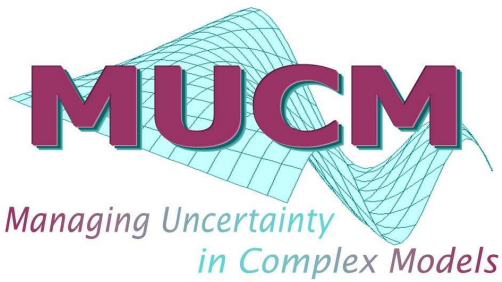
Análise Estatística de Simuladores
Lecture 1: Introduction to Computer Experiments

Leonardo Bastos[1] and Richard Wilkinson[2]

[1]Universidade Federal Fluminense
[2]University of Nottingham, UK

19º SINAPE, São Pedro, 26 July, 2010



Computer experiments

Statistics is the scientific discipline that creates methodology for empirical research.

The gold-standard of empirical research is the designed experiment, studied in depth by R.A. Fisher, and involves concepts such as

- Replication
- Blocking
- Randomization

Computer experiments

Statistics is the scientific discipline that creates methodology for empirical research.

The gold-standard of empirical research is the designed experiment, studied in depth by R.A. Fisher, and involves concepts such as

- Replication
- Blocking
- Randomization

In the past three decades, computer experiments (*in silico* experiments) have become commonplace in nearly all areas of human endeavour.

Although the statistical challenges posed by physical experiments are well known, the challenges posed by computer experiments are somewhat different and have only recently begun to be tackled by statisticians.

Computer experiments

For example, for deterministic computer experiments, the concepts of replication, randomization and blocking are irrelevant because a computer model will give identical answers if run multiple times.

Methods to quantify, analyse and reduce uncertainty in the application of computer experiments are attracting increasing attention amongst users of simulation, and in this course we will describe some of the challenges faced, and introduce some of the developing methodology for dealing with these issues.

Plan of lectures

1. Lecture 1: Introduction to computer experiments
2. Lecture 2: Introduction to meta-modelling
3. Lecture 3: Design of experiments and multidimensional emulators
4. Lecture 4: Calibration
5. Lecture 5: Validation and Sensitivity Analysis
6. Lecture 6: Approximate Bayesian computation (ABC)

Plan of lectures

1. Lecture 1: Introduction to computer experiments
2. Lecture 2: Introduction to meta-modelling
3. Lecture 3: Design of experiments and multidimensional emulators
4. Lecture 4: Calibration
5. Lecture 5: Validation and Sensitivity Analysis
6. Lecture 6: Approximate Bayesian computation (ABC)

Please ask questions whenever you like to ask for further information or clarifications etc.

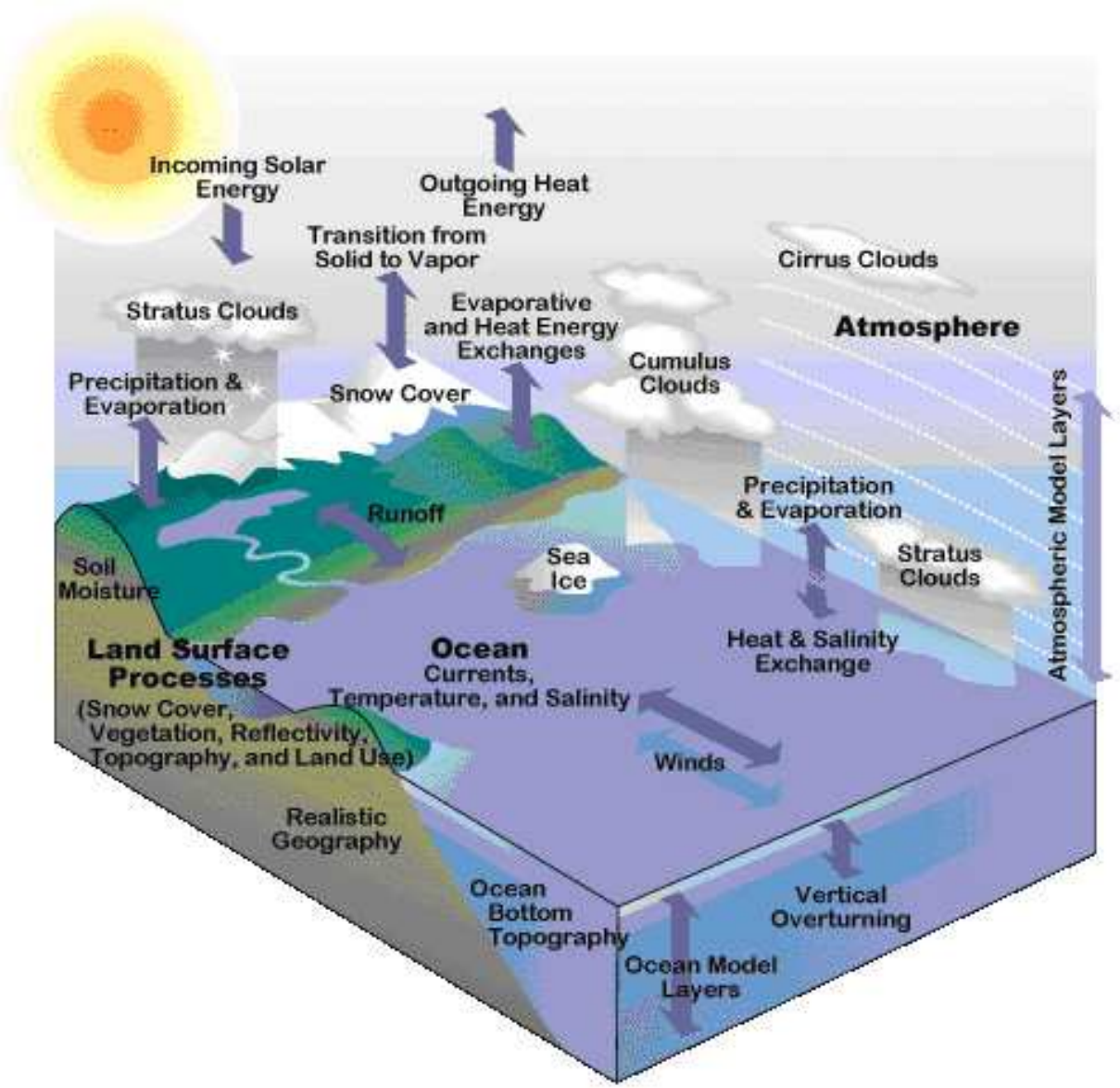
Plan of lectures

1. Lecture 1: Introduction to computer experiments
2. Lecture 2: Introduction to meta-modelling
3. Lecture 3: Design of experiments and multidimensional emulators
4. Lecture 4: Calibration
5. Lecture 5: Validation and Sensitivity Analysis
6. Lecture 6: Approximate Bayesian computation (ABC)

Please ask questions whenever you like to ask for further information or clarifications etc. We will start with some examples we have worked on.

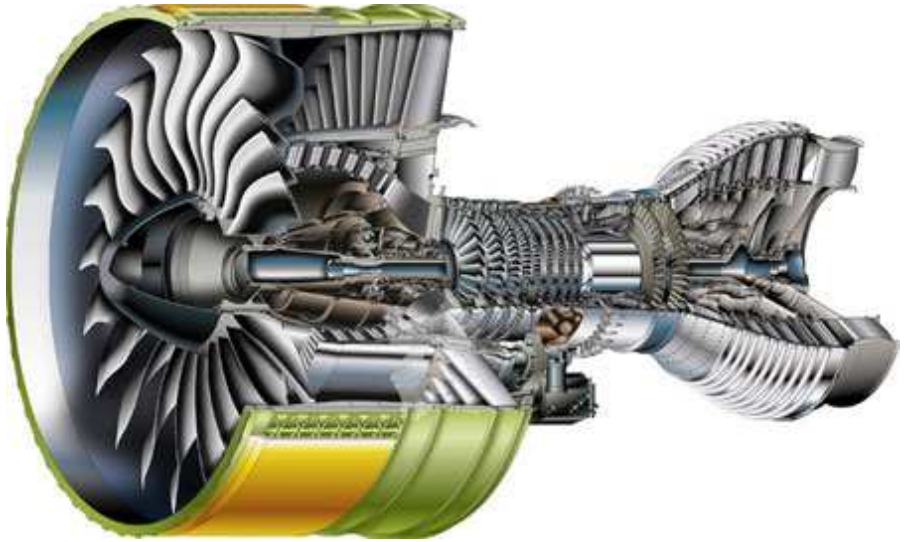
Climate Science

Predicting future climate



Engineering

Engine Health System Monitoring - Rolls Royce



We can't monitor the internals of the engine - only the outputs.

Public policy decisions

NHS drug funding decisions

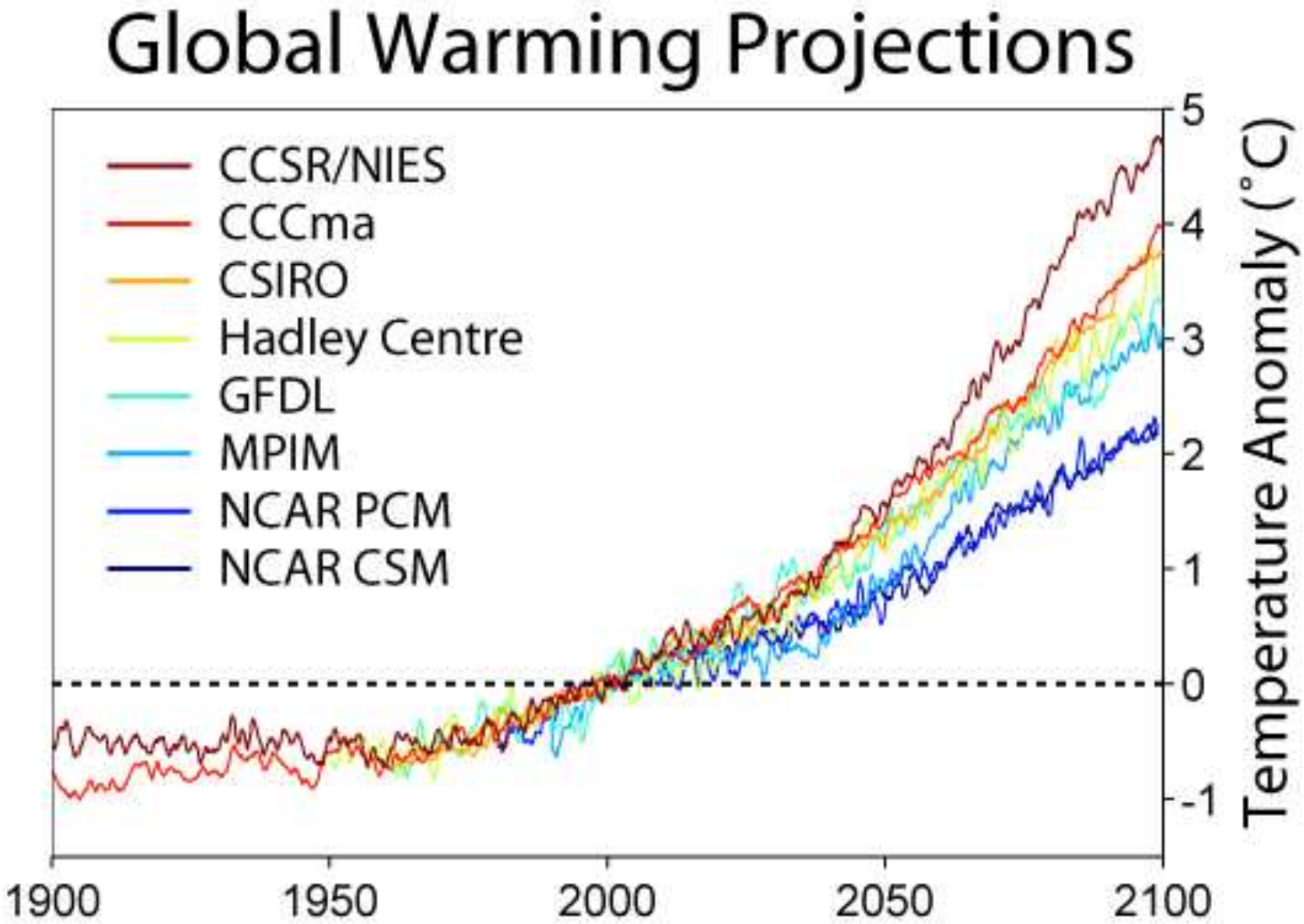
The National Health Service in the UK is a publically funded health care system. It has a finite amount of money and must decide what drugs to provide.
Complex models are used which simulate

- efficacy
- cost
- usage

and this feeds into a decision problem.
The process receives a lot of media attention and so must be transparent and carefully account for possible errors.

Challenges of computer experiments

Climate Predictions



Challenges of computer experiments

Each model is largely based on the same physics, yet the predictions all differ.
Why?

Challenges of computer experiments

Each model is largely based on the same physics, yet the predictions all differ.
Why?

- The models rely on unknown parameters - each prediction above is obtained using a single parameter value. The predictions have not accounted for parametric error.

Challenges of computer experiments

Each model is largely based on the same physics, yet the predictions all differ.
Why?

- The models rely on unknown parameters - each prediction above is obtained using a single parameter value. The predictions have not accounted for parametric error.
- Because each model can take several months to run on a supercomputer, it is not known whether the different models could all similar predictions if the right parameter value was used.

Challenges of computer experiments

Each model is largely based on the same physics, yet the predictions all differ.
Why?

- The models rely on unknown parameters - each prediction above is obtained using a single parameter value. The predictions have not accounted for parametric error.
- Because each model can take several months to run on a supercomputer, it is not known whether the different models could all similar predictions if the right parameter value was used.
- No model is perfect - all models are wrong to a certain degree. None of the predictions have accounted for model error.

Challenges of computer experiments

Models for climate change produce different predictions for the extent of global warming and other consequences

- Which predictions should we believe?
- What error bounds should we put around the predictions?
- Are model differences consistent with the error bounds?

Until we can answer such questions convincingly, governments can continue to dismiss the science.

Challenges of computer experiments

Models for climate change produce different predictions for the extent of global warming and other consequences

- Which predictions should we believe?
- What error bounds should we put around the predictions?
- Are model differences consistent with the error bounds?

Until we can answer such questions convincingly, governments can continue to dismiss the science.

There is increasing concern about uncertainty in model outputs (cf. swine flu and climate predictions etc)

- particularly where model predictions are used to inform scientific debate or environmental policy

Methods that can be used to assess whether model predictions are robust enough for high stakes decision-making are being developed in the statistics community - we shall present a few ideas that are becoming increasingly popular.

Notation

Throughout this course, we think of the simulator as a function

$$\eta : \mathcal{X} \to \mathcal{Y}$$

Typically both the input and output space will be subsets of $\mathbb{R}^n$ for some n.

Notation

Throughout this course, we think of the simulator as a function

$$\eta : \mathcal{X} \to \mathcal{Y}$$

Typically both the input and output space will be subsets of $\mathbb{R}^n$ for some n.

Sometimes it will be necessary to split the input into two types

- Control inputs x - e.g., location, time, output index etc
- Model parameters θ - e.g., model constants, unknown initial conditions, fudge factors

and write $\eta(x, \theta)$ for the model output.

Notation

Throughout this course, we think of the simulator as a function

$$\eta : \mathcal{X} \to \mathcal{Y}$$

Typically both the input and output space will be subsets of $\mathbb{R}^n$ for some n.

Sometimes it will be necessary to split the input into two types

- Control inputs x - e.g., location, time, output index etc
- Model parameters θ - e.g., model constants, unknown initial conditions, fudge factors

and write $\eta(x, \theta)$ for the model output.

The model output will sometimes be univariate $y = \eta(x, \theta) \in \mathbb{R}$, but will often be multivariate, such as a spatial temporal field.

Challenges

Some typical problems we might be interested in:

- Calibration
  - How do we estimate unknown model parameters θ given observations $\mathcal{D}$ of the physical system?

Challenges

Some typical problems we might be interested in:

- Calibration
  - How do we estimate unknown model parameters θ given observations $\mathcal{D}$ of the physical system?
- Prediction
  - Given all the unknowns, what is our best prediction and how confident are we in it?

Challenges

Some typical problems we might be interested in:

- Calibration
  - How do we estimate unknown model parameters θ given observations $\mathcal{D}$ of the physical system?
- Prediction
  - Given all the unknowns, what is our best prediction and how confident are we in it?
- Uncertainty analysis
  - How does uncertainty about the model inputs θ feed through the model?

Challenges

Some typical problems we might be interested in:

- Calibration
  - How do we estimate unknown model parameters θ given observations $\mathcal{D}$ of the physical system?
- Prediction
  - Given all the unknowns, what is our best prediction and how confident are we in it?
- Uncertainty analysis
  - How does uncertainty about the model inputs θ feed through the model?
- Sensitivity analysis
  - How can we apportion variation in the output to variation in the input parameters? In other words, what inputs are driving the output?

Challenges

Some typical problems we might be interested in:

- Calibration
  - How do we estimate unknown model parameters θ given observations $\mathcal{D}$ of the physical system?
- Prediction
  - Given all the unknowns, what is our best prediction and how confident are we in it?
- Uncertainty analysis
  - How does uncertainty about the model inputs θ feed through the model?
- Sensitivity analysis
  - How can we apportion variation in the output to variation in the input parameters? In other words, what inputs are driving the output?
- Dealing with long runs times
  - If a model takes a week to run, we can't use a brute force Monte Carlo approach.

Incorporating and accounting for uncertainty

Perhaps the biggest challenge faced is incorporating uncertainty in computer experiments.

We are used to dealing with uncertainty in physical experiments. But if your computer model is deterministic, there is no natural source of variation and so the experimenter must carefully assess where errors might arise.

Incorporating and accounting for uncertainty

Perhaps the biggest challenge faced is incorporating uncertainty in computer experiments.

We are used to dealing with uncertainty in physical experiments. But if your computer model is deterministic, there is no natural source of variation and so the experimenter must carefully assess where errors might arise.

Types of uncertainty we will consider are

- Parametric uncertainty
- Model inadequacy
- Observation errors
- Code uncertainty

Before we look at these different sources, we shall make a few comments on the Bayesian approach to uncertainty.

Representation of uncertainty

We shall use the Bayesian approach to statistics throughout this course, although you might not notice this until later lectures. Implicit in everything we do we shall assume

Representation of uncertainty

We shall use the Bayesian approach to statistics throughout this course, although you might not notice this until later lectures. Implicit in everything we do we shall assume

- Uncertainty can be measured using probability.
  - Under minimal rationality assumptions, probability can be shown to be the only rational way to represent uncertainty.

Representation of uncertainty

We shall use the Bayesian approach to statistics throughout this course, although you might not notice this until later lectures. Implicit in everything we do we shall assume

- Uncertainty can be measured using probability.
  - 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!

Representation of uncertainty

We shall use the Bayesian approach to statistics throughout this course, although you might not notice this until later lectures. Implicit in everything we do we shall assume

- Uncertainty can be measured using probability.
  - 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!
- All uncertainty quantities θ can be given distributions $\pi(\theta)$ that represent our (an experts?) uncertainty about their value - this doesn't mean that they are random quantities, just that we don't know their value.
  - Even unknown function will be described by probability distributions across a class of unknown functions

Representation of uncertainty

We shall use the Bayesian approach to statistics throughout this course, although you might not notice this until later lectures. Implicit in everything we do we shall assume

- Uncertainty can be measured using probability.
  - 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!
- All uncertainty quantities θ can be given distributions $\pi(\theta)$ that represent our (an experts?) uncertainty about their value - this doesn't mean that they are random quantities, just that we don't know their value.
  - Even unknown function will be described by probability distributions across a class of unknown functions
- We shall use the principle of conditionality, and always (where possible) condition on our data.

Bayesian Inference

The basics of Bayesian inference are very simple.

- If quantity θ is unknown we describe our uncertainty by prior distribution $\pi(\theta)$.
- Suppose we have data D from model $\pi(D|\theta)$
- Then conditional on observing this data, our posterior distribution is

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

Bayesian Inference

The basics of Bayesian inference are very simple.

- If quantity θ is unknown we describe our uncertainty by prior distribution $\pi(\theta)$.
- Suppose we have data D from model $\pi(D|\theta)$
- Then conditional on observing this data, our posterior distribution is

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

In general, we just use the relationship

$$\text{posterior} \propto \text{prior} \times \text{posterior (model)}$$

Bayesian Inference

The basics of Bayesian inference are very simple.

- If quantity θ is unknown we describe our uncertainty by prior distribution $\pi(\theta)$.
- Suppose we have data D from model $\pi(D|\theta)$
- Then conditional on observing this data, our posterior distribution is

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

In general, we just use the relationship

$$\text{posterior} \propto \text{prior} \times \text{posterior (model)}$$

Given a model and prior the entire Bayesian statistical approach is fully specified. In practice, computational difficulties add interest!

Parametric uncertainty

Suppose we have computer model $\eta(\theta)$ which depends on unknown parameters θ.
We can describe any beliefs about θ using a distribution $\pi(\theta)$ - for example,

- we may have previous experimental evidence to suggest θ takes a particular value (within some confidence interval) and decided that $\theta \sim N(\hat{\theta}, \sigma^2)$
- we may know hard physical bounds on the value of θ but not know where it lies within these bounds and thus decide a uniform distribution is the best description of our uncertainty $\theta \sim U[a, b]$
- we may have no knowledge at all and wish to use a flat improper prior $\pi(\theta) \propto 1$ - although in practice we are rarely in this situation.
- ...

Elicitation is the process of extracting distributions for unknown quantities from experts. Garthwaite *et al.* provides a good introduction to elicitation.

Representation of uncertainty II

Returning to our list of challenges for computer experiments

- Calibration

Representation of uncertainty II

Returning to our list of challenges for computer experiments

- Calibration
  - The inverse problem is the raison d'être of the Bayesian approach. Given observations of the physical system $\mathcal{D}$ we can find the posterior $\pi(\theta|\mathcal{D})$ distribution for the unknown parameters.

Representation of uncertainty II

Returning to our list of challenges for computer experiments

- Calibration
  - The inverse problem is the raison d'être of the Bayesian approach. Given observations of the physical system $\mathcal{D}$ we can find the posterior $\pi(\theta|\mathcal{D})$ distribution for the unknown parameters.
  - The posterior contains all the information about the parameter - the science from the model, empirical information from the data, and expert opinion from the data.

Representation of uncertainty II

Returning to our list of challenges for computer experiments

- Calibration
  - The inverse problem is the raison d'être of the Bayesian approach. Given observations of the physical system $\mathcal{D}$ we can find the posterior $\pi(\theta|\mathcal{D})$ distribution for the unknown parameters.
  - The posterior contains all the information about the parameter - the science from the model, empirical information from the data, and expert opinion from the data.
  - This requires careful specification of how the simulator relates to the physical system (see lecture 4).

Representation of uncertainty II

Returning to our list of challenges for computer experiments

- Calibration
  - The inverse problem is the raison d'être of the Bayesian approach. Given observations of the physical system $\mathcal{D}$ we can find the posterior $\pi(\theta|\mathcal{D})$ distribution for the unknown parameters.
  - The posterior contains all the information about the parameter - the science from the model, empirical information from the data, and expert opinion from the data.
  - This requires careful specification of how the simulator relates to the physical system (see lecture 4).
- Uncertainty Analysis

Representation of uncertainty II

Returning to our list of challenges for computer experiments

- Calibration
  - The inverse problem is the raison d'être of the Bayesian approach. Given observations of the physical system $\mathcal{D}$ we can find the posterior $\pi(\theta|\mathcal{D})$ distribution for the unknown parameters.
  - The posterior contains all the information about the parameter - the science from the model, empirical information from the data, and expert opinion from the data.
  - This requires careful specification of how the simulator relates to the physical system (see lecture 4).
- Uncertainty Analysis
  - Given $\theta \sim \pi(\theta)$ what is $\pi(\eta(\theta))$
  - i.e., what is the distribution of the simulator output given a distribution for the input.

Representation of uncertainty II

- Prediction
  - Superficially similar to uncertainty estimation, in that we want to know the distribution of the simulator prediction given a distribution for θ.

Representation of uncertainty II

- Prediction
  - Superficially similar to uncertainty estimation, in that we want to know the distribution of the simulator prediction given a distribution for θ.
  - If we have data, we will want to look at the distribution of Y given $\theta \sim \pi(\theta|\mathcal{D})$, i.e., we want to use the calibrated distribution for θ.

Representation of uncertainty II

- Prediction
  - Superficially similar to uncertainty estimation, in that we want to know the distribution of the simulator prediction given a distribution for θ.
  - If we have data, we will want to look at the distribution of Y given $\theta \sim \pi(\theta|\mathcal{D})$, i.e., we want to use the calibrated distribution for θ.
  - However to predict a physical system from a simulator, we need to relate the simulator to the system by including a specification of the model error. More on this in lecture 4.

Representation of uncertainty II

- Prediction
  - Superficially similar to uncertainty estimation, in that we want to know the distribution of the simulator prediction given a distribution for θ.
  - If we have data, we will want to look at the distribution of Y given $\theta \sim \pi(\theta|\mathcal{D})$, i.e., we want to use the calibrated distribution for θ.
  - However to predict a physical system from a simulator, we need to relate the simulator to the system by including a specification of the model error. More on this in lecture 4.
  - Sometimes we can just use the average prediction $\mathbb{E}(Y|\mathcal{D}) = \int \eta(\theta)\pi(\theta|\mathcal{D})\mathrm{d}\theta$ or perhaps a plug-in estimate $\eta(\hat{\theta})$ where $\hat{\theta} = \arg\max \pi(\theta|\mathcal{D})$.

Code uncertainty

These tasks can be computed using a brute force approach (Monte Carlo) as long as sufficient computational resources are available.
For example, to perform UA

- draw a sample of parameter values from the prior $\theta_1, \ldots, \theta_N \sim \pi(\theta)$,
- Look at $\eta(\theta_1), \ldots, \eta(\theta_N)$ to find the distribution $\pi(\eta(\theta))$.

Code uncertainty

These tasks can be computed using a brute force approach (Monte Carlo) as long as sufficient computational resources are available.
For example, to perform UA

- draw a sample of parameter values from the prior $\theta_1, \ldots, \theta_N \sim \pi(\theta)$,
- Look at $\eta(\theta_1), \ldots, \eta(\theta_N)$ to find the distribution $\pi(\eta(\theta))$.

However, for complex simulators, run times might be long.
Consequently, we will only know the simulator output at a finite number of points.

We call this *code uncertainty*.

Code uncertainty

Code uncertainty

For slow simulators, we are uncertain about the simulator value at all points except those in a finite set.

Code uncertainty

For slow simulators, we are uncertain about the simulator value at all points except those in a finite set.

- All inference must be done using a finite ensemble of model runs

$$\mathcal{D}_{sim} = \{(\theta_i, \eta(\theta_i))\}_{i=1,\ldots,N}$$

Code uncertainty

For slow simulators, we are uncertain about the simulator value at all points except those in a finite set.

- All inference must be done using a finite ensemble of model runs

$$\mathcal{D}_{sim} = \{(\theta_i, \eta(\theta_i))\}_{i=1,\ldots,N}$$

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

Code uncertainty

For slow simulators, we are uncertain about the simulator value at all points except those in a finite set.

- All inference must be done using a finite ensemble of model runs

$$\mathcal{D}_{sim} = \{(\theta_i, \eta(\theta_i))\}_{i=1,\ldots,N}$$

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

If θ is multidimensional, then even short run times can rule out brute force approaches

- $\dim(\theta) \in \mathbb{R}^{10}$ then 1000 simulator runs is only enough for one point in each corner of the design space.

Code uncertainty

For slow simulators, we are uncertain about the simulator value at all points except those in a finite set.

- All inference must be done using a finite ensemble of model runs

$$\mathcal{D}_{sim} = \{(\theta_i, \eta(\theta_i))\}_{i=1,\ldots,N}$$

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

If θ is multidimensional, then even short run times can rule out brute force approaches

- $\dim(\theta) \in \mathbb{R}^{10}$ then 1000 simulator runs is only enough for one point in each corner of the design space.

The design of computational experiments is an active field in statistics (see lecture 3).

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.

We use the emulator as a cheap approximation to the simulator.

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

There are many ways to build emulators, but we will focus on the most popular type, the Gaussian process emulator.

Summary

Computer experiments are increasingly being used to learn about the world and to make decisions.

The statistical analysis of computer experiments is vitally important in many cases, yet the theory is only just being developed.

This is an exciting and fast moving research area with lots of opportunities for methodological development and collaborative research with other research communities.

References

- MUCM toolkit, available at `www.mucm.sheffield.ac.uk`
- O'Hagan A 2004 Bayesian Analysis of Computer Code Outputs: A Tutorial.
- Sacks J, Welch WJ, Mitchell TJ and Wynn HP 1989 Design and analysis of computer experiments. *Statistical Science* **4**, 409–423.
- Santner TJ, Williams BJ and Notz W 2003 *The Design and Analysis of Computer Experiments*, Springer.
- Garthwaite PH, Kadane JB and O'Hagan A 2005 Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association* **100**, 680–701.