Informative and adaptive distances and summary statistics in sequential approximate Bayesian computation

Yannik Schälte [1,2,3] and Jan Hasenauer [1,2,3,*]

[1] Faculty of Mathematics and Natural Sciences, Rheinische Friedrich-Wilhelms-Universität Bonn, 53115 Bonn, Germany

[2] Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany

[3] Center for Mathematics, Technische Universität München, 85748 Garching, Germany

[*] To whom correspondence should be addressed (jan.hasenauer@uni-bonn.de)

Abstract

Calibrating model parameters on heterogeneous data can be challenging and inefficient. This holds especially for likelihood-free methods such as approximate Bayesian computation (ABC), which rely on the comparison of relevant features in simulated and observed data and are popular for otherwise intractable problems. To address this problem, methods have been developed to scale-normalize data, and to derive informative low-dimensional summary statistics using inverse regression models of parameters on data. However, while approaches only correcting for scale can be inefficient on partly uninformative data, the use of summary statistics can lead to information loss and relies on the accuracy of employed methods.

In this work, we first show that the combination of adaptive scale normalization with regression-based summary statistics is advantageous on heterogeneous parameter scales. Second, we present an approach employing regression models not to transform data, but to inform sensitivity weights quantifying data informativeness. Third, we discuss problems for regression models under non-identifiability, and present a solution using target augmentation. We demonstrate improved accuracy and efficiency of the presented approach on various problems, in particular robustness and wide applicability of the sensitivity weights. Our findings demonstrate the potential of the adaptive approach. The developed algorithms have been made available in the open-source Python toolbox pyABC.

1 Introduction

Mechanistic models are important tools in systems biology and many other research areas to describe and study real-world systems, allowing to understand underlying mechanisms [Gershenfeld and Gershenfeld, 1999, Kitano, 2002]. Commonly, they are subject to parameters that need to be estimated by comparison of model outputs to observed data [Tarantola, 2005]. The Bayesian framework allows doing so by combining the likelihood of data and prior information on parameters. However, for complex stochastic models, e.g. used in systems biology to describe multi-cellular systems, evaluating the likelihood is often computationally infeasible [Hasenauer et al., 2015, Tavaré et al., 1997]. Therefore, likelihood-free methods such as approximate Bayesian computation (ABC) have been developed [Beaumont et al., 2002, Pritchard et al., 1999]. In a nutshell, in ABC likelihood evaluation is circumvented by simulating data, and accepting these depending on their proximity to observed data, according to a distance measure and an acceptance threshold. This way, it generates samples from an approximation to the posterior distribution. ABC is frequently combined with a sequential Monte-Carlo scheme (ABC-SMC) [Del Moral et al., 2006, Sisson et al., 2007], which allows gradually reducing the acceptance threshold while maintaining high acceptance rates.

ABC relies on the comparison of relevant features in simulated and observed data. Prangle [2017] demonstrate superior performance of distances that adaptively weight model outputs to normalize contributions on different scales, exploiting the structure of ABC-SMC algorithms. In Schälte et al. [2021], we extend this approach to outlier-corrupted data. However, an implicit assumption of scale normalization is that all model outputs are similarly informative of the parameters. It can worsen performance, e.g. when inflating the impact of data points underlying only background noise. Therefore, it would be preferable to either only consider informative statistics, or to account for informativeness in the weighting scheme.

Especially for noise-corrupted high-dimensional data, often lower-dimensional summary statistics are employed [Blum et al., 2013]. Various methods to construct such statistics have been developed, e.g. via subset selection or auxiliary likelihoods [Drovandi et al., 2011, Nunes and Balding, 2010]. A popular line of approaches uses as statistics the outputs of inverse regression models of parameters on simulated data [Borowska et al., 2021, Fearnhead and Prangle, 2012, Jiang et al., 2017]. Such regression models can be heuristically motivated as summarizing the information in the data in a single value per parameter. In addition, Fearnhead and Prangle [2012] argue that the resulting summary statistics effectively approximate posterior means, which conserves the true posterior mean in the ABC analysis.

To evaluate proximity of regression-based statistics, e.g. Euclidean distances have been used, or weighted Euclidean distances using weights based on calibration samples [Fearnhead and Prangle, 2012]. However, here essentially the same problems apply that motivated the use of adaptive weighting [Prangle, 2017], shifted from the level of data to the level of parameters, or regression approximations thereof. In fact, the approach by Prangle [2017] is particularly applicable to regression-based statistics, as all outputs are informative. A further problem with regression-based statistics is that an inverse mapping may not always exist, e.g. when parameters are not globally identifiable.

In this work, we present two approaches combining the concepts of adaptive distances and regression models. First, we integrate summary statistics learning in an ABC-SMC framework with scale-normalizing adaptive distances. Second, the focus of this work, we employ regression models not to transform data, but in order to inform additional sensitivity weights that account for informativeness. Moreover, we discuss the problem of non-identifiability of the inverse mapping, and present a solution using augmented regression targets. On a dedicated test problem exhibiting multiple problematic features such as partly uninformative data, heterogeneous data and parameter scales, and non-identifiability, we demonstrate how both scale-normalizing distances [Prangle, 2017], and regression-based summary statistics [Fearnhead and Prangle, 2012] fail to approximate the true posterior distribution. Then, we demonstrate substantially improved performance of the newly introduced approaches. We evaluate the proposed methods on further test problems, including a systems biology application example and outlier-corrupted data, demonstrating in particular robustness as well as wide applicability of the sensitivity-weighted distance.

2 Methods

2.1 Background

In this section, we give required background knowledge on the underlying methodology.

2.1.1 Approximate Bayesian computation

In Bayesian inference, the likelihood $\pi(y_{\text{obs}}|\theta)$ of observing data $y_{\text{obs}} \in \mathbb{R}^{n_y}$ under model parameters $\theta \in \mathbb{R}^{n_\theta}$ is combined with prior information $\pi(\theta)$, giving the posterior $\pi(\theta|y_{\text{obs}}) \propto \pi(y_{\text{obs}}|\theta) \cdot \pi(\theta)$. We assume that while numerical evaluation of $\pi(y_{\text{obs}}|\theta)$ is infeasible, the model is generative, i.e. allows to simulate data $y \ \sim \pi(y|\theta)$. The core principle of ABC consists of three steps [Pritchard et al., 1999]:

1. Sample parameters $\theta \sim \pi(\theta)$.
2. Simulate data $y \sim \pi(y|\theta)$.
3. Accept (θ, y) if $d(y, y_{\text{obs}}) \leq \varepsilon$.

Here, the distance $d : \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \to \mathbb{R}_{\geq 0}$ compares simulated and observed data, and $\varepsilon \geq 0$ an acceptance threshold. This is repeated until sufficiently many, say N, particles have been accepted. For high-dimensional data, the comparison is often in terms of summary statistics $s : \mathbb{R}^{n_y} \to \mathbb{R}^{n_s}$, as $d(s(y), s(y_{\text{obs}})) \leq \varepsilon$, with $d : \mathbb{R}^{n_s} \times \mathbb{R}^{n_s} \to \mathbb{R}_{\geq 0}$ and typically $n_s \ll n_y$. Denoting $\pi(s|\theta) \propto \int I[s(y) = s]\pi(y|\theta)\, dy$ the intractable summary statistics likelihood with I the indicator function, and $s_{\text{obs}} = s(y_{\text{obs}})$, the population of accepted particles then constitutes a sample from the approximate posterior distribution

$$\pi_{\text{ABC}}(\theta|s_{\text{obs}}) \propto \int I[d(s, s_{\text{obs}}) \leq \varepsilon]\pi(s|\theta)\, ds \cdot \pi(\theta),$$

where $\pi_{\text{ABC}}(s_{\text{obs}}|\theta) \propto \int I[d(s, s_{\text{obs}}) \leq \varepsilon]\pi(s|\theta)\, ds$ can be interpreted as an approximation to the likelihood.

Algorithm 1 A basic ABC-SMC algorithm.

```
initialize $\varepsilon_1$ via calibration samples, let $g_1(\theta) = \pi(\theta)$
for $t = 1, \dots, n_t$ do
  while less than $N$ acceptances do
    sample parameter $\theta \sim g_t(\theta)$
    simulate data $y \sim \pi(y|\theta)$
    accept $\theta$ if $d(y, y_{\text{obs}}) \leq \varepsilon_t$
  end while
  compute weights $w_i^t = \frac{\pi(\theta_i^t)}{g_t(\theta_i^t)}$, for accepted parameters $\{\theta_i^t\}_{i\leq N}$
  normalize weights $W_i^t = w_i^t / \sum_j w_j^t$
  update $g_{t+1}$ and $\varepsilon_{t+1}$ based on particles from generation $t$
end for
output: weighted samples $\{(\theta_i^{n_t}, W_i^{n_t})\}_{i\leq N}$
```

For $\varepsilon \to 0$, it holds under mild assumptions that $\pi_{\text{ABC}}(\theta|s(y_{\text{obs}})) \to \pi(\theta|s(y_{\text{obs}})) \propto \pi(s(y_{\text{obs}})|\theta)\pi(\theta)$ in an appropriate sense [Barber et al., 2015]. Compared to likelihood-based sampling, ABC introduces two approximation errors [Sisson et al., 2018, Chapter 1]. First, it accepts not only particles with $y = y_{\text{obs}}$, which occur for continuous models with probability zero, but also proximate ones according to d. Second, only for sufficient statistics, $\pi(\theta|s_{\text{obs}}) \equiv \pi(\theta|y_{\text{obs}})$, is the original posterior recovered in the approximate limit $\varepsilon \to 0$. In practice, s is however usually insufficient, only capturing essential information about y in a low-dimensional representation.

2.1.2 Sequential importance sampling

As the above vanilla ABC algorithm, also called ABC-Rejection, exhibits a trade-off between decreasing the acceptance threshold ε to improve the posterior approximation, and maintaining high acceptance rates, it is frequently combined with a sequential Monte-Carlo (SMC) importance sampling scheme [Del Moral et al., 2006, Sisson et al., 2007]. In ABC-SMC, a series of particle populations $P_t = \{(\theta_i^t, y_i^t, w_i^t)\}_{i\leq N}$, $t = 1, \dots, n_t$, are generated, with acceptance thresholds $\varepsilon_1 > \ldots > \varepsilon_{n_t}$, targeting successively better posterior approximations. Particles for generation t are sampled from a proposal distribution $g_t(\theta) \gg \pi(\theta)$ based on the previous generation's accepted particles P_{t-1}, e.g. via a kernel density estimate, only initially $g_1(\theta) = \pi(\theta)$. The importance weights w_i^t are the corresponding non-normalized Radon-Nikodym derivatives, $w_t(\theta) = \pi(\theta)/g_t(\theta)$.

The underlying ABC-SMC algorithm (Algorithm 1) used throughout this work is based on Toni and Stumpf [2010], using an adaptive threshold scheme based on the median of distances in the previous generation [Drovandi and Pettitt, 2011] and multivariate normal proposal distributions with adaptive covariance matrix [Filippi et al., 2013], see Klinger and Hasenauer [2017], Klinger et al. [2018] for details. There exist various ABC-SMC sampler variants [Sisson et al., 2018], e.g. in some cases different threshold schemes [Silk et al., 2013] or proposal distributions [Filippi et al., 2013] may be beneficial. The distances and summary statistics presented in this work are mostly independent of the sequential sampler specifics.

2.1.3 Adaptive distances

A common choice of distance d is a weighted Minkowski distance

$$d(y, y_{\text{obs}}) = \|r \cdot (y - y_{\text{obs}})\|_p = \left(\sum_{i_y=1}^{n_y} \left| r_{i_y} \cdot (y_{i_y} - y_{\text{obs},i_y}) \right|^p \right)^{1/p}, \tag{1}$$

with $p \geq 1$ and weights r_{i_y}. Frequently, simply unit weights $r = 1$ are used (e.g. Borowska et al. [2021], Fearnhead and Prangle [2012], Jiang et al. [2017], Toni and Stumpf [2010]). However, model outputs can be and vary on different scales, in which case highly variable ones dominate the acceptance decision. This can be corrected for by the choice of weights r_{i_y} in (1), commonly as inversely proportional to measures of variability,

$$r_{i_y} = 1/\sigma_{i_y}, \tag{2}$$

with σ_{i_y} e.g. given via the median absolute deviation (MAD) from the sample median [Csilléry et al., 2012]. To define weights, calibration samples can be used (e.g. Beaumont et al. [2002], Fearnhead and Prangle [2012]). However, Prangle [2017] demonstrate that in an ABC-SMC framework, the relative variability of model outputs in later generations can differ considerably from pre-calibration. Thus, they propose an iteratively updated distance d_t, defining weights for generation t based on all samples generated in generation $t - 1$.

In Schälte et al. [2021], we demonstrate the L2 norm used in (1) in Prangle [2017] to be sensitive to data outliers, and show an L1 norm to be more robust on both outlier-corrupted and outlier-free data. To further reduce the impact of outliers, we complement MAD, as a measure of sample variability, by the median absolute deviation to the observed value, as a measure of deviation, giving a normalization term PCMAD (see Schälte et al. [2021] for details).

2.1.4 Regression-based summary statistics

The comparison of simulations and data in ABC is often in terms of low-dimensional, informative summary statistics. The "semi-automatic ABC" approach by Fearnhead and Prangle [2012] uses the outputs of a regression model $s : \mathbb{R}^{n_y} \to \mathbb{R}^{n_\theta}$, predicting parameters from simulated data (Figure 1):

1. In an ABC pilot run, determine a high-density posterior region H.
2. Generate a population $P = \{(\theta_i, y_i)\}_{i \leq \tilde{N}} \sim \pi(y|\theta)I[\theta \in H]$, for some $\tilde{N} \in \mathbb{N}$.
3. Train a regressor model $s : \mathbb{R}^{n_y} \to \mathbb{R}^{n_\theta}$, $y \mapsto \theta$, on P.
4. Run the actual ABC analysis using s as summary statistics.

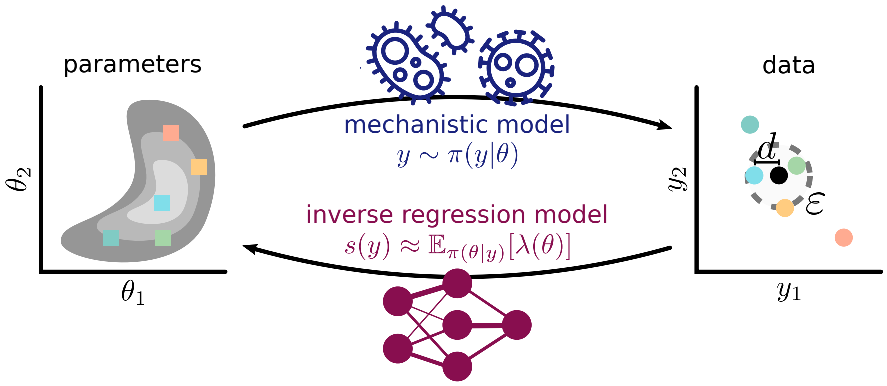

Figure 1: Concept visualization: While, given parameters θ, the mechanistic model $\pi(y|\theta)$ simulates data y, which are then compared to the observed data (black) via some distance d and threshold ε, we employ regression models to learn an inverse mapping s, to either construct summary statistics, or define sensitivity weights for distance calculation.

In step 4, the distance operates on $s(y)$. Step 1 aims to find a good training region, and can be skipped for informative priors [Fearnhead and Prangle, 2012]. In Borowska et al. [2021], $H = [0.5\tilde{\theta}, 2\tilde{\theta}]$ is used, around a literature value $\tilde{\theta}$, based on manual experimentation, which is in practice only applicable if reliable references exist. In Jiang et al. [2017], step 1 is omitted, using the prior directly, in one case constrained to an identifiable region. In step 3, Fearnhead and Prangle [2012] employ a linear regression (LR) model on potentially augmented data. Jiang et al. [2017] and Borowska et al. [2021] respectively use neural networks (NN) and Gaussian processes (GP) instead, aiming at a more accurate description of non-linear relationships, and further process automation. The sufficient performance of LR observed in Fearnhead and Prangle [2012] may be due to the substantial time spent in the pilot run, identifying a high-density region where a linear approximation suffices, while e.g. Jiang et al. [2017] observe a clearly better posterior approximation with NN, and Borowska et al. [2021] better model predictions with GP.

A theoretical justification of regression-based summary statistics is that the regression model serves as an approximation to the posterior mean, $s(y) \approx \mathbb{E}_{\pi(\theta|y)}[\theta]$, using which as statistic ensures that the ABC posterior approximation recovers the actual posterior mean as $\varepsilon \to 0$, see Fearnhead and Prangle [2012], Jiang et al. [2017], or the Supplementary Information, Theorem 1.

2.2 Adaptive and informative regression-based distances and summary statistics

In this section, we describe the novel methods introduced in this work.

2.2.1 Integrating summary statistics learning and adaptive distances

In previous studies, the regression approach from Section 2.1.4 was used together with uniform, or on a previous run pre-calibrated, distance weights [Borowska et al., 2021, Fearnhead and Prangle, 2012, Jiang et al., 2017]. However, to the regression model outputs, approximating underlying parameters, the same problems apply that motivated the adaptive approach in Prangle [2017]: Parameters varying on larger scales dominate the analysis without scale adjustment, with potentially changing levels of variability over ABC-SMC generations.

We propose to combine the regression-based summary statistics from Section 2.1.4 with the weight adaptation from Section 2.1.3. The regression model can be pre-trained as previously done. Here, we however suggest to increase efficiency and automation by integrating the training into the actual ABC-SMC run (Algorithm 2). We begin by using an adaptively scale-normalized distance on the full model outputs. Then, in a generation $t_{\text{train}} \geq 1$, the regression model $s : \mathbb{R}^{n_y} \to \mathbb{R}^{n_\theta}$ is trained on all particles $\{(\theta_i^{t_{\text{train}}-1}, y_i^{t_{\text{train}}-1})\}_{i \leq \tilde{N}}$, $\tilde{N} \geq N$, generated in the previous generation. From $t \geq t_{\text{train}}$ onward, the regression model outputs $s(y)$ are used as summary statistics, also here using a scale-normalized distance with iteratively adjusted weights. Like for adaptive distance weight calculation, the training samples also include rejected ones. First, this increases the training sample size, and second, it gives a representative sample from the joint distribution of data and parameters, focusing on a high-density region, but not confined to $y \approx y_{\text{obs}}$.

The delay of regression model training until after a few generations serves to focus on a high-density region, similar to Fearnhead and Prangle [2012], such that simpler regression models provide a sufficient description. While Fearnhead and Prangle [2012] update the prior to a typical range of values observed in the pilot run, we consider the prior as part of the problem formulation, and thus do not update it. In generations $t \geq t_{\text{train}}$ the proposal distributions g_t will usually anyway mostly suggest values within the training domain range.

2.2.2 Regression-based sensitivity weights

The adaptive scale-normalized distance approach from Prangle [2017], Schälte et al. [2021] is, operating on the full data without summary statistics, not ideal if data points are not similarly informative. The regression approach from Section 2.1.4 is one solution to focus on informative statistics. However, it performs a complex transformation of the model outputs, which can hinder interpretation, and perform badly if the regression model is inaccurate. In this section, we present an alternative approach, using the regression model to inform additional weights on the full data, instead of constructing summary statistics. The idea is to weight a data point by how informative it is of underlying parameters. We quantify informativeness via the sensitivity of how much the posterior expectation of parameters, or transformations thereof, given observed data y_{obs}, would vary under data perturbations. As in Section 2.2.1, we use a regression model to describe the inverse mapping from data to parameters.

Specifically, before a generation t_{train}, we train a regression model $s : \mathbb{R}^{n_y} \to \mathbb{R}^{n_\theta}$ on samples from the previous generation. As regression model inputs, we use normalized simulations $y/\sigma_{t_{\text{train}}}$, with $\sigma_{t_{\text{train}}}$ the measure of scale used for distance scale normalization, e.g. MAD. Further, we z-score normalize regression model targets θ, in order to render the model scale-independent. Then, we calculate the sensitivity matrix

$$S = \nabla_y s(y_{\text{obs}}) \in \mathbb{R}^{n_y \times n_\theta} \tag{3}$$

at the observed data. To robustly approximate derivatives, we employ central finite differences with

Algorithm 2 ABC-SMC algorithm with regression-based summary statistics or sensitivity-weighted distances.

```
initialize $\varepsilon_1$, $\sigma^1_{i_y}$ via calibration samples, let $g_1(\theta) = \pi(\theta)$
for $t = t_{\text{train}}, \dots, n_t$ do
  while less than $N$ acceptances do
    sample parameter $\theta \sim g_t(\theta)$
    simulate data $y \sim \pi(y|\theta)$
    if $t < t_{\text{train}}$ then
      accept $\theta$ if $d_t(y, y_{\text{obs}}) \leq \varepsilon_t$, where $d_t$ uses scale weights $r^t_{i_y} = 1/\sigma^t_{i_y}$
    else if $s$ is used as summary statistics then
      accept $\theta$ if $d_t(s(y), s(y_{\text{obs}})) \leq \varepsilon_t$, where $d_t$ uses scale weights $r^t_{i_s} = 1/\sigma^t_{i_s}$
    else if using $s$ only to define sensitivity weights $q_{i_y}$ then
      accept if $d_t(y, y_{\text{obs}}) \leq \varepsilon_t$, where $d_t$ uses scale and sensitivity weights $r^t_{i_y} = q^t_{i_y}/\sigma^t_{i_y}$
    end if
  end while
  compute importance weights $w^t_i = \frac{\pi(\theta^t_i)}{g_t(\theta^t_i)}$, for accepted parameters $\{\theta^t_i\}_{i \leq N}$
  normalize importance weights $W^t_i = w^t_i / \sum_j w^t_j$
  if $t + 1 == t_{\text{train}}$ then
    train regression model $s$ on all particles from generation $t$
    if using $s$ to weight model outputs then
      define sensitivity weights $q_1, \dots, q_{n_y}$ via $s$
    end if
  end if
  update $g_{t+1}$ and $\varepsilon_{t+1}$ based on particles from generation $t$
  update inverse scale weights $\sigma^{t+1}_{i_y}$ or $\sigma^{t+1}_{i_s}$ based on all particles from generation $t$
end for
output: weighted samples $\{(\theta^{n_t}_i, W^{n_t}_i)\}_{i \leq N}$
```

automatic step size control [Raue et al., 2013]. We define the *sensitivity weight* of model output i_y as

$$q_{i_y} = \sum_{i_\theta=1}^{n_\theta} \frac{\left|S_{i_y i_\theta}\right|}{\sum_{j_y=1}^{n_y} \left|S_{j_y i_\theta}\right|}, \tag{4}$$

i.e. as the sum over the absolute sensitivities of all parameters with respect to the model output, normalized per parameter to level their impact. The normalization can be omitted, but yields more conservative weights, accounting for the fact that the regression model may be inaccurate, by more evenly distributed weights when all sensitivities with respect to some parameters are small.

The final weight used in the distance (1) is then given as the product of scale weight (2) and sensitivity weight (4),

$$r_{i_y} = q_{i_y}/\sigma_{i_y}, \tag{5}$$

with here σ_{i_y} e.g. again given via MAD, or, also taking bias into account, PCMAD. This separate treatment of scale and sensitivity weights allows to e.g. include the error correction from Schälte et al. [2021] in the scale correction, but not in the normalized data used for regression model training, which would lead to inversely re-scaled sensitivities. Thus, we can simultaneously account for informativeness and outliers. As long as $r_{i_y} \neq 0$ for all weights, the original posterior $\pi(\theta|y_{\text{obs}})$ can be conceptually recovered for $\varepsilon \to 0$ [Barber et al., 2015, Prangle, 2017], i.e. no information is lost, unlike for insufficient summary statistics, while practical convergence is clearly weight-dependent.

2.2.3 Optimal summary statistics to recover distribution features

A problem with inverse regression models of parameters on data is that such a mapping may not exist. For example, consider a quadratic model $y \sim \mathcal{N}(\theta^2, 0.1^2)$, with prior $\theta \sim U[-1, 1]$, and observed data $y_{\text{obs}} = 0.7$. As an inverse mapping $y \mapsto \theta$ does not exist globally, a regression model $s : y \mapsto \theta$ cannot extract a meaningful relationship. Indeed, the problem is symmetric in θ, such that the posterior mean is $\mathbb{E}_{\pi(\theta|y)}[\theta] = 0$, using which as summary statistic as in Blum et al. [2013], Fearnhead and Prangle [2012] would clearly recover the true posterior mean. However, it would fail to describe the posterior shape at all.

A solution is to consider transformations $\lambda(\theta)$ of the parameters, e.g. higher-order moments $s : y \mapsto \lambda(\theta) = (\theta^1, \ldots, \theta^k)$, which may be better described as functions of the data, or identifiable in the first place. In the above example, it suffices to consider θ^2, giving a linear mapping $y \sim \theta^2$ and breaking the symmetry. While the use of parameter transformations as regression model targets is heuristically reasonable, their use can be theoretically further justified: Employing as summary statistics posterior expectations of transformations of the parameters,

$$s(y) = \mathbb{E}_{\pi(\theta|y)}[\lambda(\theta)],$$

allows under mild assumptions to recover the corresponding posterior expectations for $\varepsilon \to 0$,

$$\lim_{\varepsilon \to 0} \mathbb{E}_{\pi_{\text{ABC},\varepsilon}}[\lambda(\Theta)|s(y_{\text{obs}})] = \mathbb{E}[\lambda(\Theta)|Y = y_{\text{obs}}],$$

see Theorem 1 in the Supplementary Information for details.

Obviously, conditional posterior expectations are hardly available in practice. However, we may interpret the above regression-based summary statistics as approximations, aiming at a sufficiently accurate description of the underlying expectations by the regression model. Thus, we propose to use $\lambda(\theta)$ as targets, both for summary statistics (Section 2.2.1), and sensitivity weights (Section 2.2.2).

2.3 Implementation

We implemented all presented methods in the open-source Python package pyABC (`https://github.com/icb-dcm/pyabc`) [Klinger et al., 2018], interfacing particularly scikit-learn regression models [Pedregosa et al., 2011]. The code underlying the application study is on GitHub (`https://github.com/yannikschaelte/study_abc_slad`), a snapshot of code and data on Zenodo (`http://doi.org/10.5281/zenodo.5522919`).

3 Results

We evaluated the performance of the proposed methods on various test problems.

3.1 Distances and summary statistics

As distance to compare model outputs or summary statistics, we considered, given its robust performance in Schälte et al. [2021], an L1 norm, with adaptive MAD weights when employing scale-normalization (denoted "Ada.+MAD"). Acceptance in generation t was only based on d_t, but not previous acceptance criteria, for ease of implementation and as for an L1 norm no substantial differences were observed in Schälte et al. [2021].

As regression models, we considered LR and NN. We trained the regression model after 40% of the simulation budget. For comparison, we also considered training the regression model before the initial generation, $t_{\text{train}} = 1$, based on samples from the prior ("Init"). NN models were considered with a single hidden layer of dimension $[(n_y + n_\theta)/2]$, with ReLU activation function, using ADAM stochastic gradient descent for optimization, and early stopping to avoid overfitting, with a 10% validation set. Both regression models were computationally efficient compared to the full ABC-SMC analyses, with run-times on the order of milliseconds (LR) or few seconds (NN).

When employing parameter augmentation (Section 2.2.3), we used the first four moments, $\lambda(\theta) = (\theta^1, \ldots, \theta^4)$ ("P4"). We considered both regression to define summary statistics ("Stat", Section 2.2.1) and sensitivity weights ("Sensi", Section 2.2.2).

For example, L1+Ada.+MAD+StatNN denotes an analysis using consistently an adaptive distance with MAD normalized weights, and using a neural network to construct summary statistics after 40% of the total simulation budget, with regression targets $\lambda(\theta) = \theta$. L1+Ada.+MAD+SensiLR+P4 uses an adaptive distance with scale-normalizing weights via MAD, and a linear model to define further sensitivity weights, with regression targets $\lambda(\theta) = (\theta^1, \ldots, \theta^4)$, and L1+StatLR uses a linear model for summary statistics construction, but uses uniform distance weights.

3.2 Performance on dedicated demonstration problem

To illustrate the different problems addressed in this work, we constructed a demonstration problem with four parameters and five types of data:

- $y_1 \sim \mathcal{N}(\theta_1, 0.1^2)$ is informative of θ_1, with a relatively wide prior $\theta_1 \sim U[-7, 7]$,
- $y_2 \sim \mathcal{N}(\theta_2, 100^2)$ is informative of θ_2, with prior $\theta_2 \sim U[-700, 700]$,
- $y_3 \sim \mathcal{N}(\theta_3, 4 \cdot 100^2)^{\otimes 4} \in \mathbb{R}^4$ is informative of θ_3, with prior $\theta_3 \sim U[-700, 700]$,
- $y_4 \sim \mathcal{N}(\theta_4^2, 0.1^2)$ is informative of θ_4, with prior $\theta_4 \sim U[-1, 1]$, but quadratic in the parameter,
- $y_5 \sim \mathcal{N}(0, 10)^{\otimes 10} \in \mathbb{R}^{10}$ is uninformative.

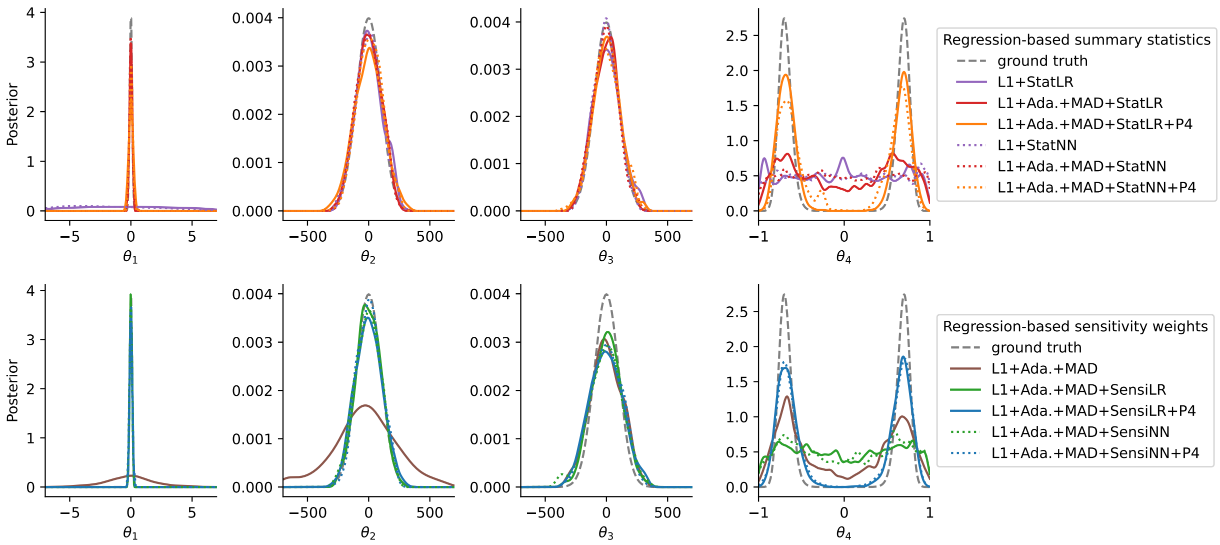

Figure 2: ABC marginal posterior approximations obtained using regression-based summary statistics (top, "Stat") or sensitivity weights (bottom, "Sensi") on the demonstration problem, using an underlying L1 norm, uniformly weighted, or MAD scale-normalized distance weights ("Ada.+MAD"), using a linear ("LR") or a neural network ("NN") regression model, and in some cases augmented regression targets $\theta^1, \ldots, \theta^4$ ("P4").

The model dynamics are purposely simple, such that inverse mappings can be captured easily. The problem exhibits the following potentially problematic features:

- A substantial part of the data, y_5, is uninformative, such that approaches ignoring data informativeness may converge slower.
- Both data and parameters are on different scales, such that approaches comparing data, or, via regression-based summary statistics, parameters, without normalization focus on large-scale variables. Further, e.g. the prior of θ_1 is relatively wide, preventing pre-calibration.
- y_4 is quadratic in θ_4, such that first-order regression models cannot capture a meaningful relationship.
- While y_2, y_3 are such that the posteriors of θ_2, θ_3 are identical, in solely scale-normalized approaches, the impact of y_4 on the distance value is roughly four times as high as that of y_3, resulting in uneven convergence.

We studied the demonstration problem with synthetic data $y_{\text{obs},1}, y_{\text{obs},2}, y_{\text{obs},3}, y_{\text{obs},5} \equiv 0$, $y_{\text{obs},4} = 0.7$, using a population size of $N = 4e3$ with a total budget of $1e6$ simulations per run. Marginal posterior approximations obtained using selected distances and summary statistics are shown in Figure 2.

Solely scale-normalized distances without informativeness assessment converge slowly

The scale-normalized adaptive distance L1+Ada.+MAD correctly captured all posterior modes and shapes, in particular the bi-modality of θ_4, however with large variances, because the uninformative model outputs y_5 were considered on the same scale as the informative ones (Figure 2 bottom). Further, while the true posteriors of θ_2 and θ_3 coincide, L1+Ada.+MAD assigned a substantially wider variance to θ_2, as only a single model output, y_2, is informative of it, while four are of θ_3, all on the same normalized scale.

Non-scale-normalized distances converge unevenly

The analyses L1+Stat{LR/NN} without scale normalization described the posteriors of θ_2 and θ_3 accurately, which are on the same scale, however yielded substantially wider variances for θ_1 (Figure 2 top), because θ_1, used as regression target, varies on a smaller scale. In contrast, all analyses employing scale normalization described θ_1, θ_2, and θ_3 roughly or almost similarly well, with the exception of L1+Ada.+MAD, as outlined above.

Regression models not accounting for non-identifiability cannot capture posterior

All analyses employing regression models but using the non-augmented regression targets $\lambda(\theta) = \theta$ failed to describe the bi-modal distribution of θ_4, because a global mapping $y_4 \mapsto \theta_4$ does not exist. In comparison, analyses considering higher-order regression targets ("P4") captured the bi-modality, as for this problem a linear mapping $\theta_4^2 \sim y_4$ exists, or a quadratic one $\theta_4^4 \sim y_4^2$.

Novel approaches fit all parameters well

The analyses L1+Ada.+MAD+{Stat{LR/NN}/Sensi{LR/NN}}+P4 combining all methods introduced in this work, i.e. scale normalization, informativeness assessment via regression-based summary statistics or sensitivity weights, and regression target augmentation, provided the overall best description of all posterior marginals, with roughly homogeneously small variances. Advantages of NN over LR were not observed.

Estimates for θ_3 were with L1+Ada.+MAD+Sensi{LR/NN} consistently slightly worse than with L1+Ada.+MAD+Stat{LR/NN}. This can be explained by the latter approaches employing a one-dimensional interpolation of $y_3 \in \mathbb{R}^4$, and thus e.g. an approximation of the sufficient statistic $\frac{1}{4}\sum_{i=1}^4 y_{3,i}$. Meanwhile, approaches that do not transform but only weight, are more subject to random noise. This illustrates that when low-dimensional sufficient statistics exist and are accurately captured, employing explicit dimension reduction can be superior to mere re-weighting.

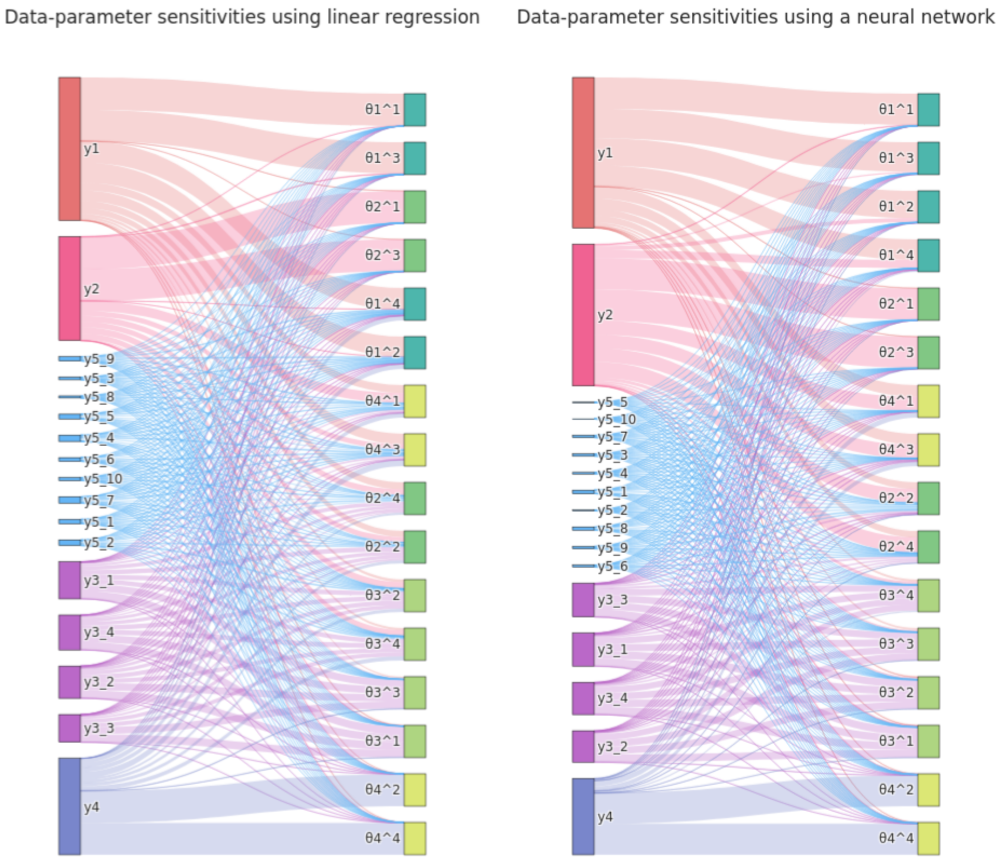

Figure 3: Exemplary normalized absolute data-parameter sensitivities for the demonstration problem, using LR (left) and NN (right), for all regression targets $\theta^1, \ldots, \theta^4$ with $\theta = (\theta_1, \ldots, \theta_4)$ (respectively on the right), with respect to all data coordinates y (respectively on the left). The absolute sensitivity matrix $|S|$ (3) was normalized per regression target, as in (4). The widths of lines connecting data and parameters, and corresponding endpoints, are proportional to their respective values. In particular, the heights of the respective left end-points are proportional to the assigned sensitivity weights q_{i_y} (4). Data types, e.g. $y_{5,1}, \ldots, y_{5,10}$, and parameters with their exponents, e.g. $\theta_1^1, \ldots, \theta_1^4$, are grouped by colors.

Sensitivity weights permit further insights

In Figure 3, normalized absolute sensitivities (4) of parameters with respect to model outputs are visualized. Overall, both regression models captured the relationship of model outputs and parameters well, and assigned large, albeit not completely homogeneous, sensitivity weights to $y_1, \ldots, y_4$, and lower ones to y_5, with roughly $q_1 \approx q_2 \approx \sum_{i=1}^4 q_{3,i}$. The description provided by NN was overall slightly better than LR, assigning lower weights to y_5, and capturing the non-linear mappings $\theta_1^2 \sim y_1$ and $\theta_1^4 \sim y_1$ better. As seen above, LR nevertheless sufficed to yield good posterior approximations. Sensitivities of θ_4^1 and θ_4^3 were, as expected, comparably small with respect to all variables.

The weight assigned to y_4 was roughly half the ones assigned to y_1, y_2 and y_3, because θ_4^1 and θ_4^3 could not be accurately described. Correspondingly, the variance of θ_4 was slightly wider under sensitivity-weighted analyses, compared to using summary statistics (Figure 2). This could be

Table 1: Test model properties: Identifier, short description, number of parameters n_θ and data points n_y, population size N and maximum number of model simulations after which an analysis was terminated.

ID	Description	n_θ	n_y	N	Max. sim.
T1	Conversion reaction ODE model	2	10	1000	250000
T2	One informative and one uninformative variable	1	2	1000	25000
T3	g-and-k distribution order statistics, small	4	7	1000	250000
T4	Lotka-Volterra Markov jump process model, small	3	32	500	125000
T5	g-and-k distribution order statistics, large	4	100	1000	250000
T6	Lotka-Volterra Markov jump process model, large	3	200	500	125000

improved by not employing parameter-wise normalization in (4), which however makes the analysis less robust to regression model misspecification, or by an alternative normalization.

An analysis such as performed here may generally allow evaluating regression model plausibility, and to obtain insights into parameter-data relationships, e.g. eliciting uninformative data.

3.3 Performance on general test problems

To evaluate robustness and general performance of the proposed methods, we next considered six test problems T1-6, not tailored to the challenges discussed in Section 3.2. Core model properties as well as employed ABC-SMC population sizes N and total budgets of numbers of simulations are given in Table 1.

T1, T3, and T4 are problems M3, M4, and M5 from Schälte et al. [2021], respectively an ODE model of a conversion reaction, and, based on application examples in Prangle [2017], g-and-k distribution samples, and a Markov jump process model of a Lotka-Volterra predator-prey process. T2 consists of two observables, thereof $y_1 \sim \mathcal{N}(\theta, 0.1^2)$ informative and $y_2 \sim \mathcal{N}(0, 1^2)$ uninformative, with wide prior $\theta \sim \mathcal{N}(0, 100^2)$, also from Prangle [2017]. T5 and T6 are variations of T3 and T4 with higher-dimensional data, based on application examples in Fearnhead and Prangle [2012]. T5 employs 100 order statistics out of 10,000 samples from a g-and-k distribution, with $U[0, 10]$ priors on the four parameters A, B, g, k, considering ground truth values $(A, B, g, k) = (3, 1, 2, 0.5)$. T6 employs noise-free observations of predators and prey at 200 evenly-spaced time-points over the interval $[0, 20]$, estimating the three reaction rate coefficients on linear scale, considering tight independent priors $\theta_1 \sim U[0, 2]$, $\theta_2 \sim U[0, 0.1]$, $\theta_3 \sim U[0, 1]$, and ground truth values $(\theta_1, \theta_2, \theta_3) = (0.5, 0.0025, 0.3)$. We ran 10 repetitions of different inference scenarios on problems T1-6 on different data sets. To measure fit quality, we reported root mean square errors (RMSE) of the weighted posterior samples from the last ABC-SMC generation, with respect to ground truth parameters (note all problem considered here are uni-modal). The results are visualized in Figure 4.

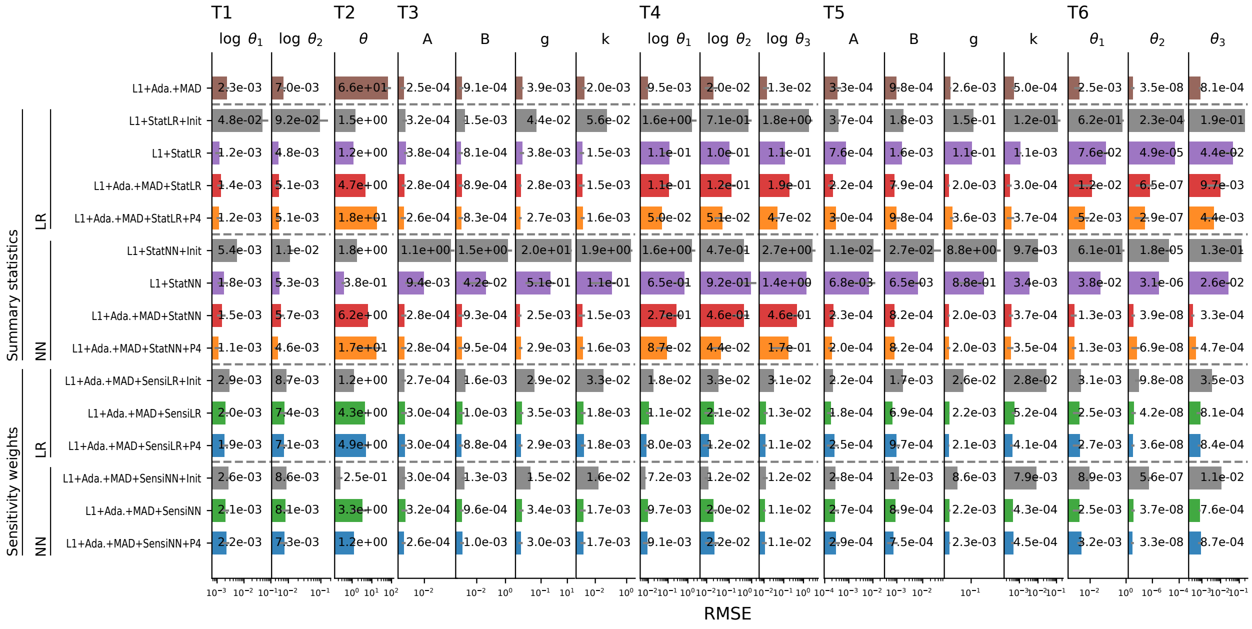

Figure 4: Median RMSE (smaller is better) for the parameters of models T1-6 (columns) obtained for 15 inference methods (rows), using an L1 distance, either uniformly weighted if unspecified, or with adaptive MAD scale normalization ("Ada.+MAD"). As regression models we considered linear regression ("LR") and neural networks ("NN"), both to define summary statistics ("Stat") and sensitivity weights ("Sensi"). Some inference settings further used parameter transformations $\lambda(\theta) = (\theta^1, \ldots, \theta^4)$ as regression targets ("P4"). In some settings, the regression model was trained before the initial generation ("Init"), or after 40% of the simulation budget if unspecified. The first row contains solely scale-normalized L1+Ada.+MAD as a reference, followed by two blocks of four rows using summary statistics, using firstly LR and secondly NN, and then by two blocks of three using sensitivity weights, using firstly LR and secondly NN. Reported values are medians over 10 replicates, with horizontal grey error lines indicating MAD.

Delay of regression model training advantageous on complex models

For the considered LR and NN models, regression model training on prior samples ("Init") gave for most problems substantially worse results than when trained after 40% of the simulation budget. One reason may be that only N prior samples were used for training, compared to potentially more samples, including rejected ones, in later generations. However, also when using only N training samples in the later-trained approach (not shown here), results were better than based on the prior. Thus, an explanation is that after multiple generations the bulk of samples is restricted to a high-density region, in which a simpler model is sufficiently accurate. This justifies empirically the approach by Fearnhead and Prangle [2012] of using a pilot run to constrain parameters. Jiang et al. [2017], who base their regression model on the prior, use firstly more complex NN models, and secondly up to 1e6 training samples, far more than entire analyses here. An exception was T2, where sometimes initial training improved performance. This can be explained by the global linear parameters-data mapping, such that accurate regression models can be easily learned and thereafter be beneficial.

Scale normalization improves performance for regression-based summary statistics

As the comparison of L1+Stat{LR/NN} and L1+Ada.+MAD+Stat{LR/NN} shows, the use of scale normalization improved performance for many problems, particularly for T5+6, while it was roughly similar for T1. An exception was again T2, where in fact a uniformly weighted L1 distance would be preferable over L1+Ada.+MAD at least in the first generations, as the uninformative observable happens to vary less there.

Sensitivity-weighted distances perform highly robustly

The approaches L1+Ada.+MAD+Sensi{LR/NN}(+P4) using regression models to define sensitivity weights performed reliably, with RMSE values generally not far higher, but in some cases consistently lower, than those obtained by L1+Ada.+MAD. This indicates that, while the sensitivity weighting could in those cases not improve performance, as sole scale normalization was efficient already, the approach is highly robust. In some cases, specifically T2, which had one clearly uninformative statistic, and arguably T5, which is a high-dimensional collection of order statistics, did the sensitivity weighting improve performance. In other cases, specifically T1, T3, T4, and T6, RMSE values for some parameters decreased, but slightly increased for others, indicating that the weighting scheme re-prioritized data points, while no overall uninformative ones could be disregarded.

Regression-based summary statistics can be superior but also less robust

In various cases, e.g. when trained in the initial generation, and consistently for T4, as well as using LR on T6, summary statistics were inferior to both L1+Ada.+MAD and sensitivity weights. Arguably, in those cases the regression model was not accurate enough to allow using its outputs as low-dimensional summaries. However, in some cases, specifically for T1, and two parameters of T6 using NN, RMSE values obtained using summary statistics were smaller than with both L1+Ada.+-MAD and sensitivity weights. This again indicates that if the lower-dimensional summary statistics representation is accurate and informative of the parameters, then its use can be beneficial and superior to mere re-weighting.

No clear preference for regression model or target augmentation

For both regression-based summary statistics and sensitivity weights, we found overall no clear preference for LR or NN, with LR more robust in many cases, but NN clearly preferable in some. Further, the use of augmented parameters as regression targets did not substantially worsen, but also not notably improve performance for any test problem, however performed inferior e.g. on T2, which has a clear linear mapping, such that the consideration of higher-order moments may have complicated the inference. This indicates that using augmented parameters as regression targets is robust, but if further information is available, a restriction to e.g. first or second order may be beneficial.

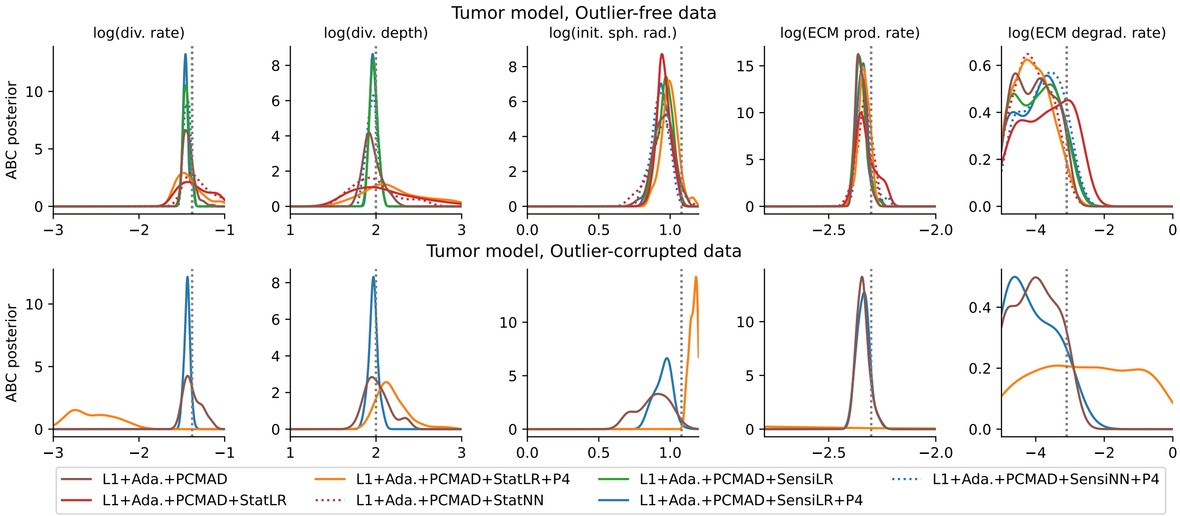

Figure 5: Posterior marginals for 5 out of the 7 model parameters of the tumor problem with interesting behavior. Without (top) and with (bottom) outliers. Ground truth parameters are indicated by vertical grey dotted lines. Plot boundaries are the employed uniform prior ranges, except the ECM production rate is zoomed in for visibility.

3.4 Performance on application example

Next, we considered an agent-based model of tumor spheroid growth (model M6 from Schälte et al. [2021]), considering both outlier-free and outlier-corrupted data. We employed the same simulated data as in Schälte et al. [2021], a population size of $N = 500$, and a computational budget of 150,000 simulations per analysis. Given its computational complexity, we considered on this problem only selected approaches: Besides the reference L1+Ada.+PCMAD, we employed, given the robust performance of LR before, L1+Ada.+PCMAD+SensiLR(+P4) using sensitivity weights, L1+Ada.+PCMAD+StatLR(+P4) using summary statistics, both with and without augmented regression targets $\lambda(\theta) = (\theta^1, \dots, \theta^4)$, further L1+Ada.+PCMAD+StatNN and L1+Ada.+PCMAD+SensiNN+P4 using NN. Here, to facilitate outlier detection, we used PCMAD instead of MAD.

Sensitivity weights identify uninformative model outputs

Using regression models to define sensitivity weights improved performance on the tumor model with outlier-free data over L1+Ada.+PCMAD, giving lower variances for the division rate and depth parameters, with otherwise similar results (Figure 5 top), and accepted simulations closely matching the observed data (Figure 6 top, simulations). No differences could be observed between using only the parameters, or also higher-order moments, as regression targets.

On this problem, regression-based summary statistics performed substantially worse, which may indicate that the employed regression models did not provide a sufficiently informative low-dimensional representation (L1+Ada.+PCMAD+StatLR(+P4), Figure 5 top), simulations did visibly not match

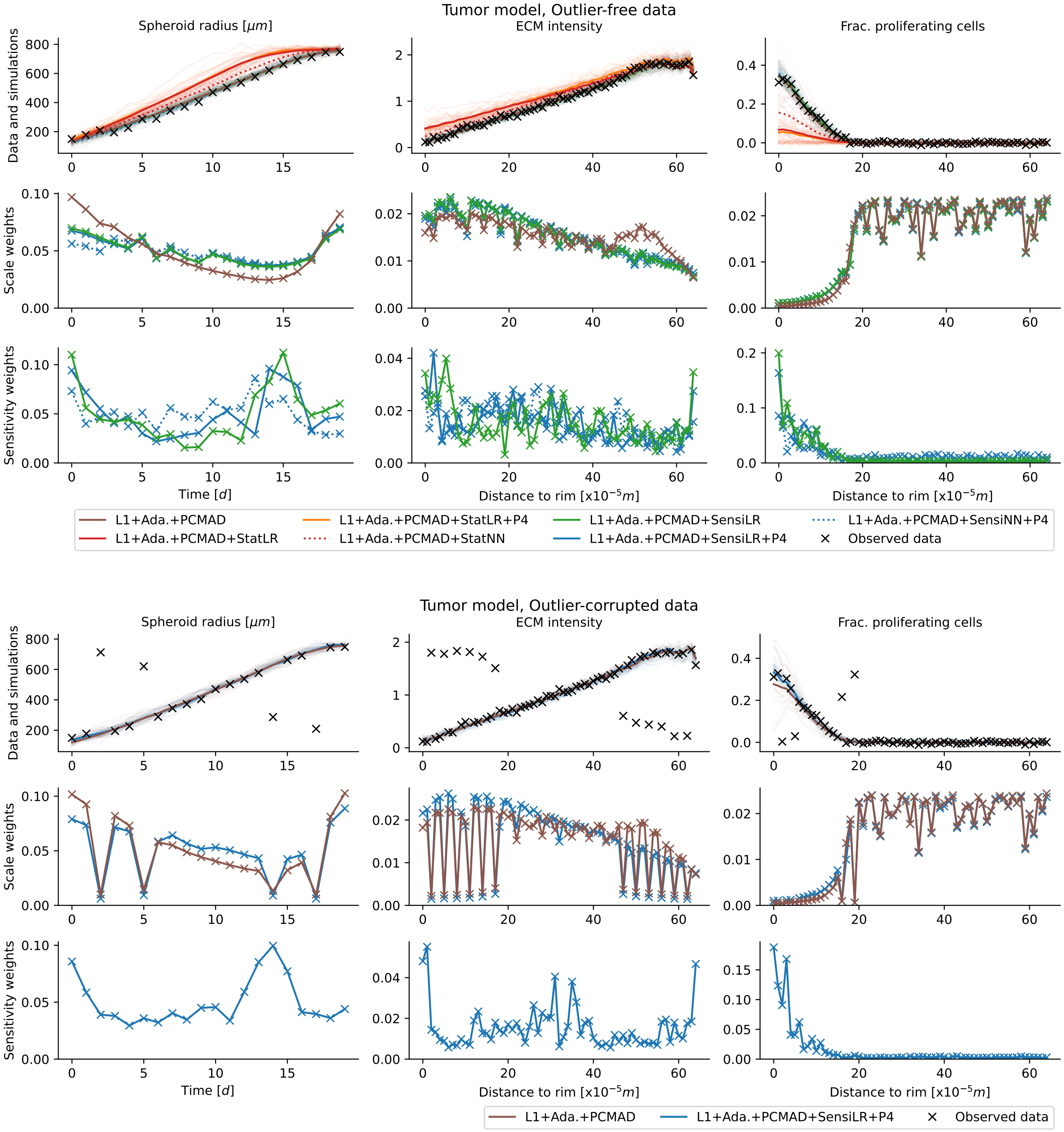

Figure 6: Fits, scale and sensitivity weights for the tumor problem on outlier-free (top) and outlier-corrupted (bottom) data. The respective upper rows show the observed data (black), and, for each approach, 20 accepted simulated data sets (light lines) as well as the sample means (darker lines) from the last ABC-SMC generation. The respective middle rows show the scale weights assigned to each data point in the last generation, normalized to unit sum, and the bottom rows the sensitivity weights, respectively only for distances employing such weights, and operating on the full data.

the observed data (Figure 6 top, 1st row).

The overall structure of sensitivity weights assigned via LR with and without parameter augmentation, as well as NN, was roughly consistent across multiple runs (Figure 6 top, 3rd row). Low weights were assigned to the fraction of proliferating cells at large distances to the rim, indicating these to be uninformative, and counteracting the large weights resulting from scale normalization (Figure 6 top, 2nd row). While the sensitivity weights exhibit some variability between adjacent points and across runs, consistent and reasonable overall patterns can be observed.

Robust on outlier-corrupted data

Using sensitivity weights improved performance also on outlier-corrupted data (Figure 5 bottom). Given its previously good performance, here we only considered L1+Ada.+PCMAD+SensiLR+P4. Accepted simulations in the final generation matched the observed data more closely than for L1+Ada.+PCMAD (Figure 6 bottom, 1st row). The PCMAD scheme assigned low weights to outliers, independent of the regression-based sensitivity weights (Figure 6 bottom, 2nd and 3rd row). Thus, the combination of both methods allowed to simultaneously account for outliers and informativeness.

4 Discussion

In this work, we discussed problems arising in ABC (1) from partly uninformative data for scale-normalized distances, (2) from heterogeneous parameter scales for regression-based summary statistics, and (3) from parameter non-identifiability for regression model adequacy. To tackle these problems, we presented multiple solutions: First, we suggested employing adaptive scale-normalizing distances on top of regression-based summary statistics, to homogenize the impact of parameters. Second, as an alternative to the first solution, we introduced novel sensitivity weights derived from regression models, measuring the informativeness of data on parameters. Third, we introduced augmented regression targets to overcome parameter non-identifiability.

We showed substantial improvements of the novel methods over established approaches on a simple demonstration problem. For the sensitivity-weighted distances, we showed robust performance on various further test problems, in particular on a complex systems biological application problem. Yet, there are numerous ways in which the presented methods can be improved:

While simple linear models often sufficed, especially when trained on a high-density region, in some cases more complex models were superior. A systematic investigation of alternative and more complex model types, e.g. neural networks tailored to the data types, as well as model selection, would be useful.

Larger training sample sets may be beneficial especially for complex models, and lead to more robust estimators. While increasing the training set is straightforward, as it only requires continued sampling from the forward model, there is a cost trade-off of the actual ABC inference and regression model training.

While in many cases delaying regression model training to later generations and a high-density region was advantageous, for simple models we observed benefits of early regression. Criteria on if and when to train or update regression models, also repeatedly, would be of interest.

This work may be regarded as an extension of the approaches of Prangle [2017], Schälte et al. [2021] as well as Fearnhead and Prangle [2012]. An alternative weighting scheme is presented by Harrison and Baker [2020], who maximize a distance between samples from the prior and the posterior approximation. While using a different notion of informativeness and a specific underlying sampler, a comparison in terms of efficiency, robustness to outliers, and information gain would be of interest.

All methods presented in this work have been implemented in the Python package pyABC, facilitating their straightforward application. We anticipate that such approaches, which automatically normalize and extract or weight features of interest without extensive manual tuning, will substantially improve performance of ABC methods on a wide range of applications problems.

Acknowledgments

We thank Dennis Prangle and Marc Vaisband for fruitful discussions, Emad Alamoudi for help with the HPC setup, and acknowledge the Gauss Centre for Supercomputing for providing computing time on the GCS Supercomputer JUWELS [Jülich Supercomputing Centre, 2019] at Jülich Supercomputing Centre.

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) (FitMultiCell/031L0159 and EMUNE/031L0293) and the German Research Foundation (DFG) under Germany's Excellence Strategy (EXC 2047 390873048 and EXC 2151 390685813)).

Author contributions

Y.S. devised the methods, wrote the implementation, and performed the case study. J.H. conceived the research question and provided supervision. Both authors discussed the results and jointly wrote the manuscript.

References

Barber, S. et al. The rate of convergence for approximate Bayesian computation. *Electronic Journal of Statistics*, 9(1):80–105, 2015.

Beaumont, M.A. et al. Approximate Bayesian Computation in Population Genetics. *Genetics*, 162 (4):2025–2035, 12 2002.

Blum, M.G. et al. A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28(2):189–208, 2013.

Borowska, A. et al. Gaussian process enhanced semi-automatic approximate Bayesian computation: parameter inference in a stochastic differential equation system for chemotaxis. *Journal of Computational Physics*, 429:109999, 2021.

Csilléry, K. et al. abc: an R package for approximate Bayesian computation (ABC). *Methods in ecology and evolution*, 3(3):475–479, 2012.

Del Moral, P. et al. Sequential Monte Carlo samplers. *J. R. Stat. Soc. B*, 68(3):411–436, 2006.

Drovandi, C.C. and Pettitt, A.N. Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, 67(1):225–233, 2011.

Drovandi, C.C. et al. Approximate bayesian computation using indirect inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(3):317–337, 2011.

Fearnhead, P. and Prangle, D. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. B*, 74(3):419–474, 2012.

Filippi, S. et al. On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Stat. Appl. Genet. Mol.*, 12(1):87–107, 2013.

Gershenfeld, N.A. and Gershenfeld, N. *The nature of mathematical modeling*. Cambridge university press, 1999.

Harrison, J.U. and Baker, R.E. An automatic adaptive method to combine summary statistics in approximate bayesian computation. *PloS one*, 15(8):e0236954, 2020.

Hasenauer, J. et al. Data-driven modelling of biological multi-scale processes. *Journal of Coupled Systems and Multiscale Dynamics*, 3(2):101–121, Sept. 2015. doi: 10.1166/jcsmd.2015.1069.

Jiang, B. et al. Learning summary statistic for approximate bayesian computation via deep neural network. *Statistica Sinica*, pages 1595–1618, 2017.

Jülich Supercomputing Centre. JUWELS: Modular Tier-0/1 Supercomputer at the Jülich Supercomputing Centre. *Journal of large-scale research facilities*, 5(A135), 2019. doi: 10.17815/jlsrf-5-171.

Kitano, H. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, Mar. 2002.

Klinger, E. and Hasenauer, J. A scheme for adaptive selection of population sizes in Approximate Bayesian Computation - Sequential Monte Carlo. In Feret, J. and Koeppl, H., editors, *Computational Methods in Systems Biology. CMSB 2017*, volume 10545 of *Lecture Notes in Computer Science*. Springer, Cham, 2017.

Klinger, E. et al. pyABC: distributed, likelihood-free inference. *Bioinformatics*, 34(20):3591–3593, 10 2018. doi: 10.1093/bioinformatics/bty361.

Nunes, M.A. and Balding, D.J. On optimal selection of summary statistics for approximate Bayesian computation. *Stat. Appl. Genet. Mol.*, 9(1), 2010.

Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Prangle, D. Adapting the ABC distance function. *Bayesian Analysis*, 12(1):289–309, 2017. doi: 10.1214/16-BA1002.

Pritchard, J.K. et al. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

Raue, A. et al. Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE*, 8(9):e74335, Sept. 2013. doi: 10.1371/journal.pone.0074335.

Schälte, Y. et al. Robust adaptive distance functions for approximate Bayesian inference on outlier-corrupted data. *bioRxiv*, 2021.

Silk, D. et al. Optimizing threshold-schedules for sequential approximate Bayesian computation: Applications to molecular systems. *Stat. Appl. Genet. Mol. Biol.*, 12(5):603–618, Oct. 2013.

Sisson, S.A. et al. Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci.*, 104(6): 1760–1765, Jan. 2007. doi: 10.1073/pnas.0607208104.

Sisson, S.A. et al. *Handbook of approximate Bayesian computation*. Chapman and Hall/CRC, 2018.

Tarantola, A. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, 2005.

Tavaré, S. et al. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

Toni, T. and Stumpf, M.P.H. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104–110, 10 2010.