Evaluating statistic appropriateness for Bayesian model choice

Jean-Michel Marin, Natesh Pillai, Christian Robert, Judith Rousseau

Evaluating statistic appropriateness for Bayesian model choice

Jean-Michel Marin

I3M, UMR CNRS 5149, Université Montpellier 2. France.

Natesh S. Pillai

Department of Statistics, Harvard University, Cambridge, USA.

Christian P. Robert

Université Paris Dauphine, CEREMADE, IUF, and CREST, Paris, France.

Judith Rousseau

ENSAE and CREST, Paris, France.

Summary. The choice of the summary statistics in Bayesian inference and in particular in ABC is paramount to produce a valid outcome. We examine necessary and sufficient conditions on those statistics for a corresponding Bayes factor to be convergent. The conditions thus obtained are then usable in ABC settings to determine which summary statistics are appropriate, following a standard Monte Carlo validation.

1. Introduction

1.1. Summary statistics

In Robert et al. (2011), the authors showed that the now popular ABC (approximate Bayesian computation) method (Tavaré et al., 1997, Pritchard et al., 1999, Toni et al., 2009, Marin et al., 2011) is not necessarily validated when applied to Bayesian model choice problems. Without embarking upon a complete description of the ABC algorithm, since it is not relevant for our purpose here, we recall that one specific feature of this approximation method is to consider simulations $\boldsymbol{\theta}$ from the prior distribution and from the corresponding sampling distribution such that a statistic $\boldsymbol{T}(\mathbf{z})$ of the simulated pseudo-data $\mathbf{z}$ is close enough to the corresponding statistic $\boldsymbol{T}(\mathbf{y})$ for the observed data $\mathbf{y}$. The amount of proximity can be controlled by an increase in the computational power, however the choice of the statistic is paramount in that the resulting inference relies on this statistic and only on this statistic. In particular, when conducting ABC model choice, the ultimate outcome of the algorithm is the Bayes factor

$$B_{12}^{\boldsymbol{T}}(\mathbf{y}) = \frac{\int \pi_1(\boldsymbol{\theta}_1) g_1^{\boldsymbol{T}}(\boldsymbol{T}(\mathbf{y})|\boldsymbol{\theta}_1)\,\mathrm{d}\boldsymbol{\theta}_1}{\int \pi_2(\boldsymbol{\theta}_2) g_2^{\boldsymbol{T}}(\boldsymbol{T}(\mathbf{y})|\boldsymbol{\theta}_2)\,\mathrm{d}\boldsymbol{\theta}_2}\,,$$

which is exactly the Bayes factor for testing $\mathfrak{M}_1$ versus $\mathfrak{M}_2$ based on the sole observation of $\boldsymbol{T}(\mathbf{y})$. This value most often differs from the Bayes factor $B_{12}(\mathbf{y})$ based on the whole data $\mathbf{y}$. As discussed in Robert et al. (2011), in the specific case when the statistic $\boldsymbol{T}(\mathbf{y})$ is sufficient for both $\mathfrak{M}_1$ and $\mathfrak{M}_2$, the difference between both Bayes factors can be expressed as

$$B_{12}(\mathbf{y}) = \frac{h_1(\mathbf{y})}{h_2(\mathbf{y})}\, B^{\boldsymbol{T}}_{12}(\mathbf{y})\,, \tag{1}$$

where the ratio of the $g_i(\mathbf{y})$'s often behave like likelihoods of order n, the data size. This discrepancy implies that ABC model choice cannot always be trusted. Indeed, even in the limiting ideal case, i.e. when the ABC algorithm uses an infinite computing power to achieve a zero tolerance, the ABC odds ratio does not take into account the features of the data besides the value of $\boldsymbol{T}(\mathbf{y})$. Robert et al. (2011) illustrates that this difference can be such that $B^{\boldsymbol{T}}_{12}(\mathbf{y})$ leads to an inconsistent model choice.

The purpose of the current paper is to study asymptotic conditions on the statistic $\boldsymbol{T}$ under which the Bayes factor for testing $\mathfrak{M}_1$ versus $\mathfrak{M}_2$ based on the sole observation of $\boldsymbol{T}(\mathbf{y})$ either converges or diverges. The main result shows that a practical choice of summary statistics providing convergent model choice is available for ABC algorithms.

1.2. Insufficient statistics

We stress that the only case when the extra term in (1) is equal to one is when the statistic $\boldsymbol{T}$ is sufficient across models $\mathfrak{M}_1$ and $\mathfrak{M}_2$, i.e. for the collection $(m, \boldsymbol{\theta}_m)$ of the model index and of the parameter. This is for instance the case in Gibbs random fields (Grelaud et al., 2009). Otherwise, the conclusion drawn on $\boldsymbol{T}(\mathbf{y})$ necessarily differs from the conclusion drawn on $\mathbf{y}$.

Example 1. To illustrate the impact of the choice of a summary statistic on the Bayes factor, we consider the comparison of model $\mathfrak{M}_1$ $\mathbf{y} \sim \mathcal{N}(\theta_1, 1)$ with model $\mathfrak{M}_2$ $\mathbf{y} \sim \mathcal{L}(\theta_2, 1/\sqrt{2})$, the Laplace or double exponential distribution with mean θ_2 and scale parameter $1/\sqrt{2}$, which has a variance equal to one.

In this formal setting, four natural statistics can be considered (as suggested by one referee of Robert et al., 2011):

(a) the sample mean $\overline{\mathbf{y}}$;
(b) the sample median med($\mathbf{y}$);
(c) the sample variance var($\mathbf{y}$);
(d) the median absolute deviation mad($\mathbf{y}$) = med($\mathbf{y}$ − med($\mathbf{y}$));

Given the models under comparison, the first statistic is sufficient for both models, but not jointly across models, the second statistic is not sufficient but its distribution depends on θ_i, while both the sample variance and the median absolute deviation are ancillary statistics. As explained later (Section 2.3), the most

important feature of those statistics is that the first three statistics have the same expectation under both models (using the pseudo-true value of θ_i under both models) while the median absolute deviation has a different expectation under model 1 and model 2.

Since we are facing standard models in this artificial example, the computation of the true Bayes factor would be possible (see Appendix 1). However, if we base our inference only on one or several of the above statistics, the computation of the corresponding Bayes factors requires an ABC step. Fig. 1 shows the distribution of the posterior probability that the model is normal (as opposed to Laplace) when the data is either normal or Laplace and when the summary statistic in the ABC algorithm is the collection of the first three statistics above. The outcome is thus that the estimated posterior probability has roughly the same predictive distribution under both models, hence is not discriminative. Fig. 2 represents the same outcome when the summary statistic used in the ABC algorithm is only made of the median absolute deviation of the sample. In this second case, the two distributions of the estimated posterior probability are quite opposed under each model, concentrating near zero and one respectively. Hence, this summary statistic is highly discriminant to compare both models. From an ABC perspective, this means that using the median absolute deviation is then satisfactory, as opposed to the first three statistics.

◀

The above example illustrates very clearly the major result of this paper, namely that the mean behaviour of the summary statistic $\boldsymbol{T}(\mathbf{y})$ under both models under comparison is fundamental for the convergence of the Bayes factor, i.e. of the Bayesian model choice based on $\boldsymbol{T}(\mathbf{y})$. This result, described in the next section, thus brings an almost definitive answer to the question raised in Robert et al. (2011) about the validation of ABC model choice, although it may require additional simulation experiments in realistic situations.

The paper is organised as follows: Section 2 contains the theoretical derivation of the asymptotic behaviour of the Bayes factor based on a summary statistic, Section 2.1 covering our main assumptions, Section 2.2 exhibiting the asymptotic behaviour of the marginal likelihods, Section 2.3 detailing the consequences of this result for model choice based on summary statistics. Section 3 illustrates the relevance of our criterion for evaluating summary statistics, with Section 3.3 deriving from the above a criterion for calibrating the tolerance in the ABC algorithm. Section 4 concludes the paper with a short discussion.

2. Convergence of Bayes Factors using summary statistics

Let $\mathbf{y} = (y_1, \ldots, y_n)$ be the observed sample, not necessarily iid. We denote by $\mathbf{y} \sim \mathbb{P}^n$ the true distribution of the sample, and by $\boldsymbol{T}(\mathbf{y}) = \boldsymbol{T}^n = (T_1(\mathbf{y}), T_2(\mathbf{y}), \cdots, T_d(\mathbf{y}))$ a d-dimensional vector of summary statistics: $\boldsymbol{T}^n \sim$

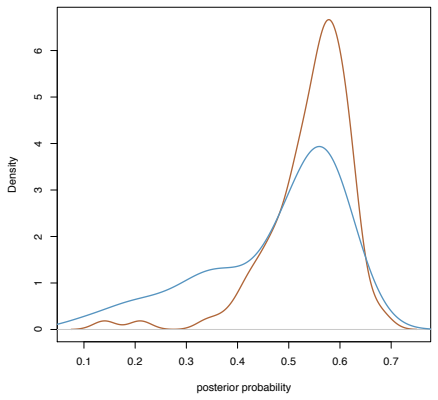

Fig. 1. Comparison of the distributions of the posterior probabilities that the data is from a normal model (rather than a Laplace model) when the data is made of 25 observations either from a normal *(brown)* or Laplace *(blue)* distribution with mean zero and when the summary statistic in the ABC algorithm is the made of the collection of the sample mean, median and variance. The ABC algorithm uses 10^5 proposals from the prior and selects the tolerance ϵ as the 1% distance quantile. The densities are estimated by a kernel estimator density() and rely on 100 replicas.

G_n by projection of $\mathbb{P}^n$. Throughout the paper the sign $\lesssim$ (resp. $\gtrsim$) means 'less than up to a multiplicative constant' and the sign $a_n \sim b_n$ for two sequences (a_n) and (b_n) means that $|a_n/b_n|$ is bounded from above and from below by a positive constant. Finally $a \wedge b$ denotes $\min(a, b)$.

In this setting, we assume that two competing models $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are under comparison in terms of providing best adequation to the sample X^n:

- under $\mathfrak{M}_1$, $\mathbf{y} \sim F_{1,n}(\cdot|\theta_1)$ where $\theta_1 \in \Theta_1$;
- under $\mathfrak{M}_2$, $\mathbf{y} \sim F_{2,n}(\cdot|\theta_2)$ where $\theta_2 \in \Theta_2$.

The corresponding distributions of $\boldsymbol{T}^n$ are denoted by $G_{1,n}(\cdot|\theta_1)$ and $G_{2,n}(\cdot|\theta_2)$, respectively, and the densities of $F_{i,n}(.|\theta_i)$ and of $G_{2,n}(\cdot|\theta_2)$ by $f_i(.|\theta_i)$ and $g_i(.|\theta_i)$, with respect to some dominating measures $\mu_{i,X}$ and $\mu_{i,T}$ $(i = 1, 2)$, respectively. Under the respective prior distributions π_1 and π_2 on θ_1 and θ_2, the posterior distributions given $\boldsymbol{T}^n$ are denoted by $\pi_1(\cdot|\boldsymbol{T}^n)$ and $\pi_2(\cdot|\boldsymbol{T}^n)$.

2.1. Distributional assumptions

Some technical assumptions on the model are necessary to establish our main result:

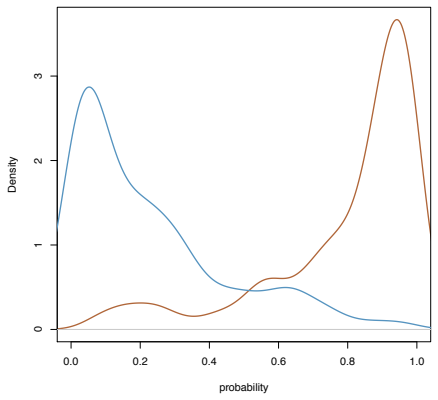

Fig. 2. Same figure as Fig. 1 when ABC is based on the median absolute deviation of the sample as the sole summary statistic.

([**A1**]) There exist a positive sequence $(v_n)_n$ converging to $+\infty$, a distribution Q absolutely continuous with respect to Lebesgue measure, a definite positive matrix V_0 and a vector μ_0 such that

$$v_n V_0^{-1/2}(\boldsymbol{T}^n - \mu_0) \to Q, \quad \text{as } n \to +\infty$$

([**A2**]) For $i \in \{1, 2\}$

$$v_n V_i(\theta_i)^{-1/2}(\boldsymbol{T}^n - \mu_i(\theta_i)) \to Q, \quad \text{as } n \to +\infty, \quad \text{under } P_{\theta_i}$$

where for each $\theta_i \in \Theta_i$, $\mu_i(\theta_i) \in \mathbb{R}^d$ and $V_i(\theta_i)$ is a positive definite matrix.

([**A3**]) The density q of Q is positive, continuous and bounded on $\mathbb{R}^d$.

([**A4**]) For each $i \in \{1, 2\}$, there exists $\mathcal{F}_{n,i}$, and $\epsilon_i, \tau_i, \alpha_i > 0$ such that for all $\tau > 0$,

$$\sup_{\theta_i \in \mathcal{F}_{n,i}} G_{i,n}\left[|\boldsymbol{T}^n - \mu(\theta_i)| > \tau |\mu_i(\theta_i) - \mu_0| \wedge \epsilon_i | \theta_i\right] \lesssim v_n^{-\alpha_i} \left(|\mu_i(\theta_i) - \mu_0| \wedge \epsilon_i\right)^{-\alpha_i} \quad (2)$$

with

$$\pi_i(\mathcal{F}_{n,i}^c) = o(v_n^{-\tau_i}). \quad (3)$$

([**A5**]) If $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} = 0$, define for $u > 0$

$$S_{n,i}(u) = \{\theta_i \in \mathcal{F}_{n,i}; |\mu(\theta_i) - \mu_0| \leq C v_n^{-1}\},$$

then there exists $d_i \leq \tau_i$ with $d_i < \alpha_i - 1$ such that

$$\pi_i(S_{n,i}(u)\}) \sim u^{d_i} v_n^{-d_i}, \quad \forall u \leq c v_n, \quad \text{for some } c > 0$$

([**A6**]) If $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} = 0$, there exists $C > 0$ such that

$$\sup_{|t-\mu_0|<Mv_n} \sup_{\theta_i \in S_{n,i}(C)} \left| |V_i(\theta_i)|^{1/2} v_n^{-d} g_i(t|\theta_i) - q(v_n V_i(\theta_i)^{-1/2}(t - \mu(\theta_i)) \right| = o(1)$$

and

$$\lim_{M\to+\infty} \limsup_n \frac{\pi_i(S_{n,i}(C) \cap \{\|V_i(\theta_i)^{-1}\| > M\}) + \pi_i(S_{n,i}(C) \cap \{\|V_i(\theta_i)\| > M\})}{\pi_i(S_{n,i}(C))} = 0$$

Conditions [**A1**]-[**A6**] are mild conditions. For instance when the summary statistics are empirical means or empirical quantiles, [**A1**]-[**A3**] are satisfied with $v_n = \sqrt{n}$ and the limiting distribution Q being the standard Gaussian distribution. Condition [**A4**] corresponds to a tail condition on the approximation of $\boldsymbol{T}^n$ by $\mu(\theta)$ in each model. For instance in the case where $\boldsymbol{T}^n$ is an empirical mean, i.e. $\boldsymbol{T}^n = n^{-1}\sum_{i=1}^n h(Y_i)$, where h is a given (possibly vector) function, then for each $\theta_i \in \Theta_i$,

$$G_{i,n}\left[\sqrt{n}|\boldsymbol{T}^n - \mu_i(\theta_i)| > u\right] \leq \frac{E\left[|\sum_{i=1}^n (h(Y_i) - \mu_i(\theta_i))|^q\right]}{u^q n^{q/2}} \leq C(\theta_i) u^{-q}, \qquad (4)$$

for potentially large values of q and under very general conditions (weaker than the independent and identically distributed case). The main difficulty in this condition comes from the fact that the term $O(u^q)$ in (4) must be uniform in the sieve sets $\mathcal{F}_{n,i}$. Set $\theta_i : |\mu_i(\theta_i) - \mu_0| \leq \epsilon$, for some positive ϵ assuming that $\mu_0 \in \{\mu_i(\theta_i), \theta_i \in \Theta_i\}$ and $u = \sqrt{n}|\mu_i(\theta_i) - \mu_0| \gtrsim 1$ (otherwise we bound the above probability by 1), then (4) implies that

$$G_{i,n}\left[|\boldsymbol{T}^n - \mu_i(\theta_i)| > |\mu_0 - \mu_i(\theta_i)|\right] \leq C(\theta_i) n^{-q/2} |\mu_0 - \mu_i(\theta_i)|^{-q} \leq (\sqrt{n}|\mu_0 - \mu_i(\theta_i)|)^{-\alpha}$$

as soon as $C(\theta_i)|\mu_0 - \mu_i(\theta_i)|^{-(q-\alpha)} \leq n^{(q-\alpha)/2}$ on $\mathcal{F}_{n,i}$. This is not problematic if the sets Θ_i are compact. If they are not compact, then this requires some tail conditions on the prior distributions π_i, see the Gaussian versus Laplace example detailed in Section 3.1.

Condition [**A5**] is a prior mass condition, as often encountered in asymptotic analysis of the posterior distribution, see for instance Ghosal and van der Vaart (2007). The exponents d_i can be viewed as effective dimensions of the parameter under the posteriors, since it corresponds to the minimal number of constraints on θ_i required to approximate μ_0 by $\mu_i(\theta_i)$.

If the relations $\theta_i \to \mu_i(\theta_i)$ can be locally inverted near μ_0, the sets $S_{n,i}(C)$ can bounded from above and below by sets in the form

$$|\theta_i - \theta_i^*| \leq c_i C v_n^{-1},$$

for some $\theta_i^* \in \Theta_i$ or by a finite collection of such sets, then if the prior density π_i is bounded from a bove and below near θ_i^*, $\pi_i(S_{n,i}(C)) \sim C^d v_n^{-d}$ and $d_i = d$.

In most cases $d_i \leq d$, since assuming that $d_i > d$ implies that the marginal prior density of $\mu(\theta)$ explodes at μ_0. The case $d_i < d$, corresponds to situations where the relation $\theta_i \to \mu_i(\theta_i)$ is not $1-1$.

Condition **[A6]** is a slightly stronger version of **[A2]**, since it is not only required that $v_n(\boldsymbol{T}^n - \mu_i)$ converges in distribution to Q but that, locally near the set of θ_i's such that $\mu_i(\theta_i) = \mu_0$, the density of $v_n(\boldsymbol{T}^n - \mu_i)$ is close to q (up to a rescaling factor). There are many instances of statistics that satisfy such an assumption. In particular empirical means of continuous variables, under moment and mixing assumptions, verify this condition uniformly over $\boldsymbol{T}^n$, see for instance Bhattacharya and Rao (1986). The (absolute) continuity of the observation is not necessary but nearly so, since the key condition to obtain uniform approximation of the densities is the so-called Cramer condition, see Bhattacharya and Rao (1986) for more details. Condition **[A6]** may become difficult to check when the sets $S_{n,i}(C)$ are not compact, typically when the sets $\mu_i^{-1}(\mu_0)$ are not compact. This essentially implies that the posterior distribution $\pi_i(.|\boldsymbol{T}^n)$ is not informative (at least no more than the prior) on the whole parameter θ_i but only on a fraction of it, sumarized by $\mu_i(\theta_i)$. In such a case, for condition **[A6]** to be nonetheless verified, it is important to have tail conditions on the prior so that $\mathcal{F}_{n,i}$ is not too large or to ensure that the distributions $G_{i,n}$ of $\boldsymbol{T}^n$ do not dependent on θ_i.

The last part of condition **[A6]** is trivially satisfied if the relation $\theta_i \to \mu_i(\theta_i)$ can be locally inverted so that the sets $S_{n,i}(C)$ can be bounded by balls in θ_i. If this is not the case, tail conditions on the prior will be enough to imply that the constraints $\|V_i(\theta_i)^{-1}\| > M$ or $\|V_i(\theta_i)^{-1}\| < M^{-1}$ can be neglected for M large enough.

2.2. Asymptotic behaviour of marginal likelihoods

The following result provides some control on the marginal likelihood. In Theorem 1, $m_1(\cdot)$ and $m_2(\cdot)$ denote the marginal densities of $\boldsymbol{T}^n$ under models $\mathfrak{M}_1$ and $\mathfrak{M}_2$, respectively, namely $(i = 1, 2)$

$$m_i(t) = \int_{\Theta_i} g_i(t|\theta_i)\, \pi_i(\theta_i)\, \mathrm{d}\theta_i\,.$$

Theorem 1. *Under assumptions* **[A1]**–**[A6]**, *for* $i = 1, 2$, *there exist constants* $C_l, C_u = O_{\mathbb{P}^n}(1)$ *such that if* $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} = 0$

$$C_l v_n^{d-d_i} \leq m_i(\boldsymbol{T}^n) \leq C_u v_n^{d-d_i} \tag{5}$$

and if

$$\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} > 0$$

$$m_i(\boldsymbol{T}^n) = o_{\mathbb{P}^n}[v_n^{d-\tau_i} + v_n^{d-\alpha_i}]. \tag{6}$$

The above theorem gives an equivalent to the marginal distribution $m_i(\boldsymbol{T}^n)$ when μ_0 can be attained by model $\mathfrak{M}_i$. Note that it does not require that G_n is in model $\mathfrak{M}_i$.

Interestingly a by-product of the proof (below) is that if $\mu_0 \in \{\mu_i(\theta_i); \theta_i \in \Theta_i\}$ then the posterior distribution of $\mu_i(\theta_i)$ given $\boldsymbol{T}^n$ is consistent at the rate $1/v_n$ as soon as $\alpha_i, \tau_i > d_i$. Indeed, with large probability

$$m_i(\boldsymbol{T}^n) \gtrsim v_n^{d-d_i}$$

and for all sequences $w_n \to +\infty$

$$\int_{S_{n,i}(w_n)^c} g_i(\boldsymbol{T}^n|\theta_i) d\pi_i(\theta_i) \lesssim w_n^{-\alpha_i} v_n^{d-\alpha_i} + v_n^{d-\tau_i} = o(v_n^{d-d_i}),$$

which implies that the posterior distribution verifies

$$\pi_i(|\mu_0 - \mu_i(\theta_i)| > w_n v_n^{-1} | \boldsymbol{T}^n) = \frac{\int_{S_{n,i}(w_n)^c} g_i(\boldsymbol{T}^n|\theta_i) d\pi_i(\theta_i)}{m_i(\boldsymbol{T}^n)} = o_{\mathbb{P}^n}(1).$$

Note also that d_i can be seen as an effective dimension of the model under the posterior $\pi_i(.|\boldsymbol{T}^n)$, since if $\mu_0 \in \{\mu_i(\theta_i); \theta_i \in \Theta_i\}$, $m_i(\boldsymbol{T}^n) \sim v_n^{d-d_i}$ and $g_n(\boldsymbol{T}^n) \sim v_n^d$. Thus $v_n^{-d_i}$ appears as the penlization coming from integrating out θ_i in model $\mathfrak{M}_i$.

We now prove Theorem 1.

Proof Recall that G_n is the true distribution of $\boldsymbol{T}^n$. Assume first that $\inf\{|\mu_0 - \mu_i(\theta_i)|; \theta_i \in \Theta_i\} = 0$ and consider $S_{n,i}$ defined in assumption **[A5]**. Let $\delta, C > 0$ and M_δ such that

$$Q(|V_0^{1/2} X| > M_\delta) = \delta,$$

note that M_δ goes to infinity as δ goes to 0. Set M_1 satisfying (see assumption **[A6]**)

$$\frac{\pi_i(S_{n,i}(C) \cap \{\|V_i(\theta_i)^{-1}\| < M_1^{-1}\}) + \pi_i(S_{n,i}(C) \cap \{\|V_i(\theta_i)^{-1}\| > M_1\})}{\pi_i(S_{n,i}(C))} \leq 1/2,$$

and $c_\delta = \inf\{q(x); |x| \leq (M_\delta + C)M_1\}$. We have

$$\begin{aligned} m_i(\boldsymbol{T}^n) &\geq \int_{S_{n,i}(C)} \mathbb{1}_{v_n |V_i(\theta_i)^{-1/2}(\boldsymbol{T}^n - \mu_i(\theta_i))| \leq (M_\delta + C)M_1} g_i(\boldsymbol{T}^n|\theta_i) d\pi_i(\theta_i) \\ &\geq \frac{c_\delta v_n^d}{2} \int_{S_{n,i}(C)} |V_i(\theta_i)|^{-1/2} \mathbb{1}[\{v_n |V_i(\theta_i)^{-1/2}(\boldsymbol{T}^n - \mu_i(\theta_i))| \leq (M_\delta + C)M_1\}] d\pi_i(\theta_i), \end{aligned}$$

where the last inequality comes from the fact that over the set of integration,

$$\begin{aligned} g_i(\boldsymbol{T}^n|\theta_i) &= |V_i(\theta_i)|^{-1/2} v_n^d [q(v_n V_i(\theta_i)^{-1/2}(t - \mu(\theta_i)) + o(1)] \\ &\geq \frac{|V_i(\theta_i)|^{-1/2} v_n^d \inf_{|x| \leq (C + M_\delta)M_1} q(x)}{2}. \end{aligned}$$

Set $\tilde{S}_{n,i} = S_{n,i}(C) \cap [\{|V_i(\theta_i)^{-1}| > M_1^{-1}\} \cup \{|V_i(\theta_i)^{-1}| < M_1\}]$,

$$m_i(\boldsymbol{T}^n) \quad \geq \quad \frac{c_\delta v_n^d M_1^{-1/2}}{2} \pi_i \left[\tilde{S}_{n,i} \cap \{v_n |V_i(\theta_i)^{-1/2} (\boldsymbol{T}^n - \mu_i(\theta_i))| > M_\delta M_1 + C\} \right].$$

Since $M_\delta > 2C$ for δ small enough, using a Markov inequality we obtain

$$\begin{aligned}
&G_n \left[\pi_i \left[\tilde{S}_{n,i} \cap \{v_n |V_i(\theta_i)^{-1/2} (\boldsymbol{T}^n - \mu_i(\theta_i))| > (M_\delta + C) M_1 \} \right] \geq \frac{\pi_i[\tilde{S}_{n,i}]}{2} \right] \\
&\leq 2 \frac{\int_{\tilde{S}_{n,i}} G_n \left[v_n |V_i(\theta_i)^{-1/2} (\boldsymbol{T}^n - \mu_i(\theta_i))| > (M_\delta + C) M_1 \right] d\pi_i(\theta_i)}{\pi_i(\tilde{S}_{n,i})} \\
&\leq 2 \frac{\int_{\tilde{S}_{n,i}} G_n \left[v_n |\boldsymbol{T}^n - \mu_0| > M_\delta \right] d\pi_i(\theta_i)}{\pi_i(\tilde{S}_{n,i})} \\
&\leq 2\delta.
\end{aligned}$$

Finally this leads to

$$m_i(\boldsymbol{T}^n) \quad \gtrsim \quad c_\delta v_n^d \pi_i \left[S_{n,i} \right] \gtrsim v_n^{d-d_i} \tag{7}$$

with probability greater than $1 - 2\delta$. We now bound from above $m_i(\boldsymbol{T}^n)$.

$$m_i(\boldsymbol{T}^n) \quad = \quad \int_{\mathcal{F}_{n,i}} g_i(\boldsymbol{T}^n|\theta_i) d\pi_i(\theta_i) + \int_{\mathcal{F}_{n,i}^c} g_i(\boldsymbol{T}^n|\theta_i) d\pi_i(\theta_i),$$

let $\delta > 0$ and M_δ defined similarly to before so that $G_n[|\boldsymbol{T}^n - \mu_0| > M_\delta v_n^{-1}] < 3\delta/2$, for n large enough. Using a Markov inequality together with assumption **[A6]** we obtain that, for all $\epsilon > 0$,

$$\begin{aligned}
&G_n \left[\int_{\mathcal{F}_{n,i}^c} g_i(\boldsymbol{T}^n|\theta_i) d\pi_i(\theta_i) > \epsilon v_n^d \pi_i(S_{n,i}) \right] \\
&\leq G_n[|\boldsymbol{T}^n - \mu_0| > M_\delta v_n^{-1}] \\
&+ \left(\sup_{|x| \leq M|V_0|} q(x) + \delta \right) \int_{\mathcal{F}_{n,i}^c} \frac{|V_0|^{-1/2} v_n^d}{\epsilon} \int g_i(t|\theta_i) dt d\pi_i(\theta_i) \\
&\lesssim \delta + \frac{v_n^d}{\epsilon} \pi(\mathcal{F}_{n,i}^c) \leq 2\delta,
\end{aligned} \tag{8}$$

when n is large enough. We then split $\mathcal{F}_{n,i}$ into the collection of $S_{n,i}((j+1)M_\delta) \cap S_{n,i}(jM_\delta)^c$, $j = 1, ..., J_n = J_0 v_n$, for some $J_0 > 0$ and $S_{n,i}(M_\delta J_n)^c$.

$$\begin{aligned}
\int_{\mathcal{F}_{n,i}} g_i(\boldsymbol{T}^n|\theta_i) d\pi_i(\theta_i) = &\sum_{j=0}^{J_n - 1} \int_{S_{n,i}((j+1)M_\delta) \cap S_{n,i}(jM_\delta)^c} g_i(\boldsymbol{T}^n|\theta_i) d\pi_i(\theta_i) \\
&+ \int_{\mathcal{F}_{n,i} \cap S_{n,i}^c(M_\delta J_n)} g_i(\boldsymbol{T}^n|\theta_i) d\pi_i(\theta_i).
\end{aligned} \tag{9}$$

If $j = 0$ and $K > d_i$, using the bound

$$\sup_{|t-\mu_0| \leq M_\delta v_n} g_n(t) \leq v_n^d |V_0|^{-1/2} [\sup_{x \in \mathbb{R}^d} q(x) + \delta] \lesssim v_n^d, \tag{10}$$

we have

$$\begin{aligned} G_n \left[\int_{S_{n,i}(M_\delta)} g_i(\boldsymbol{T}^n|\theta_i) d\pi_i(\theta_i) > M_\delta^K v_n^{d-d_i} \right] &\lesssim \frac{v_n^d}{M_\delta^K v_n^{d-d_i}} \pi_i(S_{n,i}(M_\delta)) + \delta \\ &= O(M_\delta^{d_i - K}) + \delta. \end{aligned}$$

which goes to zero as M_δ goes to infinity, i.e. if δ goes to 0. Using assumption **[A4]** and (10), we obtain

$$\begin{aligned} &G_n \left[\int_{S_{n,i}((j+1)M_\delta) \cap S_{n,i}(jM_\delta)^c} g_i(\boldsymbol{T}^n|\theta_i) d\pi_i(\theta_i) > M_\delta^K v_n^{d-d_i} \right] \\ &\lesssim \frac{v_n^{d_i}}{M_\delta^K} \int_{S_{n,i}((j+1)M_\delta) \cap S_{n,i}(jM_\delta)^c} G_{i,n} \left[|\boldsymbol{T}^n - \mu(\theta_i)| > (j - 1/2) M_\delta v_n^{-1} |\theta_i \right] d\pi_i(\theta_i) \\ &+ G_n \left[v_n |\boldsymbol{T}^n - \mu_0| > M_\delta/2 \right] \\ &\lesssim Q(|V_0^{1/2} X| > M_\delta/2) + M_\delta^{d_i - \alpha_i - K} j^{d_i - \alpha_i} \end{aligned}$$

and similarly

$$\begin{aligned} &G_n \left[\int_{S_{n,i}(J_n M_\delta)^c} g_i(\boldsymbol{T}^n|\theta_i) d\pi_i(\theta_i) > v_n^{d-d_i} \right] \\ &\lesssim v_n^{d_i} \int_{S_{n,i}(J_n M_\delta)^c} G_{i,n} \left[|\boldsymbol{T}^n - \mu(\theta_i)| > J_0/2 |\theta_i \right] d\pi_i(\theta_i) \\ &+ G_n \left[v_n |\boldsymbol{T}^n - \mu_0| > M_\delta \right] \\ &\lesssim 3\delta/2 + v_n^{d_i - \alpha_i} \leq 2\delta, \end{aligned} \tag{11}$$

for n large enough, under assumption **[A5]**. Combining the above inequalities with (9), we obtain for n large enough,

$$G_n \left[\int_{\mathcal{F}_{n,i}} g_i(\boldsymbol{T}^n|\theta_i) d\pi_i(\theta_i) > (2M_\delta^K + 1) v_n^{d-d_i} \right] \lesssim G_n \left[v_n |\boldsymbol{T}^n - \mu_0| > M_\delta/2 \right] + M_\delta^{d_i - K}$$

which can be made arbitrarily small by choosing δ small enough, which combined by (8) implies that

$$\int_{\Theta_i} g_i(\boldsymbol{T}^n|\theta_i) d\pi_i(\theta_i) = O_{\mathbb{P}^n}(v_n^{d-d_i}).$$

If $\inf\{|\mu_2(\theta_2) - \mu_0|; \theta_2 \in \Theta_2\} > 0$, there exists $j_0 > 0$ such that for all $j \leq j_0$ $S_{n,2}(j_0 v_n) = \emptyset$. This leads to, using the same computation as in (11), together with (8),

$$\begin{aligned} & G_n \left[\int_{\mathcal{F}_{n,2}} g_2(\boldsymbol{T}^n|\theta_2) d\pi_2(\theta_2) > \epsilon(v_n^{d-\tau_1} + v_n^{d-\alpha_1}) \right] \\ \lesssim \; & G_n \left[v_n |\boldsymbol{T}^n - \mu_0| > M_\delta \right] + \frac{v_n^{d_1}}{\epsilon} \int_{\mathcal{F}_{n,2}} G_{2,n} \left[|\boldsymbol{T}^n - \mu_2| > j_0 v_n/2 \right] d\pi_2(\theta_2) + 2\delta \\ \leq \; & 3\delta, \end{aligned}$$

for n large enough and all $\epsilon > 0$, and Theorem 1 is proved. □.

2.3. Consequences

Theorem 1 implies that, the asymptotic behaviour of the Bayes factor is driven by the asymptotic mean value of $\boldsymbol{T}^n$ under both models. To see this assume that the true distribution is in $\mathfrak{M}_1$ and consider first the case where $\inf\{|\mu_0 - \mu_2(\theta_2)|; \theta_2 \in \Theta_2\} = 0$ or vice-versa. Under assumptions **[A1]**-**[A6]**

$$C_l v_n^{-(d_1-d_2)} \leq \frac{m_1(\boldsymbol{T}^n)}{m_2(\boldsymbol{T}^n)} \leq C_u v_n^{-(d_1-d_2)},$$

where $C_l, C_u = O_{\mathbb{P}^n}(1)$, irrespective of the true model. The asymptotic behaviour of the Bayes factor depends solely on the difference $d_1 - d_2$ and for instance if $d_1 < d_2$ and G_n is in $\mathfrak{M}_1$, the Bayes factor goes to 0, instead of infinity. Note that the asymptotic behaviour remains the same even if G_n is in neither of the two models but if

$$\inf\{|\mu_0 - \mu_2(\theta_2)|; \theta_2 \in \Theta_2\} = \inf\{|\mu_0 - \mu_1(\theta_1)|; \theta_1 \in \Theta_1\} = 0$$

On the contrary if the true distribution is in model $\mathfrak{M}_1$ say and if $\inf\{|\mu_0 - \mu_2(\theta_2)|; \theta_2 \in \Theta_2\} > 0$, then the Bayes factor, under assumptions **[A1]**-**[A6]**, satisfies

$$\frac{m_1(\boldsymbol{T}^n)}{m_2(\boldsymbol{T}^n)} \geq C_u \min\left(v_n^{-(d_1-\alpha_2)}, v_n^{-(d_1-\tau_2)} \right),$$

and if $\min(\alpha_2, \tau_2) > d_1$,

$$\lim_{n \to +\infty} \frac{m_1(\boldsymbol{T}^n)}{m_2(\boldsymbol{T}^n)} = +\infty.$$

The conclusion of this discussion is summarized in the following result.

Theorem 2. *Under* $[\mathbf{A1}] - [\mathbf{A6}]$ *and if*

$$\{|\mu_0 - \mu_2(\theta_2)|; \theta_2 \in \Theta_2\} = \inf\{|\mu_0 - \mu_1(\theta_1)|; \theta_1 \in \Theta_1\} = 0,$$

the Bayes factor $B_{12}^{\boldsymbol{T}}$ has the same asymptotic behaviour as $v_n^{-(d_1-d_2)}$ irrespective of the true model. It is then consistent if and only if P_n is in the model having the smallest (strictly) dimension d_i.

If $\mathbb{P}^n$ belongs to one of the two models and if μ_0 cannot be attained by the other one :

$$\begin{aligned} 0 &= \min\left(\inf\{|\mu_0 - \mu_i(\theta_i)|; \theta_i \in \Theta_i\}, i = 1, 2\right) \\ &< \max\left(\inf\{|\mu_0 - \mu_i(\theta_i)|; \theta_i \in \Theta_i\}, i = 1, 2\right), \end{aligned}$$

then the Bayes factor $B_{12}^{\boldsymbol{T}}$ is consistent.

Interestingly, note that the Bayes factor is merely driven by the means $\mu_i(\theta_i)$ and the relative positive of μ_0 in both sets $\{\mu_i(\theta_i); \theta_i \in \Theta_i\}$, $i = 1, 2$. If G_n is in neither of the models but $\mu_0 \in \{\mu_1(\theta_1), \theta_1 \in \Theta_1\}$ but not in $\{\mu_2(\theta_2), \theta_2 \in \Theta_2\}$, then the Bayes factor will asymptotically favor $\mathfrak{M}_1$. If Q is the standard Gaussian distribution and if the convergence in distribution of $\sqrt{n}(\boldsymbol{T}^n - \mu_i(\theta_i))$ can be written in terms of Kullback-Leibler divergence between g_n and $g_i(.|\theta_i)$, i.e. if the Kullback-Leibler divergence between g_n and $g_i(.|\theta_i)$ is close the the Kullback-Leibler divergence between $|V_0|^{-1/2}q(\sqrt{n}V_0^{-1/2}(\boldsymbol{T}^n - \mu_0))$ and $|V_i(\theta_i)|^{-1/2}q(\sqrt{n}V_i(\theta_i)^{-1/2}(\boldsymbol{T}^n - \mu_i(\theta_i))$ then

$$\frac{1}{n}KL(g_0(\boldsymbol{T}^n), g_i(\boldsymbol{T}^n|\theta_i)) \approx \frac{(\mu_0 - \mu_i(\theta_i))^t V_i(\theta_i)^{-1}(\mu_0 - \mu_i(\theta_i))}{2} + o(1),$$

so that the difference between μ_0 and $\mu_i(\theta_i)$ is the key measure to evaluate the distance between g_n and $g_{i,n}(.|\theta_i)$.

Interestingly, the best statistics $\boldsymbol{T}^n$ to be used in an ABC - Bayes factor context are ancillary statistics which have different mean values under both models. Indeed if $\boldsymbol{T}^n$ depends asymptotically on some of the parameters of one of the models, say model $\mathfrak{M}_1$, then it is quite likely that there exists $\theta_1 \in \Theta_1$ such that $\mu_1(\theta_1) = \mu_0$ even though model $\mathfrak{M}_1$ is misspecified, specially if d the dimension of $\boldsymbol{T}^n$ is the same or smaller than the dimension of θ_1. To illustrate this remark consider the case where $d = 1$ and $\{\mu_1(\theta_1), \theta_1 \in \Theta_1\} = \mathbb{R}$ (or a large enough interval) then then $\boldsymbol{T}^n$ is not a satisfactory statistic for discriminating between models $\mathfrak{M}_1$ and $\mathfrak{M}_2$, when $\mathfrak{M}_2$ is true. Consider the example of the Laplace versus the Gaussian distribution with $\boldsymbol{T}^n = n^{-1}\sum_{i=1}^n X_i^4$, then assume that the true distribution is the Laplace with mean 0, so that $\mu_0 = 6$. Since under the Gaussian model $\mu(\theta) = 3 + \theta^4 + 6\theta^2$, the value $\theta^* = 2\sqrt{3} - 3$ leads to $\mu_0 = \mu(\theta^*)$ and a Bayes factor associated to such a statistic is not consistent (here $d_1 = d_2 = d = 1$).

However if $\boldsymbol{T}^n$ is ancillary (asymptotically), $\{\mu_1(\theta_1), \theta_1 \in \Theta_1\}$ is a singleton and it is sufficient that this singleton is different from μ_0. These remarks are illustrated in Section 3.

In the special case of $\mathfrak{M}_1$ being a submodel of $\mathfrak{M}_2$, and if the true distribution belongs to the smaller model $\mathfrak{M}_1$, any summary statistic satisfies

$\mu_0 \in \{\mu_1(\theta_1); \theta_1 \in \Theta_1\} \subset \{\mu_2(\theta_2); \theta_2 \in \Theta_2\}$, so that the Bayes factor is of order $v_n^{-(d_1-d_2)}$. If the summary statistic is informative merely on a parameter which is the same under both models, i.e if $d_1 = d_2$, then the Bayes factor is not consistent. Else, $d_1 < d_2$ and the Bayes factor is consistent under $\mathfrak{M}_1$. If the true distribution does not belong to $\mathfrak{M}_1$, then the same phenomenon as described above occures and the Bayes factor is consistent only if $\mu_1 \neq \mu_2 = \mu_0$.

3. Illustrations

3.1. Gaussian versus Laplace distributions

In this example $\theta_i \in \mathbb{R}$, both for $i = 1, 2$. We denote by $\mathfrak{M}_1$ the Gaussian model and by $\mathfrak{M}_2$ the Laplace model. In each model the prior on θ is a centered Gaussian distribution with variance 2, and in each case the data are simulated under $\theta_0 = 0$. We consider the following summary statistics :

- Fourth empirical moment : $T^n = n^{-1}\sum_{i=1}^n y_i^4$. In that case $\mu_1(\theta) = \theta^4 + 3 + 6\theta^2$, $\mu_2(\theta) = \theta^4 + 6 + 6\theta^2$ and $V_1(\theta)$ and $V_2(\theta)$ are polynomial functions in θ^2 with degree 3.
- Sixth empirical moment : $T^n = n^{-1}\sum_{i=1}^n y_i^4$ In that case $\mu_1(\theta) = \theta^6 + 15 + 45\theta^2 + 15\theta^4$, $\mu_2(\theta) = \theta^6 + 90 + 15\theta^4 + 90\theta^2$ and $V_1(\theta)$ and $V_2(\theta)$ are polynomial functions in θ^2 with degree 5.
- Sixth and fourth empirical moments : $T^n = n^{-1}\sum_{i=1}^n (y_i^4, y_i^6)$ The means and marginal variances are the same as before, and the determinant of the covariance matrix is a positive polynomial function in θ^2 with degree 8.

For each model set $\mathcal{F}_{n,1} = \mathcal{F}_{n,2} = \{|\theta| \leq C\sqrt{\log n}\}$, where $C > \sqrt{2}$ so that

$$\pi_1(\mathcal{F}_n^c) = \pi_2(\mathcal{F}_n^c) = o(n^{-C^2/4})$$

and condition **[A6]** is satisfied. Indeed, in model $\mathfrak{M}_1$, in the case of the fourth empirical moment, if $\mu_0 = 3$ (resp. 15 and $(3, 15)$ for the other summary statistics) and in model $\mathfrak{M}_2$ if $\mu_0 = 6$ (resp. 90 and $(6, 90)$), $S_{n,1}(C)$ and $S_{n,2}(C)$ can be bounded from above and below by balls in the form

$$|\theta| \leq cC^{1/2}n^{1/4},$$

so that $d_1 = d_2 = 1/2$ in those cases. Otherwise if $\mu_0 > 3$ (resp. > 15) in modem $\mathfrak{M}_1$ and $mu_0 > 6$ (resp. > 90) in model $\mathfrak{M}_2$, $S_{n,1}(C)$ and $S_{n,2}(C)$ can be bounded from above and below by balls in the form

$$|\theta^2 - \theta_*^2| \leq cCn^{-1/2}, \quad |\theta_*| > 0$$

so that $d_1 = d_2 = 0$ in those cases. For the bi-dimensional summary statistic, as soon as $\theta_0 \neq 0$ $S_{n,i}(C) \neq \emptyset$ for n large enough only if $\mathfrak{M}_i$ is the true model.

In our simulation study, we have considered $\theta_0 = 0$, so that $d_1 = d_2 = 1/2$ μ_0 is under model $\mathfrak{M}_2$ and $\inf\{|\mu_0 - \mu_2(\theta)|; \theta \in \mathbb{R}\} > 0$ if μ_0 is under model $\mathfrak{M}_1$.

Since Y^6 allows for any moment under both distributions, and since both distributions satisfy Cramer condition, $\boldsymbol{T}^n$ allows for an Edgeworth expansion under both models, which can be made uniform in sets in the form $\{|\theta| \leq Cn^{-1/4}\}$, see Bhattacharya and Rao (1986). Hence conditions **[A1]**-**[A3]** and **[A5]** are satisfied. Condition **[A4]** is verified using Tchebyshev inequalities (presented here in the case of the fourth empirical moment, but this remain valid for the other statistics) : Since for all $M > 0$ large enough $V_i(\theta) \leq M$ implies that $|\theta| > c_M$, we have if $|\theta| \leq M$

$$G_{i,n}\left[\left|n^{-1}\sum_{j=1}^{n}(y_j^4 - \mu_j(\theta))\right| > \tau|\mu_j(\theta) - \mu_0|\middle|\theta\right] \leq \frac{V_i(\theta)}{n\tau^2|\mu_j(\theta) - \mu_0|^2}$$
$$= O(n^{-1}|\mu_j(\theta) - \mu_0|^{-2}).$$

uniformly over $|\theta| \leq M$ and if $\theta > M$, then there exists $\epsilon_i > 0$ such that $|\mu_j(\theta) - \mu_0| > \epsilon_i$ and

$$G_{i,n}\left[\left|n^{-1}\sum_{j=1}^{n}(y_j^4 - \mu_j(\theta))\right| > \epsilon_i\middle|\theta\right] \leq \frac{V_i(\theta)}{n\epsilon_i^2} = O(n^{-1}(\log n)^6).$$

since in $\mathcal{F}_n$, $V_i(\theta) \leq C_i(\log n)^6$, and assumption **[A4]** is satisfied with $\alpha > 3/2$ so that **[A5]** is also satisfied.

Fig. 3. Same figure as Fig. 1 when ABC is based on the 4th empirical moment as the sole summary statistic.

Fig. 4. Same figure as Fig. 1 when ABC is based on the 4th and 6th empirical moments as summary statistic.

3.2. Quantile distributions

We consider the simulation from the four-parameter g-and-k distribution, defined through its quantile function

$$Q(p; A, B, g, k) = A + B\left(1 + \frac{1 - \exp(-gz(p))}{1 + \exp(-gz(p))}\right)\left(1 + z(p)^2\right)^k z(p)$$

where $z(p)$ is the pth standard normal quantile and the parameters A, B, g and k represent location, scale, skewness and kurtosis, respectively. The parameter

c measures the overall asymmetry and, following historical practice, is fixed at 0.8 Haynes et al. (1997). While the quantile function $F^-1(p;\theta)$ is well-defined, there is no closed-form expression for the corresponding density function, which makes the implementation of an MCMC algorithm quite delicate. We fix $A = 0$ and $B = 1$ and consider model $\mathfrak{M}_1$ such that $g = 0$ and $k \sim \mathcal{U}[-1/2, 5]$ versus model $\mathfrak{M}_2$ such that $g \sim \mathcal{U}[0, 4]$ and $k \sim \mathcal{U}[-1/2, 5]$. Model $\mathfrak{M}_1$ is a sub-model of model $\mathfrak{M}_2$. For such a case, we consider an ABC procedures which use 10^5 proposals from the prior and select the tolerance as the 1% quantile of the L_1 distances between some empirical quantiles. First, we use the empirical quantile of order 10% as summary statistics. Then, we use the empirical quantiles of order 10, 40, 60 and 90%. The results are presented in Figures 5 and 6. They are quite satisfactory when the fourth empirical quantiles are used.

Fig. 5. Comparison of the distributions of the posterior probabilities that the data is from model $\mathfrak{M}_1$ when the data is made of 100 observations either from model $\mathfrak{M}_1$ *(brown)* or $\mathfrak{M}_2$ *(blue)* distribution when the summary statistic in the ABC algorithm is the empirical quantile of order 10%. The densities are estimated by a kernel estimator density() and rely on 100 replicas.

Fig. 6. Same figure as Fig. 5 when ABC is based on the empirical quantiles of order 10, 40, 60 and 90% as set of summary statistics.

3.3. Calibration of the ABC tolerance

Let $\pi(M_1|\boldsymbol{T}^n)$ be the posterior probability of model M_1. In the ABC paradigm we approximate the previous quantity using

$$\hat{\pi}_\epsilon(M_1|\boldsymbol{T}^n) = \frac{1}{N_\epsilon}\sum_{i=1}^{N_\epsilon} \mathbb{I}_{m^{(i)}=M_1}$$

where $m^{(1)}, \ldots, m^{(N_\epsilon)}$ are iid sample sample from the $\pi_\epsilon(M_1|\boldsymbol{T}^n)$ the ABC approximation of $\pi(M_1|\boldsymbol{T}^n)$ (in the crude sense) and N_ϵ is the number of accepted ABC proposals using N sample from the prior. We have

$$\left(\pi(M_1|\boldsymbol{T}^n) - \hat{\pi}_\epsilon(M_1|\boldsymbol{T}^n)\right)^2 \leq \left(\pi(M_1|\boldsymbol{T}^n) - \pi_\epsilon(M_1|\boldsymbol{T}^n)\right)^2 + \left(\pi_\epsilon(M_1|\boldsymbol{T}^n) - \hat{\pi}_\epsilon(M_1|\boldsymbol{T}^n)\right)^2 .$$

We propose to select the value of ϵ that minimizes the sum of the two error terms above. The first one can be approximated by

$$\hat{\pi}_\epsilon(M_1|\boldsymbol{T}^n)\hat{\pi}_\epsilon(M_2|\boldsymbol{T}^n)\big/N_\epsilon\,.$$

and, the second one by

$$\left(\frac{|V_1|^{-1/2}\exp\left(-n(\boldsymbol{T}^n-\mu_1)'V_1^{-1}(\boldsymbol{T}^n-\mu_1)/2\right)}{|V_1|^{-1/2}\exp\left(-n(\boldsymbol{T}^n-\mu_1)'V_1^{-1}(\boldsymbol{T}^n-\mu_1)/2\right)+|V_2|^{-1/2}\exp\left(-n(\boldsymbol{T}^n-\mu_2)'V_2^{-1}(\boldsymbol{T}^n-\mu_2)/2\right)} - \hat{\pi}_\epsilon(M_1\right.$$

Example 2. *(Continuation of Example 1.)* In the comparison of the normal and the Laplace models, the asymptotic distributions for the empirical fourth and sixth moments can be derived:

```
mu1=c(3,15)
V1=matrix(c(96,900,900,10170),ncol=2)
V2=matrix(c(2484,112860,112860,7476300),ncol=2)
mu2=c(6,90)
```

We are thus able to use the optimal normal approximation to the true posterior, even though in practice we would have to use instead a Monte Carlo experiment under both models to approximate those normal distributions. Fig. 7 represents the distribution of the ABC tolerance ϵ in this setting when the data is either normal or Laplace, and when the summary statistics are the empirical fourth and sixth moments, separately (left and centre) or jointly (right). Both moments have different expectations under the two distributions, hence are acceptable candidates to run the test. The distributions are quite similar between the three choices. ◀

Fig. 7. Comparison of the distributions of the selected tolerances ϵ using the two-part error decomposition. The tolerance is expressed as a percentage of the maximal simulated distances under either the normal *(brown)* or the Laplace *(blue)* distribution. The data is made of 50 observations from a normal distribution with mean zero. The three graphs correspond to choices of summary statistics in the ABC algorithm equal to the fourth, the sixth, and both fourth and sixth moments. The normal approximation uses the true asymptotic means and variances of those statistics under both models. The R density() function is applied to 50 replications under each model.

4. Discussion

The fact that the means of the summary statistics under both models must differ for model choice to take place (in a convergent manner) is both natural, in that the asymptotic normality implies that only first moments matter, and fundamental, in that it drives the choice of summary statistics in practical ABC settings. Indeed, Theorem **??** implies that estimation statistics should not be used in ABC algorithms aiming at model comparison. This means that (a) different statistics should be used for estimation and for testing and (b) that they should not be mixed in a single summary statistic. Note that the distinction differs from the sufficient/ancillary opposition found in classical statistics (Cox and Hinkley, 1994) in that it is enough that the summary statistic T_n has a different asymptotic mean under both models. As shown in the normal-Laplace example, some ancillary statistics may not be appropriate for testing.

Acknowledgements

The authors are grateful to readers for helpful comments. Part of this work was done when Natesh Pillai was visiting Dauphine and CREST, he thanks these institutions for their hospitality. The first two authors and the last one are partly supported by the Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) through the 2009–2012 projects Bandhits and Emile. The second author is supported by the NSF grant 1107070.

Appendix 1

Laplace marginal likelihood

Consider a *sorted* sample $x_1, \ldots, x_n$ from the Laplace (double-exponential) $\mathcal{L}(\mu, 1/\sqrt{2})$ distribution

$$f(x|\mu) = \frac{1}{\sqrt{2}} \exp\{-\sqrt{2}|x-\mu|\}\,.$$

Under a normal $\mathcal{N}(0,\sigma^2)$ prior, the marginal likelihood is given by

$$\begin{aligned}
m_0(x_1,\ldots,x_n) &= \int 2^{-n/2} \prod_{i=1}^{n} \exp\{-\sqrt{2}|x_i-\mu|\} \exp\{-\mu^2/2\sigma^2\}\ \mathrm{d}\mu/\sqrt{2\pi}\sigma \\
&= 2^{-n/2} \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} \prod_{j=1}^{i} e^{\sqrt{2}x_j - \sqrt{2}\mu} \prod_{j=i+1}^{n} e^{-\sqrt{2}x_j + \sqrt{2}\mu} e^{-\mu^2/2\sigma^2}\ \mathrm{d}\mu/\sqrt{2\pi}\sigma \\
&= 2^{-n/2} \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} e^{\sqrt{2}\sum_{j=1}^{i} x_j - \sqrt{2}\sum_{j=i+1}^{n} x_j + \sqrt{2}(n-2i)\mu} e^{-\mu^2/2\sigma^2}\ \mathrm{d}\mu/\sqrt{2\pi}\sigma \\
&= 2^{-n/2} \sum_{i=0}^{n} e^{\sqrt{2}\sum_{j=1}^{i} x_j - \sqrt{2}\sum_{j=i+1}^{n} x_j + 2(n-2i)^2\sigma^2/2} \\
&\qquad \int_{x_i}^{x_{i+1}} e^{-\left\{\mu - \sqrt{2}(n-2i)\sigma^2\right\}/2\sigma^2}\ \mathrm{d}\mu/\sqrt{2\pi}\sigma \\
&= 2^{-n/2} \sum_{i=0}^{n} e^{\sqrt{2}\sum_{j=1}^{i} x_j - \sqrt{2}\sum_{j=i+1}^{n} x_j + 2(n-2i)^2\sigma^2/2} \\
&\qquad \left[\Phi(\{x_{i+1} - \sqrt{2}(n-2i)\sigma^2\}/\sigma) - \Phi(\{x_i - \sqrt{2}(n-2i)\sigma^2\}/\sigma)\right]
\end{aligned}$$

with usual conventions when $i = 0$ ($x_0 = -\infty$) and $i = n$ ($x_{n+1} = +\infty$).

References

Bhattacharya, R. N. and R. R. Rao (1986). *Normal Approximation and Asymptotic Expansions*. New-York: Wiley Series in Probability and Mathematical Statistics.

Cox, D. and D. Hinkley (1994). *Theoretical statistics*. Chapman & Hall.

Ghosal, S. and A. van der Vaart (2007). Convergence rates of posterior distributions for non iid observations. *Ann. Statist. 35*(1), 192–225.

Grelaud, A., J.-M. Marin, C. Robert, F. Rodolphe, and F. Tally (2009). Likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis 3(2)*, 427–442.

Haynes, M. A., H. L. MacGillivray, and K. L. Mengersen (1997). Robustness of ranking and selection rules using generalised g-and-k distributions. *J. Statist. Plann. Inference 65*(1), 45–66.

Kass, R. and A. Raftery (1995). Bayes factors. *J. American Statist. Assoc. 90*, 773–795.

Kleijn, B. and A. van der Vaart (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist. 34*(2), 837–877.

Marin, J., P. Pudlo, C. Robert, and R. Ryder (2011). Approximate Bayesian computational methods. *Statistics and Computing*. (To appear.).

Pritchard, J., M. Seielstad, A. Perez-Lezaun, and M. Feldman (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution 16*, 1791–1798.

Robert, C., J.-M. Marin, and N. Pillai (2011). Why approximate Bayesian computational methods cannot handle model choice problems. *Proc. Nat. Acad. Sci. USA*. (To appear.).

Tavaré, S., D. Balding, R. Griffith, and P. Donnelly (1997). Inferring coalescence times from DNA sequence data. *Genetics 145*, 505–518.

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

van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3. Cambridge: Cambridge University Press.