Biometrika (2020), **xx**, x, *pp.* 1–15
Printed in Great Britain

On the use of approximate Bayesian computation Markov chain Monte Carlo with inflated tolerance and post-correction

By Matti Vihola and Jordan Franks
Department of Mathematics and Statistics, University of Jyväskylä, P.O.Box 35, FI-40014 University of Jyväskylä, Finland
matti.s.vihola@jyu.fi franks@iki.fi

Summary

Approximate Bayesian computation allows for inference of complicated probabilistic models with intractable likelihoods using model simulations. The Markov chain Monte Carlo implementation of approximate Bayesian computation is often sensitive to the tolerance parameter: low tolerance leads to poor mixing and large tolerance entails excess bias. We consider an approach using a relatively large tolerance for the Markov chain Monte Carlo sampler to ensure its sufficient mixing, and post-processing the output leading to estimators for a range of finer tolerances. We introduce an approximate confidence interval for the related post-corrected estimators, and propose an adaptive approximate Bayesian computation Markov chain Monte Carlo, which finds a 'balanced' tolerance level automatically, based on acceptance rate optimisation. Our experiments show that post-processing based estimators can perform better than direct Markov chain targeting a fine tolerance, that our confidence intervals are reliable, and that our adaptive algorithm leads to reliable inference with little user specification.

Some key words: Adaptive, approximate Bayesian computation, confidence interval, importance sampling, Markov chain Monte Carlo, tolerance choice

1. Introduction

Approximate Bayesian computation is a form of likelihood-free inference (see, e.g., the reviews Marin et al., 2012; Sunnåker et al., 2013) which is used when exact Bayesian inference of a parameter $\theta \in \mathsf{T}$ with posterior density $\pi(\theta) \propto \mathrm{pr}(\theta)L(\theta)$ is impossible, where $\mathrm{pr}(\theta)$ is the prior density and $L(\theta) = g(y^* \mid \theta)$ is an intractable likelihood with data $y^* \in \mathsf{Y}$. More specifically, when the generative model of observations $g(\,\cdot \mid \theta)$ cannot be evaluated, but allows for simulations, we may perform relatively straightforward approximate inference based on the following pseudo-posterior:

$$\pi_\epsilon(\theta) \propto \mathrm{pr}(\theta)L_\epsilon(\theta), \qquad L_\epsilon(\theta) = \mathbb{E}\{K_\epsilon(Y_\theta, y^*)\}, \quad Y_\theta \sim g(\,\cdot \mid \theta), \tag{1}$$

where $\epsilon > 0$ is a 'tolerance' parameter, and $K_\epsilon : \mathsf{Y}^2 \to [0, \infty)$ is a 'kernel' function, which is often taken as a simple cut-off $K_\epsilon(y, y^*) = 1\left(\|s(y) - s(y^*)\| \le \epsilon\right)$, where $s : \mathsf{Y} \to \mathbb{R}^d$ extracts a vector of summary statistics from the observations.

The summary statistics are often chosen based on the application at hand, and reflect what is relevant for the inference task; see also (Fearnhead & Prangle, 2012; Raynal et al., to appear). Because $L_\epsilon(\theta)$ may be regarded as a smoothed version of the true likelihood

$g(y^* \mid \theta)$ using the kernel K_ϵ, it is intuitive that using a too large ϵ may blur the likelihood and bias the inference. Therefore, it is generally desirable to use as small a tolerance $\epsilon > 0$ as possible, but because the computational methods suffer from inefficiency with small ϵ, the choice of tolerance level is difficult (cf. Bortot et al., 2007; Sisson & Fan, 2018; Tanaka et al., 2006).

We discuss a simple post-processing procedure which allows for consideration of a range of values for the tolerance $\epsilon \leq \delta$, based on a single run of approximate Bayesian computation Markov chain Monte Carlo (Marjoram et al., 2003) with tolerance δ. Such post-processing was suggested in (Wegmann et al., 2009) in case of simple cut-off, and similar post-processing has been suggested also with regression adjustment (Beaumont et al., 2002) in a rejection sampling context. The method, discussed further in Section 2, can be useful for two reasons: A range of tolerances $\epsilon \leq \delta$ may be routinely inspected, which can reveal excess bias in the pseudo-posterior π_δ; and the Markov chain Monte Carlo inference may be implemented with sufficiently large δ to allow for good mixing.

Our contribution is two-fold. We suggest straightforward-to-calculate approximate confidence intervals for the posterior mean estimates calculated from the post-processing output, and discuss some theoretical properties related to it. We also introduce an adaptive approximate Bayesian computation Markov chain Monte Carlo which finds a balanced δ during burn-in, using the acceptance rate as a proxy, and detail a convergence result for it.

2. Post-processing over a range of tolerances

For the rest of the paper, we assume that the kernel function in (1) has the form

$$K_\epsilon(y, y^*) = \phi\big(d(y, y^*)/\epsilon\big),$$

where $d : \mathsf{Y}^2 \to [0, \infty)$ is any 'dissimilarity' function and $\phi : [0, \infty) \to [0, 1]$ is a non-increasing 'cut-off' function. Typically $d(y, y^*) = \|s(y) - s(y^*)\|$, where $s : \mathsf{Y}^2 \to \mathbb{R}^d$ are the chosen summaries, and in case of the simple cut-off discussed in Section 1, $\phi(t) = \phi_{\text{simple}}(t) = 1\,(t \leq 1)$. We will implicitly assume that the pseudo-posterior π_ϵ given in (1) is well-defined for all $\epsilon > 0$ of interest, that is, $c_\epsilon = \int \text{pr}(\theta) L_\epsilon(\theta) \mathrm{d}\theta > 0$.

The following summarises the approximate Bayesian computation Markov chain Monte Carlo algorithm of Marjoram et al. (2003), with proposal q and tolerance $\delta > 0$:

Algorithm 1 *(*ABC-MCMC*(δ)).* Suppose $\Theta_0 \in \mathsf{T}$ and $Y_0 \in \mathsf{Y}$ are any starting values, such that $\text{pr}(\Theta_0) > 0$ and $\phi(d(Y_0, y^*)/\delta) > 0$. For $k = 1, 2, \ldots$, iterate:

(i) Draw $\tilde{\Theta}_k \sim q(\Theta_{k-1}, \cdot\,)$ and $\tilde{Y}_k \sim g(\,\cdot \mid \tilde{\Theta}_k)$.
(ii) With probability $\alpha_\delta(\Theta_{k-1}, Y_{k-1}; \tilde{\Theta}_k, \tilde{Y}_k)$ accept and set $(\Theta_k, Y_k) \leftarrow (\tilde{\Theta}_k, \tilde{Y}_k)$; otherwise reject and set $(\Theta_k, Y_k) \leftarrow (\Theta_{k-1}, Y_{k-1})$, where

$$\alpha_\delta(\theta, y; \tilde{\theta}, \tilde{y}) = \min\left\{1, \frac{\text{pr}(\tilde{\theta}) q(\tilde{\theta}, \theta) \phi\big(d(\tilde{y}, y^*)/\delta\big)}{\text{pr}(\theta) q(\theta, \tilde{\theta}) \phi\big(d(y, y^*)/\delta\big)}\right\}.$$

Algorithm 1 may be implemented by storing only Θ_k and the related distances $T_k = d(Y_k, y^*)$, and in what follows, we regard either $(\Theta_k, Y_k)_{k\geq 1}$ or $(\Theta_k, T_k)_{k\geq 1}$ as the output of Algorithm 1. In practice, the initial values (Θ_0, Y_0) should be taken as the state of the Algorithm 1 run for a number of initial 'burn-in' iterations. We also introduce an adaptive algorithm for parameter tuning later in Section 4.

It is possible to consider a variant of Algorithm 1 where many (possibly dependent) observations $\tilde{Y}_k^{(1)}, \ldots, \tilde{Y}_k^{(m)} \sim g(\,\cdot \mid \tilde{\Theta}_k)$ are simulated in each iteration, and an average of their kernel values is used in the accept-reject step (cf. Andrieu et al., 2018). We focus here in the case of single pseudo-observation per iteration, following the asymptotic efficiency result of Bornn et al. (2017), but remark that our method may be applied in a straightforward manner also with multiple observations.

Definition 1. *Suppose* $(\Theta_k, T_k)_{k=1,\ldots,n}$ *is the output of* ABC-MCMC(δ) *for some* $\delta > 0$. *For any* $\epsilon \in (0, \delta]$ *such that* $\phi(T_k/\epsilon) > 0$ *for some* $k = 1, \ldots, n$, *and for any function* $f : \mathsf{T} \to \mathbb{R}$, *define*

$$U_k^{(\delta,\epsilon)} = \phi(T_k/\epsilon)/\phi(T_k/\delta), \qquad W_k^{(\delta,\epsilon)} = U_k^{(\delta,\epsilon)} \big/ \textstyle\sum_{j=1}^n U_j^{(\delta,\epsilon)},$$

$$E_{\delta,\epsilon}(f) = \textstyle\sum_{k=1}^n W_k^{(\delta,\epsilon)} f(\Theta_k), \qquad S_{\delta,\epsilon}(f) = \textstyle\sum_{k=1}^n (W_k^{(\delta,\epsilon)})^2 \big\{f(\Theta_k) - E_{\delta,\epsilon}(f)\big\}^2.$$

Algorithm 4 in Appendix details how $E_{\delta,\epsilon}(f)$ and $S_{\delta,\epsilon}(f)$ can be calculated in $O(n \log n)$ time simultaneously for all $\epsilon \le \delta$ in case of simple cut-off. The estimator $E_{\delta,\epsilon}(f)$ approximates $\mathbb{E}_{\pi_\epsilon}\{f(\Theta)\}$ and $S_{\delta,\epsilon}(f)$ may be used to construct a confidence interval; see Algorithm 2 below. Theorem 1 details consistency of $E_{\delta,\epsilon}(f)$, and relates $S_{\delta,\epsilon}(f)$ to the limiting variance, in case the following well-known condition ensuring a central limit theorem holds:

Assumption 1 *(Finite integrated autocorrelation).* Suppose that $\mathbb{E}_{\pi_\epsilon}\{f^2(\Theta)\} < \infty$ and $\sum_{k \ge 1} \rho_k^{(\delta,\epsilon)}$ is finite, with $\rho_k^{(\delta,\epsilon)} = \mathrm{Corr}\big\{h_{\delta,\epsilon}(\Theta_0^{(s)}, Y_0^{(s)}), h_{\delta,\epsilon}(\Theta_k^{(s)}, Y_k^{(s)})\big\}$, where $(\Theta_k^{(s)}, Y_k^{(s)})_{k \ge 1}$ is a stationary version of the ABC-MCMC(δ) chain, and

$$h_{\delta,\epsilon}(\theta, y) = w_{\delta,\epsilon}(y) f(\theta), \qquad w_{\delta,\epsilon}(y) = \phi\big(d(y, y^*)/\epsilon\big)/\phi\big(d(y, y^*)/\delta\big).$$

Theorem 1. *Suppose* $(\Theta_k, T_k)_{k \ge 1}$ *is the output of* ABC-MCMC(δ), *and denote by* $E_{\delta,\epsilon}^{(n)}(f)$ *and* $S_{\delta,\epsilon}^{(n)}(f)$ *the estimators in Definition* 1. *If* $(\Theta_k, T_k)_{k \ge 1}$ *is* φ*-irreducible (Meyn & Tweedie, 2009) then, for any* $\epsilon \in (0, \delta)$, *we have as* $n \to \infty$:

(i) $E_{\delta,\epsilon}^{(n)}(f) \to \mathbb{E}_{\pi_\epsilon}\{f(\Theta)\}$ *almost surely, whenever the expectation is finite.*
(ii) Under Assumption 1, $n^{1/2}\big[E_{\delta,\epsilon}^{(n)}(f) - \mathbb{E}_{\pi_\epsilon}\{f(\Theta)\}\big] \to N\big(0, v_{\delta,\epsilon}(f)\tau_{\delta,\epsilon}(f)\big)$ *in distribution, where* $\tau_{\delta,\epsilon}(f) = \big(1 + 2\sum_{k \ge 1} \rho_k^{(\delta,\epsilon)}\big) \in [0, \infty)$ *and* $nS_{\delta,\epsilon}^{(n)}(f) \to v_{\delta,\epsilon}(f) \in [0, \infty)$ *almost surely.*

Proof of Theorem 1 is given in Appendix. Inspired by Theorem 1, we suggest to report the following approximate confidence intervals for the suggested estimators:

Algorithm 2. Suppose $(\Theta_k, T_k)_{k=1,\ldots,n}$ is the output of ABC-MCMC(δ) and $f : \Theta \to \mathbb{R}$ is a function, then for any $\epsilon \le \delta$:

(i) Calculate $E_{\delta,\epsilon}(f)$ and $S_{\delta,\epsilon}(f)$ as in Definition 1 (or in Algorithm 4).
(ii) Calculate $\hat{\tau}_\delta(f)$, an estimate of the integrated autocorrelation of $\big(f(\Theta_k)\big)_{k=1,\ldots,n}$.
(iii) Report the confidence interval

$$\big[E_{\delta,\epsilon}(f) \pm z_q\big\{S_{\delta,\epsilon}(f)\hat{\tau}_\delta(f)\big\}^{1/2}\big],$$

where $z_q > 0$ corresponds to the desired normal quantile.

The confidence interval in Algorithm 2 is straightforward application of Theorem 1, except for using a common integrated autocorrelation estimate $\hat\tau_\delta(f)$ for all $\tau_{\delta,\epsilon}(f)$. This relies on the approximation $\tau_{\delta,\epsilon}(f) \lessapprox \tau_\delta(f)$, which may not always be entirely accurate, but likely to be reasonable, as illustrated by Theorem 2 in Section 3 below. We suggest using a common $\hat\tau_\delta(f)$ for all tolerances because direct estimation of integrated autocorrelation is computationally demanding, and likely to be unstable for small ϵ.

The classical choice for $\hat\tau_\delta(f)$ in Algorithm 2(ii) is windowed autocorrelation, $\hat\tau_\delta(f) = \sum_{k=-\infty}^{\infty} \omega(k)\hat\rho_k$, with some $0 \le \omega(k) \le 1$, where $\hat\rho_k$ is the sample autocorrelation of $\big(f(\Theta_k)\big)$ (cf. Geyer, 1992). We employ this approach in our experiments with $\omega(k) = 1\,(|k| \le M)$ where the cut-off lag M is chosen adaptively as the smallest integer such that $M \ge 5\big(1 + 2\sum_{i=1}^{M} \hat\rho_k\big)$ (Sokal, 1996). Also more sophisticated techniques for the calculation of the asymptotic variance have been suggested (e.g. Flegal & Jones, 2010).

We remark that, although we focus here on the case of using a common cut-off ϕ for both the ABC-MCMC(δ) and the post-correction, one could also use a different cut-off ϕ_s in the simulation phase, as considered by Beaumont et al. (2002) in the regression context. The extension to Definition 1 is straightforward, setting $U_k^{(\delta,\epsilon)} = \phi(T_k/\epsilon)/\phi_s(T_k/\delta)$, and Theorem 1 remains valid under a support condition.

3. Theoretical justification

The following result, whose proof is given in Appendix, gives an expression for the integrated autocorrelation in case of simple cut-off.

THEOREM 2. *Suppose Assumption* 1 *holds and* $\phi = \phi_{\mathrm{simple}}$, *then*

$$\tau_{\delta,\epsilon}(f) - 1 = \frac{\{\check\tau_{\delta,\epsilon}(f) - 1\}\mathrm{var}_{\pi_\delta}(f_{\delta,\epsilon}) + 2\int \pi_\delta(\theta)\bar w_{\delta,\epsilon}(\theta)\{1 - \bar w_{\delta,\epsilon}(\theta)\}\frac{r_\delta(\theta)}{1-r_\delta(\theta)} f^2(\theta)\mathrm{d}\theta}{\mathrm{var}_{\pi_\delta}(f_{\delta,\epsilon}) + \int \pi_\delta(\theta)\bar w_{\delta,\epsilon}(\theta)\{1 - \bar w_{\delta,\epsilon}(\theta)\} f^2(\theta)\mathrm{d}\theta},$$

where $\bar w_{\delta,\epsilon}(\theta) = L_\epsilon(\theta)/L_\delta(\theta)$, $f_{\delta,\epsilon}(\theta) = f(\theta)\bar w_{\delta,\epsilon}(\theta)$, $\check\tau_{\delta,\epsilon}(f)$ *is the integrated autocorrelation of* $\{f_{\delta,\epsilon}(\Theta_k^{(s)})\}_{k\ge1}$ *and* $r_\delta(\theta)$ *is the rejection probability of the* ABC-MCMC*(δ) chain at* θ.

We next discuss how this loosely suggests that $\tau_{\delta,\epsilon}(f) \lessapprox \tau_{\delta,\delta}(f) = \tau_\delta(f)$. The weight $\bar w_{\delta,\delta} \equiv 1$, and under suitable regularity conditions both $\bar w_{\delta,\epsilon}(\theta)$ and $\check\tau_{\delta,\epsilon}(f)$ are continuous with respect to ϵ, and $\bar w_{\delta,\epsilon}(\theta) \to 0$ as $\epsilon \to 0$. Then, for $\epsilon \approx \delta$, we have $\bar w_{\delta,\epsilon} \approx 1$ and therefore $\tau_{\delta,\delta}(f) \approx \tau_{\delta,\epsilon}(f)$. For small ϵ, the terms with $\mathrm{var}_{\pi_\delta}(f_{\delta,\epsilon})$ are of order $O(\bar w_{\delta,\epsilon}^2)$, and are dominated by the other terms of order $O(\bar w_{\delta,\epsilon})$. The remaining ratio may be written as

$$\frac{2\int \pi_\delta(\theta)\bar w_{\delta,\epsilon}(\theta)\{1 - \bar w_{\delta,\epsilon}(\theta)\}\frac{r_\delta(\theta)}{1-r_\delta(\theta)} f^2(\theta)\mathrm{d}\theta}{\int \pi_\delta(\theta)\bar w_{\delta,\epsilon}(\theta)\{1 - \bar w_{\delta,\epsilon}(\theta)\} f^2(\theta)\mathrm{d}\theta} = 2\mathbb{E}_{\pi_\delta}\Big\{\bar g_{\delta,\epsilon}^2(\Theta)\frac{r_\delta(\Theta)}{1-r_\delta(\Theta)}\Big\},$$

where $\bar g_{\delta,\epsilon} \propto \{\bar w_{\delta,\epsilon}(1-\bar w_{\delta,\epsilon})\}^{1/2} f$ with $\pi_\delta(\bar g_{\delta,\epsilon}^2) = 1$. If $r_\delta(\theta) \le r_* < 1$, then the term is upper bounded by $2r_*(1-r_*)^{-1}$, and we believe it to be often less than $\tau_{\delta,\delta}(f)$, because the latter expression is similar to the contribution of rejections to the integrated autocorrelation; see the proof of Theorem 2.

For general ϕ, it appears to be hard to obtain similar theoretical result, but we expect the approximation to be still sensible. Theorem 2 relies on $Y_k^{(s)}$ being independent of $(\Theta_k^{(0)}, Y_k^{(0)})$ conditional on $\Theta_k^{(s)}$, assuming at least single acceptance. This is not true with other cut-offs, but we believe that the dependence of $Y_k^{(s)}$ from $(\Theta_0^{(s)}, Y_0^{(s)})$ given $\Theta_k^{(s)}$ is generally weaker than dependence of $\Theta_k^{(s)}$ and $\Theta_0^{(s)}$, suggesting similar behaviour.

We conclude the section with a general (albeit pessimistic) upper bound for the asymptotic variance of the post-corrected estimators.

THEOREM 3. *For any $\epsilon \leq \delta$, denote by $\sigma^2_{\delta,\epsilon}(f) = v_{\delta,\epsilon}(f)\tau_{\delta,\epsilon}(f)$ the asymptotic variance of the estimator of Definition* 1 *(see Theorem* 1(ii)*) and $\bar{f}(\theta) = f(\theta) - \mathbb{E}_{\pi_\epsilon}[f(\Theta)]$, then for any $\epsilon \leq \delta$,*

$$\sigma^2_{\delta,\epsilon}(f) \leq (c_\delta/c_\epsilon)\big\{\sigma^2_\epsilon(f) + \tilde{\pi}_\epsilon\big(\bar{f}^2(1 - w_{\delta,\epsilon})\big)\big\},$$

where $\tilde{\pi}_\epsilon$ is the stationary distribution of the direct ABC-MCMC*(ϵ) and $\sigma^2_\epsilon(f) = \sigma^2_{\epsilon,\epsilon}(f)$ its asymptotic variance.*

Theorem 3 follows directly from (Franks & Vihola, 2017, Corollary 4). The upper bound guarantees that a moderate correction, that is, ϵ close to δ and c_δ close to c_ϵ, is nearly as efficient as direct ABC-MCMC(δ). Indeed, typically $w_{\delta,\epsilon} \to 1$ and $c_\epsilon \to c_\delta$ as $\epsilon \to \delta$, in which case Theorem 3 implies $\limsup_{\epsilon\to\delta} \sigma^2_{\delta,\epsilon}(f) \leq \sigma^2_\epsilon(f)$. However, as $\epsilon \to 0$, the bound becomes less informative.

4. TOLERANCE ADAPTATION

We propose Algorithm 3 below to adapt the tolerance δ in ABC-MCMC(δ) during a burn-in of length n_b, in order to obtain a user-specified overall acceptance rate $\alpha^* \in (0, 1)$. Tolerance optimisation has been suggested earlier based on quantiles of distances, with parameters simulated from the prior (e.g. Beaumont et al., 2002; Wegmann et al., 2009). This heuristic might not be satisfactory in the Markov chain Monte Carlo context, if the prior is relatively uninformative. We believe that acceptance rate optimisation is a more natural alternative, and Sisson & Fan (2018) suggested this as well.

Our method requires also a sequence of decreasing positive step sizes $(\gamma_k)_{k\geq 1}$. We used $\alpha^* = 0.1$ and $\gamma_k = k^{-2/3}$ in our experiments, and discuss these choices later.

Algorithm 3. Suppose $\Theta_0 \in \mathsf{T}$ is a starting value with $\mathrm{pr}(\Theta_0) > 0$. Initialise $\delta = d(Y_0, y^*) > 0$ where $Y_0 \sim g(\,\cdot \mid \Theta_0)$. For $k = 1, \ldots, n_b$, iterate:

(i) Draw $\tilde{\Theta}_k \sim q(\Theta_{k-1}, \cdot\,)$ and $\tilde{Y}_k \sim g(\,\cdot \mid \tilde{\Theta}_k)$.
(ii) With probability $A_k = \alpha_{\delta_{k-1}}(\Theta_{k-1}, Y_{k-1}; \tilde{\Theta}_k, \tilde{Y}_k)$ accept and set $(\Theta_k, Y_k) \leftarrow (\tilde{\Theta}_k, \tilde{Y}_k)$; otherwise reject and set $(\Theta_k, Y_k) \leftarrow (\Theta_{k-1}, Y_{k-1})$.
(iii) $\log \delta_k \leftarrow \log \delta_{k-1} + \gamma_k(\alpha^* - A_k)$.

In practice, we use Algorithm 3 with a Gaussian symmetric random walk proposal q_{Σ_k}, where the covariance parameter Σ_k is adapted simultaneously (Haario et al., 2001; Andrieu & Moulines, 2006); see Algorithm 2 of Supplement C. We only detail theory for Algorithm 3, but note that similar simultaneous adaptation has been discussed earlier (cf. Andrieu & Thoms, 2008), and expect that our results could be elaborated accordingly.

The following conditions suffice for convergence of the adaptation:

Assumption 2. Suppose $\phi = \phi_{\mathrm{simple}}$ and the following hold:

(i) $\gamma_k = Ck^{-r}$ with $r \in (\frac{1}{2}, 1]$ and $C > 0$ a constant.
(ii) The domain $\mathsf{T} \subset \mathbb{R}^{n_\theta}$, $n_\theta \geq 1$, is a nonempty open set and $\mathrm{pr}(\theta)$ is bounded.

(iii) The proposal q is bounded and bounded away from zero.
(iv) The distances $D_\theta = d(Y_\theta, y^*)$ where $Y_\theta \sim g(\cdot \mid \theta)$ admit densities which are uniformly bounded in θ.
(v) $(\delta_k)_{k\geq 1}$ stays in a set $[a, b]$ almost surely, where $0 < a \leq b < +\infty$.
(vi) $c_\epsilon = \int \mathrm{pr}(\mathrm{d}\theta) L_\epsilon(\theta) > 0$ for all $\epsilon \in [a, b]$.

THEOREM 4. *Under Assumption 2, the expected value of the acceptance probability, with respect to the stationary distribution of the chain, converges to α^*.*

Proof of Theorem 4 will follow from the more general Theorem 1 of Supplement A.

Polynomially decaying step size sequences as in Assumption 2 (i) are common in adaptation which is of the stochastic approximation type as our approach (Andrieu & Thoms, 2008). Slower decaying step sizes such as $n^{-2/3}$ often behave better with acceptance rate adaptation (cf. Vihola, 2012, Remark 3).

Simple random walk Metropolis with covariance adaptation (Haario et al., 2001) typically leads to a limiting acceptance rate around 0.234 (Roberts et al., 1997). In case of a pseudo-marginal algorithm such as ABC-MCMC(δ), the acceptance rate is lower than this, and decreases when δ is decreased; see Lemma 2 of Supplement B. Markov chain Monte Carlo would typically be necessary when rejection sampling is not possible, that is, when the prior is far from the posterior. In such a case, the likelihood approximation must be accurate enough to provide reasonable approximation $\pi_\delta \approx \pi_\epsilon$. This suggests that the desired acceptance rate should be taken substantially lower than 0.234.

The choice of the desired acceptance rate α^* could also be motivated by theory developed for pseudo-marginal Markov chain Monte Carlo algorithms. Doucet et al. (2015) rely on log-normality of the likelihood estimators, which is problematic in our context, because the likelihood estimators take value zero. Sherlock et al. (2015) find the acceptance rate 0.07 to be optimal under certain conditions, but also in a quite dissimilar context. Indeed, in our context, the 0.07 guideline assumes a fixed tolerance, and informs about choosing the number of pseudo-data per iteration. As we stick with single pseudo-data per iteration following (Bornn et al., 2017), the 0.07 guideline cannot be taken too informative. We recommend slightly higher α^* such as 0.1 to ensure sufficient mixing.

5. POST-PROCESSING WITH REGRESSION CORRECTION

Beaumont et al. (2002) suggested similar post-processing as in Section 2, applying a further regression correction. Namely, in the context of Section 2, we may consider a function $\tilde{f}^{(\epsilon)}(\theta, y) = f(\theta) - \bar{s}(y)^{\mathrm{T}} b_\epsilon$ where $\bar{s}(y) = s(y) - s(y^*)$ and b_ϵ is a solution of

$$\min_{a_\epsilon, b_\epsilon} \mathbb{E}_{\tilde{\pi}_\epsilon}\big[\big\{f(\Theta) - a_\epsilon - \bar{s}(Y)^{\mathrm{T}} b_\epsilon\big\}^2\big] = \min_{a_\epsilon, b_\epsilon} \mathbb{E}_{\tilde{\pi}_\delta}\big[w_{\delta,\epsilon}(Y)\big\{f(\Theta) - a_\epsilon - \bar{s}(Y)^{\mathrm{T}} b_\epsilon\big\}^2\big],$$

where $\tilde{\pi}_\delta$ is the stationary distribution of ABC-MCMC(δ), with marginal π_δ, given in Appendix. When the latter expectation is replaced by its empirical version, the solution coincides with weighted least squares $(\hat{a}_\epsilon, \hat{b}_\epsilon^{\mathrm{T}})^{\mathrm{T}} = (\mathrm{M}^{\mathrm{T}} \mathrm{W}_\epsilon \mathrm{M})^{-1} \mathrm{M}^{\mathrm{T}} \mathrm{W}_\epsilon v$, with $v_k = f(\Theta_k)$, $\mathrm{W}_\epsilon = \mathrm{diag}(W_1^{(\delta,\epsilon)}, \ldots, W_n^{(\delta,\epsilon)})$ and with matrix M having rows $[M]_{k,\cdot} = (1, \bar{s}(Y_k)^{\mathrm{T}})$.

We suggest the following confidence interval for $a_\epsilon = \mathbb{E}_{\tilde{\pi}_\epsilon}\{\tilde{f}^{(\epsilon)}(\Theta, Y)\}$ in the spirit of Algorithm 2:

$$\big[\hat{a}_\epsilon \pm z_q \big(S_{\delta,\epsilon}^{\mathrm{reg}} \hat{\tau}_\delta^{\mathrm{reg}}\big)^{1/2}\big],$$

where $\hat{\tau}_{\delta}^{\mathrm{reg}}$ is the integrated autocorrelation estimate for $(\hat{F}_k^{(\delta)})$ where $\hat{F}_k^{(\delta)} = f(\Theta_k) - \bar{s}^T \hat{b}_\delta$ and $S_{\delta,\epsilon}^{\mathrm{reg}} = [(\mathrm{M}^{\mathrm{T}}\mathrm{W}_\epsilon M)^{-1}]_{1,1} \sum_{k=1}^{n} (W_k^{(\delta,\epsilon)})^2 (\hat{F}_k^{(\epsilon)} - \hat{a}_\epsilon)^2$, where the first term is included as an attempt to account for the increased uncertainty due to estimated $\hat{b}_\epsilon$, analogous to weighted least squares. Experimental results show some promise for this confidence interval, but we stress that we do not have better theoretical backing for it, and leave further elaboration of the confidence interval for future research.

6. Experiments

We experiment with our methods on two models, a lightweight Gaussian toy example, and a Lotka-Volterra model. Our experiments focus on three aspects: can ABC-MCMC(δ) with larger tolerance δ and post-correction to a desired tolerance $\epsilon < \delta$ deliver more accurate results than direct ABC-MCMC(ϵ); does the approximate confidence interval appear reliable; how well does the tolerance adaptation work in practice. All the experiments are implemented in Julia (Bezanson et al., 2017), and the codes are available in `https://bitbucket.org/mvihola/abc-mcmc`.

Because we believe that Markov chain Monte Carlo is most useful when little is known about the posterior, we apply covariance adaptation (Haario et al., 2001; Andrieu & Moulines, 2006) throughout the simulation in all our experiments, using an identity covariance initially. When running the covariance adaptation alone, we employ the step size n^{-1} as in the original method of Haario et al. (2001), and in case of tolerance adaptation, we use step size $n^{-2/3}$.

Regarding our first question, we investigate running ABC-MCMC(δ) starting near the posterior mode with different pre-selected tolerances δ. We first attempted to perform the experiments by initialising the chains from independent samples of the prior distribution, but in this case, most of the chains failed to accept a single move during the entire run. In contrast, our experiments with tolerance adaptation are initialised from the prior, and both the tolerances and the covariances are adjusted fully automatically by our algorithm.

6·1. *One-dimensional Gaussian model*

Our first model is a toy model with $\mathrm{pr}(\theta) = N(\theta; 0, 30^2)$, $g(y \mid \theta) = N(y; \theta, 1)$ and $d(y, y^*) = |y|$. The true posterior without approximation is Gaussian. While this scenario is clearly academic, the prior is far from the posterior, making rejection sampling approximate Bayesian computation inefficient. It is clear that π_ϵ has zero mean for all ϵ by symmetry, and that π_ϵ is more spread for bigger ϵ. We experiment with both simple cut-off ϕ_{simple} and Gaussian cut-off $\phi_{\mathrm{Gauss}}(t) = e^{-t^2/2}$.

We run the experiments with 10,000 independent chains, each for 11,000 iterations including 1,000 burn-in. The chains were always started from $\theta_0 = 0$. We inspect estimates for the posterior mean $\mathbb{E}_{\pi_\epsilon}\{f(\Theta)\}$ for $f(\theta) = \theta$ and $f(\theta) = |\theta|$. Figure 1 (left) shows the estimates with their confidence intervals based on a single realisation of ABC-MCMC(3). Figure 1 (right) shows box plots of the estimates calculated from each ABC-MCMC(δ), with δ indicated by colour; the rightmost box plot (blue) corresponds to ABC-MCMC(3), the second from the right (red) ABC-MCMC(2.275) etc. For $\epsilon = 0.1$, the post-corrected estimates from ABC-MCMC(0.825) and ABC-MCMC(1.55) appear slightly more accurate than direct ABC-MCMC(0.1). Similar figure for Gaussian cut-off, with similar findings, may be found in the Supplement Figure 1.

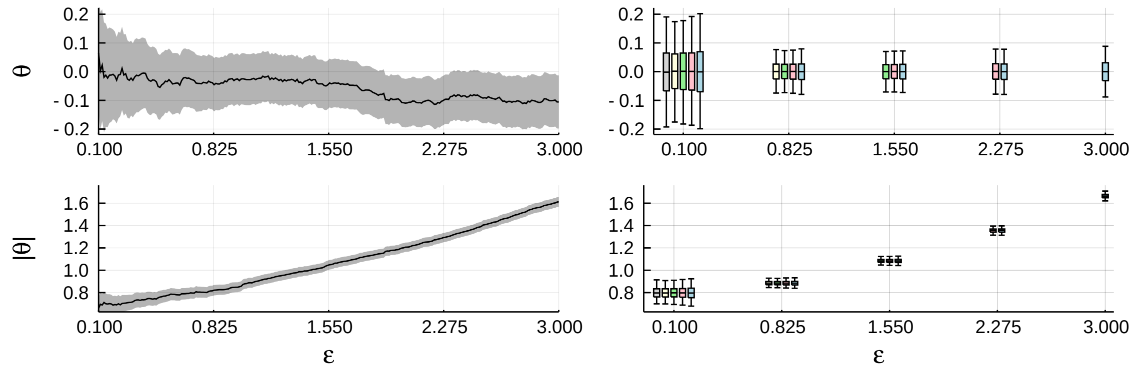

Fig. 1. Gaussian model with ϕ_{simple}. Estimates from single run of ABC-MCMC(3) (left) and estimates from 10,000 replications of ABC-MCMC(δ) for $\delta \in \{0.1, 0.825, 1.55, 2.275, 3\}$ indicated by colours.

Table 1. *Frequencies of the 95% confidence intervals, from* ABC-MCMC*(δ) to tolerances ϵ, containing the ground truth in the Gaussian model.*

		$f(x) = x$					$f(x) = \|x\|$					Acc.
Cut-off	$\delta \setminus \epsilon$	0.10	0.82	1.55	2.28	3.00	0.10	0.82	1.55	2.28	3.00	rate
	0.1	0.93					0.93					0.03
	0.82	0.97	0.95				0.95	0.94				0.22
ϕ_{simple}	1.55	0.97	0.97	0.95			0.96	0.95	0.95			0.33
	2.28	0.98	0.97	0.96	0.95		0.96	0.96	0.96	0.95		0.4
	3.0	0.98	0.98	0.97	0.97	0.95	0.96	0.96	0.96	0.95	0.95	0.43
	0.1	0.93					0.93					0.05
	0.82	0.94	0.95				0.92	0.95				0.29
ϕ_{Gauss}	1.55	0.94	0.94	0.95			0.94	0.94	0.95			0.38
	2.28	0.95	0.95	0.95	0.95		0.95	0.95	0.96	0.95		0.41
	3.0	0.95	0.95	0.95	0.95	0.95	0.95	0.96	0.95	0.95	0.95	0.42

Table 1 shows frequencies of the calculated 95% confidence intervals containing the 'ground truth', as well as mean acceptance rates. The ground truth for $\mathbb{E}_{\pi_\epsilon}\{f_1(\Theta)\}$ is known to be zero for all ϵ, and the overall mean of all the calculated estimates is used as the ground truth for $\mathbb{E}_{\pi_\epsilon}\{f_2(\Theta)\}$. The frequencies appear close to ideal with the post-correction approach, being slightly pessimistic in case of simple cut-off as anticipated by the theoretical considerations; see Theorem 2 and the related discussion.

Figure 2 shows progress of tolerance adaptations during the burn-in, and histogram of the mean acceptance rates of the chain after burn-in. The lines on the left show the median, and the shaded regions indicate the 50%, 75%, 95% and 99% quantiles. The figure suggests concentration, but reveals that the adaptation has not fully converged yet. This is also visible in the mean acceptance rate over all realisations, which is 0.17 for simple cut-off and 0.12 for Gaussian cut-off; see Figure 2 in the Supplement. Table 2 shows root mean square errors for target tolerance $\epsilon = 0.1$, with both ABC-MCMC(δ) with δ fixed as above, and for the tolerance adaptive algorithm. Here, only the adaptive chains with final tolerance ≥ 0.1 were included (9,998 and 9,993 out of 10,000 chains for

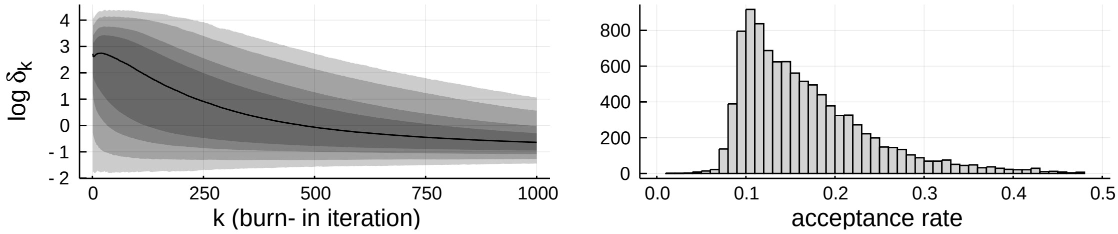

Fig. 2. Progress of tolerance adaptation (left) and histogram of acceptance rates (right) in the Gaussian model experiment with simple cut-off.

Table 2. *Root mean square errors* ($\times 10^{-2}$) *from* ABC-MCMC(δ) *for tolerance* $\epsilon = 0.1$ *with fixed tolerance and with the adaptive algorithms in the Gaussian model.*

	ϕ_{simple}						ϕ_{Gauss}					
	Fixed tolerance					Adapt	Fixed tolerance					Adapt
δ	0.1	0.82	1.55	2.28	3.0	0.64	0.1	0.82	1.55	2.28	3.0	0.28
x	9.75	8.95	9.29	9.65	10.3	9.15	7.97	7.12	7.82	8.94	9.93	7.08
$\lvert x\rvert$	5.49	5.35	5.51	5.81	6.24	5.38	4.47	4.22	4.68	5.26	5.95	4.15

ϕ_{simple} and ϕ_{Gauss}, respectively). Tolerance adaptation, started from prior distribution, appears to be competitive with 'optimally' tuned fixed tolerance ABC-MCMC(δ).

6·2. *Lotka-Volterra model*

Our second experiment is a Lotka-Volterra model suggested by Boys et al. (2008), which was considered in the approximate Bayesian computation context by Fearnhead & Prangle (2012). The model is a Markov process $(X_t, Y_t)_{t\geq 0}$ of counts, corresponding to a reaction network $X \to 2X$ with rate θ_1, $X + Y \to 2Y$ with rate θ_2 and $Y \to \emptyset$ with rate θ_3. The reaction log-rates $(\log\theta_1, \log\theta_2, \log\theta_3)^{\mathrm{T}}$ are parameters, which we equip with a uniform prior, $(\log\theta_1, \log\theta_2, \log\theta_3)^{\mathrm{T}} \sim U([-6, 0]^3)$. The data is a simulated trajectory from the model with $\theta = (0.5, 0.0025, 0.3)^{\mathrm{T}}$ until time 40. The inference is based on the Euclidean distances of five-dimensional summary statistics of the process observed every 5 time units ($\tilde{X}_k = X_{5k}$ and $\tilde{Y}_k = Y_{5k}$). The summary statistics are the sample autocorrelation of $(\tilde{X}_k)$ at lag 2 multiplied by 100, and the 10% and 90% quantiles of $(\tilde{X}_k)$ and $(\tilde{Y}_k)$. The observed summary statistics are $(-51.07, 29, 304, 65, 404)^{\mathrm{T}}$.

We first run comparisons similar to Section 6·1, but now with 1,000 independent ABC-MCMC(δ) chains with simple cut-off. We investigate the effect of post-correction, with 20,000 samples, including 10,000 burn-in, for each chain. All chains were started from near the posterior mode, from $(-0.55, -5.77, -1.09)^{\mathrm{T}}$. Figure 3 shows similar comparisons as in Section 6·1, and Figure 4 shows results for regression correction with Epanechnikov cut-off $\phi_{\text{Epa}}(t) = \max\{0, 1 - t^2\}$ (Beaumont et al., 2002). The results suggest that post-correction might provide slightly more accurate estimators, particularly with smaller tolerances. There is also some bias in ABC-MCMC(δ) with smaller δ, when compared to the ground truth calculated from ABC-MCMC(δ) chain of ten million iterations. Table 3 shows coverages of confidence intervals.

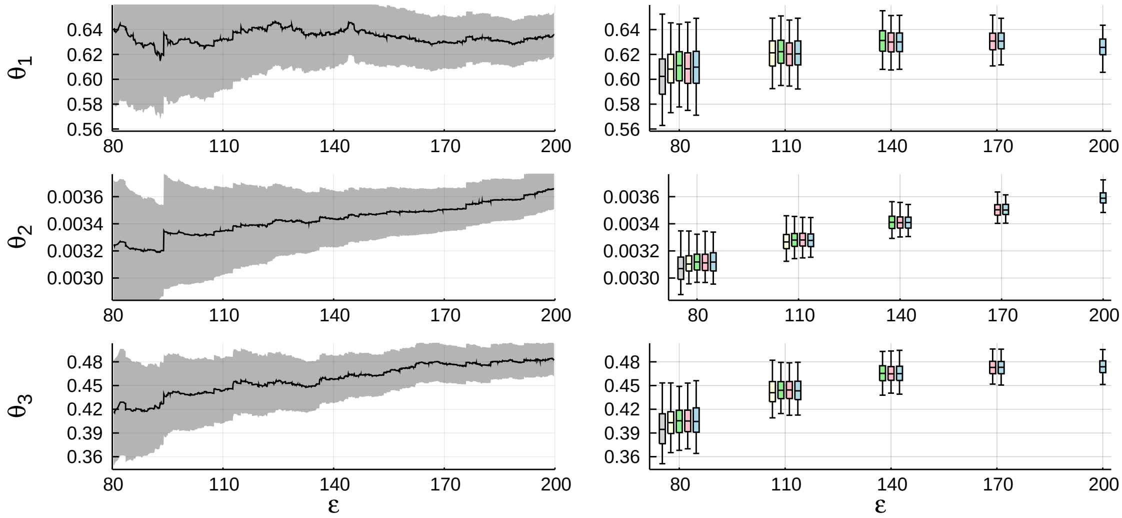

Fig. 3. Lotka-Volterra model with simple cut-off. Estimates from single run of ABC-MCMC(200) (left) and estimates from 1,000 replications of ABC-MCMC(δ) with $\delta \in \{80, 110, 140, 170, 200\}$ indicated by colour.

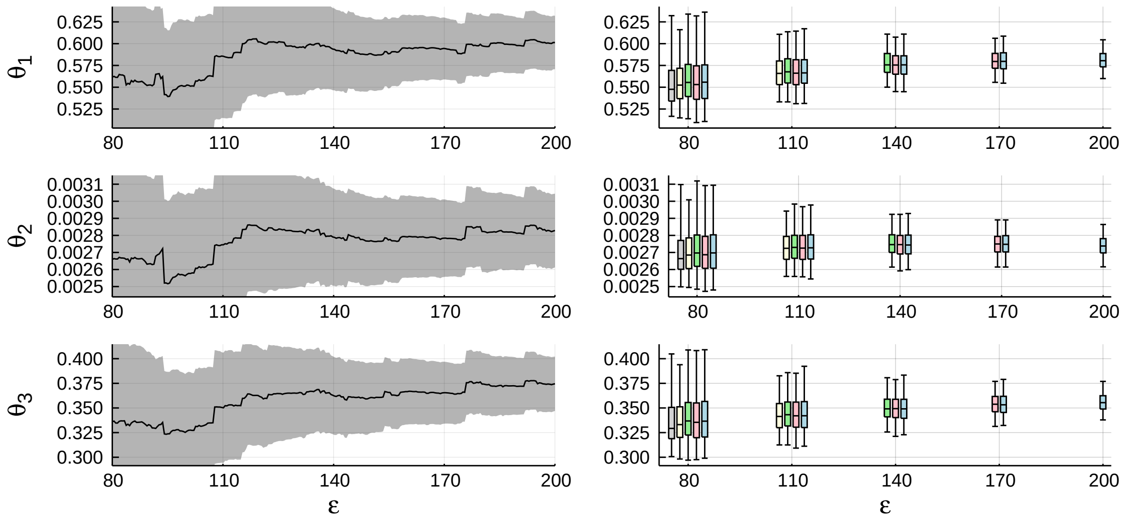

Fig. 4. Lotka-Volterra model with Epanechnikov cut-off and regression correction. Estimates from single run of ABC-MCMC(200) (left) and estimates from 1,000 replications of ABC-MCMC(δ) with $\delta \in \{80, 110, 140, 170, 200\}$ indicated by colour.

In addition, we experiment with the tolerance adaptation, using also 20,000 samples out of which 10,000 are burn-in. Figure 5 shows the progress of the log-tolerance during the burn-in, and histogram of the realised mean acceptance rates during the estimation phase. The realised acceptance rates are concentrated around the mean 0.10. Table 4 shows root mean square errors of the estimators from ABC-MCMC(δ) for $\epsilon = 80$ for fixed

Table 3. *Mean acceptance rates and frequencies of the 95% confidence intervals, from* ABC-MCMC*(δ) to tolerances ϵ, in the Lotka-Volterra model.*

		$f(\theta)=\theta_1$					$f(\theta)=\theta_2$					$f(\theta)=\theta_3$					Acc.
	$\delta \setminus \epsilon$	80	110	140	170	200	80	110	140	170	200	80	110	140	170	200	rate
ϕ_{simple}	80	0.8					0.73					0.74					0.05
	110	0.97	0.93				0.94	0.89				0.94	0.9				0.07
	140	0.99	0.97	0.93			0.98	0.96	0.92			0.98	0.96	0.94			0.1
	170	0.99	0.98	0.96	0.93		0.98	0.97	0.96	0.93		0.99	0.98	0.96	0.95		0.14
	200	1.0	0.99	0.98	0.97	0.94	0.99	0.99	0.98	0.97	0.92	0.99	0.98	0.98	0.96	0.94	0.17
regr. ϕ_{Epa}	80	0.75					0.76					0.68					0.05
	110	0.92	0.92				0.93	0.94				0.87	0.91				0.07
	140	0.93	0.94	0.94			0.94	0.96	0.97			0.9	0.92	0.94			0.1
	170	0.93	0.95	0.95	0.95		0.96	0.97	0.97	0.98		0.92	0.94	0.94	0.95		0.14
	200	0.96	0.96	0.96	0.96	0.96	0.98	0.98	0.98	0.98	0.98	0.95	0.96	0.95	0.96	0.96	0.17

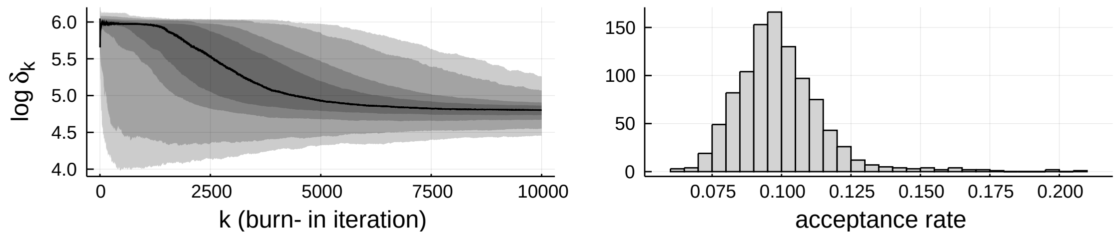

Fig. 5. Progress of tolerance adaptation (left) and histogram of acceptance rates (right) in the Lotka-Volterra experiment.

Table 4. *Root mean square errors of estimators from* ABC-MCMC*(δ) for tolerance $\epsilon = 80$, with fixed tolerance and with adaptive tolerance in the Lotka-Volterra model.*

	Post-correction, simple cut-off						Regression, Epanechnikov cut-off					
	Fixed tolerance					Adapt	Fixed tolerance					Adapt
δ	80	110	140	170	200	122.6	80	110	140	170	200	122.6
θ_1 ($\times 10^{-2}$)	2.37	1.81	1.75	1.83	1.93	1.8	3.1	2.74	3.02	3.09	3.19	2.57
θ_2 ($\times 10^{-4}$)	1.32	0.99	0.93	0.96	1.06	1.04	1.52	1.39	1.54	1.61	1.63	1.28
θ_3 ($\times 10^{-2}$)	2.94	2.26	2.11	2.14	2.37	2.34	2.77	2.53	2.76	2.85	2.91	2.34

tolerance and with tolerance adaptation. Only the adaptive chains with final tolerance ≥ 80.0 were included (999 out of 1,000 chains).

In this case, the chains run with the tolerance adaptation led to better results than those run only with the covariance adaptation and fixed tolerance. This perhaps surprising result may be due to the initial behaviour of the covariance adaptation, which may be unstable when there are many rejections. Different initialisation strategies, for instance following (Haario et al., 2001, Remark 2), might lead to more stable behaviour compared to using the adaptation of Andrieu & Moulines (2006) from the start, as we do. The different step size sequences n^{-1} and $n^{-2/3}$ could also play a rôle. We repeated the experiment for the chains with fixed tolerances, but now with covariance adaptation step size $n^{-2/3}$. This led to more accurate estimators for ABC-MCMC(δ) with higher δ, but worse behaviour with smaller δ. In any case, also here, tolerance adaptation delivered competitive results; see Supplement E.

7. Discussion

We believe that approximate Bayesian computation inference with Markov chain Monte Carlo is a useful approach, when the chosen simulation tolerance allows for good mixing. Our confidence intervals for post-processing and automatic tuning of simulation tolerance may make this approach more appealing in practice.

A related approach by Bortot et al. (2007) makes tolerance an auxiliary variable with a user-specified prior. This approach avoids explicit tolerance selection, but the inference is based on a pseudo-posterior $\breve{\pi}(\theta, \delta)$ not directly related to $\pi_\delta(\theta)$ in (1). Bortot et al. (2007) also provide tolerance-dependent analysis, showing parameter means and variances with respect to conditional distributions of $\breve{\pi}(\theta, \delta)$ given $\delta \leq \epsilon$. We believe that our approach, where the effect of tolerance in the expectations with respect π_ϵ can be investigated explicitly, can be more immediate to interpret. Our confidence interval only shows the Monte Carlo uncertainty related to the posterior mean, and we are currently investigating how the overall parameter uncertainty could be summarised in a useful manner.

The convergence rates of approximate Bayesian computation has been investigated by Barber et al. (2015) in terms of cost and bias with respect to the true posterior, and recently by Frazier et al. (2018) and Li & Fearnhead (2018b,a) in the large data limit, the latter in the context of regression. It would be interesting to consider extensions of these results in the Markov chain Monte Carlo context. In fact, Li & Fearnhead (2018b) already suggest that the acceptance rate must be lower bounded, which is in line with our adaptation rule.

Automatic selection of tolerance has been considered earlier in Ratmann et al. (2007), who propose an algorithm based on tempering and a cooling schedule. Based on our experiments, the tolerance adaptation we present in this paper appears to perform well in practice, and provides reliable results with post-correction. For the adaptation to work efficiently, the Markov chains must be taken relatively long, rendering the approach difficult for the most computationally demanding models.

We conclude with a brief discussion of certain extensions of the suggested post-correction method; more details are given in Supplement D. First, in case of non-simple cut-off, the rejected samples may be 'recycled' by using the acceptance probability as weight (Ceperley et al., 1977). The accuracy of the post-corrected estimator could be enhanced with smaller values of ϵ by performing further independent simulations from $g(\cdot \mid \Theta_k)$, which may be calculated in parallel. The estimator is rather straightforward, but requires some care because the estimators of the pseudo-likelihood take value zero. The latter extension, which involves additional simulations as post-processing, is similar to the 'lazy' version of Prangle (2016, 2015) incorporating a randomised stopping rule for simulation, and to debiased 'exact' approach of Tran & Kohn (2015), which may lead to estimators which get rid of ϵ-bias entirely.

8. Acknowledgements

This work was supported by Academy of Finland (grants 274740, 284513 and 312605). The authors wish to acknowledge CSC, IT Center for Science, Finland, for computational resources, and thank Christophe Andrieu for useful discussions.

9. Supplementary material

Supplementary material available at Biometrika online includes proofs of tolerance adaptation convergence and additional results.

References

Andrieu, C., Lee, A. & Vihola, M. (2018). Theoretical and methodological aspects of MCMC computations with noisy likelihoods. In *Handbook of Approximate Bayesian Computation*, S. A. Sisson, Y. Fan & M. Beaumont, eds. Chapman & Hall/CRC Press.

Andrieu, C. & Moulines, É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.* **16**, 1462–1505.

Andrieu, C. & Thoms, J. (2008). A tutorial on adaptive MCMC. *Statist. Comput.* **18**, 343–373.

Barber, S., Voss, J. & Webster, M. (2015). The rate of convergence for approximate Bayesian computation. *Electron. J. Statist.* **9**, 80–105.

Beaumont, M., Zhang, W. & Balding, D. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.

Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM review* **59**, 65–98.

Bornn, L., Pillai, N., Smith, A. & Woodard, D. (2017). The use of a single pseudo-sample in approximate Bayesian computation. *Statist. Comput.* **27**, 583–590.

Bortot, P., Coles, S. & Sisson, S. (2007). Inference for stereological extremes. *J. Amer. Statist. Assoc.* **102**, 84–92.

Boys, R. J., Wilkinson, D. J. & Kirkwood, T. B. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Stat. Comput.* **18**, 125–135.

Ceperley, D., Chester, G. & Kalos, M. (1977). Monte Carlo simulation of a many-fermion study. *Phys. Rev. D* **16**, 3081.

Doucet, A., Pitt, M., Deligiannidis, G. & Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika* **102**, 295–313.

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

Flegal, J. M. & Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.* **38**, 1034–1070.

Franks, J. & Vihola, M. (2017). Importance sampling correction versus standard averages of reversible MCMCs in terms of the asymptotic variance. Preprint arXiv:1706.09873v3.

Frazier, D. T., Martin, G. M., Robert, C. P. & Rousseau, J. (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika* **105**, 593–607.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statist. Sci.* , 473–483.

Haario, H., Saksman, E. & Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7**, 223–242.

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

Li, W. & Fearnhead, P. (2018b). On the asymptotic efficiency of approximate Bayesian computation estimators. *Biometrika* **105**, 285–299.

Marin, J.-M., Pudlo, P., Robert, C. P. & Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statist. Comput.* **22**, 1167–1180.

Marjoram, P., Molitor, J., Plagnol, V. & Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100**, 15324–15328.

Meyn, S. & Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability.* Cambridge University Press, 2nd ed.

Prangle, D. (2015). Lazier ABC. Preprint arXiv:1501.05144.

Prangle, D. (2016). Lazy ABC. *Statist. Comput.* **26**, 171–185.

Ratmann, O., Jørgensen, O., Hinkley, T., Stumpf, M., Richardson, S. & Wiuf, C. (2007). Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum. *PLoS Comput. Biol.* **3**, e230.

Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P. & Estoup, A. (to appear). ABC random forests for Bayesian parameter inference. *Bioinformatics* .

Roberts, G., Gelman, A. & Gilks, W. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7**, 110–120.

Roberts, G. O. & Rosenthal, J. S. (2006). Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *Ann. Appl. Probab.* **16**, 2123–2139.

Sherlock, C., Thiery, A. H., Roberts, G. O. & Rosenthal, J. S. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.* **43**, 238–275.

Sisson, S. & Fan, Y. (2018). ABC samplers. In *Handbook of Markov chain Monte Carlo*, S. Sisson, Y. Fan & M. Beaumont, eds. Chapman & Hall/CRC Press.

Sokal, A. D. (1996). Monte Carlo methods in statistical mechanics: Foundations and new algorithms. Lecture notes.

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M. & Dessimoz, C. (2013). Approximate Bayesian computation. *PLoS computational biology* **9**, e1002803.

Tanaka, M., Francis, A., Luciani, F. & Sisson, S. (2006). Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics* **173**, 1511–1520.

Tran, M. N. & Kohn, R. (2015). Exact ABC using importance sampling. Preprint arXiv:1509.08076.

Vihola, M. (2012). Robust adaptive Metropolis algorithm with coerced acceptance rate. *Statist. Comput.* **22**, 997–1008.

Vihola, M., Helske, J. & Franks, J. (2016). Importance sampling type estimators based on approximate marginal MCMC. Preprint arXiv:1609.02541v5.

Wegmann, D., Leuenberger, C. & Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihoods. *Genetics* **182**, 1207–1218.

Appendix

The following algorithm shows that in case of simple post-correction cut-off, $E_{\delta,\epsilon}(f)$ and $S_{\delta,\epsilon}(f)$ may be calculated simultaneously for all tolerances efficiently:

Algorithm 4. Suppose $\phi = \phi_{\text{simple}}$ and $(\Theta_k, T_k)_{k=1,\ldots,n}$ is the output of ABC-MCMC(δ).

(i) Sort $(\Theta_k, T_k)_{k=1,\ldots,n}$ with respect to T_k:
  - Find indices $I_1, \ldots, I_n$ such that $T_{I_k} \le T_{I_{k+1}}$ for all $k = 1, \ldots, n-1$.
  - Denote $(\hat{\Theta}_k, \hat{T}_k) \leftarrow (\Theta_{I_k}, T_{I_k})$.

(ii) For all unique values $\epsilon \in \{\hat{T}_1, \ldots, \hat{T}_n\}$, let $m_\epsilon = \max\{k \ge 1 : \hat{T}_k \le \epsilon\}$, and define

$$E_{\delta,\epsilon}(f) = m_\epsilon^{-1} \textstyle\sum_{k=1}^{m_\epsilon} f(\hat{\Theta}_k), \qquad S_{\delta,\epsilon}(f) = m_\epsilon^{-2} \textstyle\sum_{k=1}^{m_\epsilon} \{f(\hat{\Theta}_k) - E_{\delta,\epsilon}(f)\}^2.$$

(and for $\hat{T}_k < \epsilon < \hat{T}_{k+1}$, let $E_{\delta,\epsilon}(f) = E_{\delta,\hat{T}_k}(f)$ and $S_{\delta,\epsilon}(f) = S_{\delta,\hat{T}_k}(f)$.)

The sorting in Algorithm 4(i) may be performed in $O(n \log n)$ time, and $E_{\delta,\epsilon}(f)$ and $S_{\delta,\epsilon}(f)$ may all be calculated in $O(n)$ time by forming appropriate cumulative sums.

Proof of Theorem 1. Algorithm 1 is a Metropolis–Hastings algorithm with compound proposal $\tilde{q}(\theta, y; \theta', y') = q(\theta, \theta')g(y' \mid \theta')$ and with target $\tilde{\pi}_\epsilon(\theta, y) \propto \text{pr}(\theta)g(y \mid \theta)\phi\big(d(y, y^*)/\epsilon\big)$. The chain $(\Theta_k, Y_k)_{k\ge1}$ is Harris-recurrent, as a full-dimensional Metropolis–Hastings which is φ-irreducible (Roberts & Rosenthal, 2006). Because ϕ is monotone and $\epsilon \le \delta$, we have $\phi\big(d(y, y^*)/\delta\big) \ge \phi\big(d(y, y^*)/\epsilon\big)$, and therefore $\tilde{\pi}_\epsilon$ is absolutely continuous with respect to $\tilde{\pi}_\delta$, and $w_{\delta,\epsilon}(y) = c_{\delta,\epsilon}\tilde{\pi}_\epsilon(\theta, y)/\tilde{\pi}_\delta(\theta, y)$, where $c_{\delta,\epsilon} > 0$ is a constant. If we denote $\xi_k(f) = U_k^{(\delta,\epsilon)} f(\Theta_k)$ and $\xi_k(\mathbf{1}) = U_k^{(\delta,\epsilon)} = w_{\delta,\epsilon}(Y_k)$, then $E_{\delta,\epsilon}^{(n)}(f) = \sum_{k=1}^n \xi_k(f) / \sum_{j=1}^n \xi_j(\mathbf{1}) \to \mathbb{E}_{\tilde{\pi}_\epsilon}[f(\Theta)]$ almost surely by Harris recurrence and $\tilde{\pi}_\epsilon$ invariance (e.g. Vihola et al., 2016). The claim (i) follows because π_ϵ is the marginal density of $\tilde{\pi}_\epsilon$.

The chain $(\Theta_k, Y_k)_{k\geq 1}$ is reversible, so (ii) follows by (Vihola et al., 2016, Theorem 7(i)), because $m_f^{(2)}(\theta, y) = w_{\delta,\epsilon}^2(y) f^2(\theta)$ satisfies

$$\mathbb{E}_{\tilde{\pi}_\delta}\{m_f^{(2)}(\Theta, Y)\} = c_{\delta,\epsilon}\mathbb{E}_{\tilde{\pi}_\epsilon}\{w_{\delta,\epsilon}(Y)f^2(\Theta)\} \leq c_{\delta,\epsilon}\mathbb{E}_{\pi_\epsilon}\{f^2(\Theta)\} < \infty,$$

and because the asymptotic variance of the function $h_{\delta,\epsilon}$ with respect to $(\Theta_k, Y_k)_{k\geq 1}$ may be expressed as $\mathrm{var}_{\tilde{\pi}_\delta}\{h_{\delta,\epsilon}(\Theta, Y)\}\tau_{\delta,\epsilon}(f)$, so $v_{\delta,\epsilon}(f) = \mathrm{var}_{\tilde{\pi}_\delta}\{h_{\delta,\epsilon}(\Theta, Y)\}/c_{\delta,\epsilon}^2$. The convergence $nS_{\delta,\epsilon}^{(n)}(f) \to v_{\delta,\epsilon}(f)$ follows from (Vihola et al., 2016, Theorem 9). □

Proof of Theorem 2. The invariant distribution of ABC-MCMC(δ) may be written as $\tilde{\pi}_\delta(\theta, y) = \pi_\delta(\theta)\bar{g}_\delta(y \mid \theta)$ where $\bar{g}_\delta(y \mid \theta) = g(y \mid \theta)1\left(d(y, y^*) \leq \delta\right)/L_\delta(\theta)$, and that $\int \bar{g}_\delta(y \mid \theta) w_{\delta,\epsilon}^p(y)\mathrm{d}y = \bar{w}_{\delta,\epsilon}(\theta)$ for $p \in \{1, 2\}$. Consequently, $\tilde{\pi}_\delta(h_{\delta,\epsilon}) = \pi_\delta(f_{\delta,\epsilon})$ and $\tilde{\pi}_\delta(h_{\delta,\epsilon}^2) = \pi_\delta(f^2\bar{w}_{\delta,\epsilon})$, so $\mathrm{var}_{\tilde{\pi}_\delta}(h_{\delta,\epsilon}) = \mathrm{var}_{\pi_\delta}(f_{\delta,\epsilon}) + \pi_\delta\big(\bar{w}_{\delta,\epsilon}(1 - \bar{w}_{\delta,\epsilon})f^2\big)$. Hereafter, let $a_{\delta,\epsilon} = \{\mathrm{var}_{\tilde{\pi}_\delta}(h_{\delta,\epsilon})\}^{-1/2}$ and denote $\tilde{h}_{\delta,\epsilon} = a_{\delta,\epsilon}h_{\delta,\epsilon}$ and $\tilde{f}_{\delta,\epsilon} = a_{\delta,\epsilon}f_{\delta,\epsilon}$. Clearly, $\mathrm{var}_{\tilde{\pi}_\delta}(\tilde{h}_{\delta,\epsilon}) = 1$ and

$$\rho_k^{(\delta,\epsilon)} = e_k^{(\delta,\epsilon)} - \{\pi_\delta(\tilde{f}_{\delta,\epsilon})\}^2, \qquad e_k^{(\delta,\epsilon)} = \mathbb{E}\big\{\tilde{h}_{\delta,\epsilon}(\Theta_0^{(s)}, Y_0^{(s)})\tilde{h}_{\delta,\epsilon}(\Theta_k^{(s)}, Y_k^{(s)})\big\}.$$

Note that with $\phi = \phi_{\mathrm{simple}}$, the acceptance ratio is $\alpha_\delta(\theta, y; \hat{\theta}, \hat{y}) = \dot{\alpha}(\theta, \hat{\theta})1\left(d(\hat{y}, y^*) \leq \delta\right)$, where $\dot{\alpha}(\theta, \hat{\theta}) = \min[1, \mathrm{pr}(\hat{\theta})q(\hat{\theta}, \theta)/\{\mathrm{pr}(\theta)q(\theta, \hat{\theta})\}]$, which is independent of y, so $(\Theta_k^{(s)})$ is marginally a Metropolis–Hastings type chain, with proposal q and acceptance probability $\alpha(\theta, \hat{\theta})L_\delta(\hat{\theta})$, and

$$\begin{aligned} &\mathbb{E}\big\{\tilde{h}_{\delta,\epsilon}(\Theta_1^{(s)}, Y_1^{(s)}) \bigm| (\Theta_0^{(s)}, Y_0^{(s)}) = (\theta, y)\big\} - r_\delta(\theta)\tilde{h}_{\delta,\epsilon}(\theta, y) \\ &= a_{\delta,\epsilon}\int q(\theta, \hat{\theta})\dot{\alpha}(\theta, \hat{\theta})g(\hat{y} \mid \hat{\theta})w_{\delta,\epsilon}(\hat{y})f(\hat{\theta})\mathrm{d}\hat{\theta}\mathrm{d}\hat{y} = \int q(\theta, \hat{\theta})\dot{\alpha}(\theta, \hat{\theta})L_\delta(\hat{\theta})\tilde{f}_{\delta,\epsilon}(\hat{\theta})\mathrm{d}\hat{\theta}. \end{aligned}$$

Using this iteratively, we obtain that

$$e_k^{(\delta,\epsilon)} = \mathbb{E}\{\tilde{f}_{\delta,\epsilon}(\Theta_0^{(s)})\tilde{f}_{\delta,\epsilon}(\Theta_k^{(s)})\} + \textstyle\int \tilde{\pi}_\delta(\theta, y)\{\tilde{h}_{\delta,\epsilon}^2(\theta, y) - \tilde{f}_{\delta,\epsilon}^2(\theta)\}r_\delta^k(\theta)\mathrm{d}\theta\mathrm{d}y,$$

and therefore with $\gamma_k^{(\delta,\epsilon)} = a_{\delta,\epsilon}^2\mathrm{cov}\{f_{\delta,\epsilon}(\Theta_0^{(s)}), f_{\delta,\epsilon}(\Theta_k^{(s)})\}$,

$$\textstyle\sum_{k\geq 1}\rho_k^{(\delta,\epsilon)} = \sum_{k\geq 1}\gamma_k^{(\delta,\epsilon)} + a_{\delta,\epsilon}^2\int \pi_\delta(\theta)\bar{w}_{\delta,\epsilon}(\theta)\{1 - \bar{w}_{\delta,\epsilon}(\theta)\}r_\delta(\theta)\{1 - r_\delta(\theta)\}^{-1}f^2(\theta)\mathrm{d}\theta.$$

We conclude by noticing that $2\sum_{k\geq 1}\gamma_k^{(\delta,\epsilon)} = a_{\delta,\epsilon}^2\mathrm{var}_{\pi_\delta}(f_{\delta,\epsilon})\{\check{\tau}_{\delta,\epsilon}(f) - 1\}$. □