Statistical analysis of computer experiments

Richard Wilkinson
r.d.wilkinson@nottingham.ac.uk
School of Mathematical Sciences
University of Nottingham

Vet School collaboration workshop

Computer experiments

Baker 1977 (Science): ‘Computerese is the new lingua franca of science’

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? (model error)
- How do we estimate tunable parameters? (calibration)
- How do we deal with computational constraints? (stat. comp.)
- How do we make uncertainty statements about the world that combine models, data and their corresponding errors? (UQ)

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

Problem 1: Inverse Problems

Fitting models to data/calibration/parameter estimation

Most models are forwards models, i.e., specify parameters $\theta$ and i.c.s and the model generates output $D$. Usually, we are interested in the inverse-problem, i.e., observe data, want to estimate parameter values.

When did the primates evolve?
- Oldest primate fossil is 54.8 My old.
- Palaeontologists estimate the date to be 60-65 Mya.
- Geneticists estimate 80-100 Mya.

Approximate Bayesian computation

Approximate Bayesian computation (ABC) algorithms are a collection of Monte Carlo algorithms used for calibrating simulators
- they do not require explicit knowledge of the likelihood function $\pi(x|\theta)$
- instead, inference is done using simulation from the model (consequently they are sometimes called ‘likelihood-free’).

ABC methods are becoming very popular in the biological sciences.

Uniform ABC
- Draw $\theta$ from $\pi(\theta)$
- Simulate $X \sim f(\theta)$
- Accept $\theta$ if $\rho(D,X) \leq \delta$ - i.e. if $D$ is close to $X$. 
Combining uncertainties and data sets
An integrated molecular and palaeontological analysis

The primate fossil record does not precisely constrain the primate divergence time

Problem 2: Representation of uncertainty
Probability (Bayesian) can be used to represent uncertainty.
- Under minimal rationality assumptions, probability can be shown to be the only rational way to represent uncertainty.

Problem 2: Representation of uncertainty
Probability (Bayesian) can be used to represent uncertainty.
- Under minimal rationality assumptions, probability can be shown to be the only rational way to represent uncertainty.
- Probability is subjective probability - distributions represent degrees of belief of individuals. There is no escaping this interpretation in many applications!

Problem 2: Representation of uncertainty
Probability (Bayesian) can be used to represent uncertainty.
- Under minimal rationality assumptions, probability can be shown to be the only rational way to represent uncertainty.
- Probability is subjective probability - distributions represent degrees of belief of individuals. There is no escaping this interpretation in many applications!

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.

There are many types of emulator.

- ideally an emulator should come with an assessment of its accuracy
- rather than just predicting \( \eta(\theta) \) it should predict \( \pi(\eta(\theta)|D_{sim}) \) - our uncertainty about the simulator value given the ensemble \( D_{sim} \).

Gaussian process emulators are most popular choice for emulator. Built using

- an ensemble of model runs \( D_{sim} = \{(\theta_i, \eta(\theta_i))\}_{i=1}^N \)
- expert opinion about the simulator output.

Gaussian Process Illustration

Zero mean

Prior Beliefs

\[ Y \]

\[ X \]

Posterior beliefs

\[ Y \]

\[ X \]

Estimating Carbon Cycle Feedbacks using UVic

- Inputs: \( Q_{10} = \) soil respiration sensitivity to temperature (carbon source) and \( K_c = \) CO2 fertilization of photosynthesis (carbon sink).
- Output: time-series of CO2 values,

Quantifying code uncertainty, simulator error, measurement error.

Conclusions

Statistical analysis of computer experiments is a rapidly growing field.

Techniques for combining empirical data with complex models are improving all the time

Uncertainty quantification is increasingly being recognised as a crucial part of modelling:

- Design of simulation experiments
- Uncertainty analysis
- Sensitivity analysis
- Parametric uncertainty
- Parameter estimation
- Model error