Model assumptions & extending the twin model

Matthew Keller
Hermine Maes
Brad Verhulst
Lindon Eaves

Boulder 2018

Acknowledgments

- John Jinks
- David Fulker
- Robert Cloninger
- Lindon Eaves
- Nick Martin
- Andrew Heath
- Sarah Medland, Pete Hatemi, Will Coventry, Hermine Maes, Mike Neale

Third annual OpenMx HACKATHON! Friday morning (8 am) session

- Lucia and I will give you an .RData file of twin data and a specific question to test. Your job is to write an OpenMx script—from scratch—that gets the right answer!
  - Cheating isn't bad here—it's encouraged! Use your old scripts or help from anyone in the class.
  - You have an hour to write script and to produce and interpret estimates.

Files you will need are in Faculty drive: /matt/Assumptions2018

- Assumptions_mck_2018.pdf (PPT presentation)
- CTD.ACDE-param.indet_2018.R (OpenMx script)
- PDFs of papers describing details of what we go over here & that correspond to the approach/ notation I'm using here

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 VA and 20% to VC."
  - Should you believe that 50% of variation is truly additive genetic?

True parameters vs. Estimated parameters

- VA VC VD VE: true (unknowable) values in the population

- *VA', VC', VD', VE'*: **estimated** values of VA, VC, VD, VE.

- *VA', VC', VD', VE'*, will differ from VA, VC, VD, VE due to:

  1) sampling variability
  2) bias (= E[*θ'*] - θ)

- This session is about deriving biases in estimates, how to interpret them in light of these biases, and how to model in ways that minimize bias

How to derive algebraic expectations of variance component estimates

1) In an ACE model, we assume VD=0. So to get algebraic expectations of *VA'* and *VC'* in an ACE model, write down what CVmz and CVdz are composed of:

CVmz = VA + VC

CVdz = ½VA + VC

2) To get an estimate of one term (e.g., VA) try to think of possible contrasts of linear transformations that get rid of one parameter (e.g., VC) and isolate the other (e.g., VA). Thus:

CVmz – CVdz = ½VA. Thus 2(CVmz-CVdz) = VA. Thus an estimate of VA:

VA' = 2(CVmz – CVdz).

3) Similarly to get rid of VA and isolate VC:

VC' = 2CVdz - CVmz

NOTE: I'm using *VY'* rather than the usual VŶ to denote estimates of VY simply due to PPT issues!

Practical 1 – algebraic expectations of ADE

1) Use what we just learned to derive algebraic expectations of the estimates of VA and VD in an ADE model (where we assume VC=0). As a hint, in this situation, we're assuming:

$$CVmz = VA + VD$$

$$CVdz = \tfrac{1}{2}VA + \tfrac{1}{4}VD$$

2) Now to get *VA'*, think of possible contrasts of linear transformations of CVmz and CVdz that get rid of VD and isolate VA.

QUESTION1.1: What is your estimate of VA (*VA'*) in an ADE model?

3) Now do the same to get *VD'*

QUESTION1.2: What is your estimate of VD (*VD'*) in an ADE model?

How to derive algebraic expectations of bias in estimates due to misspecification

1) We want to know what happens when we misspecify the model (a parameter that is non-zero in real life is omitted in the model). To get at this, first write out your estimate. E.g., in an ACE model, *VA′* is:

$$VA' = 2*(CVmz - CVdz).$$

2) Next consider what variance components REALLY exist in your estimates. If VD is actually non-zero, then we know:

$$CVmz = VA + VD + VC$$

$$CVdz = \tfrac{1}{2}VA + \tfrac{1}{4}VD + VC$$

3) Finally, just plug in the reality to your estimates. Thus, in an ACE:

$$VA' = 2*(VA + VD + VC - \tfrac{1}{2}VA - \tfrac{1}{4}VD - VC) = VA + 3/2(VD)$$

IN word: when VD actually exists and you fit an ACE model, *VA′* is biased upwards by 1.5 of whatever VD truly is.

4) Similarly, VC′ = VC - ½VD. *VC′* is biased downward by half of VD.

Practical 2 – deriving biases of ADE

1) Use what we just learned to derive the bias in the *VA'* and *VD'* in an ADE model (where we assume VC=0). Recall that:

VA' = 4CVdz – CVmz

VD' = 2CVmz – 4CVdz

CVdz = ½VA + ¼VD

2) Now just plug in the constituent variance components into CVmz and CVdz and see how our estimates are biased.

QUESTION2.1: How is *VA'* biased in an ADE model when VC is, contrary to our assumption, non-zero?

QUESTION2.2: How is *VD'* biased in an ADE model when VC is, contrary to our assumption, non-zero?

Quiz Question 1

1) We must fix to zero (and not estimate) either *VC′* or *VD′* in an identified classical twin model because: [exactly two answers are correct]

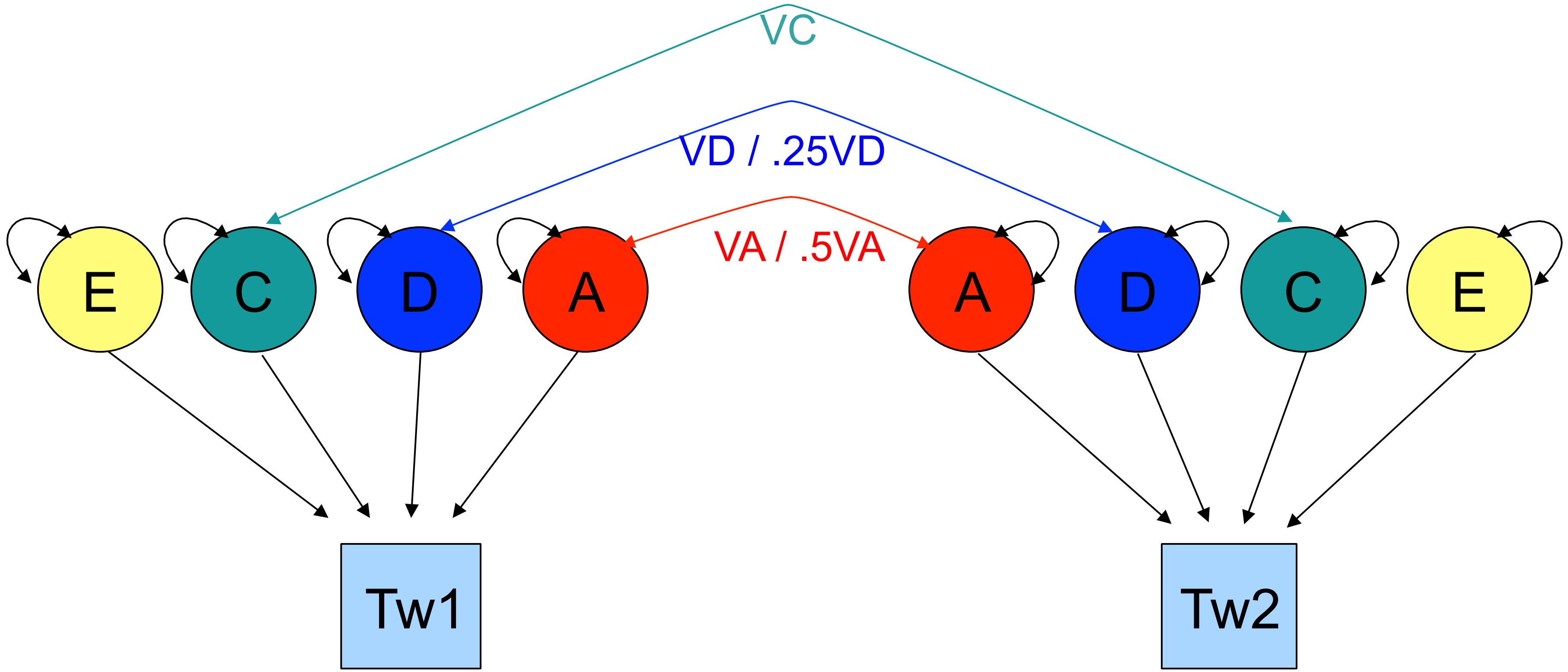
a) these estimates are too highly correlated (multicolinearity problems)

b) you **can** estimate *VC′* and *VD′* simultaneously - you just have to fix *VA′* to some specific value

c) you **can** estimate *VC′* and *VD′* simultaneously - you just have to allow them to go negative (not use path coefficient approach)

d) there are fewer informative statistics (2) than parameters to be estimated (3), thus the full ADCE model is unidentified.

The Classical Twin Design



Why can’t we estimate *VC’* & *VD’* at same time using twins only?

- Solve the following two equations for *VA’*, *VC’*, & *VD’*:

$$CVmz = VA + VD + VC$$

$$CVdz = \tfrac{1}{2} VA + \tfrac{1}{4} VD + VC$$

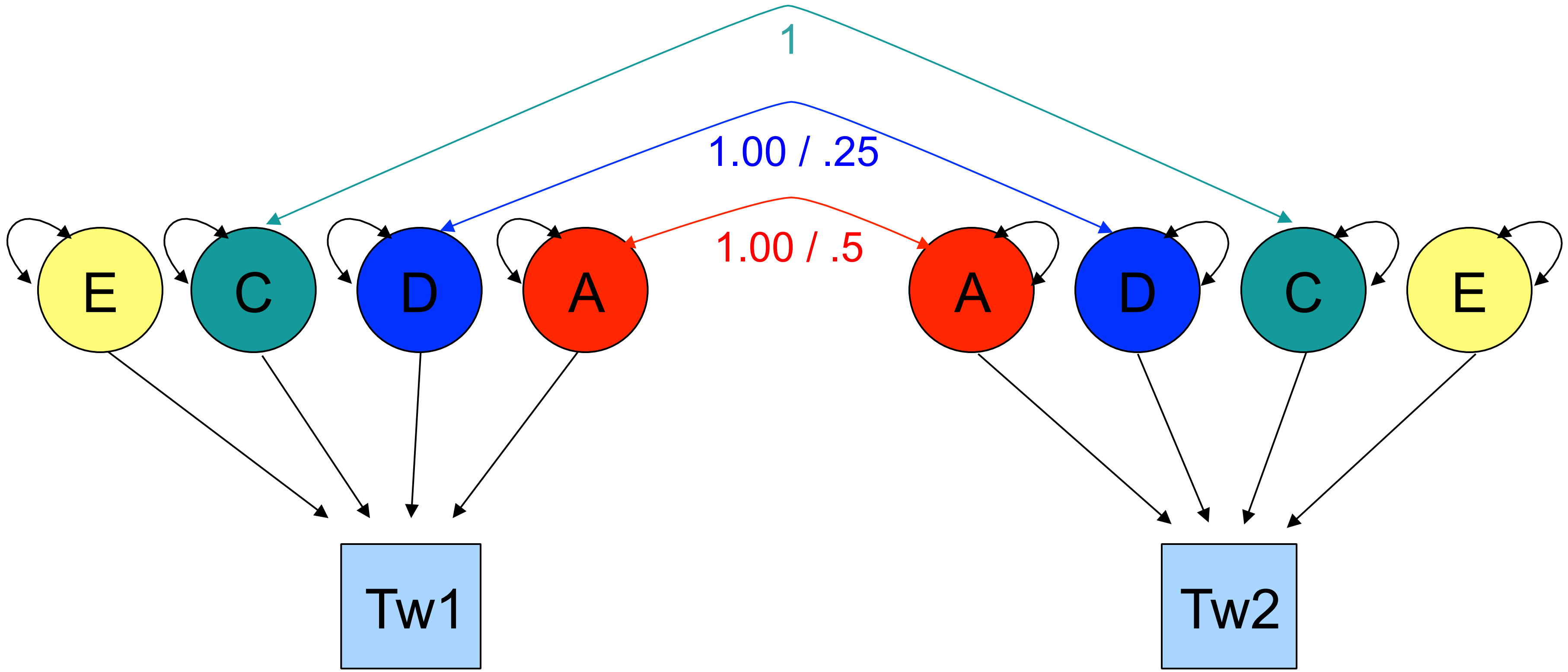
- 3 unknowns, 2 informative equations. It can't be done. There are no <u>unique</u> solutions. The model is “unidentified”.

- In practice, you can detect non-identification by noting that (a) model estimates depend on starting values AND (b) all final models have identical likelihoods

Nonidentification: Practical 3 (using R)

- Open CTD.ACDE-param.indet_2018.R in R
- Run practical 3A to simulate data where truth is VA=.4, VD=.2, VC=.05 (and thus CVmz=.65; CVdz=.3). Pause for discussion.
- Run practical 3B for ADE model on this data. Pause for discussion.
- Run practical 3C for ACE model (which we normally wouldn't do) on same data. Pause for discussion.
- Run practical 3D for ADCE model (which we definitely wouldn't normally do). Pause for discussion:
  - Write down your -2LL and your estimates of VA, VC, and VD
  - Compare these to your neighbor's
  - WHY are -2LL the same despite different *VA'*, *VC'*, and *VD'* (that depend on arbitrary start values)
- Do not close CTD.ACDE-param.indet_2018.R in R

The CTD: Two statistics give info about within-family resemblance



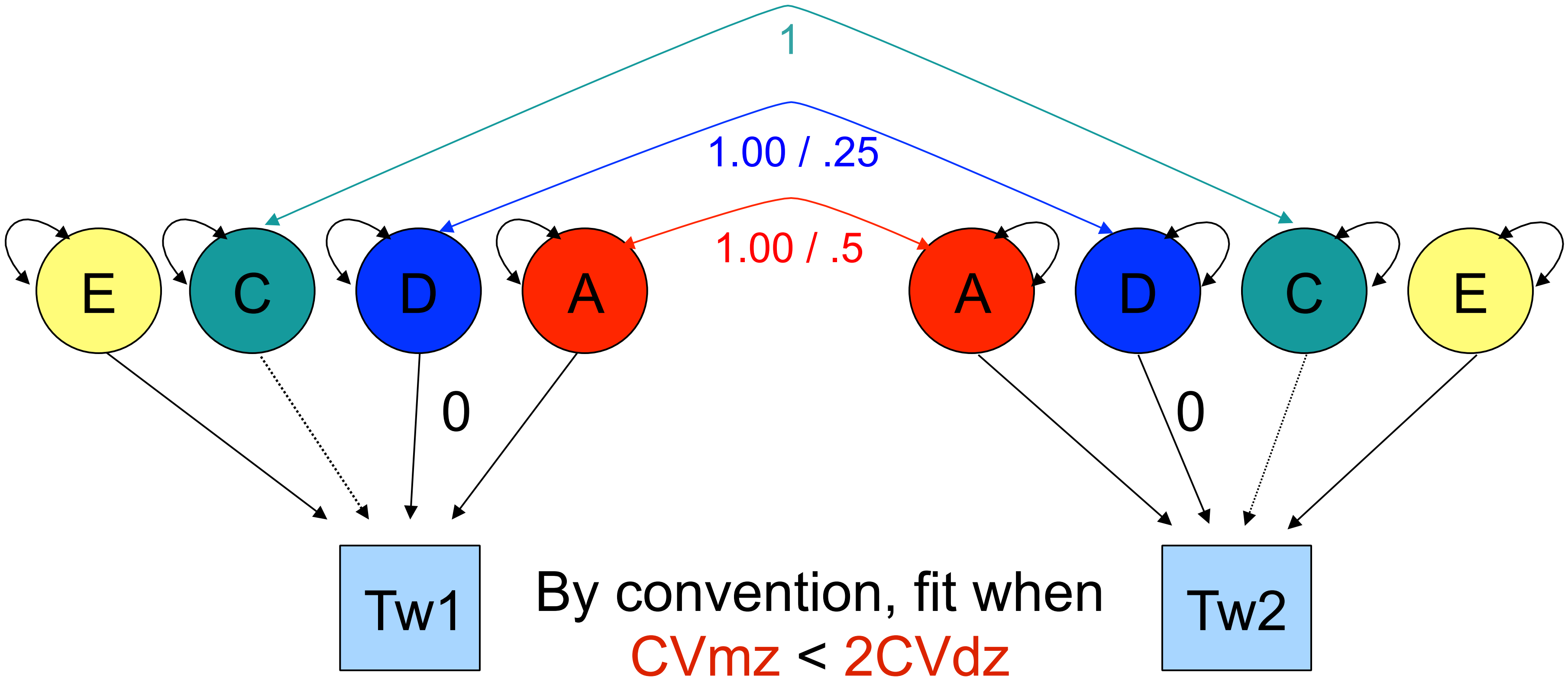
MZ covariance

Vp	
CVmz	Vp

DZ covariance

Vp	
CVdz	Vp

ACE Model

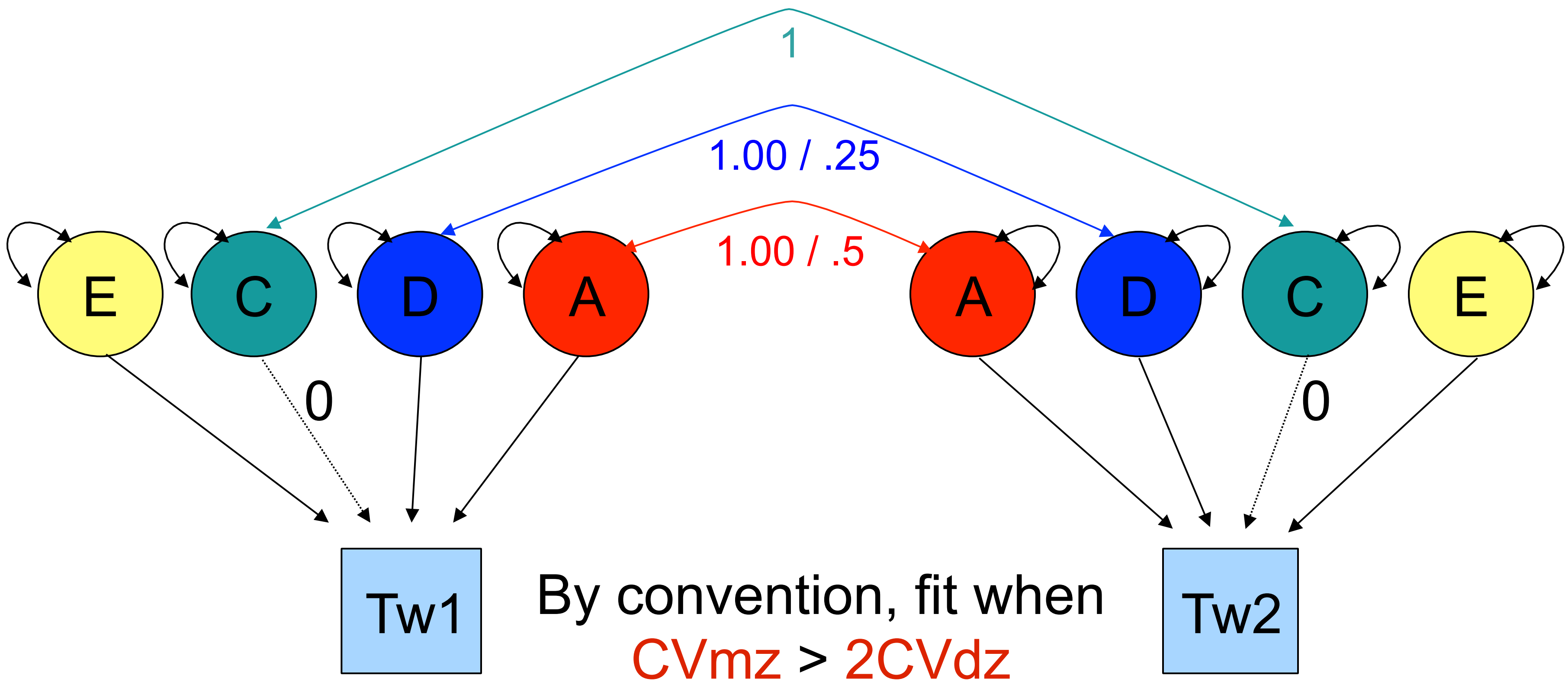


By convention, fit when
CVmz < 2CVdz

Vp	
CVmz	Vp

Vp	
CVdz	Vp

ADE Model



By convention, fit when
CVmz > 2CVdz

Vp	
CVmz	Vp

Vp	
CVdz	Vp

The CTD: Just because we cannot fit VD & VC simultaneously doesn’t mean they’re not there!

- However, when we TRY to fit an ADCE model with just twins, there are an infinite number of combinations of *VA’*, *VD’*, and *VC’* that fit the data equally well = parameter indeterminacy due to model non-identification.

- Thus, we just have to fit either an ADE or ACE model and live with potentially biased estimates.

- But it’s good to quantify this bias to help in interpreting those estimates.

Quiz Question 1 again – what do you think now?

1) We must fix to zero (and not estimate) either *VC′* or *VD′* in an identified classical twin model because: [exactly two answers are correct]

a) these estimates are too highly correlated (multicolinearity problems)

b) you **can** estimate *VC′* and *VD′* simultaneously - you just have to fix *VA′* to some specific value

c) you **can** estimate *VC′* and *VD′* simultaneously - you just have to allow them to go negative (not use path coefficient approach)

d) there are fewer informative statistics (2) than parameters to be estimated (3), thus the full ADCE model is unidentified.

So what is the advantage of estimating variances directly (without a bound) if it doesn't solve bias due to model misspecification?

- Foremost: valid p-values. If we bound estimates, the distribution of -2LL differences under null is not χ^2 (it's 50% χ^2 & 50% with point mass at lower bound; e.g., 0). Thus inflated type-II errors.

- Second: eliminates a source of bias due to sampling variability.
  - If we think about estimates being random values under repeated draws of data, whenever the estimate hits a zero bound, it creates biases in it's own estimate (up) and in other estimates (up or down).
  - This is a separate (and probably smaller) source of bias from that due to model misspecification.

- Note – when you directly estimate variances, it's easy to transform between *VC'* and *VD'*:
  - In ADE model, *VC'* you would have gotten in ACE = - ½*VD'*
  - In ACE model, *VD'* you would have gotten in ADE = -2*VC'*

Quiz Question 2

2) If the assumptions of the CTD model that either VD or VC is zero is violated (i.e., VA, VC, and VD simultaneously affect the phenotype)... [choose all that apply]

a) the interpretation of the estimated parameters should be altered; e.g., *VA'* should be considered an amalgam of VA & VD (in ACE model) or of VA & VC (in ADE model)

b) there is no point in doing the analysis

c) the point estimates of the estimated parameters will be biased

Bias in parameter estimates for violation of assumption that either VD or VC is 0

- In ACE Models (bias induced in setting $VD' = 0$):

$VA' = VA + 3/2VD$

$VC' = VC - \frac{1}{2}VD$

- In ADE Models (bias induced in setting $VC' = 0$):

$VA' = VA + 3VC$

$VD' = VD - 2VC$

Quiz Question 3

3) An ADE model finds that *VA'* = .30 and *VD'* = .10. This implies that shared environmental factors do not influence the trait in question.

a) TRUE

b) FALSE

Quiz Question 4

4) We run an ADE model and find that *VA'* = .69 and that *VD'* = .05. If in truth, VC = .10, what will the effect on the estimated parameters be? [choose all that apply]

a) *VA'* will be biased (too low)

b) *VA'* will be biased (too high)

c) *VD'* will be biased (too low)

d) *VD'* will be biased (too high)

e) there is no affect on the estimated parameters; however by not estimating VC (aka, fixing it to zero), we underestimated VC

PRACTICAL 4: Sensitivity analysis

- Sensitivity analysis: studying what the effects are on estimated parameters when assumptions are wrong
- In CTD.ACDE-param.indet_2018.R, run:

  FROM "# START PRACTICAL 4"
  TO "# END PRACTICAL 4"

- Run one section at a time and change the value of VC from 0 to other possible values in an ADE model. What happens to estimates of VA and VD depending on different assumed values of VC?

Effects of epistasis on these biases

- Epistasis (across loci interactions) can increase the degree of the biases because it can reduce the CVdz:CVmz ratio even further than the expected 1:4 under dominance.
- However, the degree of bias rests on how strong non-additive genetic influences are. This is an active area of debate.
- Epistatic effects will generally come out in the estimates of VD. Thus, interpret *VD'* broadly, as a rough estimate of VNA
- My take: VA is almost certainly greater than VNA, and evidence for much VD per se is scant. But some traits may show high enough VNA to bias estimates of VC and VD (VNA) down and VA up considerably from twin studies.

Quiz Question 5

5) What are the *typical* assumptions of a classical twin model? [choose all that apply]

a) only genetic factors cause MZ twins to be more similar to each other than DZ twins

b) either VD or VC is zero

c) no epistasis

d) no assortative mating

e) no gene-environment interactions or correlations

What are the effects of violations of assumptions in the CTD?

a) Only genetic factors cause MZ twins to be more similar to each other than DZ twins: VA and VD overestimated and VC underestimated

b) Either VD or VC is zero: VA overestimated and VD & VC underestimated

c) No epistasis: VD or VA overestimated and VC underestimated

d) No assortative mating: VA and VD underestimated and VC overestimated

e) No gene-environment interactions or correlations: AxC: VA overestimated; AxE: VE overestimated; passive Cov(A,C): VC overestimated

Assortative mating consequence on VA

- AM: phenotypic correlation between mating partners
- Many examples (e.g., height ~.2; IQ ~ .3; Social attitudes ~ .5)
- If AM leads to genetic similarity in partners (as it does if due to choice for similarity), there are genetic consequences:
  - Height VA increases in the population because 'tall' ('short') alleles are more concentrated in individuals than expected.
  - E.g., if you're a 'tall' allele sitting in an egg and are waiting around to see what other height genes you'll get paired with from that sperm swimming to you, they are more likely than chance to be other 'tall' alleles (both at the same locus and at others; & this just considers the effects on VA in 1st gen)
-

$$V_{A.equil} = \frac{V_{A0}}{1 - rh^2_{equil}}$$

AM consequence on relative covariance

- AM increases genetic covariances and correlations between relatives (e.g., sibs, parents, cousins, etc).
  - While CVmz increases, it's correlation is already 1 so it doesn't increase
- Consider again being a 'tall' allele in a zygote. This time you are watching your co-twin's zygote get formed. Regardless of whether you exist (are IBD) in your co-twin's egg, you can expect more tall alleles swimming to your co-twin's egg.
- Thus, you can also expect to share more 'tall' alleles with your sibling(s).
- The CVdz that is due to additive genetics is:

$$CV_{DZ,A.equil} = .5V_{A.equil} + .5rh^2_{equil}$$

Quiz Question 6

6) In the CTD, say that CVmz < 2CVdz, so we fit an ACE model. How would AM tend to affect parameter estimates? [choose all that apply]

a) deflates estimates of VA

b) inflates estimates of VA

c) deflates estimates of VC

d) inflates estimates of VC

Quiz Question 7

7) Say we add parents to the CTD. That gives us 2 additional relative covariance estimate to work with (parent-offspring and spousal) in addition to the normal CVmz and CVdz and allows us to ______________ [choose all that apply]

a) estimate VA, VC, & VD simultaneously

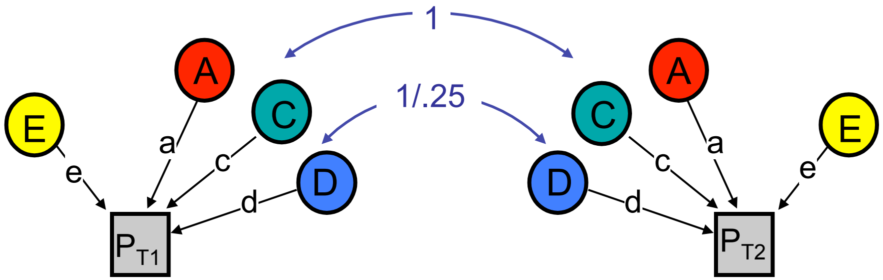
b) account for effects of assortative mating

c) account for passive G-E covariance

d) reduce the bias in estimates of VA, VC, and VD

Classical Twin Design (CTD)

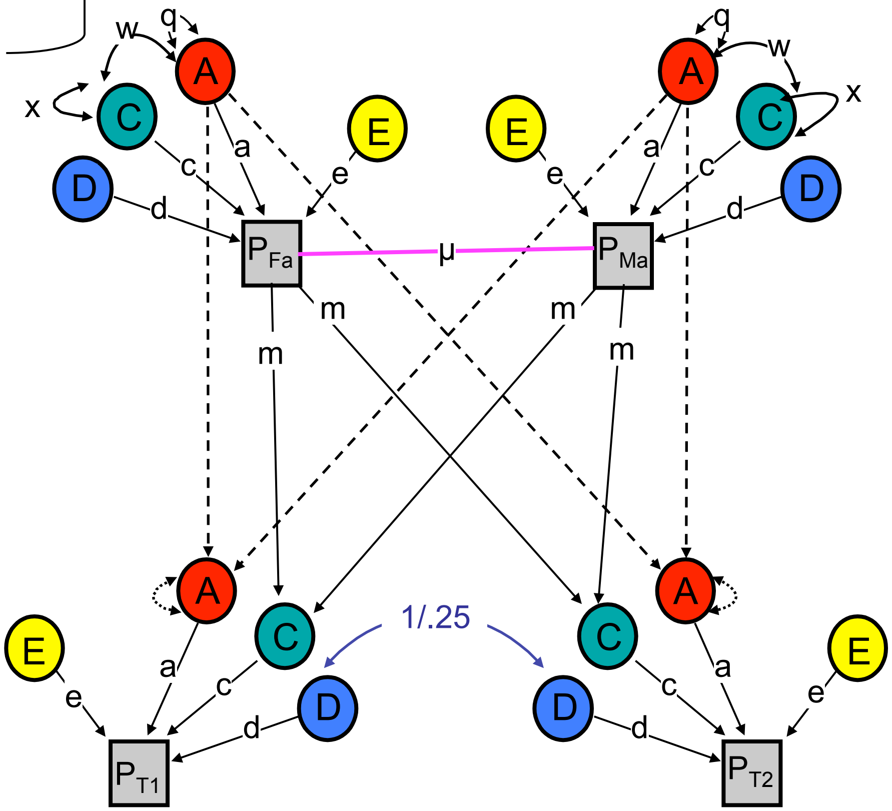
Assumption	biased up	biased down
Either VD or VC is zero	*VA’*	*VC’& VD’*
No assortative mating	*VC’*	*VD’*
No A-C covariance	*VC’*	*VD’& VA’*



Adding parents gets us around all these assumptions

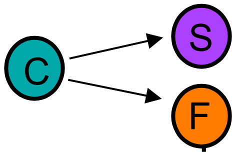
Assumption	biased up	biased down
Either VD or VC is zero		
No assortative mating		
No A-C covariance		

We don't have to make these



We can model VC as either VS or VF

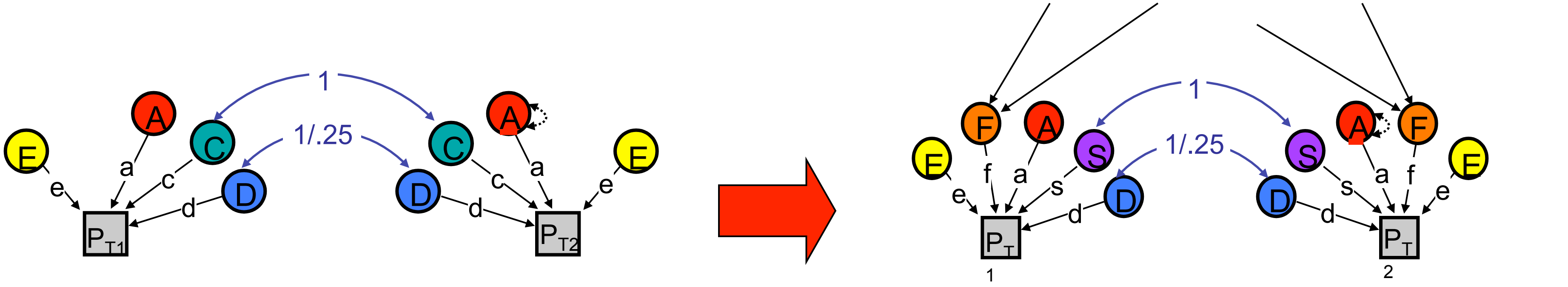
With parents, we can break “VC” up into:



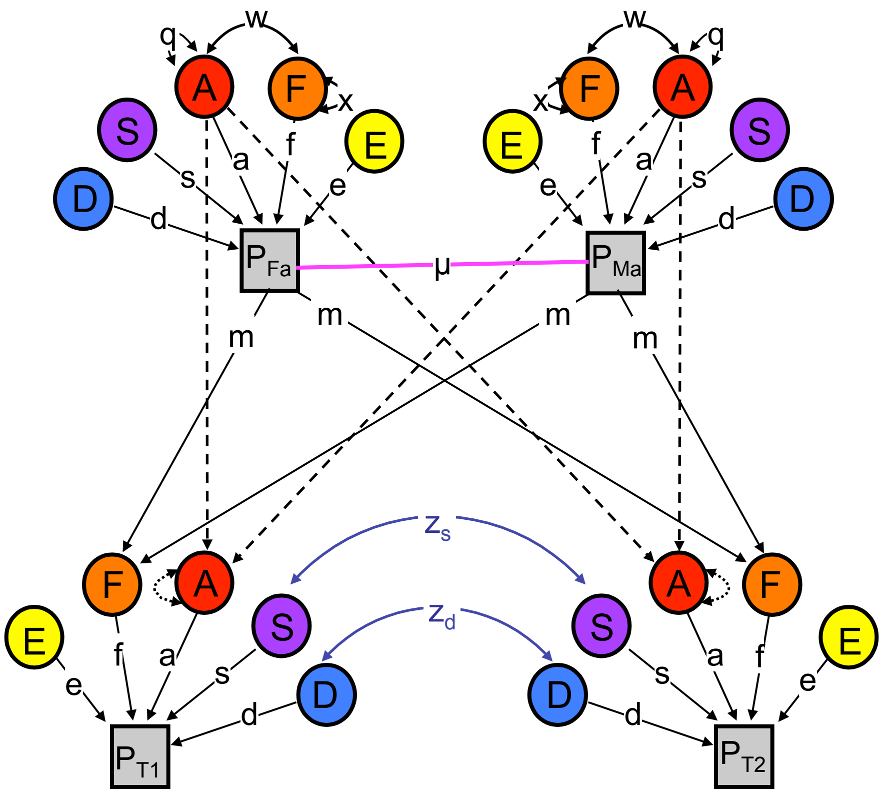
= env. factors shared only between **sibs**

= **familial** env factors passed from parents to offspring

But we can only estimate one of these (or more technically, one of VA, VS, VF, & VD)



Nuclear Twin Family Design (NTFD)



Note: m estimated and f fixed to 1

PRACTICAL 5: NTFD analysis

- In CTD.ACDE-param.indet_2018.R, run:
  FROM "# START PRACTICAL 5"
  TO "# END PRACTICAL 5"

- What are the estimated values of VA, VD, & VS? [Note: VS = sib environment, equivalent to VC in the CTD]

Simulated (true) vs. CTD vs. NTFD results

TRUE values	CTD estimates	NTFD estimates
VA = .30	*VA’* = .68	*VA’* = .32
VD = .30	*VD’* = .04	*VD’* = .29
VS = .10	*VS’* = 0	*VS’* = .13

Note: these are results from a single simulation. The estimates don’t equal the parameters here due to sampling variance. If we ran this a lot of times, NTFD estimates would be unbiased.

On average across 38 traits CTD vs. ETFD results*

- VA 65% higher in CTD
- VD 43% lower in CTD

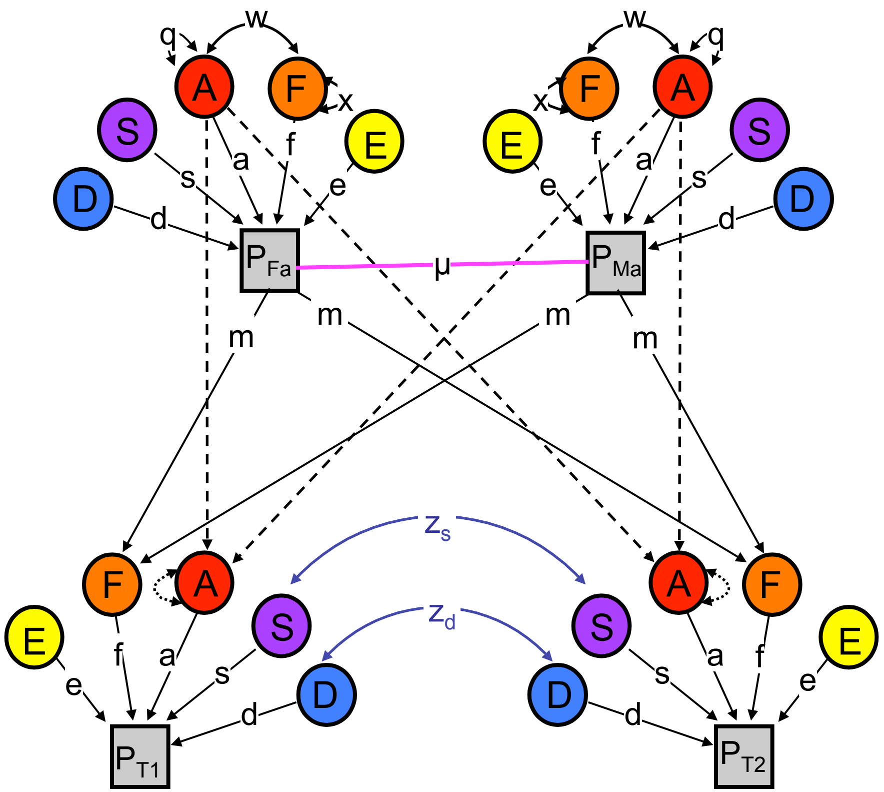
VG 18% higher in CTD

- VC 45% lower in CTD when r(spouse)~0
- VC 100% higher in CTD when r(spouse)>0

- ETFD results are not perfect, but theory and simulation suggest they are, on average, much more accurate than CTD results.
  - Accuracy across all sims: CTD=.14; NTF=.07; ETFD=.045

[*] Coventry & Keller, 2005

Nuclear Twin Family Design (NTFD)



Note: m estimated and f fixed to 1

- Assumptions:
  - Only can estimate 3 of 4: VA, VD, VS, and VF (bias is variable)
  - Assortative mating due to primary phenotypic assortment (bias is variable)

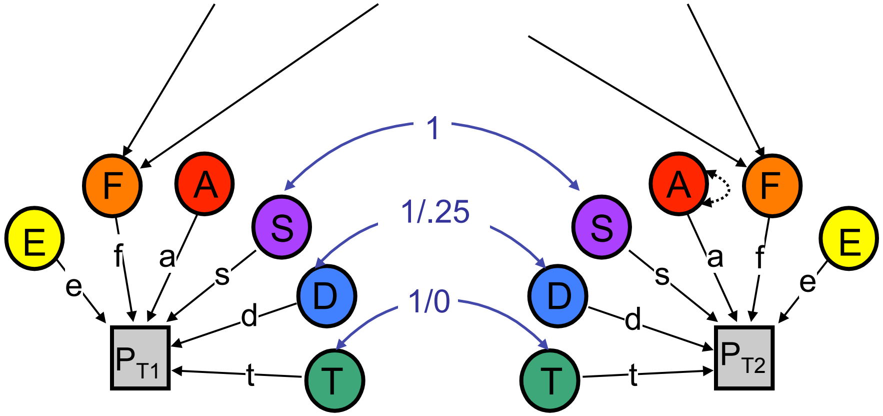
Stealth

- 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 VA, VS, VD, VF, VT

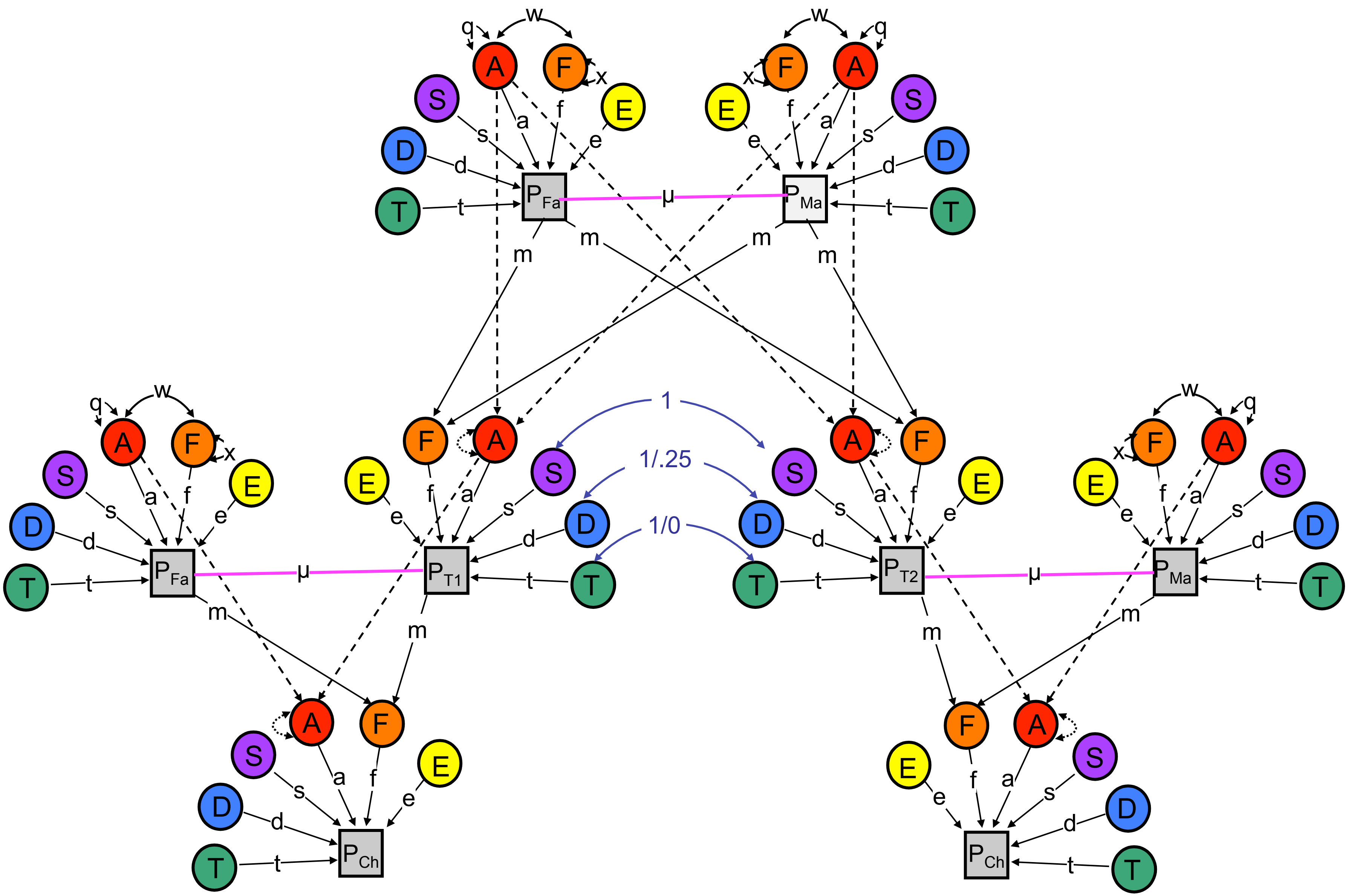
A S F D T can be estimated simultaneously

T = env. factors shared only between **twins**



(Remember: we're not just estimating more effects. More importantly, we're reducing the bias in estimated effects – although perhaps at the expense of more variance in estimates)

Stealth


w
q
A
F
x
S
E
D
T
s
a
f
e
d
t
P Fa
μ
P Ma
m
m
m
m
w
q
A
F
x
S
E
D
T
P Fa
F
A
E
S
D
T
f
a
e
s
d
t
P T1
1
1/.25
1/0
A
F
S
E
D
T
a
f
s
e
d
t
P T2
μ
P Ma
w
q
A
F
x
S
E
D
T
m
m
m
m
A
F
S
E
D
T
a
f
s
e
d
t
P Ch
F
A
E
S
D
T
f
a
e
s
d
t
P Ch

Stealth

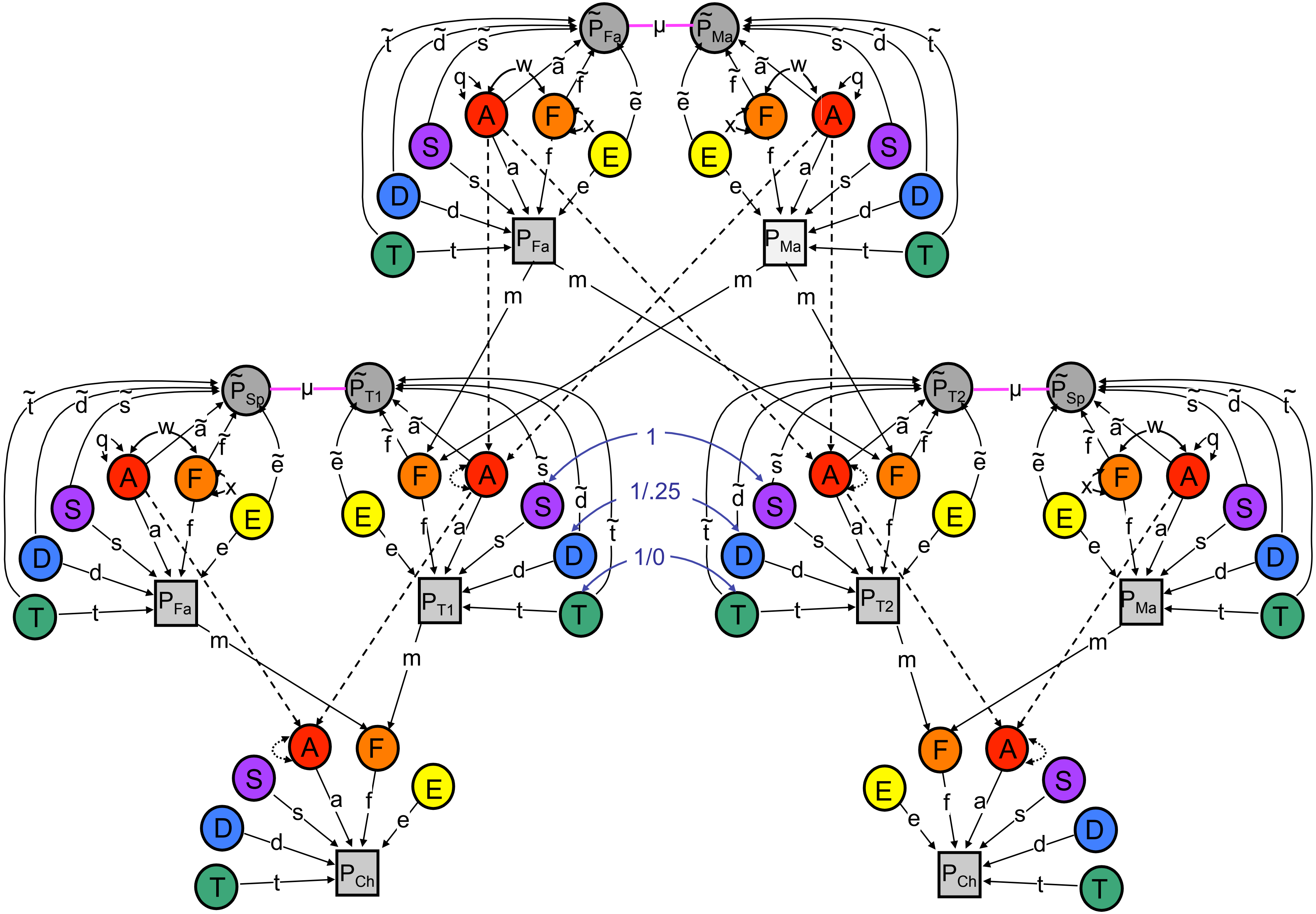
Assumption	biased up	biased down
Primary assortative mating	VA, VD, or VF	VA, VD, or VF
No epistasis	VA, VD	VS
No AxAge	VD, VS	VA

Stealth

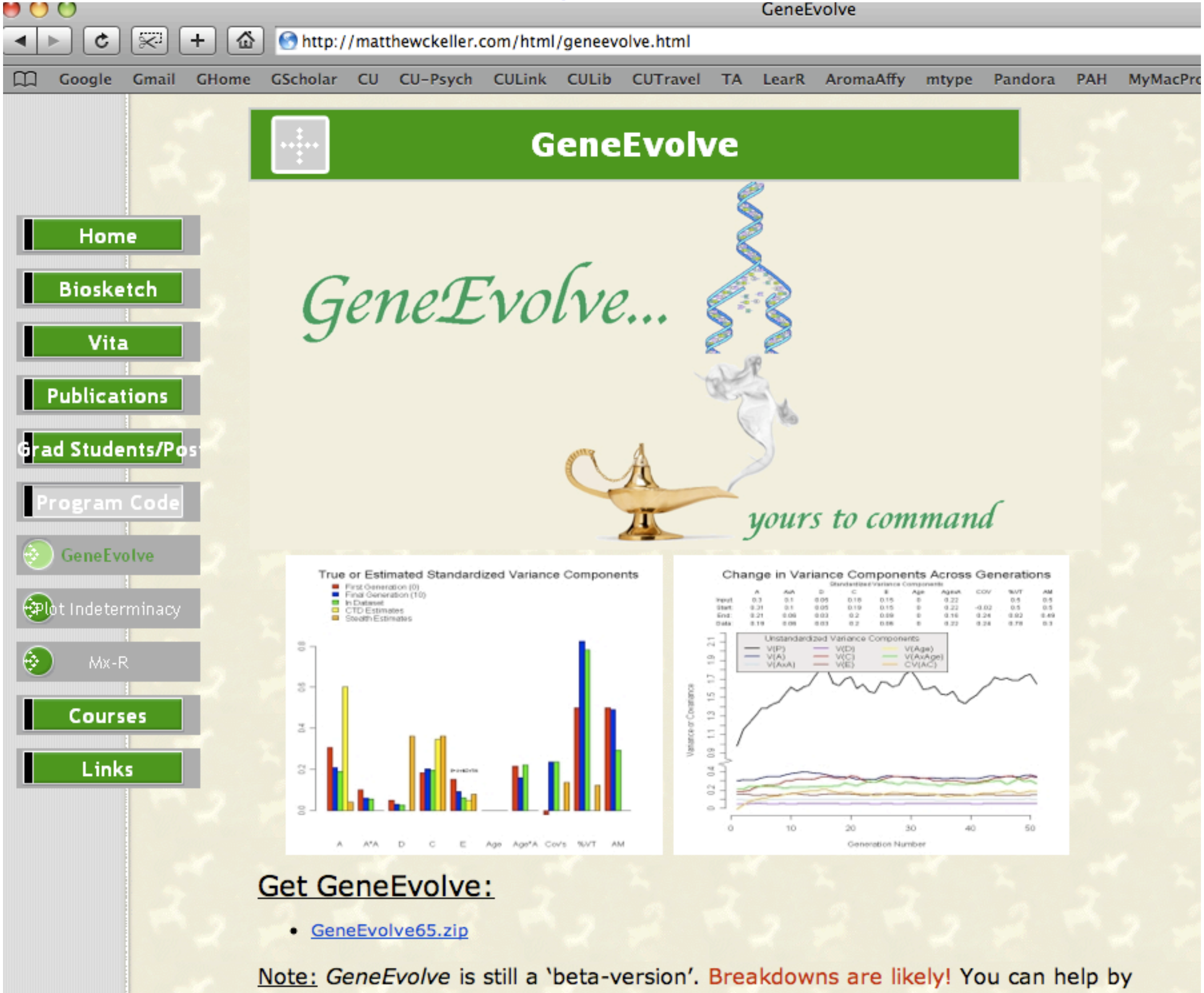
Assumption	biased up	biased down
Primary assortative mating	VA, VD, or VF	VA, VD, or VF
No epistasis	VA, VD	VS
No AxAge	VD, VS	VA

- 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 to each other (e.g., becoming more conservative when dating a conservative)

Cascade


P̃Fa
μ
P̃Ma
t̃
d̃
s̃
q
w
ã
f̃
ẽ
x
A
F
S
E
D
T
a
f
s
e
d
t
PFa
PMa
m
P̃Sp
P̃T1
P̃T2
PT1
PT2
1
1/.25
1/0
PCh

Simulation program: GeneEvolve



GeneEvolve

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

Google Gmail GHome GScholar CU CU-Psych CULink CULib CUTravel TA LearR AromaAffy mtype Pandora PAH MyMacPro

GeneEvolve

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

GeneEvolve... yours to command

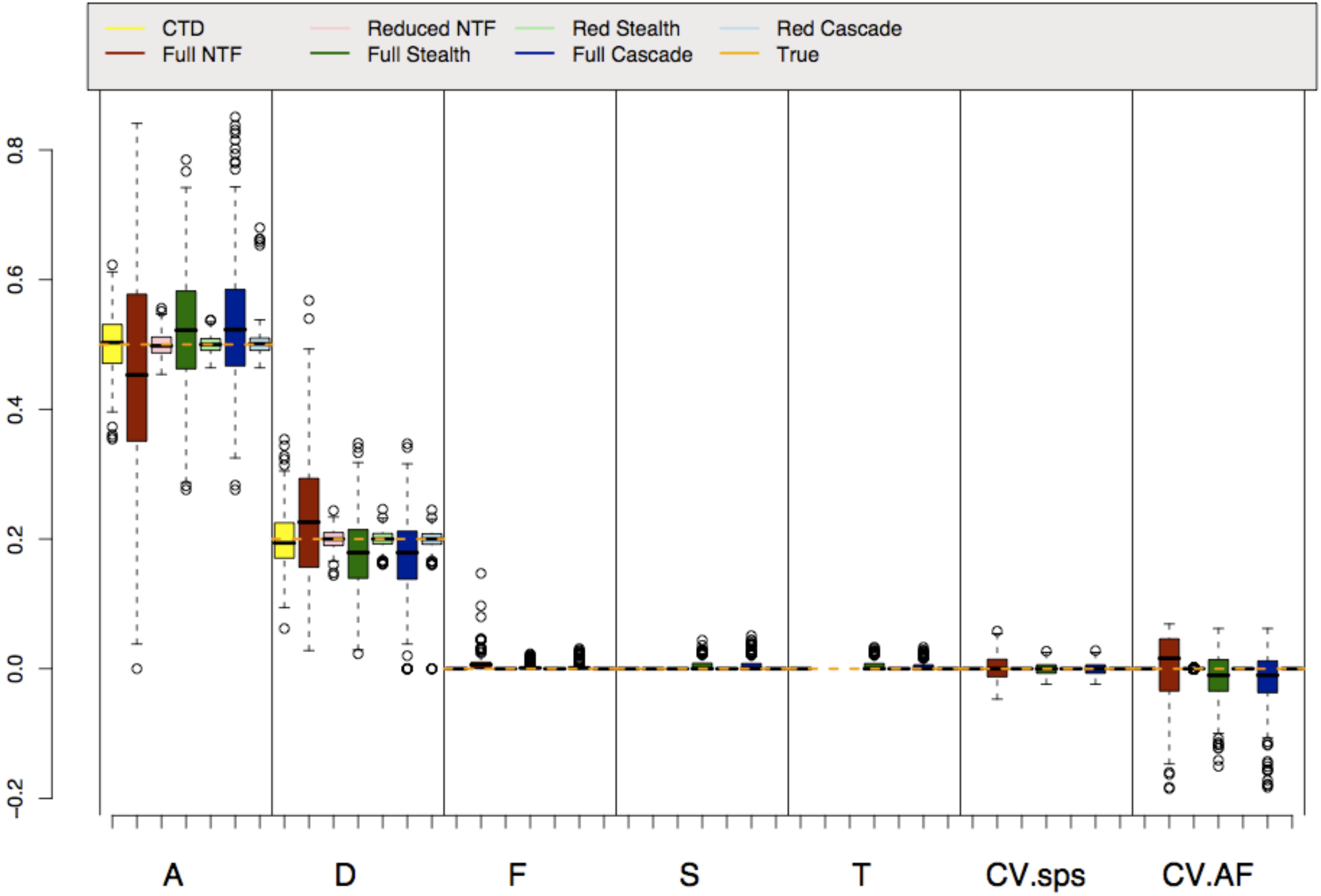
	A	AxA	D	C	E	Age	AgexA	COV	%VT	AM
Input:	0.3	0.1	0.05	0.18	0.15	0	0.22		0.5	0.5
Start:	0.31	0.1	0.05	0.19	0.15	0	0.22	-0.02	0.5	0.5
End:	0.21	0.06	0.03	0.2	0.09	0	0.16	0.24	0.82	0.49
Data:	0.19	0.06	0.03	0.2	0.06	0	0.22	0.24	0.78	0.3

Get GeneEvolve:

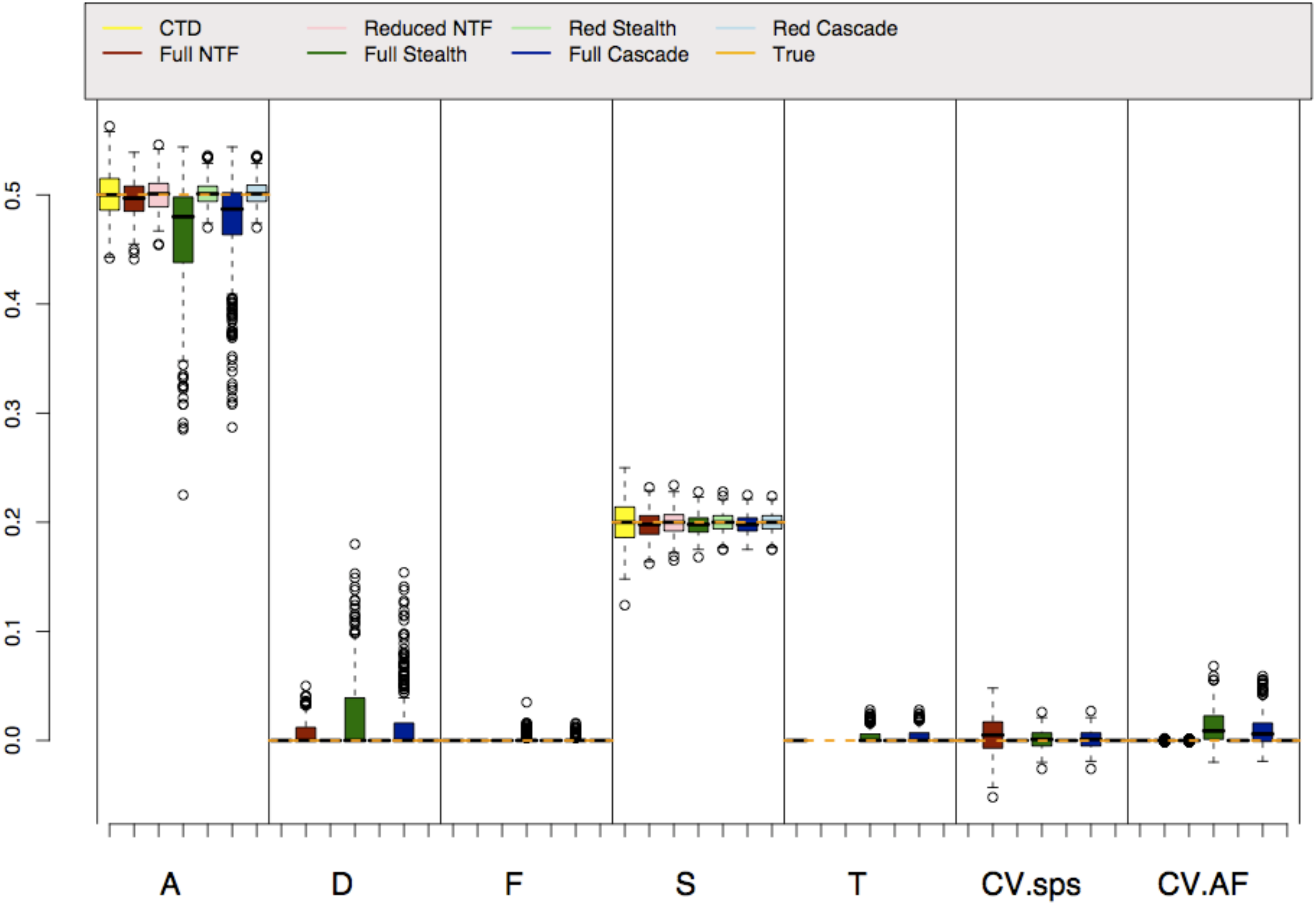
- GeneEvolve65.zip

Note: *GeneEvolve* is still a 'beta-version'. Breakdowns are likely! You can help by

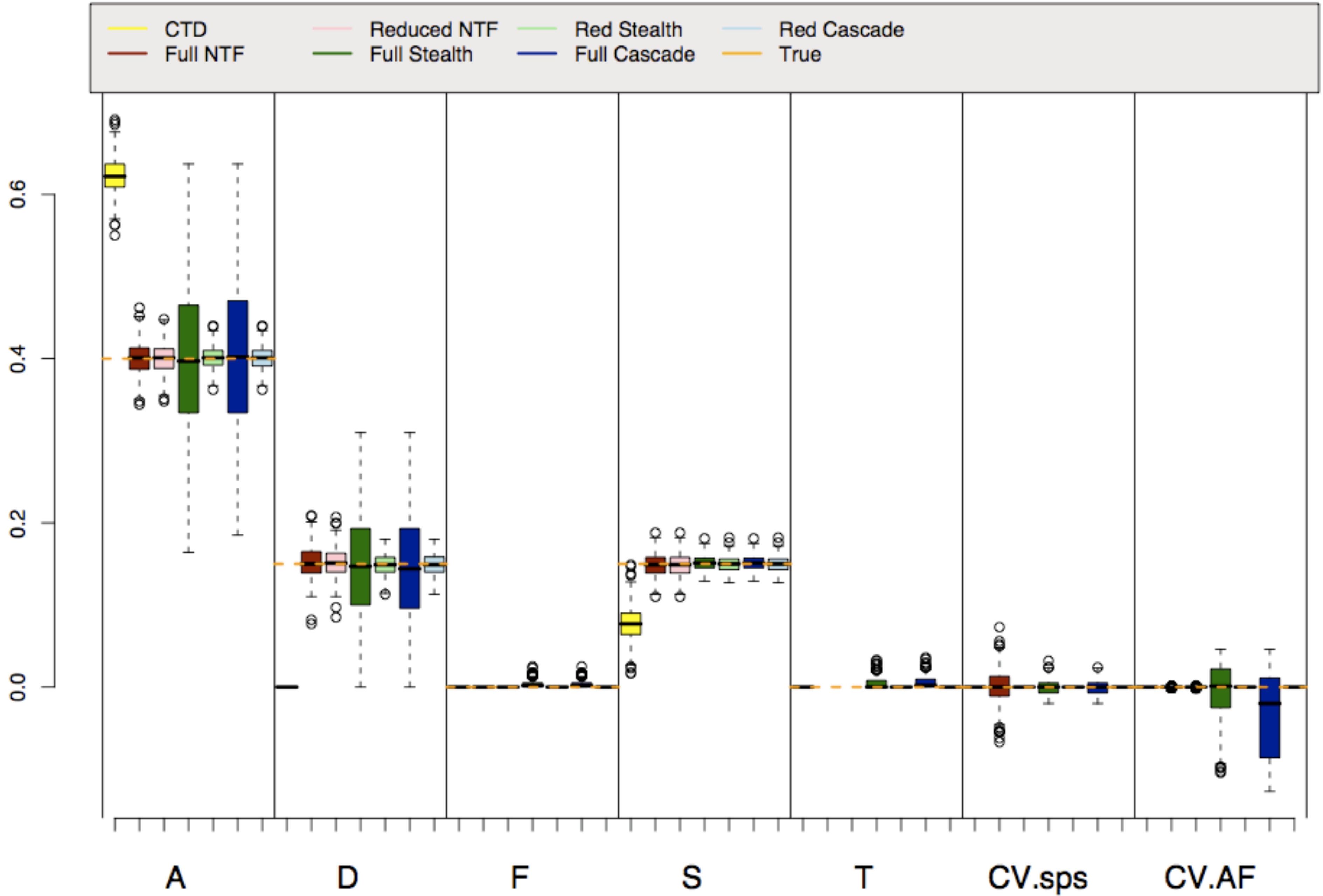
Reality: VA=.5, VD=.2



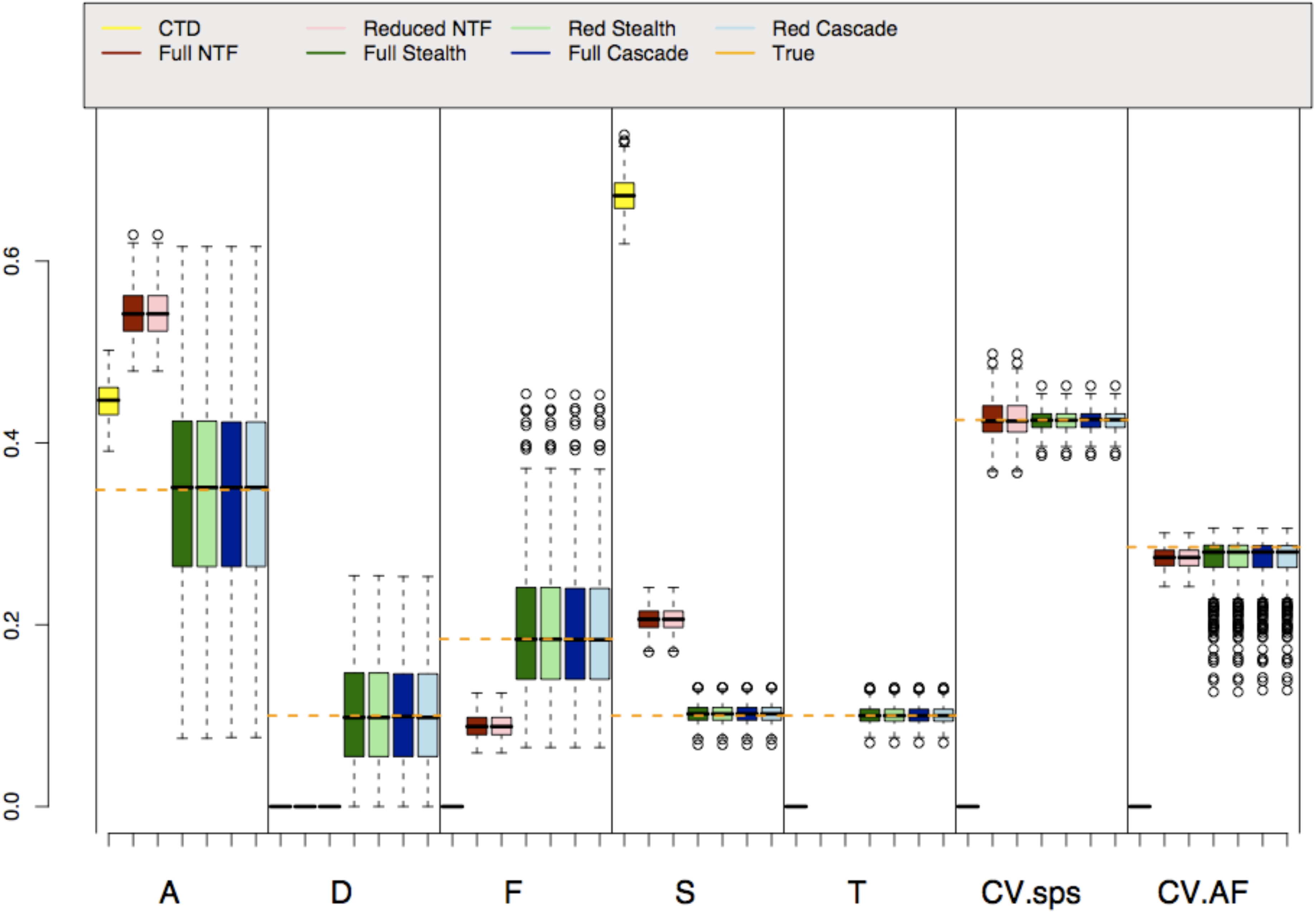
Reality: VA=.5, VS=.2



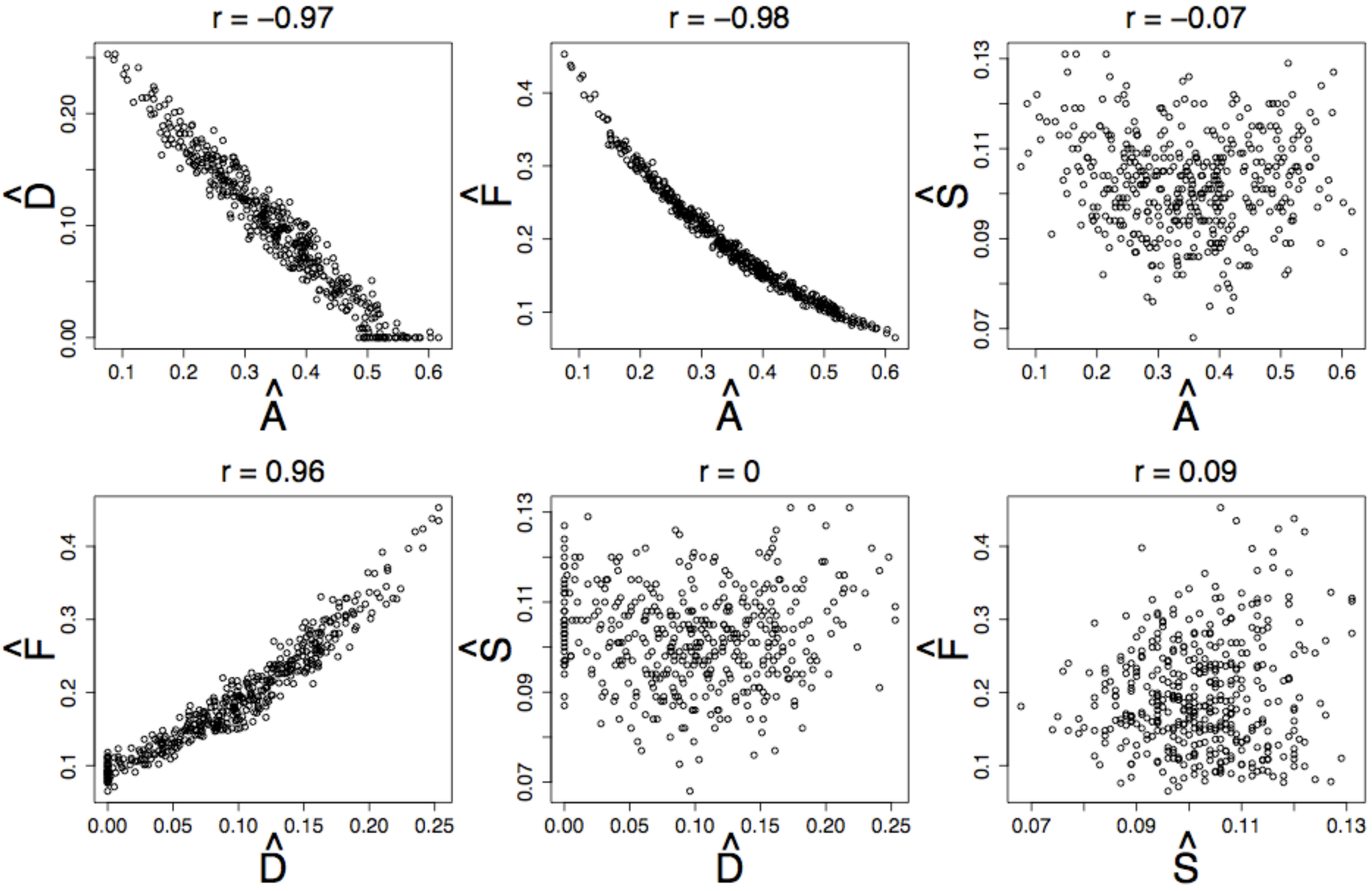
Reality: VA=.4, VD=.15, VS=.15



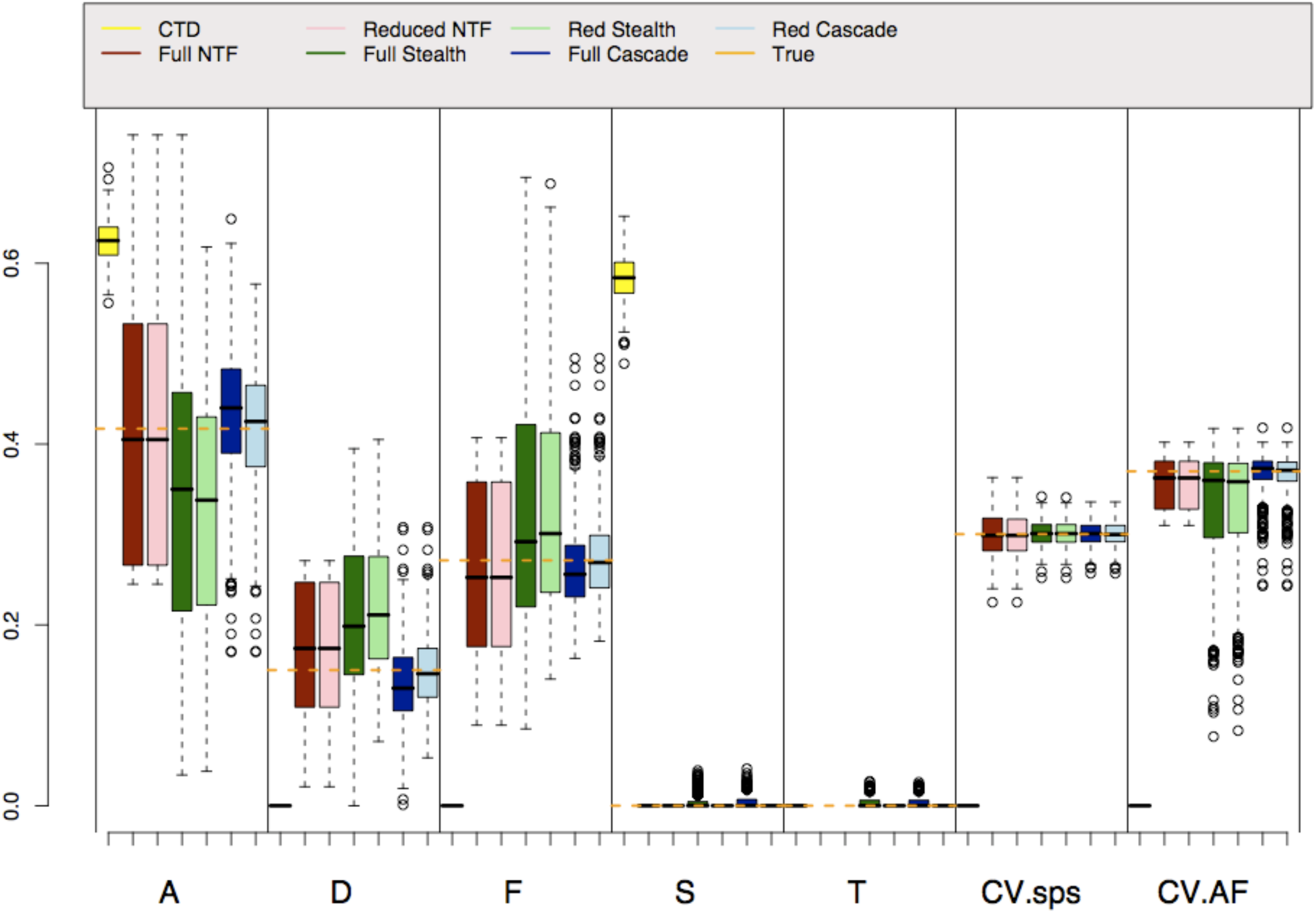
Reality: VA=.35, VD=.15, VF=.2, VS=.15, VT=.15, AM=.3



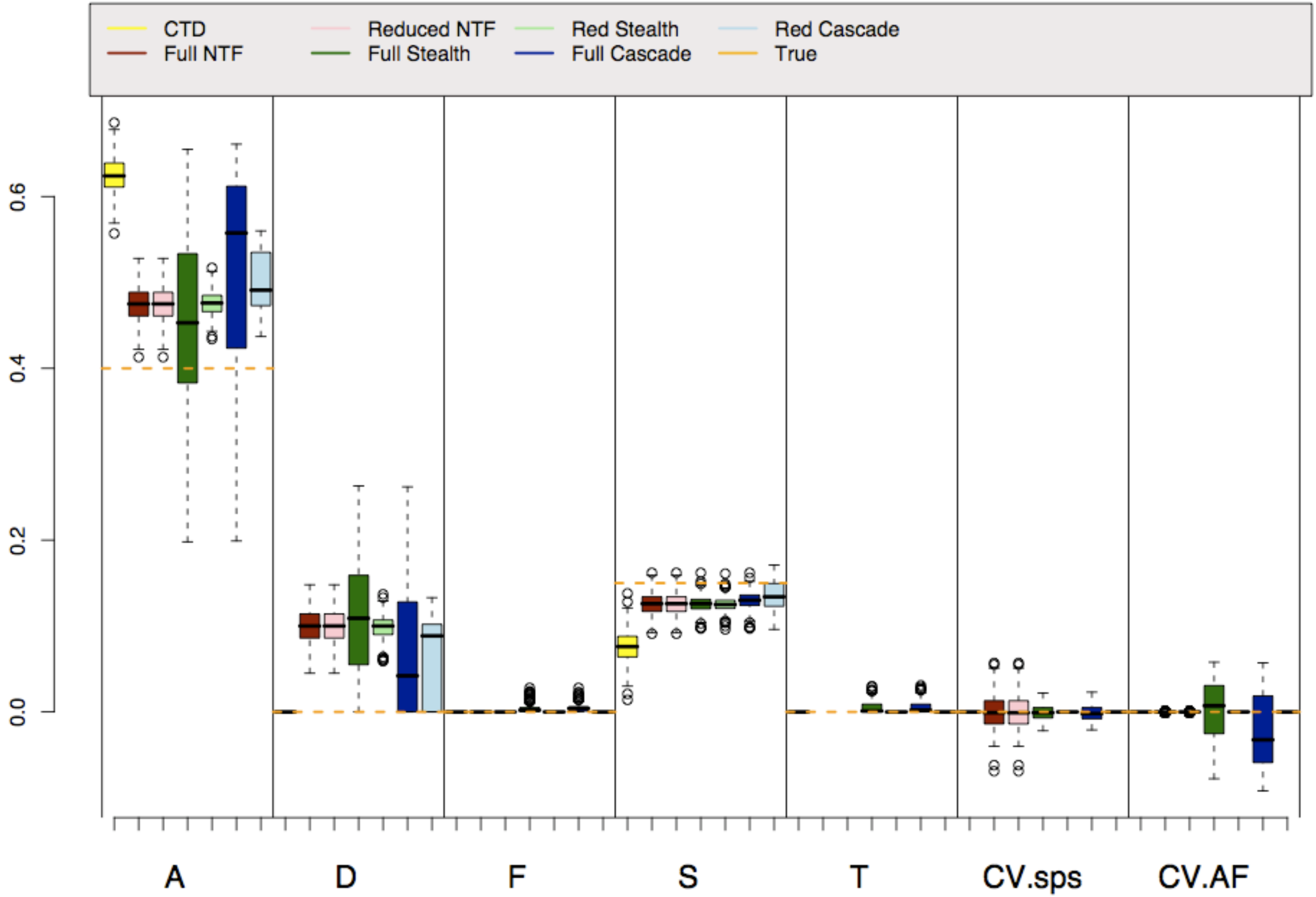
VA,VD, & VF estimates are highly correlated in Stealth & Cascade



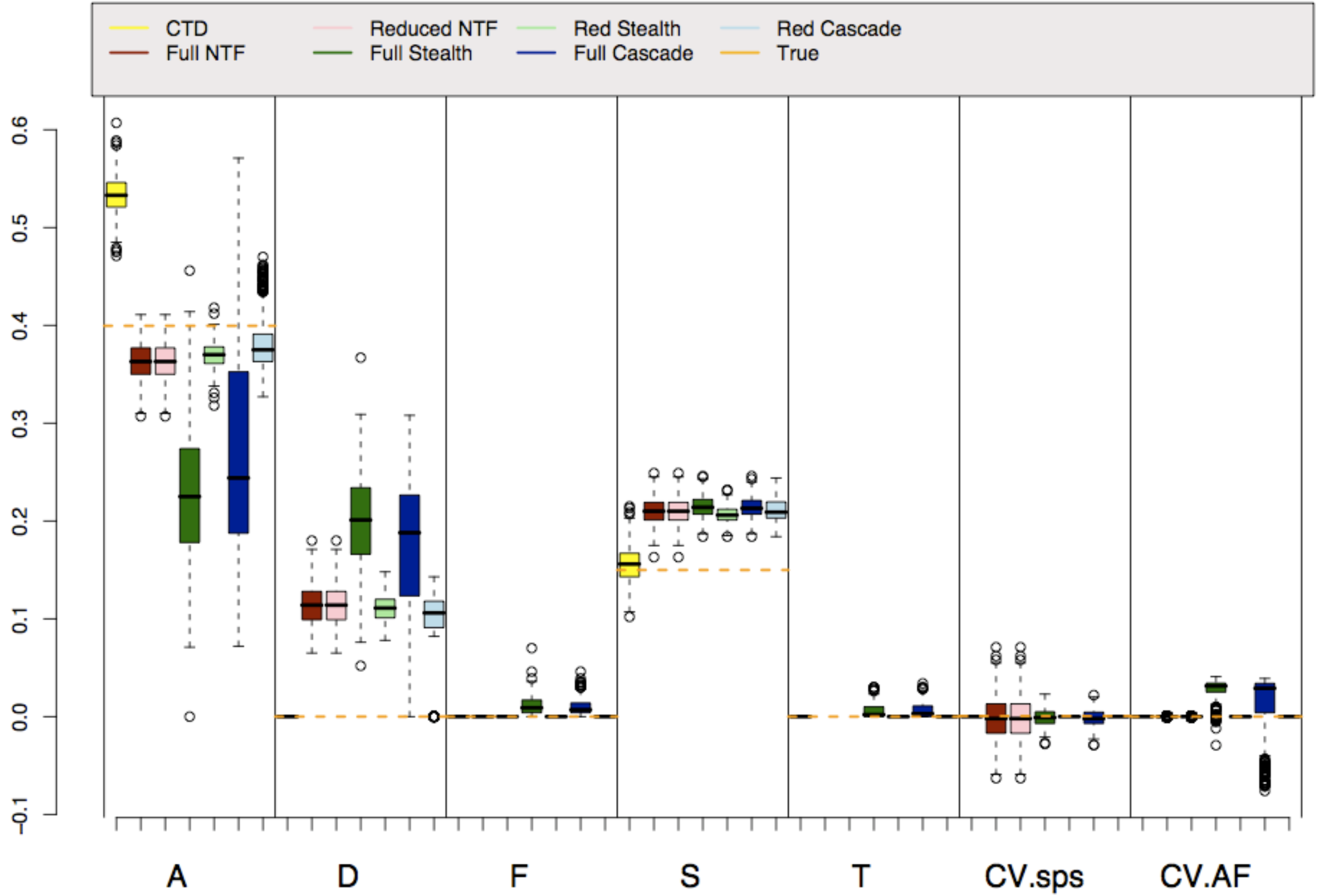
Reality: VA=.45, VD=.15, VF=.25, AM=.3 (Soc Hom)



Reality: VA=.4, VA*A=.15, VS=.15



Reality: VA=.4, VA*Age=.15, VS=.15



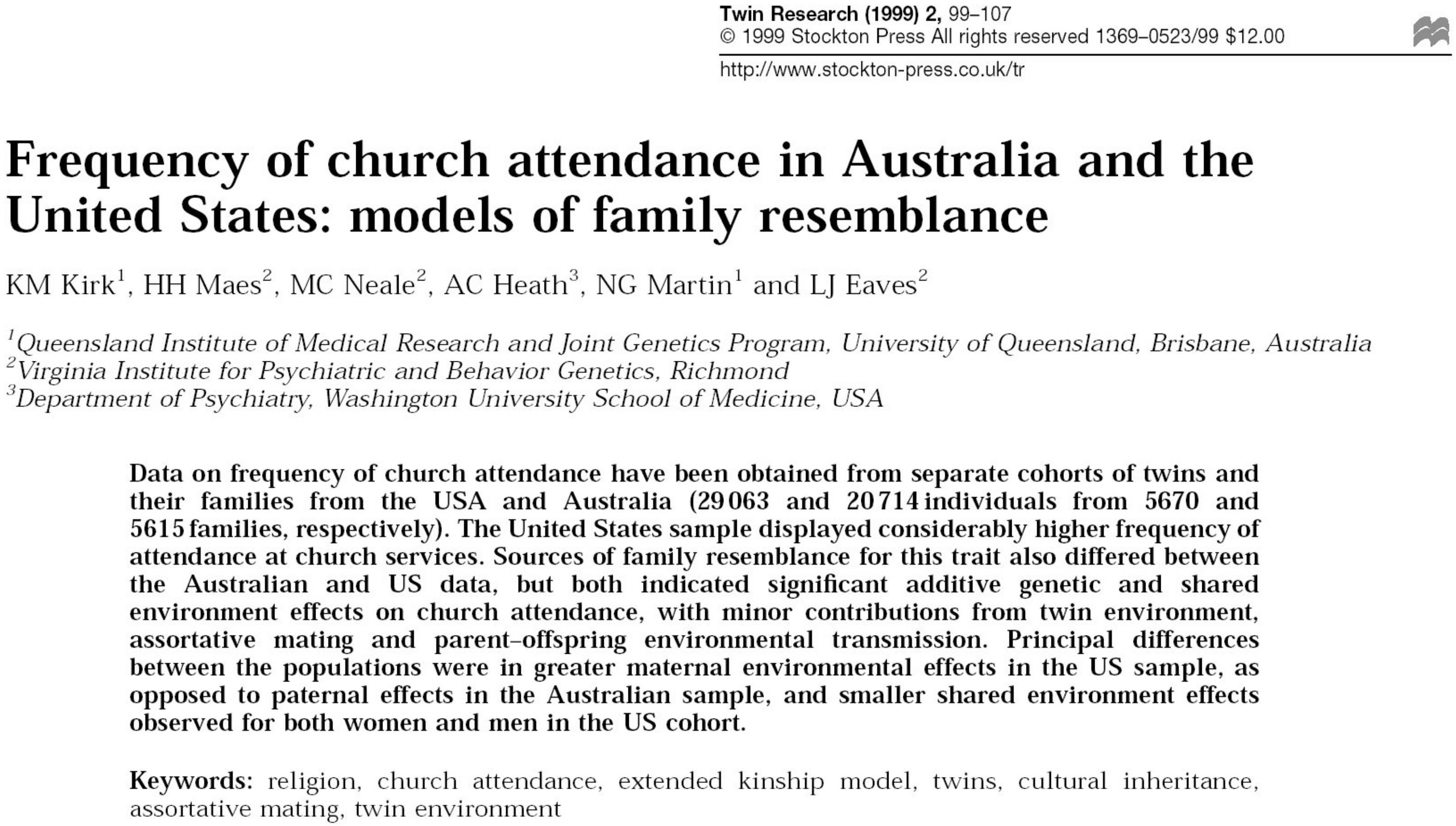
Conclusions

- All models require assumptions. Generally, more assumptions = more biased estimates
- Simulations provide independent assessments of the NTFD, *Stealth*, and *Cascade* models
  - These complicated models work as designed, but they have drawbacks
- In all models, but especially the CTD, be cautious of reifying parameter estimates!
  - *VA'* is amalgam of mostly VA but also VD & VC. *VA'/VP'* (in ACE models) or (*VA'+VD')/VP'* (in ADE models) is a decent estimate of <u>broad sense </u>h^2.
  - *VD'* & *VC'* are likely to be underestimates

Discussion questions

- Are ETFDs worth the trouble? Or should we simply adjust our interpretations of estimates from simpler models?
- How well do methods work that rely on skewness to fit *VA'*, *VD'*, and *VC'* simultaneously work?
- Should we report full or reduced parameter estimates?
- Should we fit variances of latent variables rather than pathways, and hence allow variance component estimates to go negative?

Stealth application



Twin Research (1999) 2, 99–107
© 1999 Stockton Press All rights reserved 1369–0523/99 $12.00
http://www.stockton-press.co.uk/tr

Frequency of church attendance in Australia and the United States: models of family resemblance

KM Kirk[1], HH Maes[2], MC Neale[2], AC Heath[3], NG Martin[1] and LJ Eaves[2]

[1] *Queensland Institute of Medical Research and Joint Genetics Program, University of Queensland, Brisbane, Australia*
[2] *Virginia Institute for Psychiatric and Behavior Genetics, Richmond*
[3] *Department of Psychiatry, Washington University School of Medicine, USA*

Data on frequency of church attendance have been obtained from separate cohorts of twins and their families from the USA and Australia (29 063 and 20 714 individuals from 5670 and 5615 families, respectively). The United States sample displayed considerably higher frequency of attendance at church services. Sources of family resemblance for this trait also differed between the Australian and US data, but both indicated significant additive genetic and shared environment effects on church attendance, with minor contributions from twin environment, assortative mating and parent–offspring environmental transmission. Principal differences between the populations were in greater maternal environmental effects in the US sample, as opposed to paternal effects in the Australian sample, and smaller shared environment effects observed for both women and men in the US cohort.

Keywords: religion, church attendance, extended kinship model, twins, cultural inheritance, assortative mating, twin environment

Further reading on this lecture

- Eaves LJ, Last KA, Young PA, Martin NG (1978) Model-fitting approaches to the analysis of human behaviour. *Heredity* 41:249-320
- Fulker DW (1982) Extensions of the classical twin method. Human Genetics. Part A: The Unfolding Genome (Progress in Clinical and Biological Research Vol 103A). p. 395-406
- Fulker DW (1988) Genetic and cultural transmission in human behavior. Proceedings of the Second International conference on Quantitative Genetics
- Eaves LJ, Heath AC, Martin NG, Neale MC, Meyer JM, Silberg JL, Corey LA, Truett K, Walter E (1999) Comparing the biological and cultural inheritance of stature and conservatism in the kinships of monozygotic and dizygotic twins. In: Cloninger CR (Ed) Proceedings of 1994 APPA Conference. p. 269-308
- Keller MC & Coventry WL (2005). Quantifying and addressing parameter indeterminacy in the classical twin design. *Twin Research and Human Genetics,* 8, 201-213
- Keller MC, Medland SE, Duncan LE, Hatemi PK, Neale MC, Maes HHM, Eaves LJ. Modeling extended twin family data I: Description of the Cascade Model. *Twin Research and Human Genetics*, 29, 8-18.
- Keller MC, Medland SE, & Duncan LE (2010). Are extended twin family designs worth the trouble? A comparison of the bias, precision, and accuracy of parameters estimated in four twin family models. *Behavior Genetics.*