Latent Factor Models in Twins

Brad Verhulst

Daniel Gustavson

International Statistical Genetics Workshop

March 6th, 2024

Necessary information for this session:

Copy files via SSH
Open the SSH client from the workshop hub or:
`https://workshop.colorado.edu/ssh`

Make sure you are in your home folder by typing:
`pwd`

Create a directory to hold today's work by typing:
`mkdir TwinFacMod`

Change your directory to the TwinFacMod folder:
`cd TwinFacMod`

Copy over the files/exercises from my directory into yours by typing the following (please note that there IS a period that must be included at the end of the second line):
`cp /faculty/brad/2024/TwinFacMod/* .`

Check to make sure you have the following files (with `ls`)

Utility of Structural Equation Modeling

Structural Equation Modeling attempts to explain the covariance matrix of all the variables in the analysis rather than the variation in a single dependent variable

If there is only one dependent variable and one independent variable, SEM reduces to regression

- Flexible framework to estimate a variety of causal and correlational models:
  - **Confirmatory Factor Analysis**
  - Path Analysis (mediation, feedback loops)
  - **Regression (linear, logistic, ordinal)**

Three major advantages of SEM over traditional multivariate techniques:

1. Explicit focus on measurement (error)
2. Estimate latent variables
3. Test complex theoretical structures

Factor Analysis

qualtrics XM

Factor analysis is the practice of condensing many variables into just a few, so that your research data is easier to work with.

Factor analysis is a powerful tool when you want to simplify complex data, find hidden patterns, and set the stage for deeper, more focused analysis.

Factor analysis is a way to explain the covariance between a set of observed and latent variables

- Observed variables are concepts that can be directly measured (e.g., Items from questionnaires)
- Latent variables are concepts that must be inferred (through a mathematical model) from observed variables

Confirmatory Factor Analysis (CFA)

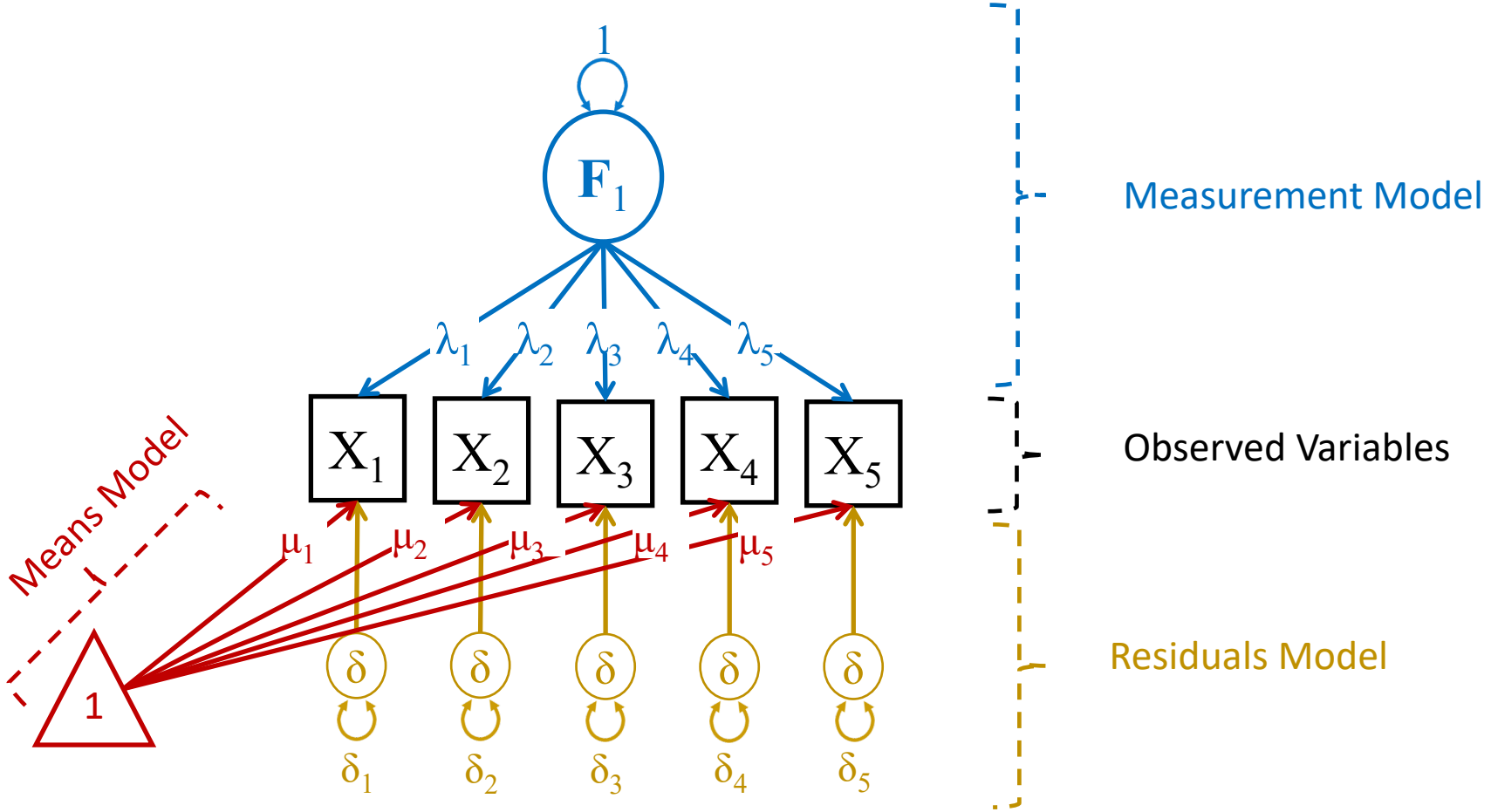
- CFA
  - Theory driven model
  - Must specify which variables are related in advance
    - Both latent and manifest

Caution: It is possible to write/draw unidentified models

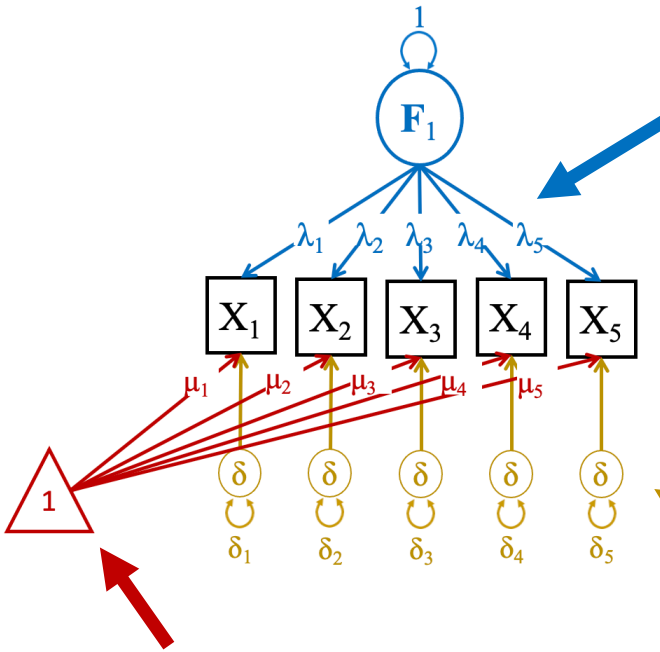
Example: Depression is the cause of the DSM Depression Symptoms:

1) Depressed mood
2) Anhedonia
3) Changes in weight
4) Insomnia or hypersomnia
5) Psychomotor agitation or retardation
6) Fatigue
7) Feelings of worthlessness or guilt
8) Diminished ability to think or concentrate
9) Suicidal ideation

Phenotypic Common Factor Model



Interpreting a CFA



Measurement Model

- Factor Loadings (Regression of the item on the latent factor)
  - A 1 unit increase in the latent factor is associated with a lambda increase in the observed variable

The larger the magnitude of the factor loading, the more central the item is to the interpretation of the latent factor

Means Model

- The expected means of the observed variables
- If we include covariates (e.g., age, sex), the means are intercepts from a regression model
- This means model is saturated and will fit almost perfectly (but it doesn't have to be)

Residuals Model

- Variance in the observed variables not explained by the latent factor
- Some items may have large residuals, or variance that cannot be explained by covariation with the other items in the model

Latent Variables and Identification

- Identification of the Scale of the Latent Factor: Two Approaches
  - Constrain the variance of the latent factor to 1
    - This standardizes the latent factor to have a unit variance
    - We typically assume that the mean is zero
    - Under these circumstances, distribution of the latent factor is assumed to be standard normal
  - Constrain one of the factor loadings
    - This fixes the scale of the latent factor to equal that of the variable with the fixed factor loading
      - A unit increase in x_i corresponds to a unit increase in the latent factor
      - The latent factor is assumed to be normally distributed (but is longer follows a standard normal distribution)
    - This is the default in many SEM programs (e.g., Mplus)

Identification of CFA Models

- **The t-Rule: $t \leq \frac{1}{2} q(q+1)$**
  - The number of free parameters t in the model must be equal to or less than the number of unique elements in the covariance matrix, q(q+1)/2.
    - 'Unique' means different expectations
  - The t-Rule is necessary but not sufficient

- **The 3 Indicator Rule:**
  - A 1 Factor model is identified if there are three indicators with non-zero loadings and a diagonal residual matrix.
  - With more than three indicators of a factor the model may be over-identified.
  - Multifactor models are identified if:
    1. Each factor has 3 indicators
    2. Each row has one and only one nonzero (free) element (This implies simple structure)
    3. The residual matrix is diagonal

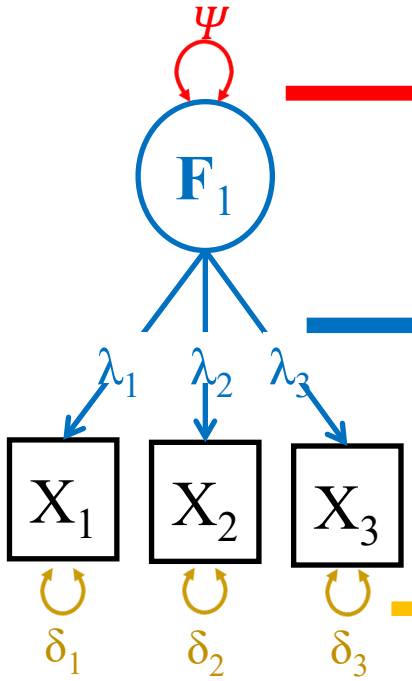
These are sufficient conditions, but they are not necessary:

Exceptions can be made (i.e., Correlated Residuals, Cross-loadings)

- **The Two-Indicator Rule:**
  - The residual matrix is diagonal
  - One loading (for each factor) is fixed (probably to 1).

Phenotypic Common Factor Model

$$\Sigma_\theta = \Lambda\Psi\Lambda' + \Delta$$

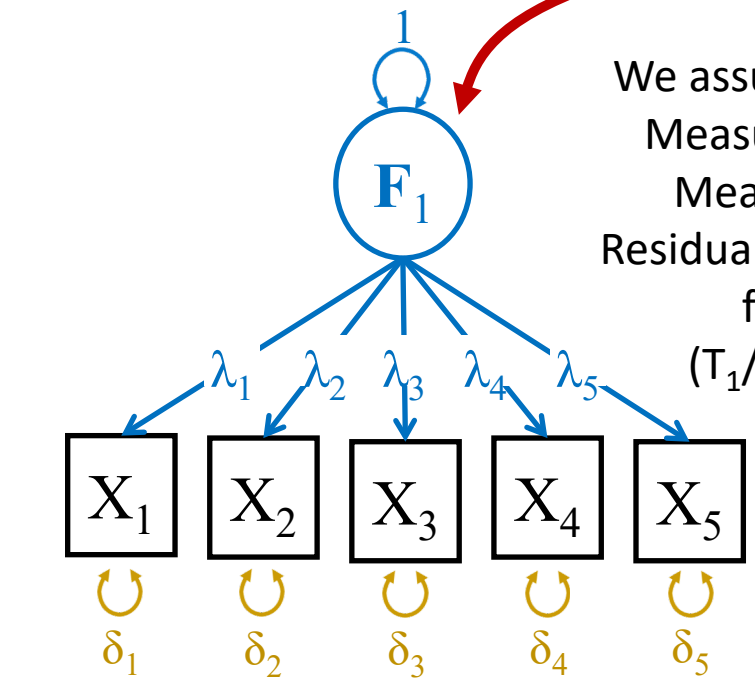


$$\begin{bmatrix}\lambda_1\\ \lambda_2\\ \lambda_3\end{bmatrix}[\Psi][\lambda_1 \quad \lambda_2 \quad \lambda_3] + \begin{bmatrix}\delta_1 & 0 & 0\\ 0 & \delta_2 & 0\\ 0 & 0 & \delta_3\end{bmatrix}$$

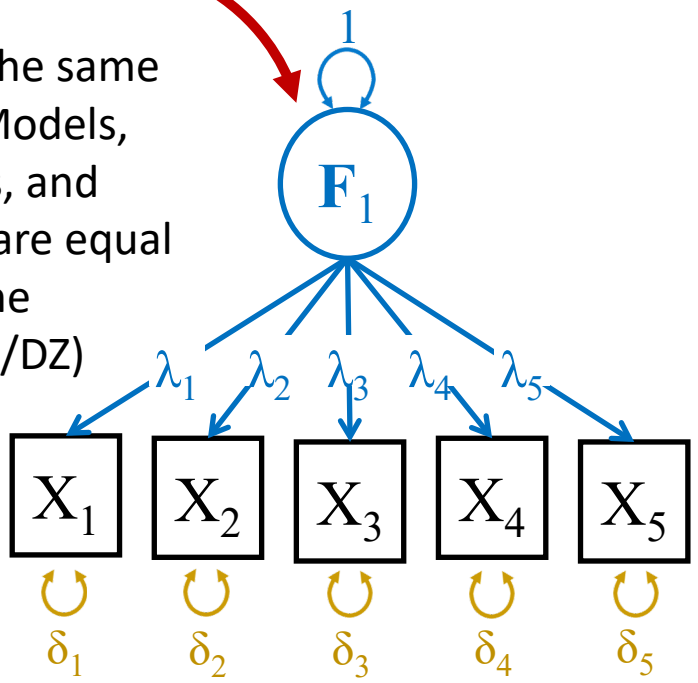
Latent Variable Models in Twins

Familial Resemblance

We assume that the same Measurement Models, Means Models, and Residuals Models are equal for everyone (T_1/T_2 and MZ/DZ)



Twin 1

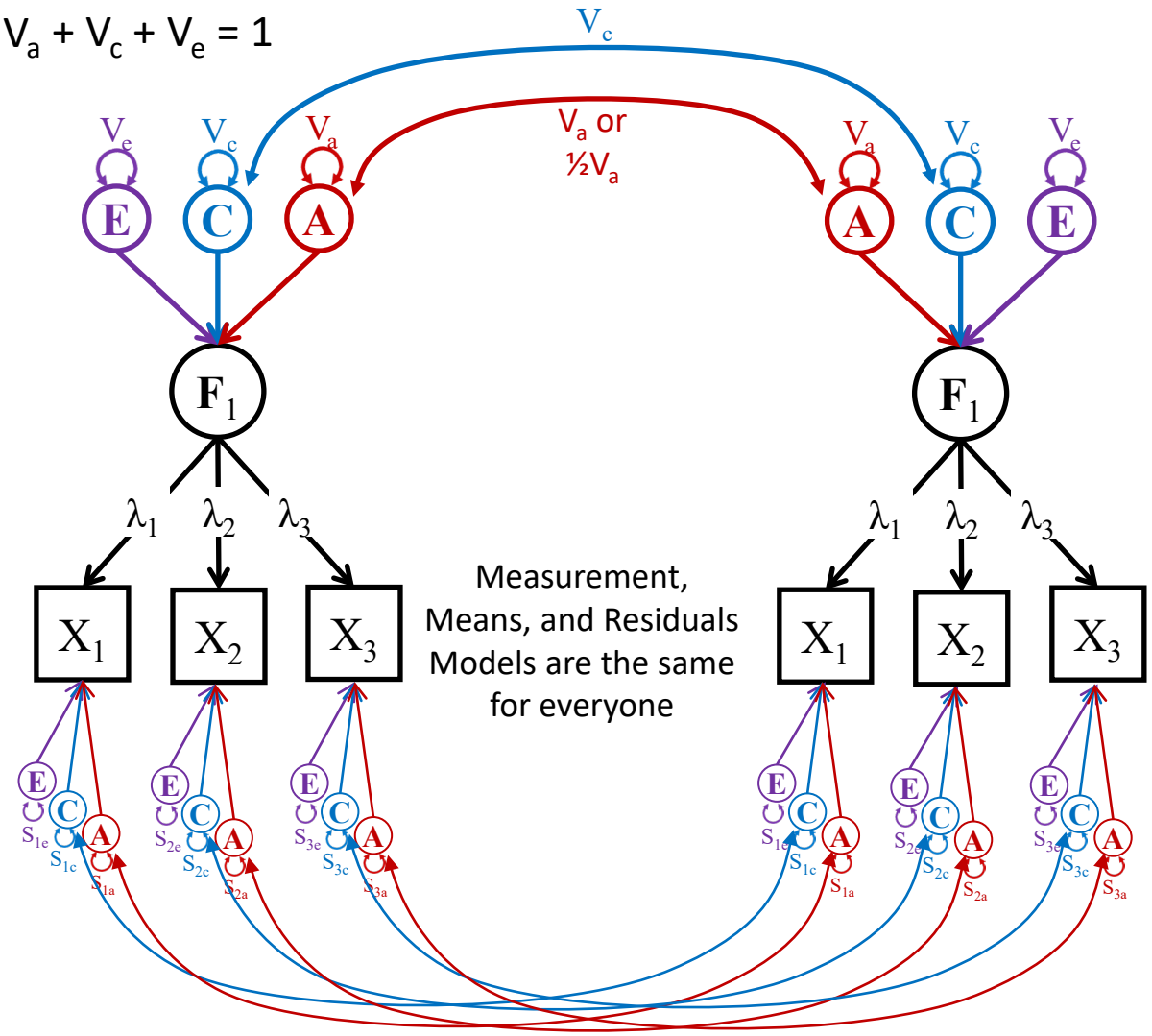


Twin 2

Common Pathway Model

Common Constraint:

$V_a + V_c + V_e = 1$



Measurement Model Variance Decomposition

Standard Measurement Model

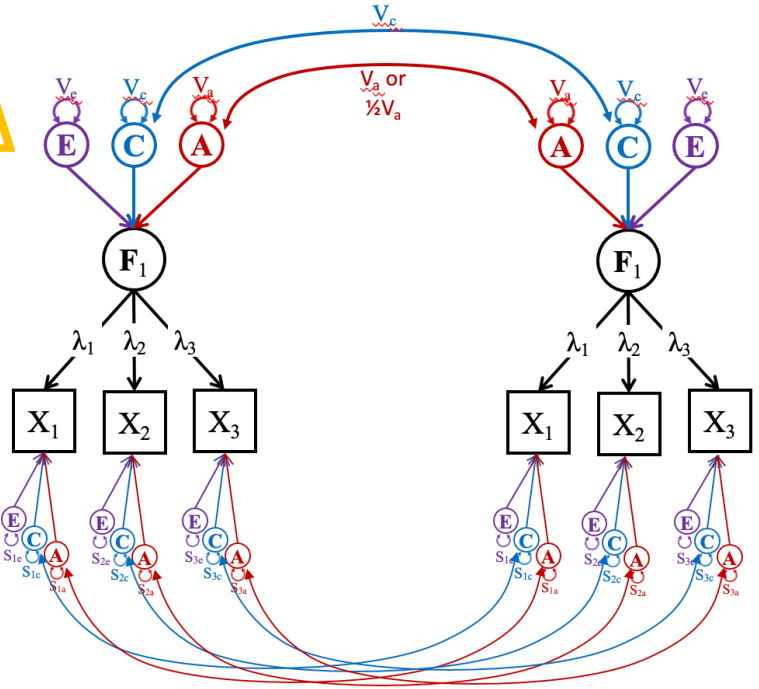
Residuals Model Variance Decomposition

Common Pathway Model Algebra

$$\Sigma_\theta = \Lambda\Psi\Lambda' + \Delta$$

$$\Psi = h \otimes V_a + s \otimes V_c + d \otimes V_e$$

$$\Delta = h \otimes S_a + s \otimes S_c + d \otimes S_e$$



$$h_{MZ} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$h_{DZ} = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$$

$$s = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$d = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Common Pathway Model Algebra

Measurement Model
Variance Decomposition

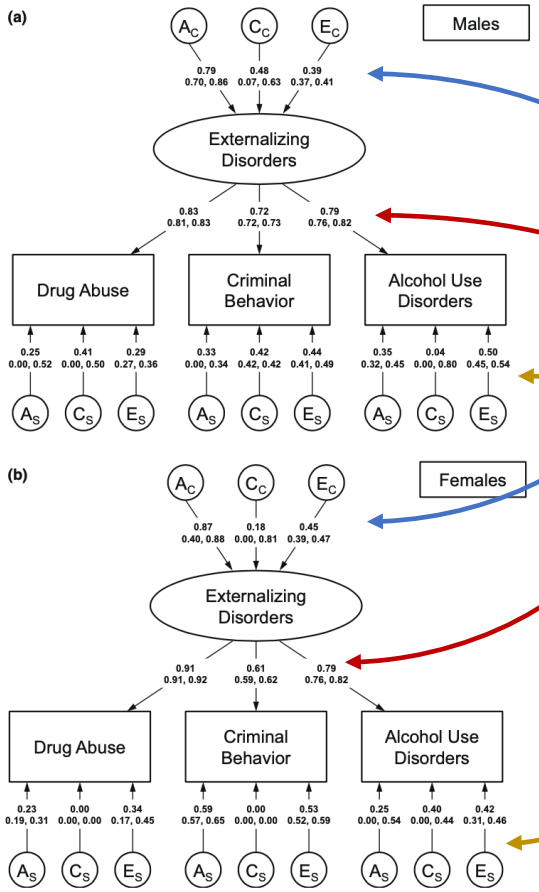
$$\begin{bmatrix} \lambda_1 & 0 \\ \lambda_2 & 0 \\ \lambda_3 & 0 \\ 0 & \lambda_1 \\ 0 & \lambda_2 \\ 0 & \lambda_3 \end{bmatrix} [h \otimes V_a + s \otimes V_c + d \otimes V_e] \begin{bmatrix} \lambda_1 & \lambda_2 & \lambda_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_1 & \lambda_2 & \lambda_3 \end{bmatrix}$$

$$+$$

$$h \otimes \begin{bmatrix} S_{1a} & 0 & 0 \\ 0 & S_{2a} & 0 \\ 0 & 0 & S_{3a} \end{bmatrix} + s \otimes \begin{bmatrix} S_{1c} & 0 & 0 \\ 0 & S_{2c} & 0 \\ 0 & 0 & S_{3c} \end{bmatrix} + d \otimes \begin{bmatrix} S_{1e} & 0 & 0 \\ 0 & S_{2e} & 0 \\ 0 & 0 & S_{3e} \end{bmatrix}$$

Residuals Model
Variance Decomposition

Common Pathway Model in Action



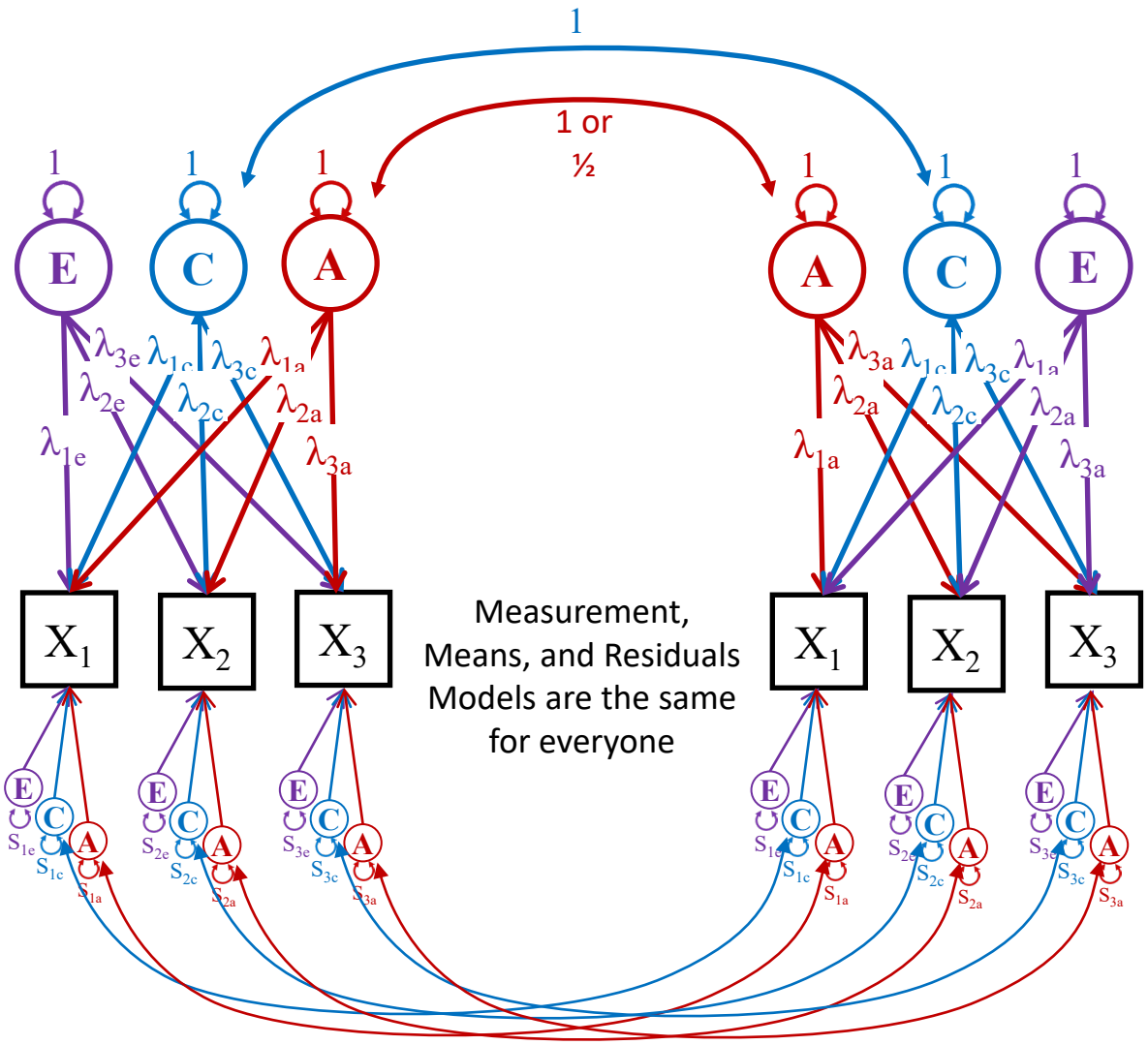
A common pathway model with quantitative but not qualitative sex effects fit best with twin resemblance for the latent liability to externalizing syndromes due to both genetic and shared environmental factors. Heritability of the liability was higher in females (76 vs. 62 %) while shared environmental influences were considerably stronger in males (23 vs. 3 %). In both sexes, this latent liability was most strongly indexed by DA and least by CB. All three syndromes had specific genetic influences (especially CB and AUD in males, and CB in females) and specific shared environmental effects (especially DA and CB in males, and AUD in females). For DA, CB and AUD in men, and DA and AUD in women, at least 75 % of the genetic risk arose through the common factor. The best fit model assumed that genetic and environmental influences on these externalizing syndromes in males and females were the same.

A Swedish Population-Based Multivariate Twin Study of Externalizing Disorders

Kenneth S. Kendler[1,2,3] · Sara Larsson Lönn[4] · Hermine H. Maes[1,3] · Paul Lichtenstein[5] · Jan Sundquist[4,6] · Kristina Sundquist[4,6]

Behav Genet (2016) 46:183–192
DOI 10.1007/s10519-015-9741-7

Independent Pathway Model

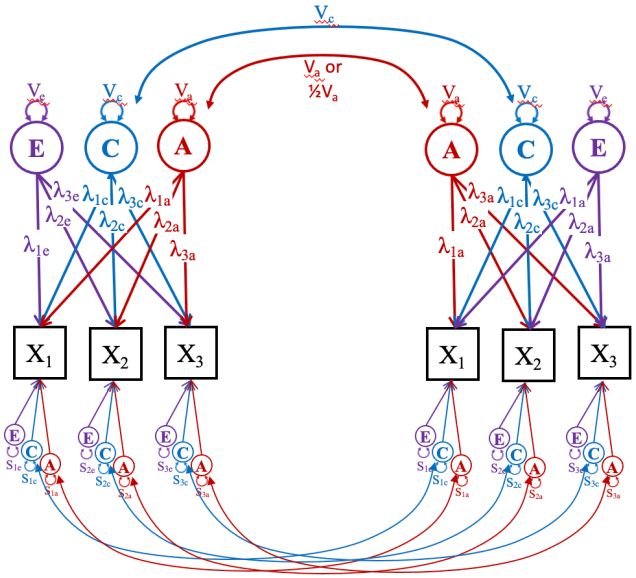


Measurement, Means, and Residuals Models are the same for everyone

Separate Measurement Models for each Variance Component

Residuals Model Variance Decomposition

Independent Pathway Model Algebra



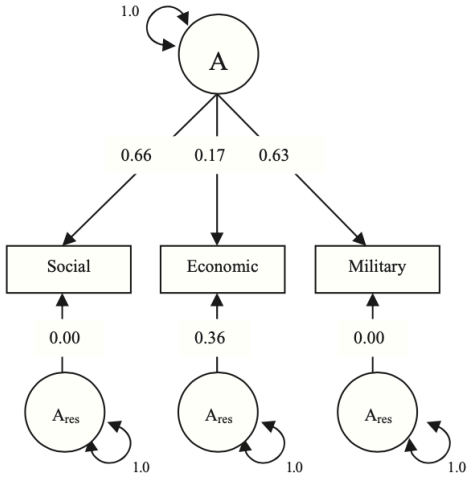
$$\Sigma_\theta = \Lambda\Psi\Lambda' + \Delta$$

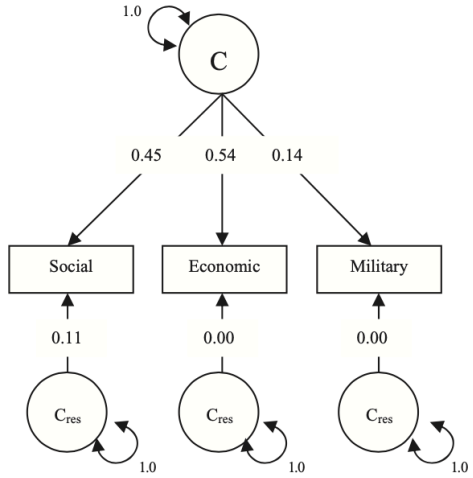
$$\Delta = h \otimes S_a + s \otimes S_c + d \otimes S_e$$

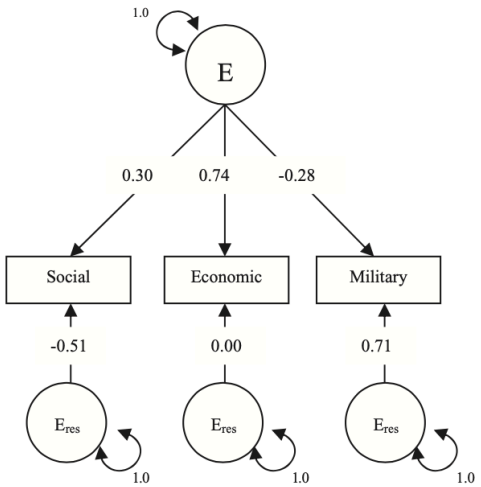
$$\Lambda\Psi\Lambda' = h \otimes \Lambda_a\Lambda'_a + s \otimes \Lambda_c\Lambda'_c + d \otimes \Lambda_e\Lambda'_e$$

Independent Pathway Model in Action

Bottom-up process that leads to genetic covariation between political attitude dimensions







Top-down processes that leads to shared environmental covariation between political attitude dimensions

Disentangling the Importance of Psychological Predispositions and Social Constructions in the Organization of American Political Ideology

Brad Verhulst
Virginia Commonwealth University

Peter K. Hatemi
University of Sydney, Australia

Lindon J. Eaves
Virginia Commonwealth University

Political Psychology, Vol. 33, No. 3, 2012
doi: 10.1111/j.1467-9221.2012.00882.x

Saturated Multivariate Twin Models

- In multivariate twin analyses, saturated models are essential comparison models to gauge the fit of hypothesis driven models (i.e. Common and Independent pathway models).

- Saturated models freely estimate all possible covariances and therefore should fit the data as accurately as possible.

- Therefore, comparing hypothesis driven models to the saturate model allows us to test how much worse the hypothesis driven models fits the observed data.

Saturated Multivariate Twin Models

Direct
Symmetric
Matrix

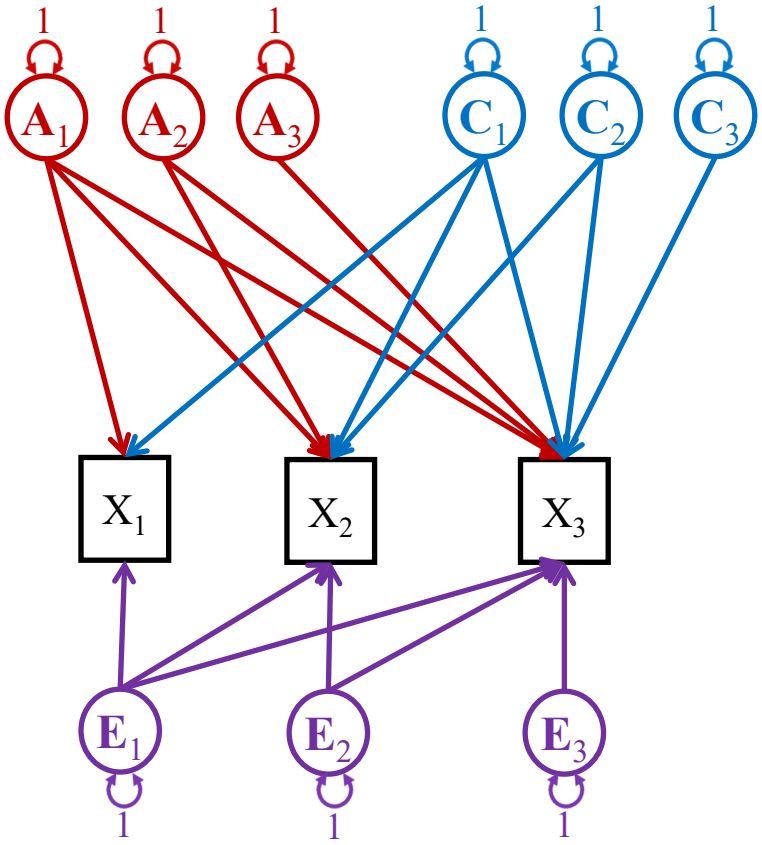
$$\begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \sigma_{1,1} & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} & \sigma_{1,2} & \sigma_{2,2} & \sigma_{2,3} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_3^2 & \sigma_{1,3} & \sigma_{2,3} & \sigma_{3,3} \\ \sigma_{1,1} & \sigma_{1,2} & \sigma_{1,3} & \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{1,2} & \sigma_{2,2} & \sigma_{2,3} & \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_{3,3} & \sigma_{1,3} & \sigma_{2,3} & \sigma_3^2 \end{bmatrix}$$

$$\begin{bmatrix} V_{a1} & a_{1,2} & a_{1,3} & a_{1,1} & a_{1,2} & a_{1,3} \\ a_{1,2} & V_{a2} & a_{2,3} & a_{1,2} & a_{2,2} & a_{2,3} \\ a_{1,3} & a_{2,3} & V_{a3} & a_{1,3} & a_{2,3} & a_{3,3} \\ a_{1,1} & a_{1,2} & a_{1,3} & V_{a1} & a_{1,2} & a_{1,3} \\ a_{1,2} & a_{2,2} & a_{2,3} & a_{1,2} & V_{a2} & a_{2,3} \\ a_{1,3} & a_{2,3} & a_{3,3} & a_{1,3} & a_{2,3} & V_{a3} \end{bmatrix} + \begin{bmatrix} V_{c1} & c_{1,2} & c_{1,3} & c_{1,1} & c_{1,2} & c_{1,3} \\ c_{1,2} & V_{c2} & c_{2,3} & c_{1,2} & c_{2,2} & c_{2,3} \\ c_{1,3} & c_{2,3} & V_{c3} & c_{1,3} & c_{2,3} & c_{3,3} \\ c_{1,1} & c_{1,2} & c_{1,3} & V_{c1} & c_{1,2} & c_{1,3} \\ c_{1,2} & c_{2,2} & c_{2,3} & c_{1,2} & V_{c2} & c_{2,3} \\ c_{1,3} & c_{2,3} & c_{3,3} & c_{1,3} & c_{2,3} & V_{c3} \end{bmatrix} + \begin{bmatrix} V_{e1} & e_{1,2} & e_{1,3} & e_{1,1} & e_{1,2} & e_{1,3} \\ e_{1,2} & V_{e2} & e_{2,3} & e_{1,2} & e_{2,2} & e_{2,3} \\ e_{1,3} & e_{2,3} & V_{e3} & e_{1,3} & e_{2,3} & e_{3,3} \\ e_{1,1} & e_{1,2} & e_{1,3} & V_{e1} & e_{1,2} & e_{1,3} \\ e_{1,2} & e_{2,2} & e_{2,3} & e_{1,2} & V_{e2} & e_{2,3} \\ e_{1,3} & e_{2,3} & e_{3,3} & e_{1,3} & e_{2,3} & V_{e3} \end{bmatrix}$$

$$SatCov = h \otimes V_a + s \otimes V_c + d \otimes V_e$$

Saturated Multivariate Twin Models

Cholesky Decomposition



$$SatCov = h \otimes AA' + s \otimes CC' + d \otimes EE'$$

$$h \otimes \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ 0 & a_{22} & a_{32} \\ 0 & 0 & a_{33} \end{bmatrix} +$$

$$s \otimes \begin{bmatrix} c_{11} & 0 & 0 \\ c_{21} & c_{22} & 0 \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{bmatrix} c_{11} & c_{21} & c_{31} \\ 0 & c_{22} & c_{32} \\ 0 & 0 & c_{33} \end{bmatrix} +$$

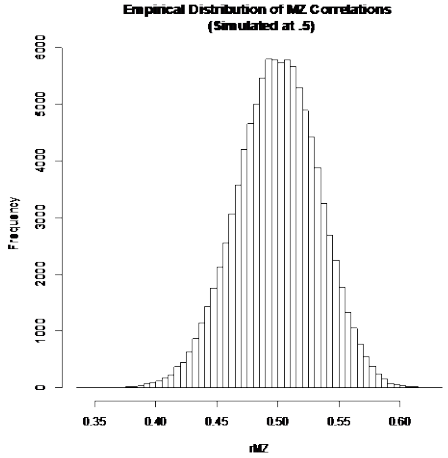
$$d \otimes \begin{bmatrix} e_{11} & 0 & 0 \\ e_{21} & e_{22} & 0 \\ e_{31} & e_{32} & e_{33} \end{bmatrix} \begin{bmatrix} e_{11} & e_{21} & e_{31} \\ 0 & e_{22} & e_{32} \\ 0 & 0 & e_{33} \end{bmatrix}$$

Cholesky Problems

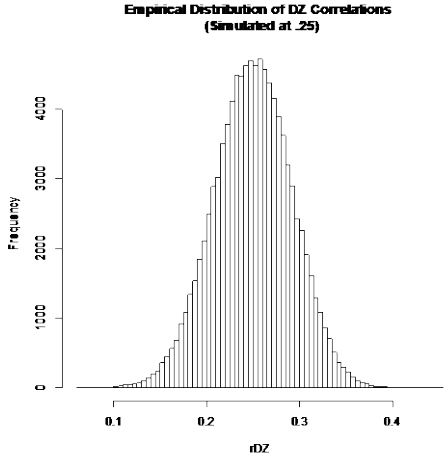
- The Cholesky Decomposition implicitly constrains all variance parameters to be positive
  - (Seems sensible)
- This constraint truncates the distribution of the variance parameters under the null distribution
  - (Meaning the p-values are wrong)
- Under the null, the test statistics is distributed as a mixture distribution of 0, $\chi^2(1)$, … $\chi^2(k)$.

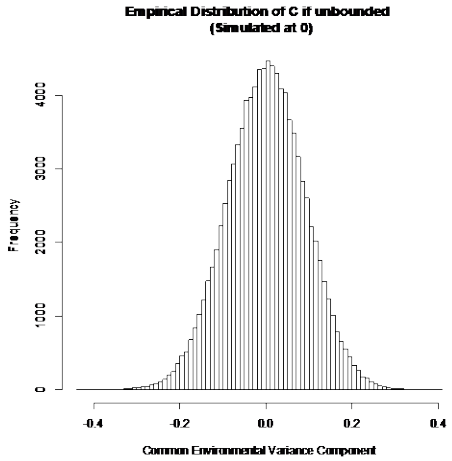
Intuition behind the Problem

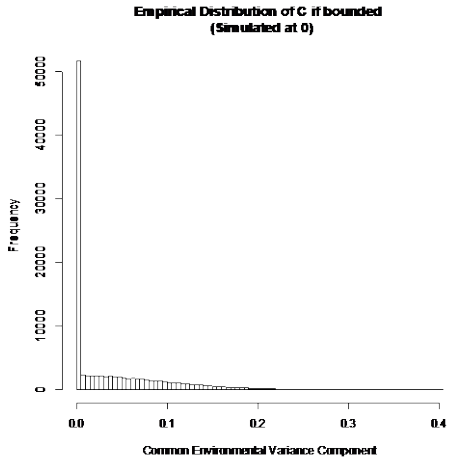
rMZ = .50



rDZ = .25







If rMZ is .50 and rDZ is .25:
Va = 0.5
Vc = 0
Ve = 0.5

In repeated sampling, sometimes rMZ will be slightly overestimated and rDZ will be slightly underestimated.

If so, C will be negative

Other times, rMZ will be slightly underestimated and rDZ will be slightly overestimated.

If so, C will be positive

Model Evaluation

Requirements for the Likelihood Ratio Test (LRT) :

- estimated from the same data (preferably using ML)
- a restricted model is nested in a more saturated model
- restricted must have fewer fitted parameters (more df) than the saturated model

$$LRT = -2ln\left(\frac{L_{simple}(\hat{\theta})}{L_{complex}(\hat{\theta})}\right)$$

Nesting: A reduced model is nested in a saturated model if the reduced model is a special case of the saturated model.

- A parameter is set to 0 (or some other value)
- Two parameters are equated

It is possible to have a complex nesting structure

- A is nested in B which is nested in C

Assumption Testing and Model Fitting

More Saturated

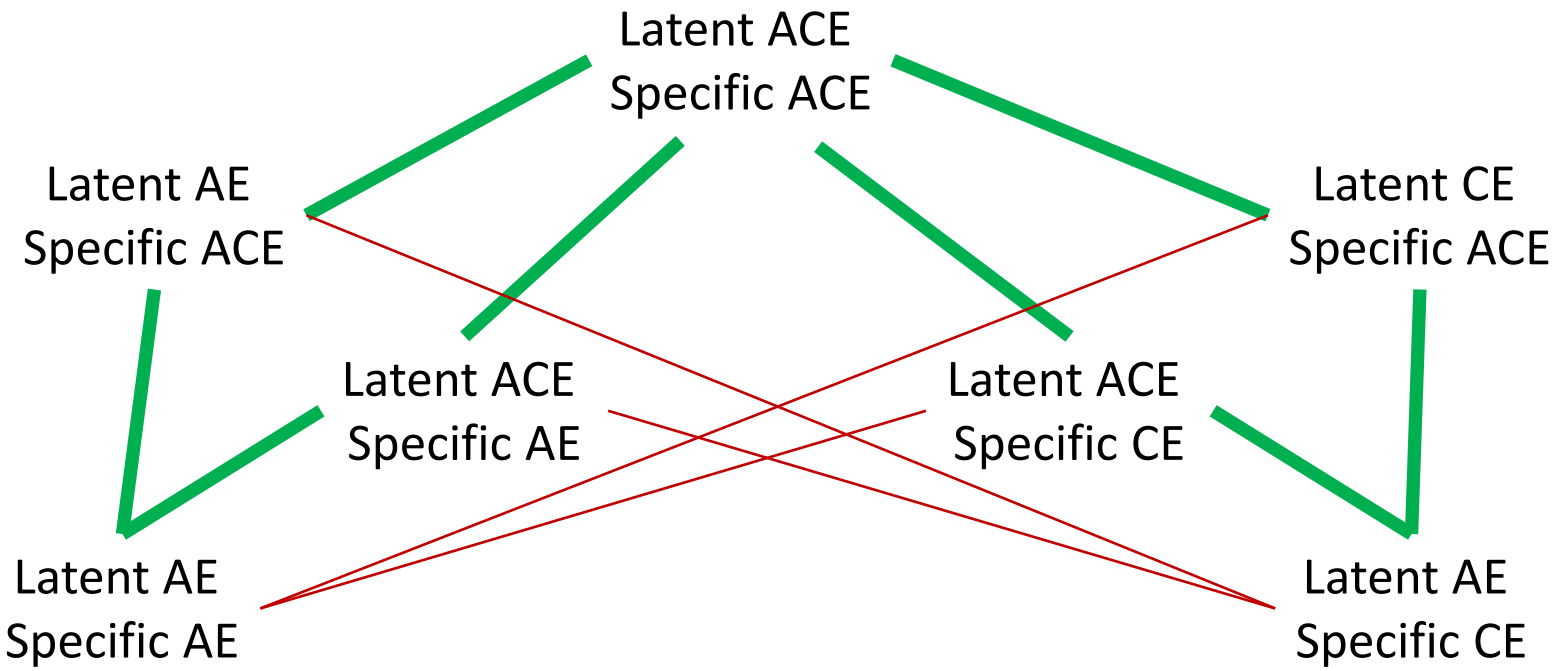
Less Saturated

- Completely Saturated Model
  - All possible variances, covariances and means are freely estimated
- Saturated Model
  - Equating cross-twin cross trait covariances within zygosity (e.g. $r_{t1v1\text{-}t2v2} = r_{t1v2\text{-}t2v1}$)
- Equal means and variances across twin order
- Equal means and variances across Zygosity

Under certain regularity conditions, the twice the negative log of the likelihoods between the saturated and restricted model will be distributed as a χ^2 with the degrees of freedom equal to the difference in the number of parameters estimated in each model

Model Fitting in MV Twin Studies

Common Pathway Model



Order Dependent Results

Caution: The interpretation of your “Best” model may depend on the order that you conducted your model comparisons.

After fitting my full CPM, I dropped the specific Cs in my model

Then I couldn’t drop the latent C!

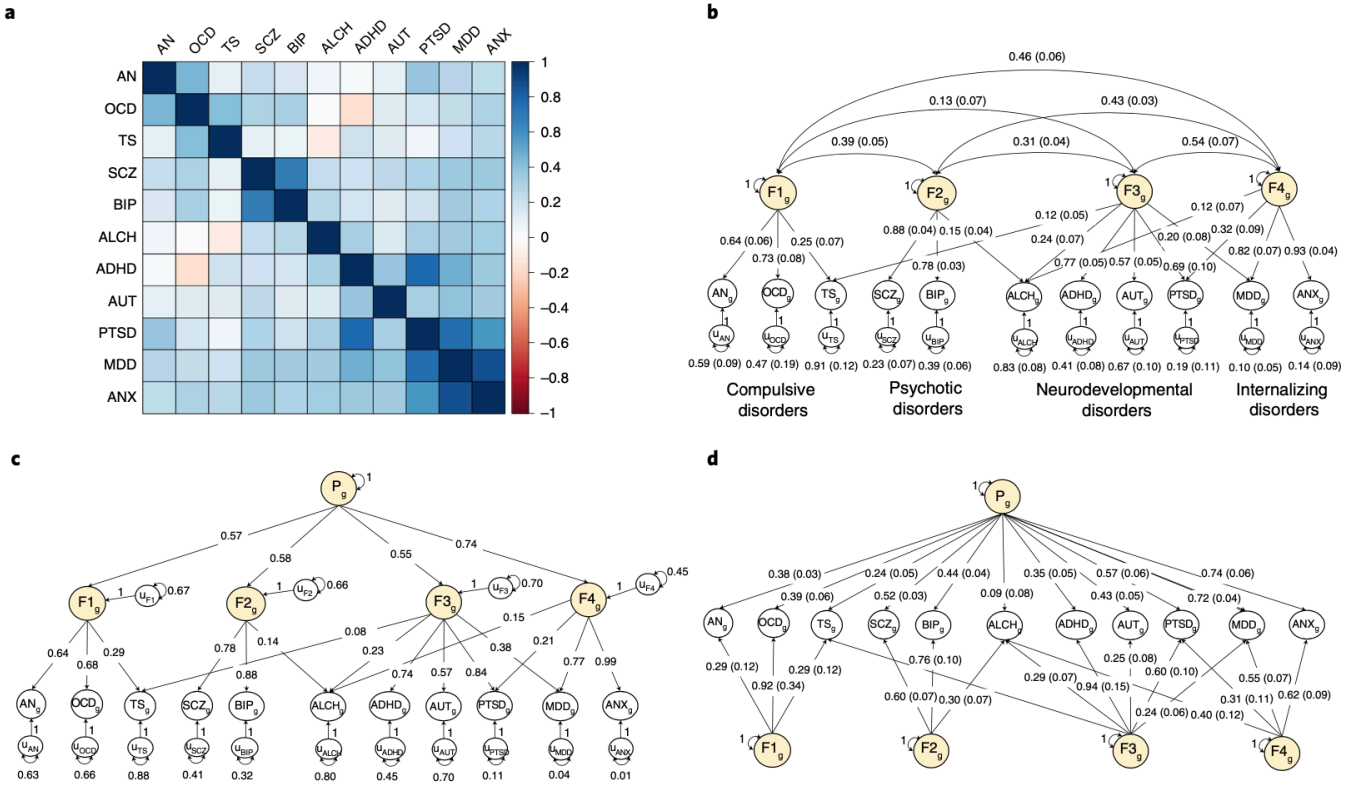
Therefore, shared environmental factors affect the latent factor

After fitting my full CPM I dropped the latent C in my model

Then I couldn’t drop the specific C!

Therefore, shared environmental factors affect the item residuals

Related Examples from the Literature



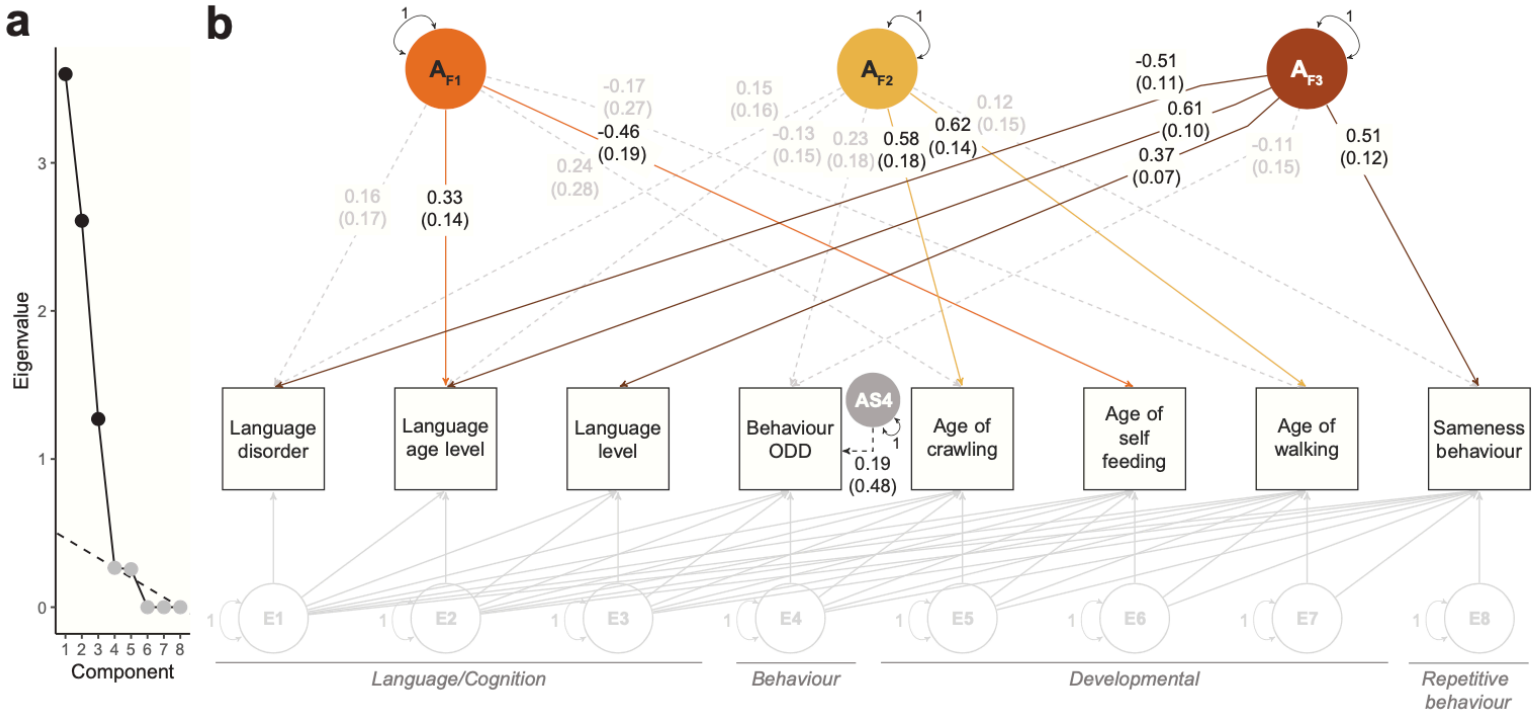
Genetic architecture of 11 major psychiatric disorders at biobehavioral, functional genomic and molecular genetic levels of analysis

Andrew D. Grotzinger[1,2], Travis T. Mallard[3], Wonuola A. Akingbuwa[4,5], Hill F. Ip[4], Mark J. Adams[6], Cathryn M. Lewis[7,8], Andrew M. McIntosh[6], Jakob Grove[9,10,11,12],

NATURE GENETICS | VOL 54 | MAY 2022 | 548–559 |

GenomicSEM

Related Examples from the Literature



GRM-SEM

Structural models of genome-wide covariance identify multiple common dimensions in autism

Lucía de Hoyos[1], Maria T. Barendse[1,2], Fenja Schlag[1], Mariolein M. J. van Donkelaar[1], Ellen Verhoef[1], Chin Yang Shapland[3,4],

Nature Communications | (2024)15:1770

Multivariate Twin Practical

- Qualitrics Link:
  https://qimr.az1.qualtrics.com/jfe/form/SV_0dmHAEyYb4bPbIG

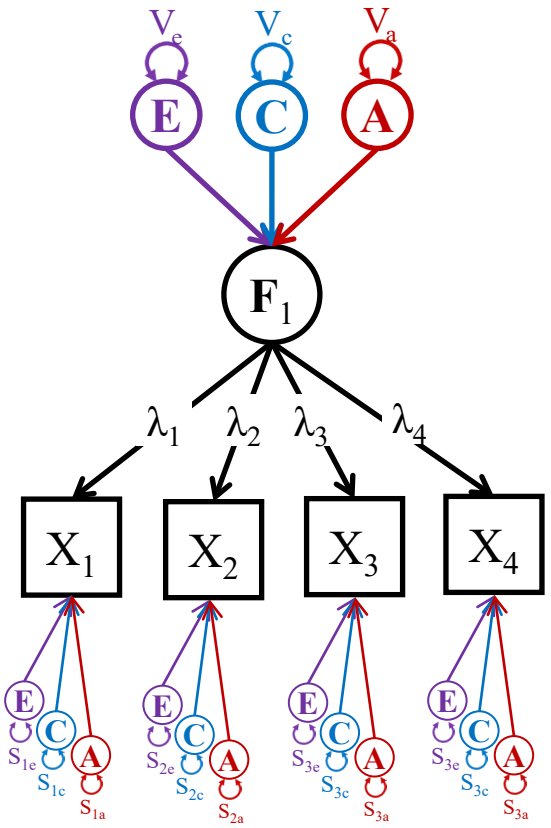
- Read in data
  - There are two sets of datasets - only select one

- Run the assumption testing models
  - There are 4 sequentially more restricted models

- Run the CPM

- Run the IPM

Dataset 1: Common Pathway Model



	Obs	Sim
Va	0.50	0.50
Vc	0	0
Ve	0.50	0.50

	Obs	Sim
λ_1	0.80	0.80
λ_2	0.75	0.75
λ_3	0.70	0.70
λ_4	0.65	0.65

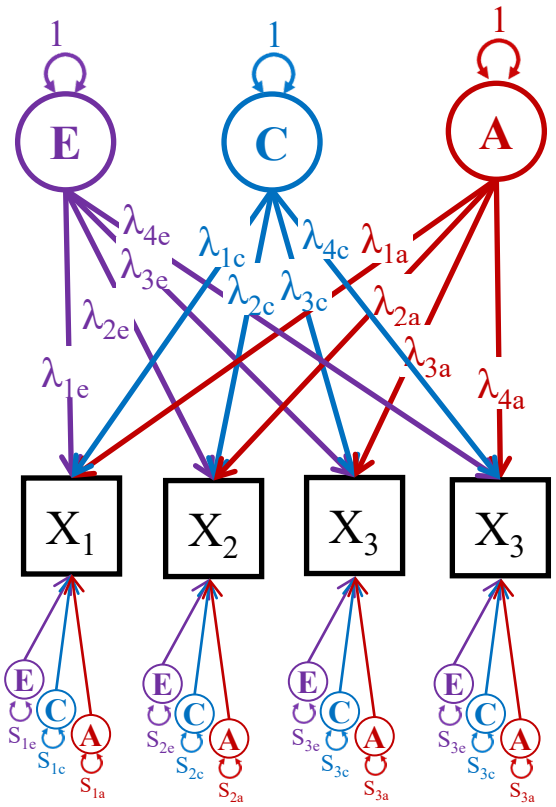
Obs

	As	Cs	Es
x_1	0.07	0.11	0.18
x_2	0.13	0.09	0.22
x_3	0.25	0	0.25
x_4	0	0.29	0.29

Sim

	As	Cs	Es
x_1	.072	.108	.180
x_2	.131	.088	.219
x_3	.255	0	.255
x_4	0	.289	.289

Independent Pathway Model



Obs

	A	C	E
λ_1	-.50	-.20	-.50
λ_2	-.40	-.25	-.55
λ_3	-.45	-.15	-.45
λ_4	-.30	-.20	-.60

Sim

	A	C	E
λ_1	.50	.20	.50
λ_2	.40	.25	.55
λ_3	.45	.15	.45
λ_4	.30	.20	.60

Obs

	As	Cs	Es
x_1	0.09	0.14	0.23
x_2	0.14	0.09	0.24
x_3	0.29	0	0.29
x_4	0	0.25	0.25

Sim

	As	Cs	Es
x_1	.092	.138	.230
x_2	.143	.095	.238
x_3	.286	0	.286
x_4	0	.255	.255

Acknowledgements