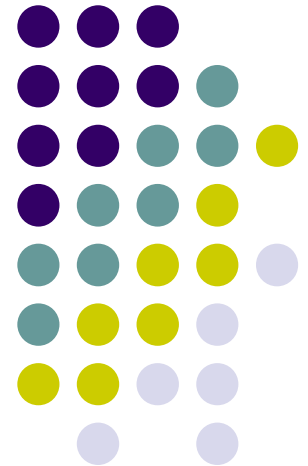


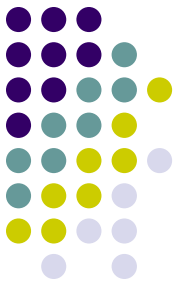
Data and Complex Models

Daniel Lawson
University of Bristol

work done at Biomathematics and
Statistics Scotland



Why bother with data?



Models are interesting in themselves – why bother with data?

Adds legitimacy to the model

Concrete example of the models relevance

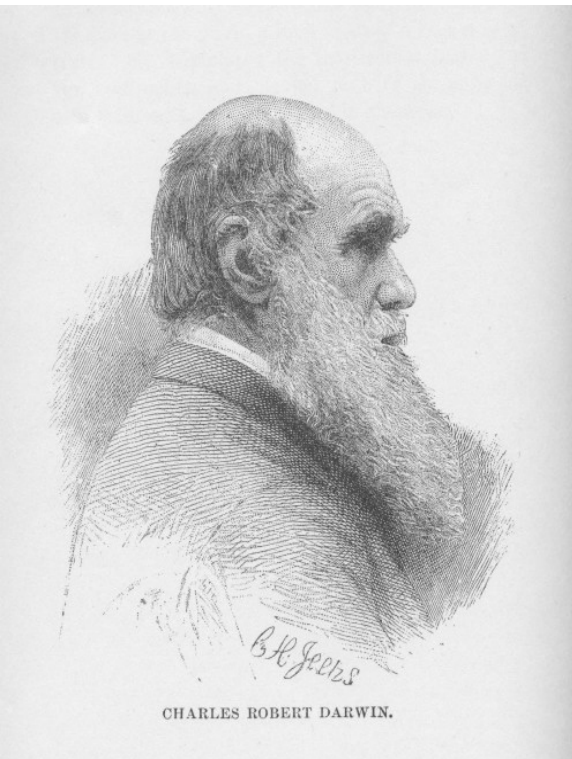
Gets the model looked at and used by other scientists

Citations!

<cynic>Applied journals produce a lot more paper volume... </cynic>

Modelling conversation

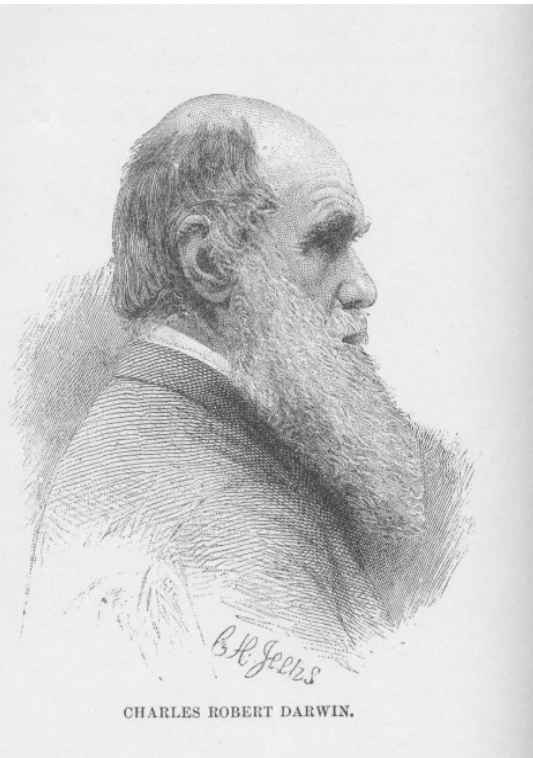
Distinguished Biologist



A. N. Other, Modeller

Modelling conversation

I've got this great model
that predicts really cool
stuff is going to happen!

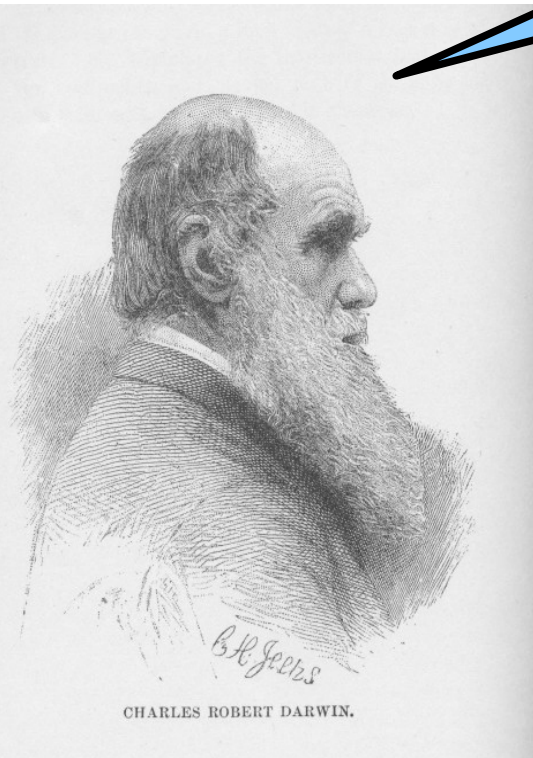


CHARLES ROBERT DARWIN.

Modelling conversation

I've got this great model
that predicts really cool
stuff is going to happen!

Yes? But does it relate to reality?



CHARLES ROBERT DARWIN.

Modelling conversation

I've got this great model
that predicts really cool
stuff is going to happen!

Yes? But does it relate to reality?


Of course! I already said it was cool.



CHARLES ROBERT DARWIN.



Modelling conversation



I've got this great model
that predicts really cool
stuff is going to happen!

Yes? But does it relate to reality?


Of course! I already said it was cool.

Really? Show me one
example where it works in
practice.



CHARLES ROBERT DARWIN.

Modelling conversation



I've got this great model
that predicts really cool
stuff is going to happen!

Yes? But does it relate to reality?

Of course! I already said it was cool.

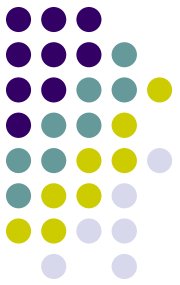
Really? Show me one
example where it works in
practice.

...



CHARLES ROBERT DARWIN.

Talk Outline



Motivation

Model testing overview

Bayesian statistics overview

- General approach

- MCMC method

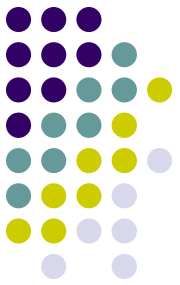
Example 1: ODE model of gut bacteria

- Uses MCMC and stats methodology

Example 2: Ecological neutral model

- True complex model with interesting predictions

- Uses Bayesian approach to do hypothesis testing



Patterns in nature

Observe some interesting pattern in nature

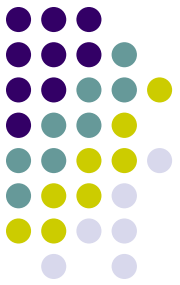
From Physics, Biology, Chemistry, etc

Create a model

Reproduce the pattern

Is the model the real process?

Many processes produce the same pattern!



What causes the pattern?

Best way to find out is with idealised experimental system

- Test model assumptions

- Knock out experiments, etc

Can't always do this!

- Often can in traditional physics

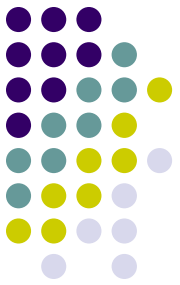
- Complex systems more difficult to study

Formal inference often needed

Mathematically interesting model is usually the limit of a more realistic & general model

- Full model more suitable for testing

Classical tests of models



Formal hypothesis testing

- Gives clearest results

- Hard to do in practice

- Hard to compare models

Information criterion (AIC, BIC)

- Informal “heuristic” for model fit

- Compare easily between related models

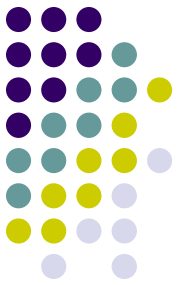
- Can be difficult to interpret for unrelated models

- Possible to “over fit” noise

Bayesian Model Selection

- Hardest to perform

Information Criterion



Akaike Information Criterion:

$$AIC = 2k - \log(P(D|M))$$

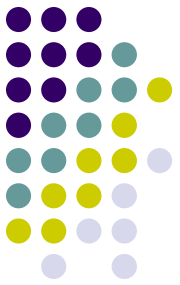
Bayesian Information Criterion:

$$BIC = k \log(n) - \log(P(D|M))$$

where k =number of parameters, n =number of datapoints

$P(D|M)$ is the probability of the data given the model (the likelihood).

Both penalise parameters against model fit, but *weight all parameters equally* (not fair on complex models)



Example of pattern matching

Ecological neutral theory - all species are “just as good”

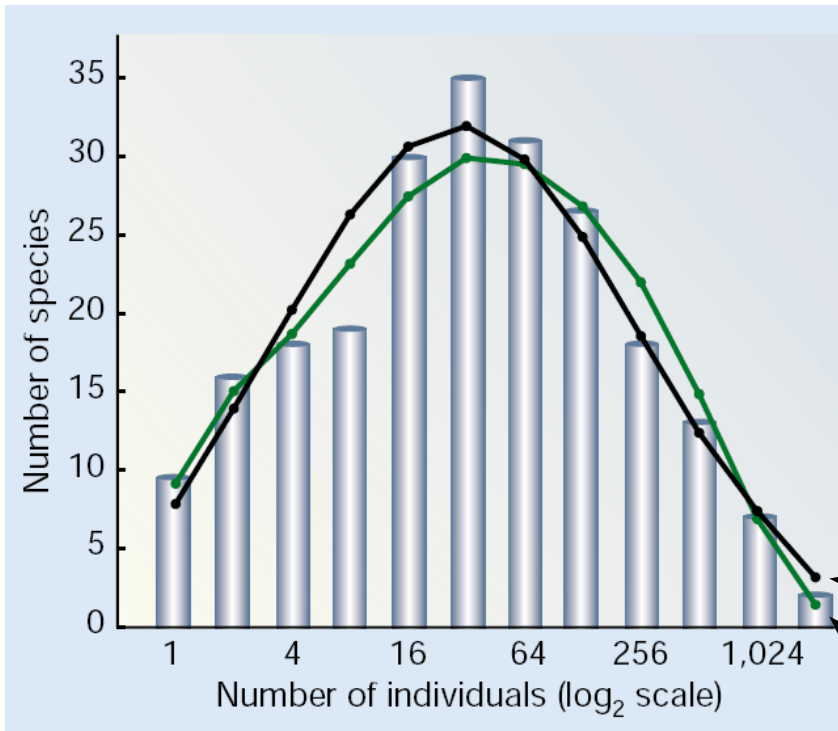
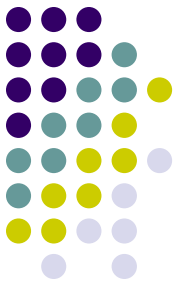
Fixed N individuals equally likely to die or reproduce in a timestep

Mutations occur at rate p on reproduction, creating a new species

What is the distribution of species sizes – the “Species Abundance Distribution”?

Does it match with data?

Species Abundance Distribution



Is this evidence of neutral dynamics?

Competing models just as good AIC

Is this even useful evidence?

Neutral prediction

Statistical curve fit

Bayesian Parameter estimation



Posterior:
$$\pi(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

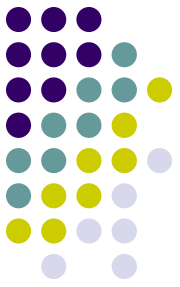
$P(D|\theta)$ is the *Likelihood* of the data given the model

$P(\theta)$ is the *Prior* probability of the model parameters

$P(D)$ is the probability of the data – requires integration over all model parameters

Usually have to evaluate $\pi(\theta)$ numerically

(Insulting) Frequentist example of “Bayesian” Estimation



Have six dice labelled 1..6, dice x has x white faces and $(6-x)$ black faces. One dice is taken at random and rolled. What is the probability the rolled dice was x given that the face observed is white?

$$p(x|\text{white}) = \frac{p(\text{white}|x) p(x)}{p(\text{white})}$$

Calculations:

$$p(x) = 1/6$$

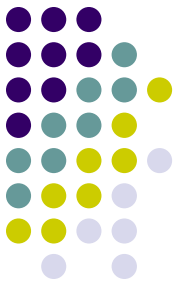
$$p(\text{white}|x) = x/6$$

$$p(\text{white}) = \sum_{x=1}^6 p(\text{white}|x) p(x) = 21/36$$

Answer:

$$p(x|\text{white}) = x/21$$

MCMC (Markov Chain Monte Carlo)



Sample from the posterior probability distribution

$$\pi(\theta) \propto P(D|\theta)P(\theta)$$

Likelihood $P(D|\theta)$ of parameters given some data

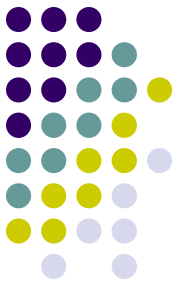
Prior $P(\theta)$: previous experiments

Metropolis-Hastings algorithm:

Select new parameters $\theta' \sim Q(\theta \rightarrow \theta')$

Where Q is the proposal distribution.

Accept θ' with probability $\min\left(1, \frac{\pi(\theta')Q(\theta', \theta)}{\pi(\theta)Q(\theta, \theta')}\right)$

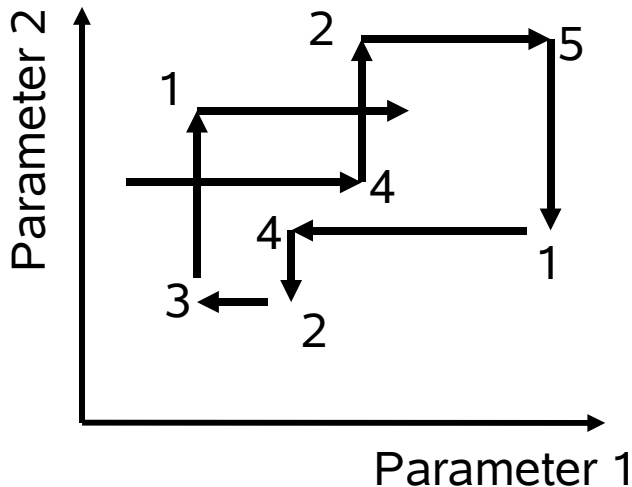


MCMC (2)

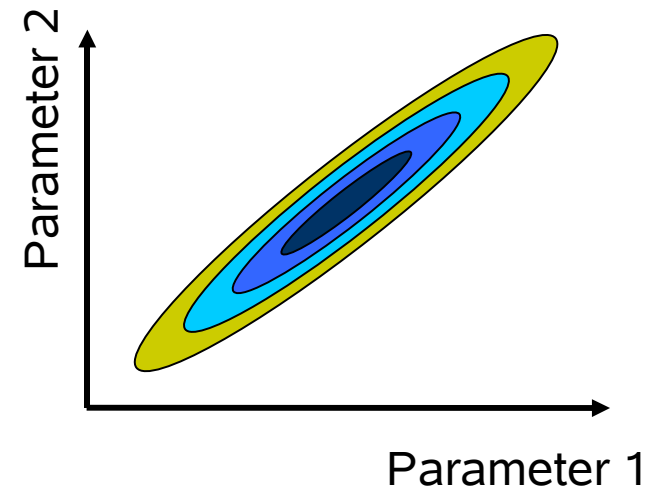
Reject proposal: add θ to the parameter set.

Accept: add θ' .

Obtain estimator $\tilde{\pi}(\theta) \sim \frac{\sum_{i=0}^N \delta(\theta - \theta_i)}{N}$ (can smooth)

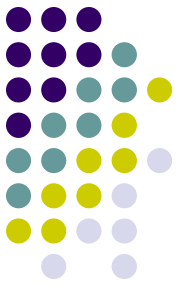


Many iterations



Build up Posterior Distribution over many iterations N

MCMC (3)



If proposal distribution Q is irreducible and aperiodic:
Guaranteed to obtain posterior as $N \rightarrow \infty$

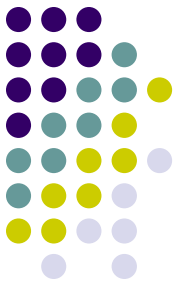
$$\Rightarrow \tilde{\pi}(\theta) \rightarrow \pi(\theta)$$

But nothing said about finite N

Efficient (i.e. good at low N) if proposal distribution matches posterior distribution

And acceptance probability is not too low

MCMC as a random walk in a potential



Random walk: probability of moving left q and right p with $p+q=1$:

$$\frac{q(x)}{p(x-1)} = \exp\left(-\frac{V(x-1) - V(x)}{T}\right)$$

MCMC: probability of moving left (Q symmetric):

$$q(x) = \frac{\min\left(1, \frac{\pi(x-1)}{\pi(x)}\right)}{\min\left(1, \frac{\pi(x-1)}{\pi(x)}\right) + \min\left(1, \frac{\pi(x+1)}{\pi(x)}\right)}$$

MCMC in a potential:

$$\frac{V(x)}{T} = -\log\left[\pi(x) \left[\min\left(1, \frac{\pi(x-1)}{\pi(x)}\right) + \min\left(1, \frac{\pi(x+1)}{\pi(x)}\right)\right]\right]$$



Obtaining faster convergence

Methods include

Reparameterisation

Reduction or simplification of parameter space

Difficult in complex models

Annealing (“heating”)

Decrease temperature with time to find high probability regions

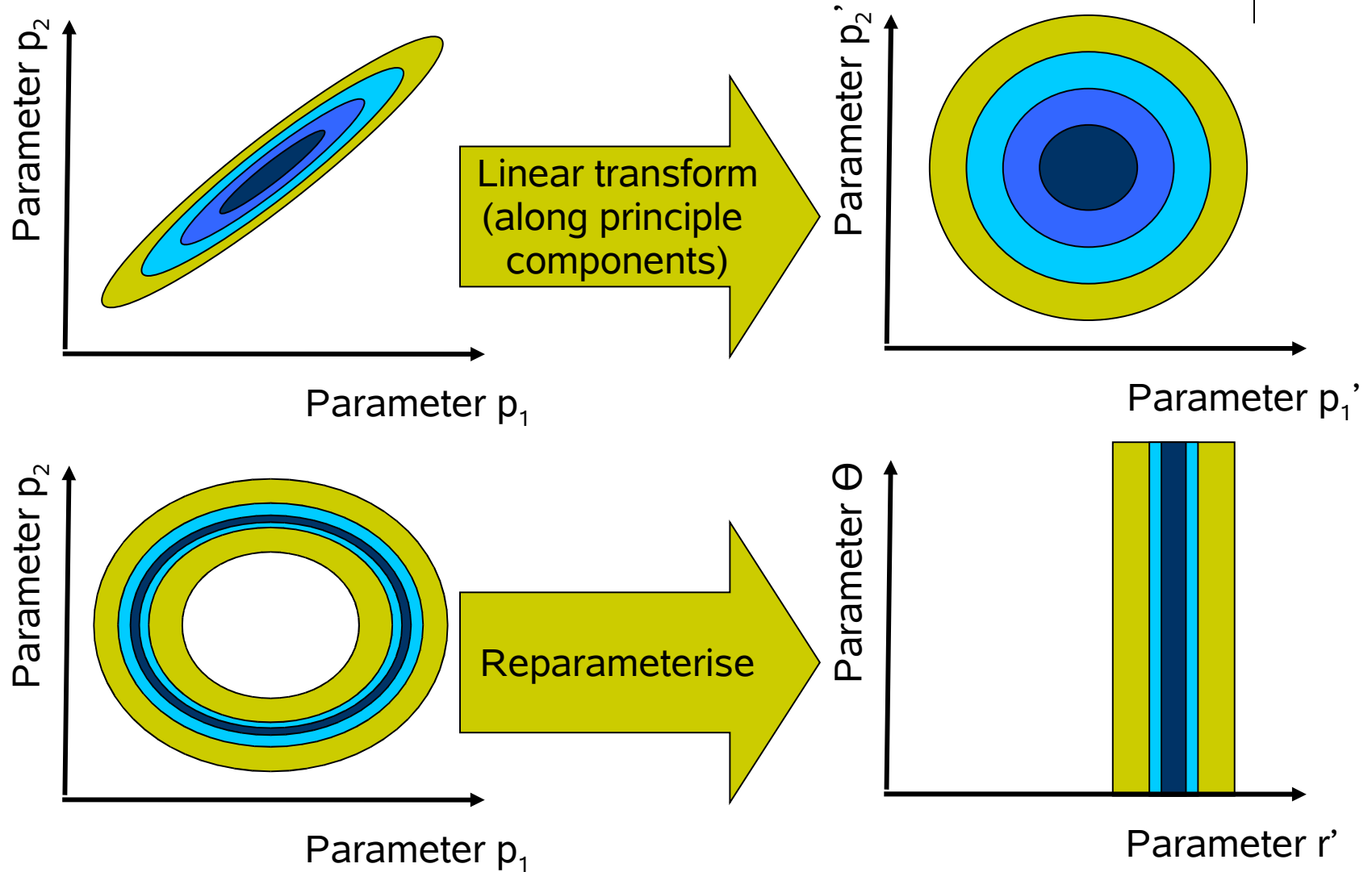
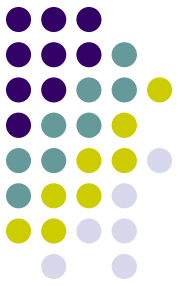
Only gets *to* equilibrium distribution, doesn’t explore it

Auxiliary variables

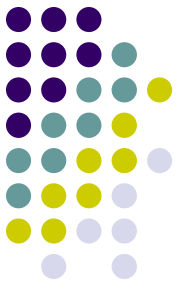
Augment parameter space to allow better mixing

Hard to apply in general case

Reparameterisation

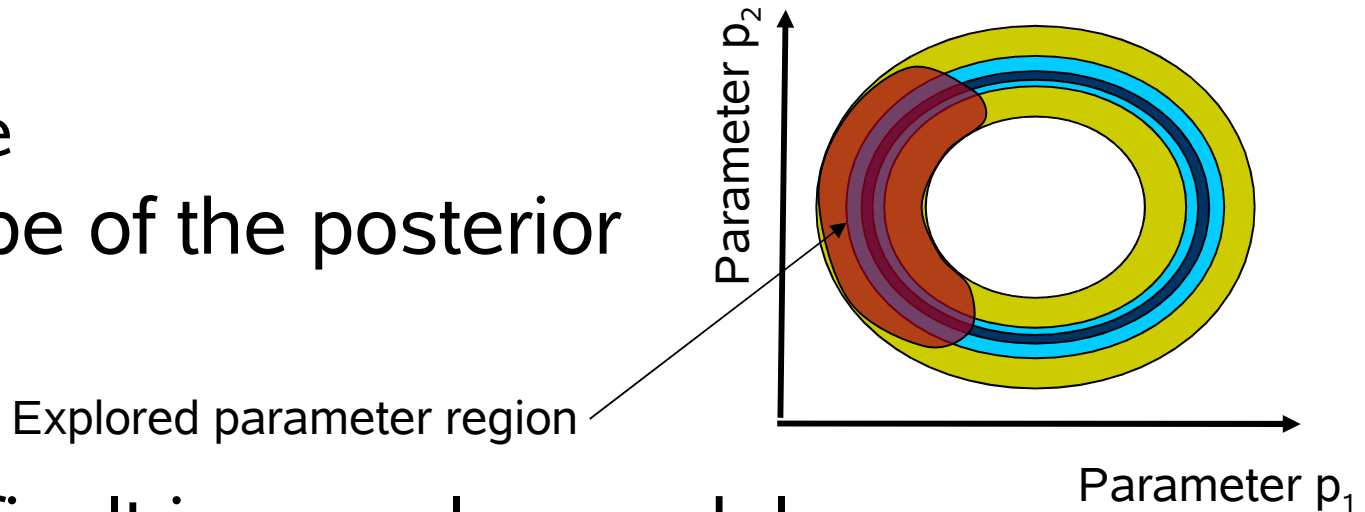


Convergence problems



Reparameterisation works if we can:

Detect
or Calculate
the shape of the posterior



This is difficult in complex models

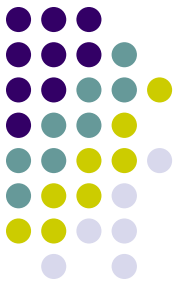
In practice, parameter region is often of too high dimension

Why: random walk in $D > 3$ doesn't fill space!

Use of a “proper prior” guarantees convergence

But is sometimes an unwarranted assumption

Example 1 – gut bacteria



Differential equation model of gut bacteria growth

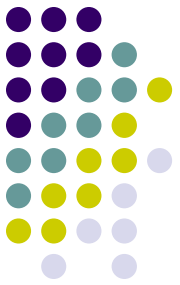
Several strains b_i competing for several “substrate” resources s_j

Interact via reaction products a_k (short chain fatty acids, or SCFA)

Flow of all materials through the gut (modelled as several compartments)

Detailed data from idealised experimental model available

General form of equations



Change of bacteria = Bacterial growth – bacterial outflow

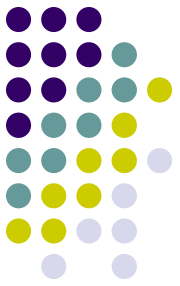
$$\frac{d B_1}{dt} = B_1 \sum_k G(B_1, S_k) \left(1 + \sum_j G(B_1, a_j) \right) - k B_1$$

Growth rate is non-linear in density of bacteria and resource

$$G(B_1, x_j) = \frac{g_{ij} B_1(t) x_j(t)}{(x_j(t) + K_{ik})}$$

Substrates and SCFA behave similarly

Unsolvable non-linear model
– equations not important!



Inference

Try to establish **qualitative** and **quantitative** behaviour

Not all parameters measurable

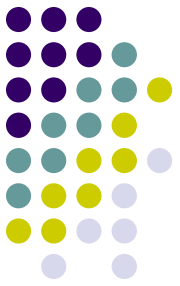
Some time series data is available

Likelihood for differential equation model?

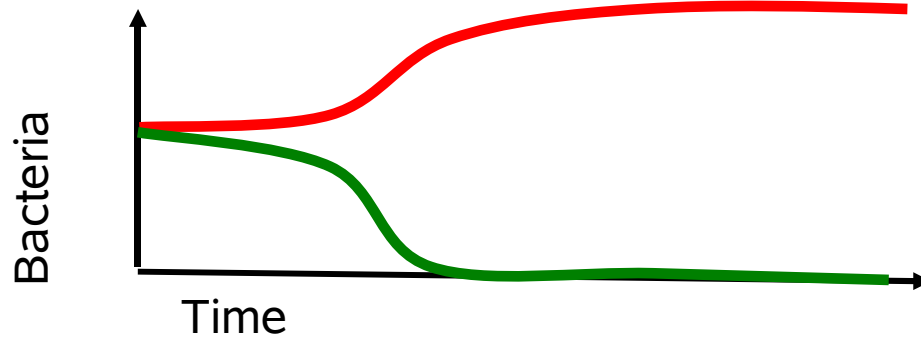
Need a probabilistic model!

Use stochastic measurement process

Model behaviours



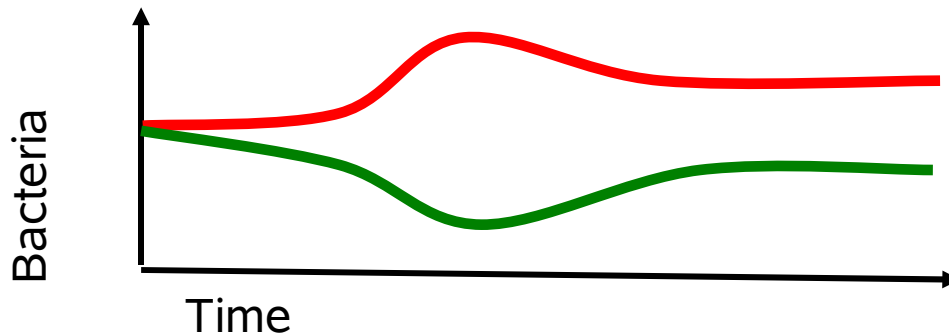
Competitive exclusion



Behaviour with single
substrate:

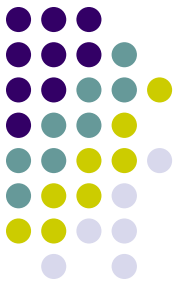
NO CROSS FEEDING

Coexistence

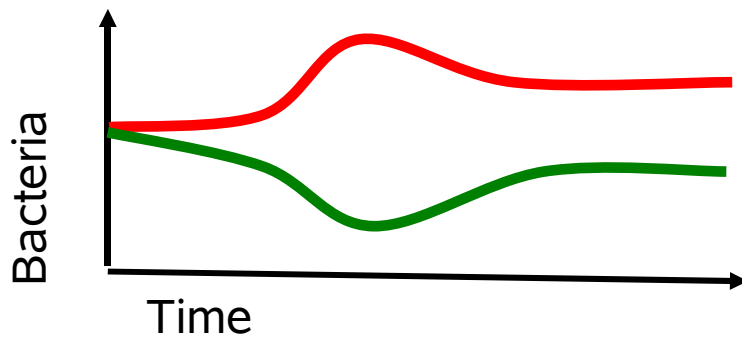


CROSS FEEDING

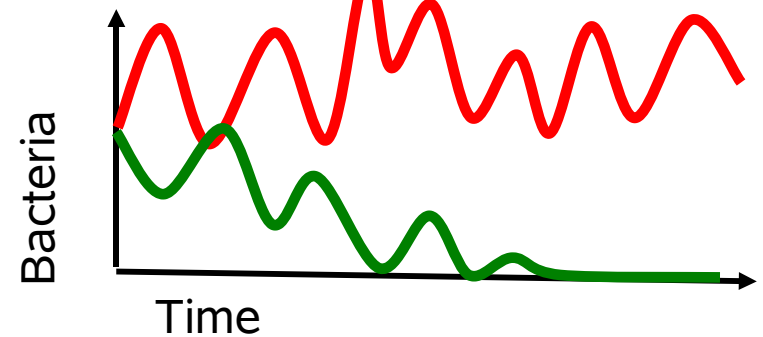
More model behaviours



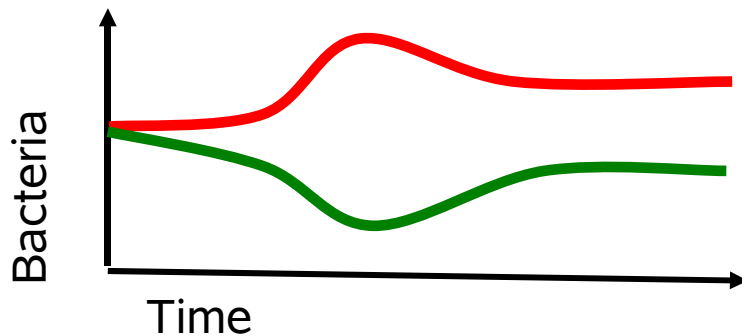
Input driven extinction



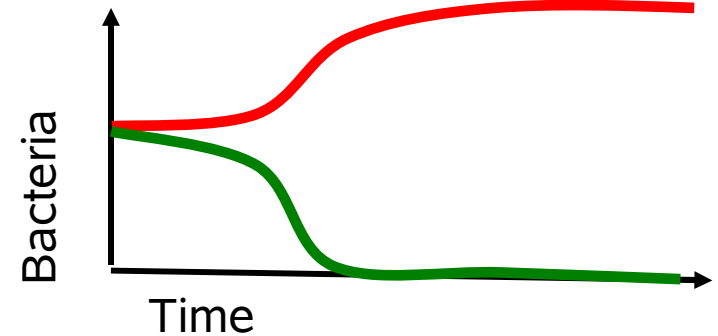
Periodic input



Host competition exclusion



Host absorbs SCFA

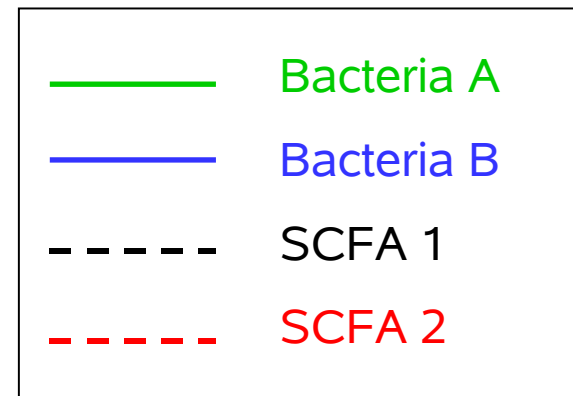
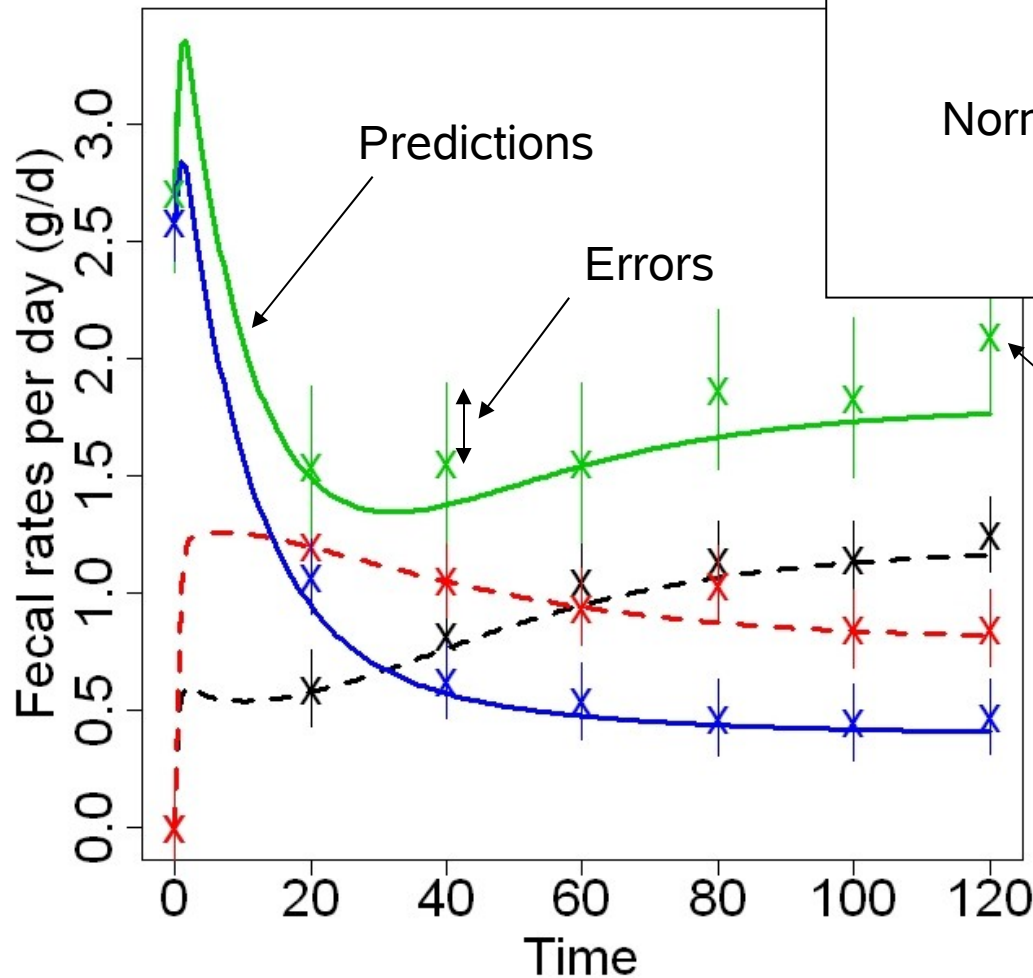


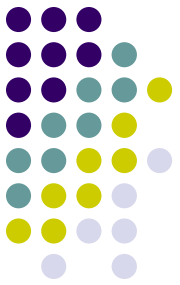
Likelihood from differential equation models



$$\log L(D|M) = \sum_i N(y_i, y(x_i), \sigma_i)$$

Normal distribution Observations Predictions Errors





Data in the gut model

Multiple experiments E_i under different conditions

Each leads to inconclusive inference

Posterior of E_1 is summarised, used as prior for E_2 , etc. Some parameter estimates never improve.

Combined approach may be better?

No need to summarise

Hierarchical statistical model combines datasets

Dangers:

Larger state space – might *increase* degrees of freedom!

Model may be inconsistent between experiments?

Hierarchical statistical model

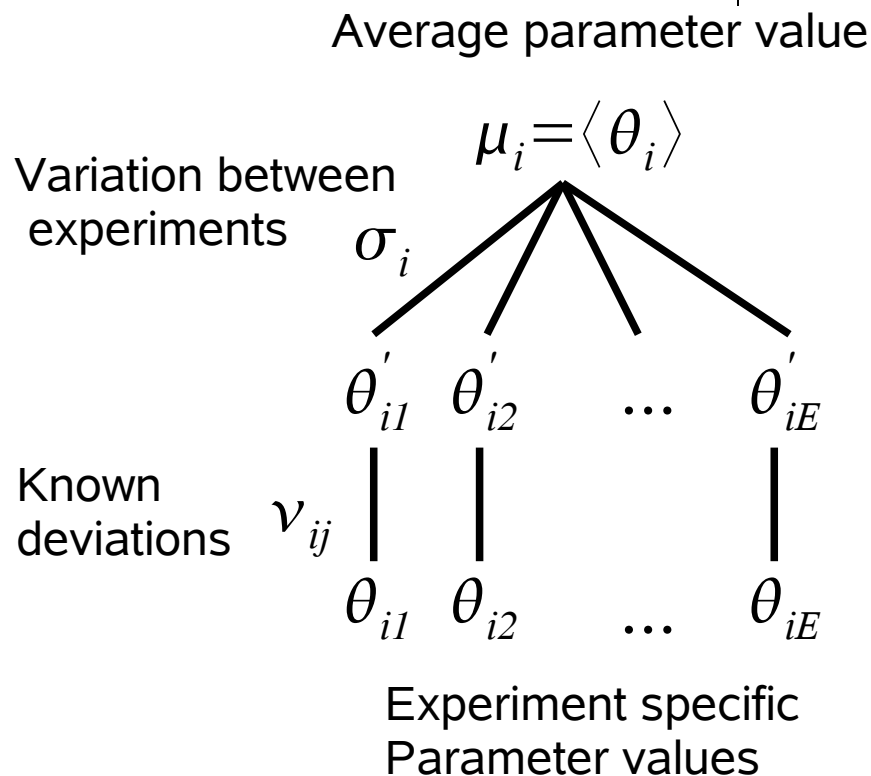


Each parameter value
hierarchically related
to others

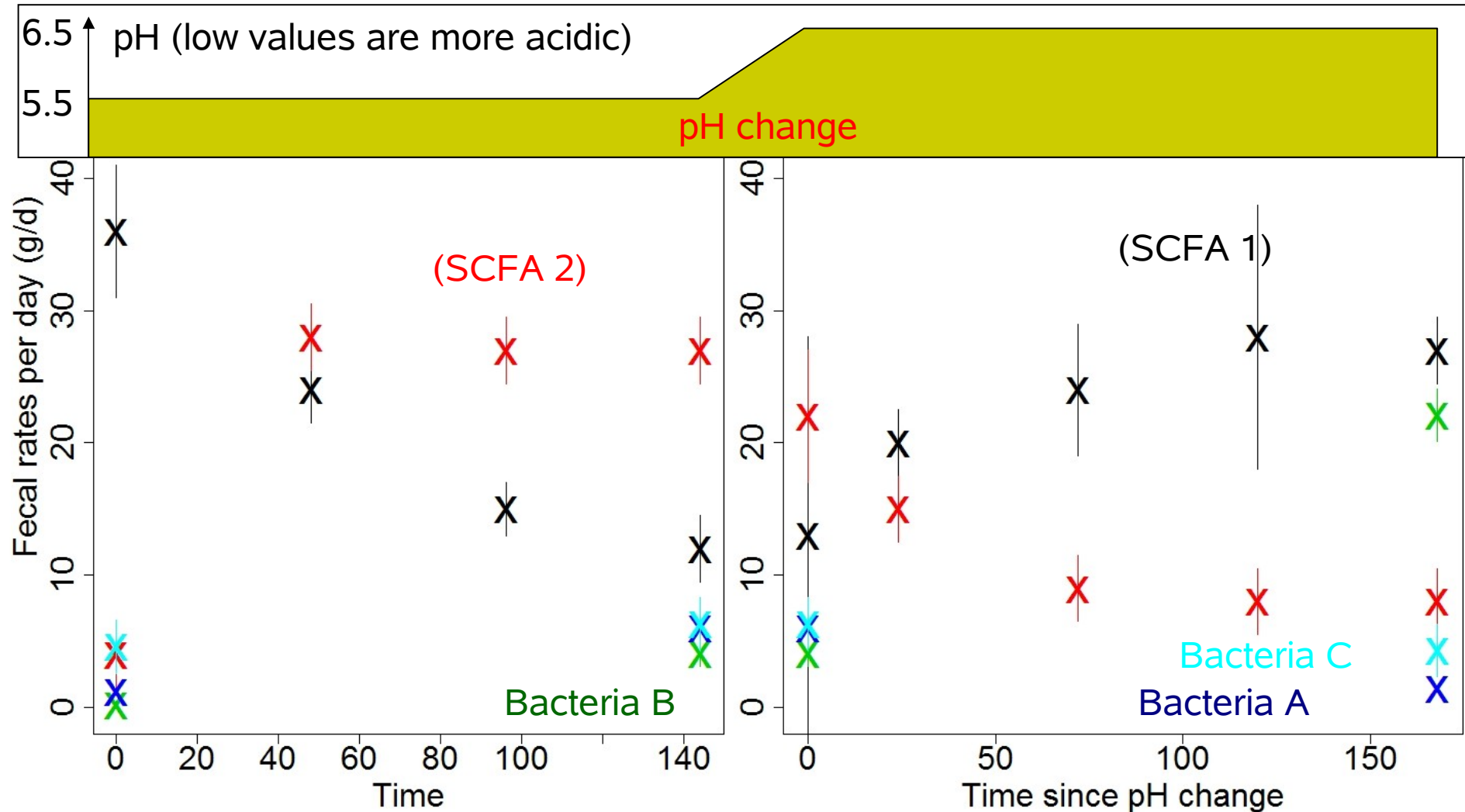
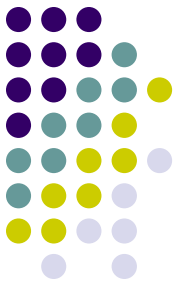
Allow for **variation** and
predictable differences
between experiments

Can in principle use full
covariance matrix over
all parameters

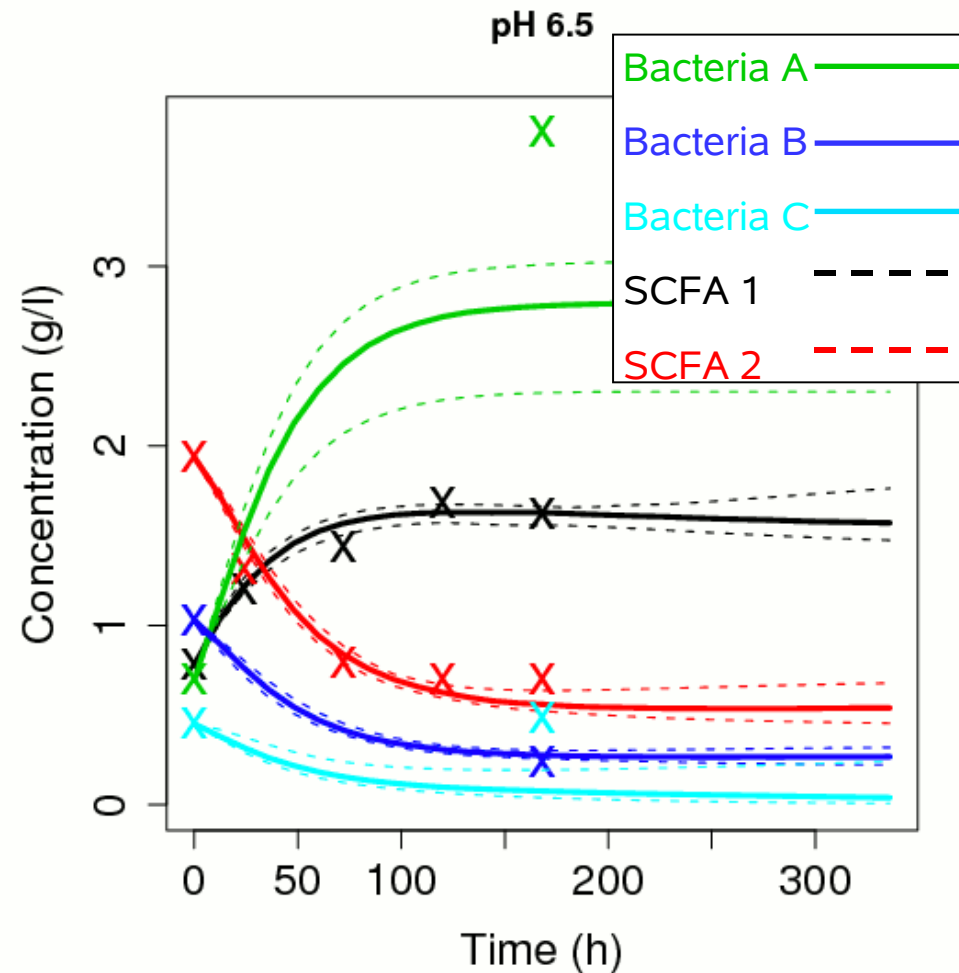
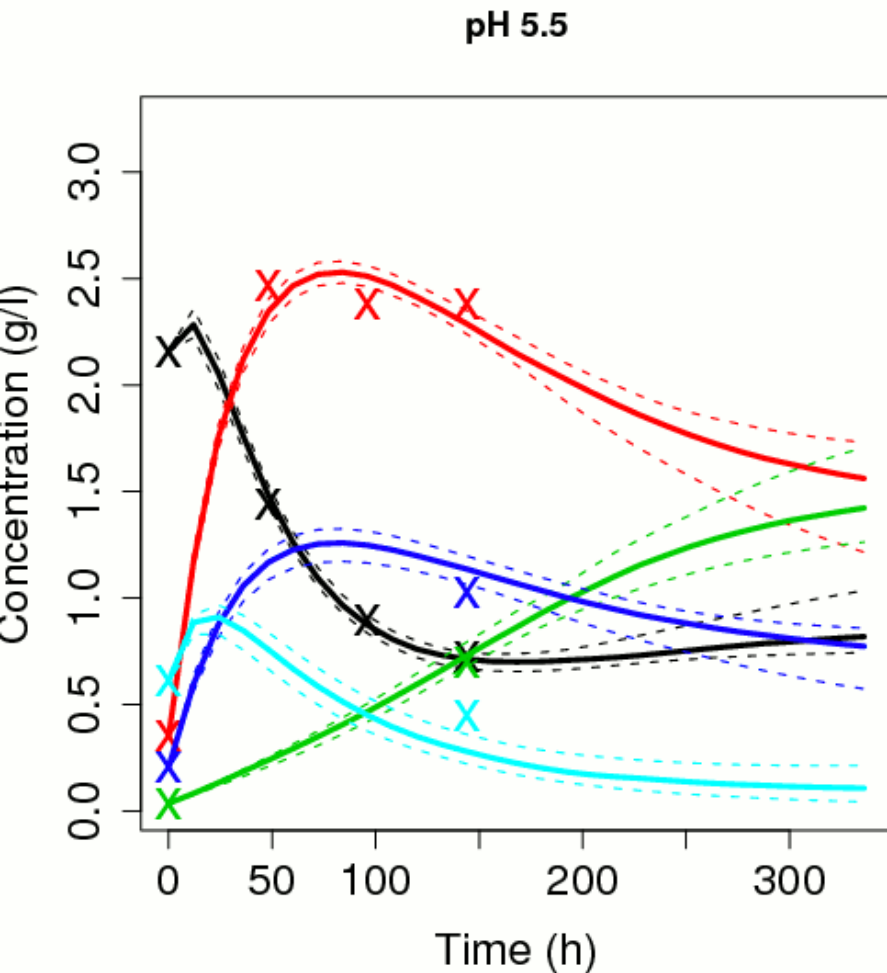
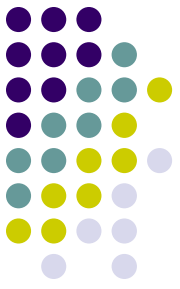
Can estimate σ_i if
enough experiments
available

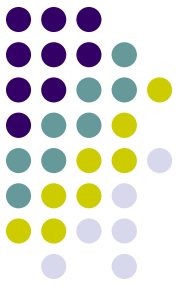


Example data



Results – prediction





Practical application

MCMC inference with hierarchical model

- Obtain full parameter distribution

- Explains experimental results in terms of fundamental bacterial properties

Connect experiment and *in-vivo* behaviour

- Extended colon

- Periodic food intake

Can predict for different scenarios

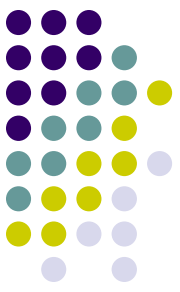
- e.g. effect of antibiotics

- Direct access to SCFA for health predictions

Care needed - model still too simple

Example 2: Ecological Spatial Pattern model

- Scottish Pine Trees



Relate mathematically interesting neutral model to real world

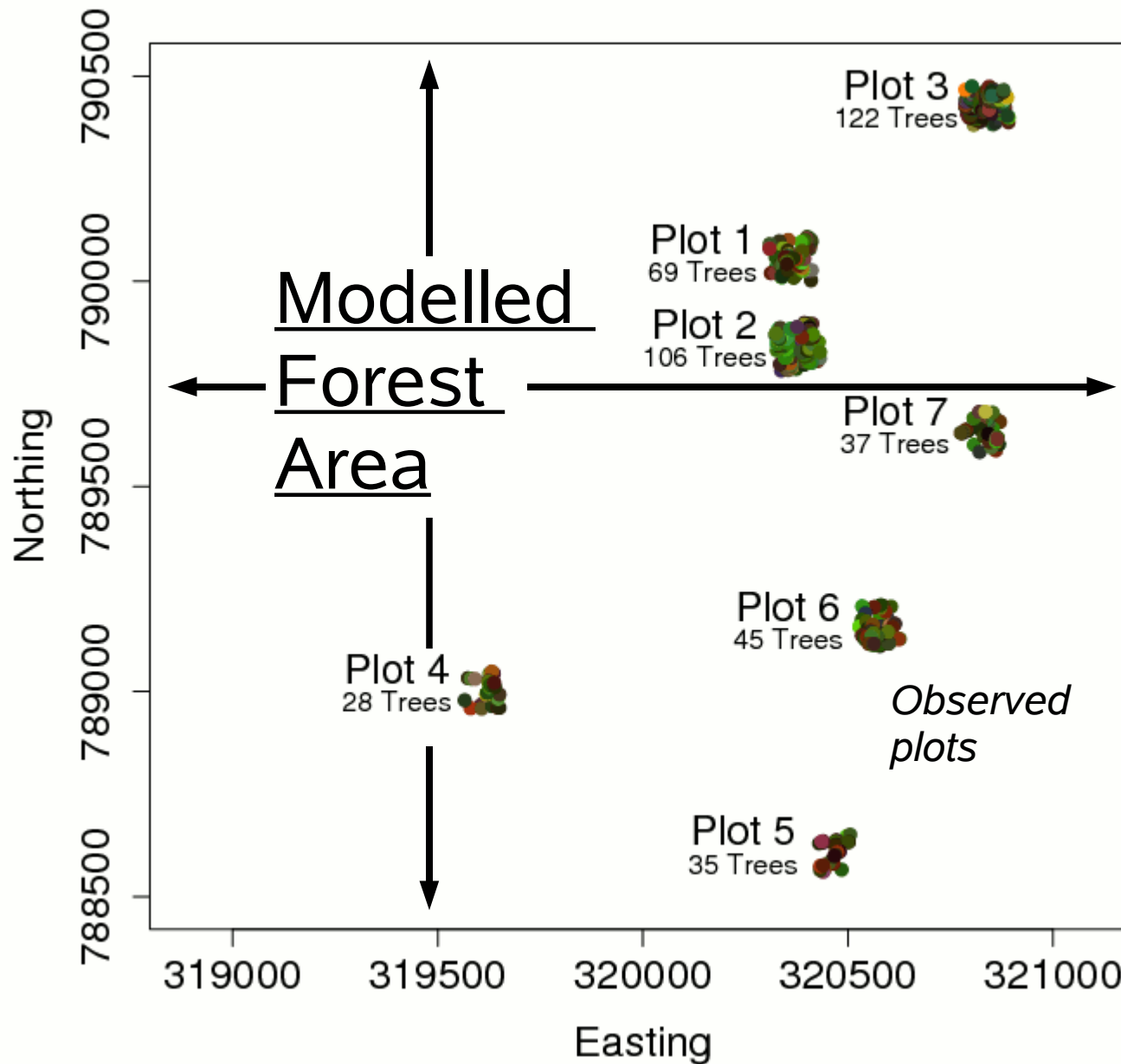
Neutral ecological model (i.e. no heritable differences)

Genetic differences observable through chemistry (monoterpenes)

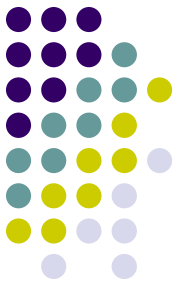
Spatial data for monoterpenes – large heterogeneity observed

Theoretical models predict this – is the prediction quantitatively correct?

If not, monoterpenes are shown to be selected

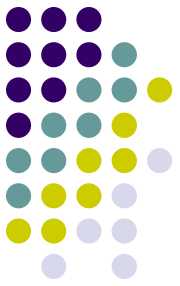


Model ingredients



- Competition for space
- Sexual reproduction of trees
- Short-ranged seed dispersal
- Longer ranged pollen dispersal
- Pollen also arrives from outside modelled area
- Monoterpenes determined genetically
- Neutrality - parenting probability independent of monoterpenes

Inference in model



Define “sufficient likelihood”: set of descriptors that capture all features

- Spatial clustering of trees

- Spatial clustering of monoterpenes

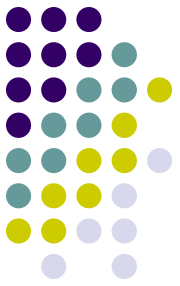
- Long ranged correlations in monoterpenes

Model too slow for MCMC

- Sample parameters using latin-hypercube

- Statistically model the likelihoods to obtain an approximation of the posterior distribution

Hypothesis test

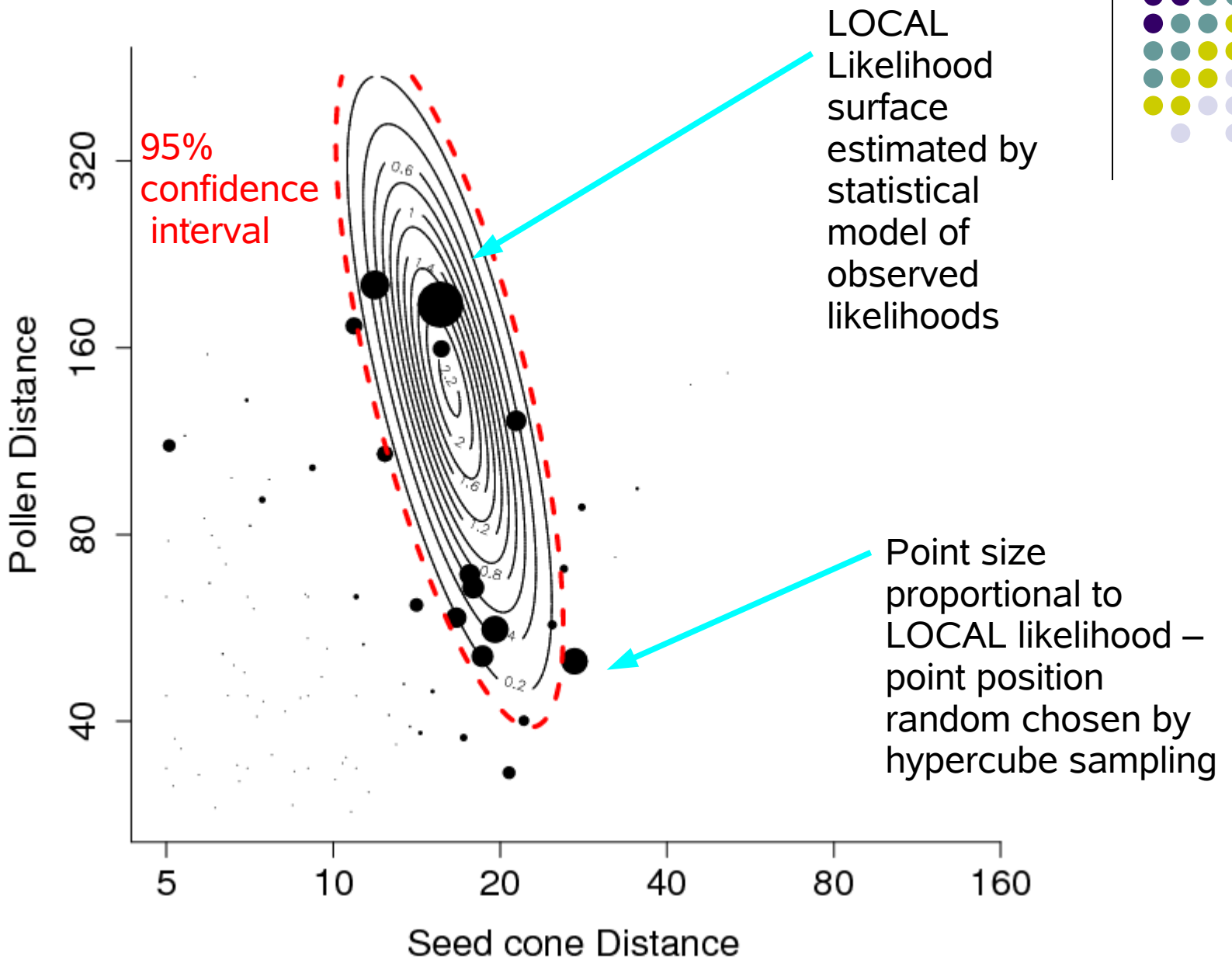


Obtain likelihood model for the local clustering of trees and monoterpenes

This is hard – requires approximation by variogram models

Obtain a sample from this LOCAL posterior

Test whether the data falls within the 95% confidence interval of the sample for the large scale clustering



Hypothesis test



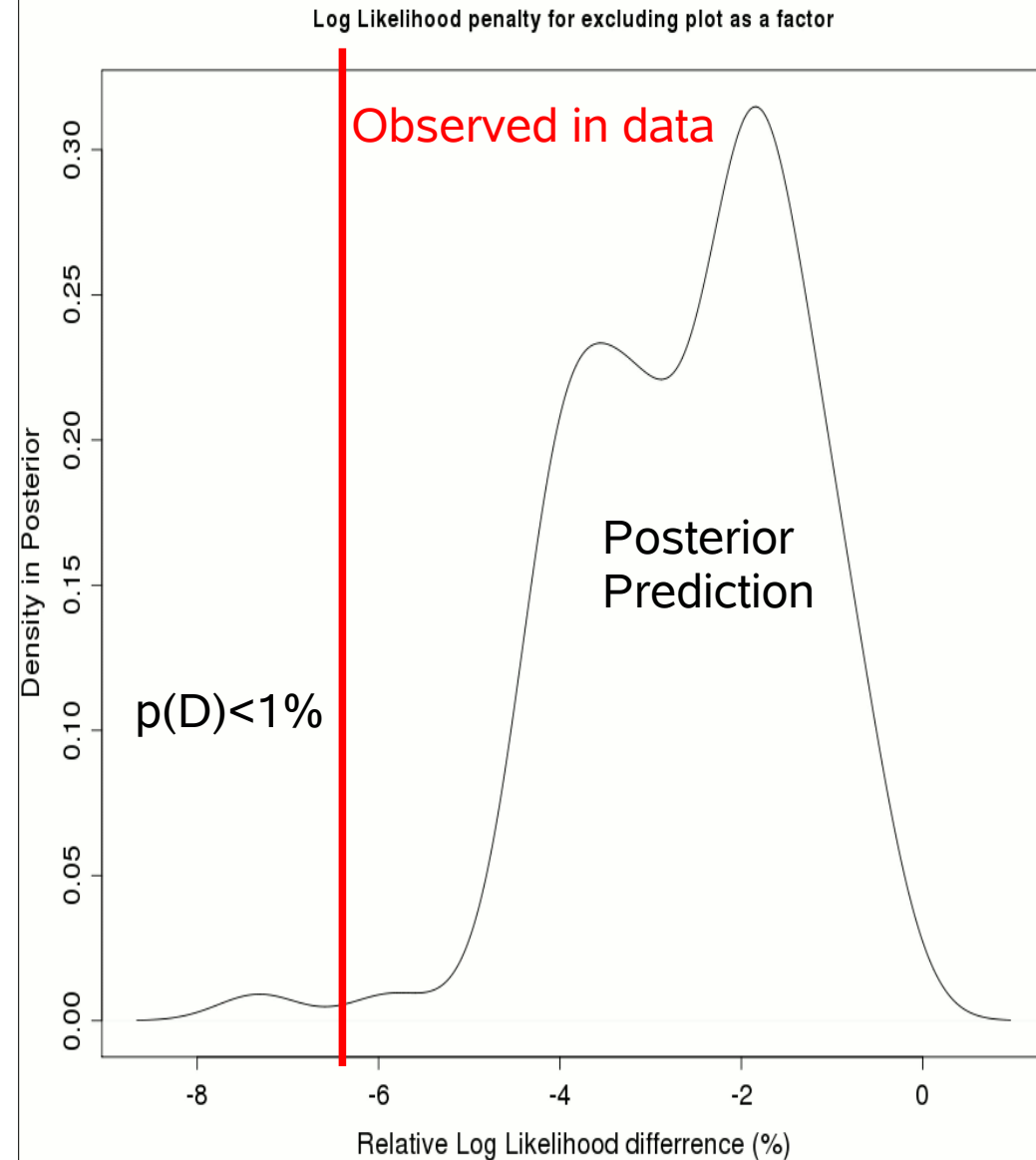
Formal test for neutrality in the data

Model parameters sampled from local effect posterior

Consider log likelihood gain from using site number as an explicit factor

Its significantly more important in real data than model

Therefore real data is not explained by a neutral model



Conclusions



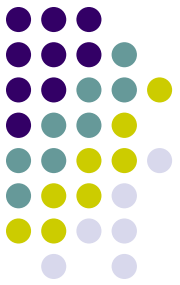
Inference is possible for complex models
Bayesian formalism is the most appropriate
MCMC allows sampling from a likelihood, if we
can write one down
Can formally test whether a model is incorrect
Lots of scope to mix statistics with complex
systems problems!

Conclusions



Inference is possible for complex models
Bayesian formalism is the most appropriate
MCMC allows sampling from a likelihood, if we
can write one down
Can formally test whether a model is incorrect
Lots of scope to mix statistics with complex
systems problems!
Thank you for listening!

Conclusions



Inference is possible for complex models
Bayesian formalism is the most appropriate
MCMC allows sampling from a likelihood, if we
can write one down

Can formally test whether a model is incorrect

Lots of scope to mix statistics with complex
systems problems!

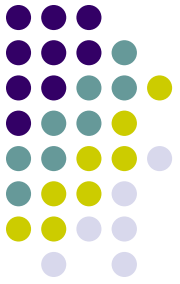
Thank you for listening!

Questions and discussion - might this work on
any of your problems?

Pre-prints available... just ask

Bristol has a complex systems group with heavy stats involvement...

Extra slides follow



Hypothesis testing



Consider $P(D;C|M) < P_0$

P_0 is the threshold for the test, C is a condition to test on the data D

Requires careful formulation:

- Probability of exactly the data is often infinitesimal

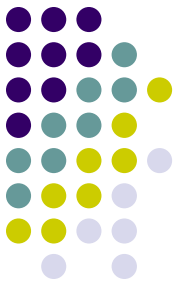
- Consider probability of exceeding some threshold

- Usually only possible for simple cases

Example: is a coin biased?

Observe 10 throws, 8 of which are heads

$$P(h \geq 8 | p(h) = 0.5) = \sum_{x=8}^{10} \binom{n}{x} 0.5^n = 0.054$$



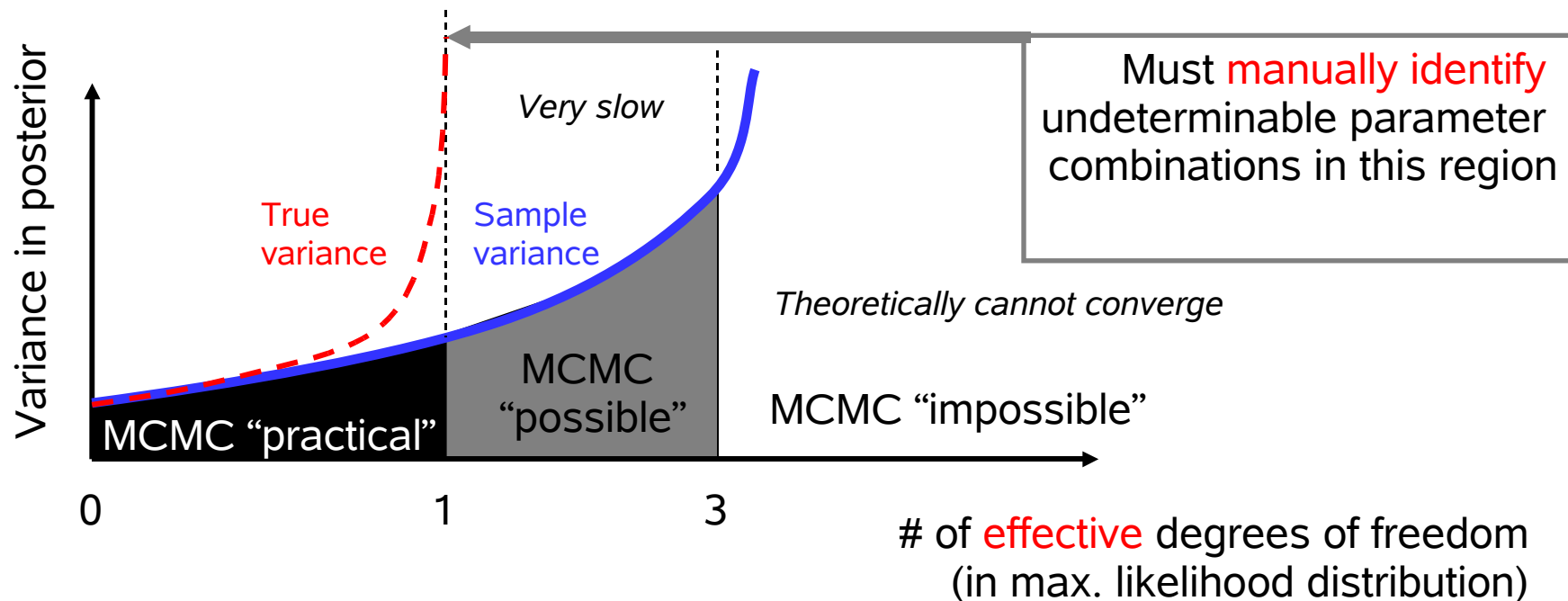
MCMC for process models

More restrictive prior, or increasing data:

Reduces parameter region size

Reduces degrees of freedom in posterior

Reduces total variance



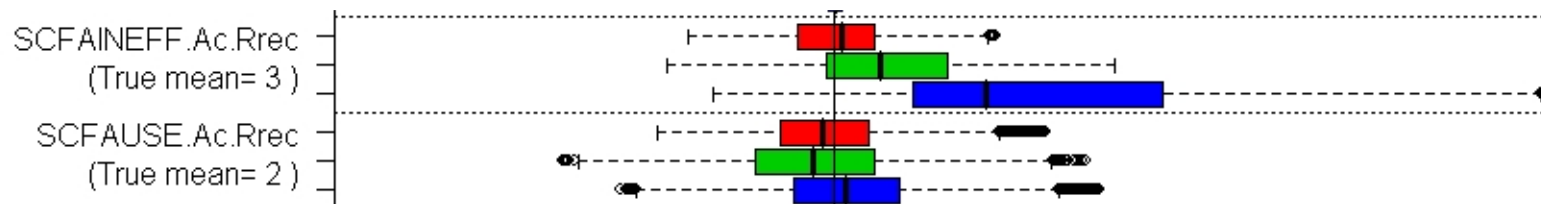
Simulation study with multiple experiments:

Competition at (a) pH 5.5 & (b) 6.5

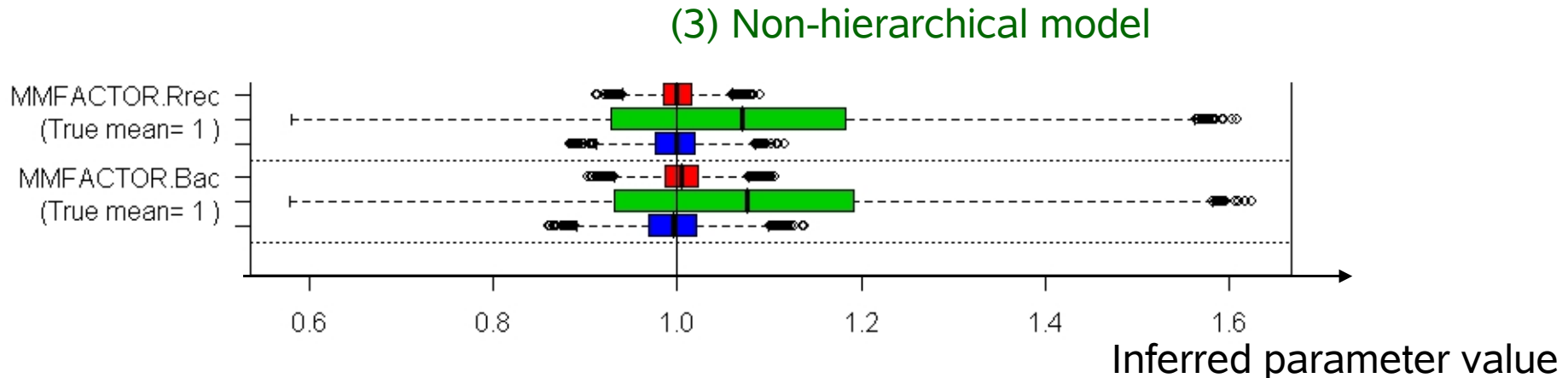
Bacteria A growth experiment at (c) pH 5.5 & (d) 6.5

3 scenarios considered: only **scenario (1)** has converged MCMC chain!

(1) Hierarchical model



(2) Less data -Not measuring substrate output

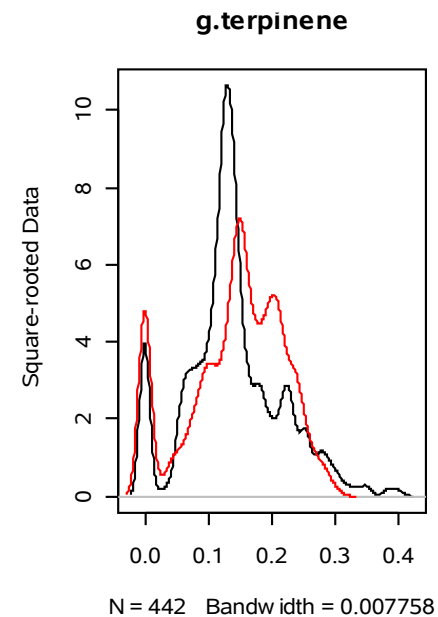
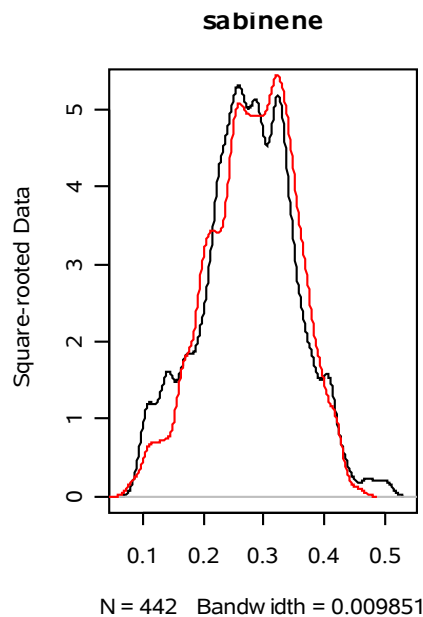
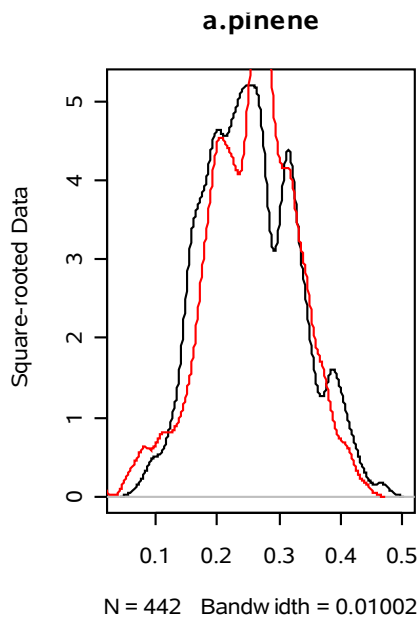




Monoterpenes are genetically controlled and heritable

Distributions can be well approximated by a weighted binomial distribution

L genes (values 0 or 1), each contributes differently to monoterpene count



Variogram models

