Part II: Approximate inference for **approximate** computer models

Richard Wilkinson
Ines Krissaane
axtis21

School of Mathematical Sciences
University of Nottingham

September 7, 2021

Talk plan

Yesterday

- Approximate Bayesian computation (ABC) for complex models
- Accelerating ABC
- Surrogate models for ABC

Today

- Inference for misspecified models
- Generalized Variational Inference

Mechanistic models

Models describe hypothesised relationships between variables.

Mechanistic model

- e.g. ODE/PDE models
- explains how/why the variables interact the way they do.
- parameters may have a physical meaning
- often imperfect representations of reality, but may be the only link between the quantity of interest and the data

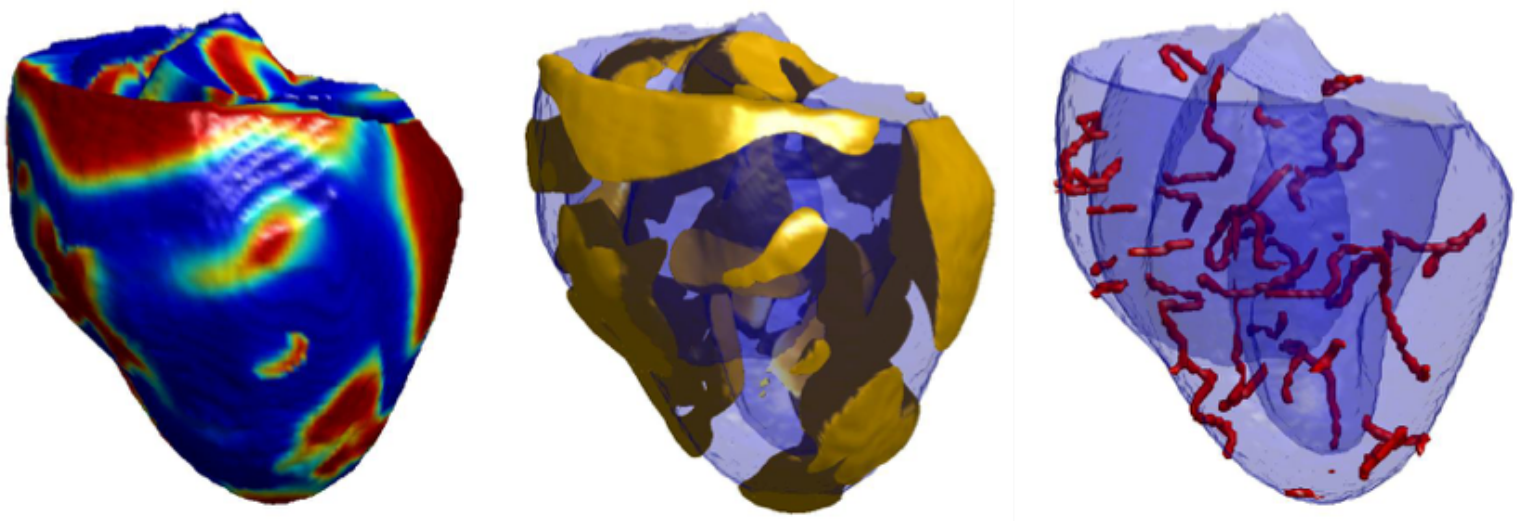
Mechanistic models

Models describe hypothesised relationships between variables.

Mechanistic model

- e.g. ODE/PDE models
- explains how/why the variables interact the way they do.
- parameters may have a physical meaning
- often imperfect representations of reality, but may be the only link between the quantity of interest and the data

e.g. **Atrial fibrillation**



UQ in Patient Specific Cardiac Models

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

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, ...

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).

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

UQ in Patient Specific Cardiac Models

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

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).

Aim: predict which AF patients will develop AT following ablation, and then treat for both in a single procedure.
We use complex electrophysiology simulations, combine these with sparse and noisy clinical data, to

- Infer tissues properties, including regions of fibrotic material
- Predict AT pathways
- Aid clinical decision making (accounting for uncertainty)

UQ in Patient Specific Cardiac Models

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

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).

Aim: predict which AF patients will develop AT following ablation, and then treat for both in a single procedure.
We use complex electrophysiology simulations, combine these with sparse and noisy clinical data, to

- Infer tissues properties, including regions of fibrotic material
- Predict AT pathways
- Aid clinical decision making (accounting for uncertainty)

However, our simulator is imperfect. How should we proceed?

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

$$y \sim G$$

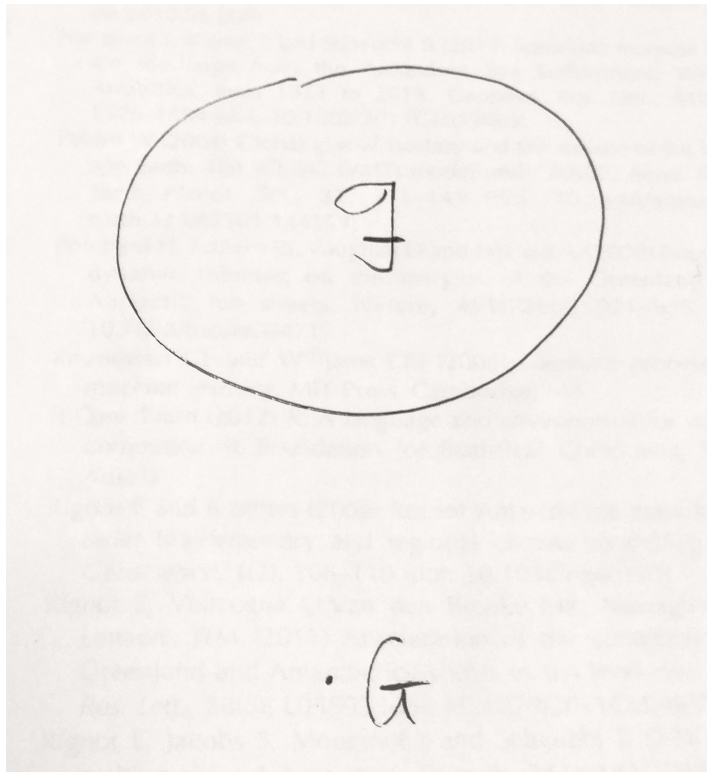
Model (complex simulator, finite dimensional parameter)

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

If $G = F_{\theta_0} \in \mathcal{F}$ then we know what to do.

How should we proceed if

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



Maximum likelihood

Maximum likelihood estimator

$$\hat{\theta}_n = \arg\max_{\theta} \log F_\theta(y_{1:n})$$

Maximum likelihood

Maximum likelihood estimator

$$\hat{\theta}_n = \arg\max_{\theta} \log F_\theta(y_{1:n})$$

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

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

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

Asymptotic consistency, efficiency, normality.

Maximum likelihood

Maximum likelihood estimator

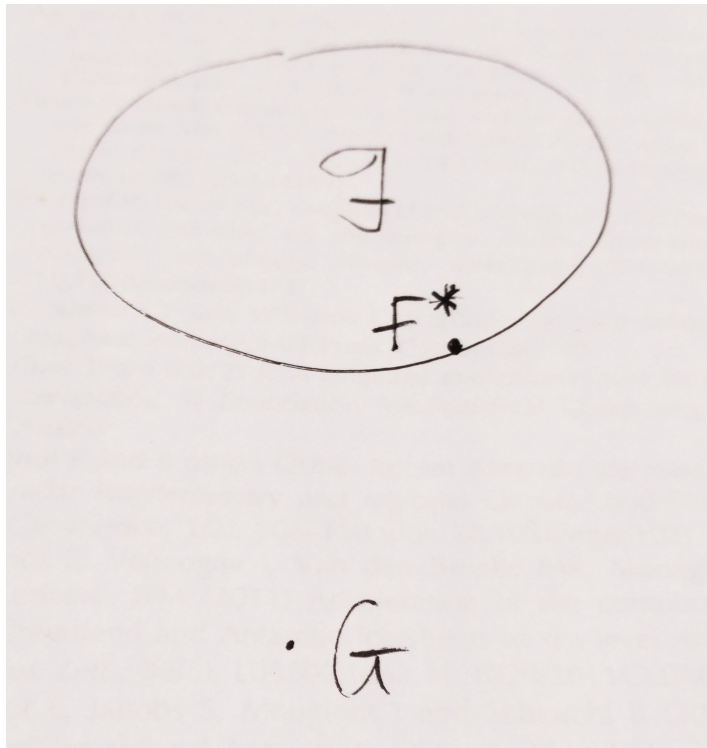
$$\hat{\theta}_n = \arg\max_{\theta} \log F_\theta(y_{1:n})$$

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 $G \notin \mathcal{F}$



Maximum likelihood

Maximum likelihood estimator

$$\hat{\theta}_n = \arg\max_{\theta} \log F_\theta(y_{1:n})$$

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

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

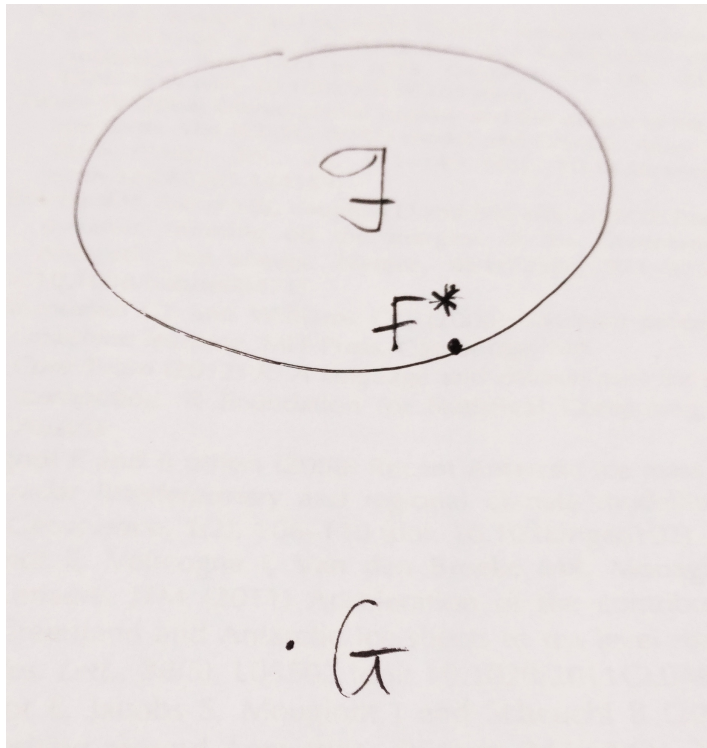
$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\Longrightarrow} N(0, \mathcal{I}^{-1}(\theta_0))$$

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

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

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

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



Bayes

Bayesian posterior

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

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

Bayes

Bayesian posterior

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

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

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

Bernstein-von Mises theorem[1]: we forget the prior, and get asymptotic concentration and normality.

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

Bayes

Bayesian posterior

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

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

$$\pi(\theta|y_{1:n}) \overset{d}{\Longrightarrow} N(\theta_0, \mathcal{I}^{-1}(\theta_0)) \text{ as } \quad n \to \infty$$

Bernstein-von Mises theorem[1]: we forget the prior, and get asymptotic concentration and normality.

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

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

Bayes

Bayesian posterior

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

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

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

Bernstein-von Mises theorem[1]: we forget the prior, and get asymptotic concentration and normality.

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

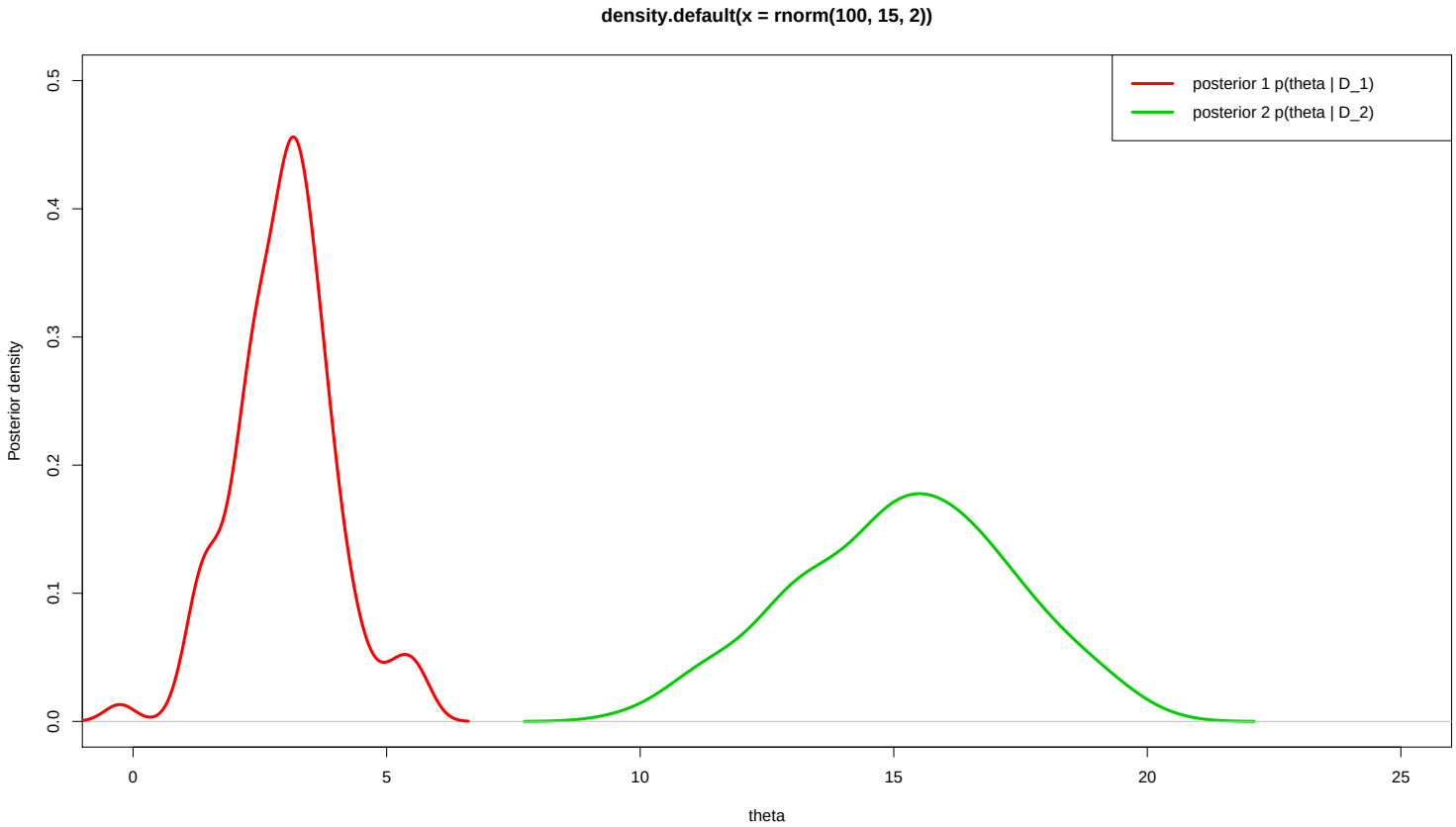
there is no obvious meaning for Bayesian analysis in this case

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

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

Different dataset, different posterior

This discrepancy often materializes when we fit a model to two different datasets.



For example, when estimating climate sensitivity from modern vs palaeo or Antarctic vs African data etc.

Inferential aims

How should we proceed when we have a misspecified model?

Inferential aims

How should we proceed when we have a misspecified model?

The approach should depend upon our goals:

- Calibrated prediction:

$$\pi(y' \mid y) = \int F_\theta(y')\pi(\theta \mid y)\mathrm{d}\theta$$

- Inference:

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

- Decision making

If the aim is prediction, perhaps we only need to statistically 'correct' the model....?

Option 1: Fix the model

- Assume the simulator is perfect
- Fit the simulator to data
- Look to falsify the simulator, and then improve it.

Option 1: Fix the model

- Assume the simulator is perfect
- Fit the simulator to data
- Look to falsify the simulator, and then improve it.

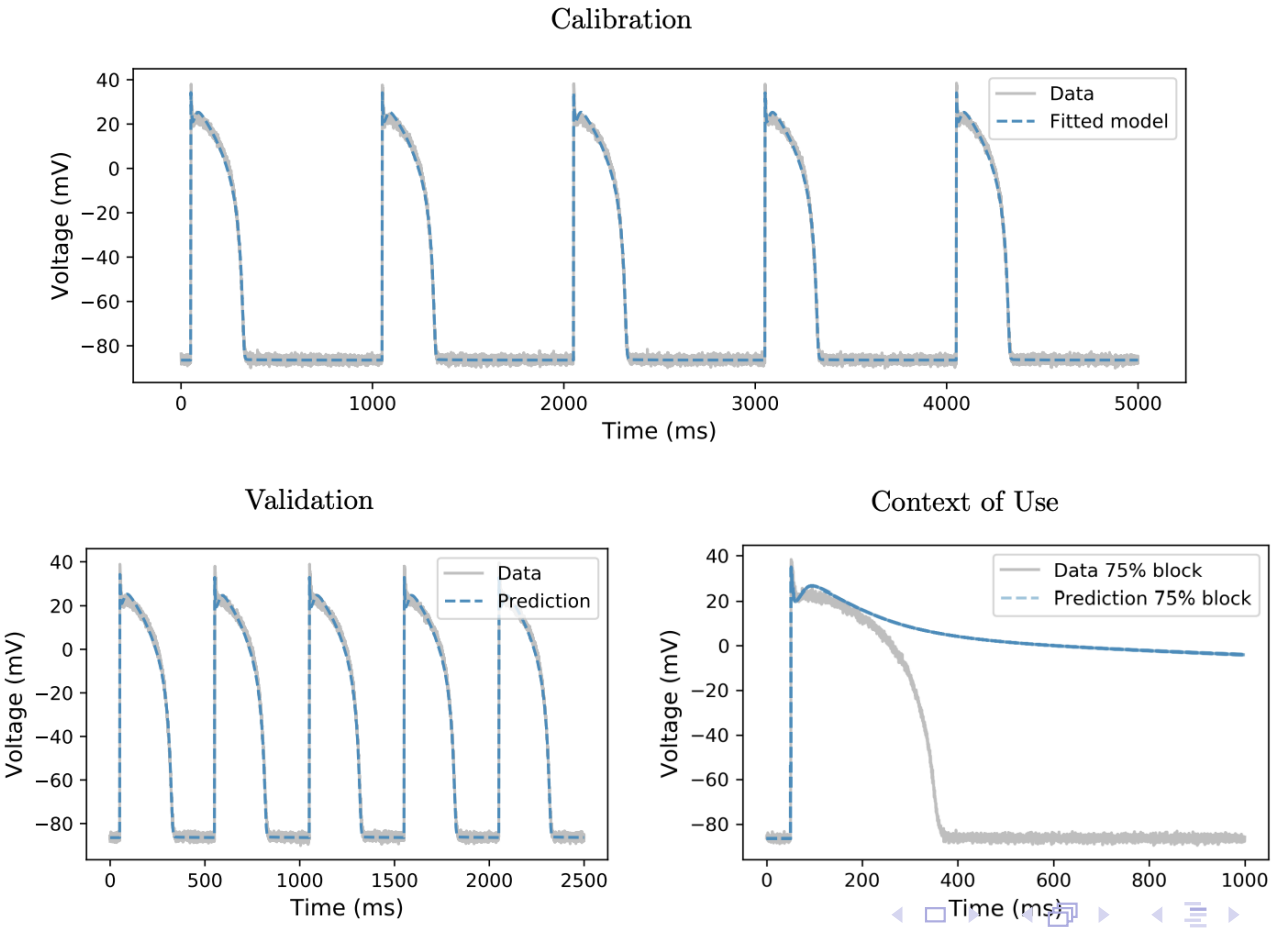
Problems with this approach include

- hard to improve simulator
- hard to spot what is wrong with a simulator even when we know it is misspecified
- In flexible models, errors can cancel to produce excellent fits to data.

Ion channel model, Lei et al. 2020

Two models of the action potential in human ventricular cells

- Generate data from model T, fit modified model F to the data.
- Test on a held-out validation set (double pacing frequency).
- Predict under an inhibited hERG channel (e.g. a common side effect and antitarget during drug development)

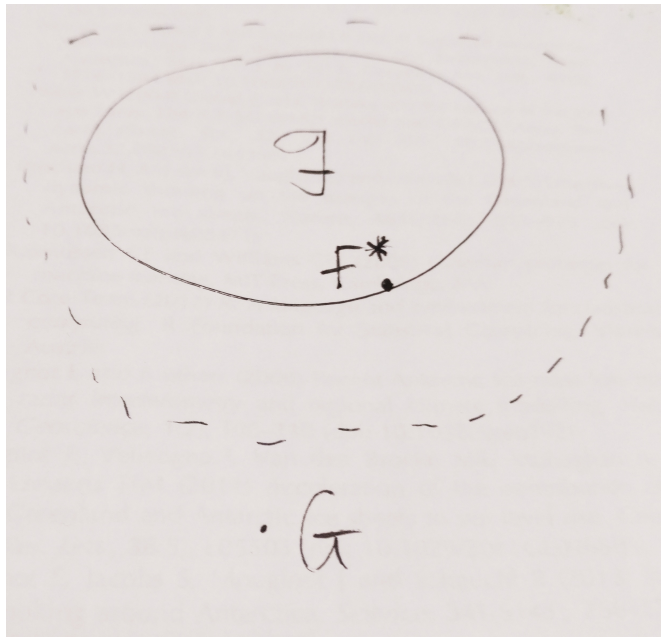


Option 2: probabilistic model of the discrepancy

Kennedy and O'Hagan 2001

Can we model our way out of trouble by expanding $\mathcal{F}$ into a non-parametric world?

- Grey-box models



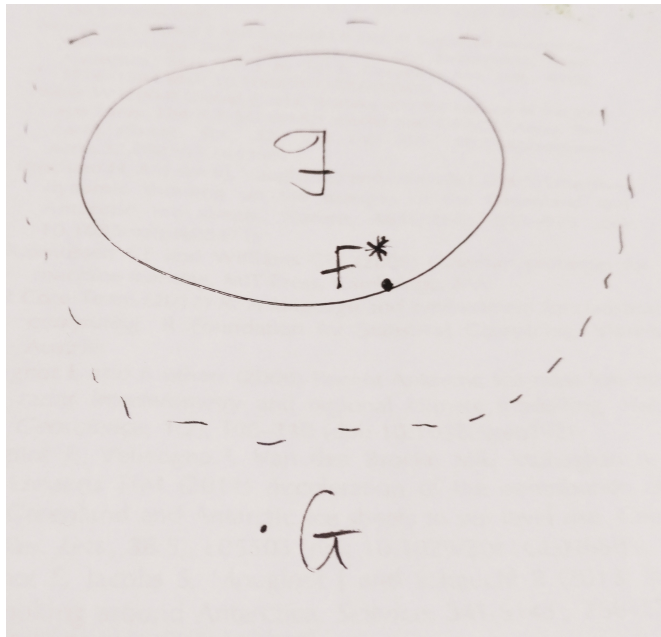
Option 2: probabilistic model of the discrepancy

Kennedy and O'Hagan 2001

Can we model our way out of trouble by expanding $\mathcal{F}$ into a non-parametric world?

- Grey-box models

One way to expand the class of models is by adding a Gaussian process (GP) to the simulator.



If $f_\theta(x)$ is our simulator, y the observation, then perhaps we can correct f using the model

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

and jointly infer θ^* and $\delta(\cdot)$

An appealing, but flawed, idea

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

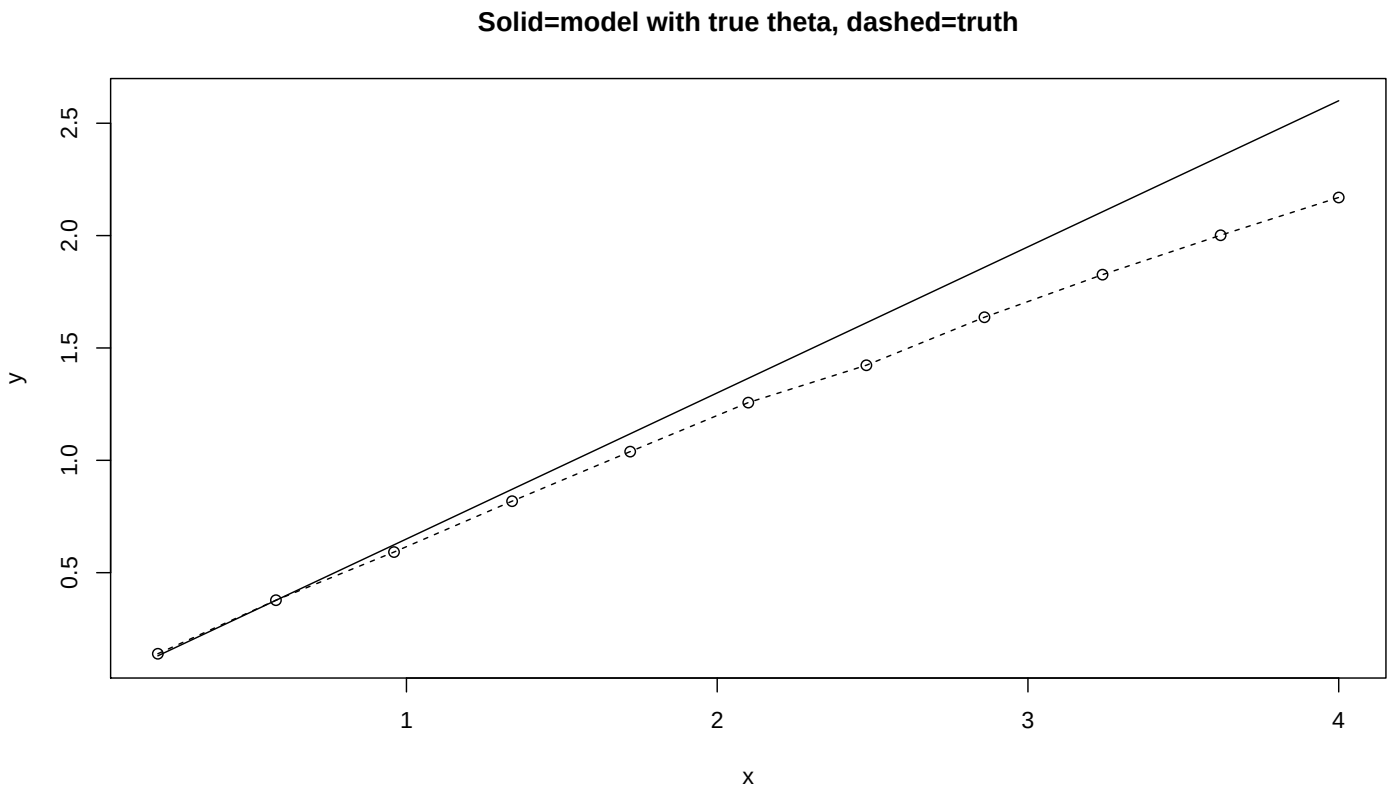
Simulator

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

Reality

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

x is a control input



An appealing, but flawed, idea

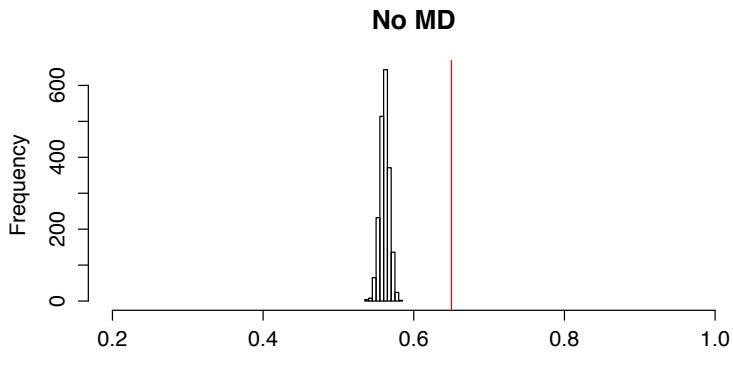
Bolting on a GP can correct your predictions[2], but won't necessarily fix your inference:

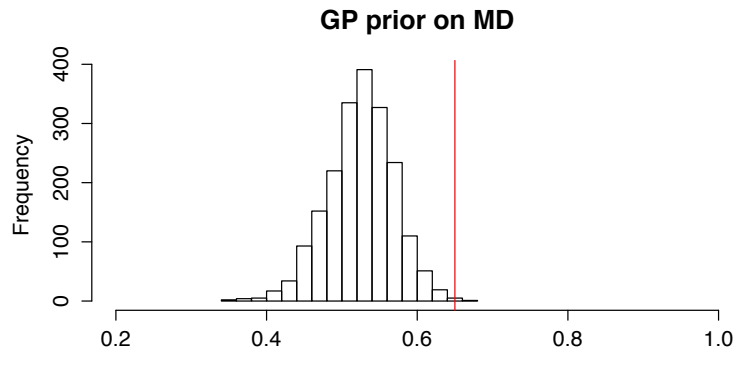
- No discrepancy:

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

- GP discrepancy:

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





[2] as long as you are not extrapolating

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
  - GPs are complex infinite dimensional models, & are 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 GP samples...

Dangers of non-parametric model extensions

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

- We may still find $G \notin \mathcal{F}$
- Identifiability
  - GPs are complex infinite dimensional models, & are 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 GP samples...
  - Brynjarsdottir and O'Hagan 2014 try to model their way out of trouble with prior information:

$$\delta(0) = 0 \qquad \delta'(x) \geq 0.$$

    Great if you have this information.

Option 3: Minimal specification of the discrepancy

Cf: Bissiri et al. 2016

Can we get away with a less complete description of the discrepancy, i.e., not a fully probabilistic model?

Option 3: Minimal specification of the discrepancy

Cf: Bissiri et al. 2016

Can we get away with a less complete description of the discrepancy, i.e., not a fully probabilistic model?

Rejection ABC

- Draw θ from $\pi(\theta)$
- Simulate $Y' \sim F_\theta(\cdot)$
- Accept θ if $\rho(Y, Y') \leq \epsilon$

ABC as a probability model

W. 2013

We wanted to solve the inverse problem

$$Y = f(\theta)$$

but instead ABC solves

$$Y = f(\theta) + e.$$

ABC as a probability model

W. 2013

We wanted to solve the inverse problem

$$Y = f(\theta)$$

but instead ABC solves

$$Y = f(\theta) + e.$$

ABC gives 'exact' inference under a different model!

Proposition

If $\rho(Y, Y') = |Y - Y'|$, then ABC samples from the posterior distribution $\pi(\theta|Y)$ where we assume $Y = f(\theta) + e$ and that

$$e \sim U[-\epsilon, \epsilon]$$

ABC as a probability model

W. 2013

We wanted to solve the inverse problem

$$Y = f(\theta)$$

but instead ABC solves

$$Y = f(\theta) + e.$$

ABC gives 'exact' inference under a different model!

Proposition

If $\rho(Y, Y') = |Y - Y'|$, then ABC samples from the posterior distribution $\pi(\theta|Y)$ where we assume $Y = f(\theta) + e$ and that

$$e \sim U[-\epsilon, \epsilon]$$

Further, we can generallze from a uniform distribution to any other distribution by introducing a random acceptance in the final ABC step.

History matching and ABC

Craig et al. 1999, 2001

History matching is an inferential approach designed for misspecified models. It seeks to find a (not ruled out yet) 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

Craig et al. 1999, 2001

History matching is an inferential approach designed for misspecified models. It seeks to find a (not ruled out yet) 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

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

for some choice of S and ϵ.

- Typically $S(\hat{F}_\theta, y) = \rho(\eta(y), \eta(y'))$ where $y' \sim F_\theta$

History matching and ABC

Craig et al. 1999, 2001

History matching is an inferential approach designed for misspecified models. It seeks to find a (not ruled out yet) 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

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

for some choice of S and ϵ.

- Typically $S(\hat{F}_\theta, y) = \rho(\eta(y), \eta(y'))$ where $y' \sim F_\theta$

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?

- 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
- Potentially use generalised scores/loss-functions
  - Potentially results in a form of robustness
- The thresholding type nature potentially makes them somewhat conservative and ensures we don't get asymptotic concentration.
  - Allow for crude/simple discrepancy characterization.

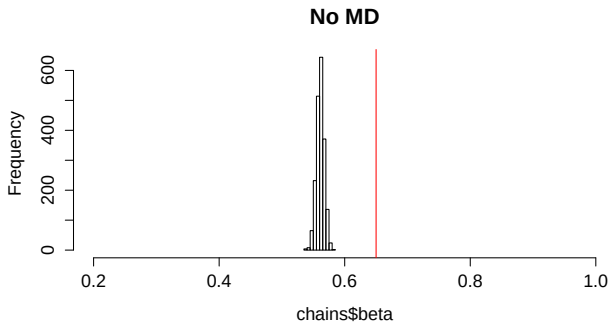
Brynjarsdottir *et al.* revisited

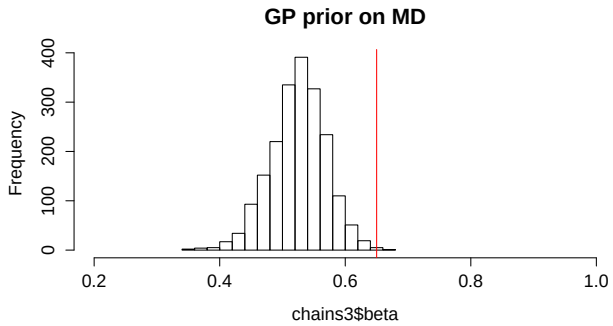
Simulator

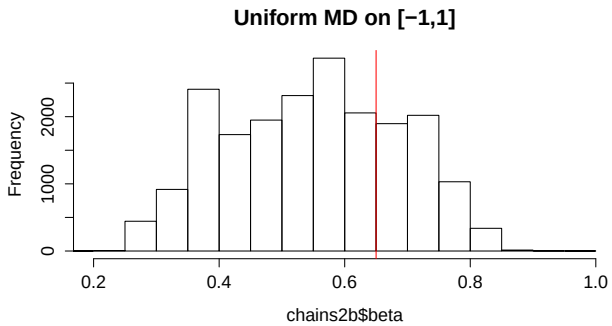
$$f_\theta(x) = \theta x$$

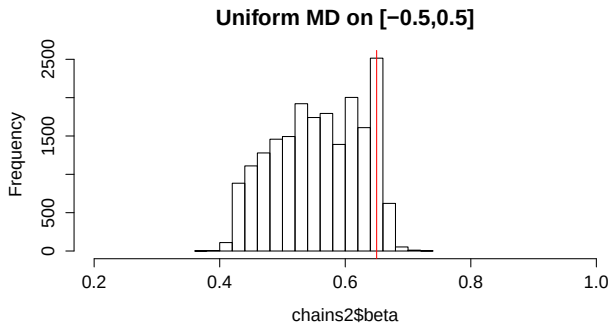
Reality

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









Option 4: changes to the inference process

Variational formulation of Bayesian inference

We can view Bayes inference as an optimization problem.
Write the log-likelihood as $\ell(\theta, y)$. Then

$$\pi(\theta|y) = \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{\theta \sim q} \ell(\theta, y) + D_{KL}(q||\pi)$$

where $\mathcal{Q}$ the set of all probability measures on Θ, and π is the prior for θ.

Option 4: changes to the inference process

Variational formulation of Bayesian inference

We can view Bayes inference as an optimization problem.
Write the log-likelihood as $\ell(\theta, y)$. Then

$$\pi(\theta|y) = \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{\theta \sim q} \ell(\theta, y) + D_{KL}(q||\pi)$$

where $\mathcal{Q}$ the set of all probability measures on Θ, and π is the prior for θ.

Proof:

$$\begin{aligned} RHS &= \arg\min \int \left(\log \exp(\ell(\theta, y)) + \log \frac{q(\theta)}{\pi(\theta)} \right) q(\theta) d\theta \\ &= \arg\min \int \log \left(\frac{q(\theta)}{\exp(-\ell(\theta, y))\pi(\theta)} \right) q(\theta) d\theta \\ &= \arg\min \int \log \frac{q(\theta)}{\pi(\theta|x)} q(\theta) d\theta \\ &= \arg\min KLD(q||\pi(\theta|y)) \end{aligned}$$

which is achieved uniquely at $q = \pi(\theta|y) = q_B(\theta)$ as required.

Generalized Variational Inference

Knoblauch *et al.* 2019

$$\pi(\theta|y) = \arg\min_{q \in \Pi} \mathbb{E}_{\theta \sim q} \ell(\theta, y) + D(q||\pi)$$

By changing

- the loss $\ell(\theta, y)$
- the prior-posterior divergence $D(q||\pi)$
- restricting $q \in \Pi \subset \mathcal{Q}$

we generate other learning algorithms that are Bayesian in flavour but which may have *nice* properties - "GVI"

Generalized Variational Inference

Knoblauch *et al.* 2019

$$\pi(\theta|y) = \arg\min_{q \in \Pi} \mathbb{E}_{\theta \sim q} \ell(\theta, y) + D(q||\pi)$$

By changing

- the loss $\ell(\theta, y)$
- the prior-posterior divergence $D(q||\pi)$
- restricting $q \in \Pi \subset \mathcal{Q}$

we generate other learning algorithms that are Bayesian in flavour but which may have *nice* properties - "GVI"

Method	$\ell(\boldsymbol{\theta}, x_i)$	D	Π
Standard Bayes	$-\log p(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Power Likelihood Bayes[1]	$-\log p(x_i\|\boldsymbol{\theta})$	$\frac{1}{w}$KLD, $w < 1$	$\mathcal{P}(\boldsymbol{\Theta})$
Composite Likelihood Bayes[2]	$-w_i \log p(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Divergence-based Bayes[3]	divergence-based ℓ	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
PAC/Gibbs Bayes[4]	any ℓ	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
VAE[5,†]	$-\log p_{\boldsymbol{\zeta}}(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{Q}$
β-VAE[6,†]	$-\log p_{\boldsymbol{\zeta}}(x_i\|\boldsymbol{\theta})$	$\beta \cdot$ KLD, $\beta > 1$	$\mathcal{Q}$
Bernoulli-VAE[7,†]	continuous Bernoulli	KLD	$\mathcal{Q}$
Standard VI	$-\log p(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{Q}$
Power VI[8]	$-\log p(x_i\|\boldsymbol{\theta})$	$\frac{1}{w}$KLD, $w < 1$	$\mathcal{Q}$
Utility VI[9]	$-\log p(x_i\|\boldsymbol{\theta}) + \log u(h, x_i)$	KLD	$\mathcal{Q}$
Regularized Bayes[10]	$-\log p(x_i\|\boldsymbol{\theta}) + \phi(\boldsymbol{\theta}, x_i)$	KLD	$\mathcal{Q}$
Gibbs VI[11]	any ℓ	KLD	$\mathcal{Q}$
Generalized VI	any ℓ	any D	$\mathcal{Q}$

History matching as GVI

HM and ABC can be written as GVI approaches.

E.g., Define the history matching loss to be

$$\ell_{HM}(\theta, y) = -\log \mathbb{I}_{S(\theta,y)<c}$$
$$= \begin{cases} 0 & \text{if } S(\theta, y) < c \\ \infty & \text{otherwise} \end{cases}$$

Denote the not ruled out yet (NROY) set by

$$\mathcal{N} = \{\theta \in \Theta : S(\theta, y) < c\}.$$

History matching as GVI

HM and ABC can be written as GVI approaches.

E.g., Define the history matching loss to be

$$\ell_{HM}(\theta, y) = -\log \mathbb{I}_{S(\theta,y)<c}$$
$$= \begin{cases} 0 & \text{if } S(\theta, y) < c \\ \infty & \text{otherwise} \end{cases}$$

Denote the not ruled out yet (NROY) set by

$$\mathcal{N} = \{\theta \in \Theta : S(\theta, y) < c\}.$$

If we use the KL divergence and set $\Pi = \mathcal{Q}$, then

$$q_{HM}(\theta) = \frac{\pi(\theta)\mathbb{I}_{\theta\in\mathcal{N}}}{\int_{\mathcal{N}} \pi(\theta)d\theta}$$

is the solution to the GVI problem.

Variational inference: computation

Cf Eliane's and Shiran's talks

Typically we'll restrict the family of target posteriors to $\Pi \subset \mathcal{Q}$, e.g.,

$$q_\psi(\theta) = N(\theta; \psi_1, \psi_2)$$

VI then involves optimizing for ψ.

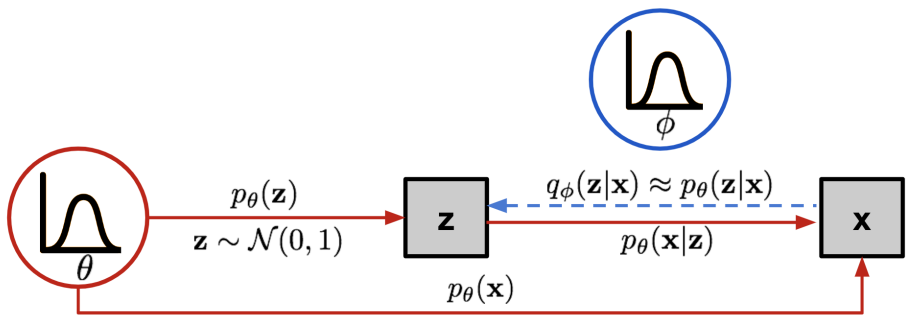
Variational inference: computation

Cf Eliane's and Shiran's talks

Typically we'll restrict the family of target posteriors to $\Pi \subset \mathcal{Q}$, e.g.,

$$q_\psi(\theta) = N(\theta; \psi_1, \psi_2)$$

VI then involves optimizing for ψ. There are many blackbox inference approaches, e.g., VAE (Kingma and Welling 2013).



- The approximate posterior $q_\psi(\theta)$ used as the encoder
- The simulator $f(\theta, U)$ used as the decoder.

Requires us to write the simulator in an autodiff language and to explicitly control the random variables U.

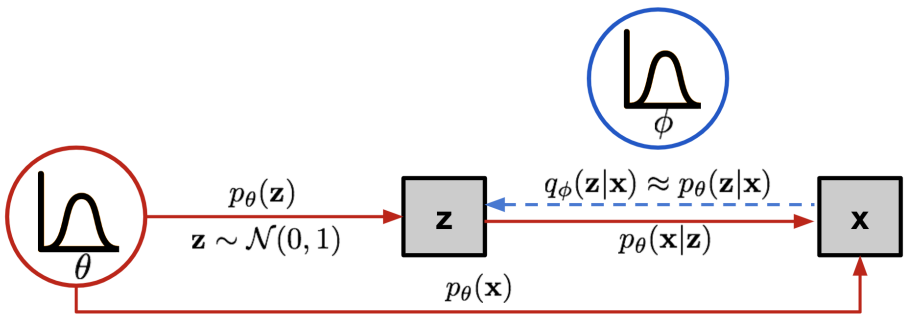
Variational inference: computation

Cf Eliane's and Shiran's talks

Typically we'll restrict the family of target posteriors to $\Pi \subset \mathcal{Q}$, e.g.,

$$q_\psi(\theta) = N(\theta; \psi_1, \psi_2)$$

VI then involves optimizing for ψ. There are many blackbox inference approaches, e.g., VAE (Kingma and Welling 2013).



- The approximate posterior $q_\psi(\theta)$ used as the encoder
- The simulator $f(\theta, U)$ used as the decoder.

Requires us to write the simulator in an autodiff language and to explicitly control the random variables U.

Trying to develop an amortized version $q_\psi(\theta|y) = N(\theta; \psi_1(y), \psi_2(y))$ for in-procedure calibration.

Discussion: What makes a good inferential approach?

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

Discussion: What makes a good inferential approach?

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

- Consistency?

Discussion: What makes a good inferential approach?

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.

Discussion: What makes a good inferential approach?

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?

Discussion: What makes a good inferential approach?

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?~~

Discussion: What makes a good inferential approach?

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?

Discussion: What makes a good inferential approach?

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.

Discussion: What makes a good inferential approach?

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?

Discussion: What makes a good inferential approach?

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?

Conclusions

Given a useful misspecified simulator, how should make good inferences, and what justifiable(?) short-cuts can we take to accelerate inference?

- Summaries - only use simulator outputs we trust.
- Include a (potentially crude) discrepancy model
- Abandon complex likelihood functions to avoid extreme sensitivity. Choose loss functions that are robust to misspecification of the discrepancy model!
- Thresholding of the score for crude UQ?

Conclusions

Given a useful misspecified simulator, how should make good inferences, and what justifiable(?) short-cuts can we take to accelerate inference?

- Summaries - only use simulator outputs we trust.
- Include a (potentially crude) discrepancy model
- Abandon complex likelihood functions to avoid extreme sensitivity. Choose loss functions that are robust to misspecification of the discrepancy model!
- Thresholding of the score for crude UQ?

Key problems:

- How do we relate the level of simulator discrepancy to the decision about the posterior approximation accuracy required?
- How do we learn the simulator discrepancy?
- 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?

Thank you for listening!