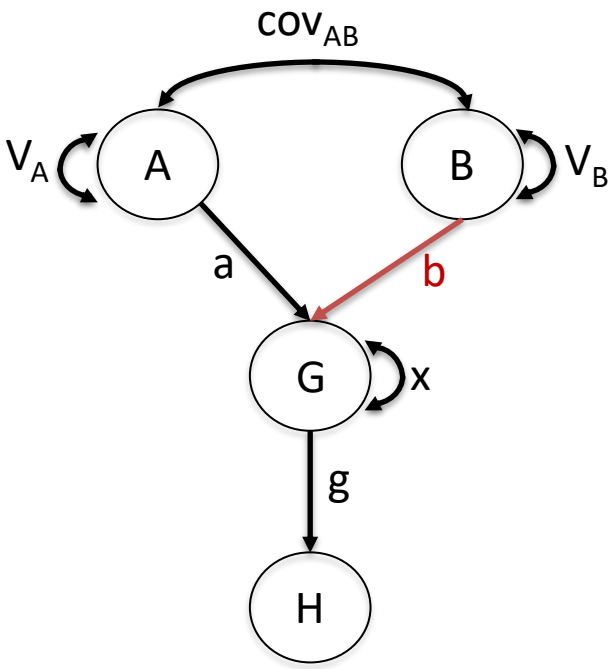
Path Tracing Example

VAR(G) = **x**



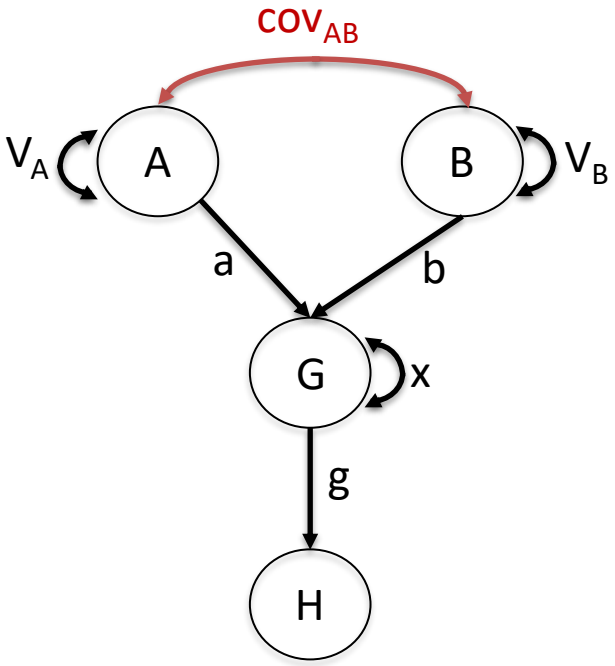
Path Tracing Example

VAR(G) = x + b



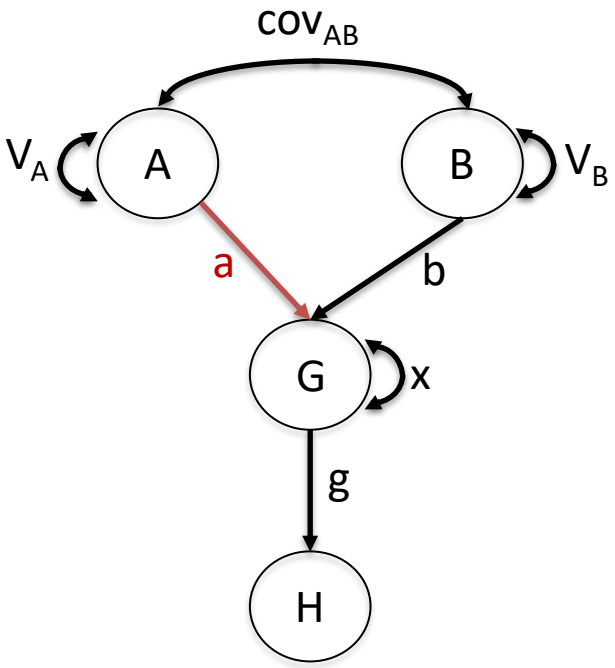
Path Tracing Example

$$VAR(G) = x + b * COV_{AB}$$



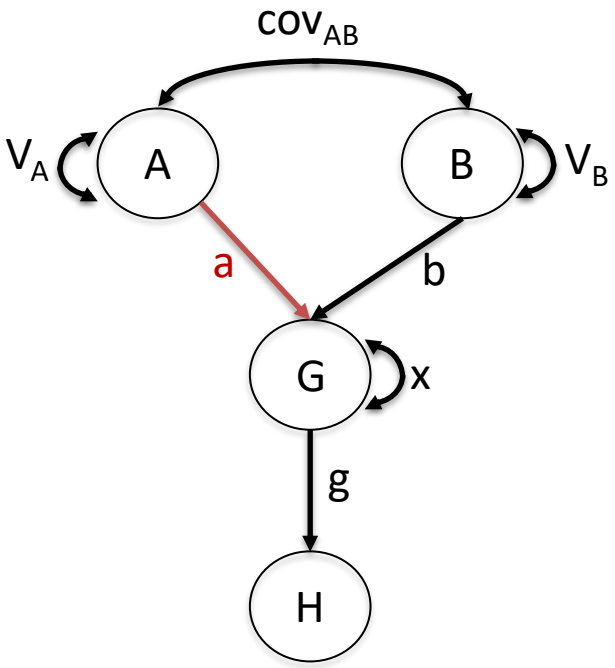
Path Tracing Example

$VAR(G) = x + b * COV_{AB} * a$



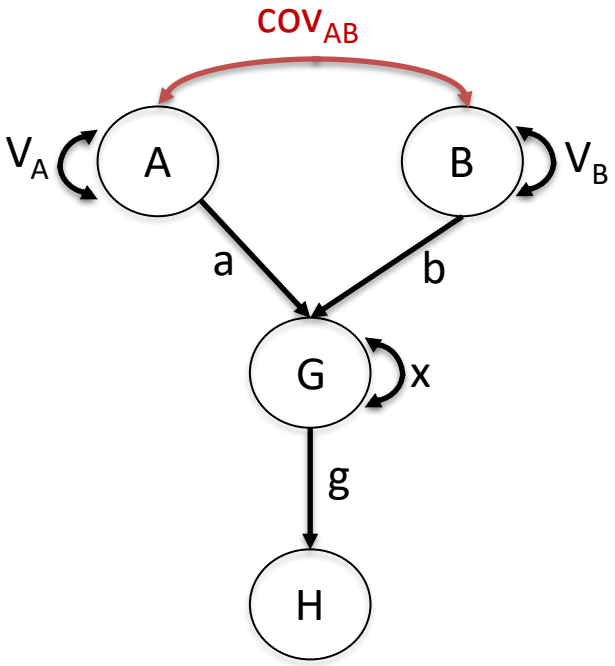
Path Tracing Example

$VAR(G) = x + b * COV_{AB} * a + a$



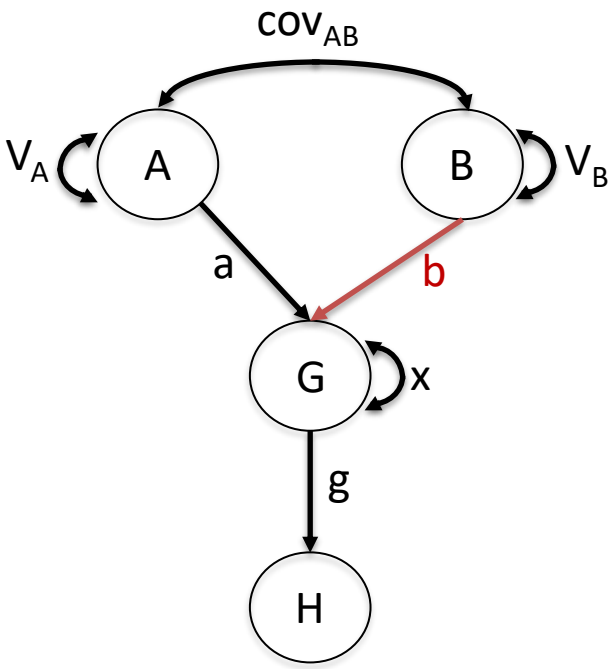
Path Tracing Example

$VAR(G) = x + b * COV_{AB} * a + a * \mathbf{COV_{AB}}$



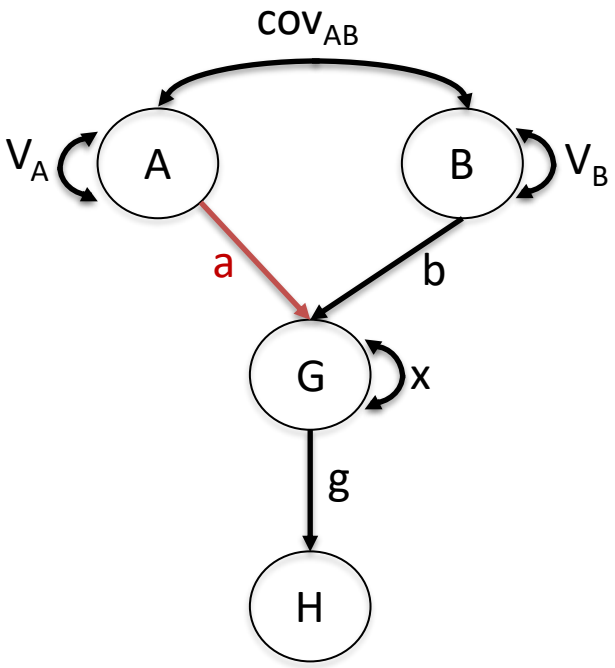
Path Tracing Example

$VAR(G) = x + b * COV_{AB} * a + a * COV_{AB} * b$



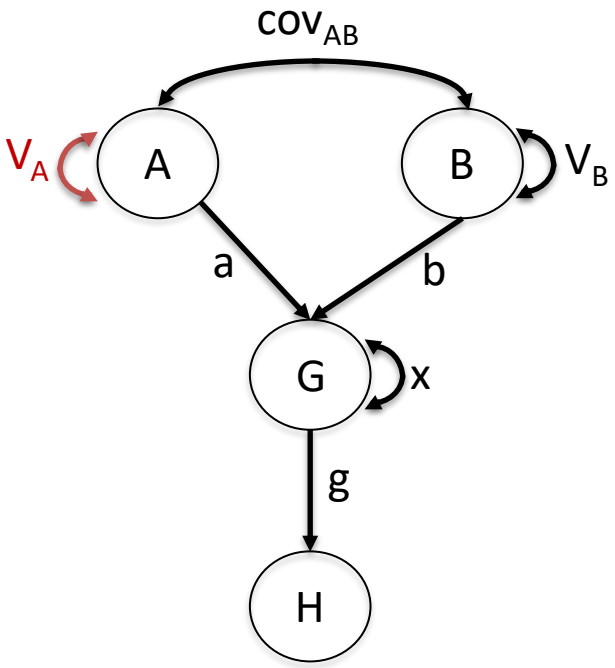
Path Tracing Example

$VAR(G) = x + b * COV_{AB} * a + a * COV_{AB} * b + a$



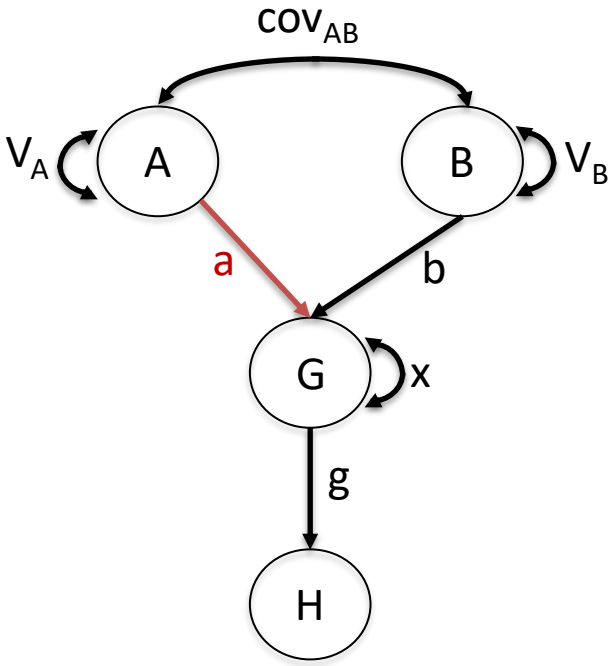
Path Tracing Example

$VAR(G) = x + b * COV_{AB} * a + a * COV_{AB} * b + a * V_A$



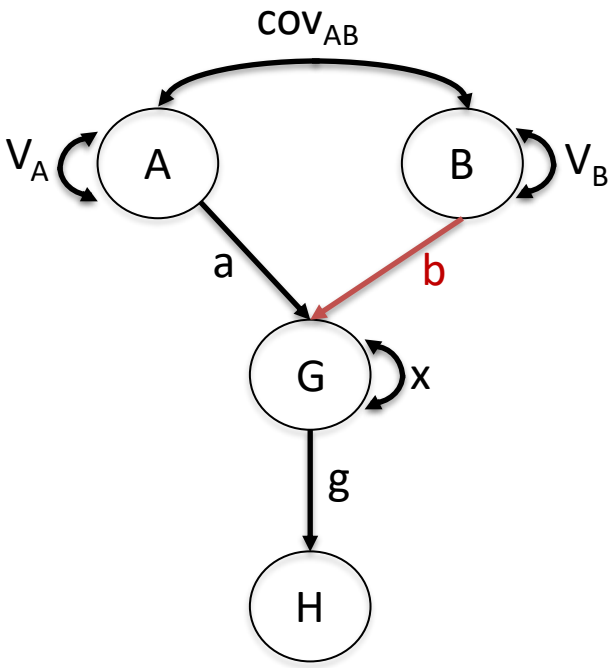
Path Tracing Example

$VAR(G) = x + b * COV_{AB} * a + a * COV_{AB} * b + a * V_A * a$



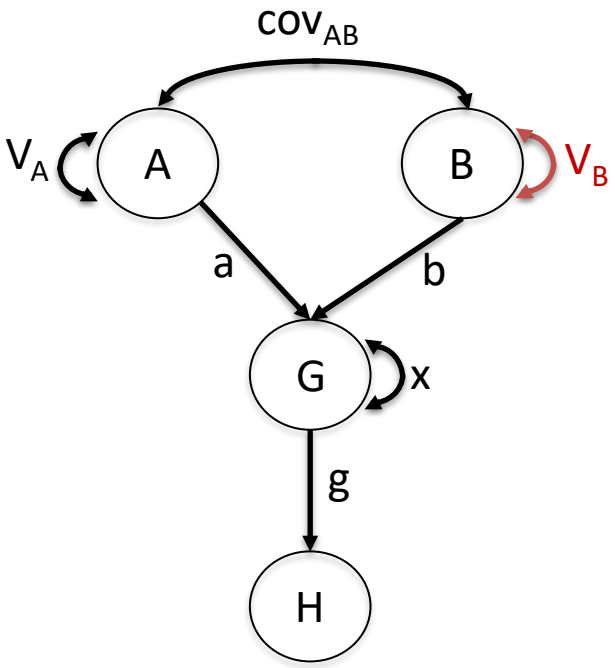
Path Tracing Example

$VAR(G) = x + b * COV_{AB} * a + a * COV_{AB} * b + a * V_A * a + b$



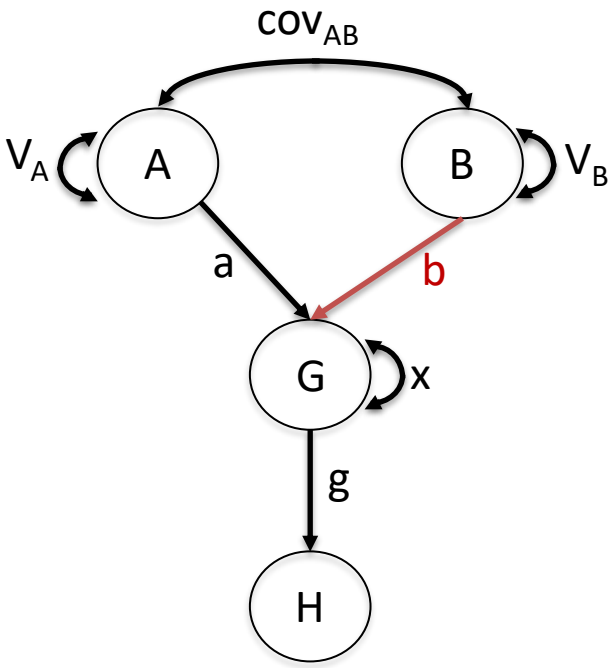
Path Tracing Example

$VAR(G) = x + b * COV_{AB} * a + a * COV_{AB} * b + a * V_A * a + b * V_B$



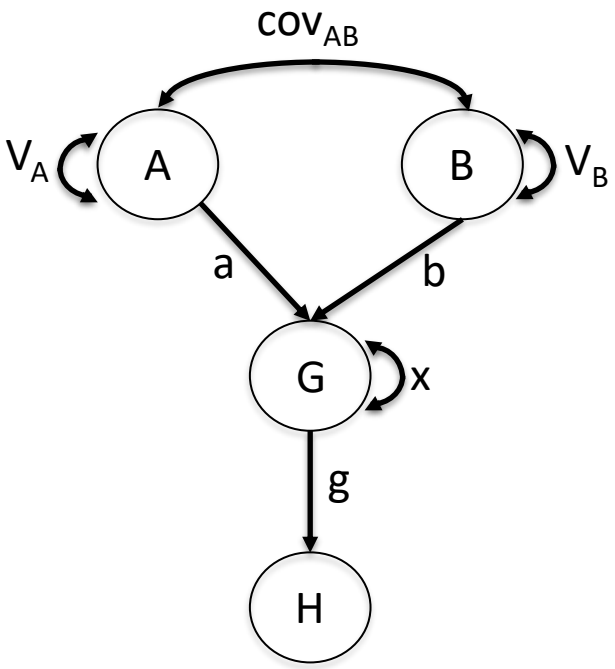
Path Tracing Example

$$VAR(G) = x + b * COV_{AB} * a + a * COV_{AB} * b + a * V_A * a + b * V_B * b$$



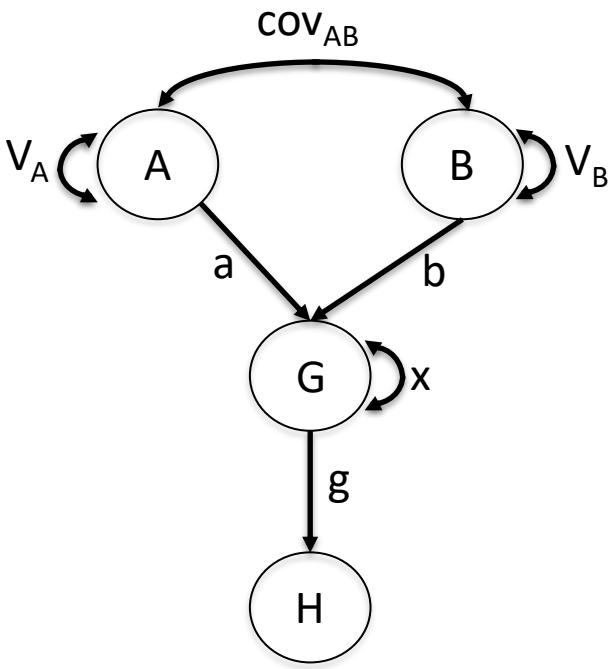
Path Tracing Example

$VAR(G) = x + 2*a*b*COV_{AB} + a^2*V_A + b^2*V_B$



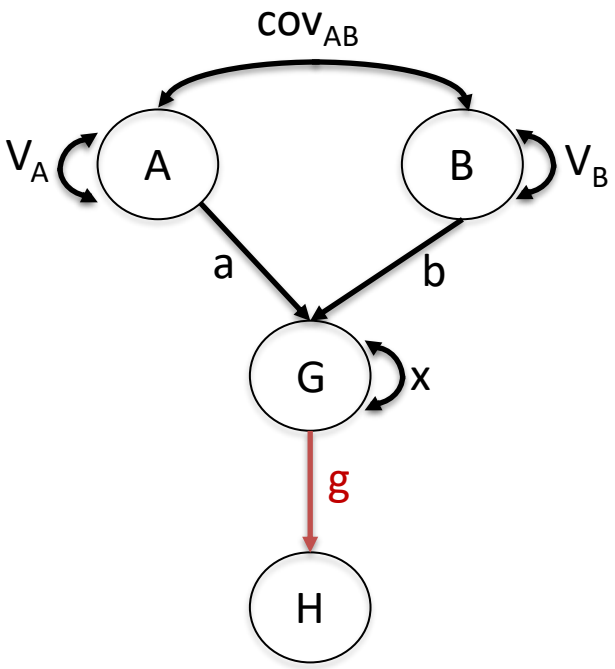
Path Tracing Example

VAR(H) =



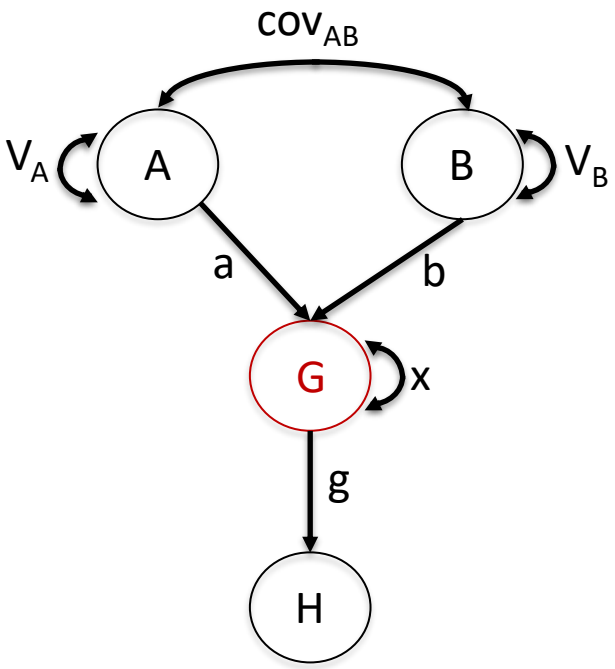
Path Tracing Example

VAR(H) = **g**



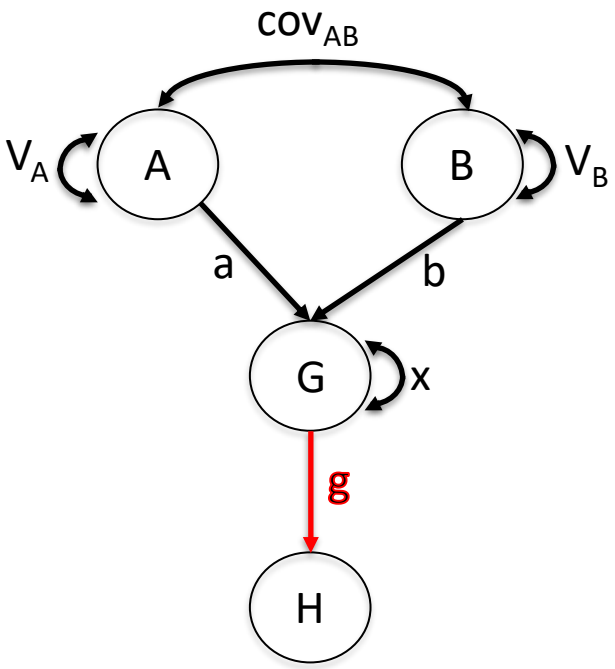
Path Tracing Example

VAR(H) = **g** * var(G)



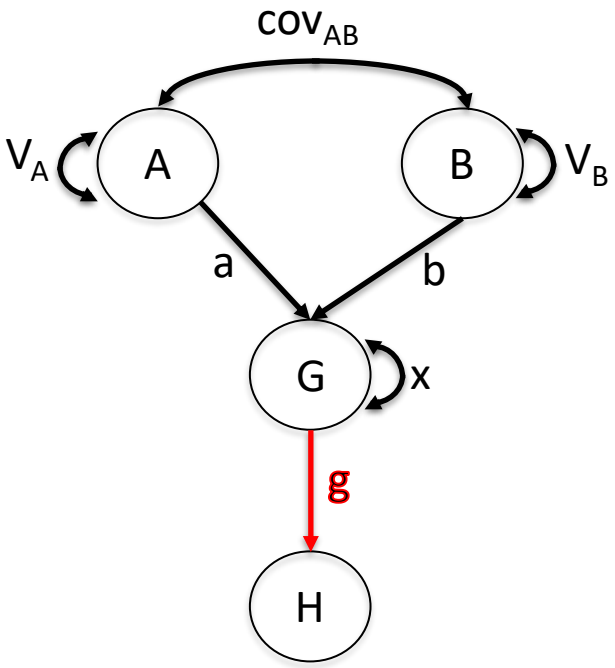
Path Tracing Example

VAR(H) = **g * var(G) * g**



Path Tracing Example

VAR(H) = g^2 * **var(G)**

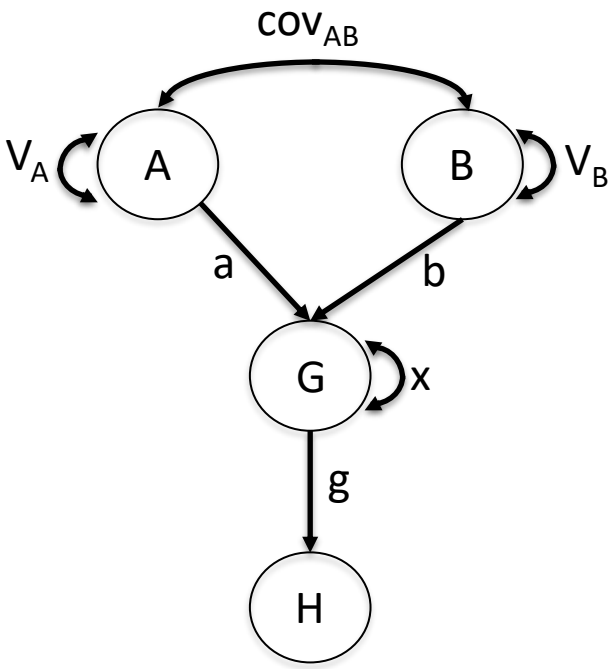


Deriving Expected Variances and Covariances Using Covariance Algebra

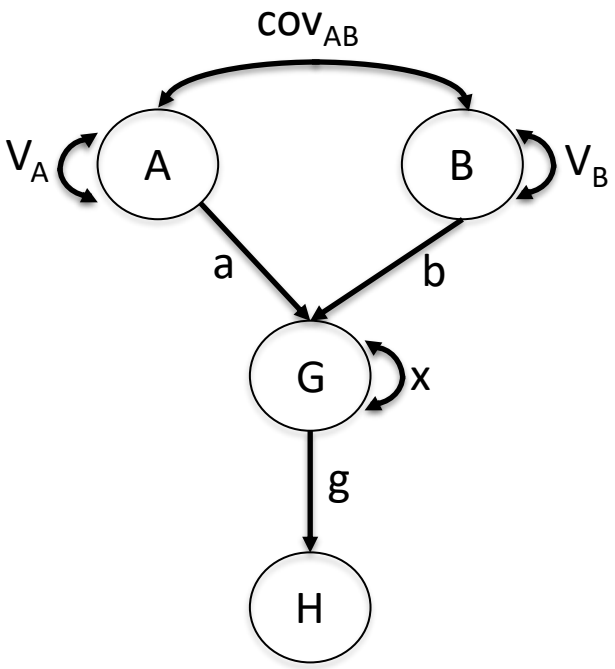
Rules of Covariance Algebra

- $COV(c, X) = 0$
- $COV(cX_1, X_2) = cCOV(X_1, X_2)$
- $COV(X_1 + X_2, X_3) = COV(X_1, X_3) + COV(X_2, X_3)$
- $VAR(X_1) = COV(X_1, X_1)$

Covariance Algebra Example



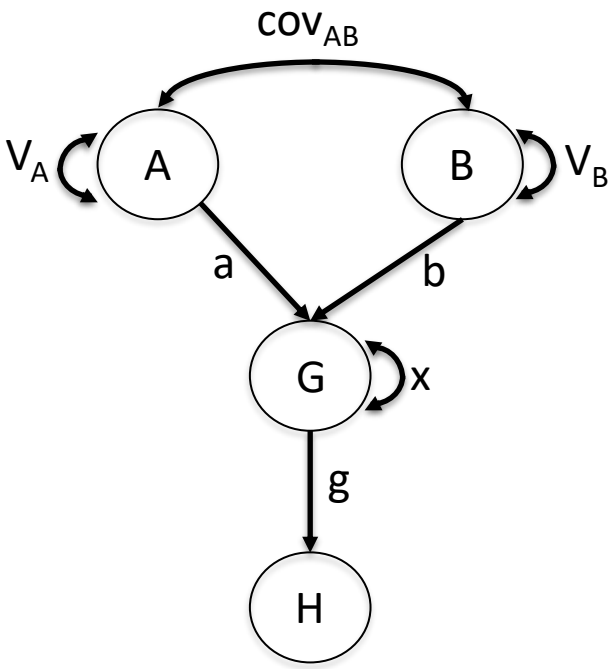
Covariance Algebra Example



$H = g*G$

$G = a*A + b*B + e_X$

Covariance Algebra Example

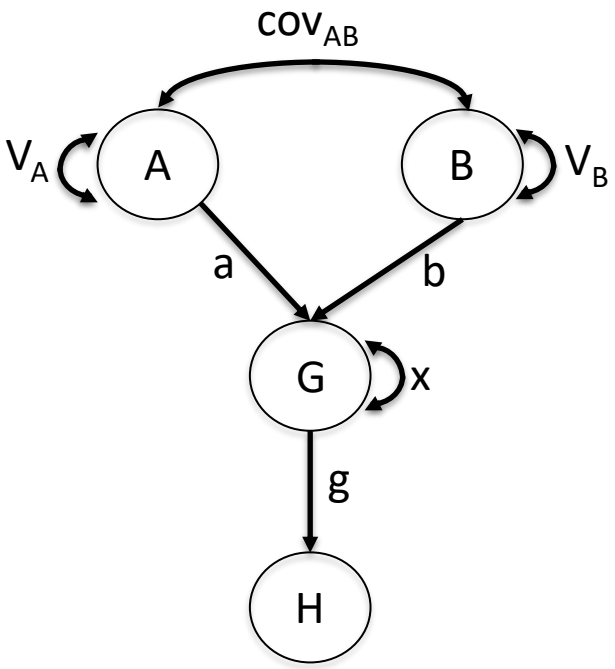


H = g*G

G = a*A + b*B + e_X

COV(H,A) = ?

Covariance Algebra Example

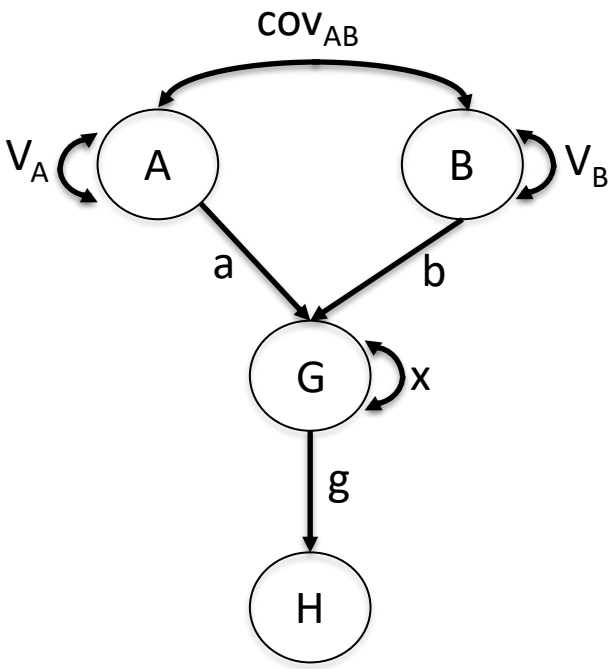


H = g*G

G = a*A + b*B + e_X

COV(H,A) = COV(g*G, A)

Covariance Algebra Example



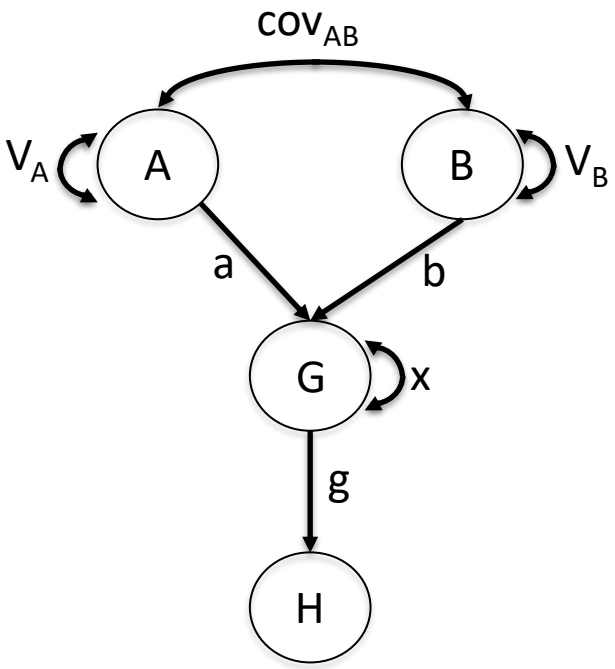
$H = g*G$

$G = a*A + b*B + e_X$

$COV(H,A) = COV(g*G, A)$

$= COV(g*(a*A + b*B + e_X), A)$

Covariance Algebra Example



$H = g*G$

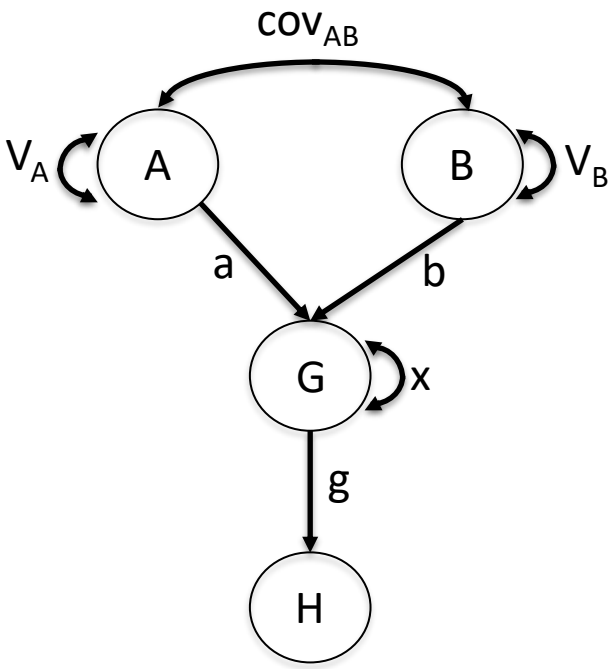
$G = a*A + b*B + e_X$

$COV(H,A) = COV(g*G, A)$

$= COV(g*(a*A + b*B + e_X), A)$

$= COV(g*a*A + g*b*B + g*e_X), A)$

Covariance Algebra Example



$H = g*G$

$G = a*A + b*B + e_X$

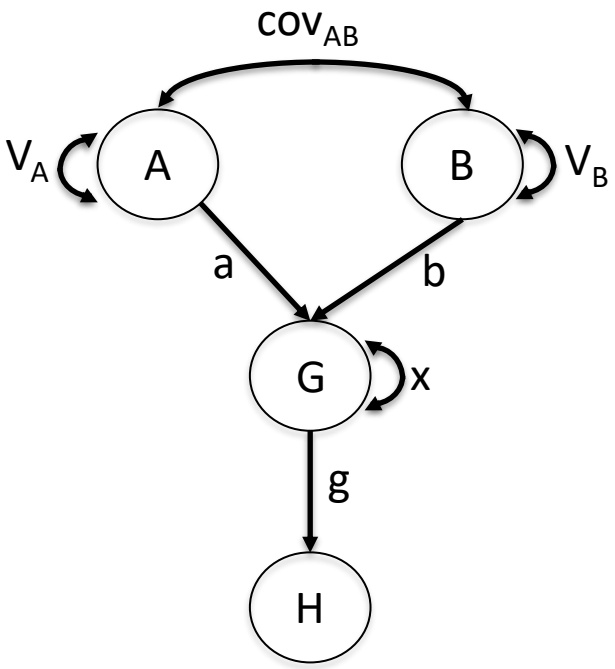
$COV(H,A) = COV(g*G, A)$

$= COV(g*(a*A + b*B + e_X), A)$

$= COV(g*a*A + g*b*B + g*e_X), A)$

$= COV(g*a*A, A) + COV(g*b*B, A) + COV(g*e_X, A)$

Covariance Algebra Example



$H = g*G$

$G = a*A + b*B + e_X$

$COV(H,A) = COV(g*G, A)$

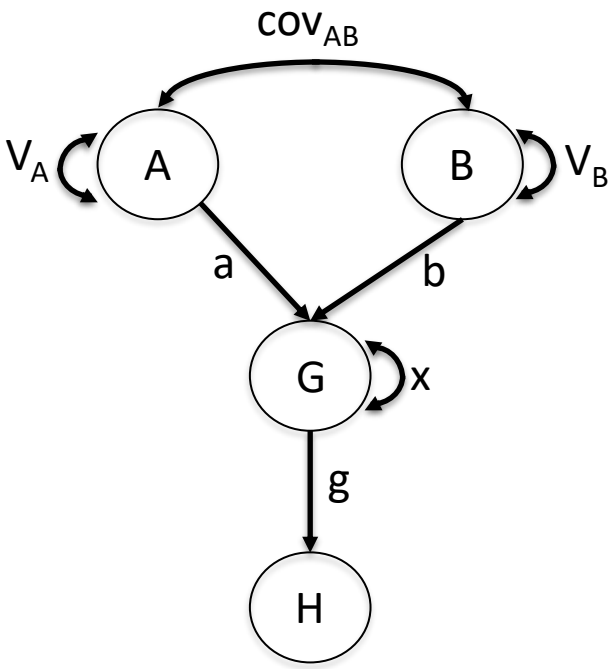
$= COV(g*(a*A + b*B + e_X), A)$

$= COV(g*a*A + g*b*B + g*e_X), A)$

$= COV(g*a*A, A) + COV(g*b*B, A) + COV(g*e_X, A)$

$= g*a*COV(A, A) + g*b*COV(B, A) + g*COV(e_X, A)$

Covariance Algebra Example



$H = g*G$

$G = a*A + b*B + e_X$

$COV(H,A) = COV(g*G, A)$

$= COV(g*(a*A + b*B + e_X), A)$

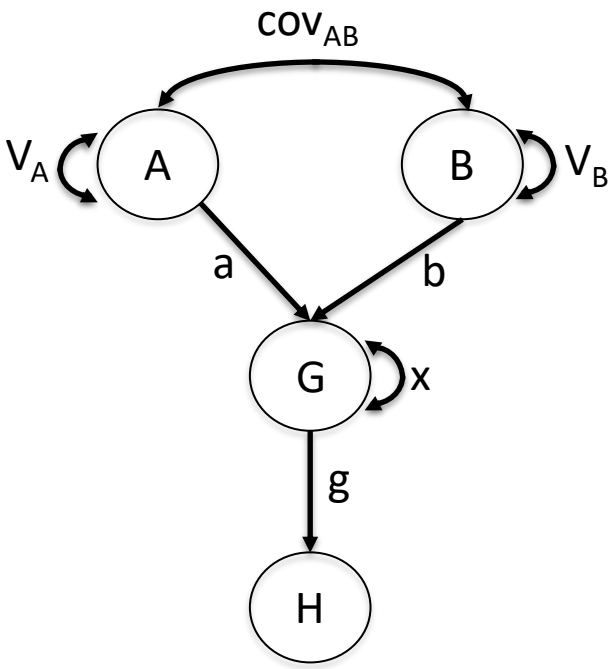
$= COV(g*a*A + g*b*B + g*e_X), A)$

$= COV(g*a*A, A) + COV(g*b*B, A) + COV(g*e_X, A)$

$= g*a*COV(A, A) + g*b*COV(B, A) + g*COV(e_X, A)$

$= g*a*VAR(A) + g*b*COV(B, A)$

Covariance Algebra Example



$H = g*G$

$G = a*A + b*B + e_X$

$COV(H,A) = COV(g*G, A)$

$= COV(g*(a*A + b*B + e_X), A)$

$= COV(g*a*A + g*b*B + g*e_X), A)$

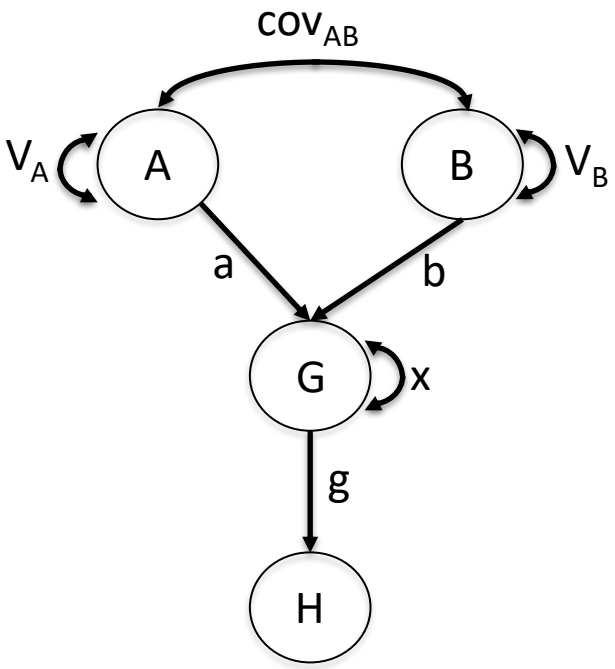
$= COV(g*a*A, A) + COV(g*b*B, A) + COV(g*e_X, A)$

$= g*a*COV(A, A) + g*b*COV(B, A) + g*COV(e_X, A)$

$= g*a*VAR(A) + g*b*COV(B, A)$

$= g*a*V_A + g*b*COV(B, A)$

Covariance Algebra Example



$H = g*G$

$G = a*A + b*B + e_X$

$COV(H,A) = COV(g*G, A)$

$= COV(g*(a*A + b*B + e_X), A)$

$= COV(g*a*A + g*b*B + g*e_X), A)$

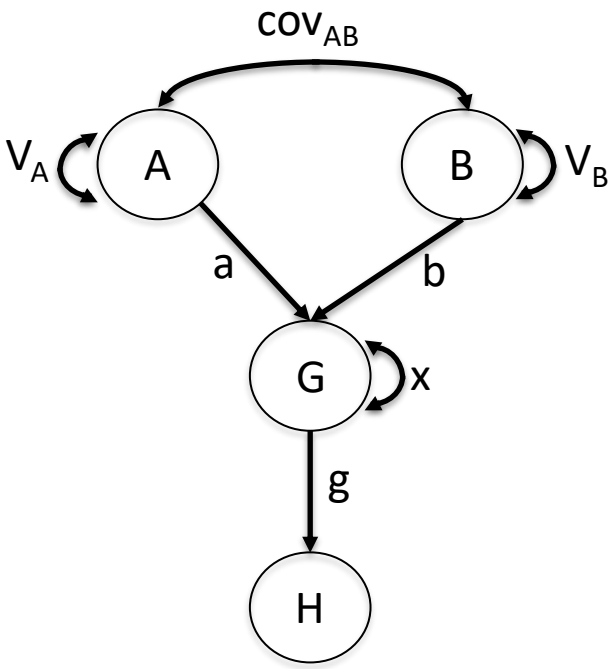
$= COV(g*a*A, A) + COV(g*b*B, A) + COV(g*e_X, A)$

$= g*a*COV(A, A) + g*b*COV(B, A) + g*COV(e_X, A)$

$= g*a*VAR(A) + g*b*COV(B, A)$

$= g*a*V_A + g*b*COV_{AB}$

Covariance Algebra Example

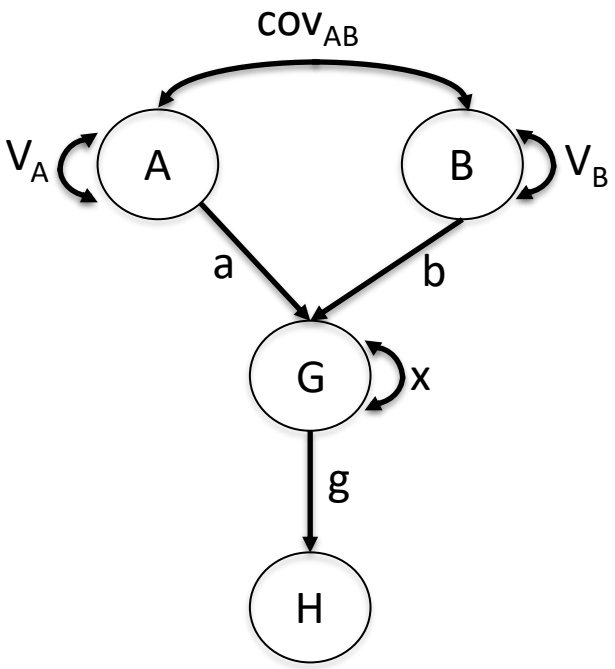


H = g*G

G = a*A + b*B

VAR(G) = ?

Covariance Algebra Example

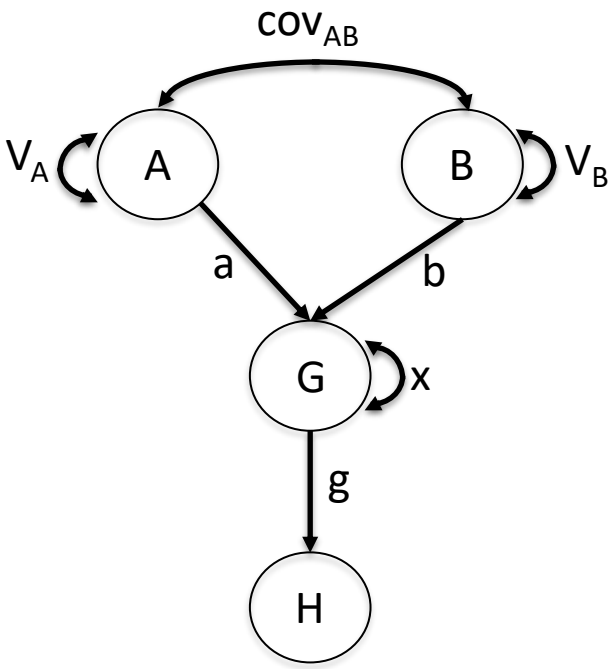


H = g*G

G = a*A + b*B

VAR(G) = COV(G, G)

Covariance Algebra Example



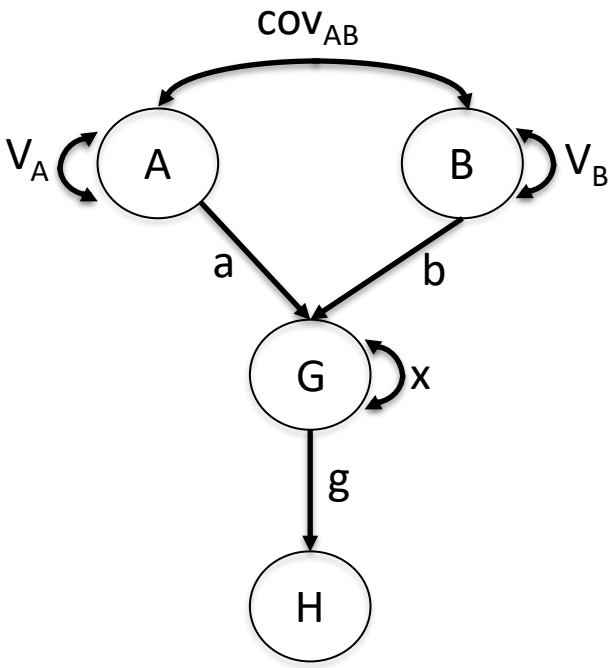
H = g*G

G = a*A + b*B

VAR(G) = COV(G, G)

= COV(a*A + b*B + e, a*A + b*B + e)

Covariance Algebra Example



H = g*G

G = a*A + b*B

VAR(G) = COV(G, G)

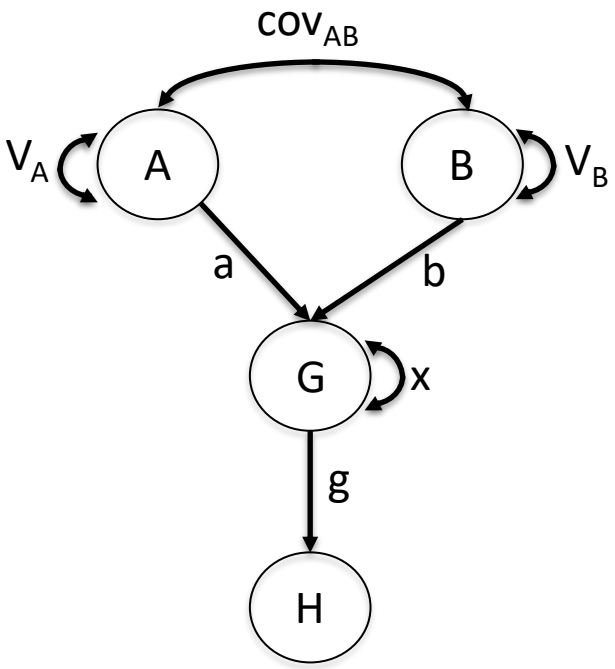
= COV(a*A + b*B + e, a*A + b*B + e)

= COV(a*A, a*A) + COV(a*A, b*B) + COV(a*A, e)

+ COV(b*B, a*A) + COV(b*B, b*B) + COV(b*B, e)

+ COV(e, a*A) + COV(e, b*B) + COV(e, e)

Covariance Algebra Example



H = g*G

G = a*A + b*B

VAR(G) = COV(G, G)

= COV(a*A + b*B + e, a*A + b*B + e)

= COV(a*A, a*A) + COV(a*A, b*B) + COV(a*A, e)

+ COV(b*B, a*A) + COV(b*B, b*B) + COV(b*B, e)

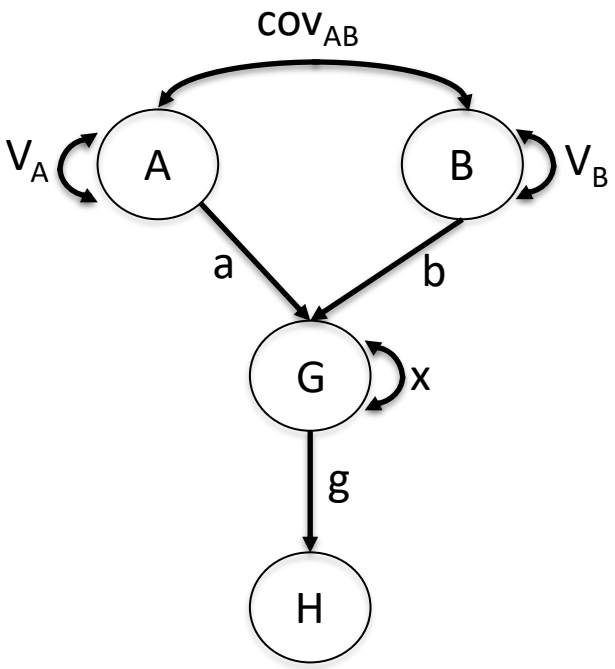
+ COV(e, a*A) + COV(e, b*B) + COV(e, e)

= a*a*COV(A, A) + a*b*COV(A, B)

+ b*a*COV(B, A) + b*b*COV(A, B)

+ COV(e, e)

Covariance Algebra Example



H = g*G

G = a*A + b*B

VAR(G) = COV(G, G)

= COV(a*A + b*B + e, a*A + b*B + e)

= COV(a*A, a*A) + COV(a*A, b*B) + COV(a*A, e)

+ COV(b*B, a*A) + COV(b*B, b*B) + COV(b*B, e)

+ COV(e, a*A) + COV(e, b*B) + COV(e, e)

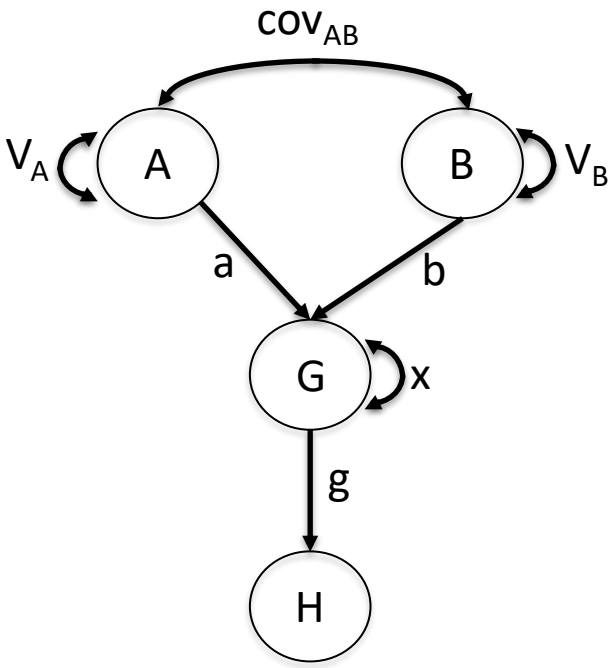
= a*a*COV(A, A) + a*b*COV(A, B)

+ b*a*COV(B, A) + b*b*COV(A, B)

+ COV(e, e)

= $a^2*V_A + b^2*V_B + 2*a*b*COV_{AB} + x$

Covariance Algebra Example



H = g*G

G = a*A + b*B

VAR(G) = COV(G, G)

= COV(a*A + b*B + e, a*A + b*B + e)

= COV(a*A, a*A) + COV(a*A, b*B) + COV(a*A, e)

+ COV(b*B, a*A) + COV(b*B, b*B) + COV(b*B, e)

+ COV(e, a*A) + COV(e, b*B) + COV(e, e)

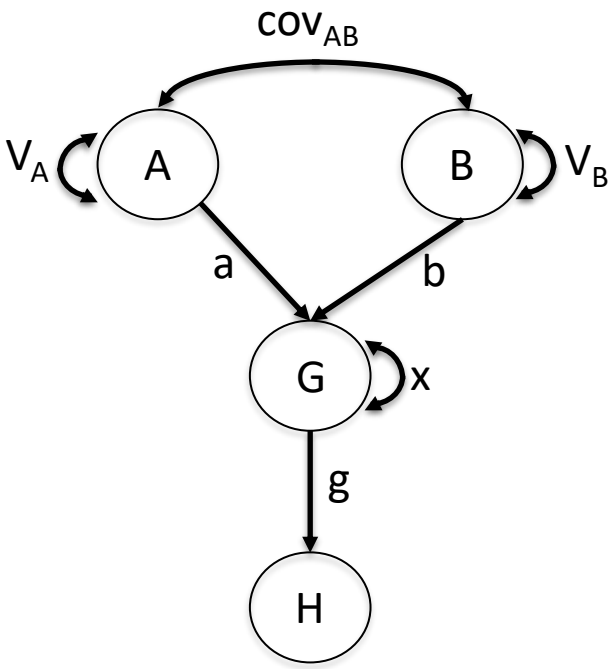
= a*a*COV(A, A) + a*b*COV(A, B)

+ b*a*COV(B, A) + b*b*COV(A, B)

+ COV(e, e)

$= a^2*V_A + b^2*V_B + 2*a*b*COV_{AB} + x$

Covariance Algebra Example

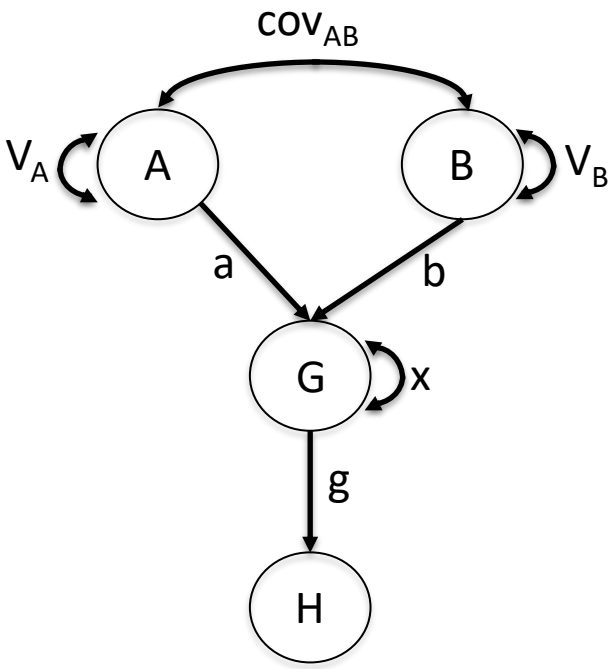


H = g*G

G = a*A + b*B

VAR(H) = ?

Covariance Algebra Example

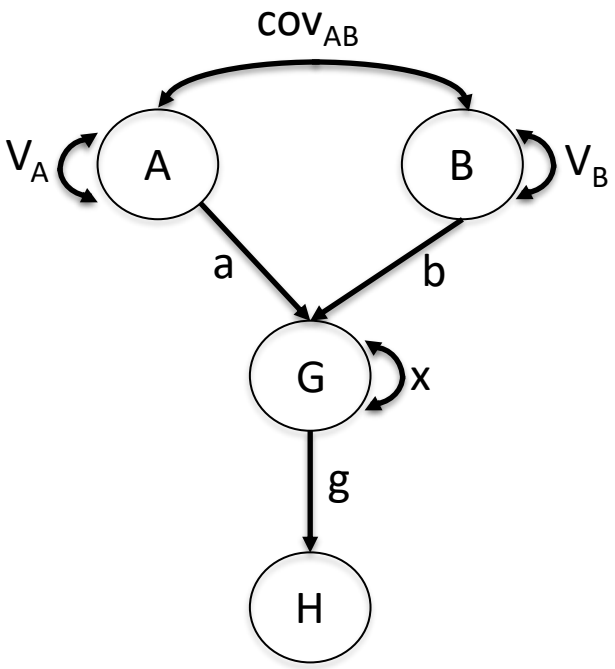


H = g*G

G = a*A + b*B

VAR(H) = COV(H, H)

Covariance Algebra Example



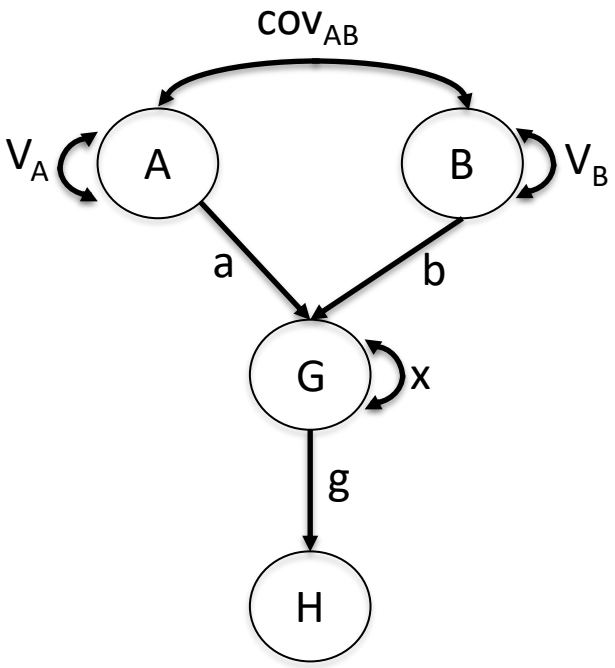
H = g*G

G = a*A + b*B

VAR(H) = COV(H, H)

= COV(g*G, g*G)

Covariance Algebra Example



H = g*G

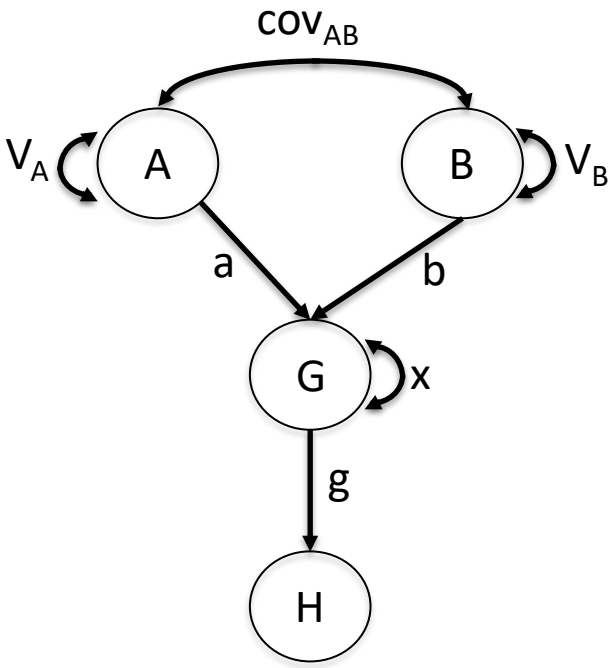
G = a*A + b*B

VAR(H) = COV(H, H)

= COV(g*G, g*G)

= g*g*COV(G, G)

Covariance Algebra Example



H = g*G

G = a*A + b*B

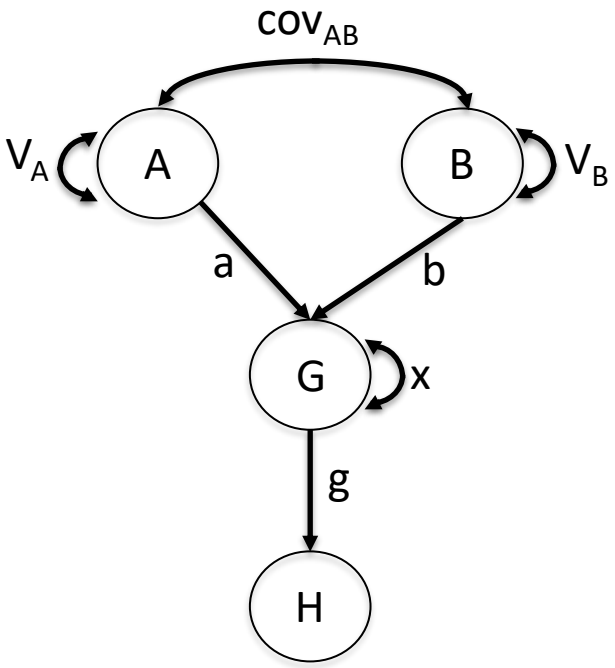
VAR(H) = COV(H, H)

= COV(g*G, g*G)

= g*g*COV(G, G)

= g^2*VAR(G)

Covariance Algebra Example



H = g*G

G = a*A + b*B

VAR(H) = COV(H, H)

= COV(g*G, g*G)

= g*g*COV(G, G)

= g^2*VAR(G)

Further Reading

- Evans DM. et al (2002). Biometrical Genetics. *Biol Psychol*, 61, 33-51.
- Bollen K. (1989). Structural equations with latent variables.
- Neale M. & Cardon L. (1992). Methodology for genetic studies of twins and families.
- Rijsdijk F.V. & Sham P.C. (2002). Analytic approaches to twin data using structural equation models. *Brief Bioinform*, 3(2), 119-33.