Modelling The Model for Approximate Bayesian Computation

Michael Gutmann

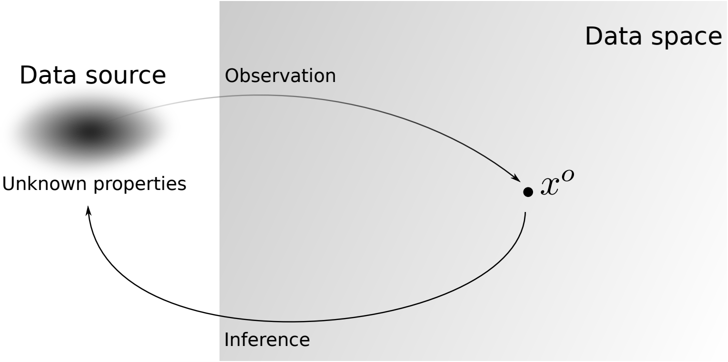
https://sites.google.com/site/michaelgutmann

University of Edinburgh

1st March 2017

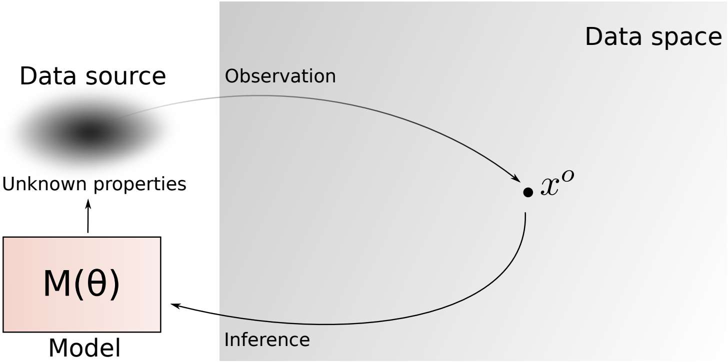
Overall goal

- Inference: Given data x^o, learn about properties of its source
- Enables decision making, predictions, ...



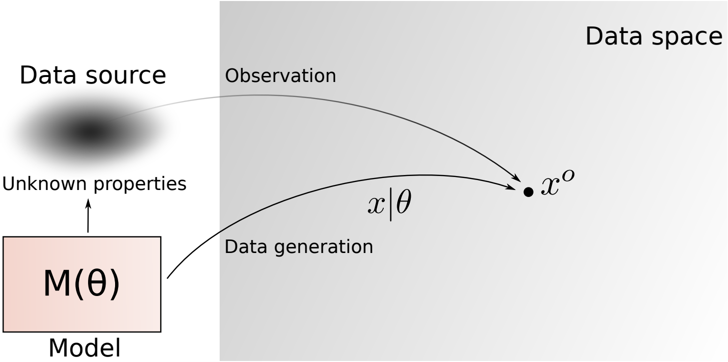
Parametric Inference

- Set up a model with potential properties θ (hypotheses)
- See which θ are in line with the observed data x^o



The likelihood function $L(\theta)$

- Measures agreement between θ and the observed data x^o
- Probability to generate data like x^o if hypothesis θ holds



Three foundational issues

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

Traditional ABC

1. How should we assess whether $x_\theta \equiv x^o$?
   $\Rightarrow$ Check whether $||T(x_\theta) - T(x^o)|| \leq \epsilon$
2. How should we compute the probability of the event $x_\theta \equiv x^o$?
   $\Rightarrow$ By counting
3. For which values of θ should we compute it?
   $\Rightarrow$ Sample from the prior

Traditional ABC

1. How should we assess whether $x_\theta \equiv x^o$?
   $\Rightarrow$ Check whether $||T(x_\theta) - T(x^o)|| \leq \epsilon$
2. How should we compute the probability of the event $x_\theta \equiv x^o$?
   $\Rightarrow$ By counting
3. For which values of θ should we compute it?
   $\Rightarrow$ Sample from the prior

Trade-off between computational cost and accuracy of the inference may be poor.

This talk

Building models of The Model is a powerful approach to address the three foundational issues and to improve the trade-off between computation and accuracy.

Models of The Model?

- "The Model" is the model of primary interest
  - We can sample sample from it
  - Likelihood function intractable
- Often a "mechanistic model" that emulates nature
- "Models of The Model" are auxiliary entities that facilitate the inference
  - Auxiliary models used in indirect inference
  - Regression models used for regression adjustment
- In what follows: "auxiliary models" instead of "models of The Model"

Program

1. Brief overview of how auxiliary models are used in ABC
2. How we used auxiliary models

Overview

Diverse use of auxiliary models (1/2)

- To define/construct summary statistics
  - indirect inference

    (e.g. Gouriéroux et al, 1993; Smith 1993; Heggland & Frigessi, 2004; Drovandi et al, 2011 & 2015)
  - semi-automatic approach by Fearnhead and Prangle, 2012
- To model the posterior $\theta|x^o$
  - linear regression adjustment by Beaumont et al, 2002
  - flexible nonlinear models

    (e.g. Blum & Francois, 2010; Papamakarios & Murray, 2016)

Diverse use of auxiliary models (2/2)

- To define a "synthetic" likelihood (Wood, 2010; Leuenberger & Wegmann, 2010)
- To reduce the number of simulations from the model
  - Surrogate models of approximate likelihoods
    (Wilkinson, 2014; Meeds & Welling, 2014)
  - Models of the discrepancy between simulated and observed data (Gutmann & Corander, 2013-2016)
- To measure the discrepancy by classification
  (Gutmann et al, 2014, 2017)

A key challenge

- We can e.g.
  - Construct summary statistics by auxiliary models of the data
  - Adjust the summary statistics by regression
  - Reduce computations by surrogate models
  - Increase accuracy by (nonlinear) regression adjustments
- Which model to use for any given purpose?
  - $\Rightarrow$ Automated model choice
  - $\Rightarrow$ Taking computational considerations into account

How we used auxiliary models

How we used auxiliary models

1. To model the discrepancy and decide where to run the simulator
2. To measure the discrepancy by classification
3. To estimate the posterior by penalised logistic regression

How we used auxiliary models

1. To model the discrepancy and decide where to run the simulator
2. To measure the discrepancy by classification
3. To estimate the posterior by penalised logistic regression

Modelling the discrepancy

(Gutmann & Corander, 2013-2016)

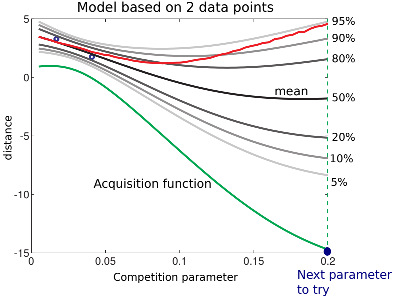
- Assume that a discrepancy measure d between simulated and observed data has been specified.
- Model the conditional distribution of the discrepancy d given θ
- Estimated model yields approximation $\hat{L}(\theta)$ for any choice of ϵ

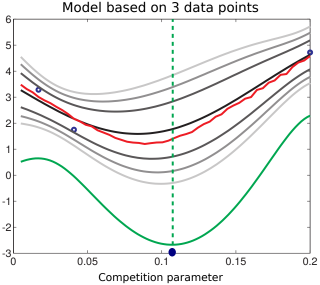
$$\hat{L}(\theta) \propto \widehat{\mathbb{P}}(d \leq \epsilon \mid \theta)$$

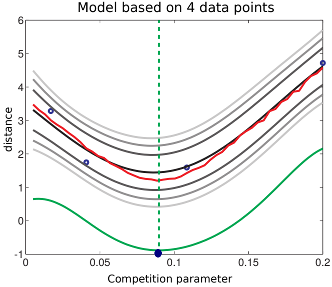
  $\widehat{\mathbb{P}}$ is probability under the estimated model.
- We used the model of $d|\theta$ to decide for which parameters to simulate the model next.

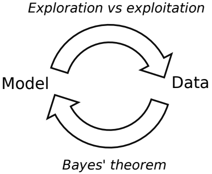
- Approach also applicable to other kernels and synthetic likelihood.

Bayesian optimisation for likelihood-free inference



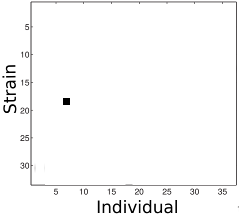




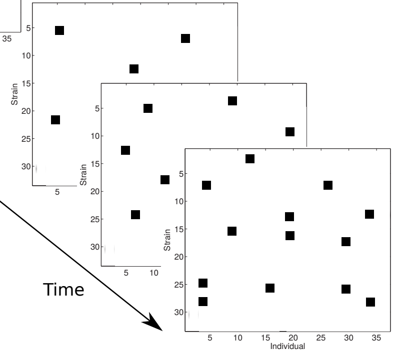


Example: Bacterial infections in child care centres

- Likelihood intractable for cross-sectional data
- But generating data from the model is possible



Parameters of interest:
- rate of infections within a center
- rate of infections from outside
- competition between the strains



(Numminen et al, 2013)

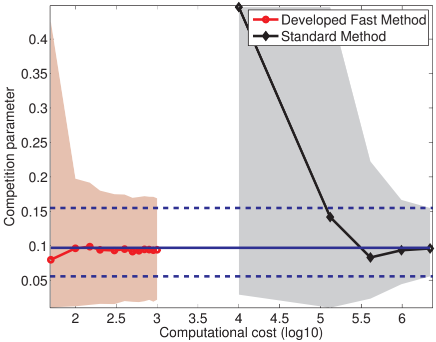
Example: Bacterial infections in child care centres

- Comparison of the proposed approach with a standard population Monte Carlo ABC approach.
- Roughly equal results using 1000 times fewer simulations.

4.5 days with 200 cores
$\downarrow$
90 minutes with seven cores

Posterior means: solid lines,

credibility intervals: shaded areas or dashed lines.



(Gutmann and Corander, 2016)

Interesting future research

- Choice of the model for $d|\theta$

  (Some results available here: arXiv:1610.06462, Järvenpää et al, 2016)

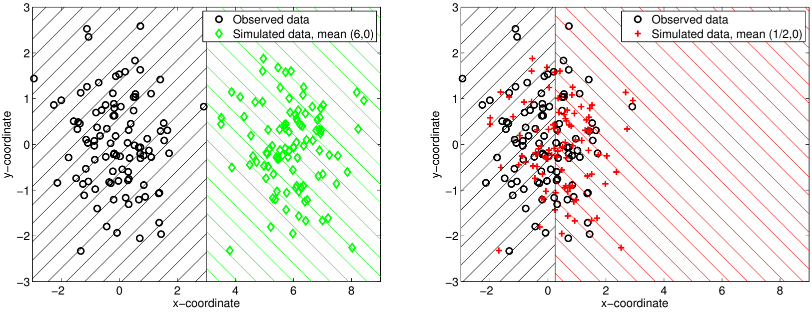
- Choice of the acquisition function

How we used auxiliary models

1. To model the discrepancy and decide where to run the simulator
2. To measure the discrepancy by classification
3. To estimate the posterior by penalised logistic regression

Classification accuracy as discrepancy measure

Correctly classifying data into two categories is easier if the two data sets were generated with very different values of θ (left) than with similar values (right).



(Gutmann et al, 2014, 2017)

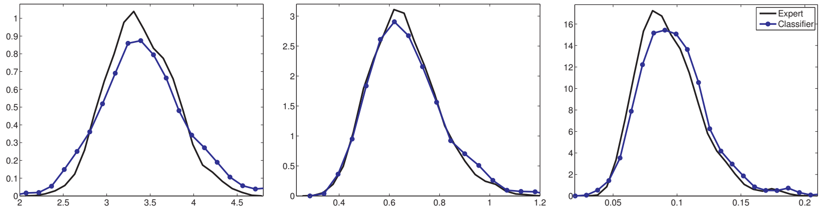
Classification accuracy as discrepancy measure

(Gutmann et al, 2014, 2017)

- Classification accuracy (discriminability) serves as distance measure.
- Value of $1/2$: close; Value of 1: far
- Complete arsenal of classification methods becomes available to inference.
- Choice of discriminative model? Use tools from classification literature.

Example: Bacterial infections in child care centres

- The classification-based distance measure does not require domain/expert knowledge.
- Performs as well as a distance measure based on domain knowledge (Numminen et, Biometrics, 2013).



(a) rate of infections within a centre (b) rate of infections due to outside source (c) competition parameter

(Gutmann et al, 2014, 2017)

How we used auxiliary models

1. To model the discrepancy and decide where to run the simulator
2. To measure the discrepancy by classification
3. To estimate the posterior by penalised logistic regression

Basic idea

(Dutta et al, 2016, arXiv:1611.10242)

- Frame posterior estimation as ratio estimation problem

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)} = p(\theta)r(x,\theta) \tag{1}$$

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

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

$$\hat{L}(\theta) \propto \hat{r}(x^o, \theta), \qquad \hat{p}(\theta|x^o) = p(\theta)\hat{r}(x^o, \theta). \tag{3}$$

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

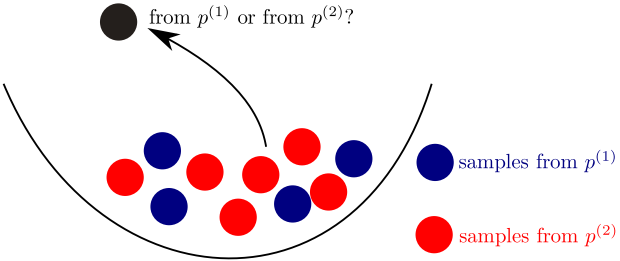
- Samples from two data sets

$$x_i^{(1)} \sim p^{(1)}, \quad i = 1, \ldots, n^{(1)} \tag{4}$$

$$x_i^{(2)} \sim p^{(2)}, \quad i = 1, \ldots, n^{(2)} \tag{5}$$

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

$$\mathbb{P}(x \sim p^{(1)}|x, h) = \frac{1}{1 + \nu \exp(-h(x))}, \qquad \nu = \frac{n^{(2)}}{n^{(1)}} \tag{6}$$



Density ratio estimation by logistic regression

- Estimate h by minimising

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

$$h_i^{(1)} = h\left(x_i^{(1)}\right) \qquad h_i^{(2)} = h\left(x_i^{(2)}\right)$$

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

- 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)

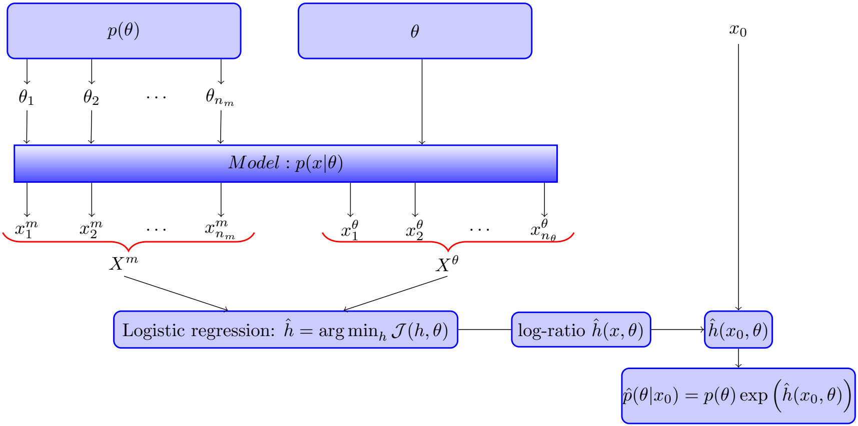
- For posterior estimation, we use
  - data generating pdf $p(x|\theta)$ for $p^{(1)}$
  - marginal $p(x)$ for $p^{(2)}$ (Other choices for $p(x)$ possible too)
  - sample sizes entirely under our control
- Logistic regression gives (point-wise in θ)

$$\hat{h}(x,\theta) \to \log \frac{p(x|\theta)}{p(x)} = \log r(x,\theta) \tag{7}$$

- Estimated posterior and likelihood function:

$$\hat{p}(\theta|x^o) = p(\theta)\exp(\hat{h}(x^o,\theta)) \qquad \hat{L}(\theta) \propto \exp(\hat{h}(x^o,\theta)) \tag{8}$$

Estimating the posterior



(Dutta et al, 2016, arXiv:1611.10242)

Auxiliary model

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

$$h(x) = \sum_{i=1}^{b} \beta_i \psi_i(x) = \beta^\top \psi(x) \tag{9}$$

where $\psi_i(x)$ are summary statistics
- More complex models possible

Exponential family approximation

- Logistic regression yields

$$\hat{h}(x;\theta) = \hat{\beta}(\theta)^\top \psi(x), \qquad \hat{r}(x,\theta) = \exp(\hat{\beta}(\theta)^\top \psi(x)) \tag{10}$$

- Resulting posterior

$$\hat{p}(\theta|x^o) = p(\theta)\exp(\hat{\beta}(\theta)^\top \psi(x^o)) \tag{11}$$

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

$$\hat{r}(x,\theta) = \frac{\hat{p}(x|\theta)}{\hat{p}(x)} \tag{12}$$

$$\hat{p}(x|\theta) = \hat{p}(x)\exp(\hat{\beta}(\theta)^\top \psi(x)) \tag{13}$$

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

Remarks

- Vector of summary statistics $\psi(x)$ 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 β for weighing and selecting summary statistics

Application to ARCH model

- Model:

$$x^{(t)} = \theta_1 x^{(t-1)} + e^{(t)} \tag{14}$$

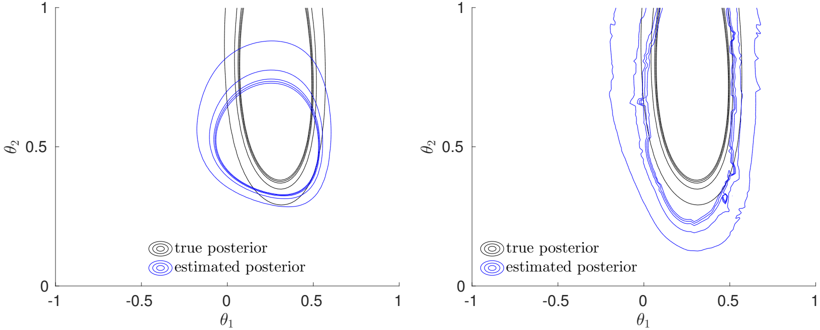
$$e^{(t)} = \xi^{(t)} \sqrt{0.2 + \theta_2 (e^{(t-1)})^2} \tag{15}$$

  $\xi^{(t)}$ 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 θ_1, θ_2

Application to ARCH model

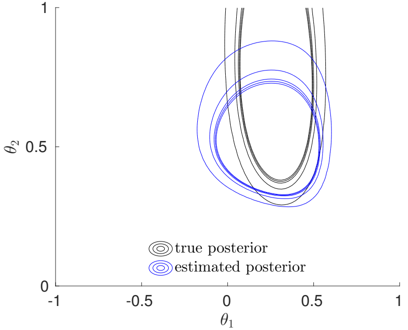
- Summary statistics:
  - auto-correlations with lag one to five
  - all (unique) pairwise combinations of them
  - a constant
- To check robustness: 50% irrelevant summary statistics (drawn from standard normal)
- Comparison with synthetic likelihood with equivalent set of summary statistics (relevant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

Example posterior

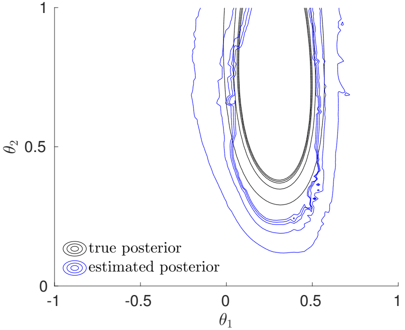


(d) synthetic likelihood (e) proposed method

Example posterior



(f) synthetic likelihood

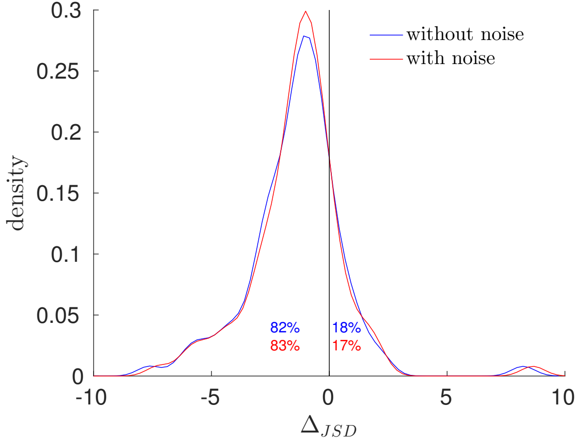


(g) proposed method subject to noise

Systematic analysis

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

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



Application to Ricker model

- Model

$$x^{(t)}|N^{(t)}, \phi \sim \mathrm{Poisson}(\phi N^{(t)}) \tag{16}$$
$$\log N^{(t)} = \log r + \log N^{(t-1)} - N^{(t-1)} + \sigma e^{(t)} \tag{17}$$
$$t = 1, \ldots, 50, \qquad N^{(0)} = 0 \tag{18}$$

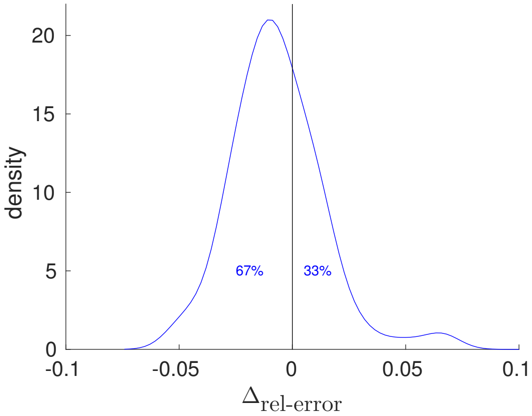
- Parameters and priors
  - log growth rate $\log r \sim \mathcal{U}(3, 5)$
  - scaling parameter $\phi \sim \mathcal{U}(5, 15)$
  - standard deviation $\sigma \sim \mathcal{U}(0, 0.6)$

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

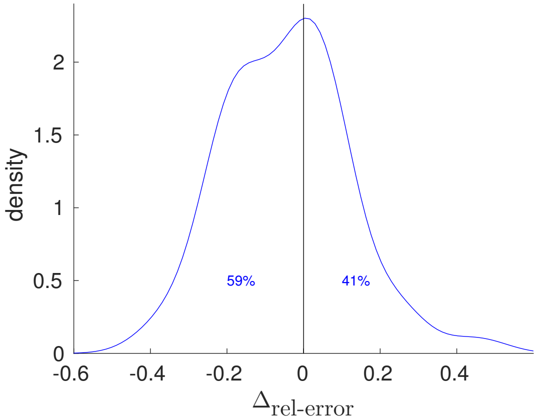
Results for log r

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



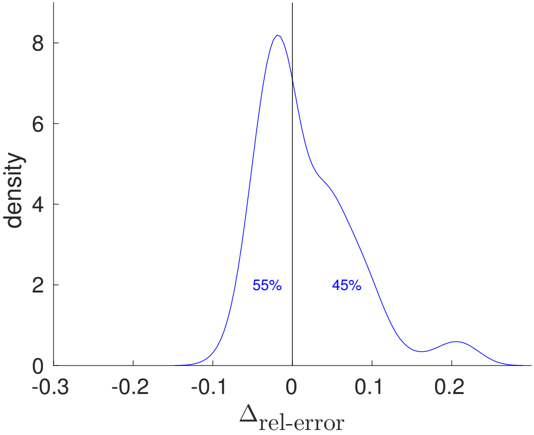
Results for σ

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



Results for ϕ

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



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

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

More results and details in arXiv:1611.10242v1

Conclusions

- Brief overview of how auxiliary models are used in ABC
- Our work on
  - modelling the discrepancy and using the model to decide for which parameter values to evaluate the model
  - discriminative modelling (classification) to measure the discrepancy
  - posterior estimation by regularised ratio estimation
- Importance of automatically controlling the complexity of the auxiliary model (model selection or regularisation)