Relevant statistics 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 **?**, 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, in the sense that the resulting Bayes factors may fail to pick the correct model even asymptotically. The ABC algorithm is getting more and more accepted as a component of the Bayesian toolbox for handling intractable likehoods. Since ABC is not the central topic of this article, but rather both a motivation and an immediate application domain, we do not embark upon a complete description of its implementation, refering to Marin et al. (2011) and **?** for details. We simply recall here that the core feature of this approximation technique is to run simulations $(\boldsymbol{\theta}, \mathbf{z})$ from the prior distribution and the corresponding sampling distribution until a statistic $\boldsymbol{T}(\mathbf{z})$ of the simulated pseudo-data $\mathbf{z}$ is close enough to the corresponding value of the statistic $\boldsymbol{T}(\mathbf{y})$ at the observed data $\mathbf{y}$. The degree of proximity (also called the tolerance) can be improved by an increase in the computational power. However the choice of the statistic $\boldsymbol{T}$ is particularly crucial in that the resulting (approximately Bayesian) inference relies on this statistic and only on this statistic. It thus impacts the resulting inference much more than the choices of the tolerance distance and of the tolerance value.

When conducting ABC model choice (Grelaud et al., 2009, **?**), the outcome of the ideal algorithm associated with zero tolerance 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 unsurprisingly is 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 Didelot et al. (2011) and **?**, 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

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

where the ratio of the $g_i(\mathbf{y})$'s often behave like likelihoods of same order as the data size n. The discrepancy revealed by the above is such that ABC model choice cannot be trusted without further checks. 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 $\boldsymbol{T}(\mathbf{y})$. **?** warn that this difference can be such that $B_{12}^{\boldsymbol{T}}(\mathbf{y})$ leads to an inconsistent model choice. (The same is obviously true for point estimation, e.g. when considering the special case of an ancillary summary statistic $\boldsymbol{T}(\mathbf{y})$.)

Beyond ABC applications, note that many fields report summary statistics in their publications rather than the raw data, for various reasons ranging from confidentiality to storage, to proprietary issues. For instance, a dataset may be replaced by several p-values, $p_i(\mathbf{y})$, against several specific hypotheses. Handling a model choice problem based solely on $T(\mathbf{y}) = (p_1(\mathbf{y}), \ldots, p_k(\mathbf{y}))$ is therefore a relevant issue, with the coherence of the correponding Bayes factor at stake.

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. We obtain a precise characterisation of consistency in terms of the limiting distributions of the summary statistic $\boldsymbol{T}(\mathbf{y})$ under both models, namely that the true asymptotic mean of the summary statistic $\boldsymbol{T}(\mathbf{y})$ cannot be recovered under the wrong model, except for nested models. As explained in the paper, this characterisation implies that using point estimation statistics as summary statistics is rarely pertinent for testing. The main result shows that a practical choice of summary statistics providing convergent model choice is available for ABC algorithms. The practical side is computational in that the mean values of the summary statistics can be checked by simulation. Further properties of the vector of summary statistics can also be tested via these simulations, including the comparison of several summary statistics or, equivalently, the selection of the most discriminant components of the above vector.

1.2. Insufficient statistics

The above connection between the Bayes factor based on the whole data $\mathbf{y}$ and the Bayes factor based on the summary $\boldsymbol{T}(\mathbf{y})$ is only valid when the latter is sufficient for both models. In this setting, and only in this setting, the extra term in (1) is equal to one solely when the statistic $\boldsymbol{T}$ is furthermore 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. A rather special instance where this occurs is the case of Gibbs random fields (Grelaud et al., 2009). Otherwise, the conclusion drawn on $\boldsymbol{T}(\mathbf{y})$ necessarily differs from the conclusion drawn on $\mathbf{y}$. The same is obviously true outside the sufficient case, which implies that the selection of a summary statistic must be evaluated against its performances for model choice, because it is not garanteed per se. The following example illustrates this point:

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 **?**):

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

Given the models under comparison, the first statistic is sufficient only for the Gaussian model, the second statistic is not sufficient but its distribution depends on θ_i in both models, 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 appropriate values of the θ_i's 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 (even in the Laplace case, 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 ABC based on those summary statistics 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 for the comparison of the two models. From an ABC perspective, this means that using the median absolute deviation is then satisfactory, as opposed to the first three statistics.

◀

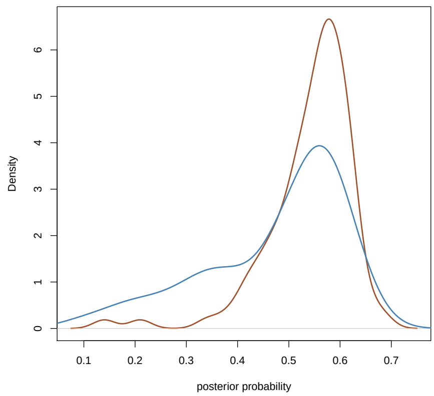

Fig. 1. Comparison of the distributions of the posterior probabilities that the data is from a normal model (rather than a Laplace model) when the data is made of 25 observations either from a normal *(brown)* or Laplace *(blue)* distribution with mean zero and when the summary statistic in the ABC algorithm 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.

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 **?** 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. Section 4 concludes the paper with a short discussion.

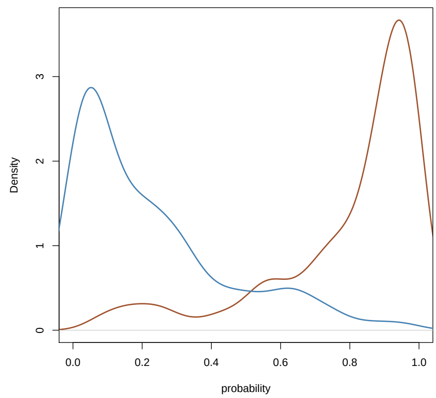

Fig. 2. Same figure as Fig. 1 when ABC is based on the median absolute deviation of the sample as the sole summary statistic.

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, $\boldsymbol{T}^n \sim G_n$. The distribution G_n is the projection of $\mathbb{P}^n$ under the map $\boldsymbol{T}^n : \mathbb{R}^n \mapsto \mathbb{R}^d$ and we denote its density by g_n.

There are two competing models $\mathfrak{M}_1$ and $\mathfrak{M}_2$ that we wish to compare:

– under $\mathfrak{M}_1$, $\mathbf{y} \sim F_{1,n}(\cdot|\theta_1)$ where $\theta_1 \in \Theta_1 \subset \mathbb{R}^{p_1}$

– under $\mathfrak{M}_2$, $\mathbf{y} \sim F_{2,n}(\cdot|\theta_2)$ where $\theta_2 \in \Theta_2 \subset \mathbb{R}^{p_2}$

The distributions of $\boldsymbol{T}^n$ under $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are denoted by $G_{1,n}(\cdot|\theta_1)$ and $G_{2,n}(\cdot|\theta_2)$, respectively. We also assume that the distribution functions $F_{i,n}(\cdot|\theta_i)$, $G_{i,n}(\cdot|\theta_i)$ have densities $f_i(\cdot|\theta_i)$ and $g_i(\cdot|\theta_i)$ 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. Assumptions

Before provinding the main result in the paper, let us state our theoretical assumptions on the models and the summary statistics under which the main result holds.

We start with a brief primer on our notations. The letter C denotes a positive constant, whose value may change from one occurrence to the next, but is independent of everything else. We write $a \wedge b$ to denote $\min(a, b)$. For two sequences $\{a_n\}, \{b_n\}$ of real numbers, $a_n \lesssim b_n$ (resp. $\gtrsim$) means $a_n \leq C b_n$ (resp. $a_n \geq C b_n$). Similarly, $a_n \sim b_n$ means that

$$1/C \leq \liminf_{n\to\infty} |a_n/b_n| \leq \limsup_{n\to\infty} |a_n/b_n| \leq C \ .$$

The symbol $\overset{n\to\infty}{\rightsquigarrow}$ denotes convergence in distribution.

The necessary assumptions are as follows:

([**A1**]) There exist a sequence of positive real numbers $\{v_n\}$ converging to $+\infty$, a distribution Q that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ with a positive, continuous and bounded version of the density function, $q(\cdot)$, a symmetric, $d \times d$ positive definite matrix V_0 and a vector $\mu_0 \in \mathbb{R}^d$, such that

$$v_n V_0^{-1/2}(\boldsymbol{T}^n - \mu_0) \overset{n\to\infty}{\rightsquigarrow} Q, \qquad \text{under } G_n \ ,$$

and for all $M > 0$

$$\sup_{v_n|t-\mu_0|<M} \left| |V_0|^{1/2} v_n^{-d} g_n(t) - q\{v_n V_0^{-1/2}(t - \mu_0)\} \right| = o(1) \ .$$

([**A2**]) For every $\theta_i \in \Theta_i$, $i \in \{1, 2\}$, there exist $d \times d$ symmetric positive definite matrices $V_i(\theta_i)$ and vectors $\mu_i(\theta_i) \in \mathbb{R}^d$ such that

$$v_n V_i(\theta_i)^{-1/2}(\boldsymbol{T}^n - \mu_i(\theta_i)) \overset{n\to\infty}{\rightsquigarrow} Q, \qquad \text{under } G_{i,n}(\cdot|\theta_i) \ .$$

([**A3**]) For every $i \in \{1, 2\}$, there exist sieves $\mathcal{F}_{n,i} \subset \Theta_i$ and constants $\epsilon_i, \tau_i, \alpha_i > 0$ such that

$$\pi_i(\mathcal{F}_{n,i}^c) = o(v_n^{-\tau_i}) \tag{2}$$

and, for all $\tau > 0$,

$$\sup_{\theta_i \in \mathcal{F}_{n,i}} \frac{G_{i,n}\Big[|\boldsymbol{T}^n - \mu(\theta_i)| > \tau|\mu_i(\theta_i) - \mu_0| \wedge \epsilon_i \,\Big|\theta_i\Big]}{(|\mu_i(\theta_i) - \mu_0| \wedge \epsilon_i)^{-\alpha_i}} \lesssim v_n^{-\alpha_i} \ . \tag{3}$$

([**A4**]) Define the sets $S_{n,i} \subset \mathcal{F}_{n,i}$ ($i \in \{1, 2\}$) as

$$S_{n,i}(u) = \big\{\theta_i \in \mathcal{F}_{n,i}; |\mu_i(\theta_i) - \mu_0| \leq u\, v_n^{-1}\big\}, \quad u > 0 \ .$$

If $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} = 0$, then there exists a constant

$$d_i < \tau_i \wedge (\alpha_i - 1)$$

such that

$$\pi_i(S_{n,i}(u)) \sim u^{d_i} v_n^{-d_i}, \quad \forall u \lesssim v_n \ , \tag{4}$$

where τ_i and α_i are defined in assumption [**A3**].

([**A5**]) If $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} = 0$, there exists $U > 0$ such that for any $M > 0$,

$$\sup_{\substack{v_n|t-\mu_0|<M \\ \theta_i \in S_{n,i}(U)}} \left| |V_i(\theta_i)|^{1/2} v_n^{-d} g_i(t|\theta_i) - q\{v_n V_i(\theta_i)^{-1/2}(t - \mu(\theta_i))\} \right| = o(1) \quad (5)$$

and

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

Here $\|V_i(\theta_i)\|$ and $|V_i(\theta_i)|$ denote the largest eigenvalue and the determinant of the matrix $V_i(\theta_i)$, respectively.

Even though these assumptions might look overwhelming, we claim that [**A1**]-[**A5**] are both mild and relatively easy to check in applications. Below we discuss briefly the implications of each of those and how to verify them. We will later (Section 3.1) illustrate why they hold in the Gaussian versus Laplace example.

Conditions [**A1**]-[**A2**] can be usually verified by means of the Central Limit theorem and are satisfied by many summary statistics. For instance, when the summary statistics are empirical means or empirical quantiles, conditions [**A1**]-[**A2**] are satisfied with $v_n = \sqrt{n}$ with the Gaussian distribution being the limiting Q (a most common occurence). Obviously, condition [**A1**] is redundant when the true distribution belongs to one of the two models under comparison.

Condition [**A3**] controls the large deviations of the estimator $\boldsymbol{T}^n$ from the estimand $\mu(\theta)$ under each model. For instance, when $\boldsymbol{T}^n$ is an empirical mean, i.e., $\boldsymbol{T}^n = n^{-1}\sum_{i=1}^n h(y_i)$ for a given function h, Markov's inequality implies that for every $\theta_i \in \Theta_i$,

$$G_{i,n}\left[\sqrt{n}|\boldsymbol{T}^n - \mu_i(\theta_i)| > u\right] \leq \frac{\mathbb{E}\Big[\ |\sum_{i=1}^n\{h(y_i) - \mu_i(\theta_i)\}|^p\ \Big]}{u^p\ n^{p/2}} \leq \kappa(\theta_i)u^{-p}, \quad (6)$$

for large values of p and under very weak assumptions (much weaker than the i.i.d case). The main difficulty in this condition comes from the fact that, for our arguments to operate, the factor $\kappa(\theta_i)$ in (6) must be controlled uniformly in θ. This is obviously much easier if the parameter space is compact. Otherwise, this control can still be achieved by choosing a power α_i that is smaller than p in the following way: consider θ_i's such that $|\mu_i(\theta_i) - \mu_0| \leq \epsilon$, 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 (6) implies that

$$\begin{aligned} G_{i,n}\left[|\boldsymbol{T}^n - \mu_i(\theta_i)| > |\mu_0 - \mu_i(\theta_i)|\right] &\leq \kappa(\theta_i)n^{-p/2}|\mu_0 - \mu_i(\theta_i)|^{-p} \\ &\leq (\sqrt{n}|\mu_0 - \mu_i(\theta_i)|)^{-\alpha}\ , \end{aligned}$$

provided $\kappa(\theta_i)|\mu_0 - \mu_i(\theta_i)|^{-(p-\alpha)} \leq n^{(p-\alpha)/2}$ on $\mathcal{F}_{n,i}$. Furthermore, if Θ_i is *not* compact, we usually have

$$\sup_{\theta \in \Theta_i} \kappa(\theta_i) = \infty \,.$$

In such situations we use the sieves $\mathcal{F}_{n,i}$ (which typically are compact subsets of Θ_i) to recover uniform bounds on the constant $\kappa(\theta_i)$ in (6). On the complement of the sieves $\mathcal{F}_{n,i}$, we need the additional assumption that the tail probability of the prior distribution decays sufficiently quickly for large n. This argument is illustrated in the Gaussian versus Laplace example detailed in Section 3.1.

Condition [**A4**] is a condition on the prior distribution under either model, as often encountered in asymptotic analyses of the posterior distribution, see for instance Ghosal and van der Vaart (2007). Usually referred to as *the prior mass condition*, it corresponds to the fact that if the prior vanishes in regions where the likelihood is not too small (i.e., near μ_0 in our case) then the marginal becomes very small. The exponents d_i can be viewed as effective dimensions of the parameter under the posterior distributions, as discussed after Corollary 1. If the maps $\theta_i \mapsto \mu_i(\theta_i)$ are locally invertible near μ_0, under the usual continuity conditions on the maps $\theta_i \mapsto |\mu_0 - \mu_i(\theta_i)|$, for any $u > 0$, there exists a finite collection of points $\theta_{ij}^* \in \Theta_i$ such that the sets $S_{n,i}(u)$ can bounded both from above and below by sets of the form

$$\bigcup_{j=1}^{J} \{\theta_j : |\theta_j - \theta_{ij}^*| \lesssim u v_n^{-1}\}, \quad J \in \mathbb{N} \,. \tag{7}$$

Thus if the prior density π_i is bounded from above and below near the points θ_{ij}^*, we immediately deduce that $\pi_i\{S_{n,i}(u)\} \sim u^d v_n^{-d}$ and $d_i = d$ verifying [**A4**]. In most cases we will have $d_i \leq d$, since assuming that $d_i > d$ implies that the prior density of $\mu(\theta)$ explodes at μ_0.

Condition [**A5**] is a slightly stronger version of [**A2**], since it not only requires that $v_n(\boldsymbol{T}^n - \mu_i)$ converges in distribution to Q but also that, 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 examples of summary statistics that satisfy this 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 observations $\mathbf{y}$ that we require is not necessary but it is *nearly* so, since the key criteria to obtain uniform approximation of the densities is the so-called Cramer condition, see Bhattacharya and Rao (1986) for details. Condition [**A5**] may become difficult to check when the sets $S_{n,i}(u)$ are not compact, which is typically the case when the sets $\{\theta_i; \mu_i(\theta_i) = \mu_0\}$ are not compact. The important point to note here is that, in such cases, the posterior distribution $\pi_i(\cdot|\boldsymbol{T}^n)$ is not informative (at least no more than the prior) on the whole parameter θ_i but only on a fraction of it, summarized by $\mu_i(\theta_i)$. In such a case, for condition **[A5]** to be nonetheless verified, it is important to impose tail conditions on the prior so that the sieves $\mathcal{F}_{n,i}$ are not too large or to ensure that the distributions $G_{i,n}$ of $\boldsymbol{T}^n$ do not depend on θ_i.

The last part of condition **[A5]** is trivially satisfied if the map $\theta_i \mapsto \mu_i(\theta_i)$ can be inverted as described above so that the sets $S_{n,i}(C)$ can be bounded (from above and below) by balls in θ_i as mentioned in (7) and if $\theta_i \to V_i(\theta_i)^{-1}$ is continuous or at least bounded on compact sets. If the map $\theta_i \mapsto \mu_i(\theta_i)$ is not invertible, tail conditions on the prior will typically 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. (See the Gaussian versus Laplace example in Section 3.1 below for the illustration of this point).

2.2. *Asymptotic behaviour of marginal likelihoods*

The following result provides some control on the marginal likelihoods. In Lemma 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 $(i = 1, 2)$

$$m_i(t) = \int_{\Theta_i} g_i(t|\theta_i)\, \pi_i(\theta_i)\, \mathrm{d}\theta_i\,. \tag{8}$$

LEMMA 1. *Under assumptions* **[A1]**–**[A5]**, *for $i = 1, 2$, there exist constants $C_l, C_u = O_{\mathbb{P}^n}(1)$ such that if* $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} = 0$

$$C_l v_n^{d-d_i} \le m_i(\boldsymbol{T}^n) \le C_u v_n^{d-d_i} \tag{9}$$

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

$$m_i(\boldsymbol{T}^n) = o_{\mathbb{P}^n}[v_n^{d-\tau_i} + v_n^{d-\alpha_i}]. \tag{10}$$

The above lemma, or more precisely (9), gives an equivalent to the marginal distribution $m_i(\boldsymbol{T}^n)$ when $\mu_0 \in \{\mu_i(\theta_i), \theta_i \in \Theta_i\}$ but it does not specifically require that G_n is in model $\mathfrak{M}_i$. See Appendix 2 for the proof of Lemma 1. The following result is a corollary on the use of $\boldsymbol{T}^n$ for estimation purposes beyond model choice:

COROLLARY 1. *Under the assumptions of Lemma 1, if $\mu_0 \in \{\mu_i(\theta_i); \theta_i \in \Theta_i\}$, the posterior distribution of $\mu_i(\theta_i)$ given $\boldsymbol{T}^n$ is consistent at the rate $1/v_n$ provided $\alpha_i, \tau_i > d_i$.*

PROOF. Indeed Equation (9) of Lemma 1 yields that

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

with large probability. For all sequences $\{w_n\}$ converging to $+\infty$, calculations performed in the proof of Lemma 1 (see Appendix 2) yield that

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

Therefore the posterior distribution of $\mu_i(\theta_i)$ has its tail probability given by

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

and the corollary follows.

Note again that d_i can be seen as an effective dimension of the model under the posterior $\pi_i(\cdot|\boldsymbol{T}^n)$, since if $\mu_0 \in \{\mu_i(\theta_i); \theta_i \in \Theta_i\}$,

$$m_i(\boldsymbol{T}^n) \sim v_n^{d-d_i} \qquad \text{and} \qquad g_n(\boldsymbol{T}^n) \sim v_n^d\ .$$

Thus $v_n^{-d_i}$ appears as the penalization coming from integrating θ_i out in model $\mathfrak{M}_i$, in the same spirit as the effective number of parameters used in DIC (Spiegelhalter et al., 2002) or as discussed in Rousseau (2007) or in **?**.

2.3. Consequences of the main result

Lemma 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]**-**[A5]**

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

where $C_l, C_u = O_{\mathbb{P}^n}(1)$, irrespective of the true model. Thus the asymptotic behaviour of the Bayes factor depends solely on the difference $d_1 - d_2$. For instance if $d_1 < d_2$ (as in the embedded case) 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 when G_n is in neither model provided

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

On the opposite 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]**-**[A5]**, satisfies

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

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 the above discussion is summarized by the following result:

THEOREM 1. *Under assumptions* $[\mathbf{A1}] - [\mathbf{A5}]$, *if*

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

then the Bayes factor $B_{12}^{\boldsymbol{T}}$ *has the same asymptotic behaviour as* $v_n^{-(d_1-d_2)}$ *irrespective of the true model. Therefore, it always asymptotically selects the model having the smallest effective dimension* d_i.

If the true distribution G_n *belongs to model* $\mathfrak{M}_\omega$ *and if* μ_0 *cannot be represented in the other model* $\mathfrak{M}_{3-\omega}$,

$$0 = \inf\{|\mu_0 - \mu_\omega(\theta_\omega)|; \theta_\omega \in \Theta_\omega\} < \inf\{|\mu_0 - \mu_{3-\omega}(\theta_{3-\omega})|; \theta_{3-\omega} \in \Theta_{3-\omega}\}$$

and if if $\min(\alpha_{3-\omega}, \tau_{3-\omega}) > d_\omega$, *then the Bayes factor* $B_{12}^{\boldsymbol{T}}$ *is consistent.*

An important practical consequence of Theorem 2 is that the Bayes factor is merely driven by the means $\mu_i(\theta_i)$ and the relative position of μ_0 in both sets $\{\mu_i(\theta_i); \theta_i \in \Theta_i\}$, $i = 1, 2$. If G_n is in neither model but μ_0 belongs to $\{\mu_1(\theta_1), \theta_1 \in \Theta_1\}$ but not to $\{\mu_2(\theta_2), \theta_2 \in \Theta_2\}$, then the Bayes factor will asymptotically favor $\mathfrak{M}_1$.

Suppose the summary statistics (appropriately rescaled) are asymptotically normal (thus Q is the standard Gaussian distribution) and assume that 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 $g_i(\cdot|\theta_i)$. That is, assume the Kullback-Leibler divergence between g_n and $g_i(\cdot|\theta_i)$ is close to the Kullback-Leibler divergence between $\sqrt{n}|V_0|^{-1/2}q(\sqrt{n}V_0^{-1/2}(\boldsymbol{T}^n - \mu_0))$ and $\sqrt{n}|V_i(\theta_i)|^{-1/2}q(\sqrt{n}V_i(\theta_i)^{-1/2}(\boldsymbol{T}^n - \mu_i(\theta_i))$. 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 $g_{i,n}(\cdot|\theta_i)$.

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 $\theta_2 \in \Theta_2$

such that $\mu_2(\theta_2) = \mu_0$ even though model $\mathfrak{M}_2$ is misspecified, specially if d the dimension of $\boldsymbol{T}^n$ is the same or smaller than the dimension of θ_2. 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 $\boldsymbol{T}^n$ 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 $\boldsymbol{T}^n = n^{-1}\sum_{i=1}^n X_i^4$, and assume that the true distribution is the Laplace with mean 1, so that $\mu_0 = 13$. Since under the Gaussian model $\mu(\theta) = \theta^4 + 3 + 6\theta^2$, the value $\theta^* = \sqrt{\sqrt{19} - 3}$ leads to $\mu_0 = \mu(\theta^*)$ and a Bayes factor associated to such a statistic is not consistent (here $d_1 = d_2 = 1$).

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 occurs 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, $\theta_i \in \mathbb{R}$, 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_i is a centered Gaussian distribution with variance 2, and in each case the data are simulated under $\theta_i = 0$. We consider the following summary statistics :

- Fourth empirical moment : $\boldsymbol{T}^n = n^{-1}\sum_{i=1}^n y_i^4$. In that case $\mu_1(\theta) = \theta^4 + 3 + 6\theta^2$, $\mu_2(\theta) = \theta^4 + 6 + 6\theta^2$, while $V_1(\theta)$ and $V_2(\theta)$ are polynomial functions in θ^2 with degree 3.

- Sixth empirical moment : $\boldsymbol{T}^n = n^{-1}\sum_{i=1}^n y_i^6$. In that case $\mu_1(\theta) = \theta^6 + 15 + 45\theta^2 + 15\theta^4$, $\mu_2(\theta) = \theta^6 + 90 + 15\theta^4 + 90\theta^2$, while $V_1(\theta)$ and $V_2(\theta)$ are polynomial functions in θ^2 with degree 5.

- Sixth and fourth empirical moments : $\boldsymbol{T}^n = n^{-1}\sum_{i=1}^n (y_i^4, y_i^6)$. 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.

We now endeavour to check that assumptions **[A1]**–**[A5]** hold for those statistics. Given that they are empirical moments, condition **[A2]** is trivially satisfied as a consequence of the Central Limit theorem, with $v_n = \sqrt{n}$ and $\mu_i(\theta_i)$, $V_i(\theta_i)$ defined above. Condition **[A1]** is redundant with **[A2]** in that we only consider the cases where one of the two models is the true model.

For both models, we set $\mathcal{F}_{n,1} = \mathcal{F}_{n,2} = \{|\theta| \leq u\sqrt{\log n}\} = \mathcal{F}_n$ for condition **[A3]**, where $u > \sqrt{2}$ so that

$$\pi_1(\mathcal{F}_{n,1}^c) = \pi_2(\mathcal{F}_{n,2}^c) = o(n^{-u^2/4})$$

which implies $\tau_i = u^2/2$. The second part of condition **[A3]** is verified using Markov inequalities. First, for $M > 0$ large enough, there exists c_M such that $(\theta) \leq M$ implies that $|\theta| < c_M$. For instance, in the case of the fourth empirical moment, if $|\theta| \leq c_M$

$$\begin{aligned} G_{i,n}\left[\left|n^{-1}\sum_{j=1}^{n}(y_j^4 - \mu_i(\theta))\right| > \tau|\mu_i(\theta) - \mu_0| \,\middle|\, \theta\right] &\leq \frac{\mathbb{E}_i[(Y^4 - \mu_i(\theta))^4|\theta]}{n^2\tau^4|\mu_i(\theta) - \mu_0|^4} \\ &\leq O(n^{-2}|\mu_i(\theta) - \mu_0|^{-4}) \end{aligned}$$

uniformly. On the other hand, if $|\theta| > c_M$, then there exists $\epsilon_i > 0$ such that $|\mu_i(\theta) - \mu_0| > \epsilon_i$ and

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

since, in $\mathcal{F}_n$, $\mathbb{E}_i[(Y^4 - \mu_i(\theta))^4|\theta] \leq C_i(\log n)^6$. Thus, assumption **[A3]** is satisfied for any $\alpha_i < 4$.

Concerning **[A4]**, 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 of 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 model $\mathfrak{M}_1$ 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 $d_1 = d_2 = 1$ 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 if the true distribution belongs to model $\mathfrak{M}_2$ (Laplace) $\mu_0 \in \{\mu_i(\theta); \theta \in \mathbb{R}\}$ for both $i = 1, 2$

and we have $d_1 = 1$ and $d_2 = 1/2$. On the other hand if the true distribution belongs to model $\mathfrak{M}_1$ (Gaussian) then $d_1 = 1/2$ and $\inf\{|\mu_0 - \mu_2(\theta_2)|; \theta_2 \in \mathbb{R}\} > 0$. Following from Theorem 2, The Bayes factor is consistent in both cases but at the rate $n^{-1/4}$ under model $\mathfrak{M}_2$ and to some extent accidentally (it is merely due to the fact that in that case $d_1 > d_2$). If $\theta_0 \neq 0$ but is small then a similar argument leads to non consistency of the Bayes factor under model $\mathfrak{M}_2$ since then $d_1 = d_2 = 0$ and $\mu_0 \in \{\mu_i(\theta); \theta \in \mathbb{R}\}$, for both $i = 1, 2$.

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 $\{|\theta_i| \leq Cn^{-1/4}\}$, see Bhattacharya and Rao (1986). Hence condition **[A5]** is satisfied.

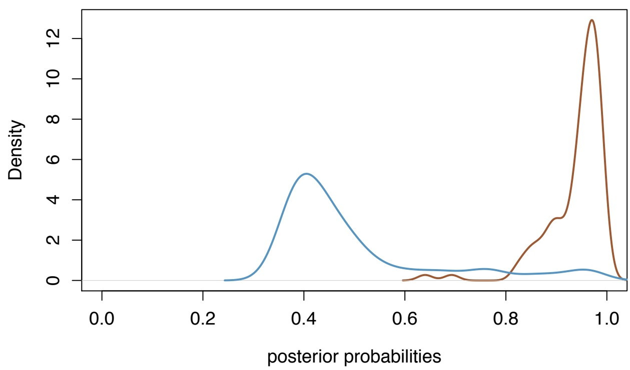

Fig. 3. Same figure as Fig. 1 when ABC is based on the 4th empirical moment as the sole summary statistic, for 250 observations simulated from either the Gaussian or the Laplace model.

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

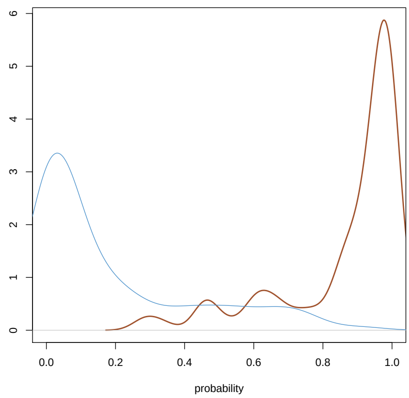

Fig. 4. Same figure as Fig. 4 when ABC is based on the 4th and 6th empirical moments as summary statistic.

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.

4. Discussion

The fact that the true asymptotic mean of the summary statistic cannot be recovered under the wrong model if model choice is 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 2 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 versus ancillary opposition found in classical statistics (Cox and Hinkley, 1994) in that it is enough that the summary statistic $\boldsymbol{T}_n$ has a different asymptotic mean under both models. As shown in the normal-Laplace example, some ancillary statistics may not be appropriate for testing.

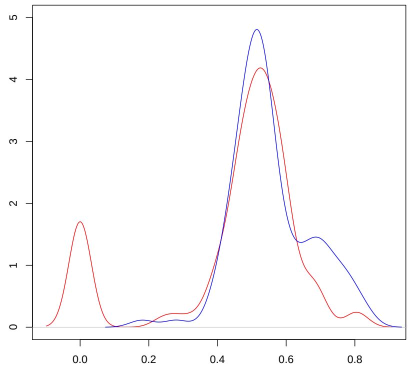

Fig. 5. Comparison of the distributions of the posterior probabilities that the data is from model $\mathfrak{M}_1$ when the data is made of 100 observations either from model $\mathfrak{M}_1$ *(brown)* or $\mathfrak{M}_2$ *(blue)* distribution when the summary statistic in the ABC algorithm is the empirical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

Acknowledgements

Part of this work was done when the second author (NSP) was visiting Dauphine and CREST and he thanks these institutions for their warm hospitality. All authors but the second author are partially 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 partially 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

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

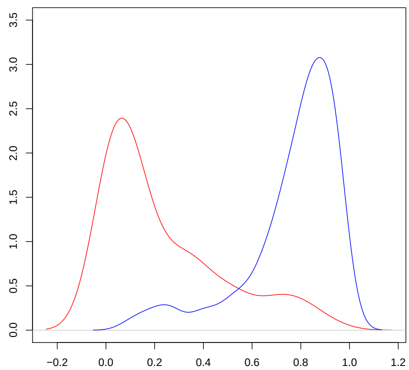

Fig. 6. Same figure as Fig. 5 when ABC is based on the empirical quantiles of order 10, 40, 60 and 90% as set of summary statistics.

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

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

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

Appendix 2

Proof of Lemma 1

Recall that G_n is the true distribution of $\boldsymbol{T}^n$. Let us first assume that $\inf\{|\mu_0-\mu_i(\theta_i)|;\theta_i\in\Theta_i\}=0$ and let $S_{n,i}$ be as defined in assumption **[A4]**. Fix constants

$\delta, U > 0$ and let M_δ be such that

$$Q\Big(\{X \in \mathbb{R}^d : |V_0^{1/2} X| > M_\delta\}\Big) = \delta \, .$$

Note that M_δ goes to infinity as δ goes to 0. Fix an M_1 satisfying (see assumption [**A5**])

$$\frac{\pi_i\Big(S_{n,i}(U) \cap \{\|V_i(\theta_i)^{-1}\| + \|V_i(\theta_i)\| > M_1\}\Big)}{\pi_i(S_{n,i}(U))} < 1/2 \, . \tag{11}$$

Set $c_\delta = \inf\{q(x); |x| \leq (M_\delta + U)M_1\}$ and define the (random) set

$$E_n = \Big\{\theta_i \in \Theta_i \, ; \; v_n |V_i(\theta_i)^{-1/2}(\boldsymbol{T}^n - \mu_i(\theta_i))| \leq (M_\delta + U)M_1\Big\} \, . \tag{12}$$

From (8) we have

$$\begin{aligned} m_i(\boldsymbol{T}^n) \quad &\geq \quad \int_{S_{n,i}(U)} \mathbb{1}_{E_n}(\theta_i)\, g_i(\boldsymbol{T}^n|\theta_i)\, \pi_i(\theta_i)\, \mathrm{d}\theta_i \\ &\geq \quad \frac{c_\delta\, v_n^d}{2} \int_{S_{n,i}(U)} |V_i(\theta_i)|^{-1/2}\, \mathbb{1}_{E_n}(\theta_i)\, \pi_i(\theta_i)\, \mathrm{d}\theta_i \, , \end{aligned} \tag{13}$$

where the last inequality follows from the fact that on the set $S_{n,i}$ (see (5) in [**A5**]) we have

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

Set

$$\tilde{S}_{n,i} = S_{n,i}(U) \cap \Big\{\|V_i(\theta_i)\| + \|V_i(\theta_i)^{-1}\| \leq M_1\Big\} \, .$$

Note that from (11) it follows that $\pi_i(\tilde{S}_{n,i}(U)) \geq \frac{1}{2}\pi_i(S_{n,i}(U))$. From (13) we deduce that,

$$m_i(\boldsymbol{T}^n) \quad \geq \quad \frac{1}{2}\, c_\delta\, v_n^d\, M_1^{-1/2}\, \pi_i\Big(\tilde{S}_{n,i} \cap E_n\Big) \, .$$

Since $M_\delta > 2U$ for δ small enough, using Markov's inequality we obtain

$$\begin{aligned} G_n\Big(\pi_i\big(\tilde{S}_{n,i} \cap E_n^c\big) \geq \frac{\pi_i(\tilde{S}_{n,i})}{2}\Big) \quad &\leq \quad 2\frac{\int_{\tilde{S}_{n,i}} G_n\big(E_n^c\big)\, \pi_i(\theta_i)\, d\theta_i}{\pi_i(\tilde{S}_{n,i})} \\ &\leq \quad 2\frac{\int_{\tilde{S}_{n,i}} G_n\big(v_n|\boldsymbol{T}^n - \mu_0| > M_\delta\big)\, \pi_i(\theta_i)\, d\theta_i}{\pi_i(\tilde{S}_{n,i})} \\ &\leq \quad 3\delta \end{aligned}$$

for n large enough. Thus we deduce that $\pi_i\big(\tilde{S}_{n,i} \cap E_n\big) \geq \frac{\pi_i(\tilde{S}_{n,i})}{2}$ with probability $(1 - 3\delta)$. Putting it all together and using (4) in [**A4**] we obtain the lower bound,

$$m_i(\boldsymbol{T}^n) \quad \gtrsim \quad c_\delta\, v_n^d\, \pi_i\big(S_{n,i}\big) \gtrsim v_n^{d-d_i} \tag{14}$$

with probability greater than $1 - 3\delta$.

We now obtain an upper bound for $m_i(\boldsymbol{T}^n)$. Using (8) we write,

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

As before fix $\delta > 0$ and let M_δ be a constant such that

$$G_n\big(|\boldsymbol{T}^n - \mu_0| > M_\delta\, v_n^{-1}\big) < 3\delta/2\,,$$

for n large enough. Note that from assumption **[A1]**

$$\sup_{|t-\mu_0|\le M_\delta v_n^{-1}} g_n(t) \lesssim v_n^d\,|V_0|^{-1/2}\,[\sup_{x\in\mathbb{R}^d} q(x) + \delta] \lesssim v_n^d. \tag{15}$$

Applying Markov's inequality together with (15) we obtain that, for all $\epsilon > 0$,

$$\begin{aligned}
&G_n\Big(\int_{\mathcal{F}_{n,i}^c} g_i(\boldsymbol{T}^n|\theta_i)\,\pi_i(\theta_i)\,d\theta_i > \epsilon\, v_n^d\,\pi_i(S_{n,i})\Big)\\
&\le G_n[|\boldsymbol{T}^n - \mu_0| > M_\delta v_n^{-1}]\\
&+ \int_{\mathcal{F}_{n,i}^c} \frac{\|V_0\|^{-1/2}\,v_n^d}{\epsilon} \int_{v_n|t-\mu_0|\le M_\delta} g_n(t) g_i(t|\theta_i) dt\,\pi_i(\theta_i)\,d\theta_i\\
&\le G_n[|\boldsymbol{T}^n - \mu_0| > M_\delta v_n^{-1}]\\
&+ \left(\sup_{|x|\le M_\delta\|V_0\|} q(x) + \delta\right) \int_{\mathcal{F}_{n,i}^c} \frac{\|V_0\|^{-1/2}\,v_n^d}{\epsilon} \int_{\mathbb{R}^d} g_i(t|\theta_i) dt\,\pi_i(\theta_i)\,d\theta_i\\
&\lesssim \delta + \frac{v_n^d}{\epsilon}\pi(\mathcal{F}_{n,i}^c) \le 2\delta,
\end{aligned} \tag{16}$$

when n is large enough.

We now express $\mathcal{F}_{n,i}$ as a finite disjoint union of the following sets:

$$\begin{aligned}
\mathcal{F}_{n,i} \quad &= \quad \bigcup_{j=0}^{J_n+1} \mathcal{H}_j, \quad J_n = J_0 v_n, \quad \text{for some } J_0 \in \mathbb{N}\,,\\
\mathcal{H}_j \quad &= \quad S_{n,i}((j+1)M_\delta) \cap S_{n,i}(jM_\delta)^c, \quad j \le J_n\,,\\
\mathcal{H}_{J_n+1} \quad &= \quad \mathcal{F}_{n,i} \cap S_{n,i}^c(M_\delta J_n)\ .
\end{aligned}$$

Now we have

$$\int_{\mathcal{F}_{n,i}} g_i(\boldsymbol{T}^n|\theta_i)\,\pi_i(\theta_i)\,d\theta_i = \sum_{j=0}^{J_n} \int_{\mathcal{H}_j} g_i(\boldsymbol{T}^n|\theta_i)\,\pi_i(\theta_i)\,d\theta_i \ . \tag{17}$$

If $j = 0$, $\mathcal{H}_0 = S_{n,i}(M_\delta)$ and if K is a constant such that $K > d_i$, using (15) we obtain

$$\begin{aligned}
G_n\left[\int_{S_{n,i}(M_\delta)} g_i(\boldsymbol{T}^n|\theta_i)\,\pi_i(\theta_i)\,d\theta_i > M_\delta^K\, v_n^{d-d_i}\right] \quad &\lesssim \quad \frac{v_n^d}{M_\delta^K v_n^{d-d_i}}\pi_i(S_{n,i}(M_\delta)) + \delta\\
&= \quad O(M_\delta^{d_i-K}) + \delta,
\end{aligned} \tag{18}$$

where the last inequality follows from (4) in **[A4]**. Since $\limsup_{\delta\to 0} M_\delta = \infty$, the bound in (18) goes to 0 as δ goes to zero. Using assumption **[A3]** and (15), we obtain that for $0 < j \le J_n$,

$$\begin{aligned}
& G_n\Big(\int_{\mathcal{H}_j} g_i(\boldsymbol{T}^n|\theta_i)\,\pi_i(\theta_i)\,d\theta_i > M_\delta^K v_n^{d-d_i}\Big) \\
& \lesssim \frac{v_n^{d_i}}{M_\delta^K}\int_{\mathcal{H}_j} G_{i,n}\left(|\boldsymbol{T}^n - \mu(\theta_i)| > (j-1/2)M_\delta v_n^{-1}|\theta_i\right)\,\pi_i(\theta_i)\,d\theta_i \\
& + G_n\left(v_n|\boldsymbol{T}^n - \mu_0| > M_\delta/2\right) \\
& \lesssim Q(\{X \in \mathbb{R}^d : |V_0^{1/2}X| > M_\delta/2\}) + M_\delta^{d_i-\alpha_i-K} j^{d_i-\alpha_i}\ ,
\end{aligned}$$

for n large enough, and similarly

$$\begin{aligned}
& G_n\Big(\int_{\mathcal{H}_{J_n+1}} g_i(\boldsymbol{T}^n|\theta_i)\,\pi_i(\theta_i)\,d\theta_i > v_n^{d-d_i}\Big) \\
& \lesssim v_n^{d_i}\int_{\mathcal{H}_{J_n+1}} G_{i,n}\left(|\boldsymbol{T}^n - \mu(\theta_i)| > J_0/2|\theta_i\right)\,\pi_i(\theta_i)\,d\theta_i \\
& + G_n\left(v_n|\boldsymbol{T}^n - \mu_0| > M_\delta\right) \\
& \lesssim 3\delta/2 + v_n^{d_i-\alpha_i} \le 2\delta,
\end{aligned} \tag{19}$$

for n large enough, under assumption **[A4]**. Combining the above inequalities with (17), we obtain for n large enough,

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

which can be made arbitrarily small by choosing δ small enough. Combining the above with (16) implies that

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

The above estimate together with the lower bound obtained in (14) proves the first claim (Equation (9)) of Lemma 1.

Now suppose $\inf\{|\mu_i(\theta_i) - \mu_0|; \theta_i \in \Theta_i\} > 0$. Then there exists $j_0 > 0$ such that $S_{n,i}(j\,v_n) = \emptyset$ for all $j \le j_0$. An identical computation as in (19), together with (16) yields

$$\begin{aligned}
& G_n\Big(\int_{\mathcal{F}_{n,i}} g_i(\boldsymbol{T}^n|\theta_i)\,\pi_i(\theta_i)\,d\theta_i > \epsilon(v_n^{d-\tau_i} + v_n^{d-\alpha_i})\Big) \\
\lesssim\ & G_n\left(v_n|\boldsymbol{T}^n - \mu_0| > M_\delta\right) + \frac{v_n^{d_i}}{\epsilon}\int_{\mathcal{F}_{n,i}} G_{i,n}\left(|\boldsymbol{T}^n - \mu_i(\theta_i)| > j_0 v_n/2\right)\pi_i(\theta_i)\,d\theta_i + 2\delta \\
\le\ & 3\delta,
\end{aligned}$$

for all n large enough and $\epsilon > 0$. This proves the second claim (Equation (10)) of Lemma 1. $\square$

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.

Didelot, X., R. Everitt, A. Johansen, and D. Lawson (2011). Likelihood-free estimation of model evidence. *Bayesian Analysis 6*(1), 1–28.

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.

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.

Rousseau, J. (2007). Approximating interval hypotheses: p-values and Bayes factors. In J. M. Bernardo, M. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics 8*. Oxford: Oxford University Press.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit (with discussion). *J. Royal Statist. Society Series B 64*(2), 583–639.

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.