Advanced Genetic Epidemiology Statistical Workshop

NIDA 2012 AGES WORKSHOP

October 22-26, 2012

Overlook

This workshop is designed to provide an overview of advanced statistical methodology for genetic studies of substance use and abuse phenotypes.

It covers analytical methods for twin and family studies, including measurement and phenotyping, development, family processes and GxE interaction.

TABLE 1. Four Major Paradigms of Psychiatric Genetics

Paradigm	Samples Studied	Method of Inquiry	Scientific Goals
1. Basic genetic epidemiology	Family, twin, and adoption studies	Statistical	To quantify the degree of familial aggregation and/or heritability
2. Advanced genetic epidemiology	Family, twin, and adoption studies	Statistical	To explore the nature and mode of action of genetic risk factors
3. Gene finding	High-density families, trios, case-control samples	Statistical	To determine the genomic location and identity of susceptibility genes
4. Molecular genetics	Individuals	Biological	To identify critical DNA variants and trace the biological pathways from DNA to disorder

Heritability	Psychiatric Disorders	Other Important Familial Traits
~zero		Language Religion
20-40%	Anxiety disorders, Depression, Bulimia, Personality Disorders	Myocardial Infarction, Normative Personality, Breast Cancer, Hip Fracture
40-60%	Alcohol Dependence Drug Dependence	Blood Pressure, Asthma Plasma cholesterol, Prostate Cancer, Adult-onset diabetes
60-80%	Schizophrenia Bipolar Illness	Weight, Bone Mineral Density
80-100%	Autism	Height, Total Brain Volume

Twin study

- Twins are a valuable source for observation because their genotypes and family environments tend to be similar.

- monozygotic (MZ) or "identical" twins, share nearly 100% of their genetic polymorphisms, which means that most variation in pairs' traits is due to their unique experiences.

- Dizygotic (DZ) or "fraternal" twins share only about 50% of their polymorphisms. Fraternal twins are helpful to study because they tend to share many aspects of their environment by virtue of being born in the same time and place.

Quick review

- Structural Equation Modeling
- Path Diagram
- ACE Model
- R & OpenMX

Structural Equation Modeling

- Structural Equation Modeling is a very general, and powerful multivariate analysis technique allow both confirmatory and exploratory modeling, meaning they are suited to both theory testing and theory development.

- Factor analysis, path analysis and regression all represent special cases of SEM.

Structural Equation Modeling

- Among the strengths of SEM is the ability to construct latent variables: variables which are not measured directly, but are estimated in the model from several measured variables.

http://en.wikipedia.org/wiki/Structural_equation_modeling

The Basic Idea Behind Structural Modeling

- One of the fundamental ideas taught in intermediate applied statistics courses is the effect of additive and multiplicative transformations on a list of numbers. Students are taught that, if you multiply every number in a list by some constant K, you multiply the *mean* of the numbers by K. Similarly, you multiply the *standard deviation* by the absolute value of K.

Structural Equation Modeling

- For example, suppose you have the list of numbers 1,2,3. These numbers have a *mean* of 2 and a *standard deviation* of 1. Now, suppose you were to take these 3 numbers and multiply them by 4. Then the mean would become 8, and the standard deviation would become 4, the variance thus 16.

Structural Equation Modeling

- The point is, if you have a set of numbers X related to another set of numbers Y by the equation Y = 4X, then the variance of Y *must* be 16 times that of X, so you can test the hypothesis that Y and X are related by the equation Y = 4X *indirectly* by comparing the variances of the Y and X variables.

Structural Equation Modeling

- This idea generalizes, in various ways, to several variables inter-related by a group of linear equations. The rules become more complex, the calculations more difficult, but the basic message remains the same -- you can test whether variables are interrelated through a set of linear relationships by examining the variances and covariances of the variables.

http://www.statsoft.com/textbook/structural-equation-modeling/

Path Diagram

- Path Diagrams play a fundamental role in structural modeling. Path diagrams are like flowcharts.

- They show variables interconnected with lines that are used to indicate causal flow.

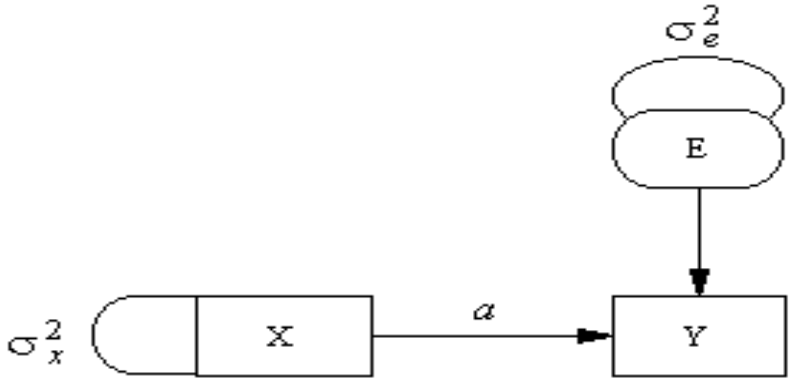
- Most structural equation models can be expressed as path diagrams.

Path Diagram

- Consider the classic linear regression equation

$$y = a x + e$$

- Any such equation may be represented in a path diagram as follows:



http://www.statsoft.com/textbook/structural-equation-modeling/

ACE model

- Typically these three components are called:
  - **A** (additive genetics)
  - **C** (common environment)
  - **E** (unique environment)

It is also possible to examine non-additive genetics effects (often denoted **D** for dominance)

ACE model

- Monozygotic (identical - MZ) twins raised in a family share both 100% of their genes, and all of the shared environment. Any differences arising between them in these circumstances are random (unique).
- The correlation we observe between identical twins provides an estimate of $A + C$.

ACE model

- Dizygous (DZ) twins have a common shared environment, and share on average 50% of their genes:
- so the correlation between fraternal twins is a direct estimate of ½ $A + C$.

ACE model

- If *r* is the correlation observed for a particular trait, then:

- $r_{mz} = A + C$
- $r_{dz} = \frac{1}{2}A + C$

- Where r_{mz} and r_{dz} are simply the correlations of the trait in identical and fraternal twins respectively.

http://en.wikipedia.org/wiki/Twin_study

R & OpenMX

- 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 Structural equation modeling (SEM) model and estimate parameters given observed data.

http://openmx.psyc.virginia.edu/

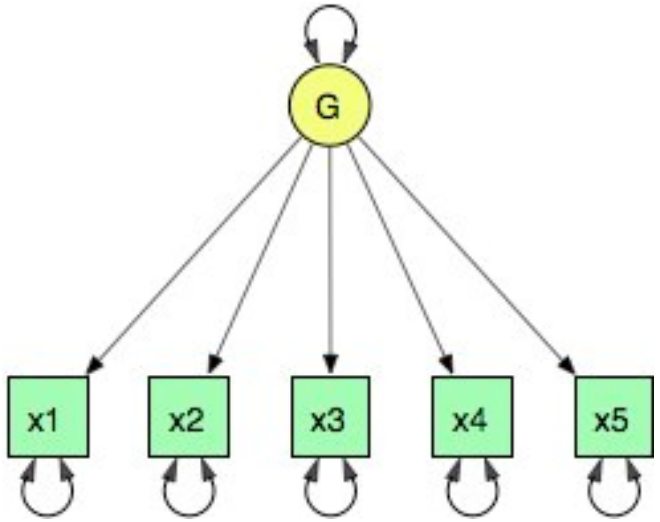
R & OpenMX

- **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.

Path Model Specification

- Here is a path diagram for a one factor path model with five indicators. Beside it is an R script using **OpenMx** path modeling commands to read the data from disk, create the one factor model, fit the model to the observed covariances, and print a summary of the results.

Path Model Specification




```
require(OpenMx)
data(demoOneFactor)
manifests <- names(demoOneFactor)
latents <- c("G")
factorModel <- mxModel("One Factor",
    type="RAM",
    manifestVars = manifests,
    latentVars = latents,
    mxPath(from=latents, to=manifests),
    mxPath(from=manifests, arrows=2),
    mxPath(from=latents, arrows=2,
        free=FALSE, values=1.0),
    mxData(cov(demoOneFactor),
type="cov",
        numObs=500))
summary(mxRun(factorModel))
```

Matrix Model Specification

- **OpenMx** can also specify models in terms of matrix algebra. On the left is an equation for the same one factor path model with five indicators. Beside it is an R script using **OpenMx** matrix modeling commands to read the data from disk, create the one factor model, fit the model to the observed covariances, and print a summary of the results.

Matrix Model Specification

$\mathbf{R} = \mathbf{ALA}' + \mathbf{U}$	data(demoOneFactor) factorModel <- mxModel("One Factor", mxMatrix("Full", 5, 1, values=0.2, free=TRUE, name="A"), mxMatrix("Symm", 1, 1, values=1, free=FALSE, name="L"), mxMatrix("Diag", 5, 5, values=1, free=TRUE, name="U"), mxAlgebra(A %*% L %*% t(A) + U, name="R"), mxMLObjective("R", dimnames = names(demoOneFactor)), mxData(cov(demoOneFactor), type="cov", numObs=500)) summary(mxRun(factorModel))

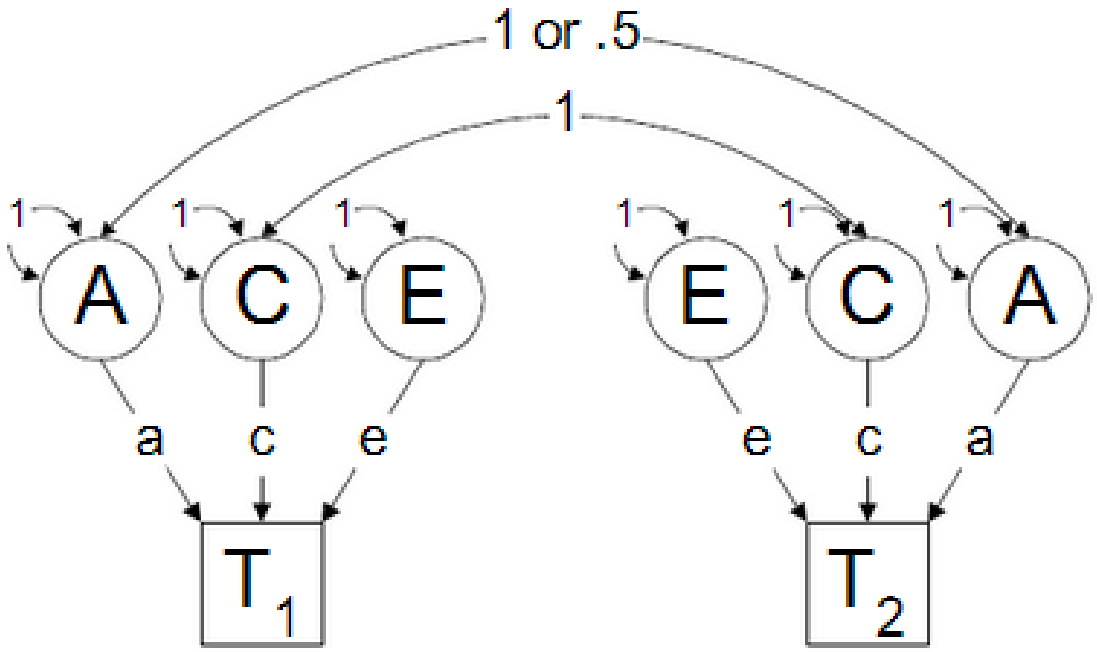
Multivariate Genetic Analysis - Question

Univariate Analysis: What are the contributions of additive genetic, dominance/ shared environmental and unique environmental factors to the variance?

Bi/Multivariate Analysis: What are the contributions of genetic and environmental factors to the covariance between two traits?

What makes sets of variables correlate or co-vary, comorbid?

Univariate ACE model



http://www.slideshare.net/devenvaija09/bivariate

Expected Covariance Matrices

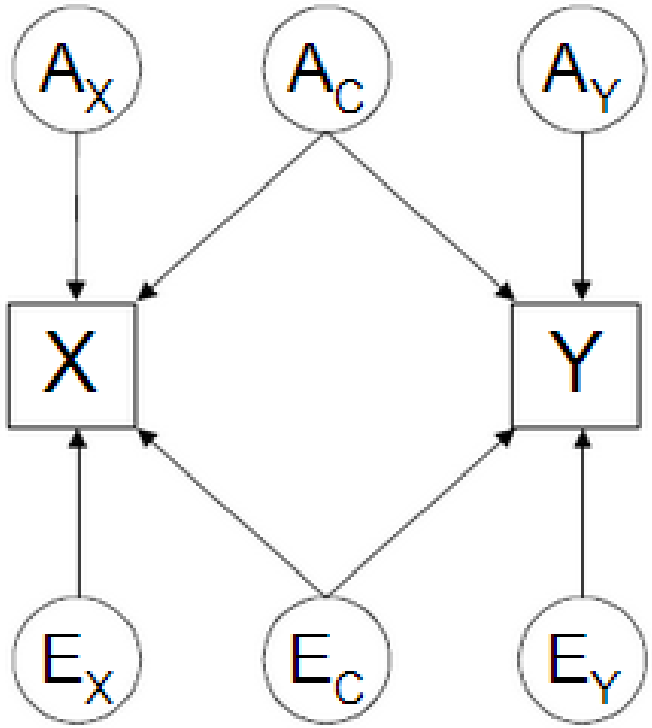
$$\Sigma \text{ MZ} = \begin{bmatrix} a^2+c^2+e^2 & a^2+c^2 \\ a^2+c^2 & a^2+c^2+e^2 \end{bmatrix} \quad 2 \times 2$$

$$\Sigma \text{ DZ} = \begin{bmatrix} a^2+c^2+e^2 & .5a^2+c^2 \\ .5a^2+c^2 & a^2+c^2+e^2 \end{bmatrix} \quad 2 \times 2$$

Bivariate Questions I

- Univariate Analysis: What are the contributions of additive genetic, dominance/shared environmental and unique environmental factors to the variance?
- Bivariate Analysis: What are the contributions of genetic and environmental factors to the covariance between two traits?

Two Traits



Bivariate Questions II

- Two or more traits can be correlated because they share common genes or common environmental influences
  - e.g. Are the same genetic/environmental factors influencing the traits?
- With twin data on multiple traits it is possible to partition the covariation into its genetic and environmental components
- Goal: to understand what factors make sets of variables correlate or co-vary

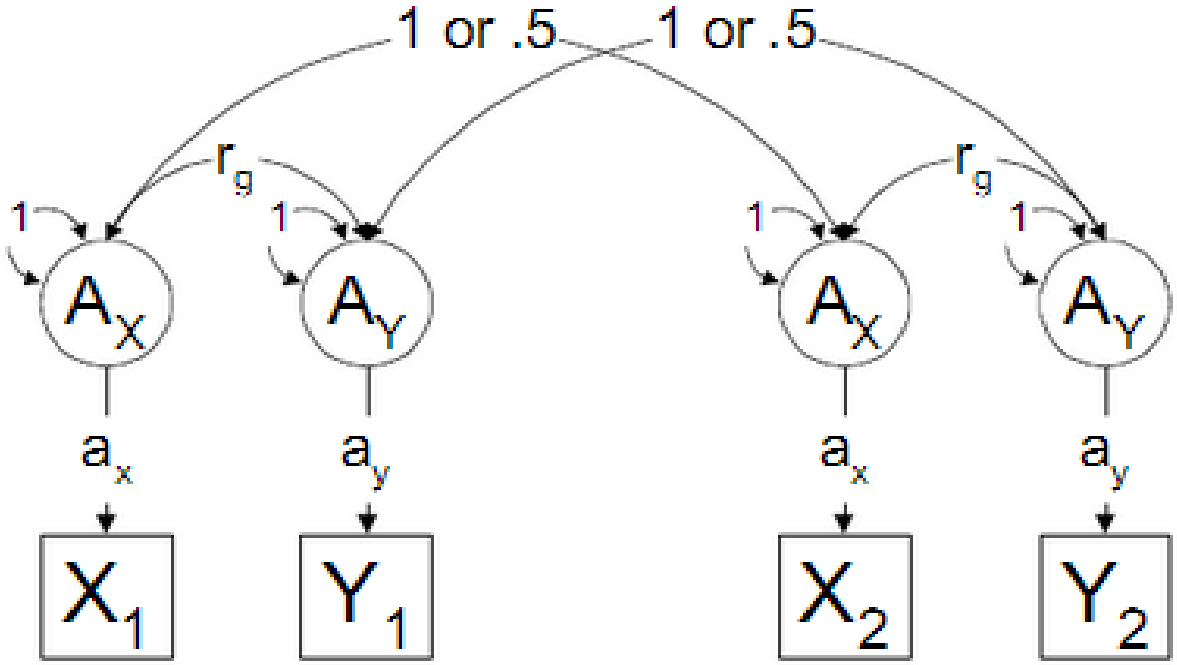
Bivariate Twin Data

		individual twin	
		within	between
trait	within	variance	twin covariance
	between	trait covariance	cross-trait twin covariance

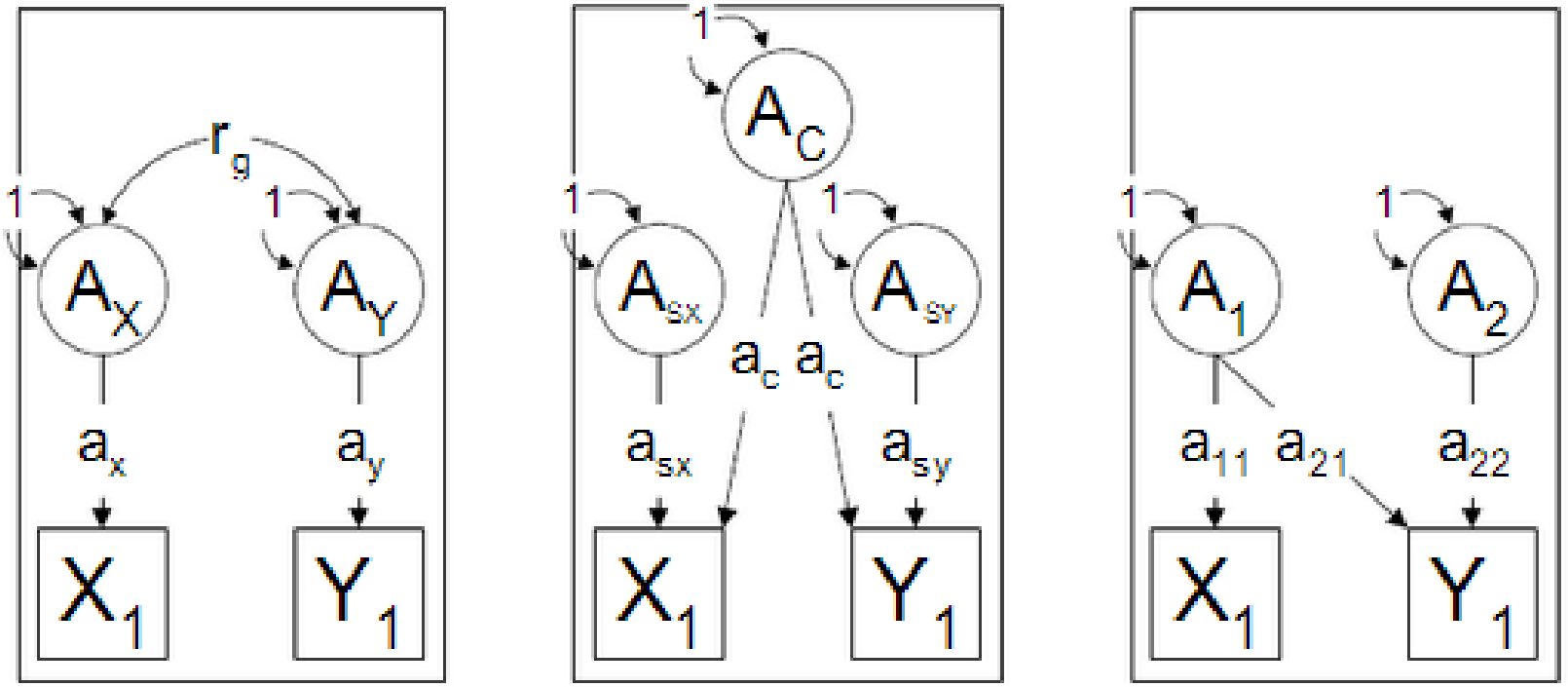
Bivariate Twin Covariance Matrix

	X_1	Y_1	X_2	Y_2
X_1	V_{X1}	C_{X1Y1}	C_{X1X2}	C_{X1Y2}
Y_1	C_{Y1X1}	V_{Y1}	C_{Y1X2}	C_{Y1Y2}
X_2	C_{X2X1}	C_{X2Y1}	V_{X2}	C_{X2Y2}
Y_2	C_{Y2X1}	C_{Y2Y1}	C_{Y2X2}	V_{Y2}

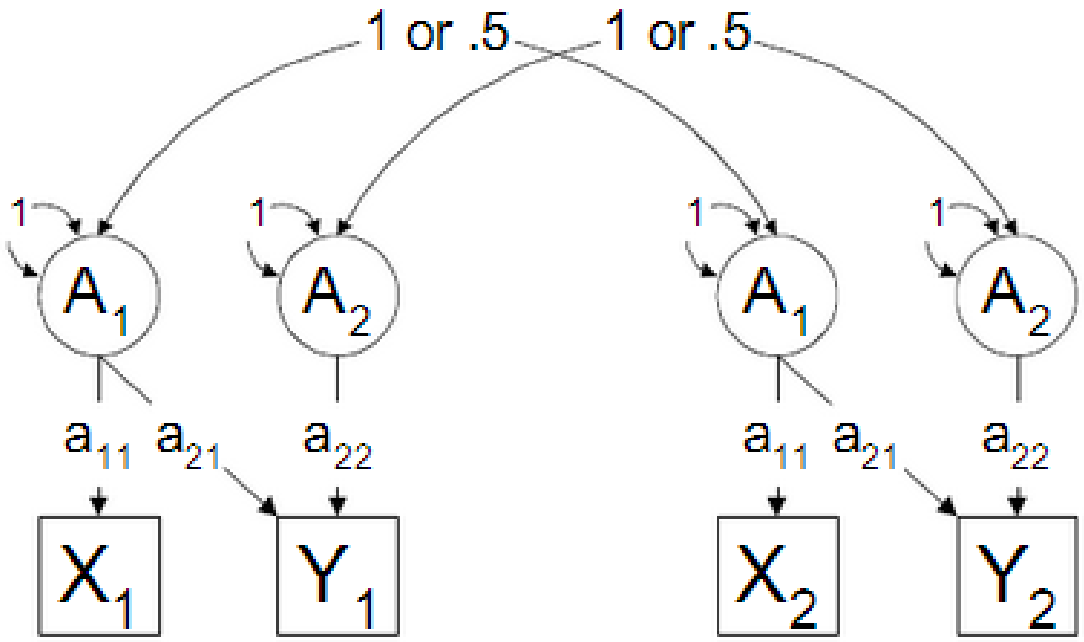
Genetic Correlation



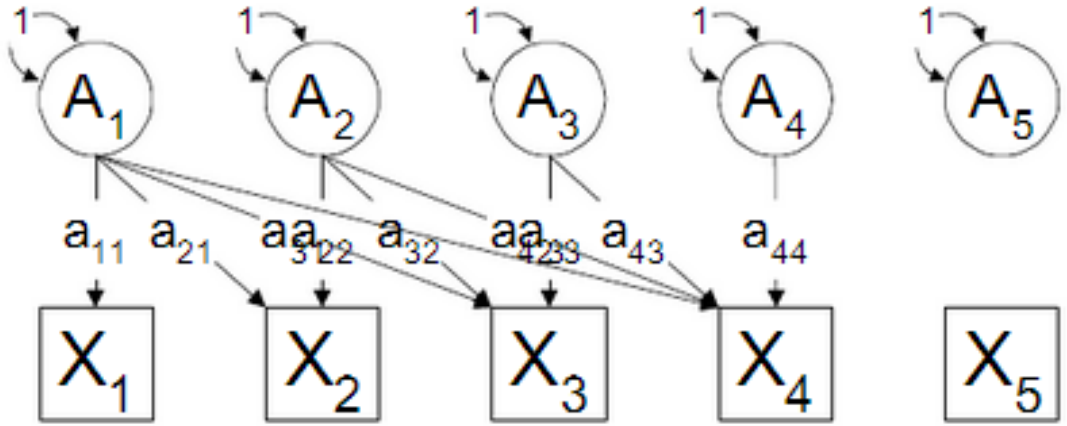
Alternative Representations



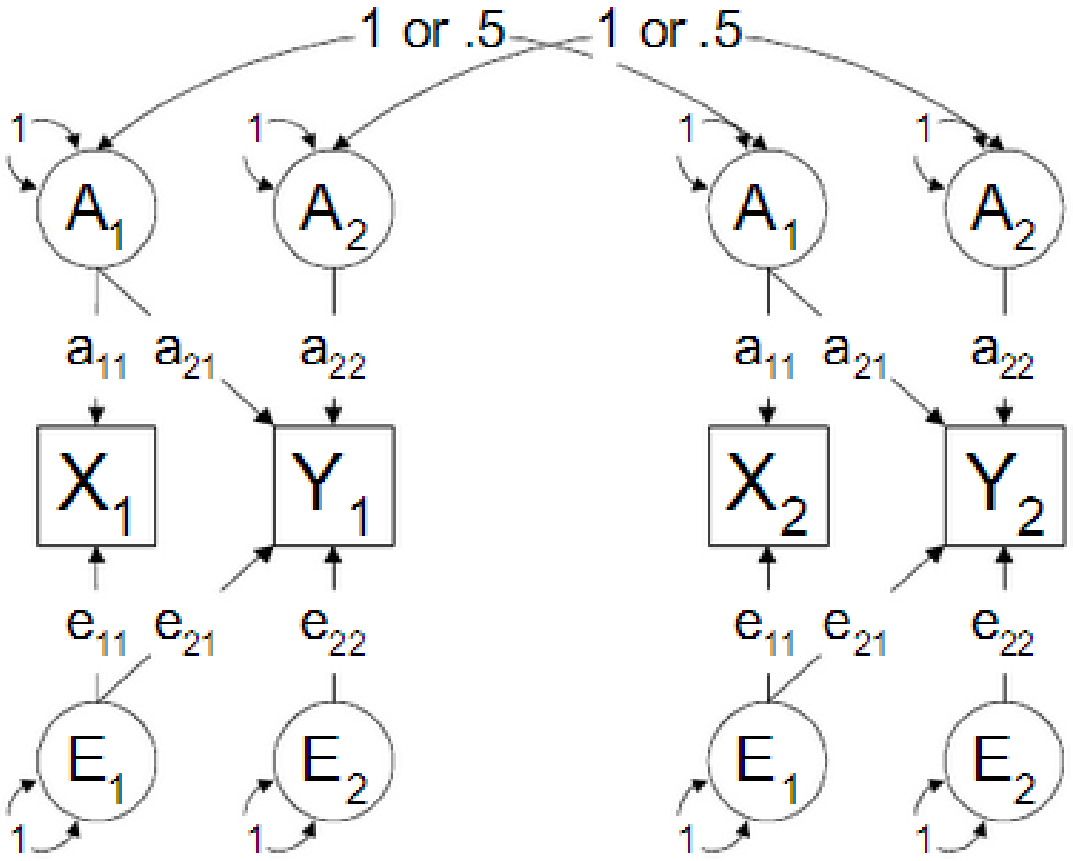
Cholesky Decomposition



More Variables



Bivariate AE Model



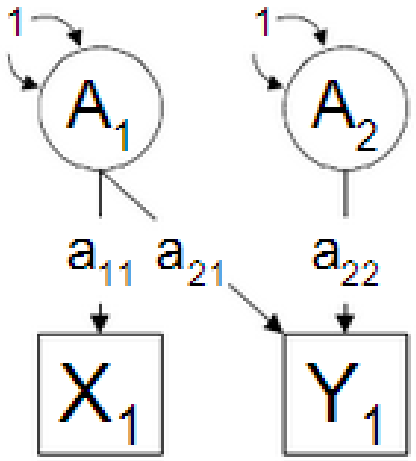
MZ Twin Covariance Matrix

	X_1	Y_1	X_2	Y_2
X_1	$a_{11}^2 + e_{11}^2$		a_{11}^2	
Y_1	$a_{21} * a_{11} + e_{21} * e_{11}$	$a_{22}^2 + a_{21}^2 + e_{22}^2 + e_{21}^2$	$a_{21} * a_{11}$	$a_{22}^2 + a_{21}^2$
X_2				
Y_2				

DZ Twin Covariance Matrix

	X_1	Y_1	X_2	Y_2
X_1	$a_{11}^2+e_{11}^2$		$.5a_{11}^2$	
Y_1	$a_{21}*a_{11}+$ $e_{21}*e_{11}$	$a_{22}^2+a_{21}^2+$ $e_{22}^2+e_{21}^2$	$.5a_{21}*a_{11}$	$.5a_{22}^2+$ $.5a_{21}^2$
X_2				
Y_2				

Within-Twin Covariances [Mx]



X Lower 2 2

	A_1	A_2
X_1	a_{11}	0
Y_1	a_{21}	a_{22}

A=X*X'

$$\Sigma A = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix} * \begin{bmatrix} a_{11} & a_{21} \\ 0 & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11}^2 & a_{11}*a_{21} \\ a_{21}*a_{11} & a_{22}^2+a_{21}^2 \end{bmatrix}$$

Within-Twin Covariances

$$\Sigma A = \begin{bmatrix} a_{11}^2 & a_{11}*a_{21} \\ a_{21}*a_{11} & a_{22}^2+a_{21}^2 \end{bmatrix}$$

$$\Sigma E = \begin{bmatrix} e_{11}^2 & e_{11}*e_{21} \\ e_{21}*e_{11} & e_{22}^2+e_{21}^2 \end{bmatrix}$$

$$\Sigma P = \Sigma A + \Sigma E = \begin{bmatrix} a_{11}^2 + e_{11}^2 & a_{11}*a_{21} + e_{11}*e_{21} \\ a_{21}*a_{11} + e_{21}*e_{11} & a_{22}^2+a_{21}^2+e_{22}^2+e_{21}^2 \end{bmatrix}$$

Cross-Twin Covariances

MZ $\quad \Sigma A = \begin{bmatrix} a_{11}^2 & a_{11}*a_{21} \\ a_{21}*a_{11} & a_{22}^2+a_{21}^2 \end{bmatrix}$

DZ $\quad .5@\Sigma A = \begin{bmatrix} .5a_{11}^2 & .5a_{11}*a_{21} \\ .5a_{21}*a_{11} & .5a_{22}^2+.5a_{21}^2 \end{bmatrix}$

Cross-Trait Covariances

- Within-twin cross-trait covariances imply common etiological influences
- Cross-twin cross-trait covariances imply familial common etiological influences
- MZ/DZ ratio of cross-twin cross-trait covariances reflects whether common etiological influences are genetic or environmental

Univariate Expected Covariances

Σ MZ =

$a^2+c^2+e^2$	a^2+c^2
a^2+c^2	$a^2+c^2+e^2$

2 x 2

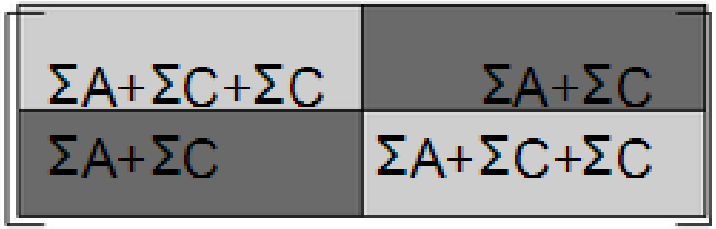
Σ DZ =

$a^2+c^2+e^2$	$.5a^2+c^2$
$.5a^2+c^2$	$a^2+c^2+e^2$

2 x 2

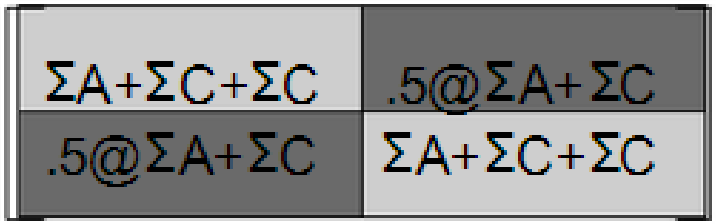
Univariate Expected Covariances II

Σ MZ =



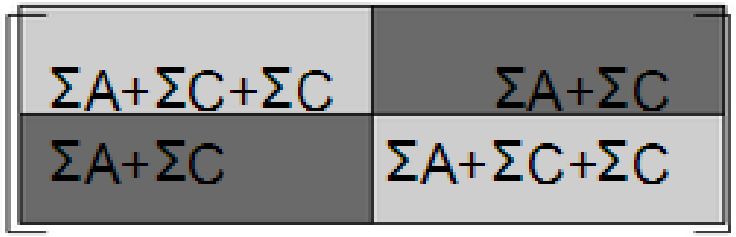
2 x 2

Σ DZ =

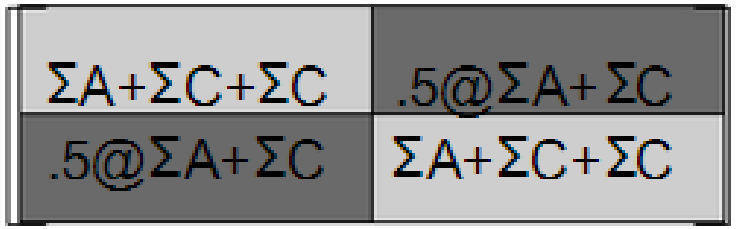


2 x 2

Bivariate Expected Covariances

Σ MZ = 

4 x 4

Σ DZ = 

4 x 4

Practical Example I

- Dataset: MCV-CVT Study
- 1983-1993
- BMI, skinfolds (bic,tri,calf,sil,ssc)
- Longitudinal: 11 years
- N MZFY: 107, DZF: 60

Practical Example II

- Dataset: NL MRI Study
- 1990's
- Working Memory, Gray & White Matter

- N MZFY: 68, DZF: 21

Cholesky decomposition

- A typical starting point in bivariate and multivariate analysis is the **Cholesky decomposition** .

http://genepi.qimr.edu.au/contents/p/staff/CV409.pdf

Cholesky decomposition

- Given a symmetric positive definite matrix A, the Cholesky decomposition is an upper triangular matrix U with strictly positive diagonal entries such that

$$A = U^T U.$$

http://mathworld.wolfram.com/CholeskyDecomposition.html

Cholesky decomposition

- The most commonly used multivariate technique in the Classical Twin Design is **Cholesky decomposition**.
- The Cholesky is a method of triangular decomposition where the first variable (y_1) is assumed to be caused by a latent factor (η_1) that can explain the variance in remaining variables $(y_2, \ldots, y_n)$ and so on.

Cholesky decomposition

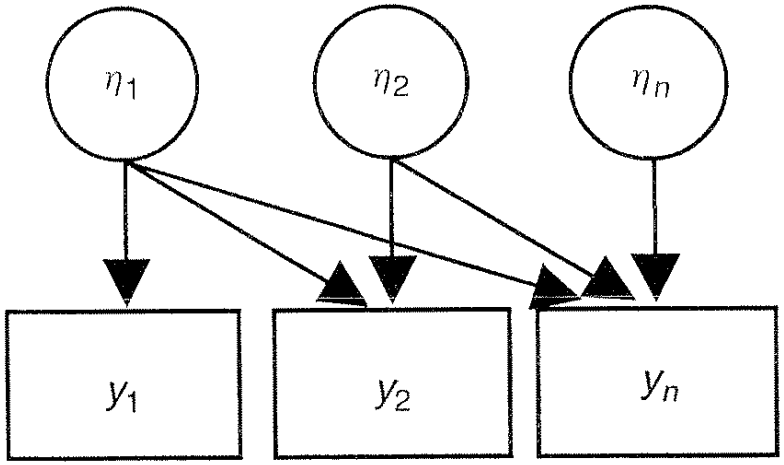


Figure 1 Multivariate Cholesky triangular decomposition, $y_1, \ldots, y_n$ = observed phenotypic variables, η_{1-n} = latent factors

Cholesky decomposition

- The expected variance-covariance matrix in the Cholesky decomposition is parameterized in terms of n latent factors (Where n is the number of variables).
- All variables load on the first latent factor, n-1 variables load on the second factor and so on, the final variable loads on the nth latent factor only. Each source of phenotypic variation (i.e A, C or D, E) is parameterized in the same way.

Cholesky decomposition

- Therefore, the full factor Cholesky does not distinguish between common factor and specific factor variance and does not estimate a specific factor effect for any variable expect the last.