Approximate Bayesian Computation with Domain Expert in the Loop

Ayush Bharti[1], **Louis Filstroff**[1], **and Samuel Kaski**[1,2]
[1]Department of Computer Science, Aalto University, Finland
[2]Department of Computer Science, University of Manchester, United Kingdom

Aalto University
School of Science

Introduction

Simulator-based models:

- mechanistic models of some physical phenomenon, easy to simulate from
- prevalent in domains such population genetics, ecology, astronomy, telecommunications, and cognitive science
- inferring their parameters θ from data $\mathbf{y}$ is challenging as their likelihood function $p(\mathbf{y}|\theta)$ is intractable
  $\Rightarrow$ maximizing likelihood or Bayesian inference not possible

Solution: Likelihood-free inference methods such as *approximate Bayesian computation* (ABC) [1] that permit sampling from the approximate posterior.

Fundamental unsolved problem in ABC: choosing summary statistics

- choice of statistics readily impacts the performance of ABC methods
- involves a non-trivial trade-off between
  - information loss due to summarization
  - curse of dimensionality
- choice depends on the model, application, and data at hand

In practice, domain experts are crucial for statistics selection.

- Experts manually handcraft and select statistics, which is laborious and time-consuming, involving multiple trial-and-error steps.

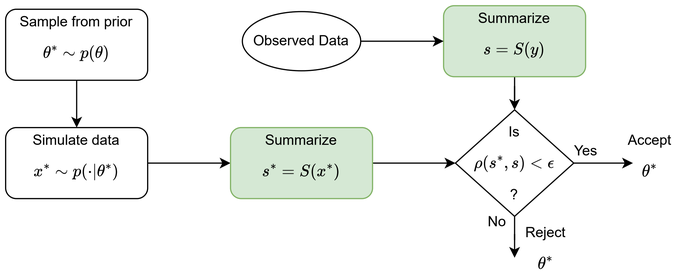
Our contributions:

- We propose an active learning method that makes the statistics selection task easier and efficient for domain experts.
- By eliciting expert knowledge, we are able to handle
  - low simulation regimes, and
  - model misspecification scenarios.

Approximate Bayesian Computation

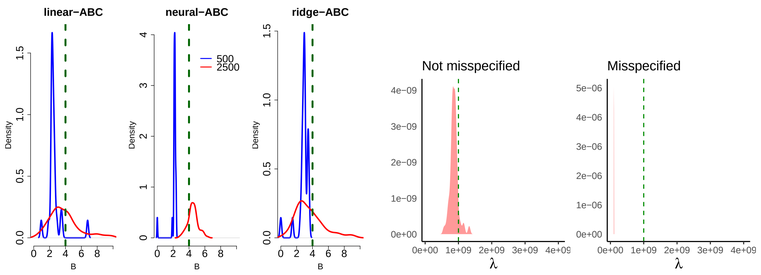
ABC is a likelihood-free inference method that permits sampling from the approximate posterior of a model, given that it is easy to simulate from.

Rejection-ABC algorithm:



Failure of regression-ABC methods

Regression-ABC methods [2] account for the difference between the simulated and observed statistics by adjusting the parameter values.

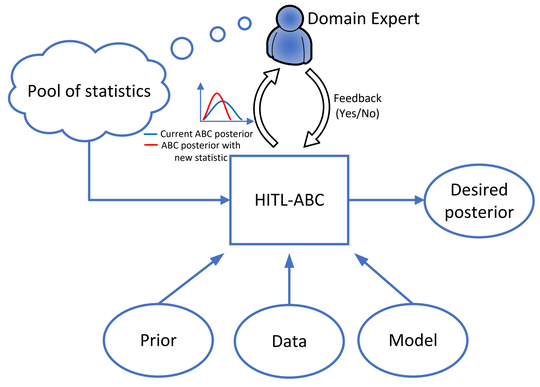


(a) Low-simulation regime

(b) Model misspecification

Method: Human-in-the-loop (HITL) ABC

- We design an experiment to find a statistic from the pool to query
- Domain expert gives feedback (Yes/No) on the usefulness of the statistic
- Expert feedback is formulated as a probabilistic modeling problem
- Knowledge of expert is treated as a latent variable



Expert feedback model

$$\gamma_j \sim \text{Bernoulli}(\rho_j) \tag{1}$$
$$f_j|\gamma_j \sim \gamma_j \text{Bernoulli}(\pi) + (1-\gamma_j)\text{Bernoulli}(1-\pi) \tag{2}$$

- $\gamma \in \{0,1\}$: statistic inclusion/exclusion variable
- $f \in \{0,1\}$: expert binary feedback
- $\pi \in [0,1]$: expert latent knowledge variable
- $\rho \in [0,1]$: prior probability of selecting a statistic

Posterior of feedback model given indices $\mathscr{J}$ of queried statistics:

$$p(\boldsymbol{\gamma}|\mathscr{F}) = \prod_{j\in\mathscr{J}} p(\gamma_j|f_j) \prod_{j\notin\mathscr{J}} p(\gamma_j) \tag{3}$$

ABC posterior based on feedback

We define the ABC posterior given a sequence of feedback $\mathscr{F} = \{f_1, \ldots, f_m\}$ as

$$p^{\epsilon}_{\text{ABC}}(\theta|\mathbf{y},\mathscr{F}) := \sum_{\boldsymbol{\gamma}\in\{0,1\}^w} p^{\epsilon}_{\text{ABC}}(\theta|\mathbf{y},\boldsymbol{\gamma}) p(\boldsymbol{\gamma}|\mathscr{F}). \tag{4}$$

Procedure for sampling from this ABC posterior:

1. Sample $\boldsymbol{\gamma}^{(i)} \sim p(\boldsymbol{\gamma}|\mathscr{F})$
2. Sample $\theta^{(i)}|\boldsymbol{\gamma}^{(i)} \sim p^{\epsilon}_{\text{ABC}}(\theta|\mathbf{y},\boldsymbol{\gamma}^{(i)})$

Utility function

Based on the KL divergence between current and future ABC posterior given feedback $\tilde{f}_j$:

$$j^* = \underset{j\notin\mathscr{J}_k}{\text{argmax}}\, \mathbb{E}_{p(\tilde{f}_j|\mathscr{F}_k)} \left[\text{KL}[p^{\epsilon}_{\text{ABC}}(\theta|\mathbf{y},\mathscr{F}_k,\tilde{f}_j) \,||\, p^{\epsilon}_{\text{ABC}}(\theta|\mathbf{y},\mathscr{F}_k)]\right] \tag{5}$$

HITL-ABC posterior

At the end of each iteration k, the HITL-ABC posterior is $p^{\epsilon}_{\text{ABC}}(\theta|\mathbf{y},\hat{\boldsymbol{\gamma}})$ where $\hat{\boldsymbol{\gamma}}$ is

$$\hat{\gamma}_{k,j} = \begin{cases} \underset{\gamma_j\in\{0,1\}}{\text{argmax}}\, p(\gamma_j|f_j), & \text{if } j \in \mathscr{J}_k \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

References

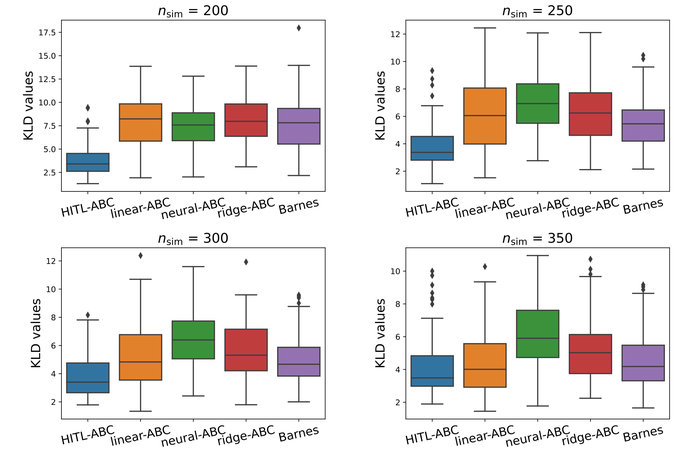
[1] Sisson, S. A, *Handbook of Approximate Bayesian Computation*, Chapman and Hall/CRC, 2018.

[2] Beaumont, M. A., Zhang, W., and Balding, D. J "Approximate Bayesian computation in population genetics", *Genetics*, 162(4):2025–2035, 2002.

Results

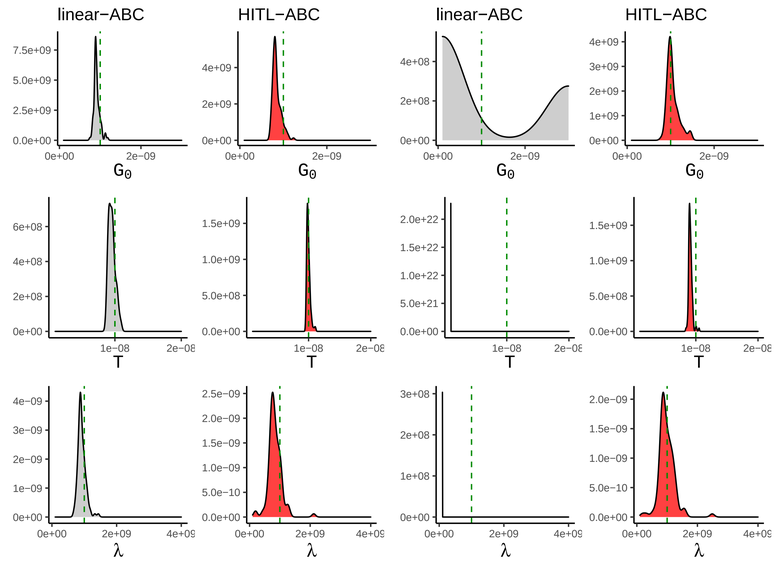
Experiment in low-simulation regime

- Model: g-and-k distribution (4 parameters)
- HITL-ABC outperforms other methods for $n_{\text{sim}} \leq 350$, otherwise at par



Experiment under model misspecification

- Radio propagation model (high-dimensional, complex-valued time series)
- Expert is shown inference results, and can detect misspecified statistics
- 3 parameters, ζ specifies misspecification level



(a) $\zeta = 0$ (b) $\zeta = 5$

Conclusion

- We introduce the first ABC method that actively leverages domain knowledge from experts in order to select summary statistics.
- With fairly limited effort from the expert, we are able to outperform the regression-ABC methods.