How approximate is *Approximate Bayesian Computation*?

Marika Di Marcantonio, Francesca Pietrobon, Raffaele Saviello

Bayesian Statistics A.Y. 2020/2021

1 Introduction

In Bayesian inference, to have a complex model and/or prior means that the posterior distribution is not available in closed form and so numerical methods are needed to proceed.

It's typical to use a Monte Carlo approach but the numerical evaluation of the likelihood function could be either computationally prohibitive or simply not possible, for the scientific problems that we have to face nowadays.

The most realistic thing to do, in this case, is to consider an approximation to the model at the expense of some error. One of the most effective method to perform an approximate Bayesian analysis, is the Bayesian computation **(ABC)** which is a particular case of *likelihood free* Bayesian method.

1.1 Likelihood-Free

'Likelihood-free' refers to any likelihood-based analysis that proceeds without direct numerical evaluation of the likelihood function.

Under the assumption of discrete data generated by the model $y \sim p(y|\theta)$, if we assume the prior is used as sampling distribution (i.e. $f(\theta) \sim \pi(\theta|y_{obs})$) then the acceptance probability will be proportional to the likelihood.
In particular, the acceptance probability is approximated without its numerical evaluation: we can accept or reject draw from the sampling density without the acceptance probability's computation.

- **Likelihood-Free Rejection Sampling Algorithm**
  For i=1:N
  1. Generate $\theta^{(i)} \sim g(\theta)$
  2. Generate $y \sim p(y|\theta^{(i)})$

3. If $\|y - y_{obs}\| \leq h$ then accept $\theta^{(i)}$ with probability $\frac{\pi(\theta^{(i)})}{Kg(\theta^{(i)})}$ otherwise go to 1.

where $K \geq max_\theta \frac{\pi(\theta)}{g(\theta)}$

The output samples now will be draws from an approximation of $\theta^{(i)} \sim g(\theta)$. Increasing h will considerably improve the acceptance rate of the algorithm but make the accuracy worse.

When the dimension of the vector y_{obs} is high, it's very difficult to generate y close to y_{obs}. This implies that the likelihood-free approximation to $p(\theta|y_{obs})$ isn't efficient and the true parameter vector θ_0 is not even close to the estimated posterior samples.
The odds of matching y with y_{obs} can be increased by reducing the dimension of the comparison $y - y_{obs}$ by using the lower dimensional statistics $s = S(y)$ sufficient, or highly informative, for θ under the model and $s_{obs} = S(y_{obs})$ such that $dim(S(y)) = dim(y)$.

Then $\|y - y_{obs}\|$ might be replaced by $\|s - s_{obs}\|$ without too much loss of information. That implies a modification in step 3. in the algorithm as follows:

3. If $\|s - s_{obs}\| \leq h$ then accept $\theta^{(i)}$ with probability $\frac{\pi(\theta^{(i)})}{Kg(\theta^{(i)})}$, else go to 1.

2 ABC

Approximate Bayesian Computation it's a likelihood-free method, it produces an approximation of the posterior distribution using $\|y - y_{obs}\|$ or summary statistics $\|s - s_{obs}\|$.
The procedure seen above is equivalent to drawing a sample (θ, y) from the joint distribution proportional to $I(\|y - y_{obs}\| \leq h)p(y|\theta)g(\theta)$.
If this sample (θ, y) is accepted with probability proportional to $\pi(\theta)/g(\theta)$, the likelihood-free rejection algorithm is sampling from the joint distribution proportional to:

$$I(\|y - y_{obs}\| \leq h)p(y|\theta)g(\theta)\frac{\pi(\theta)}{g(\theta)} = I(\|y - y_{obs}\| \leq h)p(y|\theta)\pi(\theta) \qquad (1)$$

If $h = 0$, then the θ marginal equals the true posterior so the likelihood-free rejection algorithm draws samples, (θ, y), for which the marginal distribution of the parameter vector is the true posterior, $\pi(\theta|y_{obs})$.
We want to generalize this formulation introducing a more continuous scaling because $I(\|y - y_{obs}\| \leq h)$ only takes the value 0 or 1 so it does not distinguish the samples for which $y = y_{obs}$ and samples for which y is far from y_{obs}.
This can be obtained considering indicator function with a standard smoothing kernel function, $K_h(u)$, with $u = \|y - y_{obs}\|$, where:

$$K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$$

With $K : K(u) \geq 0 \ \forall u$, $\int K(u)du = 1$, $\int uK(u)du = 0$ and $\int u^2K(u)du < \infty$ where $h > 0$ corresponds to the scale parameter of the kernel function.

Substituting the kernel function, $K_h(u)$ into the likelihood-free rejection algorithm we obtain the so called:

- **ABC Rejection Sampling Algorithm**
  For i=1:N

  1. Generate $\theta^{(i)} \sim g(\theta)$
  2. Generate $y \sim p(y|\theta^{(i)})$
  3. Accept $\theta^{(i)}$ with probability $\frac{K_h(\|y-y_{obs}\|)\pi(\theta^{(i)})}{Kg(\theta^{(i)})}$ otherwise go to 1.

  where $K \geq K_h(0)max_\theta \frac{\pi(\theta)}{g(\theta)}$

The target distribution have the following form:

$$\pi_{ABC}(\theta, y|y_{obs}) \propto K_h(\|y - y_{obs}\|)p(y|\theta)\pi(\theta) \tag{2}$$

but it works very well in case when y_{obs} is very low dimensional or when the likelihood function $p(y|\theta)$ factorises into very low dimensional components. For this reason this approximation is rarely used, indeed it is highly unlikely that $y \sim y_{obs}$ can be generated from $p(y|\theta)$ in realistic scenario.

A solution to that problem is to use **sufficient statistics** because they can be much lower dimensional than the full dataset and so, we'll achieve a greater approximation accuracy *(which is hinted at in the g-and-k distribution analysis).*

The most efficient choice is the minimal sufficient statistic, but it may still be non-viable in practice. In general, a typical ABC analysis will involve specification of a vector of summary statistics $s = S(y)$, where $dim(s) \ll dim(y)$. The rejection sampling algorithm with the contrast s with $s_{obs} = S(y_{obs})$, rather than y with y_{obs}. As a result, this procedure will produce samples form the distribution $\pi_{ABC}(\theta|s_{obs})$ as follows:

- **ABC Rejection Sampling Algorithm**
  For i=1:N

  1. Generate $\theta^{(i)} \sim g(\theta)$
  2. Generate $y \sim p(y|\theta^{(i)})$
  3. Compute summary statistic $s = S(y)$
  4. Accept $\theta^{(i)}$ with probability $\frac{K_h(\|s-s_{obs}\|)\pi(\theta^{(i)})}{Kg(\theta^{(i)})}$ otherwise go to 1.

  where $K \geq K_h(0)max_\theta \frac{\pi(\theta)}{g(\theta)}$

Similarly to the previous discussion, it can be shown that the new ABC posterior approximation is

$$\pi_{ABC}(\theta|s_{obs}) \propto \int K_h(\|s - s_{obs}\|)p(s|\theta)\pi(\theta)ds \tag{3}$$

where $p(s|\theta)$ denotes the likelihood of the summary statistic implied by $p(y|\theta)$.

$$\lim_{h\to 0} \pi_{ABC}(\theta|s_{obs}) \propto p(\theta|s_{obs})\pi(\theta) \tag{4}$$

If the vector of summary statistics is sufficient for the model parameters then $\pi(\theta|s_{obs}) \equiv \pi(\theta|y_{obs})$, thus samples are produced from the true posterior distribution. Otherwise the ABC posterior approximation is given by (3) where, in the best scenario ($h \to 0$), the approximation is given by $\pi(\theta|s_{obs})$.

2.1 Choice of summary statistics

The choice of summary statistics for an ABC analysis is an important decision that have consequences on the posterior approximation.

We have to make a trade off between the approximation of $\pi(\theta|y_{obs})$ with $\pi(\theta|s_{obs})$ and the one of s with s_{obs} (using a smoothing kernel $K_h(\|s - s_{obs}\|)$). However, the first one could imply information loss thus s_{obs} should contain high information. On the other hand, matching s with s_{obs} it's very difficult as the dimension of the summary statistics increases and, consequentially, the dimension of s should be low.

We can obtain an accurate ABC posterior approximation using a lower-dimensional non-sufficient statistics rather than use sufficient statistics (this is the case of g-and-k distribution analysis).

2.2 Example

Suppose we have $y = (y_1, \ldots, y_n)^T$ where $y_i \sim Poisson(\lambda)$ and $\lambda \sim Gamma(\alpha, \beta)$ as prior $\Rightarrow \lambda|y \sim Gamma(\alpha + n\bar{y}, \beta + n)$.
Let's consider $y = (0, 0, 0, 0, 5)^T$ so that $(\bar{y_{obs}}, v^2_{obs}) = (1, 5)$. For this example the sample mean $\bar{y}$ is a sufficient statistic.

By observing the plots of various ABC approximation of the posterior by applying a uniform kernel over $[-h, h]$ based on the euclidean distance, we note that when $s = \bar{y}$ or $s = v$ the approximation are less deviated in the case h=0. The third panel, in the figure, is clearly biased to the right, with the resulting ABC posterior approximation visually appearing to be a loose average of those distributions.

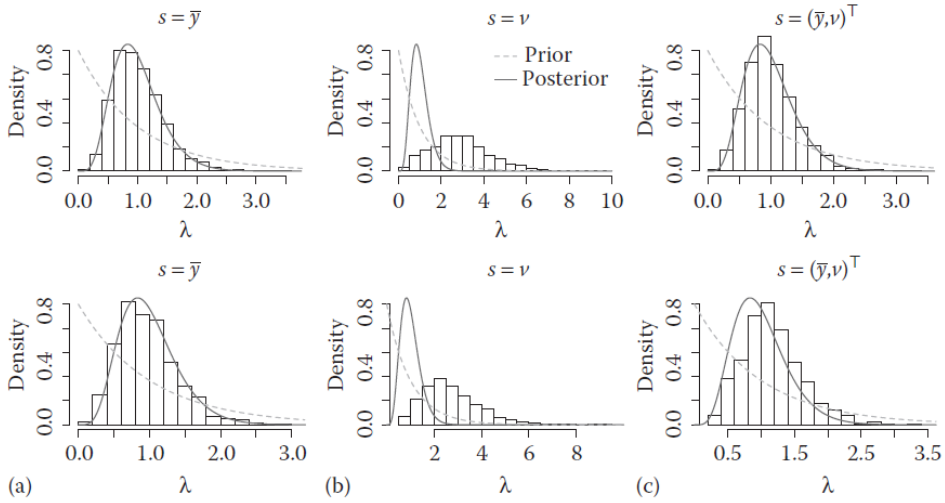


Figure 1: ABC posterior approximations for *Gamma(*$1 + \bar{y}, 1 + n$*)* with prior *Gamma(1,1)* based on (a)sample mean $s = \bar{y}$, (b)standard deviation $s = v$ and (c)$s = (\bar{y}, v)^T$ as summary statistics. Result for h=0 on the top, h=0.3 below.

2.3 Choice of distance measure

While the choice of summary statistics is itself of primary importance, it is less appreciated that the distance measure $\|\cdot\|$ can also have a substantial impact on ABC algorithm efficiency, and therefore the quality of the posterior approximation.

3 Levels of Approximation in ABC

The challenge in implementing an ABC analysis is to reduce the approximation's impact and, simultaneously, the required computation. We can summarise the nature of the approximations involved in any ABC analysis as follows:

1. *All models are approximations to the real data-generation process*
2. *Use of summary statistics rather than full datasets*
3. *Weighting of summary statistics within a region of the observed summary statistics*
4. *Approximations due to other ABC techniques*
5. *Monte Carlo error*

Problems that can arise with these kind of approximation are:

1. When the model isn't able to reproduce the observed summary statistics, meaning that the simulated data are far from the observed data, the quality of the ABC approximation may be low. *(Inflexible model)*

2. The full posterior $\pi(\theta|y_{obs})$ is approximated by $\pi(S(y_{obs})|\theta)\pi(\theta)$ and if S is sufficient for θ, then there is no approximation at this stage. More commonly for non-sufficient S, there is a loss of information.

3. The partial posterior $\pi(\theta|s_{obs})$ is approximated by

$$\pi_{ABC}(\theta|s_{obs}) = \pi(\theta) \int K_h(\|s - s_{obs}\|)p(s|\theta)ds$$

where K_h is a standard smoothing kernel with scale parameter $h \geq 0$.
If $h > 0$, as in most cases, ABC makes use of a kernel density estimate as an approximation to the true likelihood function (hence that causes problems whenever θ is large because it will implies the same for the vector of summary statistics s), otherwise h=0 and there is no further approximation.

4 Example: univariate g-and-k distribution

The *univariate g-and-k distribution* is a flexible unimodal distribution that is able to describe data with significant amounts of skewness and kurtosis. Its density function has no closed form so it's known through its quantile function:

$$Q(q|A, B, g, k) = A + B(1 + c\frac{1-e^{-g*z(q)}}{1+e^{g*z(q)}})(1 + z(q)^2)^k z(q)$$

where $B > 0, k > -1/2, z(q) = \phi^{-1}(q)$ and $c = 0.8$.
Note that when $g = k = 0$, the distribution well approximate a Normal one.

4.1 Wasserstein distance

In order to avoid loss of information by using summaries, we consider the Wasserstein distance between the empirical distributions of the observed and synthetic data.

$$\mathcal{W}_p(y_{1:n}, z_{1:n})^p = inf_{\sigma \in \mathcal{S}_n} \frac{1}{n} \sum_{i=}^{n} |y_i, z_{\sigma(i)}|^p$$

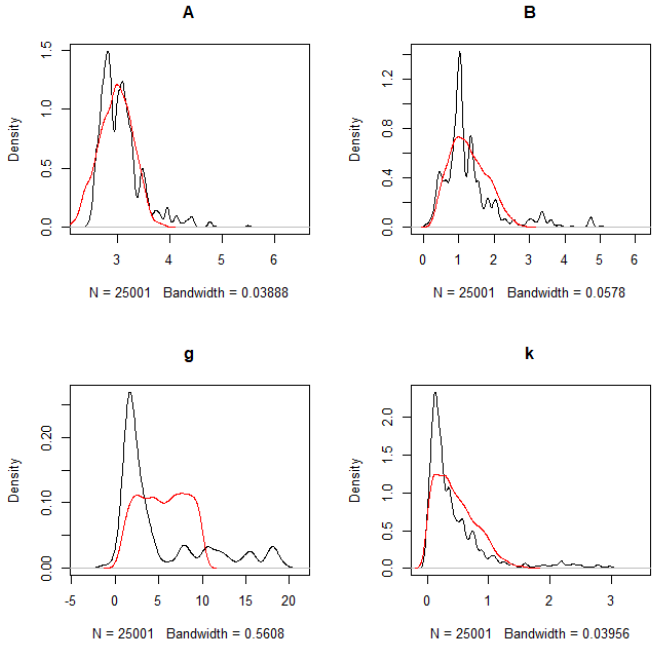
where $\mathcal{S}_n$ is the set of permutations of $\{1, \ldots, n\}$.
Consequently the infimum above is achieved by sorting $y_{1:n}$ and $z_{1:n}$ in increasing order. That can be thought as a generalization of the use of order statistics within ABC to arbitrary dimensions.

4.2 Numerical experiment

The *g-and-k distribution* probability density function is intractable but can be numerically calculated with high precision since it only involves one-dimensional

inversions and differentiations of the quantile function. Then, Bayesian inference was carried out with Markov chain Monte Carlo.

We generate $n = 250$ observations from the model using $a = 3; b = 1; g = 2; k = 0.5$ and the parameters are assigned a uniform prior on $[0; 10]^4$. For the ABC approximation we first use a rejection sampler.



Posterior marginals obtained via MCMC in black, WABC approximations with a budget of $2.4 * 10^6$ model simulations in red

The WABC posteriors appear to better approximate the marginal posterior distributions in the case of parameters a, b and k but the method fails in estimating the target distribution for g. Nonetheless, all graphics are concentrated in the region of interest on every parameter.

5 Semi-automatic ABC

The core idea behind semi-automatic ABC is that we can use simulation to estimate appropriate summary statistics that are equal to posterior means.
The steps are:

1. *simulate sets of parameter values and data*
2. *use the simulated sets of parameter values and data to estimate the summary statistics*

3. *run ABC with this choice of summary statistics*

1. Simulate M sets of parameter values from the prior and for each of them we simulate an artificial dataset.

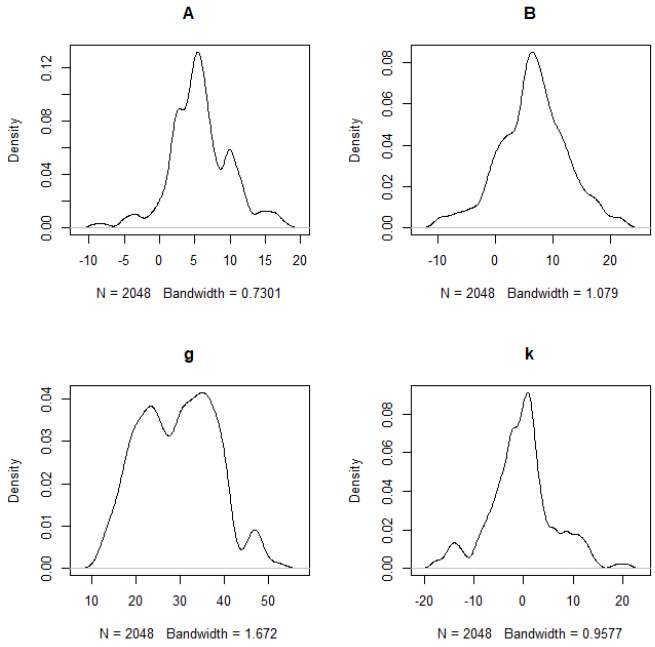
2. We consider a vector of linear transformation of the data, for example $f(y) = (y, y^2)$, that is a vector of length $4n$ composed by all the original data followed by the data transformed with all second to fourth powers of the data y. Then we fit the following model

$$\theta_i = E(\theta_i|y) + \epsilon_i = \beta_0^{(i)} + \beta^{(i)} f(y) + \epsilon_i$$

where θ_i is the ith parameter and ϵ_i is some mean-zero noise.
So $E(\theta_i|y)$ is estimated by the fitted function $\beta_0^{(i)} + \beta^{(i)} f(y)$. ABC uses only the difference in summary statistics so β_0 can be neglected and the summary statistic for ABC is just $\hat{\beta}^{(i)} f(y)$.

5.1 Numerical experiment

As we run the *g-and-k* example, we see a huge difference between the following approximation, obtained with the previous algorithm, and the one derived from the WABC acceptance-rejection sampling.



WABC semi-auto approximation

The point is that semi-automatic doesn't work very well in this case. Moreover we had to perform a regression step used in constructing the summary statistics and, even parallelization doesn't fix this problem because it requires a lot of memory (too much for a standard personal computer!).

6 ABC SMC

Approximate Bayesian Computation methods can also be applied when based on sequential Monte Carlo (SMC) to estimate parameters of dynamical models. It gives information about the inferability of parameters and the model sensitivity to changes in parameters but, more importantly for the aim of this project, it tends to perform better than other ABC approaches.

- **ABC Sequential Monte Carlo**

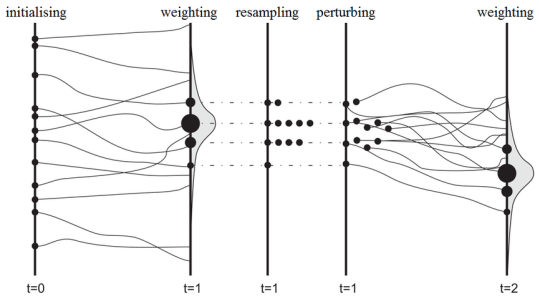
For i=1,...,N

1. Generate $\theta_0^{(i)} \sim g(\theta)$
2. Generate $y_0^{(i)}(t) \sim p(y|\theta_0^{(i)})$ and compute $s_0^{(i)}(t) = S(y_0^{(i)}), \forall t = 1, \dots, T$
3. Compute weights $w_0^{(i)} = \pi(\theta_0^{(i)})/g(\theta_0^{(i)})$ and set $m = 1$
4. Sampling:
   (a) *Reweigth*
   Determine $h_m : ESS(w_m^{(m)}, \dots, w_m^{(N)}) = \alpha ESS(w_{m-1}^{(m)}, \dots, w_{m-1}^{(N)})$ where $w_m^{(i)} = w_{m-1}^{(i)} \frac{\sum_{t=1}^{T} K_{h_m}(||s_{m-1}^{(i)}(t) - s_{obs}||)\pi(\theta_m^{(i)})}{\sum_{t=1}^{T} K_{h_{m-1}}(||s_{m-1}^{(i)}(t) - s_{obs}||)\pi(\theta_{m-1}^{(i)})}$, then compute new particle weigths and set $\theta_m^{(i)} = \theta_{m-1}^{(i)}$, $s_m^{(i)} = s_{m-1}^{(i)}$ $\forall i = 1 : n$ $\forall t = 1 : T$
   (b) *Resample*
   If $ESS(w_m^{(m)}, \dots, w_m^{(N)}) < E$ then resample N particles from $\{\theta_m^{(i)}, s_m^{(i)}(1), \dots, s_m^{(i)}(T), w_m^{(i)} / \sum_{j=1}^{N} w_m^{(j)}\}$ and set $w_m^{(i)} = 1/N$
   (c) *Move*
   For i=1,...,N
   If $w_m^{(i)} > 0$
   - Generate $\theta' \sim g_m(\theta_m^{(i)}, \theta), y'(t) \sim p(y|\theta_m^{(i)})$ and compute $s'(t) = S(y'(t)) \forall t = 1, \dots, T$
   - Accept θ' with probability $\min\{1, \frac{\sum_{t=1}^{T} K_{h_m}(||s_{m-1}^{(i)}(t) - s_{obs}||)\pi(\theta_m^{(i)})g(\theta', \theta_m^{(i)})}{\sum_{t=1}^{T} K_{h_{m-1}}(||s_{m-1}^{(i)}(t) - s_{obs}||)\pi(\theta_{m-1}^{(i)})g(\theta_m^{(i)}, \theta')}\}$ and set $\theta_m^{(i)} = \theta'$, $s_m^{(i)}(t) = s'(t) \forall t = 1, \dots, T$
   - Increment $m = m + 1$.
     If stopping rule is not satisfied, go to (a)

The stopping rule to which the algorithm refers to, is a condition that checks whether a pre-specified degree of accuracy has been achieved. Thus, the number of iterations that the algorithm ran will be an indicator of the model specification's efficiency.

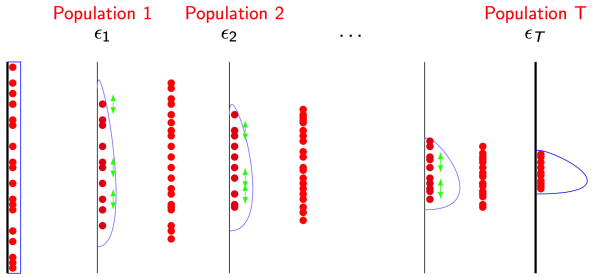
The output will be a set of weighted parameter vectors $(\theta_M^{(1)}, w_M^{(1)}), \ldots, (\theta_M^{(N)}, w_M^{(N)})$ drawn from $\pi_{ABC}(\theta|s_{obs}) \propto \int K_{h_M}(||s - s_{obs}||)p(s|\theta)\pi(\theta)$

It's useful to visualize the algorithm in order to understand it:



D. Alvares, "Sequential Monte Carlo methods in Bayesian joint models for longitudinal and time-to-event data" (2017).

If we consider $t = 1, \ldots, T$, we'll have the following procedure:



T. Toni, M. Stumpf, "Tutorial on ABC rejection and ABC SMC for parameter estimation and model selection" (2009).

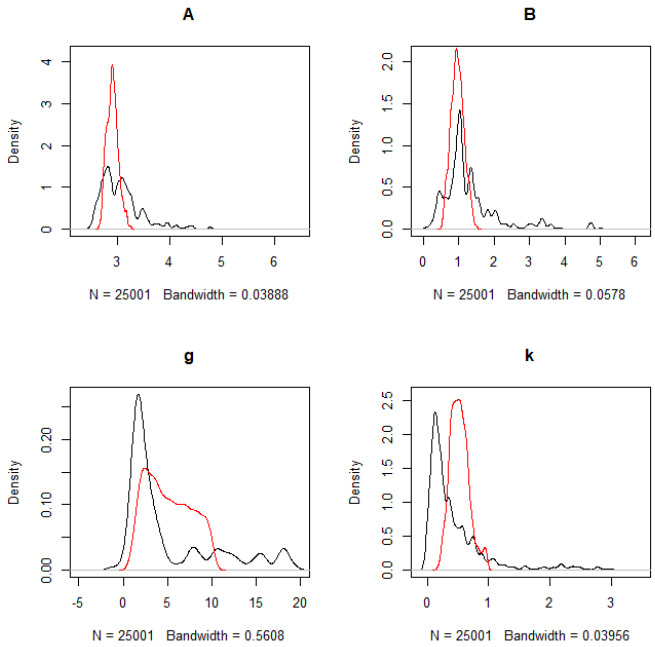
6.1 Example: univariate g-and-k distribution

We've introduced the *g-and-k distribution* since it's a classical example in the ABC literature, indeed it is a flexible shaped distribution and, for this reason, it is used to model non-standard data through a small number of parameters.

The approximate Bayesian Computation method based on sequential Monte Carlo is an alternative approach to estimate dynamical models' parameters.
In particular it can give us information about the inferability of parameters and regarding the model sensitivity but, more importantly to the objective of this

work, it has better performances with respect to other ABC approaches. Furthermore it was developed as a tool for model selection and, for this reason, it's able to choose the best model using the standard Bayesian model selection apparatus.

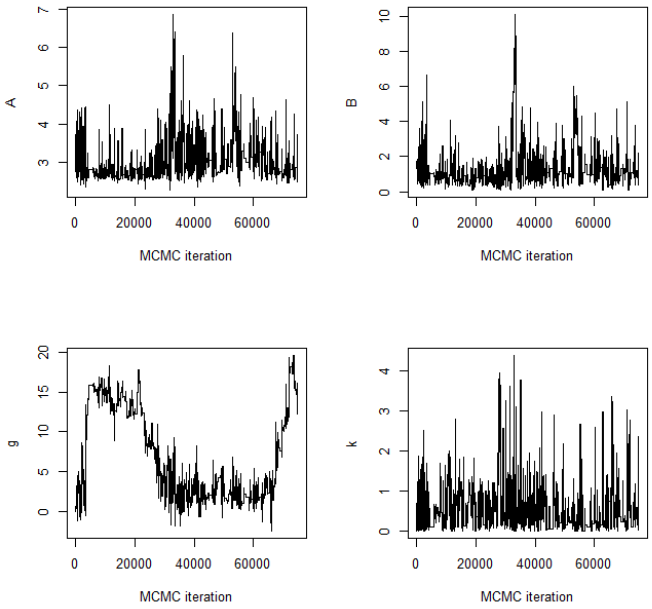
In order to check these statements, we ran the numerical experiment already seen in **4.2** by applying the sequential Monte Carlo approach to the ABC algorithm.



Posterior marginals obtained via MCMC in black, WABC approximations with a budget of $2.4 * 10^6$ model simulations in red

The obtained result suggests that, actually, the ABC SMC better estimates parameters of the model with respect to other ABC approaches. Note that the approximations of the parameter's densities have tighter ranges centered in their real values but, unfortunately, this doesn't hold for parameter g.

We think that the reason why g can't be approximated well is to search in the traceplots, indeed we can see a strange pattern in g's plots.


A
B
g
k
MCMC iteration
MCMC iteration
MCMC iteration
MCMC iteration
0
20000
40000
60000

Traceplots

References

[1] **S.A. Sisson, Y. Fan and M. Beaumont**, "Handbook of Approximate Bayesian Computation", *Chapman and Hall/CRC*, Chapter 1, pp. 3-44, (2018).

[2] **E. Bernton, P.E. Jacob, M. Gerber, C.P. Robert**, "Approximate Bayesian computation with the Wasserstein distance", *Journal of the Royal Statistical Society: Series B*, Volume 81, Issue 2, pp. 235-269, (2019).
arXiv:1905.03747

[3] **P. Del Moral, A. Doucet, A. Jasra**, "An adaptive sequential Monte Carlo method for approximate Bayesian computation", *Statistics and Computing*, 22 pp. 1009–1020, (2012).
https://link.springer.com/article/10.1007/s11222-011-9271-y

[4] **T. Toni, D. Welch, N. Strelkowa, A. Ipsen and M. Stumpf**, "Approximate Bayesian Computation scheme for parameter inference and model selection in dynamical systems", *Journal of the Royal Statistical Society*, Volume 6, Number 31, pp. 187-202, (2008).
https://doi.org/10.1098/rsif.2008.0172

[5] **P. Fearnhead and D. Prangle**, "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation", *Journal of the Royal Statistical Society: Series B*, Volume 74, Number 3, pp. 419-474, (2012).
https://doi.org/10.1111/j.1467-9868.2011.01010.x

[6] **S. Jackman**, "Bayesain Analysis for the Social Sciences", *WILEY SERIES IN PROBABILITY AND STATISTICS*, Chapter 5, (2009).
ISBN:978-0-470-01154-6

[7] **J. Alsing, B. D. Wandelt, S. M. Feeney**, "Optimal proposals for Approximate Bayesian Computation", *International Society for Bayesian Analysis*, Volume 00, Number 0, pp. 1-14, (2018).
arXiv:1808.06040v1

[8] **U. Simola**, **"Developments in Approximate Bayesian Computation and Statistical Applications in Astrostatistics**, (2018).
http://paduaresearch.cab.unipd.it/11267/