Evaluation of Tolerance Selection Strategies and Multifidelity Techniques in ABC Methods

de Sousa. Graziele D. S. [1], Silva. Renato S. [1], de Almeida. Regina C. C. [1]

[1] *Laboratório Nacional de Computação Científica*
grazidss@posgrad.lncc.br, 25651-075, RJ/Petropólis, Brazil
rssr@lncc.br, rcca@lncc.br

Abstract. Approximate Bayesian Computation (ABC) methods provide a flexible and robust framework for solving model fitting problems, particularly for complex models with intractable likelihood functions. This methodology approximates simulated parameter values using auxiliary data and evaluates the distance between this data and the true dataset. Effective implementation of ABC methods requires careful selection of techniques and algorithmic approaches to ensure computational efficiency. This study investigates the impact of tolerance selection and the integration of multifidelity techniques on the convergence and computational cost of the method. Initially, tolerance selection methods are explored, including a predefined vector, a percentile-based approach, and a percentage calculation derived from the distance vector obtained from model simulations. Subsequently, the optimal tolerance selection approach is combined with multifidelity techniques to enhance accuracy and reduce computational cost. This methodology is demonstrated using the Susceptible-Infected-Recovered (SIR) model.

Keywords: ABC, tolerance selection methods, multifidelity.

1 Introduction

Mathematical models are essential tools for understanding complex systems and predicting outcomes, such as those used in infectious disease dynamics. However, many biological systems lack reliable parameter data, which is typically acquired through calibration [1]. Approximate Bayesian Computation (ABC) methods effectively approximate posterior distributions and are widely applied in model calibration due to their simplicity and ease of implementation [2]. The key advantage of ABC is its ability to handle models that are intractable using traditional statistical approaches [3], while also offering computational efficiency compared to methods like Markov Chain Monte Carlo (MCMC) [4].

This study is divided into two stages. The first stage tests three tolerance selection methods: a predefined vector, a percentile-based calculation, and a percentage derived from the distance vector obtained from model simulations. In the second stage, the most advantageous tolerance method is combined with a multifidelity strategy, integrating both the Euler method (low-fidelity model) and the Richardson extrapolation method (high-fidelity model).

The study aims to emphasize the importance of tolerance selection in ABC methods and explore the potential of combining multifidelity strategies. To demonstrate this, an epidemic model with known parameters is simulated, and Sequential Monte Carlo ABC (SMC-ABC) is applied to re-estimate these parameters.

2 Background

2.1 ABC Methods

Approximate Bayesian computation (ABC) comprehend a set of methods based in Bayesian statistics, which avoid the need for an explicit evaluation of the likelihood function to make inferences about model parameters. The fundamental idea of ABC methods involves replacing the likelihood calculation with a process that compares observed data with simulated data. According to Toni et al. [1], given a parameter vector θ to be estimated and a prior distribution $\pi(\theta)$, the goal of the method is to approximate the posterior distribution $\pi(\theta|y)$. Generally, the ABC algorithm proceeds as follows:

1. Sample a candidate parameter (particle) vector θ^* from the prior $\pi(\theta)$;
2. Simulate a dataset $\tilde{y}$ from the model using the sampled parameter θ^*;
3. Compare the simulated dataset $\tilde{y}$ with the experimental dataset y, using a distance function d and a tolerance ϵ: if $d(y, \tilde{y}) \leq \epsilon$, then accept θ^*;
4. Repeat steps 1, 2, and 3 until the desired number of accepted samples is obtained;
5. The output is a sample of parameters from the approximate posterior distribution $\pi(\theta|d(y, \tilde{y}) \leq \epsilon)$.

In the basic Approximate Bayesian Computation (ABC) algorithm, the selection of the tolerance ϵ is critical. A sufficiently small value of ϵ ensures that the posterior distribution of the simulated parameters closely approximates the true posterior distribution. Thus, tolerance selection affects both the convergence and efficiency of the algorithm, requiring careful consideration [4, 5]. In this study, three tolerance selection strategies based on [5] are explored, utilizing the Sequential Monte Carlo (SMC) ABC approach.

In ABC-SMC, the rejection mechanism from the ABC Rejection algorithm is employed, but successive populations are used to iteratively refine the parameter distribution. This refinement is based on ensuring that the distance between the simulated data $\tilde{y}$ and the observed data y, denoted as $d(y, \tilde{y})$, remains less than or equal to a threshold ϵ_{pop} for each population, where pop refers to the number of populations used. In this study, population sizes of pop $= 2$, 3, and 5 were tested.

2.2 Epidemiological Compartmental Model SIR

One of the most commonly used approaches in the mathematical modeling of infectious diseases is the compartmental model. These models simulate the collective behavior of population subgroups through labeled compartments. In this study, we employed a Susceptible-Infectious-Recovered (SIR) compartmental model based on [6], as a simple mathematical representation of disease transmission dynamics for model testing. In this model, individuals are classified into three compartments: Susceptible (S), Infected (I), and Recovered (R). The model represents infectious diseases that confer immunity upon recovery.

The SIR model describes population dynamics based on two parameters: β and γ, which represent the infection rate and the recovery rate, respectively. The infection rate β defines the rate at which susceptible individuals become infected through contact with infected individuals, while the recovery rate γ defines the rate at which infected individuals recover and acquire immunity. The model is governed by the following system of ordinary differential equations.

$$\frac{dS}{dt} = -\beta SI, \quad \frac{dI}{dt} = \beta SI - \gamma I, \quad \frac{dR}{dt} = \gamma I. \tag{1}$$

3 Methodology

In this section, we present the methodology adopted for implementing the SMC-ABC to calibrate an SIR model. The methodology is divided into two parts: the analysis of tolerance choice strategies and the use of a multifidelity technique. The code under development is being constructed and published on *GitHub*[1].

Normally, finding an initial tolerance is necessary. Specifying a reasonable value for ϵ_0 without prior knowledge can be challenging, especially for more complex problems. This often requires extensive testing using different values. In this work, we employed an approach based on the work of Simola et al. [5], where kN samples, with $k \in \mathbb{R}^+$, are obtained from the prior and used in the model. The distances obtained are then ordered in ascending order, and we choose the value corresponding to 20% of this sorted distance vector. The remaining input parameters required for calibration using the ABC-SMC method were determined through empirical testing.

3.1 Tolerances Strategies Implementation

The choice of tolerance values significantly impacts the computational efficiency of ABC methods. The ideal scenario is to obtain a sequence of tolerances that minimizes the total number of simulations, as this step constitutes the highest computational cost in the ABC algorithm [1]. A careful balance is required between the rapid decrease in tolerance values and the particle acceptance rate. In theory, a tolerance close to zero would reduce the number of accepted parameters, increasing the number of required simulations and, consequently, the

[1] `https://github.com/grazuzu/ABC-Tolerance-Selection-MultiFidelity`

computational cost. Conversely, a very large tolerance would result in the acceptance of many particles with a less accurate approximation. The challenge, as noted by Mertens et al. [7], is to find an appropriate balance between the choice of tolerance and the desired execution time.

Assuming ϵ_0 as mentioned above, we tested three strategies: (a) fixing values in advance, (b) adaptive percentile selection, and (c) adaptive percentage selection. For the fixed vector method, the selection of the remaining values in the tolerance vector $\epsilon_{1:T}$ was based on prior knowledge. For the adaptive methods, the approach was as follows: the distance values $\{d_{t-1}^I\}_{I=1}^N$ accepted in the previous iteration $t-1$ were used to select the next tolerance value. For each new population, the previous distance vector was ordered in ascending order, and the value was chosen based on the strategy.

For adaptive percentile selection, a predefined percentile was chosen and kept constant for all populations, the data was calculated using the `numpy.percentile` function from the *NumPy library*. This percentile was then computed from the ordered vector of distances. In the case of adaptive percentage selection, the procedure was the same; however, instead of using a percentile, we chose a percentage of the size of the ordered vector.

3.2 Multifidelity Techniques

Realistic models for predicting complex interactive systems often require calibration methods with respect to observed data, such as ABC methods. A critical consideration in applying these methods is the computational cost of generating simulations. According to Fernández-Godino [8], one approach to improving the efficiency of ABC is through the use of multifidelity models, which combine less expensive models with more costly ones to enhance computational efficiency while maintaining accuracy. In this work, we utilized a multifidelity strategy that combines the Euler method, a low-fidelity approach, with the Richardson extrapolation method, a higher-fidelity technique, to enhance the accuracy of numerical solutions while effectively managing computational resources.

Euler Method

The Euler method is a straightforward numerical technique for solving ordinary differential equations which approximates the solution as:

$$y_{n+1} = y_n + hf(t_n, y_n), \tag{2}$$

where h be the step size and y_{n+1} is the approximate solution at time t_{n+1}. The Euler method is first-order accurate, meaning the global error over N steps is proportional to h.

Richardson Extrapolation

Richardson extrapolation improves the accuracy of a numerical method by combining solutions computed with different step sizes. Normally, the solutions obtained with step sizes h and $h/2$, assuming the method is of order p, yield a higher-order accurate solution as:

$$y_{extrapolated} = \frac{2^p y(h/2) - y(h)}{2^p - 1}. \tag{3}$$

As we are using the Euler method ($p = 1$), the extrapolated solution becomes:

$$y_{extrapolated} = 2y(h/2) - y(h). \tag{4}$$

The multifidelity strategy used integrates the Euler method with Richardson extrapolation to balance computational efficiency and accuracy. The approach involves the following steps:

1. **Low-Fidelity Simulation**: Perform a coarse simulation using the Euler method with a larger step size h;
2. **High-Fidelity Simulation**: Perform a finer simulation using the Euler method, if the particle is accepted, with a smaller step size $h/2$ lead to a more accurate estimate,
3. Compute the enhanced solution using a Richardson extrapolation eq. (4).

4 Results

For this study, the number of particles that meet the tolerance criteria was set to $N = 300$. We used three different population numbers: $pop = 2$ 3, and 5. The maximum number of iterations T for each tested tolerance strategy was set to $10,000$ to achieve the desired particle sample size for each population. The Euclidean distance was used as the distance metric d, and the Euler method with $h = 0.25$ was employed to solve the differential equations presented in eq. (1). The data used was obtained from the SIR model with predefined parameters, acquired through simulation over 70 days and solved with a fourth-order Runge-Kutta method. The parameters used in the model are presented in Table 1.

Table 1. Definitions and prior distributions for SIR model parameters.

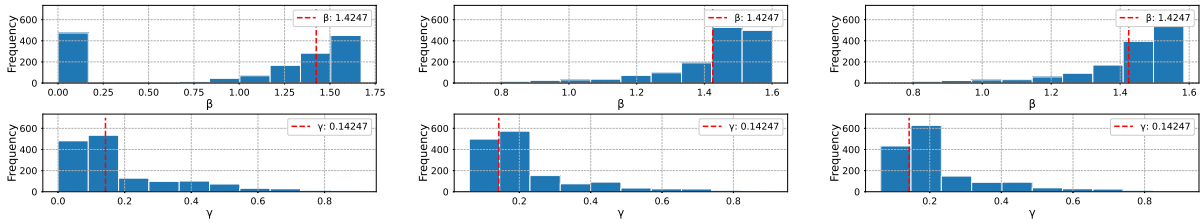
Parameter	Definition	Prior distribution	True value
β	Transmission rate	Uniform $(0.1, 1.5)$	1.4247
γ	Recovery rate	Uniform $(0.1, 1.5)$	0.14286
$S(0)$	Initial number of susceptible	$1.0 - I(0)$	
$I(0)$	Initial number of infected	10^{-06}	
$R(0)$	Initial number of recovered	0.0	

We compared tolerance choice strategies for the SMC-ABC method using the example model described in Section 2.2 to evaluate the efficiency of each strategy with the same parameters. Subsequently, the most effective tolerance strategy was applied in combination with the multifidelity technique to optimize computational cost without sacrificing accuracy. The obtained tolerance values are presented in Table 2.

Table 2. Tolerances values for the three strategies.

Fixed vector	Percentile vector	Percentage vector
10.51	10.51	10.51
5.0	4.6539	4.6492
4.0	2.4269	2.4217
2.0	1.3482	1.3439
1.0	0.8395	0.8539

Figure 1 presents the histograms for the parameters β and γ for $pop = 5$, obtained from the implementation of SMC-ABC with the three tolerance strategies. The histograms revel that the posterior parameters estimates varied in accuracy and precision across the different strategies.



(a) Fixed tolerance method. (b) Percentile tolerance method. (c) Percentage tolerance method.

Figure 1. Comparison of histograms for the posterior distribution of β and γ for each tolerance strategy with $pop = 5$ and $h = 0.25$. The red dashed lines indicate the true value of the parameters used to simulate the data.

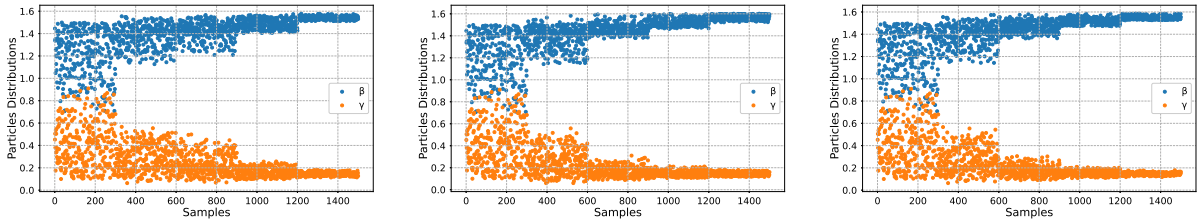
For the fixed tolerance values, the posterior distribution of β showed greater variation compared to the percentile and percentage-based methods, with the latter two yielding more consistent results for both parameters. However, the posterior distribution of γ using the fixed tolerance method was closer to the true values than those obtained with the other two strategies. The Table 3 presents the optimal parameter values for β and γ.

Table 3. A comparative analysis of the precise values of the parameters β and γ with the parameter exhibiting the shortest distance for each tolerance strategy across populations $2, 3$ and 5 e $h = 0.25$

	Pop2		Pop3		Pop5	
	β	γ	β	γ	β	γ
Exact	1.4247	0.14286				
Fixed	1.58182564	0.13764287	1.53726966	0.14464196	1.57235399	0.1429698
Percentile	1.53726966	0.14464196	1.55134924	0.15190873	1.61734834	0.14304775
Percentage	1.53726966	0.14464196	1.55134924	0.15190873	1.58335781	0.14355147

It can be observed that all three methods successfully achieved the desired number of particles for all populations. The tolerance values obtained using the percentile and percentage methods were very similar. Although the fixed tolerance method produced good results, the percentile and percentage methods demonstrated better adaptability in managing tolerance selection without requiring significant adjustments.

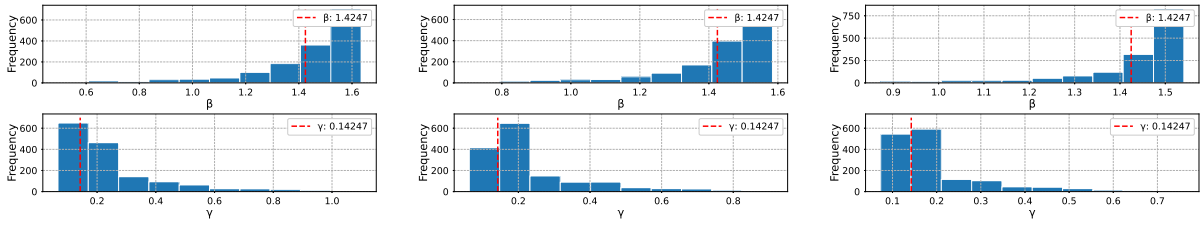
Figure 2 presents the distribution of particles for each tolerance selection strategy with $pop = 5$. The figure demonstrates that all three methods converge and were able to achieve the desired number of particles. Although the fixed tolerance method produces good results, the choice of tolerance values is made arbitrarily and may not always result in satisfactory values, leading the algorithm to converge to sub-optimal solutions rather than the true posterior distribution. Adaptive selection can help mitigate this problem by focusing on areas of the parameter space that are most likely to contain the true posterior distribution.



(a) Fixed tolerance method. (b) Percentile tolerance method. (c) Percentage tolerance method.

Figure 2. Comparison of particles distribution of β, blue dots, and γ, orange dots, for each tolerance strategy with $pop = 5$ and $h = 0.25$.

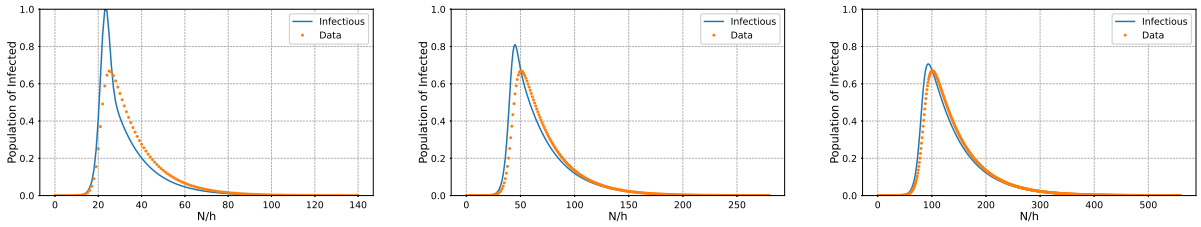
Given the similarity between the results of the percentile-based and percentage-based tolerance methods, the percentile method was selected for the continued implementation of the multifidelity strategy. Figure 3 illustrate the histograms of tests performed using the multifidelity strategy with tolerance based in percentile of 20%, and $h = 0.5, 0.25$ and 0.125. The figure shows that, although the highest frequency of values is close to the expected value in all cases, the values of γ are closer to the true value.



(a) $h = 0.5$. (b) $h = 0.25$. (c) $h = 0.125$.

Figure 3. Comparison of the histograms of β and γ for percentile tolerance strategies with multifidelity techniques for $pop = 5$ and $h = 0.5, 0.25, 0.125$. The red dashed lines indicate the true value of the parameters used to simulate the data.

Figure 4 presents the solution of the SIR model for the infected population and the data used in the calibration. The parameters were obtained with SMC-ABC, presented in Table 3, for $pop = 5$ and $h = 0.5, 0.25$ and 0.125. As can be observed, the solution improves as the value of h decreases.



(a) $h = 0.5$. (b) $h = 0.25$. (c) $h = 0.125$.

Figure 4. Comparison of the infects population curve of SIR model for percentile tolerance strategies with multifidelity techniques for $pop = 5$, $N = 70$, $h = 0.5, 0.25, 0.125$, line in blue, and the observed data, orange dots.

The values of the parameters that resulted in the smallest distance, as well as the value of this distance, in the application of the multifidelity technique for $h = 0.5, 0.25$ and 0.125 are summarized in the Table 4, being very close to the real values.

Table 4. Parameters with smallest distance with step size of $h = 0.5, 0.25$ and 0.125, and $pop = 5$.

	β	γ	dmin
$h = 0.5$	1.6333	0.1482	1.4247
$h = 0.25$	1.5832	0.1385	0.3901
$h = 0.125$	1.5176	0.1421	0.1610
$h = 0.167$	1.5516	0.1432	0.1868
$h = 0.083$	1.4855	0.1428	0.1310
$h = 0.042$	1.4538	0.1425	0.0972

Table 5 shows the number of samples necessary to reach $N = 300$ accepted values for SMC-ABC with multifidelity technique. The displayed results were obtained with $h = 0.5, 0.25$ and 0.125.

Table 5. Number of samples necessary for each population in $pop = 5$ round, with $h = 0.5, 0.25$ and 0.125.

	$pop1$	$pop2$	$pop3$	$pop4$	$pop5$
$h = 0.5$	1181	1121	892	923	915
$h = 0.25$	1770	925	848	1127	989
$h = 0.125$	2657	1117	926	1328	1157

Using our multifidelity strategy, the computational cost is proportional to the total number of Euler evaluations, which in this case will be the number of samples plus two twice the number of accepted particles. This solution has an error of $O(h^2)$. The total savings are evaluated by subtracting the number of samples (Table 5) used by Euler with step size $\frac{h}{2}$ from the number of Euler evaluations in the multifidelity strategy using h. These differences are presented in Table 6, where the minus sign represents how many Euler iterations are saved using the multifidelity strategy. Although the optimal values of β and γ, presented in Table 4 for $h = 0.5$, are further

from the expected values compared to those obtained with $h = 0.25$, this is a consequence of assuming the same ABC parameters for all values of h. In reality, all ABC parameters and h are correlated.

Table 6. Cost comparison between $h = 0.5$ and $h = 0.25$, and between $h = 0.25$ and $h = 0.125$, for Euler and multifidelity.

	$pop1$	$pop2$	$pop3$	$pop4$	$pop5$
$0.5/0.25$	-1759	-129	-204	-731	-463
$0.25/0.125$	-2944	-709	-404	-929	-725

5 Conclusions

The two adaptive strategies, based on percentile and percentage, demonstrated good results and are straightforward to implement. Initial analysis indicates that these strategies do not produce significant variations in the parameter estimates compared to the true values. For the multifidelity technique, the percentile strategy was adopted. The analysis of the multifidelity technique revealed its dependence on the step size h when solving the differential equations eq. (1). Future research could explore the parameters of ABC to enhance estimation accuracy while maintaining reductions in computational cost, as well as its application to larger models. In high-dimensional problems, the computational complexity of ABC increases significantly, introducing additional challenges. Future work aims to adapt the proposed method for more complex and high-dimensional models to improve model inference.

Acknowledgements. This work was funded in part by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – FINANCING CODE 001. Regina C. Almeida acknowledges the support provided by CNPq, Grant number 306588/2022-6.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, vol. 6, n. 31, pp. 187–202, 2009.
[2] K. Csilléry, M. G. Blum, O. E. Gaggiotti, and O. François. Approximate bayesian computation (abc) in practice. *Trends in ecology & evolution*, vol. 25, n. 7, pp. 410–418, 2010.
[3] J. Lopes and M. A. Beaumont. Abc: a useful bayesian tool for the analysis of population data. *Infection, Genetics and Evolution*, vol. 10, n. 6, pp. 825–832, 2010.
[4] S. A. Sisson, Y. Fan, and M. Beaumont. *Handbook of approximate Bayesian computation*. CRC Press, 2018.
[5] U. Simola, J. Cisewski-Kehe, M. U. Gutmann, and J. Corander. Adaptive approximate bayesian computation tolerance selection. *Bayesian analysis*, vol. 16, n. 2, pp. 397–423, 2021.
[6] E. Kuhl. *Computational Epidemiology, Data-Driven Modeling of COVID-19*. Springer, Switzerland, 2021.
[7] U. K. Mertens, A. Voss, and S. Radev. Abrox—a user-friendly python module for approximate bayesian computation with a focus on model comparison. *PloS one*, vol. 13, n. 3, pp. e0193981, 2018.
[8] M. G. Fernández-Godino. Review of multi-fidelity models. *arXiv preprint arXiv:1609.07196*, 2016.