Bayesian Analysis (0000) **00**, Number 0, pp. 1

Supplementary Material of "Adaptive Approximate Bayesian Computation Tolerance Selection"

Umberto Simola[*], Jessi Cisewski-Kehe[†], Michael U. Gutmann[‡], Jukka Corander[*,§]

Appendix A: Lotka–Volterra Model

The Lotka-Volterra model (Lotka, 1925; Volterra, 1927) describes two interacting populations and in their original ecological setting representing predators and prey. Since the true posterior distribution is not available, the ABC posteriors from Toni et al. (2009), are used as benchmarks. The interaction between the predators (y) and the prey (x) is defined by the following two differential equations:

$$\frac{dx}{dt} = ax - xy \tag{A.1}$$

$$\frac{dy}{dt} = bxy - y, \tag{A.2}$$

where the parameters of interests are a and b.

Inference on this model using ABC was considered originally in Toni et al. (2009), and we use their same model, dataset, summary statistic and distance function in order to test the performance of our proposed aABC-PMC algorithm. The same configuration was also used by Järvenpää et al. (2016) in order to test their ABC extensions. The dataset for the analysis $(x_{\text{obs}}, y_{\text{obs}})$ was obtained by using Eq. (A.1) and Eq. (A.2) with input values $a = 1$ and $b = 1$. The sample size is $n = 8$ for the two species, and the distance function for comparing real data, (x, y), with the simulated data, $(x_{\text{sim}}, y_{\text{sim}})$, is defined as:

$$d[(x_{\text{obs}}, y_{\text{obs}}), (x_{\text{sim}}, y_{\text{sim}})] = \sum_{i=1}^{n} \left[(x_{i,\text{obs}} - x_{i,\text{sim}})^2 + (y_{i,\text{obs}} - y_{i,\text{sim}})^2 \right]. \tag{A.3}$$

The forward model solves the deterministic set of differential equations defined above for x and y, then Gaussian noise is added from $N(0, 0.5^2)$ to obtain a simulated dataset, $(x_{\text{sim}}, y_{\text{sim}})$. The prior for both a and b is a Uniform distribution with the range $[-10, 10]$.

[*] Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland umberto.simola@helsinki.fi

[†]Department of Statistics and Data Science, Yale University, New Haven, CT, USA jessica.cisewski@yale.edu

[‡]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom michael.gutmann@ed.ac.uk

[§]Department of Biostatistics, University of Oslo, Oslo, Norway jukka.corander@medesin.uio.fi

 DOI: 0000

For $t > 1$ of the ABC-PMC algorithm, rather than using the Gaussian perturbation kernel as displayed in Algorithm 1, the selected particles are additively perturbed by draws from an $U(-0.1, 0.1)$, in order to avoid degeneracy of the samplers. The importance weights are calculated by taking into account of the fact that a Uniform kernel is used. In Toni et al. (2009), the series of tolerances is manually selected as listed in Table 1.

For the proposed aABC-PMC procedure, the initial number of draws sampled from the prior distributions is set at $N_{\text{init}} = 5 \times 1,000$ in order to sufficiently explore the parameter space. A comparison between the two procedures is done as before, in terms of the computational time and the total number of draws, with the results shown in Table 1. The results are based on 21 independent runs and the table includes the values for the run that produced the median number of total draws. Although aABC-PMC requires more iterations, the proposed procedure outperforms Toni et al. (2009)'s implementation of ABC-PMC in terms of total number of draws and computational time.

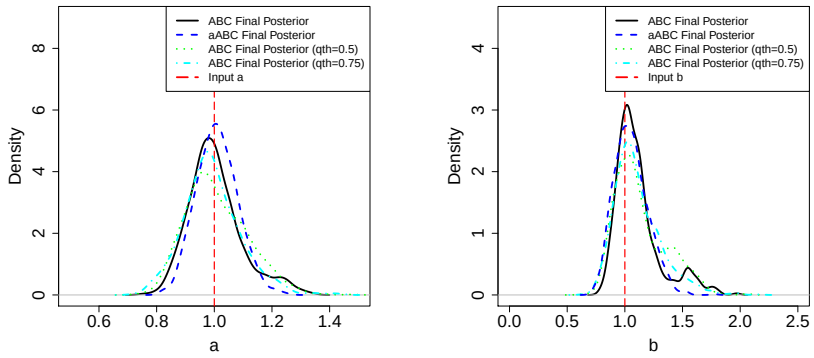
The ABC posteriors for parameters a and b for the manually-selected tolerances of Toni et al. (2009) and the proposed aABC-PMC approach are displayed in Figure 1. Additionally, ABC posteriors are displayed for two quantile-selected tolerances (0.5 and 0.75) for comparison. For each iteration $t > 1$ of the ABC–PMC procedure, in the first analysis the quantile of level 0.5 (i.e. the median) of the distances of the accepted particles from iteration $t - 1$ is used in order to provide the tolerance for the next iteration, while in the second analysis the quantile of level 0.75 of the distances of the accepted particles from iteration $t-1$ is used. The quantile-selected tolerance algorithms are stopped once the final number of draws needed by the aABC–PMC is reached. The ABC–PMC algorithm that uses a quantile-selected tolerance of 0.5 is stopped after $T = 11$ iterations, with a final tolerance equal to $\epsilon_{11} = 5.57$. When a quantile of 0.75 is used, $T = 21$ iterations are performed by the ABC–PMC algorithm and the final tolerance is $\epsilon_{21} = 5.43$. Both final tolerances are higher than the one adaptively selected by the proposed aABC-PMC approach, which is $\epsilon_9 = 4.76$. As shown in Fig. 1, when using the ABC–PMC algorithm with quantile-selected tolerances of 0.5 and 0.75, less informative posterior distributions are obtained than the posterior distributions obtained with the aABC–PMC approach.

The series of tolerances for the proposed aABC-PMC algorithm is adaptively selected in such a way that the forward model is drawn from fewer times than the manually-selected tolerances from Toni et al. (2009) and the common quantile-selected approaches. Though the final tolerance from Toni et al. (2009), $\epsilon_5 = 4.23$, is smaller than the final tolerance of aABC-PMC, $\epsilon_9 = 4.76$, the posteriors for a ($H_{\text{dist}} = 0.25$) and b ($H_{\text{dist}} = 0.34$) are comparable (Figure 1)[1].

[1]The Hellinger distances are calculated between the ABC posterior distributions found by Toni et al. (2009) and those retrieved with the proposed aABC-PMC approach.

	Toni et al. (2009)			aABC-PMC		
t	ϵ_t	D_t	t	ϵ_t	q_t	D_t
1	30	3,541	1	27.82		5,000
2	16	48,402	2	25.56	0.65	2,734
3	6	52,471	3	22.59	0.50	3,645
4	5	25,097	4	18.52	0.33	5,853
5	4.3	47,521	5	10.19	0.09	12,429
			6	6,70	0.17	12,146
			7	5,73	0,49	12,295
			8	4,99	0.51	20,890
			9	4.76	0.78	22,870
Total		177,032				97,862

Table 1: Lotka–Volterra model results. The number of draws needed in each iteration to reach $N = 1,000$ accepted values for the ABC-PMC algorithm of Toni et al. (2009) and the aABC-PMC algorithm. (The displayed results were obtained by running the procedure 21 times and using the run that produced the median number of total draws.) The aABC-PMC algorithm quantiles automatically selected through the iterations are listed under q_t. The procedure stopped once the quantile $q_{10} = 0.999$ was calculated. For the ABC–PMC algorithm a total of $177,032$ $(1,074$ sec.) draws were required, while the aABC-PMC took $97,862$ (548 sec.) draws overall.



(a) Final ABC posteriors for a.

(b) Final ABC posteriors for b.

Figure 1: Lotka–Volterra posterior distributions. Comparison between the final posterior distributions for (a) a and (b) b obtained using Toni et al. (2009)'s manually selecting the tolerances (black), by fixing the quantile equal to .50 (green) and 0.75 (cyan), and by using the aABC-PMC (blue).

Appendix B: Gaussian Mixture Model: Sedki et al. (2012)'s Results

Thanks to Sedki et al. (2012), a comparison for this example with the method proposed by Del Moral et al. (2012) was possible. (We also discuss the Del Moral et al. (2012) method in Sec. 3.3 of the main text.) The desired particle sample size was fixed at N=10,000 and their final threshold was $\epsilon_{final} = 0.09$. Using the same ϵ_{final}, the results of the analyses for the aABC-PMC approach are based on 21 independent runs with the same dataset $y_{\text{obs}} = 0$ and $k = 5$ (i.e. $N_{\text{init}} = 50000$). The run with the median number of total draws used the simulator $399,577$ times. For the ABC–SMC algorithm, we set the parameters required by the algorithm by using the specifications privided in Sedki et al. (2012): $N = 10000$, $\alpha = 0.95$, $M = 50$, $\text{nb}_{\text{threshold}} = N/2$ and $\epsilon_{\text{final}} = 0.09$ (these parameters are defined in Sec. 3.3 of the main text, while further details can be found in Del Moral et al. (2012)). We performed the analyses on 21 independent runs with the same dataset, $y_{\text{obs}} = 0$, and considered that run the produced the median number of total draws. For that run the total number of draws was equal to $3,088,550$. We note that the number of draws reported by Sedki et al. (2012) in their paper is $4.6 \cdot 10^6$ (Table 1, Sedki et al. 2012).

References

Del Moral, P., Doucet, A., and Jasra, A. (2012). "An adaptive sequential Monte Carlo method for approximate Bayesian computation." *Statistics and Computing*, 22(5): 1009–1020. 4

Järvenpää, M., Gutmann, M., Vehtari, A., and Marttinen, P. (2016). "Gaussian process modeling in approximate Bayesian computation to estimate horizontal gene transfer in bacteria." *arXiv preprint arXiv:1610.06462*. 1

Lotka, A. (1925). *Elements of physical biology*, volume 1. Baltimore, MD: Williams & Wilkins Co. 1

Sedki, M., Pudlo, P., Marin, J.-M., Robert, C. P., and Cornuet, J.-M. (2012). "Efficient learning in ABC algorithms." *arXiv preprint arXiv:1210.1388*. 4

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H. (2009). "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems." *Journal of the Royal Society, Interface / the Royal Society*, 6(31): 187–202. 1, 2, 3

Volterra, V. (1927). *Variazioni e fluttuazioni del numero d'individui in specie animali conviventi*. C. Ferrari. 1