Approximate Bayesian Computation with Domain Expert in the Loop

Ayush Bharti[1], Louis Filstroff[1], Samuel Kaski[1,2]

Department of Computer Science, Aalto University, Finland[1]
Department of Computer Science,University of Manchester, UK[2]

July 19, 2022

Likelihood-free Inference

Setting: Inference for simulator-based models with intractable likelihoods

- Let data $\mathrm{y}_{\mathrm{obs}} = \{y_{\mathrm{obs},i}\}_{i=1}^{n}$ be denoted by empirical distribution $\mathbb{Q}^n$.
- Simulator $\mathcal{M}_\Theta = \{\mathbb{P}_\theta : \theta \in \Theta \subset \mathbb{R}^q\}$ is a parametric family of distributions.

Aim: Given data $\mathrm{y}_{\mathrm{obs}}$, estimate θ s.t. $\mathbb{Q}^n$ is "closest" to $\mathbb{P}_\theta$.

Likelihood-free Inference

Setting: Inference for simulator-based models with intractable likelihoods

- Let data $\mathrm{y}_{\mathrm{obs}} = \{y_{\mathrm{obs},i}\}_{i=1}^{n}$ be denoted by empirical distribution $\mathbb{Q}^n$.
- Simulator $\mathcal{M}_\Theta = \{\mathbb{P}_\theta : \theta \in \Theta \subset \mathbb{R}^q\}$ is a parametric family of distributions.

Aim: Given data $\mathrm{y}_{\mathrm{obs}}$, estimate θ s.t. $\mathbb{Q}^n$ is "closest" to $\mathbb{P}_\theta$.

Problem: The likelihood function $p(\mathrm{y}_{\mathrm{obs}}|\theta)$ is intractable and cannot be evaluated numerically.

Therefore, classical estimation techniques such as

Maximum a Posteriori (MAP) estimate: $\hat{\theta}_{\mathsf{MAP}} = \underset{\theta}{\operatorname{argmax}}\ p(\mathrm{y}_{\mathrm{obs}}|\theta)p(\theta)$

Maximum Likelihood (ML) estimate: $\hat{\theta}_{\mathsf{ML}} = \underset{\theta}{\operatorname{argmax}}\ p(\mathrm{y}_{\mathrm{obs}}|\theta)$

are unrealizable.

Approximate Bayesian Computation (ABC)

Solution: Likelihood-free inference methods such as ABC:

- permits sampling from the approximate posterior of a generative model
- widely used in fields such as population genetics, ecology, epidemiology, astrophysics, economics, and telecommunications
- relies on comparing distance between observed and simulated statistics

Fundamental unsolved problem in ABC

Choosing summary statistics!

- choice of statistics readily impacts the performance of ABC methods
- sufficient statistics are not available in most practical cases
- 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

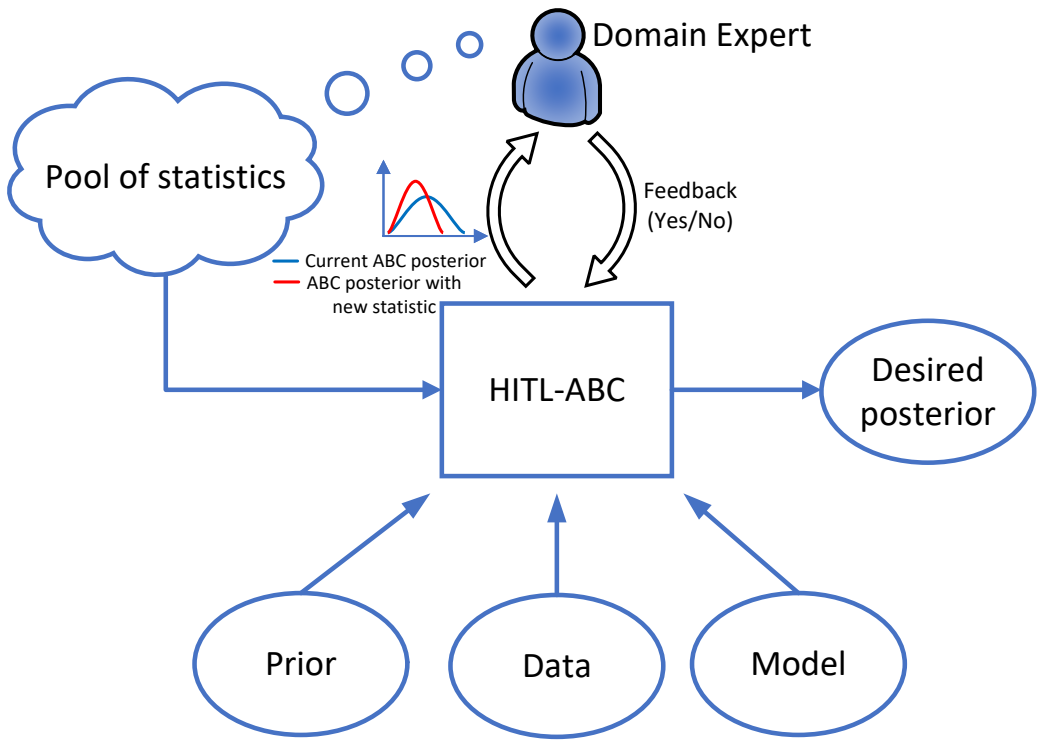
Domain knowledge is vital for constructing statistics

In practice, domain experts manually handcraft and select statistics:

- laborious and time-consuming
- involves multiple trial-and-error steps
- takes up majority of the time of likelihood-free inference projects

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

- We propose a human-in-the-loop ABC statistics selection method which considerably eases the work of domain experts
- By including the experts in the inference loop, we achieve better posterior characterization when
  - model evaluation is costly
  - model is misspecified
- Assumption: expert knowledge is tacit



To sum up...

- How to choose statistics is a fundamental unsolved problem in ABC.
- In practice, experts manually select and handcraft statistics based on domain knowledge, which is laborious.
- We propose an active statistics selection method that reduces the expert's effort.
- Find us at: Hall E. **#705**