Where have we been and we are going?

Lucía Colodro Conde,
Elizabeth Prom-Wormley, and Sarah Medland

Using SEM to model genetic and environmental effects on phenotypes

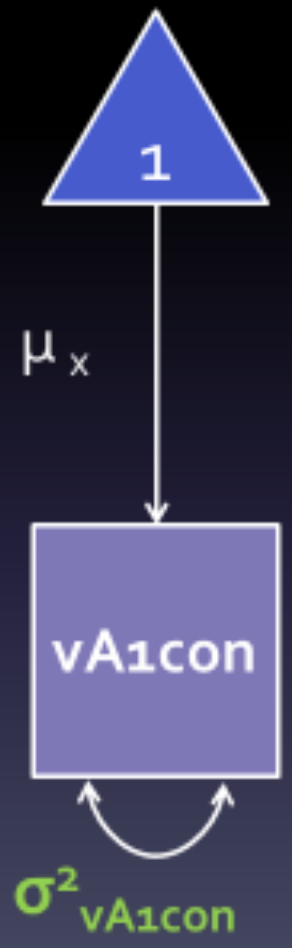
https://www.colorado.edu/ibg/international-workshop/2020-international-statistical-genetics-workshop/workshop-2020-preliminary

Revisiting SEM

In summary, we have been doing processes of:

- Specification
- Identification
- Estimation
- Evaluation

Mean and Variance as a Path Diagram



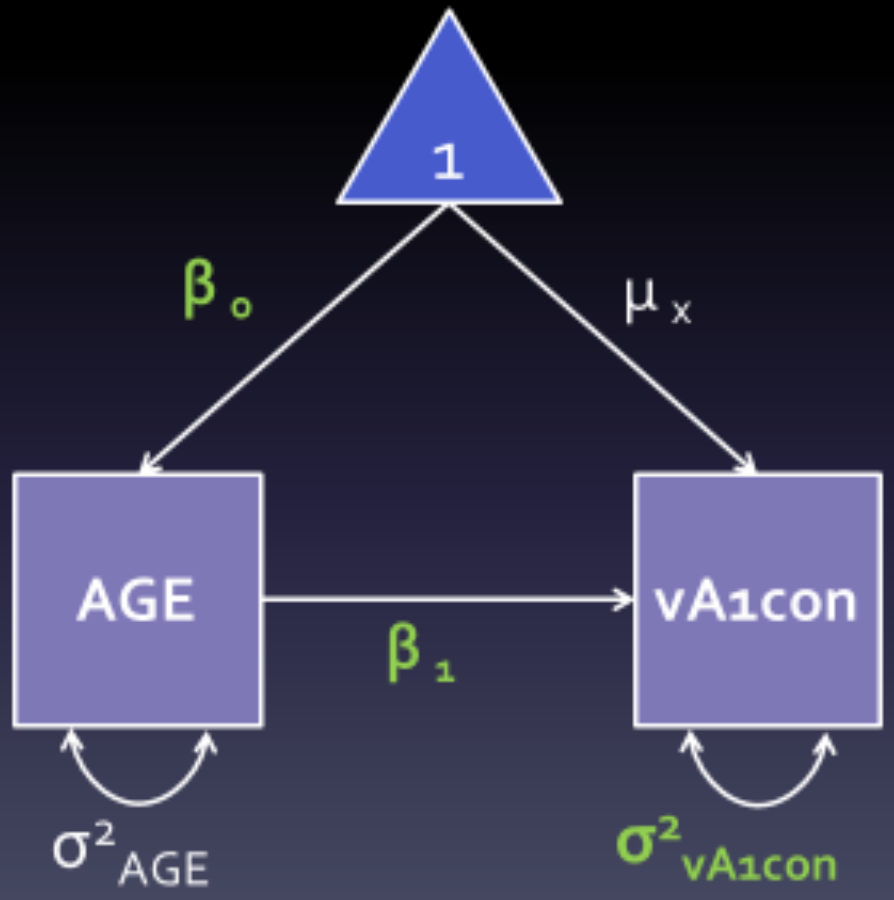
Triangle: a constant variable, usually a vector of ones (here, we use it to reflect a deviation from the mean (none for this picture)

Single-headed arrows: linear relationship between two variables. Starts from an independent variable and ends on a dependent variable

Squares or rectangular boxes: observed or manifest variables (MEASURED)

Double-headed arrows: variance of a variable or covariance between two variables

Linear Regression as a Path Diagram



Squares or rectangular boxes: observed or manifest variables

Single-headed arrows: linear relationship between two variables. Starts from an independent variable and ends on a dependent variable

Double-headed arrows: variance of a variable or covariance between two variables

Triangle: a constant variable, usually a vector of ones

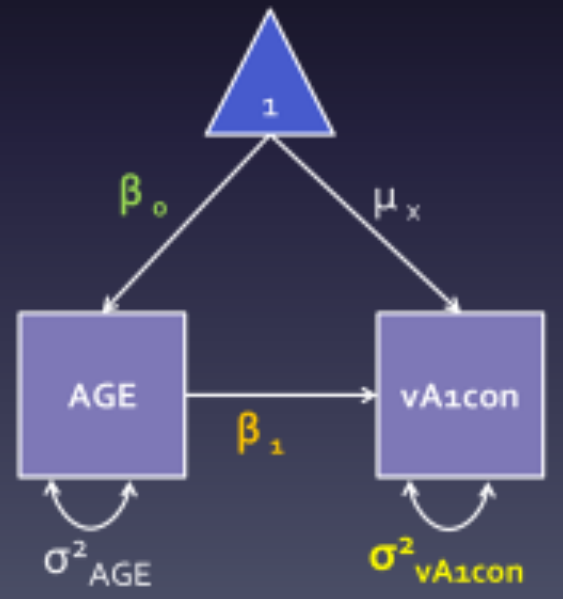
Circles or ovals: errors, factors, latent variables

Regression Across All Twin 1 Members of a Twin Pair

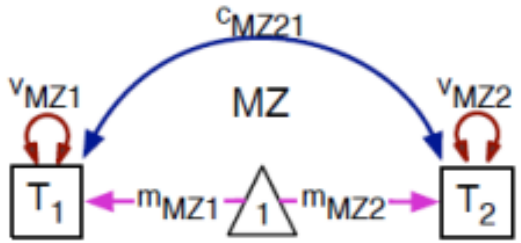
```
require (OpenMx)
depVar <- 'vA1con_1'

# Variance/Covariance matrix
Variance    <- mxMatrix( type="Full", nrow=1, ncol=1, free=TRUE,
                values=10, labels='resid', name="residualVar" )

# Regression betas
b0          <-mxMatrix(type="Full", nrow=1, ncol=1, free=T, values=30,
              labels="beta0", name="Intercept" )
b1          <-mxMatrix(type="Full", nrow=1, ncol=1, free=T, values=0,
              labels="beta1", name="bAge" )

# Independent variable
x           <-mxMatrix(type="Full", nrow=1, ncol=1, free=F,
              labels="data.AGE_1", name="Age" )
```

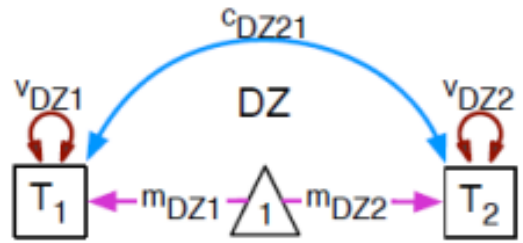
$$y = \beta X + \varepsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \end{bmatrix} = \begin{bmatrix} \beta 0 \\ \beta 1 \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_i \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \end{bmatrix}$$



SAT Deconstructed: *Covariance Matrices & Means*






```
meanMZ    <- mxMatrix( type="Full", nrow=1, ncol=ntv,
 free=TRUE, values=svMe, labels=c("mMZ1","mMZ2"),name="meanMZ" )
meanDZ    <- mxMatrix( type="Full", nrow=1, ncol=ntv,
 free=TRUE, values=svMe, labels=c("mDZ1","mDZ2"),name="meanDZ" )
```

mMZ1	mMZ2

meanMZ *1x2*

mDZ1	mDZ2

meanDZ *1x2*

```
covMZ     <- mxMatrix( type="Symm", nrow=ntv, ncol=ntv,
free=TRUE, values=svVas, lbound=lbVas,
labels=c("vMZ1","cMZ21","vMZ2"), name="covMZ" )
covDZ     <- mxMatrix( type="Symm", nrow=ntv, ncol=ntv,
free=TRUE, values=svVas, lbound=lbVas,
labels=c("vDZ1","cDZ21","vDZ2"), name="covDZ" )
```

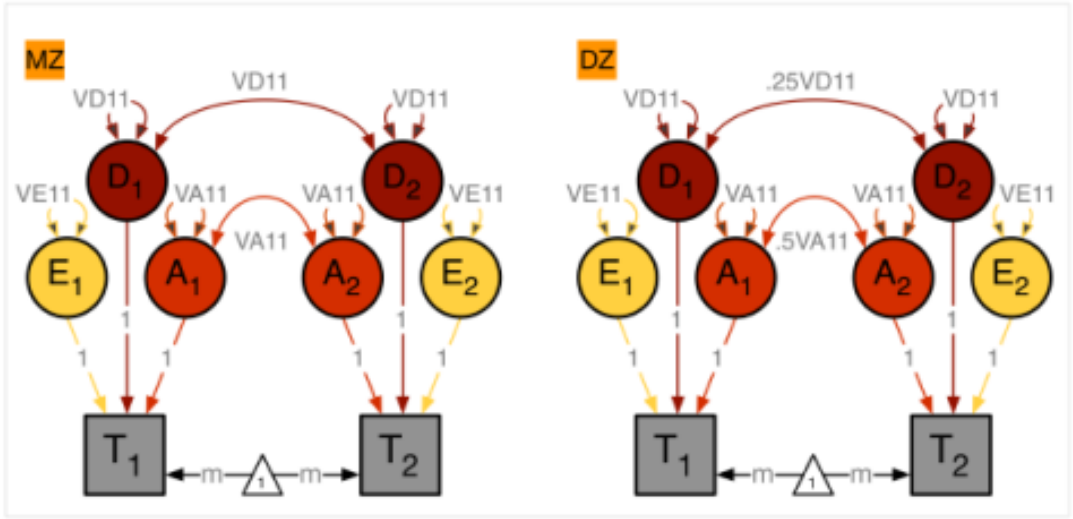
vMZ1	cMZ21
cMZ21	vMZ2

covMZ *2x2*

vDZ1	cDZ21
cDZ21	vDZ2

covDZ *2x2*

ACE Deconstructed: *Covariance Matrices & Means*




```
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" )



meanG     <- mxMatrix( type="Full", nrow=1, ncol=ntv,
 free=TRUE, values=svMe, labels="x1", name="meanG" )
```

V	cMZ
cMZ	V

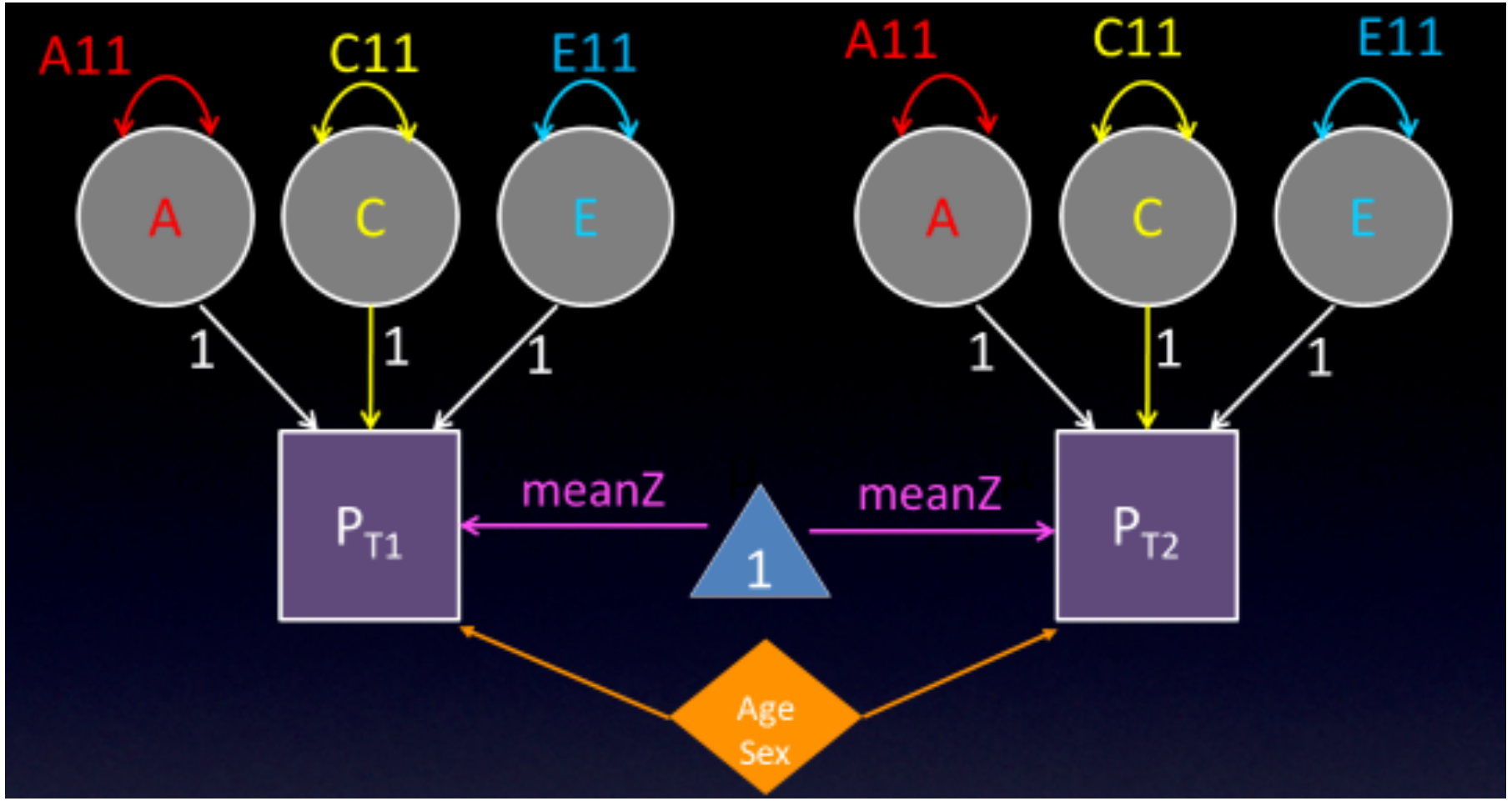
expCovMZ *2x2*

V	cDZ
cDZ	V

expCovDZ *2x2*

x_1	x_1

meanG *1x2*


A11
C11
E11
A11
C11
E11
A
C
E
A
C
E
1
1
1
1
1
1
meanZ
meanZ
P_{T1}
P_{T2}
1
Age
Sex

Conventions of path tracing

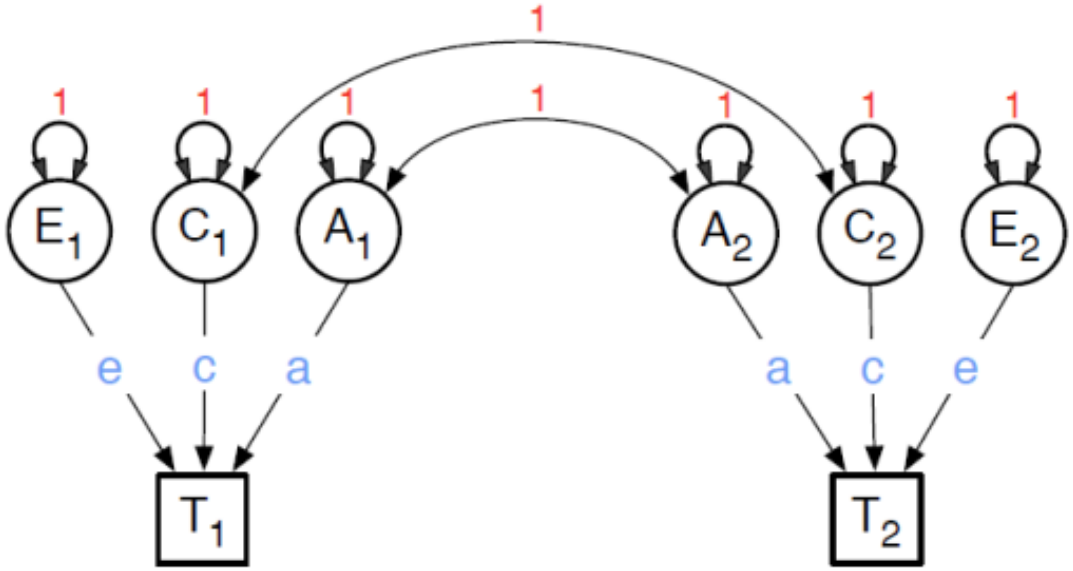
Reference – Path Tracing Rules for SEM

- 1 Find All Distinct Chains between Variables:
  - A) Go backwards along zero or more single-headed arrows
  - B) Change direction at one and only one Double-headed arrow
  - C) Trace forwards along zero or more Single-headed arrows
- 2 Multiply path coefficients in a chain
- 3 Sum the results of step 2.

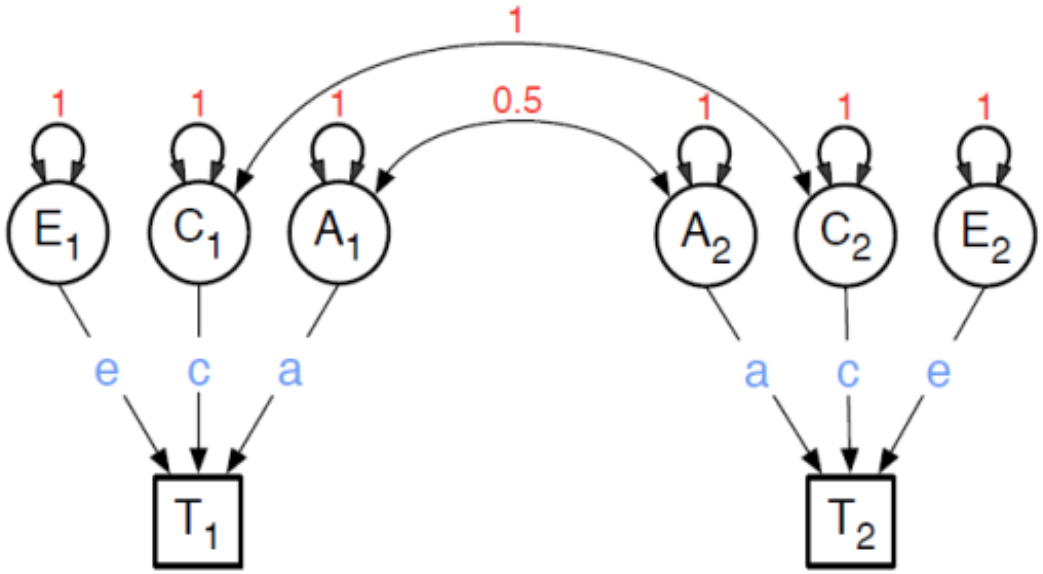
- Note- For covariance of a variable with itself (Variance), chains are distinct if they have different paths or a different order

Estimation of path coefficients

MZ

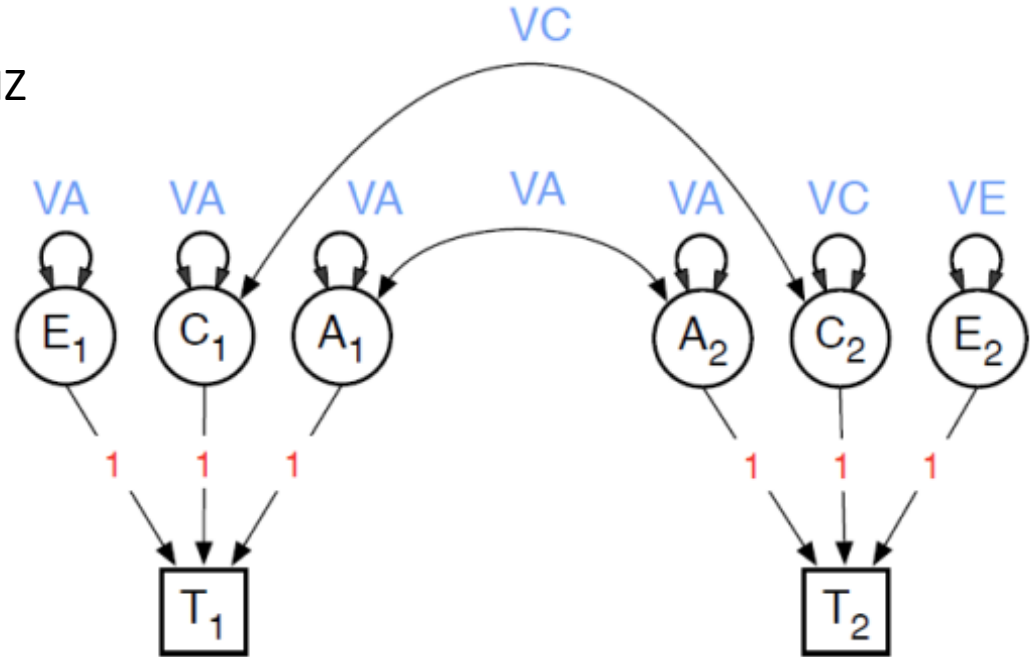


DZ

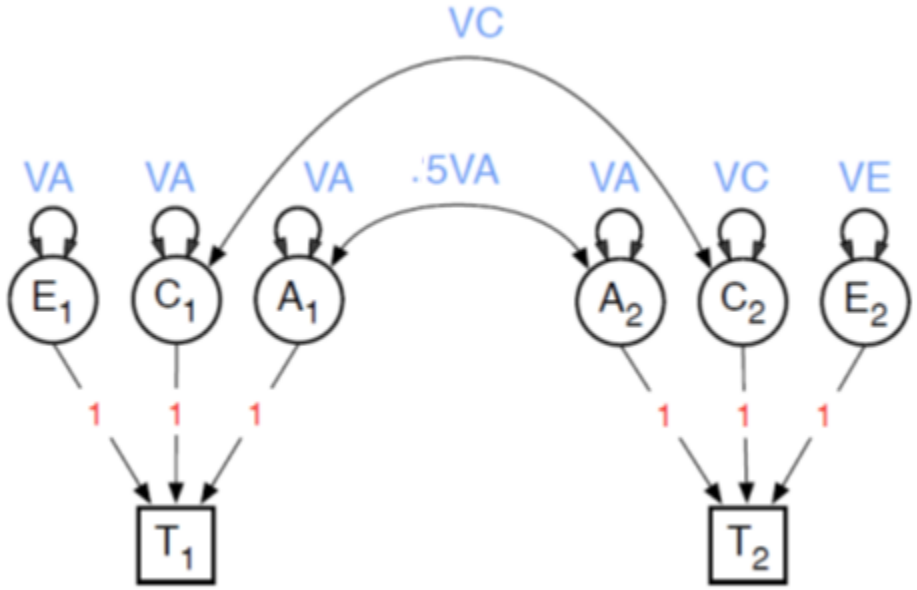


Estimation of variance components

MZ

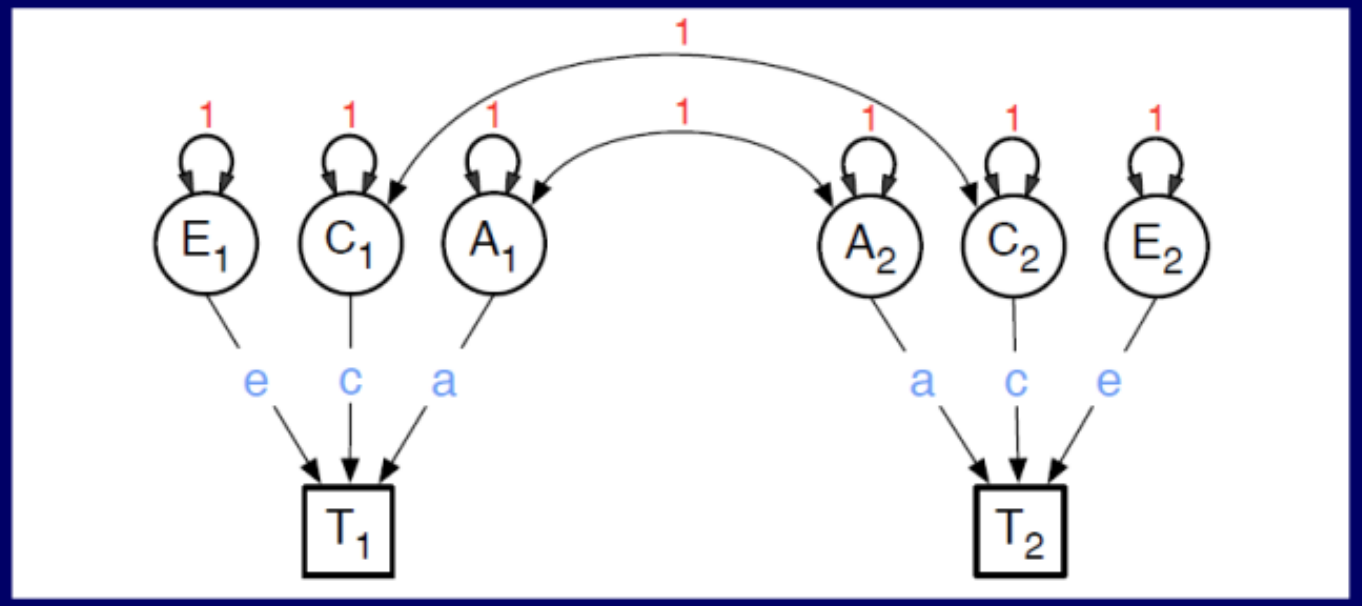


DZ



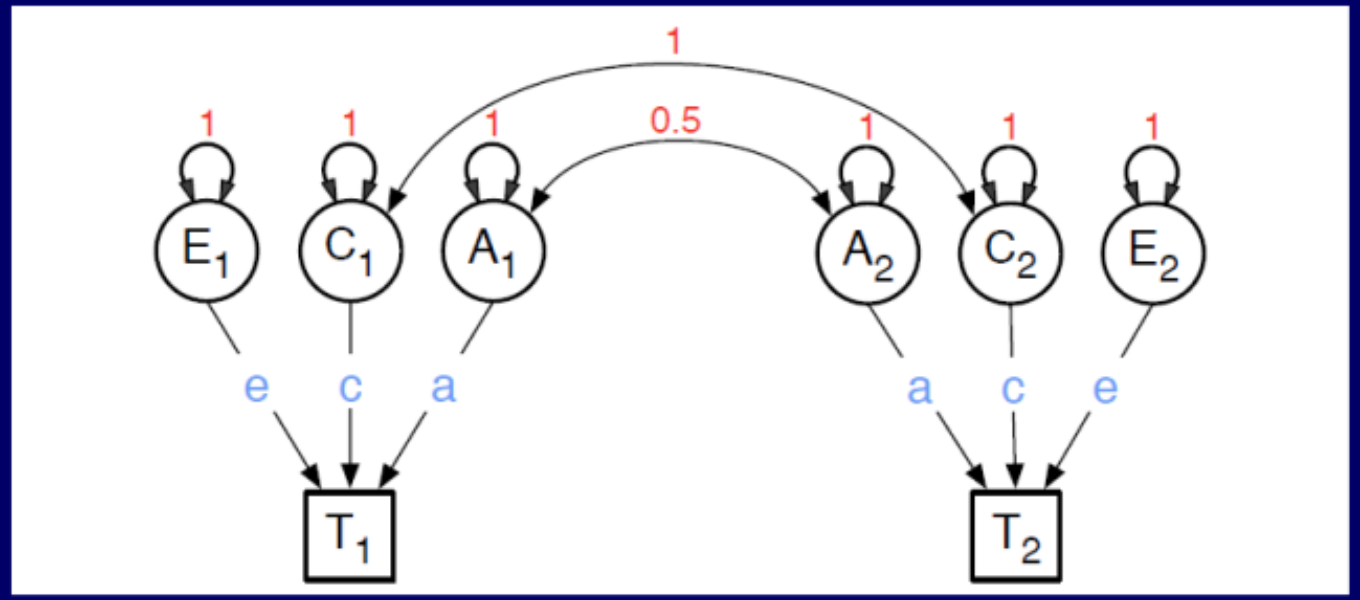
Estimation of path coefficients

Path Model for an MZ Pair



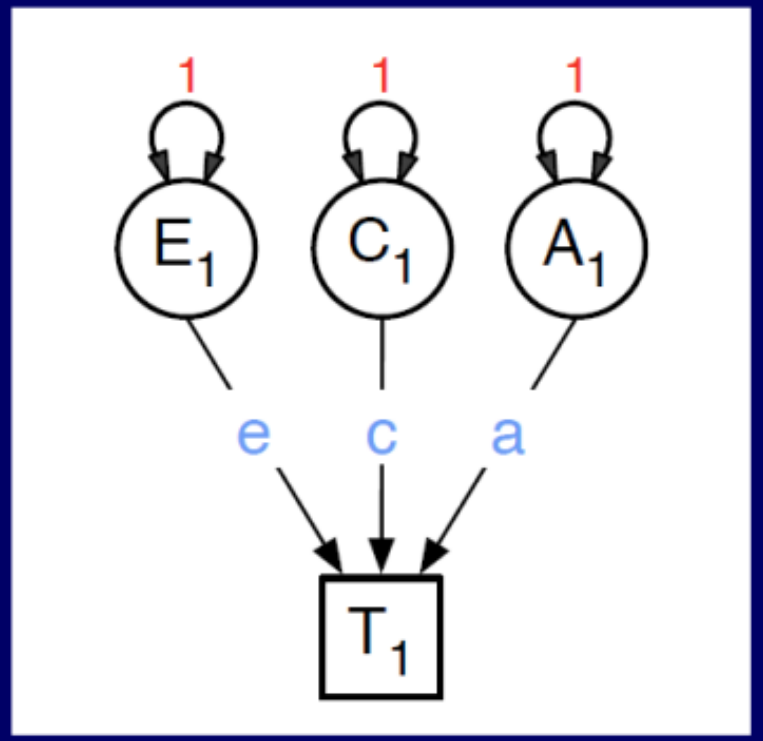
Latent variables A1 C1 and E1 have variance 1, and cause phenotype T1 via path coefficients a, c and e.
Same model for T2. Cov(A_1,A_2)=1

Path Model for a DZ Pair



Latent variables A1 C1 and E1 have variance 1, and cause phenotype T1 via regression paths a, c and e.
Same model for T2. Cov(A1,A2) = .5

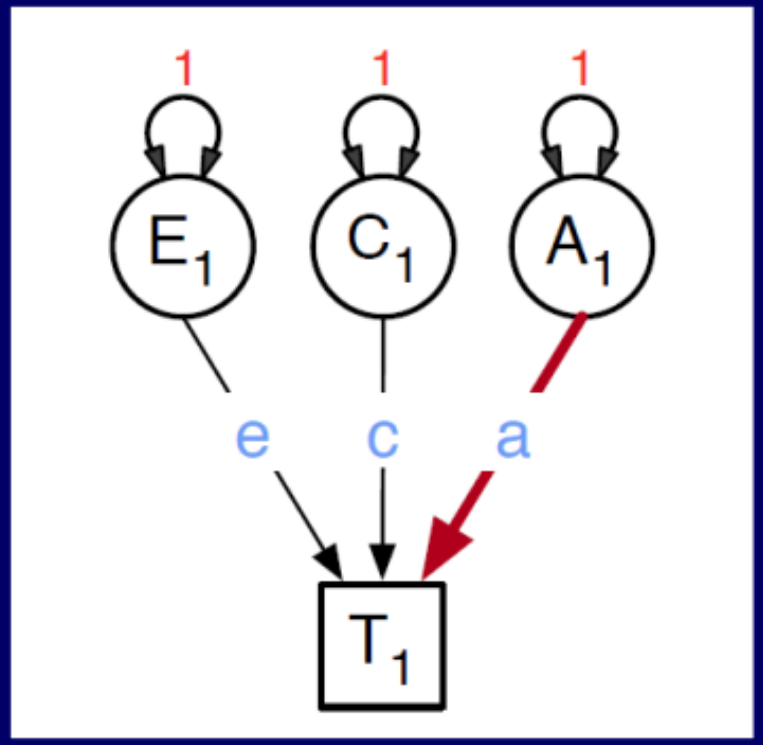
Variance of Twin 1 AND Twin 2 (for MZ and DZ pairs)



What Chains?

Total Variance = $a^2 + c^2 + e^2$

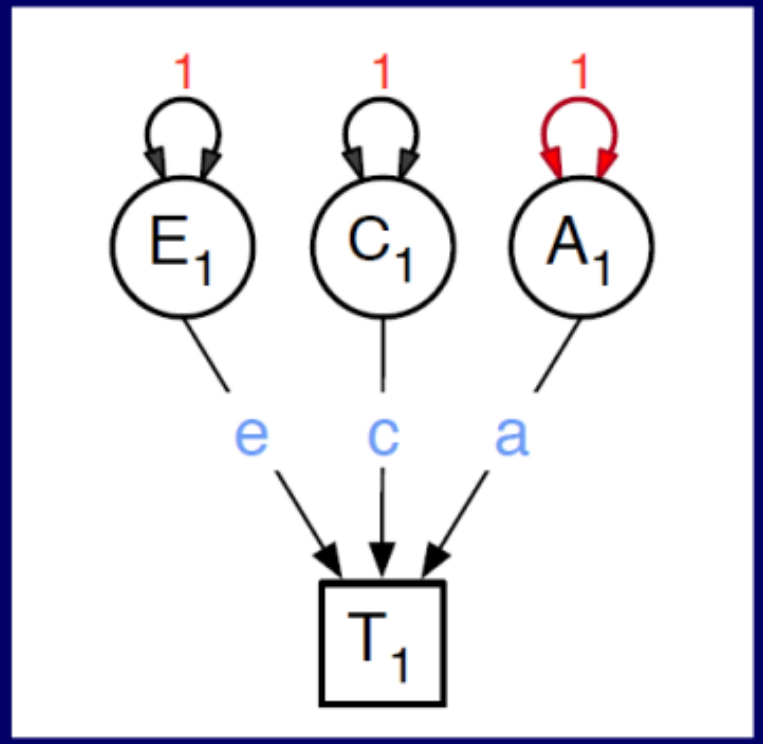
Variance of Twin 1 AND Twin 2 (for MZ and DZ pairs)



a^* =

Total Variance = $a^2 + c^2 + e^2$

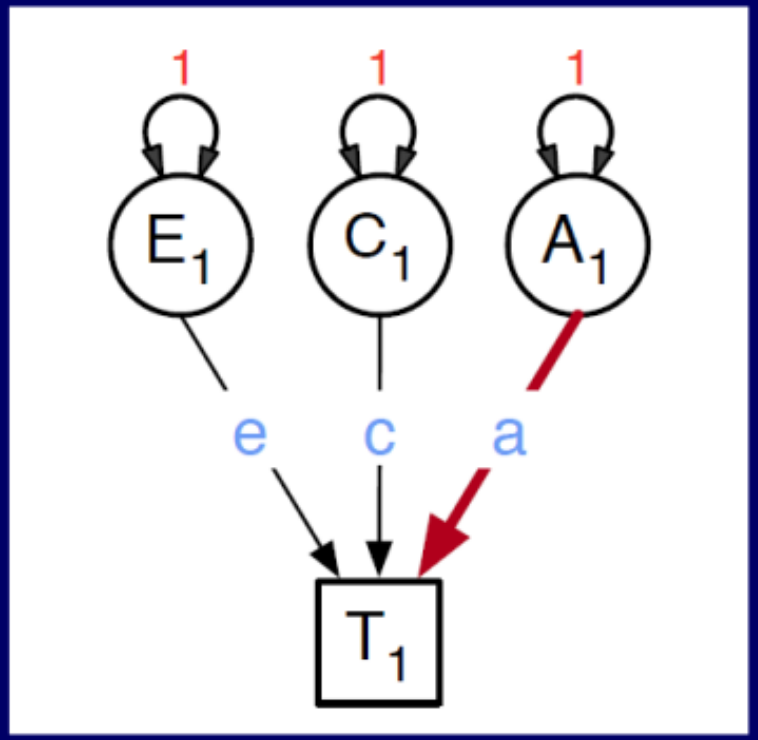
Variance of Twin 1 AND Twin 2 (for MZ and DZ pairs)



a*1

Total Variance = $a^2 + c^2 + e^2$

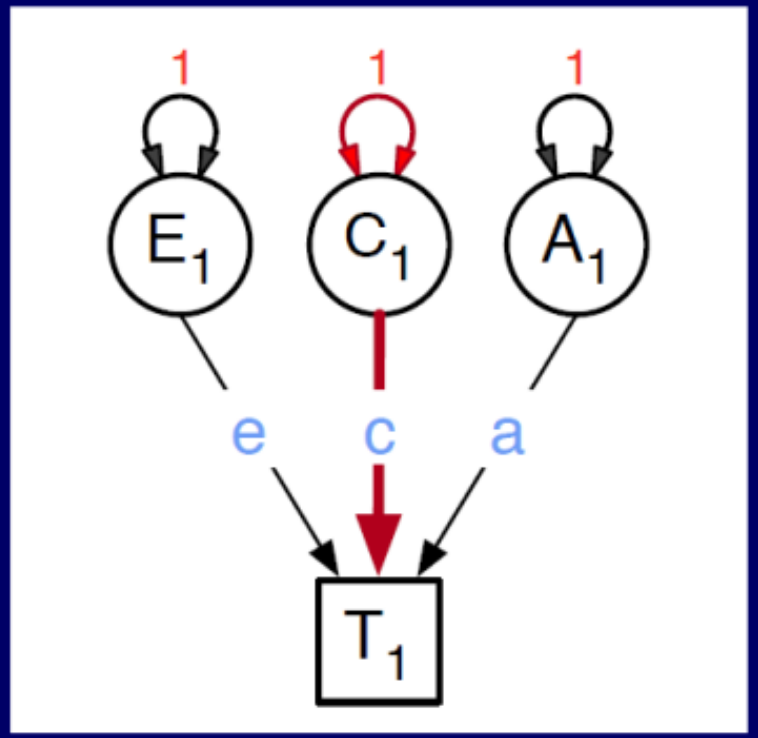
Variance of Twin 1 AND Twin 2 (for MZ and DZ pairs)



$a*1*a = a^2$

Total Variance = a^2 + …

Variance of Twin 1 AND Twin 2 (for MZ and DZ pairs)



$a*1*a = a^2$

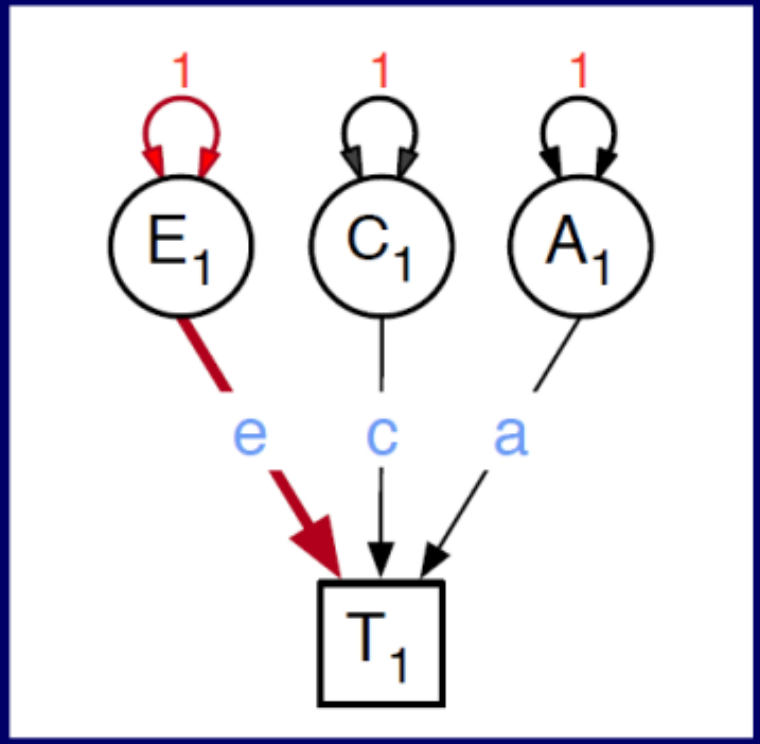
+

$c*1*c = c^2$

+

Total Variance = $a^2 + c^2 + \ldots$

Variance of Twin 1 AND Twin 2 (for MZ and DZ pairs)



$a*1*a = a^2$

+

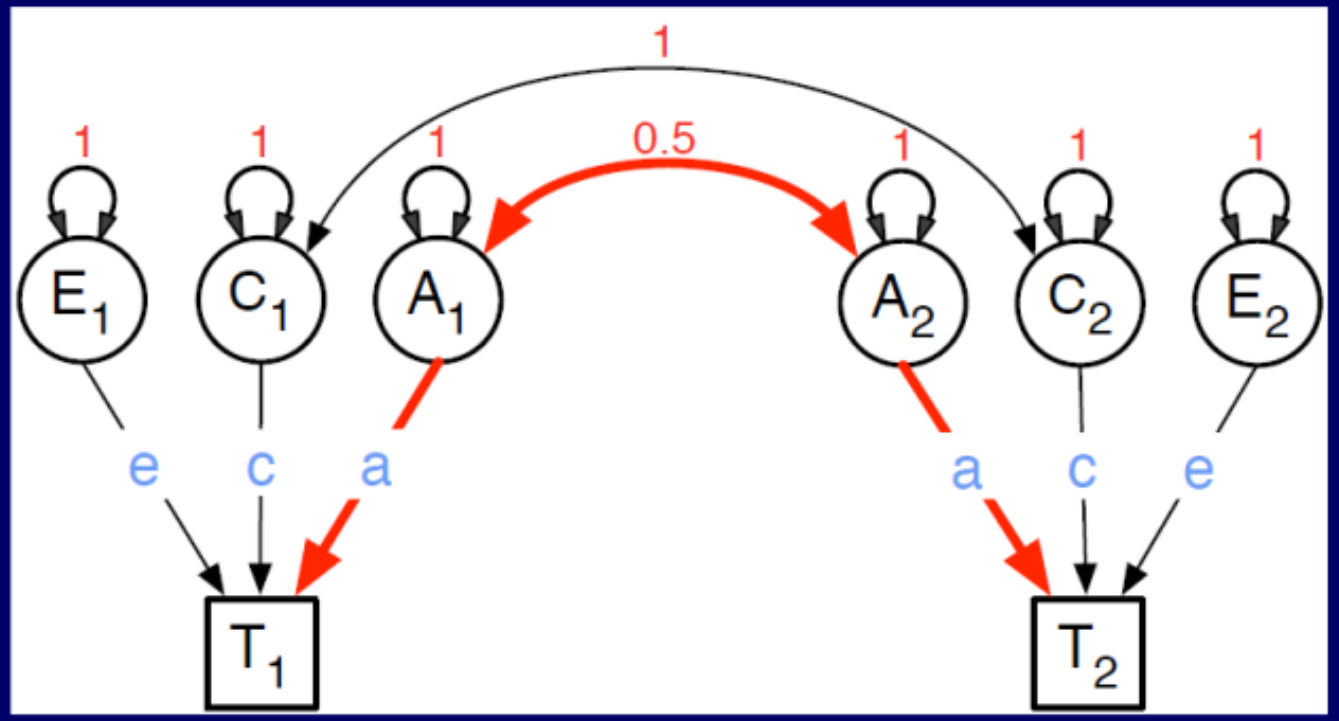
$c*1*c = c^2$

+

$e*1*e = e^2$

Total Variance = $a^2 + c^2 + e^2$

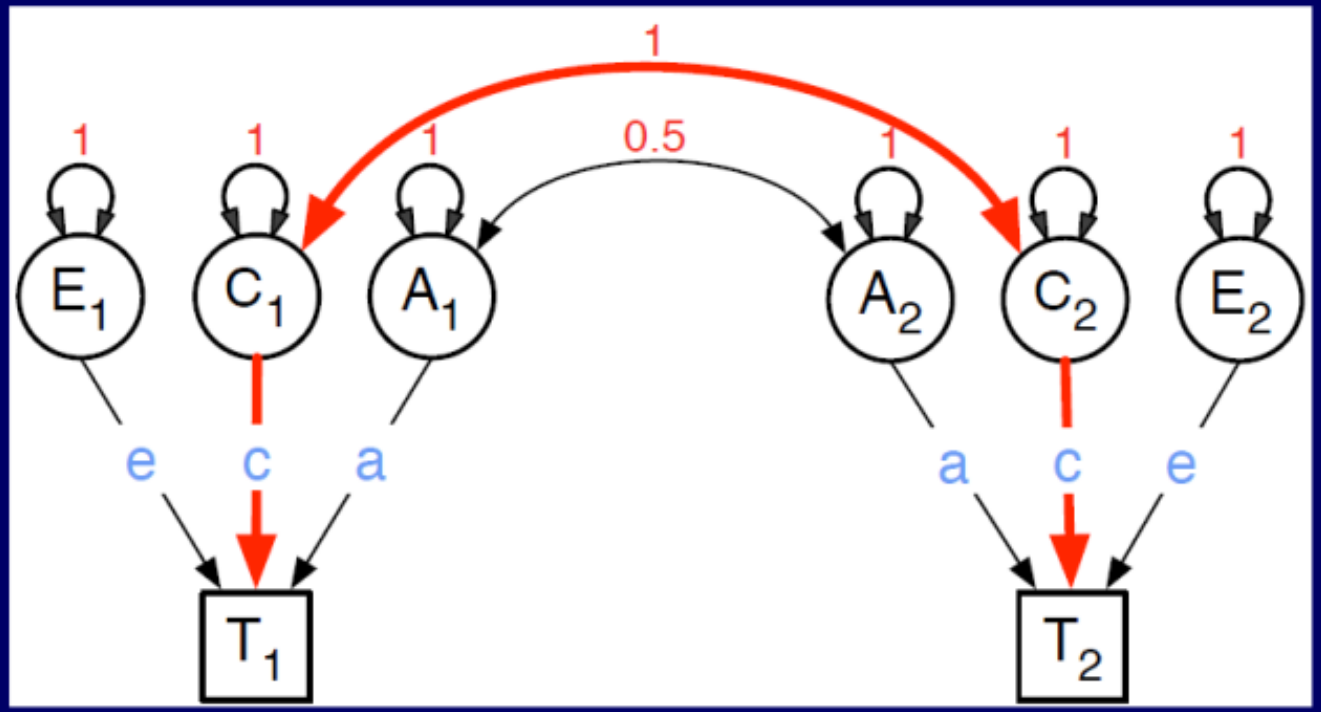
Covariance of Twin 1 AND Twin 2 (for DZ pairs)



$a*.5*a = .5a^2$
+

Covariance = $.5a^2$ + …

Covariance of Twin 1 AND Twin 2 (for DZ pairs)



$a^*.5^*a = .5a^2$
+
$c^* 1^*c = c^2$

Total Covariance = $.5a^2 + c^2$

Predicted Variance-Covariance Matrices ACE Path Model

Cov MZ

	Tw1	Tw2
Tw1	$a^2+c^2+e^2$	a^2+c^2
Tw2	a^2+c^2	$a^2+c^2+e^2$

Cov DZ

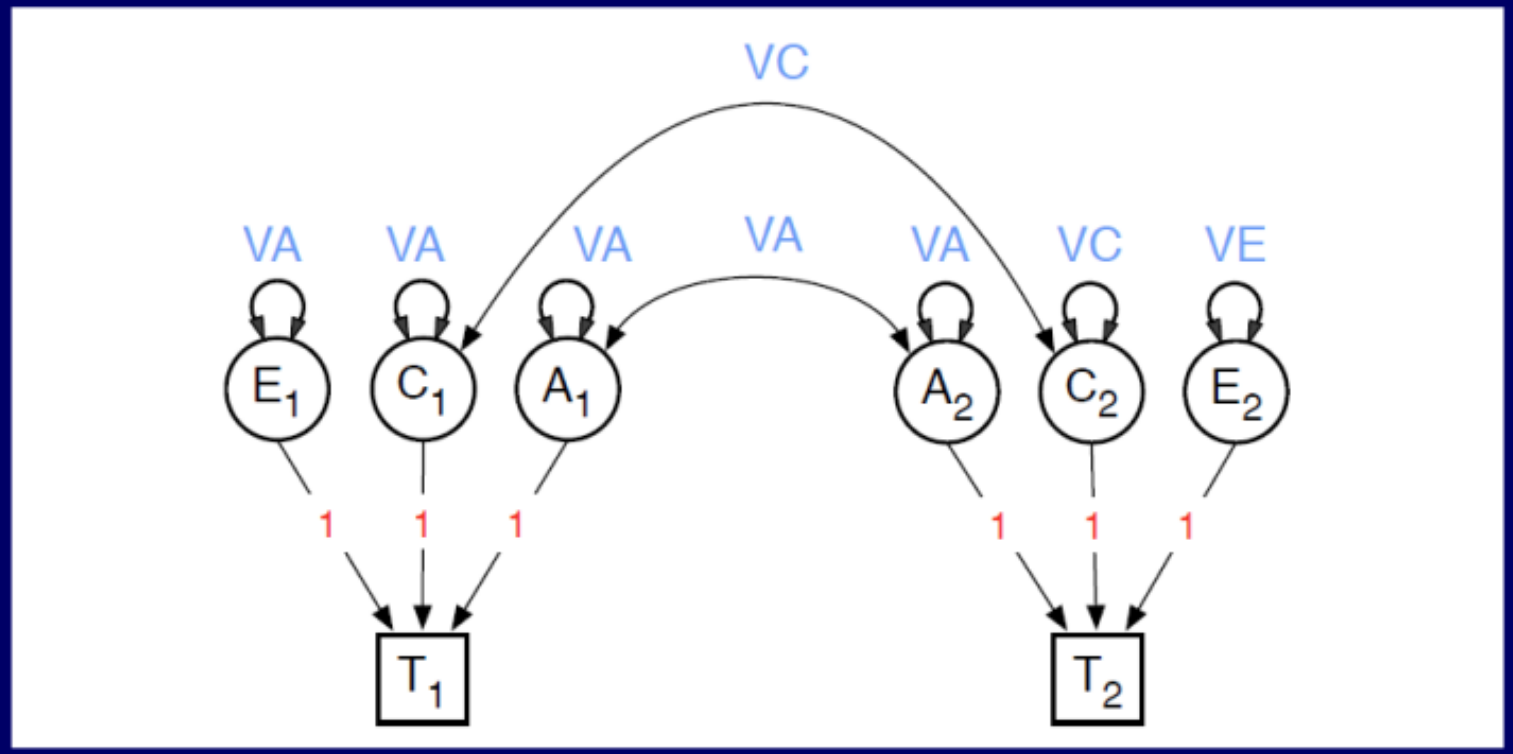
	Tw1	Tw2
Tw1	$a^2+c^2+e^2$	$\frac{1}{2}a^2+c^2$
Tw2	$\frac{1}{2}a^2+c^2$	$a^2+c^2+e^2$

Code in estimation of path coefficients

```
pathA     <- mxMatrix( type="Lower", nrow=nv, ncol=nv, free=TRUE, values=svPa, label="a11", lbound=lbPa, name="a" )
pathC     <- mxMatrix( type="Lower", nrow=nv, ncol=nv, free=TRUE, values=svPa, label="c11", lbound=lbPa, name="c" )
pathE     <- mxMatrix( type="Lower", nrow=nv, ncol=nv, free=TRUE, values=svPe, label="e11", lbound=lbPa, name="e" )

# Create Algebra for Variance Components
covA      <- mxAlgebra( expression=a %*% t(a), name="A" )
covC      <- mxAlgebra( expression=c %*% t(c), name="C" )
covE      <- mxAlgebra( expression=e %*% t(e), name="E" )

# Create Algebra for expected Variance/Covariance Matrices in MZ & DZ twins
covP      <- mxAlgebra( expression= A+C+E, name="V" )
covMZ     <- mxAlgebra( expression= A+C, name="cMZ" )
covDZ     <- mxAlgebra( expression= 0.5%x%A+ C, 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" )
```

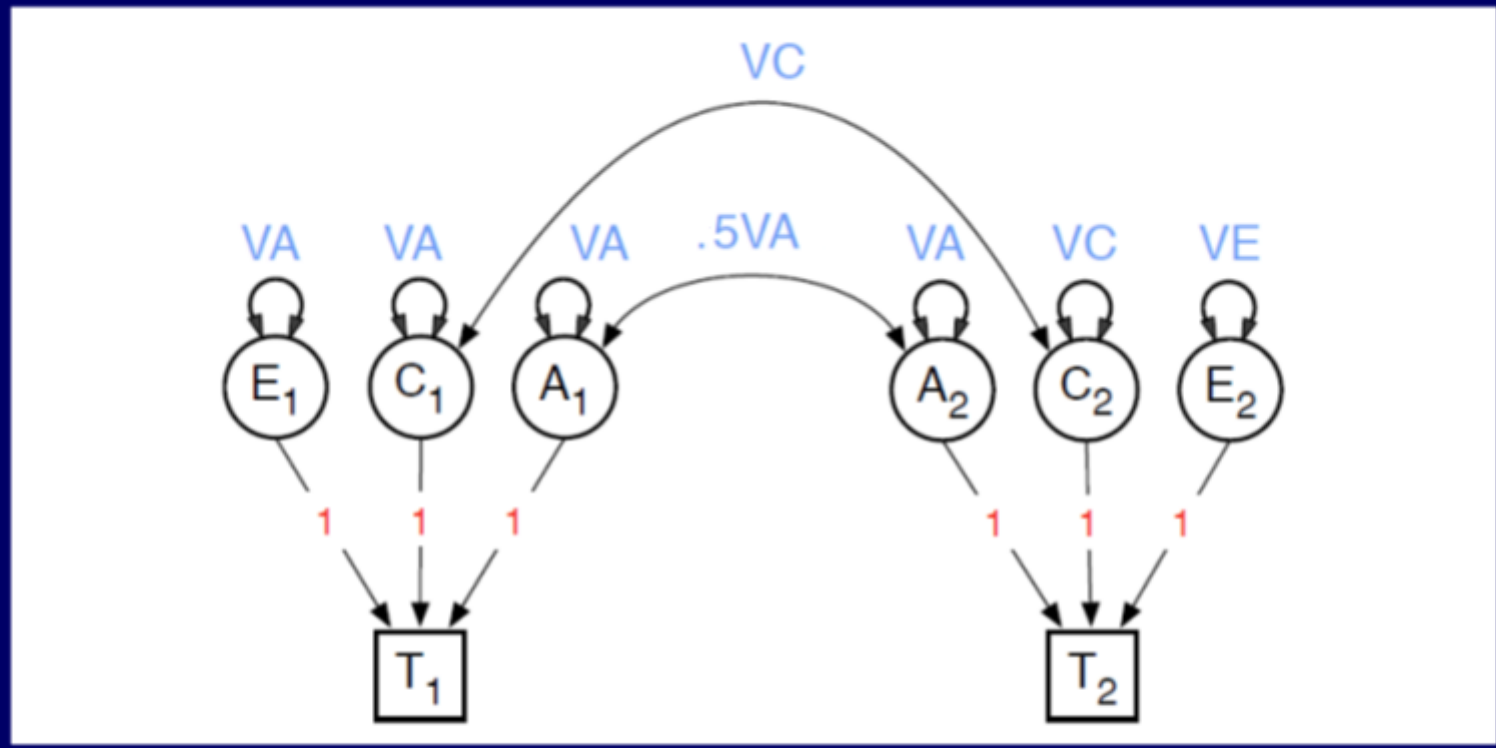
Estimation of variance components

Variance Component Model: MZ



Latent variables A1 C1 and E1 have variances VA, VC and VE, and cause phenotype T1 via regression paths 1. Same model for T2

Variance Component Model: DZ



Latent variables A1 C1 and E1 have variances VA, VC and VE, and cause phenotype T1 via regression paths 1. Same model for T2

Predicted Variance-Covariance Matrices ACE VC Model

Cov MZ

	Tw1	Tw2
Tw1	$V_A+V_C+V_E$	V_A+V_C
Tw2	V_A+V_C	$V_A+V_C+V_E$

Cov DZ

	Tw1	Tw2
Tw1	$V_A+V_C+V_E$	$.5V_A+V_C$
Tw2	$.5V_A+V_C$	$V_A+V_C+V_E$

Code in estimation of variance components

```
## Create Matrices for Variance Components
covA    <- mxMatrix( type="Symm", nrow=nv, ncol=nv, free=TRUE, values=valDiag(svPa,nv), label=labLower("VA",nv), name="VA" )
covC    <- mxMatrix( type="Symm", nrow=nv, ncol=nv, free=TRUE, values=valDiag(svPa,nv), label=labLower("VC",nv), name="VC" )
covE    <- mxMatrix( type="Symm", nrow=nv, ncol=nv, free=TRUE, values=valDiag(svPa,nv), label=labLower("VE",nv), name="VE" )

## Create Algebra for expected Variance/Covariance Matrices in MZ & DZ twins
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" )
```

What’s the difference?

- **Path**: Implicit (artificial) boundary constraint
  - Estimate a but a^2 can *never* be negative.
  - As the number of variables in a twin model increases, the number of implicit boundaries in the model increase.

- **Variance Component**: Unbounded
  - Estimates VA, VC, and VE can be positive and negative

Why do we prefer the variance component approach?

The statistical significance of the parameters from a univariate ACE model is often assessed using a likelihood ratio test.

Under certain regularity conditions this statistic is asymptotically distributed as $\chi 2$ with 1 d.f. BUT these regularity conditions are not met when models have either implicit or explicit bounds.

When boundaries are included, the numerical Type I error rates are lower than theoretically expected.

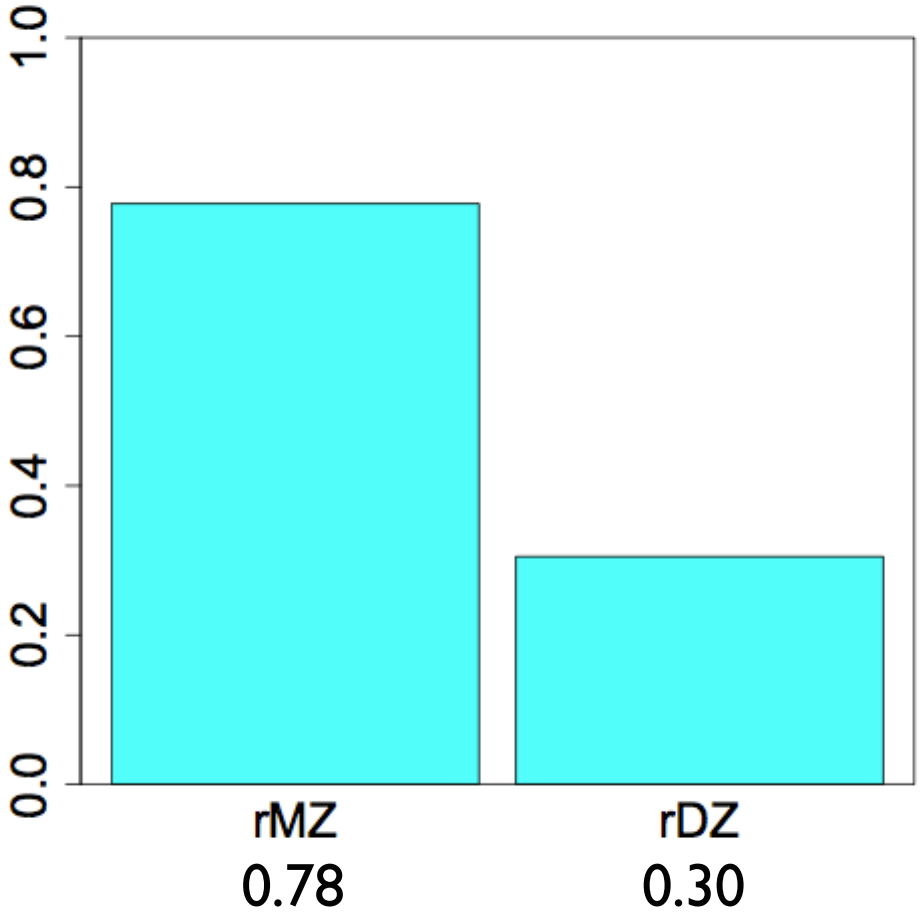
- The null hypotheses that either $a^2 = 0$ or $c^2 = 0$ are rejected less frequently than would be expected due to chance.
- This, causes an increase in Type II errors, where we can falsely conclude the variance component is not significant.

Why do we prefer the variance component approach?

- It may fit better
  - No bias from implicit boundary

- Negative variances? Model wrong?

BMI Twin Correlations



A = 2(rMZ-rDZ)
C = 2rDZ - rMZ
E = 1- rMZ

ADE or ACE?

Why do we prefer the variance component approach?

Behavior Genetics (2019) 49:99–111
https://doi.org/10.1007/s10519-018-9942-y

ORIGINAL RESEARCH

Type I Error Rates and Parameter Bias in Multivariate Behavioral Genetic Models

Brad Verhulst[1] · Elizabeth Prom-Wormley[2] · Matthew Keller[3] · Sarah Medland[4] · Michael C. Neale[5]