Modeling/Estimating Genetic and Environmental Components

40th International Statistical Genetics Online Workshop 2026

Hermine HM Maes et al.

Outline Day 3

Modeling/Estimating Genetic and Environmental Components

- **Part 1**: Ways to Estimate Heritability
- **Part 2**: Estimate heritability using data from relatives / twins
- **Part 3**: Modeling genetic and environmental sources of variation in a single trait
- **Part 4**: Modeling genetic and environmental sources of variation in multiple traits

Ways to estimate heritability

Part 1

credit to
Loic Yengo, Sarah Medland, Abdel Abdellaoui, Conor Dolan, Michael Neale & others

Outline Part 1

Ways to estimate heritability

- Recap on Quantitative Genetics Theory
- Coefficient of Genetic Relatedness > GRMs
- Haseman-Elston Regression
- HE Regression vs Classical Twin Design
- Linear Regression with one SNP & R Code
- Path Analysis
- From Equation to Path Model [to OpenMx Code]
- Introduction to OpenMx
- Linear Regression with OpenMx

Recap on Quantitative Genetics Theory

What is heritability?

- Heritability h^2 is the proportion of trait variance explained by genetic factors, thus if a trait (=phenotype) **Y** can be modeled as **Y = Genes + Environment** then h^2 = **var(G) / var(Y)**

- Alternatively, h^2 can be derived from the phenotypic correlation **corr(Y_j,Y_k)** between relatives and their (genetic) relatedness R_{jk} using the theorem: **corr(Y_j,Y_k) = h^2 R_{jk} + Residual**
  where R_{jk} = coefficient of genetic relationship between individuals **j** & **k** (e.g., R_{jk} =0.5 for full sibs)

- h^2 quantifies degree to which inter-individual differences and resemblance in populations are due to genetic factors, and can be approached in terms of
  - Differences between people in the population: h^2 = **var(G) / var(Y)**
  - Resemblance between relatives (in families): **corr(Yi,Yj) = h^2 R_{jk} + Residual**

- This fundamental theorem is used to estimate h^2 using individual-level SNP data in Haseman-Elston (HE) regression or Genome-based REstricted Maximum Likelihood (GREML)

Coefficient of Genetic Relationship > GRMs

How to estimate genetic relatedness from SNP data?

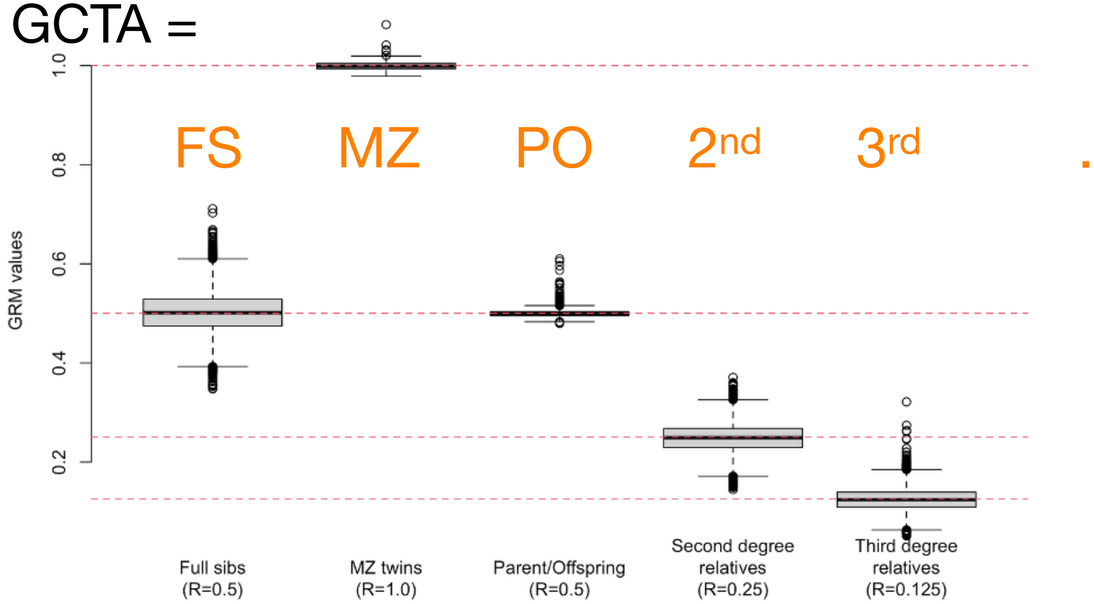
- Coefficient of genetic relationship **R_{jk}** = 2x probability that two alleles picked at random in individuals **j** and **k** are Identical-By-Descent (IBD)
- Genetic Relationship Matrix (GRM) is matrix of **R_{jk}**'s and can be quantified based on expected IBD sharing (e.g., 0.5 for full sibs) or using actual SNP data

- Standard estimator of **R_{jk}** implemented in software GCTA =

$$\hat{\pi}_{jk} = \frac{1}{m}\sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)}$$

where, x_{ij} & x_{ik} = minor allele count (0,1 or 2) at SNP i for individuals j and k respectively, p_i =minor allele frequency (MAF) of SNP i, m =number of SNPs used to calculate GRM

- Expectation of $\hat{\pi}_{jk}$ is exactly **R_{jk}**
- Observed relatedness varies by relationship



Yang, 2011

Haseman-Elston Regression

How do we use the genetic relationship coefficient to estimate SNP heritability?

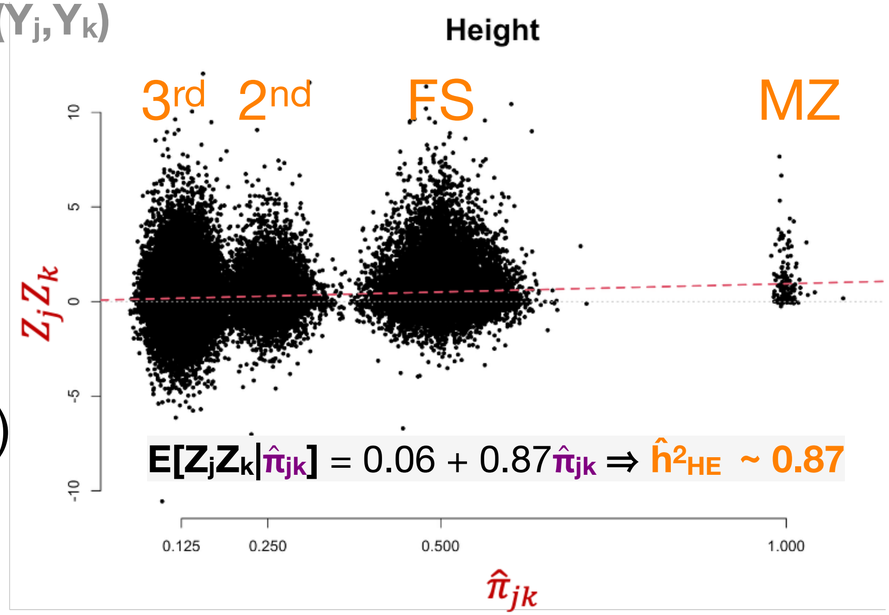
- classic statistical genetics tool to estimate heritability of quantitative traits, relies on association between phenotypic similarity and genetic sharing
  - original model regressed squared difference in siblings' traits on proportion of markers shared IBD to detect linkage
  - extended to use cross-products which increases statistical power

- estimates h^2 by regressing Z_jZ_k [phenotypic similarity] onto $\hat{\pi}_{jk}$ [genetic similarity]

  where $Z_j=(Y_j\text{-mean}(Y))/sd(Y)$ & $Z_k=(Y_k\text{-mean}(Y))/sd(Y)$, $E[Z_jZ_k]= corr(Y_j,Y_k)$

- Cross-Product (CP) method: $Z_jZ_k = \beta_0 + \beta_1\hat{\pi}_{jk} + \varepsilon$
- Slope β_1 directly estimates heritability h^2

- Similar to assumption-free estimation of heritability from genome-wide IBD sharing between full siblings (or twins)



Haseman & Elston, 1972; Elston et al, 2000; Visscher at al, 2006

HE Regression vs Classical Twin Design

How does SNP heritability relate to Twin heritability?

- Shared goal of partitioning phenotypic variance to estimate h^2, by analyzing how phenotypic similarity between individuals relates to their genetic relatedness, with HE-regression using DNA-based relatedness rather than inferred pedigree relationships in the twin design

- **Heritability Estimation**:
  - Haseman-Elston Regression **HE** (Molecular): regresses phenotypic cross-product (CP) of siblings (or pairs of unrelated individuals) on shared proportion of alleles, estimated via Identity-By-State (IBS) or Identity-By-Descent (IBD)

$$\mathbf{corr(Yi,Yj) = h^2 R_{jk} + Residual}$$

  - Classical Twin Design **CTD** (Traditional): compares monozygotic **MZ** twins to dizygotic **DZ** twins to partition variation in a phenotype into additive genetic **A**, common environmental **C**, and unique environmental **E** components

$$\mathbf{cov = VA \otimes rA + VC \otimes rC + VE \otimes rE}$$

HE vs CTD

Why may estimates of SNP heritability and Twin heritability differ?

- **Assumptions**:
  - **CTD** relies on Equal Environments Assumption (EEA) and assumptions of zygosity (100% vs. 50% relatedness **rA**)
  - **HE**-regression uses DNA markers to calculate GRM to estimate h^2 from pairs of sibs or unrelated individuals and requires that the genetic similarity measure R_{jk} truly captures the genetic variance, without confounding by shared environmental factors

- **Effect Magnitude**:
  - **CTD** often yields higher h^2 estimates than **HE**-regression based on SNPs
  - measuring different components of genetic variance (e.g., rare variants or epistatic interactions captured by twins but missed by GWAS-based HE-regression)

Linear Regression with one SNP

How to go from linear regression with `lm` to structural equation modeling (SEM)?

- **Predict Y** [height] **from X** [SNP]

- **Equation**: $\mathbf{Y_i = b_0 + b_1 * X_i + e_i}$ e.g. GWAS: $Height_i = b_0 + b_1 * SNP_i + e_i$
- variables:
  - Y_i dependent (predicted) variable in participant i (i.e. Height)
  - X_i predictor in participant i (genetic variant: a SNP)
  - e_i residual in participant i
- parameters:
  - b_0 intercept
  - b_1 slope or regression coefficient

- Y, X & e: variables because their value vary over persons
- b_0 & b_1: (fixed) parameters, with unknown values (in well defined population)

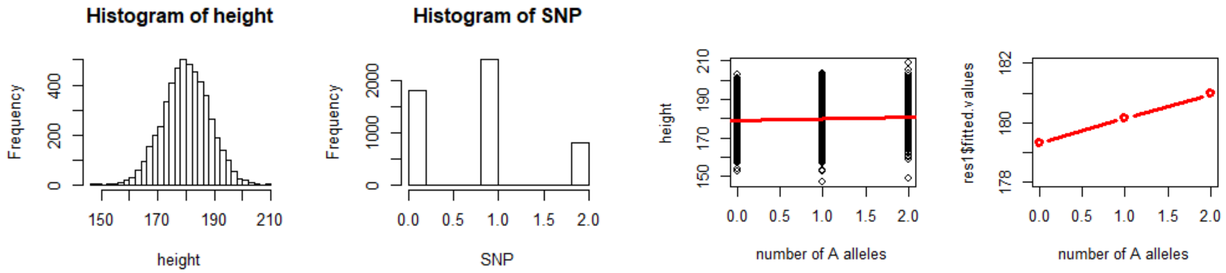
- Are **X** and **Y** linearly related? **Null hypothesis (H_0): $b_1=0$**

SNP: single nucleotide polymorphism

Linear Regression: R code

How to specify linear regression in R: Does SNP predicts height?

N=20k	mean	cov	ht	SNP
ht	180.03	**ht**	64.01	r=0.038
SNP	0.80	**SNP**	0.21	0.48



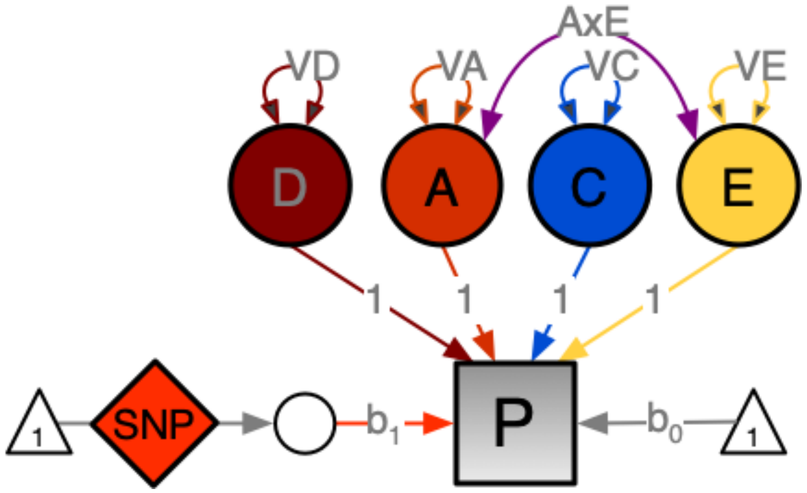

```
lin1A  <-  lm(height~ SNP, data=dataSNP)
                    Estimate    Std. Error t value Pr(>|t|)
b0       (Intercept) 179.67245    0.08635 2080.67  < 2e-16 ***
b1       SNP           0.44226    0.08160    5.42 6.02e-08 ***
Multiple R-squared:  0.0014669, Adjusted R-squared:  0.001417
F-statistic: 29.378 on 1 and 19998 DF,  p-value: 6.0232e-08
```

- Significant association? **H_0**: $b_1=0$, **H_1**: $b_1\neq0$, $p< \alpha$: reject H_0
  - SNP explains variance of height or SNP is associated with height
- Linear additive model: effect of alleles A on height is additive
  - from aa to Aa: +1A= +0.44 (b_1)
  - from aa to AA: +2A's= +0.44+0.44 (additive: $b_1+ b_1$)

SNP alleles **A-a**, genotypes **aa**, **Aa/aA** & **AA**, coded **0**, **1**, **2** (additive coding)

Path Analysis

How to specify models using path diagrams?



- Path Diagram **Conventions**:
  - Squares or rectangles denote observed variables
  - Circles or ellipses denote latent (unmeasured) variables
  - Single-headed arrows or paths (–>): causal relationships between variables where variable at tail has influence on variable at head
  - Double-headed arrows (<–>): covariance between two variables, through common causes not in modeled; or variance of a variable
  - Triangles denote means; Diamonds denote definition variables (variables on paths)

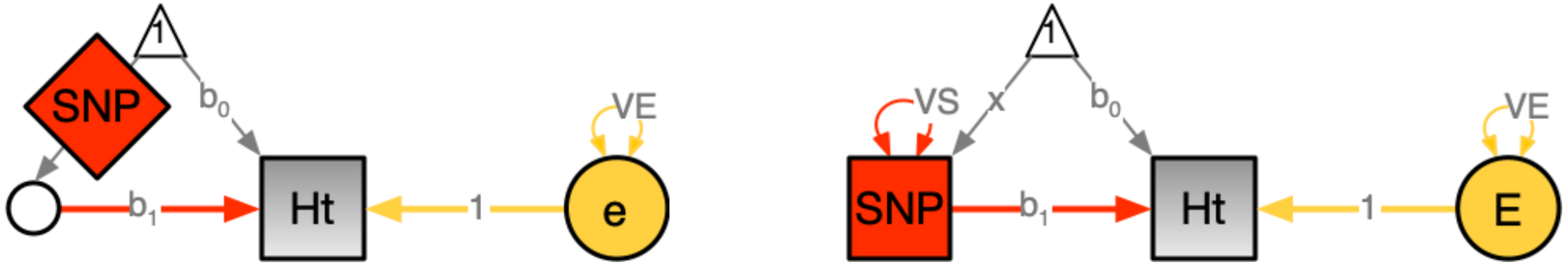
- Path Analysis **Tracing Rules**:
  - Trace backwards, change direction at a 2-headed arrow, then trace forwards (never trace through two-headed arrows in same chain)
  - Expected covariance between two variables, or variance of a variable, is computed by multiplying together all coefficients in a chain, and summing over all possible chains

Sewall Wright, 1918; https://en.wikipedia.org/wiki/Path_analysis_(statistics)

From Equation to Path Model [to OpenMx Code]

How do equation translate into path model?

- linear regression model as equation: **Height$_i$** = b_0 + b_1*SNP$_i$ + e_i
- linear regression model as path diagram



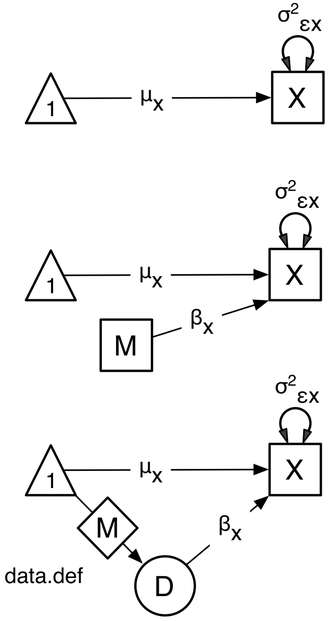
- model implied means and covariance structure

mean	Ht
	$b_0 + b_1$*SNP

mean	Ht	SNP
	$b_0 + b_1$*SNP	x

variance	Ht
Ht	VE

cov	Ht	SNP
Ht	b_1^2*VS+ VE	b_1*VS
SNP	b_1*VS	VS



definition variables on paths

Introduction to OpenMx

Why OpenMx?

Typical Script Outline

INTRO to OpenMx SCRIPT Comments Load Libraries & Options Create Output
PREPARE DATA Load Data Select Variables for Analysis Select Data for Analysis Generate Descriptive Statistics Set Starting Values
PREPARE MODEL Create Algebra for Mean Matrices Create Algebra for Variance/Covariance Matrices Create Data Objects for Multiple Groups Create Expectation Objects for Multiple Groups Create Model Objects for Multiple Groups Create Confidence Interval (CI) Objects
BUILD & RUN MODEL Build Model [with CIs] Run Model Print Goodness-of-fit Statistics Print Parameter Estimates

What is **OpenMx**?

OpenMx is free and open source software for use with **R** that allows estimation of a wide variety of advanced multivariate statistical models. **OpenMx** consists of a library of functions and optimizers that allow you to quickly and flexibly define an SEM model and estimate parameters given observed data.

OpenMx runs on MacOS, Windows, and most varieties of Linux/GNU. This means the same scripts you write in Windows will run in MacOS or Linux.

OpenMx can be used by those who think in terms of path models or by those who prefer to specify models in terms of matrix algebra. **OpenMx** is extremely powerful, taking full advantage of the **R** programming environment. This means that complicated models and data sets can be specified and modified using the **R** language. In order to give a very brief idea of what **OpenMx** looks like, here are two small demo examples: one from a path modeler's perspective and one from a matrix algebra perspective.

Syntax

```
mxData( observed=, type="" )
mxMatrix( type="Full", nrow=?, ncol=?, free=F, labels=c("data.?"), name="" )
mxMatrix( type="", nrow=?, ncol=?, free=T/F, values=?, labels="", name="" )
mxAlgebra( expression= beta0 + beta1 %*% SNP, name="" )
mxExpectationNormal( covariance="", means="", dimnames=c("") )
mxFitFunctionML()
mxModel( "", objects, …. )
mxRun( model)
mxGetExpected( fittedmodel, "mean")
mxEval( expression=), fittedmodel )
omxSetParameters( fittedmodel, label="", free=FALSE, values=?, name="")
mxCompare( model1, model2)
```

openmx.ssri.psu.edu/ Boker et al 2011; Neale et al 2016

Linear Regression: OpenMx code

How to specify linear regression in OpenMx?

- $height_i = b_0 + b_1 * SNP_i + e_i$

Data

```
dataRaw     <- mxData( observed=dataSNP[,c("SNP","height")], type="raw" )
defSNP      <- mxMatrix( type="Full", nrow=1, ncol=1, free=FALSE, labels=c("data.SNP"), name="SNP" )
```

Model

```
intercept   <- mxMatrix( type="Full", nrow=1, ncol=1, free=TRUE, values=0, labels="b0", name="beta0" )
regCoef     <- mxMatrix( type="Full", nrow=1, ncol=1, free=TRUE, values=1, labels="b1", name="beta1" )
resVar      <- mxMatrix( type="Diag", nrow=1, ncol=1, free=TRUE, values=1, labels="e", name="resVar" )
expMean     <- mxAlgebra( expression= beta0 + beta1 %*% SNP, name="expMean" )
expCov      <- mxAlgebra( expression= resVar, name="expCov" )
```

Fit Model to Data

```
exp         <- mxExpectationNormal( covariance="expCov", means="expMean", dimnames=c("height") )
funML       <- mxFitFunctionML()
modReg      <- mxModel( "Simple Regression", dataRaw, defSNP, intercept, regCoef, resVar, expMean, expCov, exp, funML )
fitReg      <- mxRun(modReg);  summary(fitReg); mxGetExpected(fitReg, "mean")
```

- compare model estimating b_1 with one fixing it to 0 to test significance of b_1

Test Regression

```
varSNP      <- var(dataSNP$SNP1)
Rsquared    <- mxEval( expression= beta1^2 %*% varSNP / (beta1^2 %*% varSNP + resVar), fitReg)
fitNob1     <- mxRun( omxSetParameters( fitReg, label="b1", free=FALSE, values=0, name="fitNob1"))
summary(fitNob1);  print(Rsquared); mxCompare(fitReg, fitNob1)
```

Linear Regression: Results in R & OpenMx

Do results from `lm` in R match those from OpenMx?

R

```
             Estimate    Std. Error t value Pr(>|t|)
b0      (Intercept) 179.67245    0.08635 2080.67  < 2e-16 ***
b1              SNP   0.44226    0.08160    5.42 6.02e-08 ***
Multiple R-squared:  0.0014669,    Adjusted R-squared:  0.001417
F-statistic: 29.378 on 1 and 19998 DF,  p-value: 6.0232e-08
```

OpenMx

```
free parameters:          name matrix row col      Estimate   Std.Error A
                     1    b0  beta0   1   1 179.67173607 0.086335752 !
                     2    b1  beta1   1   1   0.44569639 0.081569831 !
                     3     e resVar   1   1  63.91128138 0.639127709 !
Model Statistics:                    |  Parameters  |  Degrees of Freedom  |  Fit (-2lnL units)
                           Model:                3                 19997            139906.88
                     Number of observations/statistics: 20000/20000
Information Criteria:          |  df Penalty  |  Parameters Penalty  |  Sample-Size Adjusted
                     AIC:          99940.239              139940.24              139940.24
                     BIC:         -58113.705              139956.05              139949.69
R^2= 0.0014897596
                base comparison ep  minus2LL    df       AIC    diffLL diffdf             p
1 Simple Regression       <NA>  3 139906.88 19997 139912.88        NA     NA            NA
2 Simple Regression      b1Sig  2 139936.24 19998 139940.24 29.356508      1 6.0213961e-08
```

Why not OpenMx?

openmx.ssri.psu.edu/

- looks complicated, has steep learning curve
- truly atomic, can build any model you want
- can build models for different kinds of data, can mix & match
- from individual time series to population-based studies
- combine different datasets or types of data for meta-analysis
- runs on windows, Mac and linux systems
- AIs can help you code - not perfect yet but improving
- very large datasets and models possible - RAM becomes a limit eventually
- user-defined fit-functions or likelihoods for non-standard applications
- can interface to drawing software
- has its own drawSEM graphical add-on under development

- is powerful and flexible
- is free and open-source
- scripts can be put into functions, just like `lm`
- developed with genetically informative data analysis in mind

Estimate heritability using data from relatives / twins

Part 2

credit to
Elizabeth Prom-Wormley, Sarah Medland, Lucía Colodro Conde, Jose Morosoli Garcia, Michael Neale & others

Outline Part 2

Estimate heritability using data from relatives / twins

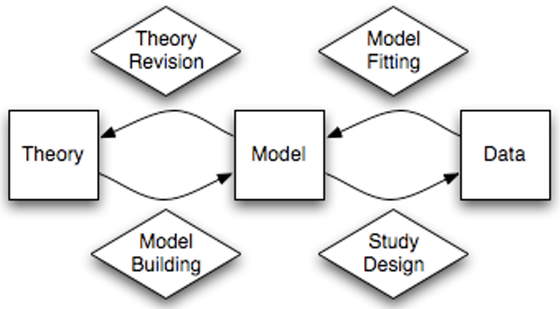
- The Classical Twin Study
- Sources of Variation
- Patterns of Twin Correlations
- Twin Model Assumptions
- Maximum Likelihood
- Saturated Models: Specification of Data & Model in OpenMx
- Goodness of Fit Statistics - Likelihood Ratio Tests
- Parameter Estimates & ML Correlations

The Classical Twin Study

How to estimate twin heritability?

- Partition phenotypic variance (V) in genetic and environmental components
  - $\mathbf{V}_{total}$ = **VG**enetic + **VE**nvironmental
  - Assumptions: additivity & independence of effects
  - Heritability ($\mathbf{h^2}$): proportion of variance due to genetic influences ($\mathbf{h^2}$ = **VG**enetic / $\mathbf{V}_{total}$)
  - Property of a group (not an individual), thus specific to that group in place & time

- Classical Twin Study/Design (CTS/CTD): monozygotic (MZ) and dizygotic (DZ) twins reared together
  - **MZ** twins share **100%** of their genes
  - **DZ** twins share on average **50%** of their genes
- Genetic factors are assumed to contribute to a phenotype when MZ twins are more similar for that phenotype than DZ twins



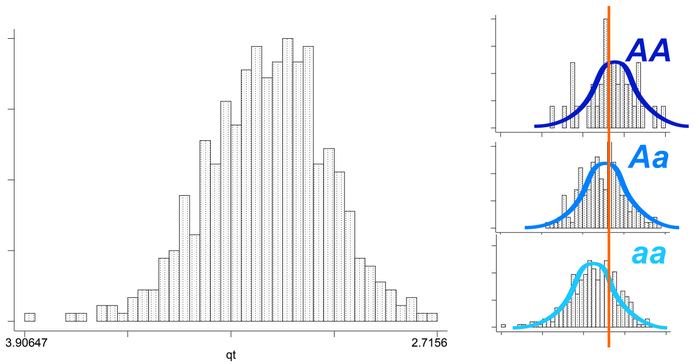
Eaves et al, 1978; Boomsma et al, 2002; Van Dongen et al, 2012; Polderman et al, 2015

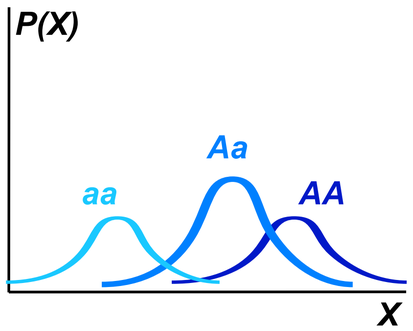
Sources of Variance

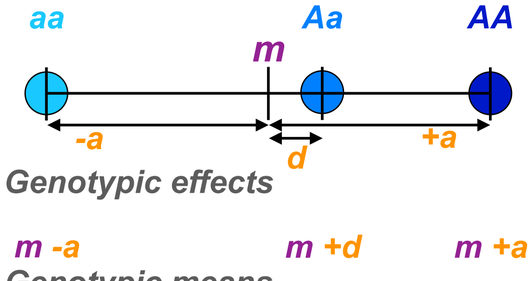
Which factors make people different from one another?

- Additive Genetic factors (**VA**, **A**, a^2): sum of all average effects of single alleles at individual loci
- Genetic Dominance (**VD**, **D**, d^2): result of interactions between alleles at same locus
- Common Environment: (**VC**, **C**, c^2): aspects of environment shared by family members, which contribute to similarity between relatives [shared]
- Unique Environment (**VE**, **E**, e^2): unique to individual, contribute to variation within family [specific, unique or within-family]

- Biometrical Model





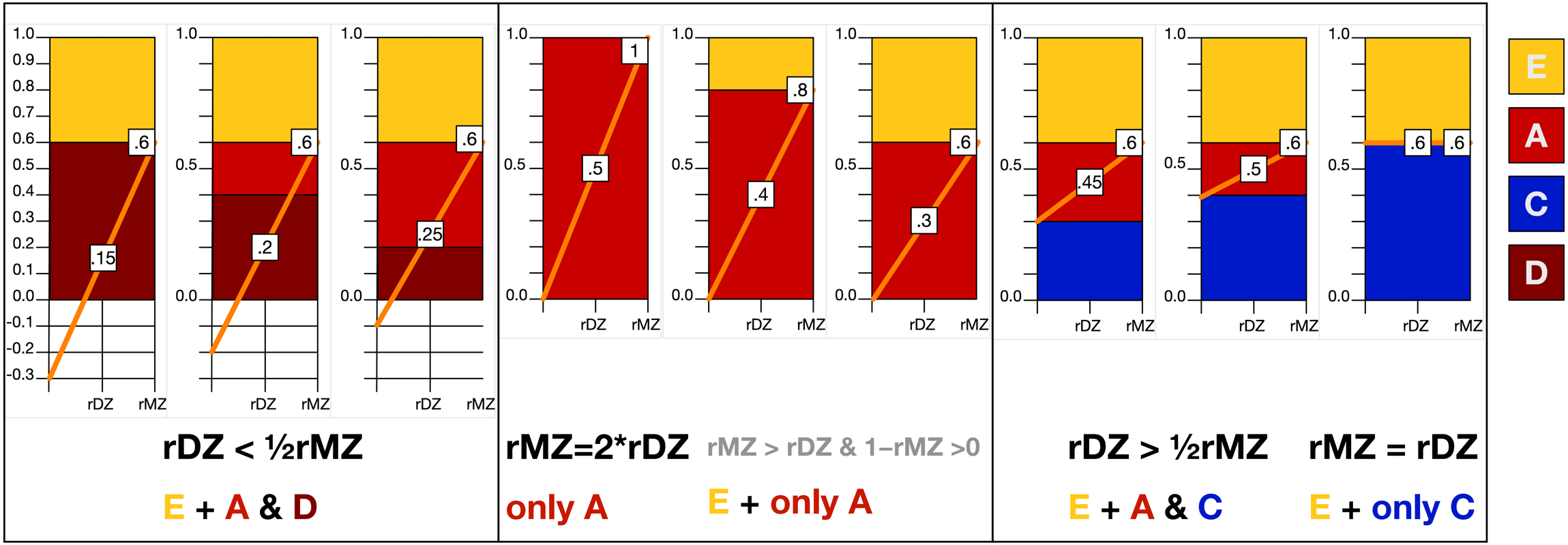


Fisher, 1918; Mather & Jinks, 1972; Martin & Eaves, 1977; Neale & Cardon, 1992

Patterns of Twin Correlations

Where does the information come from to estimate **A**, **C**, **D** & **E**?

From twin correlations to expected sources of variance



rDZ < ½rMZ

E + A & D

rMZ=2*rDZ rMZ > rDZ & 1–rMZ >0

only A **E + only A**

rDZ > ½rMZ **rMZ = rDZ**

E + A & C **E + only C**

rMZ: monozygotic twin correlation; **rDZ**: dizygotic twin correlation

A: additive genetic variance; **E**: unique environmental variance: **C**: common environmental variance; **D**: dominance genetic variance

Twin Model Assumptions

What assumptions are we making?- Important Quality Control Step!

- Equal Environments of MZ and DZ pairs (EEA)
  - MZ & DZ twins equally correlated in exposure to environmental events of etiologic importance for trait
- Random Mating, No rGE Correlation, No G x E Interaction
- No Sex Limitation, No G x Age Interaction

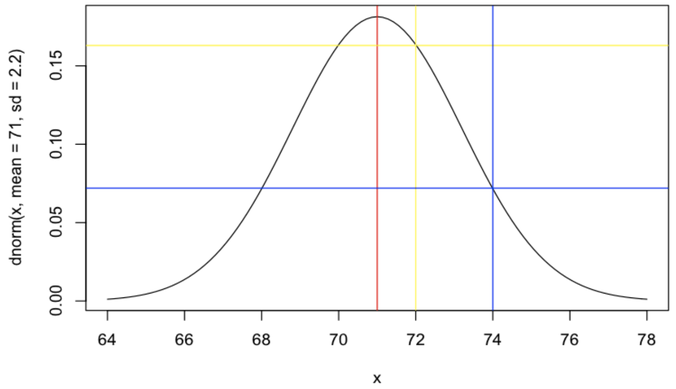
- **Basic Data Assumptions**: MZ & DZ twins sampled from same population, therefore we expect
  - Equal means/variances in Twin 1 and Twin 2 [twin order]
  - Equal means/variances in MZ and DZ twins [zygosity]
  - Further assumptions possibly needed when testing for heterogeneity by sex or adding other relatives [full siblings etc.]

Evans et al, 1999

Maximum Likelihood

How to evaluate statistically which model fits the data best?

- Likelihood (L): probability that observation (data point) is predicted by model
- Maximum Likelihood Estimates (MLE): most likely values of population parameter values (e.g, **μ**, **σ**, **β**) given observed sample values
- Likelihood Ratio (LR): test statistic comparing Log-Likelihoods under two separate models: Mu: Unconstrained & Mc: Constrained (fewer parameters): **LR**(Mc|Mu) = 2ln(L(Mu) - 2ln(L(Mc)
- Likelihood Ratio Test (LRT): asymptotically distributed as χ^2_{df} with degrees of freedom (**df**) equal to number of constraints (or parameters)
- If associated **p**-value < alpha (e.g. 0.05), Null hypothesis is rejected

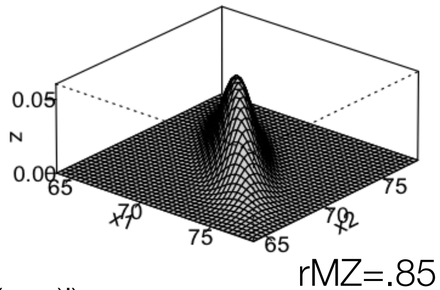


$$\Phi\left(x_i\right) = -|2\pi\sigma^2|^{-.5}\ e^{-5((x_i-\mu)^2/\sigma^2)}$$

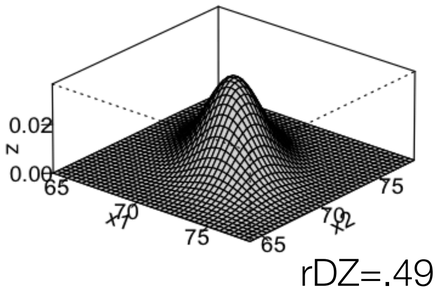
Univariate

Multivariate

$$\Phi\left(x_i\right) = -|2\pi\Sigma|^{-n/2}\ e^{-5((x_i-\mu)\Sigma^{-1}(x_i-\mu)')}$$



rMZ=.85



rDZ=.49

Saturated Models

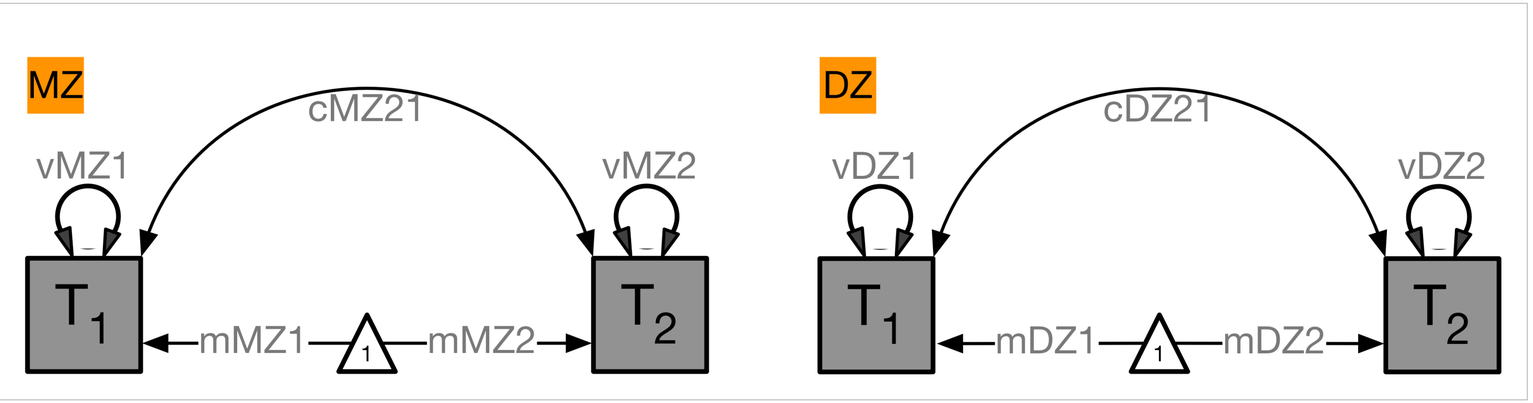
Are data assumptions valid: can we equate means/variances?

- Data: BMI in young OZ female MZ & DZ twins: observed means & variances [NH&MRC 1981 Q]

MZf twins pairs N=534		
cov	**T1**	**T2**
T1	0.73	0.59
T2	0.59	0.79
mean	21.34	21.35

DZf twins pairs N=328		
cov	**T1**	**T2**
T1	0.77	0.24
T2	0.24	0.82
mean	21.45	21.46

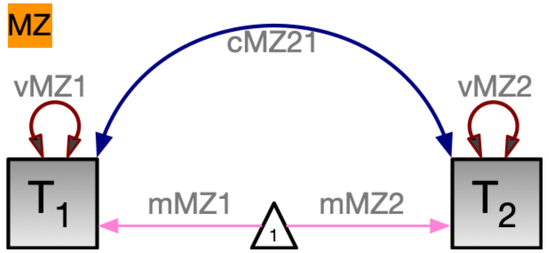
- Model: as many free parameters as statistics



- Lindon Eaves from his Hollywood Cemetery grave: ‘look at the bloody data’

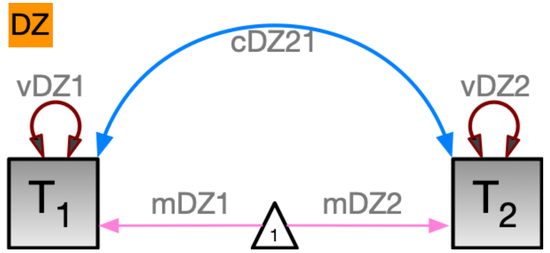
Specification of Data & Model in OpenMx

What are the key commands for OpenMx script of saturated twin model?



covMZ *2x2*	vMZ1	cMZ21
	cMZ21	vMZ2

meanMZ *1x2*	mMZ1	mMZ2



covDZ *2x2*	vDZ1	cDZ21
	cDZ21	vDZ2

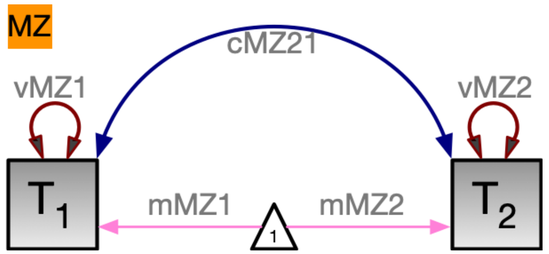
meanDZ *1x2*	mDZ1	mDZ2

```
dataMZ     <- mxData( observed=mzData, type="raw" )
meanMZ     <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=svMe, labels=c("mMZ1","mMZ2"),name="meanMZ" )
covMZ      <- mxMatrix( type="Symm", nrow=ntv, ncol=ntv, free=TRUE, values=valDiag(svVa,ntv), lbound=valDiag(lbVa,ntv),
 labels=c("vMZ1","cMZ21","vMZ2"), name="covMZ" )
```

```
dataDZ     <- mxData( observed=dzData, type="raw" )
meanDZ     <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=svMe, labels=c("mDZ1","mDZ2"),name="meanDZ" )
covDZ      <- mxMatrix( type="Symm", nrow=ntv, ncol=ntv, free=TRUE, values=valDiag(svVa,ntv), lbound=valDiag(lbVa,ntv),
 labels=c("vDZ1","cDZ21","vDZ2"), name="covDZ" )
```

Specification of Data & Model in OpenMx

What are the key commands for OpenMx script of saturated twin model?

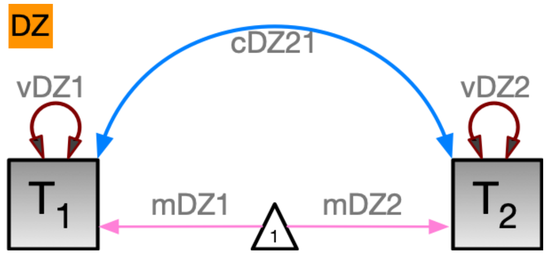


covMZ *2x2*

vMZ1	cMZ21
cMZ21	vMZ2

meanMZ *1x2*

mMZ1	mMZ2



covDZ *2x2*

vDZ1	cDZ21
cDZ21	vDZ2

meanDZ *1x2*

mDZ1	mDZ2

```
> fitSAT$MZ.meanMZ
FullMatrix 'meanMZ'

$labels
     [,1]   [,2]
[1,] "mMZ1" "mMZ2"

$values
          [,1]      [,2]
[1,] 21.344377 21.349013

$free
     [,1] [,2]
[1,] TRUE TRUE

$lbound: No lower bounds assigned.

$ubound: No upper bounds assigned.
```

```
> fitSAT$MZ.covMZ
SymmMatrix 'covMZ'

$labels
     [,1]    [,2]
[1,] "vMZ1"  "cMZ21"
[2,] "cMZ21" "vMZ2"

$values
          [,1]       [,2]
[1,] 0.72767167 0.59164085
[2,] 0.59164085 0.79320287

$free
     [,1] [,2]
[1,] TRUE TRUE
[2,] TRUE TRUE

$lbound
      [,1]  [,2]
[1,] 1e-04 0e+00
[2,] 0e+00 1e-04

$ubound: No upper bounds assigned.
```

```
> fitSAT$DZ.meanDZ
FullMatrix 'meanDZ'

$labels
     [,1]   [,2]
[1,] "mDZ1" "mDZ2"

$values
          [,1]      [,2]
[1,] 21.447522 21.457844

$free
     [,1] [,2]
[1,] TRUE TRUE

$lbound: No lower bounds assigned.

$ubound: No upper bounds assigned.
```

```
> fitSAT$DZ.covDZ
SymmMatrix 'covDZ'

$labels
     [,1]    [,2]
[1,] "vDZ1"  "cDZ21"
[2,] "cDZ21" "vDZ2"

$values
          [,1]       [,2]
[1,] 0.76919672 0.24004099
[2,] 0.24004099 0.82162754

$free
     [,1] [,2]
[1,] TRUE TRUE
[2,] TRUE TRUE

$lbound
      [,1]  [,2]
[1,] 1e-04 0e+00
[2,] 0e+00 1e-04

$ubound: No upper bounds assigned.
```

Goodness of Fit Statistics - Likelihood Ratio Tests

Are data assumptions met?

		os	ep	-2ll	df	AIC	Δ-2ll	Δdf	p
	saturated	1777	10	4055.93	1767	521.93			
m1	**meanT1=meanT2**	1777	8	4056.00	1769	518.00	0.07	2	0.97
m2	**+ varT1=varT2**	1777	6	4058.94	1771	516.94	3.01	4	0.56
m3	**+ mean/varMZ=DZ**	1777	4	4063.45	1773	517.45	7.52	6	0.28

os	observed statistics
ep	estimated parameters

-2ll	-2 LogLikelihood	
df	degrees of freedom	os - ep
AIC	Akaike's Information Criterion	-2ll -2df

Δ-2ll	likelihood ratio Chi-square
Δdf	difference in df
p	probability of Chi-square

- BMI in young OZ females (age 18-30)
  - means [m1] & variances [m2] of twin 1 & 2 not significantly different in MZ & DZ pairs
  - means & variances of MZs and DZs [m3] not significantly different from one another
- Basic data assumptions of classical twin study [CTD] met

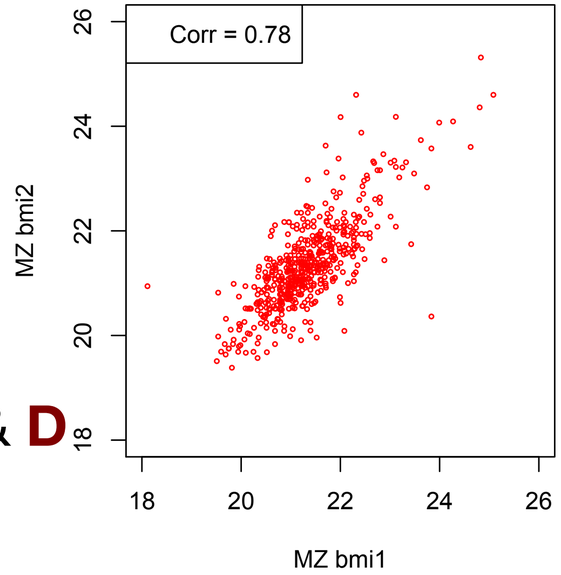
Parameter Estimates & ML Correlations

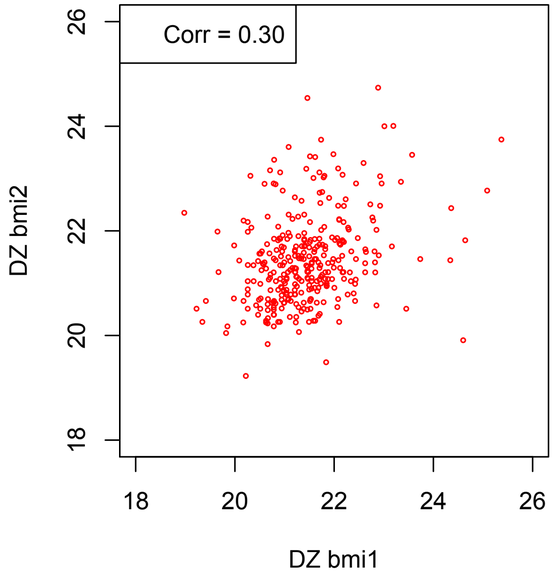
What does the pattern of MZ/DZ correlations predict about h^2?

- 4 parameters estimated in reduced model: `mZ,vZ,cMZ21 cDZ21`

MZf twins pairs N=534		
cov	**T1**	**T2**
T1	0.78	r=0.78
T2	0.61	0.78
mean	21.39	21.39

DZf twins pairs N=328		
cov	**T1**	**T2**
T1	0.78	r=0.30
T2	0.23	0.78
mean	21.39	21.39





- rMZ < 1: **E**
- rMZ > rDZ: **A** (+**D**)
- rDZ < 1/2rMZ: **A** & **D**

Modeling genetic and environmental sources of variation in a single trait

Part 3

credit to
Elizabeth Prom-Wormley, Sarah Medland, Matt Keller, Lucía Colodro Conde, Michael Neale & others

Outline Part 3

Estimate genetic and environmental sources of variation in a single trait

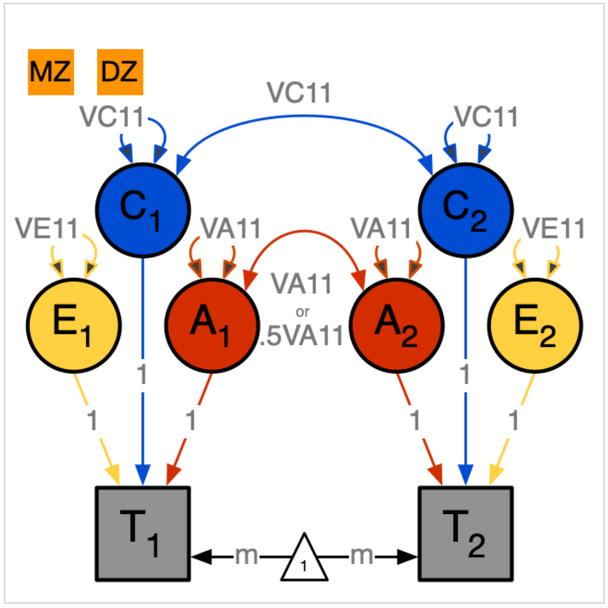
- The ACE vs ADE Model
- Path Coefficients vs Direct Variance Estimation
- One Trait - Univariate Twin Model Script | One Trait - VA ⊗ Genetic Relatedness
- Interpreting Goodness of Fit - Parameter Estimates
- Continuous vs Binary/Ordinal Data - Threshold Model
- Including Covariates - Fixed Effects
- Testing Heterogeneity - Sex Differences in (Random) Variance Components
- Including Siblings / Parents / Extended Relatives

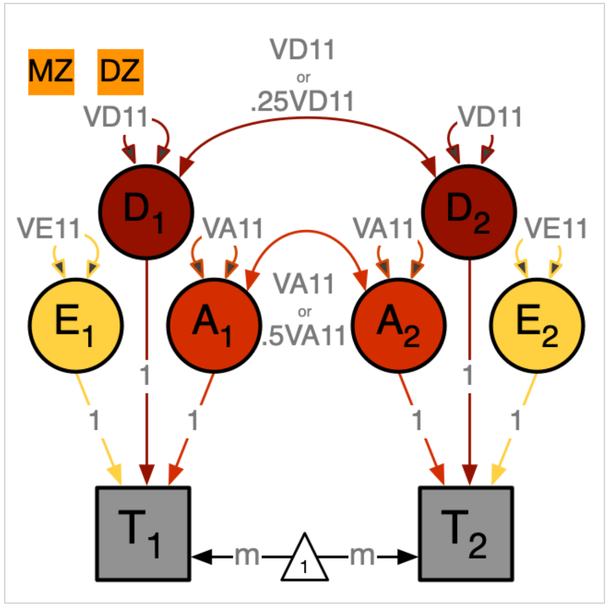
The ACE vs ADE Model

E= VE11
A= VA11
C= VC11
D= VD11

What are contributions of **A**, **C**, **D** & **E** factors to the variance?

- Classical Twin Design has three unique statistics: variance & MZ & DZ covariance, allowing estimating of three parameters: **A**, **C** & **E** or **A**, **D** & **E**
- **C** and **D** are confounded as they have opposite expectations





MZ	T1	T2
T1	A+C+E	A+C
T2	A+C	A+C+E

DZ	T1	T2
T1	A+C+E	½⊗A+C
T2	½⊗A+C	A+C+E

MZ	T1	T2
T1	A+D+E	A+D
T2	A+D	A+D+E

DZ	T1	T2
T1	A+D+E	½⊗A+¼⊗D
T2	½⊗A+¼⊗D	A+D+E

Univariate

- 1-rMZ: **E**
- rMZ > rDZ: **A** (+**D**)
- rMZ = 2*rDZ: only **A**
- rMZ = rDZ: only **C**
- 1/2rMZ < rDZ: **A** & **C**
- 1/2rMZ > rDZ: **A** & **D**

Path Coefficients vs Direct Variance Estimation

Why is Direct Symmetric Variance Estimation preferred?

- Path Estimation bounds variance components to be positive
  - May make incorrect inferences when comparing alternative models
- Direct (Symmetric) Variance Estimation allows variances to be estimated as negative
  - Makes correct inferences when comparing alternative models
  - Allows calculating **VD** when fitting **ACE** or calculating **VC** when fitting **ADE**

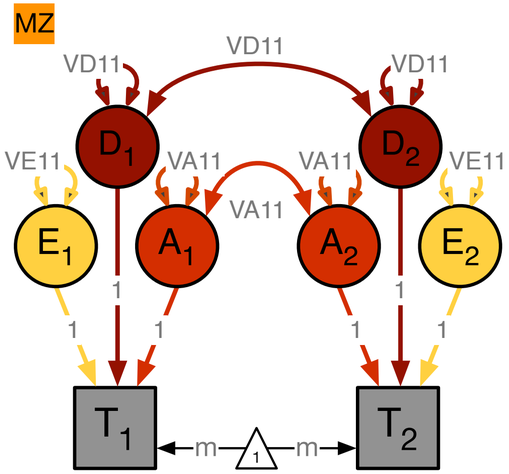
- Get **VD** when fitting **ACE** with direct variance estimation
- Three unique observed statistics
  - V = VA+VC+VE
  - cMZ = VA+VC
  - cDZ = .5VA+VC
- Three unknown parameters
  - V – cMZ = (VA+VC+VE) – (VA+VC) = VE
  - 2(cMZ – cDZ) = 2(VA+VC) – 2(.5VA+VC) = VA
  - VC = –cMZ +2cDZ, or V – VA – VE = VC
- What if negative VC?
  - VD' = –2VC
  - VA' = cMZ +2VC = VA+3VC

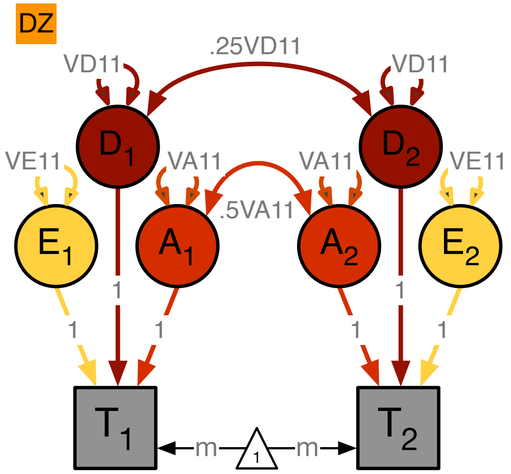
- Get **VC** when fitting **ADE** with direct variance estimation
- Three unique observed statistics
  - V = VA+VD+VE
  - cMZ = VA+VD
  - cDZ = .5VA+.25VD
- Three unknown parameters
  - V – cMZ = (VA+VD+VE) – (VA+VD) = VE
  - –(cMZ – 4cDZ) = –(VA+VD) + (2VA+VD) = VA
  - VD = 2cMZ – 4cDZ, or V – VA – VE = VD
- What if negative VD?
  - VC' = –1/2VD
  - VA' = cMZ +1/2VD = VA+3/2VD

Verhulst et al, 2019

One Trait - Univariate Twin Model Script

How to specify **ADE** model in OpenMx?





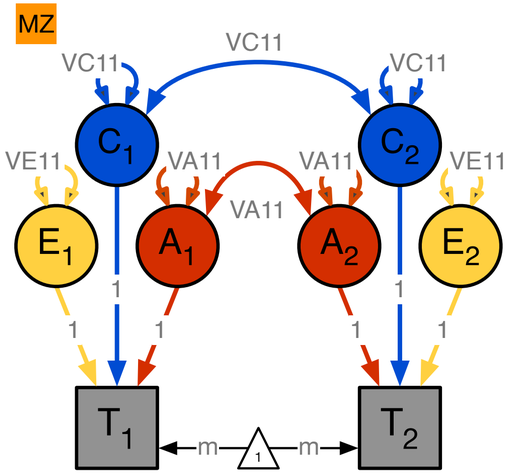
VA *1x1*	VA_{11}	***V*** *1x1*	VA+VD+VE	***expCovMZ*** *2x2*	V | cMZ cMZ | V
VC *1x1*	VD_{11}	***cMZ*** *1x1*	VA+VD	***expCovDZ*** *2x2*	V | cDZ cDZ | V
VE *1x1*	VE_{11}	***cDZ*** *1x1*	.5VA+.25VD	***meanG*** *1x2*	m | m

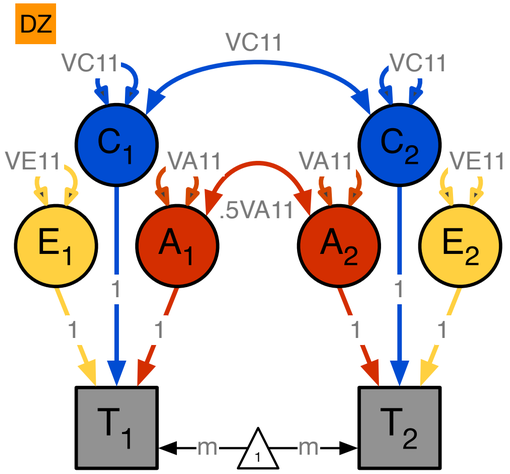
```
meanG     <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=svm, labels=labVars("mean",vars), name="meanG" )
covA      <- mxMatrix( type="Symm", nrow=nv, ncol=nv, free=TRUE, values=svp, label="VA11", name="VA" )
covD      <- mxMatrix( type="Symm", nrow=nv, ncol=nv, free=TRUE, values=svp, label="VD11", name="VD" )
covE      <- mxMatrix( type="Symm", nrow=nv, ncol=nv, free=TRUE, values=sve, label="VE11", name="VE" )
covP      <- mxAlgebra( expression= VA+VD+VE, name="V" )
covMZ     <- mxAlgebra( expression= VA+VD, name="cMZ" )
covDZ     <- mxAlgebra( expression= 0.5%x%VA+ 0.25%x%VD, name="cDZ" )
expCovMZ  <- mxAlgebra( expression= rbind( cbind(V, cMZ), cbind(t(cMZ), V)), name="expCovMZ" )
expCovDZ  <- mxAlgebra( expression= rbind( cbind(V, cDZ), cbind(t(cDZ), V)), name="expCovDZ" )
```

`nv <- 1`; `svm`, `svp`, `sve`: start values for means, VA, VD, VE; `labVars`: functions in `miFunctions.R`

One Trait - Univariate Twin Model Script

How to specify ACE model in OpenMx?





VA *1x1*	VA_{11}	***V*** *1x1*	VA+VC+VE	***expCovMZ*** *2x2*	V | cMZ cMZ | V
VC *1x1*	VC_{11}	***cMZ*** *1x1*	VA+VC	***expCovDZ*** *2x2*	V | cDZ cDZ | V
VE *1x1*	VE_{11}	***cDZ*** *1x1*	.5VA+VC	***meanG*** *1x2*	m | m

```
meanG     <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=svm, labels=labVars("mean",vars), name="meanG")
covA      <- mxMatrix( type="Symm", nrow=nv, ncol=nv, free=TRUE, values=svp, label="VA11", name="VA" )
covC      <- mxMatrix( type="Symm", nrow=nv, ncol=nv, free=TRUE, values=svp, label="VC11", name="VC" )
covE      <- mxMatrix( type="Symm", nrow=nv, ncol=nv, free=TRUE, values=sve, label="VE11", name="VE" )
covP      <- mxAlgebra( expression= VA+VC+VE, name="V" )
covMZ     <- mxAlgebra( expression= VA+VC, name="cMZ" )
covDZ     <- mxAlgebra( expression= 0.5%x%VA+ VC, name="cDZ" )
expCovMZ  <- mxAlgebra( expression= rbind( cbind(V, cMZ), cbind(t(cMZ), V)), name="expCovMZ" )
expCovDZ  <- mxAlgebra( expression= rbind( cbind(V, cDZ), cbind(t(cDZ), V)), name="expCovDZ" )
```

`nv <- 1`; `svm`, `svp`, `sve`: start values for means, VA, VC, VE; `labVars`: functions in `miFunctions.R`

One Trait - VA ⊗ Genetic Relatedness

How to specify **ACE** model in an alternative way?

	T1	T2
T1	VA+VC+VE	α⊗VA+VC
T2	α⊗VA+VC	VA+VC+VE

corr(Yi,Yj) = h^2 R_{jk} + Residual

cMZ *2x2*	VA+VC+VE	VA+VC
	VA+VC	VA+VC+VE

cDZ *2x2*	VA+VC+VE	.5⊗VA+VC
	.5⊗VA+VC	VA+VC+VE

	VA+VC+VE	α⊗VA+VC
	α⊗VA+VC	VA+VC+VE

	VA+VC+VE	α⊗VA+VC
	α⊗VA+VC	VA+VC+VE

VA⊗*rAmz* *2x2*	1	1
	1	1

VA⊗*rAdz* *2x2*	1	.5
	.5	1

VC⊗ *rC* *2x2*	1	1
	1	1

VE⊗ *rE* *2x2*	1	0
	0	1

VA *1x1*	VA_{11}	***V*** *1x1*	VA+VC+VE	***expCovMZ*** *2x2*	V	cMZ
					cMZ	V
VC *1x1*	VC_{11}	***cMZ*** *1x1*	VA+VC	***expCovDZ*** *2x2*	V	cDZ
					cDZ	V
VE *1x1*	VE_{11}	***cDZ*** *1x1*	.5VA+VC	***meanG*** *1x2*	m	m

```
expCovMZ: rbind( cbind(V, cMZ), cbind(t(cMZ), V))
expCovDZ: rbind( cbind(V, cDZ), cbind(t(cDZ), V))
```

```
rAmz: type="Symm", nrow=2, ncol=2, free=FALSE, values=c(1,1,1)
rAdz: type="Symm", nrow=2, ncol=2, free=FALSE, values=c(1,.5,1)
rC: type="Unit", nrow=2, ncol=2, free=FALSE
rE: type="Iden", nrow=2, ncol=2, free=FALSE
expCovMZ: VA%x%rAmz +VC%x%rC +VE%x%rE
expCovDZ: VA%x%rAdz +VC%x%rC +VE%x%rE
```

expCovMZ= VA⊗rAmz + VC⊗rC + VE⊗rE

expCovDZ= VA⊗rAdz + VC⊗rC + VE⊗rE

One Trait - VA ⊗ Genetic Relatedness

How to specify **ACE** model in an alternative way?

	T1	T2
T1	VA+VC+VE	α | π̂⊗VA+VC
T2	α | π̂⊗VA+VC	VA+VC+VE

corr(Yi,Yj) = h^2 R_{jk} + Residual

cMZ 2x2		
	VA+VC+VE	VA+VC
	VA+VC	VA+VC+VE

cDZ 2x2		
	VA+VC+VE	.5⊗VA+VC
	.5⊗VA+VC	VA+VC+VE

VA⊗*rAmz* 2x2		
	1	1
	1	1

VA⊗*rAdz* 2x2		
	1	.5
	.5	1

	VA+VC+VE	α⊗VA+VC
	α⊗VA+VC	VA+VC+VE

VC⊗ *rC* 2x2		
	1	1
	1	1

	VA+VC+VE	α⊗VA+VC
	α⊗VA+VC	VA+VC+VE

VE⊗ *rE* 2x2		
	1	0
	0	1

	VA+VC+VE	π̂⊗VA+VC
	π̂⊗VA+VC	VA+VC+VE

VA⊗ *rA* 2x2		
	1	1
	π̂	1

~~ cov= VA⊗rA + VC⊗rC + VE⊗rE

VA 1x1		*V* 1x1		*expCovMZ* 2x2		
	VA_{11}		VA+VC+VE		V	cMZ
					cMZ	V
VC 1x1		*cMZ* 1x1		*expCovDZ* 2x2		
	VC_{11}		VA+VC		V	cDZ
					cDZ	V
VE 1x1		*cDZ* 1x1		*meanG* 1x2		
	VE_{11}		.5VA+VC		m	m

```
expCovMZ: rbind( cbind(V, cMZ), cbind(t(cMZ), V))
expCovDZ: rbind( cbind(V, cDZ), cbind(t(cDZ), V))
```

```
rAmz: type="Symm", nrow=2, ncol=2, free=FALSE, values=c(1,1,1)
rAdz: type="Symm", nrow=2, ncol=2, free=FALSE, values=c(1,.5,1)
rC: type="Unit", nrow=2, ncol=2, free=FALSE
rE: type="Iden", nrow=2, ncol=2, free=FALSE
expCovMZ: VA%x%rAmz +VC%x%rC +VE%x%rE
expCovDZ: VA%x%rAdz +VC%x%rC +VE%x%rE
```

```
rA: type="Stand", nrow=2, ncol=2, free=FALSE, labels="data.rel"
expCovTW: VA%x%rA +VC%x%rC +VE%x%rE, name="expCovTW" )
```

rel = 1 for MZs
rel = .5 for DZs

```
rA: type="Stand", nrow=2, ncol=2, free=FALSE, labels="data.pihat"
expCovTW: VA%x%rA +VC%x%rC +VE%x%rE, name="expCovTW" )
```

One Trait - VA ⊗ Genetic Relatedness

How to specify **ACE** model in an alternative way?

- **corr(Yi,Yj) = h^2 R_{jk} + Residual**

  ~~ cov= VA⊗rA + VC⊗rC + VE⊗rE

VA+VC+VE	α⊗VA+VC
α⊗VA+VC	VA+VC+VE

VA+VC+VE	α⊗VA+VC
α⊗VA+VC	VA+VC+VE

VA+VC+VE	π̂⊗VA+VC
π̂⊗VA+VC	VA+VC+VE

VC⊗ *rC* *2x2*

1	1
1	1

VE⊗ *rE* *2x2*

1	0
0	1

VA⊗ *rA* *2x2*

1	1
π̂	1

```
rA: type="Stand", nrow=2, ncol=2, free=FALSE, labels="data.pihat"
expCovTW: VA%x%rA +VC%x%rC +VE%x%rE, name="expCovTW" )
```

Interpreting Goodness of Fit - Parameter Estimates

What is the best fitting/ most parsimonious genetic model?

		os	ep	-2ll	df	AIC	Δ-2ll	Δdf	p
	saturated	1777	10	4055.93	1767	521.93			
m1	**ADE model**	1777	4	4063.45	1773	517.45	7.51	6	0.27
m2	**AE model**	1777	3	4067.66	1774	519.66	4.21	1	0.04
m3	**E model**	1777	2	4591.79	1775	1041.79	528.34	2	0.00

- ADE model fits data not significantly worse than saturated model, is most parsimonious

		variance components			standardized components		
		VA	**VD**	**VE**	**SA**	**SD**	**SE**
m1	**ADE**	0.32 0.02-0.61	0.29 0.01-0.60	0.17 0.15-0.19	0.41	0.37	0.22
m2	**AE**	0.62 0.56-0.68	-	0.17 0.17-0.19	0.78	-	0.22
m3	**E**	-	-	0.78 0.73-0.83	-	-	1.00

- Broad heritability [**VA**+**VD** / **V**] estimated at 78%, narrow heritability [**VA** / **V**] at 41%

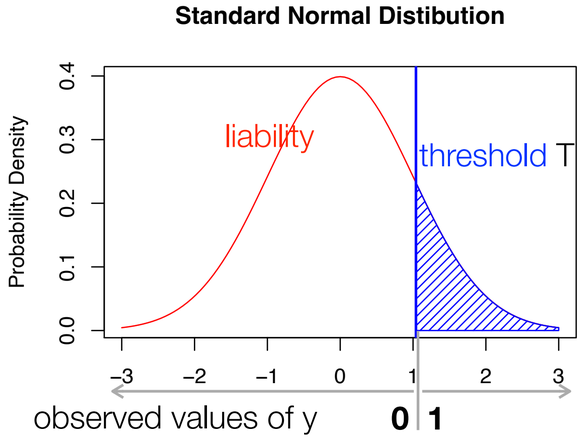
Continuous vs Binary/Ordinal Data

How to fit genetic models to categorical data?

- Measuring instrument discriminates between two or a few ordered categories : e.g. absence or presence of disease; data are counts: number of individuals in category
- Observed binary/ordinal variable assumed indicative of underlying, latent (unobserved) continuous normally distributed variable: liability: Concept of Liability Threshold Model
- Cell counts of two categorical traits, represented in contingency table, can be translated into proportions. Observed proportions relate to those expected under Bivariate Normal Distribution with certain correlation between latent variables, each cut at certain thresholds
- **A**, **C** & **E** components estimated from MZ & DZ tetrachoric/polychoric correlations

distribution of y*

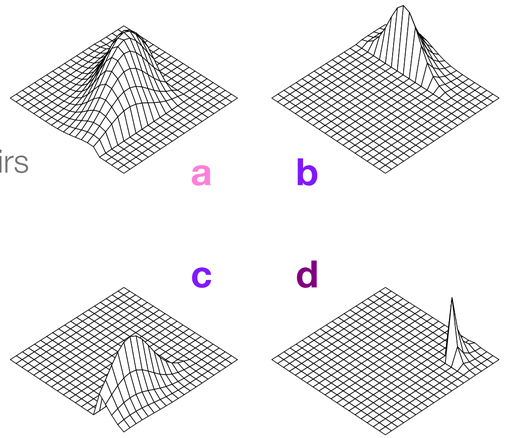
$$y = \begin{cases} 1 \text{ if } y^* > T \\ 0 \text{ if } y^* < T \end{cases}$$



T1\T2	0	1
0	00 [a]	01 [b]
1	10 [c]	11 [d]

cell **a**: number of concordant unaffected pairs
cell **d**: number of concordant affected pairs
cell **b**/**c**: number of discordant pairs

r, t_1, t_2 >> **N** in 4 cells
N in 4 cells >> ***r***, t_1, t_2

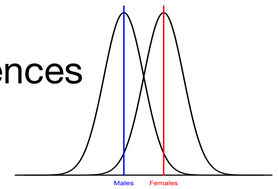


Including Covariates - Fixed Effects

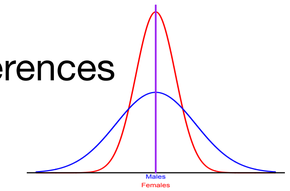
How do we take covariates into account using definition variables?

- Regression coefficients **β** capture differences in mean levels of trait values by sex
- Variance differences **σ^2** capture differences between variation around mean across sexes
- If mean differences exist, but ignored, they can induce variance differences
- Important to include covariates/definition variables for sex* in sex limitation models
- Including mean effects is analogous to including constituent terms in interaction model

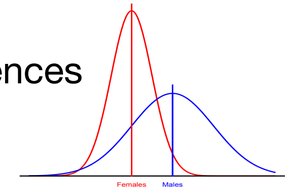
mean differences



variance differences



mean & variance differences



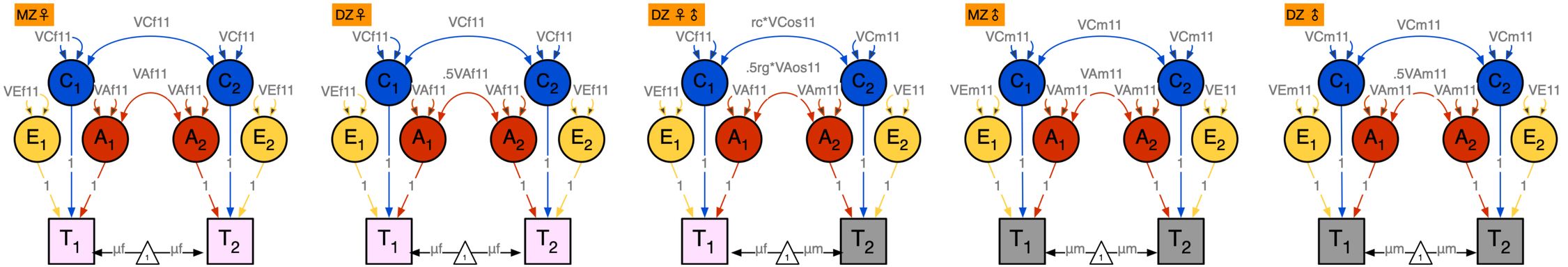

```
betaS     <- mxMatrix( type="Full", nrow=1, ncol=1, free=TRUE, values=svBe, labels="betaS", name="bS" )
betaA     <- mxMatrix( type="Full", nrow=1, ncol=1, free=TRUE, values=svBe, labels="betaA", name="bA" )
defSex    <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=FALSE, labels=c("data.sex1","data.sex2"), name="Sex" )
defAge    <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=FALSE, labels=c("data.age1","data.age2"), name="Age" )
intercept <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=svMu, labels="interC", name="intercept" )
expMean   <- mxAlgebra( expression= intercept +bS*Sex +bA*Age, name="expMean" )
```

	Twin1	Twin2
expMean	**intercept + bS*Sex + bA*Age**	**intercept + bS*Sex + bA*Age**

Heterogeneity - Sex Differences in Random Effects

Are contributions of VA, VC, VE factors equal for different groups: sex, SES, exposure, etc.?

- Independent variables, i.e. sex, can influence traits to different extents (magnitude) in different groups: quantitative sex*differences?
  - e.g. Is contribution of genetic factors greater/smaller in males than in females?
- Different kinds (nature) of independent variables can influence trait in different groups: qualitative sex differences? requires pairs discordant for group membership
  - e.g. Are there different genetic factors influencing trait in males and females?



	Tf1	Tf2
Tf1	**Vf**	
Tf2	**αVAf+VCf**	**Vf**

	Female	Male
Tf1	**Vf= VAf+VCf+VEf**	
Tm2	**0.5rg*VAos+rc*VCos**	**Vm= VAm+VCm+VEm**

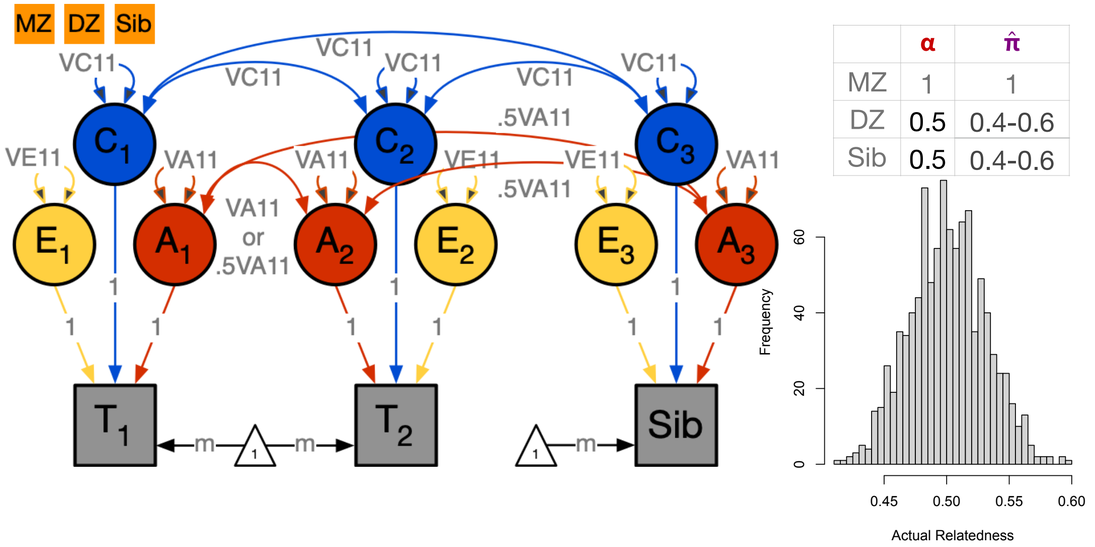
	Tm1	Tm2
Tm1	**Vm**	
Tm2	**αVAm+VCm**	**Vm**

MZs: α=1
DZs: α=0.5

Including Siblings / Parents / Extended Relatives

What is gained by adding other relatives?

expected or actual relatedness for DZ twins or full siblings



	α	π̂
MZ	1	1
DZ	0.5	0.4-0.6
Sib	0.5	0.4-0.6

VE⊗

rE	T1	T2	S
T1	1	0	0
T2	0	1	0
Sib	0	0	1

VC⊗

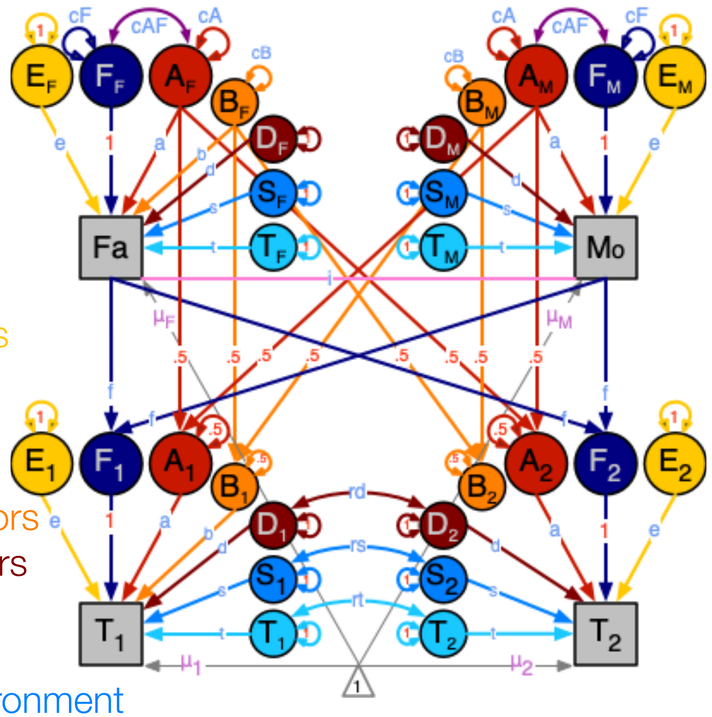
rC	T1	T2	S
T1	1	1	1
T2	1	1	1
Sib	1	1	1

VA⊗

rA	T1	T2	S
T1	1	α|π̂	α|π̂
T2	α|π̂	1	α|π̂
Sib	α|π̂	α|π̂	1

Eaves et al. 1978 and beyond

twins & parents model



E: Unique Environmental Factors (Measurement Error)

Genetic Factors

- **A**: Additive Genetic Factors
- **B**: Sex-specific Genetic Factors
- **D**: Dominance Genetic Factors

Shared Environmental Factors

- **F**: Cultural Transmission
- **S**: Non-Parental Shared Environment
- **T**: Special Twin Environment

I: Phenotypic Assortment

rGE: Genotype-Environment Covariance

Fa: father Mo: mother

Modeling genetic and environmental sources of variation in multiple traits

Part 4

credit to
Marleen de Moor, Meike Bartels, Elizabeth Prom-Wormley, Lucia Colondro Conde, Tim Bates, Michael Hunter, Michael Neale, et. al.

Outline Part 4

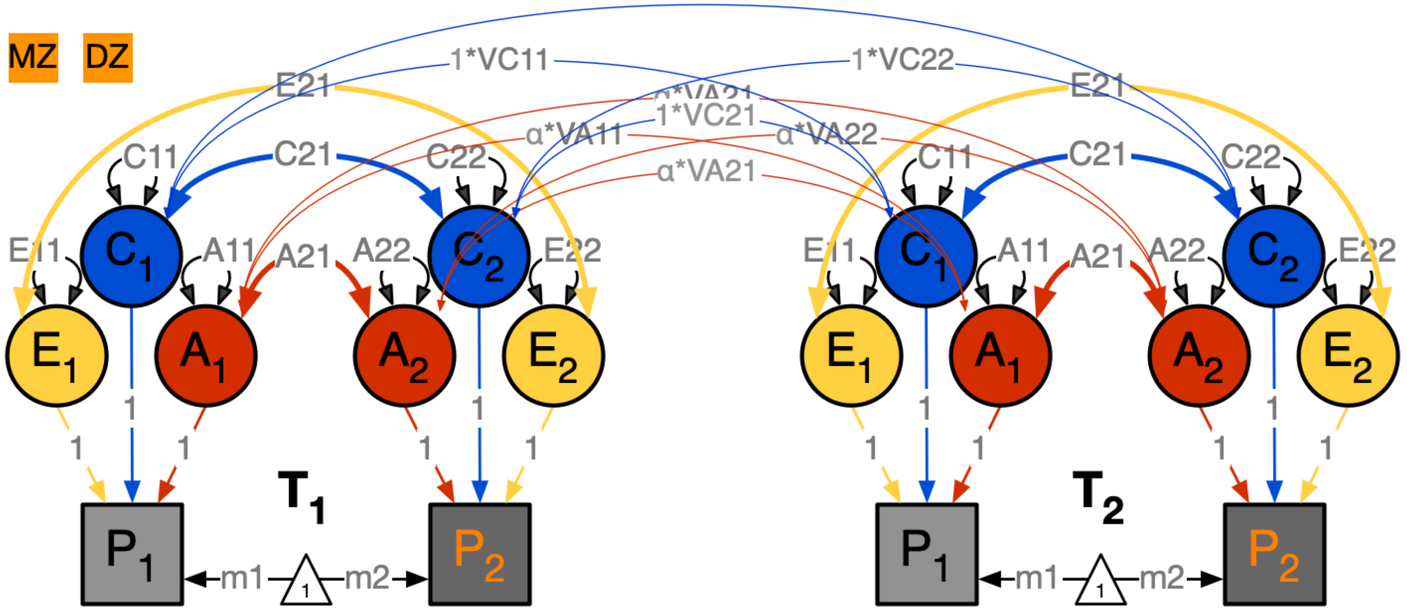
Estimate genetic and environmental sources of variation in multiple traits

- Sources of Covariation
- Patterns of Cross-Twin Cross-Trait Correlations
- Two Traits - Bivariate Twin Model Script
- Proportions of Covariation or Correlations
- Direction of Causation
- Multivariate Twin Models
- Other Genetically Informative Models

Sources of Covariation - Bi(Multi)variate Analysis

What are contributions of **A**, **C** & **E** factors to covariance between two (or more) traits?

- Two or more traits can be correlated or co-vary because they share genes or shared / unique environmental influences in common or cause one another
- With twin data on multiple traits it is possible to partition covariation into its genetic and environmental components or test direction of causation



Bivariate

- wtctMZ - ctctMZ: **E**
- ctctMZ > ctctDZ: **A** (+**D**)
- ctctMZ = 2*ctctDZ: only **A**
- ctctMZ = ctctDZ: only **C**
- ctctMZ/2 < ctctDZ: **A** & **C**
- ctctMZ/2 > ctctDZ: **A** & **D**

Patterns of Cross-Twin Cross-Trait Correlations

Which statistics provide information about causes of [co]variation?

- Cross-twin Within-trait [CTWT] covariances reflect cause of variation for each trait **P1 P2**
- Within-twin Cross-trait [WTCT] covariances imply common etiological influences
- Cross-twin Cross-trait [CTCT] covariances imply familial common etiological influences
- **MZ/DZ ratio of CTCT** covariances reflects whether common etiological influences are genetic or common environmental
- **Ratio of MZ CTCT to WTCT** covariance reflects on role of unique environment

	$P1_{T1}$	$P2_{T1}$	$P1_{T2}$	$P2_{T2}$
$P1_{T1}$	Variance $P1_{T1}$	Covariance $P2_{T1}P1_{T1}$	Within-Trait $P1_{T2}P1_{T1}$	Cross-Trait $P2_{T2}P1_{T1}$
$P2_{T1}$	Covariance $P1_{T1}P2_{T1}$	Variance $P2_{T1}$	Cross-Trait $P1_{T2}P2_{T1}$	Within-Trait $P2_{T2}P2_{T1}$
$P1_{T2}$	Within-Trait $P1_{T1}P1_{T2}$	Cross-Trait $P2_{T1}P1_{T2}$	Variance $P1_{T2}$	Covariance $P2_{T2}P1_{T2}$
$P2_{T2}$	Cross-Trait $P1_{T1}P2_{T2}$	Within-Trait $P2_{T1}P2_{T2}$	Covariance $P1_{T2}P2_{T2}$	Variance $P2_{T2}$

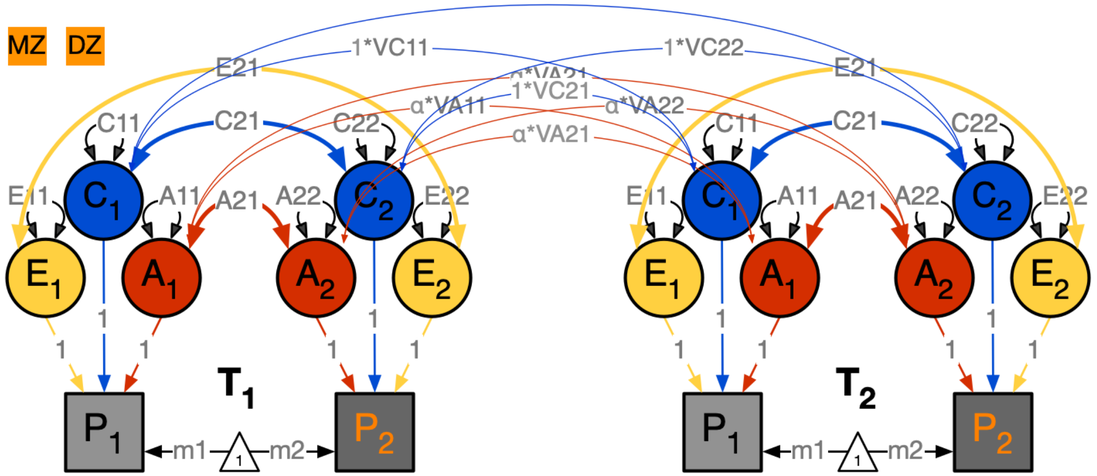
Within-Twin Covariance

	$P1_{T1}$	$P2_{T1}$	$P1_{T2}$	$P2_{T2}$
$P1_{T1}$	VA11 +VC11 +VE11	VA21 +VC21 +VE21	1/0.5 * VA11 +VC11	1/0.5 * VA21 +VC21
$P2_{T1}$	VA21 +VC21 +VE21	VA22 +VC22 +VE22	1/0.5 * VA21 +VC21	1/0.5 *VA22 +VC22
$P1_{T2}$	1/0.5 * VA11 +VC11	1/0.5 * VA21 +VC21	VA11 +VC11 +VE11	VA21 +VC21 +VE21
$P2_{T2}$	1/0.5 * VA21 +VC21	1/0.5 *VA22 +VC22	VA21 +VC21 +VE21	VA22 +VC22 +VE22

Cross-Twin Covariance

Two Traits - Bivariate Twin Model Script

How to specify a Correlated Factors **ACE** model in OpenMx?



VA *2x2*

VA_{11}	VA_{21}
VA_{21}	VA_{22}

V *2x2*

$VA_{11}+VC_{11}+VE_{11}$	$VA_{21}+VC_{21}+VE_{21}$
$VA_{21}+VC_{21}+VE_{21}$	$VA_{22}+VC_{22}+VE_{22}$

expCovMZ *4x4*

V_{11}	V_{21}	cMZ_{11}	cMZ_{21}
V_{21}	V_{22}	cMZ_{21}	cMZ_{22}
cMZ_{11}	cMZ_{21}	V_{11}	V_{21}
cMZ_{21}	cMZ_{22}	V_{21}	V_{22}

VC *2x2*

VC_{11}	VC_{21}
VC_{21}	VC_{22}

cMZ *2x2*

$VA_{11}+VC_{11}$	$VA_{21}+VC_{21}$
$VA_{21}+VC_{21}$	$VA_{22}+VC_{22}$

expCovDZ *4x4*

V_{11}	V_{21}	cDZ_{11}	cDZ_{21}
V_{21}	V_{22}	cDZ_{21}	cDZ_{22}
cDZ_{11}	cDZ_{21}	V_{11}	V_{21}
cDZ_{21}	cDZ_{22}	V_{21}	V_{22}

VE *2x2*

VE_{11}	VE_{21}
VE_{21}	VE_{22}

cDZ *2x2*

$.5VA_{11}+VC_{11}$	$.5VA_{21}+VC_{21}$
$.5VA_{21}+VC_{21}$	$.5VA_{22}+VC_{22}$

meanG *1x4*

m_1	m_2	m_1	m_2

```
meanG     <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=svm, labels=labVars("mean",vars), name="meanG")
covA      <- mxMatrix( type="Symm", nrow=nv, ncol=nv, free=TRUE, values=svp, label="VA11", name="VA" )
covC      <- mxMatrix( type="Symm", nrow=nv, ncol=nv, free=TRUE, values=svp, label="VC11", name="VC" )
covE      <- mxMatrix( type="Symm", nrow=nv, ncol=nv, free=TRUE, values=sve, label="VE11", name="VE" )
covP      <- mxAlgebra( expression= VA+VC+VE, name="V" )
covMZ     <- mxAlgebra( expression= VA+VC, name="cMZ" )
covDZ     <- mxAlgebra( expression= 0.5%x%VA+ VC, name="cDZ" )
expCovMZ  <- mxAlgebra( expression= rbind( cbind(V, cMZ), cbind(t(cMZ), V)), name="expCovMZ" )
expCovDZ  <- mxAlgebra( expression= rbind( cbind(V, cDZ), cbind(t(cDZ), V)), name="expCovDZ" )
```

`nv <- 2`; `svm`, `svp`, `sve`: start values for means, VA, VC, VE; `labVars`: functions in `miFunctions.R`

Proportions of Covariation or Correlations

What proportion of covariance is due to **A**, **C** & **E** or are **A**, **C** & **E**'s correlated?

- Bivariate heritability (**bh²**) or shared/unique environmentality (**bc²**, **be²**) indicate proportion of covariance/correlation between phenotypes accounted for by additive genetic, shared or unique environmental influences
- Genetic (**rg**), shared (**rc**) or unique environmental (**re**) correlations reflect overlap between genetic (or shared or unique environmental) factors underlying two traits
  - Calculated by dividing genetic (or environmental) covariance by square root of product of genetic (or environmental) variances of two variables: rg = VA21/ √VA11*VA22
- If bh² large & rg low, genes play large role in trait overlap, genetic influences trait specific
- Conversely, if bh² small & rg is high, genes play minor role in trait overlap, but the genes involved contribute to both traits

```
matI      <- mxMatrix( type="Iden", nrow=nv, ncol=nv, name="I" )
corA      <- mxAlgebra( expression=solve(sqrt(I*VA)) %&% VA, name ="rA" )
corA      <- mxAlgebra( expression=cov2cor(VA), name ="rA" )
estUS     <- mxAlgebra( expression=cbind(VA,VD,VE,VA/V,VD/V,VE/V), name="US" )
```

$$\begin{bmatrix} 1 & r_g \\ r_g & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{\sigma^2_{A_{11}}}} & 0 \\ 0 & \frac{1}{\sqrt{\sigma^2_{A_{22}}}} \end{bmatrix} * \begin{bmatrix} \sigma^2_{A_{11}} & \sigma^2_{A_{12}} \\ \sigma^2_{A_{21}} & \sigma^2_{A_{22}} \end{bmatrix} * \begin{bmatrix} \frac{1}{\sqrt{\sigma^2_{A_{11}}}} & 0 \\ 0 & \frac{1}{\sqrt{\sigma^2_{A_{22}}}} \end{bmatrix}$$

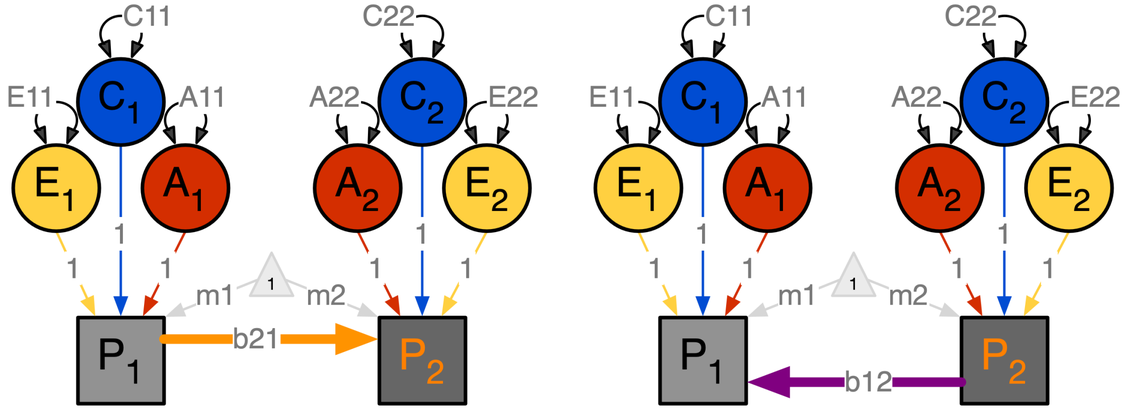
from covariance to correlation matrix

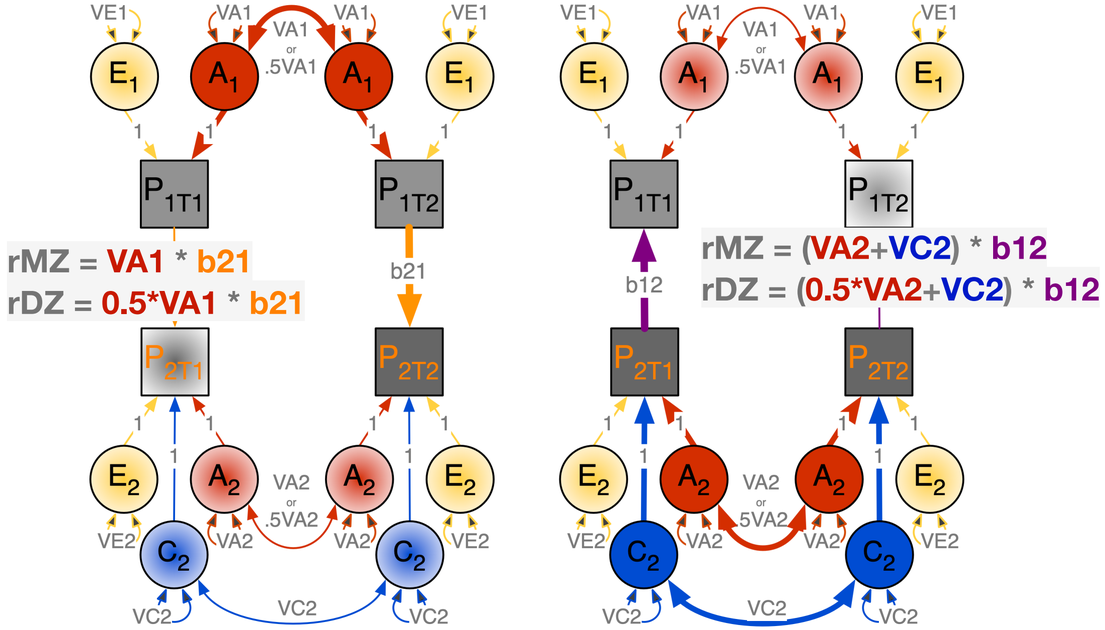
deVries et al, 2021

Direction of Causation (DOC)

Does trait 1 cause trait 2 or vice versa or is there reciprocal causation?

- Besides **rA**, **rC**, **rE** confounding, two traits can have direct causal effects on one another
- for DOC model to work **ACE** estimates of **P1** must be different from **ACE** estimates of **P2**
- If two traits have very similar **ACE**/**ADE** estimates, cross-twin cross-trait covariance is very similar under two directions of causal effects and model has little power and requires very large sample sizes to differentiate between two causal processes





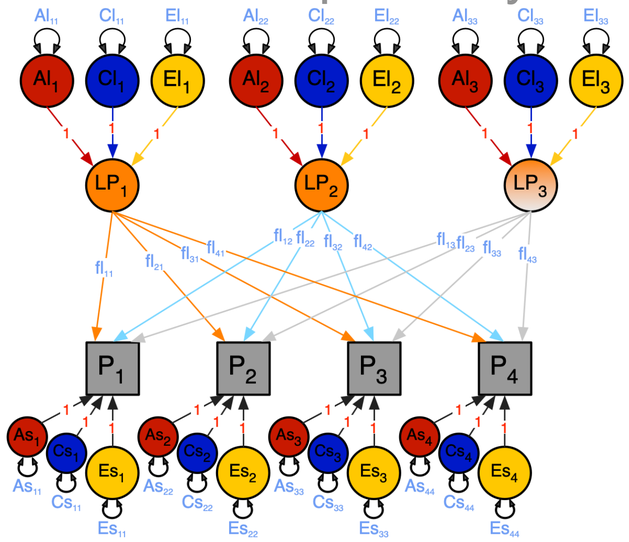
Heath et al, 1993

Multivariate Twin Models

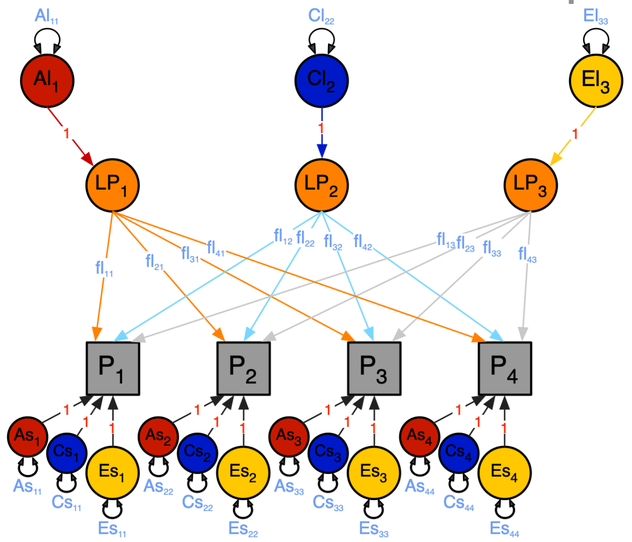
How many common **A**, **C** & **E** factors are consistent with the data?

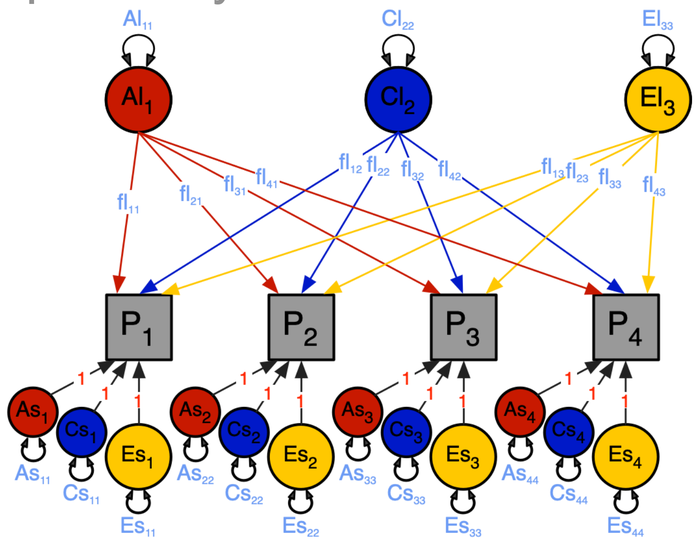
- Common Pathway Model: Same covariance structure for **A**, **C** and **E** [Psychometric]
- Independent Pathway Model: Different covariance structure for **A**, **C** and **E** [Biometric]
- Identification: be careful when adding common factors: total parameters per source of variance can not exceed (nv*(nv+1))/2

common pathway



independent pathway





Other Genetically Informative Models

What kind of questions can be addressed with genetically informative models?

- Longitudinal Data
  - Cholesky Decomposition Model: a-theoretical, potentially useful for longitudinal data
  - Simplex Model, Latent Grow Curve Model, Cross-lagged Panel Model
- Data on Stem and Probes
  - Conditional Causal Common Pathway Model
- Measured Environmental Data
  - Genotype x Environment (GxE) Moderation Models
- Data on Other Relatives: Siblings, Parents, Children, Aunts & Uncles, Nieces & Nephews
  - Extended Pedigree & Children of Twins Models
- Measured Genetic Data (SNPs, Polygenic Scores)
  - Mendelian Randomization (MR) - Direction of Causation Models
  - Trio-GCTA, Transmitted & Non-Transmitted Alleles, Intergenerational MR
- GWAS Summary Statistics
  - GenomicSEM, etc.

1987-2026

40th workshop

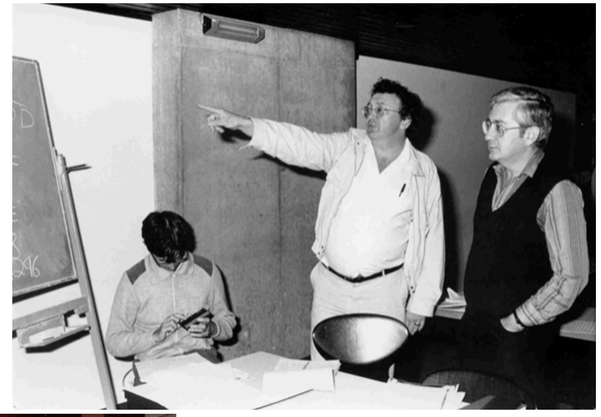
from left to right: Robert Derom†, Walter Nance†, Michael Neale, me, Joanne Meyer, Robert Vlietinck, Peter Molenaar, Andrew Heath, Karl Joreskog, Catherine Derom, Dorret Boomsma, Nick Martin, John Hewitt, Lindon Eaves†, David Fulker† [workshop founders]

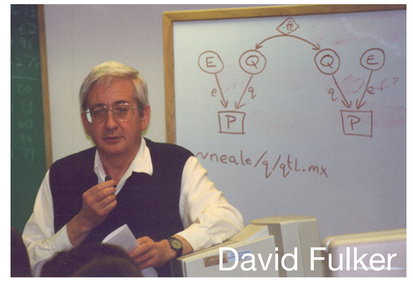
Thank you

Thank you

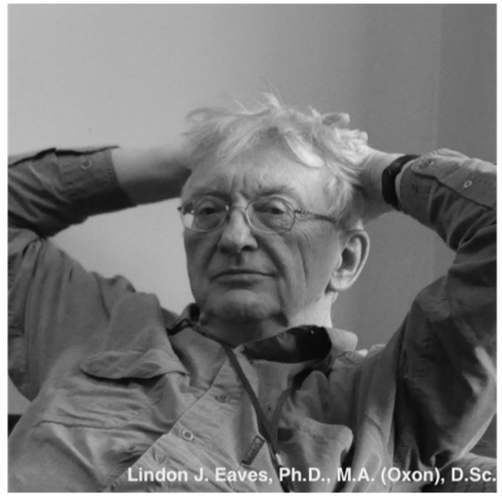
Thank you

1989 Leuven workshop



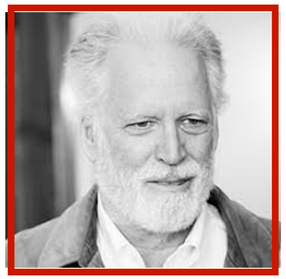


2005 Boulder workshop



Thank you

OpenMx Development Team & International Statistical Genetics Team



Steve Boker

Mike Neale

Sarah Medland

Abdel Abdellaoui

Matt Keller

Mike Hunter

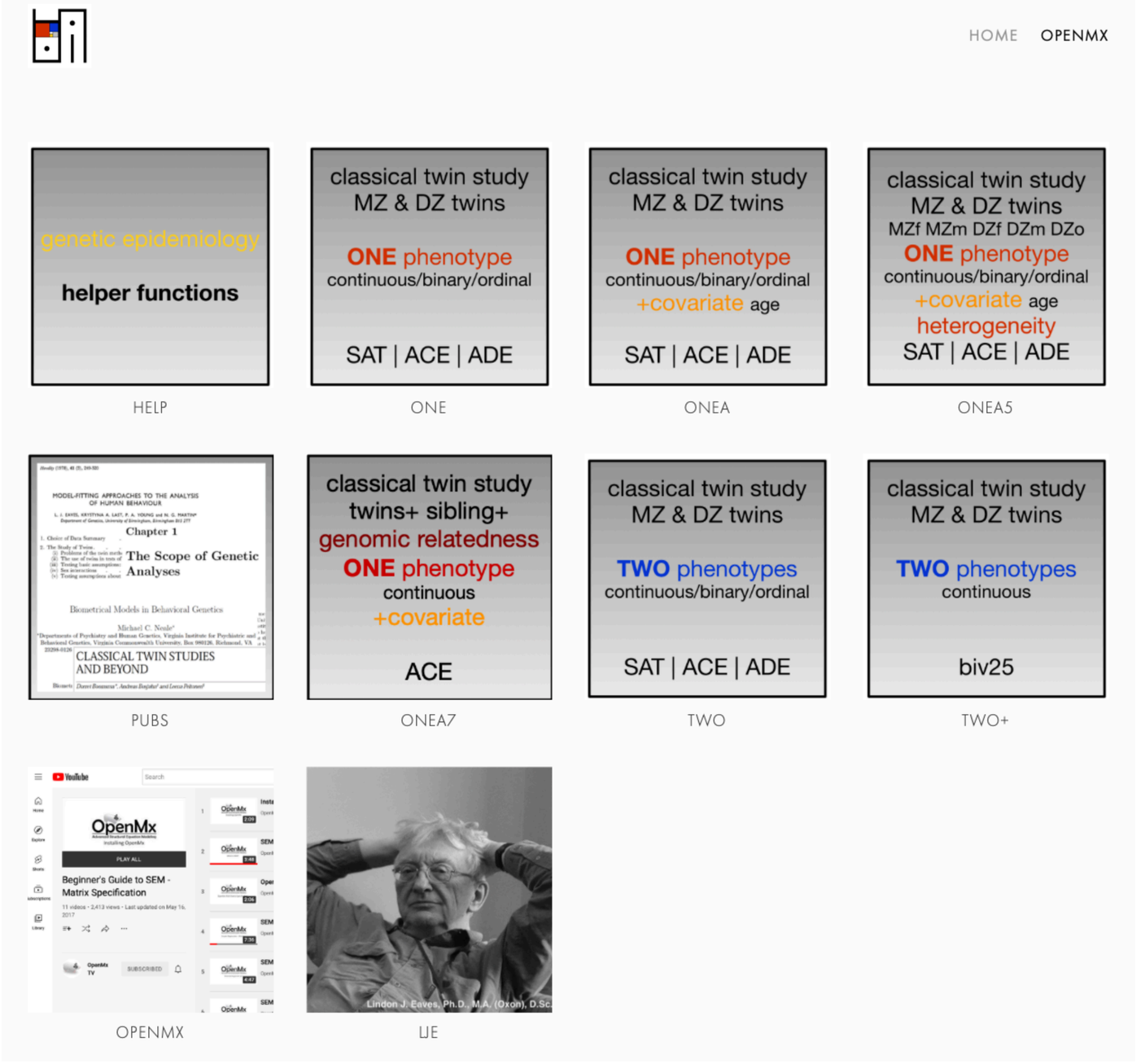
Tim Brick

Rob Kirkpatrick

John Hewitt

Jeff Lessem

hermine-maes.squarespace.com



hermine.maes@vcuhealth.org