The Direct Symmetric Matrix Approach to Fitting Twin Models

Brad Verhulst
Texas A&M University

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

ORIGINAL RESEARCH

CrossMark

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]

Phenotypic Variance

We typically discuss phenotypic variance as the sum of independent variance components:

$$V_P = V_A + V_C + V_E$$

Variance can easily be calculated with the standard equation:

$$\sigma_P^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

This encourages us to thing about the variance components as squared quantities :

$$\sigma_P^2 = \sigma_A^2 + \sigma_C^2 + \sigma_E^2$$

When we should be thinking about them differences between MZ and DZ twin groups:

$$V_P = 1 = 2(r_{MZ} - r_{DZ}) + 2r_{DZ} - r_{MZ} + 1 - (Va + Vc)$$

Who Cares?

- In most cases the estimates of A, C, and E will be very similar
- The problem is that a discrepancy emerges between the nominal and empirical Type I Error rate
- This is overwhelmingly in a conservative direction (meaning that you are likely to fail to reject the null hypothesis of genetic or common environmental variation)

Type I Error Rate in the Classical Twin Design

- Type I Error Rate: Probability of rejecting the null hypothesis by chance
  - Alpha level (0.05)
  - In repeated sampling, if a parameter is truly null, we would expect to find a significant parameter 5% of the time due to chance alone

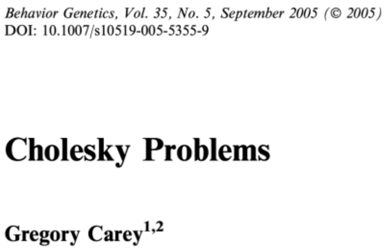
- Significance tests for ACE models are traditionally done using a Likelihood Ratio Test (LRT)
  - $\mathbf{\Delta}\text{LL} = -2\ \text{LL}_{(\text{restricted})} \quad - \quad -2\ \text{LL}_{(\text{Full})}$
  - Under certain regularity conditions, $\mathbf{\Delta}$LL is distributed as $\boldsymbol{\chi}^2$(df).
  - Twin Studies violate these regularity conditions by implicitly imposing lower boundaries on the variance components estimates

Violations of the Assumptions for the LRT

A Note on the Asymptotic Distribution of Likelihood Ratio Tests to Test Variance Components

Peter M. Visscher[1,2]

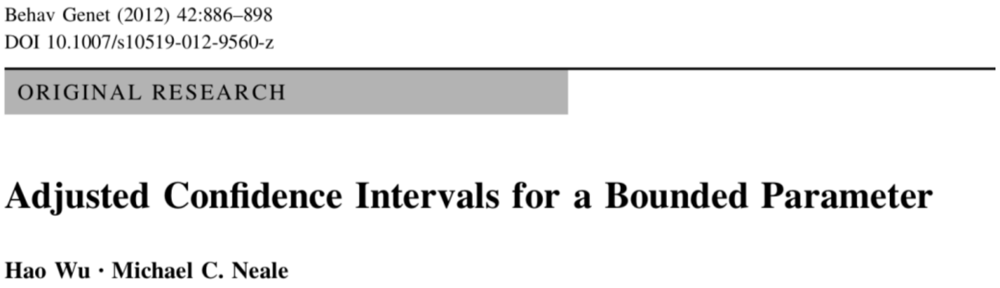
[1] Queensland Institute of Medical Research, Brisbane, Australia
[2] Institute of Evolutionary Biology, University of Edinburgh, United Kingdom



Behavior Genetics, Vol. 35, No. 5, September 2005 (© 2005)
DOI: 10.1007/s10519-005-5355-9

Cholesky Problems

Gregory Carey[1,2]



Behav Genet (2012) 42:886–898
DOI 10.1007/s10519-012-9560-z

ORIGINAL RESEARCH

Adjusted Confidence Intervals for a Bounded Parameter

Hao Wu · Michael C. Neale

Behavior Genetics, Vol. 36, No. 2, March 2006 (© 2006)
DOI: 10.1007/s10519-005-9034-7

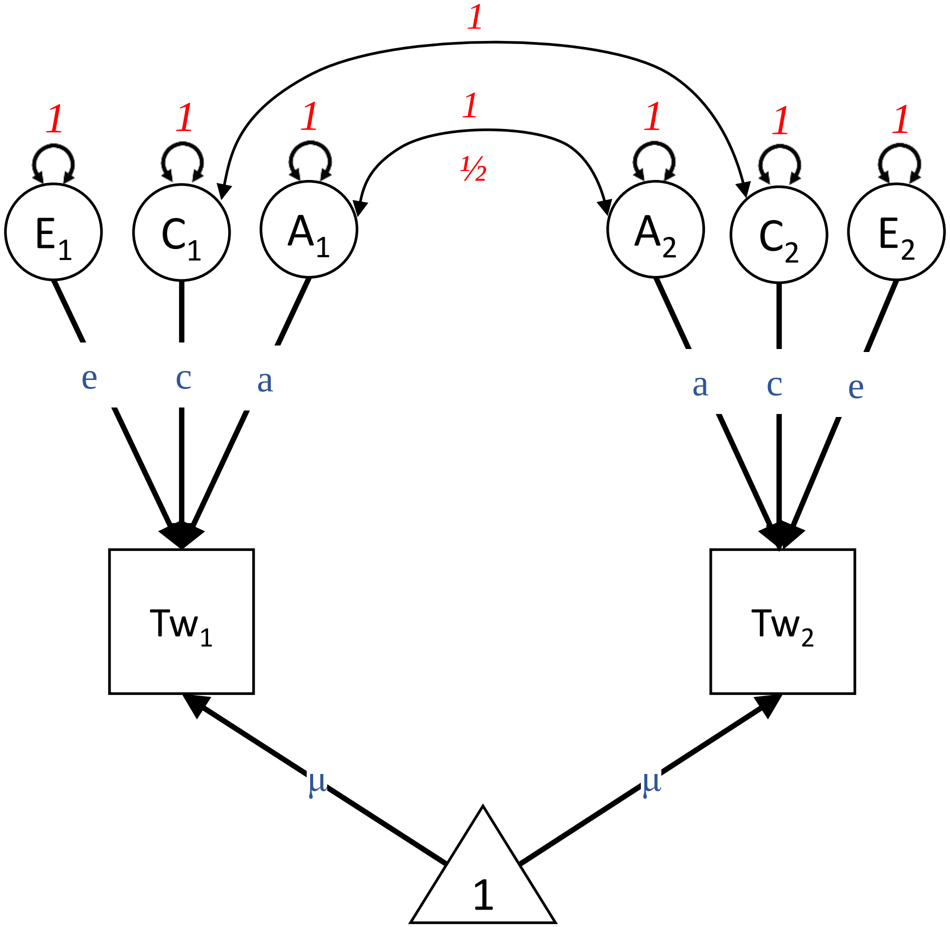
Likelihood Ratio Tests in Behavioral Genetics: Problems and Solutions

Annica Dominicus,[1,2,6] Anders Skrondal,[3,4] Håkon K. Gjessing,[4] Nancy L. Pedersen,[2,5] and Juni Palmgren[1,2]

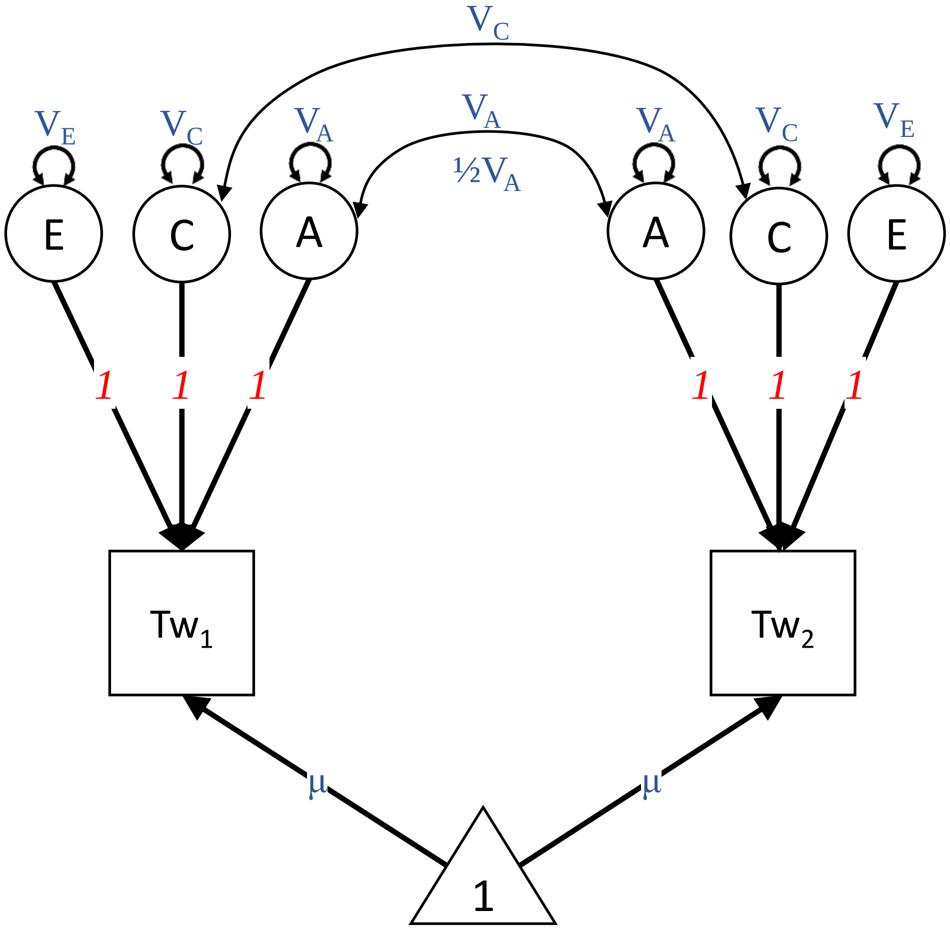
In standard twin models, under the null hypothesis that a variance component is zero, this test is distributed as a 50:50 mixture distribution of 0 and $\chi^2(1)$.

The Classical Twin Design

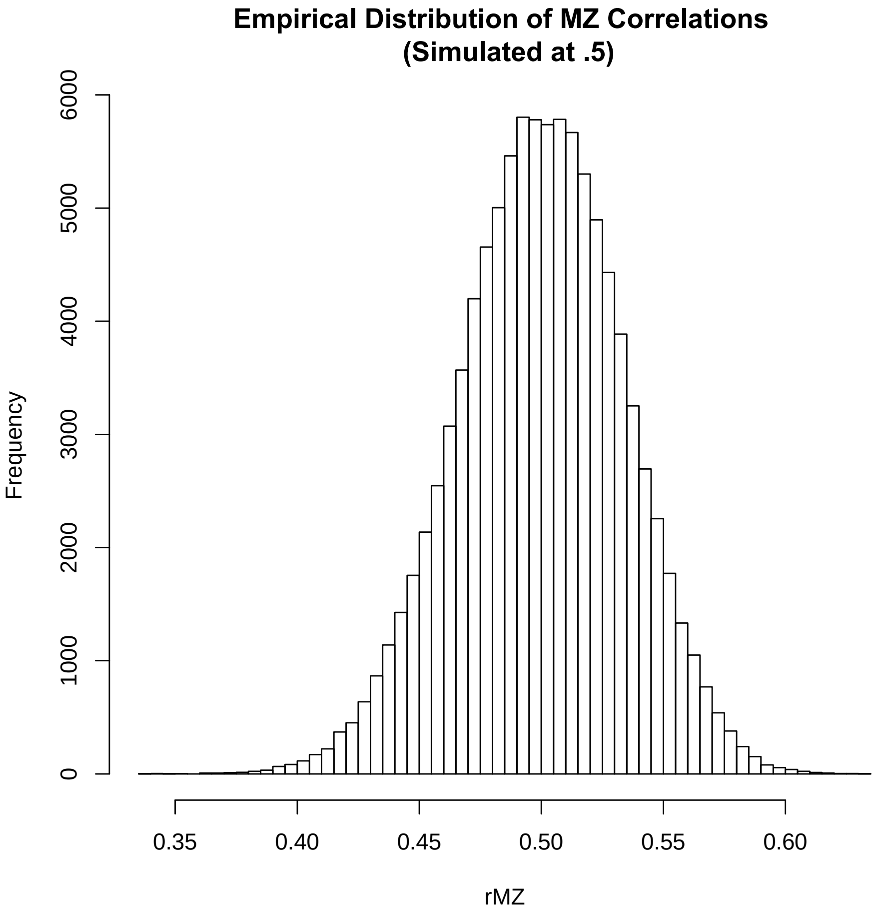
(Common Specification)

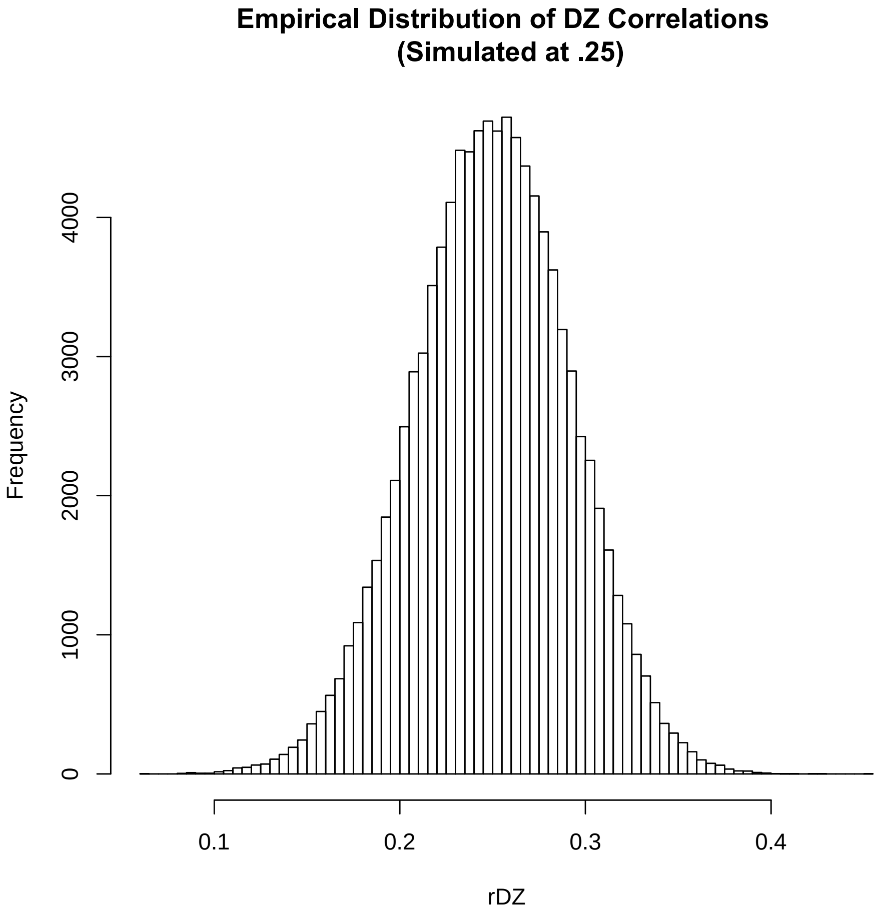


The Direct Symmetric Specification

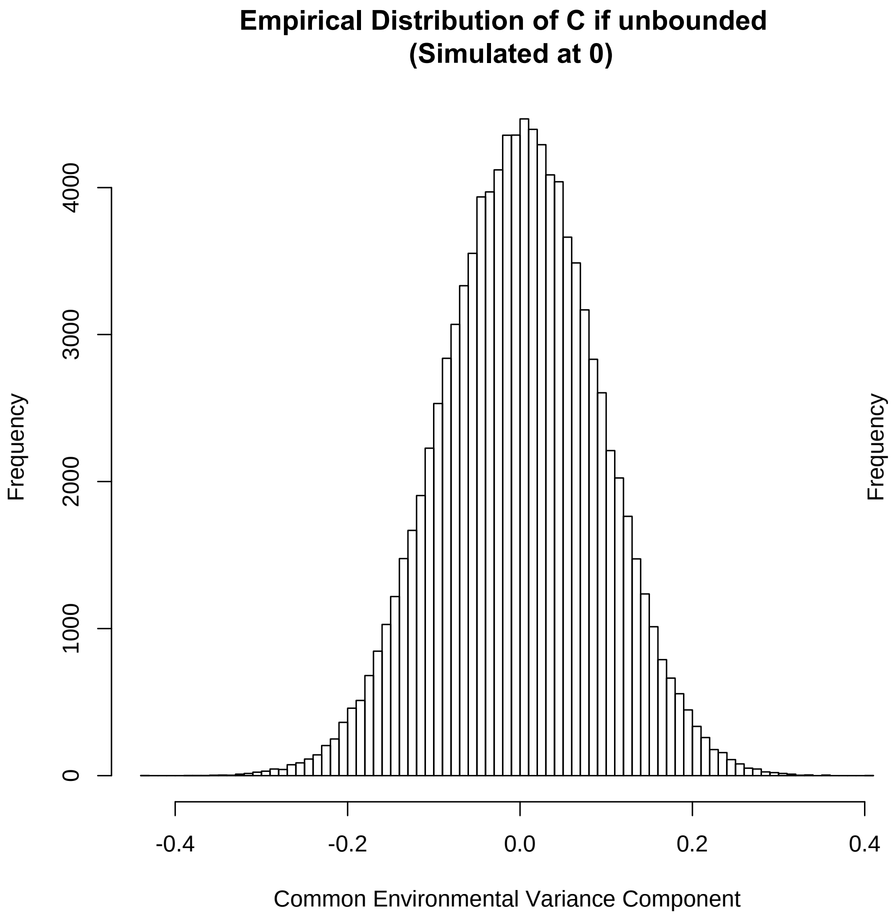


Intuition behind the Problem

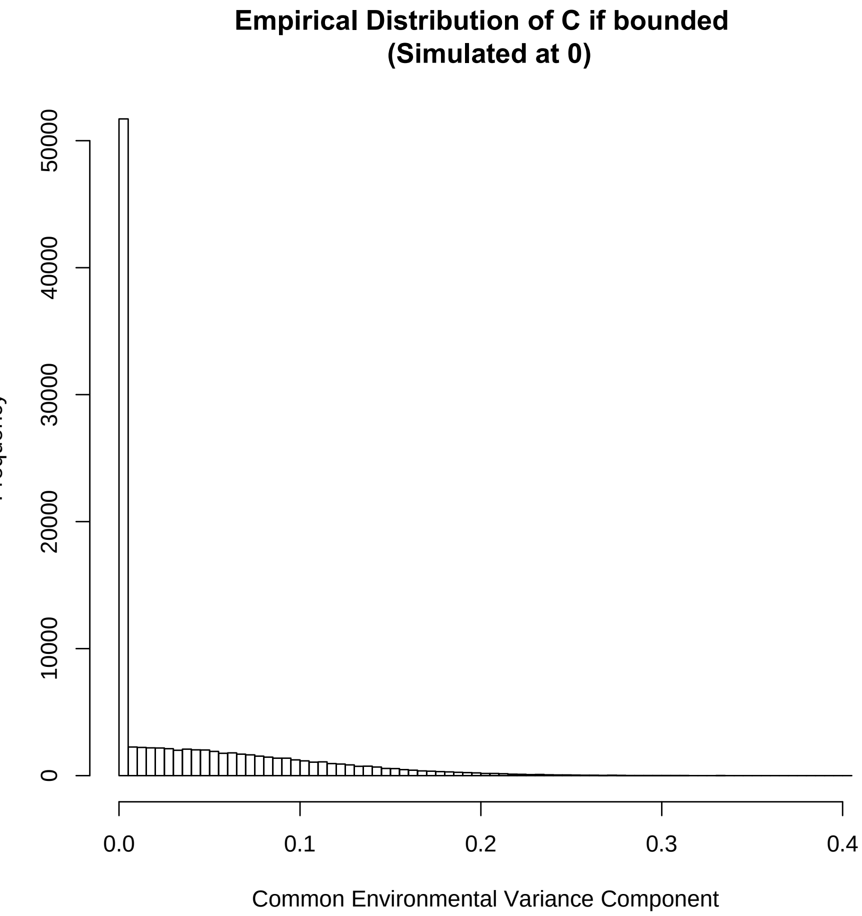




Intuition behind the Problem



Empirical Distribution of C if unbounded (Simulated at 0)



Empirical Distribution of C if bounded (Simulated at 0)

Adjusting the p-value in the Univariate Twin Model

With one variable per twin, to correct the p-value all you need to do is divide the p-value that OpenMx gives by 2 (or test at p = .10)

For bivariate twin analyses, Wu & Neale (2012) provide the solution

For more than two variables, it gets tricky

But how bad is the problem?

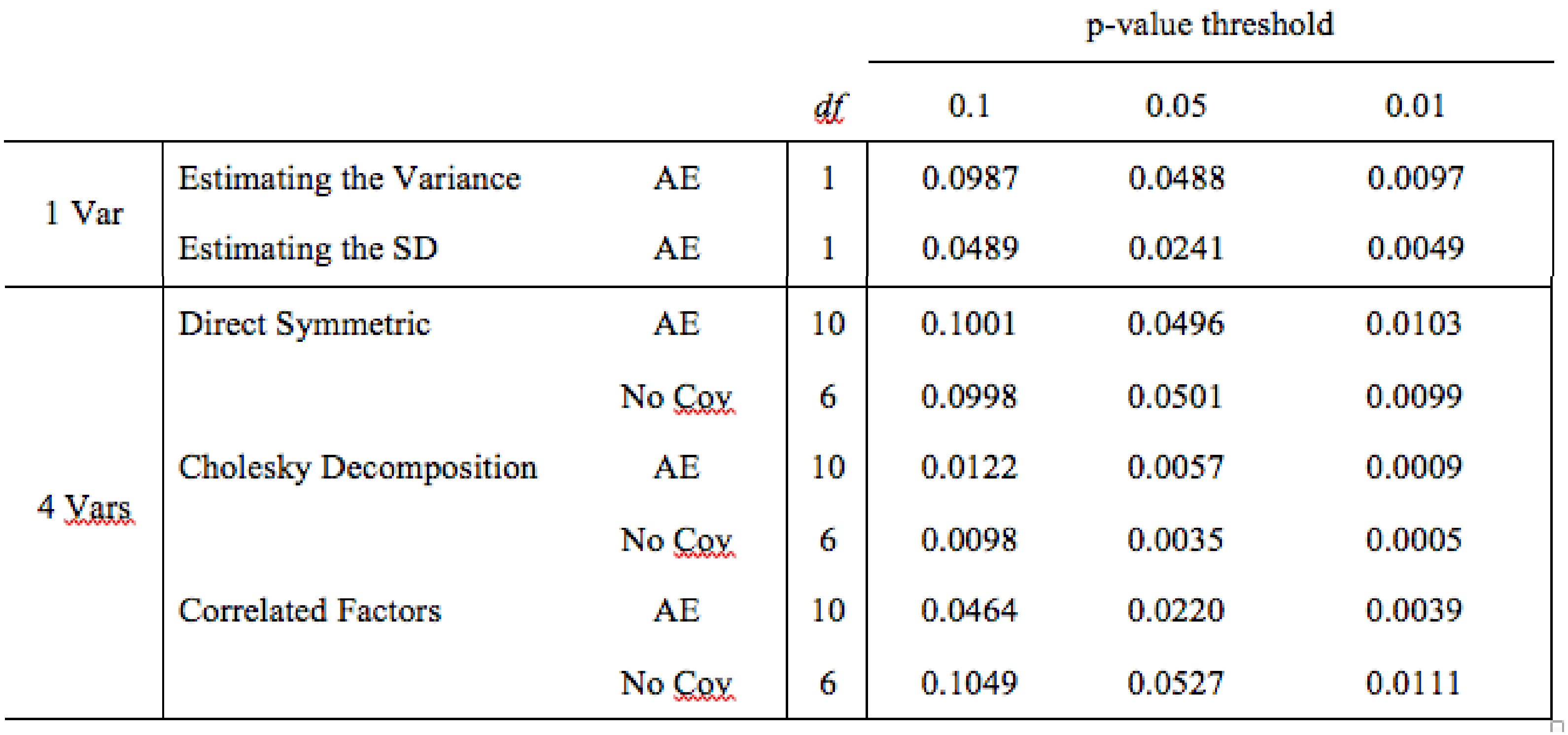
- We conducted 2 simulation studies to examine the impact of the implicit boundaries on the Type I Error rate and parameter bias

Study 1: Simulate a basic multivariate twin model where C was 0 (1 – 4 phenotypes)

- Fit the model using:
    - Cholesky Decomposition
    - Correlated Factors Model
    - Direct Symmetric Model

Study 2: Simulate a Common Pathway model

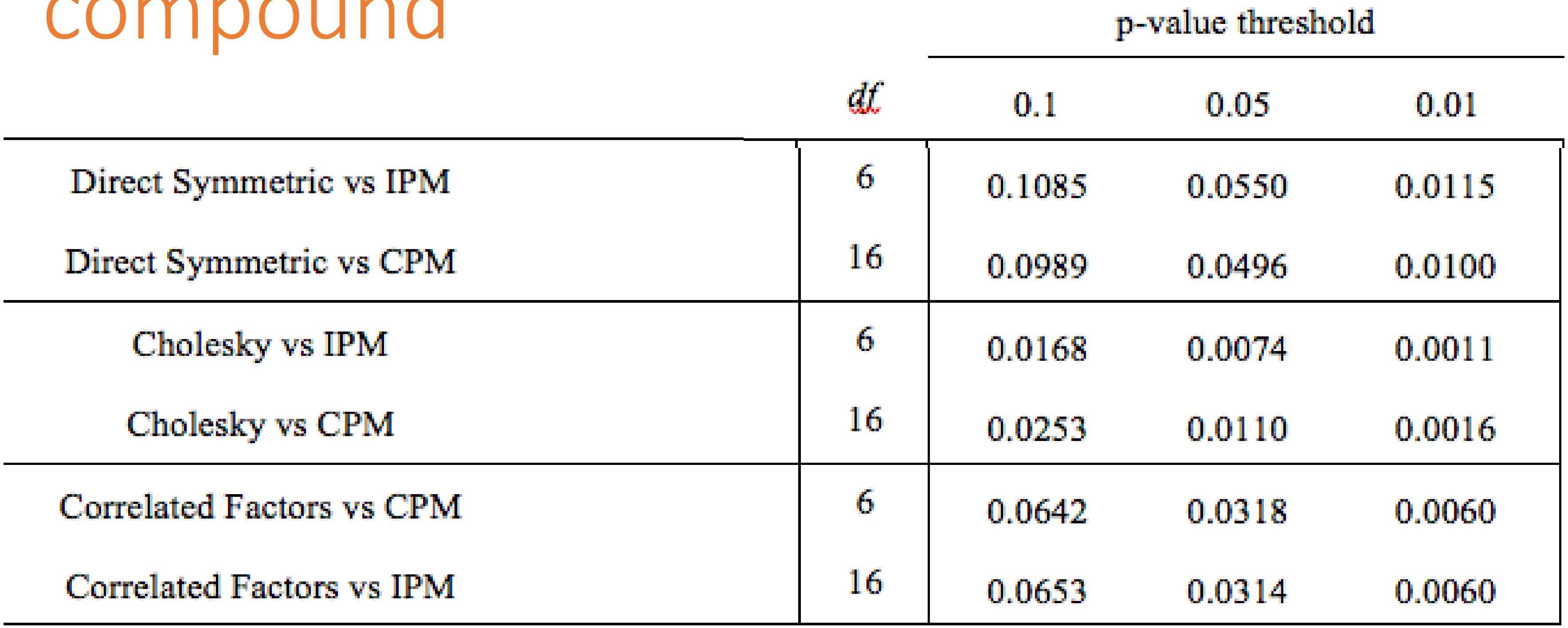
- Fit the model using:
    - Cholesky Decomposition
    - Correlated Factors Model
    - Direct Symmetric Model
    - Independent Pathway Model (IPM)
    - Common Pathway Model (CPM)



				p-value threshold		
			df	0.1	0.05	0.01
1 Var	Estimating the Variance	AE	1	0.0987	0.0488	0.0097
	Estimating the SD	AE	1	0.0489	0.0241	0.0049
4 Vars	Direct Symmetric	AE	10	0.1001	0.0496	0.0103
		No Cov	6	0.0998	0.0501	0.0099
	Cholesky Decomposition	AE	10	0.0122	0.0057	0.0009
		No Cov	6	0.0098	0.0035	0.0005
	Correlated Factors	AE	10	0.0464	0.0220	0.0039
		No Cov	6	0.1049	0.0527	0.0111

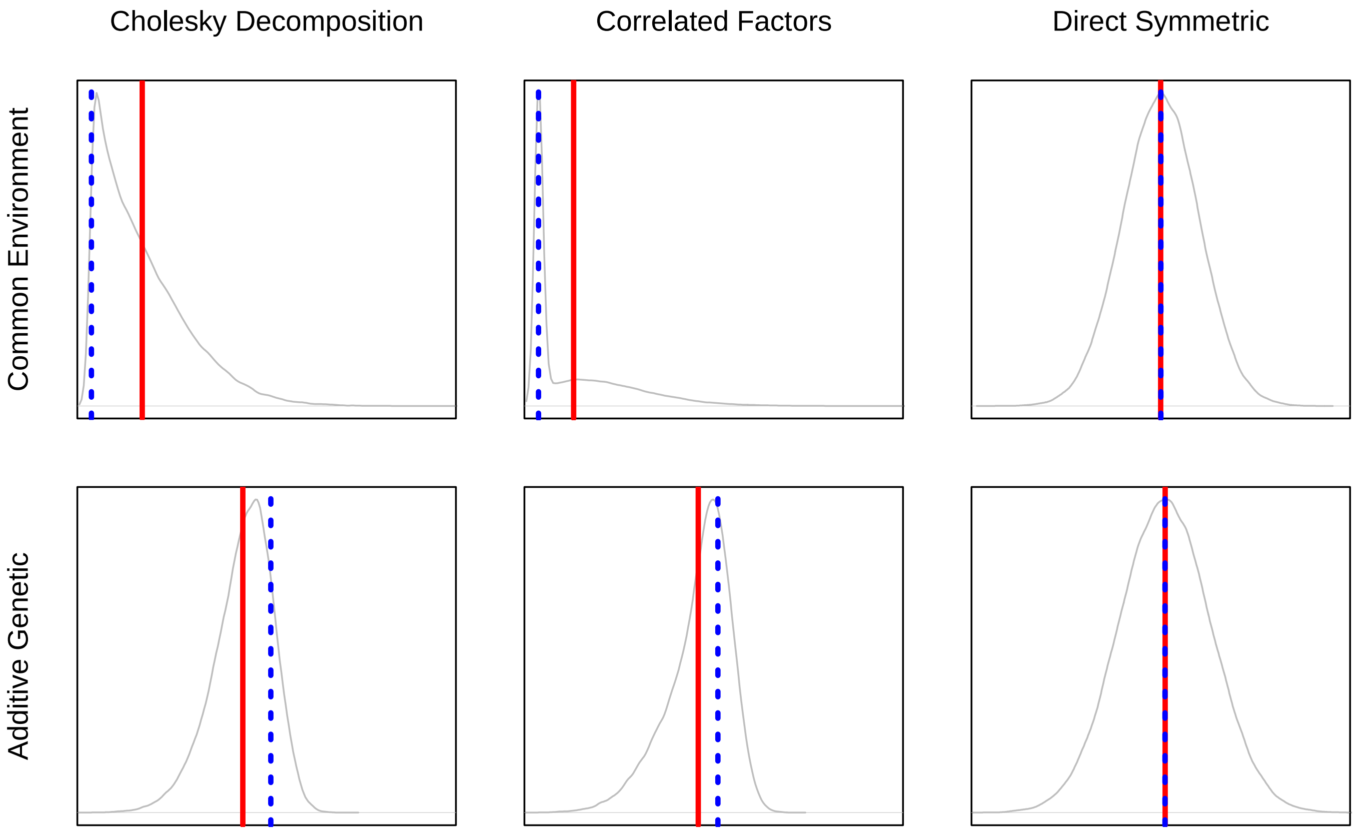
As the number of variables in the model increase, the Numerical and Theoretical Type I Error rates diverge

When the Cholesky and Correlated Factors models are used to compare hypothesis driven models, the deviations in the Type I Error Rate compound

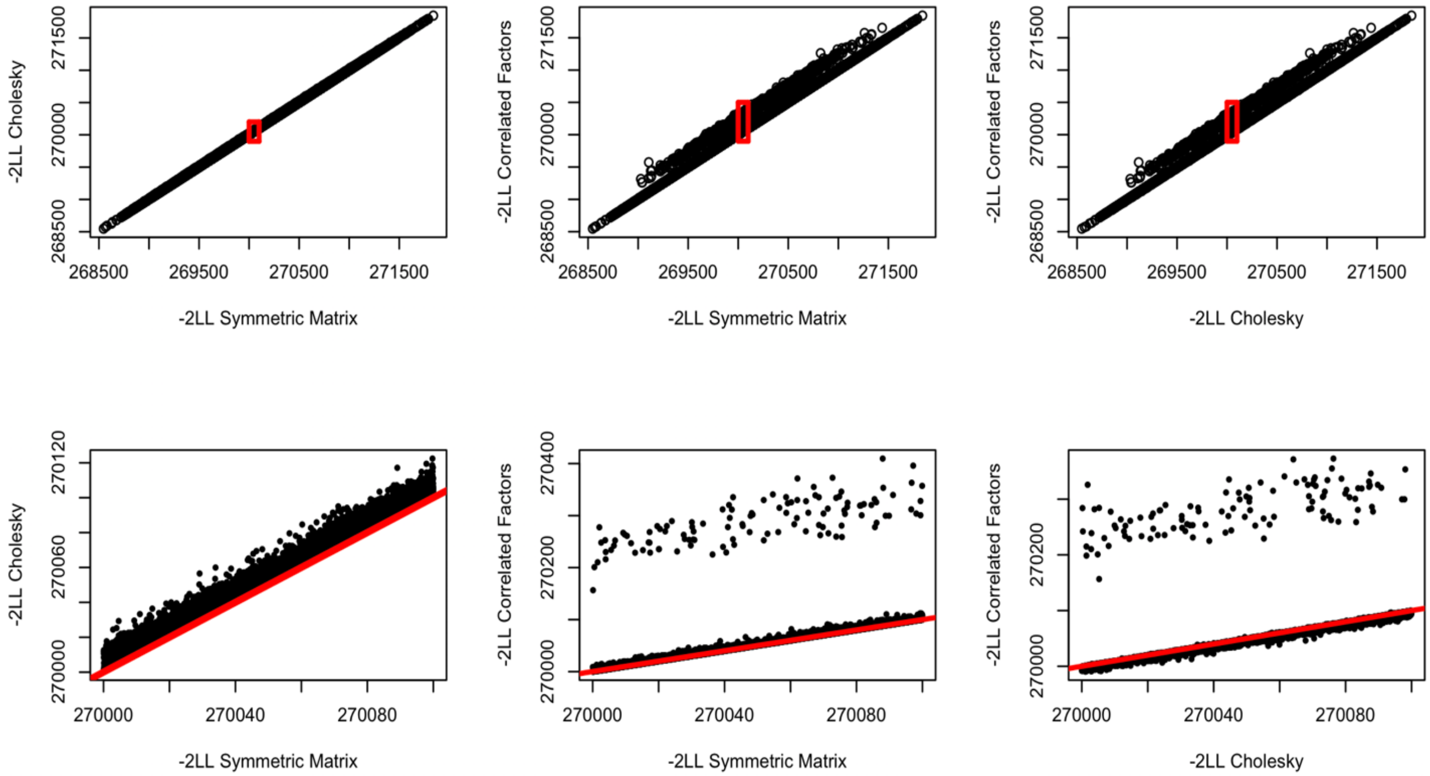


		p-value threshold		
	df	0.1	0.05	0.01
Direct Symmetric vs IPM	6	0.1085	0.0550	0.0115
Direct Symmetric vs CPM	16	0.0989	0.0496	0.0100
Cholesky vs IPM	6	0.0168	0.0074	0.0011
Cholesky vs CPM	16	0.0253	0.0110	0.0016
Correlated Factors vs CPM	6	0.0642	0.0318	0.0060
Correlated Factors vs IPM	16	0.0653	0.0314	0.0060

The lower bound of zero also causes bias in the parameter estimates



The Direct Symmetric Approach will ALWAYS fit as well or better than the other approaches



Caveat Emptor

- The Unbounded Variance Components may take negative (non-sensical) values
  - These values will likely suggest some sort of model mis-specification
  - I.E., You fit an ACE model when you should have fit an ADE model
- GxE Models can be very difficult to specify with the DSM approach

Conclusions

Implicit and explicit boundaries lead to a deviation from the expected Type I error rate and can induce bias in the parameter estimates under the null hypothesis.

The fact that the Type I error rate is conservative implies that the Type II error rate is inflated.