Bharti, Ayush; Filstroff, Louis; Kaski, Samuel

Approximate Bayesian Computation with Domain Expert in the Loop

Published in:
Proceedings of the 39th International Conference on Machine Learning

Published: 01/01/2022

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Bharti, A., Filstroff, L., & Kaski, S. (2022). Approximate Bayesian Computation with Domain Expert in the Loop. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 1893-1905). (Proceedings of Machine Learning Research; Vol. 162). JMLR. https://proceedings.mlr.press/v162/bharti22a.html

Approximate Bayesian Computation with Domain Expert in the Loop

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

Abstract

Approximate Bayesian computation (ABC) is a popular likelihood-free inference method for models with intractable likelihood functions. As ABC methods usually rely on comparing summary statistics of observed and simulated data, the choice of the statistics is crucial. This choice involves a trade-off between loss of information and dimensionality reduction, and is often determined based on domain knowledge. However, handcrafting and selecting suitable statistics is a laborious task involving multiple trial-and-error steps. In this work, we introduce an active learning method for ABC statistics selection which reduces the domain expert's work considerably. By involving the experts, we are able to handle misspecified models, unlike the existing dimension reduction methods. Moreover, empirical results show better posterior estimates than with existing methods, when the simulation budget is limited.

1. Introduction

Likelihood-free inference has considerably extended the applicability domain of probabilistic inference, to the set of problems where a simulator is available even though the likelihood function is not known or feasibly computable. Approximate Bayesian computation (ABC) (Marin et al., 2011; Lintusaari et al., 2017; Sisson, 2018; Beaumont, 2019) has emerged as a popular method for likelihood-free inference in a number of fields such as population genetics (Pritchard et al., 1999; Beaumont, 2010), cosmology (Akeret et al., 2015), and radio propagation (Bharti et al., 2021), among others. ABC permits sampling from an approximate posterior distribution of a generative model by comparing summary statistics of simulated and observed (high-dimensional) data. However, the success of ABC may hide from view the fact that major problems are still only partially solved. In this paper, we discuss the problem of choosing summary statistics, a key part of ABC which is often missed in clean theoretical works and only treated case-specifically in concrete inference studies.

Recent works have proposed to circumvent choosing statistics by learning a suitable representation from data with neural networks (Papamakarios & Murray, 2016; Lueckmann et al., 2017; 2019; Izbicki et al., 2019). However, training neural networks requires a large amount of data, which means extensive simulator runs in the ABC context. Thus, in a *low-simulation regime*, where the number of available simulations is limited, choosing the summary statistics is unavoidable – and useful in any case. This choice involves navigating a difficult trade-off between: 1) information loss due to data summarization and hence lower-quality posterior approximations, and 2) curse of dimensionality, requiring exponentially increasing numbers of simulator runs. A low-dimensional set of statistics which is highly informative about the model parameters would be ideal, but obtaining such a set is non-trivial. The only way out of this conundrum is to bring in additional domain knowledge, and hence practitioners end up spending a large proportion of the time of their likelihood-free inference projects in choosing suitable statistics.

Several methods have been proposed to automatically reduce the dimension of a given set (or pool) of available summary statistics to use in an ABC method, see Blum et al. (2013); Prangle (2015) for exhaustive surveys. These include methods based on subset selection (Joyce & Marjoram, 2008; Nunes & Balding, 2010; Blum, 2010; Barnes et al., 2012; Blum et al., 2013), projection (Wegmann et al., 2009; Fearnhead & Prangle, 2012; Aeschbacher et al., 2012; Jiang et al., 2017; Chen et al., 2020), and regression adjustment (Beaumont et al., 2002; Blum & François, 2010; Bi et al., 2021). However, none of these methods are able to handle low-simulation regimes and model misspecification, which occurs when there is a mismatch between the simulator and the true data-generating mechanism. Under model misspecification, dimension-reducing ABC methods may produce summary statistics which will never replicate the observed value irrespective of the parameter setting, which in turn causes problems in the ABC (Frazier et al., 2020). We call those statistics misspecified. In low-simulation regimes, on

[1]Department of Computer Science, Aalto University, Espoo, Finland [2]Department of Computer Science, University of Manchester, Manchester, United Kingdom. Correspondence to: Ayush Bharti <ayush.bharti@aalto.fi>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

the other hand, these methods would become susceptible to fitting to noisy, uninformative statistics (Blum et al., 2013). Therefore, existing methods cannot alone offer a sufficient solution to the statistics selection problem.

In practice, domain knowledge is brought in likelihood-free inference by experts handcrafting and selecting the statistics manually. This is necessary, as the choice depends on the model, data and application at hand, albeit laborious, as it involves multiple trial-and-error steps. In this paper, we propose a human-in-the-loop ABC statistics selection method which considerably eases the work of domain experts, extending the statistics selection method of Barnes et al. (2012). Taking the regression-based ABC methods as a case study, we show that by including the experts in the inference loop, we achieve better posterior characterization when the model is misspecified or when model evaluation is costly. We assume that expert knowledge is tacit, that is, the expert cannot easily produce an optimal set of informative statistics, but can recognize a good statistic when presented with it. Additionally, the expert can recognize potentially misspecified statistics and exclude them. We adopt a sequential Bayesian experimental design (BED) (Chaloner & Verdinelli, 1995; Ryan et al., 2016) approach to sequentially select the most informative statistics to present to the expert using a forward-stepwise selection method (Hastie et al., 2009). To the best of our knowledge, domain experts have not been formally involved in ABC methods so far. We show clearly better empirical performance than with existing methods on two models with intractable likelihoods: a quantile distribution and a radio propagation model.

2. Basics & Motivation

We introduce some basics on ABC methods in Section 2.1 and demonstrate their potential pitfalls in Section 2.2.

2.1. Approximate Bayesian computation

Let $\mathcal{Y}$ be the data space and $\mathcal{M}_\Theta = \{\mathbb{P}_\theta : \theta \in \Theta \subset \mathbb{R}^q\}$ a parametric model family of distributions $\{\mathbb{P}_\theta\}$ on $\mathcal{Y}$. We assume that $\mathbb{P}_\theta$ does not have a tractable likelihood function given observed data $\mathbf{y}_{\text{obs}}$ (comprising n_{obs} samples), but it is possible to simulate independent and identically distributed (i.i.d.) samples from $\mathcal{M}_\Theta$ given some θ. For such models, ABC methods can be used to approximate their posterior distribution $p(\theta|\mathbf{y}_{\text{obs}}) \propto p(\mathbf{y}_{\text{obs}}|\theta)p(\theta)$ in a non-parametric manner, where $p(\theta)$ denotes the prior beliefs and $p(\mathbf{y}_{\text{obs}}|\theta)$ is the joint likelihood function. We now briefly describe some of the basic ABC methods.

Rejection-ABC Consider a deterministic function η mapping the data to a lower-dimensional space of summary statistics such that $\mathbf{s}_{\text{obs}} = \eta(\mathbf{y}_{\text{obs}})$ is the vector of summary statistics of $\mathbf{y}_{\text{obs}}$. The basic rejection-ABC algorithm

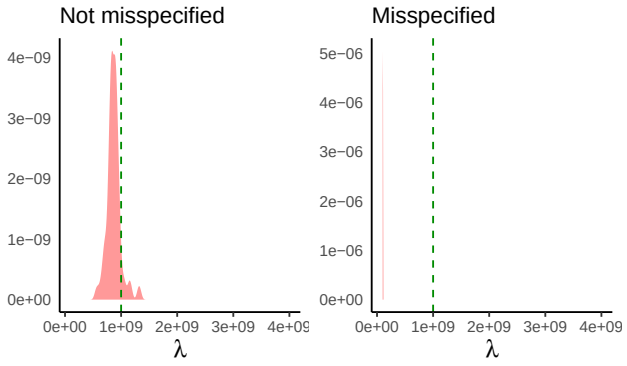


Figure 1. Catastrophic performance of linear-ABC under model misspecification for the λ parameter of the radio propagation model, see Section 4.1 for experiment details. The dashed green line denotes the true parameter value. We observe that the ABC samples concentrate far away from the true value, on the left prior boundary, when the model is misspecified.

(Pritchard et al., 1999) proceeds in the following manner: (1) Sample $\theta^* \sim p(\theta)$; (2) Simulate $\mathbf{y}^* \sim \mathbb{P}_{\theta^*}$ and compute $\mathbf{s}^* = \eta(\mathbf{y}^*)$; (3) If $\varrho(\mathbf{s}^*, \mathbf{s}_{\text{obs}}) < \epsilon$, accept θ^*. Here $\varrho(\cdot,\cdot)$ is a distance function and ϵ is a tolerance threshold. Repeating the algorithm yields a set $\{\theta_i\}_{i=1}^n$ of accepted parameter values which are i.i.d. samples from the approximate posterior $p(\theta|\varrho(\mathbf{s}, \mathbf{s}_{\text{obs}}) < \epsilon) \approx p(\theta|\mathbf{y}_{\text{obs}})$. For a fixed simulation budget of n_{sim} samples, a practical solution is to specify the tolerance as the ratio $\epsilon = n_\epsilon / n_{\text{sim}}$, where n_ϵ is the number of accepted samples out of n_{sim}.

Regression-ABC Regression adjustment approaches to ABC (Blum, 2017) aim to account for the difference between the simulated and observed statistic by adjusting the parameter values. Given samples $(\theta_i, \mathbf{s}_i)_{i=1}^{n_\epsilon}$ obtained with rejection-ABC, a homoscedastic regression model,

$$\theta_i = \varphi(\mathbf{s}_i) + \varepsilon_i, \quad i = 1, \ldots, n_\epsilon, \tag{1}$$

is fitted in the vicinity of $\mathbf{s}_{\text{obs}}$, where $\varphi(\cdot)$ is the conditional expectation $\mathbb{E}[\theta|\mathbf{s}]$, and ε_i are the residuals. The parameter samples are then adjusted as $\tilde{\theta}_i = \hat{\varphi}(\mathbf{s}_{\text{obs}}) + \hat{\varepsilon}_i$, with $\hat{\varphi}$ being the estimate of $\mathbb{E}[\theta|\mathbf{s}]$ and $\hat{\varepsilon}_i$ being the i^{th} empirical residual. Beaumont et al. (2002) assumed φ to be linear, while it was later extended to heteroscedastic non-linear adjustment in Blum & François (2010). Blum et al. (2013) proposed a regularized version of the linear-ABC method via ridge regression. We refer to these methods as linear-ABC, neural-ABC, and ridge-ABC, respectively.

2.2. Pitfalls of regression-ABC methods

Model misspecification The regression-based methods have been shown to concentrate the ABC posterior around the true value when the model is correctly specified (Li & Fearnhead, 2018). However, under model misspecification, i.e., when the true data-generating mechanism does

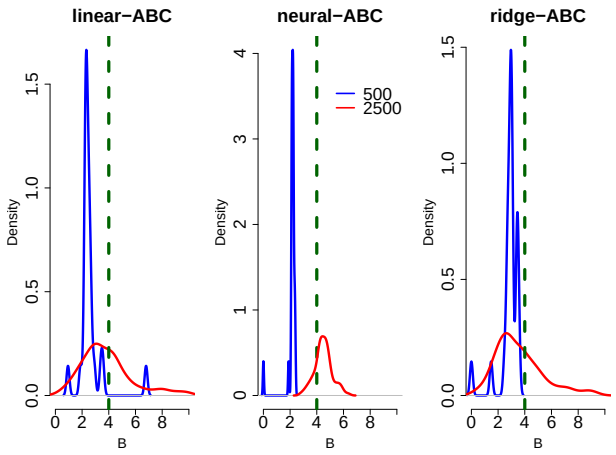


Figure 2. Inability of regression-ABC methods to handle low-simulation regime. The ABC posteriors obtained from linear-ABC, neural-ABC, and ridge-ABC for the parameter B of the g-and-k distribution (see Section 4.2 for details) using $n_{\text{sim}} = 500$ (blue) and $n_{\text{sim}} = 2500$ (red) simulated samples with $\epsilon = 5\%$. The dashed green line denotes the true parameter value. The regression layer in these methods fits to noisy, uninformative statistics, thereby leading to ABC posteriors (blue curves) concentrated away from the true value.

not belong to $\mathcal{M}_\Theta$, they can concentrate the ABC posterior on a completely different region of the parameter space than rejection-ABC, which can be outside the prior range for bounded parameters (Frazier et al., 2020). This phenomenon is demonstrated in Fig. 1 for the radio propagation model (see Section 4.1 for details) using linear-ABC. When the model is misspecified, we see the ABC posterior concentrates on the prior boundary, far away from the true parameter value. This occurs when just one of the statistics is misspecified, as is the case in Fig. 1. In this paper, we utilize the fact that the expert will be able to detect this behaviour and exclude potentially misspecified or out-of-distribution statistics from being included in the regression-ABC methods.

Low-simulation regime Inability of the regression-based methods to handle the low-simulation regime is exemplified by the g-and-k distribution (see Section 4.2) in Fig. 2. We observe that for small n_{sim}, the resulting ABC posteriors can get concentrated away from the true parameter value. As there are few samples to perform the least-squares fit in regression-ABC methods, they may overfit to uninformative statistics. In such cases, these methods may over-adjust the parameter values in the direction of such noisy statistics (Blum et al., 2013).

3. The Method

We propose including in the statistics selection loop a domain expert, who will be able to assess which statistics would be useful or misspecified, and who currently needs to do that choice completely manually. The expert may evaluate the usefulness of a given statistic by, e.g., checking relevant literature. As this is laborious, we would not want to repeat it for all possible candidates. In this section, we introduce an experimental design approach which helps reduce the expert's effort. We formulate their feedback as a probabilistic modelling problem as well, with the knowledge of the expert as a latent variable, as described in Section 3.2. This turns querying the expert into an automatic experimental design problem, which is presented in Section 3.3. Section 3.1 describes the problem set-up, and Alg. 1 outlines the proposed human-in-the-loop (HITL) ABC algorithm.

3.1. Setting

Consider a finite pool of candidate summary statistics $\mathcal{S} = \{s_1, s_2, \ldots, s_w\}$ available for ABC. We introduce a binary variable $\gamma_j \in \{0, 1\}$ to indicate the inclusion or exclusion of the statistic $s_j \in \mathcal{S}$, $j = 1, \ldots, w$ to the summarizing function $\eta(\cdot)$. Denote by $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_w]^\top$ the binary vector corresponding to a vector of statistics $\mathbf{s} = \eta(\mathbf{y})$, such that $\gamma_j = 1$ implies s_j is an element of $\mathbf{s}$. We denote the approximate posterior obtained by applying an ABC method with tolerance ϵ by $p^\epsilon_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \boldsymbol{\gamma})$. For $\boldsymbol{\gamma} = \mathbf{0}$, we set $p^\epsilon_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \boldsymbol{\gamma}) = p(\theta)$. Let $\boldsymbol{\gamma}^*$ represent the desired subset[1] of statistics $\mathbf{s}^* \subset \mathcal{S}$ to be used in the ABC method such that $p^\epsilon_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \boldsymbol{\gamma}^*)$ is the target ABC posterior. We query the expert $\mathcal{E}$ regarding the elements of $\mathcal{S}$ with the goal to converge towards $\boldsymbol{\gamma}^*$ as quickly as possible. We assume that the expert is queried only once about a given statistic, and that querying the expert is costly.

3.2. Expert feedback model

We assume that the expert provides binary feedback $f_j \in \{0, 1\}$ regarding the relevance of the j^{th} statistic s_j and interpret the answer as feedback about γ_j. More precisely, we model f_j as a noisy version of γ_j, such that

$$\gamma_j \sim \text{Bernoulli}(\rho_j), \tag{2}$$

$$f_j|\gamma_j \sim \gamma_j \text{Bernoulli}(\pi) + (1 - \gamma_j)\text{Bernoulli}(1 - \pi). \tag{3}$$

The hyperparameter $\pi \in [0, 1]$ quantifies the level of noise or uncertainty in the feedback, i.e., we have $f_j = \gamma_j$ with probability π. Of course, the method is intended to work

[1] Here, the desired subset of statistics is understood as the subset that achieves the optimal trade-off between minimum dimensionality and information regarding the parameters.

with domain experts having prior knowledge about the statistics, for whom π would be close to 1. The hyperparameter ρ_j corresponds to the prior probability of the j^{th} statistic being included. This model was first proposed by Daee et al. (2017) to get feedback on the relevance of features to be used in regression, and later applied to precision medicine by Sundin et al. (2018). Note that the feedback is independent of $\mathbf{y}_{\text{obs}}$.

By marginalization, it is straightforward to show that the feedback f_j is a Bernoulli random variable with probability of success $\omega_j = \pi\rho_j + (1-\pi)(1-\rho_j)$. We can further characterize the posterior probability of γ_j given f_j as a Bernoulli distribution of parameter ν_j, where

$$\nu_j = \frac{\pi^{f_j}(1-\pi)^{1-f_j}\rho_j}{\omega_j^{f_j}(1-\omega_j)^{1-f_j}}. \tag{4}$$

For simplicity, we assume $\rho_j = \rho$ for all j in the remainder of the paper.

Denote by $\mathcal{J} = \{j_1, j_2, \ldots, j_m\}$ the indices of the $m \leq w$ summary statistics that have been queried from the expert. The corresponding feedback sequence is denoted as $\mathcal{F} = \{f_{j_1}, f_{j_2}, \ldots, f_{j_m}\}$. Then the posterior $p(\boldsymbol{\gamma}|\mathcal{F})$ is

$$p(\boldsymbol{\gamma}|\mathcal{F}) = \prod_{j\in\mathcal{J}} p(\gamma_j|f_j) \prod_{j\notin\mathcal{J}} p(\gamma_j), \tag{5}$$

where $j \notin \mathcal{J}_k$ denotes the indices of statistics for which feedback has not been queried yet.

ABC posterior based on feedback Given that we observe expert feedback $\mathcal{F}$ and not $\boldsymbol{\gamma}$, we define the ABC posterior based on $\mathcal{F}$ as

$$p^{\epsilon}_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \mathcal{F}) := \sum_{\boldsymbol{\gamma}\in\{0,1\}^w} p^{\epsilon}_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma}|\mathcal{F}), \tag{6}$$

that is, we integrate out our current beliefs about $\boldsymbol{\gamma}$. Note that $p^{\epsilon}_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \boldsymbol{\gamma})$ does not have a closed-form expression. Nonetheless, it is possible to obtain i.i.d. samples $\theta^{(i)}$ from $p^{\epsilon}_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \mathcal{F})$ in the following manner:

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

As querying the expert is costly, we want $p^{\epsilon}_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \mathcal{F})$ to converge towards $p^{\epsilon}_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \boldsymbol{\gamma}^*)$ with the least amount of feedback.

3.3. Sequential experimental design

We design a sequential Bayesian experiment to select the next statistic to get feedback on from the expert. We refer the reader to Appendix A for background on Bayesian experimental design. Our utility function is the expected KL divergence between the ABC posteriors (defined in Eq. (6)) before and after receiving a new feedback. Denote by $\mathcal{F}_k$ the feedback collected after iteration k, and by $\mathcal{J}_k$ the indices of the queried statistics (in particular, $\mathcal{F}_0 = \emptyset$ and $\mathcal{J}_0 = \emptyset$). Thus, at iteration $k+1$, the utility maximizing statistic s_{j^*} with index

$$j^* = \arg\max_{j\notin\mathcal{J}_k} U_{k+1}(j), \tag{7}$$

is chosen. The utility function reads

$$U_{k+1}(j) = \mathbb{E}_{p(\tilde{f}_j|\mathcal{F}_k)}\left[\mathfrak{D}^{\text{KL}}_k(\tilde{f}_j)\right], \quad \text{where} \tag{8}$$

$$\mathfrak{D}^{\text{KL}}_k(\tilde{f}_j) = \text{KL}[p^{\epsilon}_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \mathcal{F}_k, \tilde{f}_j) \,||\, p^{\epsilon}_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \mathcal{F}_k)].$$

The expectation in Eq. (8) is taken w.r.t. the posterior predictive distribution $p(\tilde{f}_j|\mathcal{F}_k)$, as the feedback is only observed after actually querying the expert. Recall that in our setting, the expert can only be queried once about each statistic s_j, leading to feedback f_j. Also, f_j is independent of f'_j for $j \neq j'$. It follows that $p(f_j|\mathcal{F}_k) = p(f_j)$. Therefore, we can further write Eq. (8) as the Bernoulli expectation

$$\begin{aligned} U_{k+1}(j) = &\Pr(\tilde{f}_j = 1)\mathfrak{D}^{\text{KL}}_k(\tilde{f}_j = 1) \\ &+ \Pr(\tilde{f}_j = 0)\mathfrak{D}^{\text{KL}}_k(\tilde{f}_j = 0). \end{aligned} \tag{9}$$

Given i.i.d. samples $\{\theta^{(i)}_{k+1}\}^n_{i=1} \sim p_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \mathcal{F}_k, \tilde{f}_j)$ and $\{\theta^{(i)}_k\}^n_{i=1} \sim p_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \mathcal{F}_k)$, we estimate the KL divergence using the 1-nearest neighbour density estimate (Wang et al., 2006; Jiang, 2018)

$$\mathfrak{D}^{\text{KL}}_k(f_j) \approx \frac{q}{n}\sum_{i=1}^{n} \log \frac{\min_j \|\theta^{(i)}_k - \theta^{(j)}_{k-1}\|}{\min^n_{j\neq i} \|\theta^{(i)}_k - \theta^{(j)}_k\|} + \log\frac{n}{n-1}, \tag{10}$$

Algorithm 1 Human-in-the-loop (HITL) ABC

Input: data $\mathbf{y}_{\text{obs}}$, model $\mathcal{M}_\Theta$, expert $\mathcal{E}$, prior $p(\theta)$, pool $\mathcal{S}$, tolerance ϵ, stopping threshold δ
repeat
  Sample $\{\theta^{(i)}_k\}^n_{i=1} \sim p_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \mathcal{F}_k)$ (see Section 3.2)
  for $j \notin \mathcal{J}_k$ **do**
    $\{\theta^{(i)}_{k+1}\}^n_{i=1} \sim p_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \mathcal{F}_k, \tilde{f}_j)$ for $\tilde{f}_j = \{0, 1\}$
    Compute utility $U_{k+1}(j)$ from (8)
  end for
  Find j^* by solving (7)
  Query s_{j^*} from the expert to get feedback f_{j^*}
  $\mathcal{F}_{k+1} = \mathcal{F}_k \cup f_{j^*}$
until stopping criterion is met
Output: ABC posterior $p^{\epsilon}_{\text{ABC}}(\theta|\mathbf{y}_{\text{obs}}, \hat{\boldsymbol{\gamma}})$, where $\hat{\boldsymbol{\gamma}}$ is given by (11)

which guarantees almost-sure convergence to the true divergence (Perez-Cruz, 2008). Moreover, the estimator has a time complexity of $\mathcal{O}(n \log n)$ (Jiang, 2018), and thus scales well with the number of samples. We use k-d trees (Bentley, 1975; Maneewongvatana & Mount, 2001) to implement it with $n = 4000$ samples for all the experiments in this paper.

Stopping criterion We stop Alg. 1 as soon as the utility of the remaining statistics falls below a pre-defined threshold δ, i.e. $U_{k+1}(j) \leq \delta$ for $j \notin \mathcal{J}_k$. Addition of any new statistic from this stage onwards barely impacts the ABC posterior, indicating the absence of informative statistics in the remaining pool. We therefore assign $\gamma = 0$ to the statistics not queried before stopping the algorithm. To set the value of δ, we follow the argument by Barnes et al. (2012) and pick δ which is larger than the estimated KL divergence between samples from the same distribution. Formally, let $\mathbf{X} = (X^{(1)}, \ldots, X^{(n)})$ be a sample from a q-dimensional density p_X. We sample $\{\mathbf{X}_i\}_{i=1}^{M} \sim p_X$ and set $\delta = \max_{i,j}\ \mathrm{KL}(\mathbf{X}_i, \mathbf{X}_j)$, for $i, j = 1, \ldots, M$.

Output of the algorithm At the end of each iteration k, the current estimate of the statistics indicator vector is $\hat{\boldsymbol{\gamma}}_k = (\hat{\gamma}_{k,1}, \ldots, \hat{\gamma}_{k,w})$, where

$$\hat{\gamma}_{k,j} = \begin{cases} \underset{\gamma_j \in \{0,1\}}{\arg\max}\, p(\gamma_j | f_j), & \text{if } j \in \mathcal{J}_k \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

The final output of Alg. 1 is the ABC posterior $p^{\epsilon}_{\mathrm{ABC}}(\theta | \mathbf{y}_{\mathrm{obs}}, \hat{\boldsymbol{\gamma}})$ where $\hat{\boldsymbol{\gamma}}$ is obtained from Eq. (11) given all the collected feedback.

4. Experiments

In this section, we empirically assess the performance of the proposed HITL-ABC method against regression-ABC methods under model misspecification in Section 4.1, and in low-simulation regimes in Section 4.2. Lastly, the sensitivity to hyperparameters is analyzed in Section 4.3. The source code is available at `https://github.com/lfilstro/HITL-ABC`.

Implementation details Our algorithm can be implemented on top of any ABC method. To identify the effect of the novel contribution, we choose the same method used as a baseline method in comparisons, namely the regression adjustment approach of Beaumont et al. (2002) (linear-ABC). The regression-ABC methods are implemented using the `abc` R package (Csilléry et al., 2012). For all the experiments, we assume bounded uniform priors on the parameters and use a logit transform (Blum & François, 2010) before adjusting them to ensure adjusted parameters do not fall outside the prior range. The statistics are normalized by an estimate of their mean absolute deviation before computing the distance to account for the difference in magnitudes. The confidence in the feedback is set to $\pi = 0.95$, and the stopping criterion is $\delta = 0.06$. Assuming each statistic is equally likely to be included or excluded *a priori*, we set $\rho = 0.5$. We assume $\varrho(\cdot, \cdot)$ to be the Euclidean norm $\|\cdot\|$, as is a typical choice in ABC. Lastly, a run of the algorithm uses the same simulated data at each iteration for computational ease.

4.1. Experiment under model misspecification

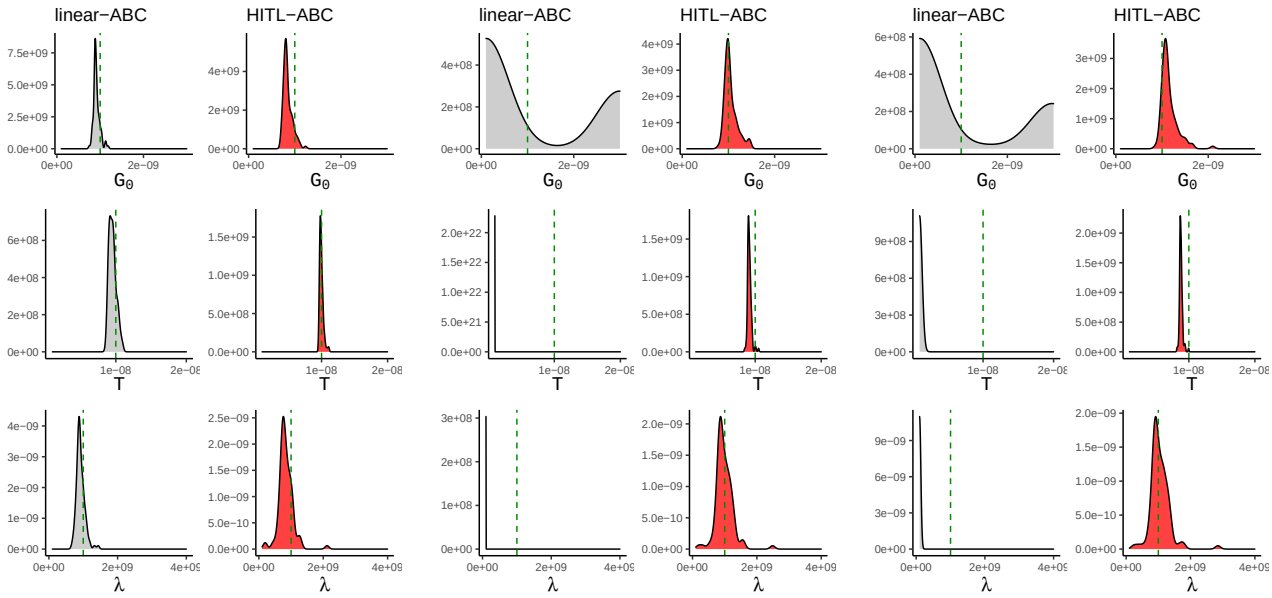
We study the performance of the HITL-ABC method against that of linear-ABC (Beaumont et al., 2002) under model misspecification. More precisely, we consider the challenging problem of estimating parameters of a stochastic radio channel model having intractable likelihood. Driven by an underlying point process, such models simulate radio propagation phenomena and are used to test and design wireless communication systems (Goldsmith, 2005). The potential of likelihood-free methods for inferring parameters of such models has been recognized recently (Bharti et al., 2020; Adeogun et al., 2021).

Data and model description Radio channel data is measured in the frequency bandwidth B at n_s equidistant points, resulting in a frequency separation of $\Delta f = B/(n_s - 1)$. The measured transfer function data is $(Y_0, Y_1, \ldots, Y_{n_s-1})$. The time-domain signal $y(t)$ is obtained by inverse Fourier transforming $\{Y_i\}_{i=0}^{n_s-1}$ to $y(t) = \frac{1}{n_s} \sum_{i=0}^{n_s-1} Y_i \exp(j2\pi i \Delta f t)$. Multiple realizations yield an $n_{\mathrm{obs}} \times n_s$-dimensional data matrix. We focus on the model by Turin et al. (1972) who define the transfer function as $Y_i = \sum_l \alpha_l \exp(-j2\pi \Delta f i \tau_l)$, where τ_l is the time-delay and α_l is the complex gain of the l^{th} component. The delays are modeled as a one-dimensional Poisson point process with arrival rate λ. The gains α_l, conditioned on τ_l, are modeled as i.i.d. zero-mean circular complex Gaussian variables with conditional variance $\mathbb{E}[|\alpha_l|^2 | \tau_l] = G_0 \exp(-\tau_l / T)/\lambda$. Therefore, the parameter vector constitutes $\theta = (G_0, T, \lambda)$. As the underlying points $\{\tau_l, \alpha_l\}$ are unobserved, the likelihood becomes intractable. The high dimensionality of the data compounds the issue, as n_s can be of the order of a few thousands.

Setting We consider the means and variances of the first three log-temporal moments m as the summary statistics,

$$m_i = \log \int_0^{\frac{1}{\Delta f}} t^i |y(t)|^2 \mathrm{dt}, \quad i = 0, 1, 2, \tag{12}$$

as they have been shown to be informative about the parameters of interest (Bharti et al., 2020; Adeogun et al., 2021). Thus, the total number of statistics is $w = 6$. To create a misspecified model, we perturb one of the temporal



(a) $\zeta = 0$ (b) $\zeta = 5$ (c) $\zeta = 10$

Figure 3. HITL-ABC radically outperforms linear-ABC under model misspecification ($\zeta > 0$, panels (b) and (c)). Approximate posteriors of the parameters of the radio propagation model obtained from HITL-ABC (red) and linear-ABC (grey) at varying levels of misspecification. The dashed green line denotes the true parameter value. For $\zeta = 0$ (panel (a)), the model is correctly specified. Prior is $\mathcal{U}([10^{-10}, 3 \times 10^{-10}] \times [10^{-9}, 2 \times 10^{-8}] \times [10^{8}, 4 \times 10^{9}])$. Settings: $B = 4 \times 10^9$, $n_s = 801$, $n_{\text{obs}} = 300$, $n_{\text{sim}} = 2000$, $\epsilon = 5\%$.

moments to produce a mismatch between observed and simulated statistics. Specifically, we compute observed statistics using $\theta_{\text{true}} = (10^{-9}, 10^{-8}, 10^{9})$, and add zero-mean Gaussian random variables with variance ζ to m_0. This leads to $\text{var}(m_0)$ being the only misspecified statistic. As a result, no setting of parameters yields temporal moments that match the observed values. We assess the performance of ABC methods by considering $\zeta = \{0, 5, 10\}$.

Expert involvement In this experiment, we involved the expert to detect misspecification by showing them inference results. To that end, we asked for real feedback from a radio propagation expert. We first asked the expert to confirm the relevance of the statistics obtained from literature, prior to running the experiment. At each iteration, presenting just the utility maximizing statistic to the expert may not be sufficient for them to qualify it as being misspecified. Thus, the expert was also presented with the ABC posterior obtained before and after including the utility maximizing statistic. In particular, this gave the expert the opportunity to observe the impact of the statistic on the ABC posteriors, and potentially exclude it if they deem it to be misspecified.

Results We observe in Fig. 3-a that when the model is correctly specified ($\zeta = 0$), both methods yield similar performance, as expected. When misspecification occurs ($\zeta > 0$), as shown in Fig. 3-b and Fig. 3-c, the performance of linear-ABC seriously degrades — posterior samples become concentrated further away from the correct value, on the prior boundary for T and λ. The posterior of G_0 is also hampered significantly. That is not the case for HITL-ABC, as the expert involved is able to observe the effect of the $\text{var}(m_0)$ statistic on the ABC posterior (as shown in Fig. 4) and exclude it from being selected. Hence, the performance of HITL-ABC remains relatively stable as the level of misspecification increases. Additional results of the experiment can be found in Appendix B.

4.2. Experiment in low-simulation regime

We now compare the performance of the proposed HITL-ABC method against linear-ABC (Beaumont et al., 2002), neural-ABC (Blum & François, 2010), and ridge-ABC (Blum et al., 2013) in low-simulation regimes. We also include the statistics selection method of Barnes et al. (2012) (implemented with $\delta = 0.1$) combined with linear-ABC for comparison. We demonstrate the results using the g-and-k distribution (Prangle, 2020), which is a flexible univariate distribution without a closed-form density. It is defined by its inverse cumulative distribution function

$$F^{-1}(x; A, B, c, g, k) = A + B\left[1 + c\frac{1 - \exp\left(-gz(x)\right)}{1 + \exp\left(-gz(x)\right)}\right](1 + z(x)^2)^k z(x),$$

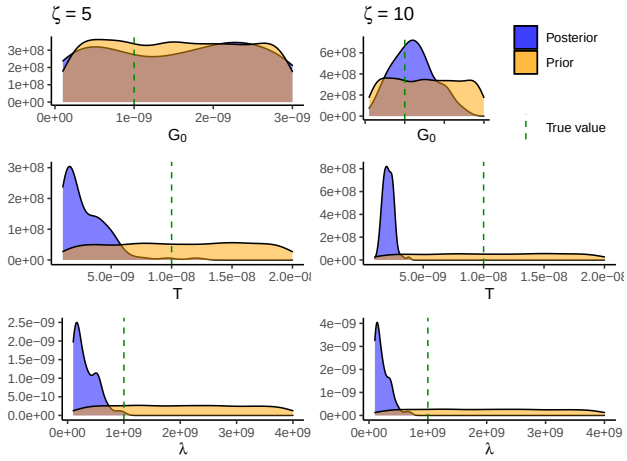


Figure 4. The expert is presented with the ABC posteriors before (yellow) and after (blue) including a statistic in the misspecification experiment. Figure shows the information provided to the expert for the statistic $\text{var}(m_0)$. As $\text{var}(m_0)$ is misspecified for $\zeta > 0$, the ABC posteriors for T and λ get concentrated far away from the true value, on the prior boundaries. Based on this information, the expert provides feedback on whether that statistic should be included.

where $z(x)$ is the x^{th} standard Gaussian quantile. Keeping $c = 0.8$ fixed (Rayner & MacGillivray, 2002), the unknown parameters $\theta = (A, B, g, k)$ govern the location, scale, skewness, and kurtosis of the distribution, respectively.

Setting The pool of statistics consists of estimates of these quantities: $s_A = L_2$, $s_B = L_3 - L_1$, $s_g = L_3 + L_1 - 2L_2)/s_B$, and $s_k = (E_7 - E_5 + E_3 - E_1)/s_B$ where L_i and E_j are the i^{th} quartile and j^{th} octile, respectively (Drovandi & Pettitt, 2011). We also include pairwise products of these four statistics and five uniform random variables $u_i \sim \mathcal{U}(0,1)$, $i = 1, \ldots, 5$ in $\mathcal{S}$, yielding a total of $w = 15$ statistics. The expert feedback is simulated using Eq. (3) with $\mathbf{s}^* = \{s_A, s_B, s_g, s_k\}$. The priors for all the parameters are set to $\mathcal{U}(0, 10)$. The true parameter value is $\theta_{\text{true}} = (3, 4, 2, 1)$. The statistics are computed using $n_{\text{obs}} = 10,000$ data points from the g-and-k distribution. We vary the simulation budget n_{sim}, and run the ABC methods 100 times for each n_{sim} with $\epsilon = 10\%$ (meaning the available simulations are different for each run). The accuracy of the obtained ABC posteriors is assessed by estimating the KL divergence between them and a reference ABC posterior using Eq. (10), as the likelihood is intractable. The reference ABC posterior is obtained using linear-ABC with $n_{\text{sim}} = 10,000$ and $\epsilon = 1\%$.

Results The results are shown in Fig. 5. We observe that the proposed HITL-ABC method outperforms the various regression-ABC methods along with Barnes' method for low values of simulation budget n_{sim}. For $n_{\text{sim}} = 400$ and above, the performance of HITL-ABC and linear-ABC is at par. As n_{sim} increases, the KL divergence values of regression-ABC methods decrease, indicating improved performance. Amongst the regression-ABC methods, linear-ABC performs the best, followed by ridge-ABC and neural-ABC for most values of n_{sim}.

For low values of n_{sim}, KL divergence estimates exhibit larger variance, leading to wrongly selecting non-informative statistics in the Barnes' method. However, in HITL-ABC, the lack of available simulations is compensated by the expert feedback, resulting in similar KL divergence values for each n_{sim}. This can be seen from Table 1, where the average number of required feedback increases as n_{sim} decreases. Moreover, we see that maximizing the utility in Eq. (8) is more efficient in terms of yielding the least number of feedback than a random query acquisition strategy.

Table 1. Average number of expert feedback required in the low-simulation regime experiment as a function of simulation budget.

n_{sim}	200	250	300	350	400	450
HITL-ABC	**10.1**	**8.5**	**8.3**	**6.3**	**6.0**	**6.3**
Random	13.8	13.6	13.4	13.3	13.1	13.4

Table 2. Number of times the optimal set of summary statistics, i.e., just the sample mean and sample variance, were selected out of 100 runs by the HITL-ABC, for varying values of hyperparameters π and ρ.

$\rho = 0.5$		$\pi = 0.95$	
π	$(\hat{\mu}, \hat{\sigma}^2)$	ρ	$(\hat{\mu}, \hat{\sigma}^2)$
1.0	100%	0.2	92%
0.95	89%	0.3	91%
0.9	72%	0.4	91%
0.85	70%	0.6	94%
0.8	50%	0.7	90%
0.75	27%	0.8	95%

4.3. Sensitivity to hyperparameter setting

Setting Finally, we perform a hyperparameter sensitivity analysis on a toy problem of estimating the parameters $\theta = (\mu, \sigma^2)$ of a Gaussian distributed random variable $y_1, \ldots, y_{n_{\text{obs}}} \sim \mathcal{N}(\mu, \sigma^2)$. In this case, the sample mean $\hat{\mu}$ and the sample variance $\hat{\sigma}^2$ are sufficient statistics for inferring θ. Additionally, we include the range $(\max_i y_i - \min_i y_i)$ and two uninformative statistics $u_1, u_2 \sim \mathcal{U}(0,1)$ in the pool of statistics, i.e., $\mathcal{S} = \{\hat{\mu}, \hat{\sigma}^2, \text{range}, u_1, u_2\}$. We set the true parameter value to

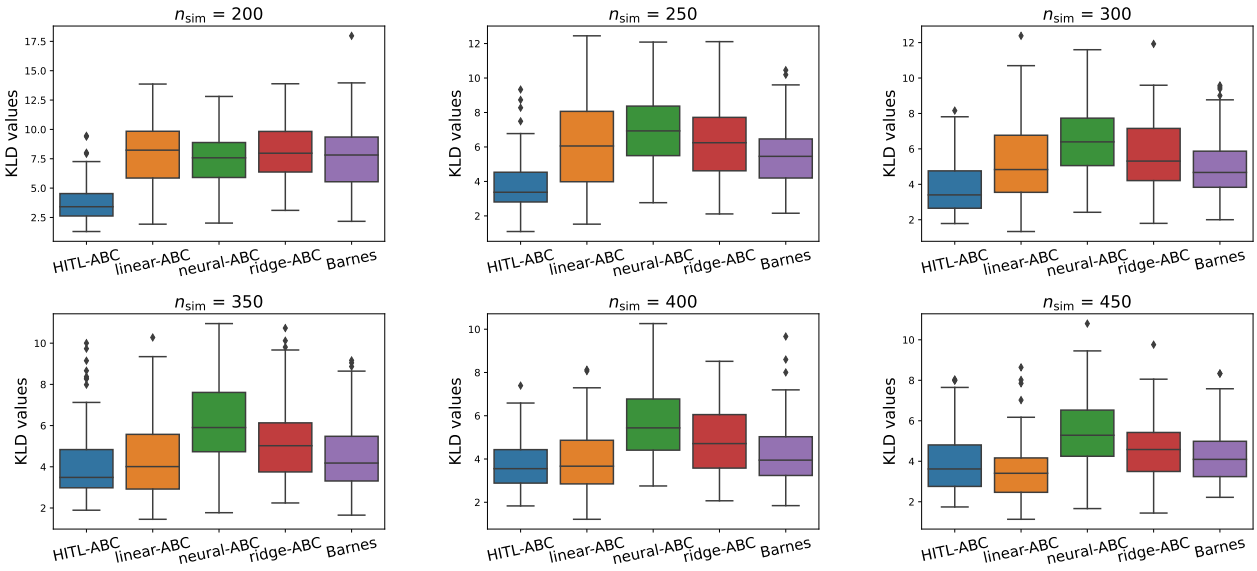


Figure 5. The proposed HITL-ABC outperforms the other regression-ABC methods which do not involve experts, on low-simulation regimes ($n_{\text{sim}} \leq 350$) and performs on-par with larger numbers of simulations. Box plots of KL divergence values between ABC posteriors from different methods at varying n_{sim} and a reference ABC posterior obtained with $n_{\text{sim}} = 10,000$. Lower values of KL divergence indicate better posterior characterization.

$\theta_{\text{true}} = (0, 2)$ and prior to $\mathcal{U}([-5, 5] \times [0, 5])$. The ABC method is run with $n_{\text{obs}} = 500$, $n_{\text{sim}} = 2000$, and $\epsilon = 5\%$.

Results We vary the values of ρ and π, and report the number of times only the sufficient statistics $(\hat{\mu}, \hat{\sigma}^2)$ are selected out of 100 runs in Table 2. Firstly, we observe that HITL-ABC is able to pick the sufficient statistics each time in case of a noiseless feedback ($\pi = 1$). As expected, when the value of π decreases, the sufficient statistics are picked less often. Additionally, we observe that varying ρ barely has any effect on the output of the algorithm. Finally, keeping $\pi = 0.95$ and $\rho = 0.5$ fixed, we vary the value of the stopping criterion δ. We report the average number of queries to the expert over 100 runs in Table 3. As the value of δ increases, the average number of feedback decreases. This could also serve as a rule of thumb on how to set δ, which could depend on how much the expert wishes to be involved, i.e., the maximum number of times they want to be queried.

Table 3. Average number of expert feedback required in the Gaussian example w.r.t. the stopping criterion δ.

δ	Avg. no. of feedback.
0.02	3.04
0.04	2.49
0.06	2.24
0.08	2.18
0.10	2.17

5. Conclusion

In this paper, we introduced the first ABC method that actively leverages domain knowledge from experts in order to select summary statistics. Involving the experts in the ABC method gives us the opportunity to handle misspecified models, something the existing methods fail in. With fairly limited effort from the expert (answering yes/no when presented with a few statistics), we are able to outperform the regression-ABC methods in situations where the simulation budget is low. This simple binary feedback could potentially be scaled to include multiple experts, albeit at the cost of added complexity to determine which expert to ask feedback from. The method also acts as an assistant for the experts to try out different statistics without much effort, however, the usefulness of this method as an AI assistant is a topic for future studies. Finally, there is avenue for further research on extending other likelihood-free inference methods to be amenable to expert's feedback.

Limitations Our method inherits the limitations of all the greedy statistics selection ABC methods, i.e., 1) due to the step-wise selection approach adopted in our method, there is no guarantee of converging to the best subset of statistics, and the method may only converge to a local optimum; and 2) applying it in combination with computationally expensive ABC methods such as ABC-MCMC (Marjoram

et al., 2003) or ABC-SMC (Beaumont et al., 2009) can be infeasible. Lastly, it might be tempting to propose eliciting feedback about a statistic by always showing the posteriors before and after including it. However, that runs the risk that the users may amplify noise in the statistics, especially in low-simulation regimes, if they are not careful. Using so-called "posterior elicitation" and inferring the priors indirectly may then be helpful (Daee et al., 2018).

Acknowledgements

This work was supported by the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence FCAI). SK was also supported by the UKRI Turing AI World-Leading Researcher Fellowship, EP/W002973/1. We acknowledge the computational resources provided by the Aalto Science-IT Project from Computer Science IT.

References

Adeogun, R., Larsen, C., Sand, D., Bovbjerg, H., Fisker, P., and Gjerde, T. Bayesian Synthetic Likelihood for Calibration of Stochastic Radio Channel Model. In *Proceedings of the IEEE Vehicular Technology Conference (VTC)*, 2021.

Aeschbacher, S., Beaumont, M. A., and Futschik, A. A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics*, 192(3):1027–1047, 2012.

Akeret, J., Refregier, A., Amara, A., Seehars, S., and Hasner, C. Approximate Bayesian computation for forward modeling in cosmology. *Journal of Cosmology and Astroparticle Physics*, 2015(08):043–043, 2015.

Barnes, C. P., Filippi, S., Stumpf, M. P. H., and Thorne, T. Considerate approaches to constructing summary statistics for ABC model selection. *Statistics and Computing*, 22(6):1181–1197, 2012.

Beaumont, M. A. Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):379–406, 2010.

Beaumont, M. A. Approximate Bayesian computation. *Annual Review of Statistics and Its Application*, 6(1):379–403, 2019.

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

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, 2009.

Bentley, J. L. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18 (9):509–517, 1975.

Bharti, A., Adeogun, R., and Pedersen, T. Learning parameters of stochastic radio channel models from summaries. *IEEE Open Journal of Antennas and Propagation*, 1:175–188, 2020.

Bharti, A., Briol, F.-X., and Pedersen, T. A general method for calibrating stochastic radio channel models with kernels. *IEEE Transactions on Antennas and Propagation*, pp. 1–1, 2021.

Bi, J., Shen, W., and Zhu, W. Random forest adjustment for approximate Bayesian computation. *Journal of Computational and Graphical Statistics*, 0(0):1–10, 2021.

Blum, M. G. Choosing the summary statistics and the acceptance rate in approximate Bayesian computation. In *Proceedings of COMPSTAT*, pp. 47–56, 2010.

Blum, M. G. Regression approaches for approximate Bayesian computation. *arXiv:1707.01254*, July 2017.

Blum, M. G. B. and François, O. Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, 20(1):63–73, 2010.

Blum, M. G. B., Nunes, M. A., Prangle, D., and Sisson, S. A. A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28(2), 2013.

Chaloner, K. and Verdinelli, I. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Chen, Y., Zhang, D., Gutmann, M., Courville, A., and Zhu, Z. Neural Approximate Sufficient Statistics for Implicit Models. *arXiv:2010.10079*, 2020.

Csilléry, K., François, O., and Blum, M. G. B. abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, 3(3):475–479, 2012.

Daee, P., Peltola, T., Soare, M., and Kaski, S. Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. *Machine Learning*, 106(9-10): 1599–1620, 2017.

Daee, P., Peltola, T., Vehtari, A., and Kaski, S. User modelling for avoiding overfitting in interactive knowledge elicitation for prediction. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, pp. 305–310, 2018.

Drovandi, C. C. and Pettitt, A. N. Likelihood-free Bayesian estimation of multivariate quantile distributions. *Computational Statistics & Data Analysis*, 55(9):2541–2556, 2011.

Fearnhead, P. and Prangle, D. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 74(3):419–474, 2012.

Frazier, D. T., Robert, C. P., and Rousseau, J. Model misspecification in approximate Bayesian computation: consequences and diagnostics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2): 421–444, 2020.

Goldsmith, A. *Wireless Communications*. Cambridge University Press, 2005.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer New York, 2009.

Izbicki, R., Lee, A. B., and Pospisil, T. ABC–CDE: Toward approximate Bayesian computation with complex high-dimensional data and limited simulations. *Journal of Computational and Graphical Statistics*, 28(3):481–492, 2019.

Jiang, B. Approximate bayesian computation with kullback-leibler divergence as data discrepancy. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1711–1721, 2018.

Jiang, B., Wu, T.-Y., Zheng, C., and Wong, W. H. Learning summary statistic for approximate Bayesian computation via deep neural network. *Statistica Sinica*, 27(4):1595–1618, 2017.

Joyce, P. and Marjoram, P. Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.

Li, W. and Fearnhead, P. Convergence of regression-adjusted approximate Bayesian computation. *Biometrika*, 105(2):301–318, 2018.

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, 66:66–82, 2017.

Lueckmann, J.-M., Gonçalves, P. J., Bassetto, G., Öcal, K., Nonnenmacher, M., and Macke, J. H. Flexible statistical inference for mechanistic models of neural dynamics. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1289–1299, 2017.

Lueckmann, J.-M., Bassetto, G., Karaletsos, T., and Macke, J. H. Likelihood-free inference with emulator networks. In *Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference*, pp. 32–53, 2019.

Maneewongvatana, S. and Mount, D. M. On the efficiency of nearest neighbor searching with data clustered in lower dimensions. In *Proceedings of the International Conference on Computational Science (ICCS)*, pp. 842–851, 2001.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2011.

Marjoram, P., Molitor, J., Plagnol, V., and Tavare, S. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

Nunes, M. A. and Balding, D. J. On optimal selection of summary statistics for approximate Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

Papamakarios, G. and Murray, I. Fast ϵ -free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1036–1044, 2016.

Perez-Cruz, F. Kullback-leibler divergence estimation of continuous distributions. In *Proceedings of the IEEE International Symposium on Information Theory*, pp. 1666–1670, 2008.

Prangle, D. Summary statistics in approximate bayesian computation. *arXiv:1512.05633*, 2015.

Prangle, D. gk: An R Package for the g-and-k and Generalised g-and-h Distributions. *The R Journal*, 12(1):7, 2020.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798, 1999.

Rainforth, T., Cornish, R., Yang, H., Warrington, A., and Wood, F. On nesting Monte Carlo estimators. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 4267–4276, 2018.

Rayner, G. D. and MacGillivray, H. L. Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing*, 12(1): 57–75, 2002.

Ryan, E. G., Drovandi, C. C., McGree, J. M., and Pettitt, A. N. A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, 84(1):128–154, 2016.

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

Sundin, I., Peltola, T., Micallef, L., Afrabandpey, H., Soare, M., Majumder, M. M., Daee, P., He, C., Serim, B., Havulinna, A., Heckman, C., Jacucci, G., Marttinen, P., and Kaski, S. Improving genomics-based predictions for precision medicine through active elicitation of expert knowledge. *Bioinformatics*, 34(13):i395–i403, 2018.

Turin, G. L., Clapp, F. D., Johnston, T. L., Fine, S. B., and Lavry, D. A statistical model of urban multipath propagation. *IEEE Transactions on Vehicular Technology*, 21(1):1–9, 1972.

Wang, Q., Kulkarni, S. R., and Verdu, S. A nearest-neighbor approach to estimating divergence between continuous random vectors. In *Proceedings of the IEEE International Symposium on Information Theory*, pp. 242–246, 2006.

Wegmann, D., Leuenberger, C., and Excoffier, L. Efficient approximate Bayesian computation coupled with markov chain monte carlo without likelihood. *Genetics*, 182(4): 1207–1218, 2009.

Supplementary Materials

A. Background on Bayesian Experimental Design

Experimental design tackles the question of selecting the most informative experimental design $d \in \mathcal{D}$ to learn about a parameter θ. To this end, we must choose a so-called utility function $U : \mathcal{D} \to \mathbb{R}$ which assesses the worth of design d, and the optimal design is then

$$d^\star = \arg\max_{d \in \mathcal{D}} U(d). \tag{13}$$

Let us assume that experimental design d leads to observation y. In *Bayesian* experimental design (Chaloner & Verdinelli, 1995; Ryan et al., 2016), we are equipped with a probabilistic model $p(y|\theta, d)$ as well as a prior distribution for the parameter of interest $p(\theta)$. A principled utility function from an information-theoretic perspective is the expected Kullback-Leibler (KL) divergence between the future posterior $p(\theta|d, y)$ and the current prior distribution $p(\theta)$:

$$U(d) = \mathbb{E}_{p(y|d)} \left[\mathrm{KL}(p(\theta|d, y)||p(\theta)) \right]. \tag{14}$$

This utility function can equivalently be presented as the so-called expected information gain, that is, the expected reduction in (differential) entropy from the prior to the posterior distributions. Another equivalent definition is the mutual information between y and θ given design d (usually denoted $\mathrm{I}(y; \theta|d)$). A closed-form expression of Eq. (14) is not available in general, and a common estimation strategy consists of Monte Carlo (MC) approximation, which is more precisely a nested MC approximation (Rainforth et al., 2018).

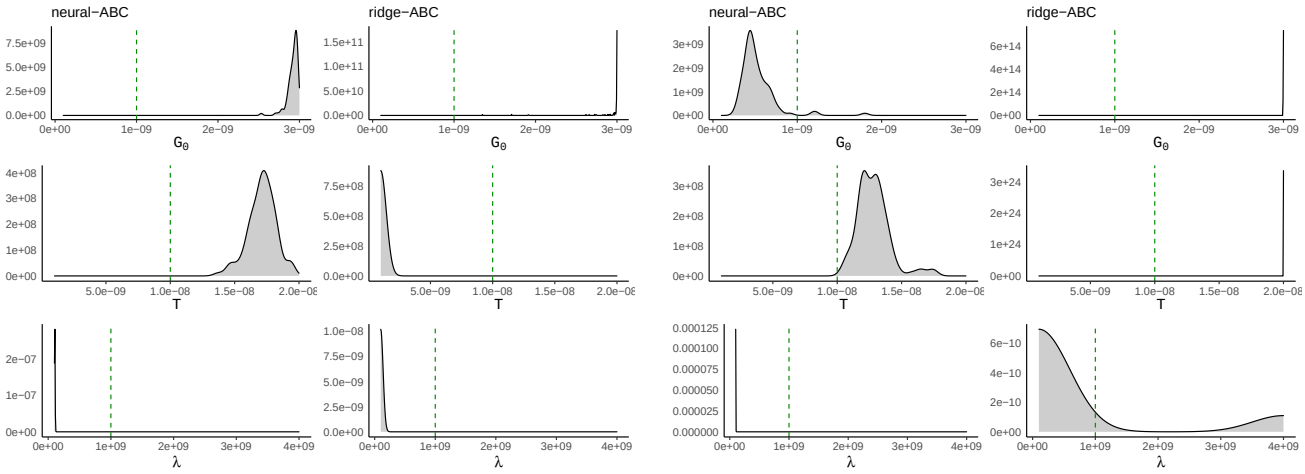
Lastly, we mention that BED is usually applied in a (myopic) sequential way. This means that once the optimal design has been found, and the associated experiment has been run, we proceed to update the posterior distribution of θ, which now acts as the prior distribution for the next step. The new utility is optimized again, and so on and so forth. Experiments are thus run one-by-one. Formally, at iteration $k + 1$, if previously obtained designs $d_1, \dots d_k$ led to observations $y_1, \dots, y_k$, respectively, we have

$$U_{k+1}(d) = \mathbb{E}_{p(y|d, d_{1:k}, y_{1:k})} \left[\mathrm{KL}(p(\theta|d, y, d_{1:k}, y_{1:k})||p(\theta|d_{1:k}, y_{1:k})) \right].$$

B. Additional Results of Misspecification Experiment

In this section, we present additional results on the misspecification experiment conducted on the Turin model (Turin et al., 1972). In Fig. 6, we show the approximate posteriors obtained from neural-ABC and ridge-ABC under model misspecification. Their results are similar to the one obtained from linear-ABC in Fig. 3, as expected, indicating the failure of all regression ABC methods in handling misspecified scenarios.

We also include the results for $\zeta = 1$ in Fig. 7, to demonstrate that even a small degree of misspecification leads to failure in linear-ABC method. On the other hand, HITL-ABC achieves better performance as the misspecified statistic is excluded by the expert. However, we remark that with lower levels of misspecification, it may become difficult for the expert to determine that a statistic is misspecified on seeing the inference results at each iteration of the sequential experiment.



(a) $\zeta = 5$ (b) $\zeta = 10$

Figure 6. Approximate posteriors of the parameters of the radio propagation model obtained from neural-ABC and ridge-ABC at varying levels of misspecification. The dashed green line denotes the true parameter value. Prior is $\mathcal{U}([10^{-10}, 3 \times 10^{-10}] \times [10^{-9}, 2 \times 10^{-8}] \times [10^{8}, 4 \times 10^{9}])$. Settings: $B = 4 \times 10^9$, $n_s = 801$, $n_{\text{obs}} = 300$, $n_{\text{sim}} = 2000$, $\epsilon = 5\%$.

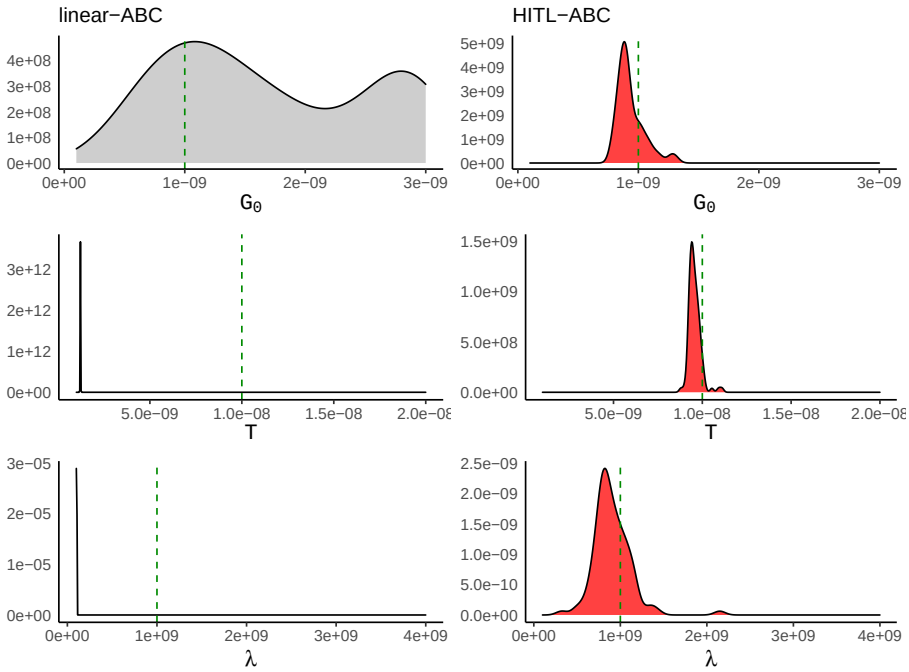


Figure 7. Approximate posteriors of the parameters of the radio propagation model obtained from linear-ABC and HITL-ABC at $\zeta = 1$. The dashed green line denotes the true parameter value. Prior is $\mathcal{U}([10^{-10}, 3 \times 10^{-10}] \times [10^{-9}, 2 \times 10^{-8}] \times [10^{8}, 4 \times 10^{9}])$. Settings: $B = 4 \times 10^9$, $n_s = 801$, $n_{\text{obs}} = 300$, $n_{\text{sim}} = 2000$, $\epsilon = 5\%$.