Introduction to Emulators

Jonathan Rougier

Maths Department, Durham University

J.C.Rougier@durham.ac.uk

22 June, 2006

A simple problem

Let's consider a simple problem of *uncertainty analysis*:

- We have a complex deterministic function

$$\theta \rightarrow g(\theta) \in \mathcal{G}$$

  with uncertainty about the 'true' parameter θ^* described by a distribution function

$$F_{\theta^*}(\theta) \triangleq \mathsf{Pr}(\theta^* \leq \theta)$$

- *Our objective is to determine* $Pr\big(g(\theta^*) \in \mathcal{Q}\big)$ *where* $\mathcal{Q} \subset \mathcal{G}$.

A simple problem

Let's consider a simple problem of *uncertainty analysis*:

- We have a complex deterministic function

$$\theta \rightarrow g(\theta) \in \mathcal{G}$$

  with uncertainty about the 'true' parameter θ^* described by a distribution function

$$F_{\theta^*}(\theta) \triangleq \mathsf{Pr}(\theta^* \leq \theta)$$

- *Our objective is to determine* $\mathsf{Pr}\big(g(\theta^*) \in \mathcal{Q}\big)$ *where* $\mathcal{Q} \subset \mathcal{G}$.

- Topical example: $g(\cdot)$ is a climate model with parameters θ; $g(\theta^*)$ is actual climate; F_{θ^*} is rectangular; $\mathcal{Q}$ is the region in which global mean temperature in 2100 is at least 2°C higher.

A simple solution

- The simplest way to estimate $q \triangleq \Pr\big(g(\theta^*) \in \mathcal{Q}\big)$ is by Monte Carlo integration:

$$q = \int \mathbb{I}\big(g(\theta) \in \mathcal{Q}\big)\, dF_{\theta^*}(\theta)$$

$$\approx n^{-1} \sum_{i=1}^{n} \mathbb{I}\big(g(\theta_i) \in \mathcal{Q}\big) \quad \text{where } \theta_i \overset{\text{iid}}{\sim} F_{\theta^*},$$

where $\mathbb{I}(\cdot)$ is the indicator function. Call this approximation $\hat{q}^{(n)}$.

A simple solution

- The simplest way to estimate $q \triangleq \Pr\big(g(\theta^*) \in \mathcal{Q}\big)$ is by Monte Carlo integration:

$$q = \int \mathbb{I}\big(g(\theta) \in \mathcal{Q}\big)\, dF_{\theta^*}(\theta)$$

$$\approx n^{-1} \sum_{i=1}^{n} \mathbb{I}\big(g(\theta_i) \in \mathcal{Q}\big) \quad \text{where } \theta_i \overset{\text{iid}}{\sim} F_{\theta^*},$$

  where $\mathbb{I}(\cdot)$ is the indicator function. Call this approximation $\hat{q}^{(n)}$.

- By the CLT and Binomial distribution, we have

$$\hat{q}^{(n)} \overset{\text{asy}}{\sim} \text{Gaussian}\big(q, q(1-q)/n\big)$$

- If we wanted an accuracy of ± 5 percentage points (± 2 sd), then we'd need $n \approx 400$, conservatively.

Some issues . . .

Generally-speaking, it seems exceedingly reckless to select the evaluations randomly if $g(\cdot)$ is anything other than a very cheap function.

More particularly,

1. We may not be able to afford 400 evaluations, or may require more accuracy, or may want to proceed sequentially, to get greater accuracy;

Some issues . . .

Generally-speaking, it seems exceedingly reckless to select the evaluations randomly if $g(\cdot)$ is anything other than a very cheap function.

More particularly,

1. We may not be able to afford 400 evaluations, or may require more accuracy, or may want to proceed sequentially, to get greater accuracy;
2. We may want to include some evaluations from another experiment where the parameters were not $\overset{\text{iid}}{\sim} F_{\theta^*}$, or may want to try several choices for F_{θ^*};

Some issues . . .

Generally-speaking, it seems exceedingly reckless to select the evaluations randomly if $g(\cdot)$ is anything other than a very cheap function.

More particularly,

1. We may not be able to afford 400 evaluations, or may require more accuracy, or may want to proceed sequentially, to get greater accuracy;
2. We may want to include some evaluations from another experiment where the parameters were not $\overset{\text{iid}}{\sim} F_{\theta^*}$, or may want to try several choices for F_{θ^*};
3. We may want to learn about $g(\cdot)$ at the same time, e.g., to check our implementation.

In these cases an emulator provides an alternative approach.

Introducing the emulator

- An emulator is a distribution function for $g(\theta)$, informed by evaluations of $g(\cdot)$ at—hopefully!—carefully-chosen values $\theta_1, \ldots, \theta_n$.
- Within the Bayesian approach we would compute

$$F_{g(\theta)}(v) \triangleq \Pr\big(g(\theta) \leq v \mid \theta, G; T\big)$$

where $T \triangleq \big(\theta_1, \ldots, \theta_n\big)$ and $G \triangleq \big(g(\theta_1), \ldots, g(\theta_n)\big)$; here $(G; T)$ comprises the *ensemble* of evaluations.

- For $\theta_i = T_i \in T$ we have

$$F_{g(T_i)}(v) = \mathbb{I}\big(G_i \leq v\big) \in \{0, 1\},$$

but generally $g(\theta)$ is an uncertain vector quantity with a non-zero variance matrix.

Calculations with the emulator

- These proceed slightly differently:

$$
\begin{aligned}
q &= \int \Pr(g(\theta) \in \mathcal{Q} \mid \theta, G; T)\, dF_{\theta^*}(\theta) \\
&= \int \left\{ \int_{\mathcal{Q}} f_{g(\theta)}(v)\, dv \right\} dF_{\theta^*}(\theta) \\
&\approx m^{-1} \sum_{j=1}^{m} \left\{ \int_{\mathcal{Q}} f_{g(\theta_j)}(v)\, dv \right\} \quad \text{where } \theta_j \overset{\text{iid}}{\sim} F_{\theta^*};
\end{aligned}
$$

call this approximation $\hat{q}^{(m)}$.

Calculations with the emulator

- These proceed slightly differently:

$$
\begin{aligned}
q &= \int \Pr(g(\theta) \in \mathcal{Q} \mid \theta, G; T)\, dF_{\theta^*}(\theta) \\
&= \int \left\{ \int_{\mathcal{Q}} f_{g(\theta)}(v)\, dv \right\} dF_{\theta^*}(\theta) \\
&\approx m^{-1} \sum_{j=1}^{m} \left\{ \int_{\mathcal{Q}} f_{g(\theta_j)}(v)\, dv \right\} \quad \text{where } \theta_j \overset{\text{iid}}{\sim} F_{\theta^*};
\end{aligned}
$$

  call this approximation $\hat{q}^{(m)}$.

- Note that we may be able to take $m \gg n$, if $\int_{\mathcal{Q}} f_{g(\theta)}(v)\, dv$ is much quicker to evaluate than $g(\theta)$. In this case we can take $\hat{q}^{(m)}$ and q to be the same, but of course *the usefulness of q depends on the reliability of the emulator as a representation of our uncertainty about $g(\cdot)$.*

Advantages of a reliable emulator

1. We can often compute q exactly, as already mentioned.
2. We are not restricted to sampling the parameter values in our ensemble of evaluations independently from a specified distribution, so we can
   (a) Use evaluations from other experiments (and our evaluations contribute to the common pool),
   (b) Select the evaluations non-randomly in order to improve accuracy,
   (c) Explore the behaviour of $g(\cdot)$, e.g., to check our implementation and develop our understanding,
   (d) Try lots of different choices for F_{θ^*}.
3. We can control the *human resources* used in the analysis, through the detail with which we construct and analyse the emulator.

So where's the catch?

Emulators are complicated objects, and will usually need to be constructed by two specialists: a scientist and a statistician, working together. The Monte Carlo approach, on the other hand, can be implemented directly by the scientist alone.

So where's the catch?

Emulators are complicated objects, and will usually need to be constructed by two specialists: a scientist and a statistician, working together. The Monte Carlo approach, on the other hand, can be implemented directly by the scientist alone.

Difficulties:

1. 'Gold-standard' emulators based on gaussian process priors cost $\mathrm{O}(n^3)$ to build;
2. If $\dim \theta$ is large then most of the parameter space of $g(\cdot)$ will be an *extrapolation* from the convex hull of T, which means
   (a) Low robustness to statistical modelling choices, *but at the same time . . .*
   (b) Hard to diagnose mis-specification.
3. Multivariate emulators are highly-constrained by tractable parametric forms.

Example of using emulators for design

Experimental Design is choosing sets of parameters at which to evaluate the model. In the simplest case it is selecting T_{n+1} on the basis of the ensemble $(G; T)$. We can use our emulator to 'tune' our selection to be highly informative, using *pseudo data*.

- For a succession of cadidate values θ', θ'', ...
  1. Sample $G' \sim F_{g(\theta')}$; these are the 'pseudo-data'
  2. Add (G', θ') to the current n-point ensemble and build a new emulator
  3. Score in terms of the resulting prediction, e.g. reduction in predictive variance
  4. Do this with lots of sampled G's, to estimate the expected score for θ'

The best θ so-chosen is necessary informative both in terms of the behaviour of $g(\cdot)$ and the inference about $\mathcal{Q}$.

When to use an emulator?

1. *When the decisions that follow are really important*, so that sufficient resources are available to hire a statistician, and more resources could be made available if warranted;

When to use an emulator?

1. *When the decisions that follow are really important*, so that sufficient resources are available to hire a statistician, and more resources could be made available if warranted;
2. When the function $g(\cdot)$ is expensive to evaluate, e.g., relative to the accuracy required for decisions;
3. When scientists are well-informed about the behaviour of $g(\cdot)$ and/or poorly-informed about the 'true' value θ^*;
4. When insights into the behaviour of $g(\cdot)$ will help scientists to improve $g(\theta^*)$ as a representation of the underlying system;

When to use an emulator?

1. *When the decisions that follow are really important*, so that sufficient resources are available to hire a statistician, and more resources could be made available if warranted;
2. When the function $g(\cdot)$ is expensive to evaluate, e.g., relative to the accuracy required for decisions;
3. When scientists are well-informed about the behaviour of $g(\cdot)$ and/or poorly-informed about the 'true' value θ^*;
4. When insights into the behaviour of $g(\cdot)$ will help scientists to improve $g(\theta^*)$ as a representation of the underlying system;
5. (not covered) *Reified analysis:* when we want to combine information from more than one representation of the underlying system.

Brief references

For a more general approach to linking models and systems,

- J.C. Rougier (2006), Probabilistic Inference for Future Climate Using an Ensemble of Climate Model Evaluations, *Climatic Change*, in press.
- M. Goldstein and J.C. Rougier (2004), Probabilistic Formulations for Transferring Inferences from Mathematical Models to Physical Systems, *SIAM Journal on Scientific Computing*, **26(2)**, 467-487.
- M. Goldstein and J.C. Rougier (2006), Reified Bayesian Modelling and Inference for Physical Systems, *Journal of Statistical Planning and Inference*, forthcoming as a discussion paper.

For the current 'gold-standard' in gaussian process emulators,

- M.C. Kennedy and A O'Hagan (2001), Bayesian Calibration of Computer Models, *Journal of the Royal Statistical Society, Series B*, **63**, 425-464. With discussion.

For our alternative Bayes linear treatment,

- P.S. Craig, M. Goldstein, A.H. Seheult and J.A. Smith, (1997), Pressure Matching for Hydrocarbon Reservoirs: A Case Study in the Use of Bayes Linear Strategies for Large Computer Experiments, C. Gatsonis et al (eds), *Case Studies in Bayesian Statistics III*, New York: Springer-Verlag. With discussion.
- P.S. Craig, M. Goldstein, J.C. Rougier and A.H. Seheult (2001), Bayesian Forecasting for Complex Systems Using Computer Simulators, *Journal of the American Statistical Association*, **96**, 717-729.
- M. Goldstein and J.C. Rougier (2006), Bayes Linear Calibrated Prediction for Complex Systems, *Journal of the American Statistical Association*, forthcoming.

My work-in-progress,

- "Lightweight emulators for complex multivariate functions." A step away from the gold-standard towards something a bit more 'quick and dirty' that builds in $\mathrm{O}(n)$ not $\mathrm{O}(n^3)$.
- "Emulating the sensitivity of the HadAM3 climate model using ensembles from different but related experiments". Building an emulator for a *very* complicated scalar function. Joint work with David Sexton, James Murphy, and Dave Stainforth.