THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

THE UNIVERSITY OF QUEENSLAND AUSTRALIA | IMB Institute for Molecular Bioscience

Introduction to Structural Equation Modelling (SEM)

Baptiste Couvy-Duchesne[1,2], Geng Wang1, Nicole Warrington1, David Evans[1,3,4]

1 Institute for Molecular Bioscience, University of Queensland

2 Paris Brain Institute, INRIA

3 University of Queensland Diamantina Institute

4 MRC Integrative Epidemiology Unit, University of Bristol

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

What is SEM?

- A statistical framework 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

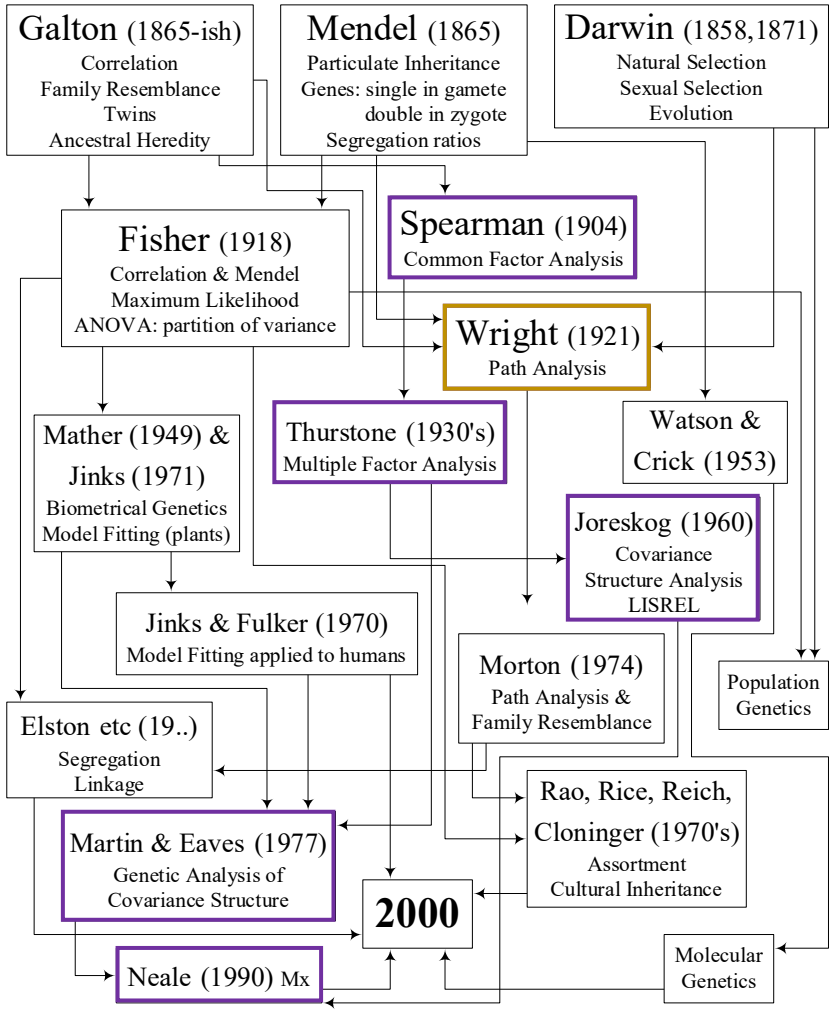
Useful for deriving expected variances/covariances in genetics

SEM means that you can think about a problem multiple ways

Advantages for modelling human genetic data:

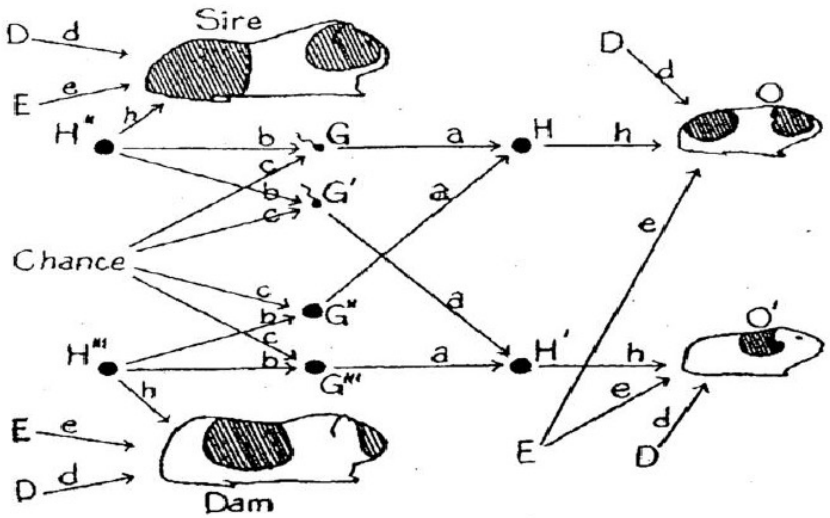
- Latent variables
- Multivariate phenotypes
- Feedback loops
- Assortative mating
- Vertical transmission
- Gene-environment covariance
- Non-linear constraints

SEM and Genetics

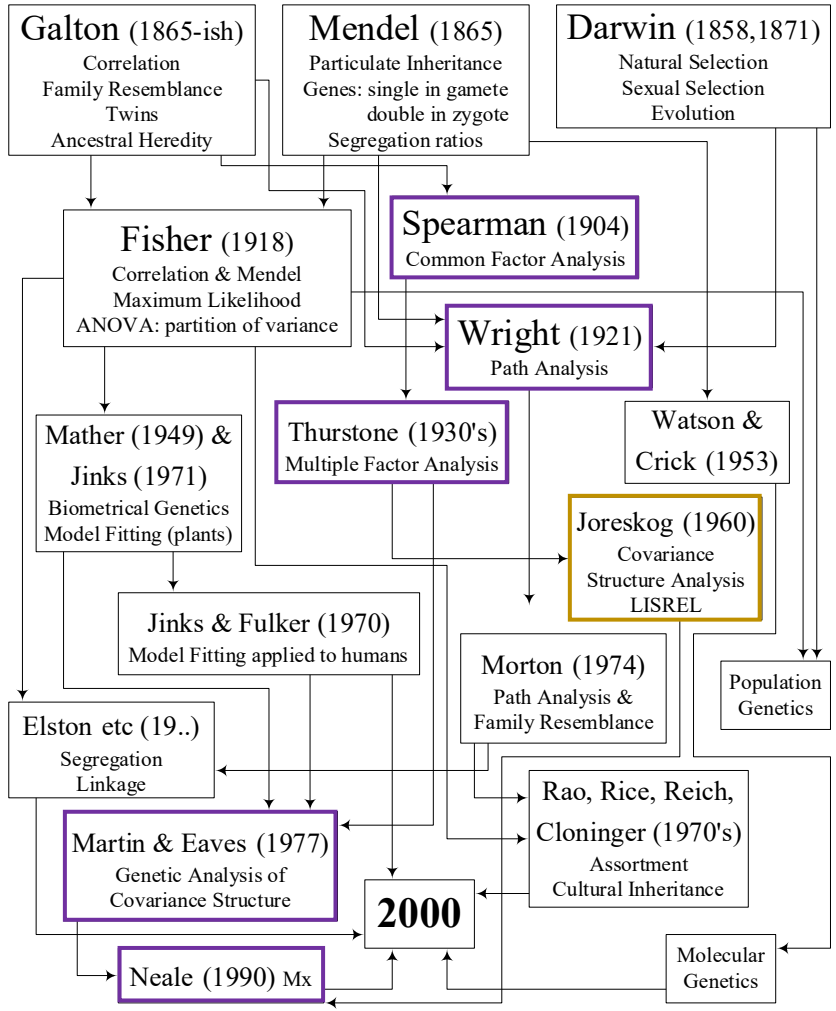


Neale & Cardon (1992)

Sewall Wright

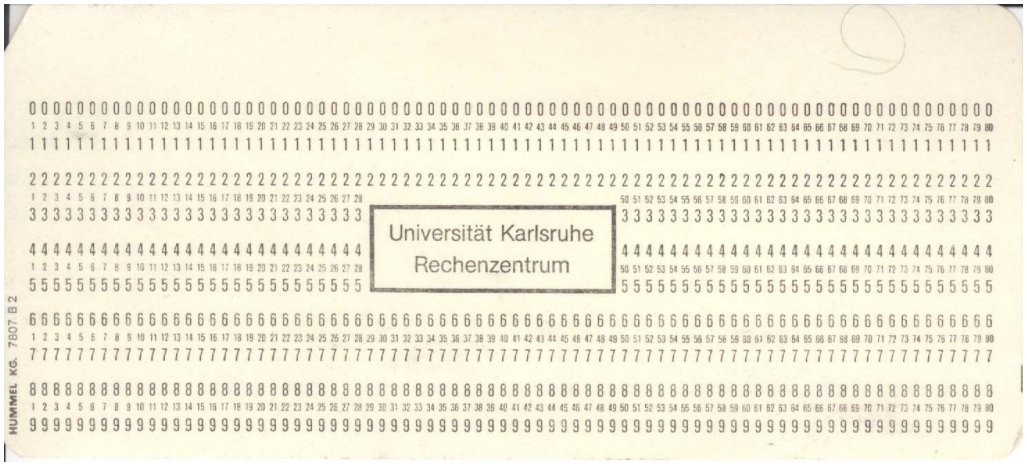


SEM and Genetics



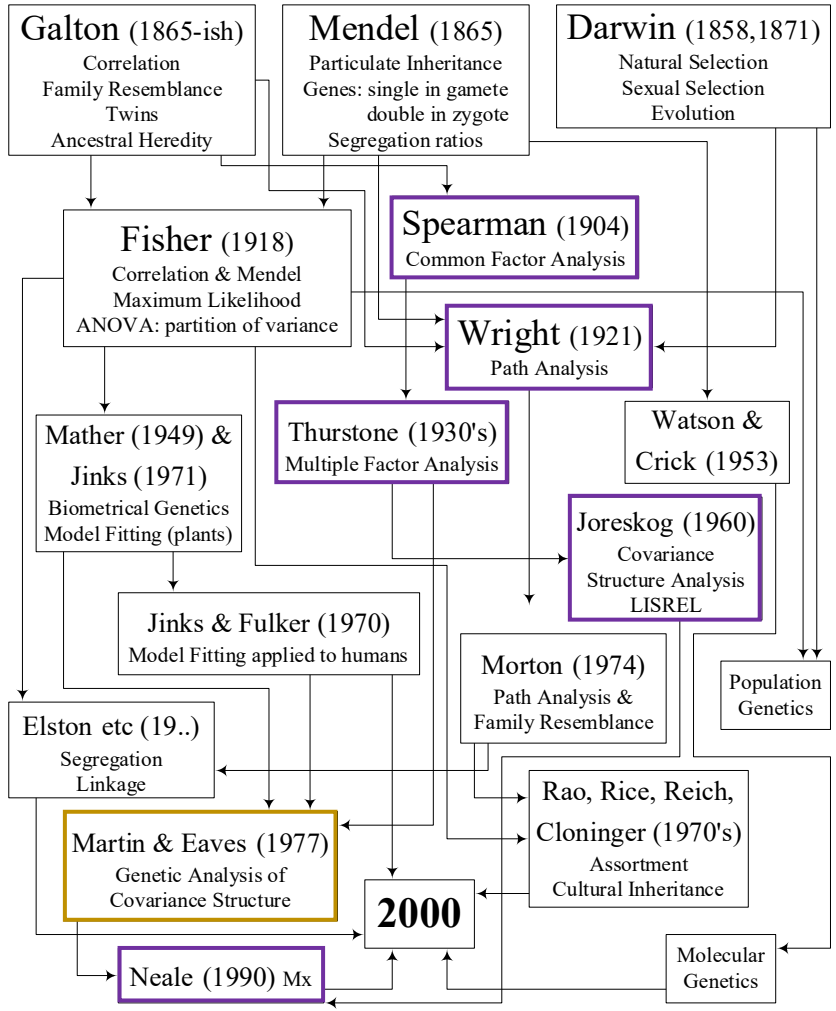
Neale & Cardon (1992)

Karl Jöreskog



A punch card – circa 1970

SEM and Genetics



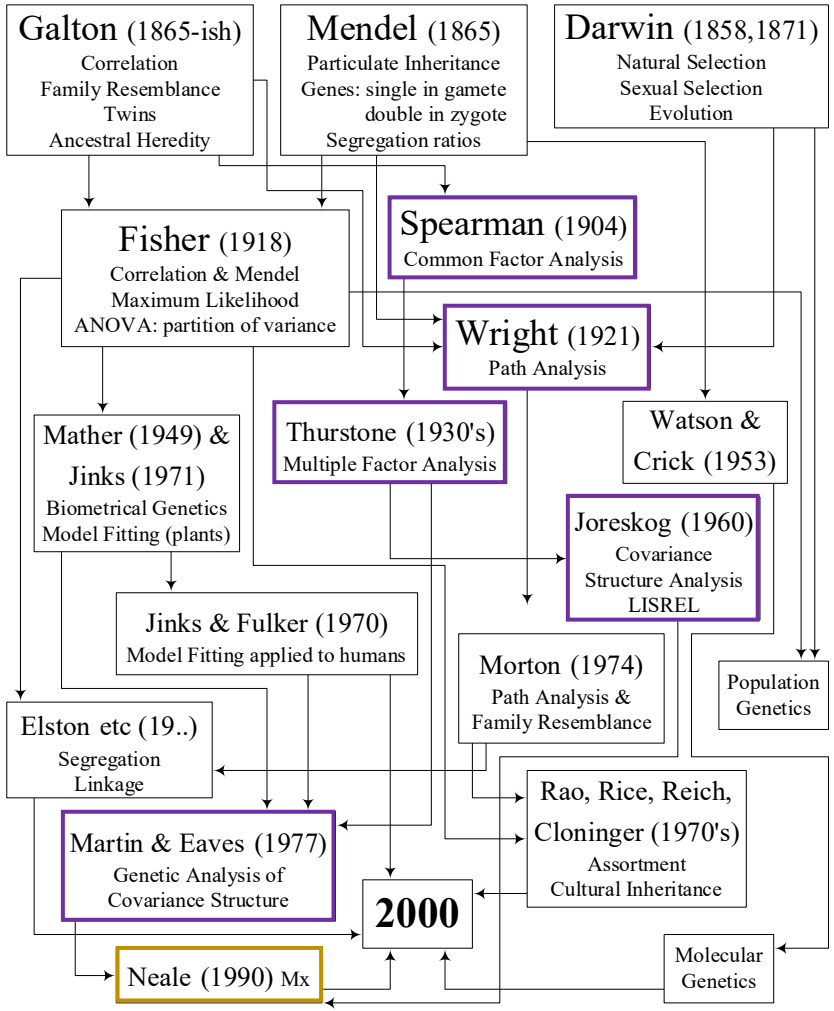
Neale & Cardon (1992)

Nick Martin Lindon Eaves
QIMRB, Brisbane, Australia

Multivariate twin model of 5 ability domains:

Numerical ability
Verbal comprehension
Spatial ability
Word fluency
Reasoning ability

SEM and Genetics



Neale & Cardon (1992)

Mike Neale
VCU – Virginia, USA

OpenMx

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

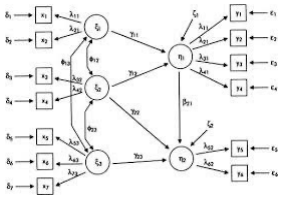
(4) FIT THE MODEL TO THE DATA. OBTAIN PARAMETER ESTIMATES AND A MEASURE OF HOW WELL THE MODEL FITS THE DATA.

(5) REVISE THE THEORY/MODEL

How does SEM Work?

THEORY

BUILD MODEL



Expected covariance structure

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

FIT TO DATA

$\boldsymbol{\Sigma}$ Observed covariance structure

OBTAIN PARAMETER ESTIMATES AND CIs

REVISE or REFINE MODEL/THEORY

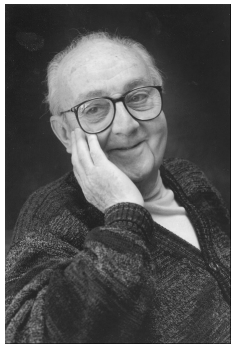
"All Models and Wrong – Some Models are Useful"

- 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 parameter estimation and confidence intervals
- SEM and model falsification

Which model is "correct"?

$$\mathbf{Y} = b_x \mathbf{X} + \mathbf{e} \qquad b_x = \mathrm{cov}(X,Y) / \mathrm{sd}(X)$$

$$\mathbf{X} = b_y \mathbf{Y} + \mathbf{e} \qquad b_y = \mathrm{cov}(X,Y) / \mathrm{sd}(Y)$$



George Box

SEM-Assumptions

Linearity

Multivariate normality (normality of residuals)

- Binary/ordinal variables can be modelled assuming an underlying normal distribution of liability
- Methods exist for combining binary and continuous variables

Identifiability

- **Means that all parameters in a model can be estimated 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 identifiable, then the model is identifiable

- Even though the model as a whole may be unidentifiable you may be able to estimate some of the parameters (*partial identifiability*) or locate them in the parameter space (*set identifiability*)

Identifiable or Not ?

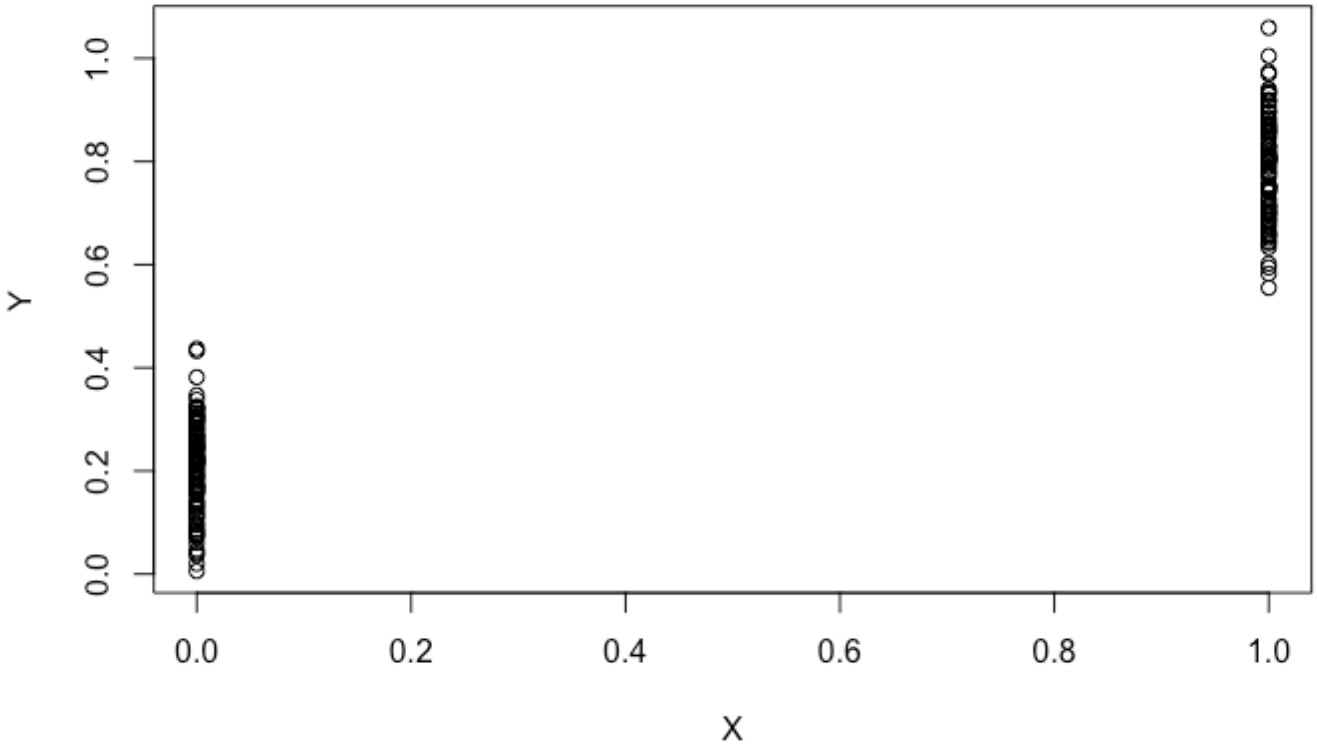
$$\mathbf{Y} = u + b\,\mathbf{X} + \mathbf{e}$$

$\mathbf{X}$ in $\{0,1\}$

$\mathbf{Y}$ continuous

Y continuous outcome (e.g. response to treatment)
X dose of treatment

u effect of no treatment
b effect attributable to treatment



Identifiable or Not ?

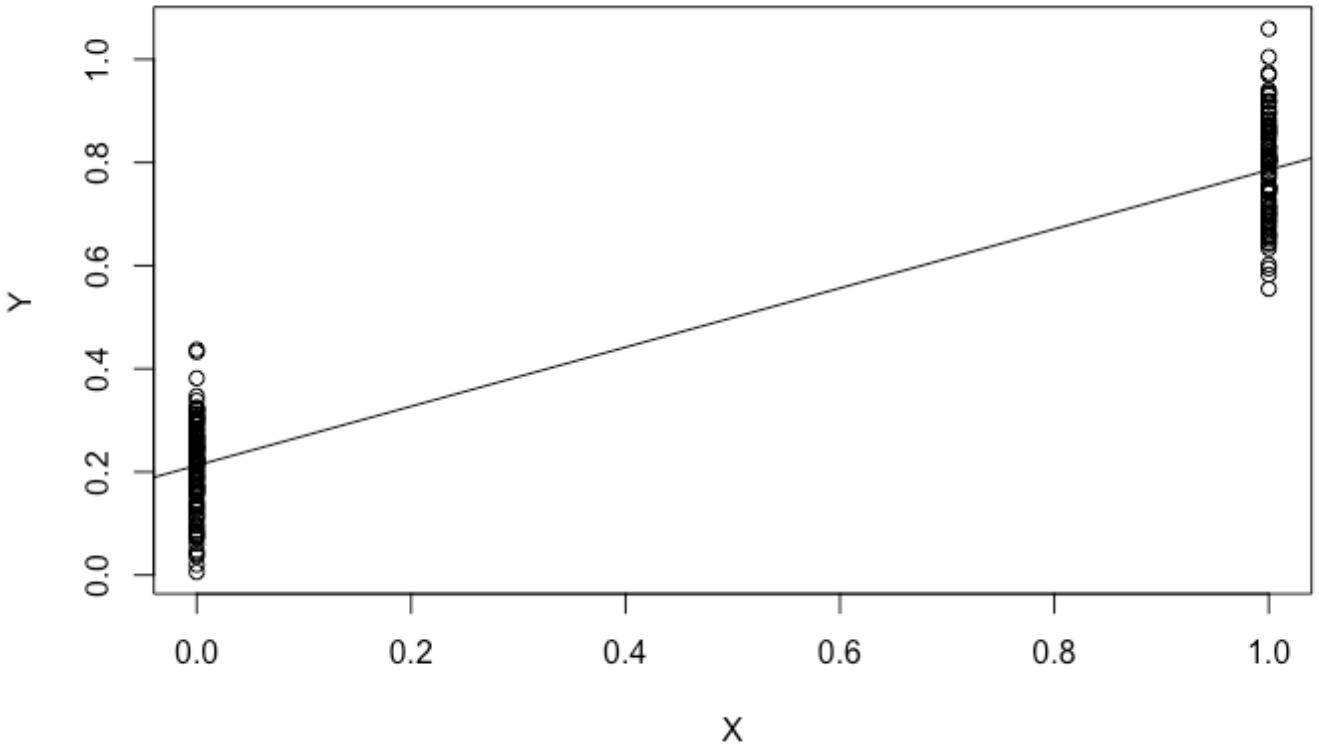
$\mathbf{Y} = u + b\,\mathbf{X} + \mathbf{e}$

$\mathbf{X}$ in $\{0,1\}$

$\mathbf{Y}$ continuous

```
Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)            X
     0.2126       0.5728
```



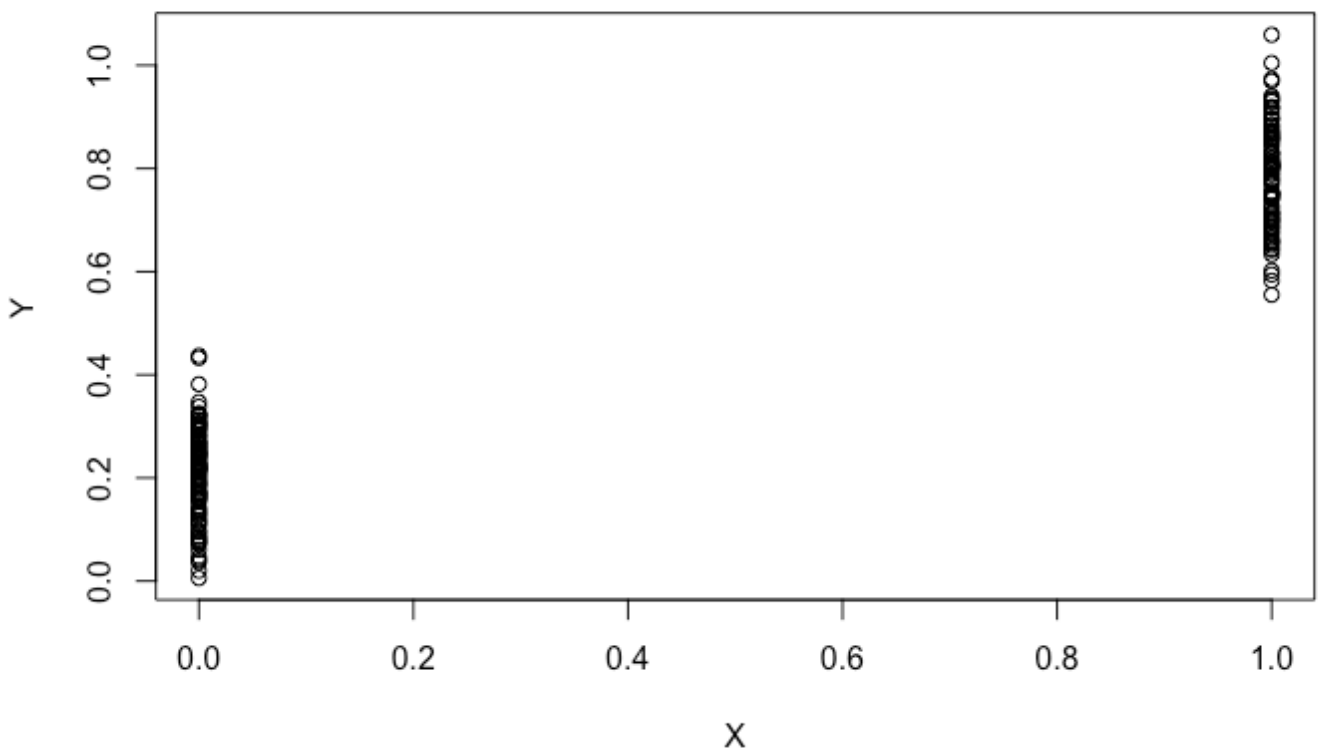
Identifiable or Not ?

$$\mathbf{Y} = u + b\,\mathbf{X} / (\mathbf{X} + c) + \mathbf{e}$$

$$\mathbf{X} \text{ in } \{0,1\}$$

EMAX model:
Y continuous outcome (e.g. response to treatment)
X dose of treatment
u effect of no treatment
b maximal effect attributable to treatment

c exposure that produces half of b



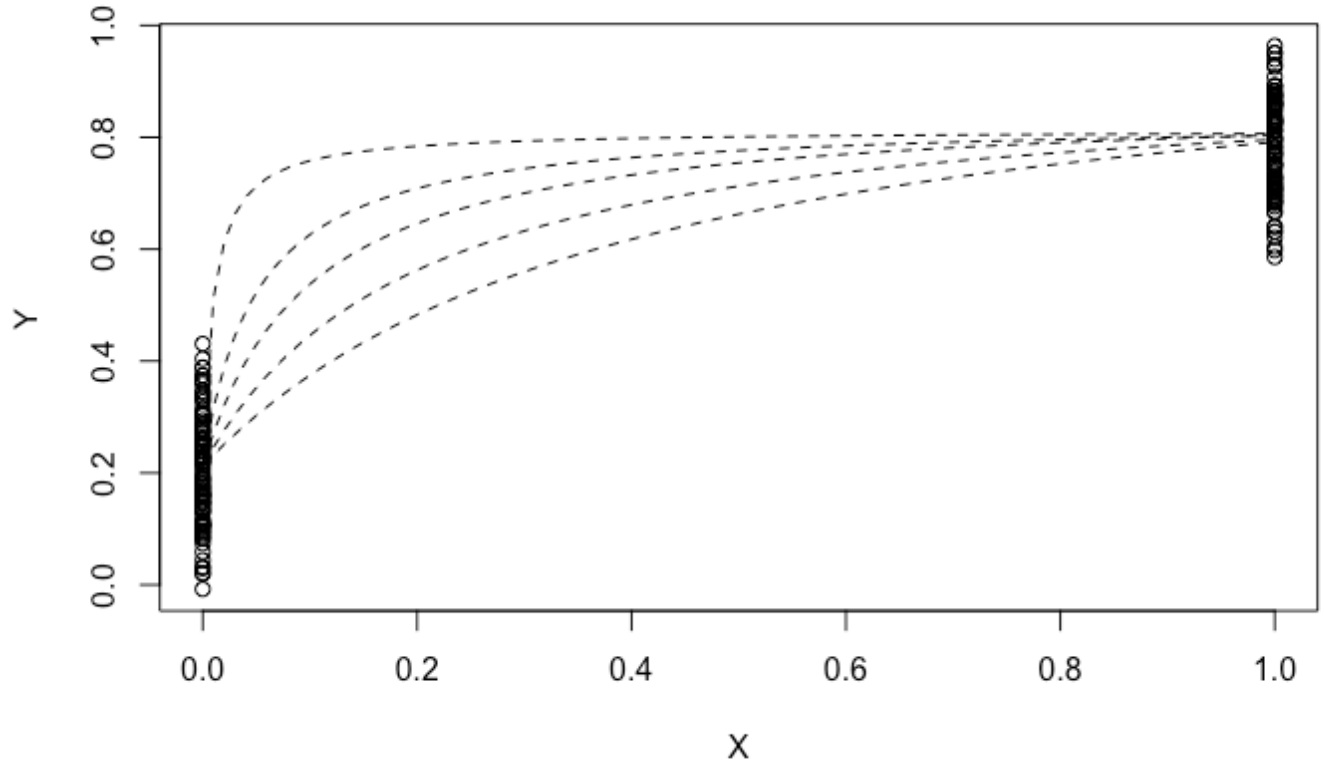
Identifiable or Not ?

$$\mathbf{Y} = u + b\,\mathbf{X} / (\mathbf{X} + c) + \mathbf{e}$$

$$\mathbf{X} \text{ in } \{0,1\}$$

EMAX model:
Y continuous outcome (e.g. response to treatment)
X dose of treatment
u effect of no treatment
b maximal effect attributable to treatment

c exposure that produces half of b



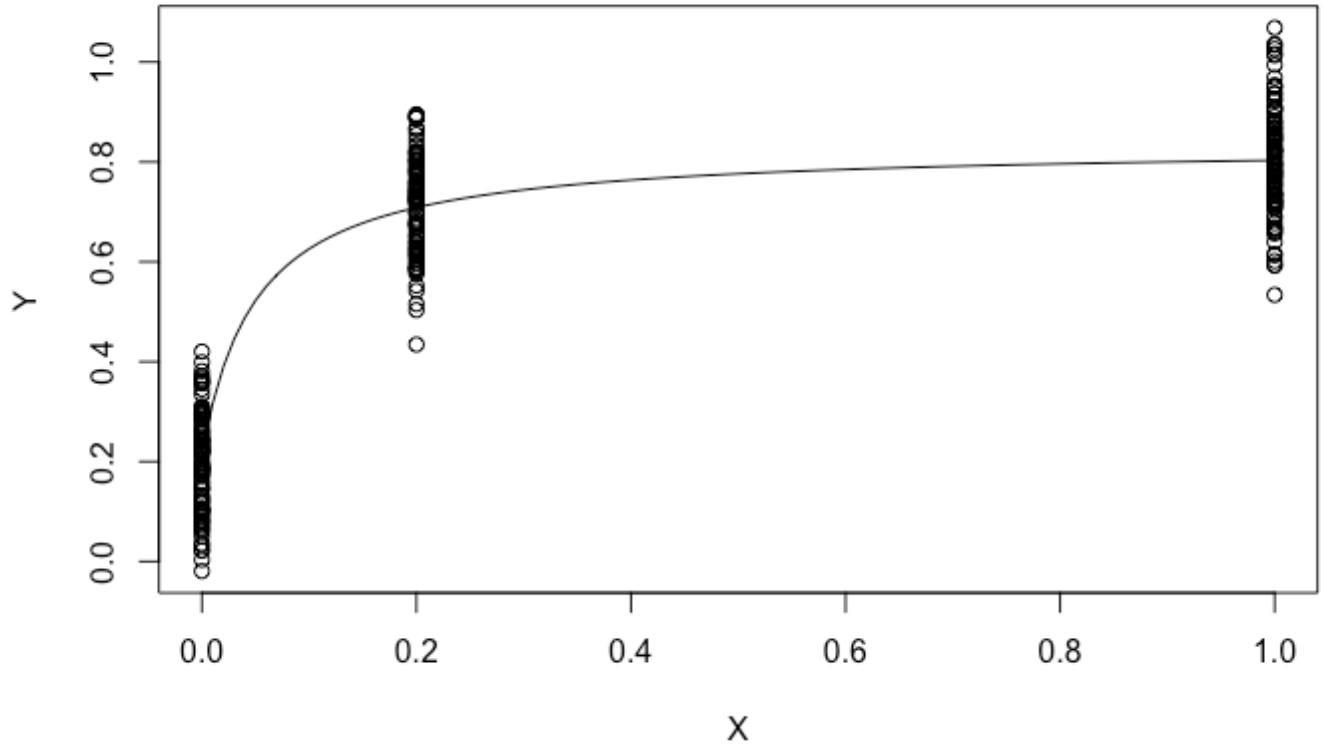
Identifiable or Not ?

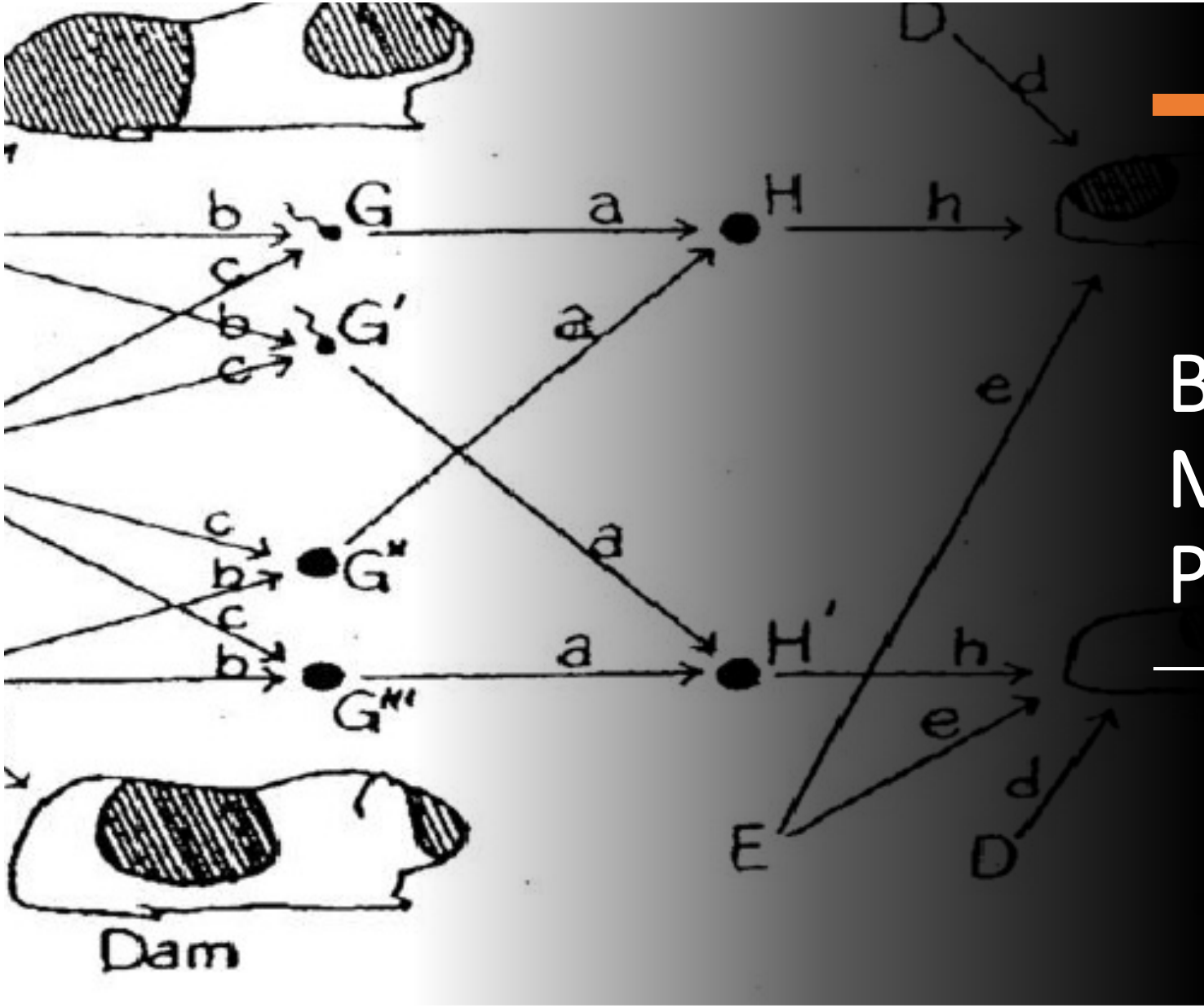
$$\mathbf{Y} = u + b\,\mathbf{X} / (\mathbf{X} + c) + \mathbf{e}$$

$$\mathbf{X} \text{ in } \{0, 0.2, 1\}$$

EMAX model:
Y continuous outcome (e.g. response to treatment)
X dose of treatment

u effect of no treatment: **0.21**
b maximal effect attributable to treatment: **0.62**
c exposure that produces half of b: **0.05**

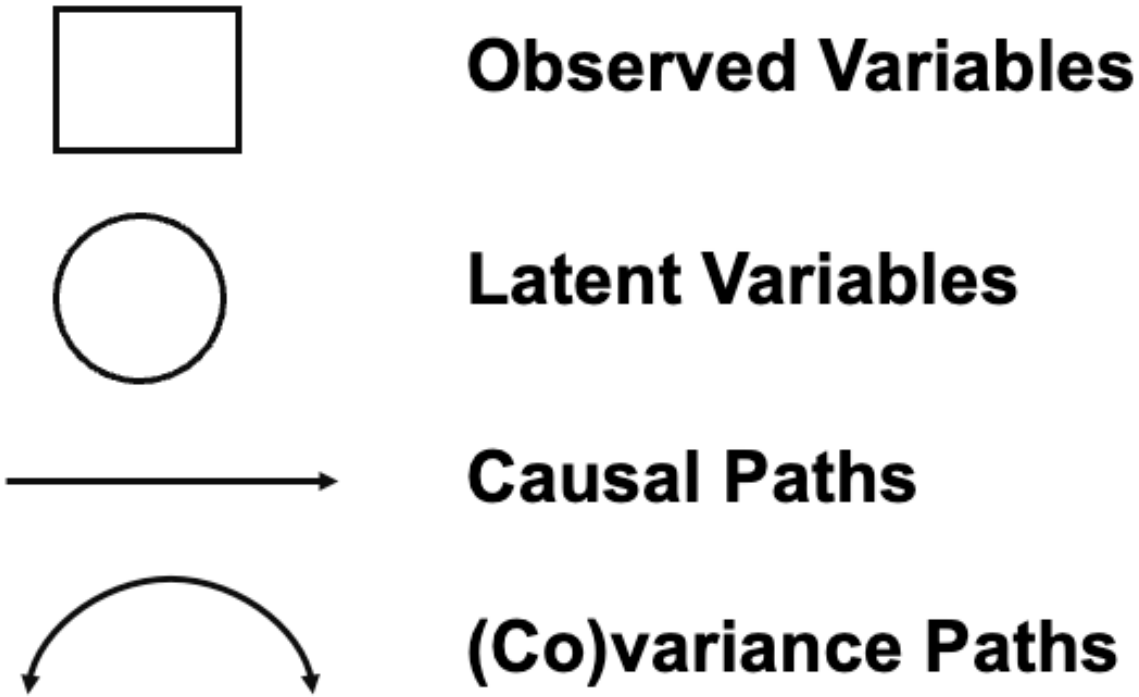




Building Models With Path Diagrams

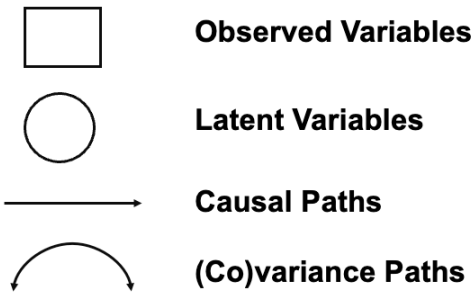
Path diagrams pictorially represent "causal" models.

They aid in deriving the variances and covariances implied by the model.

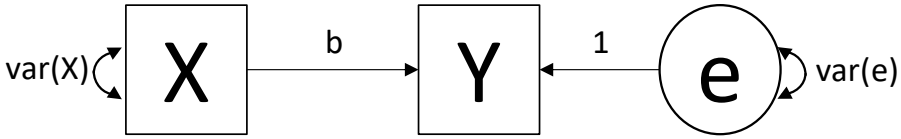


Latent variables are variables that can only be inferred indirectly through a mathematical model from other observable variables that can be directly observed or measured

Linear regression

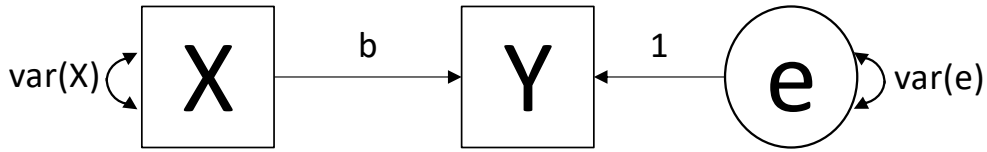


- **Y** = b **X** + **e**
- b is represented as a **path coefficient**
- b quantifies the expected change in Y for every unit change in X



Linear regression - assumptions

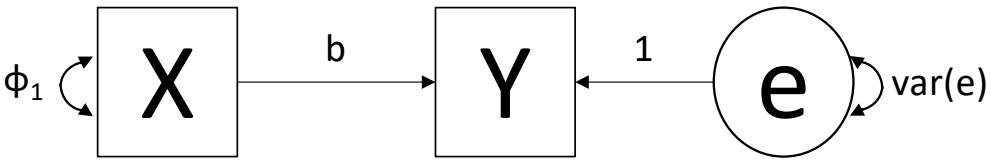
- Y = bX + e (explicit)
- Measurement error (e) in Y (explicit)
- No measurement error in X (explicit)
- No covariance between X and epsilon (explicit)
- Covariance between X and Y is b*var(X) (explicit)
- Linear relationships between the variables (implicit)
- Multivariate normality (implicit)



Back to Identifiability

General rule

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



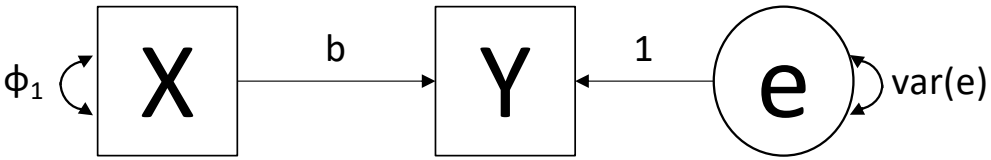
$$\mathbf{Y} = b\mathbf{X} + e$$

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

Number of observed variables: 2
2*3/2 =3

NB: Intercept does not count/matter towards indentifiability

Why n(n+1)/2 ?



$$\mathbf{Y} = b\mathbf{X} + \mathbf{e}$$

Number of observed statistics: 3

Observed Covariance Matrix:

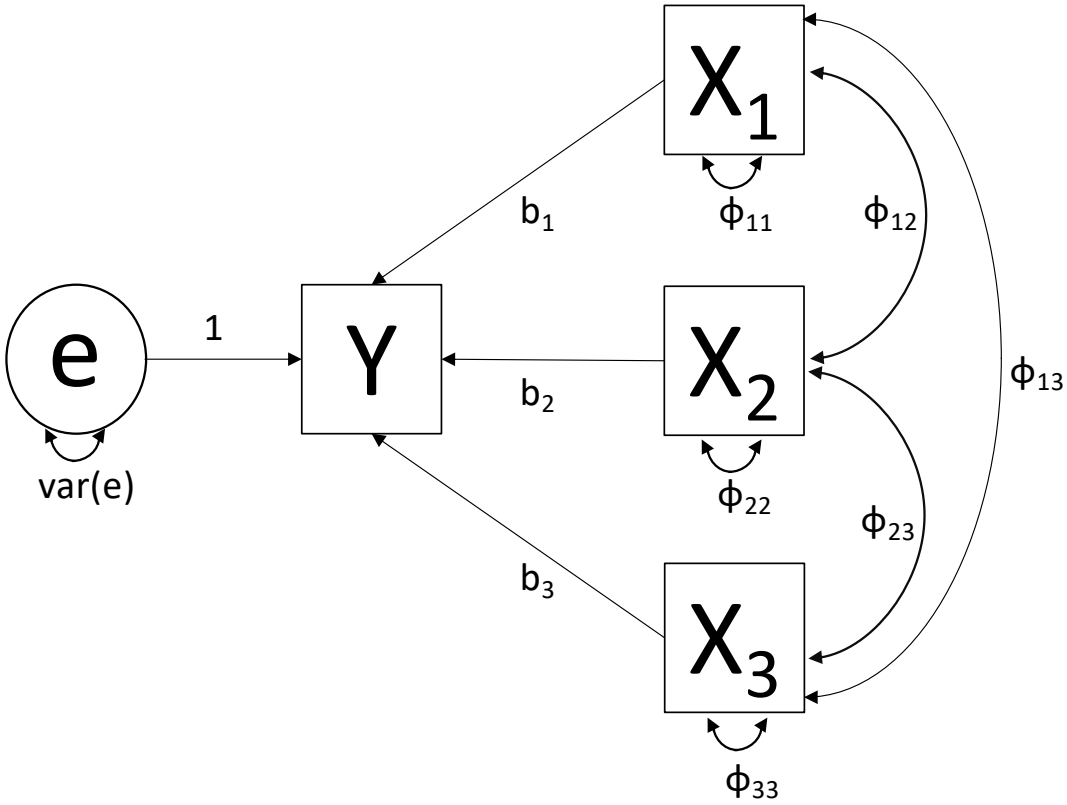
$$\Sigma = \begin{matrix} \text{VAR(X)} & \text{COV(X,Y)} \\ \text{COV(X,Y)} & \text{VAR(Y)} \end{matrix}$$

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

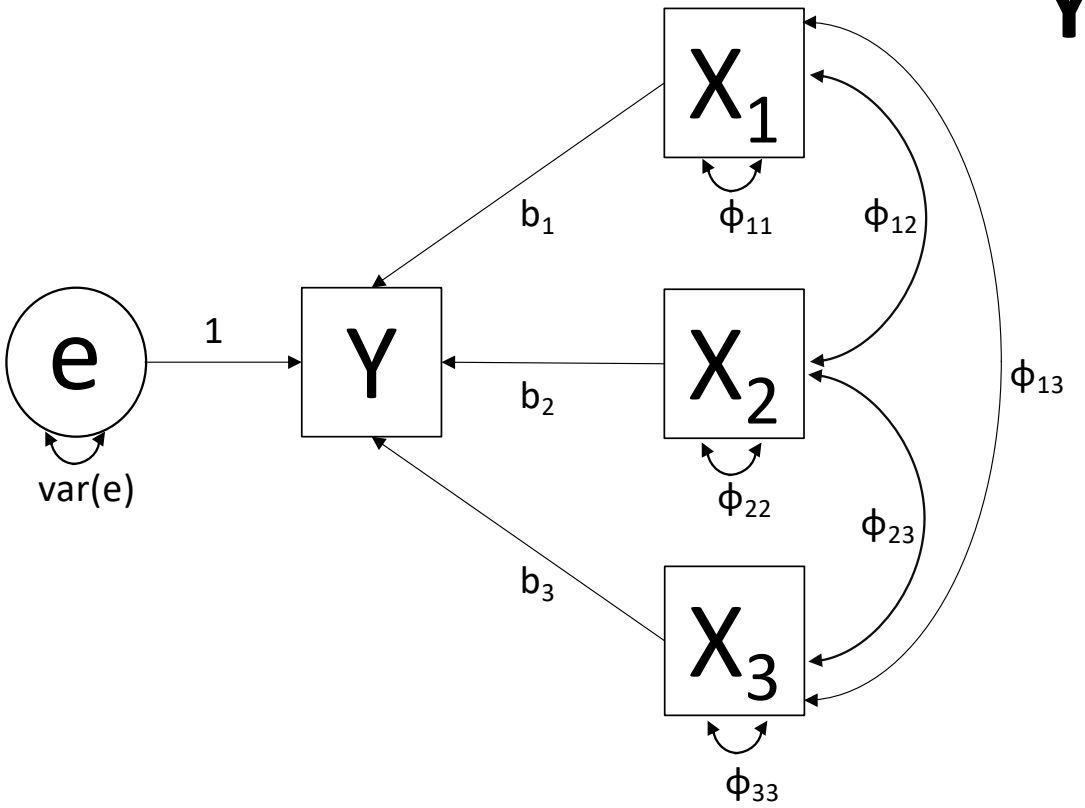
Expected/Implied Covariance Matrix:

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

A more complex model



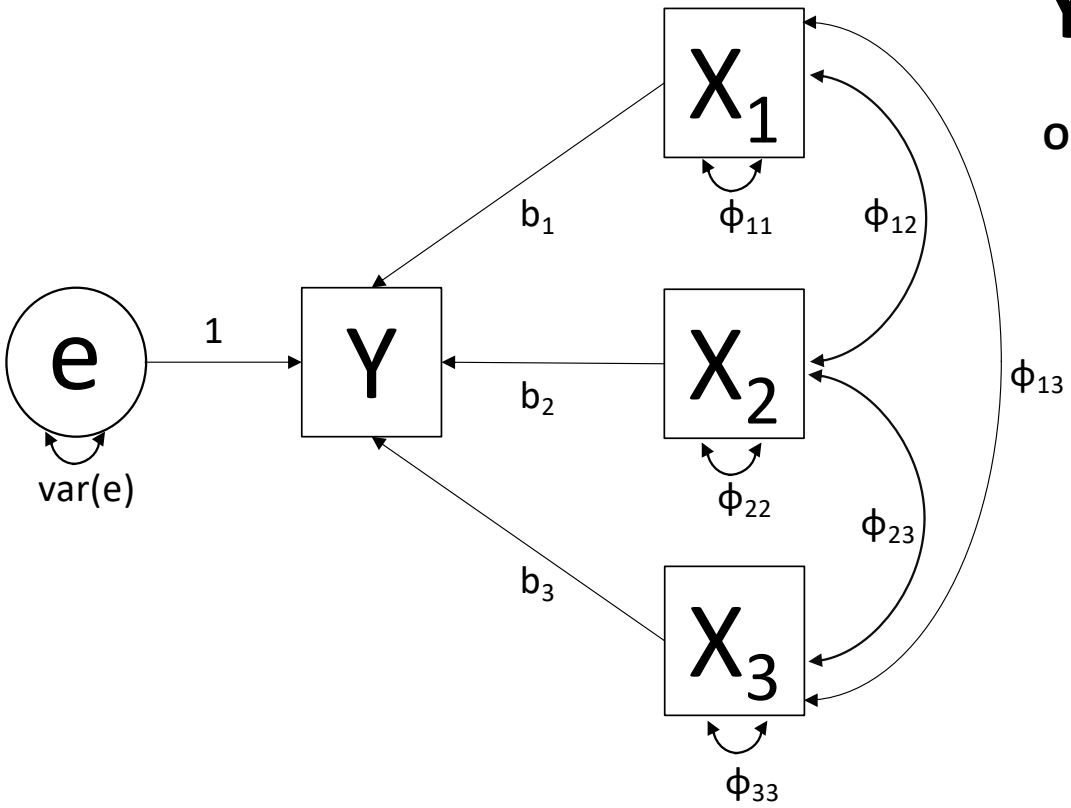
Multivariable (or multiple) Regression



Structural Equation:

$$\mathbf{Y} = b_1\mathbf{X_1} + b_2\mathbf{X_2} + b_3\mathbf{X_3} + \mathbf{e}$$

Multivariable (or multiple) Regression



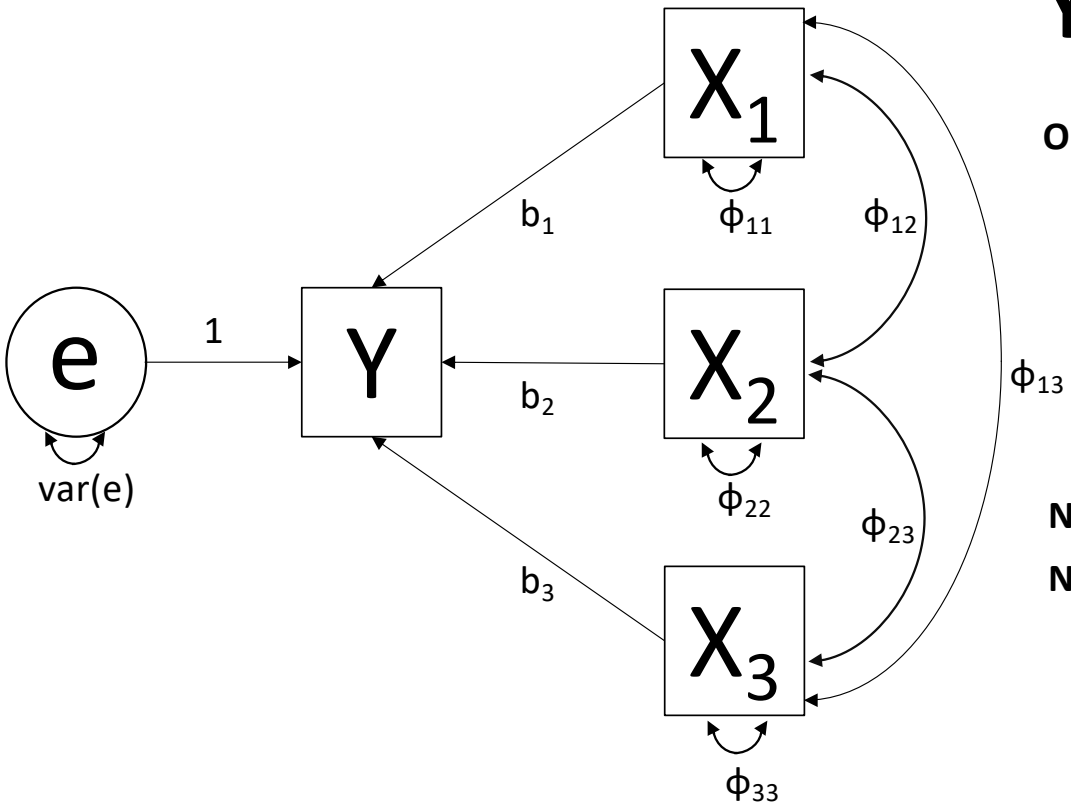
Structural Equation:

$$\mathbf{Y} = b_1\mathbf{X_1} + b_2\mathbf{X_2} + b_3\mathbf{X_3} + \mathbf{e}$$

Observed Covariance Matrix:

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

Multivariable (or multiple) Regression



Structural Equation:

$$\mathbf{Y} = b_1\mathbf{X_1} + b_2\mathbf{X_2} + b_3\mathbf{X_3} + \mathbf{e}$$

Observed Covariance Matrix:

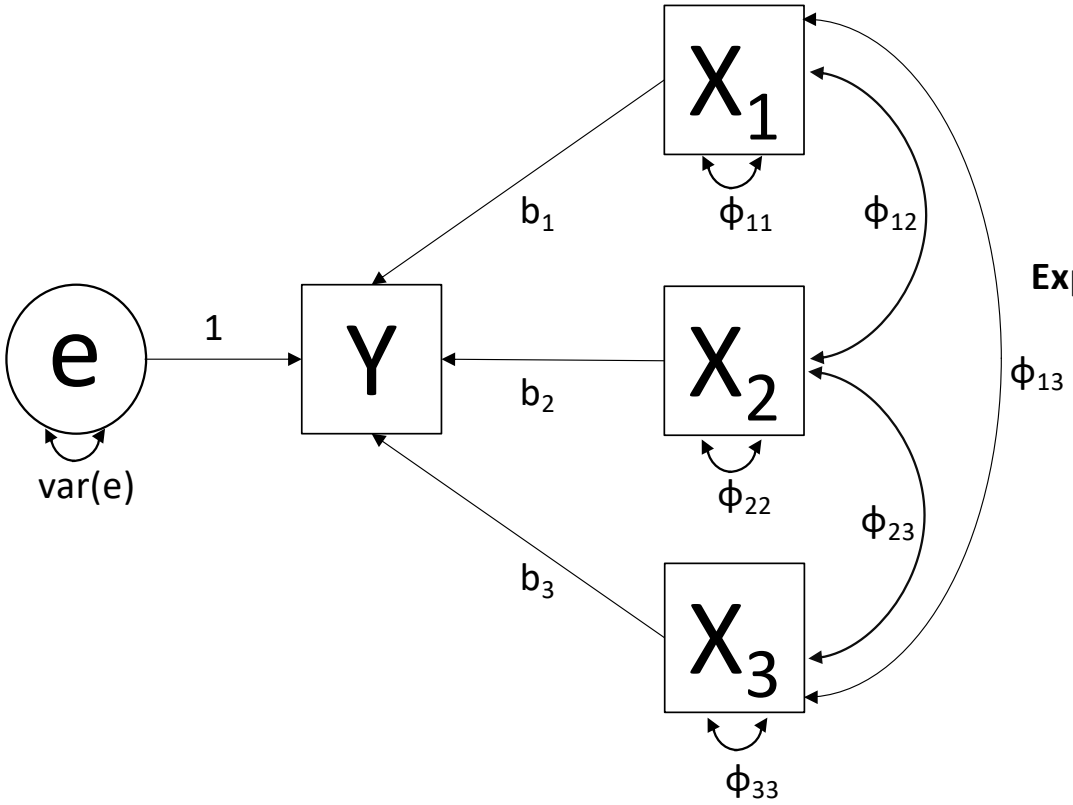
$$\Sigma = \begin{matrix} VAR(X_1) & COV(X_1,X_2) & COV(X_1,X_3) & COV(X_1,Y) \\ COV(X_2,X_1) & VAR(X_2) & COV(X_2,X_3) & COV(X_2,Y) \\ COV(X_3,X_1) & COV(X_3,X_2) & VAR(X_3) & COV(X_3,Y) \\ COV(Y,X_1) & COV(Y,X_2) & COV(Y, X_3) & VAR(Y) \end{matrix}$$

Number of observed statistics: 4*5/2=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)

Multivariable (or multiple) Regression



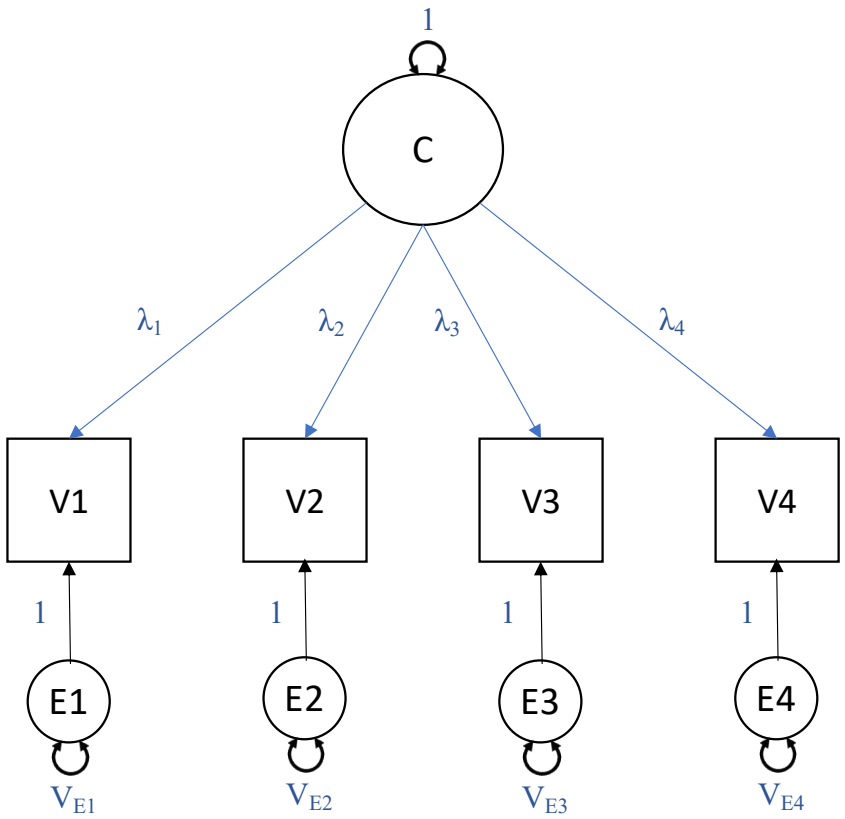
Observed Covariance Matrix:

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

Expected Covariance Matrix:

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

A multivariate model



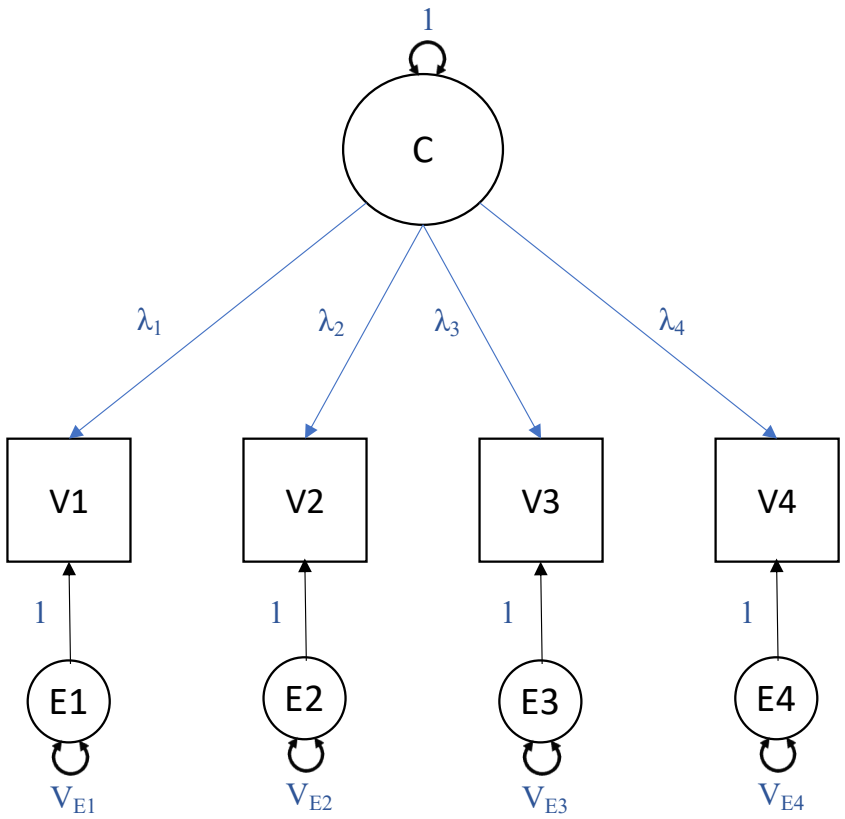
Structural Equations:

Observed Covariance Matrix:

Number of observed statistics:

Number of estimated parameters:

Common Factor Model



Structural Equations:

$V_1 = \lambda_1 C + E_1$

$V_2 = \lambda_2 C + E_2$

$V_3 = \lambda_3 C + E_3$

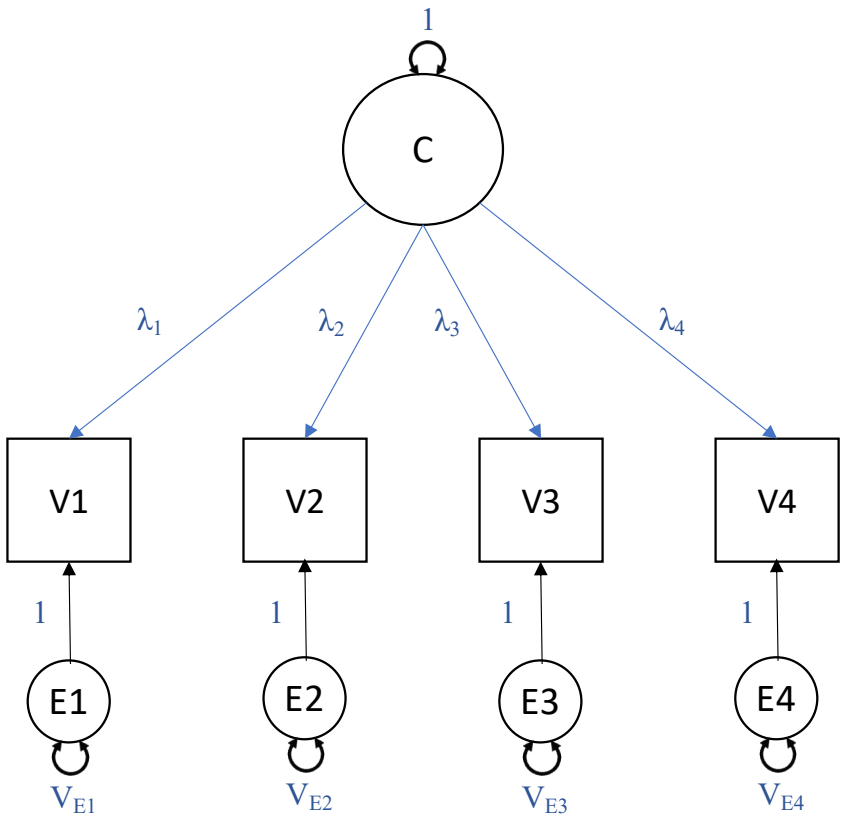
$V_4 = \lambda_4 C + E_4$

Observed Covariance Matrix:

Number of observed statistics:

Number of estimated parameters:

Common Factor Model



Structural Equations:

$V_1 = \lambda_1 C + E_1$

$V_2 = \lambda_2 C + E_2$

$V_3 = \lambda_3 C + E_3$

$V_4 = \lambda_4 C + E_4$

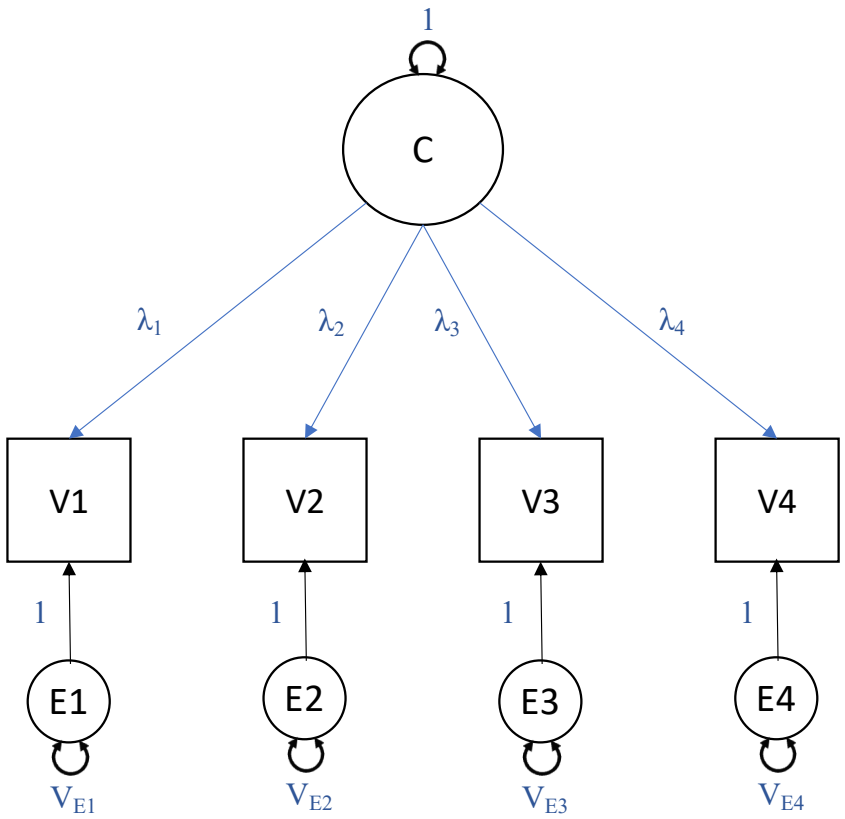
Observed Covariance Matrix:

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

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

Number of estimated parameters:

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:

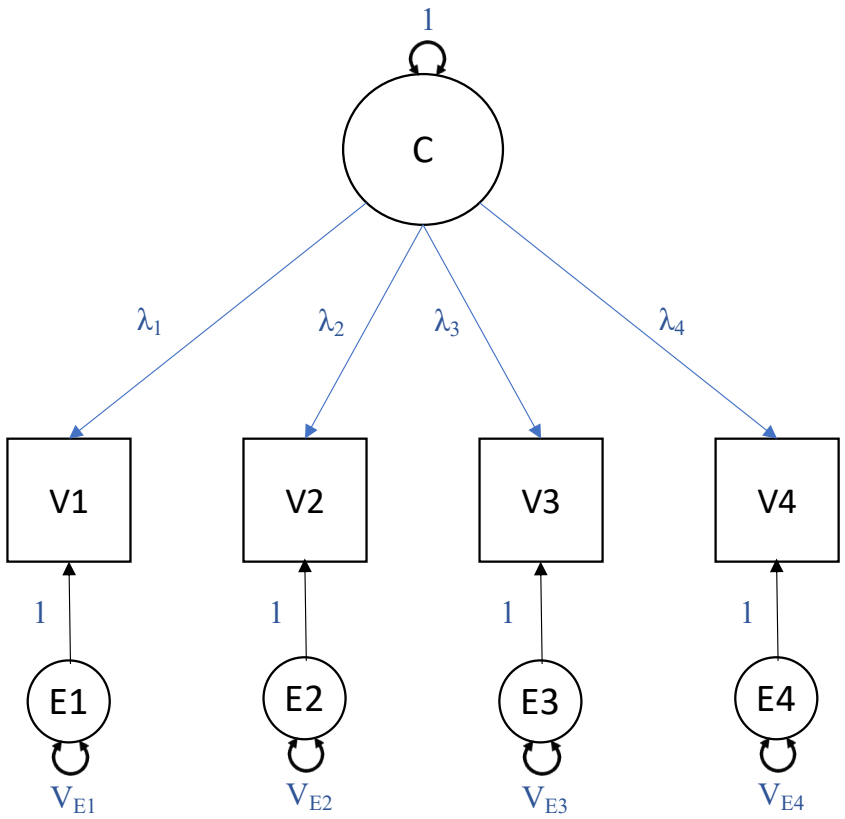
$$\Sigma = \begin{matrix} VAR(V_1) & COV(V_1,V_2) & COV(V_1,V_3) & COV(V_1,V_4) \\ COV(V_2,V_1) & VAR(V_2) & COV(V_2,V_3) & COV(V_2,V_4) \\ COV(V_3,V_1) & COV(V_3,V_2) & VAR(V_3) & COV(V_3,V_4) \\ COV(V_4,V_1) & COV(V_4,V_2) & COV(V_4,V_3) & VAR(V_4) \end{matrix}$$

Number of observed statistics: 10

Number of estimated parameters: 8

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

Common Factor Model



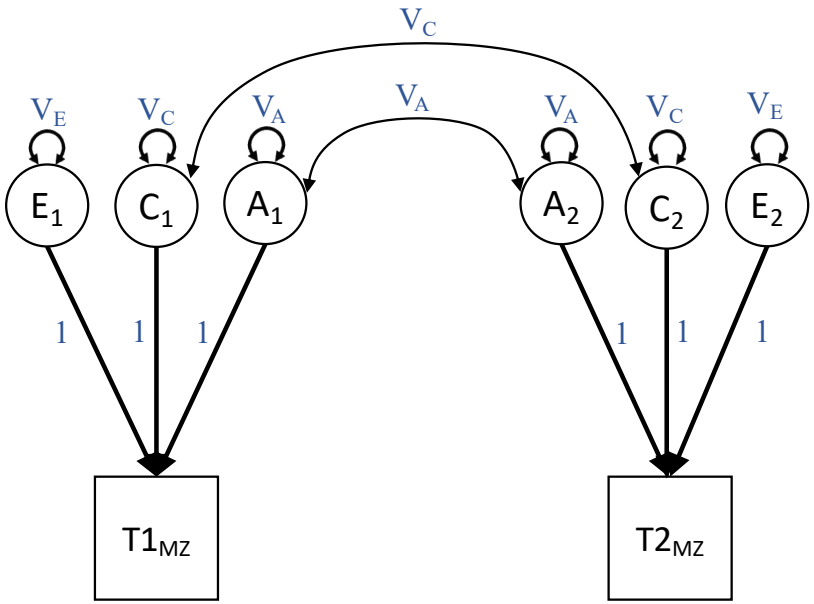
Observed Covariance Matrix:

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

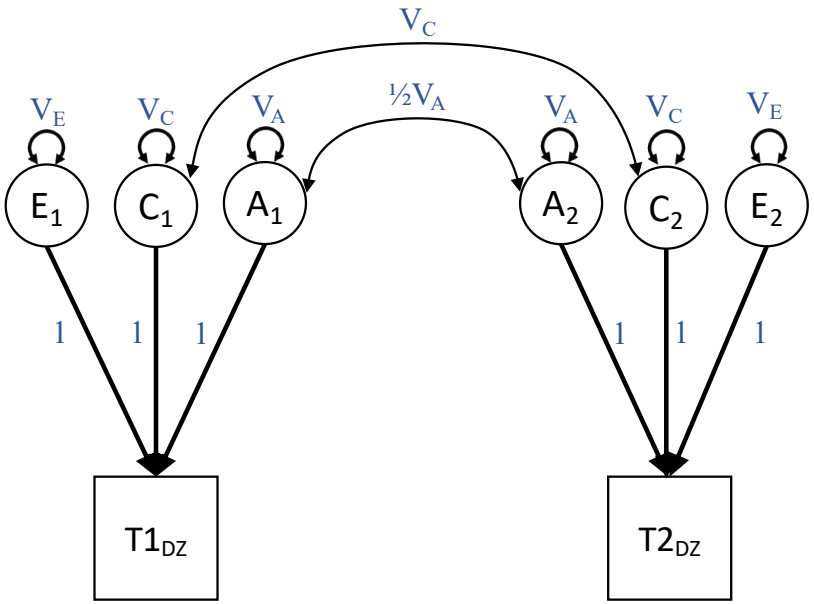
Expected Covariance Matrix:

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

Classical Twin Design



Monozygotic Twins



Dizygotic Twins

Structural Equations:

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

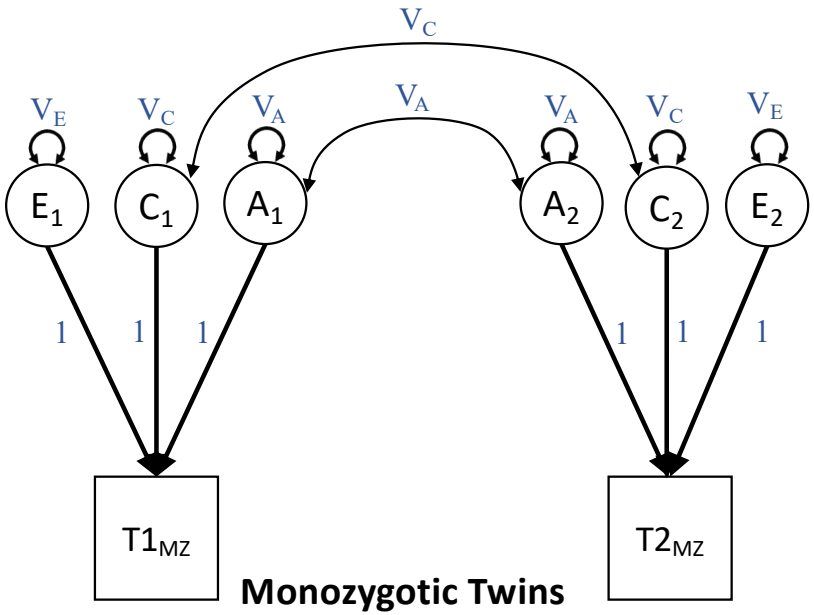
$T2_{MZ} = A_2 + C_2 + E_2$

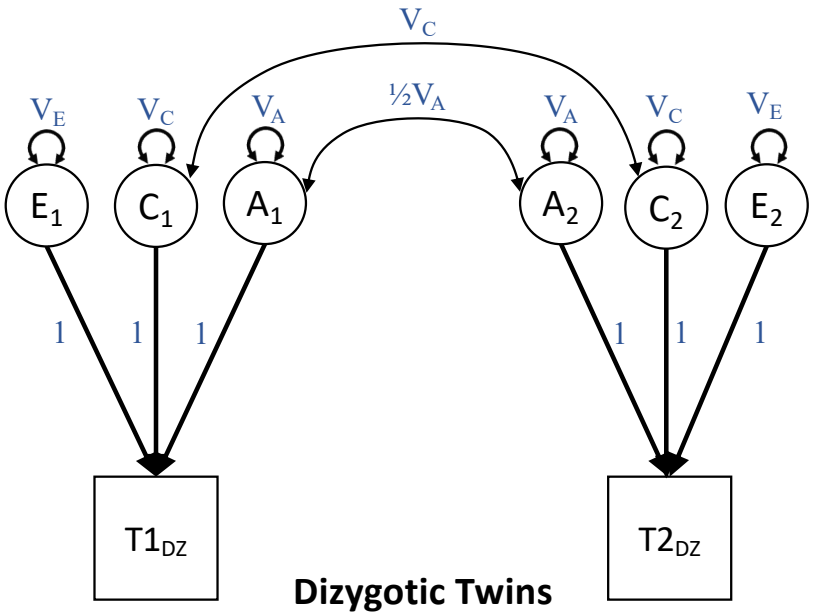
$Cov(A_1,A_2)=V_A$ if MZ pair; $1/2V_A$ if DZ
$Cov(C_1, C_2)=V_C$

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

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

Classical Twin Design





Expected Covariance Matrices:

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

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

Latent variables are random effects

Structural Equations:

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

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

$Cov(A_1,A_2)=V_A$ if MZ pair; $1/2V_A$ if DZ
$Cov(C_1, C_2)=V_C$

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

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

$T1_{MZ}$ $T2_{MZ}$ $T1_{DZ}$ $T2_{DZ}$

Model reformulation

$T1 = A_1 + C_1 + E_1$

$T2 = A_2 + C_2 + E_2$

$Cov(A_1,A_2)=V_A$ if MZ pair; $1/2V_A$ if DZ
$Cov(C_1, C_2)=V_C$

T1 T2

Model reformulation as a "classical" random effect model

T = A + C + E
A ~ N(0, GRM. VA)
C ~ N(0, CRM. VC)
E ~ N(0, I. VE)

GRM and CRM – matrices of variance – covariances of individuals for A and C.

T

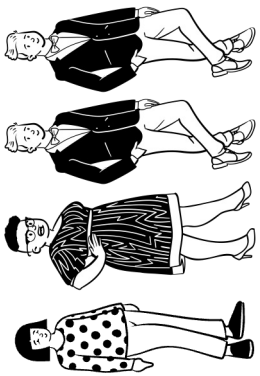
Latent variables are random effects

$Cov(A_1,A_2)=V_A$ if MZ pair; $1/2V_A$ if DZ
$Cov(C_1, C_2)=V_C$

Covariance of A between individuals is 0 unless they are a twin pair, in which case it is Va or 1/2Va depending on zygosity.

T = A + C + E
A ~ N(0, GRM. VA)
C ~ N(0, CRM. VC)
E ~ N(0, I. VE)

GRM =



1			
1	1		
0	0	1	
0	0	1/2	1

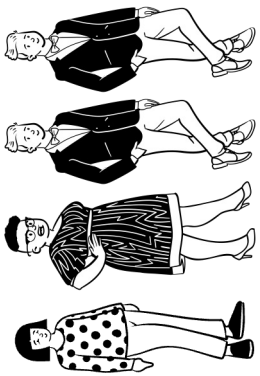
Latent variables are random effects

$Cov(A_1,A_2)=V_A$ if MZ pair; $1/2V_A$ if DZ
$Cov(C_1, C_2)=V_C$

Covariance of C between individuals is 0 unless they are a twin pair, in which case it is VC.

T = A + C + E
A ~ N(0, GRM. VA)
C ~ N(0, CRM. VC)
E ~ N(0, I. VE)

CRM =



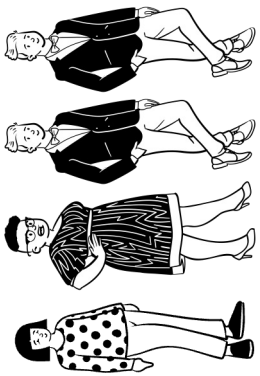
1			
1	1		
0	0	1	
0	0	1	1

Latent variables are random effects

Replace pedigree information (expected relatedness coefficients between related individuals) with observed cryptic relatedness in general population

T = A + E
A ~ N(0, GRM. VA)
E ~ N(0, I. VE)

GRM =

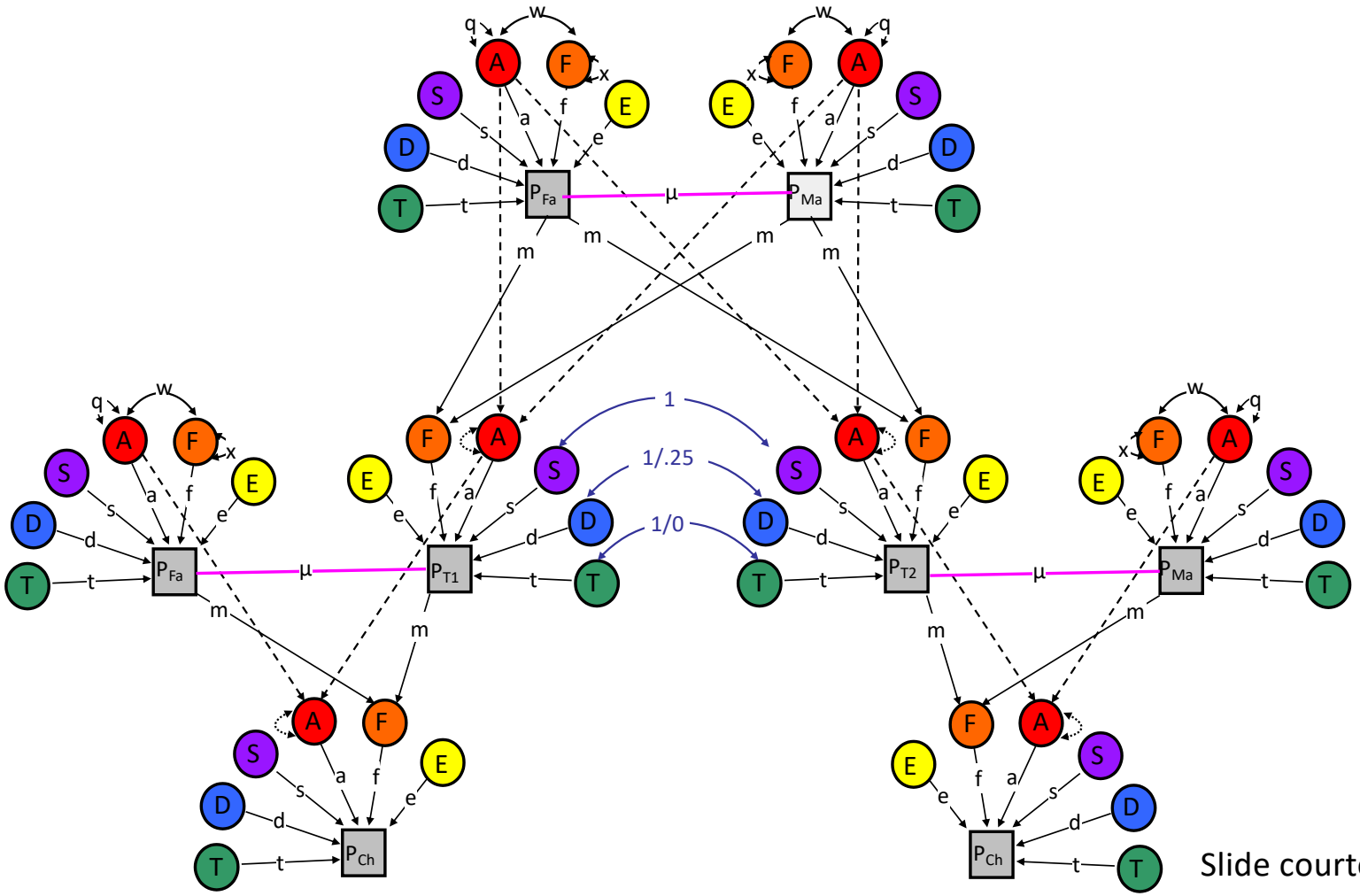


1			
0.01	1		
-0.02	0.007	1	
0.009	-0.006	0.01	1

"GREML
Or GCTA model"

VA additive variance attributable to common SNPs

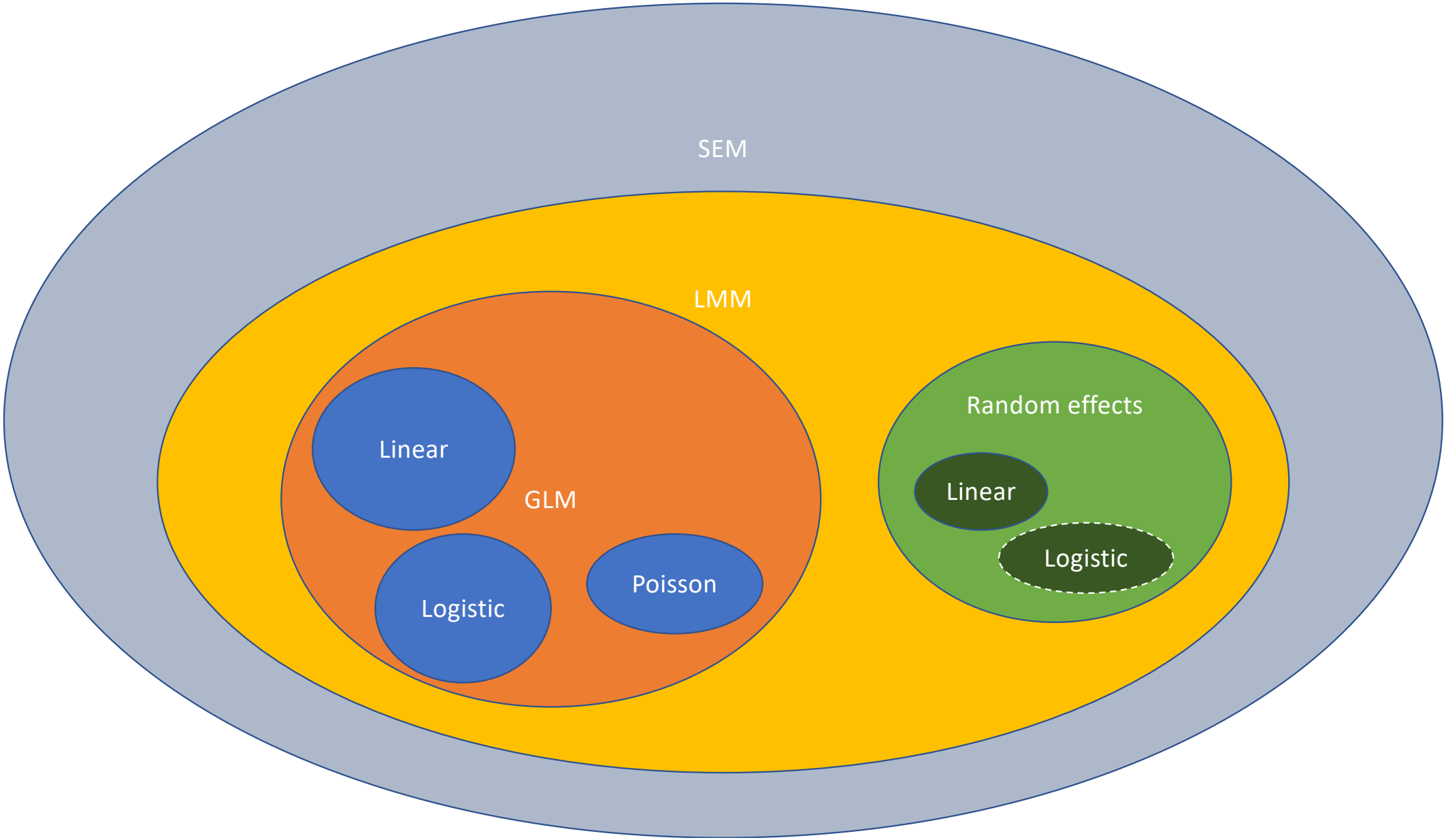
Extended Twin Design – multi generational



Slide courtesy of Matt Keller

Other Path Models

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


SEM
LMM
GLM
Linear
Logistic
Poisson
Random effects
Linear
Logistic

Model	Formula	Application	In R
Generalised Linear model	**Y** = **Za** + **e**	Test association / correlation Haseman Elston regression	Lm() Glm()
Random effect model	**Y = Xb + e** b~N(0, I. sG2 / 2)	AE or ACE model Longitudinal model Model site effect	Lme4() openMx() heritability() qgg()
Linear Mixed Model	**Y = Za + Xb + e** b~N(0, I . sG2 / p)	ACE with covariates SNP h2 with covariates Longitudinal with covariates Quadratic, interactions More…	Lme4() nlme() openMx() Umx() heritability() qgg()
Structural equation modelling	Set of GLM or LMM	Complex multivariate LMMs models Genetic correlation (rG) Comnnon pathway / independent pathway	lavaan() openMx()

Summary

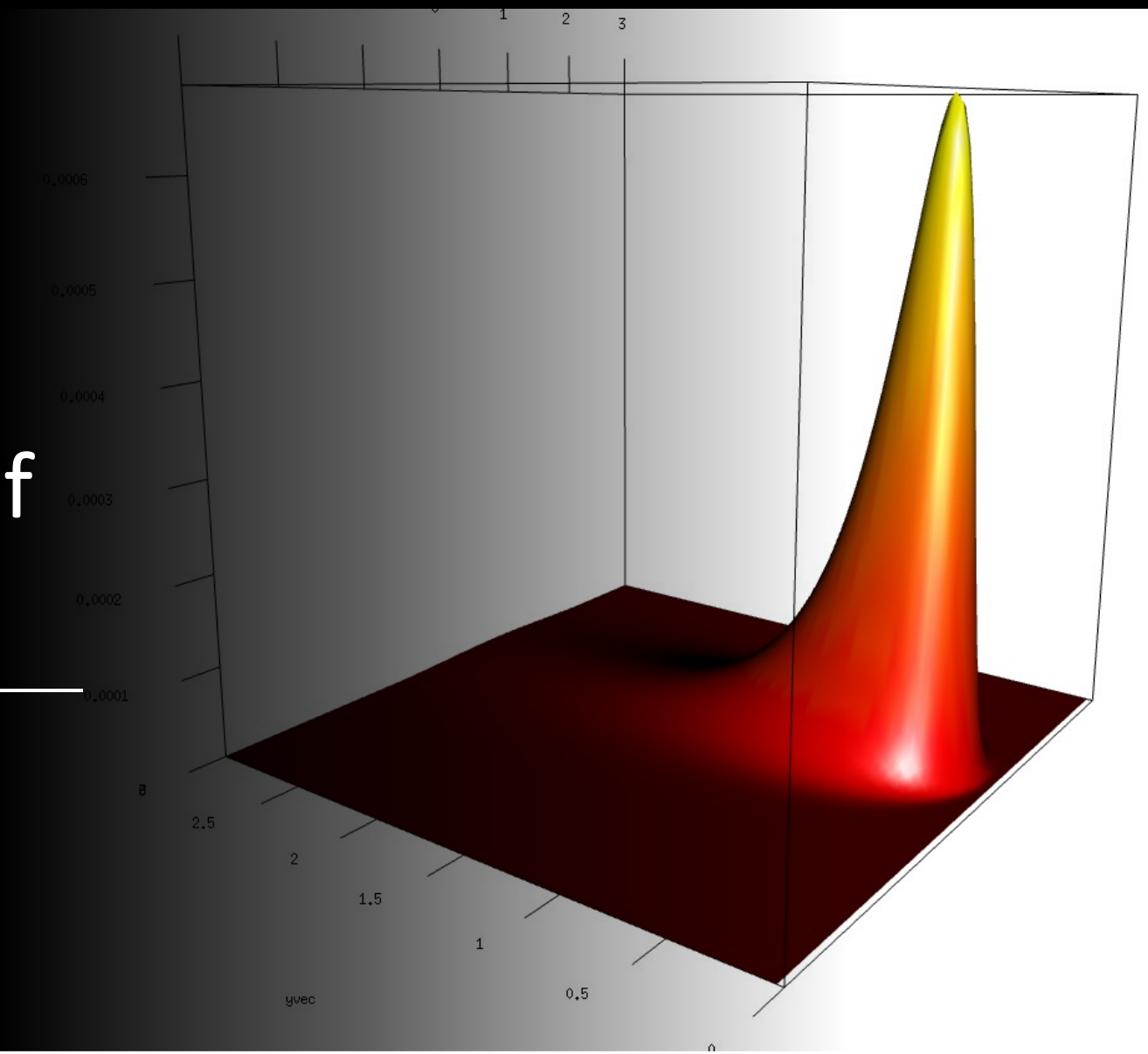
SEM and packages are very powerful

- extremely general and flexible
- but for simple(r) models (e.g. GLM, simple LMM) other packages may be more straighforward
- use of SEM is quite field dependent

Path diagram can help visualise and explain model

- can be complex to get right
- may not fit on one page
- often used with incorrect formalism – to improve readability

Estimating parameters of the model



Likelihood (function)

The likelihood function (often simply called the likelihood) is the **joint probability** of the **observed data** viewed as a **function of the parameters of a statistical model.**

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

Assuming observations are i.i.d

It is not a probability density over the parameter θ
It is not the posterior probability of θ given the data x

Likelihood (function) - example

θ: *probability of heads*

x : $(head, heads, tails)$

$$\mathcal{L}(\theta \mid x) = \prod_{j=1}^{N} P_\theta\left(x_j\right) = P_\theta\,(head).P_\theta\,(head).P_\theta\,(tails)$$

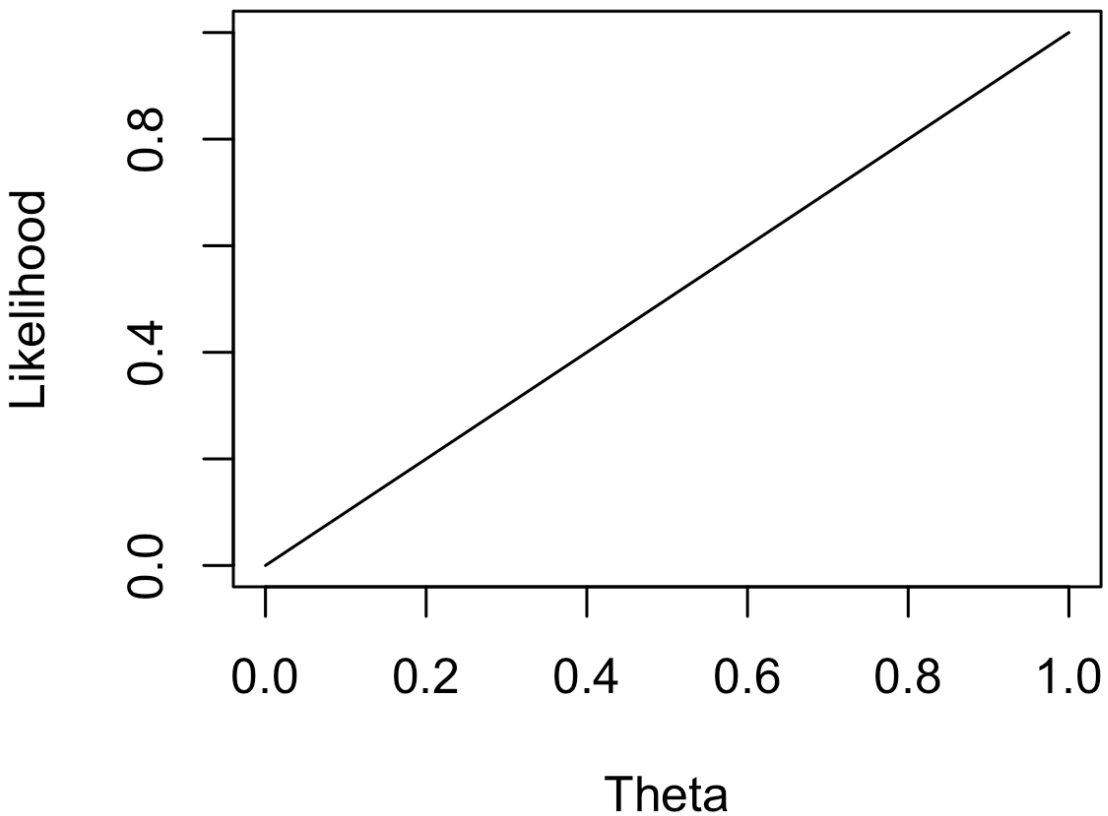
$$= \theta\,.\,\theta\,.\,(1\text{-}\,\theta)$$

Likelihood (function) – visual example

θ: *probability of heads*

x : *head*

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

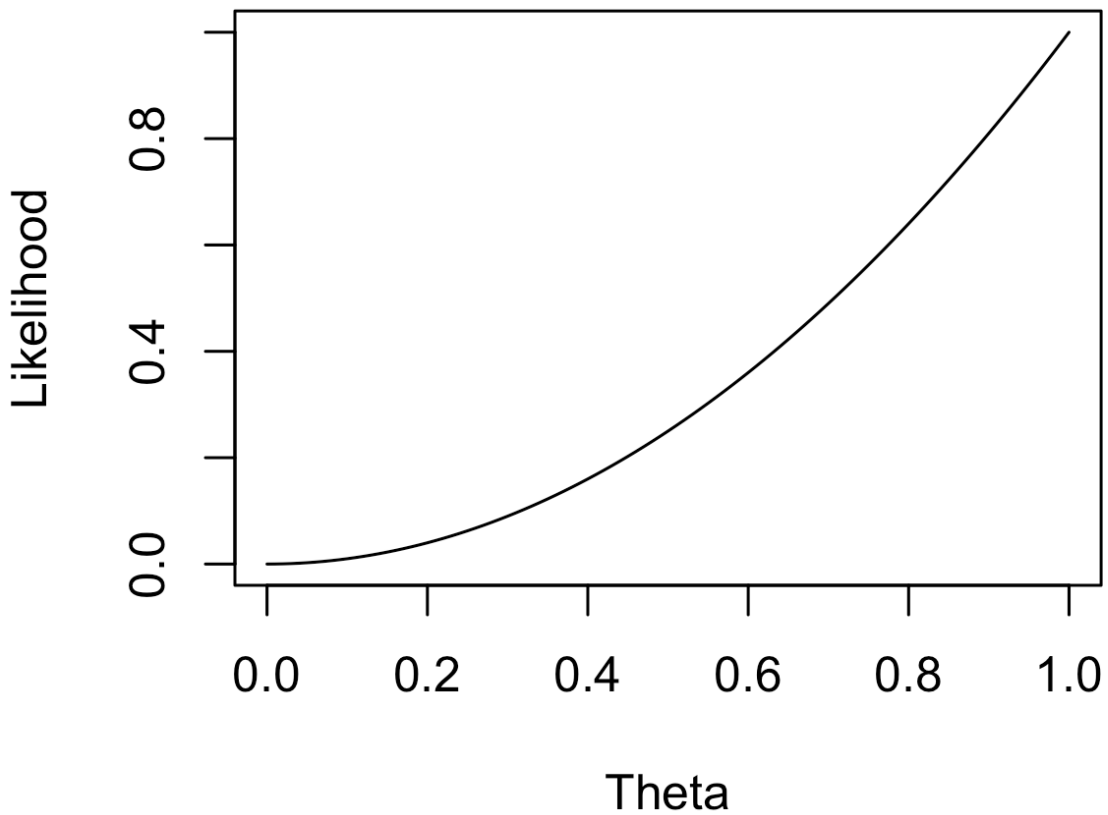


Likelihood (function) – visual example

θ: *probability of heads*

x : *heads, heads*

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

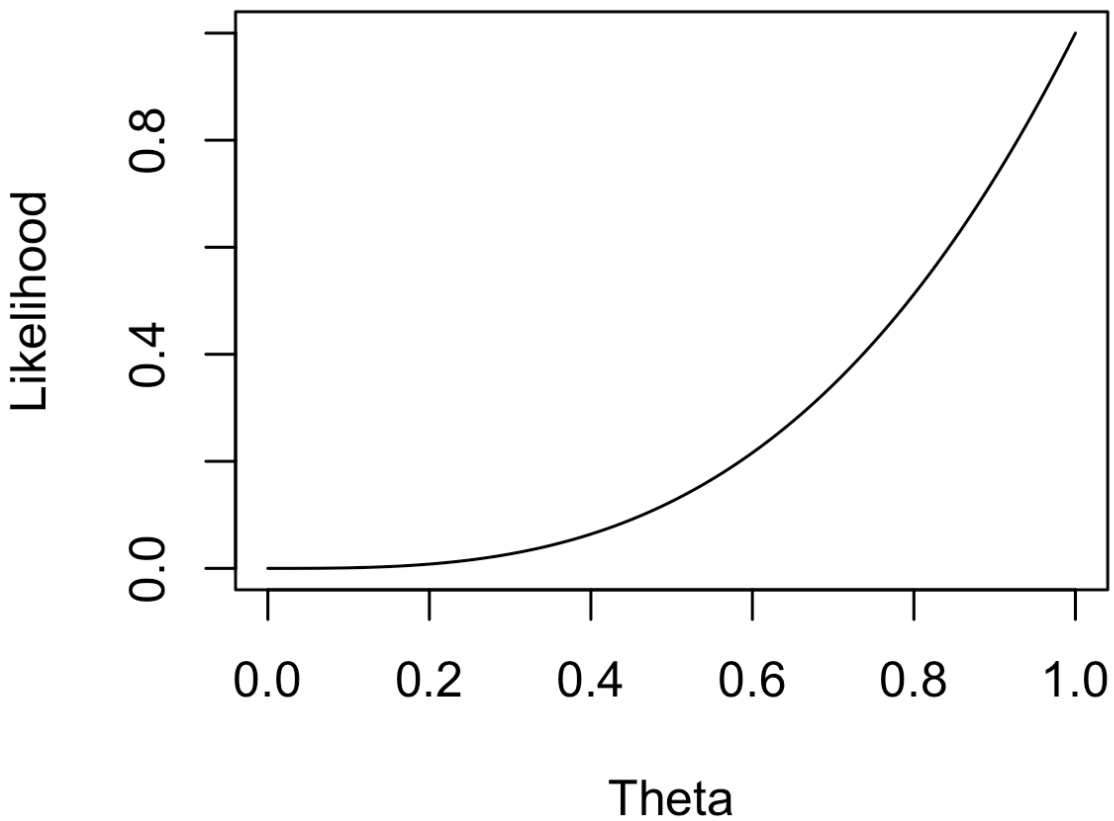


Likelihood (function) – visual example

θ: *probability of heads*

x : *heads, heads, heads*

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

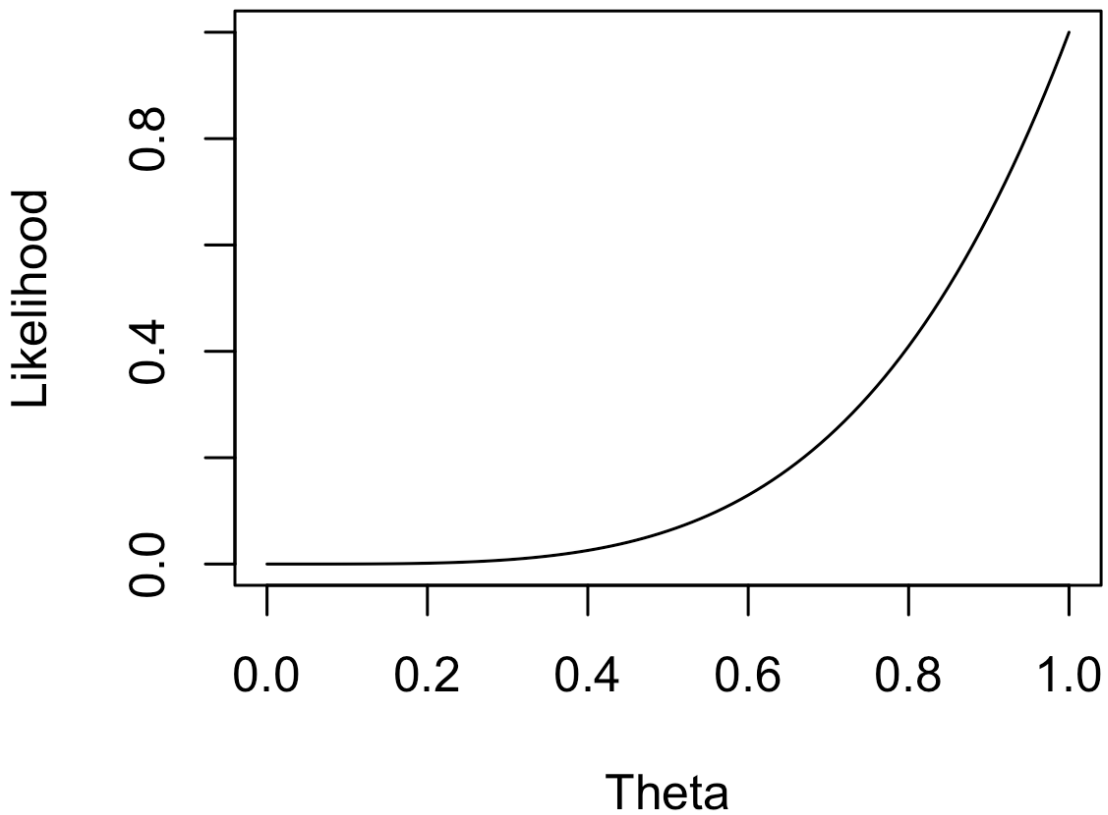


Likelihood (function) – visual example

θ: *probability of heads*

x : *heads, heads, heads, heads*

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

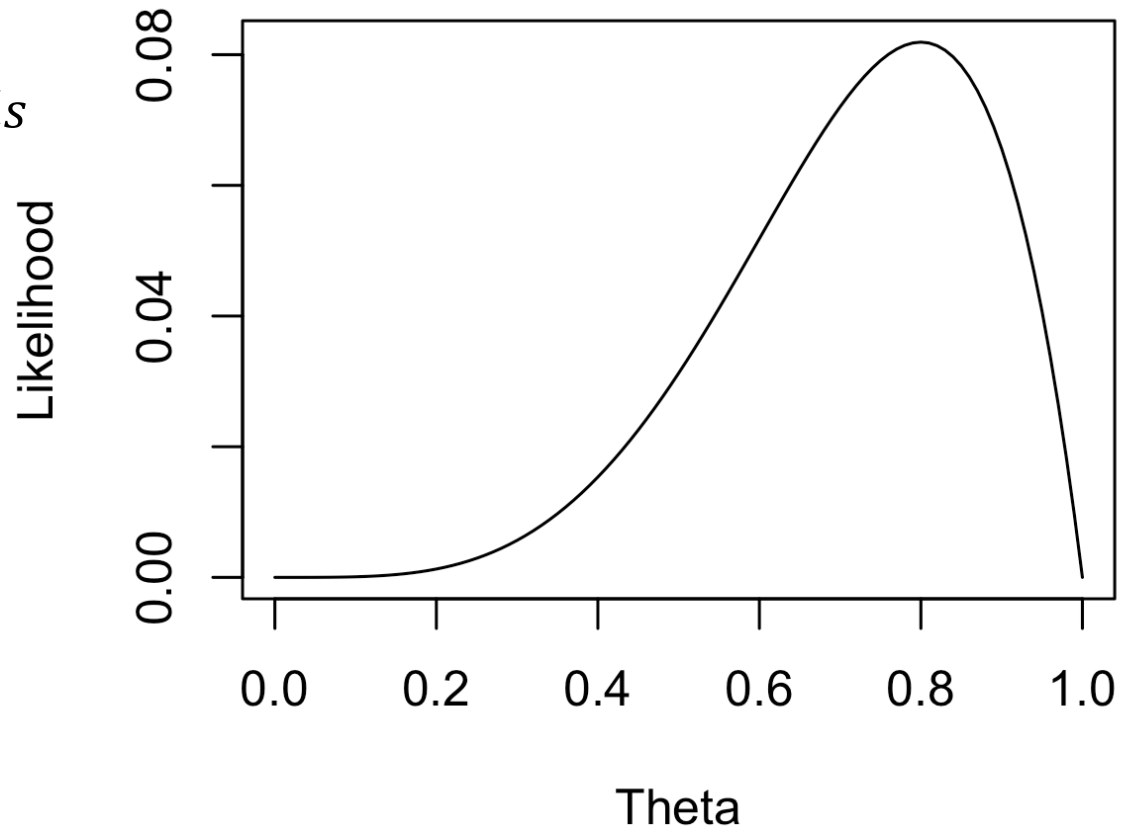


Likelihood (function) – visual example

θ: *probability of heads*

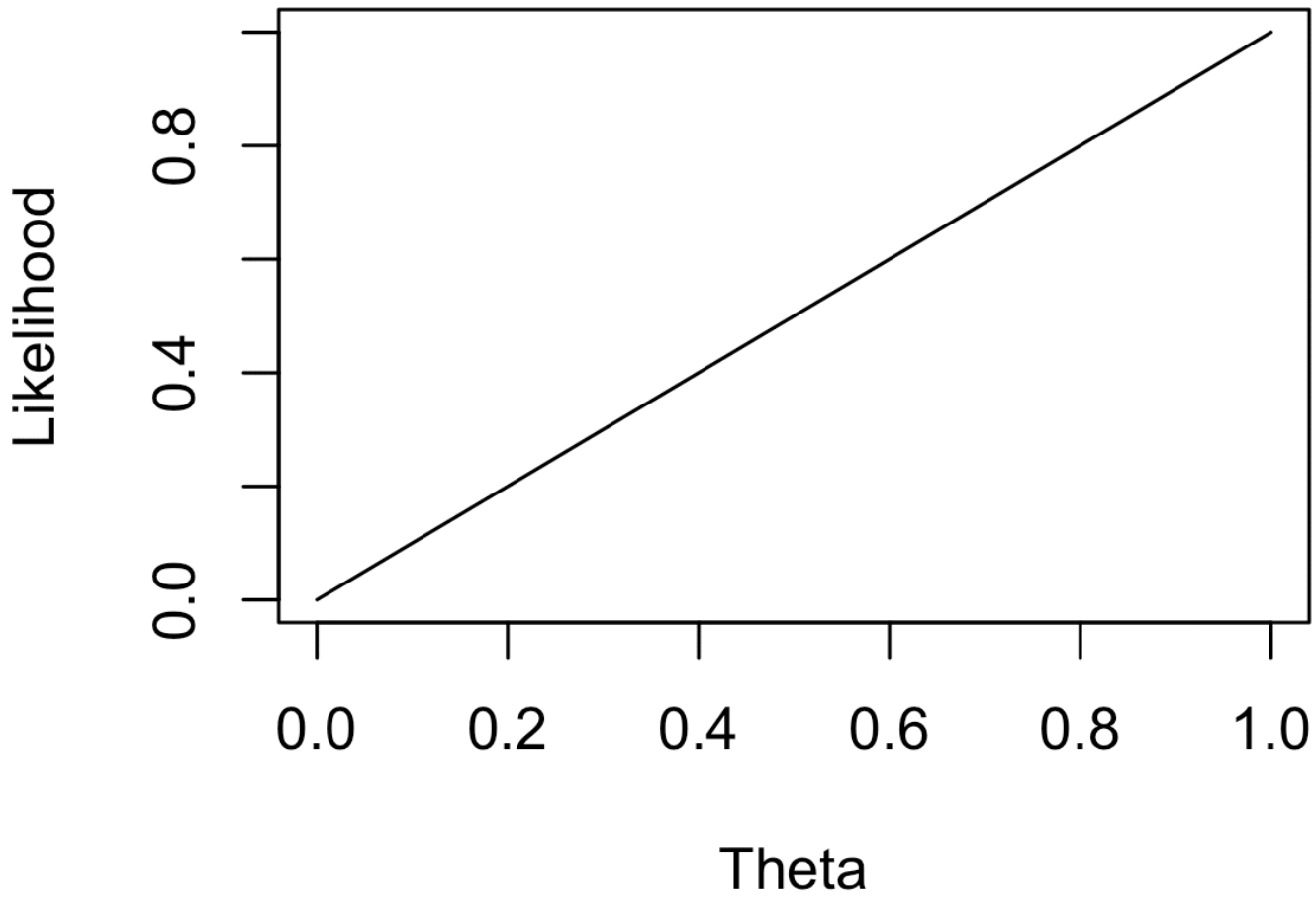
x : *heads, heads, heads, heads, tails*

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



Likelihood (function) – visual example

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

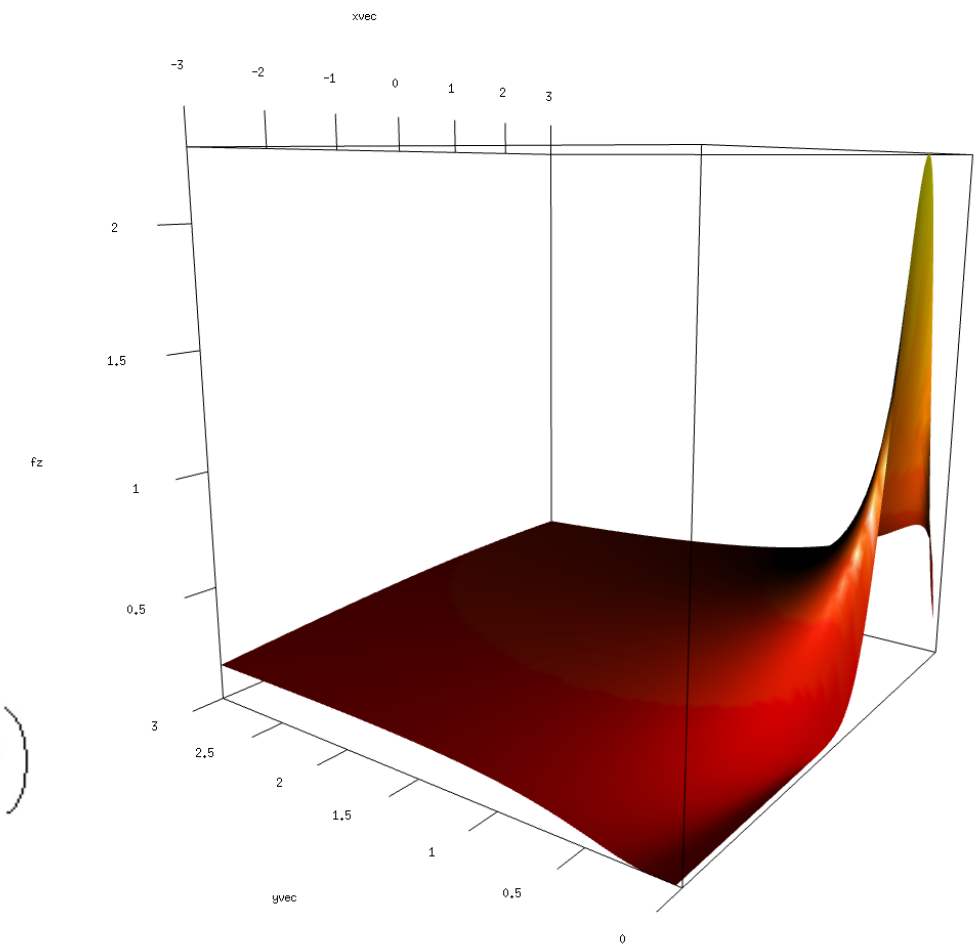


Likelihood (function) – of linear model

$$y = X\beta_0 + \varepsilon$$

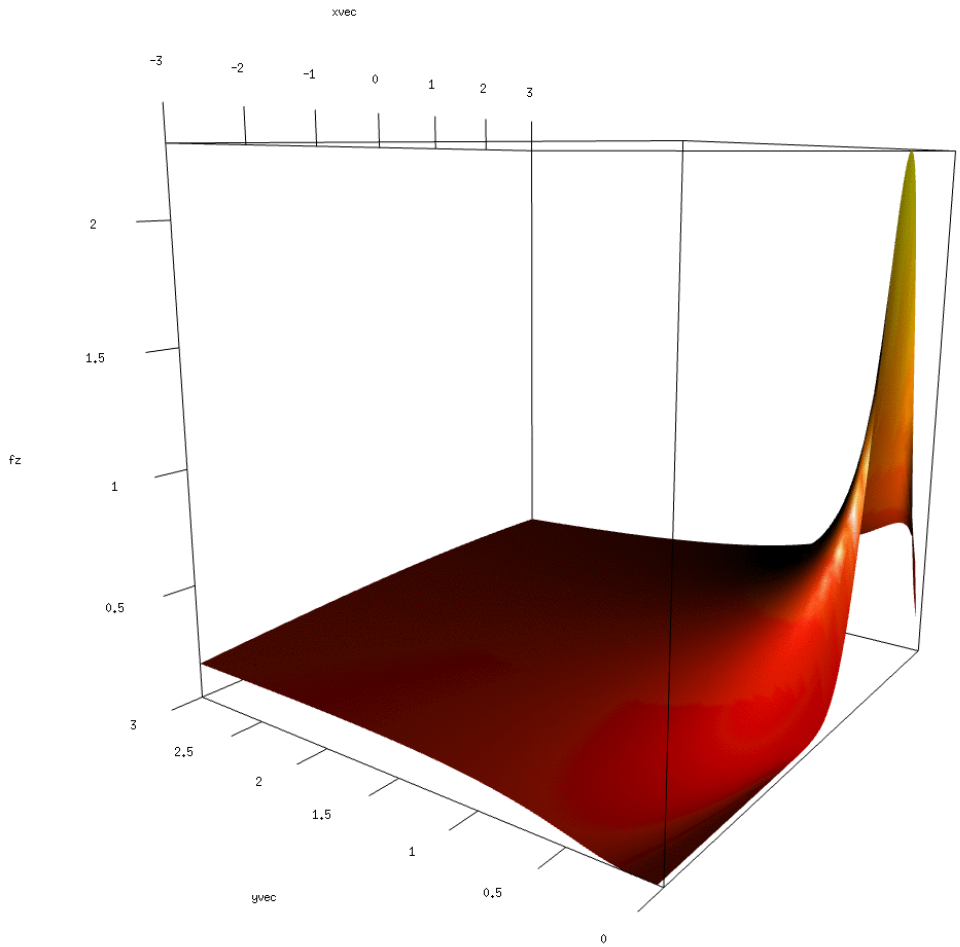
$$\sigma_0^2 = \mathrm{Var}[\varepsilon_i | X]$$

$$L(\beta, \sigma^2; y, X) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x_i\beta)^2 \right)$$



Likelihood (function) – of linear model

Evolution of the likelihood function as we add more data (from 1 to 30 observations)



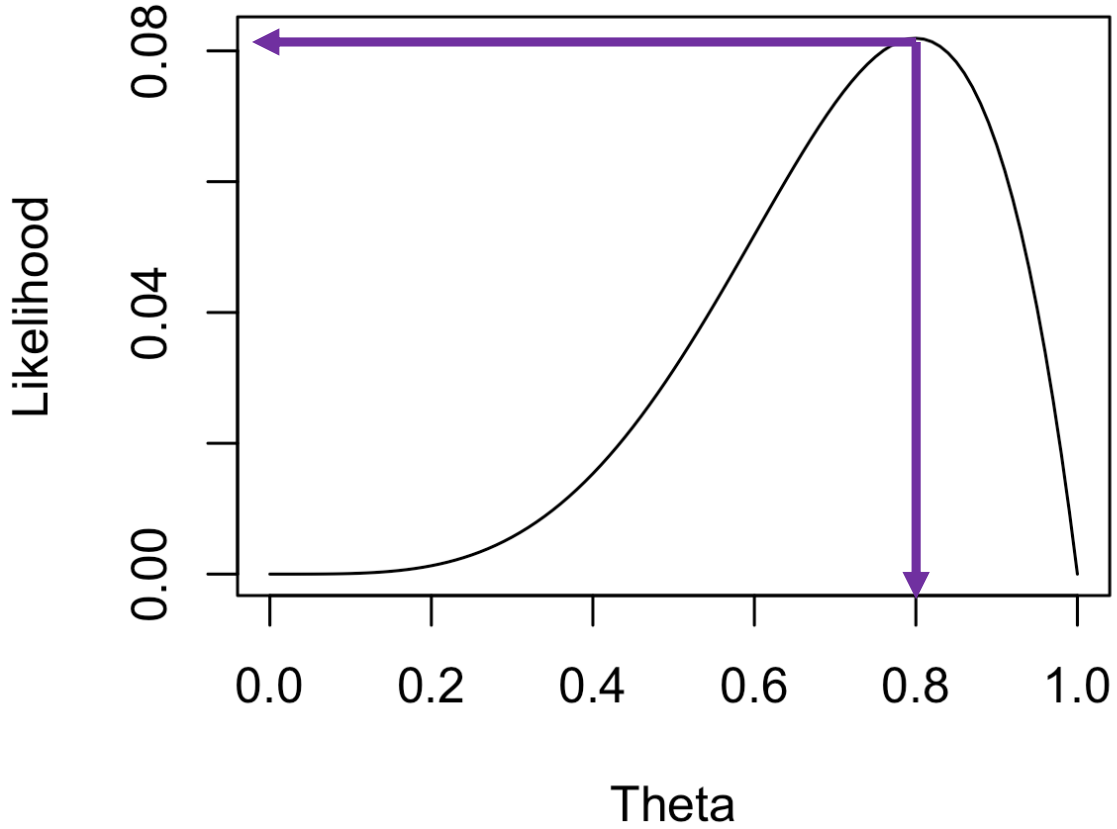
Maximum Likelihood estimate

θ: *probability of heads*

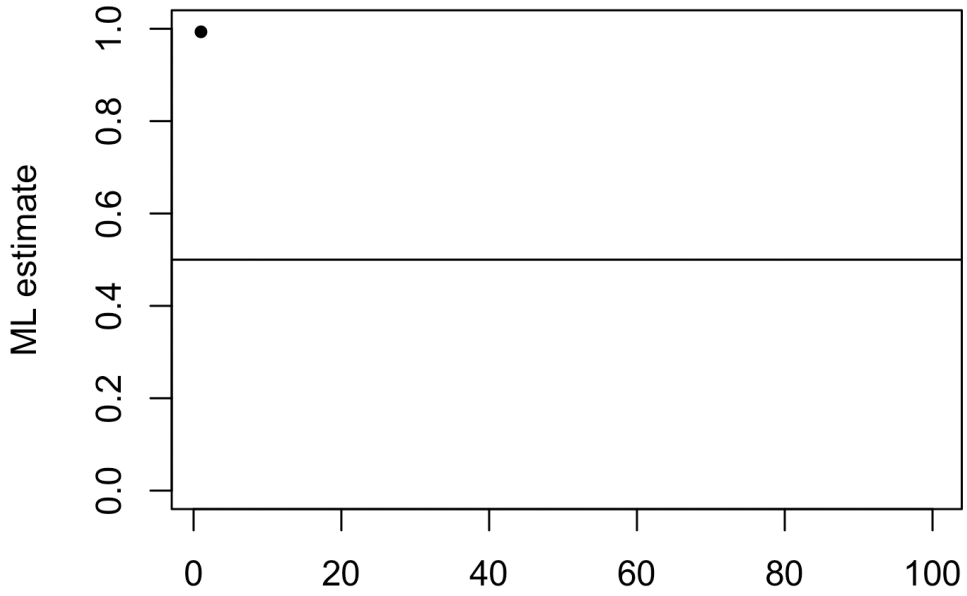
x : *heads, heads, heads, heads, tails*

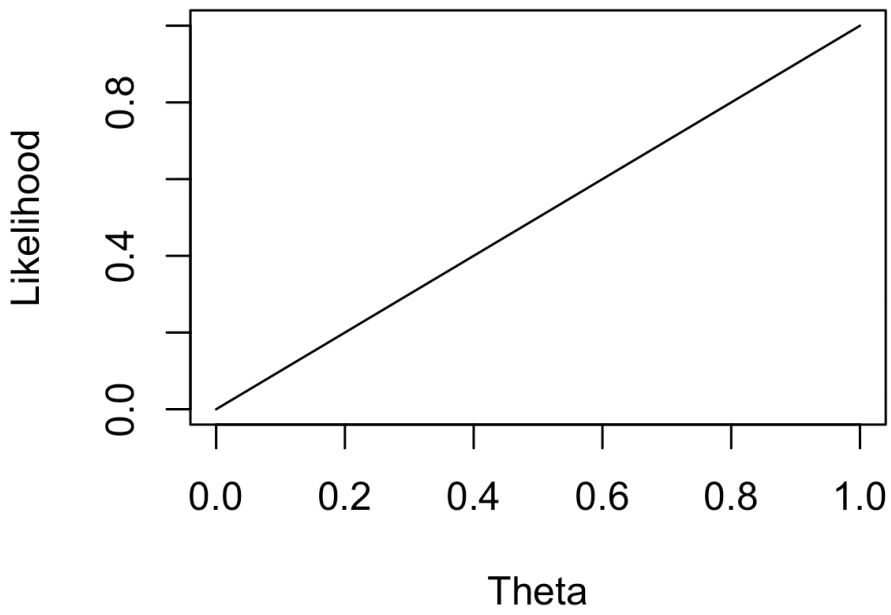
$\mathcal{L}(\theta \mid x) = \theta . \theta . \theta . \theta . (1 - \theta)$

Parameter that maximises the probability of the observed data



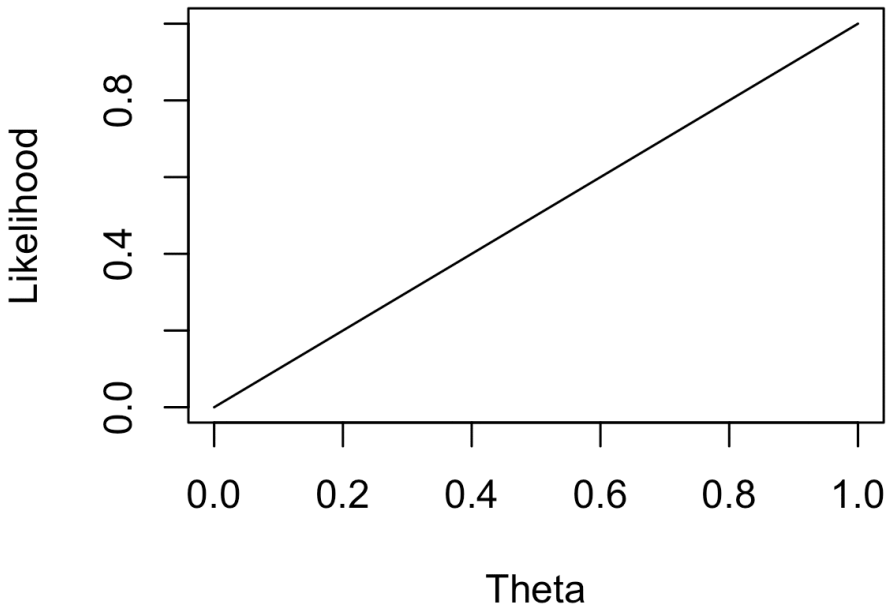
Maximum Likelihood estimate

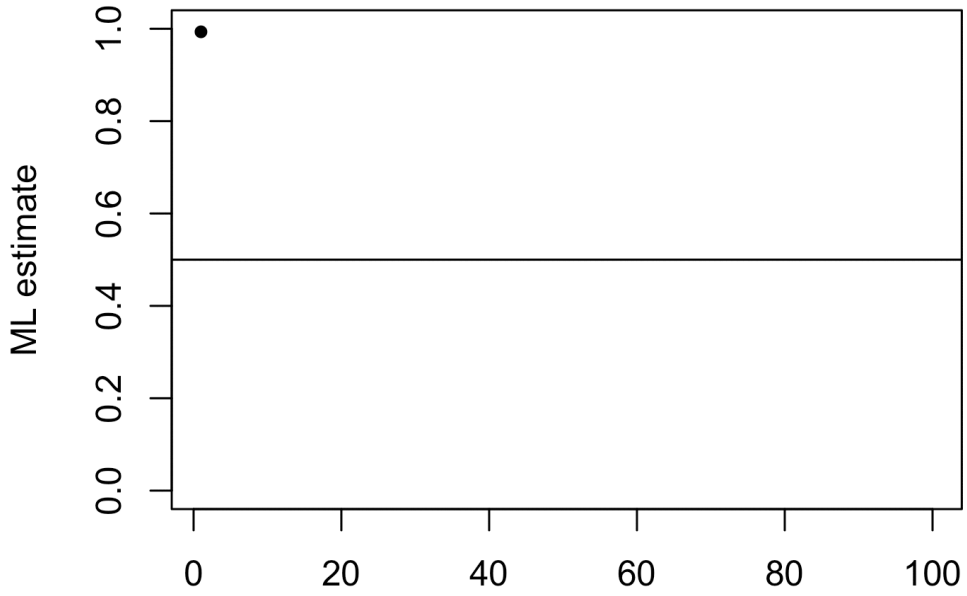




For each new data point
The likelihood function gets updated
And the ML estimate gets updated

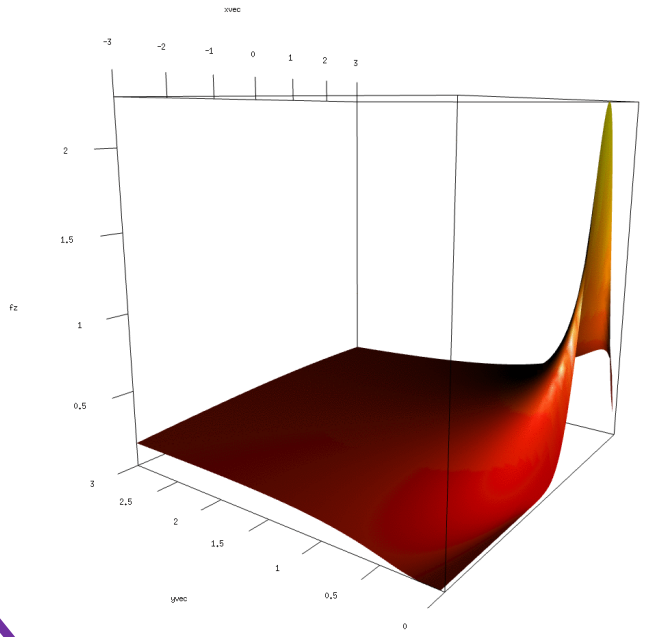
Maximum Likelihood estimate

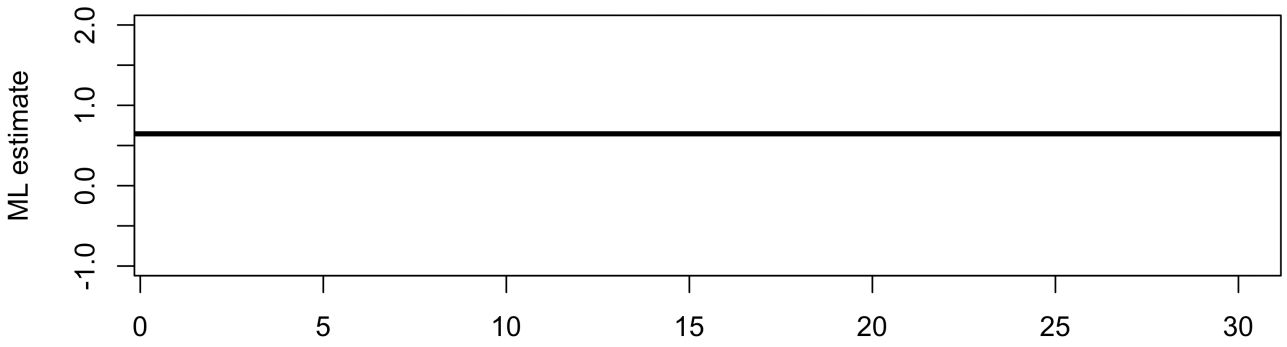




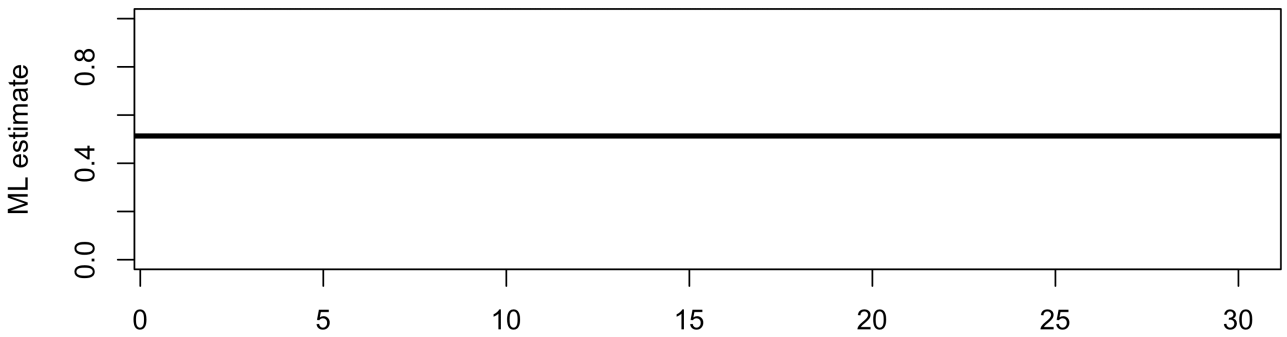
- <u>Asymptotically</u> unbiased
- Consistent
- Efficient
- Scale Invariant
- Sampling distribution of estimates is asymptotically normal

Maximum Likelihood estimate





Beta



Var(e)

Optimization

Maximum likelihood estimates can sometimes be solved in closed form

MLE of coin toss = Number of heads / number of toss

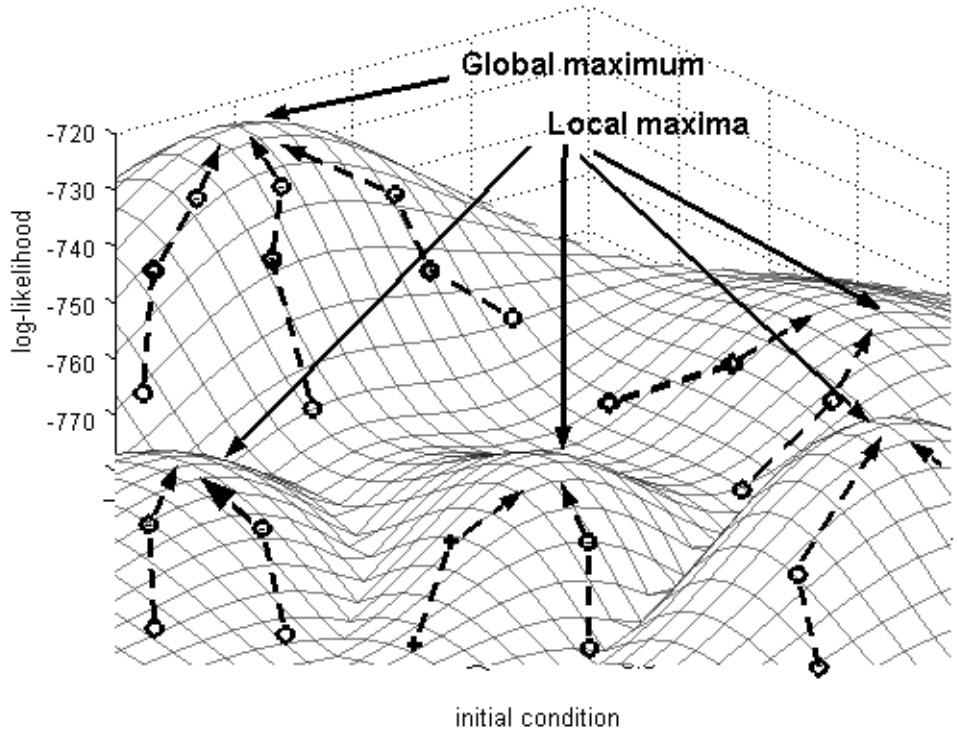
MLE of linear regression :

$$\hat{\beta}_N = (X^\top X)^{-1} X^\top y$$

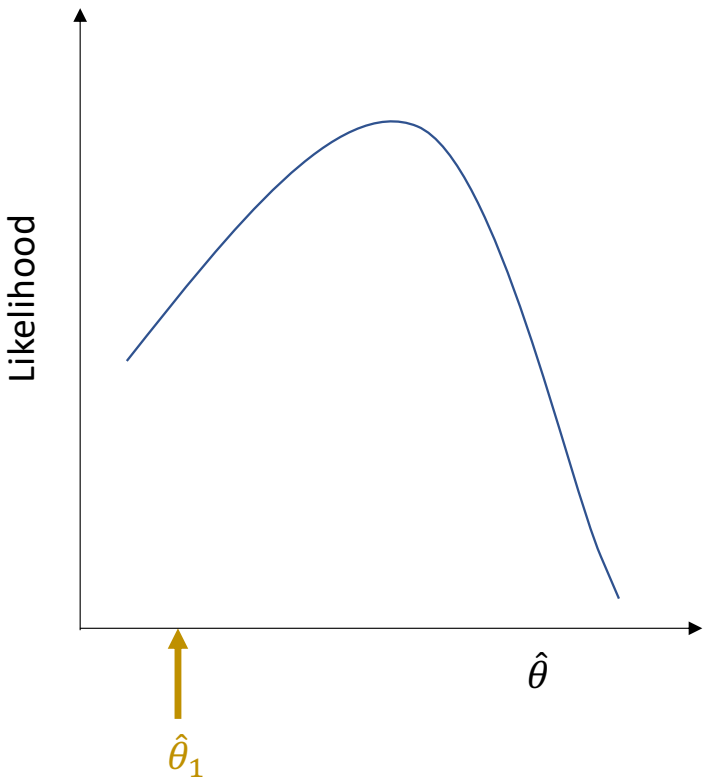
$$\hat{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - x_i \hat{\beta}_N \right)^2$$

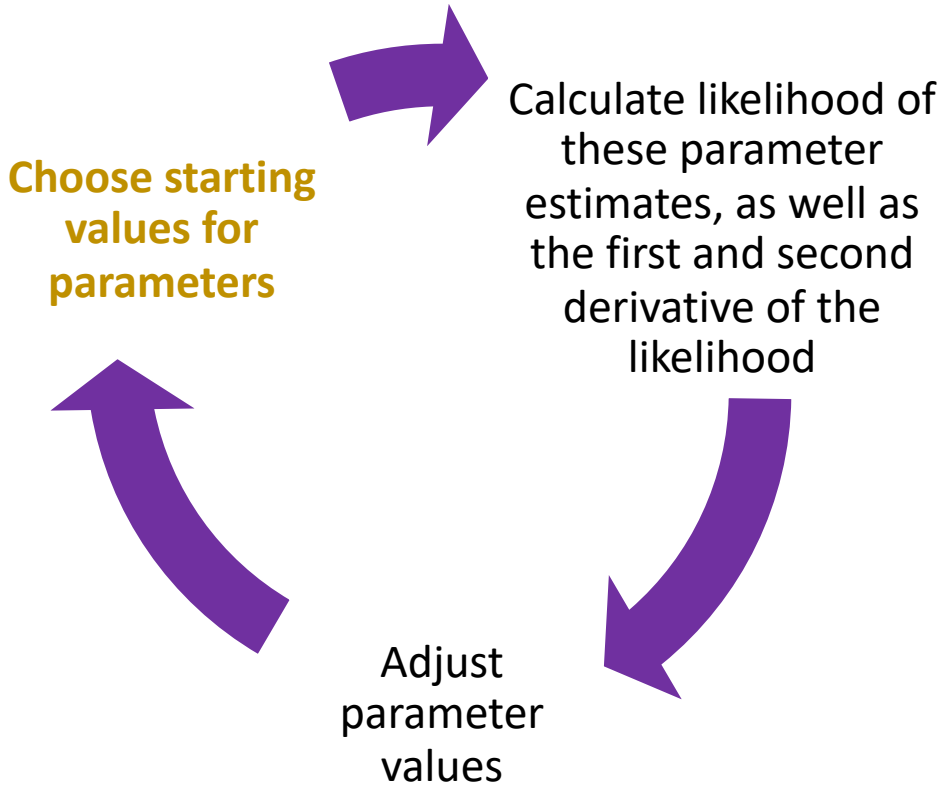
Optimization

For more complex models solutions can rarely be solved in closed form - rather iterative optimization procedures are commonly needed

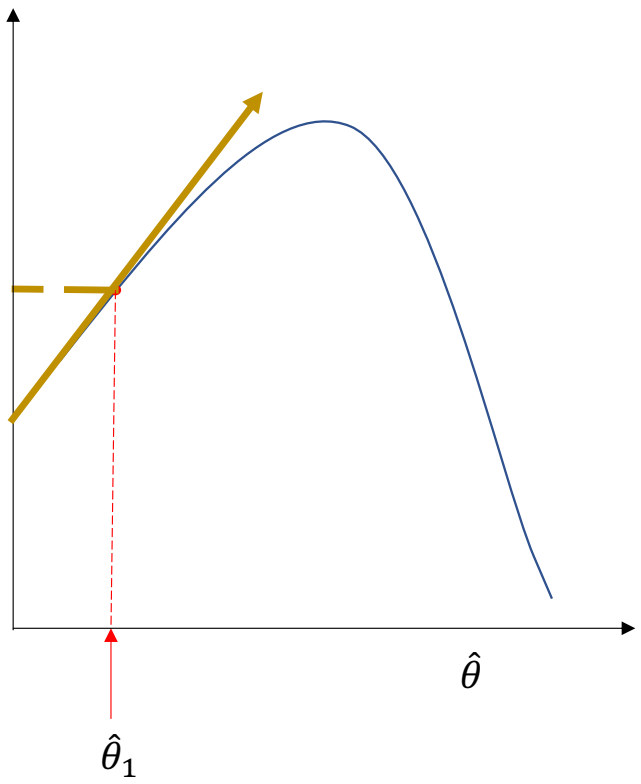


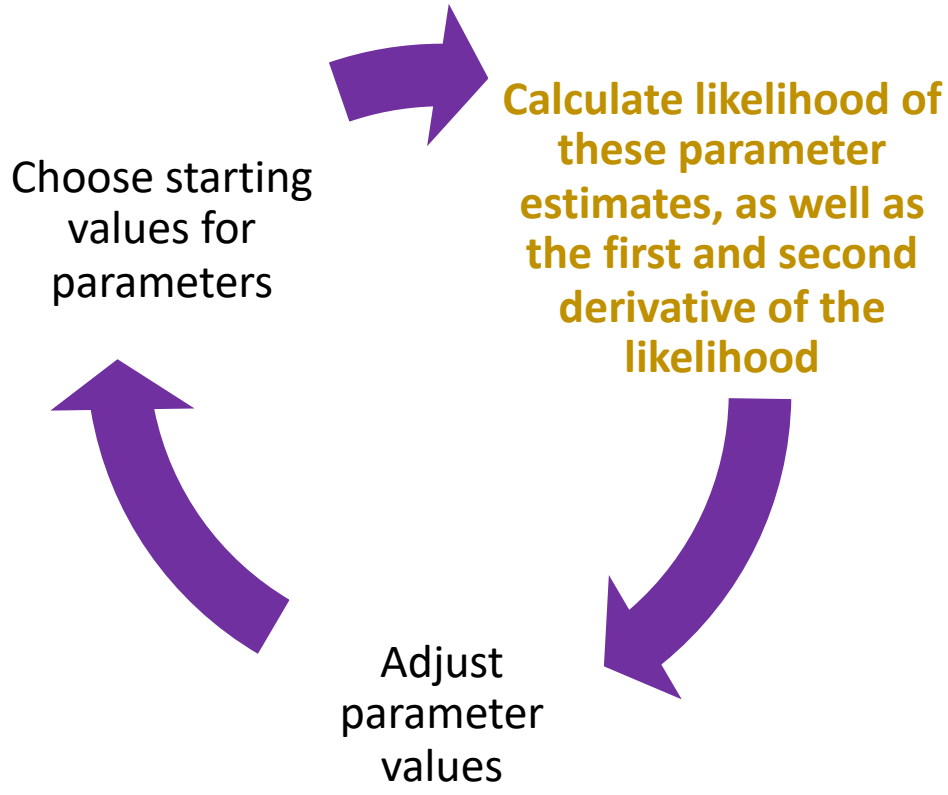
Optimization



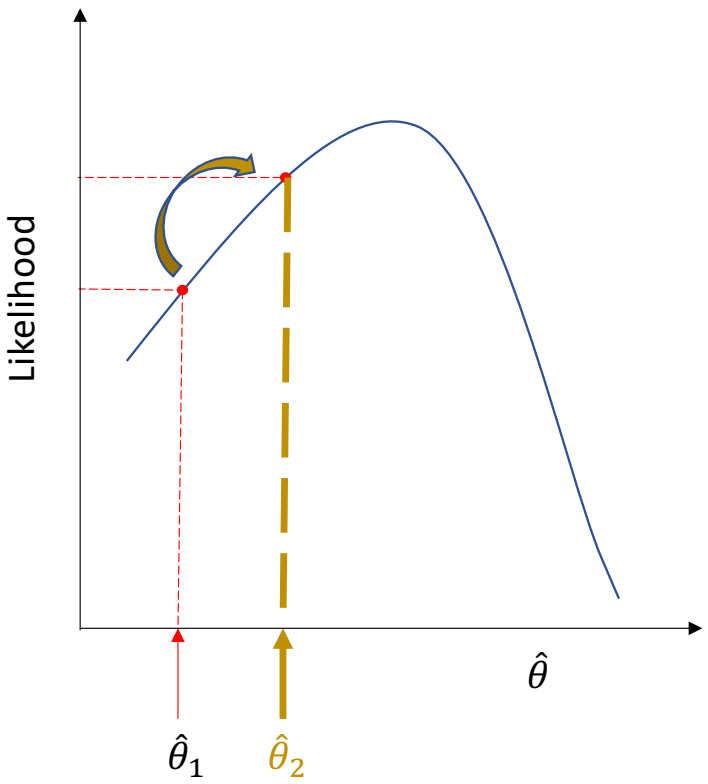


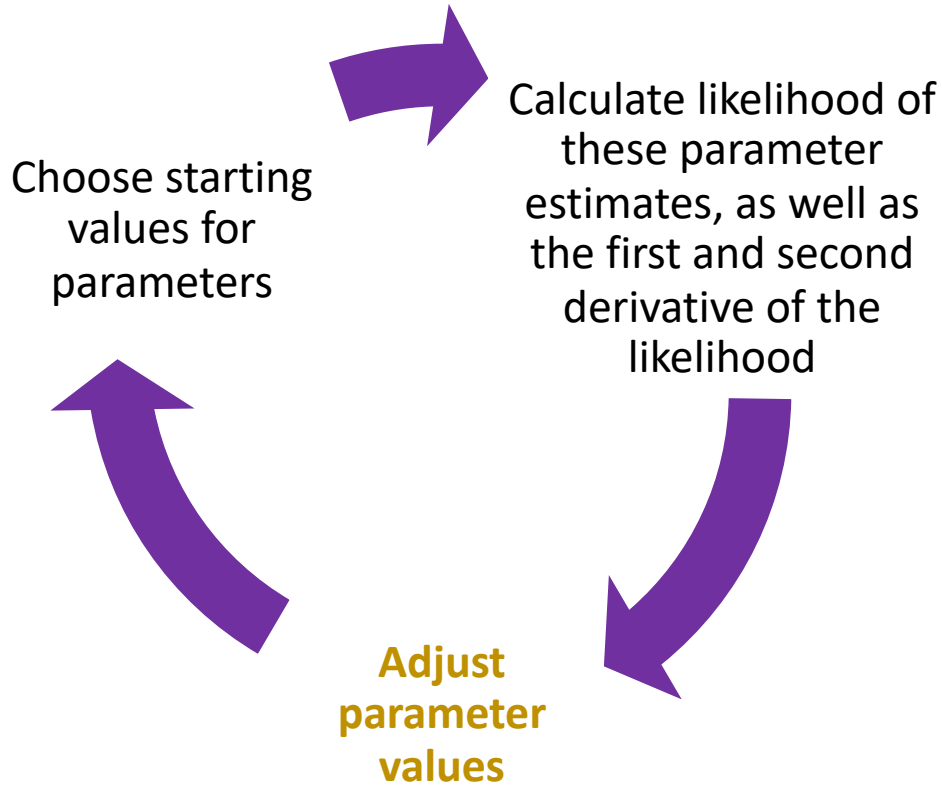
Optimization



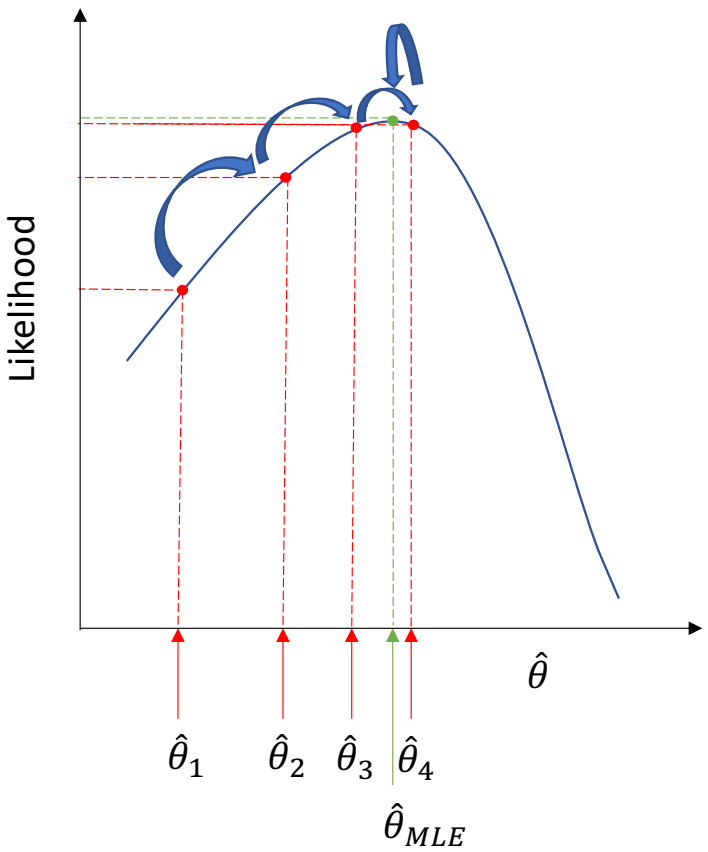


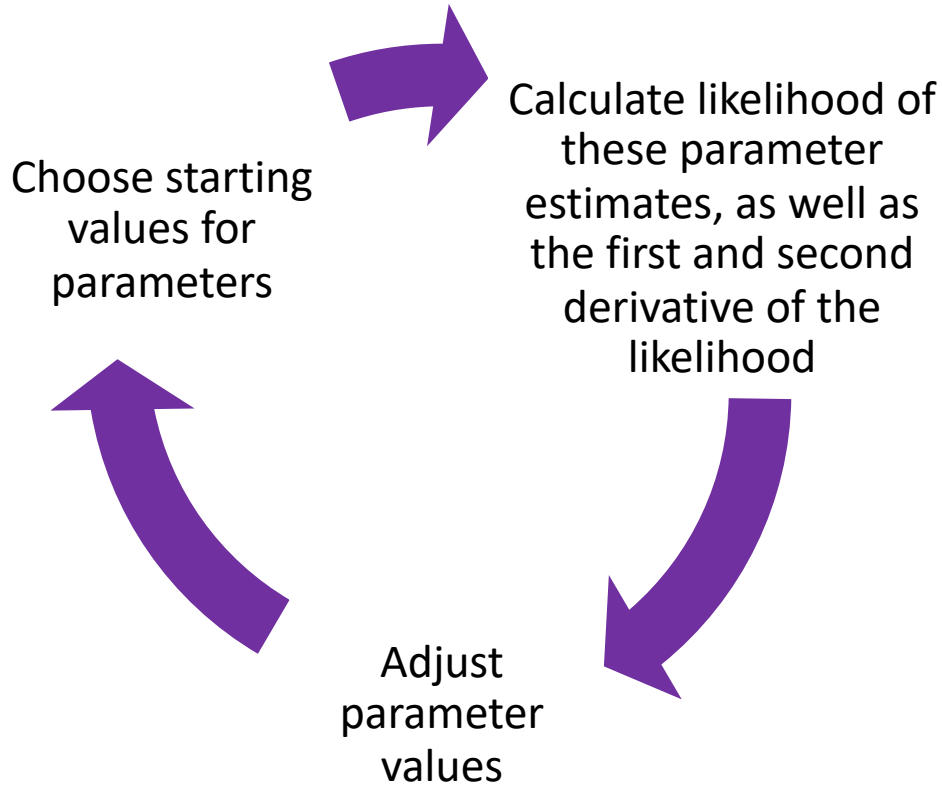
Optimization





Optimization





Repeat process until stopping criterion is reached

Likelihood ratio test

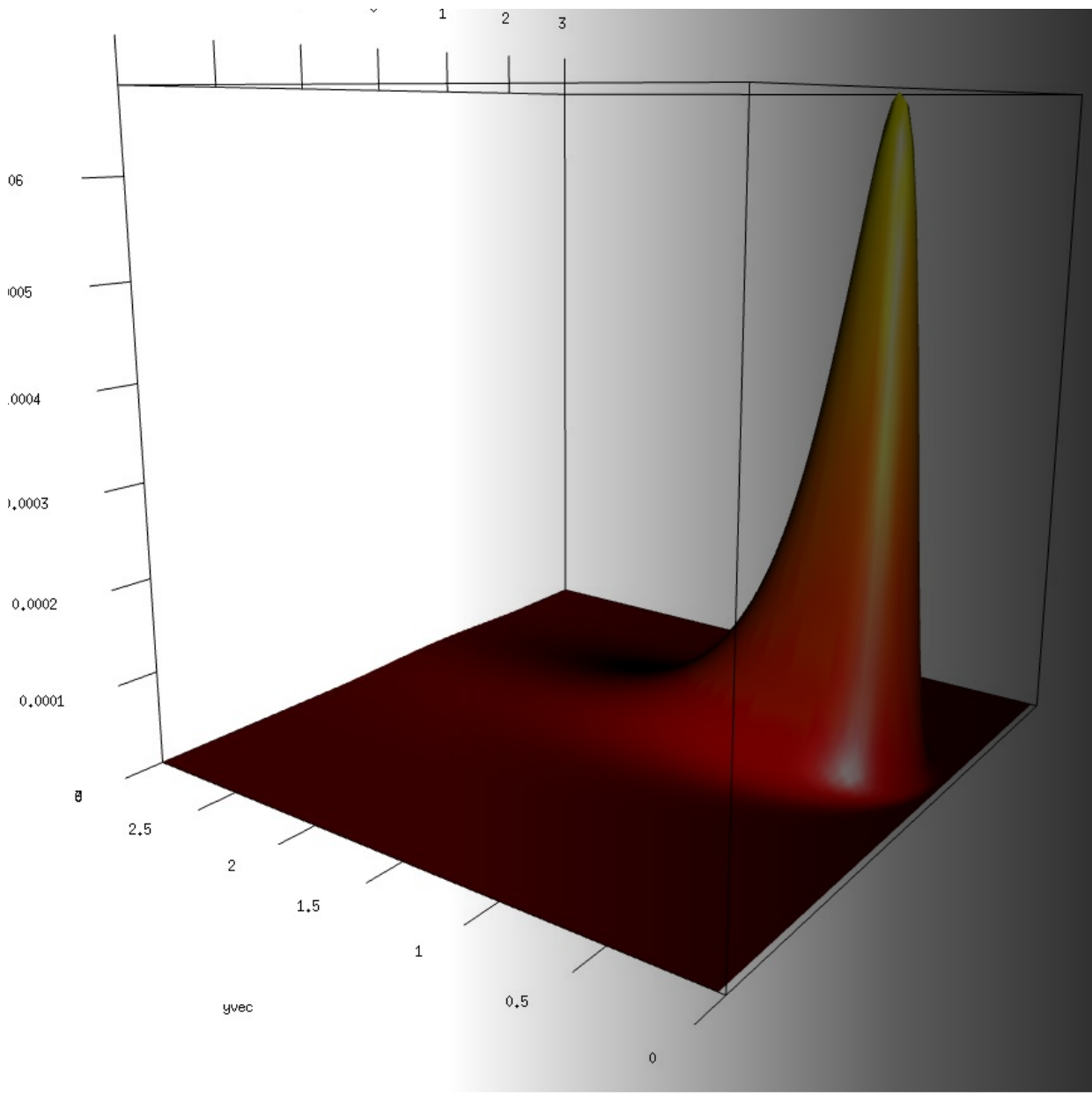
- Twice the difference in log-likelihood between nested models is distributed as chi-square

$$\lambda_{\mathrm{LR}} = -2\left[\,\ell(\theta_0) - \ell(\hat{\theta})\,\right]$$

e.g. Consider $\boldsymbol{\theta}_{\mathbf{F}}$ = (a, b, c); $\boldsymbol{\theta}_{\mathbf{R}}$ = (a, b, c=0)- twice the difference in log-likelihoods between the models would be distributed as χ^2_1

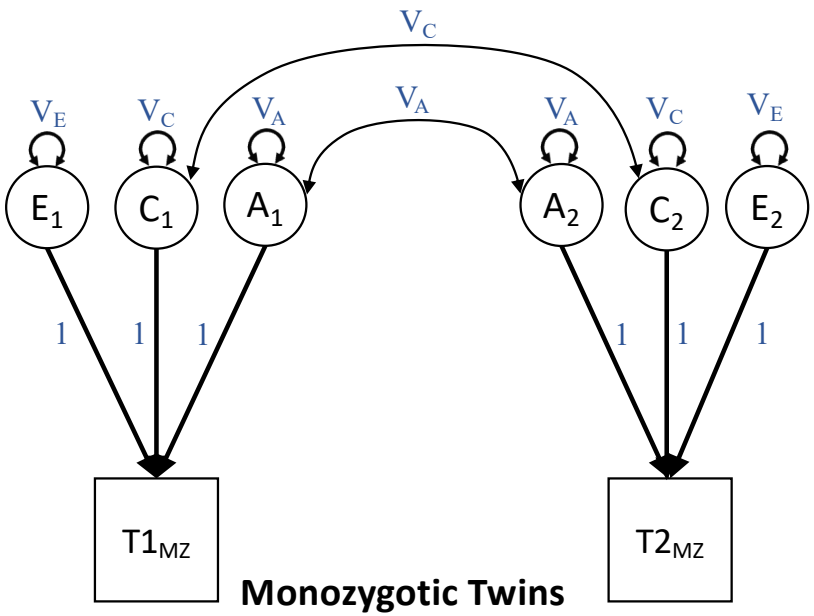
Model comparison

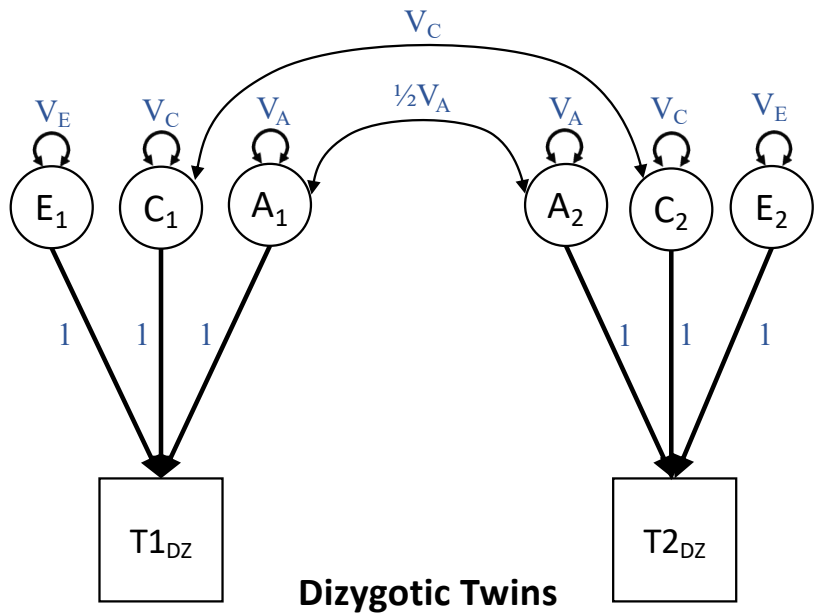
e.g. ACE vs. CE => significance test of heritability



Likelihood of Genetic model(s)

ACE model as path diagram





ACE model

Covariates (age, sex)

Genetic

Shared environment

Trait of interest

Residuals (incl. environment)

Y = **Xb** + **a** + **c** + **e**
a ~ N(0, **G** . Va)
c ~ N(0, **C** . Vc)
e ~ N(0, **I** . Vc)
Var(Y) = Va + Vc + Ve
Va/(Va+Vc+Ve) : heritability
Vc/(Va+Vc+Ve) : shared E

Main parameters of interest:

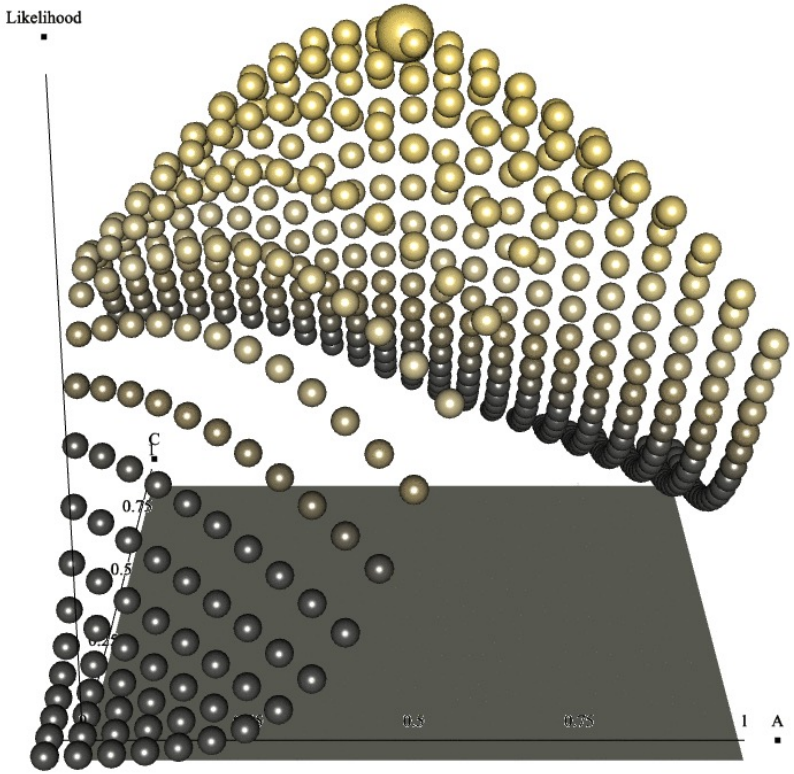
Va and Vc

Likelihood as a function of Va and Vc

"Real data" with 500 MZ + 500 DZ pairs
covariates

Fitted model in OpenMx
Estimated likelihood for a range of set Va and Vc values

```
> fitACE$output$estimate
     interC       betaS       betaA        VA11        VC11        VE11
-3.26259017  0.09401213  0.16099604  0.48645944  0.23707928  0.25814284
```

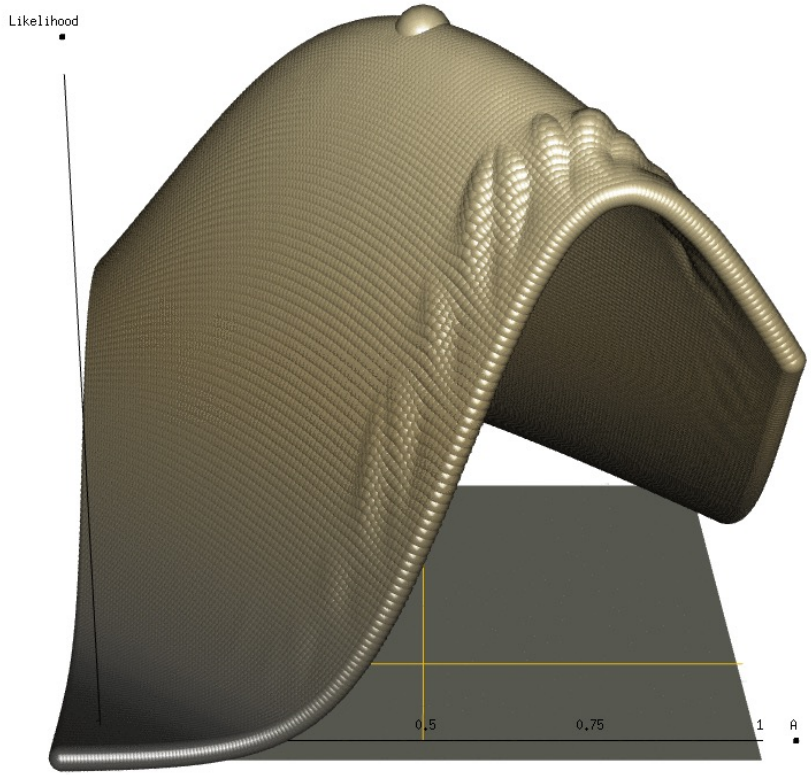


Likelihood (interpolated)

ML estimates :
Va =0.48
Vc = 0.24

Likelihood can be estimated for Va, Vc < 0. But note what happens near boudary of parameter space

```
> fitACE$output$estimate
     interC       betaS       betaA        VA11        VC11        VE11
-3.26259017  0.09401213  0.16099604  0.48645944  0.23707928  0.25814284
```

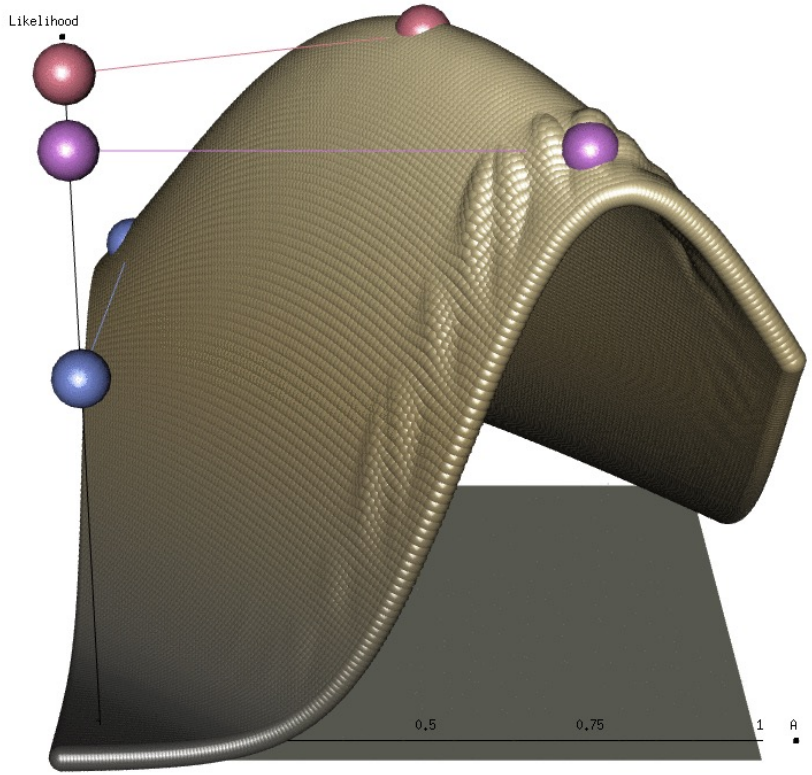


Likelihood ratio test

ACE model
AE model
CE model

Test statistic : twice the difference of log-likelihoods

```
> mxCompare(fitACE, c(fitAE, fitCE))
   base comparison ep  minus2LL   df       AIC     diffLL diffdf             p
1 ACEvc       <NA>  6 10220.367 3994 10232.367         NA     NA            NA
2 ACEvc         AE  5 10242.711 3995 10252.711  22.343713      1 2.2795805e-06
3 ACEvc         CE  5 10334.400 3995 10344.400 114.032618      1 1.2818252e-26
```

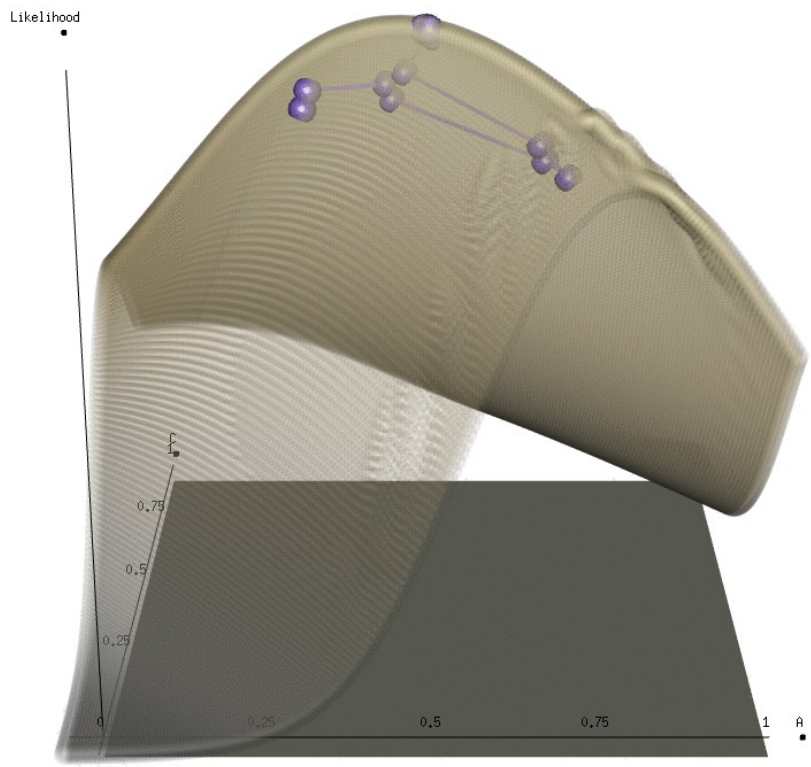


Optimization

SLSQP optimizer

Started at
Vc=Va=0.3

Found ML in 18
interations



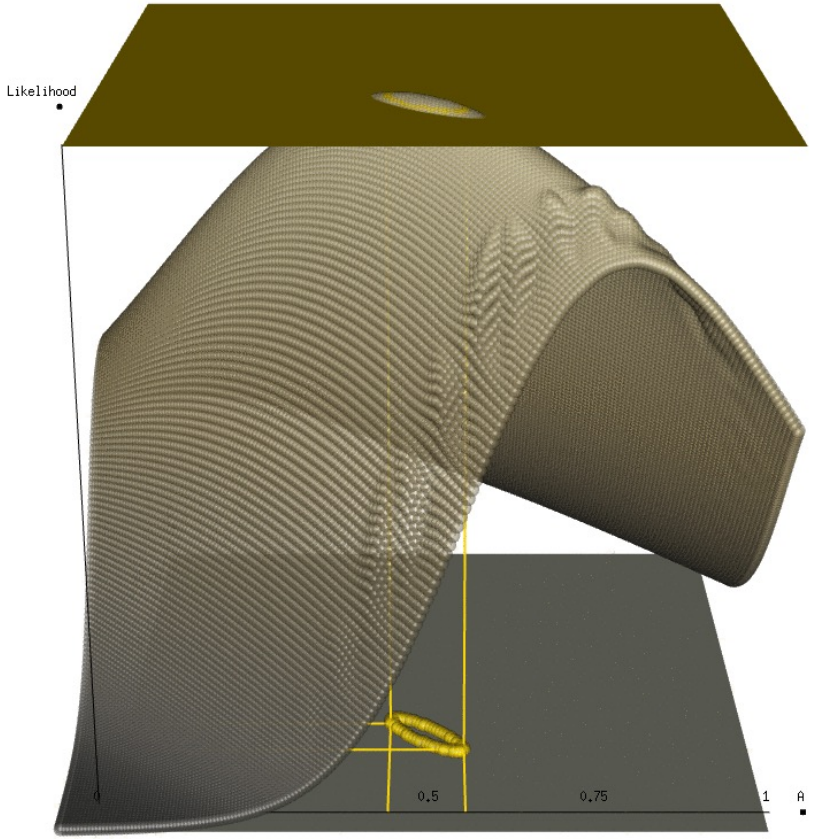
Confidence intervals

Start from maximum likelihood

Degrade (lower) the likelihood so that difference is significant (chi2 test) at 1-CI

For 95% CI : $chi^2 = 3.84 \Leftrightarrow$ pvalue=0.05

```
> fitACE$output$confidenceIntervals
                    lbound  estimate    ubound
ACEvc.VarC[1,4] 0.4005072 0.4955369 0.5960494
ACEvc.VarC[1,5] 0.1466367 0.2415032 0.3290960
```



ML, FIML, REML

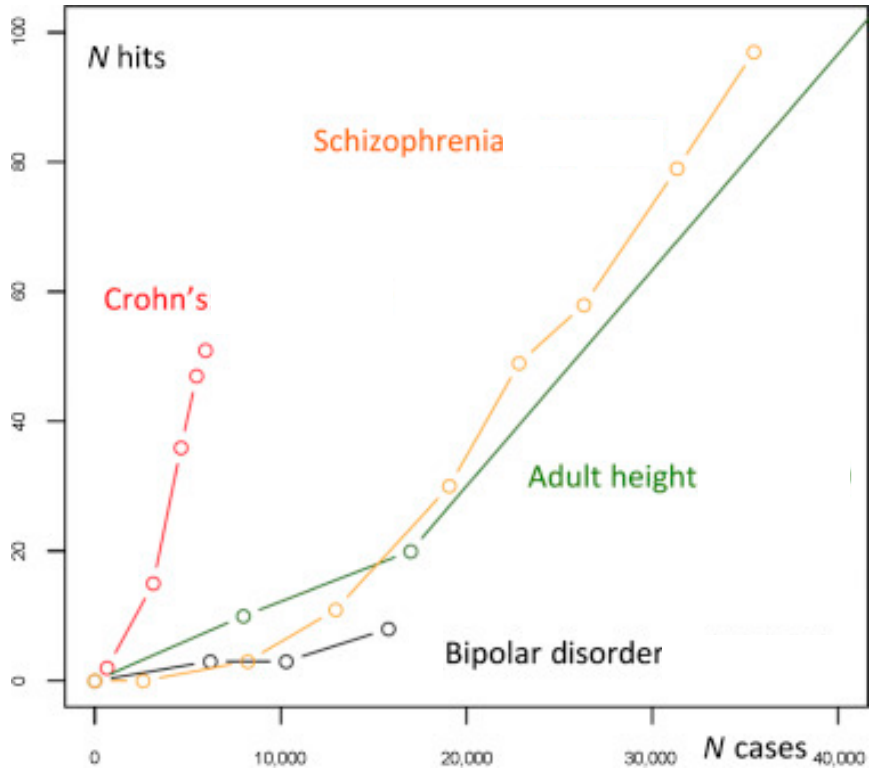
ML: Maximum likelihood	Fine for fixed effect models
FIML: Full Information Maximum Likelihood	Handles missing values
REML: Restricted Maximum Likelihood	Minimises bias in variance estimation of mixed models

Also pseudo likelihood, or quasi-likelihood..

Genomic SEM

Genomic SEM – Why Genomic SEM?

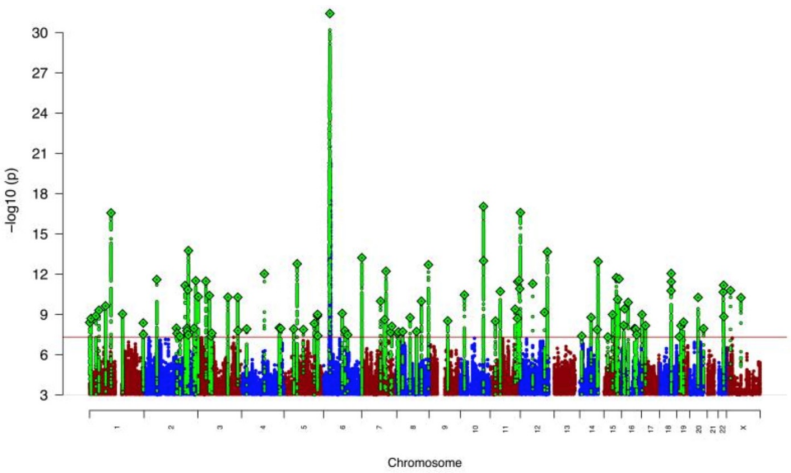
- Human complex traits/diseases are associated with many genes



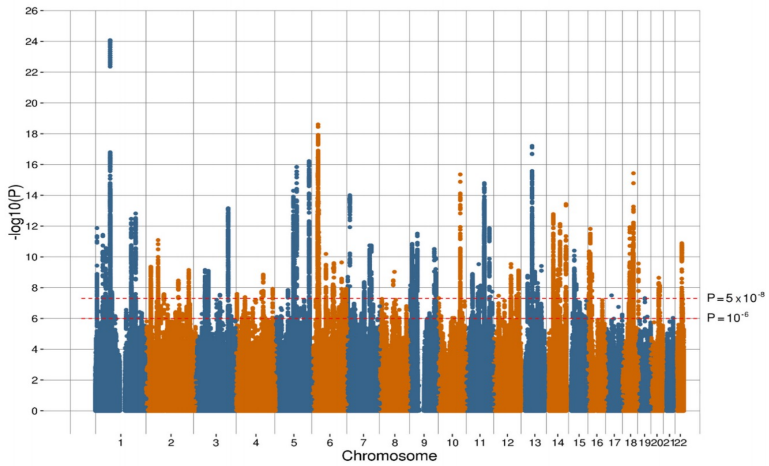
S. Cichon, S. Ripke, 2016

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*†

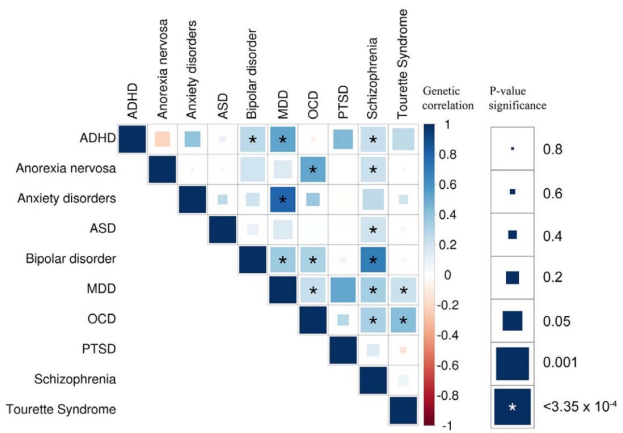


Fig. 1. Genetic correlations across psychiatric phenotypes. The color of each box indicates the magnitude of the correlation, and the size of the box indicates its significance (LDSC), with significant correlations filling each square completely. Asterisks indicate genetic correlations that are significantly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

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

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

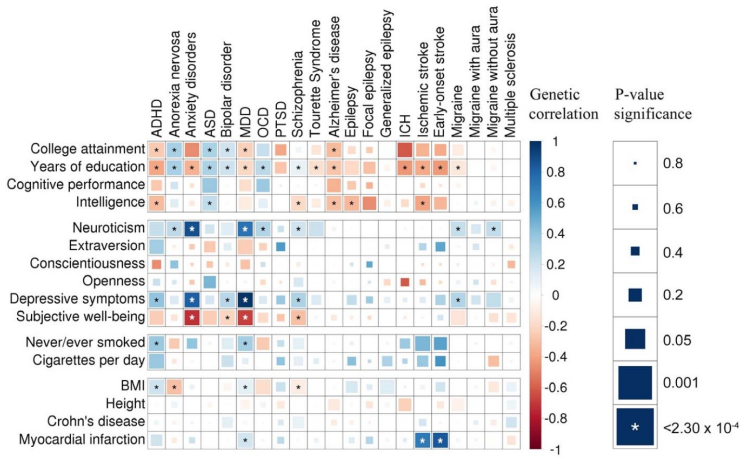


Fig. 4. Genetic correlations across brain disorders and behavioral-cognitive phenotypes. The color of each box indicates the magnitude of the correlation, and the size of the box indicates its significance (LDSC), with significant correlations filling each square completely. Asterisks indicate genetic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

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

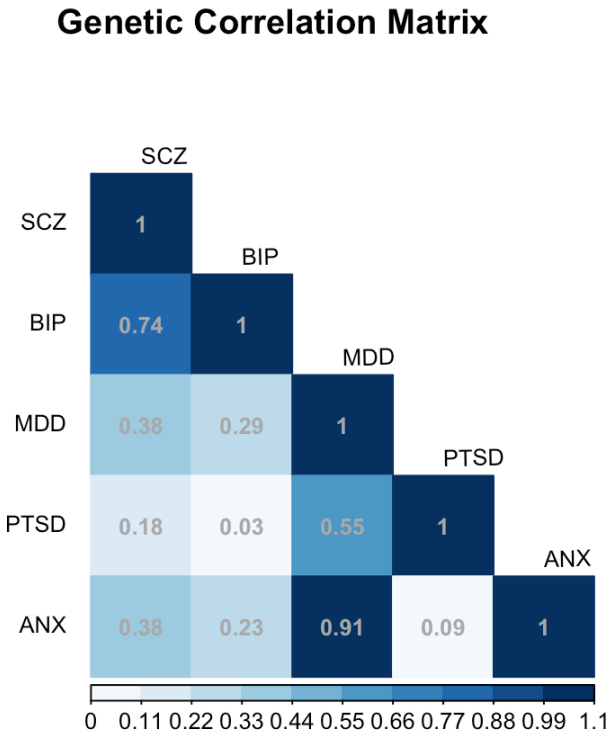
Also produced is a second matrix, V, of squared standard errors and the dependencies between estimation errors

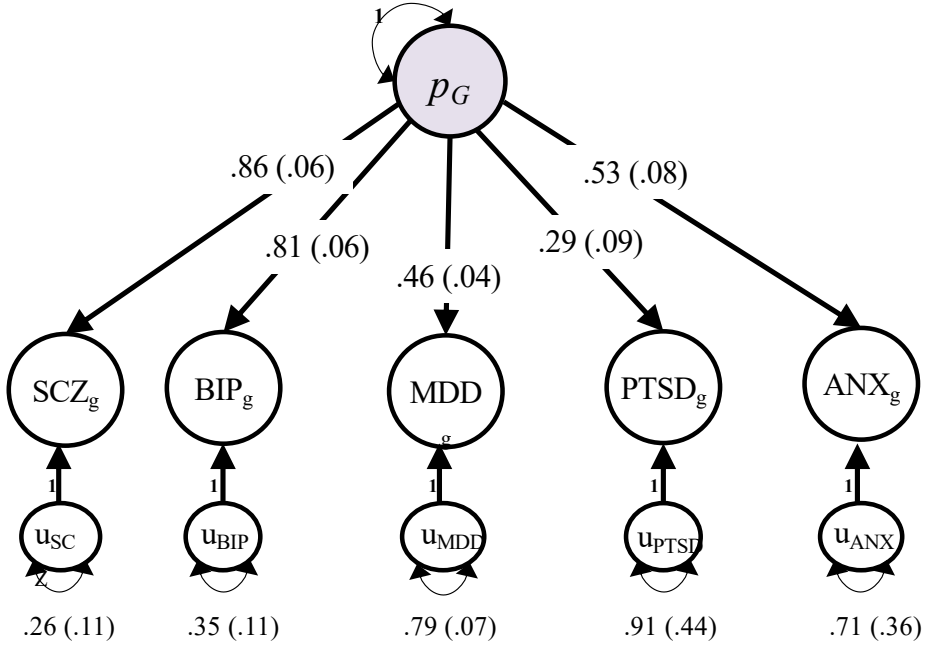
Diagonal elements are squared standard errors of genetic variances and covariances

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

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

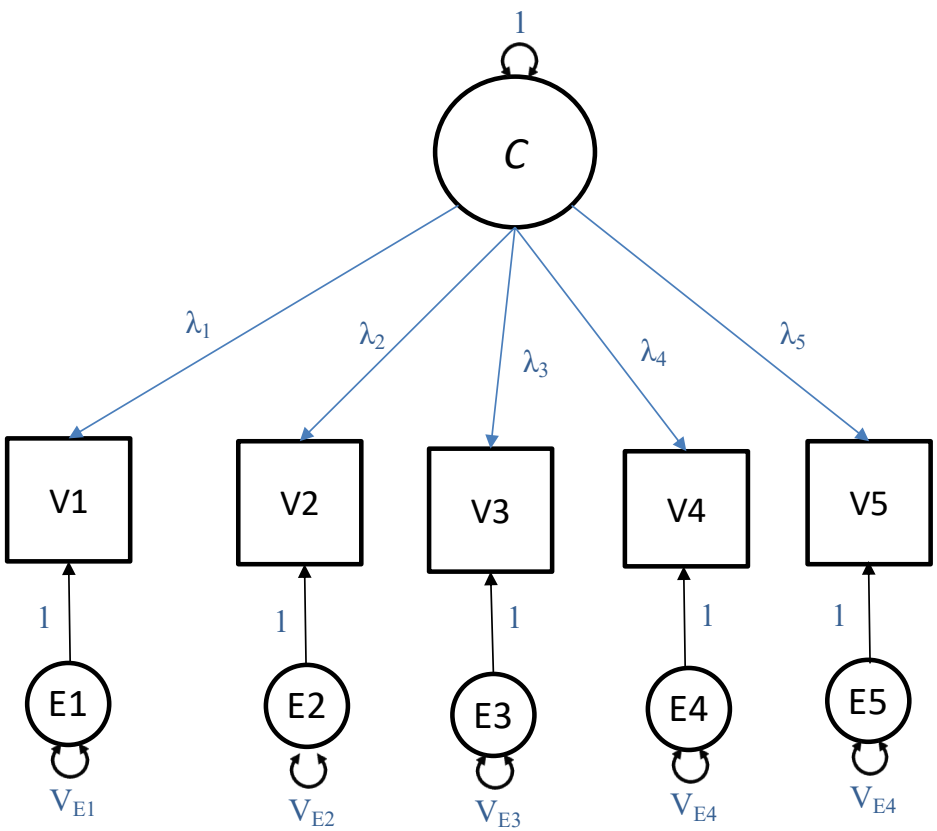
Genomic SEM – Common factor model





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

SEM – Common factor model



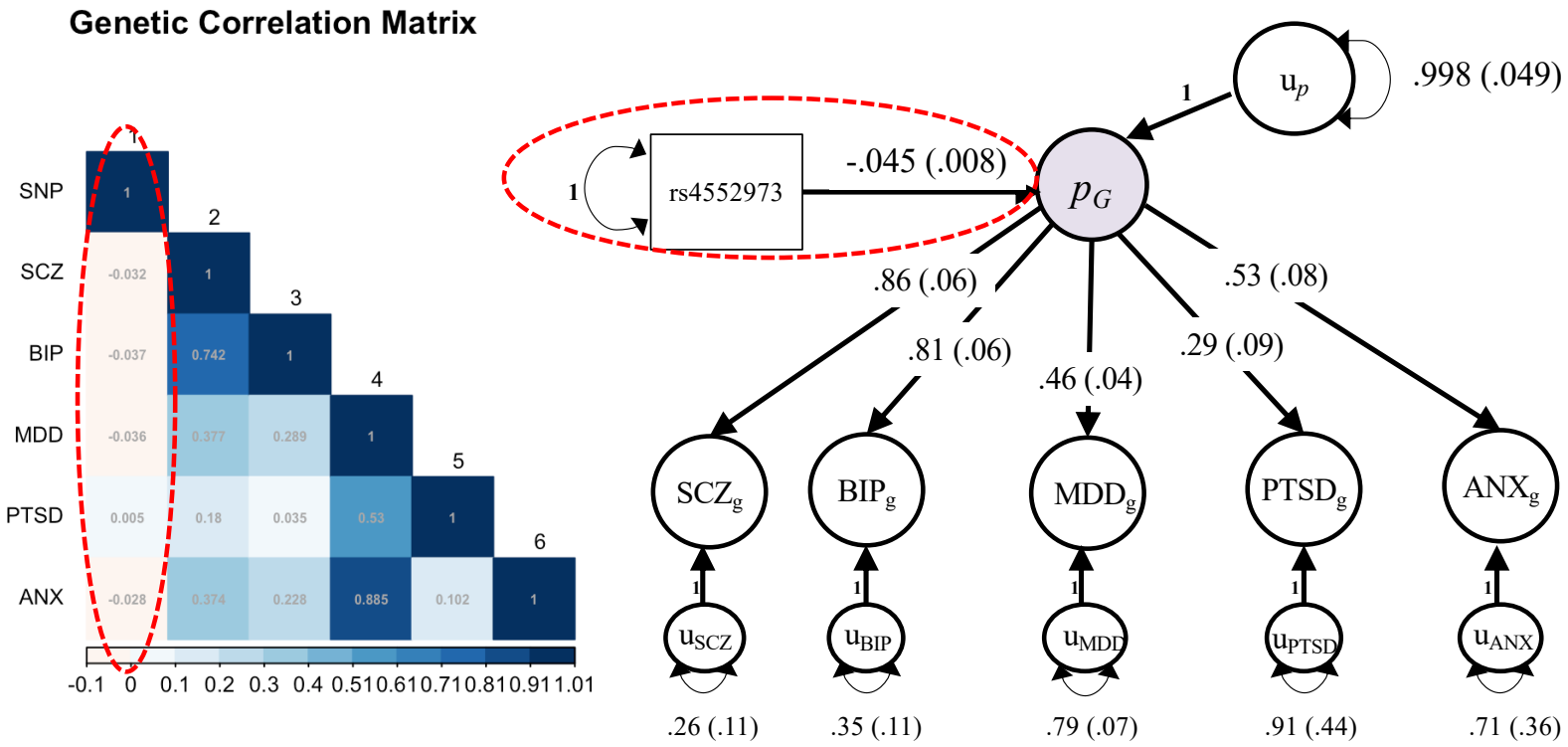
Observed Covariance Matrix:

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

Expected Covariance Matrix:

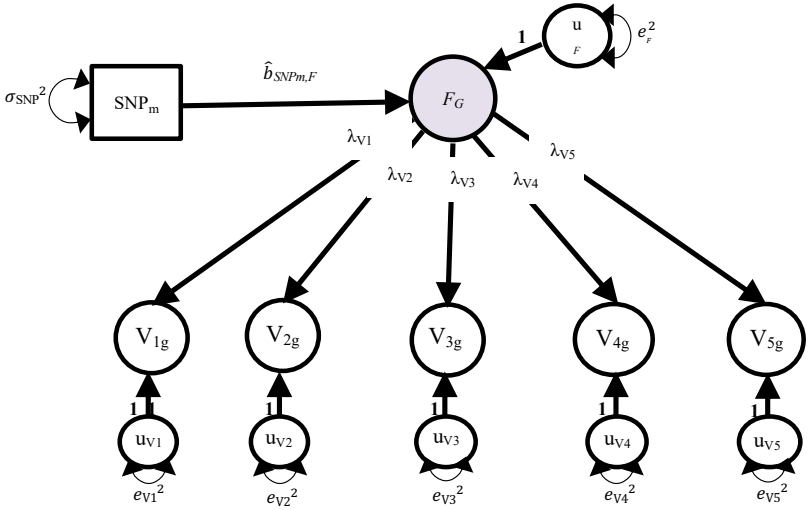
$$\Sigma(\theta) = \begin{matrix} \lambda_1^2+V_{E1} & \lambda_1\lambda_2 & \lambda_1\lambda_3 & \lambda_1\lambda_4 & \lambda_1\lambda_5 \\ \lambda_2\lambda_1 & \lambda_2^2+V_{E2} & \lambda_2\lambda_3 & \lambda_2\lambda_4 & \lambda_2\lambda_5 \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3^2+V_{E3} & \lambda_3\lambda_4 & \lambda_3\lambda_5 \\ \lambda_4\lambda_1 & \lambda_4\lambda_2 & \lambda_4\lambda_3 & \lambda_4^2+V_{E4} & \lambda_4\lambda_5 \\ \lambda_5\lambda_1 & \lambda_5\lambda_2 & \lambda_5\lambda_3 & \lambda_5\lambda_4 & \lambda_5^2+V_{E5} \end{matrix}$$

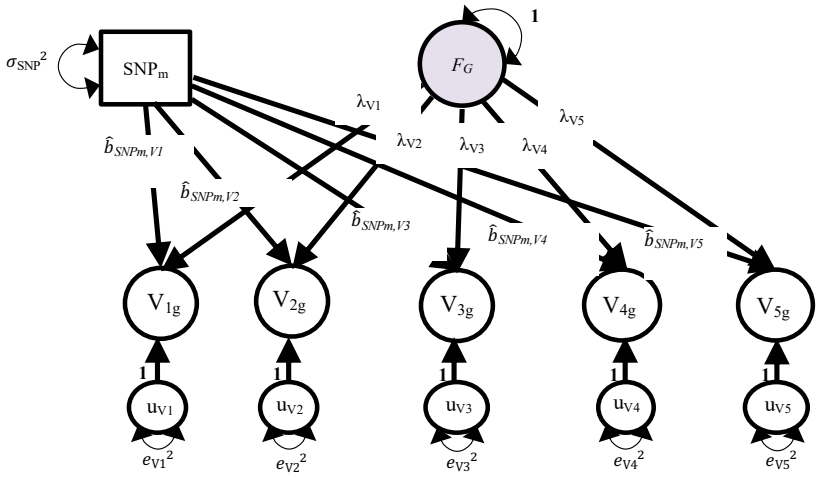
Genomic SEM – GWAS of a Latent Factor



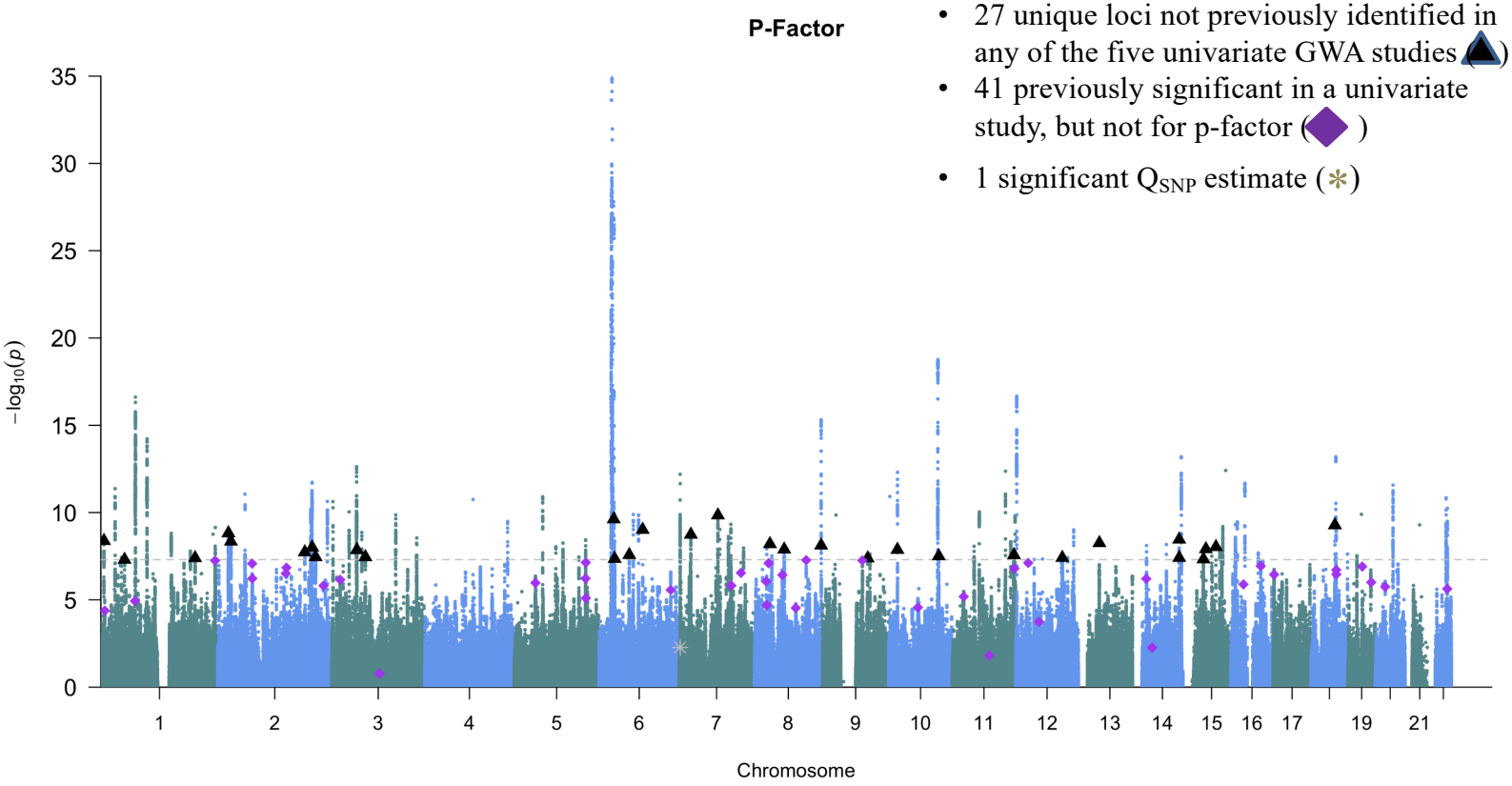
Genomic SEM - Estimates of SNP level heterogeneity (Q_{SNP})

- 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.

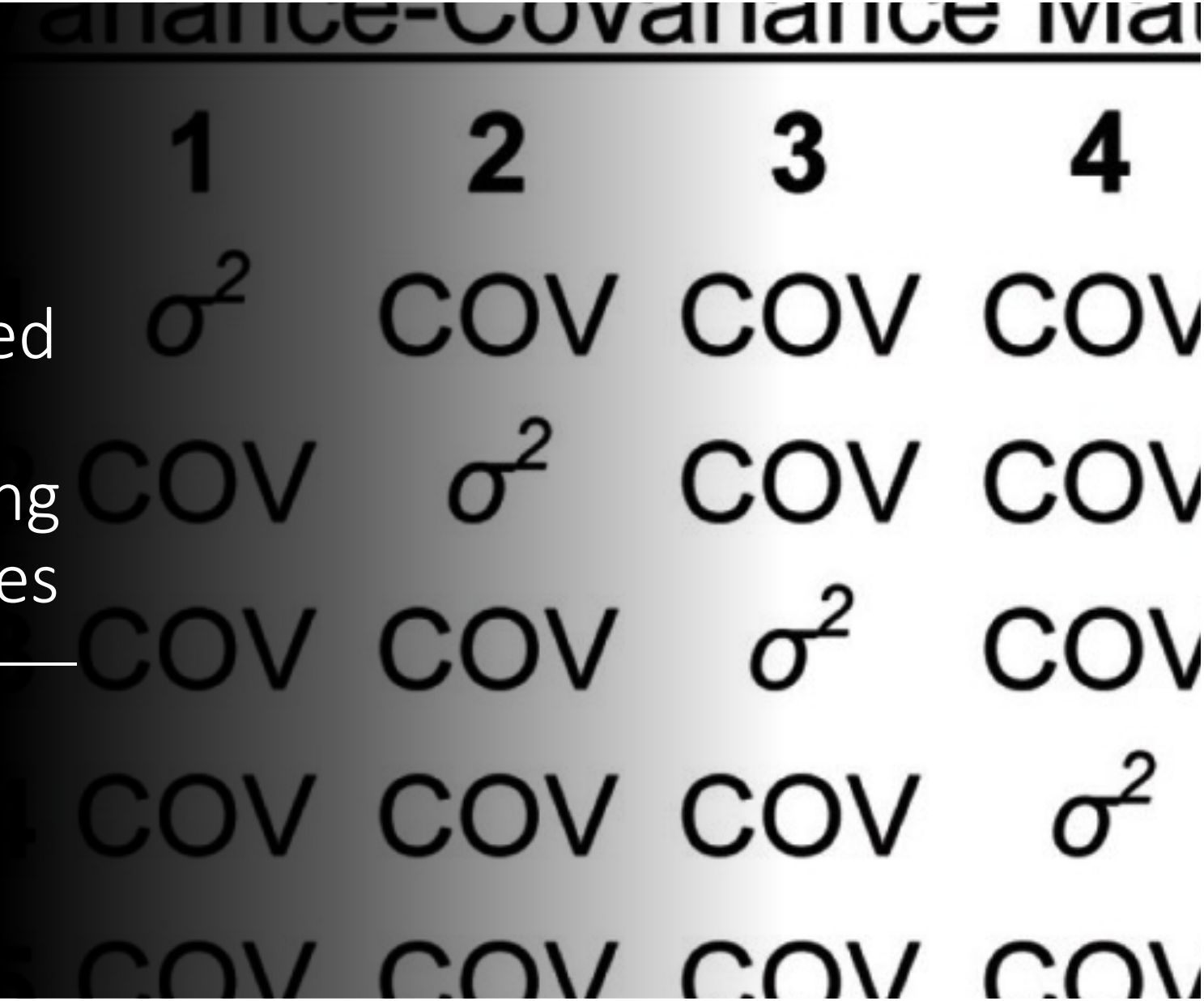
Thank you for your attention

The University of Queensland Australia | IMB
Institute for Molecular Bioscience

Members of the Centre for Population and Disease Genomics

David Evans
Nicole Warrington
Geng Wang
Mike Hunter (openMx)

Deriving Expected Variances and Covariances Using 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)

The expected value of a chain is the product of all coefficients associated with each path making up that chain

The final expected variance or covariance equals the sum of the values of all legitimate chains

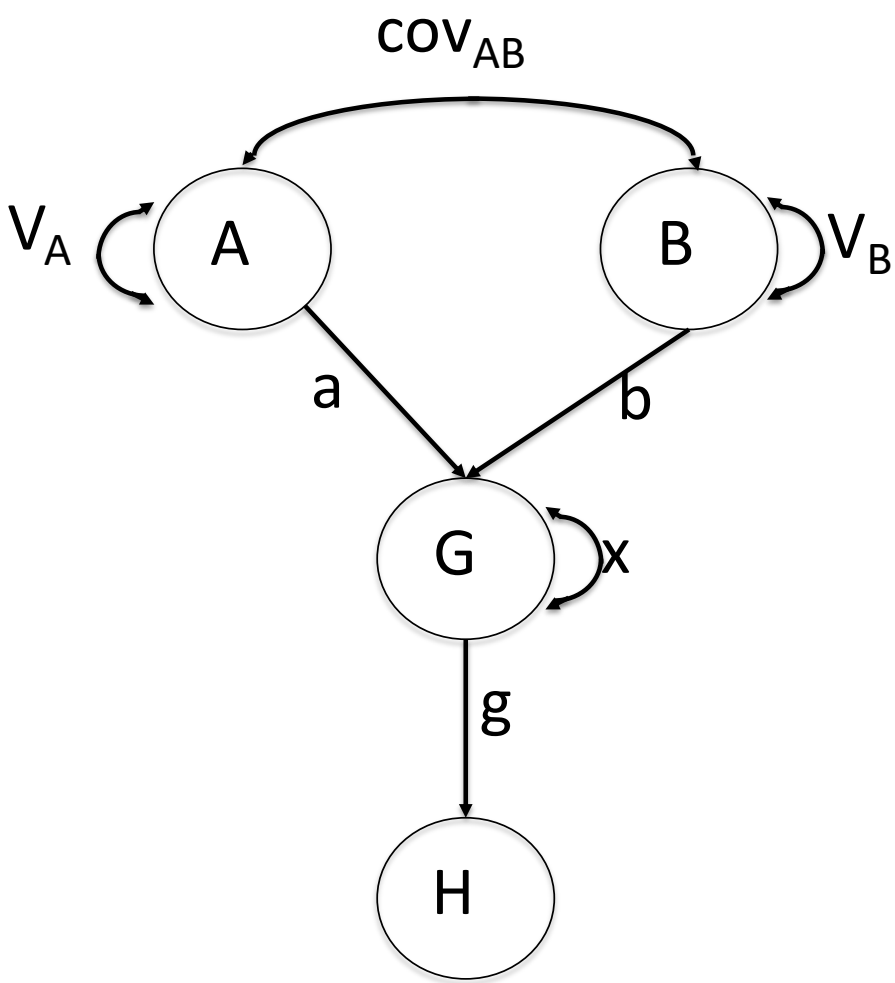
Path Tracing Rules. Legitimate chains:

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

Once a double headed arrow has been traversed, the direction reverses such that the chain travels forward

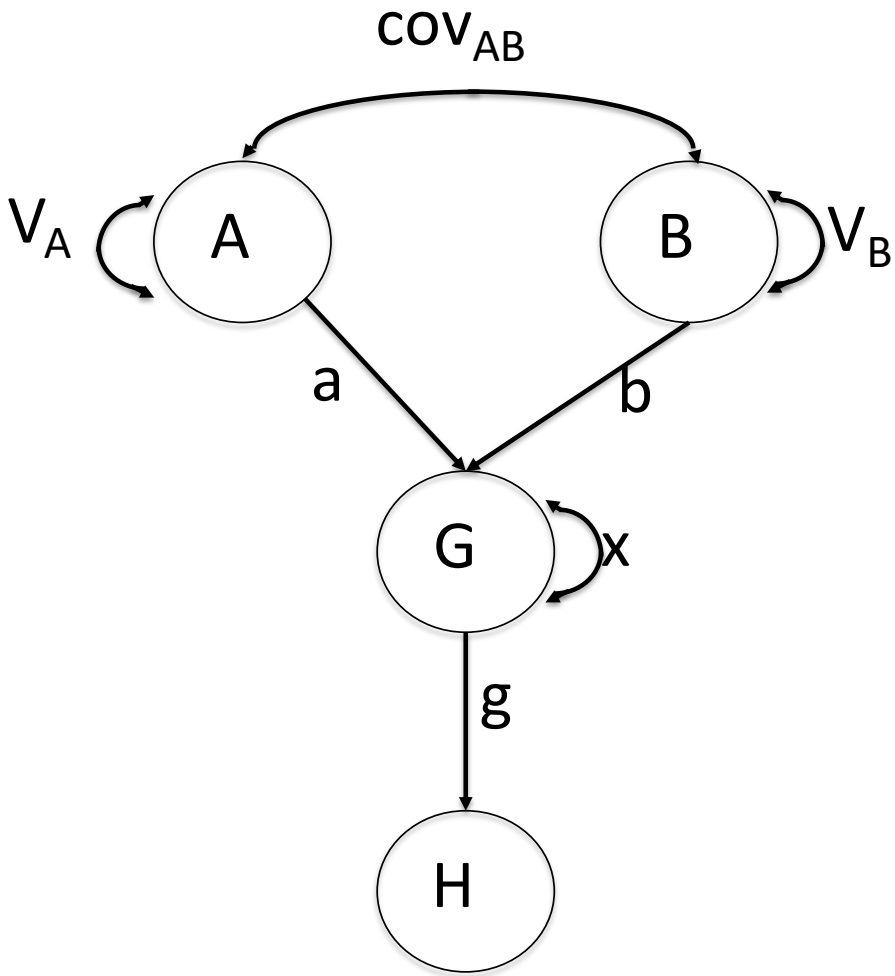
All chains must include exactly one double-headed arrow. This implies a chain must change directions exactly once.

All chains must be counted exactly once and each must be unique. However, order matters: *abc* is a distinct chain from *cba*.



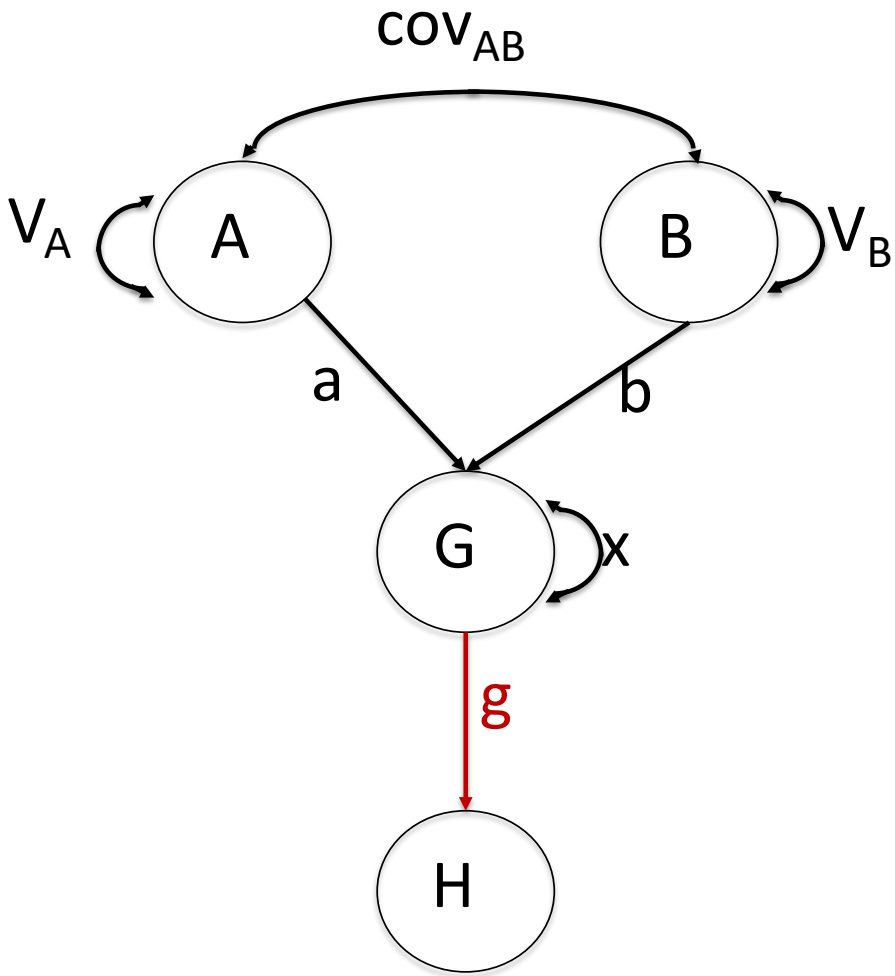
Expected covariance

- COV(H,A) =



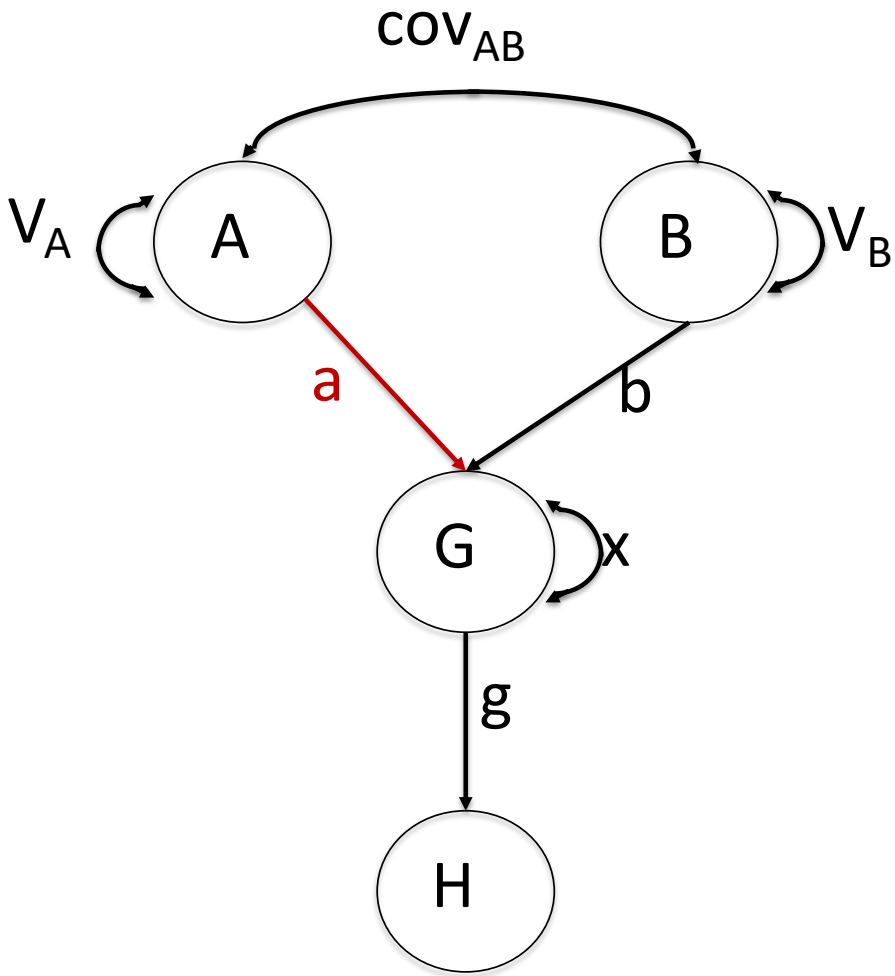
Expected covariance

- COV(H,A) =



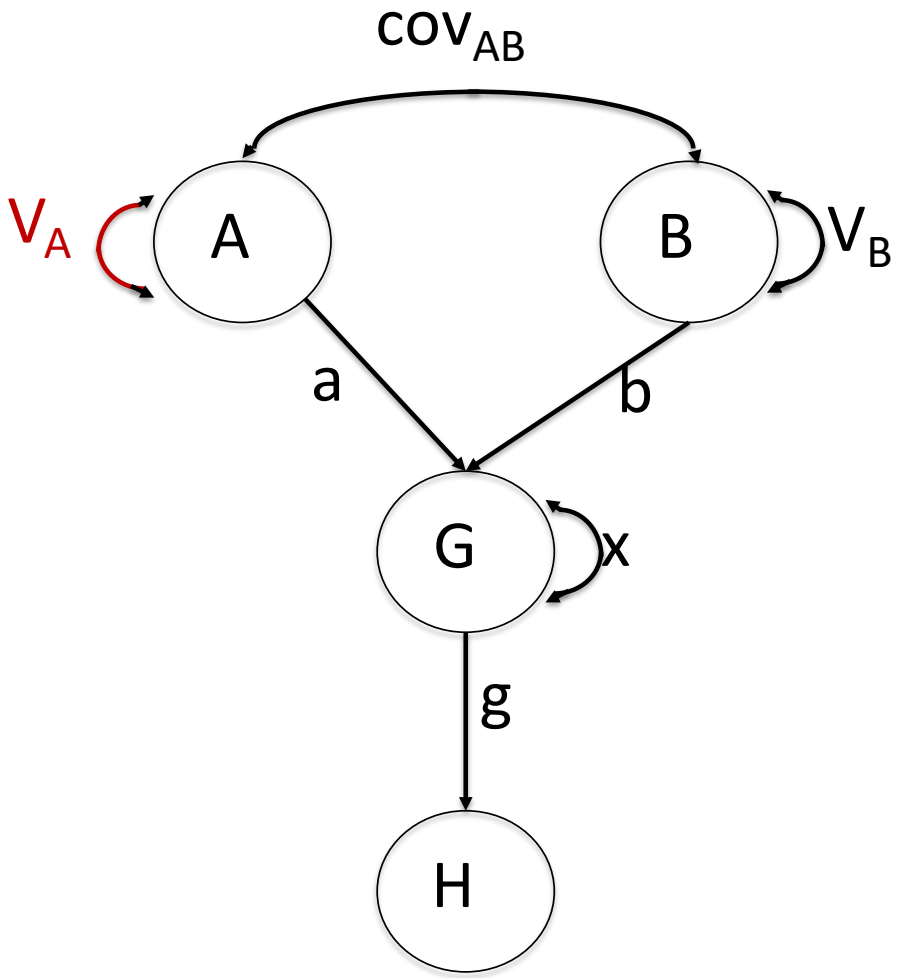
Expected covariance

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



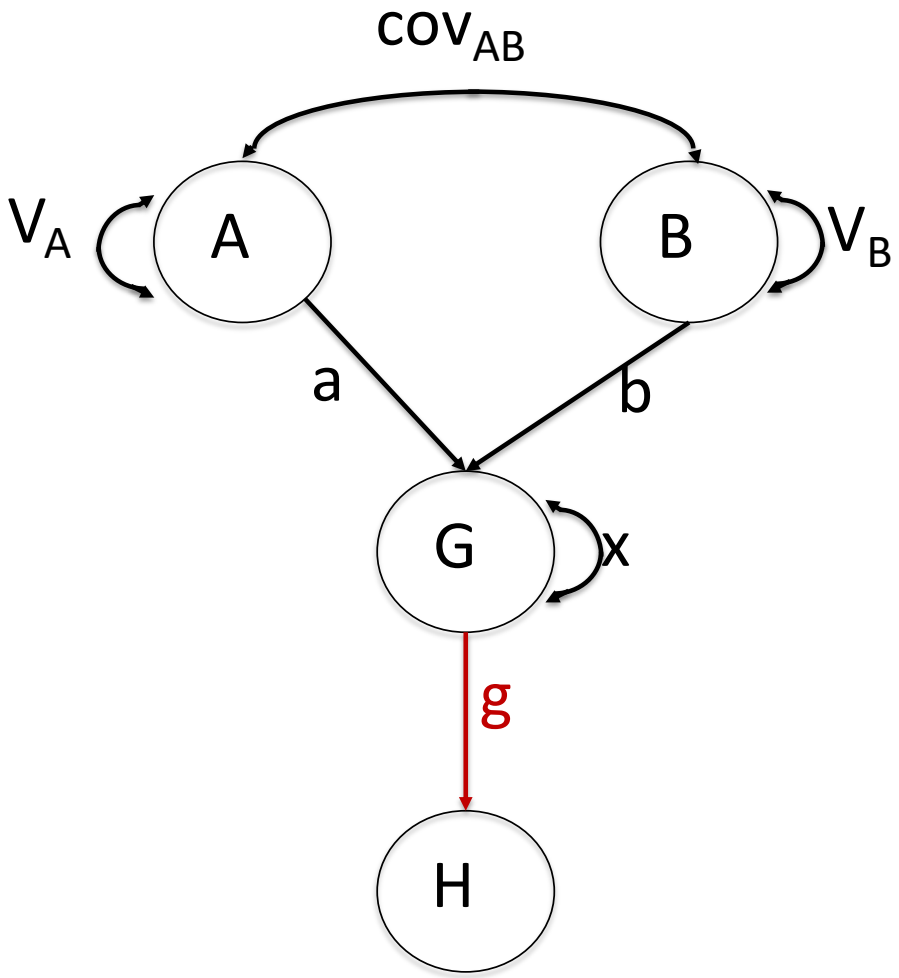
Expected covariance

COV(H,A) = g *a



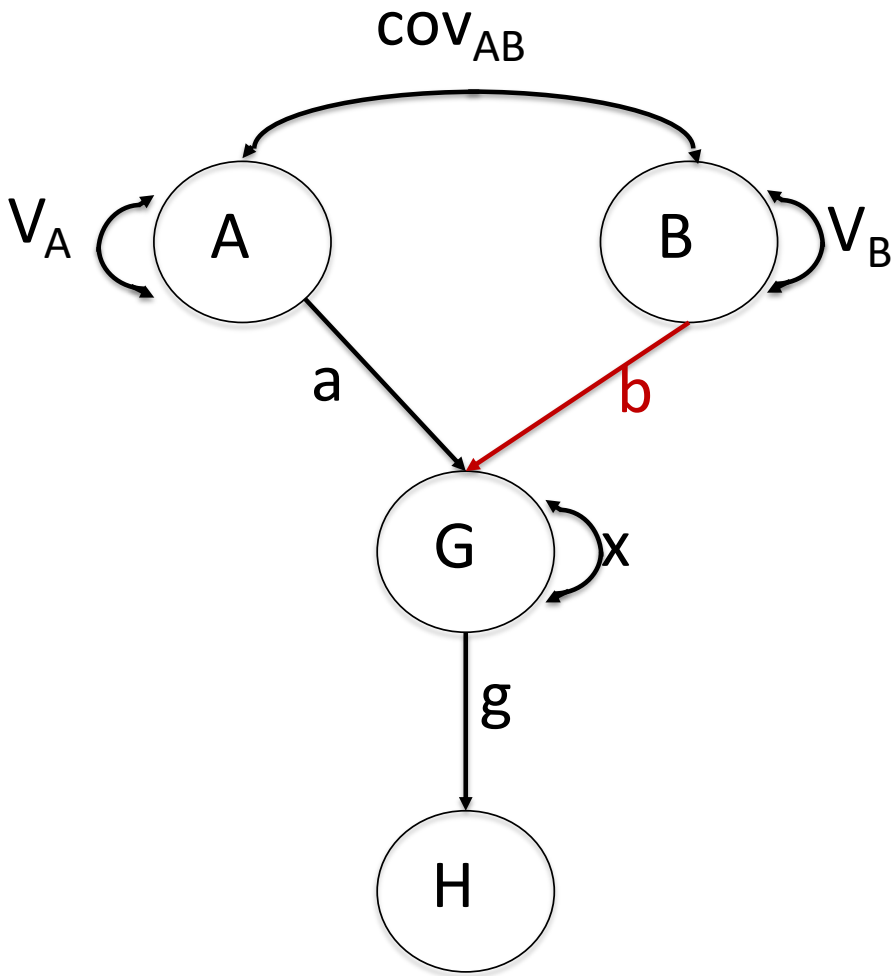
Expected covariance

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



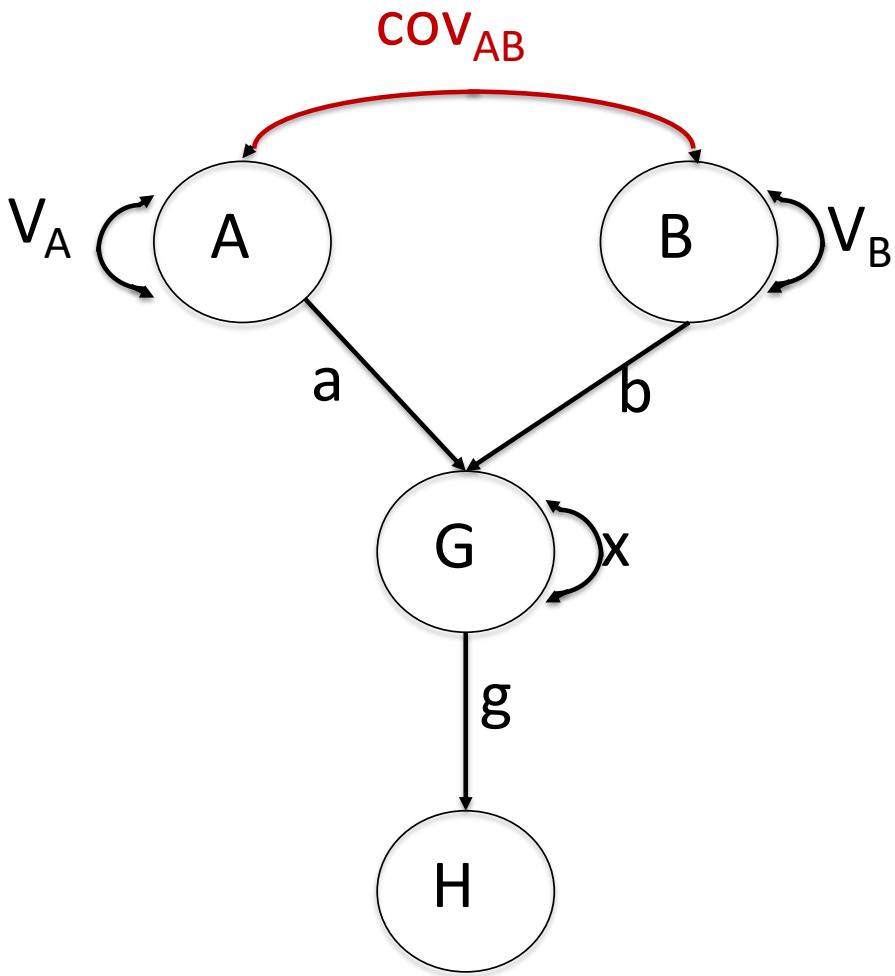
Expected covariance

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



Expected covariance

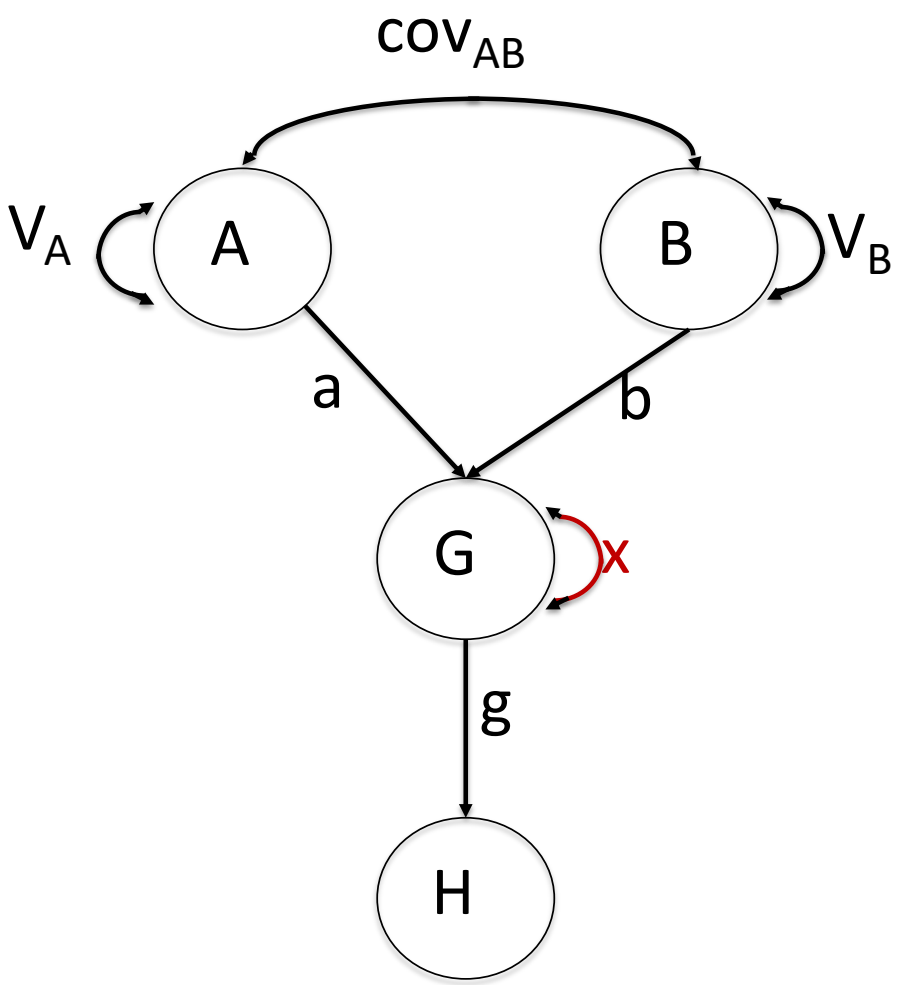
$COV(H,A) = g * a * V_A + g * b$



Expected covariance

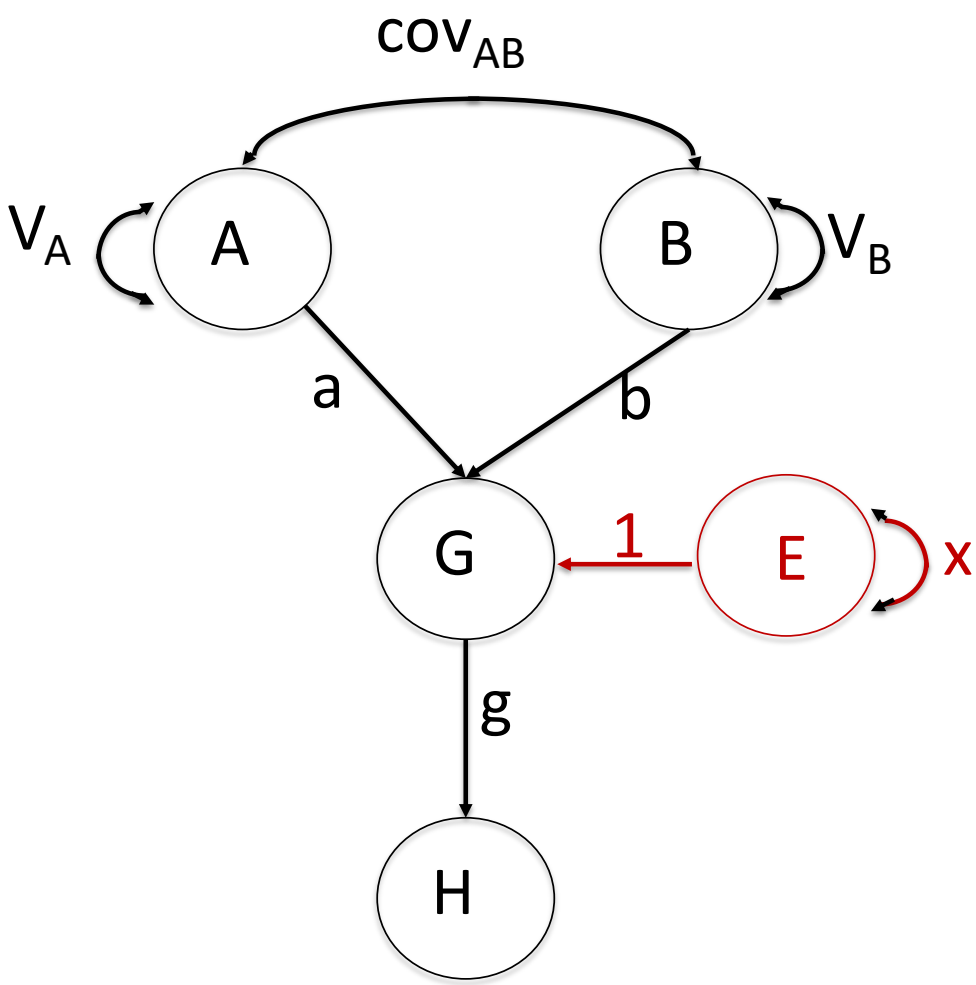
$COV(H,A) = g * a * V_A + g * b * COV_{AB}$

A visual/graphical way of deriving covariances between variables of a model!



Expected variance

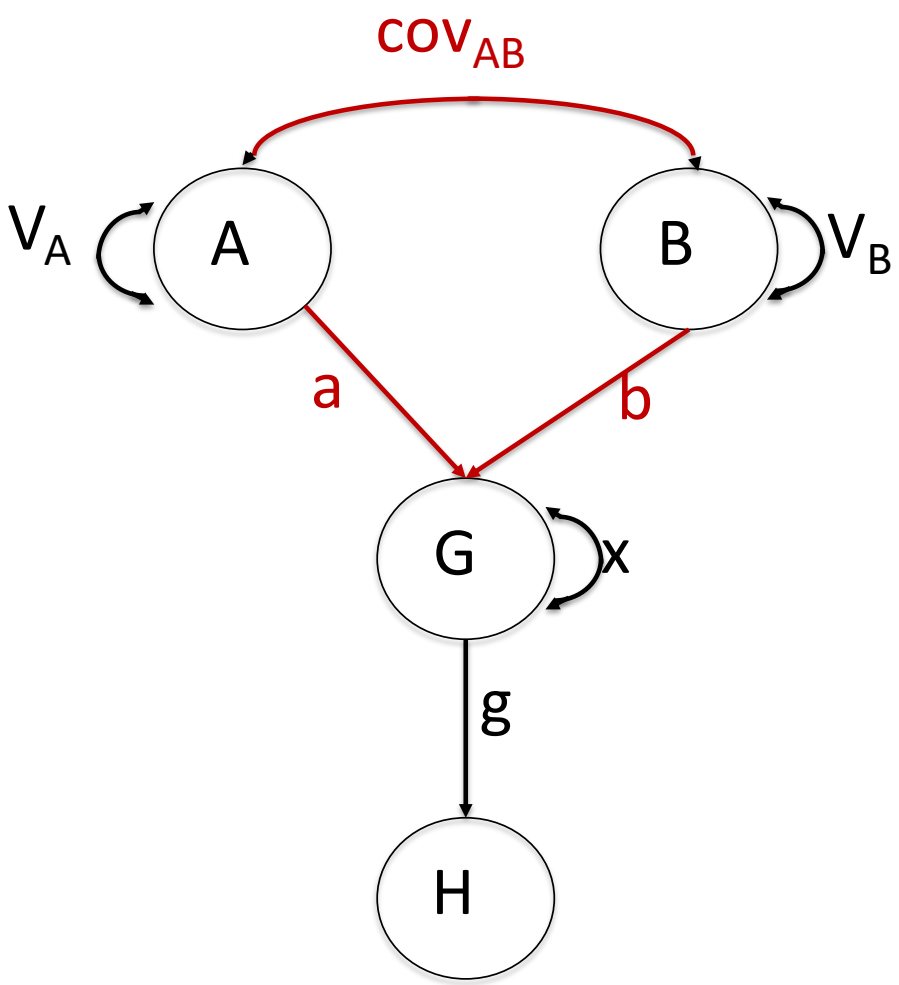
VAR(G) = x



Expected variance

VAR(G) = x

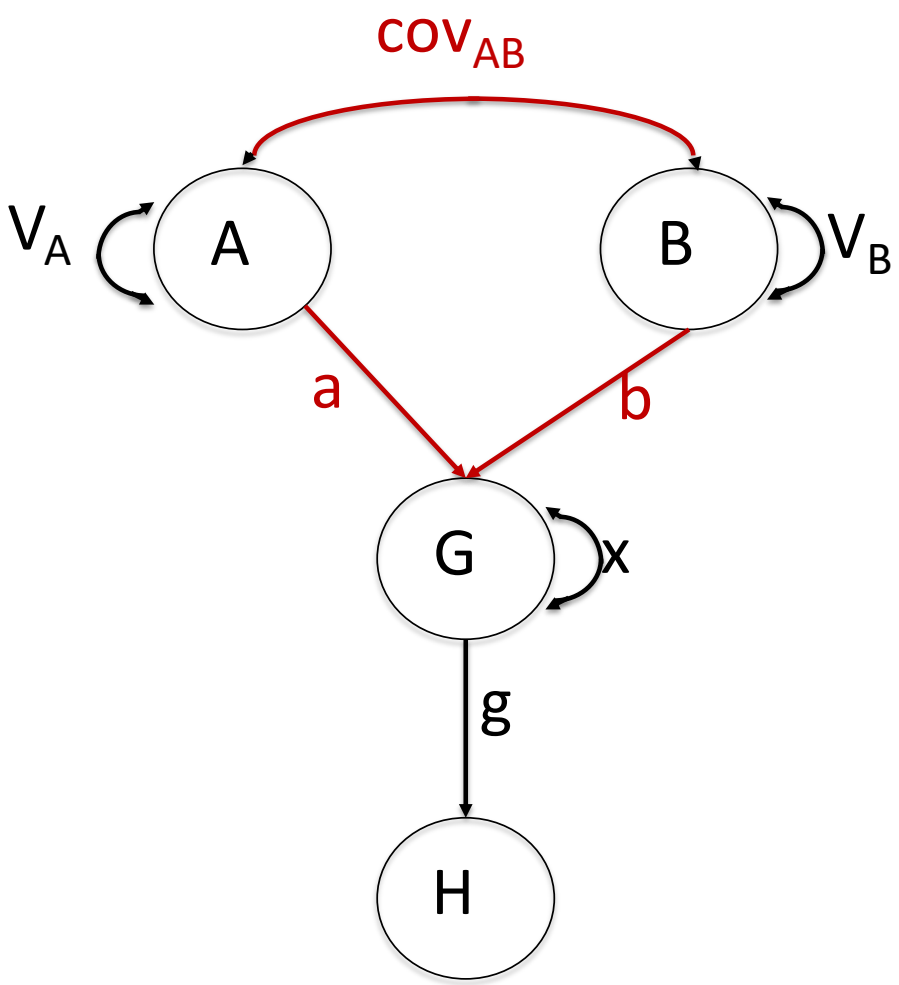
(Residual variance or measurement error)



Expected variance

VAR(G) = x +

$b * COV_{AB} * a$

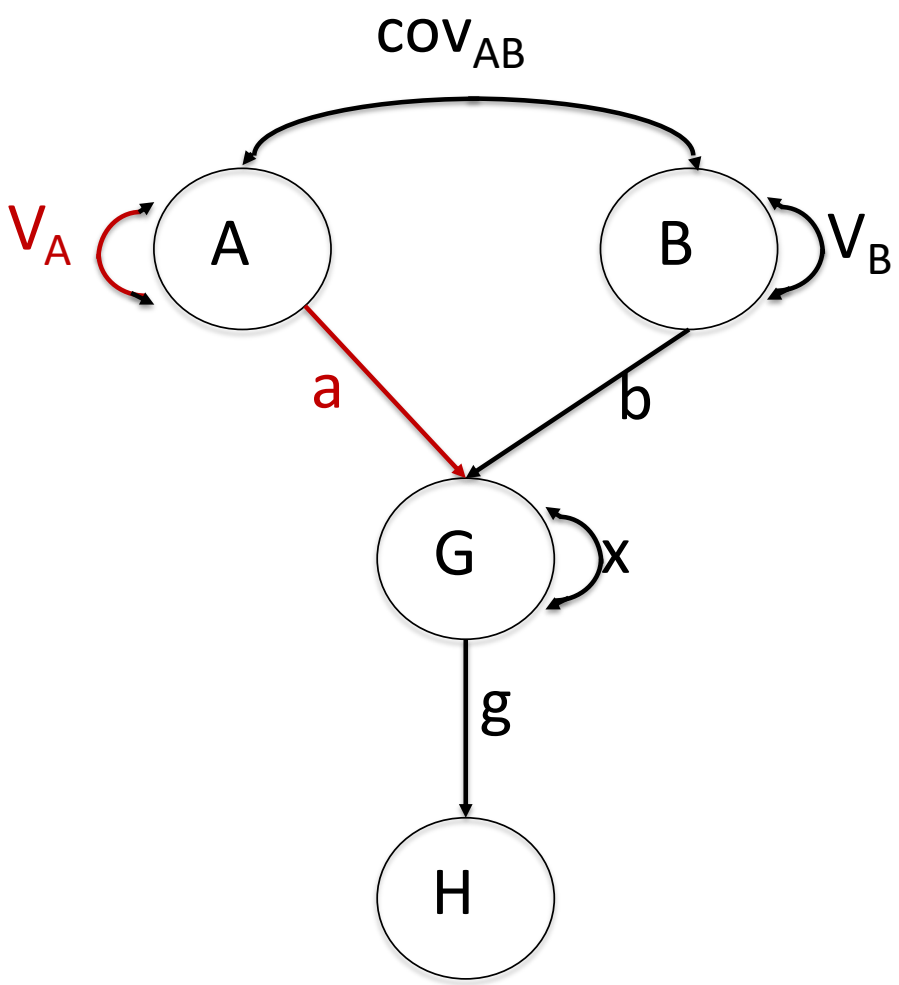


Expected variance

$VAR(G) = x +$

$b * COV_{AB} * a +$

$a * COV_{AB} * b$



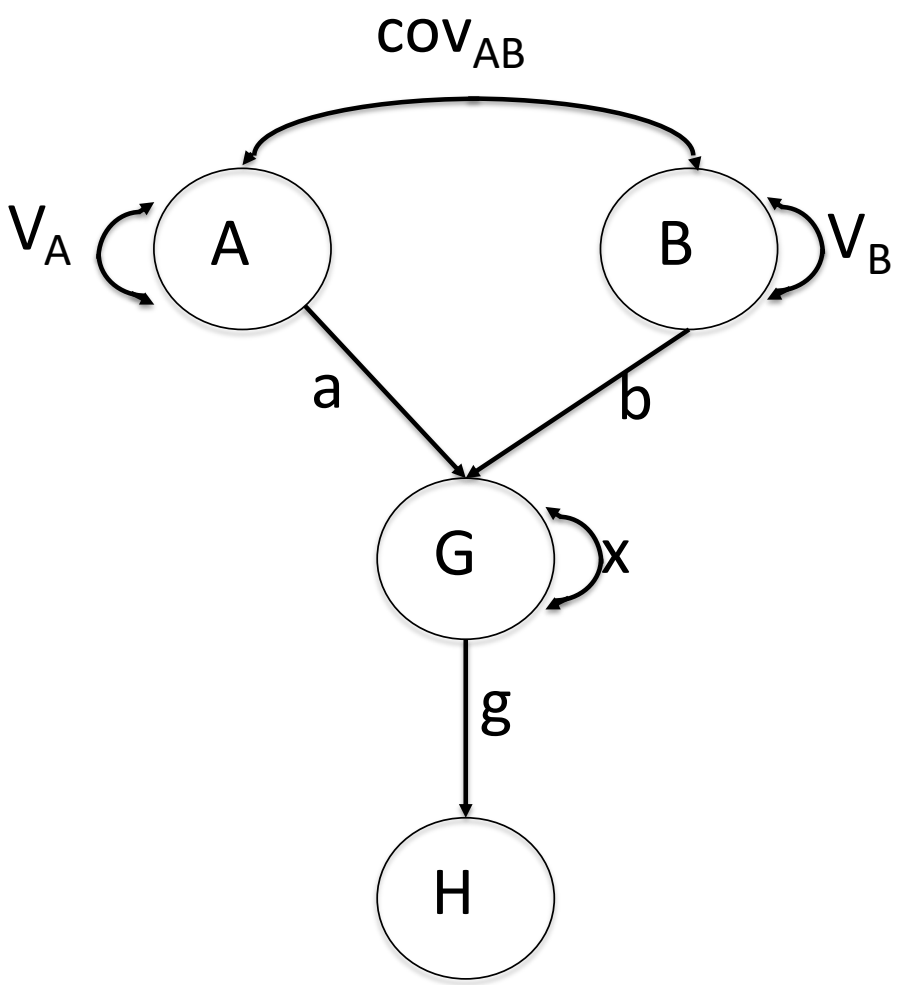
Expected variance

$VAR(G) = x +$

$b * COV_{AB} * a +$

$a * COV_{AB} * b +$

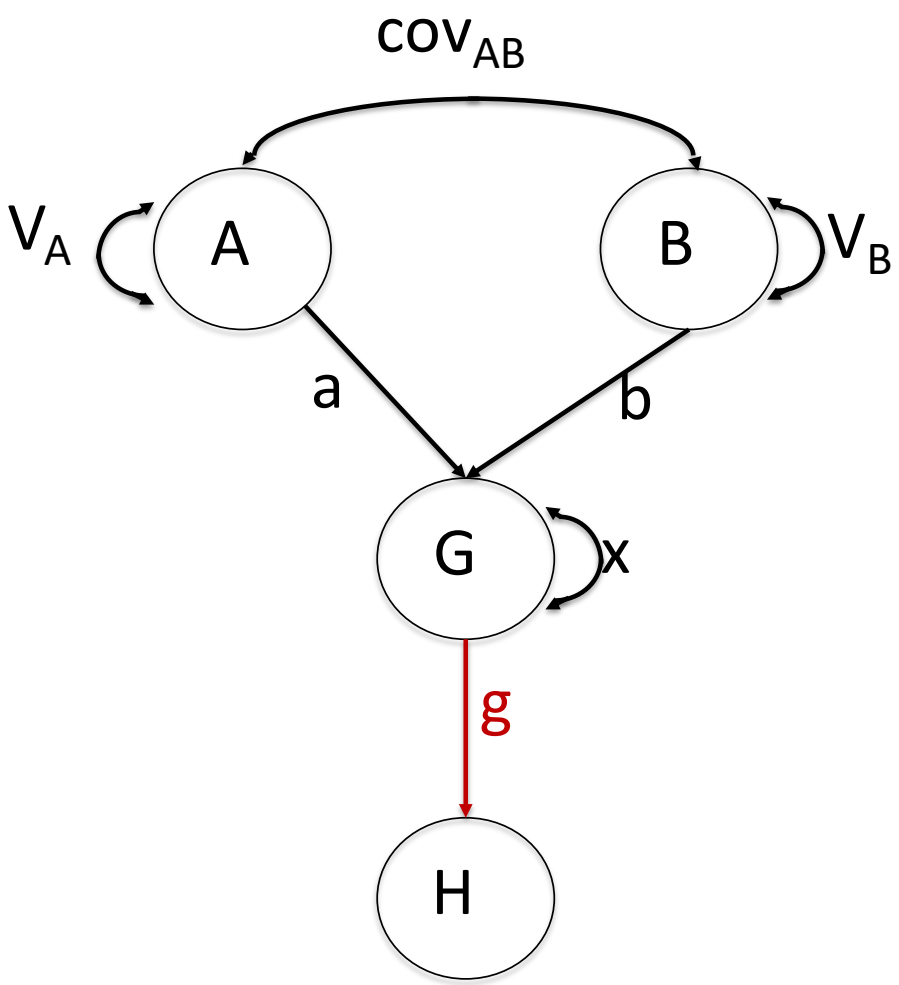
$a * V_A * a$



Expected variance

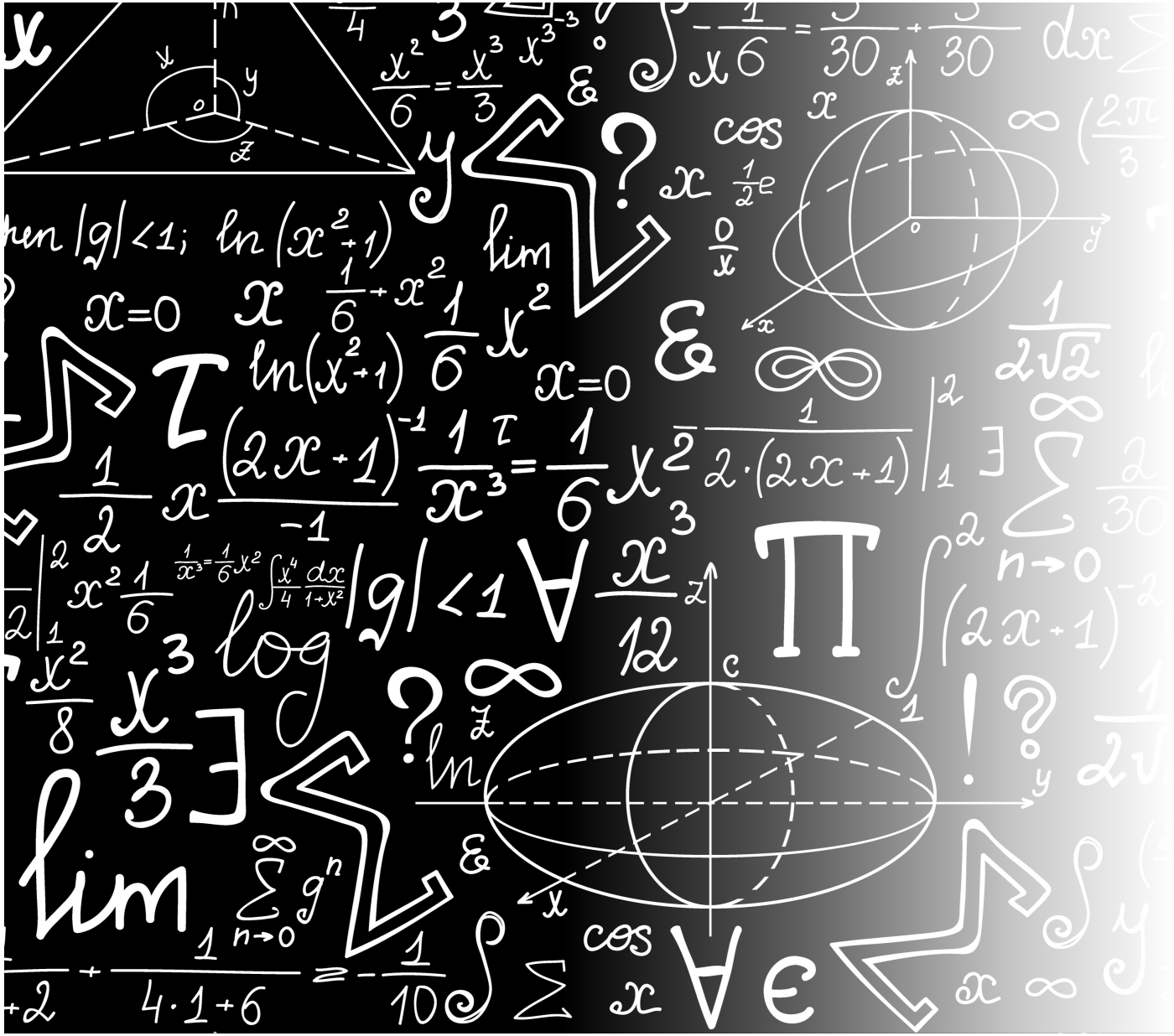
$$VAR(G) = x + b * COV_{AB} * a + a * COV_{AB} * b + a * V_A * a + b * V_B * b$$

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



Expected variance

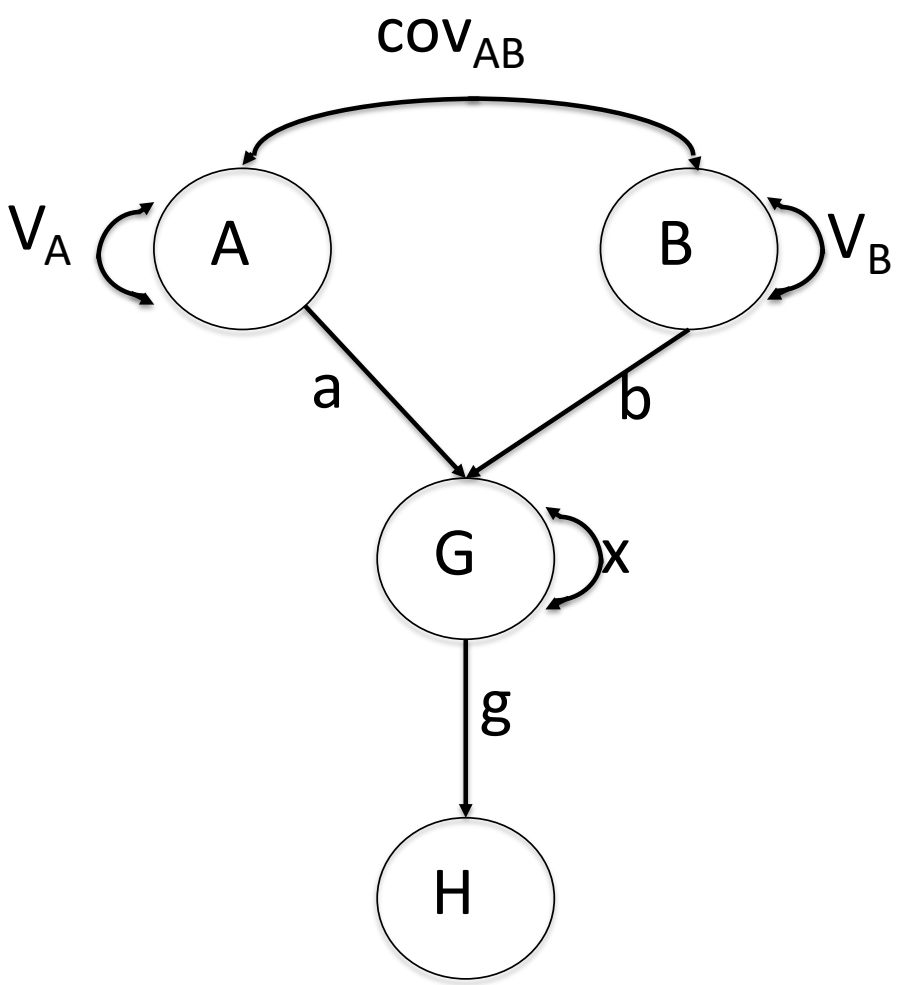
VAR(H) = g * VAR(G) * g



Deriving Expected Variances and Covariances Using Covariance Algebra

Rules of Covariance Algebra

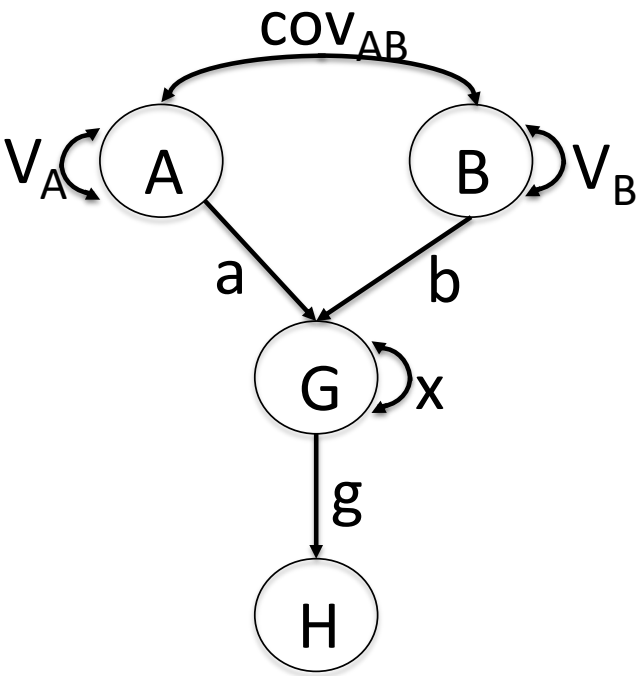
- $COV(c, X) = 0$
- $COV(cX_1, X_2) = cCOV(X_1, X_2)$
- $COV(X_1 + X_2, X_3) = COV(X_1, X_3) + COV(X_2, X_3)$
- $VAR(X_1) = COV(X_1, X_1)$



SEM model

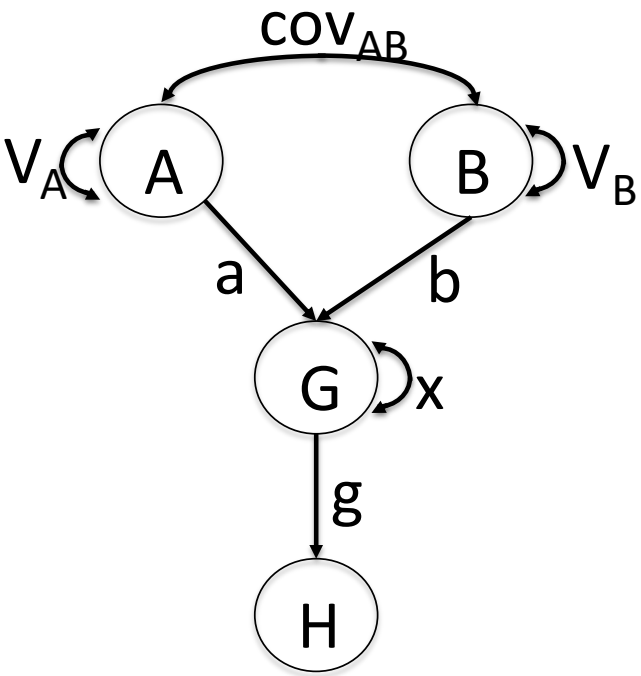
$$\begin{cases} H = g*G \\ G = a*A + b*B + e_X \end{cases}$$

Expected Variance



$$\begin{cases} H = g*G \\ G = a*A + b*B + e_X \end{cases}$$

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

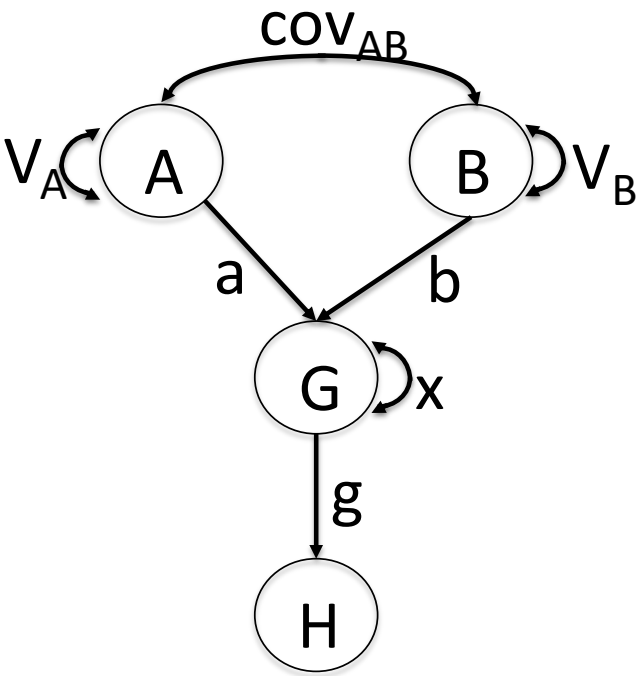


$$\begin{cases} H = g*G \\ G = a*A + b*B + e_X \end{cases}$$

Expected Variance

$$\begin{aligned}
VAR(G) &= COV(G, G) \\
&= COV(a*A + b*B + e, a*A + b*B + e) \\
&= COV(a*A, a*A) + COV(a*A, b*B) + COV(a*A, e) \\
&+ COV(b*B, a*A) + COV(b*B, b*B) + COV(b*B, e) \\
&+ COV(e, a*A) + COV(e, b*B) + COV(e, e) \\
&= a*a*COV(A, A) + a*b*COV(A, B) \\
&+ b*a*COV(B, A) + b*b*COV(A, B) \\
&+ COV(e, e) \\
&\boxed{= a^2*V_A + b^2*V_B + 2*a*b*COV_{AB} + x}
\end{aligned}$$

Expected covariance



$$\begin{cases} H = g*G \\ G = a*A + b*B + e_X \end{cases}$$

$$\begin{aligned}
COV(H,A) &= COV(g*G, A) \\
&= COV(g*(a*A + b*B + e_X), A) \\
&= COV(g*a*A + g*b*B + g*e_X), A) \\
&= COV(g*a*A, A) + COV(g*b*B, A) + COV(g*e_X, A) \\
&= g*a*COV(A, A) + g*b*COV(B, A) + g*COV(e_X, A) \\
&= g*a*VAR(A) + g*b*COV(B, A) \\
&\boxed{= g*a*V_A + g*b*COV_{AB}}
\end{aligned}$$

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.