Introduction

Monte Carlo Ensembles
Emulators: Designed Ensembles

Inference In Ensemble Experiments

Jonathan Rougier

Department of Mathematics, University of Bristol, UK
Currently visiting SAMSII

ISDS, Duke University
23 March, 2007
Ensemble experiments

- Ensemble experiments examine the behaviour of a function $g(x)$ over a variety of inputs; typically $x$ represents
  - Initial conditions (weather forecasting), or
  - Parameterisation of physical processes (climate prediction), or
  - Forcing (palæoclimate prediction).
Ensemble experiments

- Ensemble experiments examine the behaviour of a function $g(x)$ over a variety of inputs; typically $x$ represents
  - Initial conditions (weather forecasting), or
  - Parameterisation of physical processes (climate prediction), or
  - Forcing (palaeoclimate prediction).

- One application of ensemble experiments is **Uncertainty Analysis**: What is the distribution of $g(x^*)$ for some specified distribution function $F_{x^*}$?

- A simple interpretation of $x^*$ is as the 'best' input, in a framework relating the model $f$ and the system $y$:

\[
y = g(x^*) + \epsilon \quad \text{where } \epsilon \perp \perp \{g, x^*\}.
\]

This framework is rather limiting: the **Reified modelling** approach is one generalisation (Goldstein and Rougier, 2004, 2007).
If $g(\cdot)$ represents a model of a complex system it can be extremely expensive to evaluate.

Our illustration will use HadSM3: a dynamic atmosphere coupled with a simple (slab) ocean, a cryosphere module, and a land-surface module. After three years we have an ensemble of about 300 evaluations of this model. These evaluations form the basis of the next UK Climate Impacts Programme, UKCIP08 (http://www.ukcip.org.uk/).

We varied 31 physical parameters, but, unless the model is much simpler than we think, our ensemble is not large enough adequately to span this input space.

**Code uncertainty** is the contribution to our uncertainty that we could reduce by doing more evaluations.
The distribution function of sensitivity

We will focus on a single model output, climate sensitivity. This is the equilibrium response of mean global temperature to a doubling of atmospheric CO$_2$ levels from pre-industrial levels.

- If we write sensitivity as $y^* \triangleq g(x^*)$, our uncertainty is fully described by the distribution function

$$F_{y^*}(v) \triangleq \Pr[y^* \leq v] = E_{F_{x^*}}[\mathbb{1}(g(x^*) \leq v)] .$$
The distribution function of sensitivity

We will focus on a single model output, climate sensitivity. This is the equilibrium response of mean global temperature to a doubling of atmospheric CO₂ levels from pre-industrial levels.

- If we write sensitivity as $y^* \triangleq g(x^*)$, our uncertainty is fully described by the distribution function

$$F_{y^*}(v) \triangleq \Pr[y^* \leq v] = \mathbb{E}_{F_{x^*}}[\mathbb{I}(g(x^*) \leq v)].$$

- The simplest way to approximate $F_{y^*}(v)$ is to use Monte Carlo (MC) integration

$$F_{y^*}^n(v) \triangleq n^{-1} \sum_{i=1}^{n} \mathbb{I}(y_i \leq v)$$

where $y_i \triangleq g(x^{(i)})$ and $x^{(i)} \overset{iid}{\sim} F_{x^*}$. 

Jonathan Rougier  
Inference In Ensemble Experiments
Together, \((Y; X)\) constitute our ensemble of model evaluations. Note, however, that for the inference about \(y^*\) only \(Y\) is used: generating the \(x^{(i)}\) is simply a step in the process of sampling from \(F_{y^*}\). We refer to this as a Monte Carlo ensemble.
Together, \((Y; X)\) constitute our ensemble of model evaluations. Note, however, that for the inference about \(y^*\) only \(Y\) is used: generating the \(x^{(i)}\) is simply a step in the process of sampling from \(F_{y^*}\). We refer to this as a Monte Carlo ensemble.

We can construct an estimate of the entire distribution function for \(y^*\) from one sample of size \(n\). Usually this would be plotted as a step-function showing the proportions \((0), 1/n, 2/n, \ldots, 1\) against \(y(1), \ldots, y(n)\), where \(y(i)\) is the \(i\)th order statistic of \(Y\).

For not-large \(n\), sampling effects will tend to shift this empirical distribution function around, inducing uncertainty for quantiles such as the 90th percentile.

One approach to quantifying this uncertainty is to invert the KS test.
The empirical distribution function, uniform margins

Sample size of 30

Sensitivity, degrees K

Cumulative probability

Jonathan Rougier

Inference In Ensemble Experiments
The empirical distribution function, uniform margins

Sample size of 30

Cumulative probability

Sensitivity, degrees K

Jonathan Rougier
Inference In Ensemble Experiments
The empirical distribution function, uniform margins

Sample size of 30

Cumulative probability

Sensitivity, degrees K

Jonathan Rougier

Inference In Ensemble Experiments
The empirical distribution function, uniform margins

Sample size of 30

Cumulative probability

Sensitivity, degrees K

Jonathan Rougier

Inference In Ensemble Experiments
The empirical distribution function, uniform margins

Sample size of 30

Sample size of 90

Sample size of 180

Sample size of 300
The uncertainty in MC integration goes down as $1/\sqrt{n}$. For the KS test, for example, the uncertainty is $\pm 1.36/\sqrt{n}$ when $n$ is greater than about 40. This has lead people to claim that the accuracy of MC integration does not depend on $d$, the number of components of $x$.

This confuses ‘vertical’ and ‘horizontal’ uncertainty. The vertical uncertainty is invariant to $d$, but the horizontal uncertainty, e.g. uncertainty about the 90th percentile, depends on both vertical uncertainty ($n$), and the slope of $F_y^*$. The slope of $F_y^*$ depends on the number of active inputs. If we introduce a new input that is important in determining $g(x)$, then uncertainty in this input will tend to flatten $F_y^*$ and widen our uncertainty about, e.g., the 90th percentile.
Importance sampling

Typically, it is not easy to choose the distribution $F_{x^*}$. If we want to examine the effect of perturbations to this distribution, we can use Importance Sampling (IS).

Suppose we wanted to sample from $F'_{x^*}$, but our MC ensemble came from $F_{x^*}$. We have (skipping over some technicalities)

$$F'_{y^*}(v) = E_{F'_{x^*}}[\mathbb{I}(g(x^*) \leq v)] = E_{F_{x^*}}[\mathbb{I}(g(x^*) \leq v) w(x)]$$

where $w(x) \triangleq f'_{x^*}(x)/f_{x^*}(x)$. 
Importance sampling

Typically, it is not easy to choose the distribution $F_{x^*}$. If we want to examine the effect of perturbations to this distribution, we can use Importance Sampling (IS).

- Suppose we wanted to sample from $F'_{x^*}$, but our MC ensemble came from $F_{x^*}$. We have (skipping over some technicalities)

$$ F'_{y^*}(v) = E_{F'_{x^*}}[\mathbb{I}(g(x^*) \leq v)] = E_{F_{x^*}}[\mathbb{I}(g(x^*) \leq v) w(x)] $$

where $w(x) \triangleq f'_{x^*}(x)/f_{x^*}(x)$.

- Our estimator of the distribution function becomes

$$ F'_{y^*}(v) \approx n^{-1} \sum_{i=1}^{n} \mathbb{I}(y_i \leq v) w_i $$

where $w_i \triangleq w(x^{(i)})$. 
Importance sampling (cont)

- After normalising the weights, our estimate of $F_{y^*}$ is plotted as a step-function showing the proportions $(0), w_1, w_1 + w_2, \ldots$ against $y_1, y_2, \ldots$.
- Borrowing a term from Sequential Importance Sampling (SIS), we can compute the Effective Sample Size (ESS)

$$\text{ESS} \triangleq \left\{ \sum_{i=1}^{n} (w_i)^2 \right\}^{-1}.$$  

As a first approximation, we might use this in place of $n$ to derive vertical bands.
After normalising the weights, our estimate of $F_{y^*}$ is plotted as a step-function showing the proportions
$(0), w(1), w(1) + w(2), \ldots$ against $y(1), y(2), \ldots$.

Borrowing a term from Sequential Importance Sampling (SIS), we can compute the **Effective Sample Size** (ESS)

$$
\text{ESS} \triangleq \left\{ \sum_{i=1}^{n} (w_i)^2 \right\}^{-1}
$$

As a first approximation, we might use this in place of $n$ to derive vertical bands.

Note that there is a dimensional effect with IS, because a small change to each margin in $x^*$ means a huge change to the joint distribution if $d$ is large, and usually a very skewed distribution of weights (i.e., small ESS).
Importance sampling, some triangular margins

Sample size of 180 (ESS = 38)

Sample size of 300 (ESS = 69)
MC Ensembles: Summary

Good things about MC ensembles:

1. Simple to understand and implement;
2. Sequential (unlike, e.g., Gaussian quadrature);
3. Relatively easy to compute a measure of code uncertainty.
MC Ensembles: Summary

Good things about MC ensembles:

1. Simple to understand and implement;
2. Sequential (unlike, e.g., Gaussian quadrature);
3. Relatively easy to compute a measure of code uncertainty.

Bad things about MC ensembles:

1. Hard to examine alternative choices for $F_{x^*}$;
2. Expensive: Can’t we do better than $\propto 1/\sqrt{n}$?
3. Reckless to make ‘random’ choices for $X$ with an expensive model.
An emulator is a stochastic representation of $g(\cdot)$, constructed by conditioning prior judgements about $g(\cdot)$ on the outcome of a carefully-chosen ensemble of evaluations, $(Y; X)$:

$$F_{g(x)}(v) \triangleq \Pr[g(x) \leq v \mid Y; X].$$
An emulator is a stochastic representation of $g(\cdot)$, constructed by conditioning prior judgements about $g(\cdot)$ on the outcome of a carefully-chosen ensemble of evaluations, $(Y; X)$:

$$F_{g(x)}(v) \triangleq \Pr[g(x) \leq v \mid Y; X].$$

While we aspire to compute $F_{y^*}(v) \triangleq \Pr[y^* \leq v \mid g(\cdot)]$, what we can actually compute is

$$\hat{F}_{y^*}(v) \triangleq \Pr[y^* \leq v \mid Y; X] = E_{F_{x^*}}[F_{g(x^*)}(v)].$$

Compared with MC, the indicator function $\mathbb{I}(g(x^*) \leq v)$ has been replaced by the distribution function $F_{g(x^*)}(v)$.

Then: Exact answer by running the model;
Now: probabilistic answer by consulting the emulator.
If we stick with MC integration, our new approximation is

$$
\hat{F}_{y^*}(v) \triangleq m^{-1} \sum_{j=1}^{m} F_{g(x^{(j)})}(v) \quad \text{where } x^{(j)} \overset{iid}{\sim} F_{x^*}.
$$

But as we are evaluating the emulator (cheap) not the model (expensive) we can typically take $m$ large.

$\Rightarrow$ We can try a variety of choices for $F_{x^*}$. 
If we stick with MC integration, our new approximation is

\[ \hat{F}_{y^*}(v) \triangleq m^{-1} \sum_{j=1}^{m} F_{g(x(j))}(v) \text{ where } x^{(j)} \overset{iid}{\sim} F_{x^*}. \]

But as we are evaluating the emulator (cheap) not the model (expensive) we can typically take \( m \) large.

\[ \Rightarrow \] We can try a variety of choices for \( F_{x^*} \).

Now we can contrast the sources of code uncertainty in the two cases:

- **MC** Approximate calculation of exact value of \( \Pr[y^* \leq v] \);
- **Emulator** Exact calculation of the approximation \( \Pr[y^* \leq v \mid Y; X] \).

Which approach wins (where both are tenable) will depend on the predictability of \( g(\cdot) \), and our ability to come up with an informative ensemble, \((Y; X)\).
An example with quite different choices for $F_{x^*}$.
**Summary: MC Ensembles vs Designed Ensembles**

- **MC ensembles are safe.** It’s hard to go wrong, and hurtful criticism is avoided.
- **Designed ensembles are powerful** and flexible, but
  
  *With Power Comes Responsibility.*

Choices must be made:
- When designing the ensemble;
- When constructing the emulator.

- **Ensembles of opportunity.** In climate studies we may *have* to apply emulator-based methods, because most available ensembles are pieced-together, so that no specific $F_{x^*}$ is discernible.
This talk is based on the paper:


The reified modelling approach:


The ‘Best Input’ approach:


The Bayes Linear approach to emulation and ensemble design:
