The Error in ABC

Richard Wilkinson

Department of Probability and Statistics
University of Sheffield

Paris - June 2009

Managing Uncertainty in Complex Models (MUCM)

Four year project across 6 universities: Sheffield, Durham, LSE, Southampton, Aston, Bristol.

- PIs: Tony O'Hagan, Peter Challenor, Jonty Rougier, Henry Wynn, Dan Cornford, Jeremy Oakley, Michael Goldstein.
- 8 RAs, 5 PhD students.

Aim: to develop some of the statistical technology required when analysing computer experiments.

- Focused on expensive deterministic models
- Based around the use of *emulators*
  - cheap statistical surrogates (meta-models) of the *simulator*
- Aim to account for all sources of uncertainty in model predictions. Including uncertainty in
  - Initial conditions
  - Model parameters
  - Imperfect/incomplete science
  - Approximate solutions to model equations
  - Code uncertainty

Calibration

For forwards models we specify parameters θ and i.c.s and the model generates output X. We are interested in the inverse-problem, i.e., observe data $\mathcal{D}$, want to estimate parameter values.
As Bayesians, we are used to thinking of this as $\pi(\theta|\mathcal{D}, \mathcal{M})$.

Calibration

For forwards models we specify parameters θ and i.c.s and the model generates output X. We are interested in the inverse-problem, i.e., observe data $\mathcal{D}$, want to estimate parameter values.
As Bayesians, we are used to thinking of this as $\pi(\theta|\mathcal{D},\mathcal{M})$.
What does this represent? Or rather, what do we believe we are doing?

- Does θ have a physical interpretation, i.e., are we estimating physical parameters?
- Or is θ interpretted statistically? i.e., θ is the value that best explains the data given the model - cf. the coefficients in a linear regression.

e.g., ocean physics models must be run with viscosity several orders of magnitude too large.

Calibration

For forwards models we specify parameters θ and i.c.s and the model generates output X. We are interested in the inverse-problem, i.e., observe data $\mathcal{D}$, want to estimate parameter values.
As Bayesians, we are used to thinking of this as $\pi(\theta|\mathcal{D}, \mathcal{M})$.
What does this represent? Or rather, what do we believe we are doing?

- Does θ have a physical interpretation, i.e., are we estimating physical parameters?
- Or is θ interpretted statistically? i.e., θ is the value that best explains the data given the model - cf. the coefficients in a linear regression.

e.g., ocean physics models must be run with viscosity several orders of magnitude too large.
When can we interpret the value found for θ as a physical value?

Calibration

For forwards models we specify parameters θ and i.c.s and the model generates output X. We are interested in the inverse-problem, i.e., observe data $\mathcal{D}$, want to estimate parameter values.
As Bayesians, we are used to thinking of this as $\pi(\theta|\mathcal{D}, \mathcal{M})$.
What does this represent? Or rather, what do we believe we are doing?

- Does θ have a physical interpretation, i.e., are we estimating physical parameters?
- Or is θ interpretted statistically? i.e., θ is the value that best explains the data given the model - cf. the coefficients in a linear regression.

e.g., ocean physics models must be run with viscosity several orders of magnitude too large.
When can we interpret the value found for θ as a physical value?

- If the model is a perfect representation of the system

Calibration

For forwards models we specify parameters θ and i.c.s and the model generates output X. We are interested in the inverse-problem, i.e., observe data $\mathcal{D}$, want to estimate parameter values.
As Bayesians, we are used to thinking of this as $\pi(\theta|\mathcal{D}, \mathcal{M})$.
What does this represent? Or rather, what do we believe we are doing?

- Does θ have a physical interpretation, i.e., are we estimating physical parameters?
- Or is θ interpretted statistically? i.e., θ is the value that best explains the data given the model - cf. the coefficients in a linear regression.

e.g., ocean physics models must be run with viscosity several orders of magnitude too large.
When can we interpret the value found for θ as a physical value?

- If the model is a perfect representation of the system
- When the model is imperfect, but we have a description (that we believe) of the discrepancy between model and system.

Bayesian Calibration Framework I

Kennedy and O'Hagan 2001, RSS B

Bayesian Calibration Framework I

Kennedy and O'Hagan 2001, RSS B

- Suppose we have a computer model $\eta(t,\theta)$ that we wish to use to make predictions of a physical system $\zeta(t)$ using observations $D(t)$.
  - θ are model parameters we wish to learn
  - t are control/index parameters, e.g., time, location etc.

Bayesian Calibration Framework I

Kennedy and O'Hagan 2001, RSS B

- Suppose we have a computer model $\eta(t, \theta)$ that we wish to use to make predictions of a physical system $\zeta(t)$ using observations $D(t)$.
  - θ are model parameters we wish to learn
  - t are control/index parameters, e.g., time, location etc.
- Standard approach is the *best-input* approach, where we assume there is a single 'best' value of θ, which we call $\hat{\theta}$. The model run at $\hat{\theta}$, the hat-run $\eta(\hat{\theta})$, is the best model prediction.

Bayesian Calibration Framework I

Kennedy and O'Hagan 2001, RSS B

- Suppose we have a computer model $\eta(t, \theta)$ that we wish to use to make predictions of a physical system $\zeta(t)$ using observations $D(t)$.
  - θ are model parameters we wish to learn
  - t are control/index parameters, e.g., time, location etc.
- Standard approach is the *best-input* approach, where we assume there is a single 'best' value of θ, which we call $\hat{\theta}$. The model run at $\hat{\theta}$, the hat-run $\eta(\hat{\theta})$, is the best model prediction.
- The standard assumption that

$$D(t) = \eta(t, \hat{\theta}) + e_t$$

  where $e_.$ is a white noise error process is a poor assumption for most models. If the model is imperfect, then residuals $D - \eta(\theta)$ may be correlated, even if the measurement error process is white.

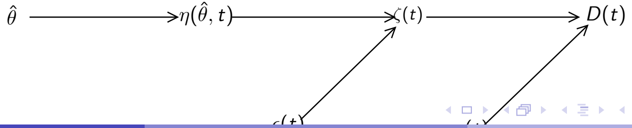
Bayesian Calibration Framework II

Kennedy and O'Hagan 2001, RSS B

- Instead, assume that we observe reality plus measurement error.

$$D(t) = \zeta(t) + e(t)$$

Often, $e(\cdot)$ will be a white noise process with known mean and variance.



Bayesian Calibration Framework II

Kennedy and O'Hagan 2001, RSS B

- Instead, assume that we observe reality plus measurement error.

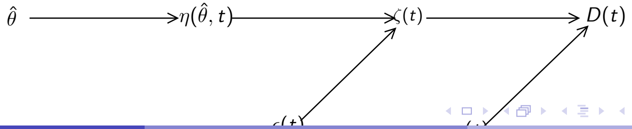
$$D(t) = \zeta(t) + e(t)$$

  Often, $e(\cdot)$ will be a white noise process with known mean and variance.

- Introduce a model error (discrepancy) term. Assume that reality is the best model prediction plus an error

$$\zeta(t) = \eta(t, \hat{\theta}) + \epsilon(t).$$

  Note ϵ does not depend on θ.



Bayesian Calibration Framework II

Kennedy and O'Hagan 2001, RSS B

- Instead, assume that we observe reality plus measurement error.

$$D(t) = \zeta(t) + e(t)$$

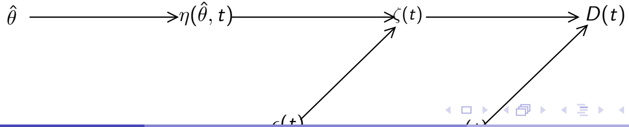
  Often, $e(\cdot)$ will be a white noise process with known mean and variance.

- Introduce a model error (discrepancy) term. Assume that reality is the best model prediction plus an error

$$\zeta(t) = \eta(t, \hat{\theta}) + \epsilon(t).$$

  Note ϵ does not depend on θ.

- Argue that $\eta(\cdot, \hat{\theta})$ and $\epsilon(\cdot)$ are independent. Kennedy and O'Hagan use Gaussian processes to model both the model η and the error ϵ. Allows a rich structure to be learnt for $\epsilon(\cdot)$.



Rejection based ABC

Approximate Rejection Algorithm

- Draw θ from $\pi(\theta)$
- Simulate $X \sim \eta(\theta)$
- Accept θ if $\rho(\mathcal{D}, X) \leq \delta$

- What is the approximation?
  - We wish to solve $\mathcal{D} = \eta(\theta)$.
  - Accepted θ are not from $\pi(\theta|\mathcal{D}, \eta)$, but from some approximation to it.
- How do we choose
  - distance measure $\rho(\cdot, \cdot)$
  - tolerance δ
  - summary statistic $S(\cdot)$, etc?

The error in ABC

The error in ABC

Approximate Rejection Algorithm

- Draw θ from $\pi(\theta)$
- Simulate $X \sim \eta(\theta)$
- Accept θ if $\rho(\mathcal{D}, X) \leq \delta$

It is possible to show that output from this algorithm is an exact draw from the posterior when we assume that the measurement is made in the presence of a uniform additive error term.

$$D = \eta(\theta) + \epsilon$$

The error in ABC

Approximate Rejection Algorithm

- Draw θ from $\pi(\theta)$
- Simulate $X \sim \eta(\theta)$
- Accept θ if $\rho(\mathcal{D}, X) \leq \delta$

It is possible to show that output from this algorithm is an exact draw from the posterior when we assume that the measurement is made in the presence of a uniform additive error term.

$$D = \eta(\theta) + \epsilon$$

If $\rho(x, y) = |x - y|$, then this is equivalent to assuming uniform error on $[-\delta, \delta]$. Accepted θ are from the posterior

$$\pi(\theta | D, \eta, \epsilon \sim U[-\delta, \delta])$$

The error in ABC

Approximate Rejection Algorithm

- Draw θ from $\pi(\theta)$
- Simulate $X \sim \eta(\theta)$
- Accept θ if $\rho(\mathcal{D}, X) \leq \delta$

It is possible to show that output from this algorithm is an exact draw from the posterior when we assume that the measurement is made in the presence of a uniform additive error term.

$$D = \eta(\theta) + \epsilon$$

If $\rho(x, y) = |x - y|$, then this is equivalent to assuming uniform error on $[-\delta, \delta]$. Accepted θ are from the posterior

$$\pi(\theta | D, \eta, \epsilon \sim U[-\delta, \delta])$$

ABC gives 'exact' inference under a different model!

A general error structure

Suppose ϵ is distributed with density $\pi_\epsilon(\cdot)$. We can modify the ABC rejection algorithm to give perform inference from the model $D = \eta(\theta) + \epsilon$ where we now control the distribution of the error.

A general error structure

Suppose ϵ is distributed with density $\pi_\epsilon(\cdot)$. We can modify the ABC rejection algorithm to give perform inference from the model $D = \eta(\theta) + \epsilon$ where we now control the distribution of the error.

Generalized ABC

- Draw $\theta \sim \pi(\theta)$
- Simulate X from model $X \sim \eta(\theta)$
- Accept θ with probability $r = \frac{\pi_\epsilon(D-X)}{c}$

Here, c is a constant chosen to maximise the acceptance probability, and guarantee $r \leq 1$. Typically, $c = \pi_e(0)$ is the best we can do.

A general error structure

Suppose ϵ is distributed with density $\pi_\epsilon(\cdot)$. We can modify the ABC rejection algorithm to give perform inference from the model $D = \eta(\theta) + \epsilon$ where we now control the distribution of the error.

Generalized ABC

- Draw $\theta \sim \pi(\theta)$
- Simulate X from model $X \sim \eta(\theta)$
- Accept θ with probability $r = \frac{\pi_\epsilon(D-X)}{c}$

Here, c is a constant chosen to maximise the acceptance probability, and guarantee $r \leq 1$. Typically, $c = \pi_e(0)$ is the best we can do.

Proposition

Accepted θ are samples from the posterior distribution $\pi(\theta | D, \epsilon \sim \pi_\epsilon)$ where $D = \eta(\theta) + \epsilon$.

A general error structure

Suppose ϵ is distributed with density $\pi_\epsilon(\cdot)$. We can modify the ABC rejection algorithm to give perform inference from the model $D = \eta(\theta) + \epsilon$ where we now control the distribution of the error.

Generalized ABC

- Draw $\theta \sim \pi(\theta)$
- Simulate X from model $X \sim \eta(\theta)$
- Accept θ with probability $r = \frac{\pi_\epsilon(D-X)}{c}$

Here, c is a constant chosen to maximise the acceptance probability, and guarantee $r \leq 1$. Typically, $c = \pi_e(0)$ is the best we can do.

Proposition

Accepted θ are samples from the posterior distribution $\pi(\theta | D, \epsilon \sim \pi_\epsilon)$ where $D = \eta(\theta) + \epsilon$.

This imples that using a 0-1 cutoff corresponds to assuming a uniformly distributed error term.

Proof

Let

$$I = \begin{cases} 1 \text{ if } \theta \text{ is accepted} \\ 0 \text{ otherwise.} \end{cases}$$

Proof

Let
$$I = \begin{cases} 1 \text{ if } \theta \text{ is accepted} \\ 0 \text{ otherwise.} \end{cases}$$

Then,
$$\mathbb{P}(I = 1|\theta) = \int \mathbb{P}(I = 1|\eta(\theta) = x, \theta)\pi(x|\theta)\mathrm{d}x$$
$$= \int \frac{\pi_\epsilon(D - x)}{c}\pi(x|\theta)\mathrm{d}x.$$

Proof

Let

$$I = \begin{cases} 1 \text{ if } \theta \text{ is accepted} \\ 0 \text{ otherwise.} \end{cases}$$

Then,

$$\begin{aligned} \mathbb{P}(I = 1|\theta) &= \int \mathbb{P}(I = 1|\eta(\theta) = x, \theta)\pi(x|\theta)\mathrm{d}x \\ &= \int \frac{\pi_\epsilon(D - x)}{c}\pi(x|\theta)\mathrm{d}x. \end{aligned}$$

So the distribution of accepted θ is

$$\pi(\theta|I = 1) = \frac{\pi(\theta)\int \pi_\epsilon(D - x)\pi(x|\theta)\mathrm{d}x}{\int \pi(\theta)\int \pi_\epsilon(D - x)\pi(x|\theta)\mathrm{d}x\mathrm{d}\theta}.$$

Proof

Let

$$I = \begin{cases} 1 \text{ if } \theta \text{ is accepted} \\ 0 \text{ otherwise.} \end{cases}$$

Then,

$$\mathbb{P}(I = 1|\theta) = \int \mathbb{P}(I = 1|\eta(\theta) = x, \theta)\pi(x|\theta)\mathrm{d}x$$
$$= \int \frac{\pi_\epsilon(D - x)}{c}\pi(x|\theta)\mathrm{d}x.$$

So the distribution of accepted θ is

$$\pi(\theta|I = 1) = \frac{\pi(\theta)\int \pi_\epsilon(D - x)\pi(x|\theta)\mathrm{d}x}{\int \pi(\theta)\int \pi_\epsilon(D - x)\pi(x|\theta)\mathrm{d}x\mathrm{d}\theta}.$$

Conversely, assuming $D = \eta(\theta) + \epsilon$, calculate the posterior directly:

$$\pi(D|\theta) = \int \pi(D|\eta(\theta) = x, \theta)\pi(x|\theta)\mathrm{d}x = \int \pi_\epsilon(D - x)\pi(x|\theta)\mathrm{d}x.$$

Consequently,

$$\pi(\theta|D) = \frac{\pi(\theta)\int \pi_\epsilon(D - x)\pi(x|\theta)\mathrm{d}x}{\int \pi(\theta)\int \pi_\epsilon(D - x)\pi(x|\theta)\mathrm{d}x\mathrm{d}\theta}.$$

Choosing discrepancies

How can we choose a distribution for ϵ?

Choosing discrepancies

How can we choose a distribution for ϵ?

- Let ϵ be measurement error on D - unlikely to be large sufficient. NB this may be built into models already and can be removed and dealt with analytically.

Choosing discrepancies

How can we choose a distribution for ϵ?

- Let ϵ be measurement error on D - unlikely to be large sufficient. NB this may be built into models already and can be removed and dealt with analytically.
- Let ϵ be the discrepancy between the model and reality
  - In a deterministic model setting, Goldstein and Rougier 2008 (amongst others), have offered advice about thinking about discrepancies.
  - In a stochastic model setting, what the model error is is much less clear. (Rougier 2008 gives a Bayes Linear approach in a simple model)

Choosing discrepancies

How can we choose a distribution for ϵ?

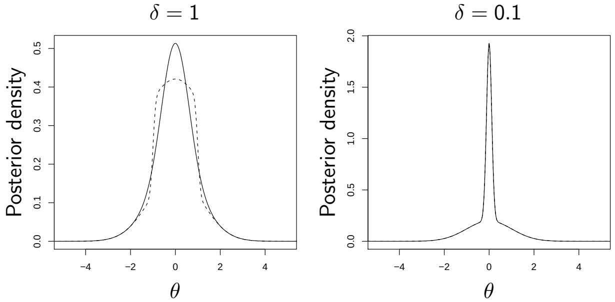
- Let ϵ be measurement error on D - unlikely to be large sufficient. NB this may be built into models already and can be removed and dealt with analytically.
- Let ϵ be the discrepancy between the model and reality
  - In a deterministic model setting, Goldstein and Rougier 2008 (amongst others), have offered advice about thinking about discrepancies.
  - In a stochastic model setting, what the model error is is much less clear. (Rougier 2008 gives a Bayes Linear approach in a simple model)

NB We may need to compromise on our beliefs about the error structure in order to achieve an acceptable acceptance rate in the inference.

Mixture of Normals

Sisson *et al.* 2007, Beaumont *et al.* 2008

$$\eta(\theta) \sim \frac{1}{2}\mathcal{N}(\theta, 1) + \frac{1}{2}\mathcal{N}(\theta, \frac{1}{100}), \qquad \theta \sim \mathcal{U}[-10, 10], \qquad D = 0$$



The posterior distributions found when using ABC with uniform error $\epsilon \sim U[-\delta, \delta]$ (solid line) and ABC with a Gaussian acceptance kernel $\epsilon \sim N(0, \delta^2/3)$ (dashed line).

Generalized ABC-MCMC

Build an exact MCMC scheme for the discrepancy model.

ABC-MCMC I

Suppose we are currently at θ.

1. Propose θ' from density $q(\theta, \theta')$.
2. Simulate X from $\eta(\theta')$.
3. Accept move with probability

$$r(\theta, \theta') = \frac{\pi_e(D - X')}{c} \min\left(1, \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}\right).$$

Else stay at θ'.

Generalizes ABC-MCMC II

Or an alternative version is to augment the sample space.

ABC-MCMC II

1. At time t, propose a move from $\psi_t = (\theta_t, X_t)$ to $\psi' = (\theta', X')$ with θ' drawn from transition kernel $q(\theta_t, \theta')$, and X' simulated from the model using θ':

$$X' \sim \eta(\theta')$$

2. Set $\psi_{t+1} = (\theta', X')$ with probability

$$r((\theta_t, X_t), (\theta', X')) = \min\left(1, \frac{\pi_\epsilon(D - X')q(\theta', \theta_t)\pi(\theta')}{\pi_\epsilon(D - X_t)q(\theta_t, \theta')\pi(\theta_t)}\right), \qquad (1)$$

otherwise set $\psi_{t+1} = \psi_t$.

Future work

- Model error
  - When should it be included
  - How to model and think about it
  - Can we learn the error?
    - ★ Dynamic model setting, sequential observations, learn the discrepancy through time.
    - ★ Prior and posterior specification of the error (cf. Ratmann *et al.*). Eg Gaussian processes, t-processes?

Future work

- Model error
  - When should it be included
  - How to model and think about it
  - Can we learn the error?
    - ★ Dynamic model setting, sequential observations, learn the discrepancy through time.
    - ★ Prior and posterior specification of the error (cf. Ratmann *et al.*). Eg Gaussian processes, t-processes?
- When can models be rewritten to take account of known structure.

Future work

- Model error
  - When should it be included
  - How to model and think about it
  - Can we learn the error?
    - ★ Dynamic model setting, sequential observations, learn the discrepancy through time.
    - ★ Prior and posterior specification of the error (cf. Ratmann *et al.*). Eg Gaussian processes, t-processes?
- When can models be rewritten to take account of known structure.
- Generalize ABC-SMC methods.

Future work

- Model error
  - When should it be included
  - How to model and think about it
  - Can we learn the error?
    - Dynamic model setting, sequential observations, learn the discrepancy through time.
    - Prior and posterior specification of the error (cf. Ratmann *et al.*). Eg Gaussian processes, t-processes?
- When can models be rewritten to take account of known structure.
- Generalize ABC-SMC methods.
- ABC as an conservative method $\mathbb{V}\text{ar}(\theta|D) \leq \mathbb{V}\text{ar}(\theta|D, \epsilon \sim \pi_\epsilon)$

Future work

- Model error
  - When should it be included
  - How to model and think about it
  - Can we learn the error?
    - Dynamic model setting, sequential observations, learn the discrepancy through time.
    - Prior and posterior specification of the error (cf. Ratmann *et al.*). Eg Gaussian processes, t-processes?
- When can models be rewritten to take account of known structure.
- Generalize ABC-SMC methods.
- ABC as an conservative method $\mathbb{V}\text{ar}(\theta|D) \leq \mathbb{V}\text{ar}(\theta|D, \epsilon \sim \pi_\epsilon)$
- Measure of the distance between the desired distribution and the approximation $\text{TVD}(\pi(\theta|D), \pi(\theta|D, \epsilon \sim \pi_\epsilon))$

Future work

- Model error
  - When should it be included
  - How to model and think about it
  - Can we learn the error?
    - Dynamic model setting, sequential observations, learn the discrepancy through time.
    - Prior and posterior specification of the error (cf. Ratmann *et al.*). Eg Gaussian processes, t-processes?
- When can models be rewritten to take account of known structure.
- Generalize ABC-SMC methods.
- ABC as an conservative method $\mathbb{V}\text{ar}(\theta|D) \leq \mathbb{V}\text{ar}(\theta|D, \epsilon \sim \pi_\epsilon)$
- Measure of the distance between the desired distribution and the approximation $\text{TVD}(\pi(\theta|D), \pi(\theta|D, \epsilon \sim \pi_\epsilon))$
- Effect of summary statistics.
  - We/the modellers believe that certain summaries will be more accurate than others.

Conclusions

Approximate Bayesian Computation gives exact inference for the wrong model!

- To move beyond inference conditioned on the truth of model, we must account for model error.
- ABC algorithms can be considered to include an additive noise term.
- For a given metric and tolerance, we can interpret the result.
- We can generalise ABC algorithms to move beyond the use of uniform error structures to account for errors closer to our beliefs.

Conclusions

Approximate Bayesian Computation gives exact inference for the wrong model!

- To move beyond inference conditioned on the truth of model, we must account for model error.
- ABC algorithms can be considered to include an additive noise term.
- For a given metric and tolerance, we can interpret the result.
- We can generalise ABC algorithms to move beyond the use of uniform error structures to account for errors closer to our beliefs.

Thank you for listening!