# Inference for complex models

Richard Wilkinson
r.d.wilkinson@sheffield.ac.uk

School of Maths and Statistics
University of Sheffield

23 April 2015

# Computer experiments

Rohrlich (1991): Computer simulation is

> *'a key milestone somewhat comparable to the milestone that started the empirical approach (Galileo) and the deterministic mathematical approach to dynamics (Newton and Laplace)'*

Challenges for statistics:
How do we make inferences about the world from a simulation of it?

# Computer experiments

Rohrlich (1991): Computer simulation is

> *'a key milestone somewhat comparable to the milestone that started the empirical approach (Galileo) and the deterministic mathematical approach to dynamics (Newton and Laplace)'*

Challenges for statistics:
How do we make inferences about the world from a simulation of it?

- how do we relate simulators to reality?
- how do we estimate tunable parameters?
- how do we deal with computational constraints?
- how do we make uncertainty statements about the world that combine models, data and their corresponding errors?

There is an inherent a lack of quantitative information on the uncertainty surrounding a simulation - unlike in physical experiments.

# Bayesian statistics

Represent all uncertainties as probability distributions:

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

# Bayesian statistics

Represent all uncertainties as probability distributions:

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

- $\pi(\theta|D)$ is the posterior distribution
  - Always hard to compute: SMC$^2$, PGAS, Tempered NUTS-HMC

# Bayesian statistics

Represent all uncertainties as probability distributions:

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

- $\pi(\theta|D)$ is the posterior distribution
  - Always hard to compute: SMC$^2$, PGAS, Tempered NUTS-HMC
- $\pi(D|\theta)$ is the likelihood function.
  - For complex models can be slow to compute: GP emulators
  - Can also be impossible to compute in some cases: ABC
  - $$\pi(D|\theta) = \int \pi(D|X)\pi(X|\theta)\mathrm{d}X$$

    Relating simulator to reality can make specifying $\pi(D|\theta)$ particularly difficult: Simlator discrepancy modelling

# Bayesian statistics

Represent all uncertainties as probability distributions:

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

- $\pi(\theta|D)$ is the posterior distribution
  - Always hard to compute: SMC$^2$, PGAS, Tempered NUTS-HMC
- $\pi(D|\theta)$ is the likelihood function.
  - For complex models can be slow to compute: GP emulators
  - Can also be impossible to compute in some cases: ABC
  - $$\pi(D|\theta) = \int \pi(D|X)\pi(X|\theta)\mathrm{d}X$$

    Relating simulator to reality can make specifying $\pi(D|\theta)$ particularly difficult: Simlator discrepancy modelling
- $\pi(D)$ is the model evidence or normalising constant.
  - Requires us to integrate, and is thus harder to compute than $\pi(\theta|D)$: SMC$^2$, nested sampling

# Uncertainty Quantification (UQ) for computer experiments

- Calibration
  - Estimate unknown parameters $\theta$
  - Usually via the posterior distribution $\pi(\theta|D)$
  - Or history matching

# Uncertainty Quantification (UQ) for computer experiments

- Calibration
  - Estimate unknown parameters $\theta$
  - Usually via the posterior distribution $\pi(\theta|D)$
  - Or history matching
- Uncertainty analysis
  - $f(x)$ a complex simulator. If we are uncertain about $x$, e.g., $X \sim \pi(x)$, what is $\pi(f(X))$?

# Uncertainty Quantification (UQ) for computer experiments

- Calibration
  - Estimate unknown parameters $\theta$
  - Usually via the posterior distribution $\pi(\theta|D)$
  - Or history matching
- Uncertainty analysis
  - $f(x)$ a complex simulator. If we are uncertain about $x$, e.g., $X \sim \pi(x)$, what is $\pi(f(X))$?
- Sensitivity analysis
  - $X = (X_1, \ldots, X_d)^\top$. Can we decompose $\mathbb{V}\mathrm{ar}(f(X))$ into contributions from each $\mathbb{V}\mathrm{ar}(X_i)$?
  - If we can improve our knowledge of any $X_i$, which should we choose to minimise $\mathbb{V}\mathrm{ar}(f(X))$?

# Uncertainty Quantification (UQ) for computer experiments

- Calibration
  - Estimate unknown parameters $\theta$
  - Usually via the posterior distribution $\pi(\theta|D)$
  - Or history matching
- Uncertainty analysis
  - $f(x)$ a complex simulator. If we are uncertain about $x$, e.g., $X \sim \pi(x)$, what is $\pi(f(X))$?
- Sensitivity analysis
  - $X = (X_1, \ldots, X_d)^\top$. Can we decompose $\mathbb{V}\mathrm{ar}(f(X))$ into contributions from each $\mathbb{V}\mathrm{ar}(X_i)$?
  - If we can improve our knowledge of any $X_i$, which should we choose to minimise $\mathbb{V}\mathrm{ar}(f(X))$?
- Simulator discrepancy
  - $f(x)$ is imperfect. How can we quantify or correct simulator discrepancy.

# Uncertainty Quantification (UQ) for computer experiments

- Calibration
  - Estimate unknown parameters $\theta$
  - Usually via the posterior distribution $\pi(\theta|D)$
  - Or history matching
- Uncertainty analysis
  - $f(x)$ a complex simulator. If we are uncertain about $x$, e.g., $X \sim \pi(x)$, what is $\pi(f(X))$?
- Sensitivity analysis
  - $X = (X_1, \ldots, X_d)^\top$. Can we decompose $\mathbb{V}\mathrm{ar}(f(X))$ into contributions from each $\mathbb{V}\mathrm{ar}(X_i)$?
  - If we can improve our knowledge of any $X_i$, which should we choose to minimise $\mathbb{V}\mathrm{ar}(f(X))$?
- Simulator discrepancy
  - $f(x)$ is imperfect. How can we quantify or correct simulator discrepancy.
- Data assimilation
  - Find $\pi(x_{1:t}|y_{1:t})$

# Meta-modelling
# Surrogate modelling
# Emulation

# Code uncertainty

For complex simulators, run times might be long, ruling out brute-force approaches such as Monte Carlo methods.

# Code uncertainty

For complex simulators, run times might be long, ruling out brute-force approaches such as Monte Carlo methods.

Consequently, we will only know the simulator output at a finite number of points.

- We call this *code uncertainty*.
- All inference must be done using a finite ensemble of model runs

$$D_{sim} = \{(\theta_i, f(\theta_i))\}_{i=1,\dots,N}$$

- If $\theta$ is not in the ensemble, then we are uncertainty about the value of $f(\theta)$.

# Meta-modelling

**Idea:** If the simulator is expensive, build a cheap model of it and use this in any analysis.

'a model of the model'

We call this meta-model an *emulator* of our simulator.

# Meta-modelling

**Idea:** If the simulator is expensive, build a cheap model of it and use this in any analysis.

'a model of the model'

We call this meta-model an *emulator* of our simulator.

Gaussian process emulators are most popular choice for emulator.

- Built using an ensemble of model runs $D_{sim} = \{(\theta_i, f(\theta_i))\}_{i=1,\dots,N}$
- They give an assessment of their prediction accuracy $\pi(f(\theta)|D_{sim})$

# Meta-modelling

## Gaussian Process Emulators

Gaussian processes provide a flexible nonparametric distributions for our prior beliefs about the functional form of the simulator:

$$f(\cdot) \sim GP(m(\cdot), \sigma^2 c(\cdot, \cdot))$$

where $m(\cdot)$ is the prior mean function, and $c(\cdot, \cdot)$ is the prior covariance function (semi-definite).
Gaussian processes are invariant under Bayesian updating.

# Meta-modelling

## Gaussian Process Emulators

Gaussian processes provide a flexible nonparametric distributions for our prior beliefs about the functional form of the simulator:

$$f(\cdot) \sim GP(m(\cdot), \sigma^2 c(\cdot, \cdot))$$

where $m(\cdot)$ is the prior mean function, and $c(\cdot, \cdot)$ is the prior covariance function (semi-definite).
Gaussian processes are invariant under Bayesian updating.

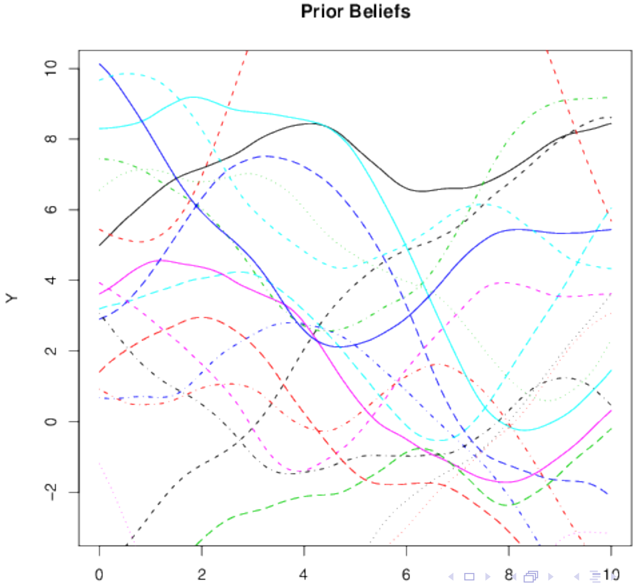
**Definition** If $f(\cdot) \sim GP(m(\cdot), c(\cdot, \cdot))$ then for any collection of inputs $x_1, \ldots, x_n$ the vector

$$(f(x_1), \ldots, f(x_n))^T \sim MVN(m(\mathbf{x}), \sigma^2 \mathbf{\Sigma})$$

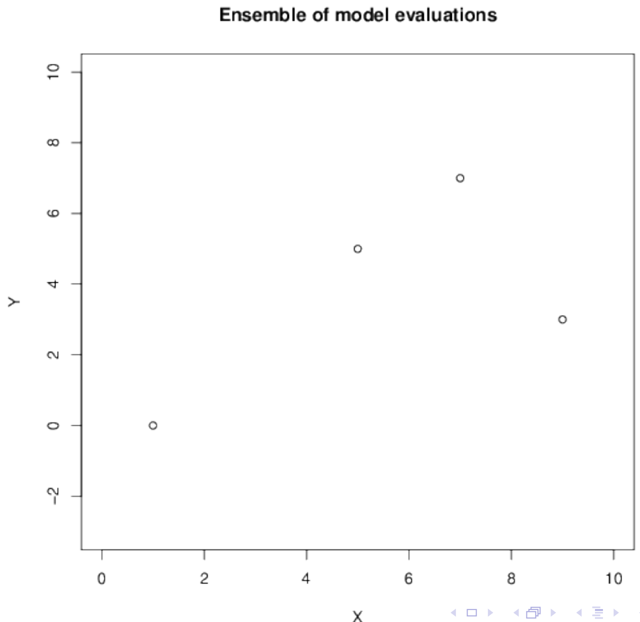
where $\Sigma_{ij} = c(x_i, x_j)$.

# Gaussian Process Illustration

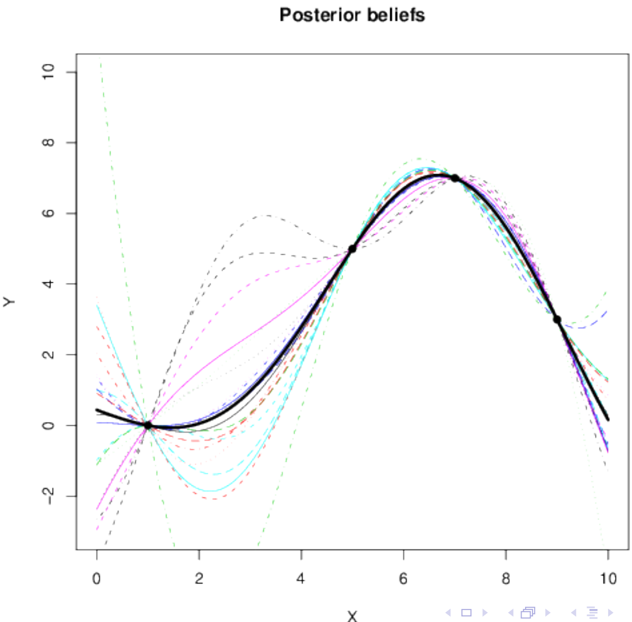
## Zero mean



# Gaussian Process Illustration



# Gaussian Process Illustration

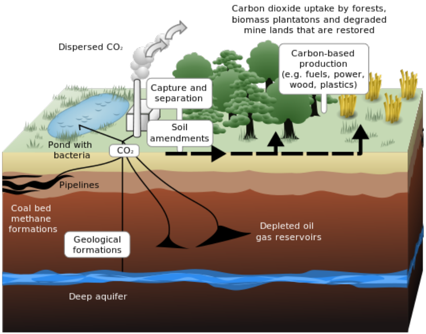


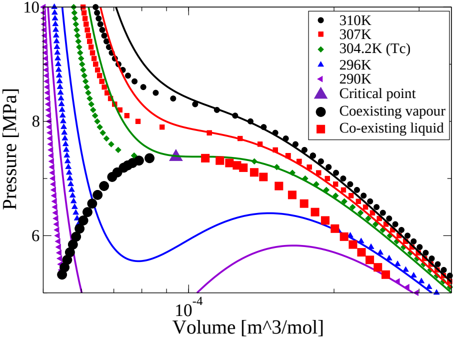
# Challenges

- Design: if we can afford $n$ simulator runs, which parameters should we run it at?
- High dimensional inputs
  - If $\theta$ is multidimensional, then even short run times can rule out brute force approaches
- High dimensional outputs
  - Spatio-temporal.
- Incorporating physical knowledge
- Difficult behaviour, e.g., switches, step-functions, non-stationarity...

# Uncertainty quantification for Carbon Capture and Storage

## EPSRC: transport




Technical challenges:

- How do we find non-parametric Gaussian process models that i) obey the fugacity constraints ii) have the correct asymptotic behaviour
- How do we fit parametric equations of state (Peng-Robinson and variants) - tempered NUTS-HMC.

# Storage Panacea

Knowledge of the physical problem is encoded in a simulator $f$

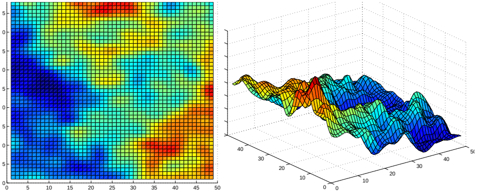
Inputs:

Permeability field, K
(2d field)

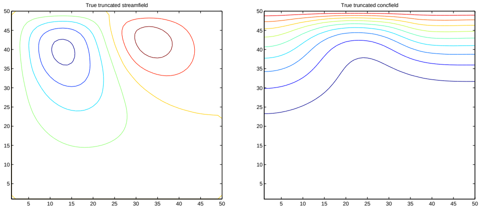
$f(K)$

Outputs:

Stream func. (2d field),
concentration (2d field),
surface flux (1d scalar),
$\vdots$



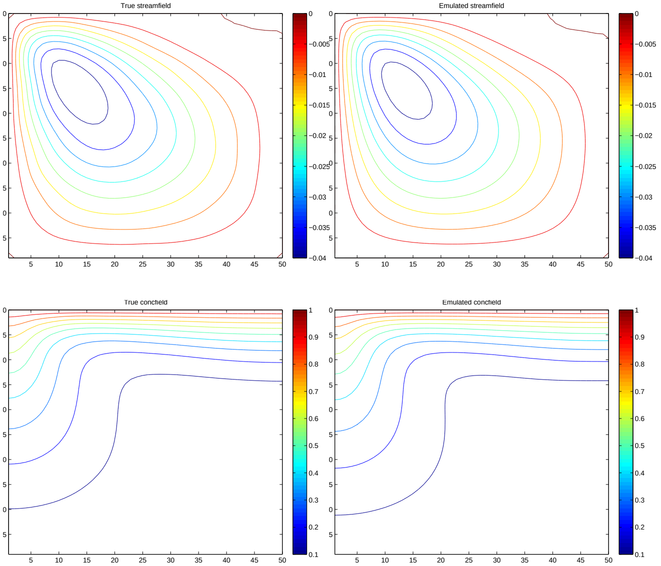
$\downarrow f(K)$




Surface Flux $= 6.43, \ldots$

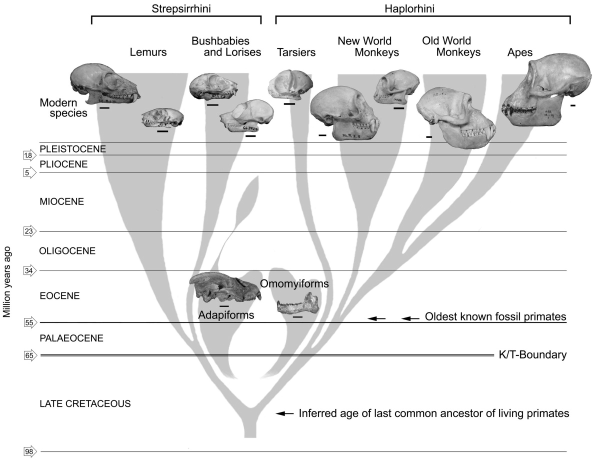
# CCS examples

Left=true, right = emulated, 118 training runs, held out test set.



# ABC: inference for complex stochastic models

# Estimating Divergence Times



# Forward simulation

Model evolution and fossil finds

- Let $\tau$ be the temporal gap between the divergence time and the oldest fossil.

The posterior for $\tau$ is then used as a prior for a genetic analysis.

The likelihood function $\pi(D|\theta)$ is intractable, but it is cheap to simulate.

# Approximate Bayesian Computation (ABC)

Wilkinson 2008/2013, Wilkinson and Tavaré 2009

If the likelihood function is intractable, then ABC is one of the few approaches we can use to do inference.

### Uniform Rejection Algorithm

- Draw $\theta$ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept $\theta$ if $\rho(D, X) \leq \epsilon$

$\epsilon$ reflects the tension between computability and accuracy.

- As $\epsilon \to \infty$, we get observations from the prior, $\pi(\theta)$.
- If $\epsilon = 0$, we generate observations from $\pi(\theta \mid D)$.

# Approximate Bayesian Computation (ABC)

Wilkinson 2008/2013, Wilkinson and Tavaré 2009

If the likelihood function is intractable, then ABC is one of the few approaches we can use to do inference.

## Uniform Rejection Algorithm

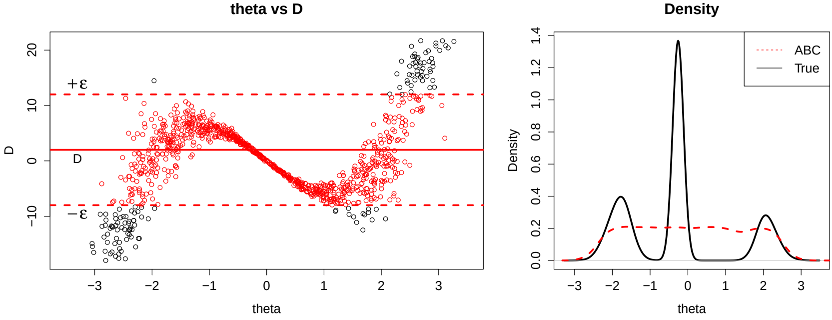
- Draw $\theta$ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept $\theta$ if $\rho(D, X) \leq \epsilon$

$\epsilon$ reflects the tension between computability and accuracy.

- As $\epsilon \to \infty$, we get observations from the prior, $\pi(\theta)$.
- If $\epsilon = 0$, we generate observations from $\pi(\theta \mid D)$.

ABC does not require explicit knowledge of the likelihood function

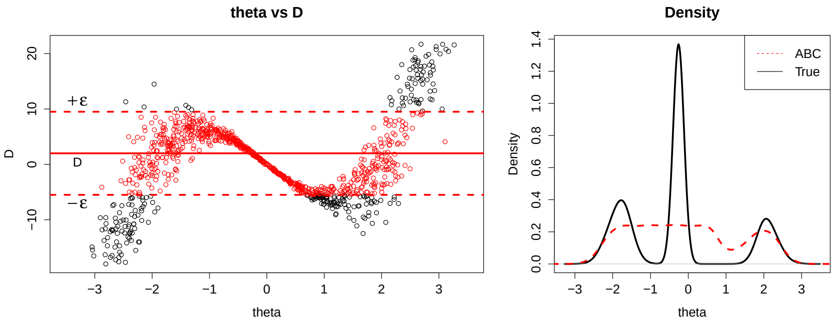
# $\epsilon = 10$



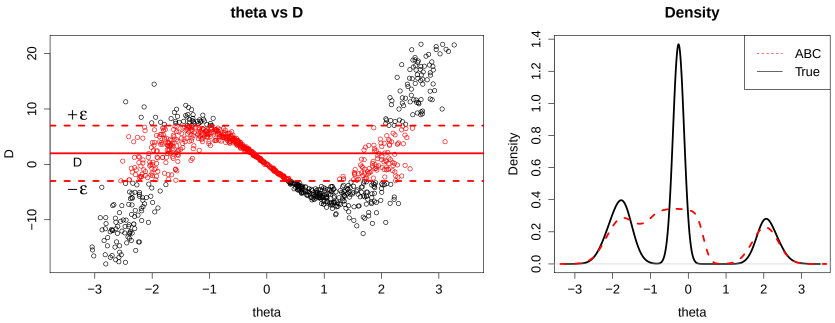

$$\theta \sim U[-10, 10], \qquad X \sim N(2(\theta+2)\theta(\theta-2), 0.1+\theta^2)$$

$$\rho(D, X) = |D - X|, \qquad D = 2$$

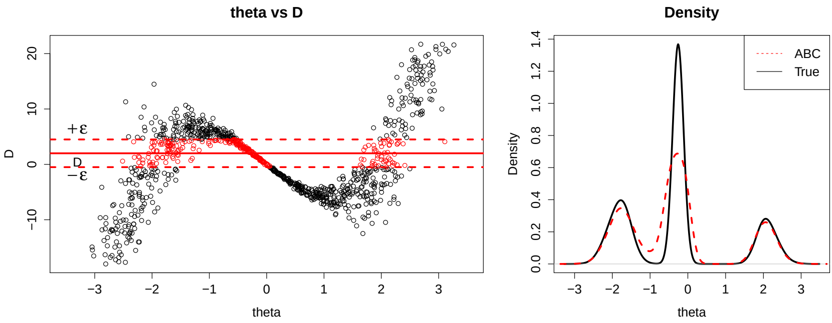
# $\epsilon = 7.5$


theta vs D
+ε
D
−ε
D
20
10
0
−10
−3
−2
−1
0
1
2
3
theta
Density
ABC
True
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
theta

## $\epsilon = 5$

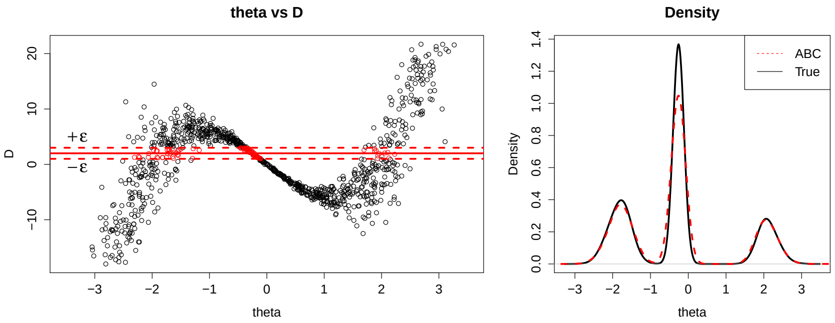


# $\epsilon = 2.5$



theta vs D
+ε
D
−ε
20
10
0
−10
−3
−2
−1
0
1
2
3
theta
Density
ABC
True
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

# $\epsilon = 1$


theta vs D
D
20
10
0
−10
+ε
−ε
−3
−2
−1
0
1
2
3
theta
Density
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
ABC
True
−3
−2
−1
0
1
2
3
theta

# Rejection ABC

If the data are too high dimensional we never observe simulations that are 'close' to the field data - curse of dimensionality

Reduce the dimension using summary statistics, $S(D)$.

### Approximate Rejection Algorithm With Summaries

- Draw $\theta$ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept $\theta$ if $\rho(S(D), S(X)) < \epsilon$

If $S$ is sufficient this is equivalent to the previous algorithm.

# Rejection ABC

If the data are too high dimensional we never observe simulations that are 'close' to the field data - curse of dimensionality

Reduce the dimension using summary statistics, $S(D)$.

### Approximate Rejection Algorithm With Summaries

- Draw $\theta$ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept $\theta$ if $\rho(S(D), S(X)) < \epsilon$

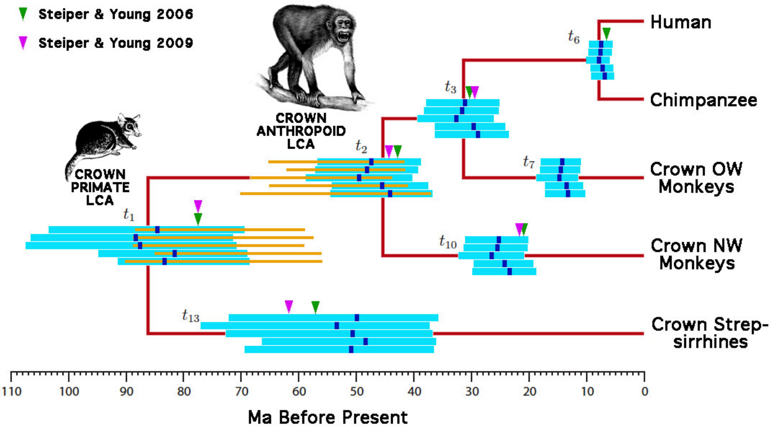
If $S$ is sufficient this is equivalent to the previous algorithm.

Simple $\rightarrow$ Popular with non-statisticians

$\exists$ many extensions and improvements

- How to choose $S(D)$
- How to efficiently sample $\theta$

# An integrated molecular and palaeontological analysis




The fossil record does not constrain the primate divergence time as closely as previously believed.

- Genetic and palaeontology estimates unified
- Human-chimp divergence time pushed further back.

Wilkinson *et al.* 2011, Bracken-Grissom *et al.* 2014.

# Accelerating ABC: GP-ABC

Monte Carlo methods (such as ABC) are costly and can require more simulation than is possible. However,

- most methods sample naively - they don't learn from previous simulations
- they don't exploit known properties of the likelihood function, such as continuity
- they sample randomly, rather than using careful design.

Emulators are the usual approach to dealing with complex models. But, emulating stochastic simulators is problematic.

# Accelerating ABC: GP-ABC

Monte Carlo methods (such as ABC) are costly and can require more simulation than is possible. However,

- most methods sample naively - they don't learn from previous simulations
- they don't exploit known properties of the likelihood function, such as continuity
- they sample randomly, rather than using careful design.

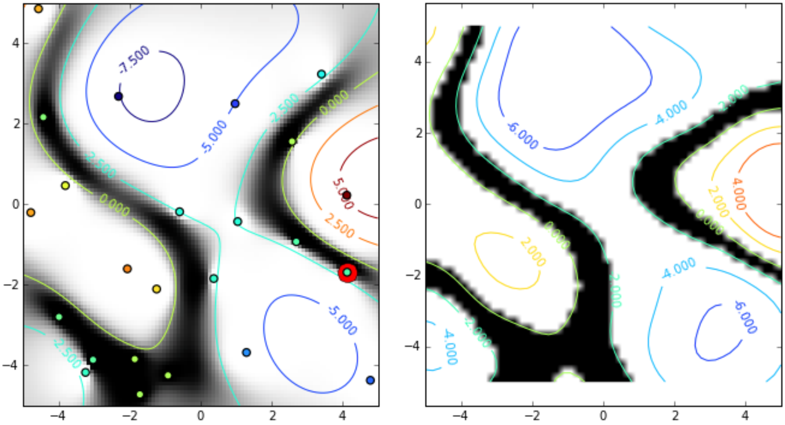
Emulators are the usual approach to dealing with complex models. But, emulating stochastic simulators is problematic.

Instead of modelling the simulator output, we can instead model $L(\theta) = \pi(D|\theta)$

- $D$ remains fixed: we only need learn $L$ as a function of $\theta$
- 1d response surface
- But, it can be hard to model.

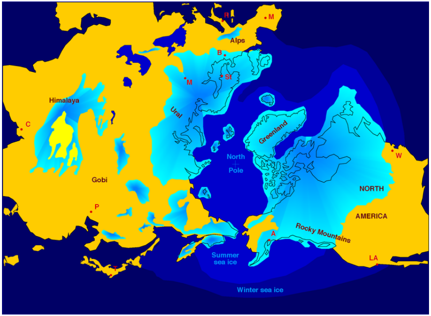
# Iteration 24

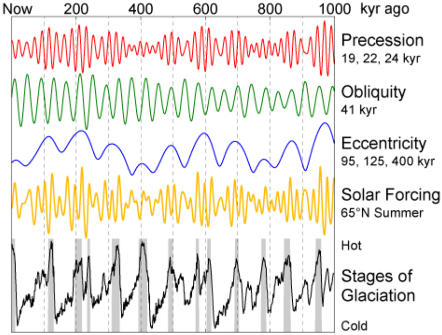
Left=estimate, right = truth



# Climate science

## What drives the glacial-interglacial cycle?




Eccentricity: orbital departure from a circle, controls duration of the seasons
Obliquity: axial tilt, controls amplitude of seasonal cycle
Precession: variation in Earth's axis of rotation, affects difference between seasons

# Model selection

What drives the glacial-interglacial cycle?

- Which aspect of the astronomical forcing is of primary importance?
- Which models best represent the cycle?

> *Most simple models of the [...] glacial cycles have at least four degrees of freedom [parameters], and some have as many as twelve. Unsurprisingly [...this is] insufficient to distinguish between the skill of the various models (Roe and Allen 1999)*

# Model selection

What drives the glacial-interglacial cycle?

- Which aspect of the astronomical forcing is of primary importance?
- Which models best represent the cycle?

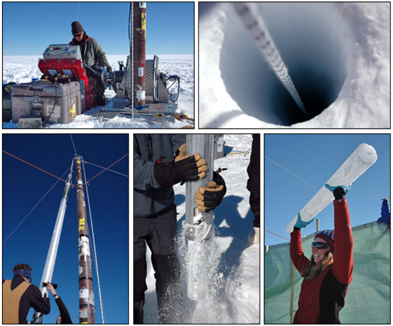
> *Most simple models of the [...] glacial cycles have at least four degrees of freedom [parameters], and some have as many as twelve. Unsurprisingly [...this is] insufficient to distinguish between the skill of the various models (Roe and Allen 1999)*

Bayesian model selection revolves around the use of the Bayes factor, which are notoriously difficult to compute.

- Model selection for stochastic differential equations
- 1000 observations, 3000 unknown state variables, 1000 unknown times, 17 unknown parameters, choice of 5 different simulators.

Simulation studies show we can accurately choose between competing models, and identify the correct forcing.

# Age model



Can we also quantify chronological uncertainty?

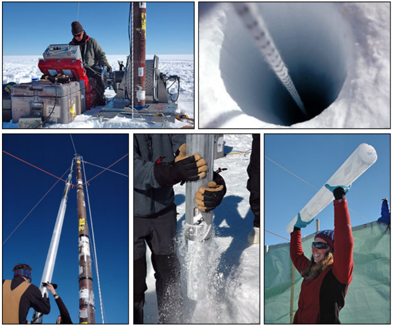
$$\begin{aligned} \mathrm{d}X_t &= g(X_t, \theta)\mathrm{d}t + F(t, \gamma)\mathrm{d}t + \Sigma \mathrm{dW} \\ Y_t &= d + sX_{1,t} + \epsilon_t \end{aligned}$$

Plus an age model

$$\mathrm{d}H = -\mu_s \mathrm{d}T + \sigma \mathrm{d}W$$

I.e., can we simultaneously date the stack, do climate reconstruction, fit the model, and choose between models?

# Age model



Target

Can we also quantify chronological uncertainty?

$$
\begin{aligned}
\mathrm{d}X_t &= g(X_t, \theta)\mathrm{d}t + F(t, \gamma)\mathrm{d}t + \Sigma \mathrm{dW} \\
Y_t &= d + sX_{1,t} + \epsilon_t
\end{aligned}
$$

Plus an age model

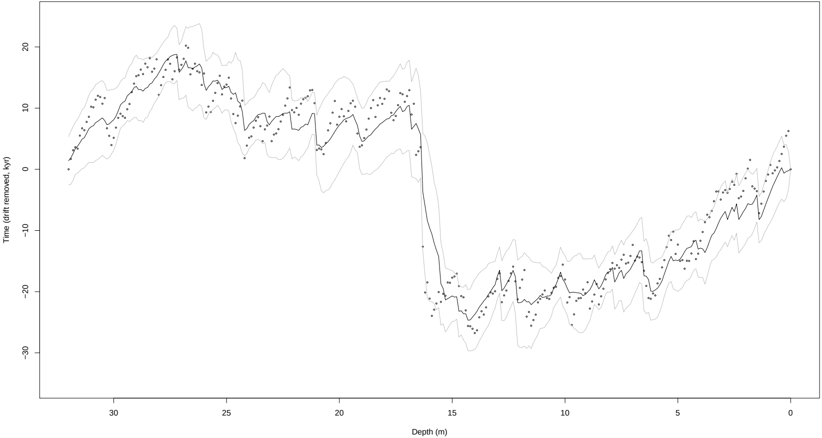
$$\mathrm{d}H = -\mu_s \mathrm{d}T + \sigma \mathrm{d}W$$

$$\pi(\theta, T_{1:N}, X_{1:N}, \mathcal{M}_k | y_{1:N})$$

where $T_{1:N}$ are the unknown times of the observations $Y_{1:N}$, $X_{1:N}$ are the climate state variables through time, $\mathcal{M}_k$ is the simulation model used, and $\theta$ is the corresponding parameter.
I.e., can we simultaneously date the stack, do climate reconstruction, fit the model, and choose between models?

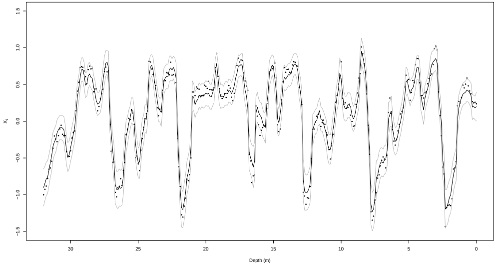
# Simulation study results - age vs depth (trend removed)

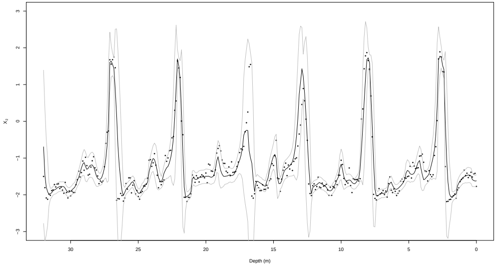
Dots = truth, black line = estimate, grey = 95% CI



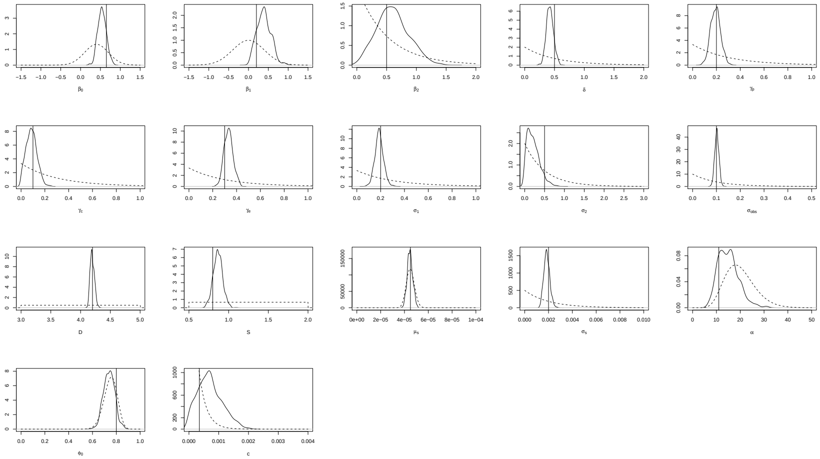
# Simulation study results - climate reconstruction

## Dots = truth, black line = estimate, grey = 95% CI



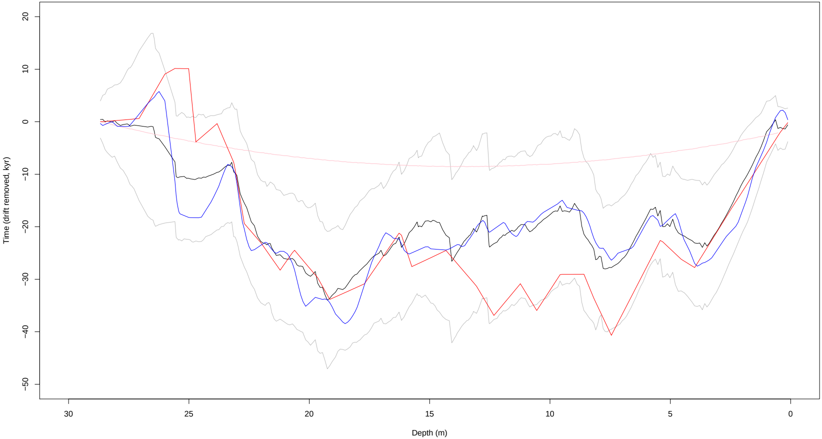
# Simulation study results - parameter estimation




Simultaneous inference of the choice between 5 models, 17 parameters, 800 ages, 2400 climate variables, using just 800 observations.

# Results for ODP846 - age vs depth (trend removed)

Black = posterior mean, grey = 95%CI, red = Huybers 2007, blue = Lisieki and Raymo 2004




Advantages: full UQ, model selection, simultaneous parameter estimation and climate reconstruction

Ignoring uncertainty leads to incorrect conclusions

# Model discrepancy

Consider the state space model:

$$x_{t+1} = f_\theta(x_t) + e_t, \qquad y_t = g(x_t) + \epsilon_t$$

$$e_t \sim p(\cdot), \qquad \epsilon_t \sim q(\cdot)$$

How do we correct errors in $f$ or $g$?

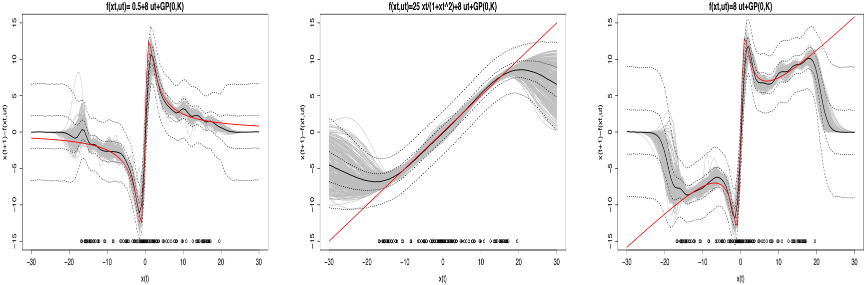
# Model discrepancy

Consider the state space model:

$$x_{t+1} = f_\theta(x_t) + e_t, \qquad y_t = g(x_t) + \epsilon_t$$

$$e_t \sim p(\cdot), \qquad \epsilon_t \sim q(\cdot)$$

How do we correct errors in $f$ or $g$?
Use a GP discrepancy model - eg, $x_{t+1} = f_\theta(x_t) + \delta(x_t) + e_t$




Technical challenge: inference using PGAS works but is expensive. A variational approach looks more promising.

# Conclusions

- UQ can be vital: ignoring uncertainty can lead to incorrect conclusions, often in subtle ways.
- Computational tractability is one of the key bottlenecks: big simulation and big data.
- Methods from machine learning have the potential to help us make large advances in statistical methodology.

# Conclusions

- UQ can be vital: ignoring uncertainty can lead to incorrect conclusions, often in subtle ways.
- Computational tractability is one of the key bottlenecks: big simulation and big data.
- Methods from machine learning have the potential to help us make large advances in statistical methodology.

Thank you for listening!