

# Uncertainty quantification: learning from flawed mechanistic models

Richard Wilkinson

University of Sheffield

# What is Uncertainty Quantification (UQ)

Uncertainty Quantification (UQ)  $\equiv$  statistics with complex models

- determining statistical information about the uncertainty in an output of interest that depends upon the complex model
- A 'complex model' is one that is expensive to evaluate.

# What is Uncertainty Quantification (UQ)

Uncertainty Quantification (UQ)  $\equiv$  statistics with complex models

- determining statistical information about the uncertainty in an output of interest that depends upon the complex model
- A 'complex model' is one that is expensive to evaluate.

Typical tasks

- **Uncertainty propagation**

$$\theta \sim \pi(\cdot), \text{ and } Y = f(\theta), \quad \text{then } Y \sim ?$$

# What is Uncertainty Quantification (UQ)

Uncertainty Quantification (UQ)  $\equiv$  statistics with complex models

- determining statistical information about the uncertainty in an output of interest that depends upon the complex model
- A 'complex model' is one that is expensive to evaluate.

Typical tasks

- **Uncertainty propagation**

$$\theta \sim \pi(\cdot), \text{ and } Y = f(\theta), \quad \text{then } Y \sim ?$$

- **Parameter estimation**

$$\text{Observe } Y = f(\theta) + \epsilon, \text{ then } \theta = ?$$

# What is Uncertainty Quantification (UQ)

Uncertainty Quantification (UQ)  $\equiv$  statistics with complex models

- determining statistical information about the uncertainty in an output of interest that depends upon the complex model
- A 'complex model' is one that is expensive to evaluate.

Typical tasks

- **Uncertainty propagation**

$$\theta \sim \pi(\cdot), \text{ and } Y = f(\theta), \quad \text{then } Y \sim ?$$

- **Parameter estimation**

Observe  $Y = f(\theta) + \epsilon$ , then  $\theta = ?$

- **Sensitivity analysis**:  $\theta = (\theta_1, \dots, \theta_p)^\top \sim \pi(\cdot)$ . If we can measure one component of  $\theta$ , which should we choose to minimize  $\text{Var}(Y)$ ?
- **Design**: what data should we collect to learn  $\theta$ , improve  $f$ , etc?
- **Decision making**: my model is uncertain, the parameters are uncertain, the data is noisy, but I need to make a decision...

# What is Uncertainty Quantification (UQ)

Uncertainty Quantification (UQ)  $\equiv$  statistics with complex models

- determining statistical information about the uncertainty in an output of interest that depends upon the complex model
- A 'complex model' is one that is expensive to evaluate.

Typical tasks

- **Uncertainty propagation**

$$\theta \sim \pi(\cdot), \text{ and } Y = f(\theta), \quad \text{then } Y \sim ?$$

- **Parameter estimation**

$$\text{Observe } Y = f(\theta) + \epsilon, \text{ then } \theta = ?$$

- **Sensitivity analysis**:  $\theta = (\theta_1, \dots, \theta_p)^\top \sim \pi(\cdot)$ . If we can measure one component of  $\theta$ , which should we choose to minimize  $\text{Var}(Y)$ ?
- **Design**: what data should we collect to learn  $\theta$ , improve  $f$ , etc?
- **Decision making**: my model is uncertain, the parameters are uncertain, the data is noisy, but I need to make a decision...

UQ should be a synergy between statistics, applied mathematics and domain sciences

## Mechanistic models

Models describe hypothesised relationships between variables.

### **Mechanistic model**

- explains how/why the variables interact the way they do.
- parameters may have physically meaning

e.g. ODE/PDE models

⋮

### **Phenomenological/statistical/empirical/machine learning models**

- the model seeks only to best describe the relationships
- models don't explain why the variables interact the way they do

e.g. Regression models

# Mechanistic models

Models describe hypothesised relationships between variables.

## **Mechanistic model**

- explains how/why the variables interact the way they do.
- parameters may have physically meaning

e.g. ODE/PDE models

⋮

## **Phenomenological/statistical/empirical/machine learning models**

- the model seeks only to best describe the relationships
- models don't explain why the variables interact the way they do

e.g. Regression models

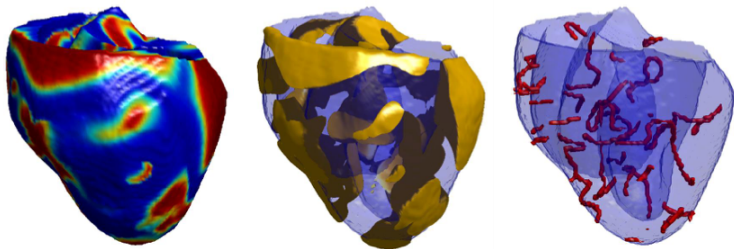
Typical problems with mechanistic models

- computationally expensive
- difficult to work with, e.g., black boxes
- often inaccurate - model misspecification
- often deterministic



# Why do we need UQ?

## Atrial fibrillation



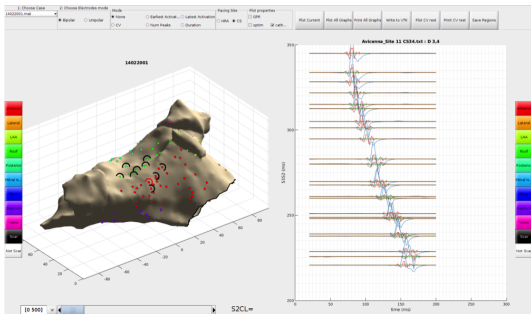
Atrial fibrillation (AF) - rapid and uncoordinated electrical activation (arrhythmia) leading to poor mechanical function.

- Affects around 610,000 people in UK.
- Catheter ablation removes/isolates pathological tissue that sustain/initiate AF.
- 40% of patients subsequently experience atrial tachycardia (AT).

# UQ in Patient Specific Cardiac Models

With Sam Coveney, Richard Clayton, Steve Neiderer, Jeremy Oakley, ...

Aim: predict which AF patients will develop AT following ablation, and then treat for both in a single procedure.



Use complex electrophysiology simulation using monodomain eqn on shell anatomy.

Accurate predictions require patient specific models, but clinical data is sparse and noisy.

We need to

- Estimate conduction velocity on the atrium using ECG measurements
- Infer tissues properties, including regions of fibrotic material
- Predict AT pathways
- Aid clinical decision making (accounting for uncertainty)

# Inference under discrepancy

How should we do inference if the model is imperfect?

# Inference under discrepancy

How should we do inference if the model is imperfect?

Data generating process

$$y \sim G$$

# Inference under discrepancy

How should we do inference if the model is imperfect?

Data generating process

$$y \sim G$$

Model (complex simulator, finite dimensional parameter)

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\}$$

If  $G = F_{\theta_0} \in \mathcal{F}$  then we know what to do<sup>1</sup>.

---

<sup>1</sup>Even if we can't agree about it!

# Inference under discrepancy

How should we do inference if the model is imperfect?

Data generating process

$$y \sim G$$

Model (complex simulator, finite dimensional parameter)

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\}$$

If  $G = F_{\theta_0} \in \mathcal{F}$  then we know what to do<sup>1</sup>.

How should we proceed if

$$G \notin \mathcal{F}$$

---

<sup>1</sup>Even if we can't agree about it!

# Inference under discrepancy

How should we do inference if the model is imperfect?

Data generating process

$$y \sim G$$

Model (complex simulator, finite dimensional parameter)

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\}$$

If  $G = F_{\theta_0} \in \mathcal{F}$  then we know what to do<sup>1</sup>.

How should we proceed if

$$G \notin \mathcal{F}$$

Interest lies in inference of  $\theta$  not calibrated prediction.

---

<sup>1</sup>Even if we can't agree about it!

# An appealing idea

Kennedy and O'Hagan 2001

Most models are imperfect. . .



# An appealing idea

Kennedy and O'Hagan 2001

Most models are imperfect. . .

Can we expand the class of models by adding a Gaussian process (GP) to our simulator?

If  $f_{\theta}(x)$  is our simulator,  $y$  the observation, then perhaps we can correct  $f$  by modelling

$$y = f_{\theta^*}(x) + \delta(x) \quad \text{where} \quad \delta(\cdot) \sim GP$$

# An appealing idea

Kennedy and O'Hagan 2001

Most models are imperfect. . .

Can we expand the class of models by adding a Gaussian process (GP) to our simulator?

If  $f_{\theta}(x)$  is our simulator,  $y$  the observation, then perhaps we can correct  $f$  by modelling

$$y = f_{\theta^*}(x) + \delta(x) \quad \text{where} \quad \delta(\cdot) \sim GP$$

This greatly expands  $\mathcal{F}$  into a non-parametric world.

# An appealing, but flawed, idea

Kennedy and O'Hagan 2001, Brynjarsdottir and O'Hagan 2014

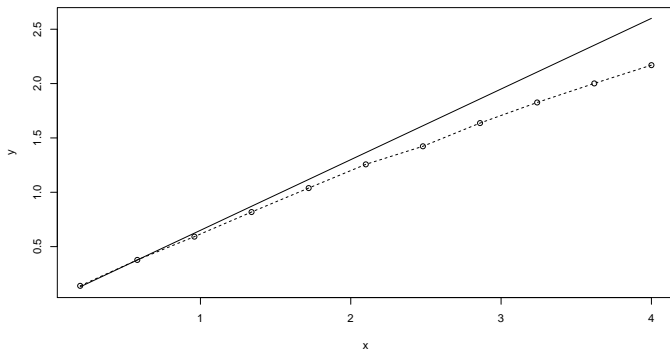
Simulator

$$f_{\theta}(x) = \theta x$$

Reality

$$g(x) = \frac{\theta x}{1 + \frac{x}{a}} \quad \theta = 0.65, a = 20$$

Solid=model with true theta, dashed=truth



## An appealing, but flawed, idea

Bolting on a GP can correct your predictions, but won't necessarily fix your inference,

## An appealing, but flawed, idea

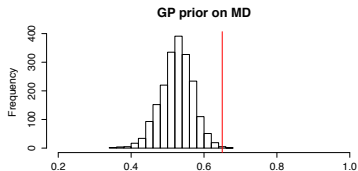
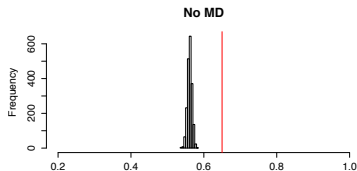
Bolting on a GP can correct your predictions, but won't necessarily fix your inference, e.g.

- No discrepancy:

$$y = f_{\theta}(x) + N(0, \sigma^2),$$
$$\theta \sim N(0, 100), \sigma^2 \sim \Gamma^{-1}(0.001, 0.001)$$

- GP discrepancy:

$$y = f_{\theta}(x) + \delta(x) + N(0, \sigma^2),$$
$$\delta(\cdot) \sim GP(\cdot, \cdot) \text{ with objective priors}$$



# Dangers of non-parametric model extensions

There are (at least) two problems with this approach:

- We may still find  $G \notin \mathcal{F}$
- Identifiability

# Dangers of non-parametric model extensions

There are (at least) two problems with this approach:

- We may still find  $G \notin \mathcal{F}$
  - Identifiability
    - ▶ A GP is an incredibly complex infinite dimensional model, which is not necessarily identified even asymptotically. The posterior can concentrate not on a point, but on some sub manifold of parameter space, and the projection of the prior on this space continues to impact the posterior even as more and more data are collected.
- ie We never forget the prior, but the prior is too complex to understand

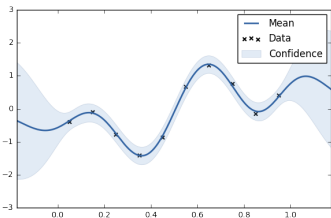
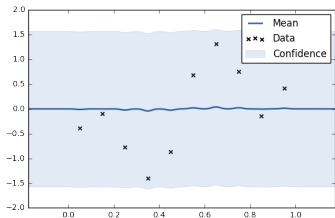
# Dangers of non-parametric model extensions

There are (at least) two problems with this approach:

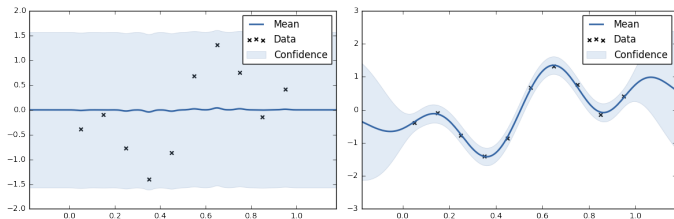
- We may still find  $G \notin \mathcal{F}$
- Identifiability
  - ▶ A GP is an incredibly complex infinite dimensional model, which is not necessarily identified even asymptotically. The posterior can concentrate not on a point, but on some sub manifold of parameter space, and the projection of the prior on this space continues to impact the posterior even as more and more data are collected.  
ie We never forget the prior, but the prior is too complex to understand
  - ▶ Brynjarsdottir and O'Hagan 2014 try to model their way out of trouble with prior information - which is great if you have it.



- We can also have problems finding the true optima for the hyperparameters, even in 1d problems:

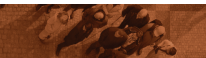


- We can also have problems finding the true optima for the hyperparameters, even in 1d problems:



- Wong et al 2017 impose identifiability (for  $\delta$  and  $\theta$ ) by giving up and identifying

$$\theta^* = \arg \min_{\theta} \int (\zeta(x) - f_{\theta}(x))^2 d\pi(x)$$



# Inferential approaches

- Maximum likelihood/minimum-distance
- Bayes(ish)
- History matching (HM)/ABC type methods (thresholding)

# Inferential approaches

- Maximum likelihood/minimum-distance
- Bayes(ish)
- History matching (HM)/ABC type methods (thresholding)

We'll consider how they behave for well-specified and mis-specified models.

Try to understand why (at least anecdotally) HM and ABC seem to work well in mis-specified cases.

# Inferential approaches

- Maximum likelihood/minimum-distance
- Bayes(ish)
- History matching (HM)/ABC type methods (thresholding)

We'll consider how they behave for well-specified and mis-specified models.

Try to understand why (at least anecdotally) HM and ABC seem to work well in mis-specified cases.

Big question<sup>2</sup> is what properties would we like our inferential approach to possess.

---

<sup>2</sup>To which I have no answer

# Maximum likelihood

Maximum likelihood estimator

$$\hat{\theta}_n = \arg \max_{\theta} l(y|\theta)$$

If  $G = F_{\theta_0} \in \mathcal{F}$ , then (under some conditions)

$$\hat{\theta}_n \rightarrow \theta_0 \text{ almost surely as } n \rightarrow \infty$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \mathcal{I}^{-1}(\theta_0))$$

Asymptotic consistency, efficiency, normality.

# Maximum likelihood

Maximum likelihood estimator

$$\hat{\theta}_n = \arg \max_{\theta} l(y|\theta)$$

If  $G = F_{\theta_0} \in \mathcal{F}$ , then (under some conditions)

$$\hat{\theta}_n \rightarrow \theta_0 \text{ almost surely as } n \rightarrow \infty$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \mathcal{I}^{-1}(\theta_0))$$

Asymptotic consistency, efficiency, normality.

If  $G \notin \mathcal{F}$

$$\hat{\theta}_n \rightarrow \theta^* = \arg \min_{\theta} D_{KL}(G, F_{\theta}) \text{ almost surely}$$

$$= \arg \min_{\theta} \int \log \frac{dG}{dF_{\theta}} dG$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, V^{-1})$$

# Bayes

Bayesian posterior

$$\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$$

If  $G = F_{\theta_0} \in \mathcal{F}$

$$\pi(\theta|y) \xrightarrow{d} N(\theta_0, n^{-1}\mathcal{I}^{-1}(\theta_0)) \text{ as } n \rightarrow \infty$$

Bernstein-von Mises theorem: we forget the prior, and get asymptotic concentration and normality.

This also requires (a long list of) identifiability conditions to hold.



# Bayes

Bayesian posterior

$$\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$$

If  $G = F_{\theta_0} \in \mathcal{F}$

$$\pi(\theta|y) \xrightarrow{d} N(\theta_0, n^{-1}\mathcal{I}^{-1}(\theta_0)) \text{ as } n \rightarrow \infty$$

Bernstein-von Mises theorem: we forget the prior, and get asymptotic concentration and normality.

This also requires (a long list of) identifiability conditions to hold.

If  $G \notin \mathcal{F}$ , we still get asymptotic concentration (and possibly normality) but to  $\theta^*$  (the pseudo-true value).

*“there is no obvious meaning for Bayesian analysis in this case”*

# Bayes

Bayesian posterior

$$\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$$

If  $G = F_{\theta_0} \in \mathcal{F}$

$$\pi(\theta|y) \xrightarrow{d} N(\theta_0, n^{-1}\mathcal{I}^{-1}(\theta_0)) \text{ as } n \rightarrow \infty$$

Bernstein-von Mises theorem: we forget the prior, and get asymptotic concentration and normality.

This also requires (a long list of) identifiability conditions to hold.

If  $G \notin \mathcal{F}$ , we still get asymptotic concentration (and possibly normality) but to  $\theta^*$  (the pseudo-true value).

*“there is no obvious meaning for Bayesian analysis in this case”*

Often with non-parametric models (eg GPs), we don't even get this convergence to the pseudo-true value due to lack of identifiability.

# ABC (Approximate Bayesian computation)

## Rejection Algorithm

- Draw  $\theta$  from prior  $\pi(\cdot)$
- Accept  $\theta$  with probability  $\propto \pi(y | \theta)$

Accepted  $\theta$  are independent draws from the posterior distribution,  $\pi(\theta | D)$ .

# ABC (Approximate Bayesian computation)

## Rejection Algorithm

- Draw  $\theta$  from prior  $\pi(\cdot)$
- Accept  $\theta$  with probability  $\propto \pi(y | \theta)$

Accepted  $\theta$  are independent draws from the posterior distribution,  $\pi(\theta | D)$ .

If the likelihood,  $\pi(D|\theta)$ , is unknown:

## 'Mechanical' Rejection Algorithm

- Draw  $\theta$  from  $\pi(\cdot)$
- Simulate  $y' \sim \pi(y|\theta)$  from the computer model
- Accept  $\theta$  if  $y = y'$ , i.e., if computer output equals observation

# Rejection ABC

If  $\mathbb{P}(D)$  is small (or  $D$  continuous), we will rarely accept any  $\theta$ . Instead, there is an approximate version:

## Uniform Rejection Algorithm

- Draw  $\theta$  from  $\pi(\theta)$
- Simulate  $y' \sim \pi(y|\theta)$
- Accept  $\theta$  if  $\rho(y, y') \leq \epsilon$

# Rejection ABC

If  $\mathbb{P}(D)$  is small (or  $D$  continuous), we will rarely accept any  $\theta$ . Instead, there is an approximate version:

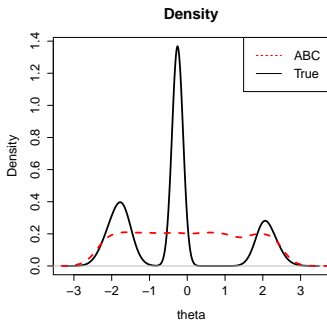
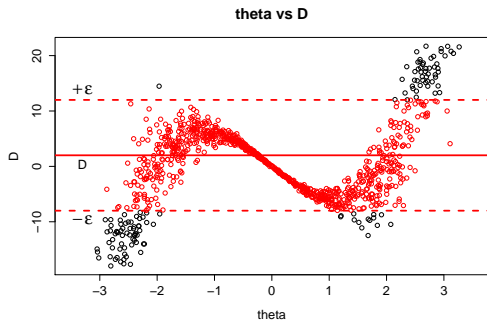
## Uniform Rejection Algorithm

- Draw  $\theta$  from  $\pi(\theta)$
- Simulate  $y' \sim \pi(y|\theta)$
- Accept  $\theta$  if  $\rho(y, y') \leq \epsilon$

$\epsilon$  reflects the tension between computability and accuracy.

- As  $\epsilon \rightarrow \infty$ , we get observations from the prior,  $\pi(\theta)$ .
- If  $\epsilon = 0$ , we generate observations from  $\pi(\theta | y)$ .

$$\epsilon = 10$$

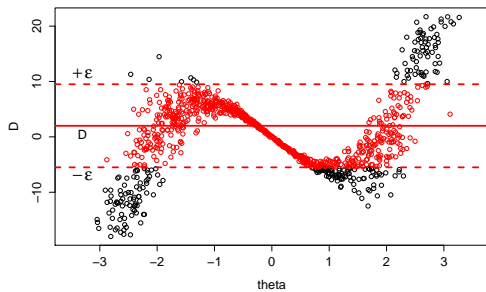


$$\theta \sim U[-10, 10], \quad y \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

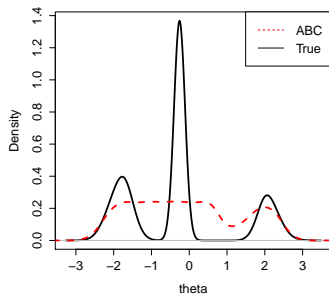
$$\rho(y, y') = |y - y'|, \quad y = 2$$

$$\epsilon = 7.5$$

theta vs D



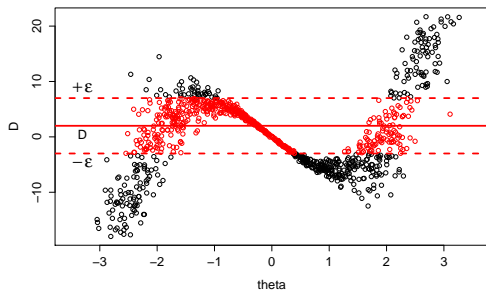
Density



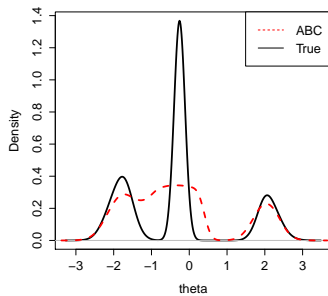


$$\epsilon = 5$$

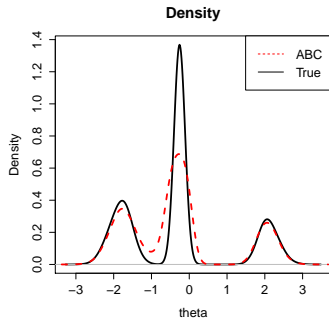
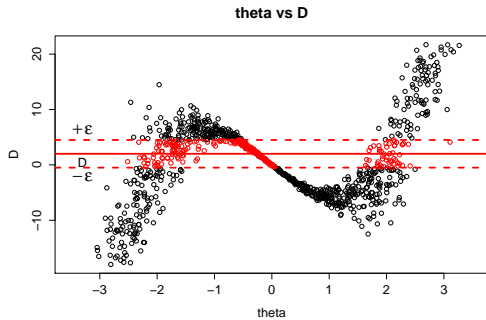
theta vs D



Density

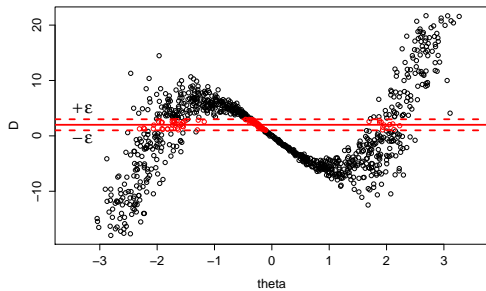


$$\epsilon = 2.5$$

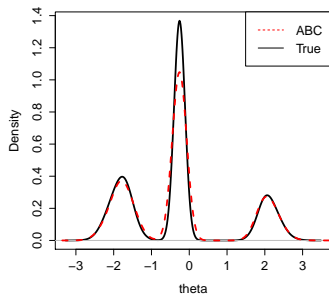


$$\epsilon = 1$$

theta vs D



Density



# History matching and ABC

History matching seeks to find a NROY set

$$\mathcal{P}_\theta = \{\theta : S_{HM}(\hat{F}_\theta, y) \leq 3\}$$

where

$$S_{HM}(F_\theta, y) = \frac{|\mathbb{E}_{F_\theta}(Y) - y|}{\sqrt{\text{Var}_{F_\theta}(Y)}}$$

## History matching and ABC

History matching seeks to find a NROY set

$$\mathcal{P}_\theta = \{\theta : S_{HM}(\hat{F}_\theta, y) \leq 3\}$$

where

$$S_{HM}(F_\theta, y) = \frac{|\mathbb{E}_{F_\theta}(Y) - y|}{\sqrt{\text{Var}_{F_\theta}(Y)}}$$

ABC approximates the posterior as

$$\pi_\epsilon(\theta) \propto \pi(\theta) \mathbb{E}(\mathbb{I}_{S(\hat{F}_\theta, y) \leq \epsilon})$$

for some choice of  $S$  and  $\epsilon$ , and where  $\hat{F}_\theta$  is estimated from the simulated  $y'$ .

For ABC, typically  $S(\hat{F}_\theta, y) = \rho(\eta(y), \eta(y'))$ , and  $\eta(\cdot)$  is a lower dimensional summary.

## History matching and ABC

History matching seeks to find a NROY set

$$\mathcal{P}_\theta = \{\theta : S_{HM}(\hat{F}_\theta, y) \leq 3\}$$

where

$$S_{HM}(F_\theta, y) = \frac{|\mathbb{E}_{F_\theta}(Y) - y|}{\sqrt{\text{Var}_{F_\theta}(Y)}}$$

ABC approximates the posterior as

$$\pi_\epsilon(\theta) \propto \pi(\theta) \mathbb{E}(\mathbb{I}_{S(\hat{F}_\theta, y) \leq \epsilon})$$

for some choice of  $S$  and  $\epsilon$ , and where  $\hat{F}_\theta$  is estimated from the simulated  $y'$ .

For ABC, typically  $S(\hat{F}_\theta, y) = \rho(\eta(y), \eta(y'))$ , and  $\eta(\cdot)$  is a lower dimensional summary.

They have thresholding of a score in common and are algorithmically comparable.

# History matching and ABC

These methods (anecdotally) seem to work better in mis-specified situations.

Why?

# History matching and ABC

These methods (anecdotally) seem to work better in mis-specified situations.

Why?

They differ from likelihood based approaches in that

- They only use some aspect of the simulator output
  - ▶ Typically we hand pick which simulator outputs to compare, and weight them on a case by case basis.



# History matching and ABC

These methods (anecdotally) seem to work better in mis-specified situations.

Why?

They differ from likelihood based approaches in that

- They only use some aspect of the simulator output
  - ▶ Typically we hand pick which simulator outputs to compare, and weight them on a case by case basis.
- Potentially use generalised scores/loss-functions

# History matching and ABC

These methods (anecdotally) seem to work better in mis-specified situations.

Why?

They differ from likelihood based approaches in that

- They only use some aspect of the simulator output
  - ▶ Typically we hand pick which simulator outputs to compare, and weight them on a case by case basis.
- Potentially use generalised scores/loss-functions
- The thresholding type nature potentially makes them somewhat conservative

# History matching and ABC

These methods (anecdotally) seem to work better in mis-specified situations.

Why?

They differ from likelihood based approaches in that

- They only use some aspect of the simulator output
  - ▶ Typically we hand pick which simulator outputs to compare, and weight them on a case by case basis.
- Potentially use generalised scores/loss-functions
- The thresholding type nature potentially makes them somewhat conservative
- They don't require a fully specified discrepancy model.

## What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

# What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?

# What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?
  - ▶ I don't want inconsistency.

# What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?
  - ▶ I don't want inconsistency.
- Asymptotic concentration or normality?

# What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?
  - ▶ I don't want inconsistency.
- ~~Asymptotic concentration or normality?~~



# What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?
  - ▶ I don't want inconsistency.
- ~~Asymptotic concentration or normality?~~
- Frequency properties?

# What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?
  - ▶ I don't want inconsistency.
- ~~Asymptotic concentration or normality?~~
- Frequency properties?
  - ▶ I wouldn't object but seems impossible for subjective priors.

# What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?
  - ▶ I don't want inconsistency.
- ~~Asymptotic concentration or normality?~~
- Frequency properties?
  - ▶ I wouldn't object but seems impossible for subjective priors.
- Coherence?

# What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?
  - ▶ I don't want inconsistency.
- ~~Asymptotic concentration or normality?~~
- Frequency properties?
  - ▶ I wouldn't object but seems impossible for subjective priors.
- Coherence?
- Robustness to small mis-specifications?

# What properties do we want?

Do any of these approaches have favourable properties/characteristics for inference under discrepancy? Particularly when the discrepancy model is crude?

- Consistency?
  - ▶ I don't want inconsistency.
- ~~Asymptotic concentration or normality?~~
- Frequency properties?
  - ▶ I wouldn't object but seems impossible for subjective priors.
- Coherence?
- Robustness to small mis-specifications?
- Ease of specification?

## Generalized scores

Likelihood based methods are notoriously sensitive to mis-specification.

## Generalized scores

Likelihood based methods are notoriously sensitive to mis-specification. Consider scoring rules instead. If we forecast  $F$ , observe  $y$ , then we receive score

$$S(F, y)$$

## Generalized scores

Likelihood based methods are notoriously sensitive to mis-specification. Consider scoring rules instead. If we forecast  $F$ , observe  $y$ , then we receive score

$$S(F, y)$$

$S$  is a proper score if

$$G = \arg \min_F \mathbb{E}_{Y \sim G} S(F, Y)$$

i.e. predicting  $G$  gives the best possible score.

- Encourages honest reporting



## Generalized scores

Likelihood based methods are notoriously sensitive to mis-specification. Consider scoring rules instead. If we forecast  $F$ , observe  $y$ , then we receive score

$$S(F, y)$$

$S$  is a proper score if

$$G = \arg \min_F \mathbb{E}_{Y \sim G} S(F, Y)$$

i.e. predicting  $G$  gives the best possible score.

- Encourages honest reporting

Examples:

- Log-likelihood  $S(F, y) = -\log f(y)$
- Tsallis-score  $(\gamma - 1) \int f(x)^\alpha dx - \gamma f(y)^{\alpha-1}$

Minimum scoring rule estimation (Dawid *et al.* 2014 etc) uses

$$\hat{\theta} = \arg \min_{\theta} S(F_{\theta}, y)$$

Minimum scoring rule estimation (Dawid *et al.* 2014 etc) uses

$$\hat{\theta} = \arg \min_{\theta} S(F_{\theta}, y)$$

For proper scores

$$\begin{aligned} \mathbb{E}_{\theta_0} \left( \left. \frac{\partial}{\partial \theta} S(F_{\theta}, y) \right|_{\theta=\theta_0} \right) &= \left. \frac{\partial}{\partial \theta} \mathbb{E}_{\theta_0} S(F_{\theta}, y) \right|_{\theta=\theta_0} \\ &= 0 \end{aligned}$$

so we have an unbiased estimating equation, and hence get asymptotic consistency for well-specified models. We also get asymptotic normality.

Dawid *et al.* 2014 show that if

- $\nabla_{\theta} f_{\theta}(x)$  is bounded in  $x$  for all  $\theta$
- Bregman gauge of scoring rule is locally bounded

then the minimum scoring rule estimator  $\hat{\theta}$  is B-robust

- i.e. it has bounded influence function

$$IF(x; \hat{\theta}, F_{\theta}) = \lim_{\epsilon \rightarrow 0} \frac{\hat{\theta}(\epsilon \delta_x + (1 - \epsilon)F_{\theta}) - \hat{\theta}(F_{\theta})}{\epsilon}$$

i.e. if  $F_{\theta}$  is infected by outlier at  $x$ , this doesn't unduly affect the inference.

Note both ABC and HM are B-robust in this sense, but using the log-likelihood is not.

Dawid *et al.* 2014 show that if

- $\nabla_{\theta} f_{\theta}(x)$  is bounded in  $x$  for all  $\theta$
- Bregman gauge of scoring rule is locally bounded

then the minimum scoring rule estimator  $\hat{\theta}$  is B-robust

- i.e. it has bounded influence function

$$IF(x; \hat{\theta}, F_{\theta}) = \lim_{\epsilon \rightarrow 0} \frac{\hat{\theta}(\epsilon \delta_x + (1 - \epsilon)F_{\theta}) - \hat{\theta}(F_{\theta})}{\epsilon}$$

i.e. if  $F_{\theta}$  is infected by outlier at  $x$ , this doesn't unduly affect the inference.

Note both ABC and HM are B-robust in this sense, but using the log-likelihood is not.

What type of robustness do we want here?

Dawid *et al.* 2014 show that if

- $\nabla_{\theta} f_{\theta}(x)$  is bounded in  $x$  for all  $\theta$
- Bregman gauge of scoring rule is locally bounded

then the minimum scoring rule estimator  $\hat{\theta}$  is B-robust

- i.e. it has bounded influence function

$$IF(x; \hat{\theta}, F_{\theta}) = \lim_{\epsilon \rightarrow 0} \frac{\hat{\theta}(\epsilon \delta_x + (1 - \epsilon)F_{\theta}) - \hat{\theta}(F_{\theta})}{\epsilon}$$

i.e. if  $F_{\theta}$  is infected by outlier at  $x$ , this doesn't unduly affect the inference.

Note both ABC and HM are B-robust in this sense, but using the log-likelihood is not.

What type of robustness do we want here?

## Bayes like approaches

What about Bayes like approaches with generalized scores?

# Bayes like approaches

What about Bayes like approaches with generalized scores?



*J. R. Statist. Soc. B* (2016)  
**78**, Part 5, pp. 1103–1130

## **A general framework for updating belief distributions**

Bissiri et al. 2016 consider updating prior beliefs when parameter  $\theta$  is connected to observations via a loss function  $L(\theta, y)$ .



# Bayes like approaches

What about Bayes like approaches with generalized scores?



*J. R. Statist. Soc. B* (2016)  
**78**, Part 5, pp. 1103–1130

## A general framework for updating belief distributions

Bissiri et al. 2016 consider updating prior beliefs when parameter  $\theta$  is connected to observations via a loss function  $L(\theta, y)$ .

They argue the update must be of the form

$$\pi(\theta|x) \propto \exp(-L(\theta, x))\pi(\theta)$$

via coherency arguments.

Note using log-likelihood as the loss function ( $L(\theta, x) = -\log f_{\theta}(x)$ ) recovers Bayes.

# Bayes like approaches

What about Bayes like approaches with generalized scores?



*J. R. Statist. Soc. B* (2016)  
**78**, Part 5, pp. 1103–1130

## A general framework for updating belief distributions

Bissiri et al. 2016 consider updating prior beliefs when parameter  $\theta$  is connected to observations via a loss function  $L(\theta, y)$ .

They argue the update must be of the form

$$\pi(\theta|x) \propto \exp(-L(\theta, x))\pi(\theta)$$

via coherency arguments.

Note using log-likelihood as the loss function ( $L(\theta, x) = -\log f_{\theta}(x)$ ) recovers Bayes.

## Advantages of this include

- Allows focus solely on the quantities of interest.
  - ▶ Full Bayesian inference requires us to model the complete data distribution even when we're only interested in a low-dimensional summary statistic of the population.
- Deals better with mis-specification

Advantages of this include

- Allows focus solely on the quantities of interest.
  - ▶ Full Bayesian inference requires us to model the complete data distribution even when we're only interested in a low-dimensional summary statistic of the population.
- Deals better with mis-specification

Presumably the posterior may inherit some form of robustness from certain choices for the loss function, e.g., the bounded robust proper scores of Dawid *et al.* .

Advantages of this include

- Allows focus solely on the quantities of interest.
  - ▶ Full Bayesian inference requires us to model the complete data distribution even when we're only interested in a low-dimensional summary statistic of the population.
- Deals better with mis-specification

Presumably the posterior may inherit some form of robustness from certain choices for the loss function, e.g., the bounded robust proper scores of Dawid *et al.* .

Relates to the Bayes linear approach of Goldstein and Wooff which is also motivated by difficulties with specifying a complete model for the data.

## HM and ABC thresholding

History matching was an approach designed for inference for mis-specified models.

$$S_{HM}(F_\theta) = \frac{|\mathbb{E}_{F_\theta}(Y) - y|}{\sqrt{\text{Var}_{F_\theta}(y)}}$$

Often applied in a Bayes linear type setting, with  $\text{Var}_{F_\theta}(y)$  broken down into constituent parts

$$\text{Var}_{F_\theta}(y) = \text{Var}_{sim} + \text{Var}_{discrep} + \text{Var}_{emulator}$$

## HM and ABC thresholding

History matching was an approach designed for inference for mis-specified models.

$$S_{HM}(F_\theta) = \frac{|\mathbb{E}_{F_\theta}(Y) - y|}{\sqrt{\text{Var}_{F_\theta}(y)}}$$

Often applied in a Bayes linear type setting, with  $\text{Var}_{F_\theta}(y)$  broken down into constituent parts

$$\text{Var}_{F_\theta}(y) = \text{Var}_{sim} + \text{Var}_{discrep} + \text{Var}_{emulator}$$

Combined with the thresholding nature

$$\mathcal{P}_\theta = \{\theta : S_{HM}(\hat{F}_{\theta,y}) \leq 3\}$$

means we don't get asymptotic concentration.

- ABC shares similar properties if  $\epsilon$  fixed at something reasonable.

$$\pi_\epsilon(\theta) \propto \pi(\theta) \mathbb{I}_{S(\hat{F}_\theta, y) \leq \epsilon}$$

The indicator functions acts to add a ball of radius  $\epsilon$  around the data, so that we only need to get within it.

- $\epsilon$  plays the same role as  $\text{Var}_{discrep}$  in HM.



- ABC shares similar properties if  $\epsilon$  fixed at something reasonable.

$$\pi_\epsilon(\theta) \propto \pi(\theta) \mathbb{I}_{S(\hat{F}_{\theta, y}) \leq \epsilon}$$

The indicator functions acts to add a ball of radius  $\epsilon$  around the data, so that we only need to get within it.

- $\epsilon$  plays the same role as  $\text{Var}_{discrep}$  in HM.

Both approaches also allow the user to focus on aspects/summaries of the simulator output that either are of interest, or for which we believe the simulator is better specified.

- We discard information by only using some aspects of the simulator output, but perhaps to benefit of the inference

- ABC shares similar properties if  $\epsilon$  fixed at something reasonable.

$$\pi_\epsilon(\theta) \propto \pi(\theta) \mathbb{I}_{S(\hat{F}_{\theta,y}) \leq \epsilon}$$

The indicator functions acts to add a ball of radius  $\epsilon$  around the data, so that we only need to get within it.

- $\epsilon$  plays the same role as  $\text{Var}_{discrep}$  in HM.

Both approaches also allow the user to focus on aspects/summaries of the simulator output that either are of interest, or for which we believe the simulator is better specified.

- We discard information by only using some aspects of the simulator output, but perhaps to benefit of the inference

Also

- Allow for crude/simple discrepancy characterization.
- Some form of robustness arises from the scores used.

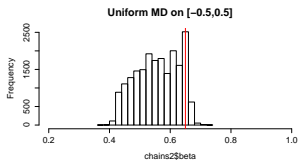
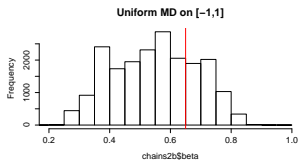
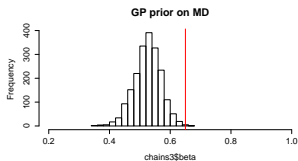
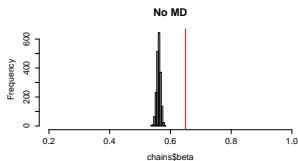
# Brynjarsdottir *et al.* revisited

Simulator

$$f_{\theta}(x) = \theta x$$

Reality

$$g(x) = \frac{\theta x}{1 + \frac{x}{a}} \quad \theta = 0.65, a = 20$$



## Recent work in ABC

Recent work on ABC has sought to move away from the use of summaries

- Bernton *et al.* 2017 look at Bayes like procedures based on the Wasserstein distance (get different pseudo-true value)
- Park *et al.* 2015 look at using kernel mean embeddings of distributions to also avoid the need to summarize outputs.

## Recent work in ABC

Recent work on ABC has sought to move away from the use of summaries

- Bernton *et al.* 2017 look at Bayes like procedures based on the Wasserstein distance (get different pseudo-true value)
- Park *et al.* 2015 look at using kernel mean embeddings of distributions to also avoid the need to summarize outputs.

Several papers (Frazier *et al.* 2017, Ridgeway 2017, ...) have studied asymptotic properties of ABC

- Consider version of ABC where we accept or reject according to

$$\rho(\eta(y), \eta(y'))$$

where  $y' \sim F_{\theta}(\cdot)$

Then if  $b_0$  is limit of  $\eta(y)$  and  $b(\theta)$  the limit of  $\eta(y')$ , then they've studied convergence to

$$\theta^* = \arg \inf_{\theta} \rho(b_0, b(\theta))$$

as  $\epsilon \rightarrow 0$ .

This focus is again on prediction not inference.

## Discussion

What properties do we want our inference scheme to possess?

## Discussion

What properties do we want our inference scheme to possess?

- Is coherence the best we can hope for or is there a form of robustness that is achievable and useful for slightly mis-specified models?

## Discussion

What properties do we want our inference scheme to possess?

- Is coherence the best we can hope for or is there a form of robustness that is achievable and useful for slightly mis-specified models?
- If  $G \notin \mathcal{F}$  can we ever hope to learn precisely about  $\theta$ ?  
If not we shouldn't use methods that converge/concentrate asymptotically.



## Discussion

What properties do we want our inference scheme to possess?

- Is coherence the best we can hope for or is there a form of robustness that is achievable and useful for slightly mis-specified models?
- If  $G \notin \mathcal{F}$  can we ever hope to learn precisely about  $\theta$ ?  
If not we shouldn't use methods that converge/concentrate asymptotically.
- Whilst modelling our way out of trouble sounds attractive, in practice it often fails (rarely works?) due to lack of identifiability.
  - ▶ Simple specification of discrepancies (Bayes linear?) look attractive in most cases. Should we just use inferential approaches that allow for this type of simple specification (ie which allow us to avoid full probabilistic models)?

## Discussion

What properties do we want our inference scheme to possess?

- Is coherence the best we can hope for or is there a form of robustness that is achievable and useful for slightly mis-specified models?
- If  $G \notin \mathcal{F}$  can we ever hope to learn precisely about  $\theta$ ?  
If not we shouldn't use methods that converge/concentrate asymptotically.
- Whilst modelling our way out of trouble sounds attractive, in practice it often fails (rarely works?) due to lack of identifiability.
  - ▶ Simple specification of discrepancies (Bayes linear?) look attractive in most cases. Should we just use inferential approaches that allow for this type of simple specification (ie which allow us to avoid full probabilistic models)?

No one trusts a model except the person who wrote it; everyone trusts an observation except the person who made it, Harlow Shapely.

Thank you for listening!