Introduction to Structural Equation Modelling (SEM)

Geng Wang

Research Fellow | Evans Group
Institute for Molecular Bioscience
The University of Queensland

9-July-2026

Acknowledgement of Country

The University of Queensland (UQ) acknowledges the Traditional Owners and their custodianship of the lands on which we meet.

We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.

We recognise their valuable contributions to Australian and global society.

Image: Digital reproduction of *A guidance through time* by Casey Coolwell and Kyra Mancktelow

Content

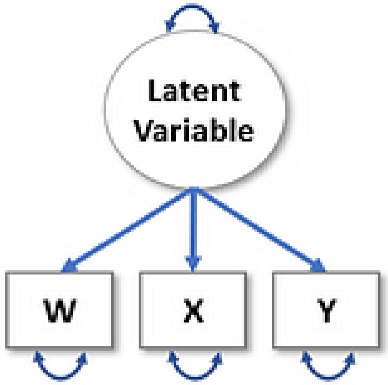
- **Part I (9:00-10:00 am)**
  - SEM basics
  - Path diagrams

Short break (5mins)

- **Part II (10:05 -10:30 am)**
  - Genomic SEM
  - Q&A (5-10mins)

Background - What is SEM?

- Structural Equation Modelling (SEM) is a statistical method for analysing 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 observed data, and obtaining estimates of parameters.
- The causal connections are represented using equations, but the postulated structuring can be illustrated by a path diagram.

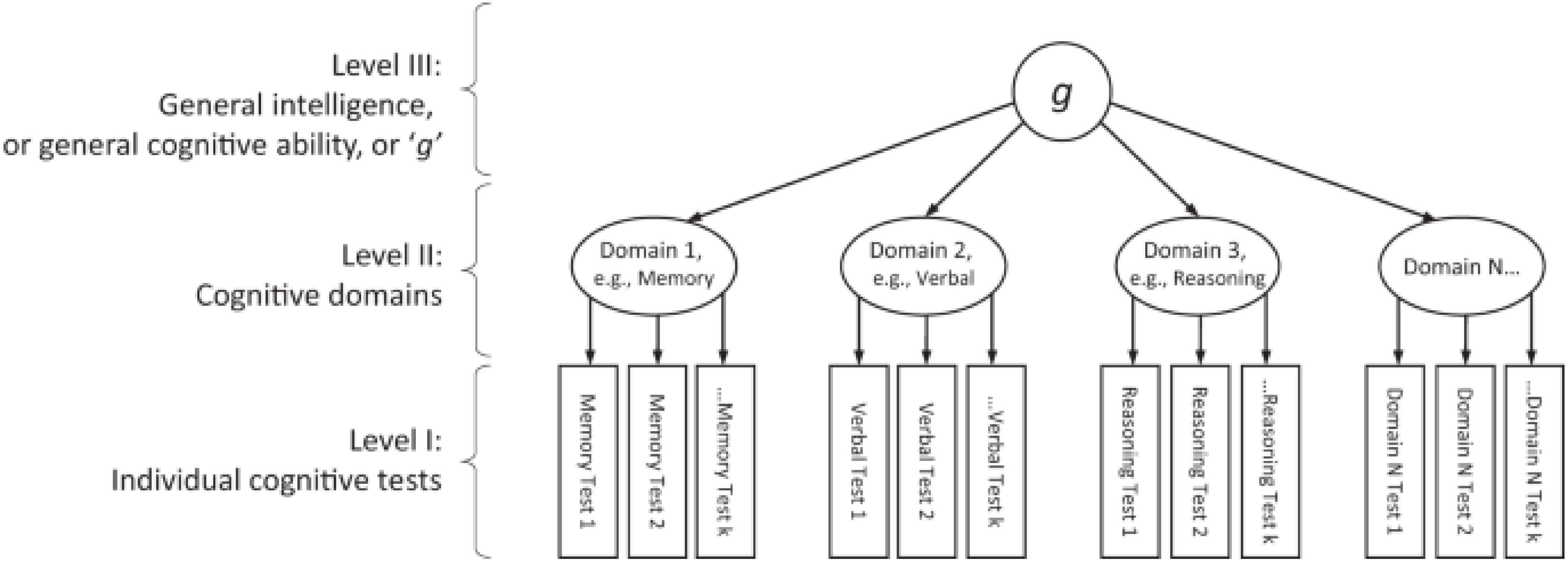


Also known as:

- **Confirmatory Factor Analysis**
- **Analysis of covariance structure**
- **Path analysis**

Background - What is latent variable?

- **For example, General Intelligence (*g*)…**

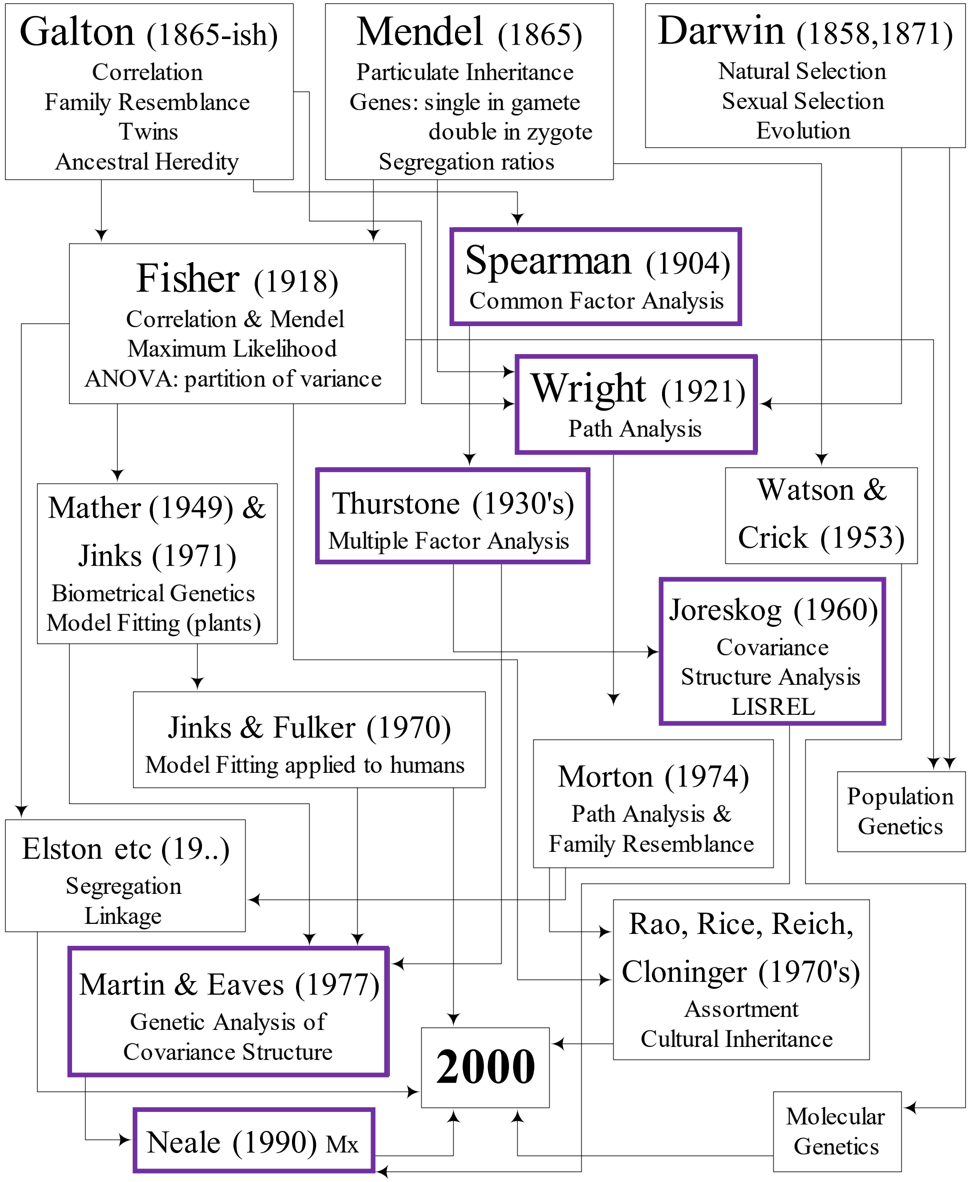


Spearman 1904
Deary et al. 2021

Background - Why SEM?

- **Flexibility - almost any linear model can be written as an SEM.**
- **Simplicity - SEM makes it easy to create new models/methods.**
- **Useful - super useful for deriving expected variances/covariances in genetics**
- **Versatility - SEM means that you can think about a problem in 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

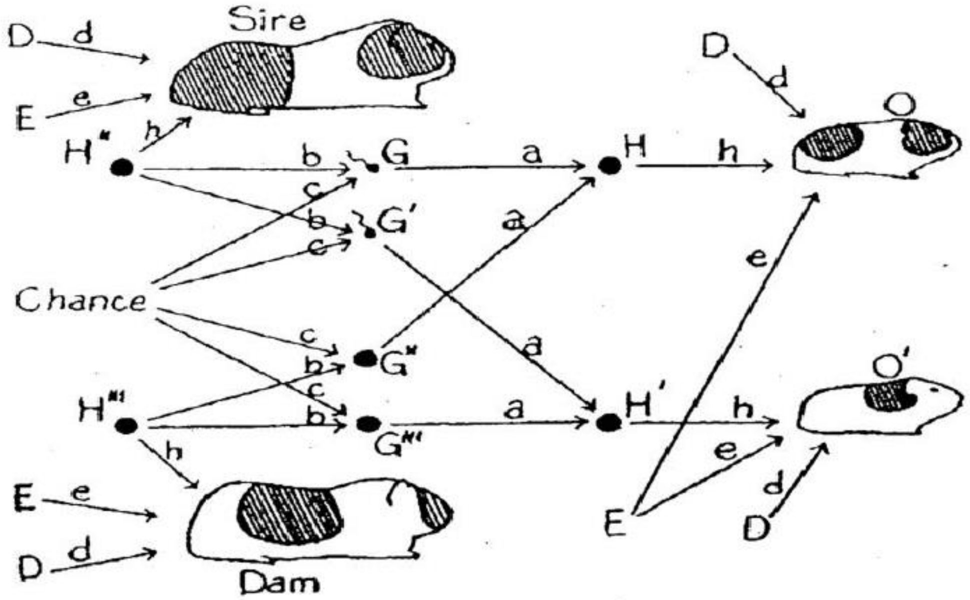
2000

Molecular
Genetics

Neale (1990) Mx

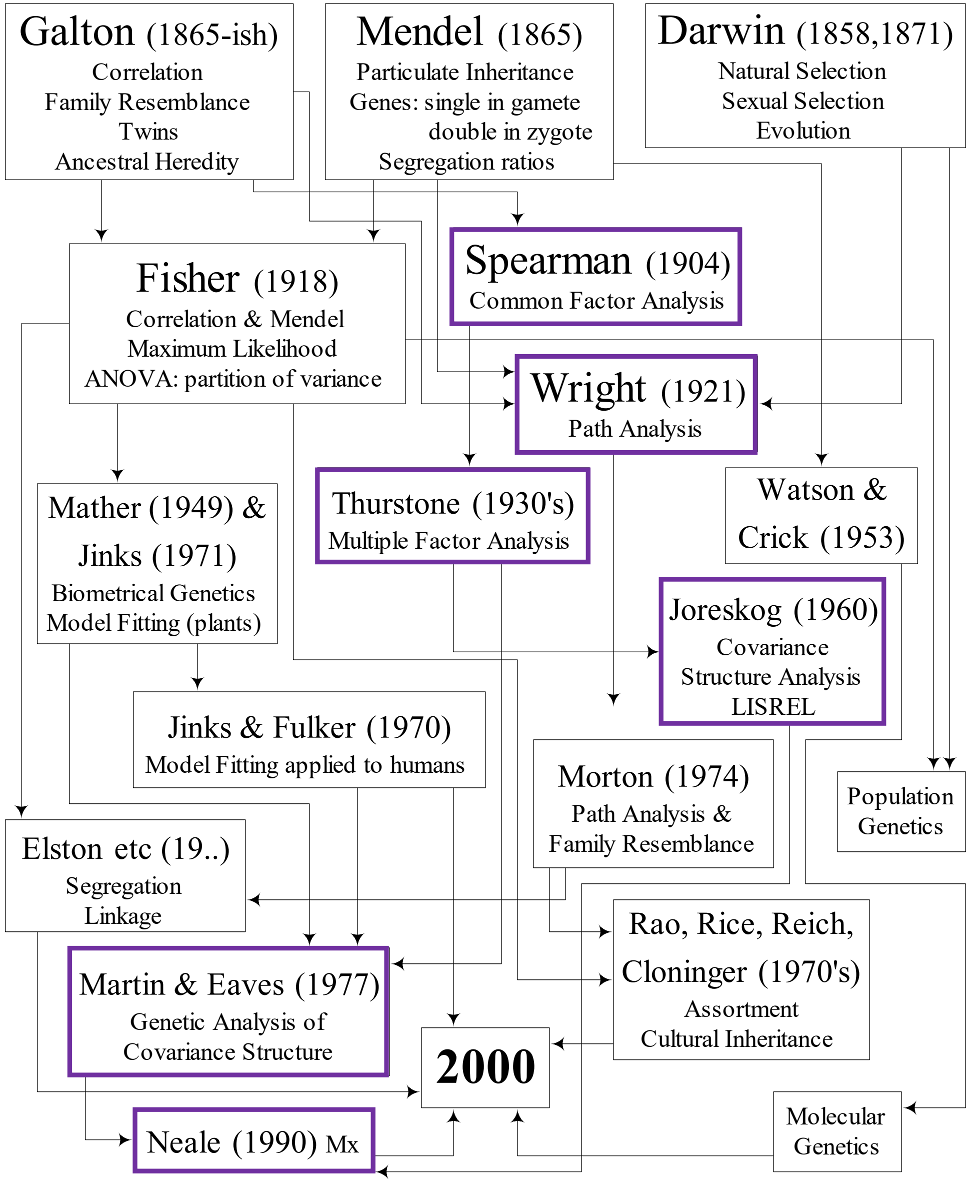
Neale & Cardon (1992)

Sewall Wright



Piebald patterns in guinea pigs

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

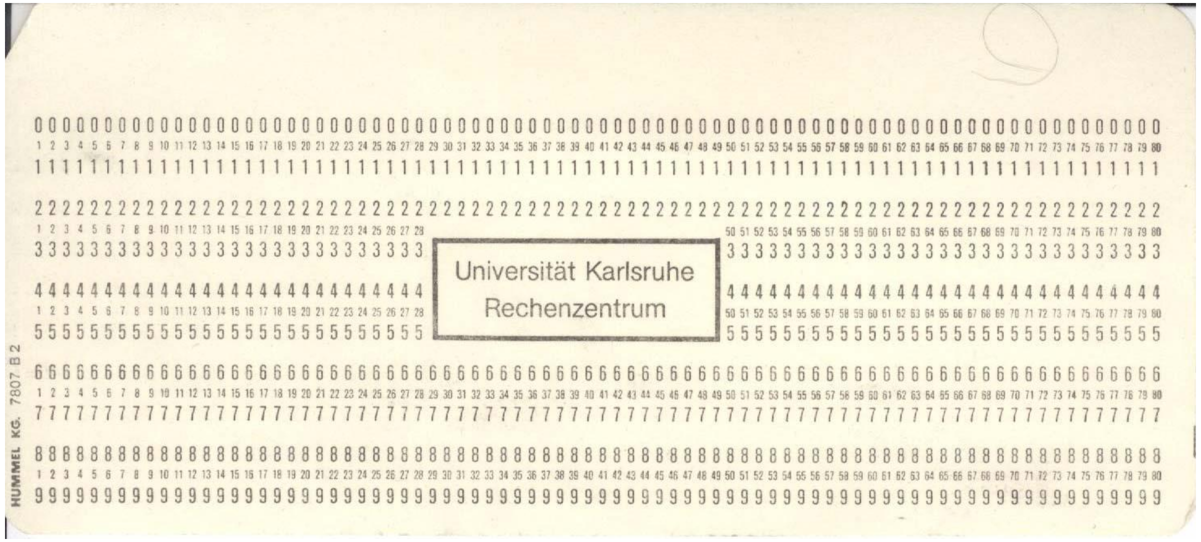
2000

Molecular
Genetics

Neale (1990) Mx

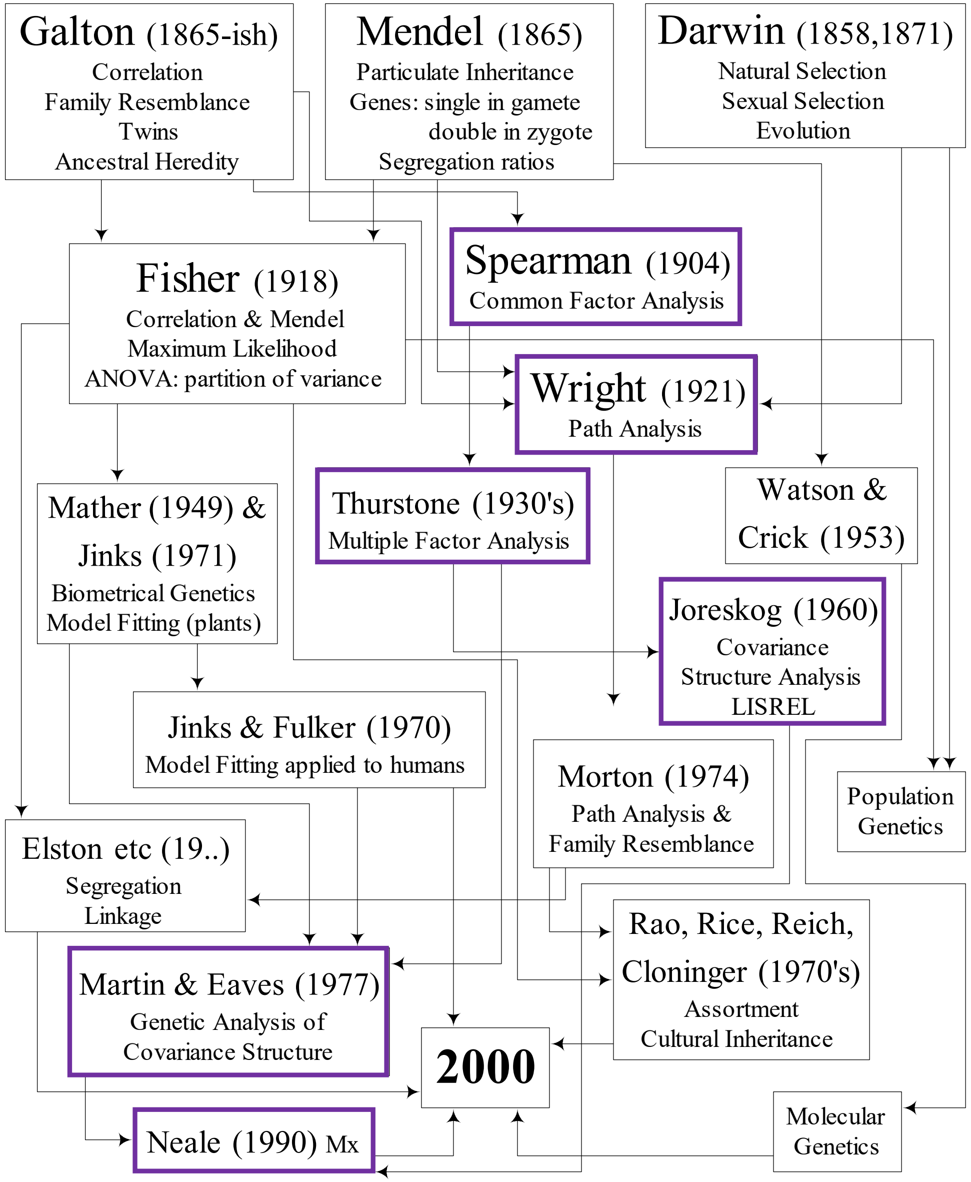
Neale & Cardon (1992)

Karl Jöreskog



LISREL program and computer

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

Molecular
Genetics

Neale (1990) Mx

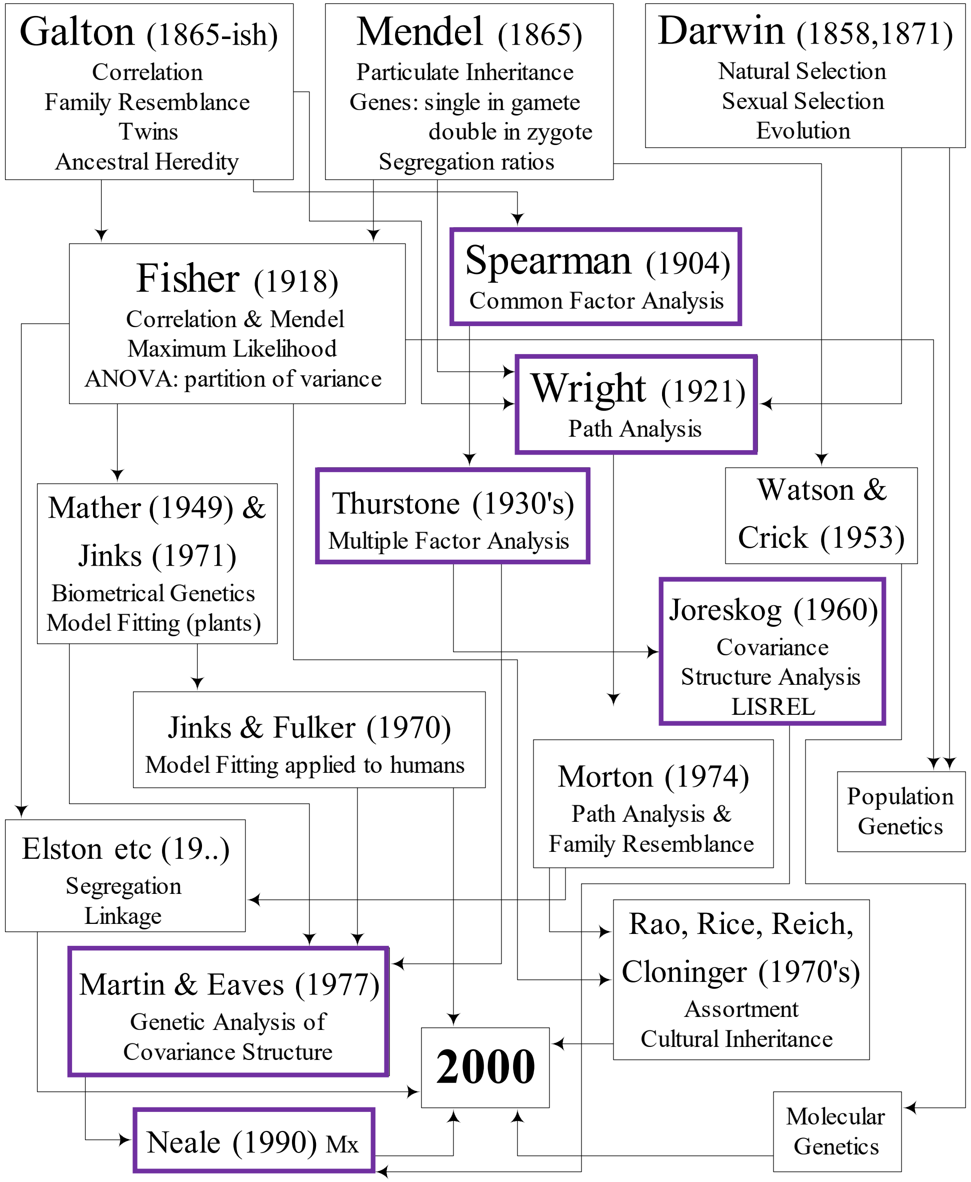
Neale & Cardon (1992)

Nick Martin
QIMRB, Brisbane,
Australia

Lindon
Eaves

Behaviour genetics community (classical twin design)

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

Molecular
Genetics

Neale (1990) Mx

Neale & Cardon (1992)

Mike Neale
VCU – Virginia, USA

Other software:
LISREL
EQS
Mplus
lavaan

SEM basics - How does SEM Work?

THEORY → BUILD MODEL

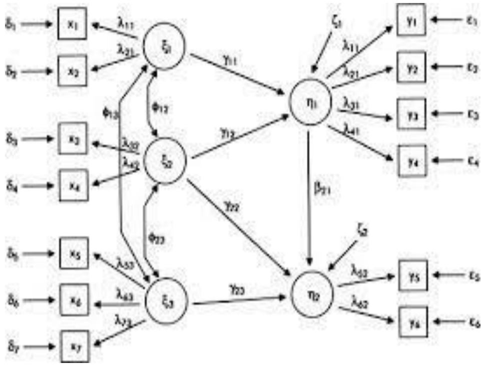
Expected covariance structure $\boldsymbol{\Sigma}(\boldsymbol{\theta})$

→ FIT TO DATA

Observed covariance structure $\boldsymbol{S}$

Yes → OBTAIN PARAMETER ESTIMATES AND CIs

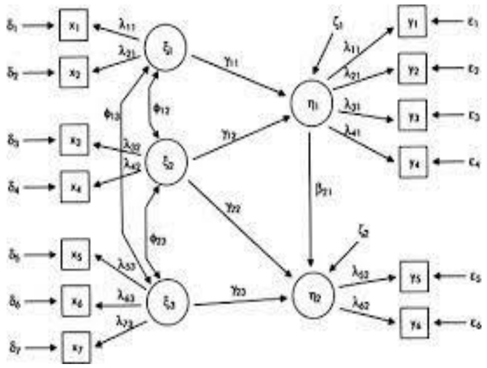
No → REVISE MODEL/ THEORY → THEORY



SEM basics - How does SEM Work?

THEORY

BUILD
MODEL



Expected
covariance
structure $\boldsymbol{\Sigma}(\boldsymbol{\theta})$

FIT
TO DATA

Observed
covariance
structure $\boldsymbol{S}$

Yes

OBTAIN PARAMETER
ESTIMATES
AND CIs

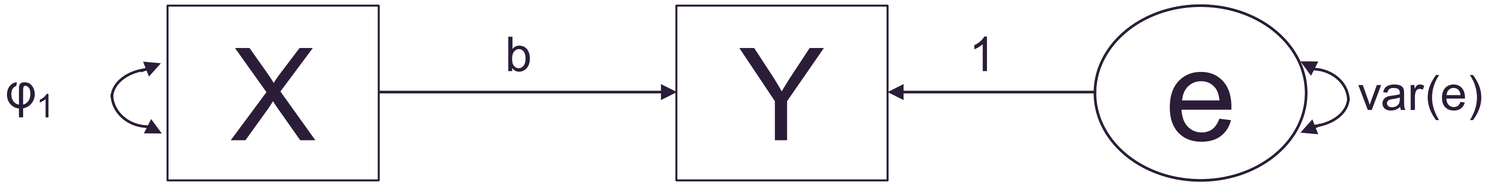
No

REVISE MODEL/
THEORY

SEM basics - Theory (e.g. Univariate Regression)

Theory: there is a linear relationship between the independent variable X and the dependent variable Y.

Structural Equation: $Y = bX + e$



Assume variables measured in deviation form

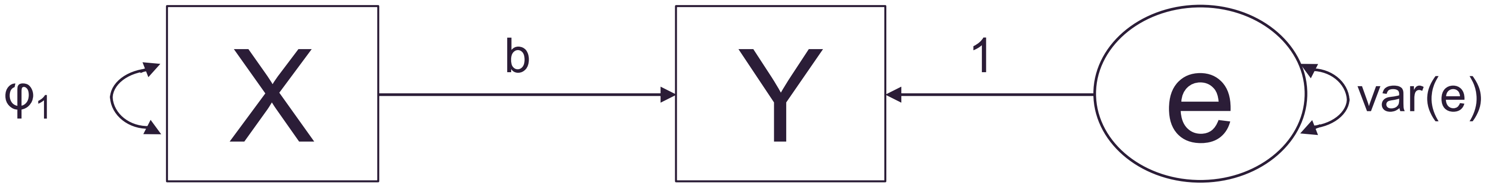
“b” is a path coefficient/regression coefficient.

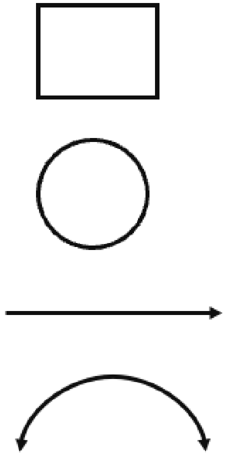
It quantifies the expected change in Y for every unit change in X is “b”

“e” is the error term (residual)

SEM basics - Path diagram elements

- A path diagram is a pictorial representation of an SEM





Observed Variables

Latent Variables

Causal Paths

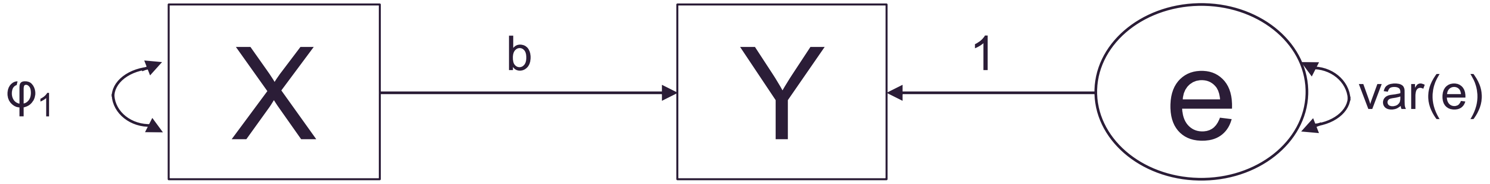
(Co)variance Paths

(Co)variance: the parameters on co/variance paths, e.g. φ_1

Path coefficient: the number on the causal path, e.g. b

*Latent variables are variables that can only be inferred indirectly through a mathematical model from other observable variables that can be directly observed or measured

SEM basics - Build Model (Univariate Regression)



- Y = bX + e (explicit)
- Measurement error in Y is e (explicit)
- No measurement error in X (explicit)
- No covariance between X and e (explicit)
- Covariance between X and Y is $b*\varphi_1$ (explicit)
- Linear relationships between the variables (implicit)
- Multivariate normality (normality of residuals; implicit)

SEM basics - Concepts

1. Identification

SEM basics - 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

SEM basics – Identified or Not?

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

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

$\theta_1 - \theta_2 = 0$

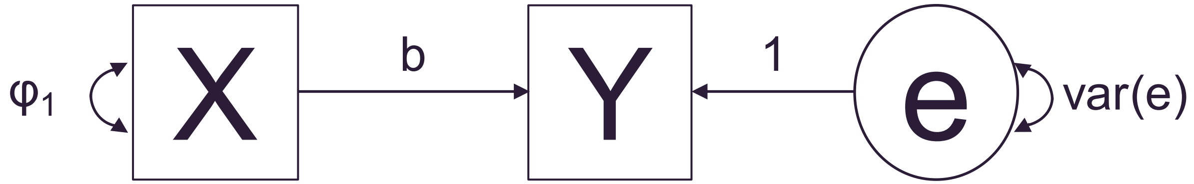
(3) $\theta_1 + \theta_2 = 10$

$2\theta_1 + 2\theta_2 = 20$

SEM basics – Identification

General rule

$t \leq n(n+1)/2$
t number of parameters to estimate
n number of observed variables



$$Y = bX + e$$

Number of estimated parameters: 3

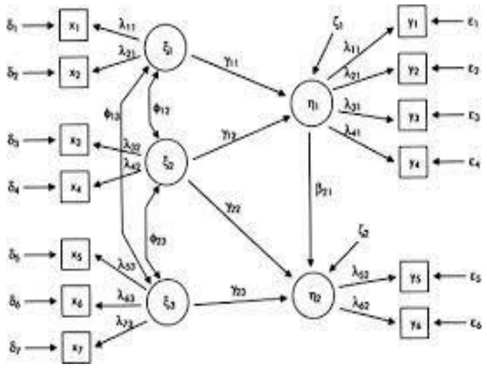
φ_1, b, var(e)

Number of observed variables: 2
Number of observed statistics:
2*3/2 =3 (var(X), cov(X,Y), var(Y))

SEM basics - How does SEM Work?

THEORY

BUILD
MODEL



Expected
covariance
structure $\boldsymbol{\Sigma}(\boldsymbol{\theta})$

FIT
TO DATA

Observed
covariance
structure $\boldsymbol{S}$

Yes

OBTAIN PARAMETER
ESTIMATES
AND CIs

No

REVISE MODEL/
THEORY

SEM basics - Concepts

1. Identification

2. Maximum Likelihood (Fit to data)

SEM basics - Likelihood (function)

The likelihood function (often simply called the likelihood) is the joint **probability** ***P()*** of the **observed data** (**x**) viewed as a function of the parameter(s) (**θ**) of a statistical model.

$$\mathcal{L}(\theta \mid x) = \prod_{j=1}^{N} P_{\theta}(x_j)$$

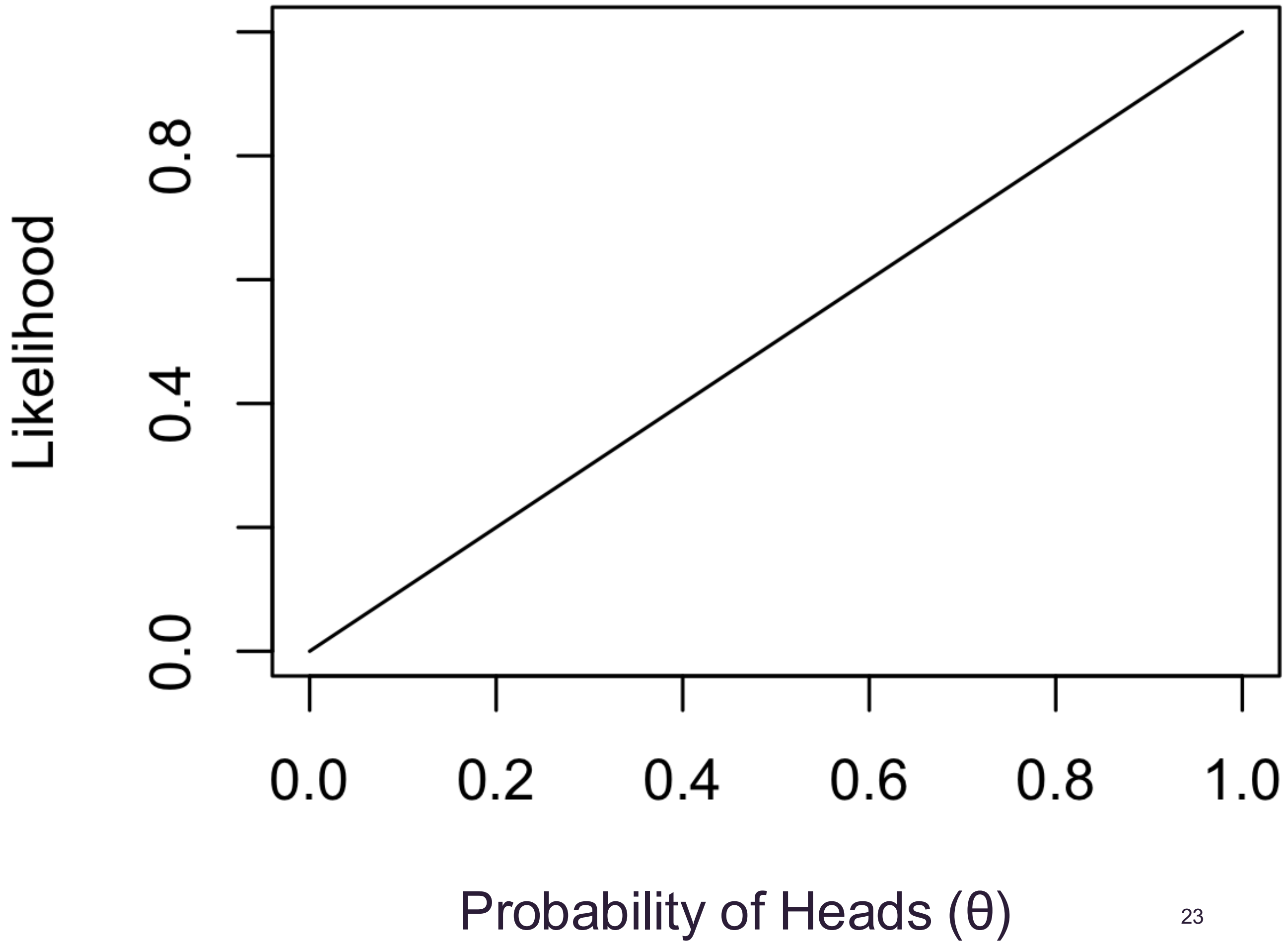
θ represents the parameters of the model.
x is the observed data
L(θ|x) represents the likelihood of the parameter θ given the observed data x.

SEM basics - Likelihood (function)

θ: *probability of heads*

x : *head*

$\mathcal{L}(\theta \mid x) = \theta$

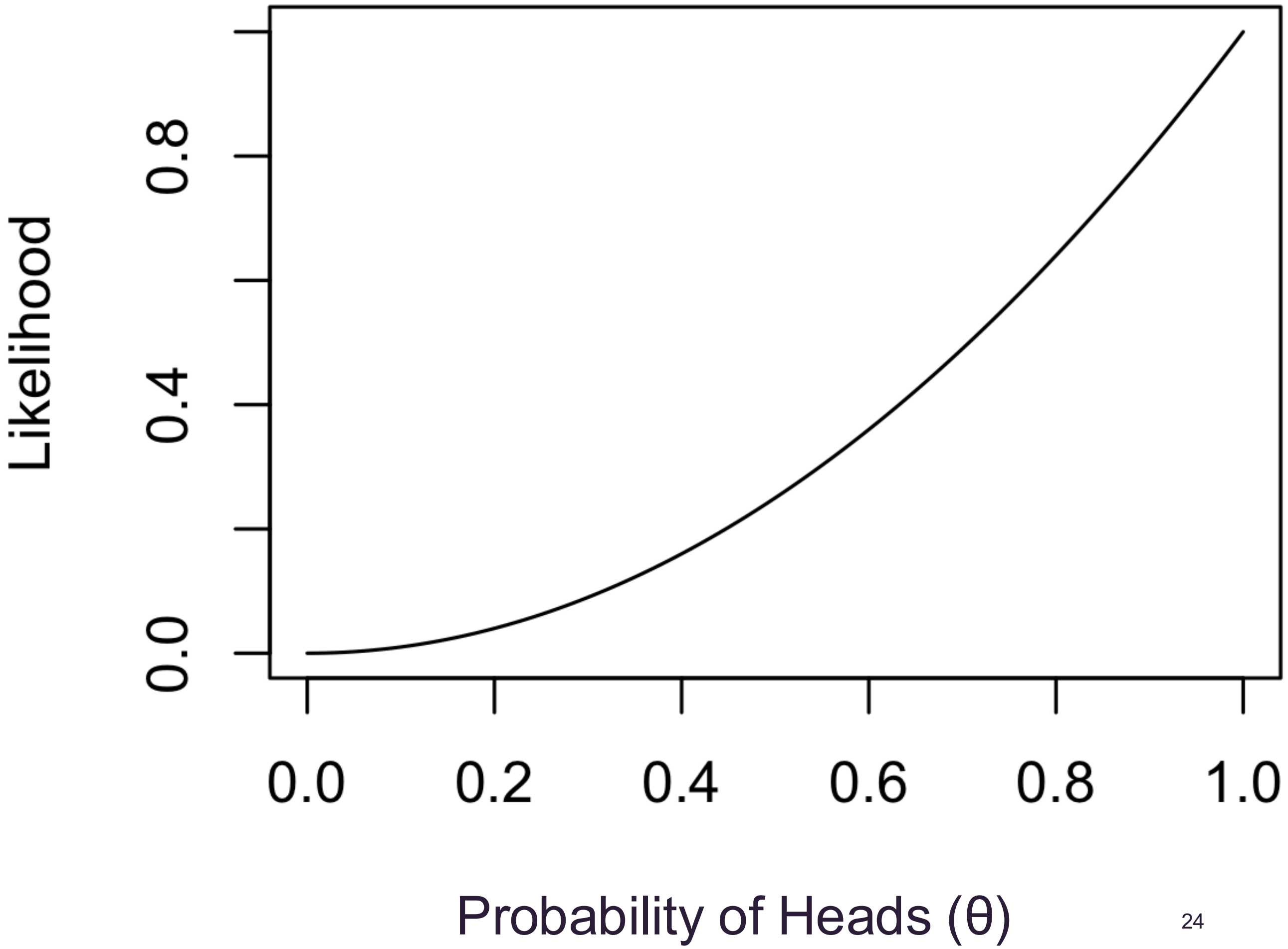


SEM basics - Likelihood (function)

$\theta: probability\ of\ heads$

$x : heads, heads$

$\mathcal{L}(\theta \mid x) = \theta . \theta$

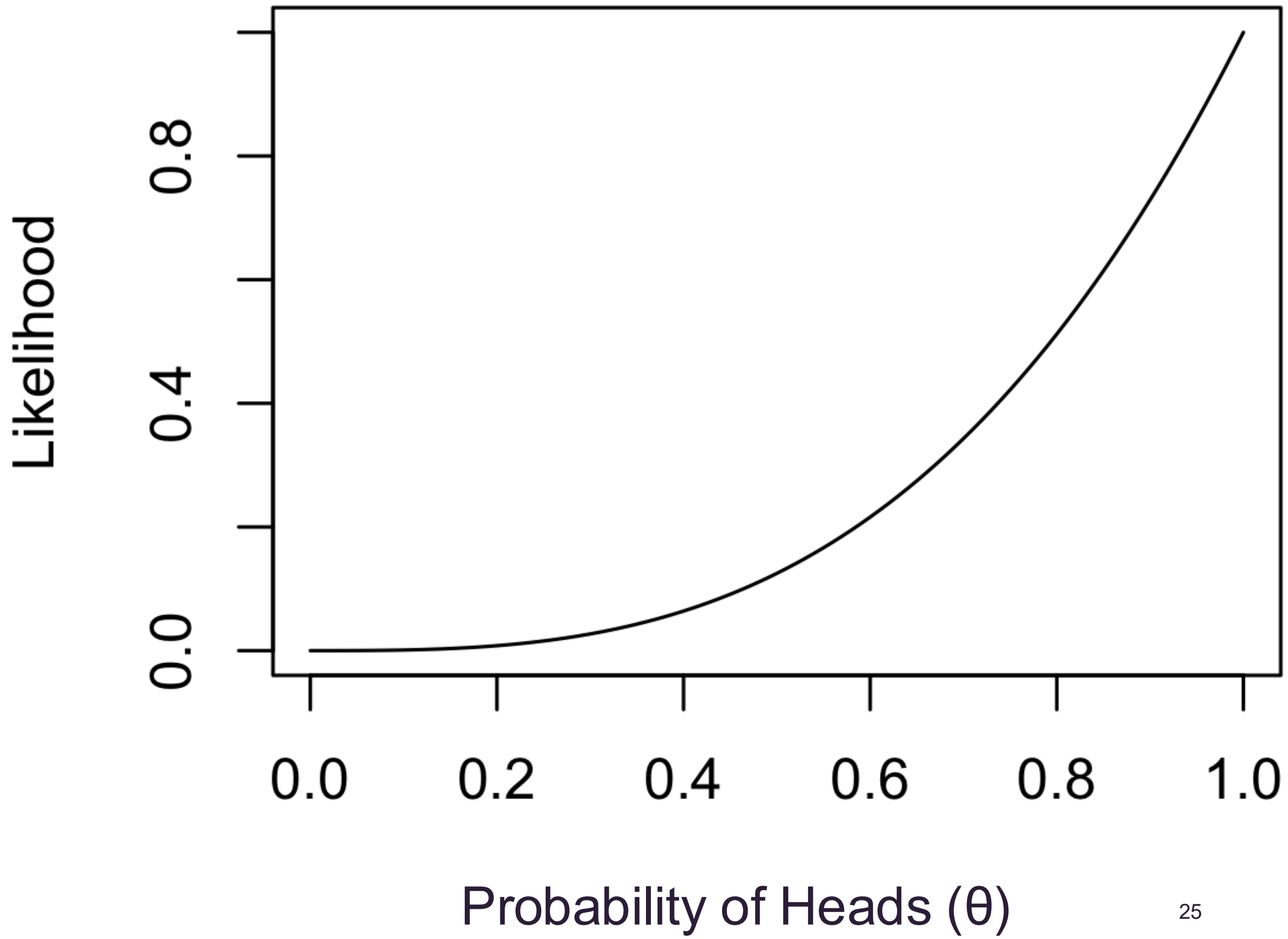


SEM basics - Likelihood (function)

$\theta: probability\ of\ heads$

$x : heads, heads, heads$

$\mathcal{L}(\theta \mid x) = \theta.\,\theta.\,\theta$

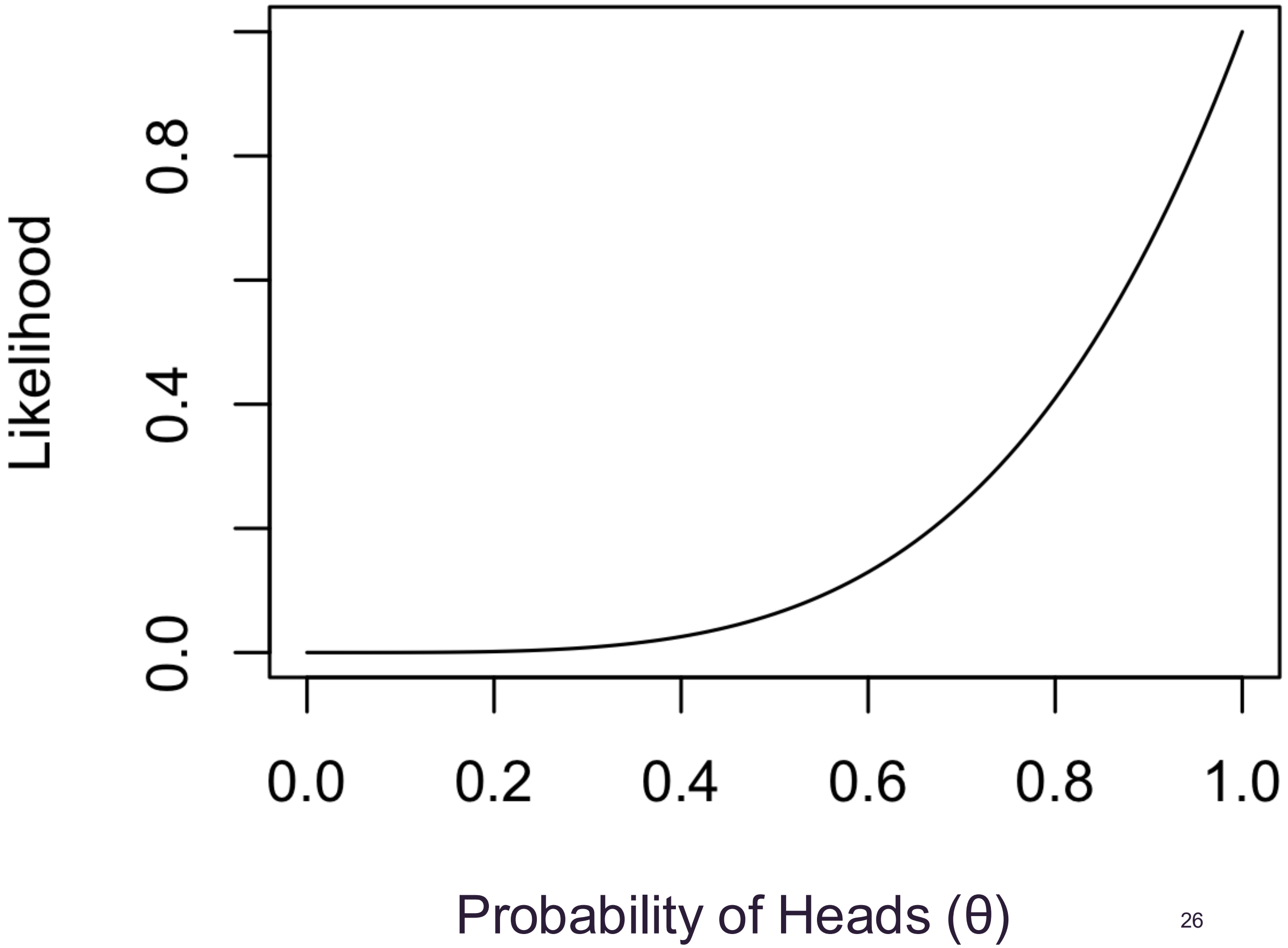


SEM basics - Likelihood (function)

$\theta: probability\ of\ heads$

$x: heads, heads, heads, heads$

$\mathcal{L}(\theta \mid x) = \theta.\theta.\theta.\theta$

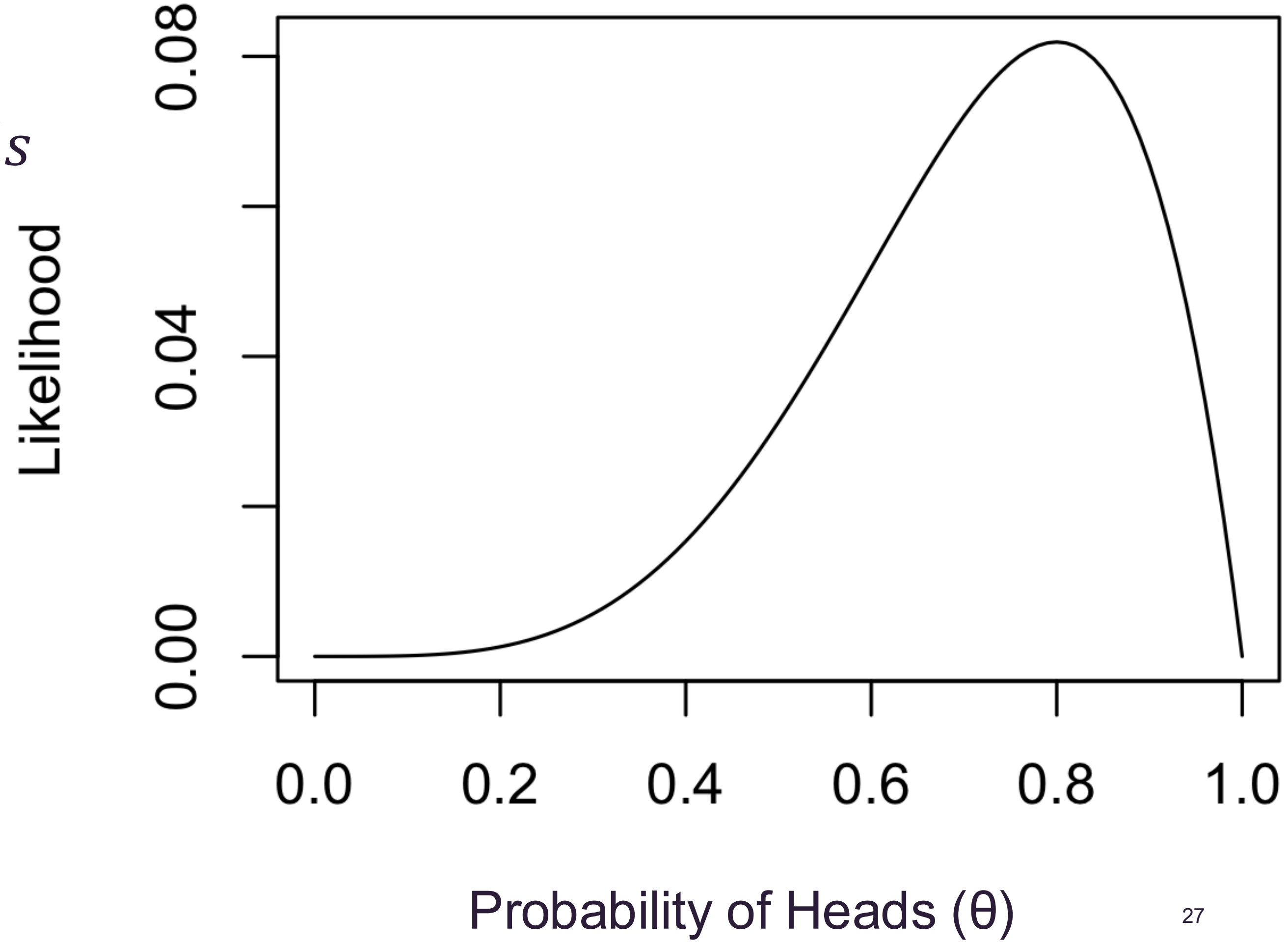


SEM basics - Likelihood (function)

θ: *probability of heads*

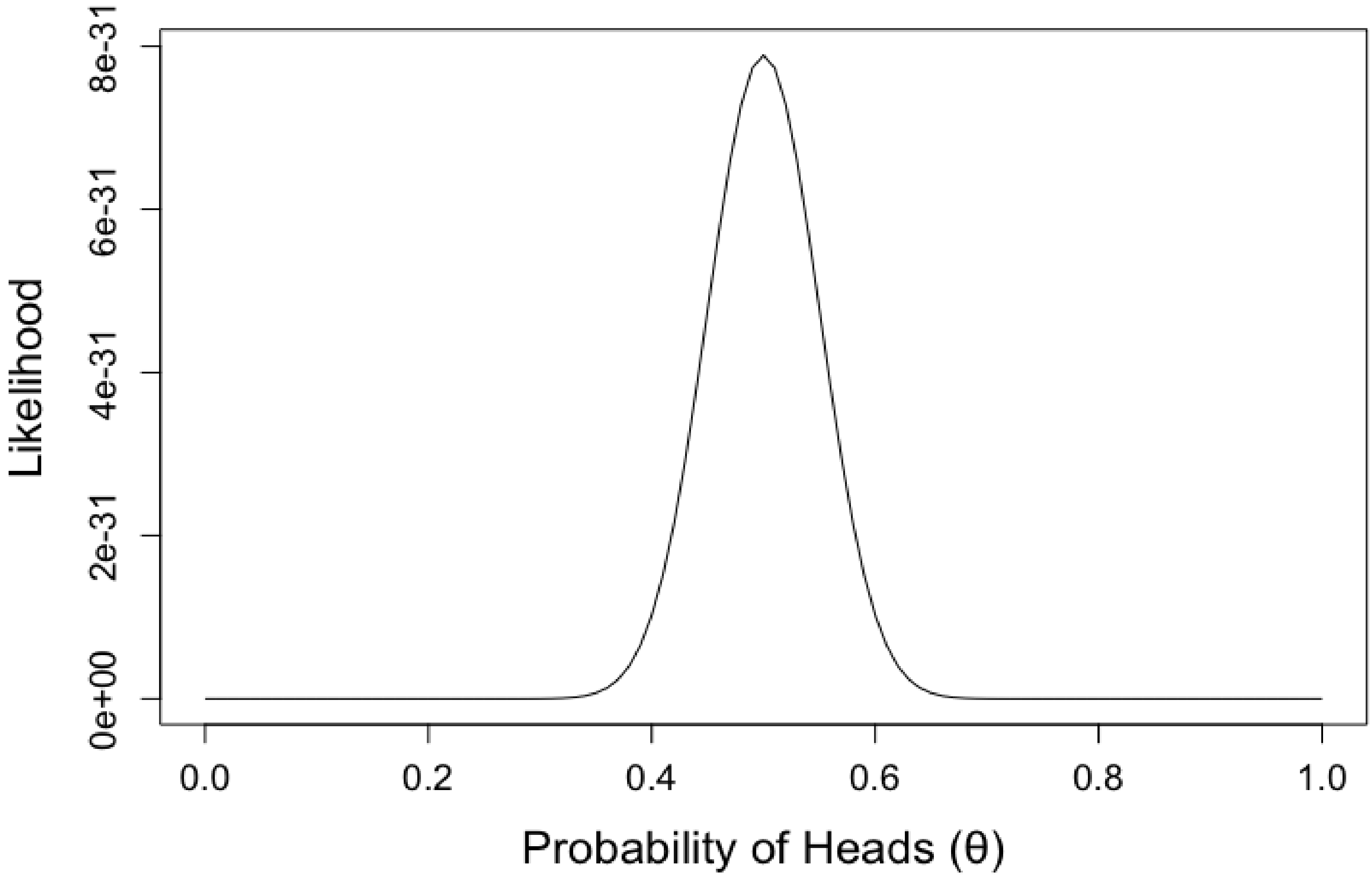
x : *heads, heads, heads, heads, tails*

$\mathcal{L}(\theta \mid x) = \theta.\,\theta.\,\theta.\theta.(1-\theta)$



SEM basics - Likelihood (function)

Evolution of Likelihood function as you add observations (1 to 100 coin flips)



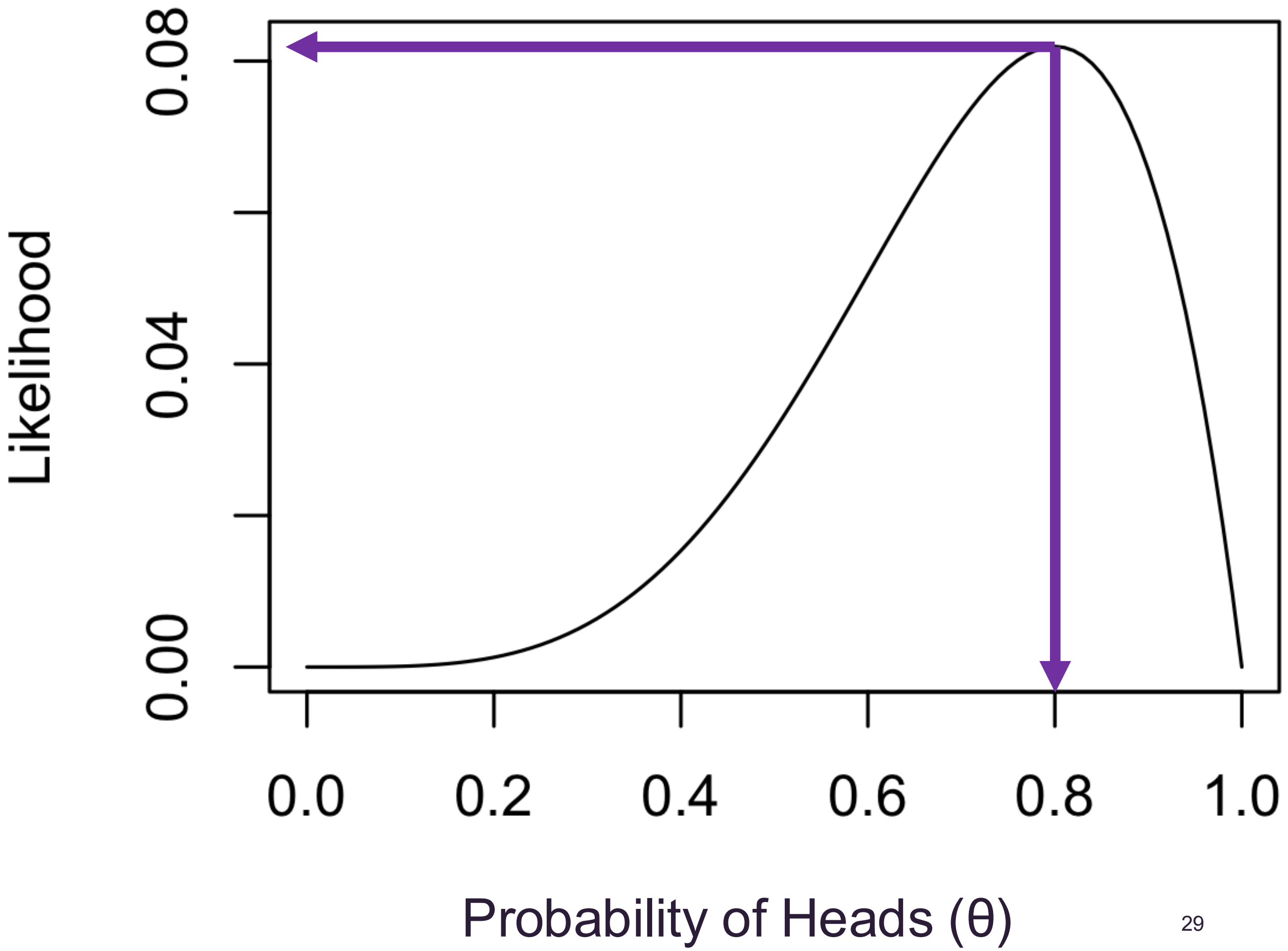
SEM basics - Maximum likelihood estimation (MLE)

Parameter that maximises the probability of the observed data

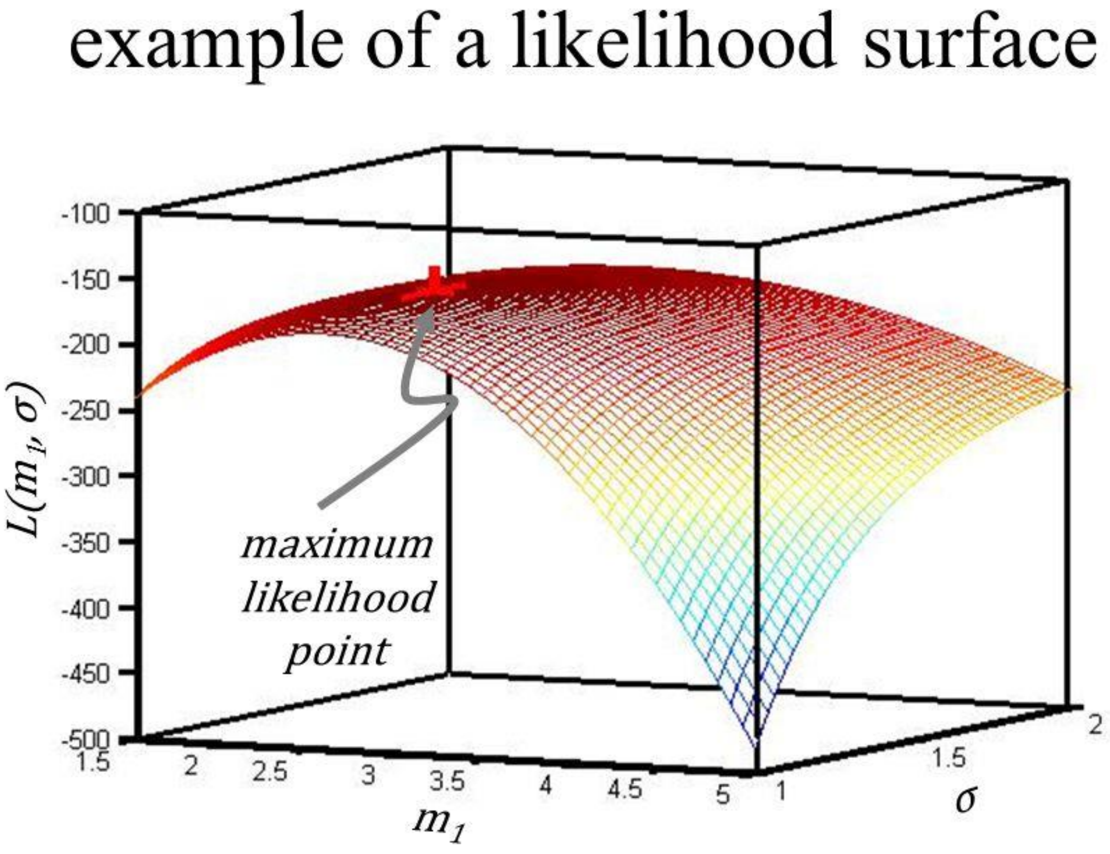
$\theta: probability\ of\ heads$

$x: heads, heads, heads, heads, tails$

$\mathcal{L}(\theta \mid x) = \theta.\theta.\theta.\theta.(1-\theta)$



SEM basics – Maximum Likelihood (>1 parameter)



- mean (m1)
- standard deviation (σ)
- Likelihood(L(m1, σ))

SEM basics – Modelling Both Means and Covariances Simultaneously

- **Maximize the multivariate normal likelihood according to the model for the means and the model for the covariances.**
- **Minimize the difference between the covariance matrix implied by the SEM (the "expected covariance matrix") and the observed covariance matrix**

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

Observed Covariance Matrix

Expected Covariance Matrix

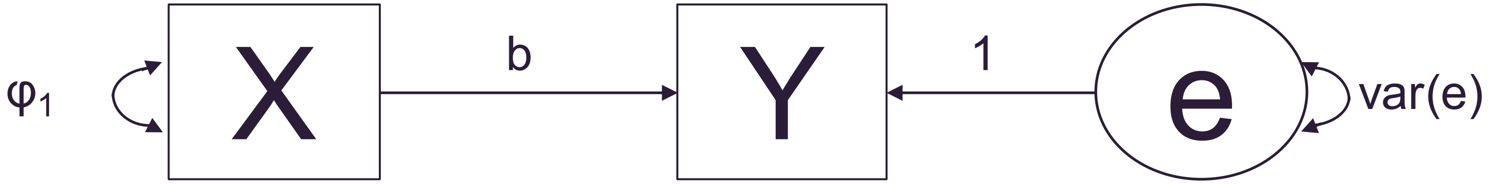
SEM basics – Understanding SEM

- Expected covariance matrix is a function of model parameters
- Parameters chosen to minimise the difference between observed and expected covariance matrices (MLEs)

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

Observed Covariance Matrix	Expected Covariance Matrix

SEM basics – Simple example (Univariate Regression)



Observed Covariance Matrix:

$$S = \begin{array}{cc} VAR(X) & COV(X,Y) \\ COV(X,Y) & VAR(Y) \end{array}$$

Expected/Implied Covariance Matrix:

$$\Sigma(\theta) = \begin{array}{cc} \varphi_1 & b\varphi_1 \\ b\varphi_1 & b^2\varphi_1 + var(e) \end{array}$$

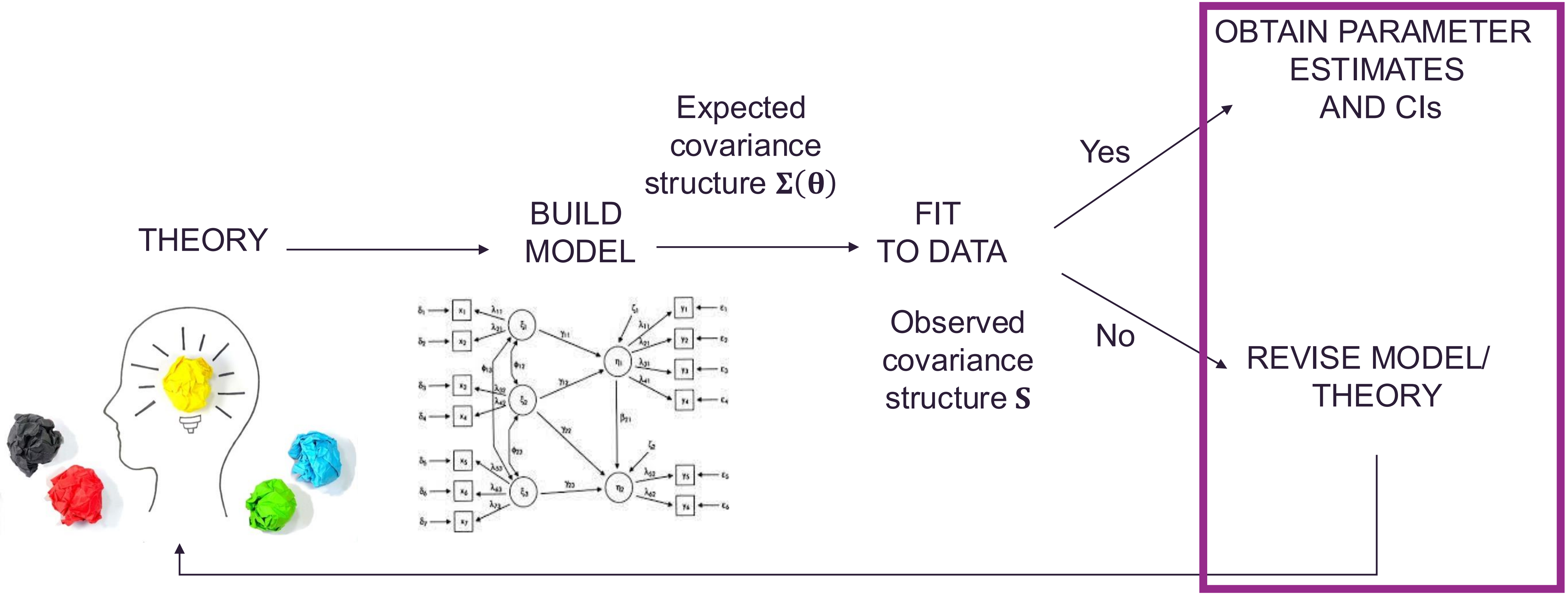
Number of observed variables: 2
Number of observed statistics: 3
(var(X), cov(X,Y), var(Y))

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

ML, FIML, REML

ML: Maximum likelihood	Fine for fixed effect models
FIML: Full Information Maximum Likelihood	Handles missing values
REML: Restricted Maximum Likelihood	Minimises bias in variance estimation of mixed models

SEM basics - How does SEM Work?



THEORY

BUILD
MODEL

Expected
covariance
structure $\boldsymbol{\Sigma}(\boldsymbol{\theta})$

FIT
TO DATA

Observed
covariance
structure $\boldsymbol{S}$

Yes

No

OBTAIN PARAMETER
ESTIMATES
AND CIs

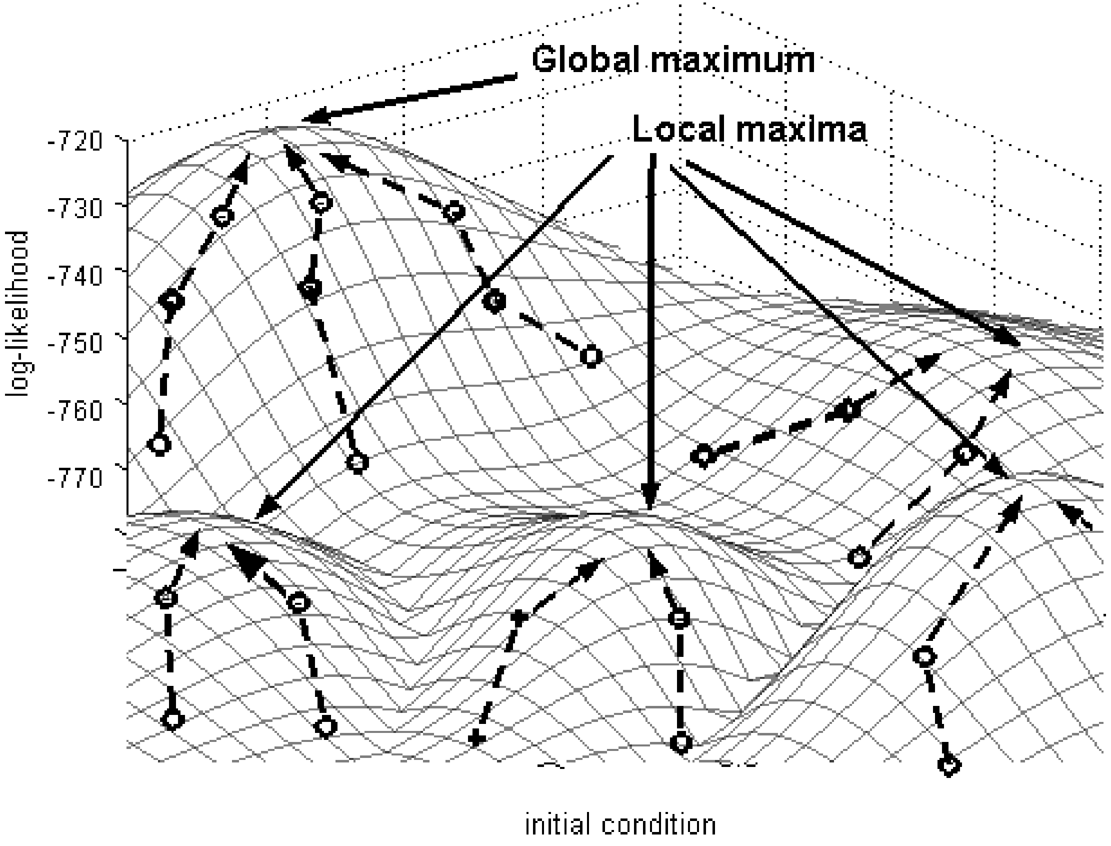
REVISE MODEL/
THEORY

SEM basics - Concepts

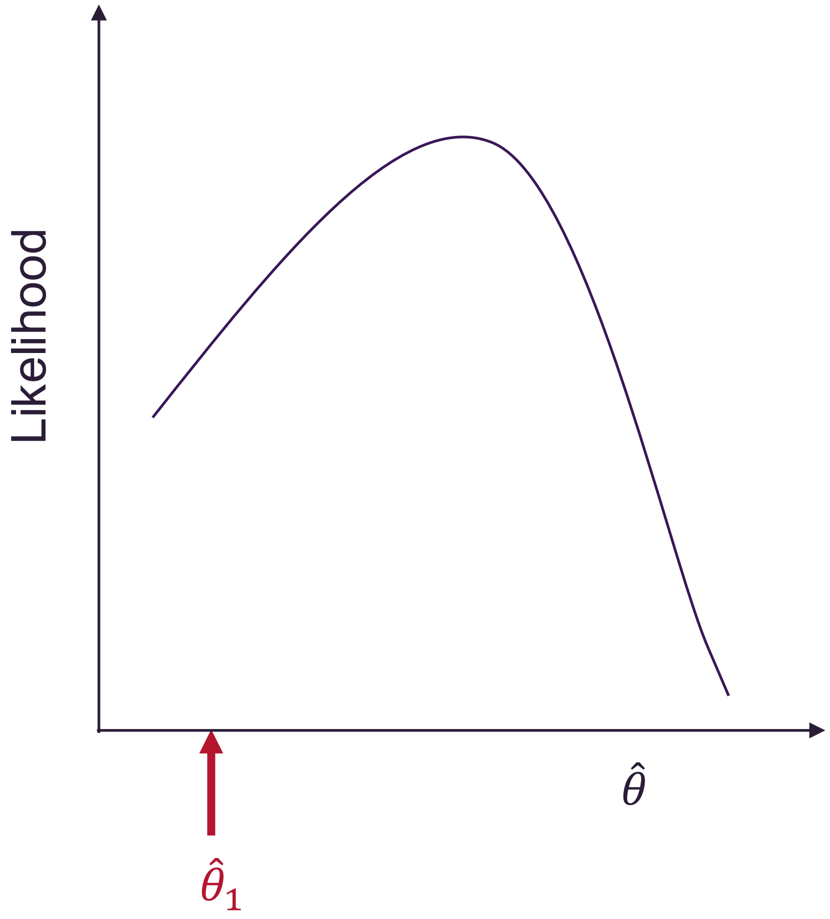
1. Identification
2. Maximum Likelihood
3. Optimization

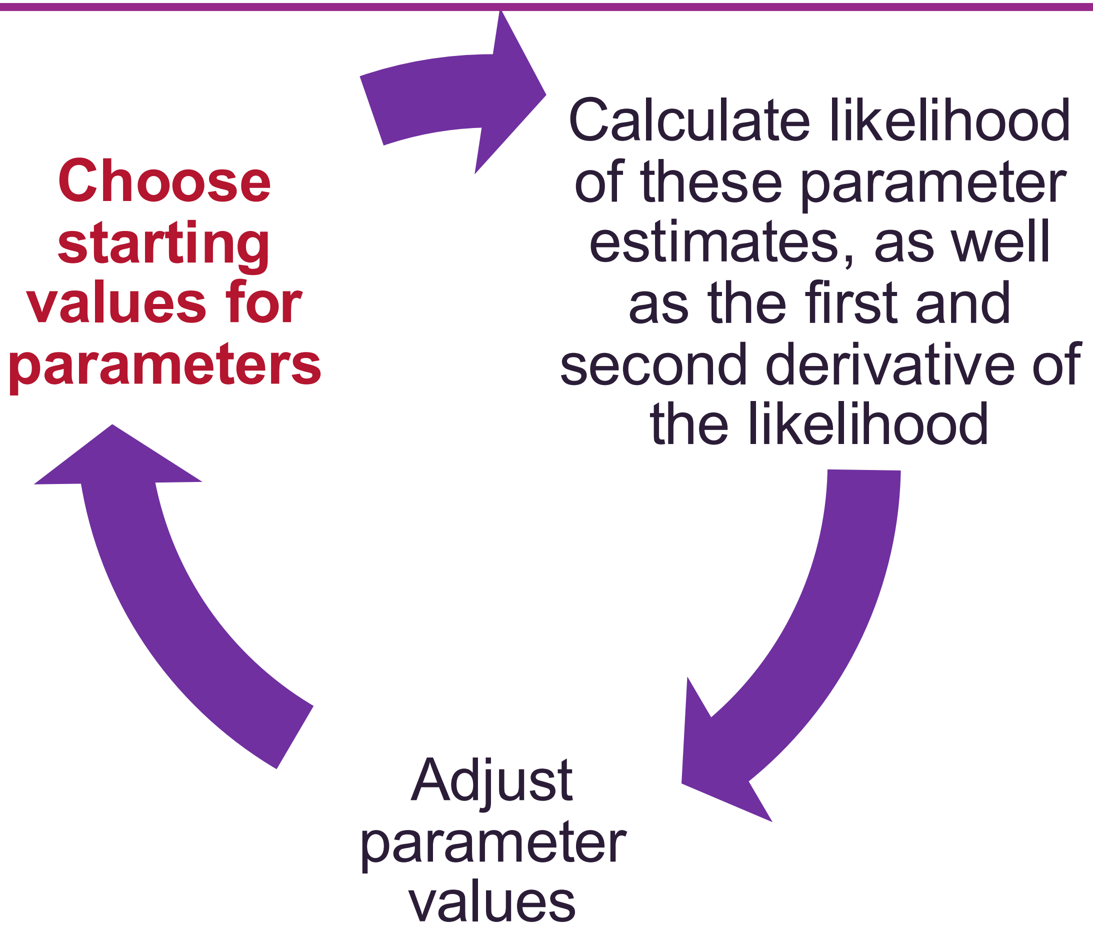
SEM basics - Optimization

- Maximum likelihood of complex model solutions can rarely be solved in closed form - rather, iterative optimization procedures are commonly needed.



SEM basics - Optimization

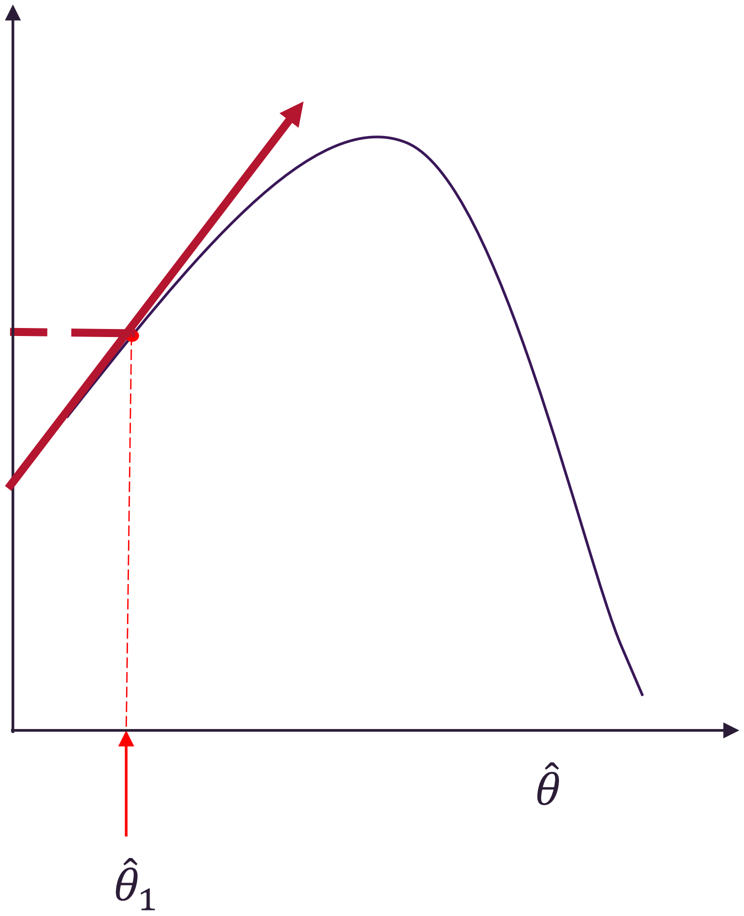


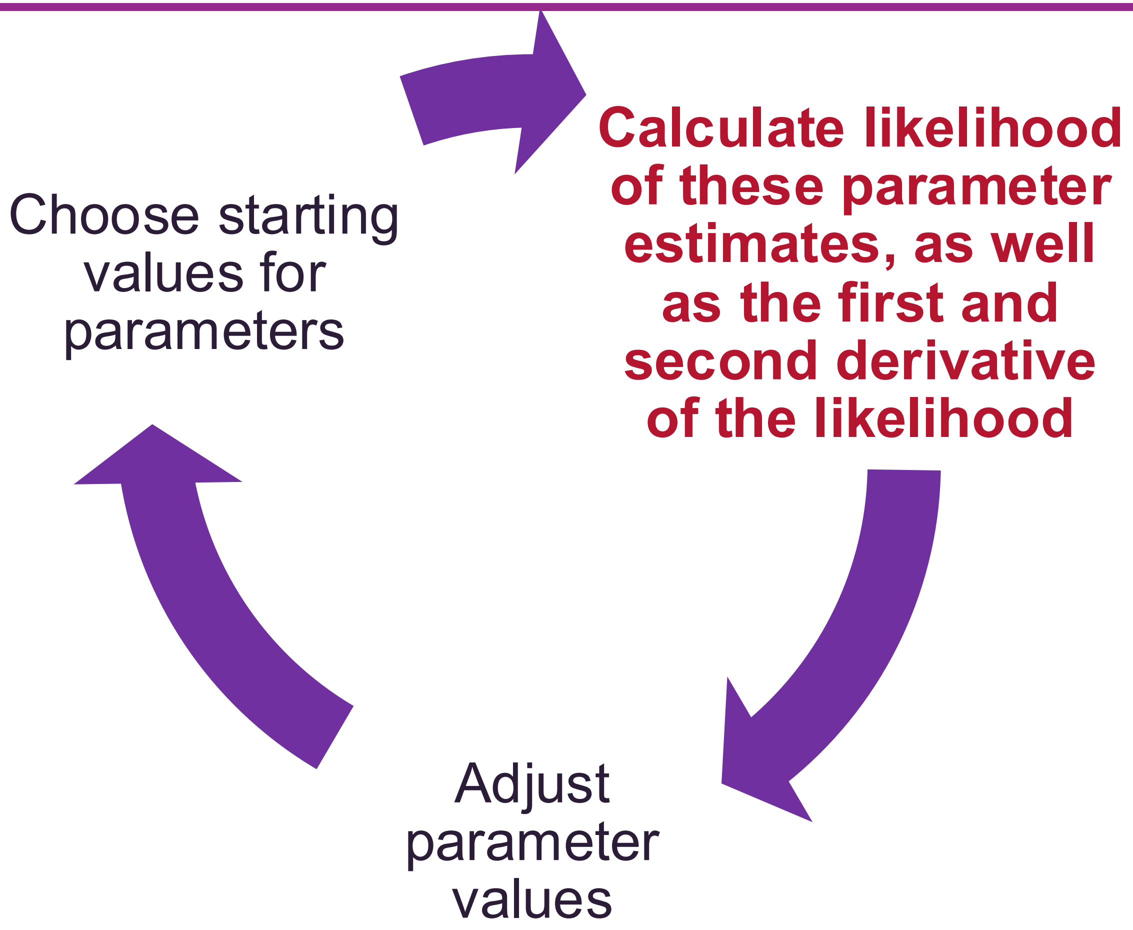

Choose starting values for parameters

Calculate likelihood of these parameter estimates, as well as the first and second derivative of the likelihood

Adjust parameter values

SEM basics - Optimization

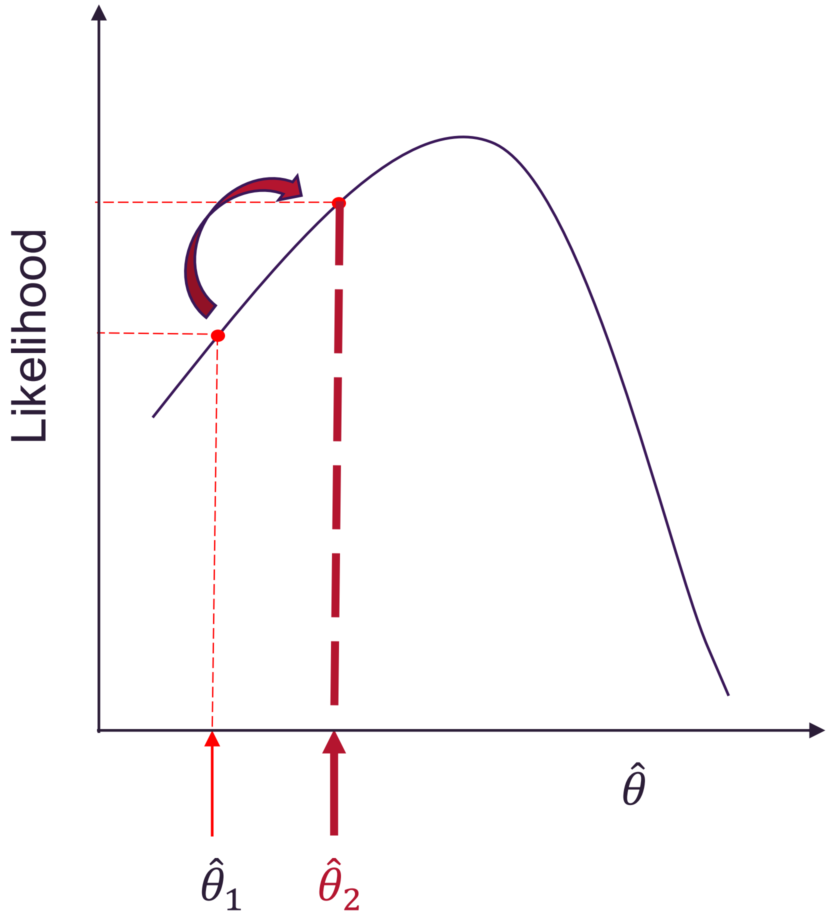


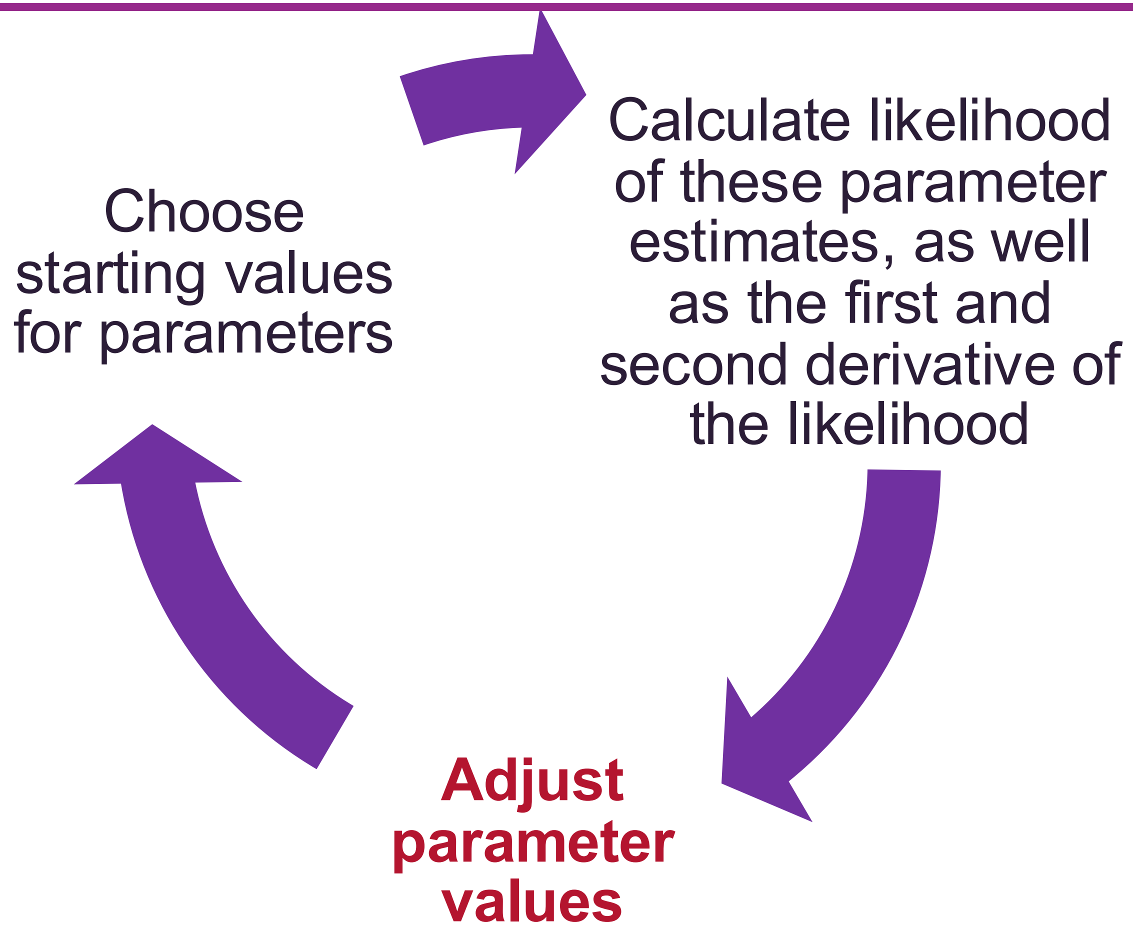

Choose starting values for parameters

Calculate likelihood of these parameter estimates, as well as the first and second derivative of the likelihood

Adjust parameter values

SEM basics - Optimization



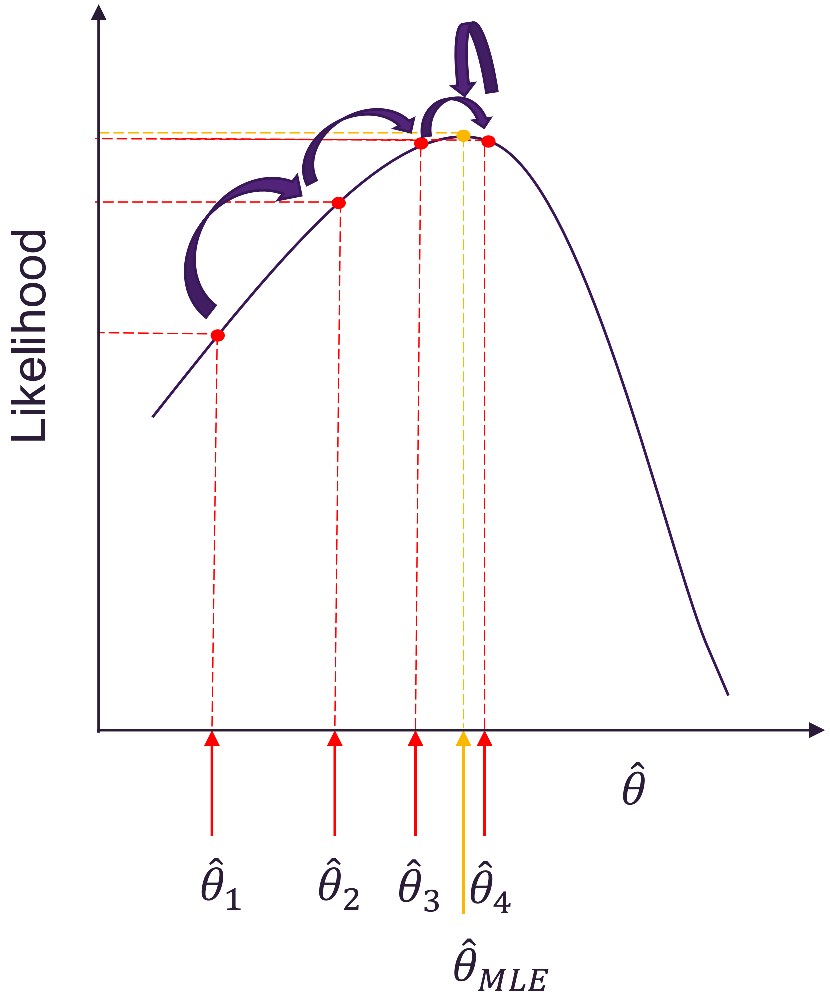


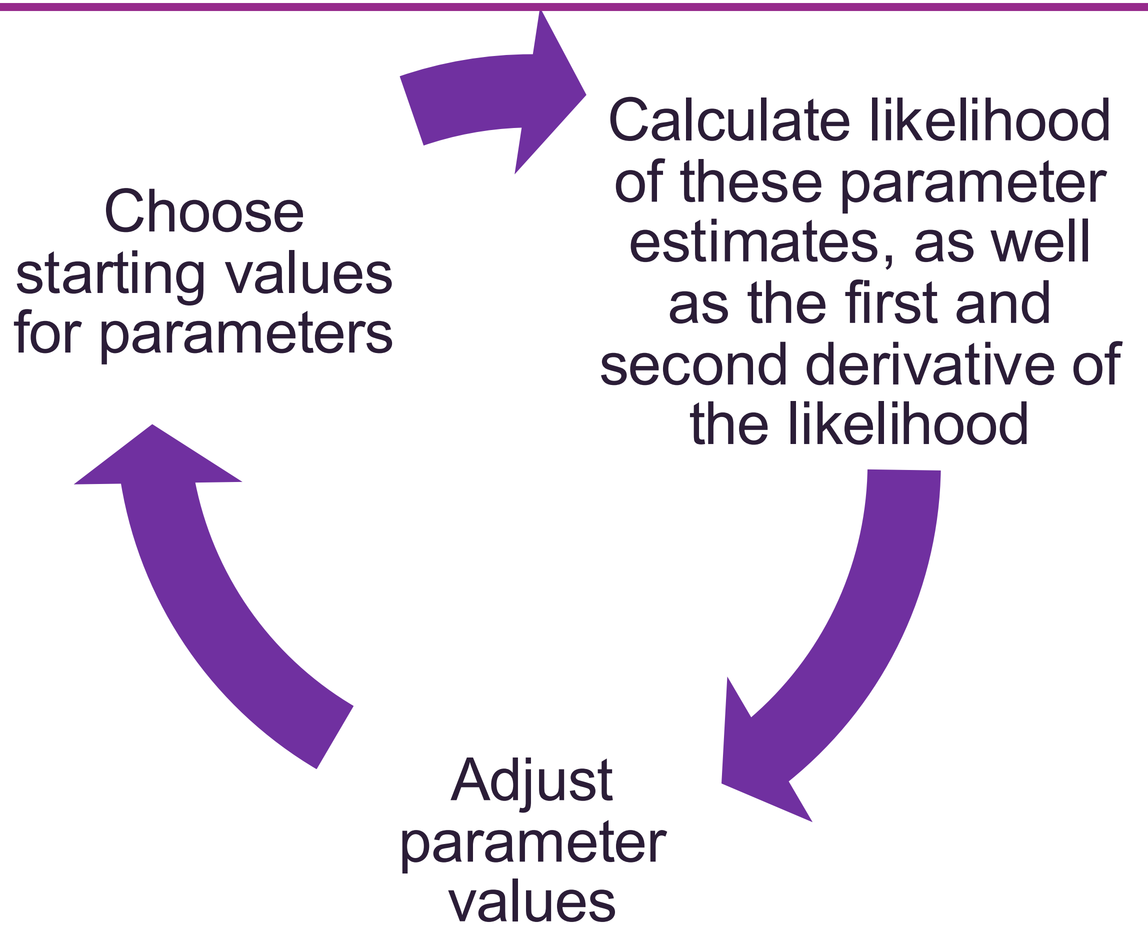
Choose starting values for parameters

Calculate likelihood of these parameter estimates, as well as the first and second derivative of the likelihood

Adjust parameter values

SEM basics - Optimization





Choose starting values for parameters

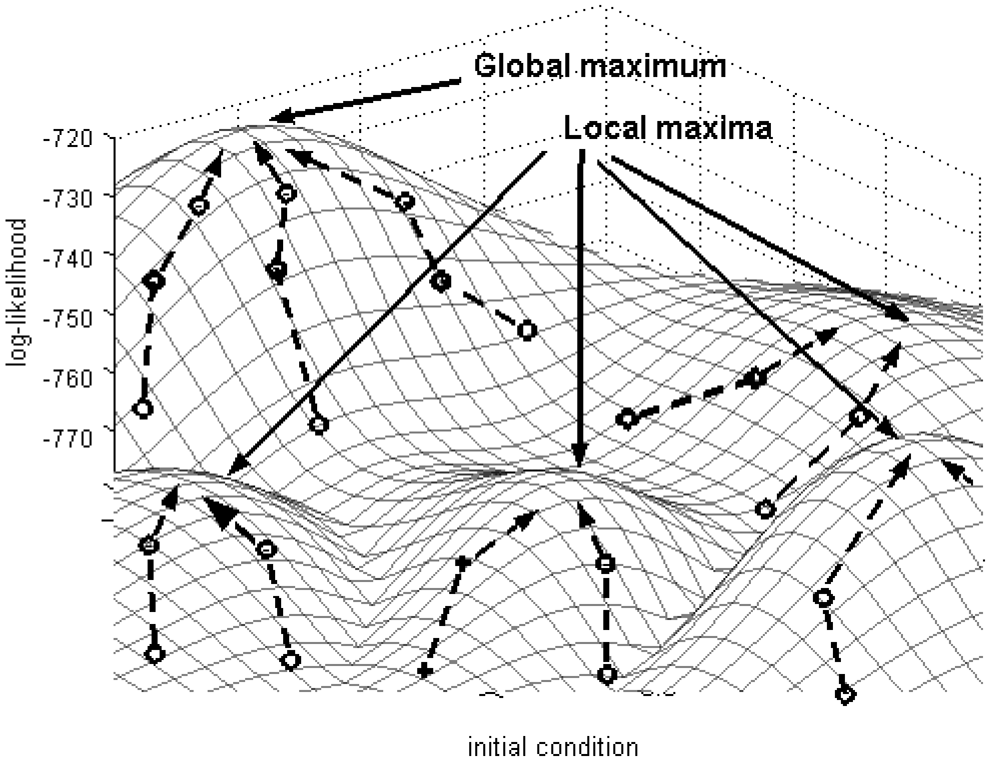
Calculate likelihood of these parameter estimates, as well as the first and second derivative of the likelihood

Adjust parameter values

Repeat process until stopping criterion is reached

SEM basics - Optimization

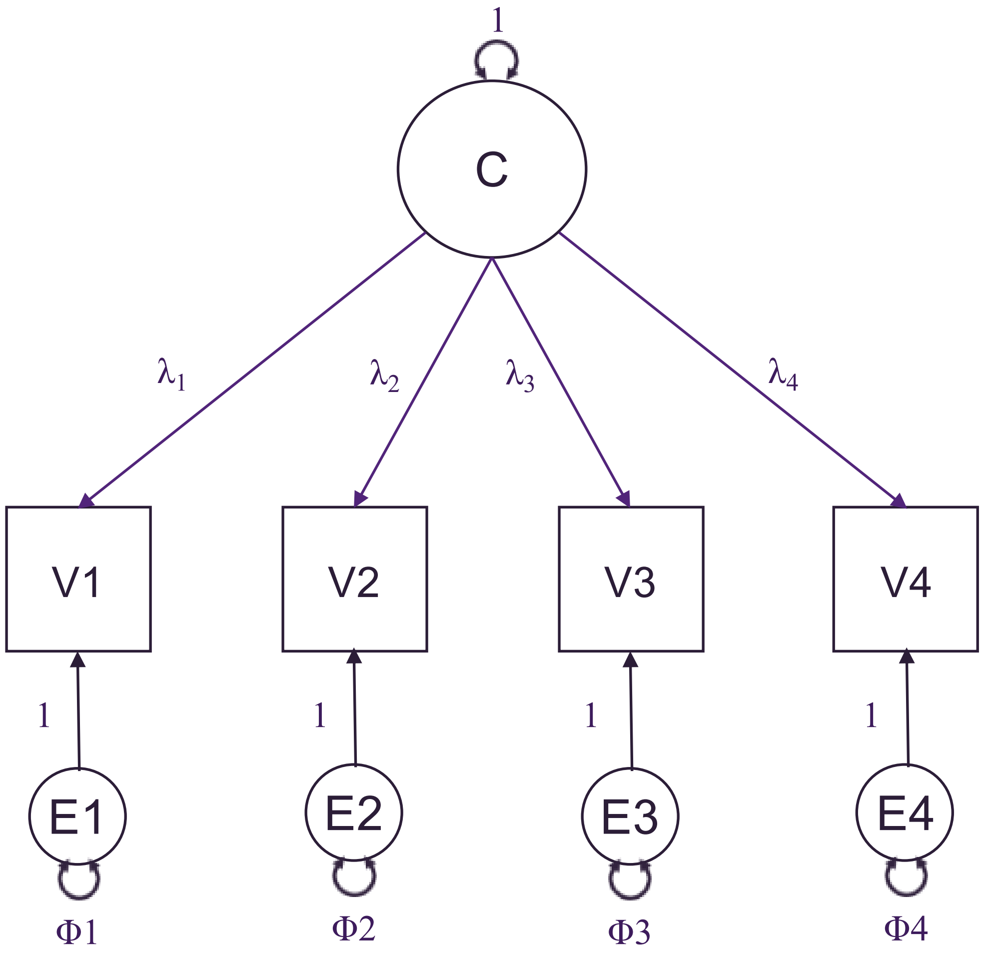
- Typically, we maximize the log-likelihood because computers find it easier to add rather than multiply.
- The likelihood surface may be complicated with one or more local maxima
- Choosing different starting values can increase confidence in a global solution
- In general, it is good practice to choose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



SEM basics - Fit indices

- **Chi-square test:** A low chi-square value relative to its degrees of freedom and a non-significant p-value suggest a good model fit.
- **Aikake Information Criterion (AIC):** Lower AIC values indicate a better model fit relative to other models.
- **Comparative Fit Index (CFI):** CFI values range from 0 to 1, with values closer to 1 indicating a better fit.
- **Standardized Root Mean Square Residual (SRMR):** SRMR values range from 0 to 1, with lower values indicating a better fit. An SRMR value less than 0.08 is generally considered a good fit.
- **Root Mean Square Error of Approximation (RMSEA):** Fit index where a value of zero indicates the best fit

Path diagram – Common Factor Model



Structural Equations:

$$V_1 = \lambda_1 C + E_1$$

$$V_2 = \lambda_2 C + E_2$$

$$V_3 = \lambda_3 C + E_3$$

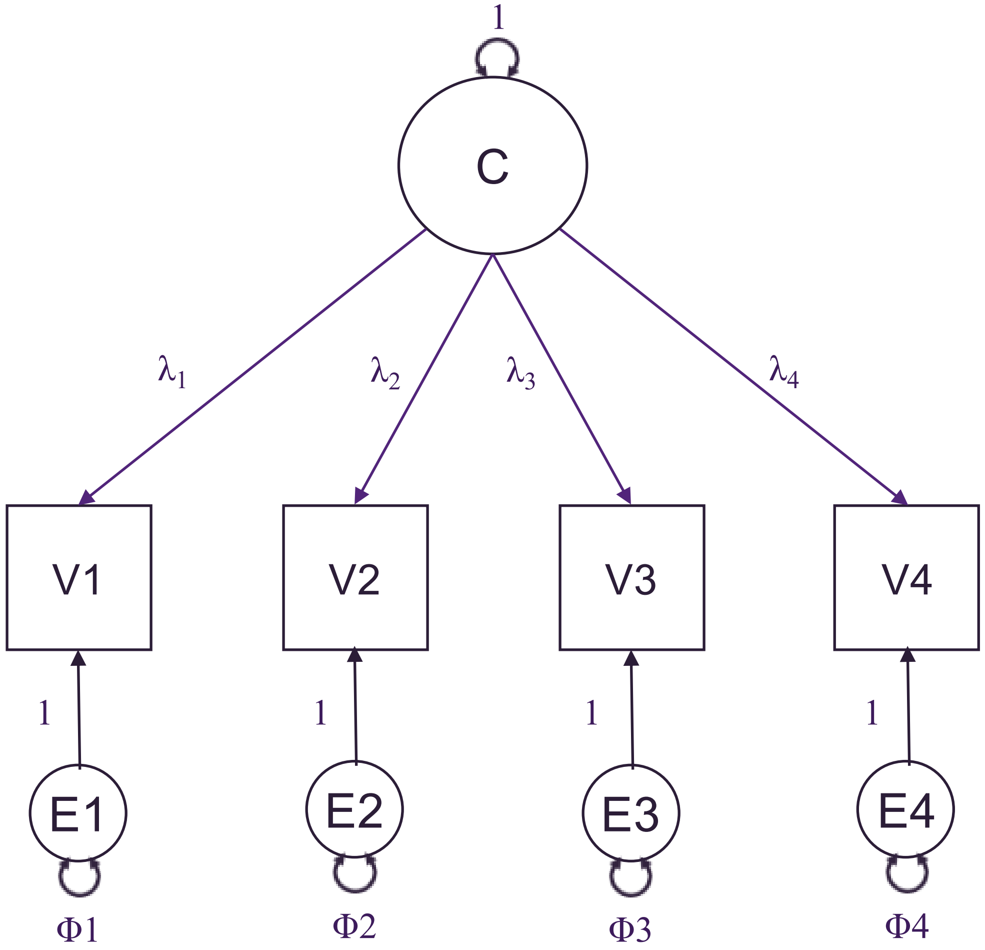
$$V_4 = \lambda_4 C + E_4$$

Observed Covariance Matrix:

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

Number of observed variable: 4
Number of observed statistics: 4*5/2=10
Number of estimated parameters:

Path diagram – 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:

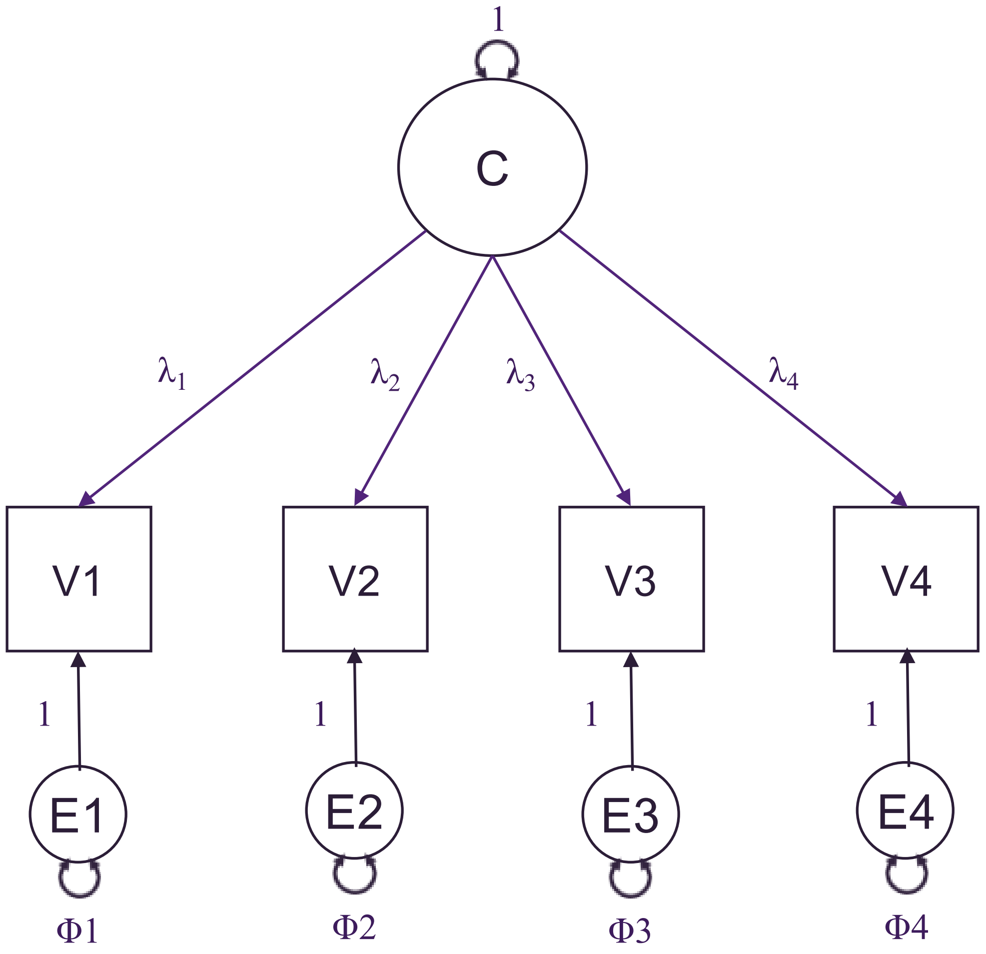
$$S = \begin{matrix} VAR(V_1) & COV(V_1, V_2) & COV(V_1, V_3) & COV(V_1, V_4) \\ COV(V_2, V_1) & VAR(V_2) & COV(V_2, V_3) & COV(V_2, V_4) \\ COV(V_3, V_1) & COV(V_3, V_2) & VAR(V_3) & COV(V_3, V_4) \\ COV(V_4, V_1) & COV(V_4, V_2) & COV(V_4, V_3) & VAR(V_4) \end{matrix}$$

Number of observed statistics: 10

Number of estimated parameters: 8 < 10

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \Phi_1, \Phi_2, \Phi_3, \Phi_4)$

Path diagram – Common Factor Model



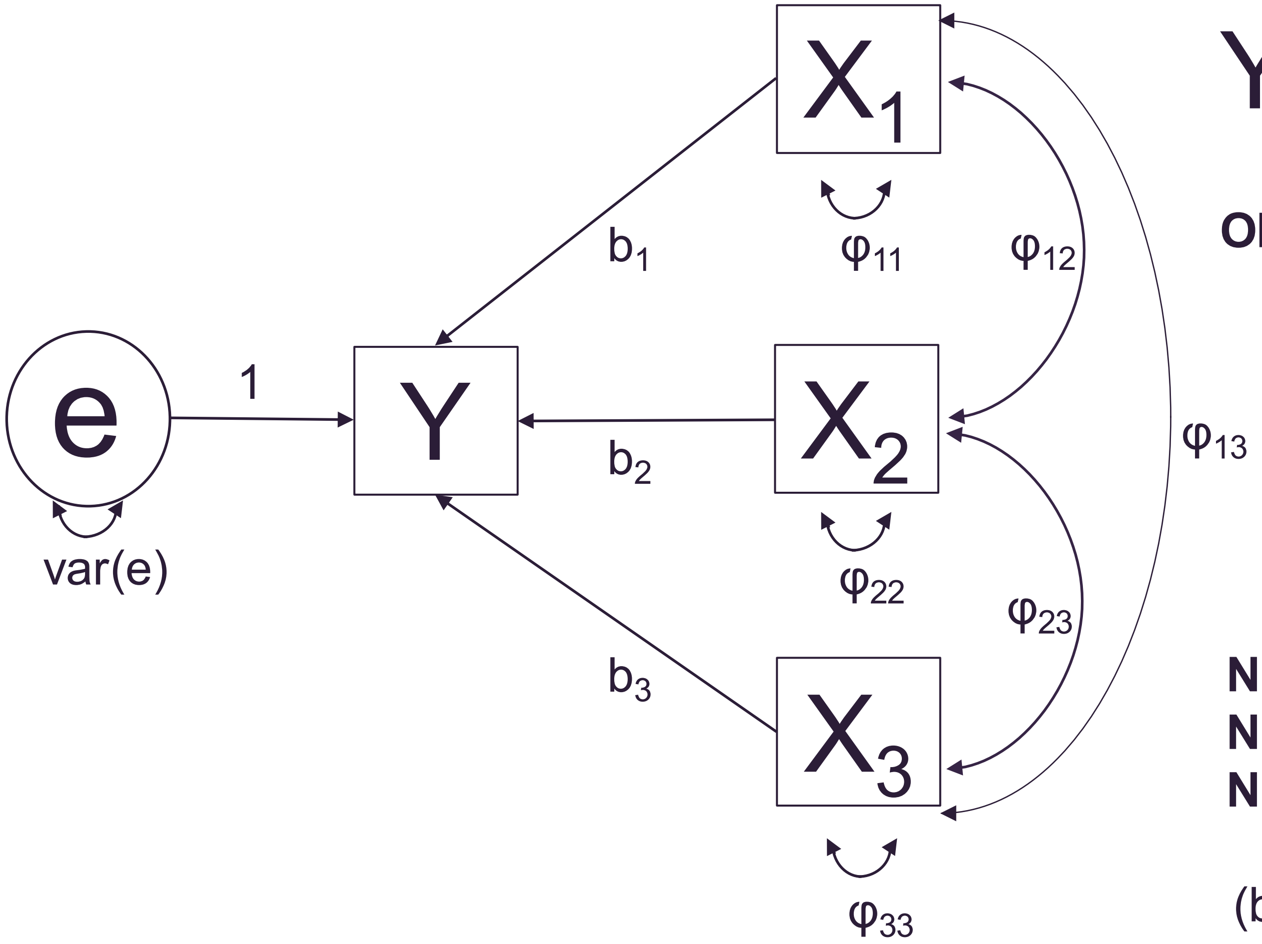
Observed Covariance Matrix:

$$S = \begin{matrix} VAR(V_1) & COV(V_{1,}V_2) & COV(V_{1,}V_3) & COV(V_{1,}V_4) \\ COV(V_2,V_1) & VAR(V_2) & COV(V_{2,}V_3) & COV(V_2,V_4) \\ COV(V_3,V_1) & COV(V_3,V_2) & VAR(V_3) & COV(V_3,V_4) \\ COV(V_4,V_1) & COV(V_4,V_2) & COV(V_4,\ V_3) & VAR(V_4) \end{matrix}$$

Expected Covariance Matrix:

$$\boldsymbol{\Sigma(\theta)} = \begin{matrix} \lambda_1{}^2+\Phi_1 & \lambda_1\lambda_2 & \lambda_1\lambda_3 & \lambda_1\lambda_4 \\ \lambda_2\lambda_1 & \lambda_2{}^2+\Phi_2 & \lambda_2\lambda_3 & \lambda_2\lambda_4 \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3{}^2+\Phi_3 & \lambda_3\lambda_4 \\ \lambda_4\lambda_1 & \lambda_4\lambda_2 & \lambda_4\lambda_3 & \lambda_4{}^2+\Phi_4 \end{matrix}$$

Path diagram – A more complex model



Structural Equation:

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

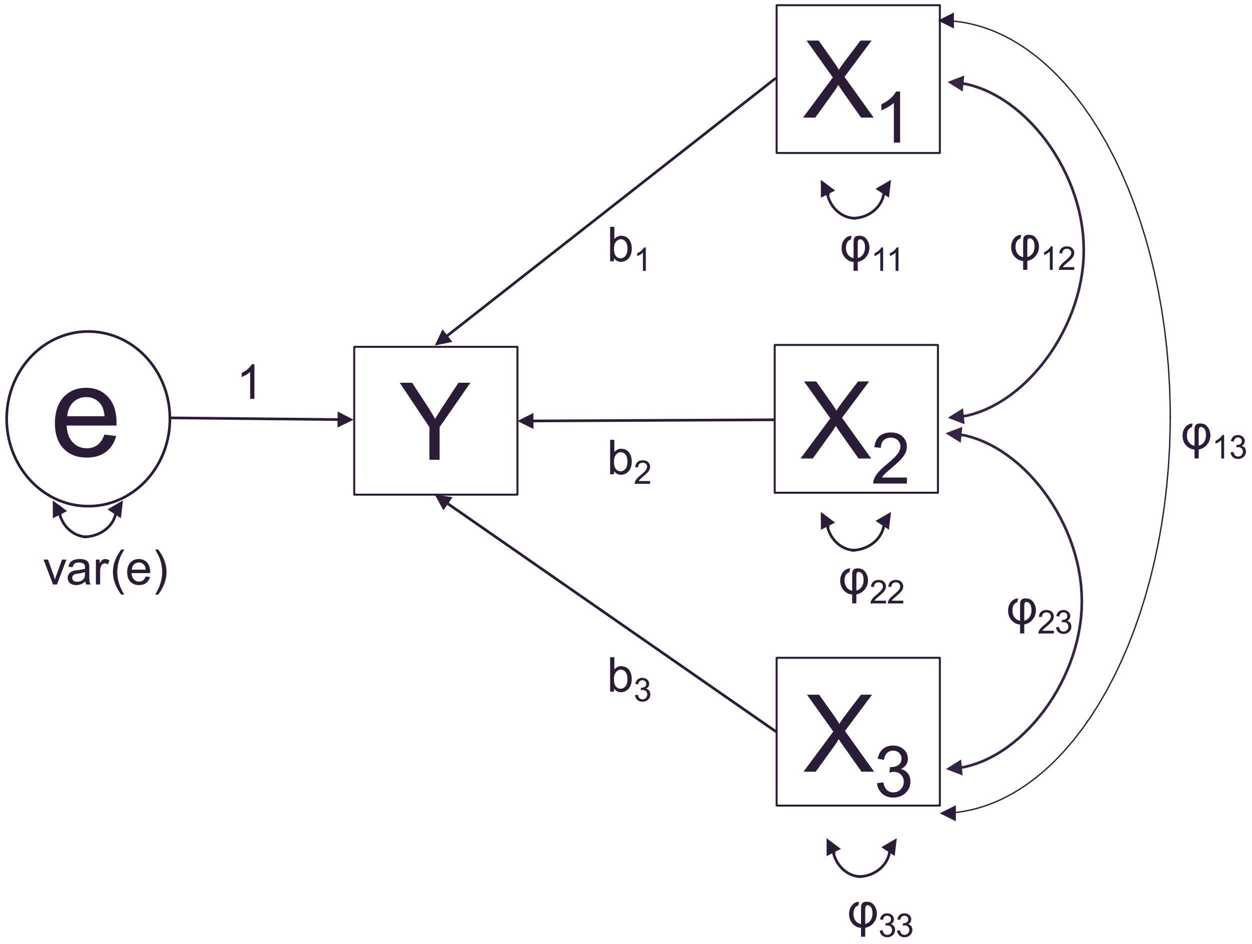
Observed Covariance Matrix:

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

Number of observed variables: 4
Number of observed statistics: (4*5)/2=10
Number of estimated parameters: 10

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

Path diagram – A more complex model



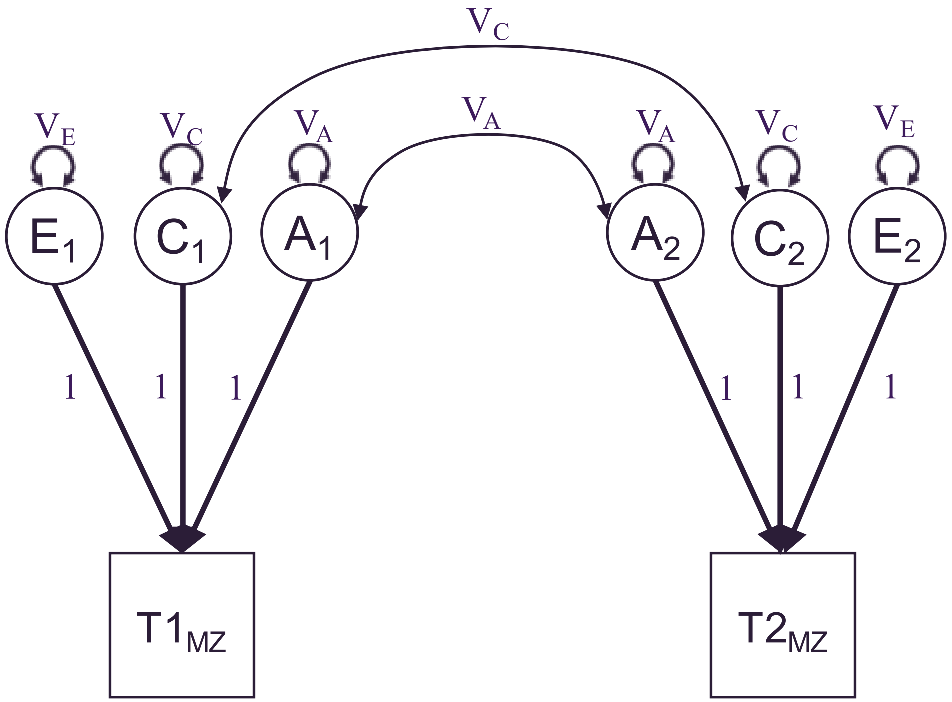
Observed Covariance Matrix:

$$S = \begin{array}{cccc} VAR(X_1) & COV(X_1,X_2) & COV(X_1,X_3) & COV(X_1,Y) \\ COV(X_2,X_1) & VAR(X_2) & COV(X_2,X_3) & COV(X_2,Y) \\ COV(X_3,X_1) & COV(X_3,X_2) & VAR(X_3) & COV(X_3,Y) \\ COV(Y,X_1) & COV(Y,X_2) & COV(Y, X_3) & VAR(Y) \end{array}$$

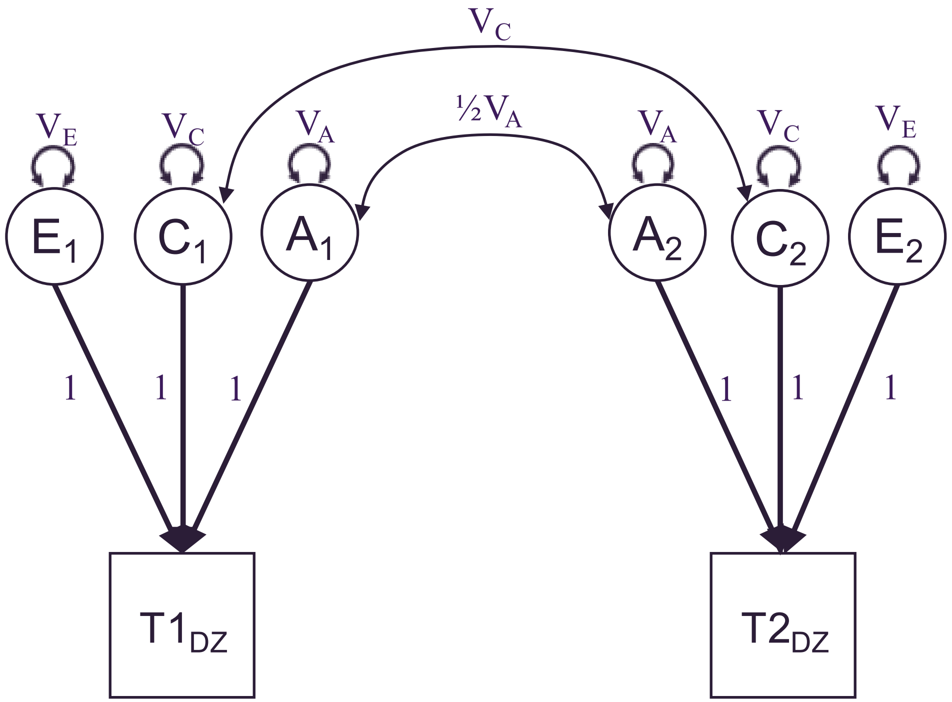
Expected Covariance Matrix:

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

Path diagram – Classical Twin Design



Monozygotic Twins



Dizygotic Twins

Additive genetic effects (A)
Common environmental effects (C)
Unique Environmental effects (E)

Structural Equations:

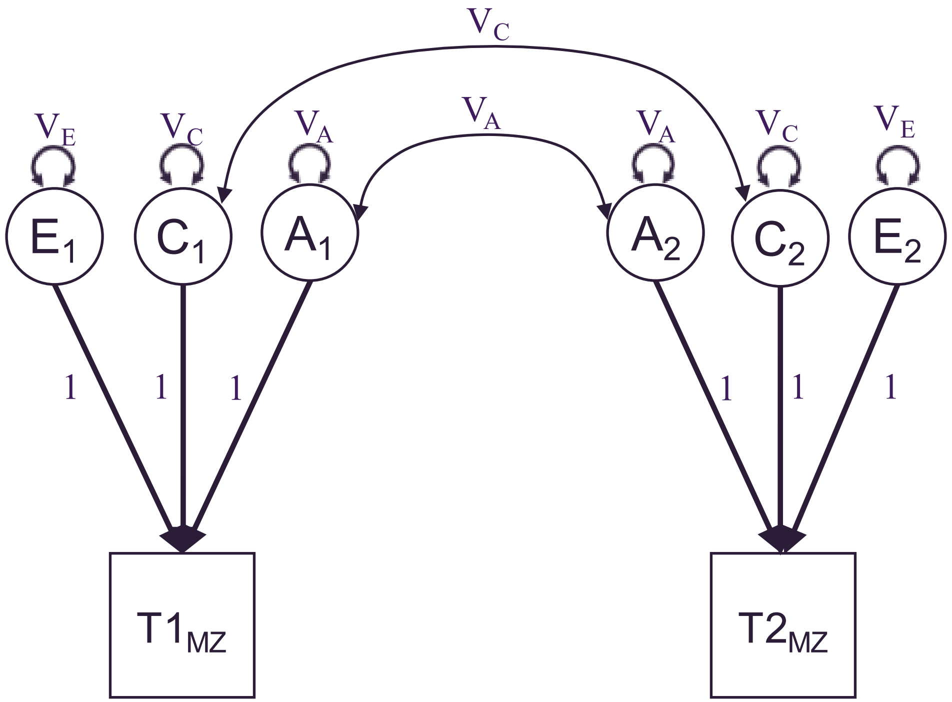
$T1_{MZ} = A_1 + C_1 + E_1$

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

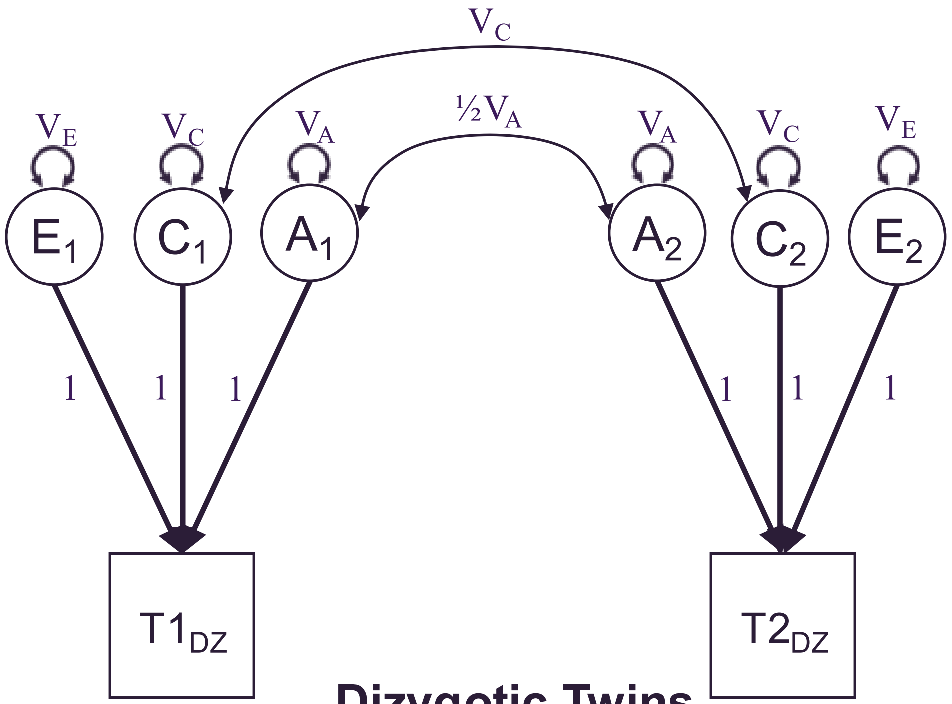
$T1_{DZ} = A_1 + C_1 + E_1$

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

Path diagram – Classical Twin Design



Monozygotic Twins



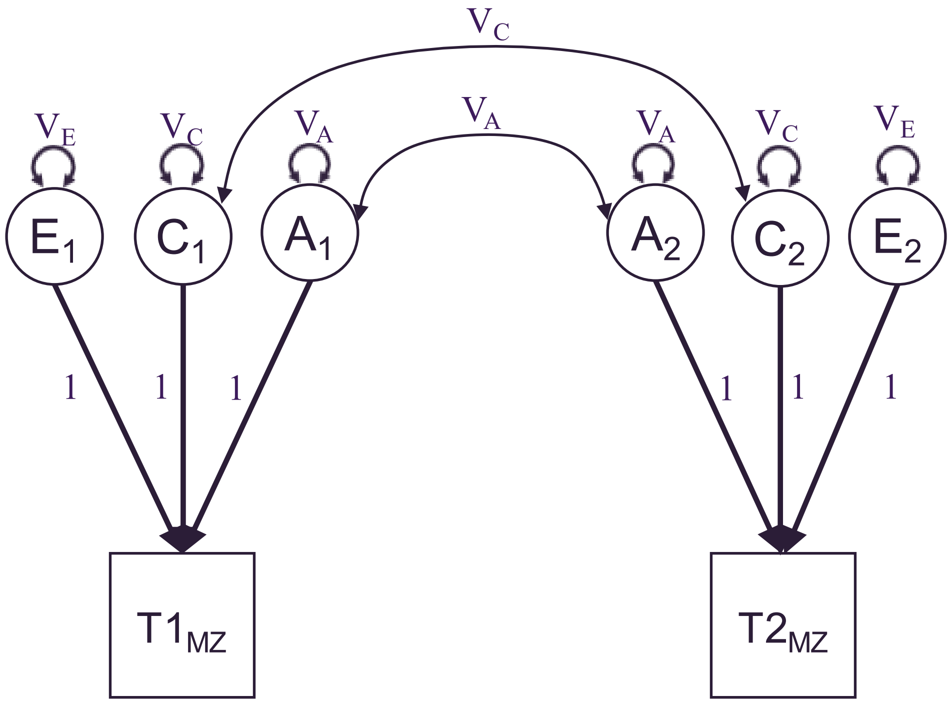
Dizygotic Twins

Observed Covariance Matrices:

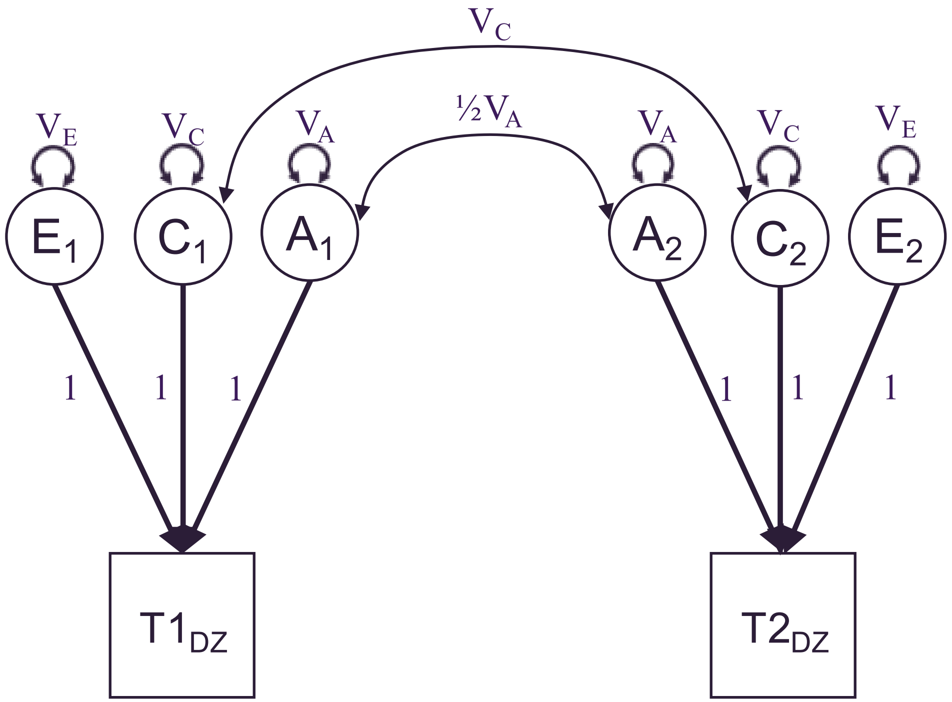
$$S_{MZ} = \begin{array}{cc} \mathrm{VAR}(T1_{MZ}) & \mathrm{COV}(T1_{MZ},T2_{MZ}) \\ \mathrm{COV}(T1_{MZ},T2_{MZ}) & \mathrm{VAR}(T2_{MZ}) \end{array}$$

$$S_{DZ} = \begin{array}{cc} \mathrm{VAR}(T1_{DZ}) & \mathrm{COV}(T1_{DZ},T2_{DZ}) \\ \mathrm{COV}(T1_{DZ},T2_{DZ}) & \mathrm{VAR}(T2_{DZ}) \end{array}$$

Path diagram – Classical Twin Design



Monozygotic Twins



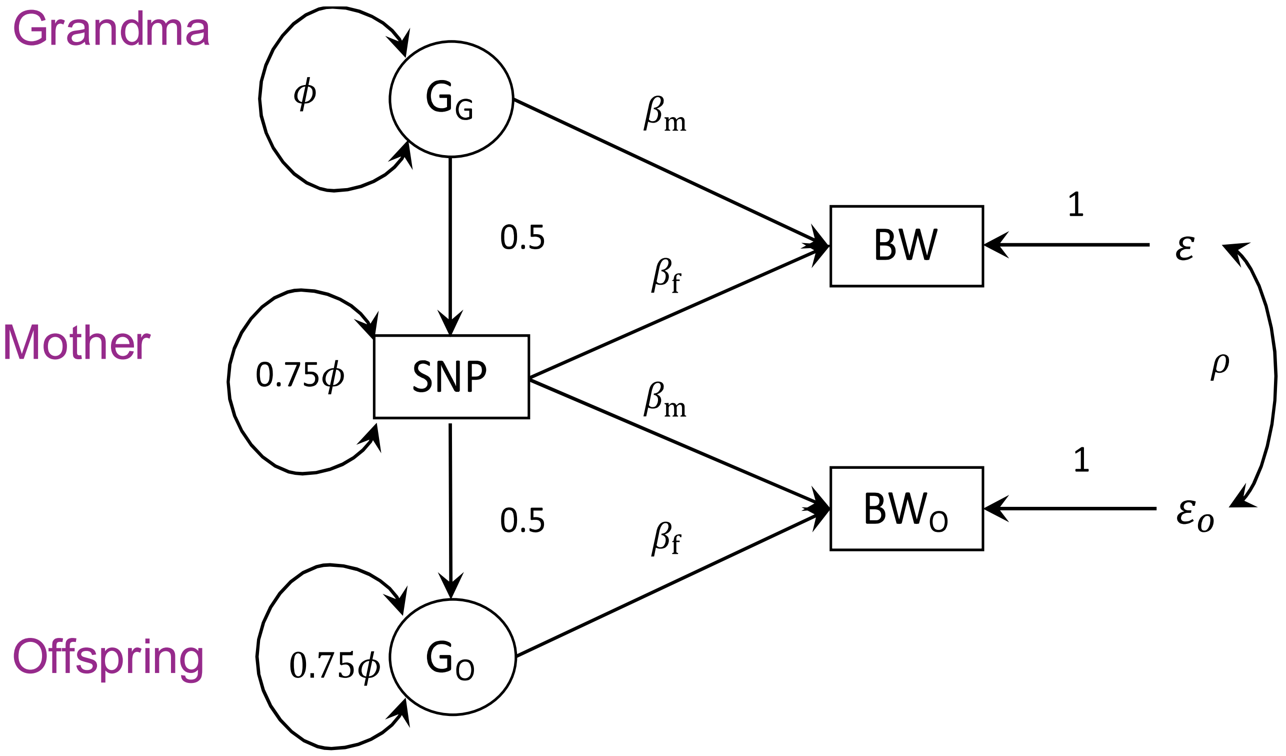
Dizygotic Twins

Expected Covariance Matrices:

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

$$\Sigma_{DZ} = \begin{matrix} V_A+V_C+V_E & \frac{1}{2}V_A+V_C \\ \frac{1}{2}V_A+V_C & V_A+V_C+V_E \end{matrix}$$

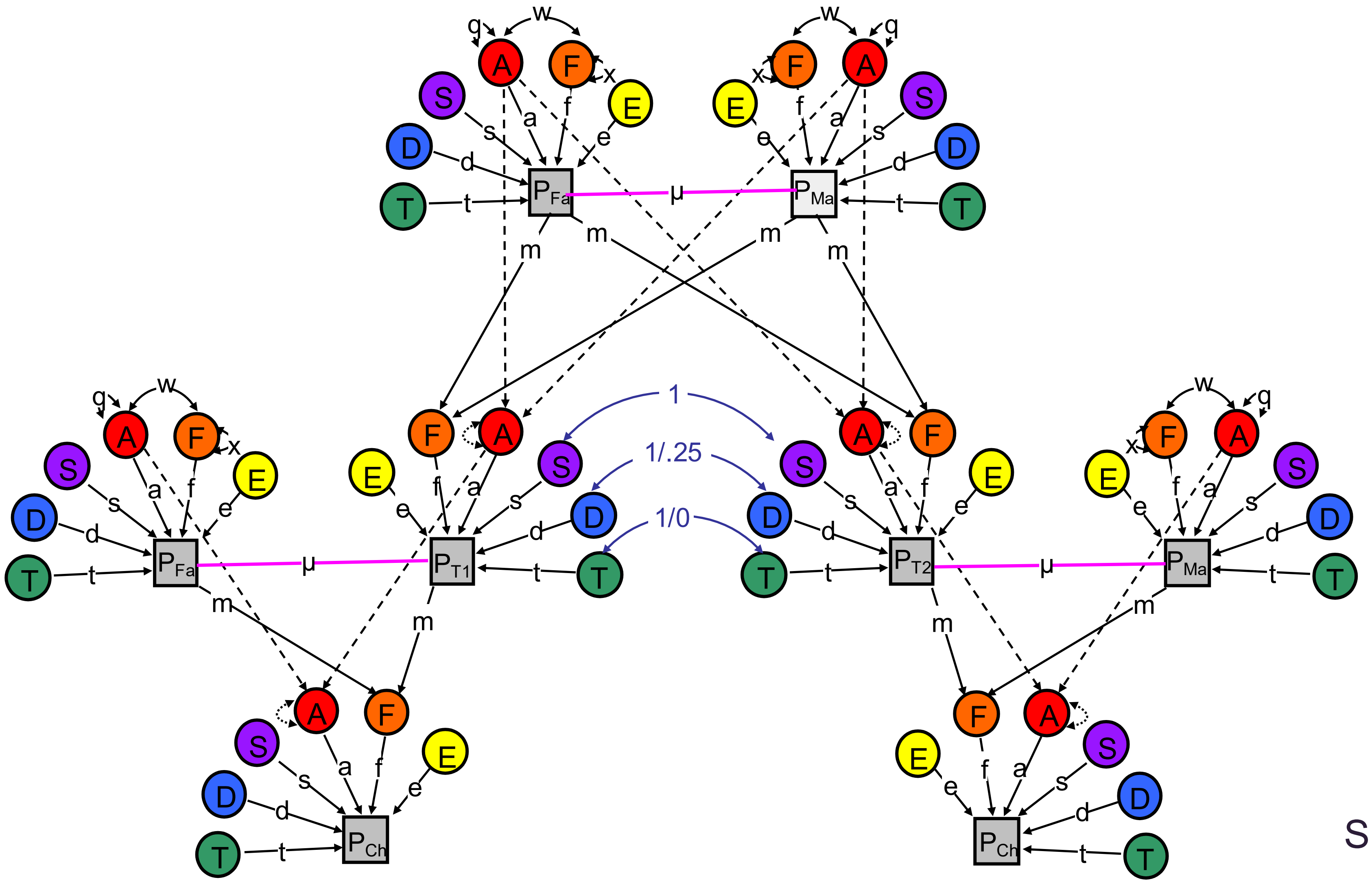
Path diagram – Indirect vs direct genetic effects



BW: Own birth weight
BW_O: Offspring birth weight
SNP: Mother's genotype
G_G: Grandmother's genotype
G_O: Offspring's genotype
β_m maternal genetic effect (indirect)
β_f fetal genetic effect (direct)

Warrington et al., Int J Epidemiol., 2018

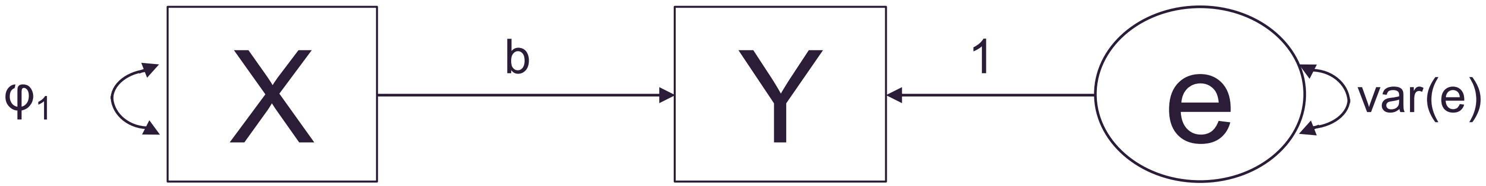
Path diagram – More complicated model



Slide courtesy of Matt Keller

Path Tracing Rules

- **Deriving Expected Variances and Covariances Using Path Tracing Rules**



Observed Covariance Matrix:

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

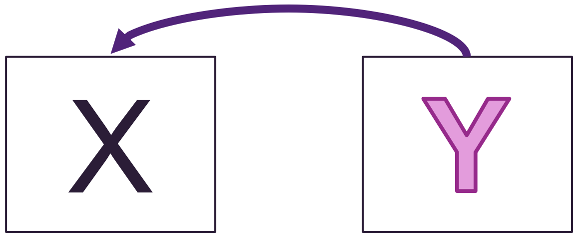
Expected/Implied Covariance Matrix:

$$\Sigma(\theta) = \begin{matrix} \varphi_1 & b\varphi_1 \\ b\varphi_1 & b^2\varphi_1+var(e) \end{matrix}$$

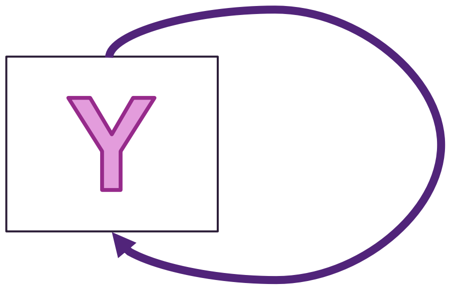
Path Tracing Rules - Deriving variances & covariances

- Identify all legitimate chains (a series of paths) that connect one variable to another (covariances) or connect a variable back to itself (variances)

Covariance: cov(x,y)

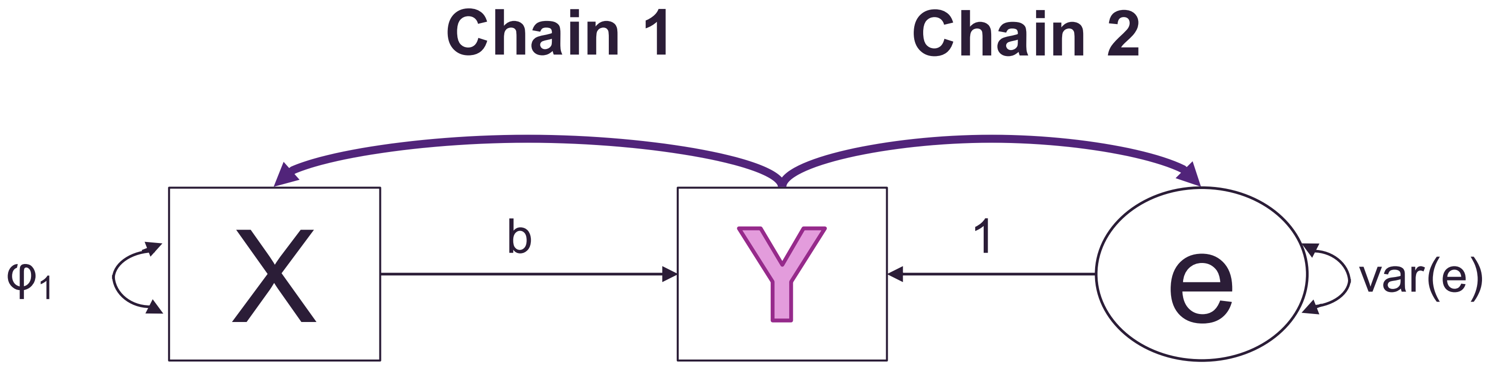


Variance: var(y)



Path Tracing Rules - Deriving variances & covariances

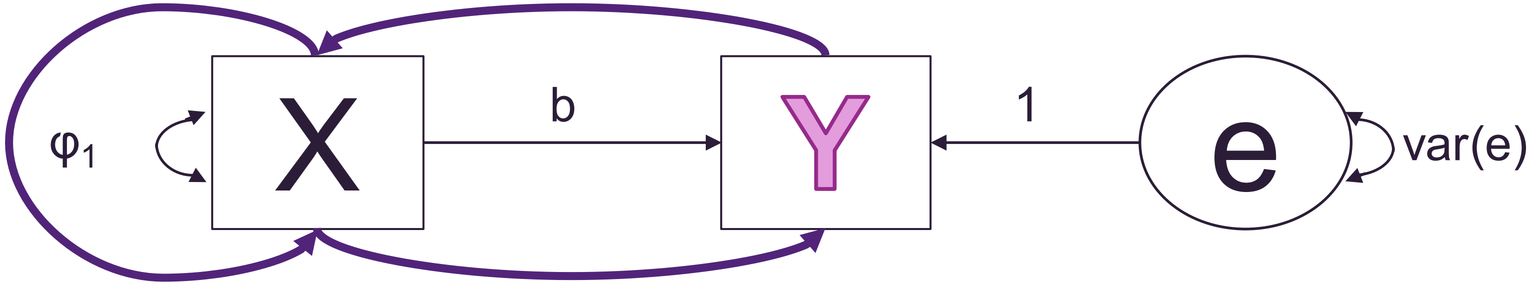
- **All chains begin by travelling backwards against the direction of a (single or double-headed) arrow, head to tail.**
- **e.g. expected variance of Y**



Path Tracing Rules - Deriving variances & covariances

- Once a double-headed arrow has been traversed, the direction reverses such that the chain travels forward.
- e.g. expected variance of Y

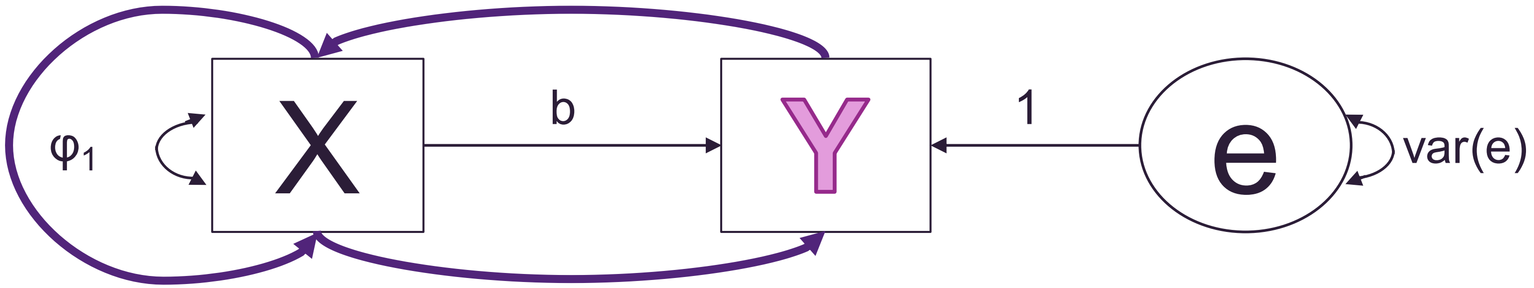
Chain 1



Path Tracing Rules - Deriving variances & covariances

- The expected value of a chain is the product of all coefficients associated with each path making up that chain.
- e.g. expected variance of Y

Chain 1 $b^*\varphi_1{}^*b$

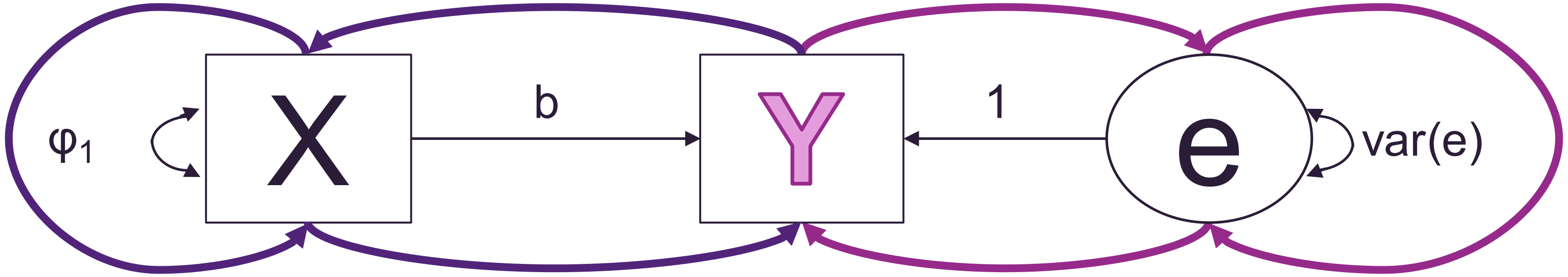


Path Tracing Rules - Deriving variances & covariances

- The final expected variance or covariance equals the sum of the values of all legitimate chains
- e.g. expected variance of Y

$b*\varphi_1*b$ + $1*var(e)*1$

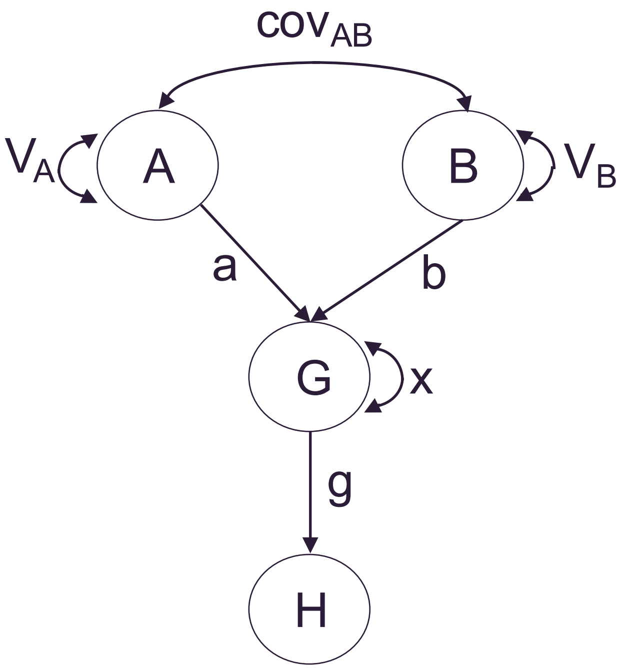
Chain 1 Chain 2



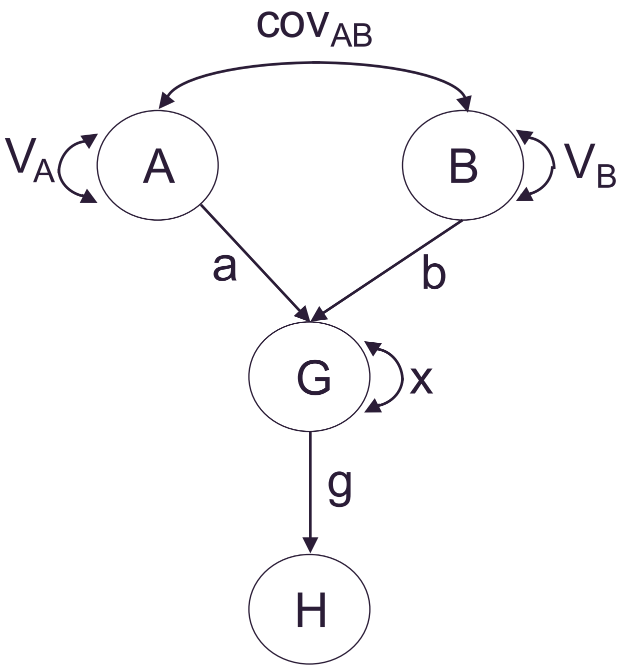
Path Tracing Rules - Deriving variances & covariances

- All chains must include exactly one double-headed arrow. This implies a chain must change directions exactly once.
- All chains must be counted exactly once, and each must be unique.
- However, order matters: *a->b->c* is a distinct chain from *c->b->a*.
- See example in the next slide

Path Tracing Rules - Deriving variances & covariances



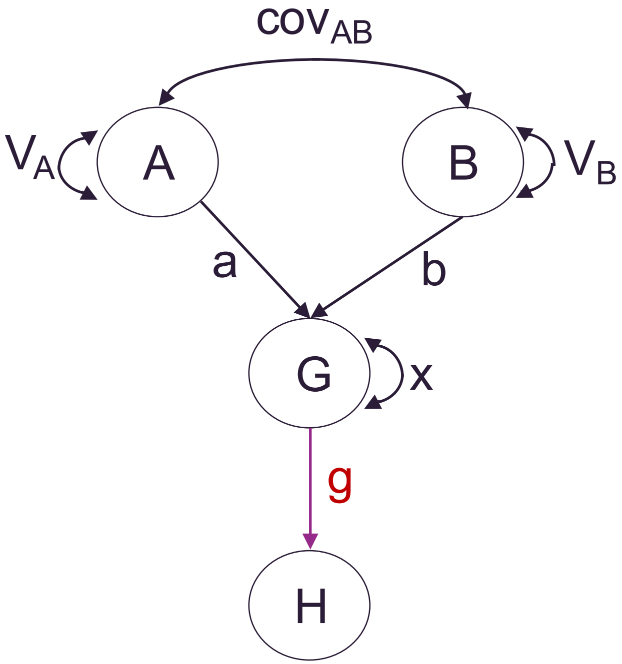
Path Tracing Rules - Deriving variances & covariances



COV(H,A) =

Path Tracing Rules - Deriving variances & covariances

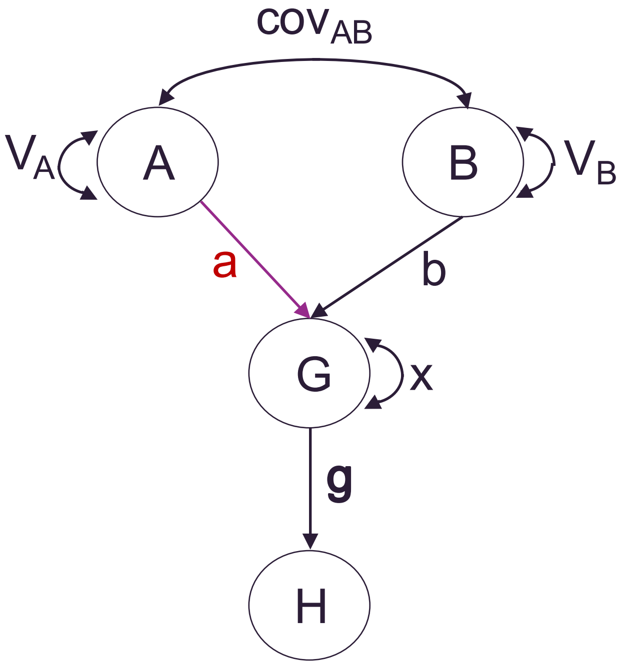
COV(H,A) = g



- **Identify all legitimate chains (a series of paths) that connect one variable to another (covariances) or connect a variable back to itself (variances)**
- **All chains begin by travelling backwards against the direction of a (single or double-headed) arrow, head to tail.**
- The expected value of a chain is the product of all coefficients associated with each path making up that chain
- Once a double-headed arrow has been traversed, the direction reverses such that the chain travels forward
- The final expected variance or covariance equals the sum of the values of all legitimate chains
- All chains must include exactly one double-headed arrow. This implies a chain must change directions exactly once.
- All chains must be counted exactly once, and each must be unique. However, order matters: *a->b->c* is a distinct chain from *c->b->a*.

Path Tracing Rules - Deriving variances & covariances

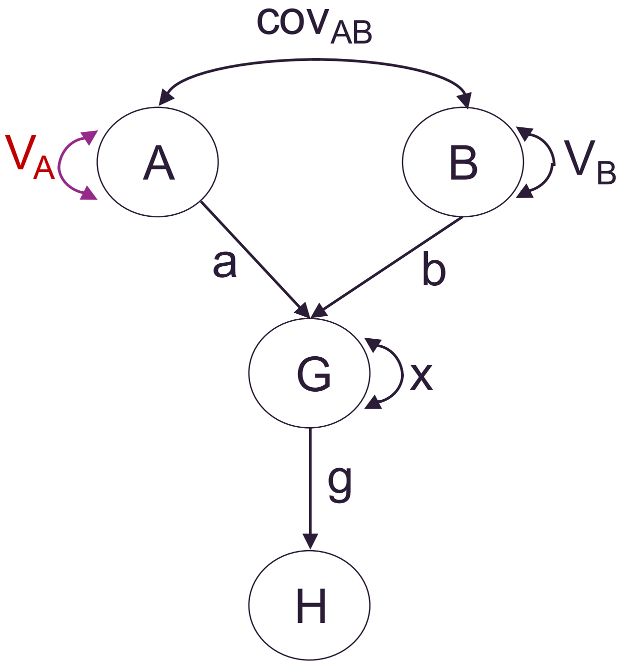
$$COV(H,A) = \mathbf{g} * a$$



- **Identify all legitimate chains (a series of paths) that connect one variable to another (covariances) or connect a variable back to itself (variances)**
- **All chains begin by travelling backwards against the direction of a (single or double-headed) arrow, head to tail.**
- **The expected value of a chain is the product of all coefficients associated with each path making up that chain**
- **Once a double-headed arrow has been traversed, the direction reverses such that the chain travels forward**
- **The final expected variance or covariance equals the sum of the values of all legitimate chains**
- **All chains must include exactly one double-headed arrow. This implies a chain must change directions exactly once.**
- **All chains must be counted exactly once, and each must be unique. However, order matters: *a->b->c* is a distinct chain from *c->b->a*.**

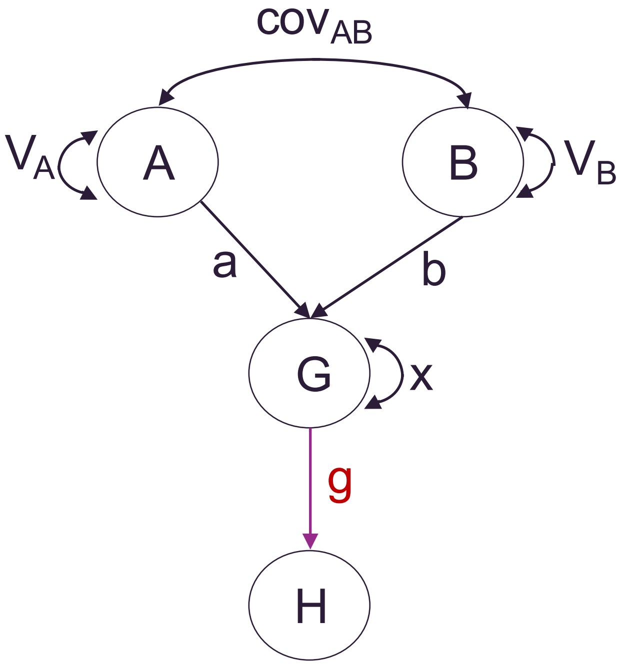
Path Tracing Rules - Deriving variances & covariances

$\mathrm{COV(H,A)} = \mathbf{g} * \mathbf{a} * V_A$



- **Identify all legitimate chains (a series of paths) that connect one variable to another (covariances) or connect a variable back to itself (variances)**
- **All chains begin by travelling backwards against the direction of a (single or double-headed) arrow, head to tail.**
- **The expected value of a chain is the product of all coefficients associated with each path making up that chain.**
- **Once a double-headed arrow has been traversed, the direction reverses such that the chain travels forward**
- **The final expected variance or covariance equals the sum of the values of all legitimate chains**
- **All chains must include exactly one double-headed arrow. This implies a chain must change directions exactly once.**
- **All chains must be counted exactly once, and each must be unique. However, order matters: *a->b->c* is a distinct chain from *c->b->a*.**

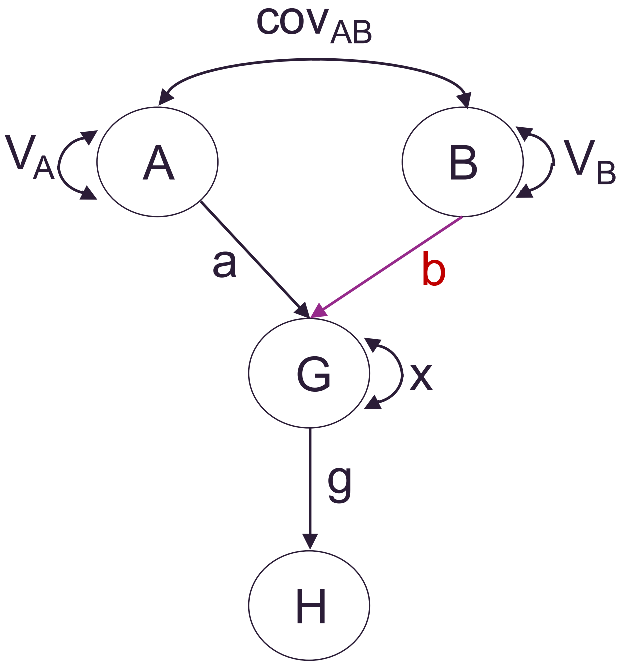
Path Tracing Rules - Deriving variances & covariances



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

- Identify all legitimate chains (a series of paths) that connect one variable to another (covariances) or connect a variable back to itself (variances)
- All chains begin by travelling backwards against the direction of a (single or double-headed) arrow, head to tail.
- The expected value of a chain is the product of all coefficients associated with each path making up that chain.
- Once a double-headed arrow has been traversed, the direction reverses such that the chain travels forward
- The final expected variance or covariance equals the sum of the values of all legitimate chains
- All chains must include exactly one double-headed arrow. This implies a chain must change directions exactly once.
- All chains must be counted exactly once, and each must be unique. However, order matters: *a->b->c* is a distinct chain from *c->b->a*.

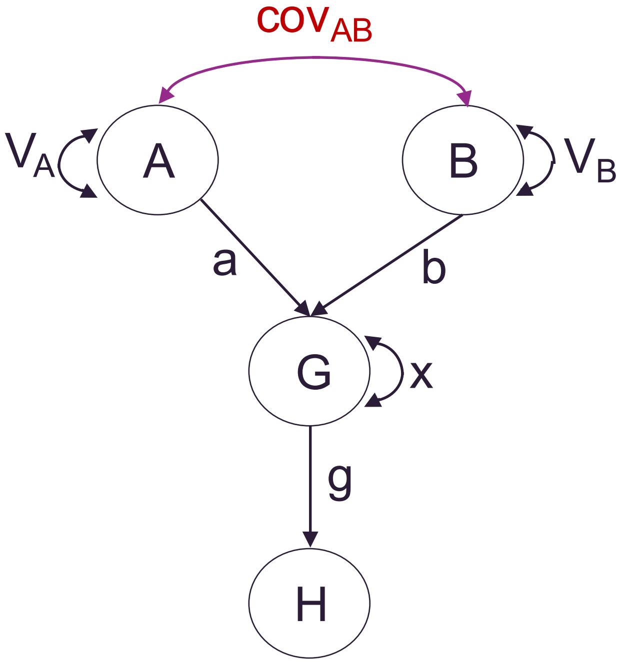
Path Tracing Rules - Deriving variances & covariances



$$COV(H,A) = \mathbf{g * a * V_A} \\ \mathbf{+ g * b}$$

- **Identify all legitimate chains (a series of paths) that connect one variable to another (covariances) or connect a variable back to itself (variances)**
- **All chains begin by travelling backwards against the direction of a (single or double-headed) arrow, head to tail.**
- **The expected value of a chain is the product of all coefficients associated with each path making up that chain.**
- **Once a double-headed arrow has been traversed, the direction reverses such that the chain travels forward**
- **The final expected variance or covariance equals the sum of the values of all legitimate chains**
- **All chains must include exactly one double-headed arrow. This implies a chain must change directions exactly once.**
- **All chains must be counted exactly once, and each must be unique. However, order matters: *a->b->c* is a distinct chain from *c->b->a*.**

Path Tracing Rules - Deriving variances & covariances

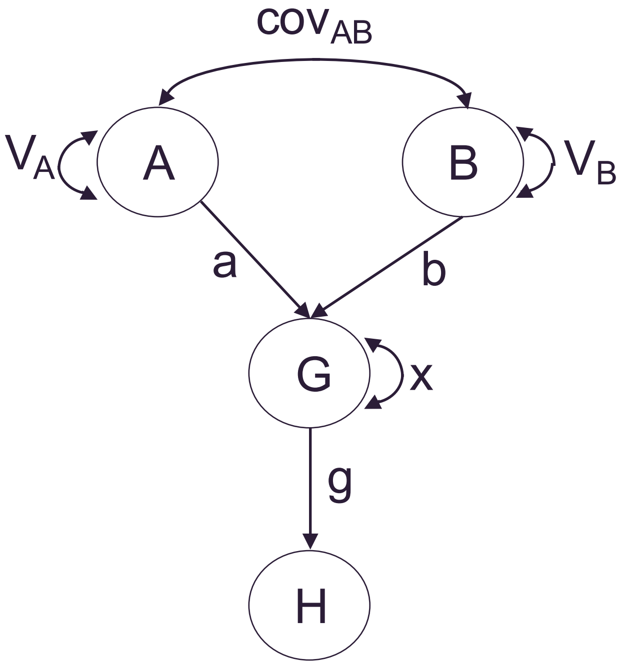


$$COV(H,A) = g * a * V_A + g * b * COV_{AB}$$

- **Identify all legitimate chains (a series of paths) that connect one variable to another (covariances) or connect a variable back to itself (variances)**
- **All chains begin by travelling backwards against the direction of a (single or double-headed) arrow, head to tail.**
- **The expected value of a chain is the product of all coefficients associated with each path making up that chain.**
- **Once a double-headed arrow has been traversed, the direction reverses such that the chain travels forward**
- **The final expected variance or covariance equals the sum of the values of all legitimate chains**
- **All chains must include exactly one double-headed arrow. This implies a chain must change directions exactly once.**
- **All chains must be counted exactly once, and each must be unique. However, order matters: *a->b->c* is a distinct chain from *c->b->a*.**

Path Tracing Rules - Deriving variances & covariances

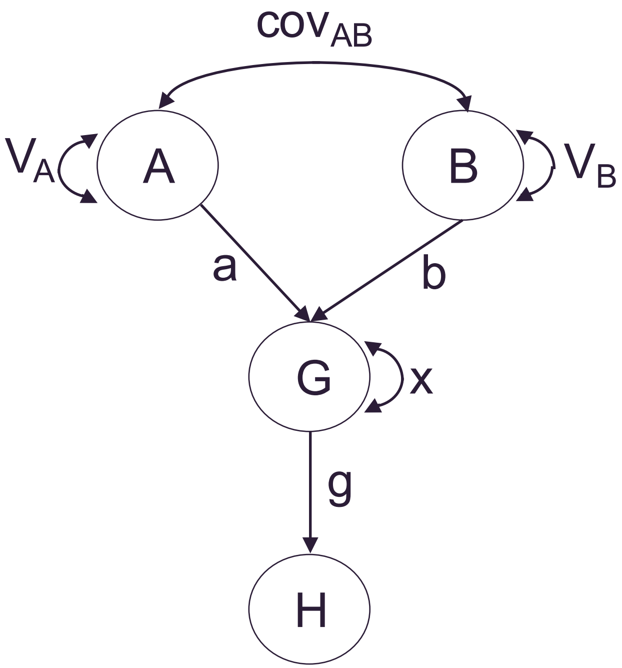
COV(H,A) = $g * a * V_A + g * b * COV_{AB}$



- Identify all legitimate chains (a series of paths) that connect one variable to another (covariances) or connect a variable back to itself (variances)
- All chains begin by travelling backwards against the direction of a (single or double-headed) arrow, head to tail.
- The expected value of a chain is the product of all coefficients associated with each path making up that chain.
- Once a double-headed arrow has been traversed, the direction reverses such that the chain travels forward
- The final expected variance or covariance equals the sum of the values of all legitimate chains
- All chains must include exactly one double-headed arrow. This implies a chain must change directions exactly once.
- All chains must be counted exactly once, and each must be unique. However, order matters: *a->b->c* is a distinct chain from *c->b->a.*

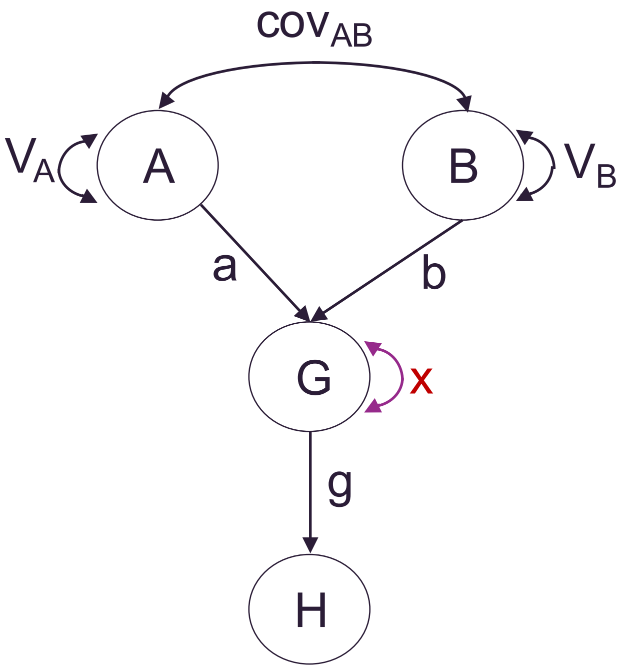
Path Tracing Rules - Deriving variances & covariances

VAR(G) =

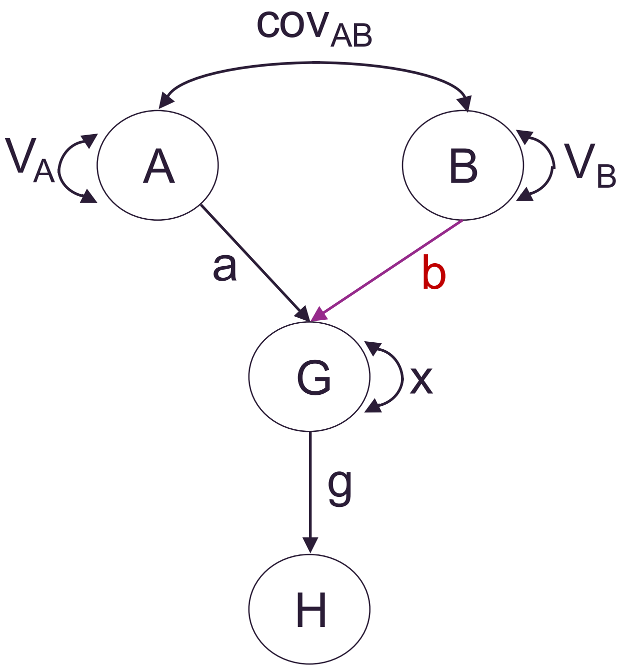


Path Tracing Rules - Deriving variances & covariances

VAR(G) = x



Path Tracing Rules - Deriving variances & covariances

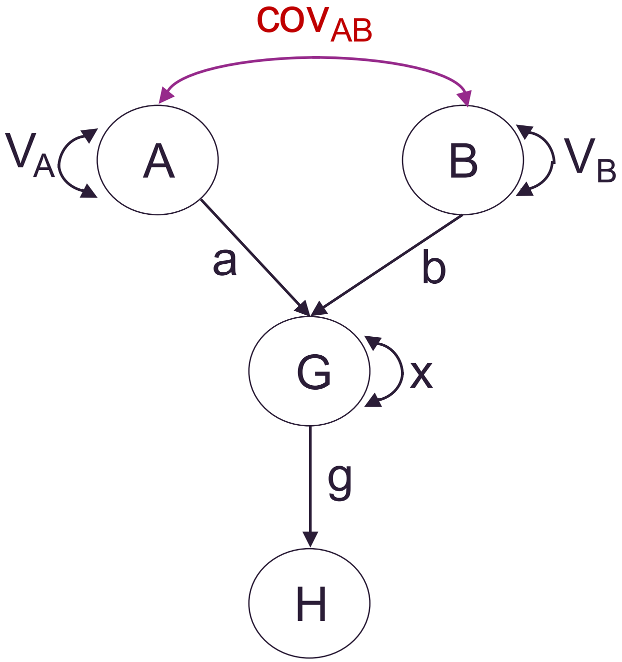


VAR(G) = x

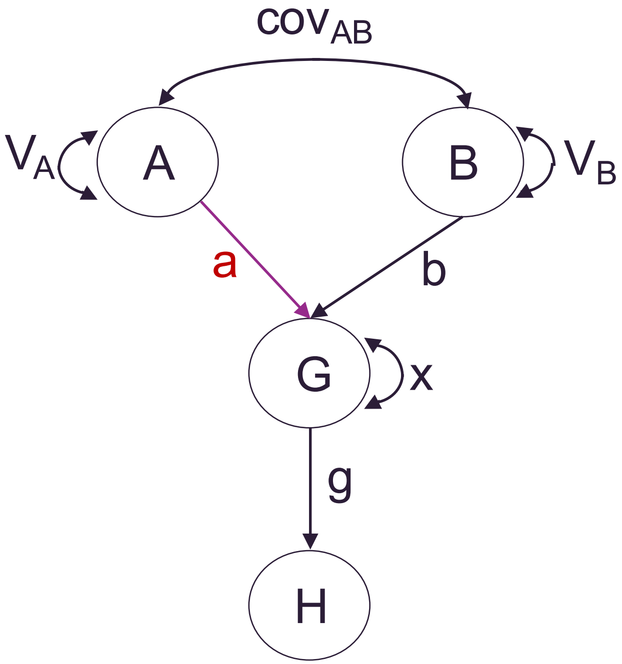
+ b

Path Tracing Rules - Deriving variances & covariances

$$\mathrm{VAR}(G) = x \\ + b * \mathrm{COV}_{AB}$$

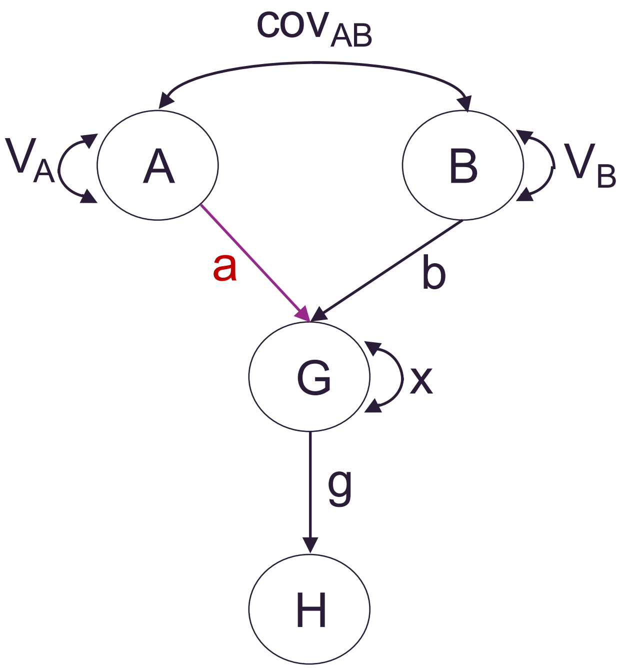


Path Tracing Rules - Deriving variances & covariances



$$\begin{aligned} VAR(G) &= x \\ &\quad + b * COV_{AB} * a \end{aligned}$$

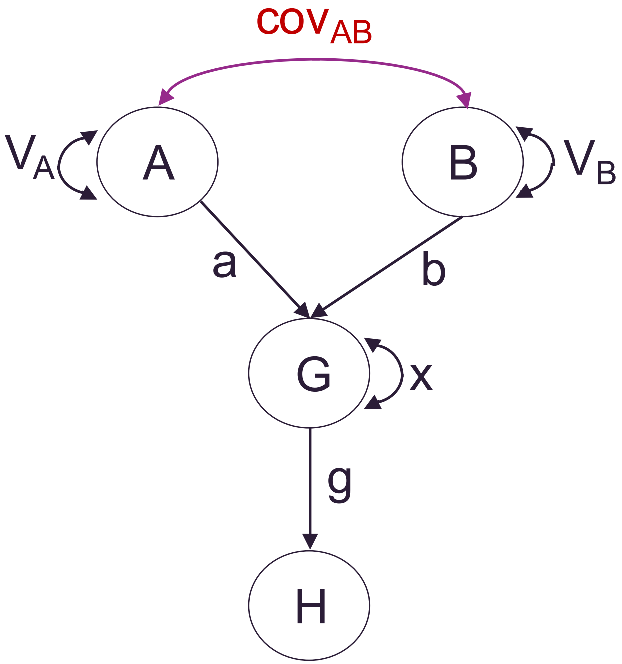
Path Tracing Rules - Deriving variances & covariances



$$\begin{aligned} VAR(G) = {} & x \\ & + b * COV_{AB} * a \\ & + a \end{aligned}$$

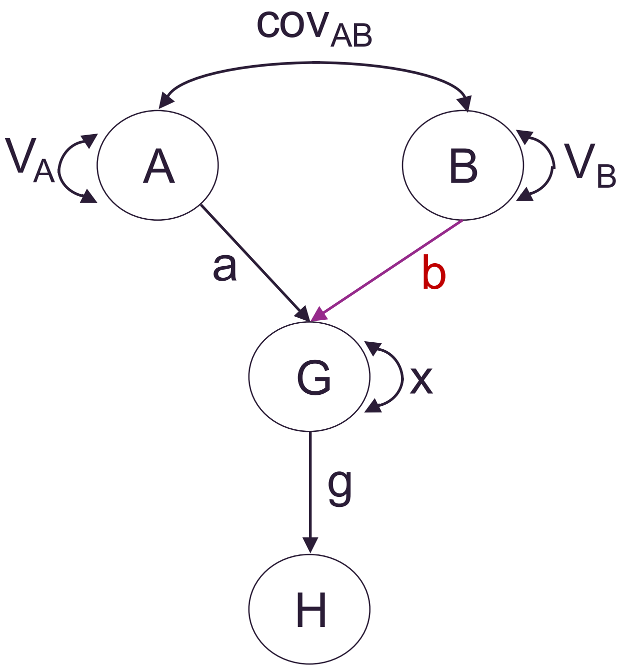
- **All chains must be counted exactly once, and each must be unique. However, order matters: *a->b->c* is a distinct chain from *c->b->a*.**

Path Tracing Rules - Deriving variances & covariances



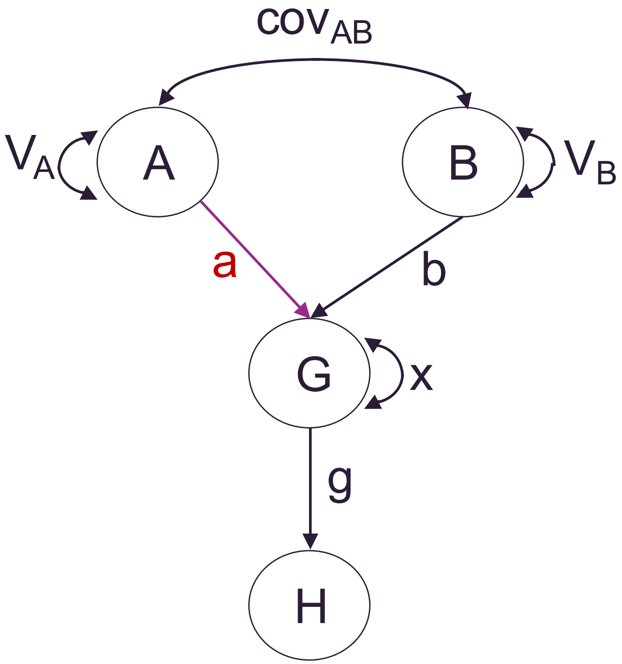
$$\begin{aligned} VAR(G) = {} & x \\ & + b * COV_{AB} * a \\ & + a * COV_{AB} \end{aligned}$$

Path Tracing Rules - Deriving variances & covariances



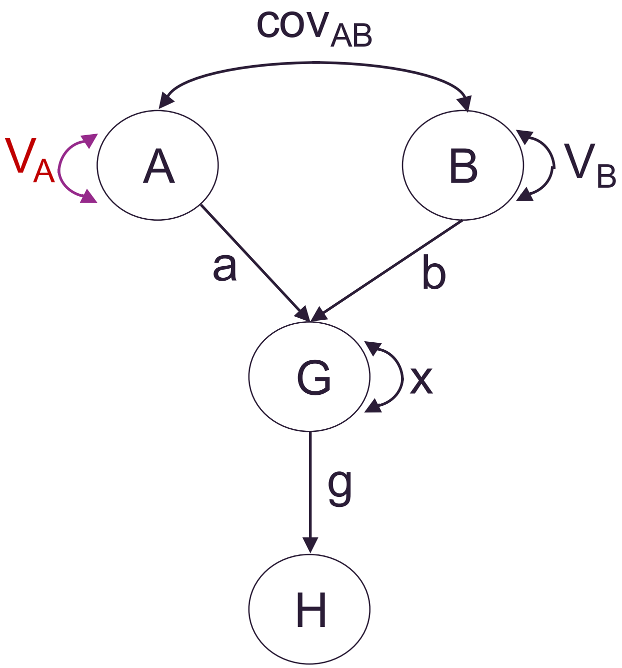
$$\begin{aligned} VAR(G) = &\ x \\ &+ b * COV_{AB} * a \\ &+ a * COV_{AB} * b \end{aligned}$$

Path Tracing Rules - Deriving variances & covariances



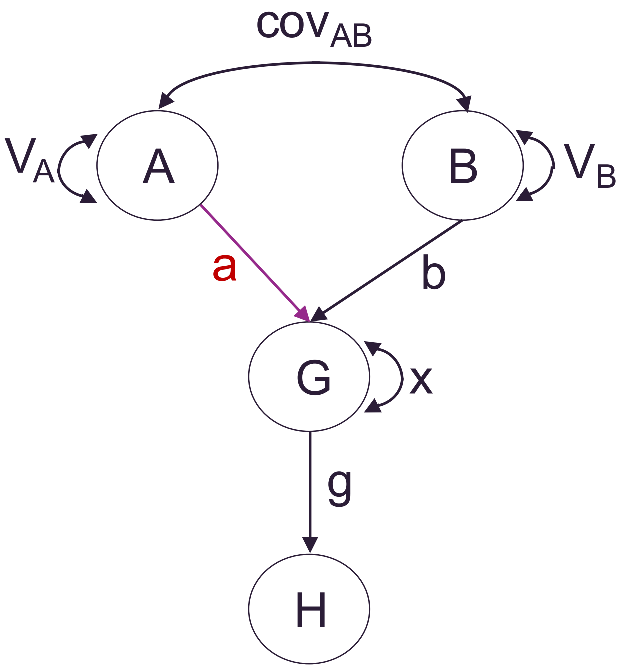
$$\begin{aligned} VAR(G) = {} & x \\ & + b * COV_{AB} * a \\ & + a * COV_{AB} * b \\ & + a \end{aligned}$$

Path Tracing Rules - Deriving variances & covariances



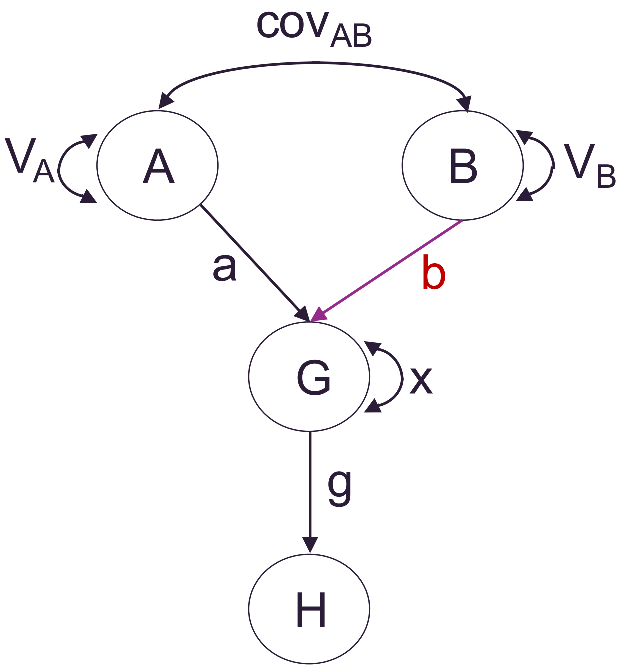
$$
\begin{aligned}
VAR(G) = x \\
&+ b * COV_{AB} * a \\
&+ a * COV_{AB} * b \\
&+ a * V_A
\end{aligned}
$$

Path Tracing Rules - Deriving variances & covariances



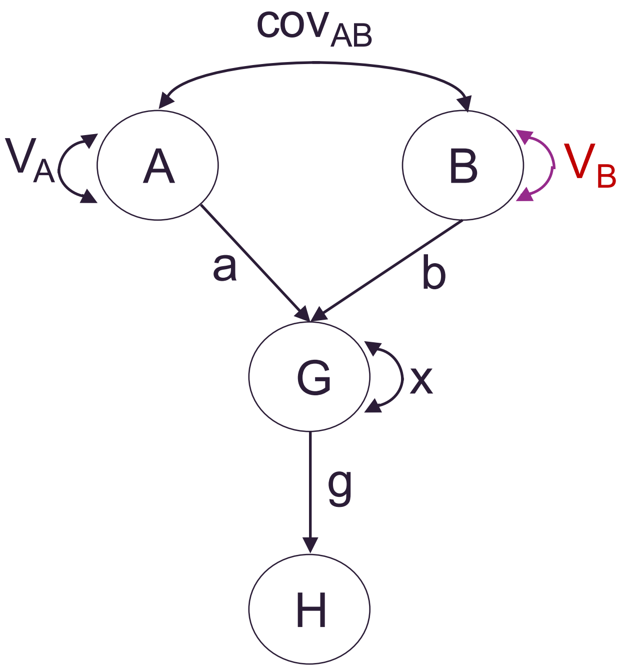
$VAR(G) = x$
$+ b * COV_{AB} * a$
$+ a * COV_{AB} * b$
$+ a * V_A * a$

Path Tracing Rules - Deriving variances & covariances



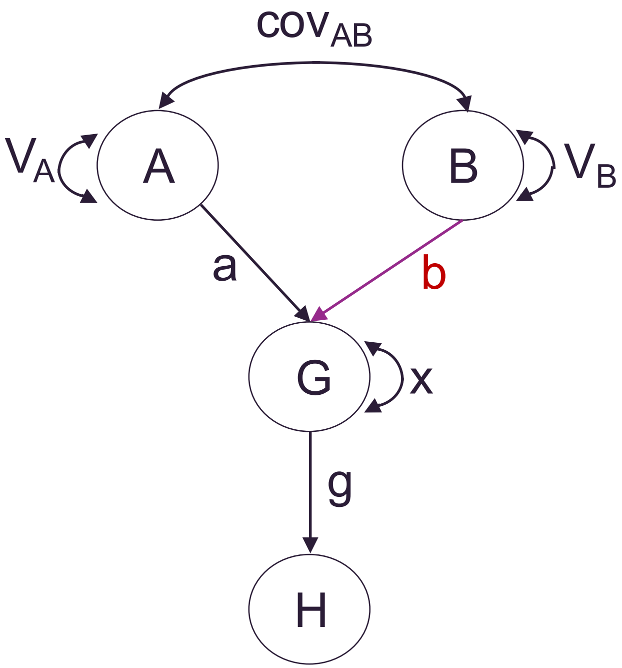
$VAR(G) = x$
$+ b * COV_{AB} * a$
$+ a * COV_{AB} * b$
$+ a * V_A * a$
$+ b$

Path Tracing Rules - Deriving variances & covariances



$$
\begin{aligned}
VAR(G) = {} & x \\
& + b * COV_{AB} * a \\
& + a * COV_{AB} * b \\
& + a * V_A * a \\
& + b * V_B
\end{aligned}
$$

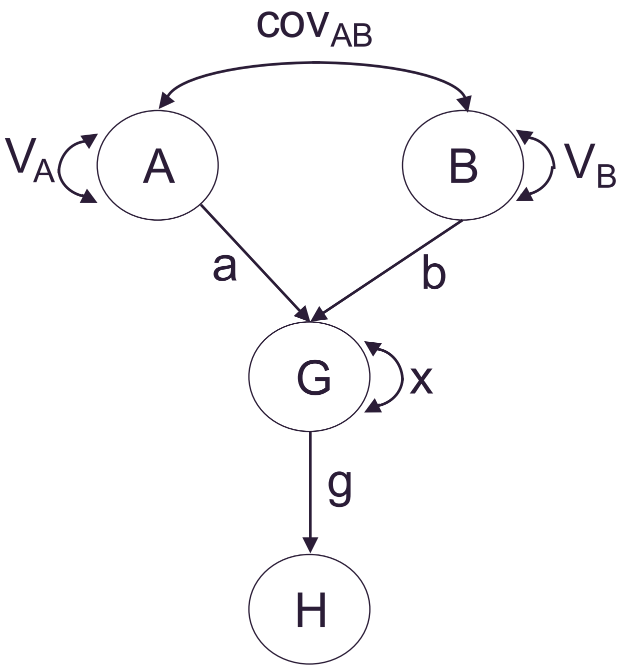
Path Tracing Rules - Deriving variances & covariances



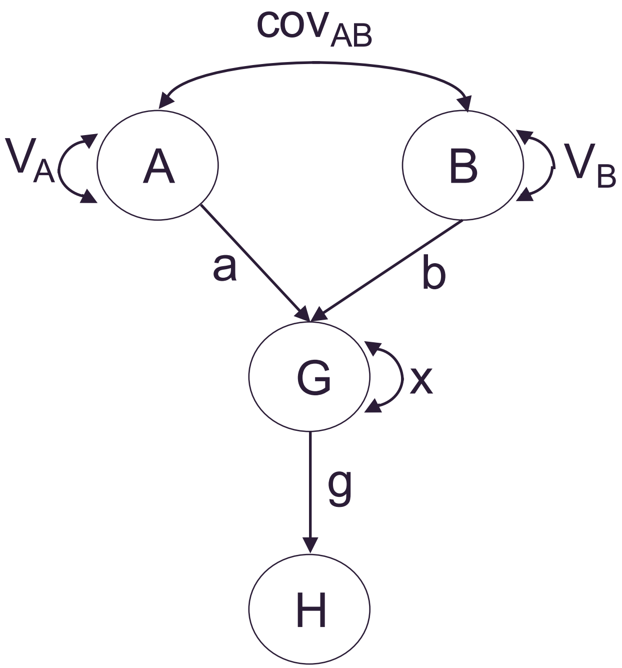
$$
\begin{aligned}
VAR(G) = x \\
&+ b * COV_{AB} * a \\
&+ a * COV_{AB} * b \\
&+ a * V_A * a \\
&+ b * V_B * b
\end{aligned}
$$

Path Tracing Rules - Deriving variances & covariances

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



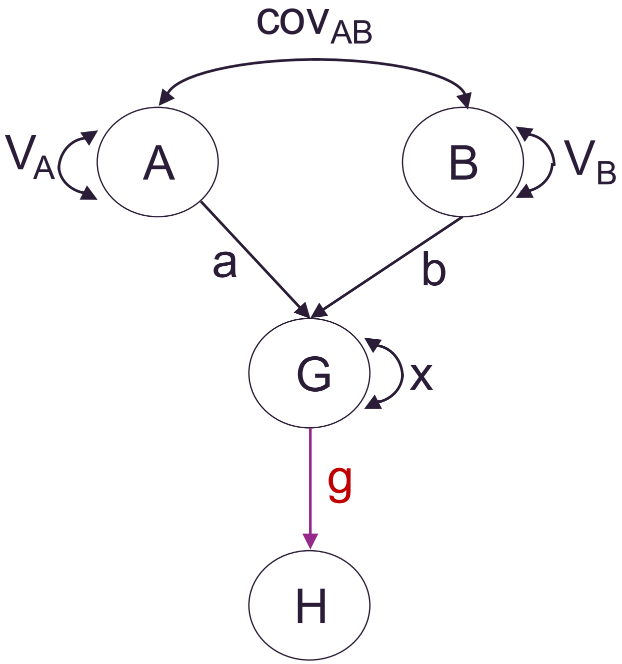
Path Tracing Rules - Deriving variances & covariances



VAR(H) =

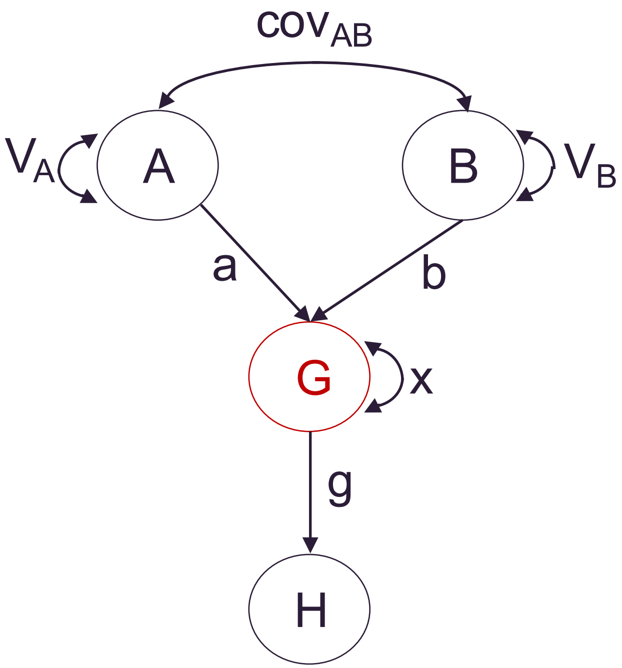
Path Tracing Rules - Deriving variances & covariances

VAR(H) = **g**



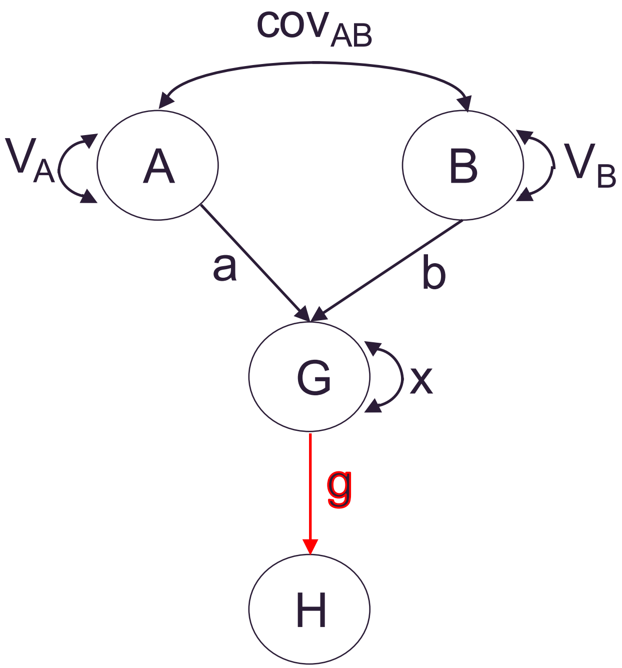
Path Tracing Rules - Deriving variances & covariances

$\mathrm{VAR}(H) = \mathbf{g} * \mathrm{var}(G)$

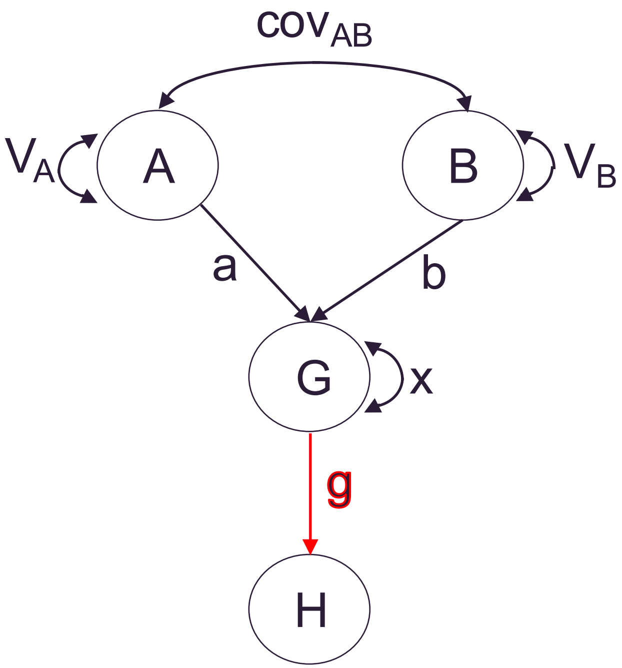


Path Tracing Rules - Deriving variances & covariances

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



Path Tracing Rules - Deriving variances & covariances



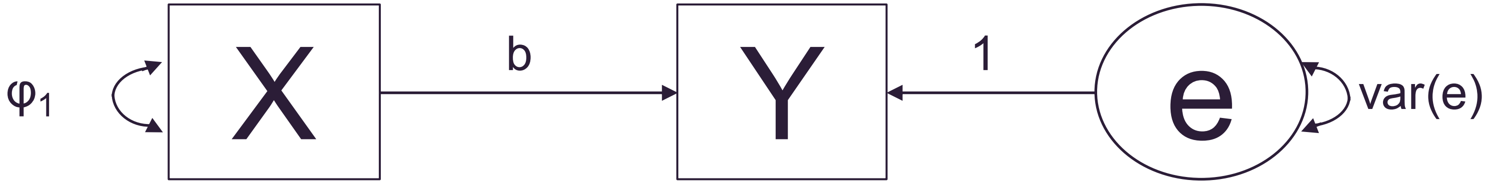
VAR(H) = $\mathbf{g^2 * var(G)}$

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

You can also derive expected variances and covariances using covariance algebra!

(Feel free to try it after the class, see slides at the end).

SEM basics – Simple example (Univariate Regression)



Observed Covariance Matrix:

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

Number of observed variables: 2
Number of observed statistics: 3 (var(X), cov(X,Y), var(Y))

Expected/Implied Covariance Matrix:

$$\Sigma(\theta) = \begin{matrix} \varphi_1 & b\varphi_1 \\ b\varphi_1 & b^2\varphi_1+\text{var(e)} \end{matrix}$$

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

Take home messages – Part I

- **Structural Equation Modeling (SEM)** is a statistical method that analyzes relationships between observed and latent variables.
- **Path diagram** is a visual representation of an SEM, which is usually found more intuitive than collections of structural equations, especially as the models grow complicated.
- **Path tracing rules** are useful for deriving variance/covariance.
- Fitting data to an SEM model involves adjusting model parameters to minimize the difference between the observed covariance matrix and the expected covariance matrix.

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.

Content

- **Part I (9:00-10:00 am)**
  - SEM basics
  - Path diagrams

Short break (5mins)

- **Part II (10:05 -10:30 am)**
  - Genomic SEM
  - Q&A (5-10mins)

Content

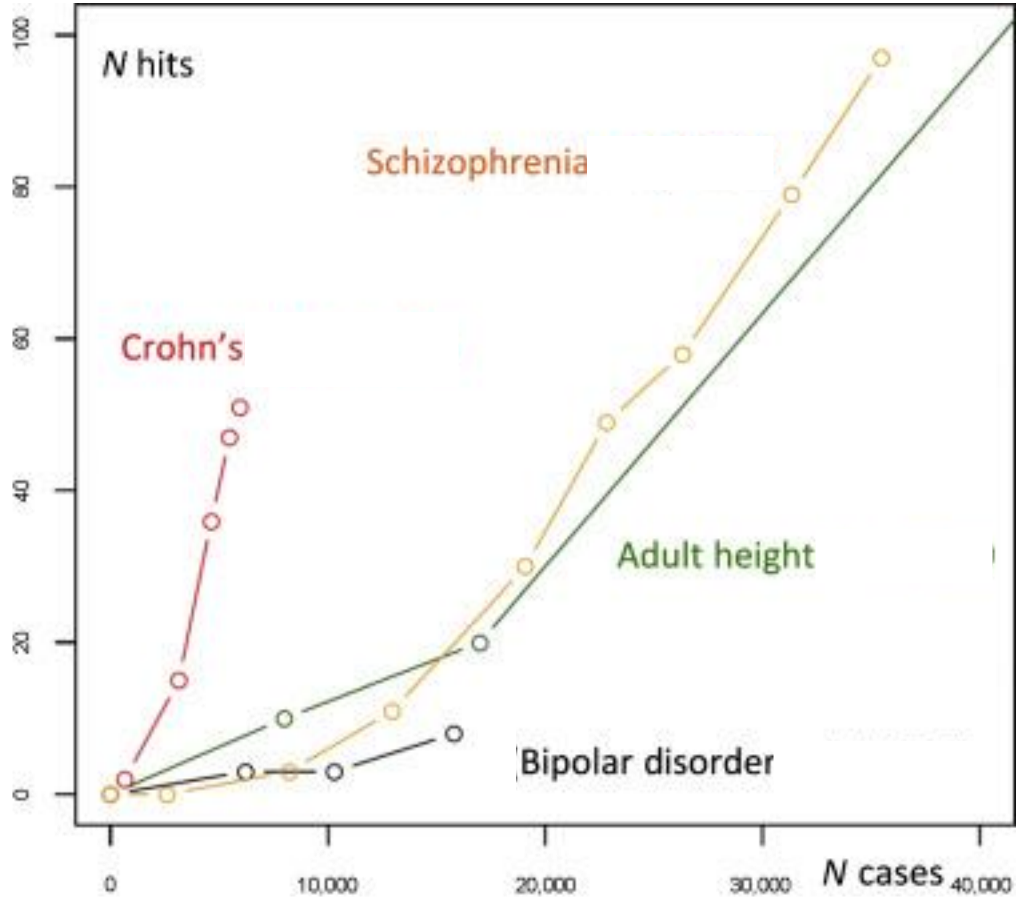
- **Part I (9:00-10:00 am)**
  - SEM basics
  - Path diagrams

Short break (5mins)

- **Part II (10:05 -10:30 am)**
  - Genomic SEM
  - Q&A (5-10mins)

Genomic SEM – Why Genomic SEM?

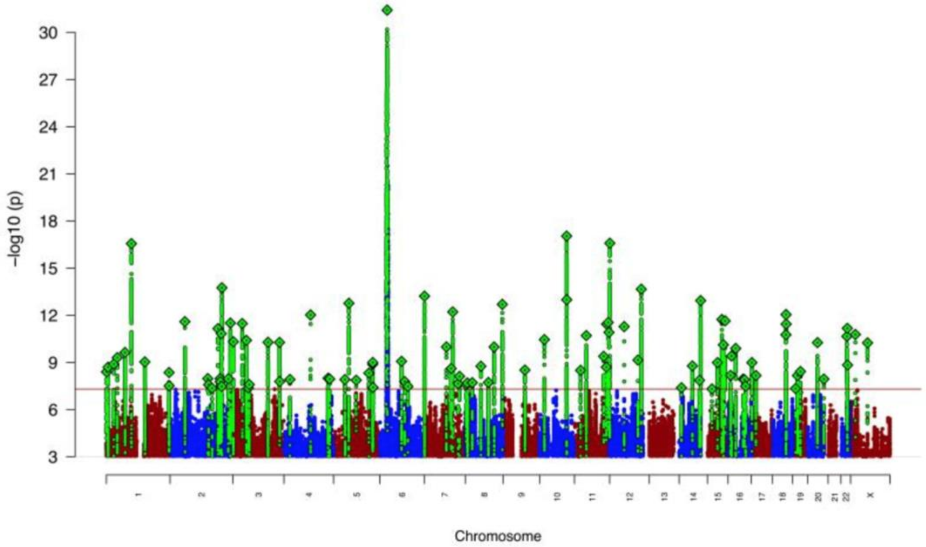
- Human complex traits/diseases are associated with **many** genes



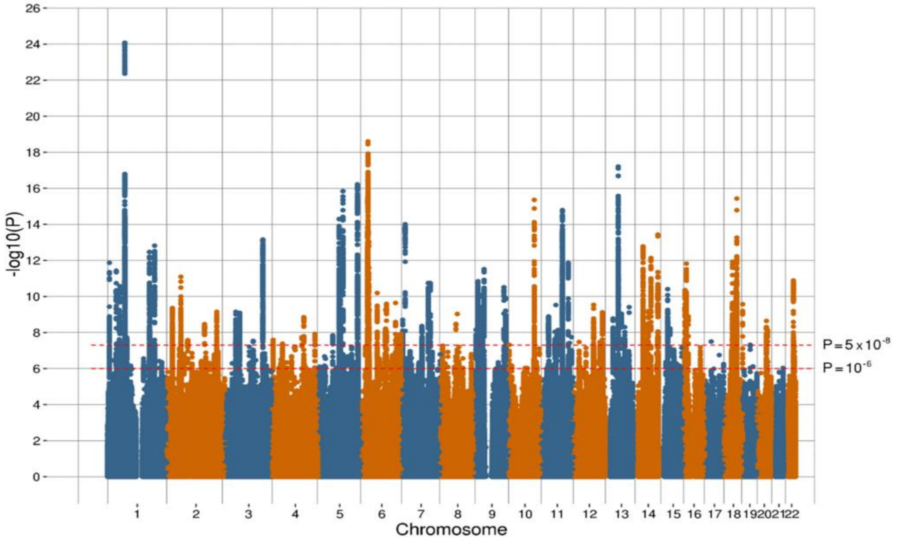
S. Cichon, S. Ripke, 2016

Genomic SEM – Why Genomic SEM?

Schizophrenia



Depression

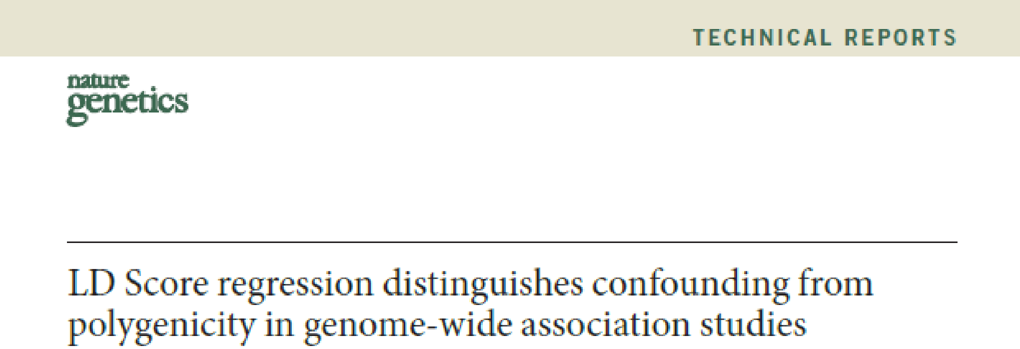


Traits are highly polygenic, so not simply a matter of identifying ~5 overlapping genes

Slide courtesy of Andrew Grotzinger

Genomic SEM – LD score regression (LDSC)

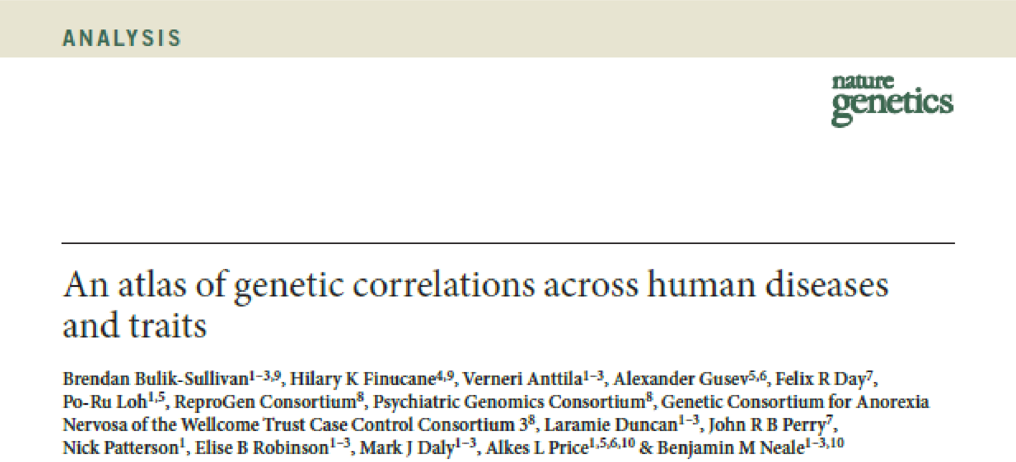
Estimates genetic correlations between samples with varying degrees of sample overlap using publicly available data



TECHNICAL REPORTS

nature genetics

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies



ANALYSIS

nature genetics

An atlas of genetic correlations across human diseases and traits

Brendan Bulik-Sullivan[1–3,9], Hilary K Finucane[4,9], Verneri Anttila[1–3], Alexander Gusev[5,6], Felix R Day[7], Po-Ru Loh[1,5], ReproGen Consortium[8], Psychiatric Genomics Consortium[8], Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3[8], Laramie Duncan[1–3], John R B Perry[7], Nick Patterson[1], Elise B Robinson[1–3], Mark J Daly[1–3], Alkes L Price[1,5,6,10] & Benjamin M Neale[1–3,10]

- To estimate **SNP Heritability**:
  - Regress GWAS test statistic against LD Scores for all SNPs (not just significant ones)
- To estimate **Genetic Correlation**:
  - Regress product of GWAS test statistics for two different phenotypes against LD Scores

Genomic SEM – Why Genomic SEM?

Analysis of shared heritability in common disorders of the brain

The Brainstorm Consortium*†

Pervasive (Statistical) Pleiotropy Necessitates Methods for Analyzing Joint Genetic Architecture

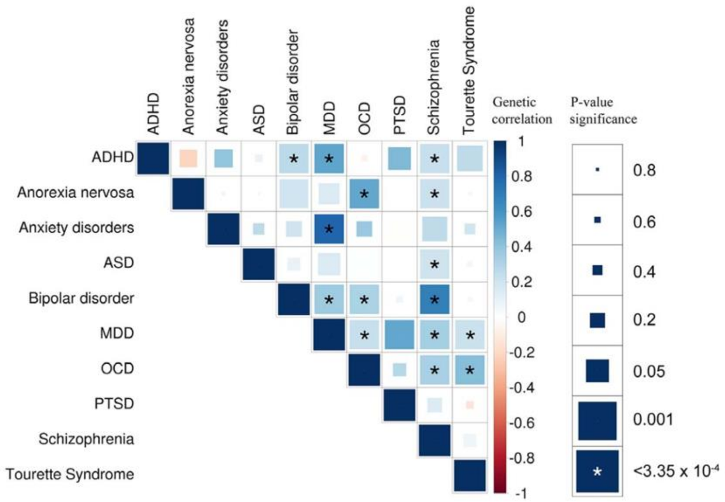


Fig. 1. Genetic correlations across psychiatric phenotypes. The color of each box indicates the magnitude of the correlation, and the size of the box indicates its significance (LDSC), with significant correlations filling each square completely. Asterisks indicate genetic correlations that are significantly different from zero after Bonferroni correction.

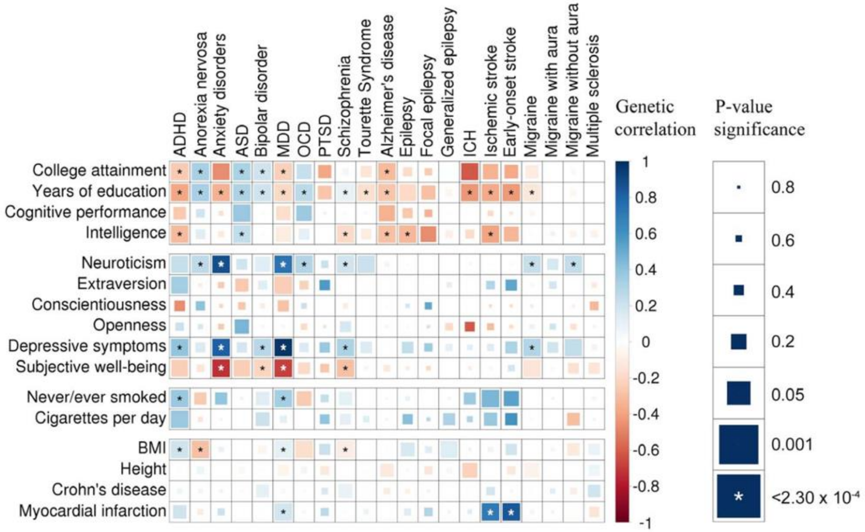
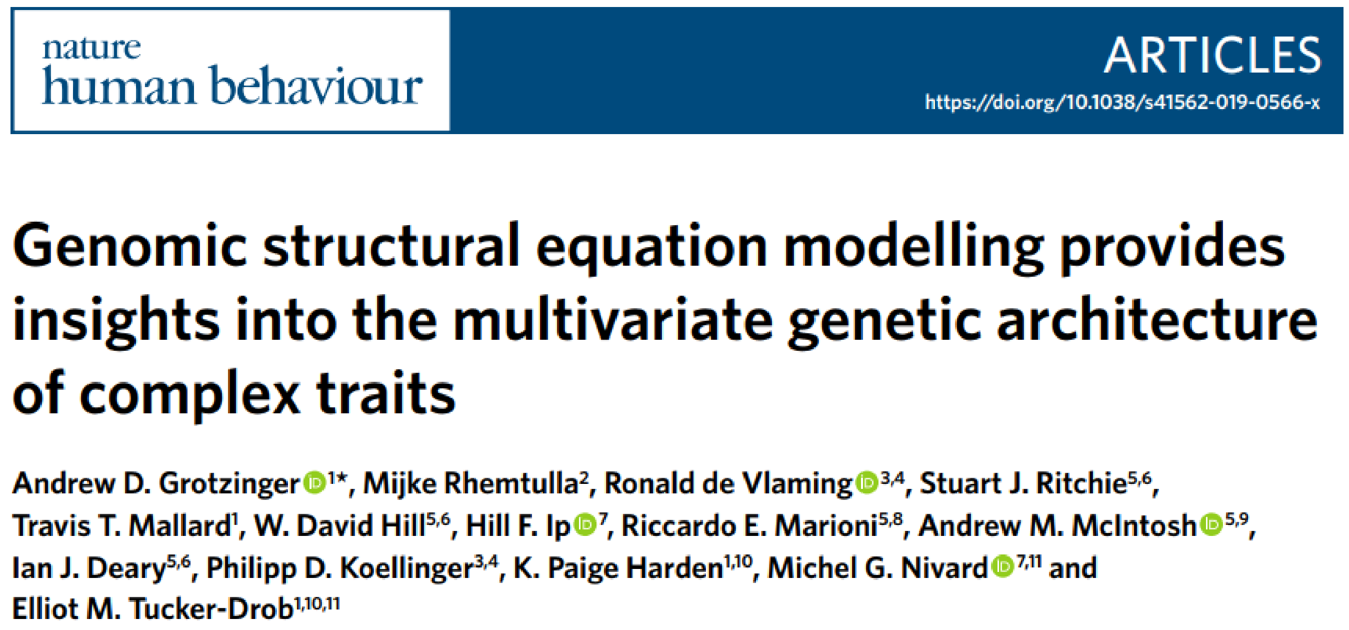


Fig. 4. Genetic correlations across brain disorders and behavioral-cognitive phenotypes. The color of each box indicates the magnitude of the correlation, and the size of the box indicates its significance (LDSC), with significant correlations filling each square completely. Asterisks indicate genetic correlations that are significantly different from zero after Bonferroni correction.

The Brainstorm Consortium, *Science* **360**, 1313 (2018) 22 June 2018

Genomic SEM – Genomic SEM



nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-019-0566-x

Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits

Andrew D. Grotzinger[1*], Mijke Rhemtulla[2], Ronald de Vlaming[3,4], Stuart J. Ritchie[5,6], Travis T. Mallard[1], W. David Hill[5,6], Hill F. Ip[7], Riccardo E. Marioni[5,8], Andrew M. McIntosh[5,9], Ian J. Deary[5,6], Philipp D. Koellinger[3,4], K. Paige Harden[1,10], Michel G. Nivard[7,11] and Elliot M. Tucker-Drob[1,10,11]

Grotzinger

Nivard

Tucker-Drob

Genomic SEM – Genomic SEM

- **Apply structural equation model to estimated genetic covariance matrices**
- Allow users to examine traits that could not be measured in the same sample

- **Genomic SEM provides a flexible framework for estimating a limitless number of structural equation models using multivariate genetic data from GWAS summary statistics .**
- Can be applied to summary stats with varying and unknown degrees of overlap

Genomic SEM – Genomic SEM

Genomic SEM fits structural equation models to genetic covariance matrices derived from GWAS summary statistics using 2 Stage Estimation.

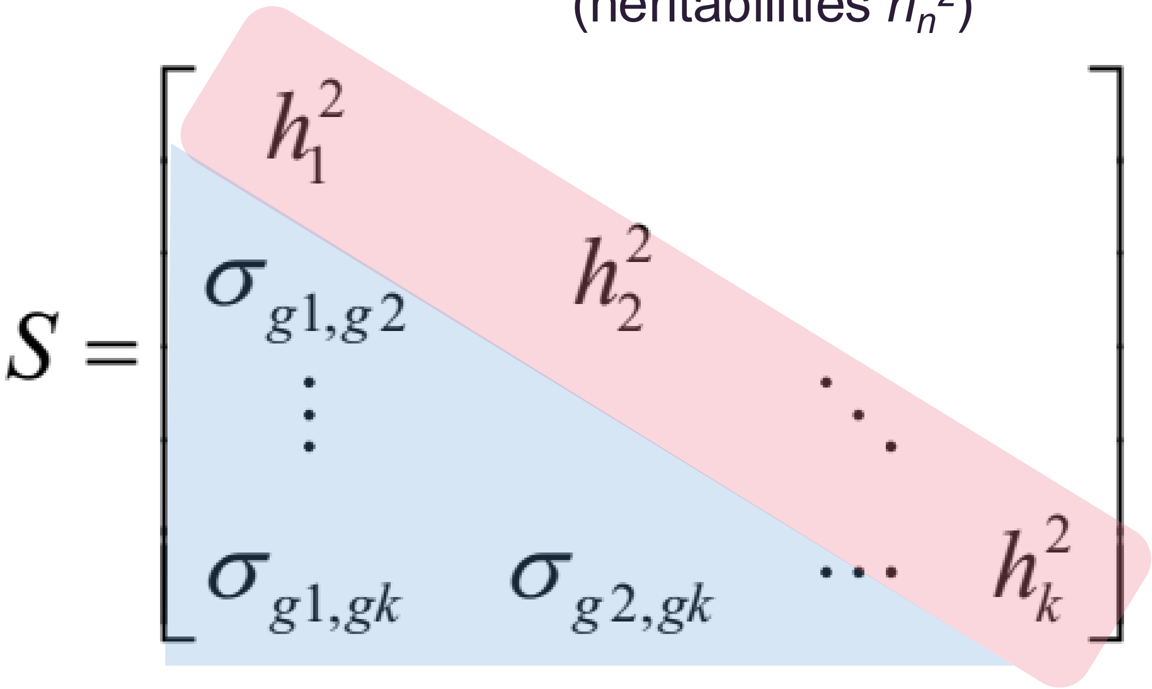
- **Stage 1: Estimate Genetic Covariance Matrix and associated matrix of standard errors and their co-dependencies**
  - We use LD Score Regression, but any method for estimating this matrix (e.g. GREML) and its sampling distribution can be used.

- **Stage 2: Fit a Structural Equation Model to the Matrices from Stage 1**

Genomic SEM – Stage 1 Estimation: Multivariable LDSC

Create a genetic covariance matrix, *S*: an “atlas of genetic correlations”

Diagonal elements are
(heritabilities h_n^2)

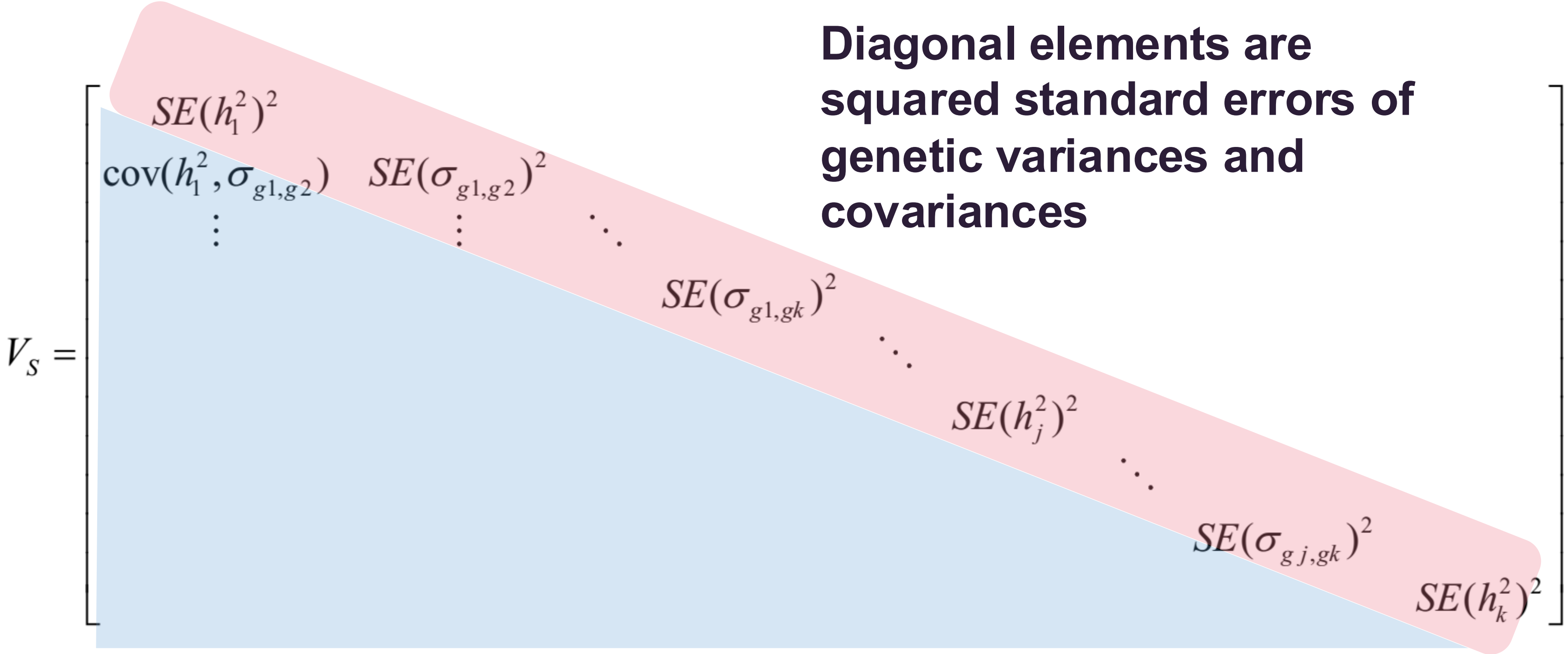


$$S = \begin{bmatrix} h_1^2 & & & \\ \sigma_{g1,g2} & h_2^2 & & \\ \vdots & & \ddots & \\ \sigma_{g1,gk} & \sigma_{g2,gk} & \cdots & h_k^2 \end{bmatrix}$$

Off-diagonal elements are
Coheritabilities ($\sigma_{gn,gm}$)

Genomic SEM – Stage 1 Estimation: Multivariable LDSC

Also produced is a second matrix, V, of squared standard errors and the dependencies between estimation errors



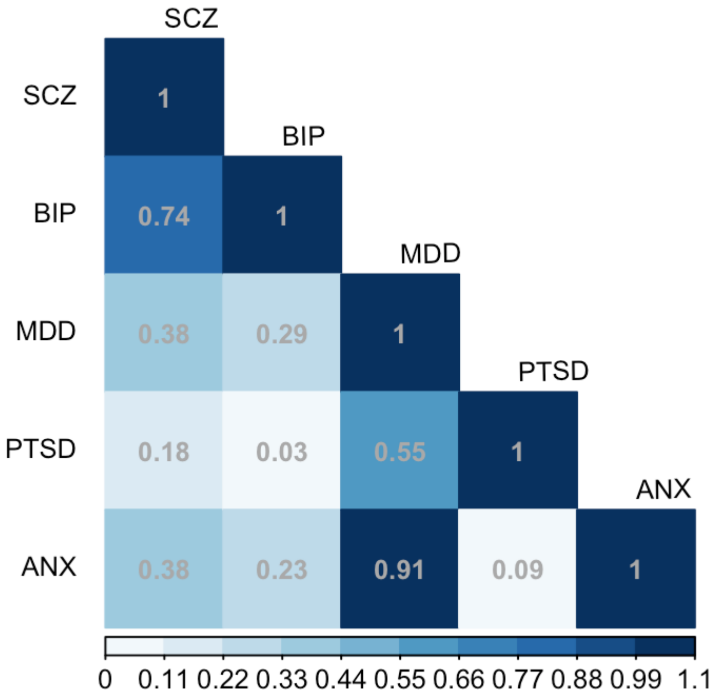
$$V_S = \begin{bmatrix} SE(h_1^2)^2 & & & & & & & \\ \mathrm{cov}(h_1^2, \sigma_{g1,g2}) & SE(\sigma_{g1,g2})^2 & & & & & & \\ \vdots & \vdots & \ddots & & & & & \\ & & & SE(\sigma_{g1,gk})^2 & & & & \\ & & & & \ddots & & & \\ & & & & & SE(h_j^2)^2 & & \\ & & & & & & \ddots & \\ & & & & & & SE(\sigma_{gj,gk})^2 & \\ & & & & & & & SE(h_k^2)^2 \end{bmatrix}$$

Diagonal elements are squared standard errors of genetic variances and covariances

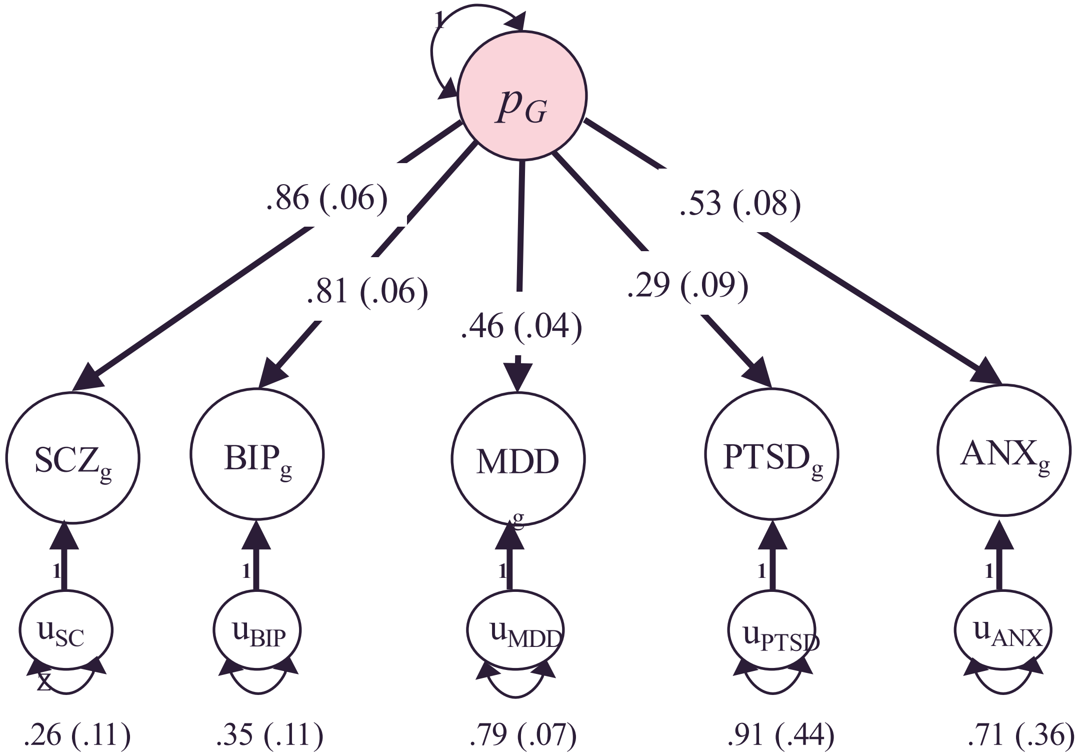
Off-diagonal elements are dependencies between estimation errors used to directly model dependencies that occur due to sample overlap from contributing GWASs

Genomic SEM – Common factor model

Genetic Correlation Matrix

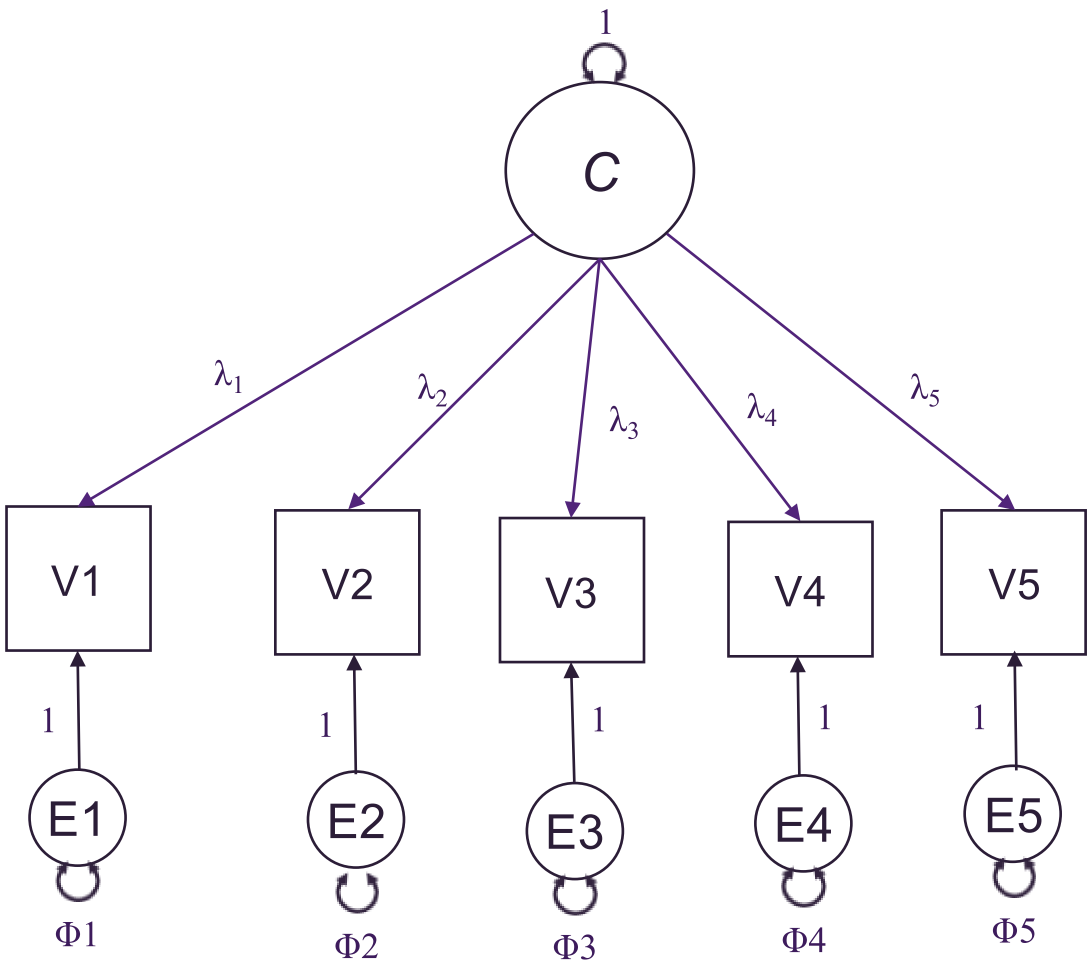


	SCZ	BIP	MDD	PTSD	ANX
SCZ	1				
BIP	0.74	1			
MDD	0.38	0.29	1		
PTSD	0.18	0.03	0.55	1	
ANX	0.38	0.23	0.91	0.09	1



Schizophrenia (SCZ), bipolar disorder (BIP), major depressive disorder (MDD), post-traumatic stress disorder (PTSD), and anxiety disorder (ANX).

SEM – Common factor model



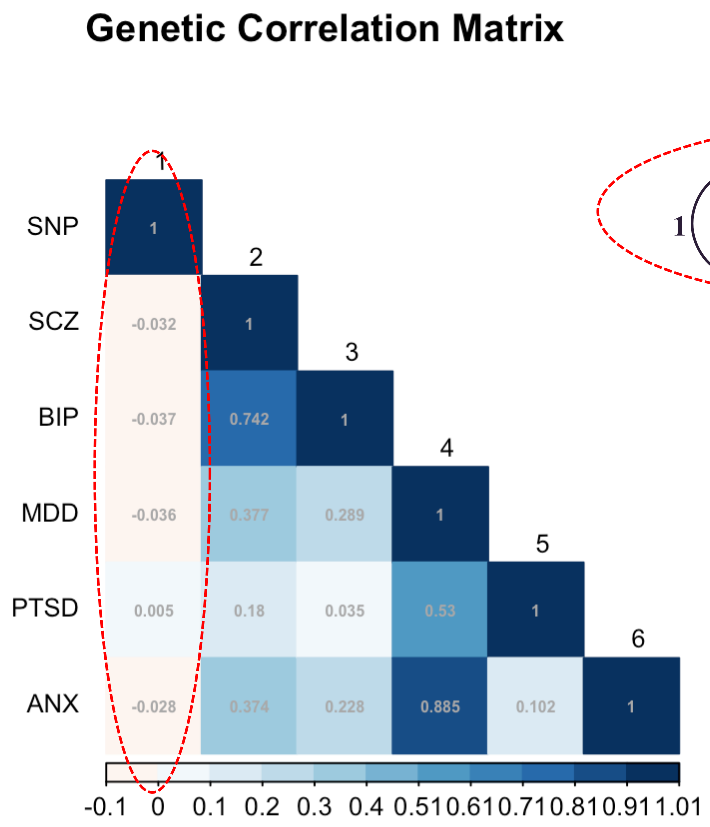
Observed Covariance Matrix:

$$S = \begin{array}{ccccc} VAR(V_1) & COV(V_1,V_2) & COV(V_1,V_3) & COV(V_1,V_4) & COV(V_5,V_1) \\ COV(V_2,V_1) & VAR(V_2) & COV(V_2,V_3) & COV(V_2,V_4) & COV(V_5,V_2) \\ COV(V_3,V_1) & COV(V_3,V_2) & VAR(V_3) & COV(V_3,V_4) & COV(V_5,V_3) \\ COV(V_4,V_1) & COV(V_4,V_2) & COV(V_4, V_3) & VAR(V_4) & COV(V_5, V_4) \\ COV(V_5,V_1) & COV(V_5,V_2) & COV(V_5,V_3) & COV(V_5, V_4) & VAR(V_5) \end{array}$$

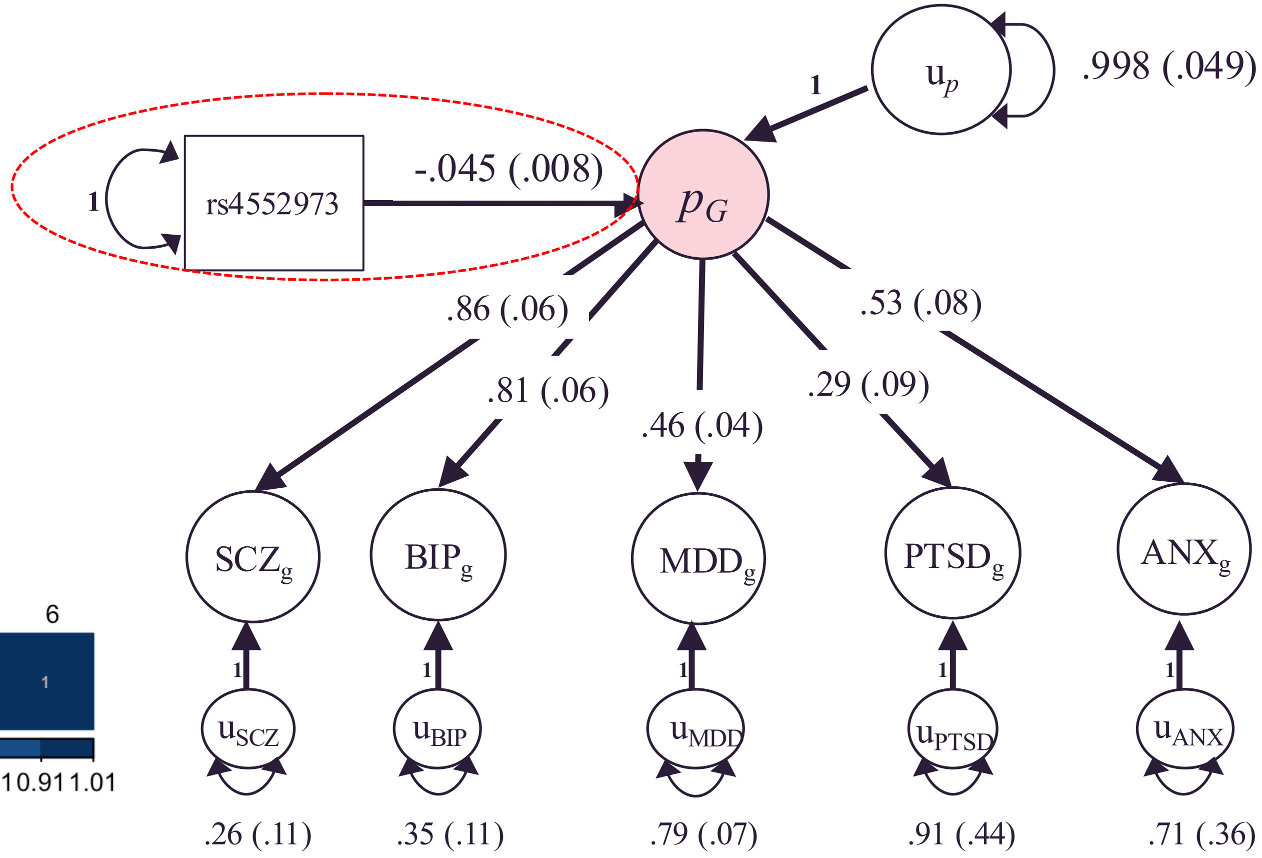
Expected Covariance Matrix:

$$\Sigma(\theta) = \begin{array}{ccccc} \lambda_1^2+\Phi_1 & \lambda_1\lambda_2 & \lambda_1\lambda_3 & \lambda_1\lambda_4 & \lambda_1\lambda_5 \\ \lambda_2\lambda_1 & \lambda_2^2+\Phi_2 & \lambda_2\lambda_3 & \lambda_2\lambda_4 & \lambda_2\lambda_5 \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3^2+\Phi_3 & \lambda_3\lambda_4 & \lambda_3\lambda_5 \\ \lambda_4\lambda_1 & \lambda_4\lambda_2 & \lambda_4\lambda_3 & \lambda_4^2+\Phi_4 & \lambda_4\lambda_5 \\ \lambda_5\lambda_1 & \lambda_5\lambda_2 & \lambda_5\lambda_3 & \lambda_5\lambda_4 & \lambda_5^2+\Phi_5 \end{array}$$

Genomic SEM – GWAS of a Latent Factor

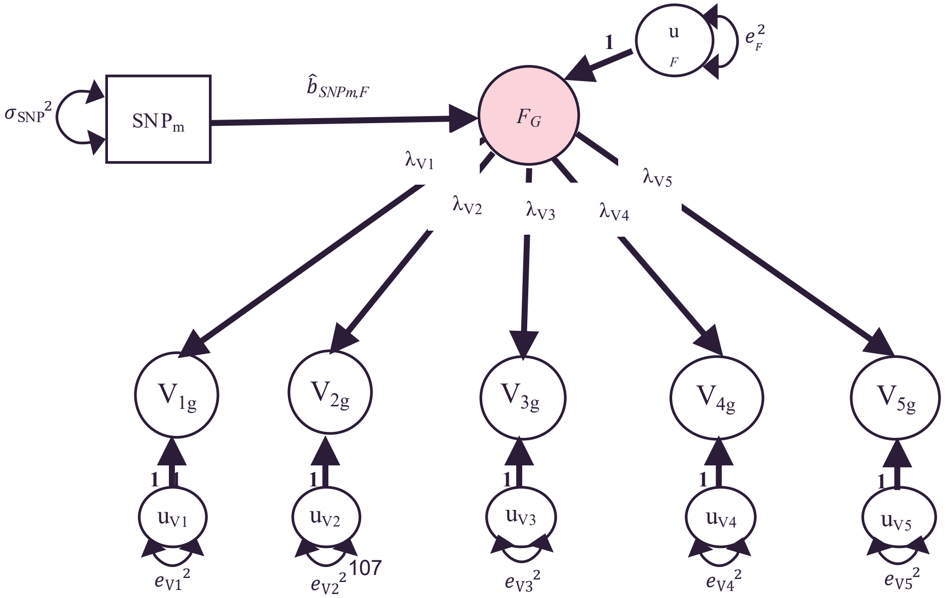

Genetic Correlation Matrix

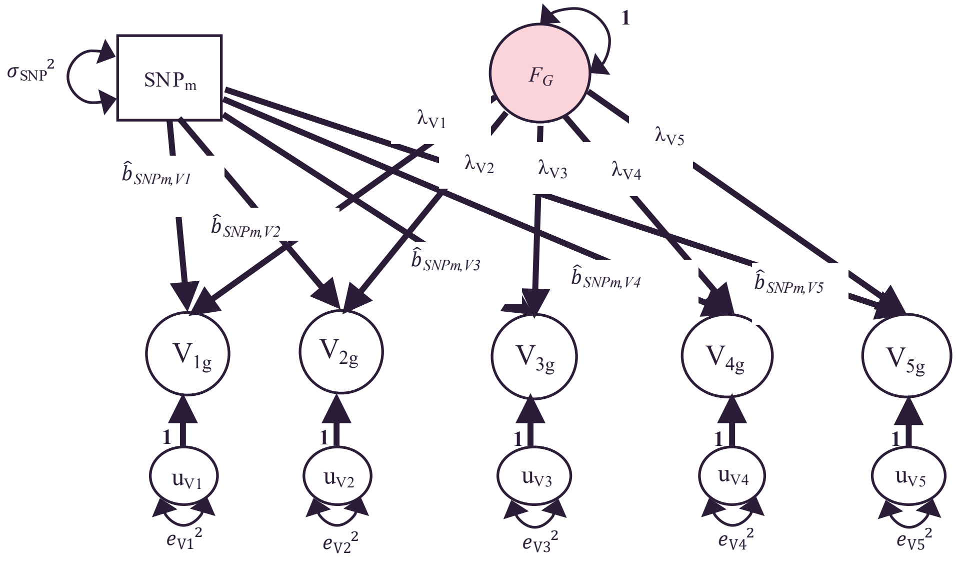
	1	2	3	4	5	6
SNP	1					
SCZ	-0.032	1				
BIP	-0.037	0.742	1			
MDD	-0.036	0.377	0.289	1		
PTSD	0.005	0.18	0.035	0.53	1	
ANX	-0.028	0.374	0.228	0.885	0.102	1



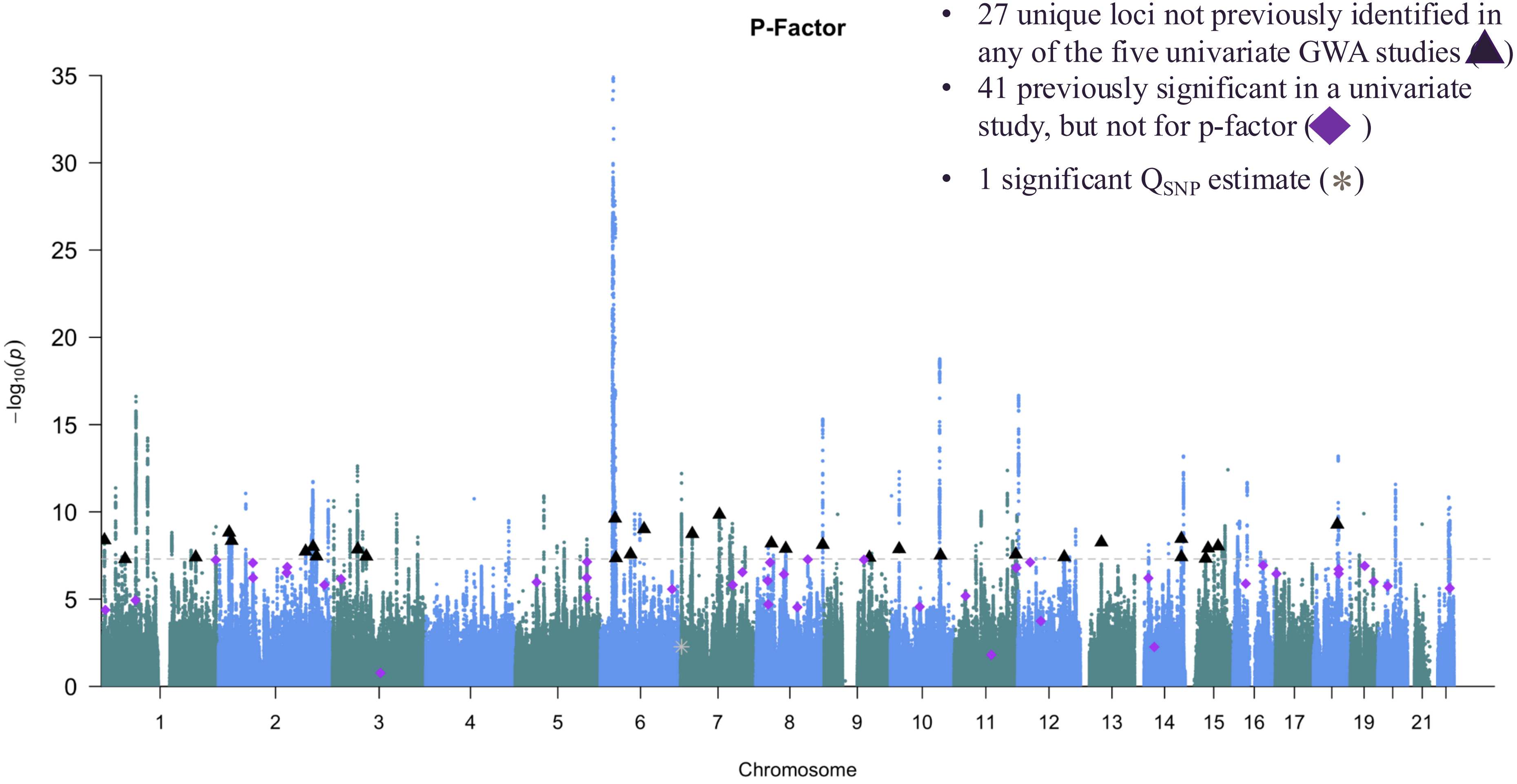
Genomic SEM - Estimates of SNP level heterogeneity (Q_{SNP})

- Asks to what extent the effect of the SNP operates through the common factor
- χ^2 distributed test statistic, indexing fit of the common pathways model against independent pathways model





Genomic SEM - Manhattan Plot (Latent Factor)



- 128 lead SNPs
- 27 unique loci not previously identified in any of the five univariate GWA studies (▲)
- 41 previously significant in a univariate study, but not for p-factor (◆)
- 1 significant Q_{SNP} estimate (*)

Take home messages – Part II

- **Genetic correlations from GWASs show widespread pleiotropy across various phenotypes.**
- **GenomicSEM is a multivariate method introduced for analyzing the joint genetic architecture of complex traits.**
- **It utilises genetic correlations and SNP heritabilities from GWAS summary statistics (i.e. LDSC), even from samples with unknown or varying overlap.**
- **It applies structural equation model to estimated genetic covariance matrices, which allow users to examine traits that could not be measured in the same sample.**

Further Reading

- **Bulik-Sullivan B. et al (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*, 47(3), 291-295.**
- **Bulik-Sullivan B. et al (2015). An atlas of genetic correlations across human diseases and traits. *Nat Genet*, 11, 1236-41.**
- **Demange PA. et al (2021). Investigating the genetic architecture of noncognitive skills using GWAS-by-subtraction. *Nat Genet*, 53(1), 35-44.**
- **Grotzinger A. et al (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat Hum Behav*, 3(5), 513-525.**
- **Warrington NM. et al (2021). Estimating direct and indirect genetic effects on offspring phenotypes using genome-wide summary results data. *Nat Commun*, 12(1), 5420.**

Acknowledgements

Institute for Molecular Bioscience

geng.wang@uq.edu.au

Prof David Evans
Dr Baptiste Couvy-Duchesne
Dr Nicole Warrington
Evans Group and CPDG

Members of the Centre for Population and Disease Genomics

Deriving Expected Variances and Covariances Using Covariance Algebra

Rules of Covariance Algebra

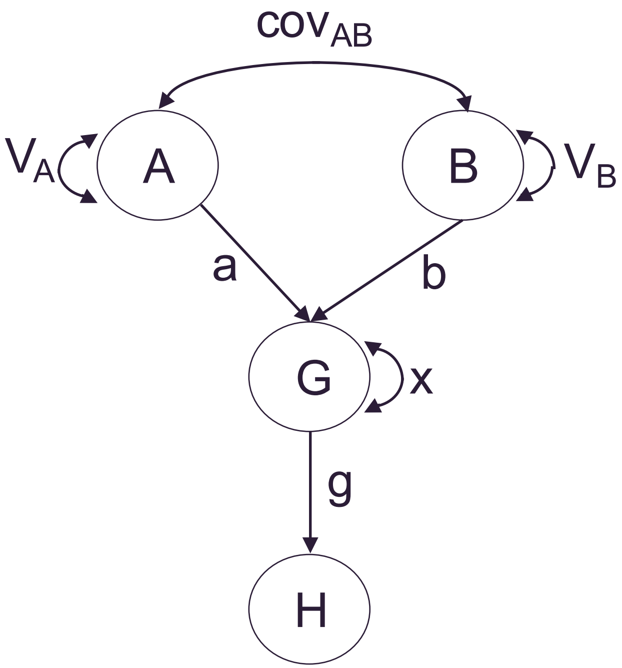
$COV(c, X) = 0$ (c is a constant)

$COV(cX_1, X_2) = cCOV(X_1, X_2)$ (c is a constant)

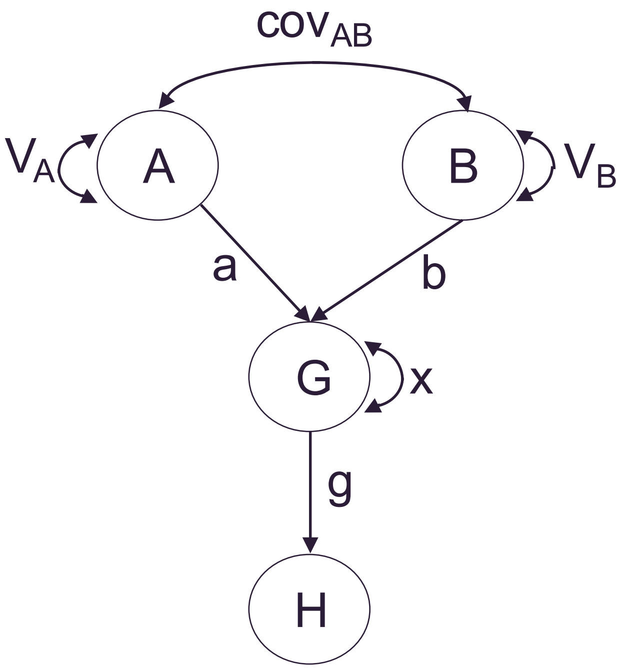
$COV(X_1 + X_2, X_3) = COV(X_1, X_3) + COV(X_2, X_3)$

$VAR(X_1) = COV(X_1, X_1)$

Covariance Algebra Example



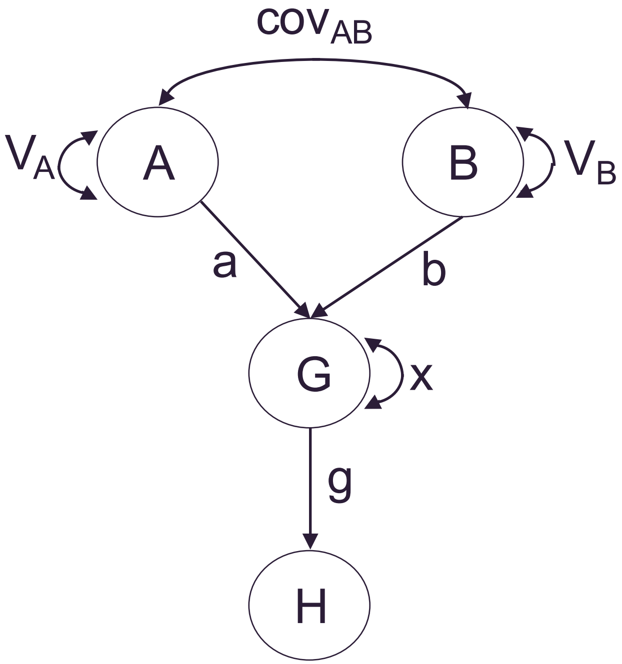
Covariance Algebra Example



$H = g*G$

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

Covariance Algebra Example

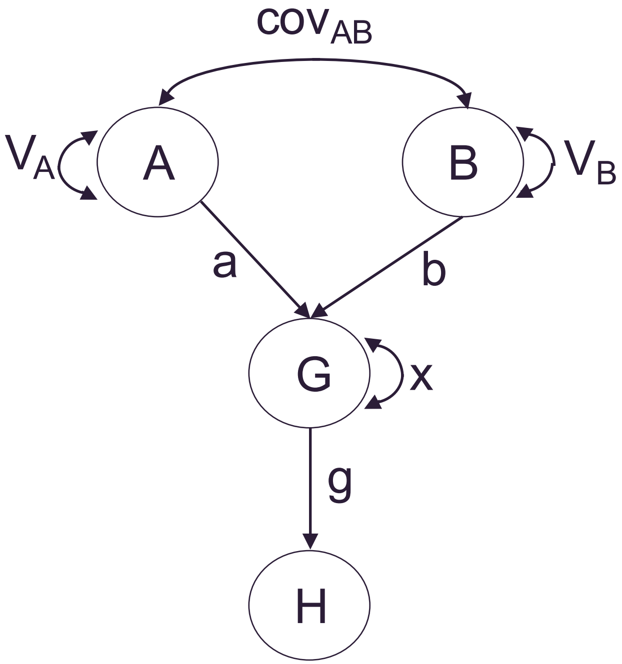


$H = g*G$

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

$COV(H,A) = ?$

Covariance Algebra Example

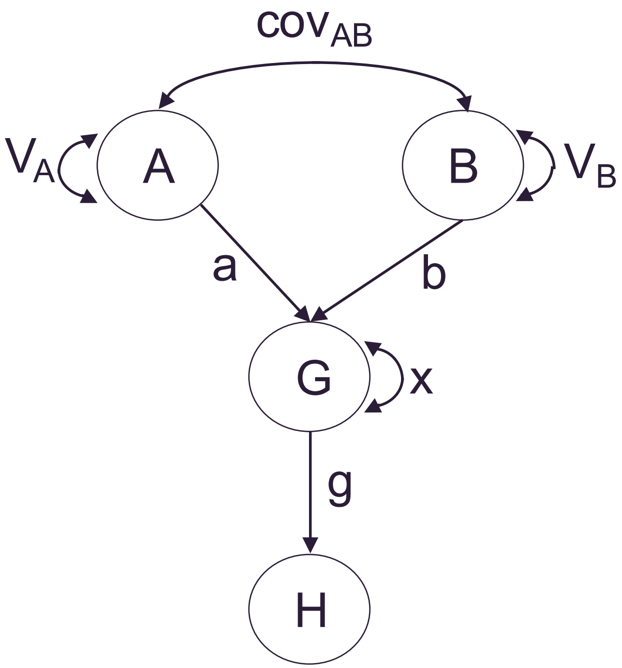


$H = g*G$

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

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

Covariance Algebra Example



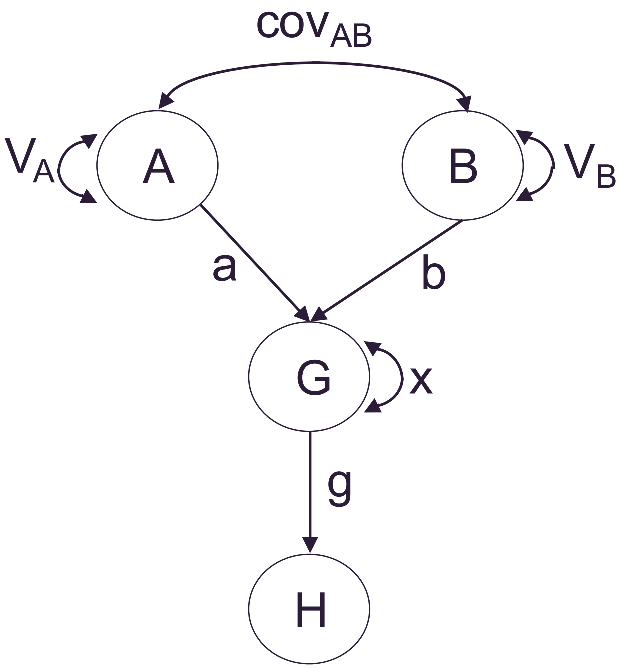
$H = g^*G$

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

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

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

Covariance Algebra Example



$H = g*G$

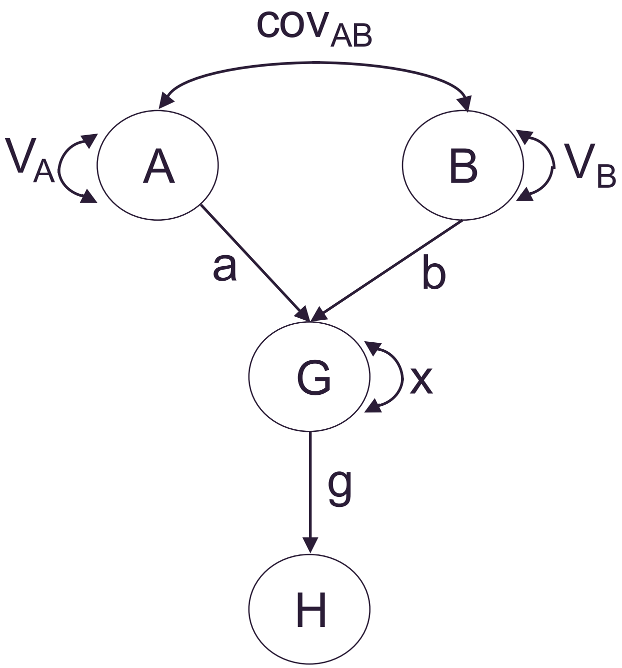
$G = a*A + b*B + e_X$

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

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

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

Covariance Algebra Example



$H = g*G$

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

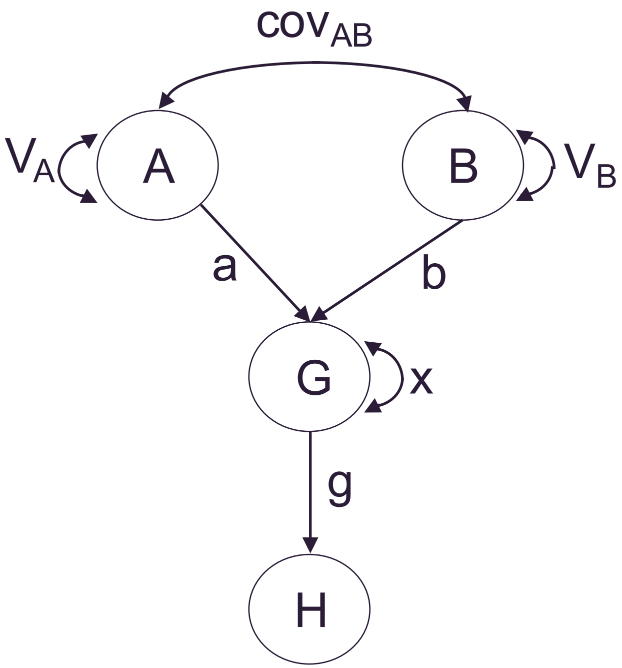
$COV(H,A) = COV(g*G, A)$

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

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

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

Covariance Algebra Example



$H = g*G$

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

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

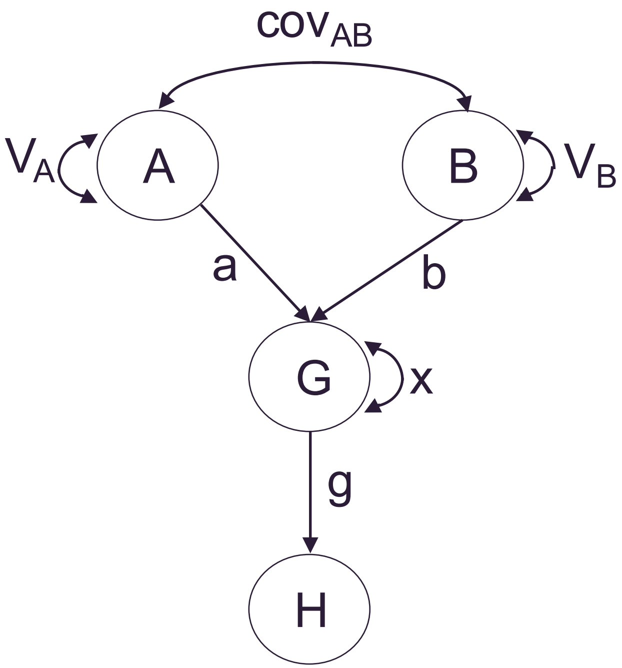
$= COV(g*(a*A + b*B + e_X), A)$

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

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

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

Covariance Algebra Example



$H = g*G$

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

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

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

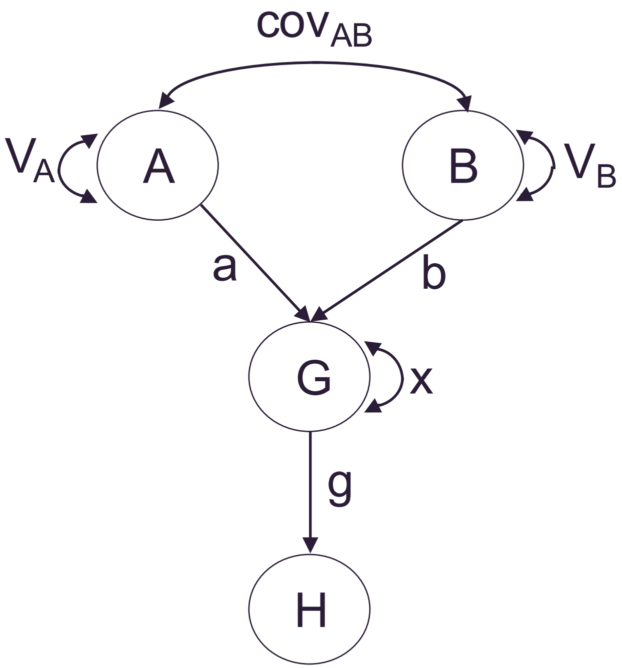
$= COV(g*a*A + g*b*B + g*e_X), A)$

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

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

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

Covariance Algebra Example



$H = g^*G$

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

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

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

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

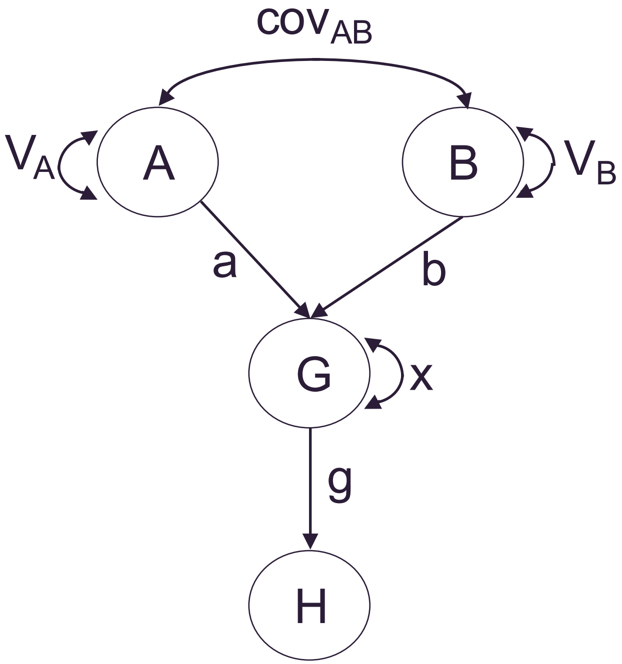
$= COV(g^*a^*A, A) + COV(g^*b^*B, A) + COV(g^*e_X, A)$

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

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

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

Covariance Algebra Example



$H = g*G$

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

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

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

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

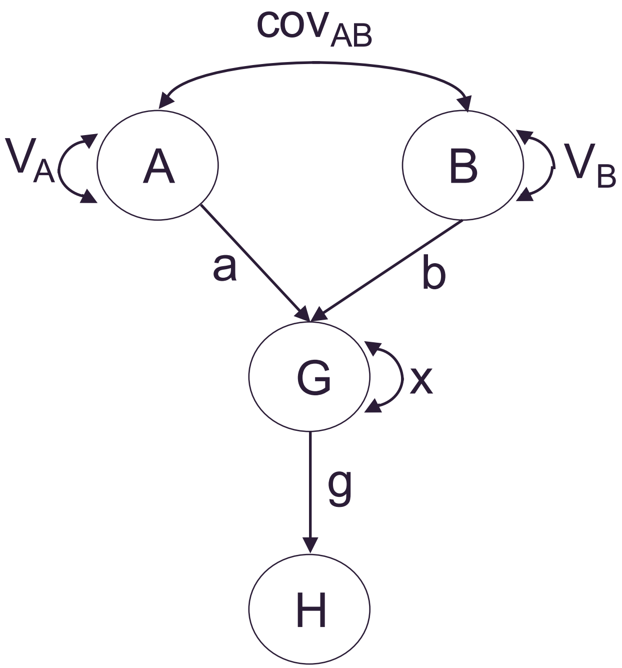
$= COV(g*a*A, A) + COV(g*b*B, A) + COV(g*e_X, A)$

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

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

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

Covariance Algebra Example

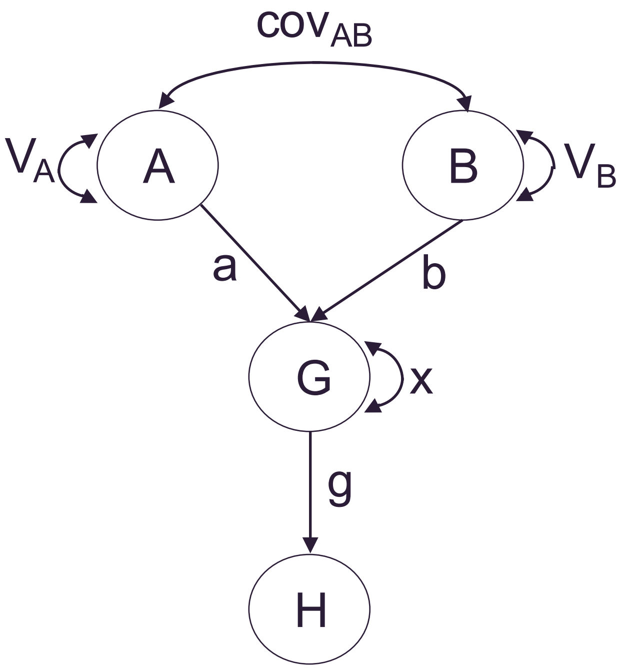


H = g*G

G = a*A + b*B

VAR(G) = ?

Covariance Algebra Example

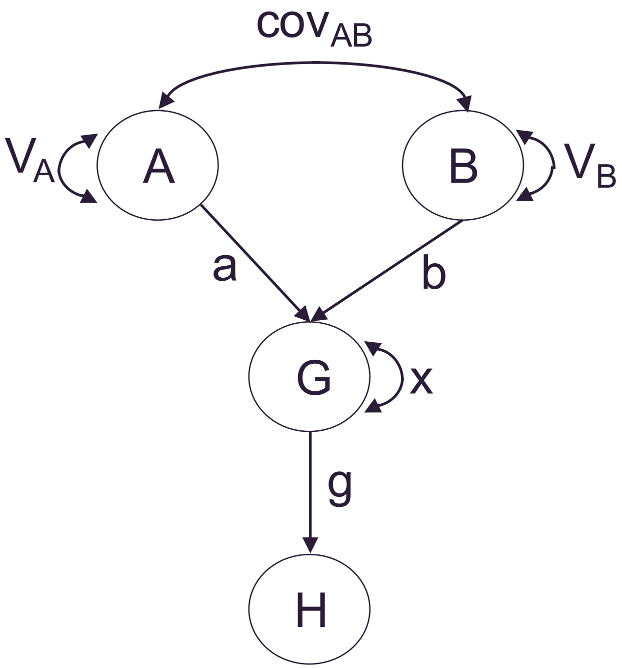


$H = g*G$

$G = a*A + b*B$

$VAR(G) = COV(G, G)$

Covariance Algebra Example



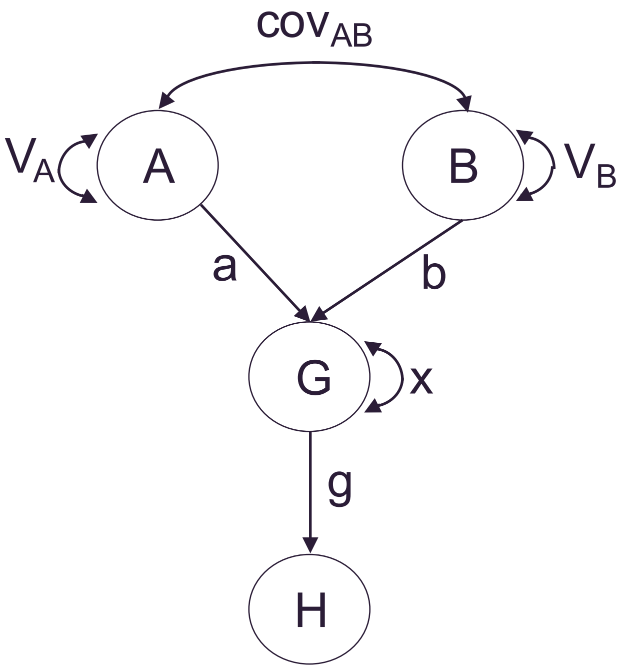
$H = g*G$

$G = a*A + b*B$

$VAR(G) = COV(G, G)$

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

Covariance Algebra Example

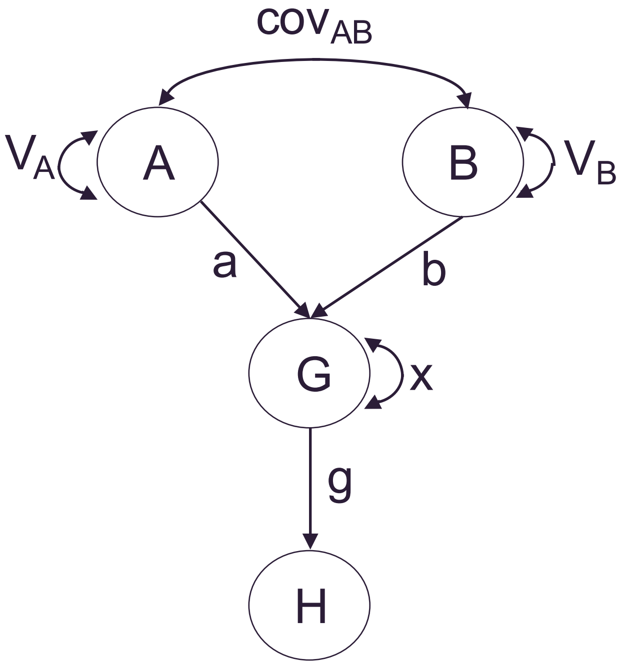


$$H = g*G$$

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

$$\begin{aligned}VAR(G) &= COV(G, G)\\ &= COV(a*A + b*B + e, a*A + b*B + e)\\ &= COV(a*A, a*A) + COV(a*A, b*B) + COV(a*A, e)\\ &+ COV(b*B, a*A) + COV(b*B, b*B) + COV(b*B, e)\\ &+ COV(e, a*A) + COV(e, b*B) + COV(e, e)\end{aligned}$$

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)

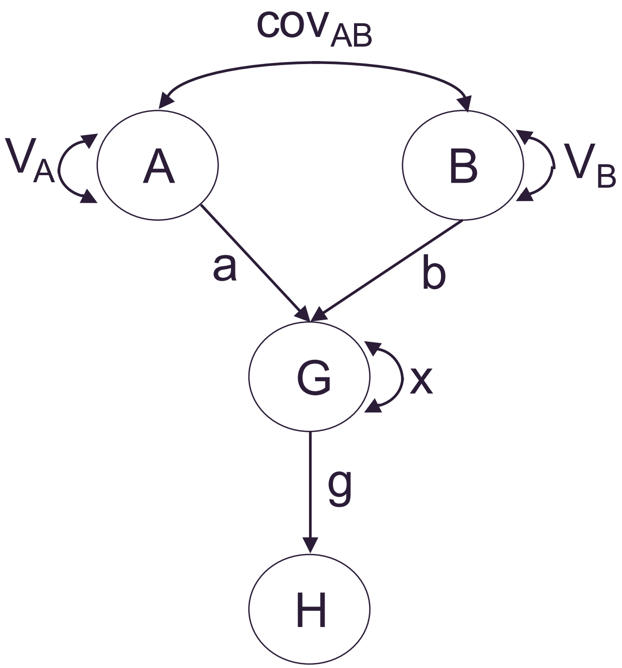
+ COV(e, a*A) + COV(e, b*B) + COV(e, e)

= 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



$H = g*G$

$G = a*A + b*B$

$VAR(G) = COV(G, G)$

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

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

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

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

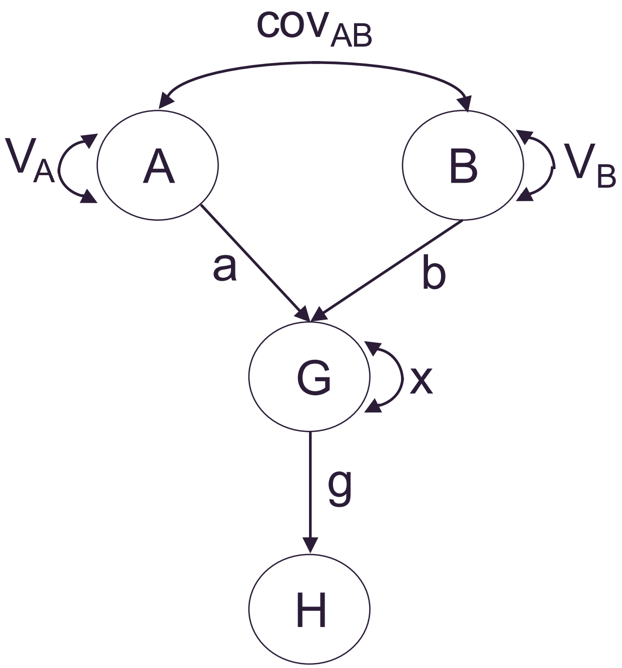
$= a*a*COV(A, A) + a*b*COV(A, B)$

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

$+ COV(e, e)$

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

Covariance Algebra Example



$H = g*G$

$G = a*A + b*B$

$\mathrm{VAR}(G) = \mathrm{COV}(G, G)$

$= \mathrm{COV}(a*A + b*B + e, a*A + b*B + e)$

$= \mathrm{COV}(a*A, a*A) + \mathrm{COV}(a*A, b*B) + \mathrm{COV}(a*A, e)$

$+ \mathrm{COV}(b*B, a*A) + \mathrm{COV}(b*B, b*B) + \mathrm{COV}(b*B, e)$

$+ \mathrm{COV}(e, a*A) + \mathrm{COV}(e, b*B) + \mathrm{COV}(e, e)$

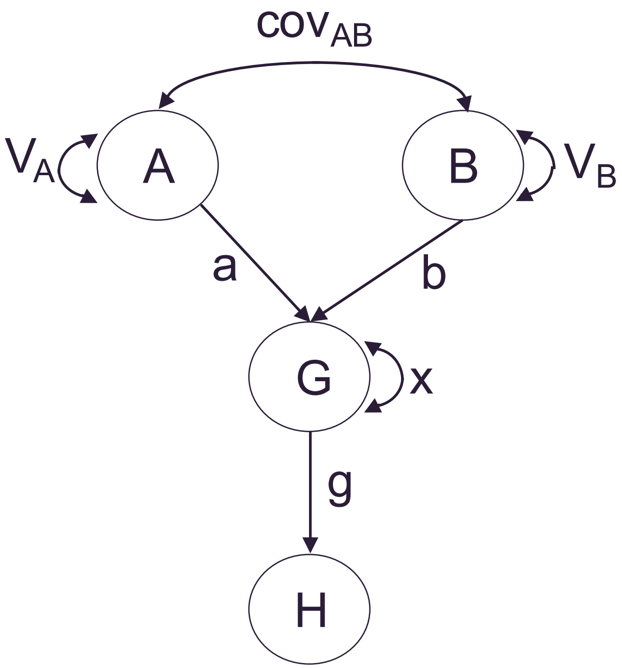
$= a*a*\mathrm{COV}(A, A) + a*b*\mathrm{COV}(A, B)$

$+ b*a*\mathrm{COV}(B, A) + b*b*\mathrm{COV}(A, B)$

$+ \mathrm{COV}(e, e)$

$$\boxed{= a^2*V_A + b^2*V_B + 2*a*b*\mathrm{COV}_{AB} + x}$$

Covariance Algebra Example

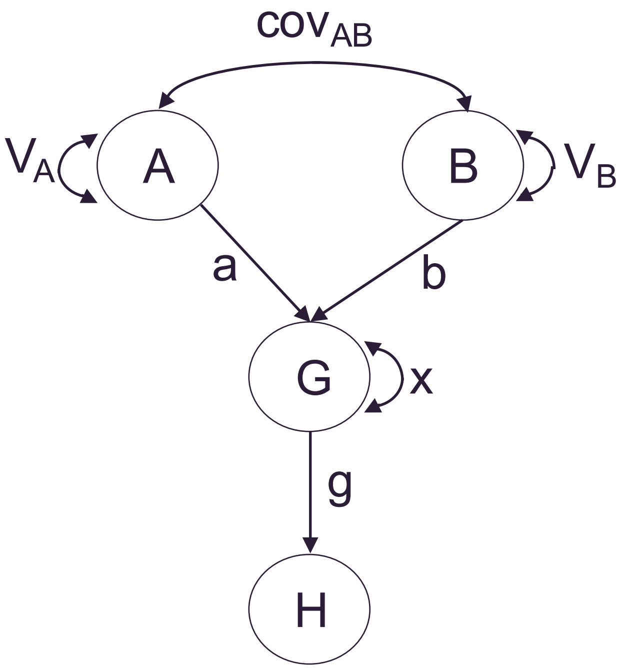


$H = g*G$

$G = a*A + b*B$

$VAR(H) = ?$

Covariance Algebra Example

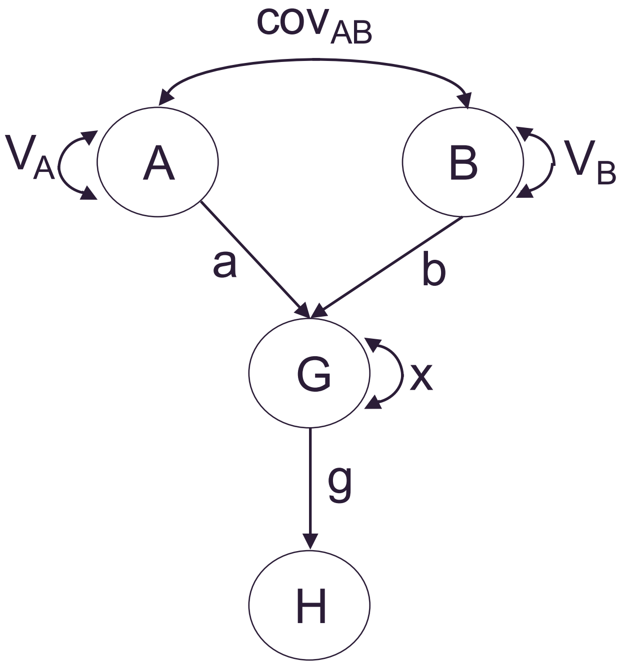


H = g*G

G = a*A + b*B

VAR(H) = COV(H, H)

Covariance Algebra Example



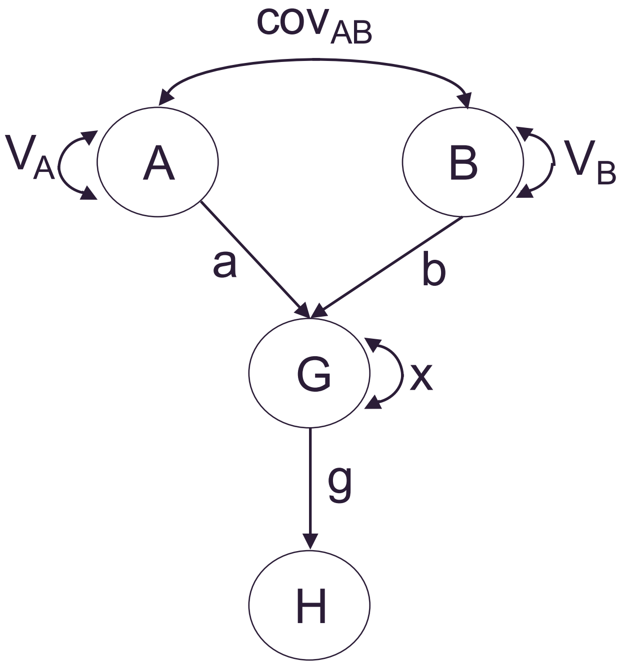
H = g*G

G = a*A + b*B

VAR(H) = COV(H, H)

= COV(g*G, g*G)

Covariance Algebra Example



H = g*G

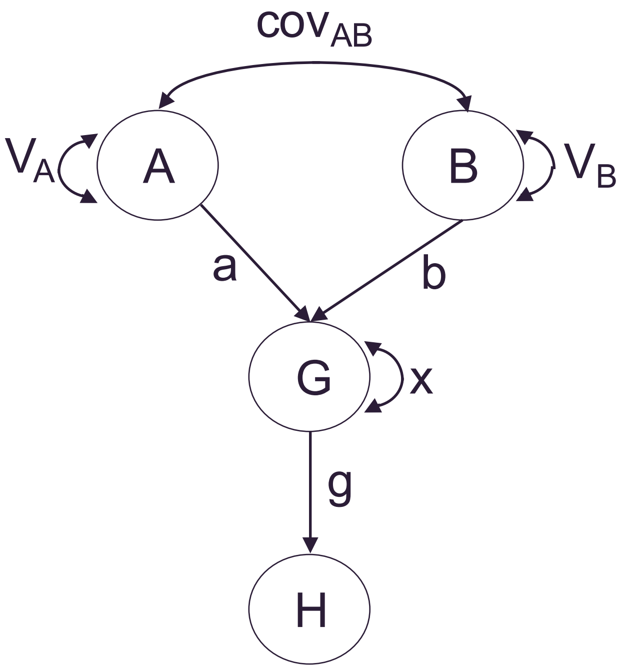
G = a*A + b*B

VAR(H) = COV(H, H)

= COV(g*G, g*G)

= g*g*COV(G, G)

Covariance Algebra Example

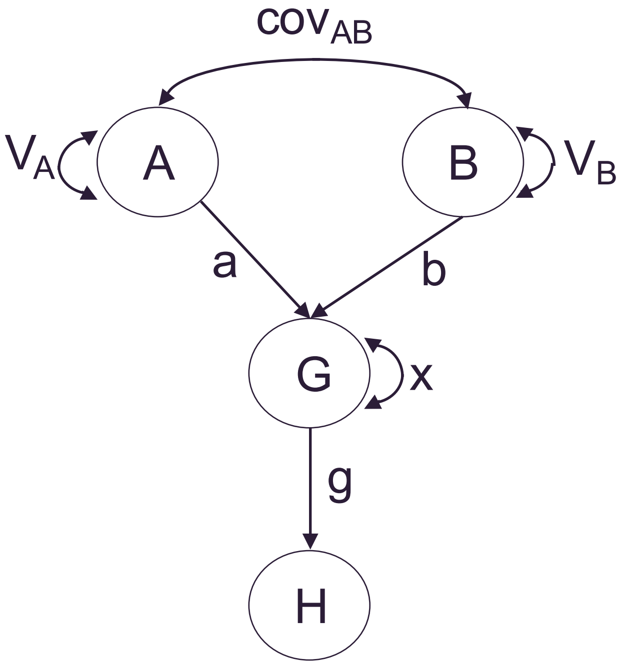


$H = g*G$

$G = a*A + b*B$

$$\begin{aligned} VAR(H) &= COV(H, H) \\ &= COV(g*G, g*G) \\ &= g*g*COV(G, G) \\ &= g^2*VAR(G) \end{aligned}$$

Covariance Algebra Example



$H = g*G$

$G = a*A + b*B$

$VAR(H) = COV(H, H)$

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

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

$\boxed{= g^2*VAR(G)}$