ABC model choice via mixture weight estimation

ABC model choice mediante stima del peso di mistura

Gianmarco Caruso, Luca Tardella, Christian P. Robert

Abstract Approximate Bayesian Computation (ABC) methods are widely employed to obtain approximations of posterior distributions without having to calculate likelihood functions. Nevertheless, the general impossibility to find statistics which are sufficient across models leads to unreliability of the classical tools for ABC model choice. To overcome this issue, a different kind of modelling is here proposed by replacing the traditional comparison between posterior probabilities of candidate models with posterior estimates of the weights of a mixture of these models. A simulation study highlights several strengths of this alternative approach, presenting it as a robust and flexible extension of the classical one.

Abstract *I metodi di ABC (*Approximate Bayesian Computation*) sono largamente utilizzati per ottenere approssimazioni di distribuzioni a posteriori senza dover calcolare funzioni di verosimiglianza. Tuttavia, in generale, non è possibile trovare statistiche che siano sufficienti tra modelli e ciò rende poco attendibili gli strumenti classici su cui si basa l'*ABC model choice*. Al fine di superare tale problema, si propone di rimpiazzare il tradizionale confronto tra probabilità a posteriori dei modelli candidati con la stima a posteriori dei pesi di una mistura di tali modelli. Uno studio di simulazione mette in luce diversi punti di forza di questo approccio alternativo, presentandolo come una robusta e flessibile estensione di quello classico.*

Key words: Approximate Bayesian Computation, Bayesian statistics, model choice, Bayes Factor, mixture models, intractable likelihoods, model misspecification

Gianmarco Caruso
Dipartimento di Scienze Statistiche, La Sapienza Università di Roma, Italy
e-mail: gianmarco.caruso@uniroma1.it

Luca Tardella
Dipartimento di Scienze Statistiche, La Sapienza Università di Roma, Italy
e-mail: luca.tardella@uniroma1.it

Christian P. Robert
CEREMADE, Université Paris Dauphine, PSL Research University, France; University of Warwick, UK; Università Ca' Foscari di Venezia, Italy
e-mail: xian@ceremade.dauphine.fr

1 Introduction

In the last decades, *Approximate Bayesian Computation* (ABC, henceforth) methods have become popular as a class of likelihood-free algorithms which aim to draw samples from an approximate posterior distribution, in the cases where the likelihood is unavailable or intractable, but it is still possible to generate data from the corresponding distribution.

Suppose to observe a sample $\mathbf{y} = (y_1, \ldots, y_n)$ of realizations from iid random variables Y_i, $i = 1, \ldots, n$, with a (*complex*) density $p(\cdot|\theta)$, where θ is an unknown parameter of interest with prior distribution $\pi(\theta)$. In order to sample from an approximate posterior distribution of θ, one can use a basic version of ABC rejection sampling algorithm (Pritchard et al. 1999): the idea is to draw iid parameter values $\theta_1^*, \ldots, \theta_N^*$ from $\pi(\theta)$, and to use each of these values to generate a synthetic sample of iid pseudo-observations, $\mathbf{z}$, from the sampling distribution $p(\cdot|\theta)$. If $\mathbf{z}$ is *similar* to the observed data $\mathbf{y}$, the corresponding θ^* is accepted as a value generated from the posterior distribution $\pi(\theta|\mathbf{y})$.

The concept of *similarity* between two datasets is expressed by three tools: a vector of statistics (or *summaries*), a distance and a tolerance threshold. The vector of summaries $\eta(\cdot) = \big(\eta_1(\cdot), \ldots, \eta_k(\cdot)\big)$ is used to summarise the information contained in a dataset, so that a pair of vectors of k statistics is compared instead of a pair of vectors of n observations (being $k \ll n$). This vector of statistics is most often not sufficient, but the consequent loss of information is tolerated with the idea to avoid the *curse of dimensionality* and to reduce the running time of the ABC algorithms. The distance $\rho(\cdot)$ quantifies how much $\eta(\mathbf{z})$ is close to $\eta(\mathbf{y})$. The threshold ε allows to accept all the θ^*'s which generate datasets whose associated vector of summaries is *close enough* to the observed vector of summaries. Therefore, the posterior sample is generated from $\pi\big(\theta\big|\rho(\eta(\mathbf{y}), \eta(\mathbf{z})) < \varepsilon\big)$ as surrogate of $\pi(\theta|\mathbf{y})$: the more informative the vector of statistics $\eta(\mathbf{y})$ and the smaller ε, the better the approximation.

1.1 *Classical ABC model choice and drawbacks*

When M models are compared, one considers the model index $\mathscr{M}$ as an additional unknown parameter, with prior distribution $\pi(\mathscr{M} = m)$, $m = 1, \ldots, M$ (Grelaud et al. 2009). The classical ABC model choice algorithm is summarised in **Algorithm 1**.

Algorithm 1 classical ABC model choice (ABC-mc)

```
for $i = 1$ to $N$ do
    repeat
        Generate $m^*$ from the prior $\pi(\mathscr{M} = m)$
        Generate $\theta^*_{m^*}$ from the prior $\pi_{m^*}(\cdot)$
        Generate $\mathbf{z}$ from the sampling distribution $p_{m^*}\left(\cdot\middle|\theta^*_{m^*}\right)$
    until $\rho\big(\eta(\mathbf{z}), \eta(\mathbf{y})\big) < \varepsilon$
    Set $m^{(i)} = m^*$ and $\theta_i = \theta^*$
end for
```

The posterior probability of model m can be estimated with the frequency of acceptances from model m, namely $\hat{\pi}_{\varepsilon}(\mathscr{M}=m|\mathbf{y})=\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(m^{(i)}=m)$, while $\hat{B}_{12,\varepsilon}(\mathbf{y})=\frac{\pi(\mathscr{M}=2)}{\pi(\mathscr{M}=1)}\frac{\sum_{i=1}^{N}\mathbb{I}(m^{(i)}=1)}{\sum_{i=1}^{N}\mathbb{I}(m^{(i)}=2)}$ can be used to approximate the Bayes Factor (BF),

$$B_{12}(\mathbf{y})=\int_{\Theta_1}p_1(\mathbf{y}|\theta_1)\pi_1(\theta_1)\,d\theta_1\bigg/\int_{\Theta_2}p_2(\mathbf{y}|\theta_2)\pi_2(\theta_2)\,d\theta_2\,, \tag{1}$$

where $p_m(\mathbf{y}|\theta)$ is the likelihood function associated with the m-th model, $m=1,2$. These approximations are valid as long as identical summaries, distance and tolerance threshold are used over both models.
Robert et al. (2011) thoroughly investigate the drawbacks of model choice in ABC. First of all, they show that the approximated BF, $\hat{B}_{12}(\mathbf{y})$, converges to the BF based on the vector of observed statistics, $\eta(\mathbf{y})$,

$$B_{12}^{\eta}(\mathbf{y})=\int\pi_1(\theta_1)p_1^{\eta}\big(\eta(\mathbf{y})|\theta_1\big)\,d\theta_1\bigg/\int\pi_2(\theta_2)p_2^{\eta}\big(\eta(\mathbf{y})|\theta_2\big)\,d\theta_2\,, \tag{2}$$

as ε goes to zero. This quantity is only based on the observed vector of statistics and, therefore, insufficient statistics yield a BF which converges to a quantity different from (1). Even in the favourable case where $\eta(\mathbf{y})$ is sufficient for both models, sufficiency *across* models can be hardly obtained and this leads to a discrepancy between (2) and (1) which cannot be computed. Insufficiency of the statistics for models or across models is thus the main source of unreliability of the ABC model choice based on the estimated BF.

2 ABC model choice via mixture weight estimation

The lack of confidence on the classical ABC model choice may be solved by introducing a different kind of modelling: the idea is to replace the inference on the posterior probabilities of the models with the posterior estimate of the weights of a mixture of the candidate models. This approach is an extension to the ABC realm of the inferential procedure proposed by Kamary et al. (2014). Here one considers the case of two candidate models (i.e. $M=2$) sharing an unknown parameter of interest θ which has a common meaning for both models.
One considers the data $\mathbf{y}$ as produced by a mixture of $\mathscr{M}_1$ and $\mathscr{M}_2$, namely

$$\mathscr{M}_w:\mathbf{y}\sim w\,p_1(\mathbf{y}|\theta)+(1-w)\,p_2(\mathbf{y}|\theta),\quad 0\le w\le 1\,, \tag{3}$$

with θ following a prior distribution $\pi(\theta)$ and associated to the sampling distribution $p_w(\cdot|\theta)$. The parameters w and θ will be considered independent *a priori*. Notice that this mixture model is an *encompassing* model since it contains both models as special cases: for $w=1$ it is equivalent to $\mathscr{M}_1$, while for $w=0$ it is equivalent to $\mathscr{M}_2$. The weight w represents the probability that an observation is sampled from p_1, so that it may be interpreted as the proportion of data which support $\mathscr{M}_1$.

A posterior inference on w may then offer interesting information about which one of the two models is the most suitable according to the observed data as well as the degree of support of one model against the other.
Algorithm 2 shows the way the ABC rejection sampling algorithm is applied on this mixture model in order to estimate θ and w.

Algorithm 2 ABC model choice via mixture weight estimation (ABC-mix)

```
for $i = 1$ to $N$ do
    repeat
        Generate $\theta^*$ and $w^*$ from their respective prior distributions
        Generate $\mathbf{z}$ from the sampling distribution $p_{w^*}(\cdot|\theta^*)$ of the model $\mathscr{M}_{w^*}$
    until $\rho\big(\eta(\mathbf{z}), \eta(\mathbf{y})\big) < \varepsilon$
    set $w_i = w^*$ and $\theta_i = \theta^*$
end for
```

This kind of modelling has several advantages. First of all, a whole posterior distribution on w allows, *inter alia*, for posterior point estimates, region of credibility, quantification of the uncertainty on the results and sensitivity analysis. On the other hand, the BF offers poor information since it is merely a scalar which suggests which model is more adequate and the relative degree of evidence of models. In addition, in the case where a same scalar value provided by the BF may seek either a misspecification of both models or a general acceptance of both - perhaps suggesting a more cautious approach via model averaging -, this new model choice approach may allow to conclude that both models or none could be acceptable, in the sense that a mixture of them may be a better choice. In particular, Kamary et al. (2014) show that this approach leads to a consistent testing procedure, not only when one of the two models is the true one, but also when neither are correct, since the posterior on w tends to concentrate around the value which minimizes the Kullback-Leibler divergence from the true distribution. A further attractive feature of this approach is the fact that it allows for improper priors on θ, where the BF totally prohibits this kind of assumption.
A standard prior for w is a symmetric $Beta(a_0, a_0)$, with $a_0 \in \mathbb{R}^+$ representing the degree of uncertainty *a priori* about the fact that one of the two candidate models is indeed the true one. A small a_0 may offer a regularization tool, being most of the density placed around the boundary values 0 and 1. For $a_0 \downarrow 0$, the proposed values from the prior distribution of w tends to be only 0's and 1's, so that the ABC-mix behaves like the ABC-mc with symmetric prior probabilities on the models. In this sense, the ABC-mc may be substantially seen as a limiting case of the ABC-mix.

3 Simulation study

One considers the case of comparison between two models, that is, the α-stable distribution (Borak, Härdle, and Weron 2005) and the skew-Normal distribution (Azzalini 2013), where the common unknown parameter is the location one. A reparametrisation of the location parameter of the skew-Normal distribution is considered, so that θ corresponds to the first moment of the two families of distributions. One considers the case where both models are wrong, since the data are iid simulations from a skew-*t* distribution (Azzalini 2013) with expected value equal to 0. The likelihood of the α-stable model is not available in closed form but it is possible to simulate from this model: the inference on the unknown parameter θ and the model choice are thus carried out through ABC-mc and ABC-mix. *Fig. 1* shows that the approximated posterior density of θ provided by ABC-mc is vague and tends to recover the corresponding prior, while ABC-mix produces posterior density estimates which put their masses around 0, in most of the cases.

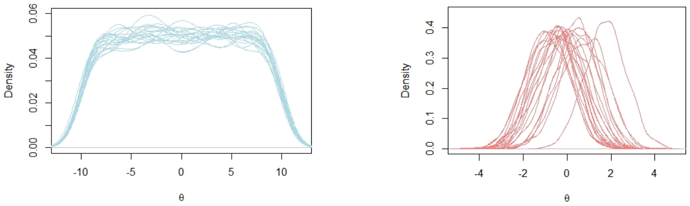

Fig. 1: Density estimate of the posterior distributions of θ computed via ABC-mc (left panel) *and ABC-mix* (right panel) *over* 20 *different datasets of* $n = 2000$ *iid samples from a skew-t with mean* $\theta = 0$*. Both algorithms are based on* 10^5 *simulations from* $\theta \sim Unif[-10,10]$*. The prior probabilities of the models are* 0.5 *(for ABC-mc) and* $w \sim Beta(0.5,0.5)$ *(for ABC-mix). The tolerance corresponds to an acceptance rate of* 0.01*. The distance used to measure the discrepancy between the observed and the j-th simulated vector of summaries* ($j = 1,\dots,N$) *is* $\rho(\eta(y),\eta(z^j)) = \sum_{i=1}^{k} \frac{|\eta_i(z^j)-\eta_i(y)|}{\mathrm{mad}_{j=1}^{N}\eta_i(z^j)}$ *, where the denominator is the median absolute deviation* (mad) *of the i-th summary statistic over all the N simulations: this avoids a distance dominated by the variable with the greatest magnitude. The two summaries (i.e.* $k = 2$*) are the* mad*, which is a good option for ABC model choice (Marin et al. 2014), and the median.*

The difficulty for ABC-mc to provide reasonable posterior density estimates for θ can be understood by analysing the asymptotic behaviour of the posterior probability of the first model (i.g. the α-stable). In fact, *Fig. 2* shows that the posterior probability of the first model seems to converge to 0, by always supporting the second model. On the other hand, the weight of the first component of the mixture of the models, w, does not concentrate near one of the boundary values as the sample size increases. In particular, as n increases, the posterior medians of w tends to concentrate around 0.2, regardless of the prior specifications.

4 Conclusions

ABC-mix has shown better performances than ABC-mc when both models are misspecified, which is the most likely situation in real applications. ABC-mix behaves as a flexible extension of ABC-mc, since a tuning of the hyperparameter a_0 of the prior of the weight w offers a further dimension to the prior specification of the model. This allows to specify the degree of prior uncertainty about the true model and this may be useful as regularization tool or for carrying out sensitivity analysis.

A whole posterior on a mixture weight offers richer information than the one provided by the posterior probability of a model, it allows for measure of uncertainty on the estimates and it may lead to propose a mixture of the candidate models as better choice. Finally, it may be a valid solution to overcome the problem of insufficiency of the summary statistics *for* models and - above all - *across* models which dramatically affects the classical ABC model choice based on the BF.

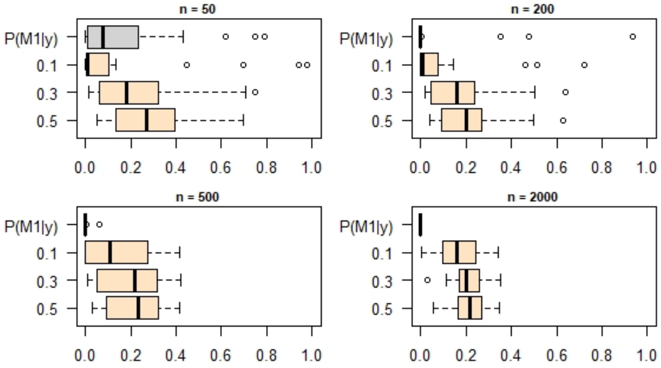

Fig. 2: Boxplot of the posterior medians of w for different values of a_0 (wheat) *- where* a_0 *is the hyperparameter of the prior* $Beta(a_0,a_0)$ *of w - and of the posterior probabilities* (grey) *of the* α*-stable model computed over* 20 *iid datasets from a skew-t with mean* $\theta = 0$ *for different sample sizes (*$n = 50,100,500,2000$*). The posterior of w and the posterior probability of the first model have been estimated via ABC-mix and ABC-mc (respectively), both based on* 10^5 *simulations and an acceptance rate set to* 0.01.

References

Azzalini, Adelchi (2013). *The skew-normal and related families*. Vol. 3. Cambridge University Press.

Borak, Szymon, Wolfgang Härdle, and Rafał Weron (2005). “Stable distributions”. In: *Statistical tools for finance and insurance*. Springer, pp. 21–44.

Grelaud, Aude et al. (2009). “ABC likelihood-free methods for model choice in Gibbs random fields”. In: *Bayesian Analysis* 4.2, pp. 317–335.

Kamary, Kaniav et al. (2014). “Testing hypotheses via a mixture estimation model”. In: *arXiv preprint arXiv:1412.2044*.

Marin, Jean-Michel et al. (2014). “Relevant statistics for Bayesian model choice”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.5, pp. 833–859.

Pritchard, Jonathan K et al. (1999). “Population growth of human Y chromosomes: a study of Y chromosome microsatellites.” In: *Molecular biology and evolution* 16.12, pp. 1791–1798.

Robert, Christian P et al. (2011). “Lack of confidence in approximate Bayesian computation model choice”. In: *Proceedings of the National Academy of Sciences* 108.37, pp. 15112–15117.