# Bayesian Inference by Density Ratio Estimation

Michael Gutmann

https://sites.google.com/site/michaelgutmann

Institute for Adaptive and Neural Computation
School of Informatics, University of Edinburgh

20th June 2017

# Task

Perform Bayesian inference for models where

1. the likelihood function is too costly to compute
2. sampling – simulating data – from the model is possible

# Program

Background

Previous work

Proposed approach

# Program

Background

Previous work

Proposed approach

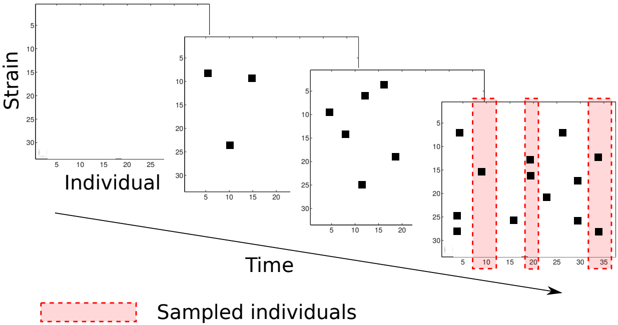
# Simulator-based models

- Goal: Inference for models that are specified by a mechanism for generating data
  - e.g. stochastic dynamical systems
  - e.g. computer models / simulators of some complex physical or biological process
- Such models occur in multiple and diverse scientific fields.
- Different communities use different names:
  - Simulator-based models
  - Stochastic simulation models
  - Implicit models
  - Generative (latent-variable) models
  - Probabilistic programs

# Examples

Simulator-based models are widely used:

- Evolutionary biology:
  Simulating evolution
- Neuroscience:
  Simulating neural circuits
- Ecology:
  Simulating species migration
- Health science:
  Simulating the spread of an infectious disease

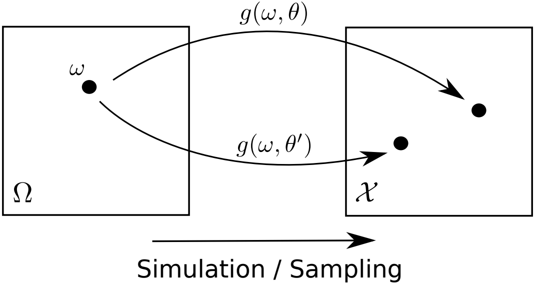


# Definition of simulator-based models

- Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space.
- A simulator-based model is a collection of (measurable) functions $g(., \boldsymbol{\theta})$ parametrised by $\boldsymbol{\theta}$,

$$\boldsymbol{\omega} \in \Omega \mapsto \boldsymbol{x}_{\boldsymbol{\theta}} = g(\boldsymbol{\omega}, \boldsymbol{\theta}) \in \mathcal{X} \tag{1}$$

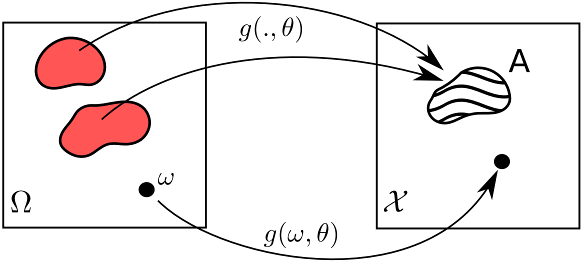
- For any fixed $\boldsymbol{\theta}$, $\boldsymbol{x}_{\boldsymbol{\theta}} = g(., \boldsymbol{\theta})$ is a random variable.



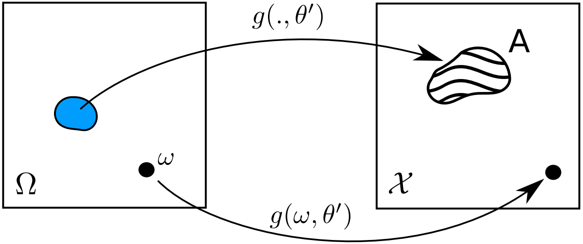
# Implicit definition of the model pdfs

$$\Pr(x \in A \mid \theta) = \mathcal{P}(\{\omega : g(\omega, \theta) \in A\})$$

Parameter value $\theta$




Parameter value $\theta'$



# Advantages of simulator-based models

- Direct implementation of hypotheses of how the observed data were generated.
- Neat interface with scientific models (e.g. from physics or biology).
- Modelling by replicating the mechanisms of nature that produced the observed/measured data. ("Analysis by synthesis")
- Possibility to perform experiments in silico.

# Disadvantages of simulator-based models

- Generally elude analytical treatment.
- Can be easily made more complicated than necessary.
- Statistical inference is difficult.

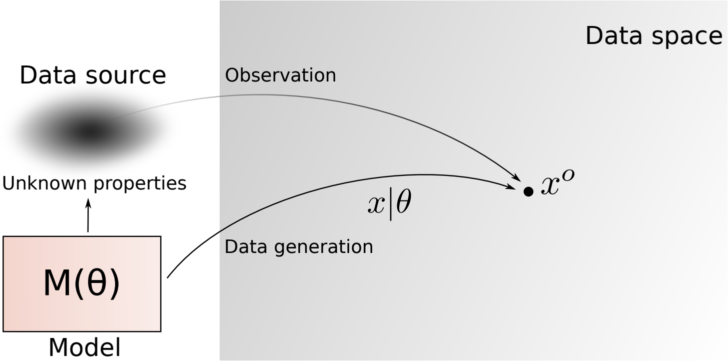
# Disadvantages of simulator-based models

- Generally elude analytical treatment.
- Can be easily made more complicated than necessary.
- Statistical inference is difficult.

  Main reason: *Likelihood function is intractable*

# The likelihood function $L(\boldsymbol{\theta})$

- Probability that the model generates data like $\boldsymbol{x}^o$ when using parameter value $\boldsymbol{\theta}$
- Generally well defined but intractable for simulator-based / implicit models



# Program

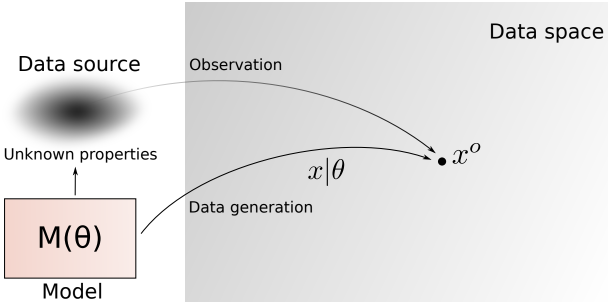
Background

Previous work

Proposed approach

# Three foundational issues

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
2. How should we compute the probability of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
3. For which values of $\boldsymbol{\theta}$ should we compute it?




Likelihood: Probability that the model generates data like $\boldsymbol{x}^o$ for parameter value $\boldsymbol{\theta}$

# Approximate Bayesian computation

1. How should we assess whether $\mathbf{x}_{\boldsymbol{\theta}} \equiv \mathbf{x}^o$?
   $\Rightarrow$ Check whether $||T(\mathbf{x}_{\boldsymbol{\theta}}) - T(\mathbf{x}^o)|| \leq \epsilon$
2. How should we compute the proba of the event $\mathbf{x}_{\boldsymbol{\theta}} \equiv \mathbf{x}^o$?
   $\Rightarrow$ By counting
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   $\Rightarrow$ Sample from the prior (or other proposal distributions)

# Approximate Bayesian computation

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ Check whether $||T(\boldsymbol{x}_{\boldsymbol{\theta}}) - T(\boldsymbol{x}^o)|| \leq \epsilon$
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   $\Rightarrow$ By counting
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   $\Rightarrow$ Sample from the prior (or other proposal distributions)

Difficulties:

- Choice of $T()$ and $\epsilon$
- Typically high computational cost

For recent review, see: Lintusaari et al (2017) "Fundamentals and recent developments in approximate Bayesian computation", Systematic Biology

# Synthetic likelihood

(Simon Wood, Nature, 2010)

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   - $\Rightarrow$ Compute summary statistics $\boldsymbol{t}_{\boldsymbol{\theta}} = T(\boldsymbol{x}_{\boldsymbol{\theta}})$
   - $\Rightarrow$ Model their distribution as a Gaussian
   - $\Rightarrow$ Compute likelihood function with $T(\boldsymbol{x}^o)$ as observed data
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   - $\Rightarrow$ Use obtained "synthetic" likelihood function as part of a Monte Carlo method

# Synthetic likelihood

(Simon Wood, Nature, 2010)

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
   - ⇒ Compute summary statistics $\boldsymbol{t}_{\boldsymbol{\theta}} = T(\boldsymbol{x}_{\boldsymbol{\theta}})$
   - ⇒ Model their distribution as a Gaussian
   - ⇒ Compute likelihood function with $T(\boldsymbol{x}^o)$ as observed data
3. For which values of $\boldsymbol{\theta}$ should we compute it?
   - ⇒ Use obtained "synthetic" likelihood function as part of a Monte Carlo method

Difficulties:

- Choice of $T()$
- Gaussianity assumption may not hold
- Typically high computational cost

# Program

Background

Previous work

Proposed approach

# Overview of some of my work

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?

   $\Rightarrow$ Use classification (Gutmann et al, 2014, 2017)

2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
3. For which values of $\boldsymbol{\theta}$ should we compute it?

   $\Rightarrow$ Use Bayesian optimisation (Gutmann and Corander, 2013-2016)
   Compared to standard approaches: speed-up by a factor of 1000 more

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?

   $\Rightarrow$ Use density ratio estimation (Dutta et al, 2016, arXiv:1611.10242)

# Overview of some of my work

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?

   $\Rightarrow$ Use classification (Gutmann et al, 2014, 2017)

2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
3. For which values of $\boldsymbol{\theta}$ should we compute it?

   $\Rightarrow$ Use Bayesian optimisation (Gutmann and Corander, 2013-2016)
   Compared to standard approaches: speed-up by a factor of 1000 more

1. How should we assess whether $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?
2. How should we compute the proba of the event $\boldsymbol{x}_{\boldsymbol{\theta}} \equiv \boldsymbol{x}^o$?

   $\Rightarrow$ Use density ratio estimation (Dutta et al, 2016, arXiv:1611.10242)

# Basic idea

(Dutta et al, 2016, arXiv:1611.10242)

- Frame posterior estimation as ratio estimation problem

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})}{p(\mathbf{x})} = p(\boldsymbol{\theta})r(\mathbf{x},\boldsymbol{\theta}) \tag{2}$$

$$r(\mathbf{x},\boldsymbol{\theta}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})}{p(\mathbf{x})} \tag{3}$$

- Estimating $r(\mathbf{x},\boldsymbol{\theta})$ is the difficult part since $p(\mathbf{x}|\boldsymbol{\theta})$ unknown.
- Estimate $\hat{r}(\mathbf{x},\boldsymbol{\theta})$ yields estimate of the likelihood function and posterior

$$\hat{L}(\boldsymbol{\theta}) \propto \hat{r}(\mathbf{x}^o,\boldsymbol{\theta}), \qquad \hat{p}(\boldsymbol{\theta}|\mathbf{x}^o) = p(\boldsymbol{\theta})\hat{r}(\mathbf{x}^o,\boldsymbol{\theta}). \tag{4}$$

# Estimating density ratios in general

- Relatively well studied problem (Textbook by Sugiyama et al, 2012)
- Bregman divergence provides general framework
  (Gutmann and Hirayama, 2011; Sugiyama et al, 2011)
- Here: density ratio estimation by logistic regression

# Density ratio estimation by logistic regression

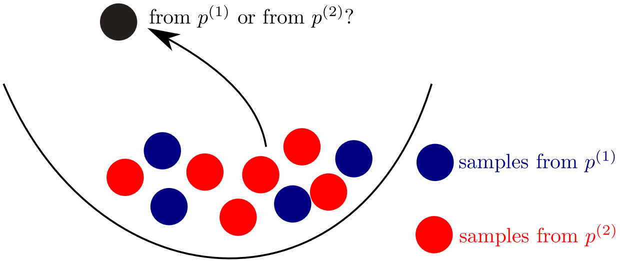
- Samples from two data sets

$$\boldsymbol{x}_i^{(1)} \sim p^{(1)}, \quad i = 1, \ldots, n^{(1)} \tag{5}$$

$$\boldsymbol{x}_i^{(2)} \sim p^{(2)}, \quad i = 1, \ldots, n^{(2)} \tag{6}$$

- Probability that a test data point $\boldsymbol{x}$ was sampled from $p^{(1)}$

$$\mathbb{P}(\boldsymbol{x} \sim p^{(1)} | \boldsymbol{x}, h) = \frac{1}{1 + \nu \exp(-h(\boldsymbol{x}))}, \qquad \nu = \frac{n^{(2)}}{n^{(1)}} \tag{7}$$



# Density ratio estimation by logistic regression

- Estimate $h$ by minimising

$$\mathcal{J}(h) = \frac{1}{n}\left\{\sum_{i=1}^{n^{(1)}} \log\left[1 + \nu \exp\left(-h_i^{(1)}\right)\right] + \sum_{i=1}^{n^{(2)}} \log\left[1 + \frac{1}{\nu}\exp\left(h_i^{(2)}\right)\right]\right\}$$

$$h_i^{(1)} = h\left(\boldsymbol{x}_i^{(1)}\right) \qquad h_i^{(2)} = h\left(\boldsymbol{x}_i^{(2)}\right)$$

$$n = n^{(1)} + n^{(2)}$$

- Objective is the re-scaled negated log-likelihood.
- For large $n^{(1)}$ and $n^{(2)}$

$$\hat{h} = \mathrm{argmin}_h\, \mathcal{J}(h) = \log \frac{p^{(1)}}{p^{(2)}}$$

without any constraints on $h$

# Estimating the posterior

- Property was used to estimate unnormalised models
  (Gutmann & Hyvärinen, 2010, 2012)
- It was used to estimate likelihood ratios
  (Pham et al, 2014; Cranmer et al, 2015)
- For posterior estimation, we use
  - data generating pdf $p(\boldsymbol{x}|\boldsymbol{\theta})$ for $p^{(1)}$
  - marginal $p(\boldsymbol{x})$ for $p^{(2)}$ (Other choices for $p(\boldsymbol{x})$ possible too)
  - sample sizes entirely under our control

# Estimating the posterior

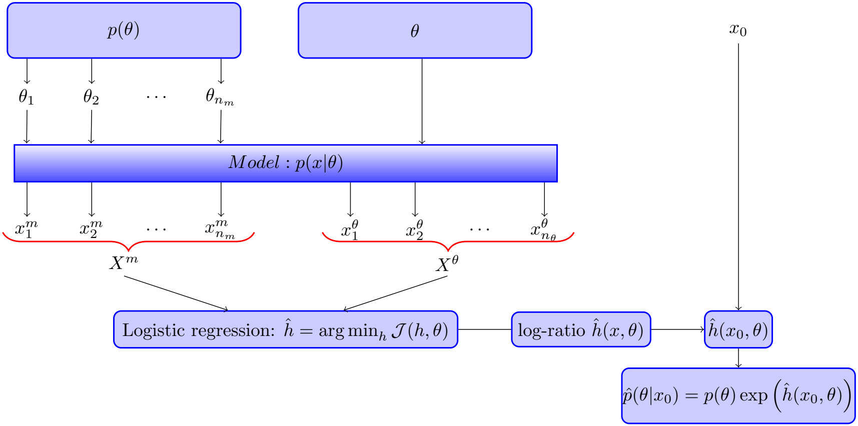
- Logistic regression gives (point-wise in $\boldsymbol{\theta}$)

$$\hat{h}(\mathbf{x}, \boldsymbol{\theta}) \to \log \frac{p(\mathbf{x}|\boldsymbol{\theta})}{p(\mathbf{x})} = \log r(\mathbf{x}, \boldsymbol{\theta}) \tag{8}$$

- Estimated posterior and likelihood function:

$$\hat{p}(\boldsymbol{\theta}|\mathbf{x}^o) = p(\boldsymbol{\theta}) \exp(\hat{h}(\mathbf{x}^o, \boldsymbol{\theta})) \quad \hat{L}(\boldsymbol{\theta}) \propto \exp(\hat{h}(\mathbf{x}^o, \boldsymbol{\theta})) \tag{9}$$

# Estimating the posterior




(Dutta et al, 2016, arXiv:1611.10242)

# Auxiliary model

- We need to specify a model for $h$.
- For simplicity: linear model

$$h(\boldsymbol{x}) = \sum_{i=1}^{b} \beta_i \psi_i(\boldsymbol{x}) = \beta^\top \psi(\boldsymbol{x}) \tag{10}$$

where $\psi_i(\boldsymbol{x})$ are summary statistics

- More complex models possible

# Exponential family approximation

- Logistic regression yields

$$\hat{h}(\boldsymbol{x};\boldsymbol{\theta}) = \hat{\beta}(\boldsymbol{\theta})^\top \psi(\boldsymbol{x}), \quad \hat{r}(\boldsymbol{x},\boldsymbol{\theta}) = \exp(\hat{\beta}(\boldsymbol{\theta})^\top \psi(\boldsymbol{x})) \tag{11}$$

- Resulting posterior

$$\hat{p}(\boldsymbol{\theta}|\boldsymbol{x}^o) = p(\boldsymbol{\theta}) \exp(\hat{\beta}(\boldsymbol{\theta})^\top \psi(\boldsymbol{x}^o)) \tag{12}$$

- Implicit exponential family approximation of $p(\boldsymbol{x}|\boldsymbol{\theta})$

$$\hat{r}(\boldsymbol{x},\boldsymbol{\theta}) = \frac{\hat{p}(\boldsymbol{x}|\boldsymbol{\theta})}{\hat{p}(\boldsymbol{x})} \tag{13}$$

$$\hat{p}(\boldsymbol{x}|\boldsymbol{\theta}) = \hat{p}(\boldsymbol{x}) \exp(\hat{\beta}(\boldsymbol{\theta})^\top \psi(\boldsymbol{x})) \tag{14}$$

- Implicit because $\hat{p}(\boldsymbol{x})$ never explicitly constructed.

# Remarks

- Vector of summary statistics $\psi(\boldsymbol{x})$ should include a constant for normalisation of the pdf (log partition function)
- Normalising constant is estimated via the logistic regression
- Simple linear model leads to a generalisation of synthetic likelihood
- $L_1$ penalty on $\beta$ for weighing and selecting summary statistics

# Application to ARCH model

- Model:

$$x^{(t)} = \theta_1 x^{(t-1)} + e^{(t)} \tag{15}$$

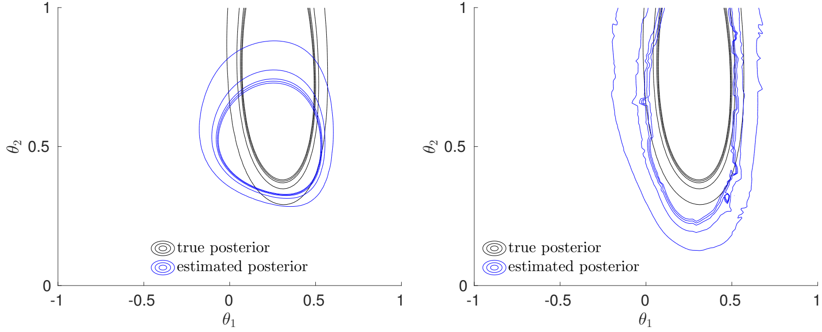
$$e^{(t)} = \xi^{(t)} \sqrt{0.2 + \theta_2 (e^{(t-1)})^2} \tag{16}$$

  $\xi^{(t)}$ and $e^{(0)}$ independent standard normal r.v., $x^{(0)} = 0$
- 100 time points
- Parameters: $\theta_1 \in (-1, 1), \quad \theta_2 \in (0, 1)$
- Uniform prior on $\theta_1, \theta_2$

# Application to ARCH model

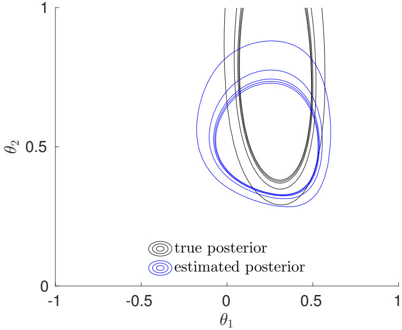
- Summary statistics:
  - auto-correlations with lag one to five
  - all (unique) pairwise combinations of them
  - a constant
- To check robustness: 50% irrelevant summary statistics (drawn from standard normal)
- Comparison with synthetic likelihood with equivalent set of summary statistics (relevant sum. stats. only)

# Example posterior

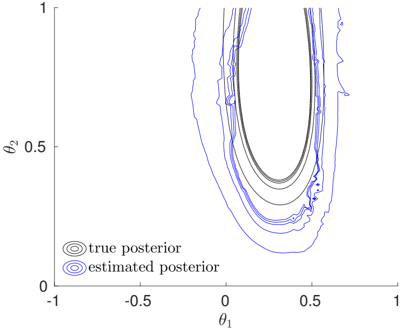



(a) synthetic likelihood (b) proposed method

# Example posterior




(c) synthetic likelihood

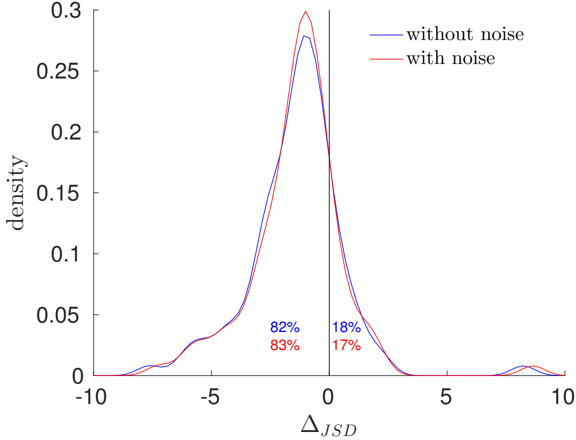



(d) proposed method subject to noise

# Systematic analysis

- Jensen-Shannon div between estimated and true posterior
- Point-wise comparison with synthetic likelihood (100 data sets)

$$\Delta_{\mathrm{JSD}} = \text{JSD for proposed method} - \text{JSD for synthetic likelihood}$$



# Application to Ricker model

- Model

$$x^{(t)}|N^{(t)}, \phi \sim \mathrm{Poisson}(\phi N^{(t)}) \tag{17}$$

$$\log N^{(t)} = \log r + \log N^{(t-1)} - N^{(t-1)} + \sigma e^{(t)} \tag{18}$$

$$t = 1, \ldots, 50, \qquad N^{(0)} = 0 \tag{19}$$

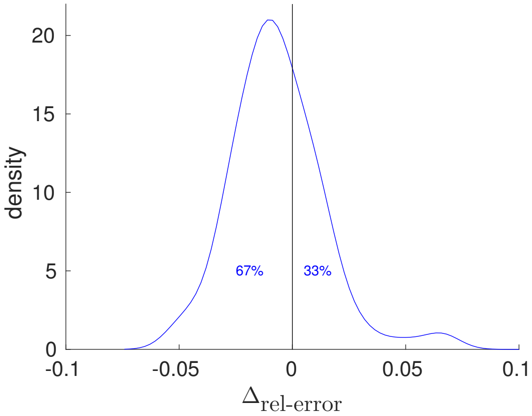
- Parameters and priors
  - log growth rate $\log r \sim \mathcal{U}(3, 5)$
  - scaling parameter $\phi \sim \mathcal{U}(5, 15)$
  - standard deviation $\sigma \sim \mathcal{U}(0, 0.6)$

# Application to Ricker model

- Summary statistics: same as Simon Wood (Nature, 2010)
- 100 inference problems
- For each problem, relative errors in posterior means were computed
- Point-wise comparison with synthetic likelihood

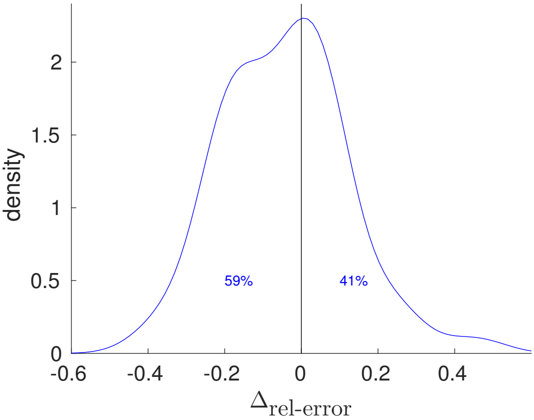
# Results for log $r$

$$\Delta_{\text{rel error}} = \text{rel error proposed method} - \text{rel error synth likelihood}$$



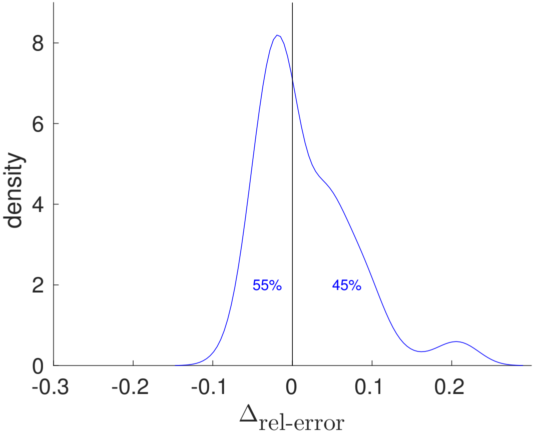
# Results for $\sigma$

$\Delta_{\text{rel error}} = \text{rel error proposed method} - \text{rel error synth likelihood}$



# Results for $\phi$

$$\Delta_{\text{rel error}} = \text{rel error proposed method} - \text{rel error synth likelihood}$$



# Observations

- Compared two auxiliary models: exponential vs Gaussian family
- For same summary statistics , typically more accurate inferences for the richer exponential family model
- Robustness to irrelevant summary statistics thanks to $L_1$ regularisation

# Conclusions

- Background and previous work on inference with simulator-based / implicit statistical models
- Our work on:
  - Framing the posterior estimation problem as a density ratio estimation problem
  - Estimating the ratio with logistic regression
  - Using regularisation to automatically select summary statistics
- Multitude of research possibilities:
  - Choice of the auxiliary model
  - Choice of the loss function used to estimate the density ratio
  - Combine with Bayesian optimisation framework to reduce computational cost

# Conclusions

- Background and previous work on inference with simulator-based / implicit statistical models
- Our work on:
  - Framing the posterior estimation problem as a density ratio estimation problem
  - Estimating the ratio with logistic regression
  - Using regularisation to automatically select summary statistics
- Multitude of research possibilities:
  - Choice of the auxiliary model
  - Choice of the loss function used to estimate the density ratio
  - Combine with Bayesian optimisation framework to reduce computational cost

More results and details in arXiv:1611.10242v1