Topics in uncertainty quantification

Richard Wilkinson

School of Maths and Statistics
University of Sheffield

July 2018

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
- Parameter estimation
- Sensitivity analysis
- Prediction
- Decision making

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

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

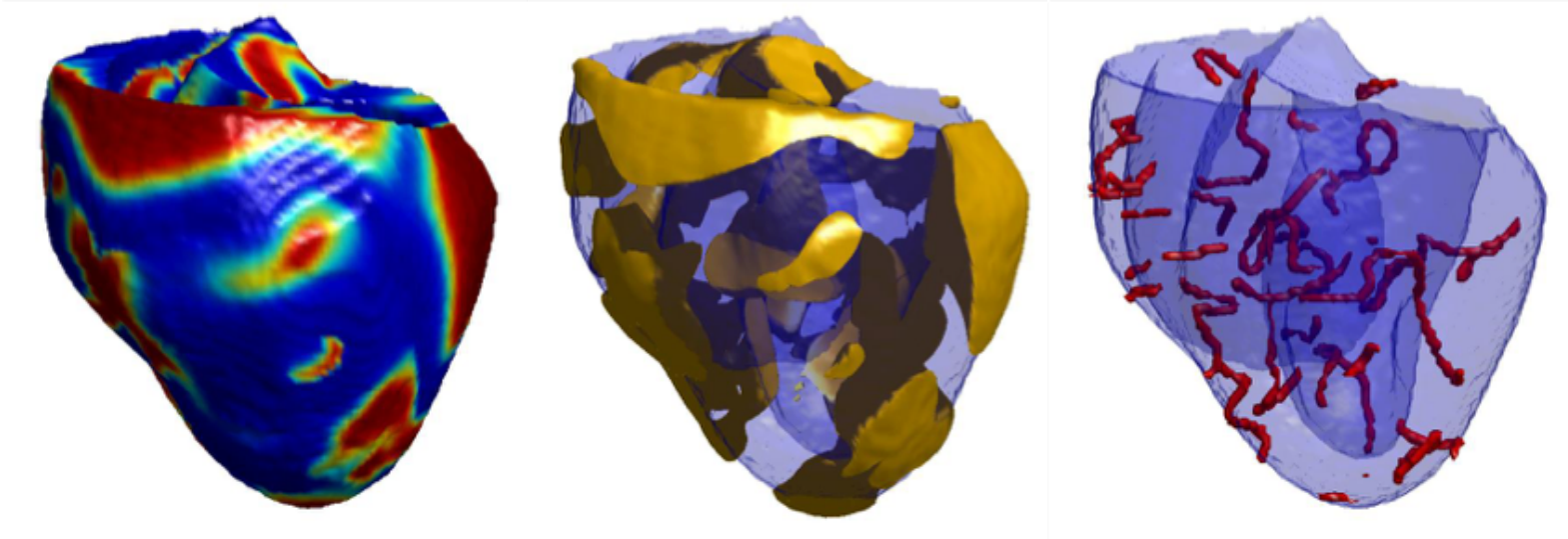
- Uncertainty propagation
- Parameter estimation
- Sensitivity analysis
- Prediction
- Decision making

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

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

Why do we need UQ?

Atrial fibrillation



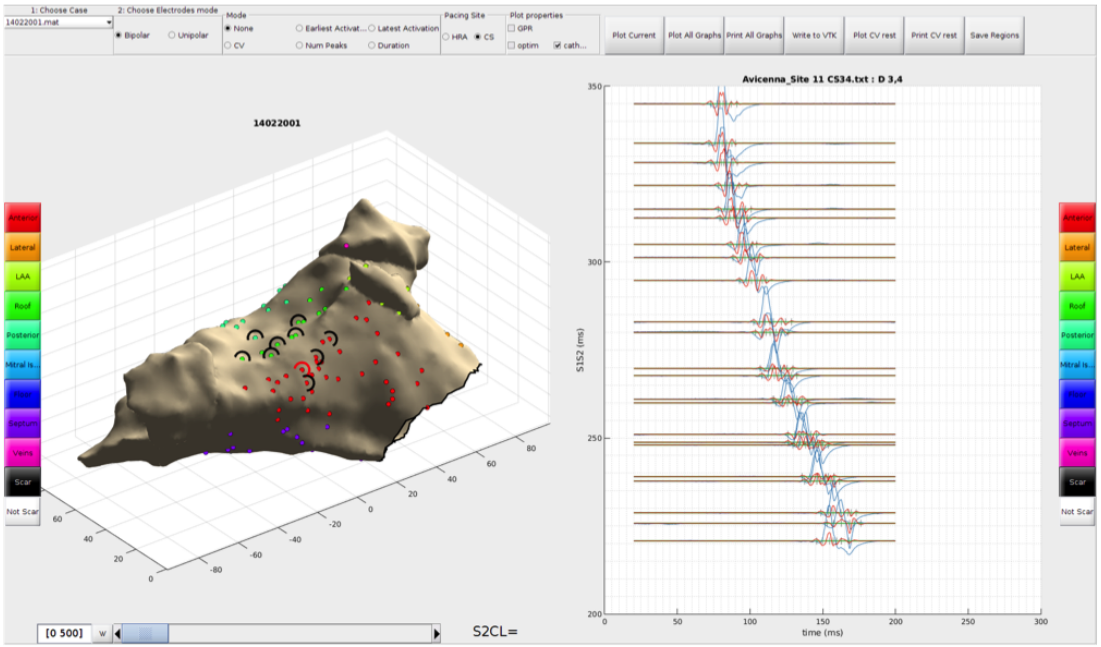
Atrial fibrillation (AF) - rapid and uncoordinated electrical activation (arrhythmia) leading to poor mechanical function.

- 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 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)

Recent progress in UQ

A good many times I have been present at gatherings of [highly-educated] people... who have with considerable gusto been expressing their incredulity at the illiteracy of scientists. Once or twice I have been provoked and have asked the company how many of them could describe the Second Law of Thermodynamics. The response was cold...Yet I was asking something which is the scientific equivalent of: Have you read a work of Shakespeare's? C.P. Snow, 'The Two Cultures'

Recent progress in UQ

A good many times I have been present at gatherings of [highly-educated] people... who have with considerable gusto been expressing their incredulity at the illiteracy of scientists. Once or twice I have been provoked and have asked the company how many of them could describe the Second Law of Thermodynamics. The response was cold...Yet I was asking something which is the scientific equivalent of: Have you read a work of Shakespeare's? C.P. Snow, 'The Two Cultures'

- Statisticians: 'What about the real world?'
- Applied maths: 'Where is the theory? Error guarantees?'
- Machine learning: 'Why weren't we invited?'

Hot topics

- Surrogate models
- Calibration/parameter estimation
- Model discrepancy
- Multi-fidelity models
- High dimensional problems
- Machine learning models
- Communicating uncertainty

I: Surrogate models

Code uncertainty

Think of the simulator as a function

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

Monte Carlo (brute force) can be used for most tasks if sufficient computational resource is available. But for long run times, we will only know the simulator output at a small number of points:

I: Surrogate models

Code uncertainty

Think of the simulator as a function

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

Monte Carlo (brute force) can be used for most tasks if sufficient computational resource is available. But for long run times, we will only know the simulator output at a small number of points:

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

$$D_{sim} = \{(x_i, f(x_i))\}_{i=1,\ldots,N}$$

I: Surrogate models

Code uncertainty

Think of the simulator as a function

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

Monte Carlo (brute force) can be used for most tasks if sufficient computational resource is available. But for long run times, we will only know the simulator output at a small number of points:

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

$$D_{sim} = \{(x_i, f(x_i))\}_{i=1,\ldots,N}$$

- If θ is not in the ensemble, then we are uncertainty about $f(x)$ - *code uncertainty*
- $\mathcal{X} \subset \mathbb{R}^{10}$ then 1000 simulator runs is only enough for one point in each corner of the design space.

Surrogate models

If the simulator is expensive, look to approximate it with a surrogate

Surrogate models

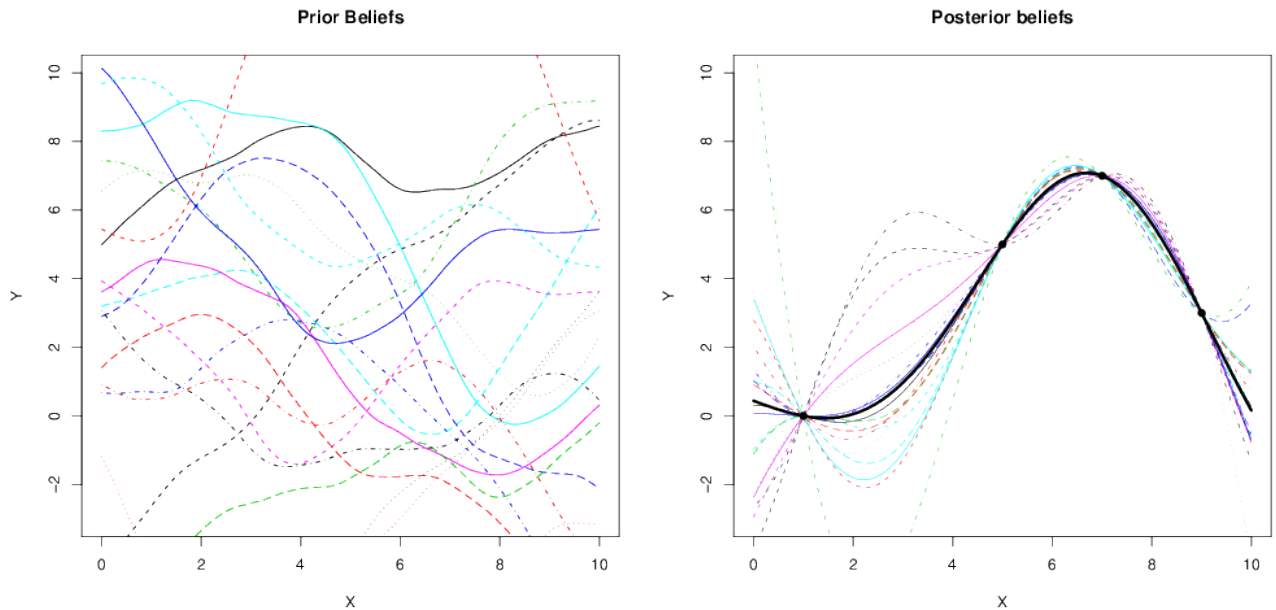
If the simulator is expensive, look to approximate it with a surrogate

- coarse-grid approximations, projection-based reduced models, and simplified physics models

Surrogate models

If the simulator is expensive, look to approximate it with a surrogate

- coarse-grid approximations, projection-based reduced models, and simplified physics models
- Data-fit regression models - primarily Gaussian processes



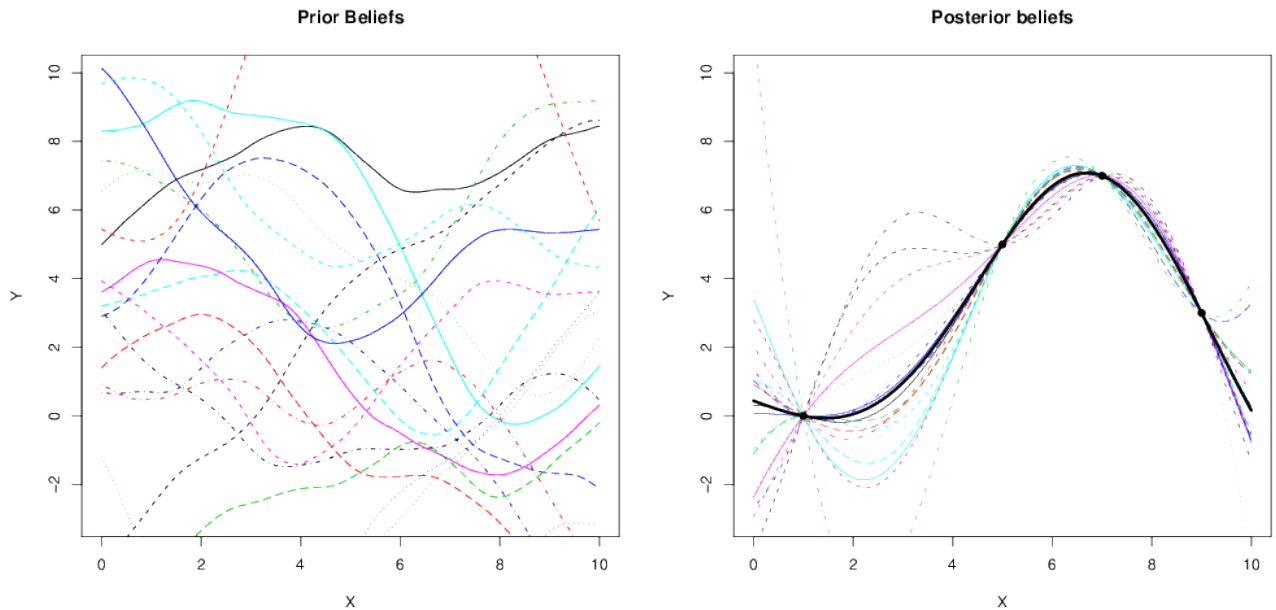
A GP is a random process indexed by $x \in \mathcal{X}$ say, such that for every finite set of indices, $x_1, \ldots, x_n$,

$$\mathbf{f} = (f(x_1), \ldots, f(x_n)) \sim \text{ multivariate Gaussian distribution}$$

Surrogate models

If the simulator is expensive, look to approximate it with a surrogate

- coarse-grid approximations, projection-based reduced models, and simplified physics models
- Data-fit regression models - primarily Gaussian processes



A GP is a random process indexed by $x \in \mathcal{X}$ say, such that for every finite set of indices, $x_1, \ldots, x_n$,

$$\mathbf{f} = (f(x_1), \ldots, f(x_n)) \sim \text{ multivariate Gaussian distribution}$$

Why would we want to use this very restricted model?

Answer 1

Class of models is closed under various operations.

Answer 1

Class of models is closed under various operations.

- Closed under addition

$$f_1(\cdot), f_2(\cdot) \sim GP \quad \text{then} \quad (f_1 + f_2)(\cdot) \sim GP$$

Answer 1

Class of models is closed under various operations.

- Closed under addition

$$f_1(\cdot), f_2(\cdot) \sim GP \quad \text{then} \quad (f_1 + f_2)(\cdot) \sim GP$$

- Closed under Bayesian conditioning, i.e., if we observe

$$\mathbf{D} = (f(x_1), \ldots, f(x_n))$$

then

$$f|D \sim GP$$

but with updated mean and covariance functions.

Answer 1

Class of models is closed under various operations.

- Closed under addition

$$f_1(\cdot), f_2(\cdot) \sim GP \quad \text{then} \quad (f_1 + f_2)(\cdot) \sim GP$$

- Closed under Bayesian conditioning, i.e., if we observe

$$\mathbf{D} = (f(x_1), \ldots, f(x_n))$$

then

$$f|D \sim GP$$

but with updated mean and covariance functions.

- Closed under any linear operation. If $\mathcal{L}$ is a linear operator, then

$$\mathcal{L}f \sim GP(\mathcal{L}m, \mathcal{L}k\mathcal{L}^\top)$$

e.g. $\frac{df}{dx}$, $\int f(x)dx$, Af are all GPs

Answer 2: non-parametric/kernel regression

- Linear regression $y = x^\top \beta + \epsilon$ can be written solely in terms of inner products $x^\top x$.

$$\begin{aligned}\hat{\beta} &= \arg\min ||y - X\beta||_2^2 + \sigma^2 ||\beta||_2^2 \\ &= X^\top (XX^\top + \sigma^2 I)^{-1} y \quad \text{(the dual form)}\end{aligned}$$

Answer 2: non-parametric/kernel regression

- Linear regression $y = x^\top \beta + \epsilon$ can be written solely in terms of inner products $x^\top x$.

$$\begin{aligned}\hat{\beta} &= \arg\min ||y - X\beta||_2^2 + \sigma^2 ||\beta||_2^2 \\ &= X^\top (XX^\top + \sigma^2 I)^{-1} y \quad \text{(the dual form)}\end{aligned}$$

- We know that we can replace x by a feature vector in linear regression, e.g., $\phi(x) = (1\ x\ x^2\ \cos(x))^\top$ etc.

Answer 2: non-parametric/kernel regression

- Linear regression $y = x^\top \beta + \epsilon$ can be written solely in terms of inner products $x^\top x$.

$$\begin{aligned}\hat{\beta} &= \arg\min ||y - X\beta||_2^2 + \sigma^2 ||\beta||_2^2 \\ &= X^\top (XX^\top + \sigma^2 I)^{-1} y \quad \text{(the dual form)}\end{aligned}$$

- We know that we can replace x by a feature vector in linear regression, e.g., $\phi(x) = (1\ x\ x^2\ \cos(x))^\top$ etc.
- For some features, inner product is equivalent to evaluating a kernel

$$\phi(x)^\top \phi(x') \equiv k(x, x')$$

where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a semi-positive definite function.

Answer 2: non-parametric/kernel regression

- Linear regression $y = x^\top \beta + \epsilon$ can be written solely in terms of inner products $x^\top x$.

$$\begin{aligned}\hat{\beta} &= \arg\min ||y - X\beta||_2^2 + \sigma^2 ||\beta||_2^2 \\ &= X^\top (XX^\top + \sigma^2 I)^{-1} y \quad \text{(the dual form)}\end{aligned}$$

- We know that we can replace x by a feature vector in linear regression, e.g., $\phi(x) = (1\ x\ x^2\ \cos(x))^\top$ etc.
- For some features, inner product is equivalent to evaluating a kernel

$$\phi(x)^\top \phi(x') \equiv k(x, x')$$

where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a semi-positive definite function.
Kernel trick: lift x into infinite dimensional feature space by replacing inner products $x^\top x'$ by $k(x, x')$, but never evaluate these features, only the $n \times n$ kernel matrix.

$$\hat{y}' = m(x') = \sum_{i=1}^{n} \alpha_i k(x, x_i)$$

Generally, we don't think about features, we just choose a kernel. But choosing a kernel is implicitly choosing features, and our model only includes functions that are linear combinations of this set of features (the Reproducing Kernel Hilbert Space (RKHS) of k).

Generally, we don't think about features, we just choose a kernel. But choosing a kernel is implicitly choosing features, and our model only includes functions that are linear combinations of this set of features (the Reproducing Kernel Hilbert Space (RKHS) of k).

Example: If (modulo some detail)

$$\phi(x) = (e^{-\frac{(x-c_1)^2}{2\lambda^2}}, \ldots, e^{-\frac{(x-c_N)^2}{2\lambda^2}})$$

then as $N \to \infty$ then

$$\phi(x)^\top \phi(x) = \exp\left(-\frac{(x-x')^2}{2\lambda^2}\right)$$

Generally, we don't think about features, we just choose a kernel. But choosing a kernel is implicitly choosing features, and our model only includes functions that are linear combinations of this set of features (the Reproducing Kernel Hilbert Space (RKHS) of k).

Example: If (modulo some detail)

$$\phi(x) = (e^{-\frac{(x-c_1)^2}{2\lambda^2}}, \ldots, e^{-\frac{(x-c_N)^2}{2\lambda^2}})$$

then as $N \to \infty$ then

$$\phi(x)^\top \phi(x) = \exp\left(-\frac{(x-x')^2}{2\lambda^2}\right)$$

Although our simulator may not lie in the RKHS defined by k, this space is much richer than any parametric regression model (and can be dense in some sets of continuous bounded functions), and is thus more likely to contain an element close to the simulator than any class of models that contains only a finite number of features.

Answer 3: Naturalness of GP framework

Why use **Gaussian** processes as non-parametric models?

Answer 3: Naturalness of GP framework

Why use **Gaussian** processes as non-parametric models?

One answer might come from Bayes linear methods[1].
If we only knew the expectation and variance of some random variables, X and Y, then how should we best do statistics?

[1] statistics without probability

Answer 3: Naturalness of GP framework

Why use **Gaussian** processes as non-parametric models?

One answer might come from Bayes linear methods[1].
If we only knew the expectation and variance of some random variables, X and Y, then how should we best do statistics?

It can been shown, that the best second-order inference we can do to update our beliefs about X given Y is

$$\mathbb{E}(X|Y) = \mathbb{E}(X) + \mathbb{C}\text{ov}(X, Y)\mathbb{V}\text{ar}(Y)^{-1}(Y - \mathbb{E}(Y))$$

which is exactly the Gaussian process update for the posterior mean.

So GPs are in some sense very natural approaches.

[1] statistics without probability

Grey box models: physically obedient GPs

With Nigel Clarke

Black box methods use no knowledge of the underlying equations in the model

Intrusive methods require complete knowledge

Grey box models: physically obedient GPs

With Nigel Clarke

Black box methods use no knowledge of the underlying equations in the model

Intrusive methods require complete knowledge

Can we develop 'grey-box' methods?

E.g. suppose model output is $f(x)$ where f is the solution of

$$\mathcal{F}_x^1[f] = 0$$
$$\mathcal{F}_x^2[f] = w(x)$$
$$\vdots$$

Can we find GP emulators that obey simpler constraints exactly, and use data to train to the other constraints?

Grey box models: physically obedient GPs

With Nigel Clarke

Black box methods use no knowledge of the underlying equations in the model

Intrusive methods require complete knowledge

Can we develop 'grey-box' methods?

E.g. suppose model output is $f(x)$ where f is the solution of

$$\mathcal{F}_x^1[f] = 0$$
$$\mathcal{F}_x^2[f] = w(x)$$
$$\vdots$$

Can we find GP emulators that obey simpler constraints exactly, and use data to train to the other constraints?

E.g., guarantee that $\nabla . f = 0$ or $\nabla \times f = 0$ etc.

Grey box models: physically obedient GPs

Jidling *et al.* 2017

Simple idea: Suppose $f = \mathcal{G}_x[g]$ for some linear operator $\mathcal{G}_x$ so that for any function g, f satisfies $\mathcal{F}_x[f] = 0$ for linear operator $\mathcal{F}_x$.

Grey box models: physically obedient GPs

Jidling *et al.* 2017

Simple idea: Suppose $f = \mathcal{G}_x[g]$ for some linear operator $\mathcal{G}_x$ so that for any function g, f satisfies $\mathcal{F}_x[f] = 0$ for linear operator $\mathcal{F}_x$.
e.g. if $f : \mathbb{R}^2 \to \mathbb{R}^2$ and

$$\mathcal{F}_x = \begin{pmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} \end{pmatrix} \qquad \text{ie} \quad \mathcal{F}_x[f] = \nabla . f$$

Grey box models: physically obedient GPs

Jidling *et al.* 2017

Simple idea: Suppose $f = \mathcal{G}_x[g]$ for some linear operator $\mathcal{G}_x$ so that for any function g, f satisfies $\mathcal{F}_x[f] = 0$ for linear operator $\mathcal{F}_x$.
e.g. if $f : \mathbb{R}^2 \to \mathbb{R}^2$ and

$$\mathcal{F}_x = \begin{pmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} \end{pmatrix} \qquad \text{ie} \quad \mathcal{F}_x[f] = \nabla . f$$

then if

$$\mathcal{G}_x = \begin{pmatrix} -\frac{\partial}{\partial y} \\ \frac{\partial}{\partial x} \end{pmatrix}$$

we have $f = \mathcal{G}_x[g]$ satisfies $\mathcal{F}_x f = 0$ for all functions $g : \mathbb{R}^2 \to \mathbb{R}$.

Grey box models: physically obedient GPs

Jidling *et al.* 2017

Simple idea: Suppose $f = \mathcal{G}_x[g]$ for some linear operator $\mathcal{G}_x$ so that for any function g, f satisfies $\mathcal{F}_x[f] = 0$ for linear operator $\mathcal{F}_x$.
e.g. if $f : \mathbb{R}^2 \to \mathbb{R}^2$ and

$$\mathcal{F}_x = \begin{pmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} \end{pmatrix} \qquad \text{ie} \quad \mathcal{F}_x[f] = \nabla . f$$

then if

$$\mathcal{G}_x = \begin{pmatrix} -\frac{\partial}{\partial y} \\ \frac{\partial}{\partial x} \end{pmatrix}$$

we have $f = \mathcal{G}_x[g]$ satisfies $\mathcal{F}_x f = 0$ for all functions $g : \mathbb{R}^2 \to \mathbb{R}$.
If $g \sim GP(m(\cdot), k(\cdot, \cdot))$ then

$$f = \mathcal{G}_x[g] \sim GP(\mathcal{G}_x[m], \mathcal{G}_x k \mathcal{G}_x'^{\top})$$

So we can train emulators of f that satisfy part of the model equations.

Grey box models: physically obedient GPs

Jidling *et al.* 2017

Simple idea: Suppose $f = \mathcal{G}_x[g]$ for some linear operator $\mathcal{G}_x$ so that for any function g, f satisfies $\mathcal{F}_x[f] = 0$ for linear operator $\mathcal{F}_x$.
e.g. if $f : \mathbb{R}^2 \to \mathbb{R}^2$ and

$$\mathcal{F}_x = \begin{pmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} \end{pmatrix} \qquad \text{ie} \quad \mathcal{F}_x[f] = \nabla . f$$

then if

$$\mathcal{G}_x = \begin{pmatrix} -\frac{\partial}{\partial y} \\ \frac{\partial}{\partial x} \end{pmatrix}$$

we have $f = \mathcal{G}_x[g]$ satisfies $\mathcal{F}_x f = 0$ for all functions $g : \mathbb{R}^2 \to \mathbb{R}$.
If $g \sim GP(m(\cdot), k(\cdot, \cdot))$ then

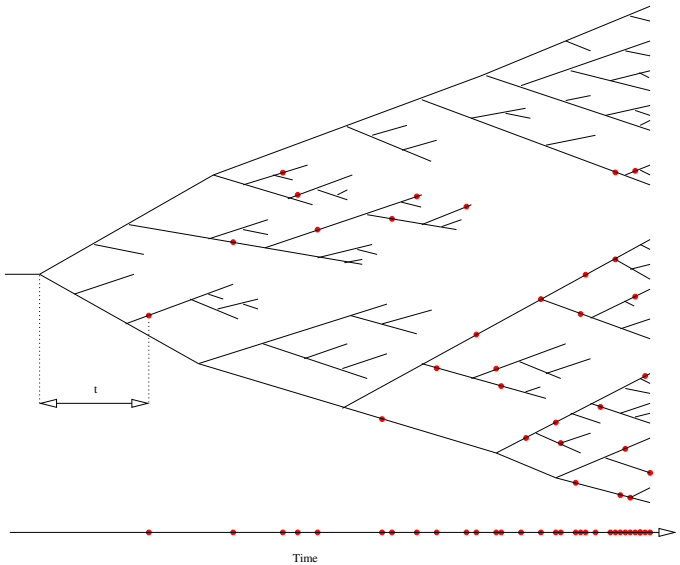
$$f = \mathcal{G}_x[g] \sim GP(\mathcal{G}_x[m], \mathcal{G}_x k \mathcal{G}_x'^{\top})$$

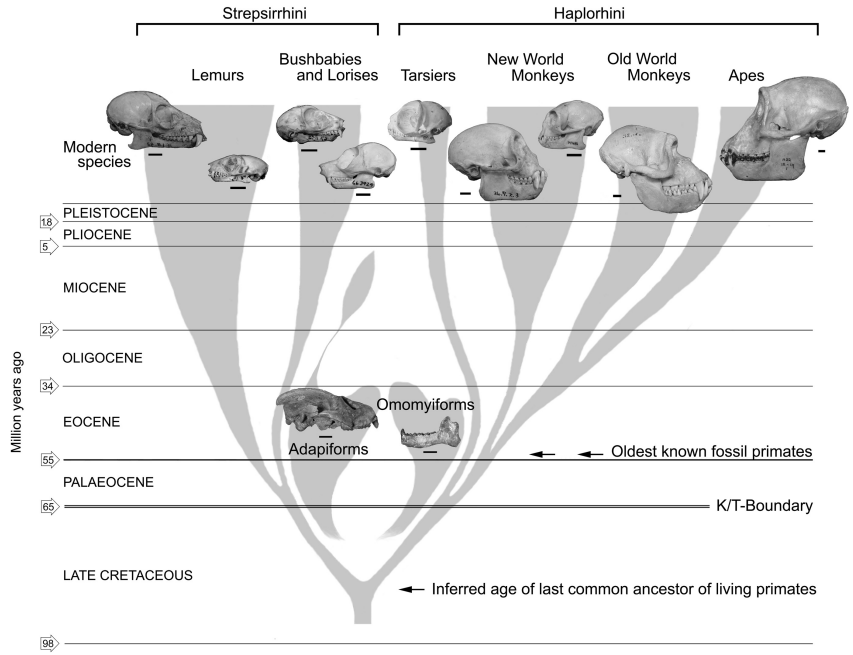
So we can train emulators of f that satisfy part of the model equations.
To find $\mathcal{G}_x$ such that $\mathcal{F}_x \mathcal{G}_x$ we look for the null space of the operator $\mathcal{F}_x$....

II: Calibration

Inverse problems/Calibration/Parameter estimation/...

- For most simulators we specify parameters θ and i.c.s and the simulator, $f(\theta)$, generates output X.
- The inverse-problem: observe data D, estimate parameter values θ which explain the data.





Major sub-discipline within statistics.

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.

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.

How should we proceed if

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

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.

How should we proceed if

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

Interest lies in inference of θ not calibrated prediction.

An appealing idea

Kennedy an O'Hagan 2001

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

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

An appealing idea

Kennedy an O'Hagan 2001

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

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

An appealing idea

Kennedy an O'Hagan 2001

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

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

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

An appealing, but flawed, idea

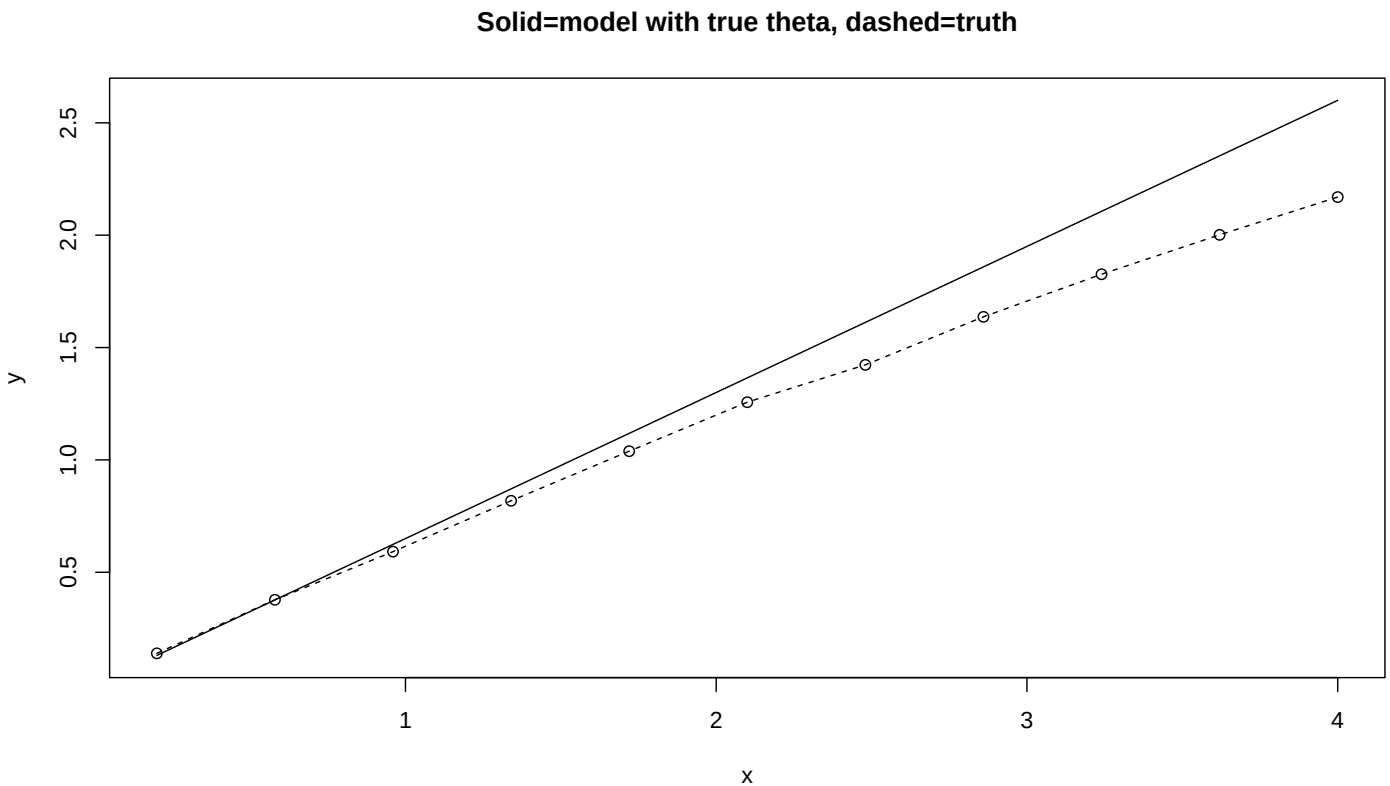
Kennedy and O'Hagan 2001, Brynjarsdottir and O'Hagan 2014

Simulator

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

Reality

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



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

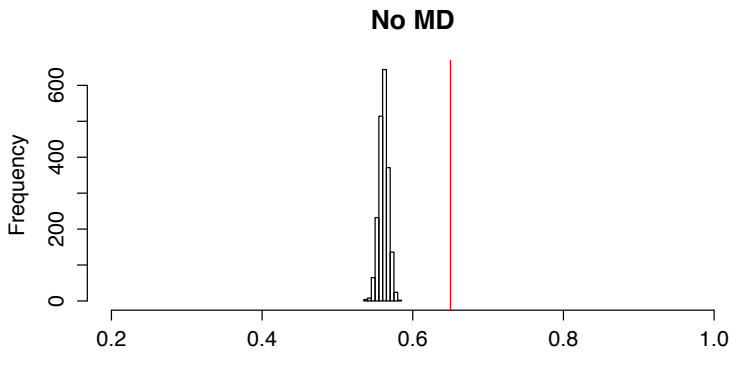
Bolting on a GP can correct your predictions, but won't necessarily fix your inference,e.g.

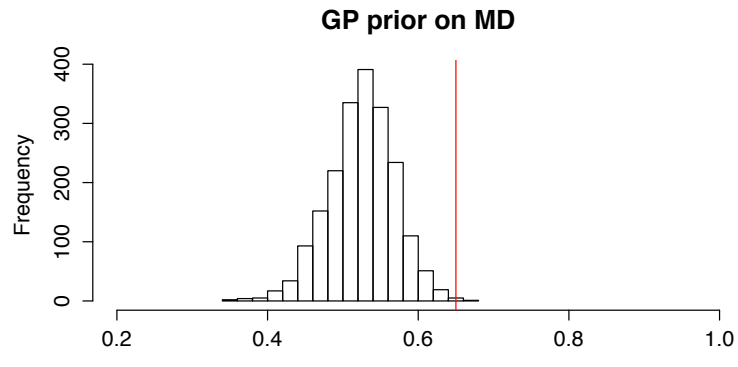
- No discrepancy:

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

- GP discrepancy:

$$y = f_\theta(x) + \delta(x) + N(0, \sigma^2),$$
$$\delta(\cdot) \sim GP(\cdot, \cdot)$$





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

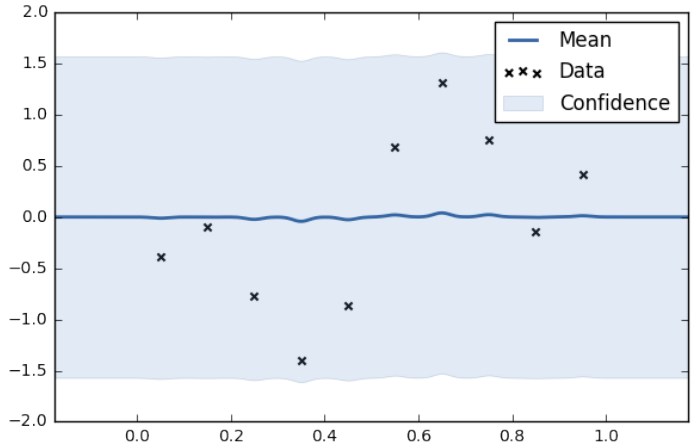
Dangers of non-parametric model extensions

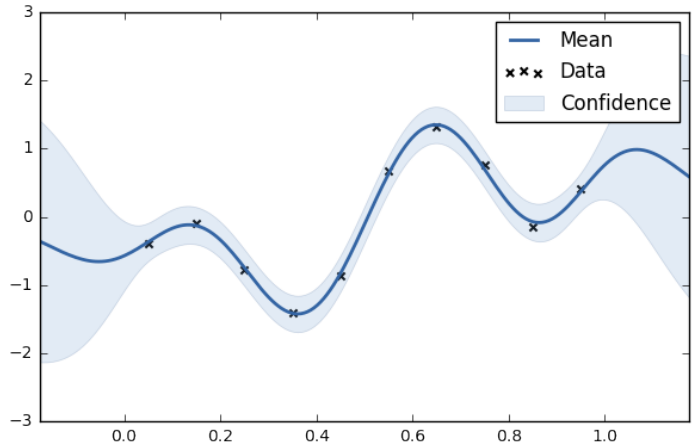
There are (at least) two problems with this approach:

- 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.

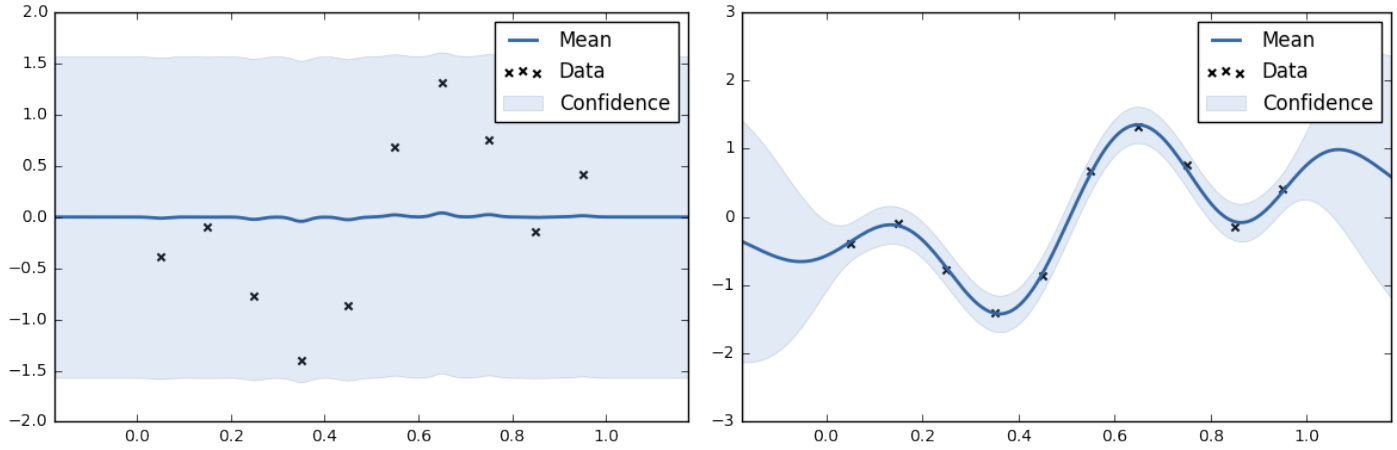
    ie We never forget the prior, but the prior is too complex to understand
  - Brynjarsdottir and O'Hagan 2014 try to model their way out of trouble with prior information - which is great if you have it.

- 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 δ and θ) by giving up and identifying

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

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)
79, *Part* 2, *pp.* 635–648

A frequentist approach to computer model calibration

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)

How do these approaches 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)

How do these approaches 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.

What properties would we like our inferential approach to possess?

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)

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

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

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)

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

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

Asymptotic consistency, efficiency, normality.
If $G \notin \mathcal{F}$

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

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

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

Bayes

Bayesian posterior

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

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

$$\pi(\theta|y) \stackrel{d}{\Rightarrow} N(\theta_0, \mathcal{I}^{-1}(\theta_0)) \text{ as } \quad n \to \infty$$

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

$$\pi(\theta|y) \stackrel{d}{\Rightarrow} N(\theta_0, \mathcal{I}^{-1}(\theta_0)) \text{ as } \quad n \to \infty$$

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 θ^* (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 $G = F_{\theta_0} \in \mathcal{F}$

$$\pi(\theta|y) \stackrel{d}{\Rightarrow} N(\theta_0, \mathcal{I}^{-1}(\theta_0)) \text{ as } \quad n \to \infty$$

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 θ^* (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 θ from prior $\pi(\cdot)$
- Accept θ with probability $\propto \pi(y \mid \theta)$

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

ABC (Approximate Bayesian computation)

Rejection Algorithm

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

Accepted θ are independent draws from the posterior distribution, $\pi(\theta \mid D)$.
If the likelihood, $\pi(D|\theta)$, is unknown:

'Mechanical' Rejection Algorithm

- Draw θ from $\pi(\cdot)$
- Simulate $y' \sim \pi(y|\theta)$ from the computer model
- Accept θ 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 θ. Instead, there is an approximate version:

Uniform Rejection Algorithm

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

Rejection ABC

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

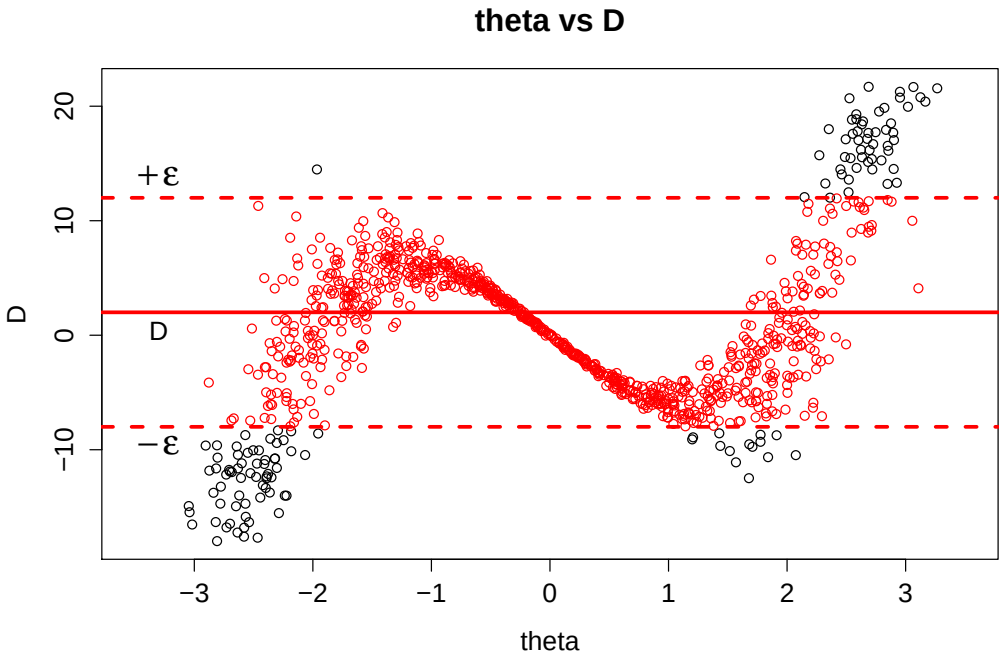
Uniform Rejection Algorithm

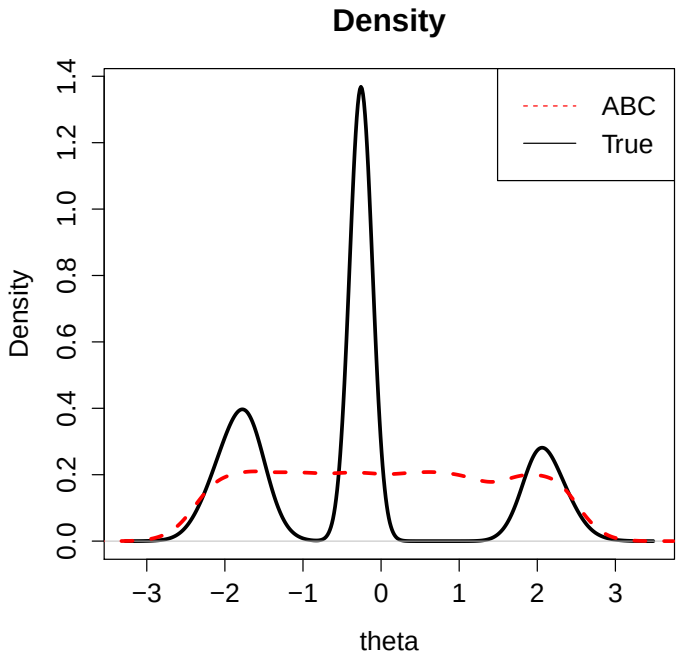
- Draw θ from $\pi(\theta)$
- Simulate $y' \sim \pi(y|\theta)$
- Accept θ if $\rho(y, y') \leq \epsilon$

ϵ reflects the tension between computability and accuracy.

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

$\epsilon = 10$

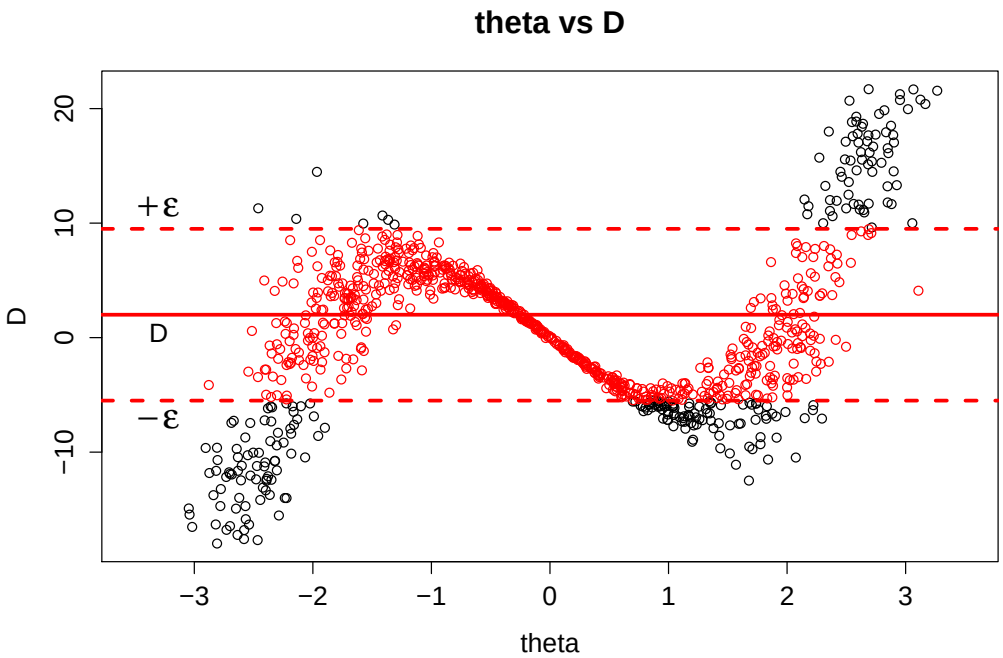


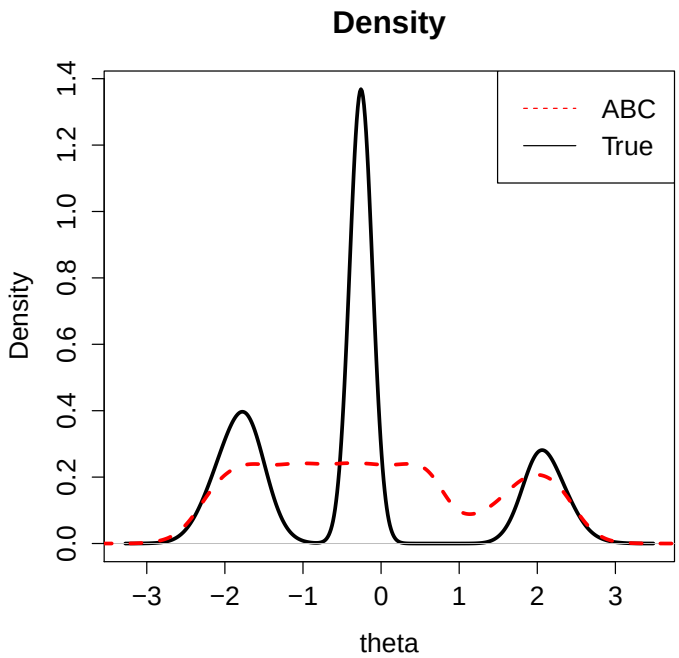


$$\theta \sim U[-10, 10], \qquad y \sim N(2(\theta + 2)\theta(\theta - 2), 0.1 + \theta^2)$$

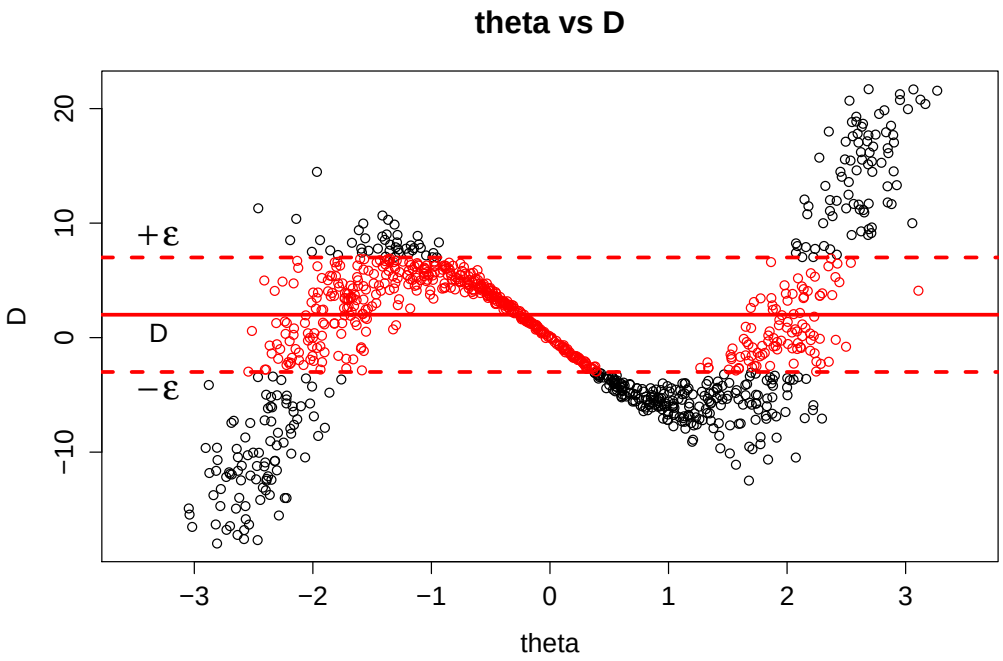
$$\rho(y, y') = |y - y'|, \qquad y = 2$$

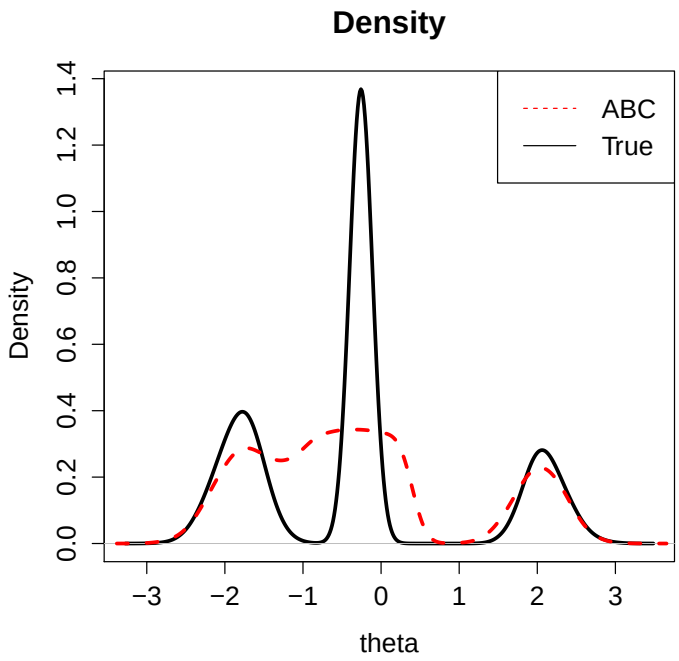
$\epsilon = 7.5$


theta vs D
D
theta
+ε
D
−ε
20
10
0
−10
−3
−2
−1
0
1
2
3

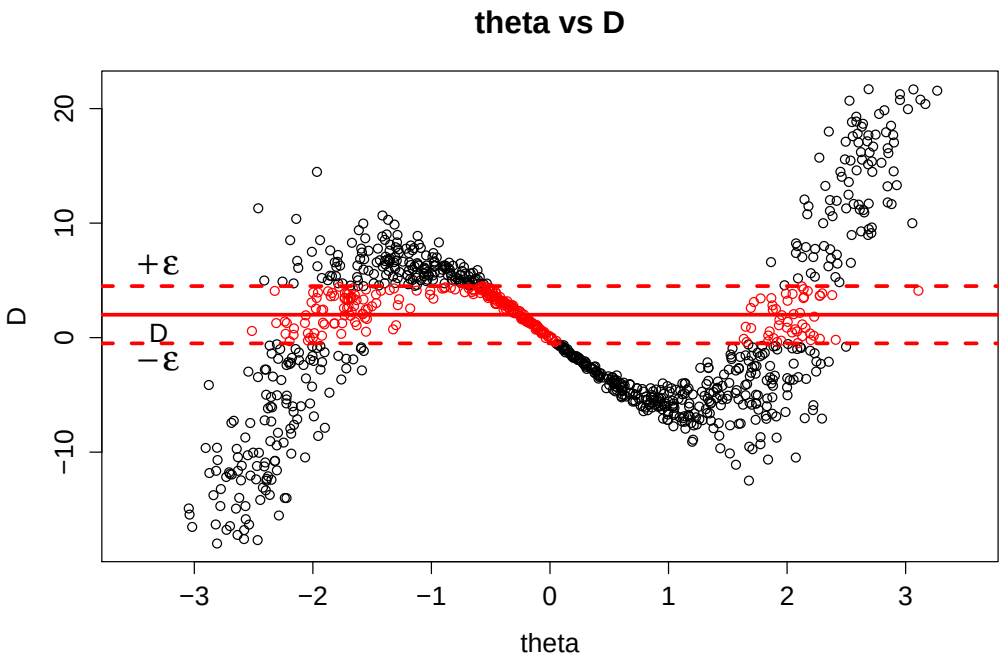

Density
Density
theta
ABC
True
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

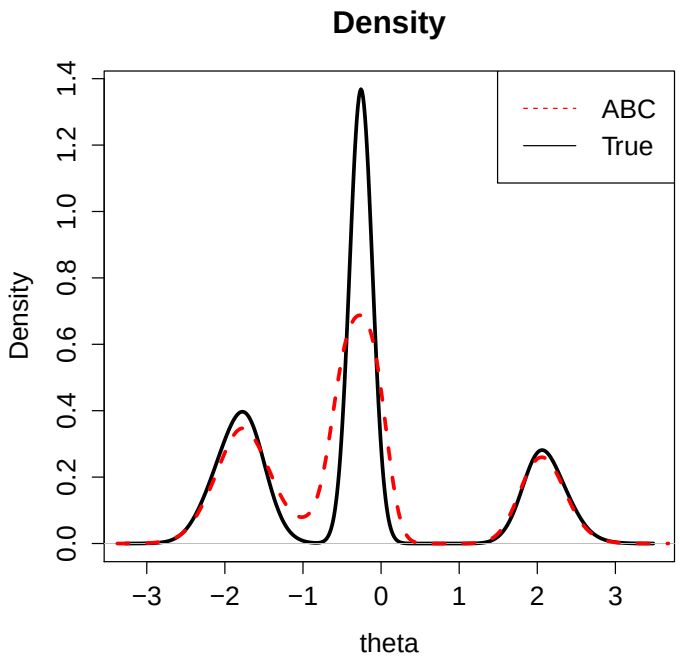
$\epsilon = 5$


theta vs D
D
theta
+ε
D
−ε
20
10
0
−10
−3
−2
−1
0
1
2
3

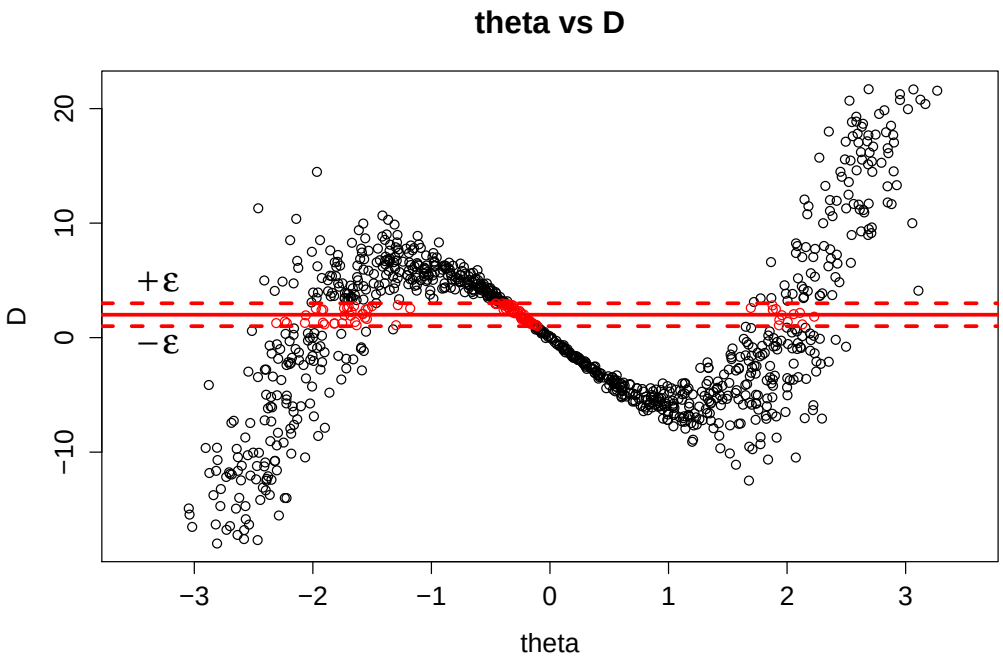

Density
Density
theta
ABC
True
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

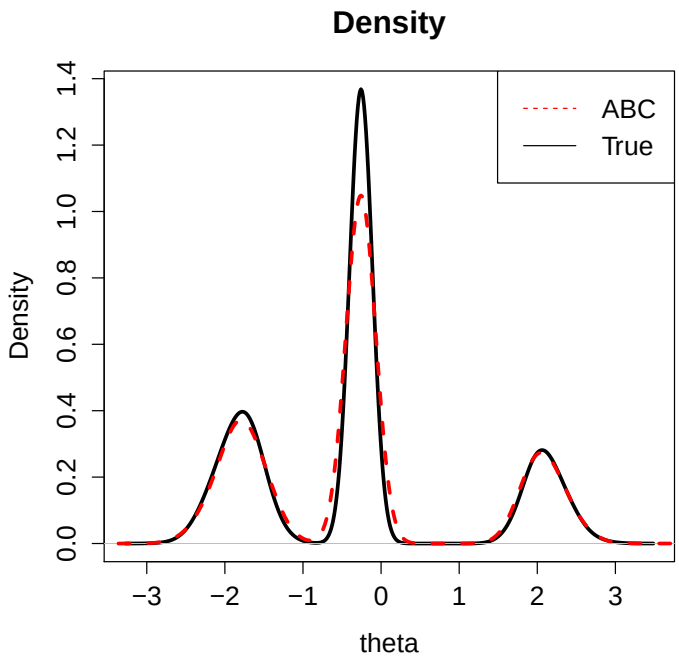
$\epsilon = 2.5$


theta vs D
D
theta
+ε
D
−ε
20
10
0
−10
−3
−2
−1
0
1
2
3


Density
Density
theta
ABC
True
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

$\epsilon = 1$


theta vs D
D
theta
+ε
−ε
20
10
0
−10
−3
−2
−1
0
1
2
3


Density
Density
theta
ABC
True
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

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

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

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

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

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 ϵ, 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

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

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 ϵ, 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.
- 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

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?

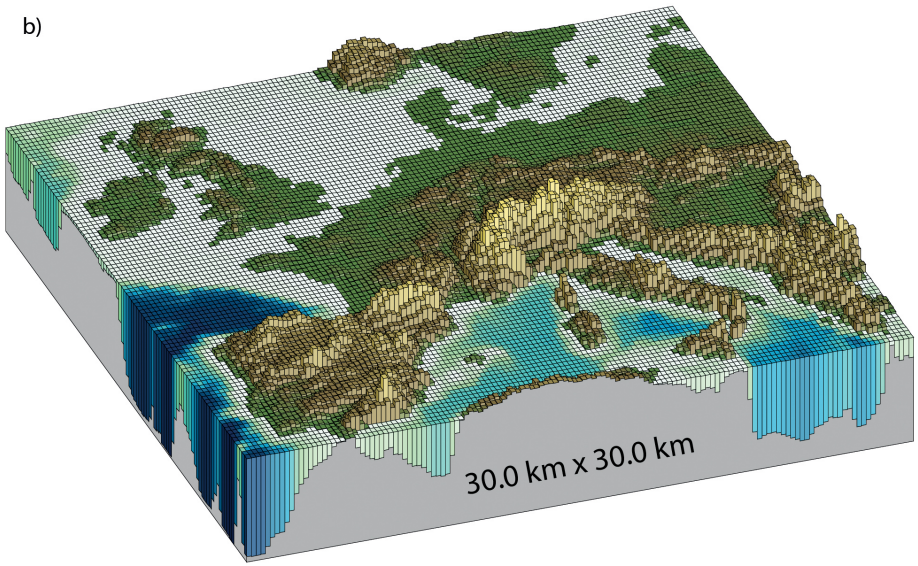
III: Multi-fidelity models

Sequence of models, $f^{(i)} : \mathcal{X} \to \mathcal{Y}$ for $i = 1, \ldots, k$ of decreasing fidelity

High-fidelity model

$$f_{hi} = f^{(1)} : \mathcal{X} \to \mathcal{Y}$$

Accurate(?) and costly



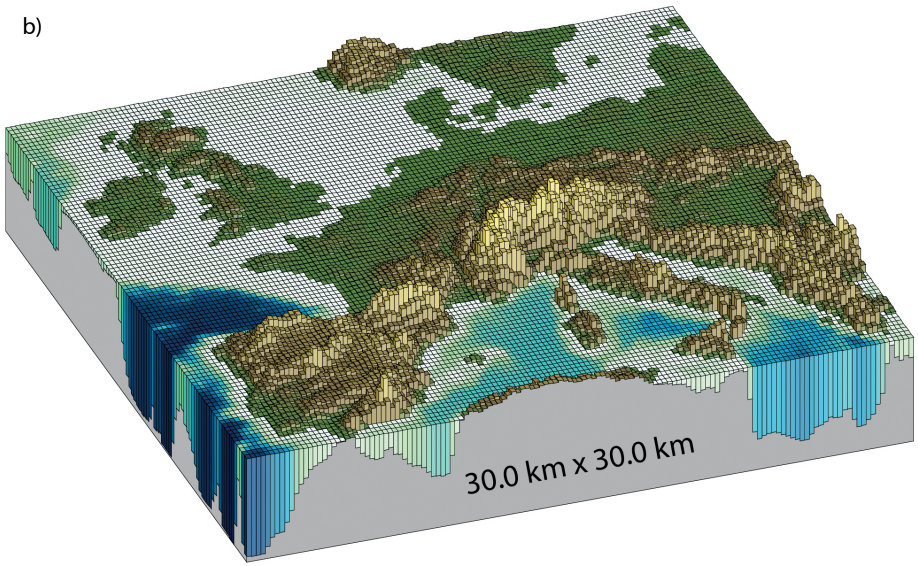
III: Multi-fidelity models

Sequence of models, $f^{(i)} : \mathcal{X} \to \mathcal{Y}$ for $i = 1, \ldots, k$ of decreasing fidelity

High-fidelity model

$$f_{hi} = f^{(1)} : \mathcal{X} \to \mathcal{Y}$$

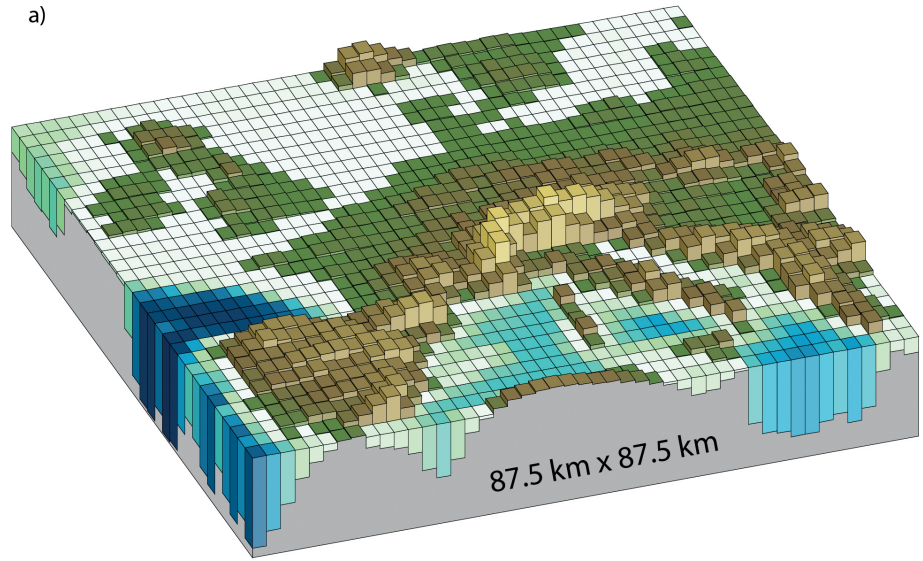
Accurate(?) and costly



Low-fidelity models

$$f_{lo} = f^{(i)} : \mathcal{X} \to \mathcal{Y}$$

Less accurate and less costly



The low-fidelity models estimate the same quantity from the same inputs, but with lower cost and lower accuracy.

Example: Multi-fidelity Uncertainty Propagation

Control variates

Basic idea:

- m an unbiased estimator of μ so that $\mathbb{E}(m) = \mu$
- t a random variable with $\mathbb{E}(t) = \tau$
- Then

$$m^* = m + c(t - \tau)$$

is also an unbiased estimator of μ for any c.

Example: Multi-fidelity Uncertainty Propagation

Control variates

Basic idea:

- m an unbiased estimator of μ so that $\mathbb{E}(m) = \mu$
- t a random variable with $\mathbb{E}(t) = \tau$
- Then

$$m^* = m + c(t - \tau)$$

  is also an unbiased estimator of μ for any c.

- The optimal choice is $c = -\mathbb{C}\text{ov}(m, t)/\mathbb{V}\text{ar}(t)$ and then

$$\mathbb{V}\text{ar}(m^*) = (1 - \rho^2)\mathbb{V}\text{ar}(m)$$

  where $\rho = \text{corr}(m, t)$

So if we can find an estimator t that is highly correlated with m we can greatly improve our estimator.

Example: Multi-fidelity Uncertainty Propagation

Peherstorfer, Willcox, Gunzburger 2016

Target: $s = \mathbb{E} f^{(1)}(X)$

Example: Multi-fidelity Uncertainty Propagation

Peherstorfer, Willcox, Gunzburger 2016

$$\text{Target: } s = \mathbb{E} f^{(1)}(X)$$

$$\approx \frac{1}{m} \sum_{j=1}^{m} f^{(1)}(x_j) := \bar{y}_m^{(1)}$$

Example: Multi-fidelity Uncertainty Propagation

Peherstorfer, Willcox, Gunzburger 2016

$$\text{Target: } s = \mathbb{E} f^{(1)}(X)$$

$$\approx \frac{1}{m} \sum_{j=1}^{m} f^{(1)}(x_j) := \bar{y}_m^{(1)}$$

We create control variates by nesting the evaluation of the low-fidelity simulators.

Example: Multi-fidelity Uncertainty Propagation

Peherstorfer, Willcox, Gunzburger 2016

$$\text{Target: } s = \mathbb{E} f^{(1)}(X)$$
$$\approx \frac{1}{m} \sum_{j=1}^{m} f^{(1)}(x_j) := \bar{y}_m^{(1)}$$

We create control variates by nesting the evaluation of the low-fidelity simulators.
We'll do m_1 evaluations of $f^{(1)}$, m_2 evaluations of $f^{(2)}$ etc with $m_i < m_{i+1}$
Given random samples $X_1, \ldots, X_{m_1}, \ldots, X_{m_2}, \ldots, X_{m_k}$ form estimator

$$\hat{s} = \bar{y}_{m_1}^{(1)} + \sum_{i=2}^{k} \alpha_i (\bar{y}_{m_i}^{(i)} - \bar{y}_{m_{i-1}}^{(i)})$$

Example: Multi-fidelity Uncertainty Propagation

Peherstorfer, Willcox, Gunzburger 2016

$$\text{Target: } s = \mathbb{E} f^{(1)}(X)$$
$$\approx \frac{1}{m} \sum_{j=1}^{m} f^{(1)}(x_j) := \bar{y}_m^{(1)}$$

We create control variates by nesting the evaluation of the low-fidelity simulators.
We'll do m_1 evaluations of $f^{(1)}$, m_2 evaluations of $f^{(2)}$ etc with $m_i < m_{i+1}$
Given random samples $X_1, \ldots, X_{m_1}, \ldots, X_{m_2}, \ldots, X_{m_k}$ form estimator

$$\hat{s} = \bar{y}_{m_1}^{(1)} + \sum_{i=2}^{k} \alpha_i (\bar{y}_{m_i}^{(i)} - \bar{y}_{m_{i-1}}^{(i)})$$

$\hat{s}$ is obviously unbiased for s.

Peherstorfer *et al.* solve the optimization problem

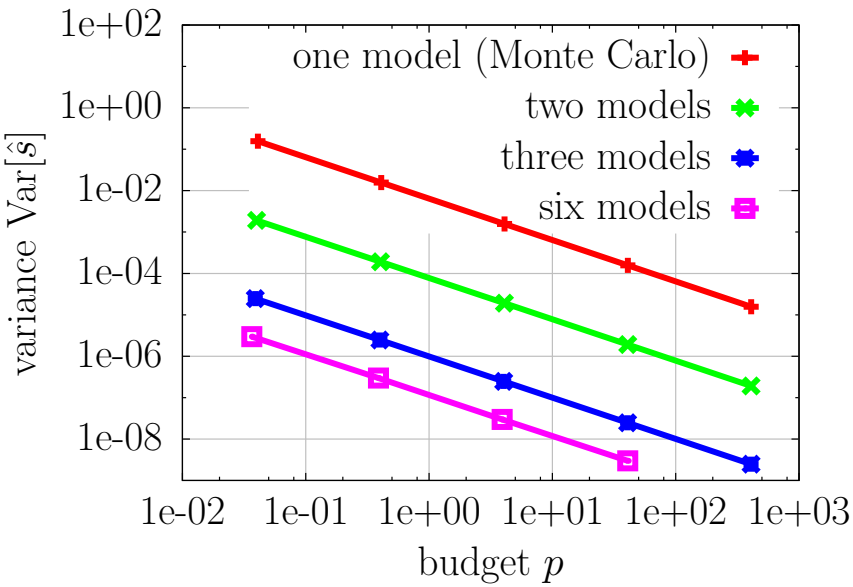
$$\begin{aligned} \min_{\mathbf{m}\in\mathbb{R}^k,\alpha_2,\ldots,\alpha_k\in\mathbb{R}} \quad & \mathbb{V}\mathrm{ar}(\hat{s}) \\ \text{s.t.} \quad & m_1 > 0 \\ & m_i > m_{i-1} \\ & \mathbf{m}^\top \mathbf{c} = \text{budget} \end{aligned}$$

for given simulator costs $c_1, \ldots, c_k$.
The solution is a function of correlations $\rho_{1,i} = \mathrm{cor}(f^{(1)}(X), f^{(i)}(X))$.

Peherstorfer *et al.* solve the optimization problem

$$\begin{aligned} \min_{\mathbf{m}\in\mathbb{R}^k,\alpha_2,\ldots,\alpha_k\in\mathbb{R}} \quad & \mathbb{V}\mathrm{ar}(\hat{s}) \\ \text{s.t.} \quad & m_1 > 0 \\ & m_i > m_{i-1} \\ & \mathbf{m}^\top \mathbf{c} = \text{budget} \end{aligned}$$

for given simulator costs $c_1, \ldots, c_k$.
The solution is a function of correlations $\rho_{1,i} = \mathrm{cor}(f^{(1)}(X), f^{(i)}(X))$.

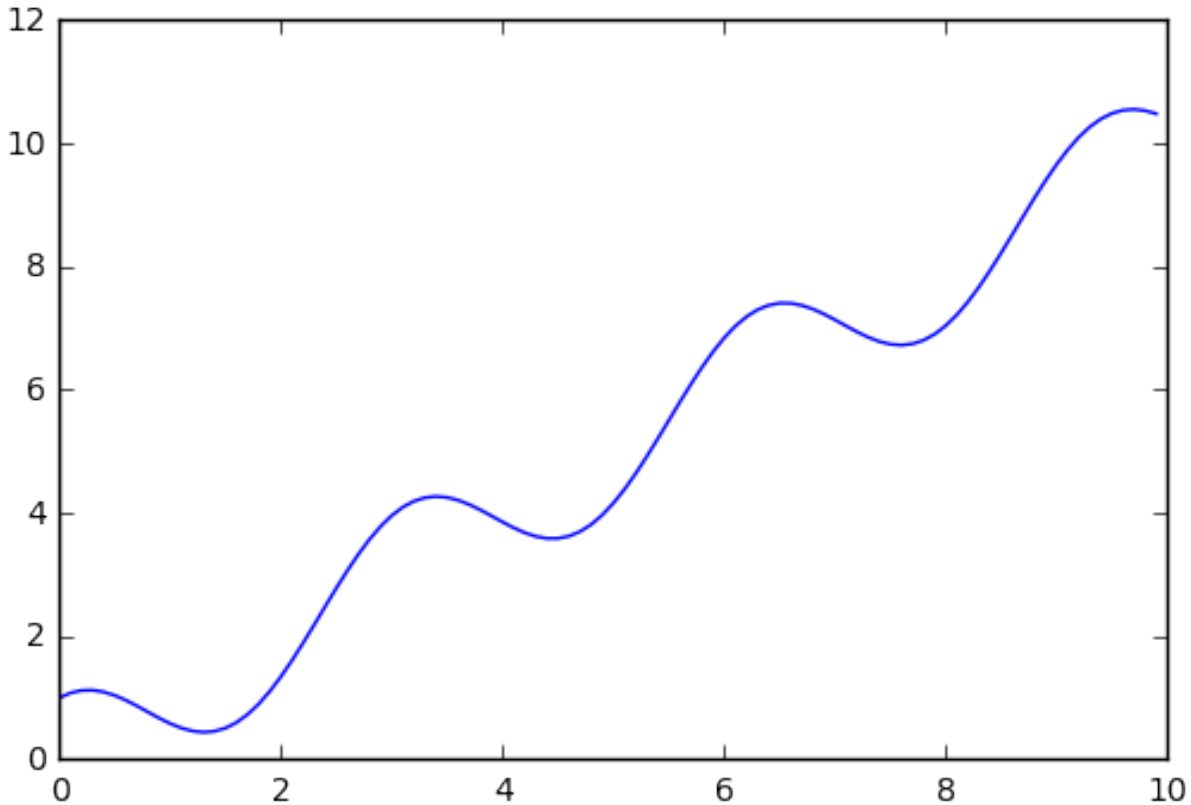


- Plate bending problem with $c_1/c_2 = 10^2$ and $\rho_{1,2} = 0.99999$
- Note there are no assumptions on the surrogate, i.e., no bounds on
$$|f^{(1)}(x) - f^{(i)}(x)|$$
- Only require the correlations $\rho_{1,i}$

Combining multifidelity MC with GP emulation

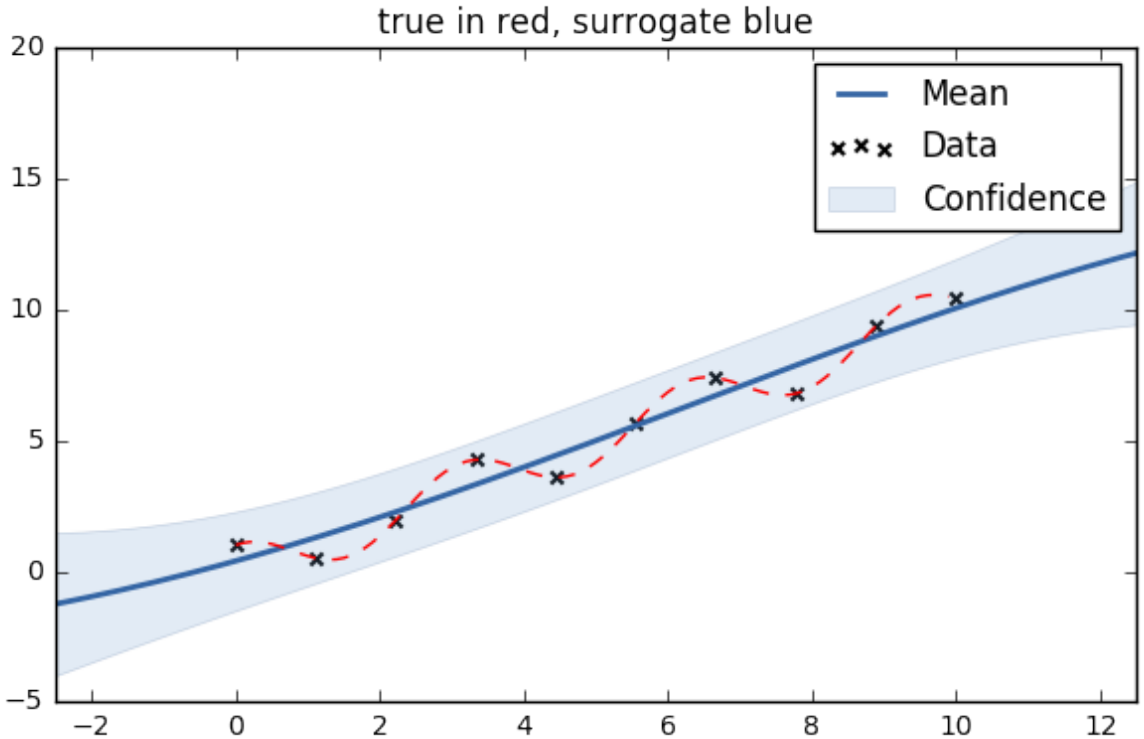
Imagine we have a expensive function f for which we want to estimate

$$\mathbb{E}f(X) = \int_0^{10} \frac{f(x)}{10} dx$$



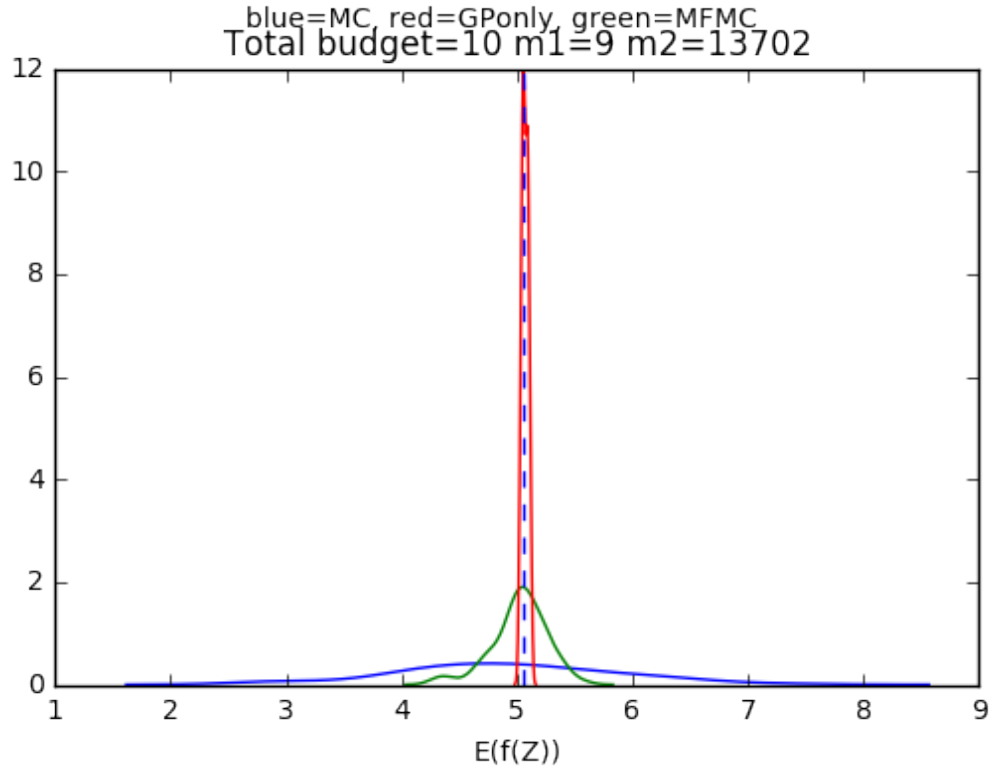
Combining multifidelity MC with GP emulation

Build a GP emulator:



Combining multifidelity MC with GP emulation

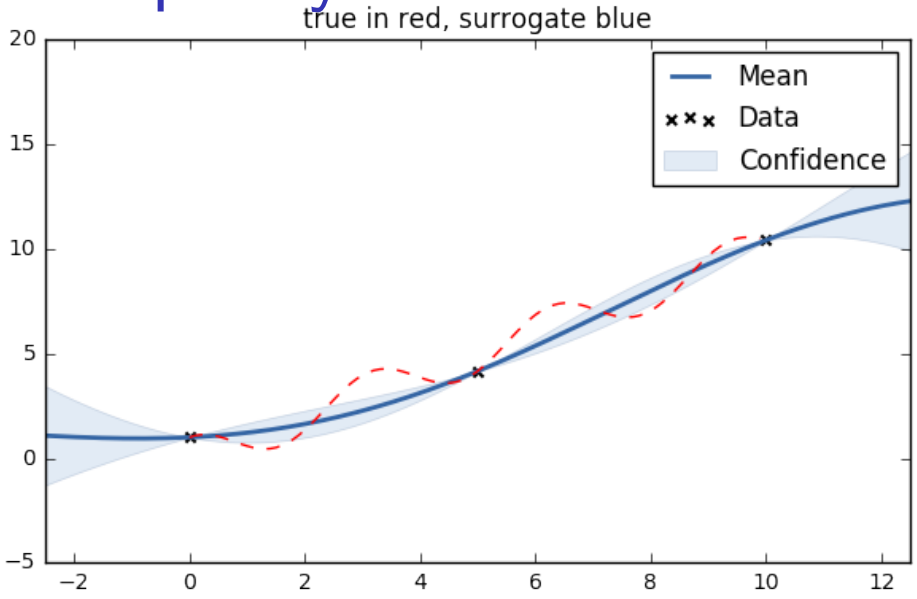
Use i) Monte Carlo, ii) just the GP, and iii) multifidelity Monte Carlo to estimate the expectation. Repeat the procedure to get an idea of sampling variation.

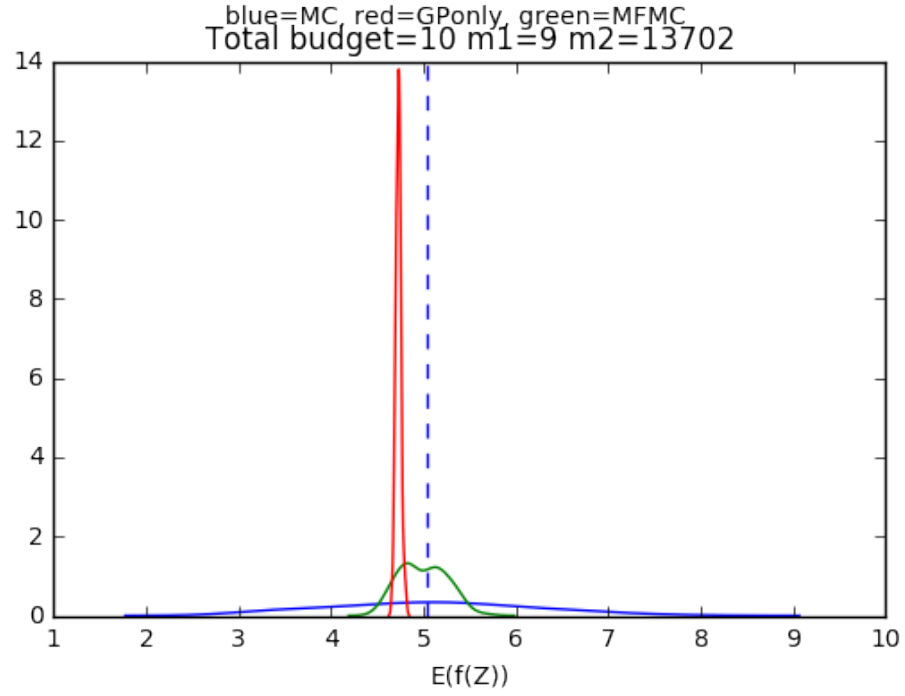


Total budget here is 10 expensive simulator evaluations, and I've assumed

$$\frac{c_1}{c_2} = 10^5$$

Lower quality emulator





For a good emulator, the MFMC estimate is worse than the estimate which just naively uses the GP.

However, the uncertainty estimates for GP emulators are often poor, particularly for high dimensional problems.

For a poor emulator, MFMC unbiases the estimate.

Problems of using GPs with MFMC

- The method requires $\sigma_i^2 = \mathbb{V}\text{ar}f^{(i)}(X)$ and $\rho_{1,i}$.
  - Estimating these is harder than estimating $s = \mathbb{E}(f^{(1)}(x))$
  - Do poor estimates reduce or eliminate the benefit of MFMC?

Problems of using GPs with MFMC

- The method requires $\sigma_i^2 = \mathbb{V}\text{ar}f^{(i)}(X)$ and $\rho_{1,i}$.
  - Estimating these is harder than estimating $s = \mathbb{E}(f^{(1)}(x))$
  - Do poor estimates reduce or eliminate the benefit of MFMC?
- When using GP emulators, for MFMC we'd need two or three training sets
  - train the emulator
  - estimate the correlations and variances
  - form the MFMC estimator

  We can't directly use the emulator training set to estimate correlations.

Problems of using GPs with MFMC

- The method requires $\sigma_i^2 = \mathbb{V}\text{ar}f^{(i)}(X)$ and $\rho_{1,i}$.
  - Estimating these is harder than estimating $s = \mathbb{E}(f^{(1)}(x))$
  - Do poor estimates reduce or eliminate the benefit of MFMC?
- When using GP emulators, for MFMC we'd need two or three training sets
  - train the emulator
  - estimate the correlations and variances
  - form the MFMC estimator

  We can't directly use the emulator training set to estimate correlations.
  - Can bootstrapping approaches reduce the number of simulator evaluations necessary and give a MFMC-GP approach which is guaranteed to be unbiased?

IV: Communicating uncertainty is hard

Maths at Sheffield
@mathsatshefuni

Following

Professor Richard Wilkinson worked with The Open University to work out the probability that governments will meet key carbon emissions targets to prevent dangerous climate change

Dangerous climate change is likely, concludes new research

A new study has revealed sensitive regions of the world are still at risk from the dangerous and potentially irreversible effects of climate change; even if we meet t...

sheffield.ac.uk

Health & Medicine | Mind & Brain | Plants & Animals | Earth & Climate | Space & Time | Matter & E

Science News

Share Blog Cite

New Statistical Model Moves Human Evolution Back Three Million Years

ScienceDaily (Nov. 5, 2010) — Evolutionary divergence of humans and chimpanzees likely occurred some 8 million years ago rather than the 5 million year estimate widely accepted by scientists, a new statistical model suggests.

A new statistical model suggests that evolutionary divergence of humans from chimpanzees likely occurred some 8 million years ago, rather than the 5 million year estimate widely accepted by scientists. (Credit: iStockphoto/Eric Gevaert)

See Also:

Plants & Animals
- Evolutionary Biology
- Nature

Computers & Math
- Statistics
- Computer Modeling

Fossils & Ruins
- Fossils
- Evolution

Reference
- Hominidae
- Multiregional hypothesis

The revised estimate of when the human species parted ways from its closest primate relatives should enable scientists to better interpret the history of human evolution, said Robert D. Martin, curator of biological anthropology at the Field Museum, and a co-author of the new study appearing in the journal *Systematic Biology*.

Working with mathematicians, anthropologists and molecular biologists, Martin has long sought to integrate evolutionary information derived from genetic material in various species with the fossil record to get a more complete picture.

Conclusion

- UQ requires a synergy between statistics, applied maths, and domain knowledge.
  - Huge unexplored gap for stats-applied math cross over.
  - Introducing physics based knowledge in ML also increasingly seen as important
- Probabilistic methods (primarily Bayesian methods) of UQ are the mainstream - venture at your peril.
- Escaping from 'model-land' is challenging.

Conclusion

- UQ requires a synergy between statistics, applied maths, and domain knowledge.
  - Huge unexplored gap for stats-applied math cross over.
  - Introducing physics based knowledge in ML also increasingly seen as important
- Probabilistic methods (primarily Bayesian methods) of UQ are the mainstream - venture at your peril.
- Escaping from 'model-land' is challenging.

Uncertainty is an uncomfortable position. But certainty is an absurd one. Voltaire

As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality. Einstein

Prediction is very difficult, especially if its about the future. Niels Bohr.

Thank you for listening!