And now….

…for something completely different

(Monty Python)

The Post-Genomic Heuristic

$$
\begin{aligned}
&\text{T.B}(_{\text{s}}) \quad = \\
&\Sigma\big(\text{Gcrap} + \text{Eshit} \\
&\qquad + \text{GcrapxEshit}\text{) dGrap dEshit} \\
&\qquad\qquad + \text{Godnoeswot}
\end{aligned}
$$

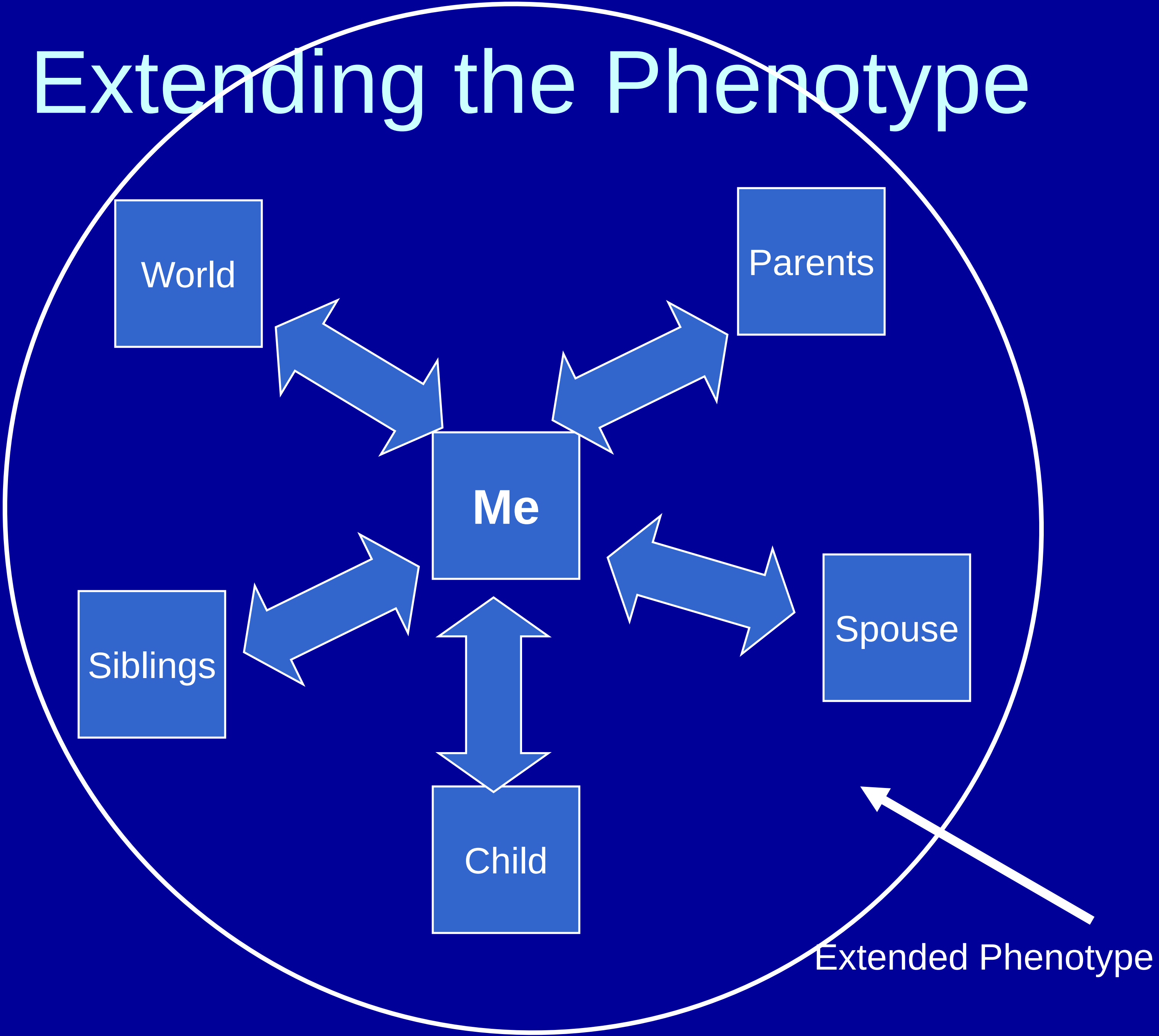
OR:

It’s pretty complicated

Seriously, though,

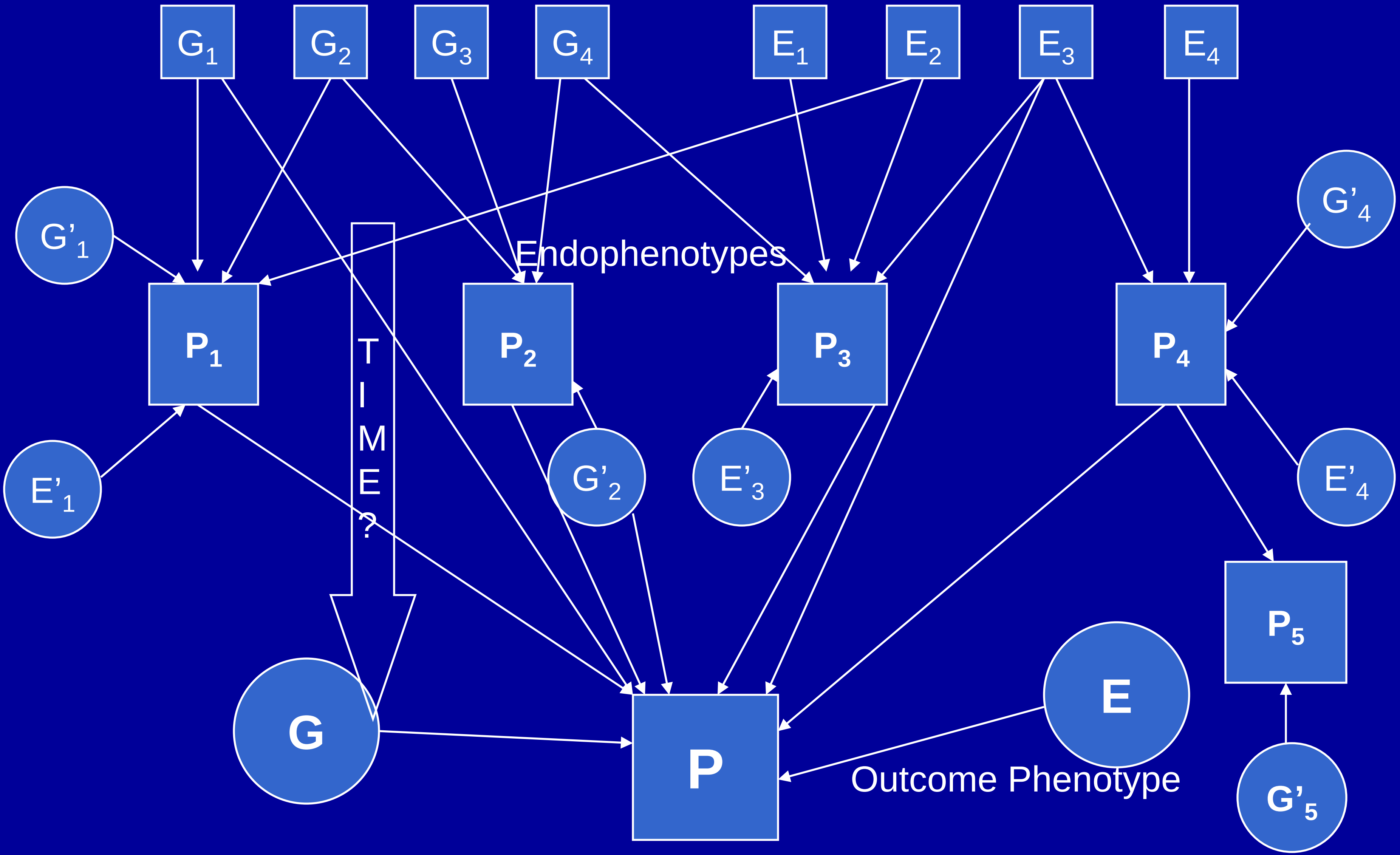
“It’s been a great week”

Extending the Phenotype



Measured Genotypes

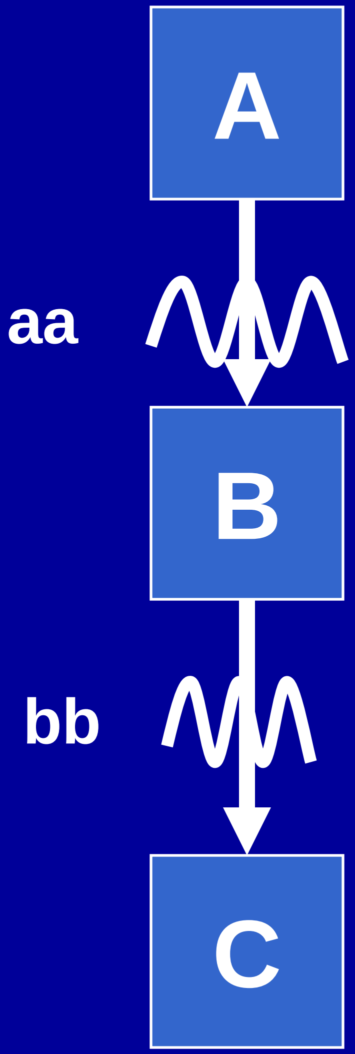
Measured Environments



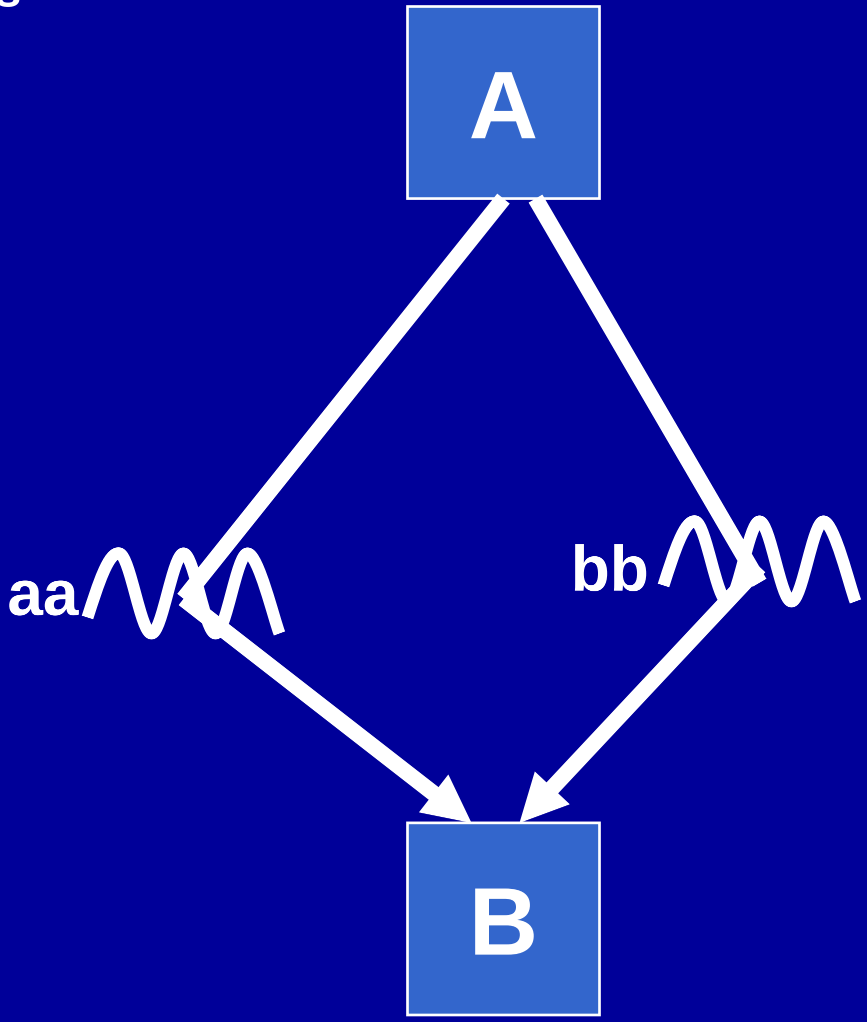
Endophenotypes

Outcome Phenotype

Sequential (“complementary”) genes

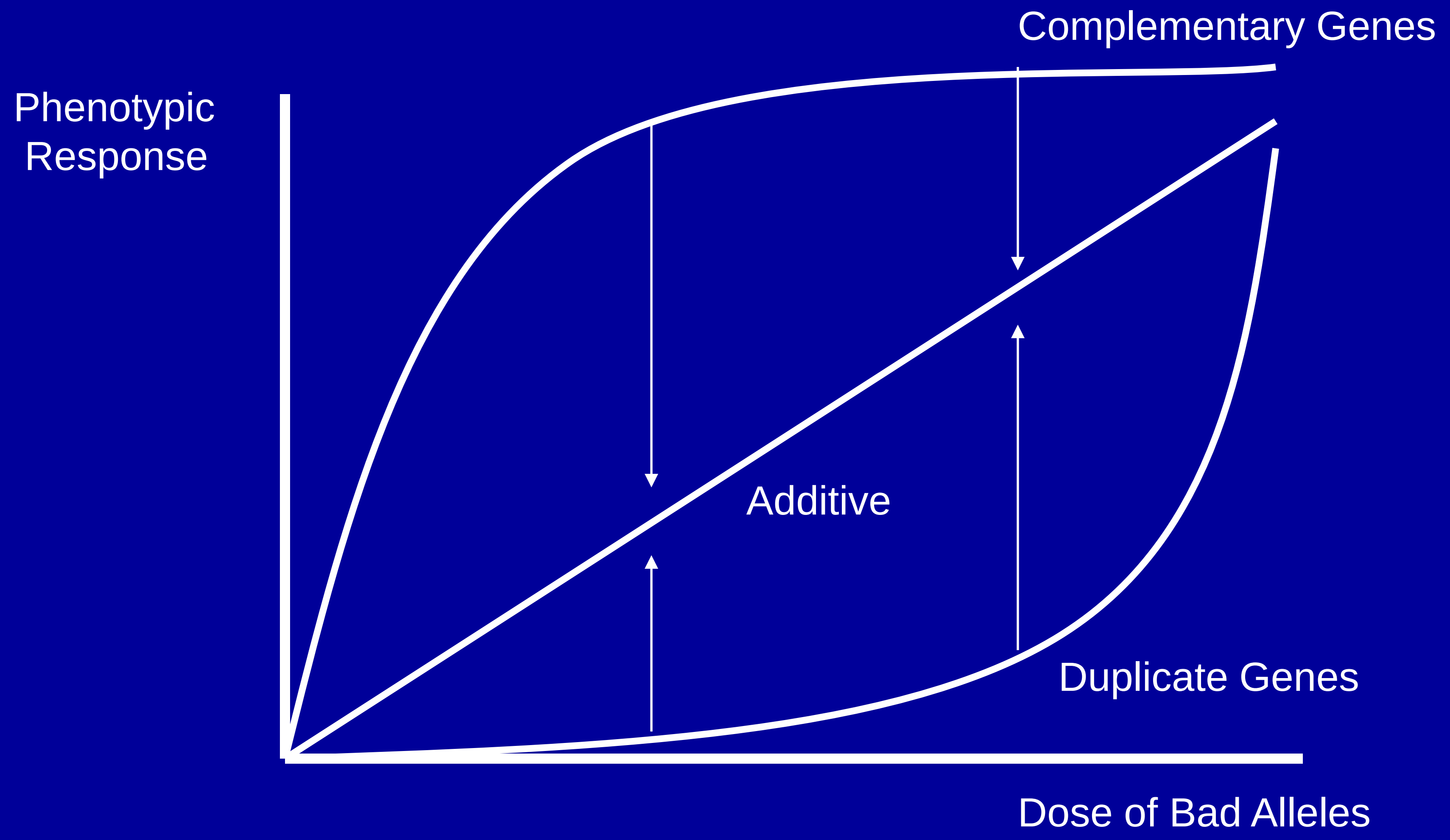


Parallel (“duplicate”) genes



= Pathway blocked by mutant gene

Combining pathways



Introduction to BUGS in Genetic Epidemiology

“**B**ayesian Inference **U**sing **G**ibbs **S**ampling”

Lindon Eaves

Boulder, March 2009.

Thanks

Allattin Erkanli

Nick Martin

Staff and students QIMR

Critical Source:
MRC BUGS Project

http://www.mrc-bsu.cam.ac.uk/bugs/

OpenBUGS

http://mathstat.helsinki.fi/openbugs/Home.html

Apology

I don’t know much and probably don’t know what I am talking about. But I hope others will see possibilities and help deepen understanding.

Why bother?

- Intellectual challenge. Different ("Bayesian") way of thinking about statistics and statistical modeling.
- "Model Liberation" – can do a bunch of cool stuff much more easily
- Learn more about data for less computational effort
- "Fast"

Payoff

- Estimate parameters of complex models
- Obtain subject parameters (e.g. “genetic and environmental factor scores”) at no extra-cost - estimating scores for GWAS
- Obtain confidence intervals and other summary statistics (s.e’s, quantiles etc) at no extra cost.
- Automatic imputation of missing data (“data augmentation”)
- Fast (35 item, IRT in 500 twin pairs with covariates takes about 1-2 hours on laptop).
- Insight, flexibility

Generally:

Seems to help with models that require multi-dimensional integration to compute likelihood.

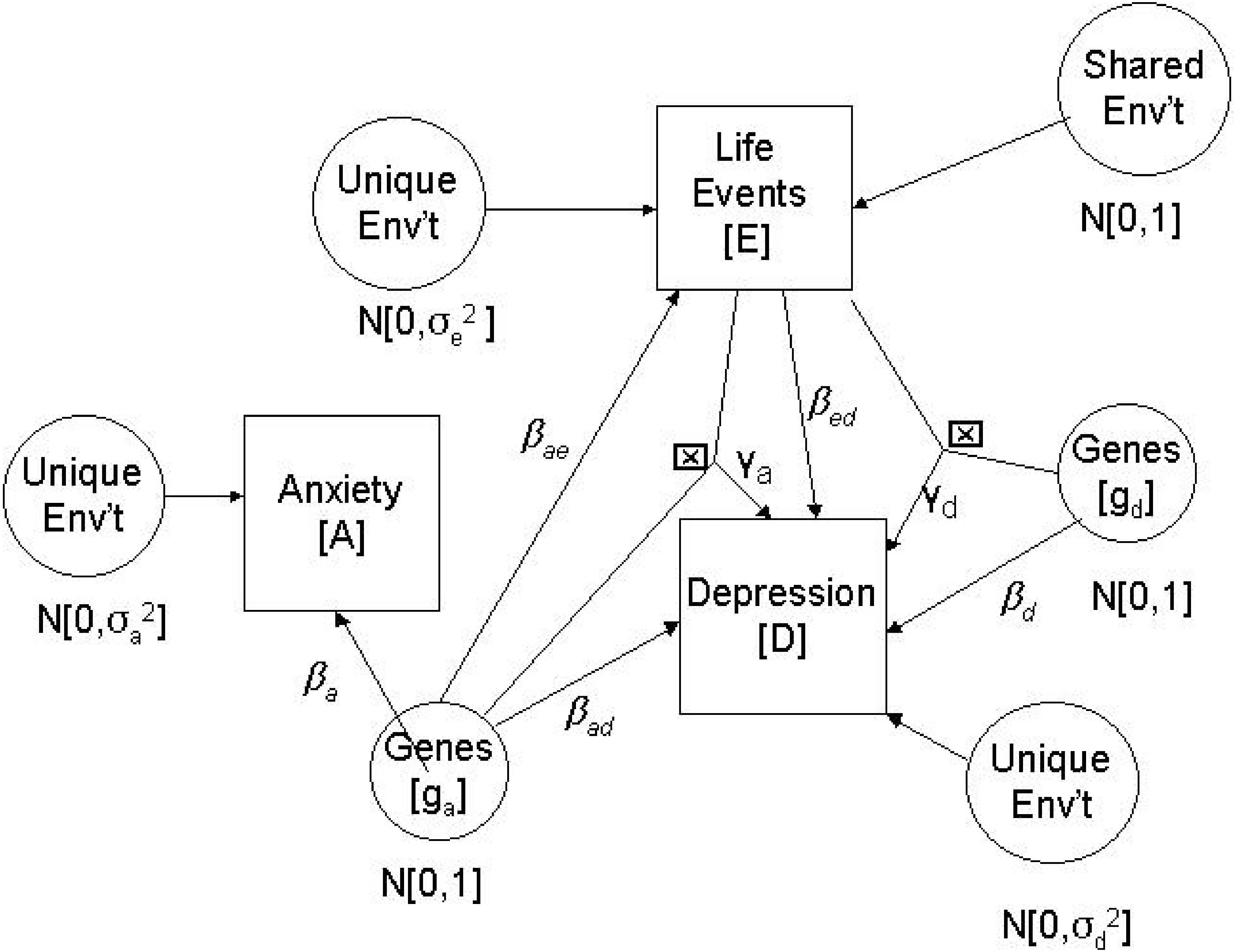
Some applications

- Non-linear latent variables (GxE interaction).
- Multidimensional, multi-category, multi-item IRT in twins.
- Genetic effects on developmental change in multiple indicators of puberty (latent growth curves).
- Hierarchical mixed models for fixed and random effects of G, E and GxE in multi-symptom (“IRT”) twin data.
- Genetic survival models
- Mixture models

Resolving multiple epigenetic pathways to adolescent depression

Lindon Eaves,1 Judy Silberg,1 and Alaattin Erkanli2
1Virginia Institute for Psychiatric and Behavioral Genetics, USA;
2Epidemiology Program, Duke University Medical Center, USA

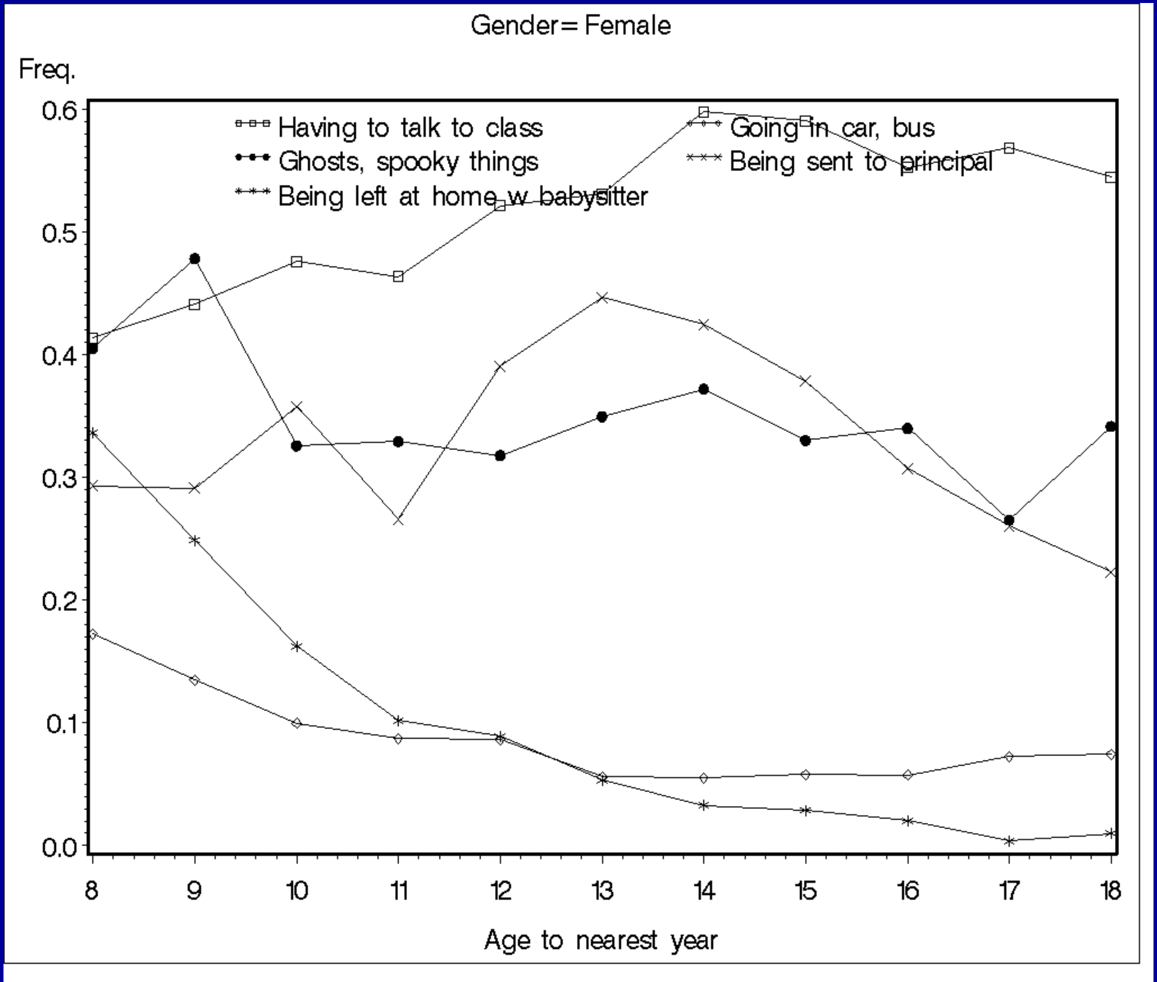
Journal of Child Psychology and Psychiatry 44:7 (2003), pp 1006–1014


Shared Env't
N[0,1]
Unique Env't
N[0,σ_e^2]
Life Events [E]
β_ae
β_ed
γ_a
γ_d
Unique Env't
N[0,σ_a^2]
Anxiety [A]
Genes [g_d]
N[0,1]
Depression [D]
β_d
β_a
β_ad
Genes [g_a]
N[0,1]
Unique Env't
N[0,σ_d^2]

Developmental-genetic effects on level and change in childhood fears of twins during adolescence

Lindon J. Eaves and Judy L. Silberg
Virginia Institute for Psychiatric and
Behavioral Genetics, Department of
Human and Molecular Genetics,
Virginia Commonwealth University,
Richmond, VA, USA

Journal of Child Psychology and Psychiatry
:* (2008), pp **–


Gender= Female
Freq.
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Having to talk to class
Ghosts, spooky things
Being left at home w babysitter
Going in car, bus
Being sent to principal
8
9
10
11
12
13
14
15
16
17
18
Age to nearest year

RJMCMC

“Reversible Jump MCMC”

Samples models as well as parameters – ranks posterior probabilities of models|data

RJMCMC

- Variable selection – developing “best” model for patterns of interaction between top candidates’ effects on phenotype

- Ranking alternative networks of pathways between multiple “endophenotypes” (genes, microarrays, proteins, metabolites, ROIs, environments…etc.)

References

Lunn, D. J., Best, N. and Whittaker, J. (2008) Generic reversible jump MCMC using graphical models, *Statistics and Computing*, DOI: 10.1007/s11222-008-9100-0

Lunn, D. J., Whittaker, J. C. and Best, N. (2006) A Bayesian toolkit for genetic association studies, *Genetic Epidemiology* **30**: 231-247.

This introduction

- Introduce ideas
- Practice use of WinBUGS
- Run some basic examples
- Look at application to genetic IRT
- Other stuff?

Some references:

Gilks WR, Richardson S, Spiegelhalter DJ (1996)
Markov Chain Monte Carlo in Practice. Boca Raton,
Chapman & Hall,

Gelman A, Carlin JB, Stern HS, Rubin DB. (2004)
Bayesian Data Analysis (2nd Ed,) Boca Raton,
Chapman & Hall.

Spiegelhalter DJ, Thomas A, Best N, Lunn D. (2004). *WinBUGS User Manual Version 1.4.1.* Cambridge, England. MRC BUGS project. [Downloaded with WinBUGS – also Examples Vols. I and II]

Maris, G and Bechger, T.M. (2005). An Introduction to the DA-T Gibbs Sampler for the Two-Parameter Logistic (2PL) Model and Beyond. *Psicol´ogica:* ***26****, 327-352.*

http://www.uv.es/~revispsi/articulos2.05/8-MARIS.pdf

Basic Ideas

- Bayesian Estimation (vs. ML)
- “Monte Carlo”
- “Markov Chain”
- Gibbs sampler

(Maximum) Likelihood

- Compute (log-) likelihood of getting data given values of model parameters and assumed distribution
- Search for parameter values that maximize likelihood (“ML” estimates)
- Compare models by comparing likelihoods
- Obtain confidence intervals by contour plots (i.e. repeated ML conditional on selected parameters)
- Obtain s.e.’s by differentiating L

Problem with ML

- Many models require integration over values of latent variables (e.g. non-linear random effects)

- Integrate to evaluate each likelihood and derivatives for each parameter

- “Expensive” when number of dimensions is large (?days), especially for confidence intervals.

Maximum Likelihood (ML) “Thinks” (*theoretically*) about parameters and data separately: P(data|parameters)

“Thinks” (*practically*) of integration, searching and finding confidence intervals as separate numerical problems (quadrature, e.g. Newton-Raphson, numerical differentiation).

Markov Chain Monte Carlo (MCMC, MC^2) “Thinks” (*theoretically)* that there is no difference between parameters and data – seeks *distribution* of parameters given data – P(parameters|data) {*Bayesian* estimation}

“Thinks” *(practically)* that integration, search and interval estimation constitute a single process addressed by a single unifying algorithm {*Gibbs Sampling*}

“Parameter”

Anything that isn’t data: means, components of variance, correlations. But also subjects’ scores (not just distributions) on latent traits (genetic liabilities, factor scores), missing data points.

Basic approach

- "Bayesian" = Considers joint distribution of parameters and data
- "Monte Carlo" = Simulation
- "Markov Chain" = Sequence of simulations ("time series") designed to converge on samples from posterior distribution of θ given *D*
- "Gibbs sampler" = method of conducting simulations – cycles through all parameters simulating new value of parameter conditional on *D* and every other parameter

“Bayesian”

- Considers joint distribution of all parameters (θ) and data (*D*): P(θ.D)
- Seeks “posterior distribution” of θ given *D:*

  P(θ|*D*)
- Need to know “prior” distribution P(θ), but don’t.
- Start out by assuming some prior distribution (“uniformative” priors – i.e. encompassing wide range of possible parameter values) and seek to refine using data.

“Monte Carlo”

- Computer simulation of unknown parameters (“nodes”) from assumed distribution (“computer intensive”).
- If distribution is assumed [e.g. mean and variance] then successive simulations represent samples from the assumed distribution i.e. Can estimate “true” distribution from large number of (simulated) samples – can get any properties of distribution (means, s.d.s, quantiles) to any desired degree of precision.

“Markov Chain”

- “True” prior distribution P(θ) unknown.
- “Markov Chain” – series of outcomes (e.g. sets of data points) each contingent on previous outcome – under certain conditions reach a sequence where underlying distribution does not change (“stationary distribution”).
- Start with assumed prior distribution and construct (simulate) Markov Chain for given *D* that converges to samples from posterior distribution: P(θ|D).
- Then use (large enough) set of samples from stationary distribution to characterize properties of desired posterior distribution.

“Gibbs Sampler”

- Algorithm for generating Markov chains from multiple parameters conditional on *D*.
- Takes each parameter in turn and generates new value conditional on the data and every other parameter. Cycle through all parameters (“one iteration”) and repeat until converge to stationary distribution.

WinBUGS

- Free
- Simple language (“R-like”)
- [“Open” Version: “Open BUGS”]
- PC version (BUGS also available for mainframe)
- Graphical interface (OK for beginners, but confining and usually easier to write code)
- Well documented, good examples

Problems

- Convergence criteria (“mixing”)
- Model comparison
- ? Sensitivity to priors
- Model identification
- Error messages sometimes obscure
- Data set-up can be a pain
- Can have problems (Latent class analysis)

Bottom line

If you can figure how you would simulate it, you can probably “BUGS-it.”

Need to be clear and explicit about model and assumptions.

Today

- Tour BUGS
- Run some simple examples:
- Complex data structure (laboratory batch effects in twin studies)
- “Genetic IRT model”
- “GxE” for candidate locus

“Ladies and gentlemen…
start your engines…”