

Simulating and Modeling Extended Twin Family Data



Matthew C. Keller

Sarah E. Medland

Hermine Maes

(also, Lindon Eaves, Mike Neale,
Pete Hatemi, Laramie Duncan)

Outline

- I. □ Motivation for using extended pedigrees
 - Briefly introduce the NTFD, Stealth, & Cascade models
- II. Simulation using GeneEvolve
 - Practical looking at changes in genetic variance across time
- III. Use GeneEvolve to simulate extended twin family data & run in Mx
 - Practical getting sensitivity analysis of CTD & NTFD

Goals

- □ Understand in general terms the reason for extended twin family designs
- Learn how to use GeneEvolve
- Understand how to derive biases & sampling distributions from simulation

NON-Goal

- Fully understand the logic, path diagrams, and scripting of extended twin family models

Structural Equation Modeling (SEM) in BG

■ SEM is great because...

- Directs focus to effect sizes, not “significance”
- Forces consideration of causes and consequences
- Explicit disclosure of assumptions

■ Potential weakness...

- Parameter reification: “Using the CTD we found that 50% of variation is due to A and 20% to C.”

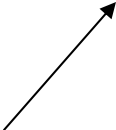
Structural Equation Modeling (SEM) in BG

■ SEM is great because...

- Directs focus to effect sizes, not “significance”
- Forces consideration of causes and consequences
- Explicit disclosure of assumptions

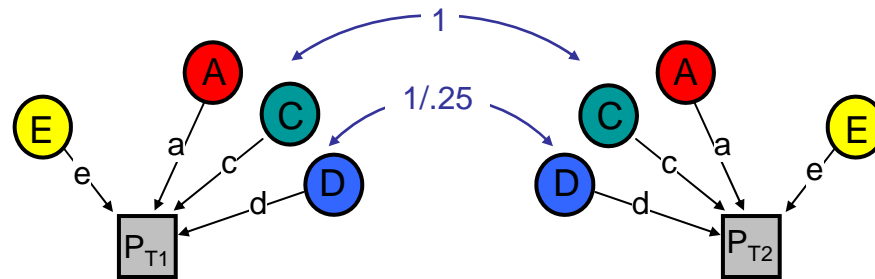
■ Potential weakness...

- Parameter reification: “Using the CTD we found that 50% of variation is due to A and 20% to C.”



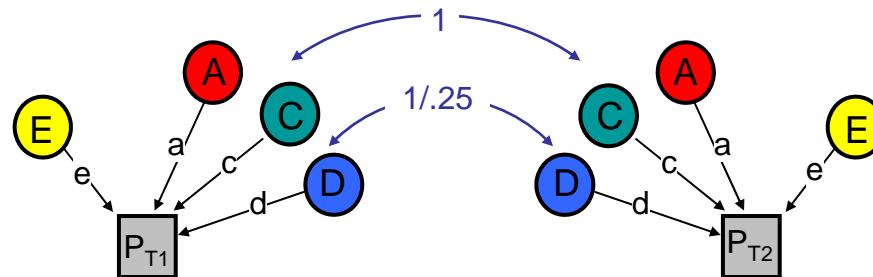
Not necessarily. Only true under assumptions that may often be unmet (e.g., $D=0$) and usually go untested. To the degree assumptions wrong, estimates are biased.

Classical Twin Design (CTD)



Classical Twin Design (CTD)

<u>Assumption</u>	<u>biased up</u>	<u>biased down</u>
Either D or C is zero	A	C & D
No assortative mating	C	D
No A-C covariance	C	D & A



Why can't we estimate A, C & D at same time using twins only?

- Solve the following two equations for A, C, & D:

$$CVmz = A + D + C$$

$$CVdz = 1/2A + 1/4D + C$$

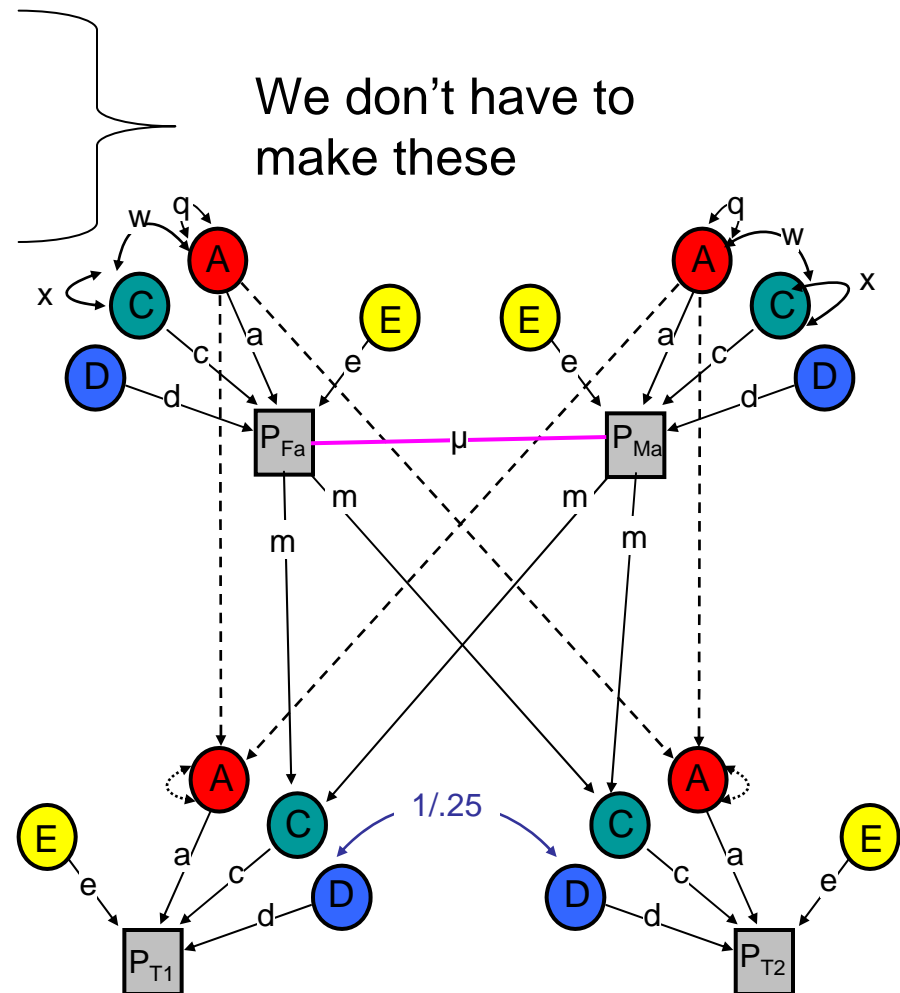
Why does simply setting D or C to zero bias C & D down and A up?

- Information to estimate A comes from the ratio $CV_{mz} : 2 * CV_{dz}$. The closer this ratio is to unity, the higher A is.
- If D & C both exist at the same time, D drives the ratio up, C drives it down. To the degree these effects 'cancel each other out,' it looks like A at the expense of D & C.

Adding parents gets us around these assumptions

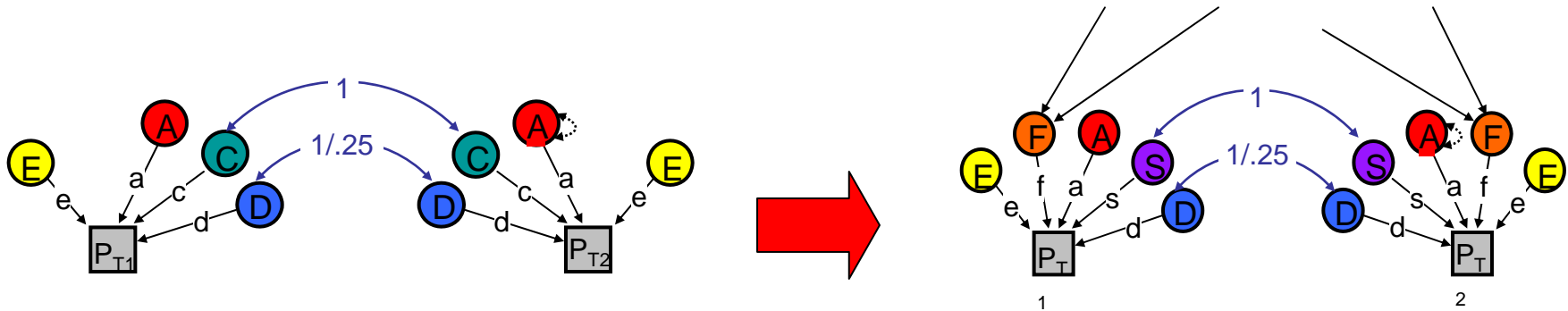
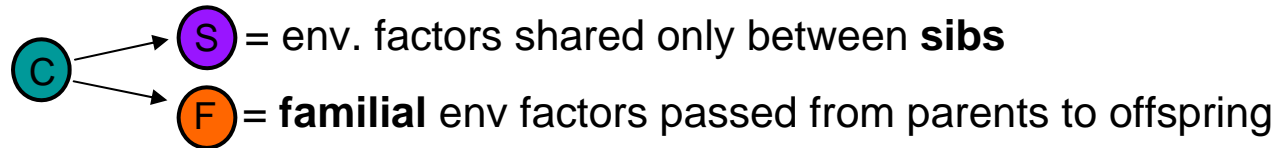
Assumption biased up biased down

- Either D or C is zero
- No assortative mating
- No A-C covariance

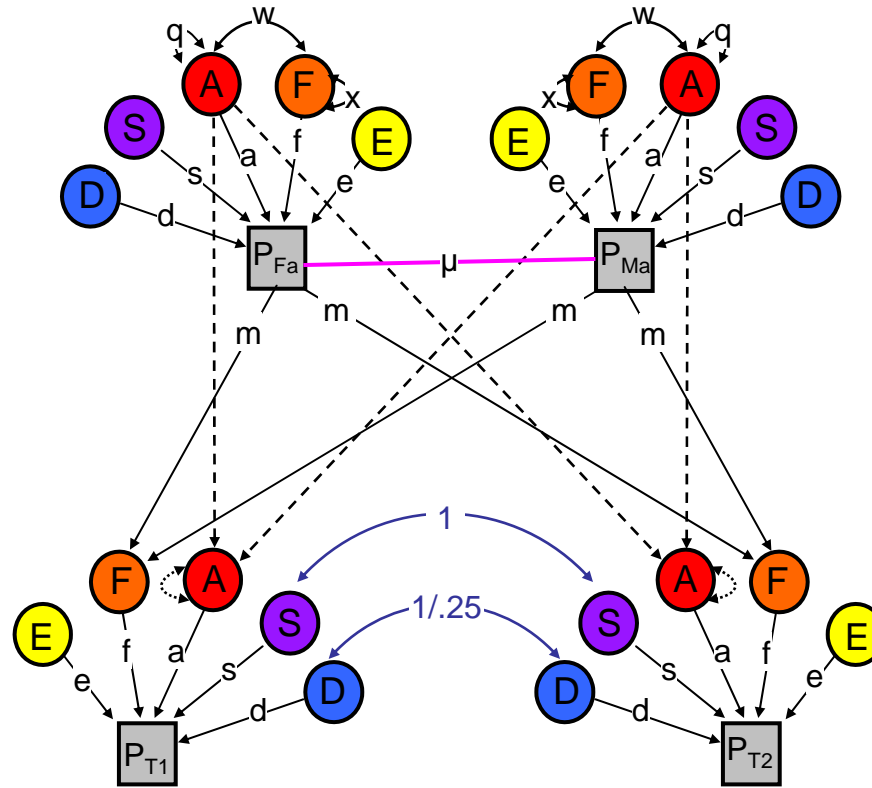


Parents also allow differentiation of C into S & F

With parents, we can break “C” up into:



Nuclear Twin Family Design (NTFD)



Note: m estimated
and f fixed to 1

Assumptions:

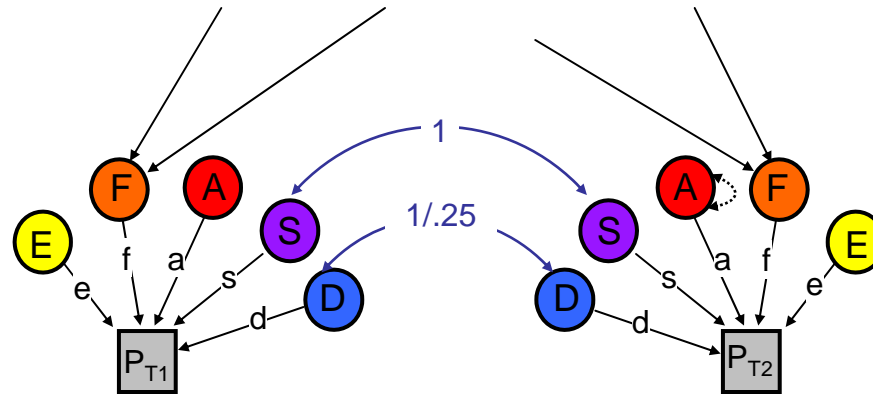
- Only can estimate 3 of 4: A, D, S, and F (bias is variable)
- Assortative mating due to primary phenotypic assortment (bias is variable)

Stealth

- Include twins and their sibs, parents, spouses, and offspring...
 - Gives 17 unique covariances (MZ, DZ, Sib, P-O, Spousal, MZ avunc, DZ avunc, MZ cous, DZ cous, GP-GO, and 7 in-laws)
 - 88 covariances with sex effects

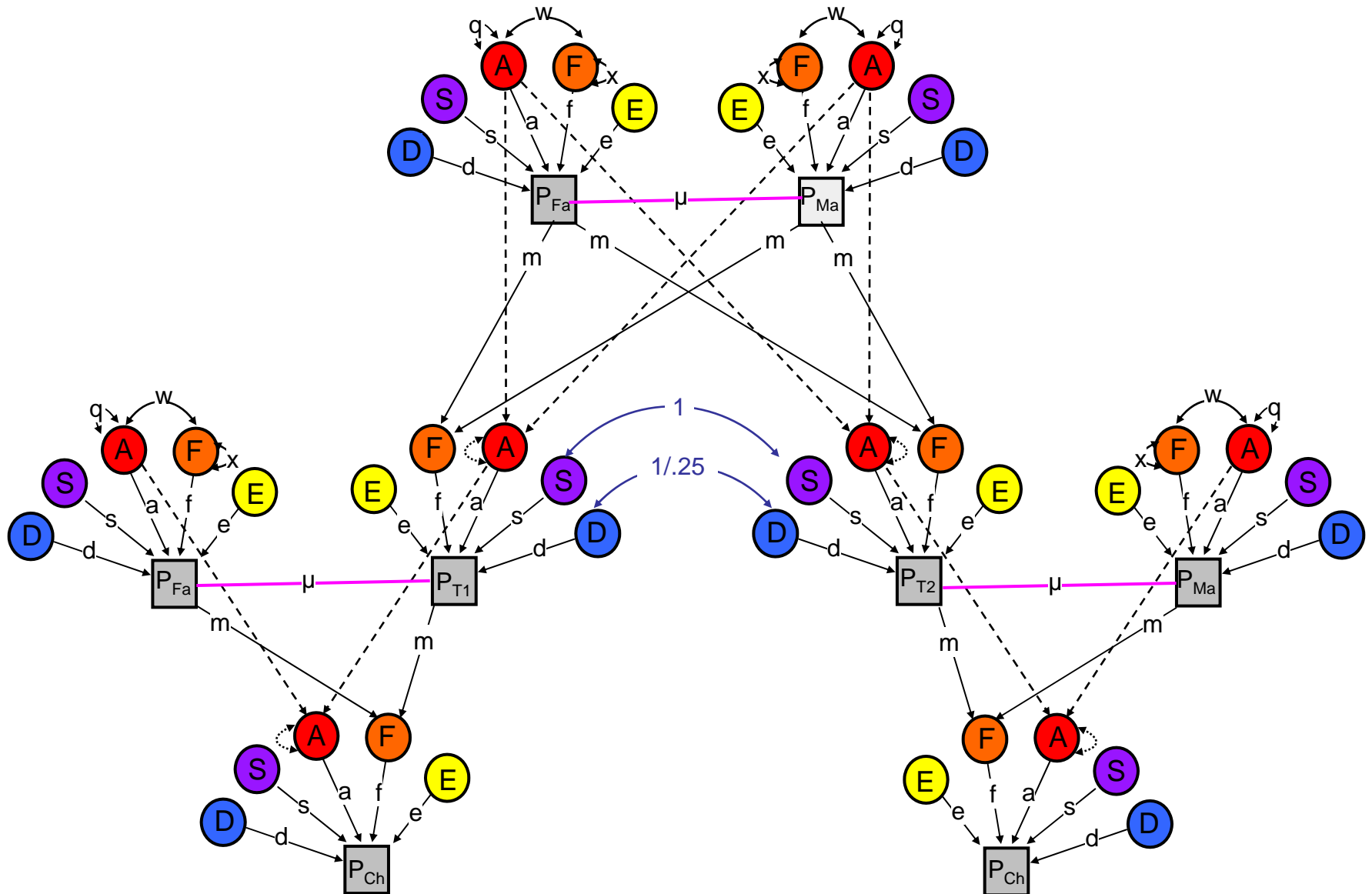
Additional obs. covs with *Stealth* allow estimation of A, S, D, & F

A **S** **F** **D** can be estimated simultaneously



(Remember: we're not just estimating more effects. More importantly, we're reducing the bias in estimated effects!)

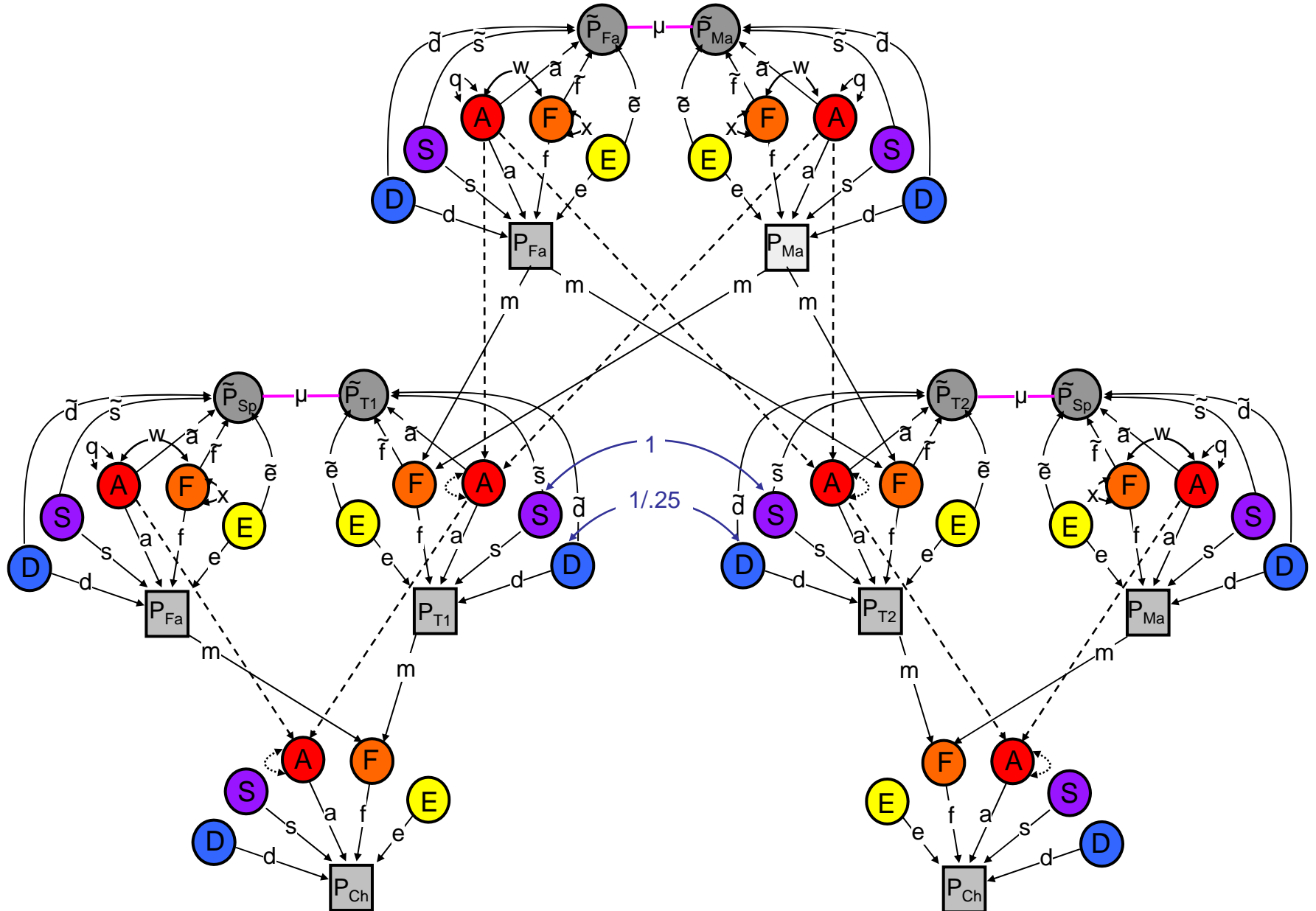
Stealth



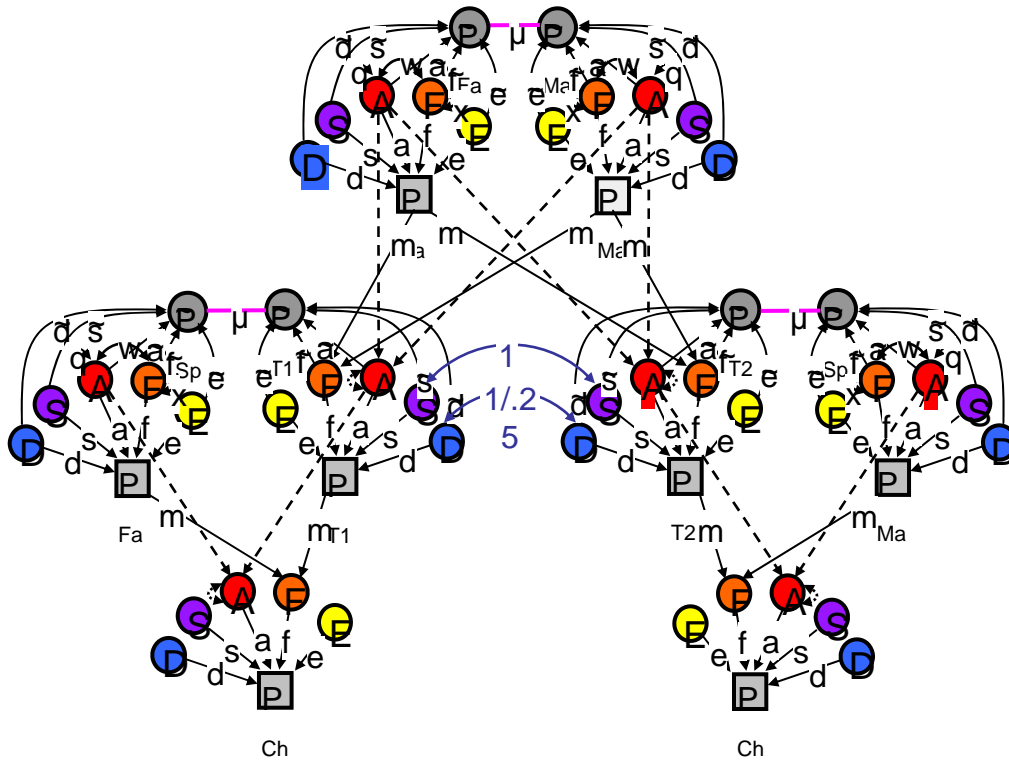
Stealth

- | <u>Assumption</u> | <u>biased up</u> | <u>biased down</u> |
|----------------------------|------------------|--------------------|
| Primary assortative mating | A, D, or F | A, D, or F |
- Primary AM: mates choose each other based on phenotypic similarity
 - Social homogamy: mates choose each other due to environmental similarity (e.g., religion)
 - Convergence: mates become more similar over time

Cascade



Modeling complexity



- The good: Tend to be less biased
- The bad: Easy to make scripting or theoretical mistakes. Are we *really* modeling what we think we are?? How to know?

Part II: Simulating complex models

Simulation provides knowledge about processes that are difficult/impossible to figure out analytically

- Independent check of models:
 - Model validation: Check that your models work as they are supposed to and check the statistical properties of estimates
 - Sensitivity analysis: Check the effect on parameter estimates when assumptions are violated (e.g., different modes of assortative mating, genetic action, etc.)
- Method for predicting complex dynamics in population genetics

Simulation program: GeneEvolve

GeneEvolve

http://matthewckeller.com/html/geneevolve.html

Google Gmail GHome GScholar CU CU-Psych CU-HRC CULink CULib CULearn CUTravel EdMed TA LearR AromaAffy

GeneEvolve

GeneEvolve...

yours to command



- Home
- Biosketch
- Vita
- Publications
- Grad Students
- Program Code
- GeneEvolve
- Plot Indeterminacy
- Mx-R
- R
- Courses
- Links

True or Estimated Standardized Variance Components

Component	First Generation (0)	Final Generation (10)
A	0.32	0.20
A^A	0.10	0.05
D	0.05	0.02
C	0.35	0.20
E	0.15	0.10
Age	0.20	0.20
Age^A	0.20	0.20
CoV	0.15	0.15
%VT	0.60	0.60
AM	0.45	0.45

Change in Variance Components Across Generations

Component	0	10	20	30	40	50
V(P)	0.1	0.1	0.1	0.1	0.1	0.1
V(A)	0.21	0.21	0.21	0.21	0.21	0.21
V(A^A)	0.21	0.21	0.21	0.21	0.21	0.21
V(D)	0.05	0.05	0.05	0.05	0.05	0.05
V(C)	0.19	0.19	0.19	0.19	0.19	0.19
V(E)	0.2	0.2	0.2	0.2	0.2	0.2
V(Age)	0	0.22	0.22	0.22	0.22	0.22
V(Age^A)	0	-0.02	-0.02	-0.02	-0.02	-0.02
V(CoV)	0.15	0.15	0.15	0.15	0.15	0.15
V(AM)	0.45	0.45	0.45	0.45	0.45	0.45

Downloads:

Program: [GeneEvolve73.zip](#)

Manual: [GeneEvolveManual](#)

GeneEvolve 0.73

- Implemented in R, open-source, user modifiable
- User specifies 31 basic parameters up front (and 17 advanced ones); no need to alter script after that.

How GeneEvolve works:

User specifies:

- population size, # generations for population to evolve, threshold effects, mechanisms of assortative mating, vertical transmission, etc.
- 3 types of genetic effects
- 5 types of environmental effects
- 13 types of moderator/covariate effects

How GeneEvolve works:

- Parameters of interest for present simulations
 - A = additive genetic effects
 - D = dominance genetic effects
 - U = unique environmental effects
 - F = familial environmental effects
 - S = sibling environmental effects
 - AM = correlation between spouses
 - am.model = “I”: primary phenotypic
“II”: social homogamy

How GeneEvolve works (cont):

- At adulthood, $\sim x\%$ find mates s.t. correlation b/w mating phenotypes = AM:
- Pairs have children :
 - ◆ Rate determined by user-specified population growth
- Process iterated n times (Markov Chain)

How GeneEvolve works (cont):

- After n iterations, population splits into two:
 - ◆ Parents of spouses
 - ◆ Parents of twins
- Parents of twins have offspring (MZ/DZ twins & their sibs)
- Twins mate with spousal population & have offspring

What you get:

- 3 generations (grandparents, parents, & offspring) of phenotypic data written out, one row per family, potentially across repeated measures
- This data can be entered into structural models for model validation and sensitivity analysis
- A summary PDF at end shows:
 - Basic simulation statistics
 - Changes in variance components across time
 - Correlations between 10 relative types

Why go through the trouble to simulate a population's evolution rather than to simulate in one step (or analytically)?

- Many parameters change dynamically (evolutionarily) as functions of other parameters in models that include assortative mating and vertical transmission. Predicting changes in such parameters is impractical and approaches impossible in models where many things simultaneously going on.

GeneEvolve Practical

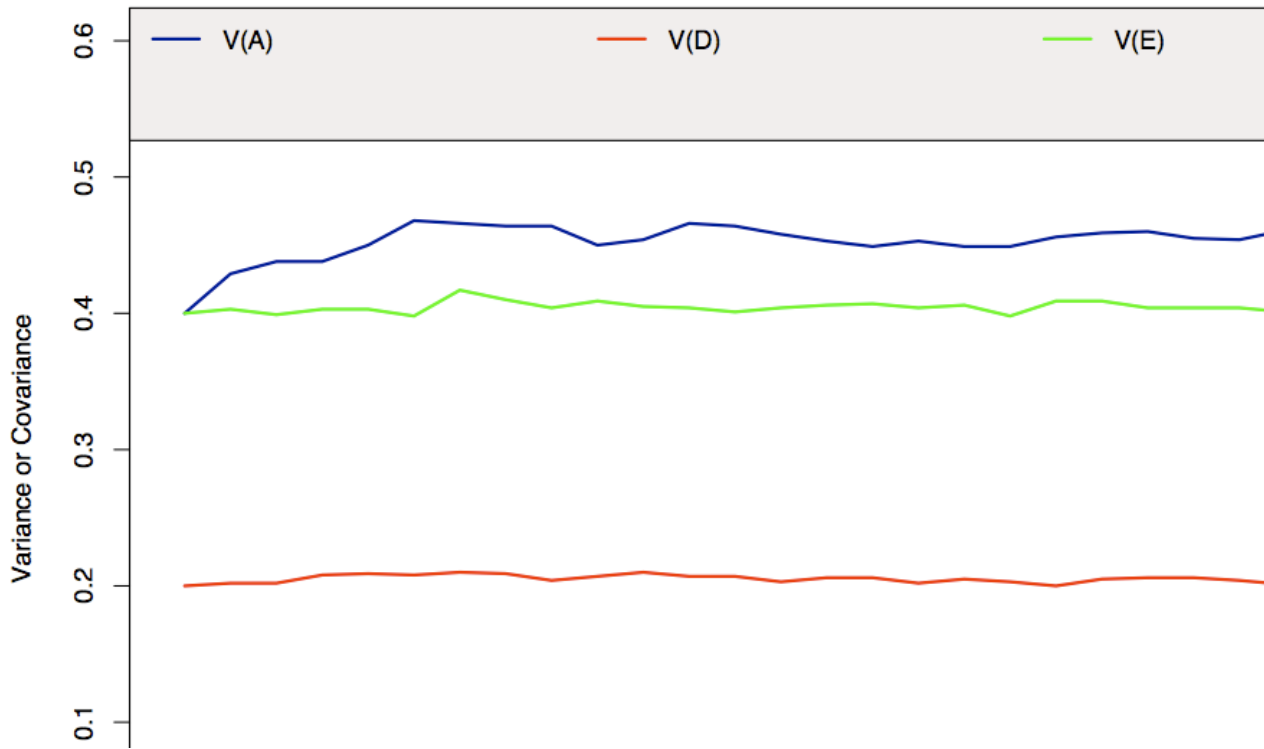
Getting Started

- Copy F:/matt/GE folder into your home directory
- Start R. Then File -> Open Script -> home:GE/GE-73.R

Running GeneEvolve

1. Create a reality where $A=.3$, $D=.2$, $F=.1$, $S=.1$, $U=.3$ & $AM = .2$ (all other parameters = 0). After it runs (~1.5 min), open the resulting PDF. What happens to the A variation across 10 generations? D? F? S? Why?
2. Do the same thing but change the mode of mating to social homogamy (`am.model <- "II"`). What happens?
3. Run another model you find interesting & see what happens

Why does variance of A increase in presence of AM?



Part III: Model validation and sensitivity analysis of extended twin family models

Using complex models without independent validation (e.g., simulation) is like...

QuickTime™ and a
decompressor
are needed to see this picture.

Process of model validation

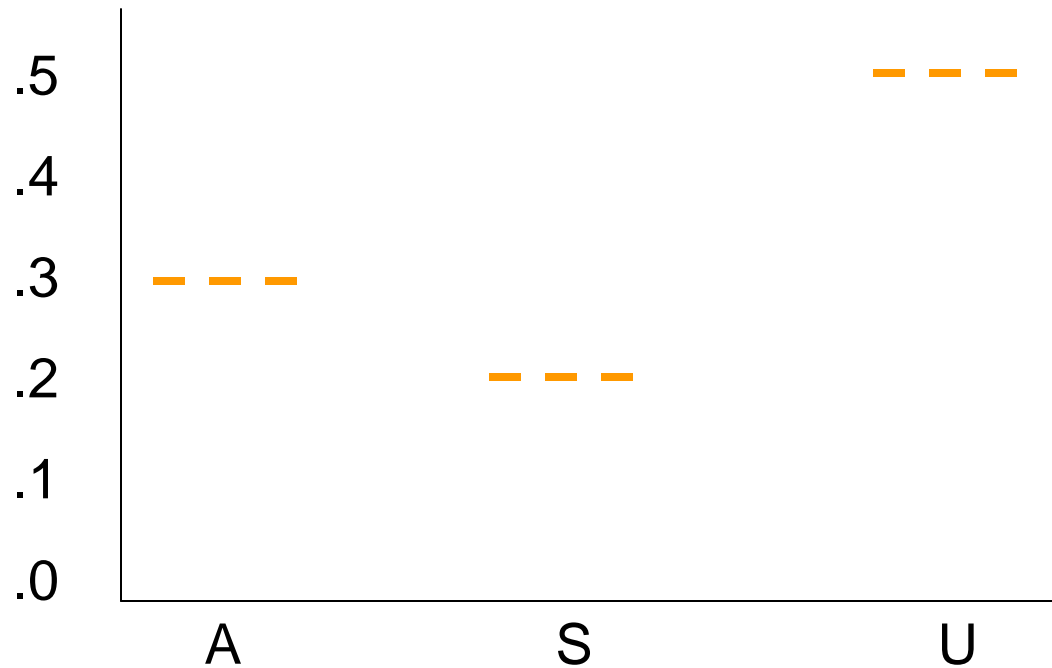
1. Simulate a dataset that has parameters that your model can estimate.
2. Run your model on the simulated dataset
3. Obtain and store parameter estimates
4. Repeat steps 1-3 many (e.g., 1000) times

Results of model validation

- If the mean parameter estimate = the simulated parameter estimate, the estimate is *unbiased*. If your model has no mistakes, parameters should generally be unbiased (there are exceptions)
- The standard deviation of an estimates corresponds to its *standard error* and its distribution to its *sampling distribution*
- You can also easily study the *multivariate sampling distribution and statistics*. E.g., how correlated parameters are.

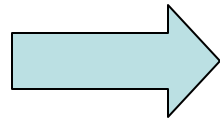
Graphical representation: model validation

Simulate
parameters &
get simulated
dataset



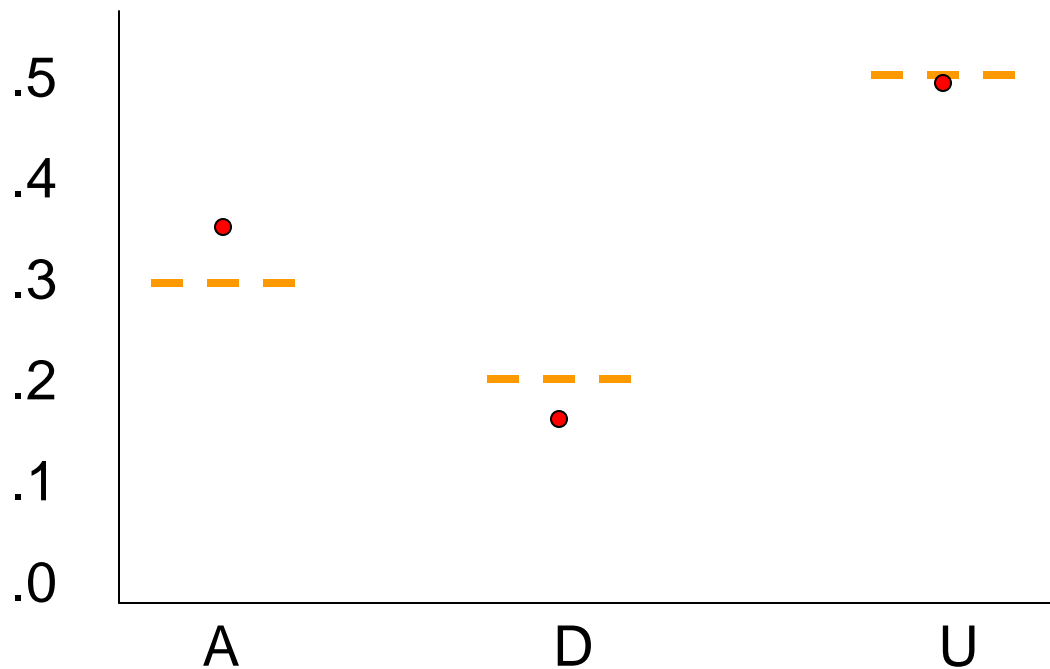
Graphical representation: model validation

Simulate parameters & get simulated dataset



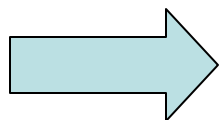
Run Mx, get estimates

e.g., $A=.34$, $D=.17$, $U=.49$



Graphical representation: model validation

Simulate
parameters &
get simulated
dataset

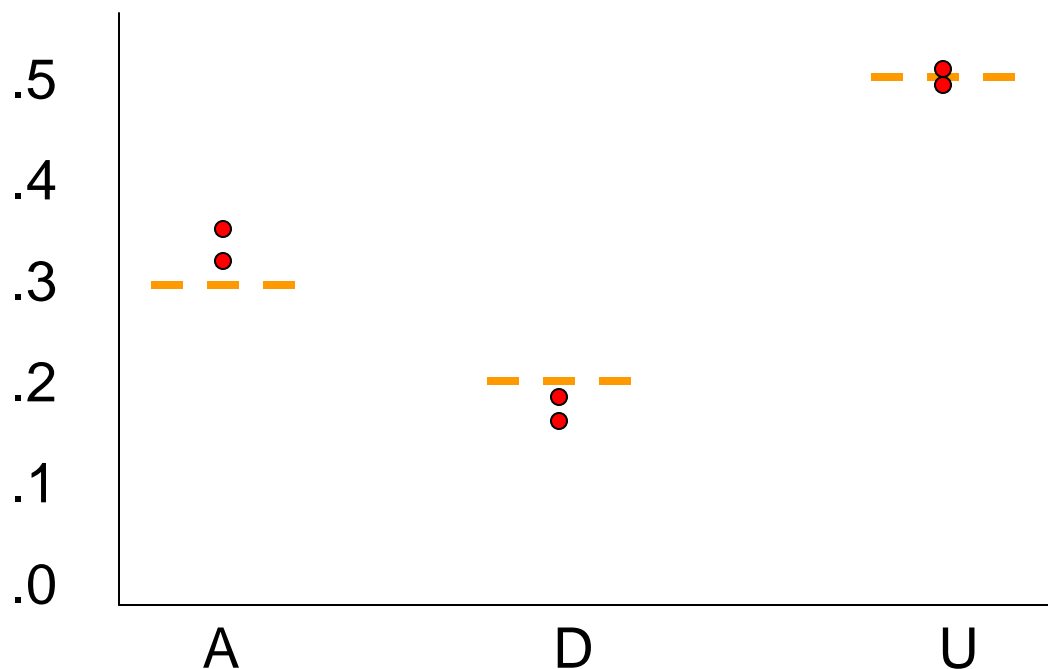


Run Mx, get
estimates

e.g., $A=.34, D=.17, U=.49$
e.g., $A=.31, D=.19, U=.50$

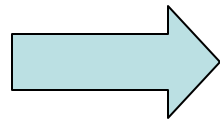


Repeat 1



Graphical representation: model validation

Simulate
parameters &
get simulated
dataset

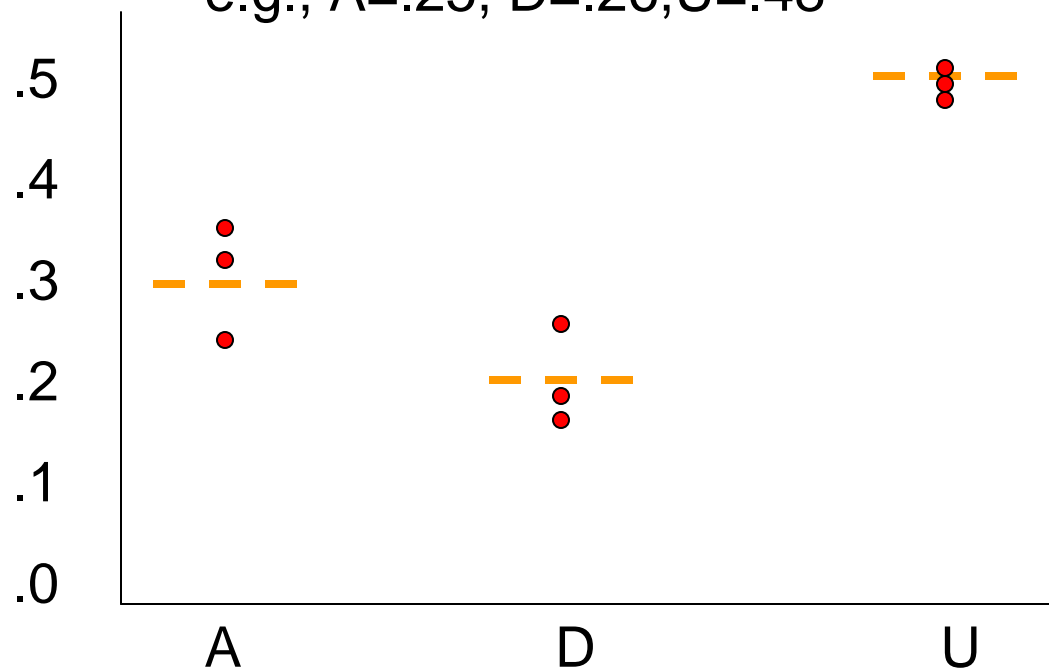


Run Mx, get
estimates



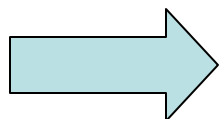
Repeat 1
Repeat 2

e.g., $A=.34, D=.17, U=.49$
e.g., $A=.31, D=.19, U=.50$
e.g., $A=.25, D=.26, U=.48$



Graphical representation: model validation

Simulate
parameters &
get simulated
dataset

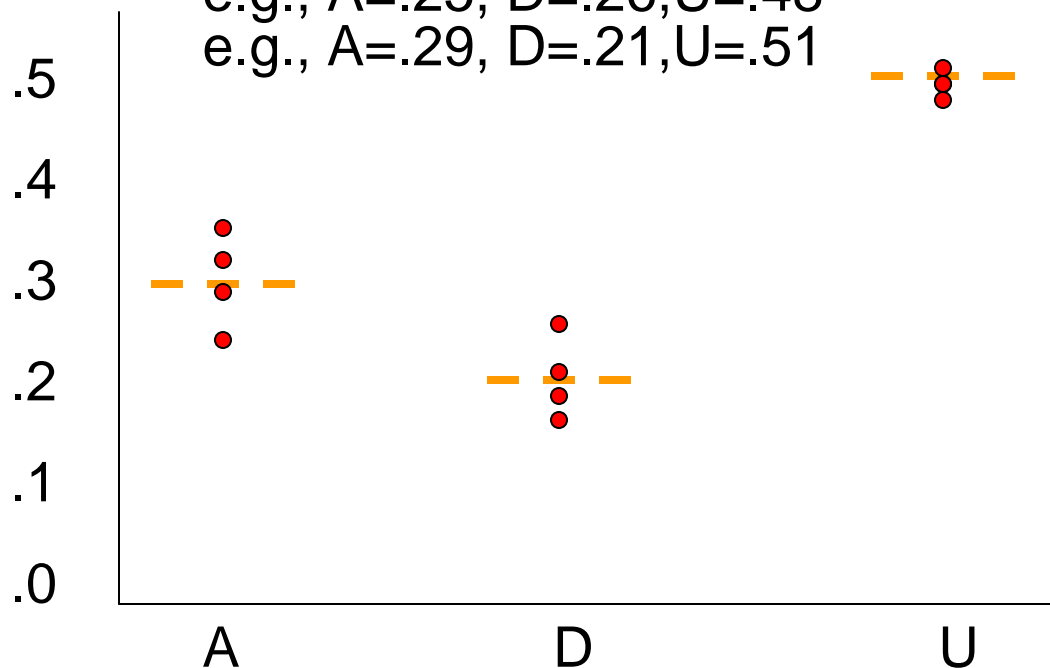


Run Mx, get
estimates



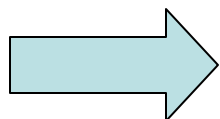
Repeat 1
Repeat 2
Repeat 3

e.g., $A=.34, D=.17, U=.49$
e.g., $A=.31, D=.19, U=.50$
e.g., $A=.25, D=.26, U=.48$
e.g., $A=.29, D=.21, U=.51$



Graphical representation: model validation

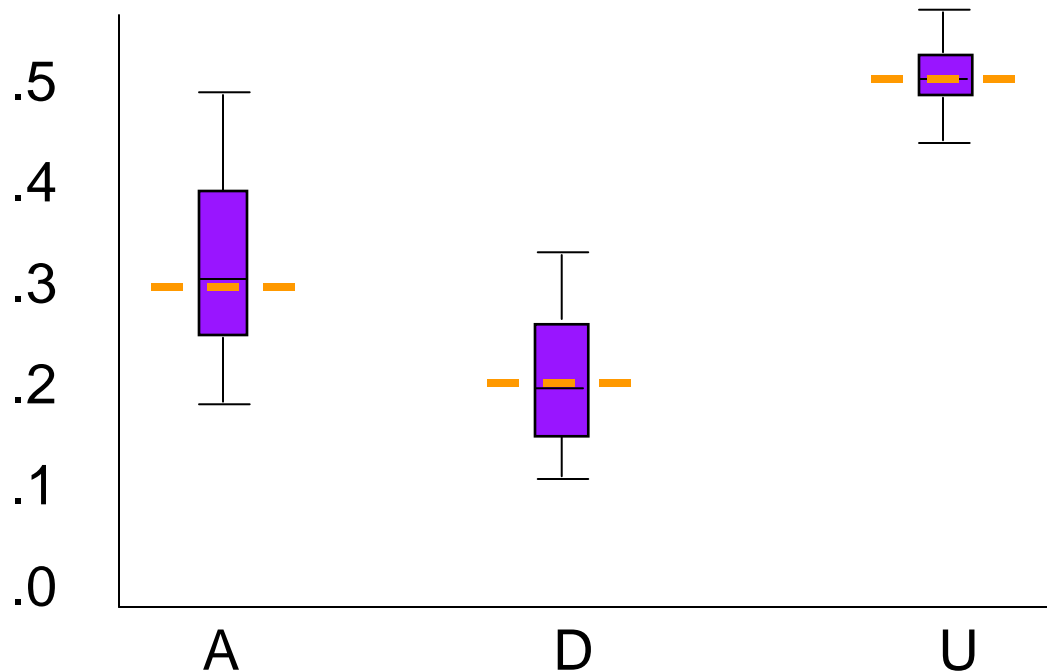
Simulate parameters & get simulated dataset



Run Mx, get estimates



Repeat 1
Repeat 2
Repeat 3

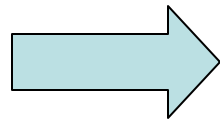


-
-
-

Repeat 1000

Graphical representation: model validation

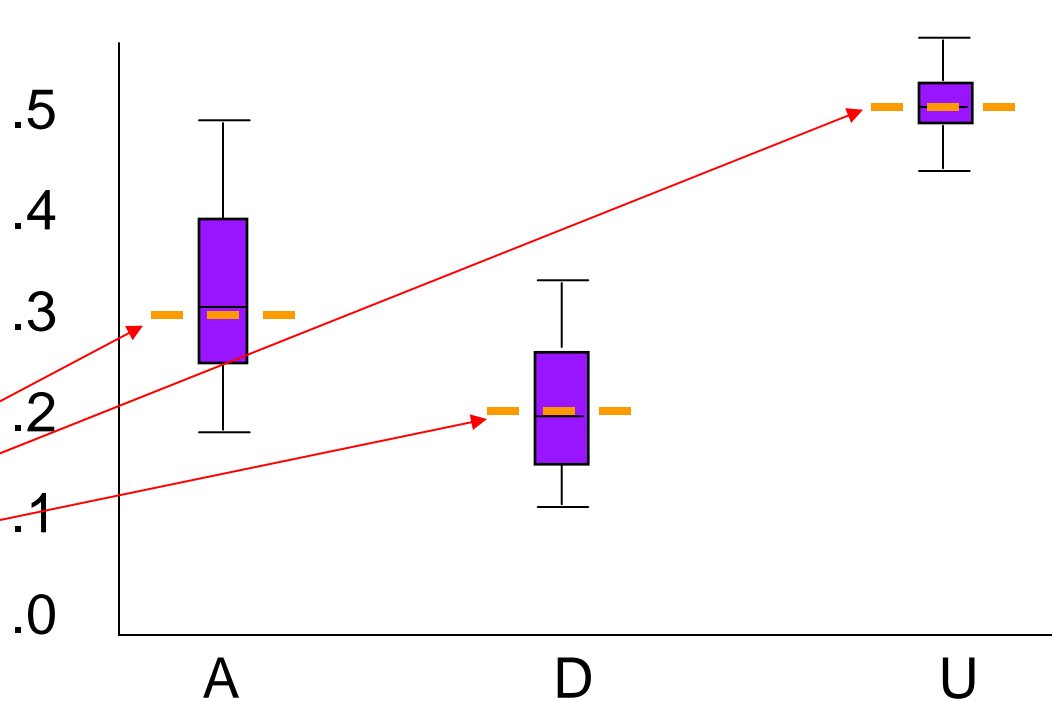
Simulate parameters & get simulated dataset



Run Mx, get estimates



Repeat 1
Repeat 2
Repeat 3



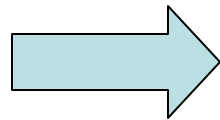
-
-
-

Repeat 1000

Mean estimate ~ true estimate: Unbiased!

Graphical representation: model validation

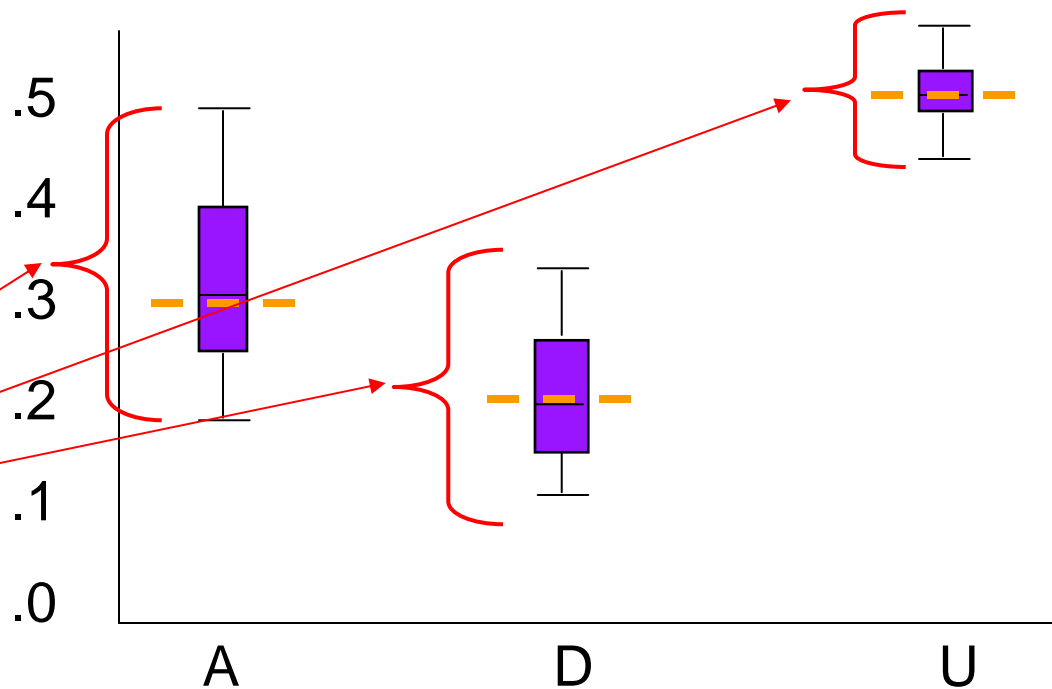
Simulate parameters & get simulated dataset



Run Mx, get estimates



Repeat 1
Repeat 2
Repeat 3



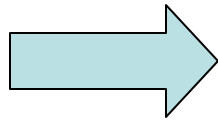
-
-
-

Repeat 1000

Boxplots give idea about variance and shape of sampling distributions

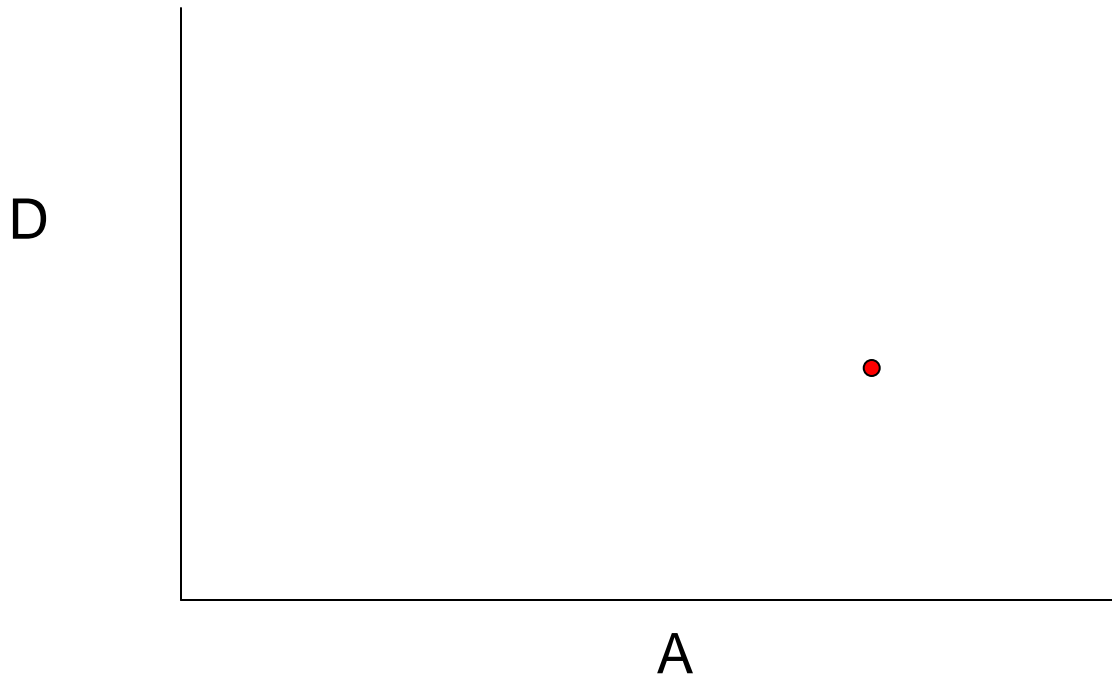
Graphical representation: model validation & multivariate distributions

Simulate parameters & get simulated dataset



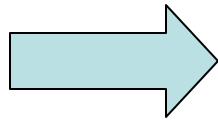
Run Mx, get estimates

e.g., $A=.34$, $D=.17$, $U=.49$



Graphical representation: model validation & multivariate distributions

Simulate parameters & get simulated dataset

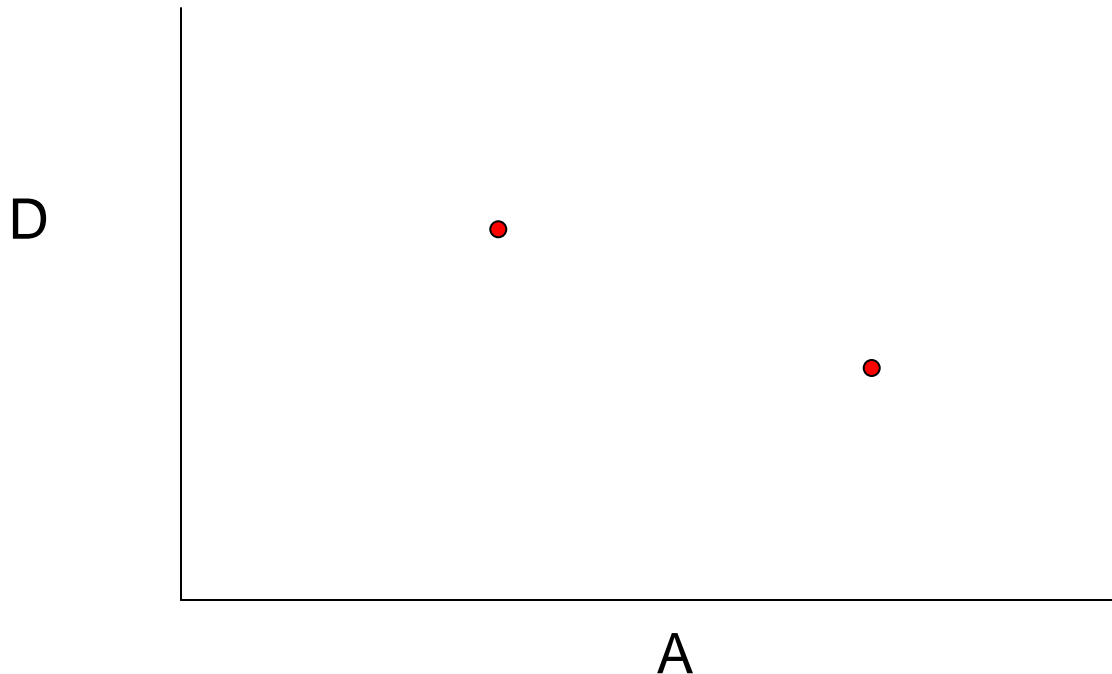


Run Mx, get estimates



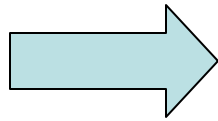
Repeat 1

e.g., $A=.34, D=.17, U=.49$
e.g., $A=.31, D=.19, U=.50$



Graphical representation: model validation & multivariate distributions

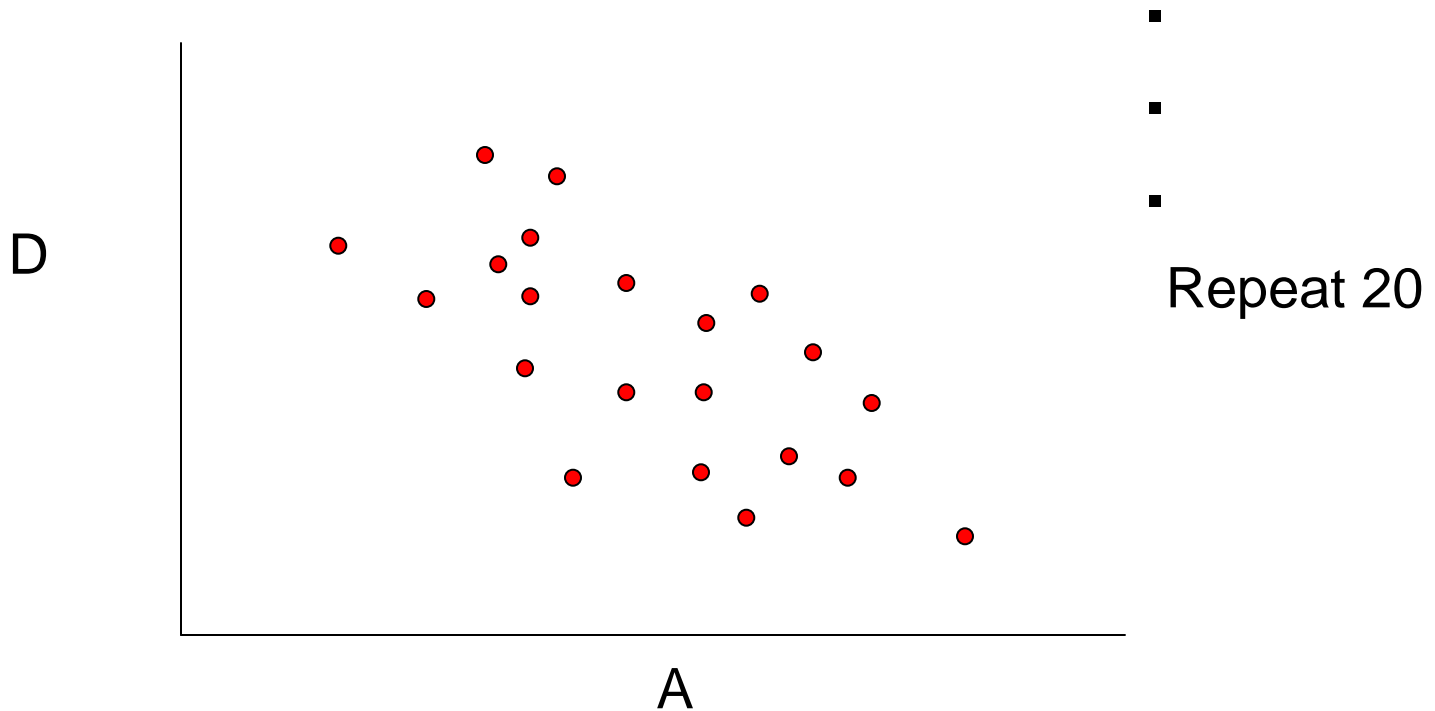
Simulate parameters & get simulated dataset



Run Mx, get estimates

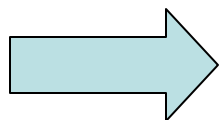


Repeat 1
Repeat 2
Repeat 3



Graphical representation: model validation & multivariate distributions

Simulate parameters & get simulated dataset

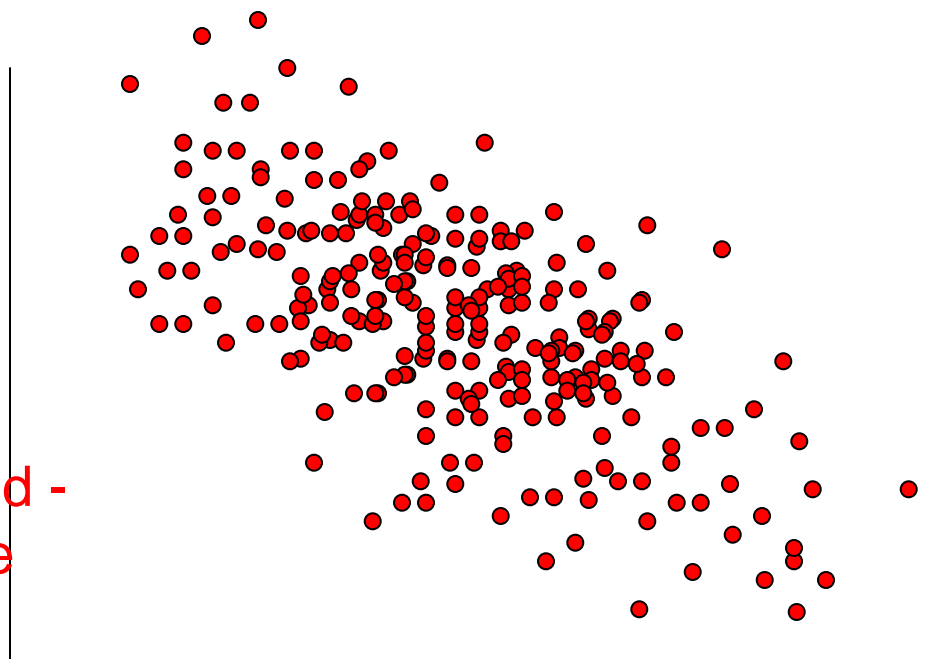


Run Mx, get estimates



Repeat 1
Repeat 2
Repeat 3

D



▪
▪
▪

Repeat 1000

A & D estimates negatively correlated - suggesting they use overlapping information to be estimated

Process of sensitivity analysis

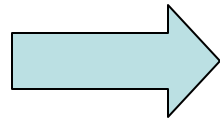
1. Simulate a dataset that has one or more parameters that your model *cannot* estimate.
2. Run your model on the simulated dataset
3. Obtain and store parameter estimates
4. Repeat steps 1-3 many (e.g., 1000) times

Results of sensitivity analysis

- Because we are simulating *violations of assumptions*, we expect parameters to be biased. The question becomes: *how* biased? I.e., how big of a deal are these violations? We should be able to quantify the answers to these questions.

Graphical representation: model sensitivity

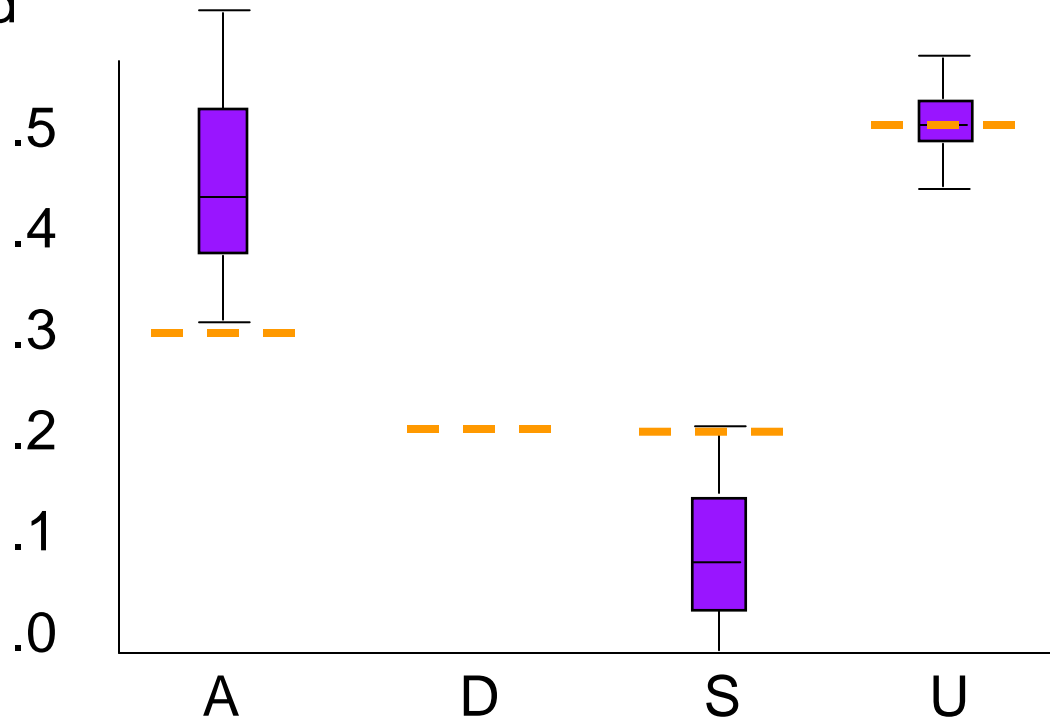
Simulate parameters that include a violation (here, both D & S exist simultaneously) & get simulated dataset



Run Mx, get estimates



Repeat 1
Repeat 2
Repeat 3



-
-
-

Repeat 1000

Sensitivity analysis practical

Run GeneEvolve

1. Create a reality where $A=.4$, $D=.1$, $S=.2$, $U=.3$ & $AM = 0$. Datasets MZM, MZF, DZM, DZF, DZOS are made automatically.

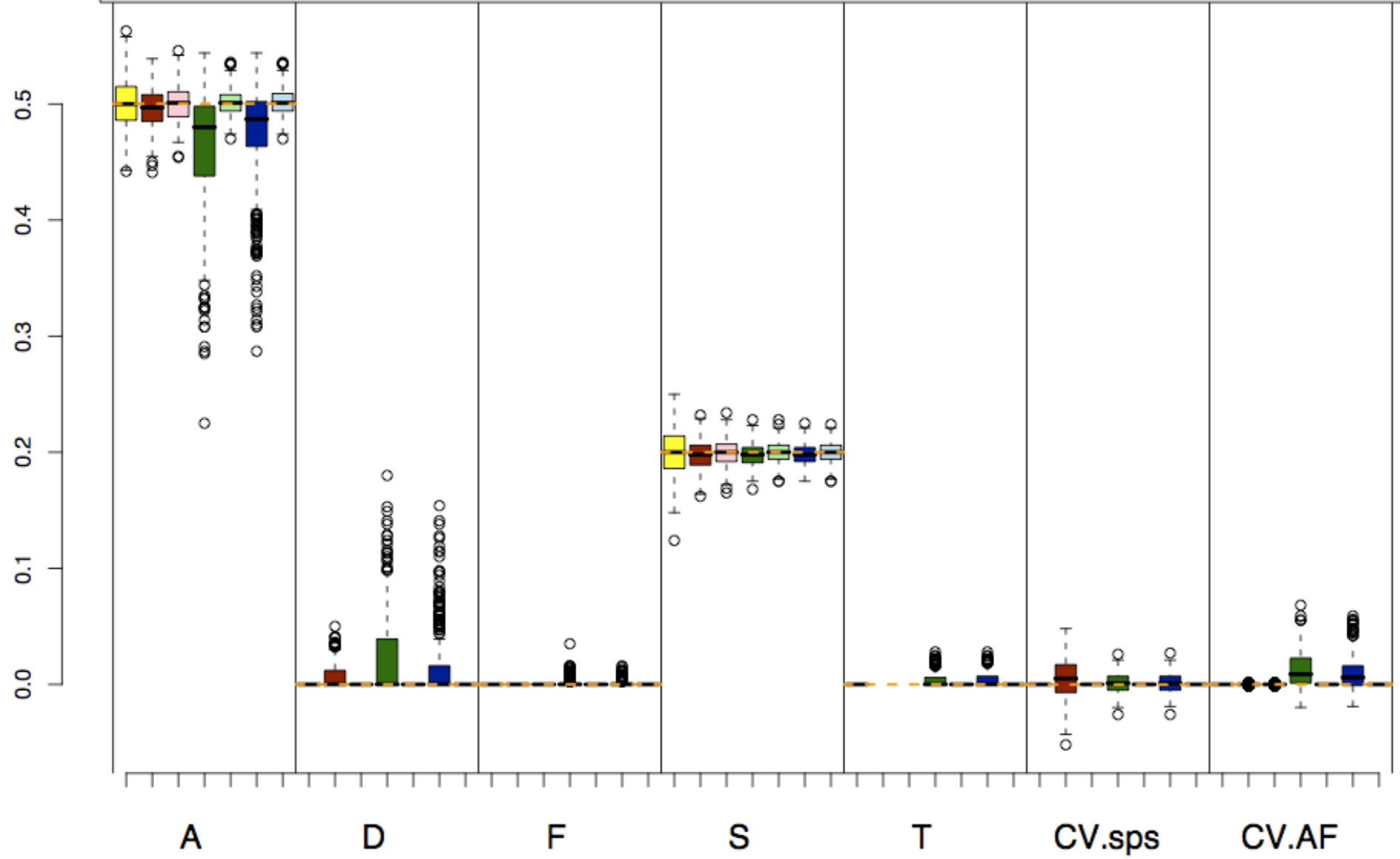
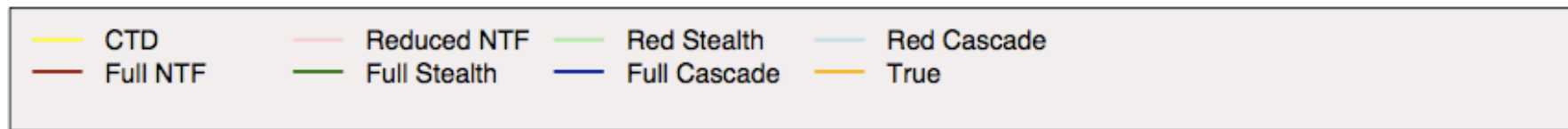
Run Mx

1. Run the script “GE.Twin_ASE.mx”. This is an ASE script where D is fixed to 0
2. Run the script “NTF.mx” This is a nuclear twin family script where A, D, and S are simultaneously estimated.
3. Once you have estimates of A, D, and S from both scripts, come up and write them into the Excel spreadsheet. They are found in the 7th, 8th, and 9th elements of the P matrix in “GE.Twin_ASE.mx” and ??? In the ??? matrix in “NTF.mx”

If there's time...

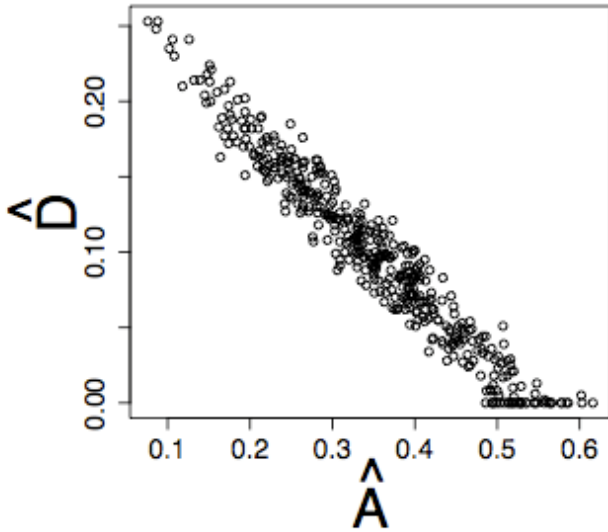
Model validation and sensitivity
results for 4 models

Reality: $A=.5, S=.2$

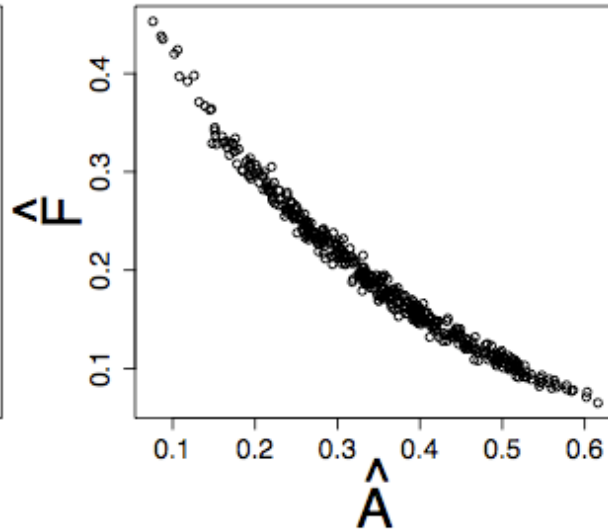


A, D, & F estimates are highly correlated in Stealth & Cascade models

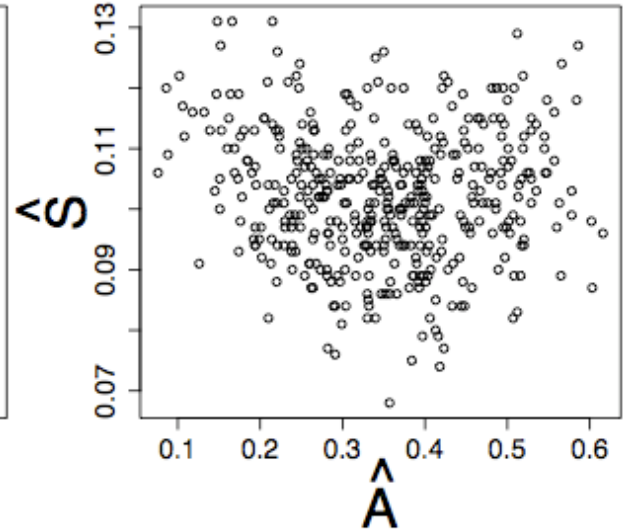
$r = -0.97$



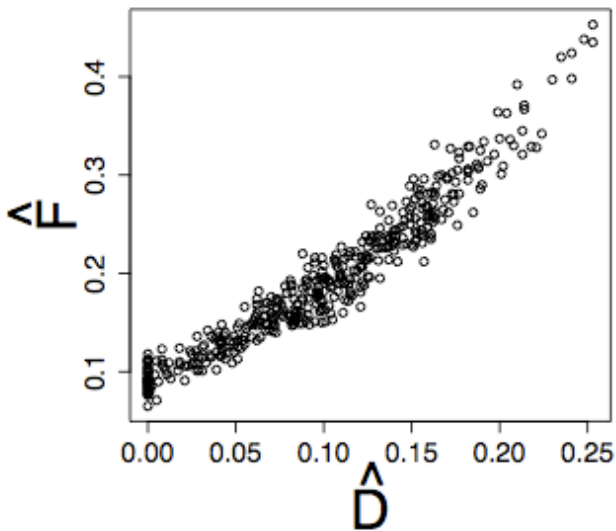
$r = -0.98$



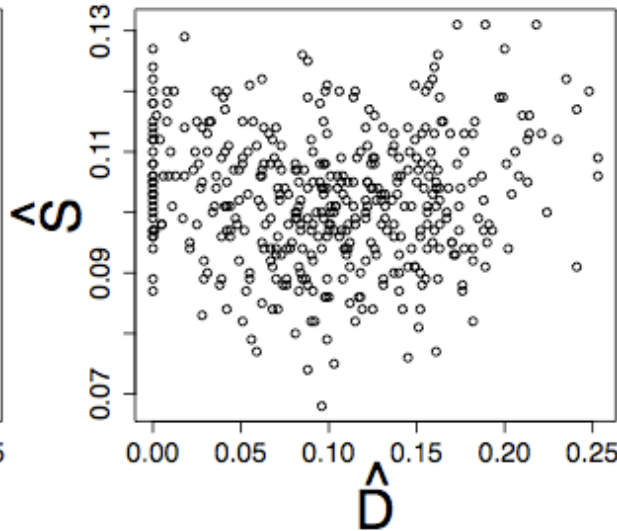
$r = -0.07$



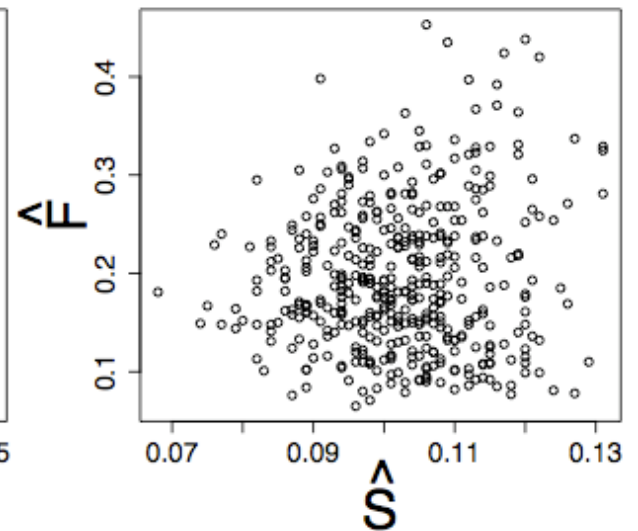
$r = 0.96$



$r = 0$



$r = 0.09$



Simulation is not a panacea

- Simulation can be said to provide “knowledge without understanding.” It is a helpful tool for understanding, but doesn’t provide understanding in and of itself.
- Simulations themselves rely on assumptions about how processes work. If these are wrong, our simulation results may not reflect reality.

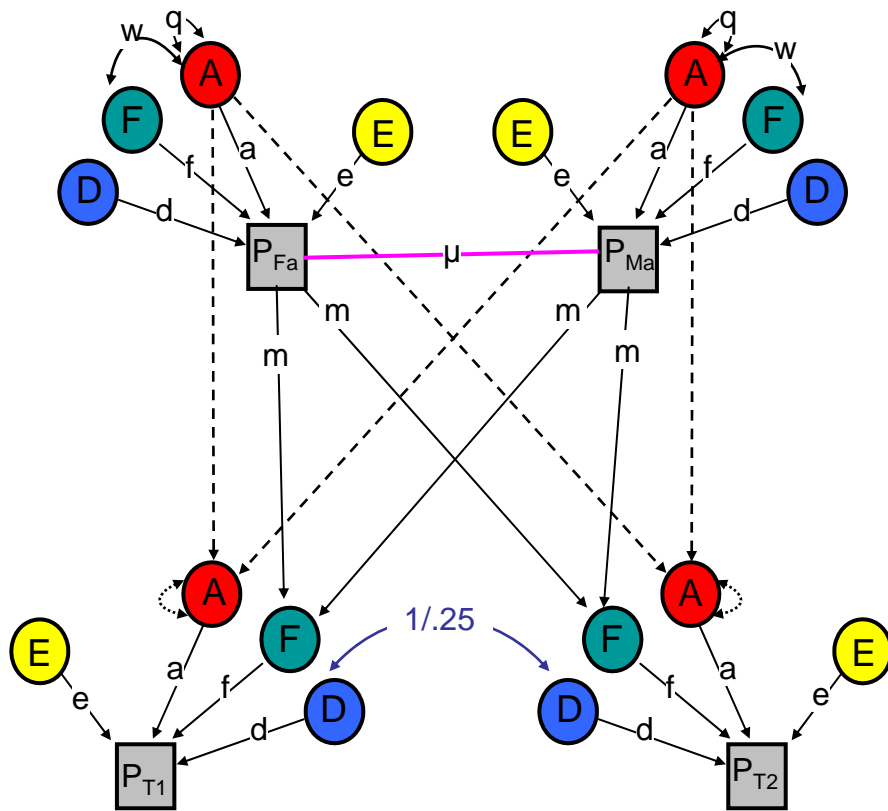
GeneEvolve Limitations

- Sex limitation possible only for A at the moment
- No multivariate except for longitudinal

Conclusions

- All models require assumptions. Generally, more assumptions = more biased estimates
- Extended twin family designs require fewer assumptions and tend to be less biased
- Simulation is a powerful tool for checking complex models (and not just extended twin family models)

Why does variance of A increase in presence of AM?



Answer: When two spouses are phenotypically similar, they also tend to have similar A effects.

Offspring A is a weighted sum of parental A.

Therefore, variance of A increases for same reason that the variance of any sum increases when components are correlated.

For similar reasons, variance of the other transmitted parameter, F, also increases.