Inference for ecological process models of gut bacteria

Daniel Lawson

Biomathematics and Statistics Scotland

Aberdeen

Bristol University from October

Biological problem

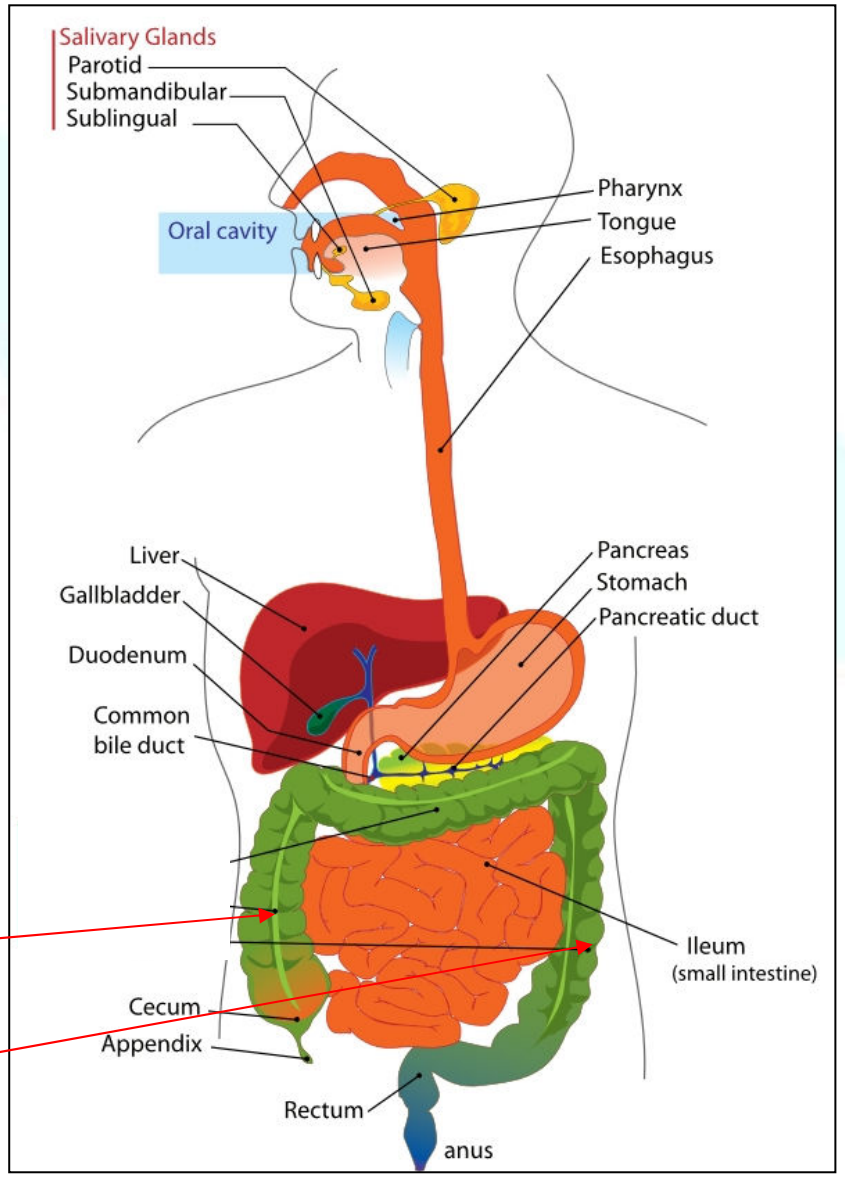
- Up to 20% of our energy is produced by fermentation of indigestible products by bacteria
- Bacteria produce Short Chain Fatty Acids (SCFA)
- Different SCFA used in different parts of the body: liver, gut wall, brain, etc.
- Each has different health properties e.g. Acetate linked to some cancers
- Ultimately want to understand & control SCFA products

Physical Flow

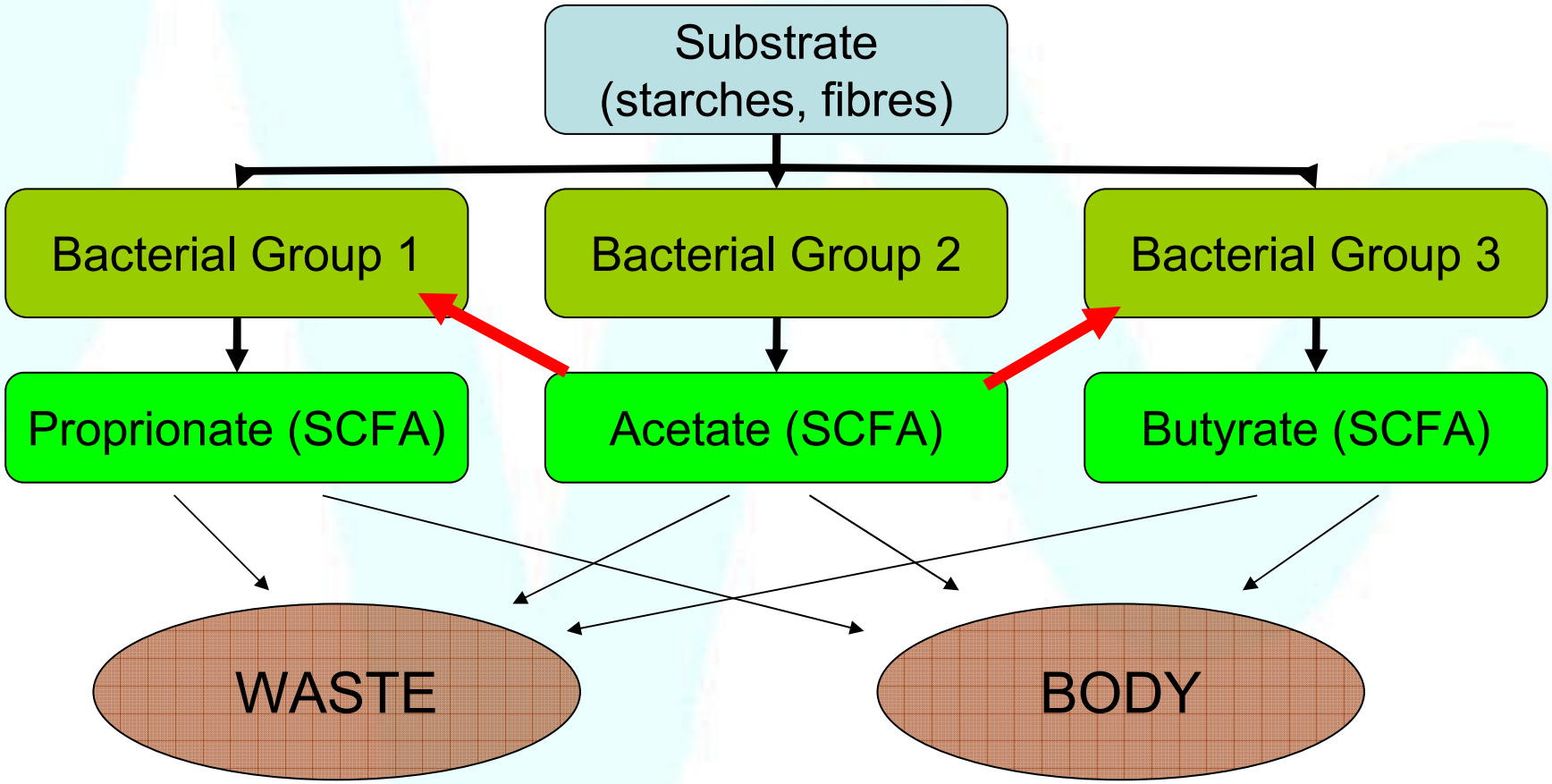
- Consider bacteria in the colon
- Proximal colon acidic but has best substrate
- Distal colon more hospitable but poorer substrate
- Important ecological trade off

Proximal colon

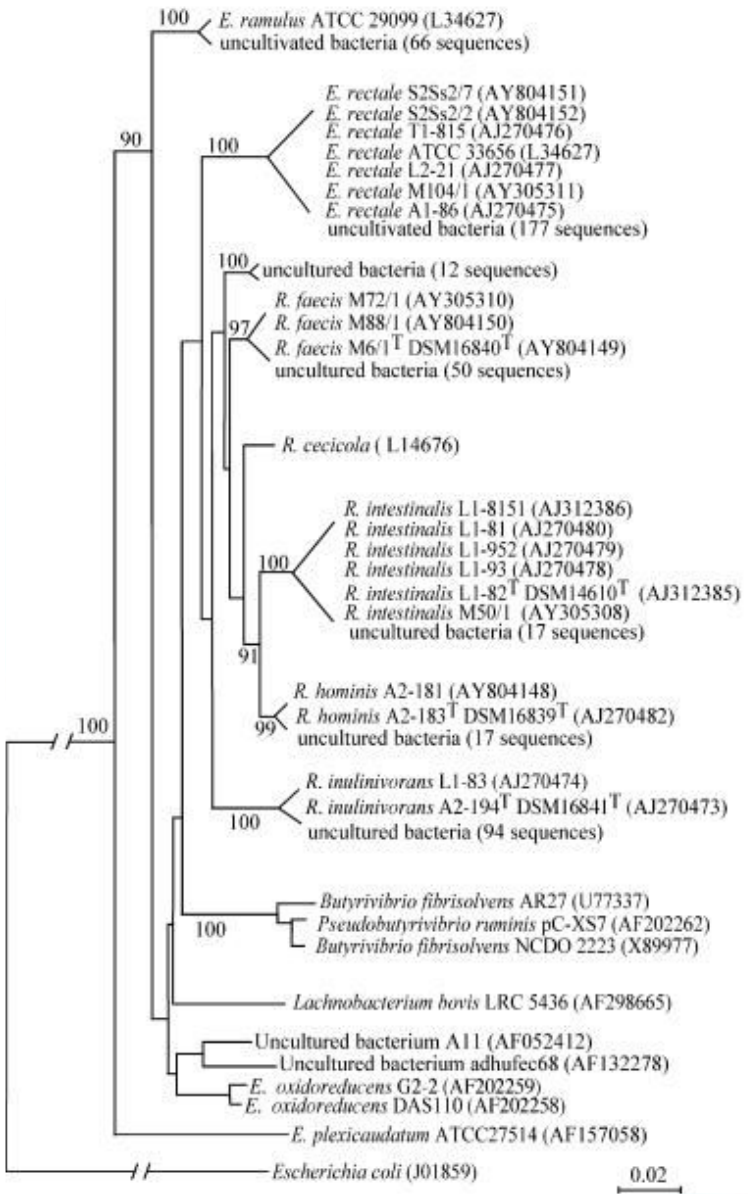
Distal colon



Conversion flow



Functional bacterial groups



Measured by FISH (Fluorescence In Situ Hybidization)

Roseburia rectalis: probe measures this whole group

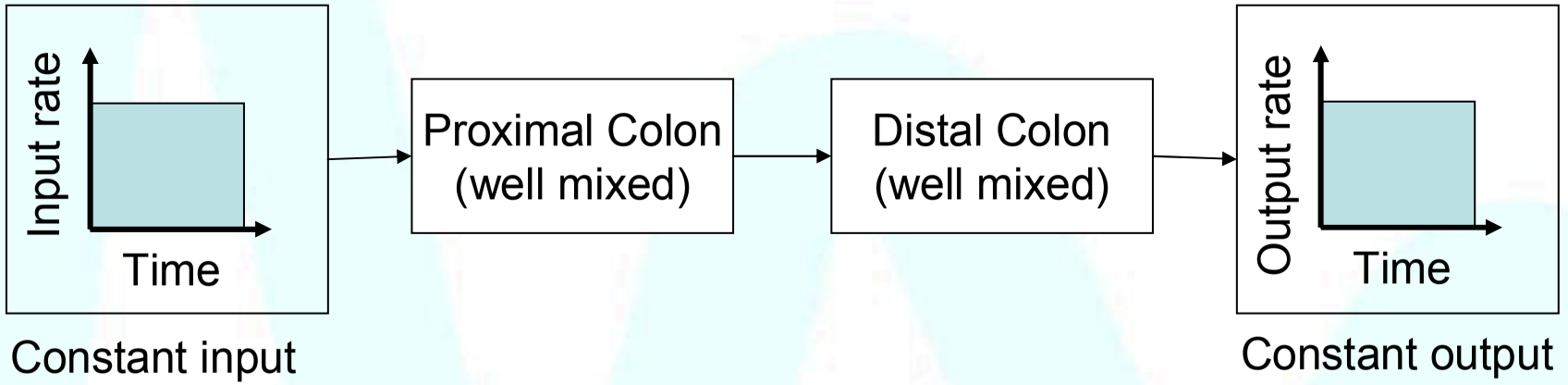
E. rectalis: Another probe measures this group

Rrec consumes acetate and produces butyrate
The rest of *Erec* typically does not

(Unfortunately FISH probe groups and functional groups don't overlap perfectly!)

Experiments, and reality

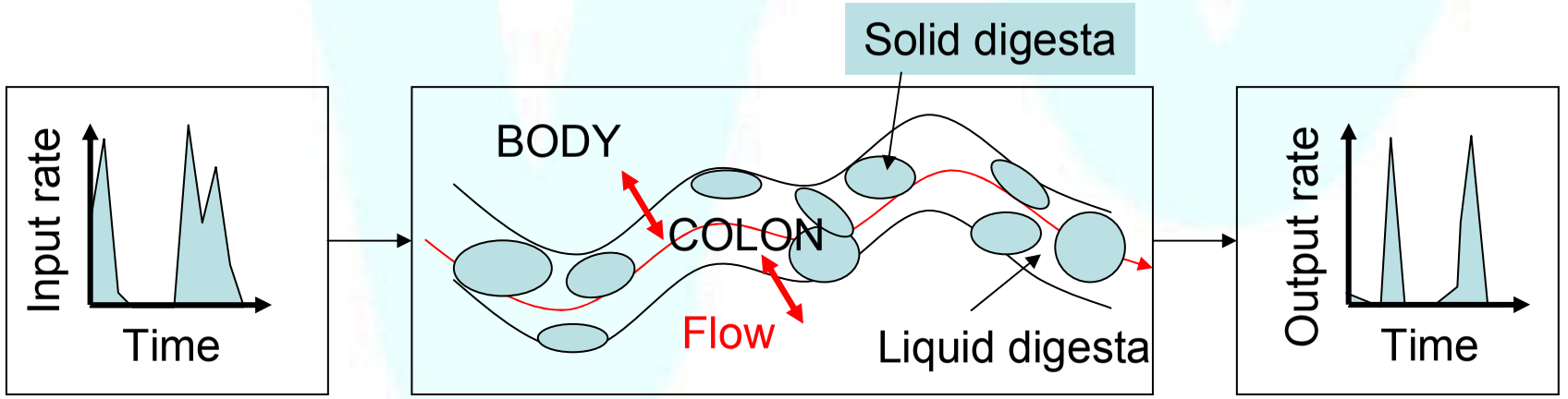
Experiment



Constant input

Constant output

Human body



Stochastic input

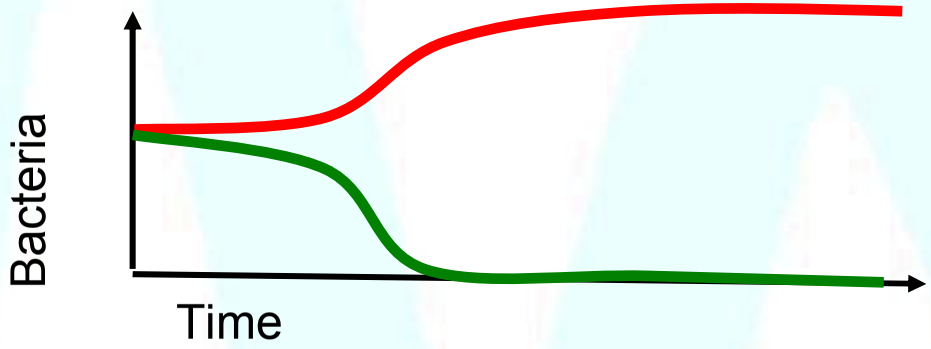
Stochastic output

Model structure

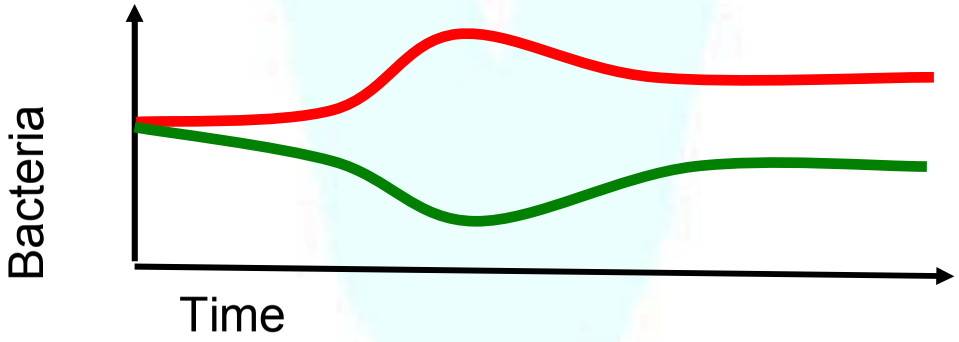
- Differential equation model of experiments
- Large number of compartments represents continuous system
- Continuous/stochastic/periodic inputs
- Account for SCFA absorption by body

Model behaviours

- Competitive exclusion



- Coexistence



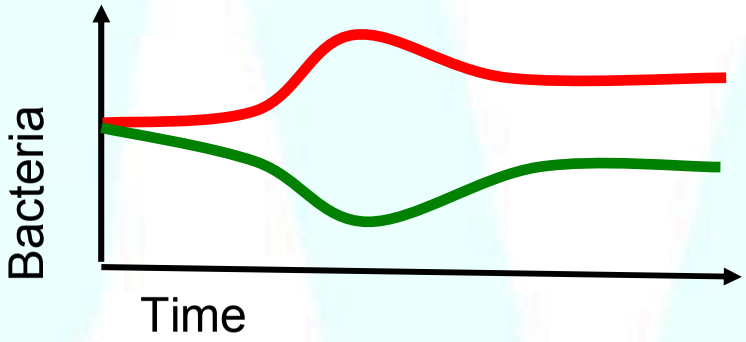
Behaviour with single substrate:

NO CROSS FEEDING

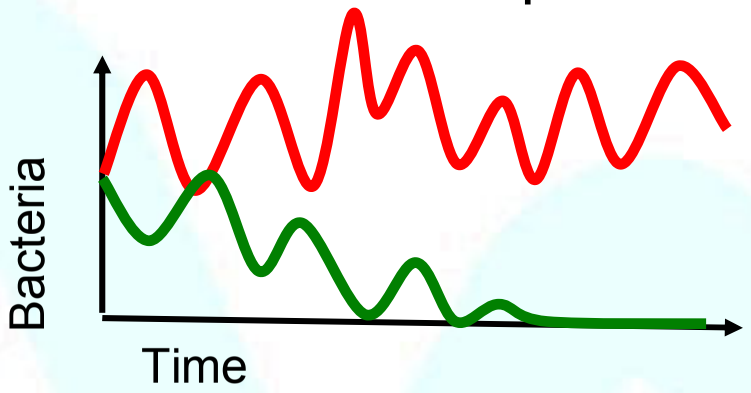
CROSS FEEDING ON ACETATE

More model behaviours

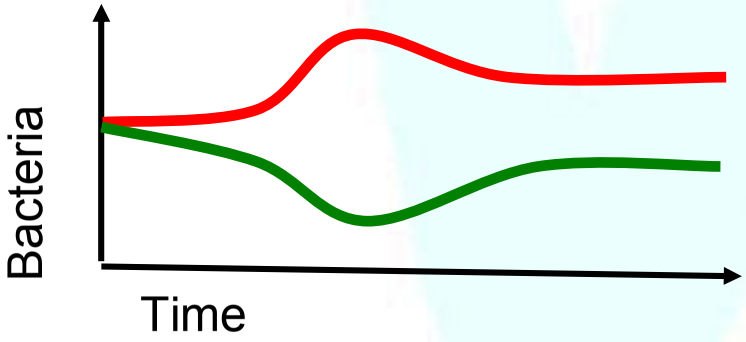
- Input driven extinction



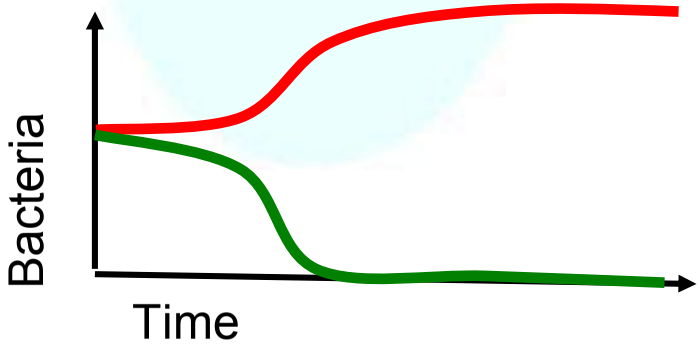
Periodic input



- Host competition exclusion



Host absorbs SCFA



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!
    - Stochastic measurement process

Likelihood from differential equation models

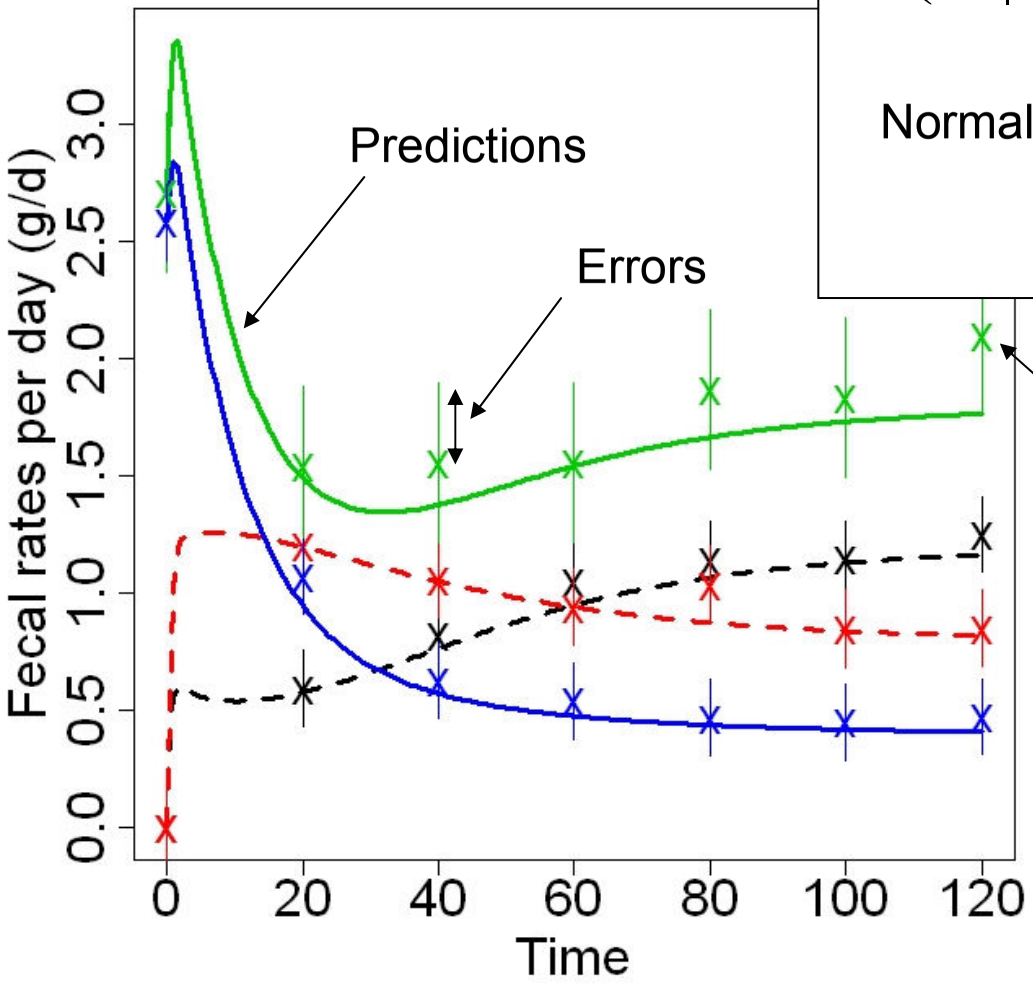
$$L(D \mid M) = \prod_i N(y_i, y(x_i), \sigma_i)$$

Normal distribution

Observations

Predictions

Errors



MCMC

- Sample from the posterior probability distribution

$$\pi(\theta) \propto L(D \mid \theta) P(\theta)$$

- Likelihood *L* of parameters: time series data
- Prior *P*: previous experiments
- Metropolis-Hastings algorithm:

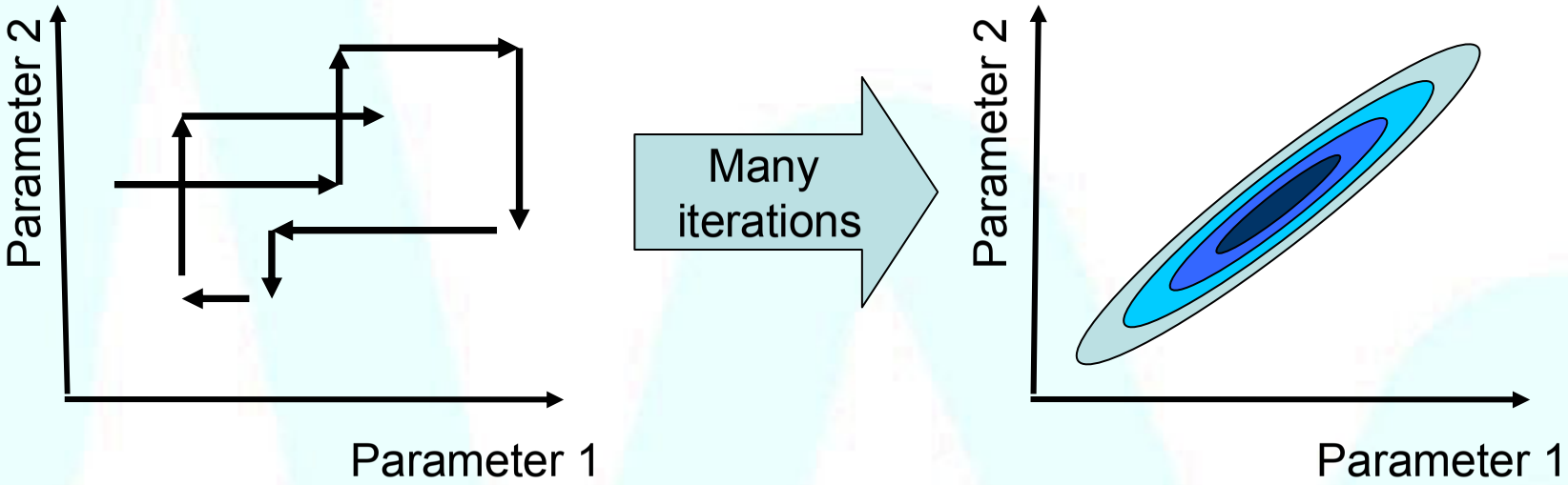
Select new parameters $\theta' \sim q(\theta', \theta)$

Accept θ'

Proposal distribution

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

MCMC (2)

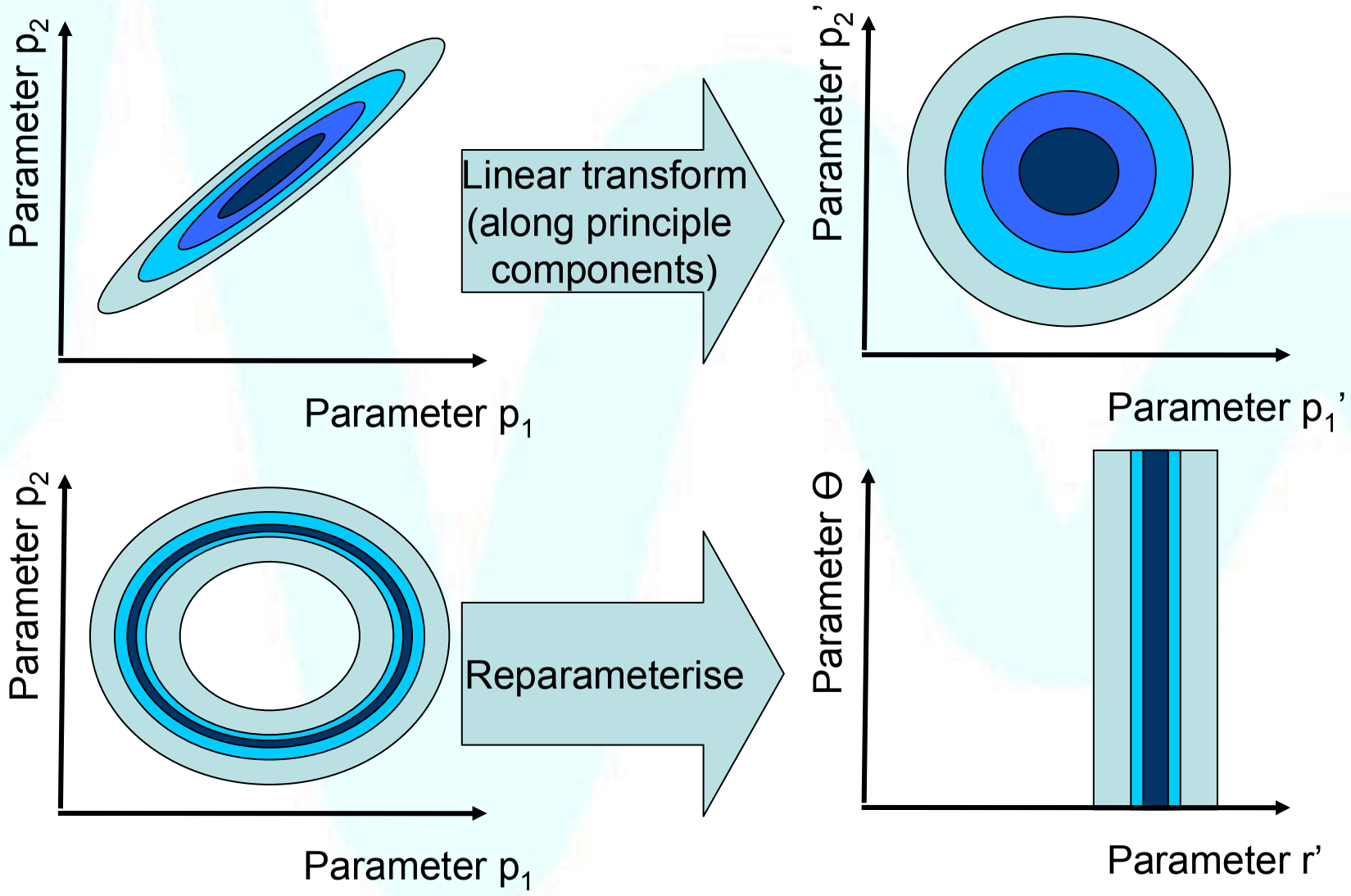


- Build up Posterior Distribution
- Efficient if proposal distribution matches posterior distribution

Obtaining faster convergence

- Methods include
  - Reparameterisation
    - Difficult in process models
    - Directional randomisation is easily applied though
  - Annealing (“heating”)
    - Only gets *to* equilbrium distribution, doesn’t explore it
  - Auxiliary variables
    - Not clearly applicable…

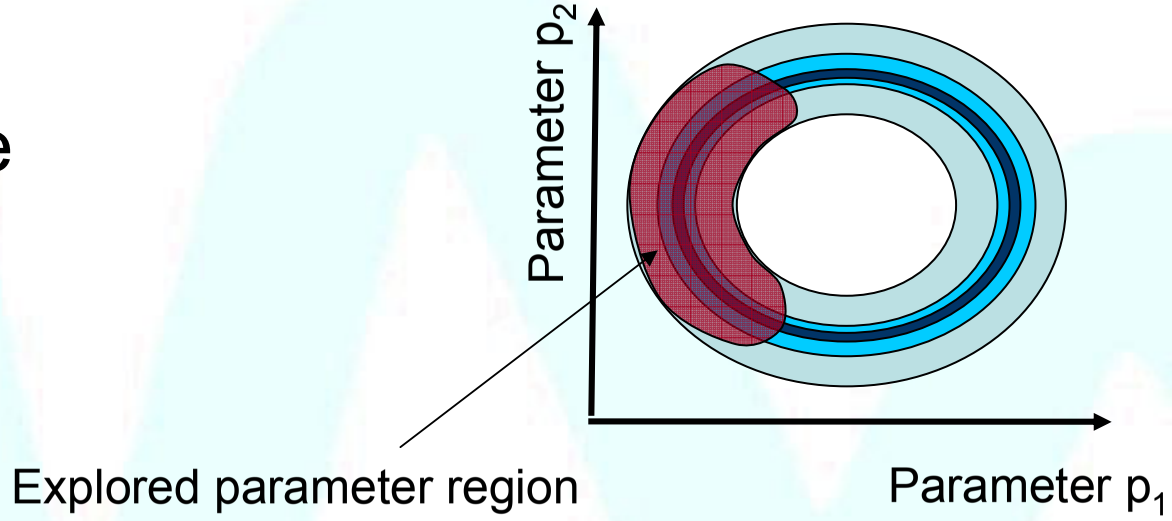
Reparameterisation



Convergence problems

- These solutions work if we can:
  - Detect
  - or Deduce

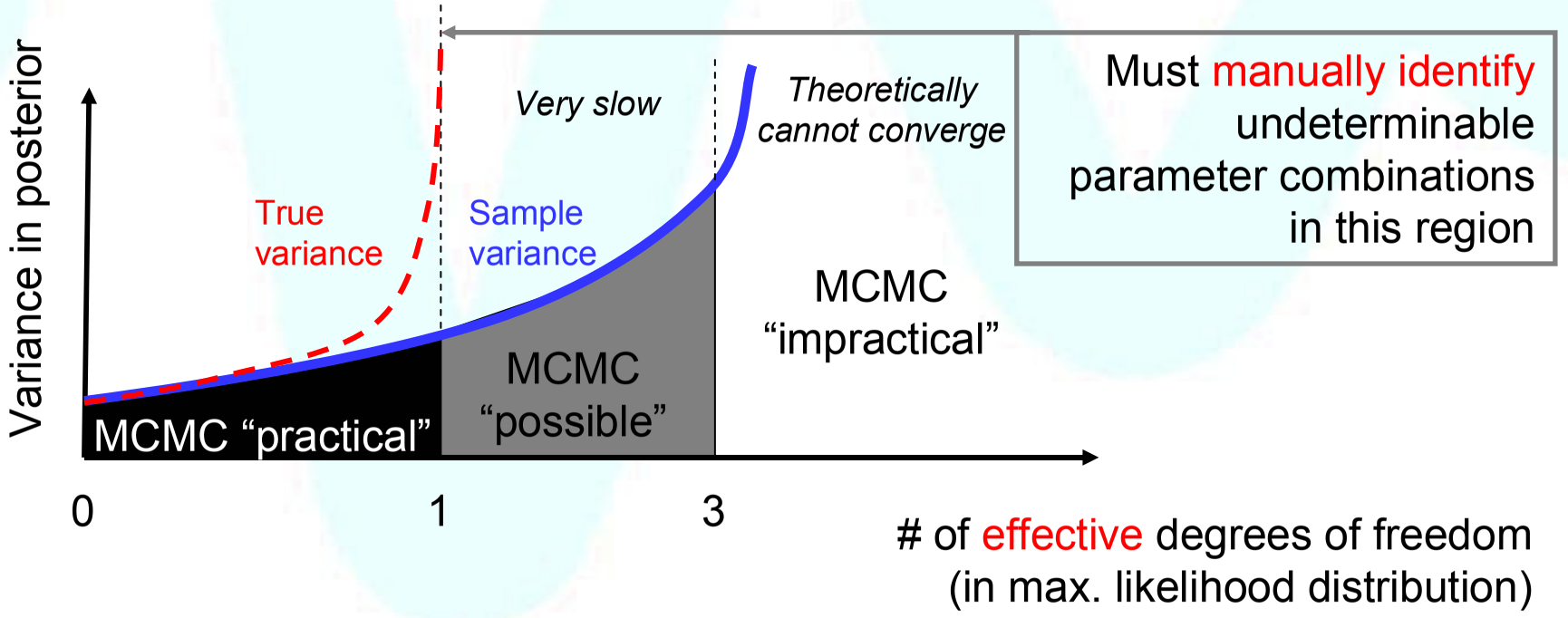
  the shape



- This is difficult in process models
- In practice, not enough data
  - Parameter region is of too high dimension
  - Why: random walk in D>3 doesn't fill space!

MCMC for process models

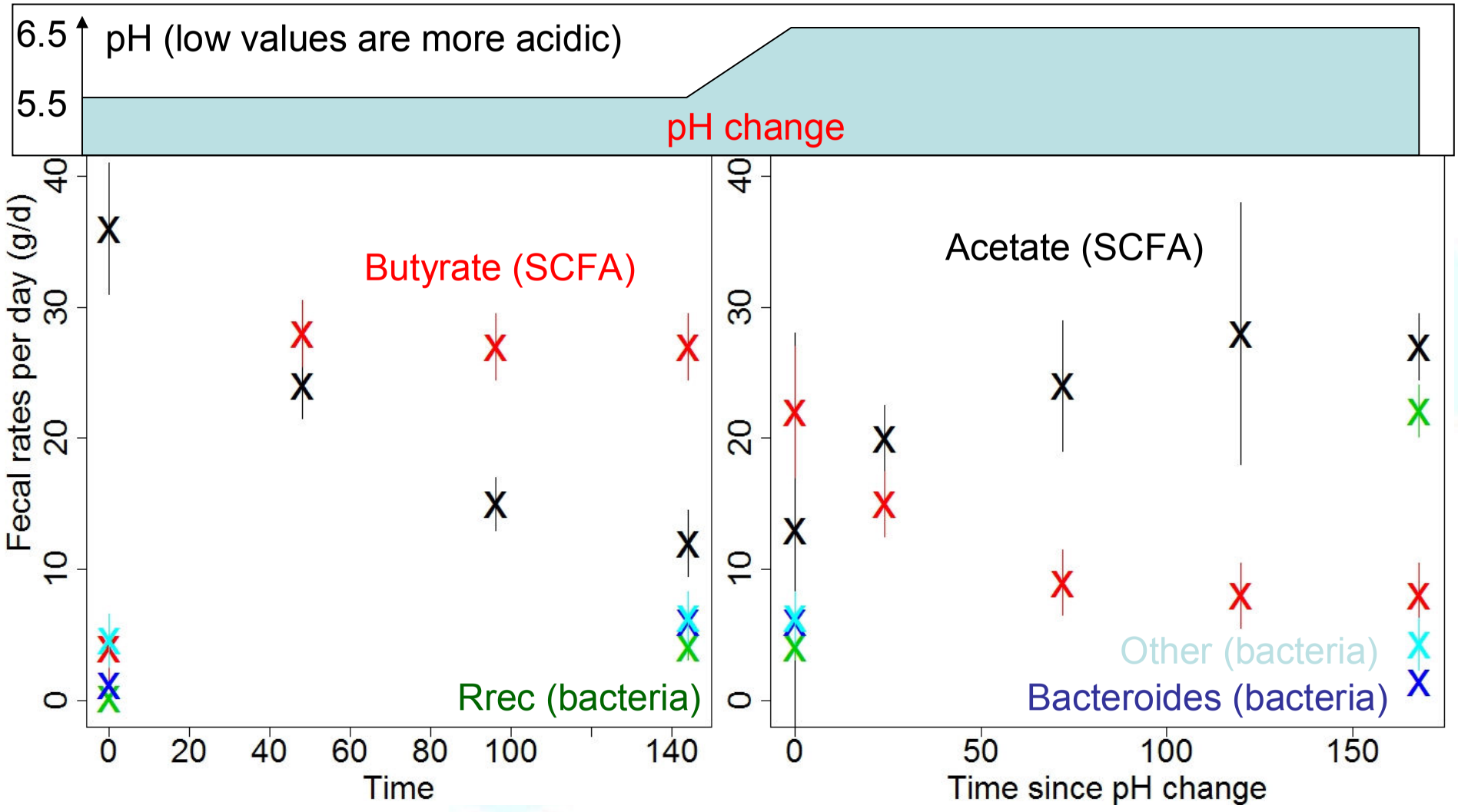
- Increasing data:
  - Reduces parameter region size
  - Reduces degrees of freedom in posterior
  - Reduces total variance



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
- Dangers:
  - Larger state space – might *increase* degrees of freedom!
  - Model may be inconsistent?

Example data



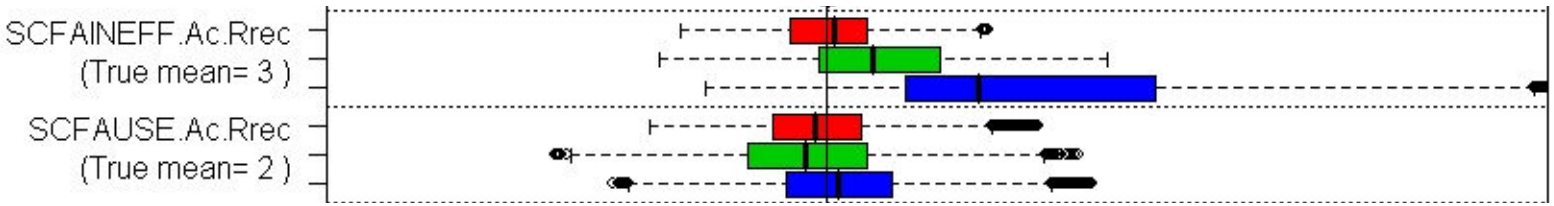
Inference using multiple (simulated) experiments:

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

Bacteroides growth experiment at (c) pH 5.5 & (d) 6.5

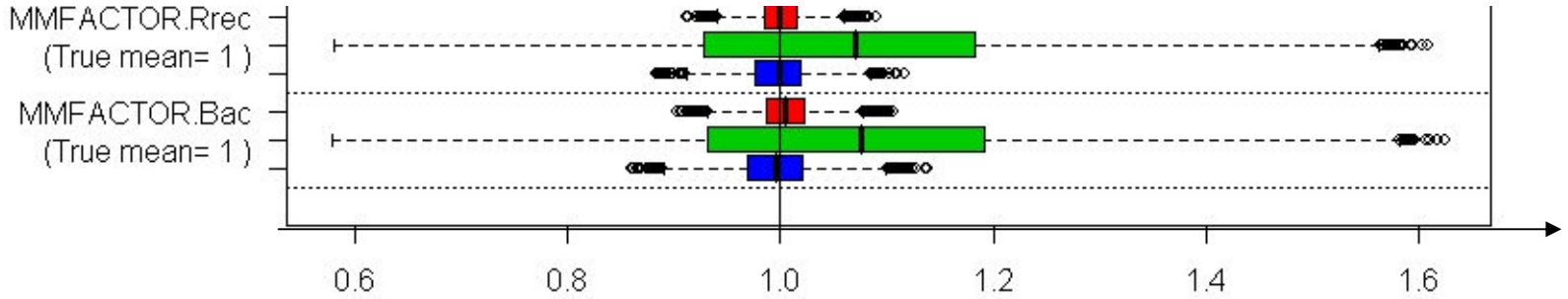
3 scenarios considered: only scenario (1) has converged MCMC chain!

(1) Simultaneous 4 experiments



(2) Not measuring substrate output

(3) Using summary of growth experiments



Scenario (4): True experimental data combines (2) & (3)

Inferred parameter value

Practical application

- Current experimental data
  - MCMC inference fails
  - Gives “reasonable” parameters (maximum likelihood?)
- More data needed: which data?
  - Model predicts if given experiments and measurements will help
- Also predicts *in-vivo* behaviour
  - Establish many possible qualitative behaviours
  - But cannot guarantee that all are found
  - Predictions not quantitative

Conclusions

- Link between quantitative and qualitative process models
- Generally applicable method to other problems
- Successful in principle
  - Needs feedback loop for full application