Approximate integrated likelihood via ABC methods[*]

Clara Grazian and Brunero Liseo[†]

We propose a novel use of a recent new computational tool for Bayesian inference, namely the Approximate Bayesian Computation (ABC) methodology. ABC is a way to handle models for which the likelihood function may be intractable or even unavailable and/or too costly to evaluate; in particular, we consider the problem of eliminating the nuisance parameters from a complex statistical model in order to produce a likelihood function depending on the quantity of interest only. Given a proper prior for the entire vector parameter, we propose to approximate the integrated likelihood by the ratio of kernel estimators of the marginal posterior and prior for the quantity of interest. We present several examples.

1. INTRODUCTION

Given a statistical model with generic density $p(x|\boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, one is often interested in a low dimensional function $\boldsymbol{\psi}$ of the parameter vector $\boldsymbol{\theta}$, say $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta}) \in \mathbb{R}^k$, with $k < d$. Modern parametric or semi-parametric statistical theories, at least the approaches based on likelihood and Bayesian theories, aim at constructing a likelihood function which depends on $\boldsymbol{\psi}$ only. There is much literature on the problem of eliminating nuisance parameters, and we do not even try to summarize it. Interested readers may refer to Berger et al. [7] and Liseo [18] for a Bayesian perspective, and to the comprehensive books by Pace and Salvan [21] and Severini [28] or to Lancaster [17] for a more classical point of view. In a Bayesian framework the problem of eliminating the nuisance parameters is, at least in principle, trivial. Let $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\theta})$ the complementary parameter transformation, such that $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$ and let

$$(1) \qquad \pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\psi}, \boldsymbol{\lambda}) = \pi(\boldsymbol{\psi})\pi(\boldsymbol{\lambda}|\boldsymbol{\psi})$$

the prior distribution. Then, after assuming we observe a data set $\mathbf{x} = (x_1, \ldots, x_n)$ from our working model, and computed the likelihood function $L(\boldsymbol{\psi}, \boldsymbol{\lambda}; \mathbf{x}) \propto p(\mathbf{x}; \boldsymbol{\psi}, \boldsymbol{\lambda})$, the marginal posterior distribution of $\boldsymbol{\psi}$ is

$$(2) \qquad \pi(\boldsymbol{\psi}|\mathbf{x}) = \frac{\int_\Lambda \pi(\boldsymbol{\psi}, \boldsymbol{\lambda})L(\boldsymbol{\psi}, \boldsymbol{\lambda}; \mathbf{x})d\boldsymbol{\lambda}}{\int_\Psi \int_\Lambda \pi(\boldsymbol{\psi}, \boldsymbol{\lambda})L(\boldsymbol{\psi}, \boldsymbol{\lambda}; \mathbf{x})d\boldsymbol{\lambda}d\boldsymbol{\psi}}$$
$$\propto \pi(\boldsymbol{\psi}) \int_\Lambda \pi(\boldsymbol{\lambda}|\boldsymbol{\psi})L(\boldsymbol{\psi}, \boldsymbol{\lambda}; \mathbf{x})d\boldsymbol{\lambda}.$$

The integral in the right-hand side of (2) is, by definition, the integrated likelihood for the parameter of interest $\boldsymbol{\psi}$, where "integration" is meant with respect to the conditional prior distribution $\pi(\boldsymbol{\lambda}|\boldsymbol{\psi})$; it will be denoted by $\tilde{L}(\boldsymbol{\psi}; \mathbf{x})$. The use of integrated likelihoods has become popular also among non Bayesian statisticians; there are several examples in which its use is clearly superior, or at least equivalent, even from a repeated sampling perspective, in reporting the actual uncertainty associated to the estimates. See for example, Severini [25], Severini [26] and Severini [27].

However the explicit calculation of the above integral might not be so easy, especially when the dimension $d - k$ is large. Notice that the dimension d may also include a possible latent structure which, from a strictly probabilistic perspective, is not different from a parameter vector. In this paper we are interested to explore the use of approximate Bayesian computation (ABC, henceforth) methods in producing an approximate integrated likelihood function, in situations where a closed form expression of $\tilde{L}(\boldsymbol{\psi}; \mathbf{x})$ is not available, or it is too costly even to evaluate the "global" likelihood function $L(\boldsymbol{\psi}, \boldsymbol{\lambda}; \mathbf{x})$, like, for example, in many genetic applications or in the hidden (semi)-Markov literature. These are situations where MCMC methods may not be satisfactory and completely reliable.

Another class of problems where an integrated likelihood would be of primary interest is that of semi-parametric problems, where the parameter of interest is a scalar — or a vector — quantity and the nuisance parameter is represented by the nonparametric part of the model; in such cases the integration step over the Λ space would be infinite dimensional, and very often infeasible to be solved in a closed form; we will discuss this issue in §4.

Approximate Bayesian computation has now become an essential tool for the analysis of complex stochastic models when the likelihood function is unavailable. It can be

considered as a (class of) popular algorithms that achieves posterior simulation by avoiding the computation of the likelihood function (see Beaumont [6] for a recent survey). A crucial condition for the use of ABC algorithms is that it must be relatively easy to generate new pseudo-observations from the working model, for a fixed value of the parameter vector. In its simplest form, the ABC algorithm is as follows (Algorithm 1 in Marin et al. [19])

Algorithm 1 Likelihood-free Rejection algorithm

for $i = 1$ **to** N **do**
repeat
Generate θ from the prior distribution $\pi(\theta)$
Generate z from the likelihood function $f(\cdot \mid \theta)$
until $z = y$ (or some statistics η is such that $\eta(z) \approx \eta(y)$)
set $\theta_i = \theta$
end for

In this paper we will argue, through several examples of increasing complexity, how the approximate integrated likelihood produced by ABC algorithms performs when compared with the existing methods. We will also explore its use in particular examples where other methods simply fail to produce a useful and easy-to-use likelihood function for the parameter of interest. The paper is organized as follows. In the next section we describe our proposal in detail. Section 3 discusses some theoretical issues related to the precision of the ABC approximation. Section 4 compares the ABC integrated likelihood with other available approaches in a series of examples. Section 5 concludes with a final discussion of pros and cons of the method.

2. THE PROPOSED METHOD

The main goal of the paper is to obtain an approximation of the integrated likelihood $\tilde{L}(\boldsymbol{\psi}; \mathbf{x})$, for $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta}) \in \mathbb{R}^k$. From expression (2) it is easy to see that

$$(3) \qquad \tilde{L}(\boldsymbol{\psi}; \mathbf{x}) \propto \frac{\pi(\boldsymbol{\psi}|\mathbf{x})}{\pi(\boldsymbol{\psi})},$$

that is the integrated likelihood function may be interpreted as the amount of experimental evidence which transforms our prior knowledge into posterior knowledge about the parameter of interest: from this perspective, we can interpret (3) as the Bayesian definition of the integrated likelihood function.

Suppose that $\tilde{L}(\boldsymbol{\psi}; \mathbf{x})$ is hard or impossible to obtain in a closed form. For example the nuisance parameter $\boldsymbol{\lambda}$ might be infinite dimensional (see Example 4.4) or it may represent the non observable latent structure associated to the statistical model as in Hidden Markov or semi-Markov set-ups.

In these situations one can exploit the alternative expression (3) of $\tilde{L}(\boldsymbol{\psi}; \mathbf{x})$. Of course, if $\tilde{L}(\boldsymbol{\psi}; \mathbf{x})$ is not available, neither $\pi(\boldsymbol{\psi}|\mathbf{x})$ will be. However it is possible to obtain an

approximate posterior distribution $\tilde{\pi}(\boldsymbol{\psi}|\mathbf{x})$, by using some standard ABC algorithm; in §3 we will discuss some issues related to the precision of this approximation; for now, we describe the practical implementation of the method. As in any ABC approach for the estimation of the posterior distribution, one has to

- select a number of summary statistics $\eta_1(\mathbf{x}), \ldots, \eta_h(\mathbf{x})$;
- select a distance $\rho(\cdot, \cdot)$ to measure the distance between "true" and proposed data, or their summary statistics;
- select a tolerance threshold ε;
- choose a (MC)MC algorithm which proposes values for the parameter vector $\boldsymbol{\theta}$.

Once the posterior is approximated by a size M ABC posterior sample $(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \ldots \boldsymbol{\theta}_M^*)$, one can produce a non parametric kernel based density approximation of the marginal posterior distribution of $\boldsymbol{\psi}$, say $\tilde{\pi}^{ABC}(\boldsymbol{\psi}|\mathbf{x})$. A similar operation can be done with the marginal prior $\pi(\boldsymbol{\psi})$, by performing another — cheap — simulation from $\pi(\boldsymbol{\psi})$ to get another density approximation, say $\tilde{\pi}(\boldsymbol{\psi})$. Notice that one is bound to use proper priors for all the involved parameters.

Then one can define the ABC integrated likelihood

$$(4) \qquad \tilde{L}^{ABC}(\boldsymbol{\psi}; \mathbf{x}) \propto \frac{\tilde{\pi}^{ABC}(\boldsymbol{\psi}|\mathbf{x})}{\tilde{\pi}(\boldsymbol{\psi})}.$$

3. THE QUALITY OF APPROXIMATION

The gist of this note is to propose an approximate method for producing a likelihood function for a quantity of interest when the usual road of integrating with respect to the nuisance parameters cannot be followed. There are two sources of error in (4). The first type of approximate error is introduced by the ABC approximation in the numerator so the level of accuracy of (4) is of the same order of any ABC-type approximation. We believe that the main difficulty with ABC methods is the choice of summary statistics. However, while generic ABC methods have the goal of producing a "global" approximation to the posterior distribution, our particular use of the ABC approximation may suggest some alternative strategies for the choice of summary statistics. Classical statistical theory on the elimination of nuisance parameters can be in fact of some guidance in the selection of summary statistics which are partially or conditionally sufficient for the parameter of interest. Basu [5] represents an excellent reading on these topics. In particular, his Definition 5 of "Specific Sufficiency" can be used in semi-parametric set-ups, like Example 4.4 below, where the selected summary statistics are oriented towards the preservation of information about the parameter of interest. In our notation a statistic T is specific sufficient for $\boldsymbol{\psi}$ if, for each fixed value of the nuisance parameter $\boldsymbol{\lambda}$, T is sufficient for the restricted statistical model in which $\boldsymbol{\lambda}$ is held fixed and known.

Another source of error in ABC is given by the tolerance threshold ε. As stressed in Marin et al. [19], the choice of the

tolerance level is mostly a matter of computational power: smaller ε's are associated with higher computational costs and more precision. It is enough to reproduce the argument in §1.2 of Sisson and Fan [29] to see that for $\varepsilon \to 0$, the error in (4), which is due to the tolerance, vanishes.

Then, there is a balance between the fact that ε has to be small and the fact that the simulation has to be practicable. It could be useful to choose ε in a recursive way, by realizing a first simulation with a high tolerance level and then by choosing it in the left tail of the thresholds related to the accepted values. However, it is always recommended to compare different levels.

The second main source of error is due to the kernel approximation step. A second order expansion for a Gaussian kernel estimator provides that

(5)
$$\mathbb{E}\left[\tilde{\pi}^{ABC}\left(\psi|\mathbf{x}\right)\right] = \pi\left(\psi|\mathbf{x}\right) + \frac{1}{2}\frac{\partial^2}{\partial\psi^2}\pi\left(\psi|\mathbf{x}\right)h_x^2 k_2 + \mathcal{O}\left(h_x^4\right)$$

where h_x is the bandwidth and $k_2 = 1$ in the case of Gaussian kernel.

A similar approximation holds for the prior distribution. Then, using general results on a first order approximation for the ratio of functions of random variables (Kendall et al. [16], pag. 351), one has

(6)
$$\mathbb{E}\left[\frac{\tilde{\pi}^{ABC}\left(\psi|\mathbf{x}\right)}{\tilde{\pi}\left(\psi\right)}\right] = \frac{\pi\left(\psi|\mathbf{x}\right) + \frac{1}{2}\frac{\partial^2}{\partial\psi^2}\pi\left(\psi|\mathbf{x}\right)h_x^2 + \mathcal{O}\left(h_x^4\right)}{\pi\left(\psi\right) + \frac{1}{2}\frac{\partial^2}{\partial\psi^2}\pi\left(\psi\right)h_\pi^2 + \mathcal{O}\left(h_\pi^4\right)}$$

where h_x is the bandwidth chosen for the approximation of the posterior distribution and h_π is the one chosen for the approximation of the prior. The prior distribution is often known in closed form or may be easily approximated with a higher accuracy than the posterior distribution.

The previous formula ensures that our estimator will be consistent provided that a sample size dependent bandwidth h_n, converging to 0, is adopted.

It is a matter of calculation to show that the variance of the estimator is

(7)
$$\mathbb{V}\left[\frac{\tilde{\pi}^{ABC}\left(\psi|\mathbf{x}\right)}{\tilde{\pi}\left(\psi\right)}\right]$$
$$= \left[\frac{\pi\left(\psi|\mathbf{x}\right) + C_{\mathbf{x}}}{\pi\left(\psi\right) + C}\right]^2$$
$$\times \left[\frac{\frac{\pi(\psi|\mathbf{x})}{2nh_x\sqrt{\pi}} + \mathcal{O}\left(n^{-1}\right)}{\left[\pi\left(\psi|\mathbf{x}\right) + C_{\mathbf{x}}\right]^2} + \frac{\frac{\pi(\psi)}{2nh_\pi\sqrt{\pi}} + \mathcal{O}\left(n^{-1}\right)}{\left[\pi\left(\psi\right) + C\right]^2}\right]$$

where

$$C_{\mathbf{x}} = \frac{h_x^2}{2}\frac{\partial^2}{\partial\psi^2}\pi\left(\psi|\mathbf{x}\right) + \mathcal{O}\left(h_x^4\right)$$

and

$$C = \frac{h_\pi^2}{2}\frac{\partial^2}{\partial\psi^2}\pi\left(\psi\right) + \mathcal{O}\left(h_\pi^4\right).$$

Again, using a bandwidth h_n, such that $h_n \to 0$, as $n \to \infty$, one can see that the first factor of the variance is asymptotically equal to the square of the true unknown value, while the second factor vanishes like n^{-1}.

In conclusion, the ABC approximation of the integrated likelihood function mainly depends on the ABC approximation and the kernel density estimate of the posterior distribution, whereas the prior distribution may be considered known, in general. Blum [8] shows that the asymptotic variance of the kernel density estimator of the posterior distribution inversely depends on the number of simulations n and on the kernel bandwidth, while the bias is proportional to the bandwidth. The mean squared error is minimized by

(8)
$$h_n = \mathcal{O}\left(n^{-\frac{1}{d+5}}\right)$$

where d is the dimension of the summary statistics. Then the minimal MSE is

(9)
$$MSE^* = \mathcal{O}\left(n^{-\frac{4}{d+5}}\right)$$

which shows that the accuracy in the approximation decreases as the dimension of the summary statistics increases. This result may be used to define the number of simulations (and the burn-in) needed to reach the desired level of accuracy.

4. EXAMPLES

In this section we illustrate our proposal throughout several examples of increasing complexity. The first one is a toy example and it is included only to show — in a very simple situation — which are the crucial steps of the algorithm.

Example 4.1. [Poisson means]. Suppose we observe a sample of size n from $X \sim \text{Poi}(\theta_1)$ and, independently of it, another sample of size n from $Y \sim \text{Poi}(\theta_2)$. The parameter of interest is $\psi = \theta_1/\theta_2$. This is considered a benchmark example in partial likelihood literature since the conditional likelihood (see Kalbfleisch and Sprott [15]), the profile likelihood and the integrated likelihood obtained using the conditional reference prior (Berger et al. [7]) are all proportional to

$$\tilde{L}(\psi; \mathbf{x}, \mathbf{y}) \propto \frac{\psi^{n\bar{x}}}{(1+\psi)^{n(\bar{x}+\bar{y})}},$$

with the obvious meaning of the symbols above. Without loss of generality, set $\lambda = \theta_2$ as the nuisance parameter.

In this situation, the ABC approximation of the integrated likelihood is, in some sense, not comparable with the "correct" integrated likelihood because the latter is obtained through the use of an improper conditional reference prior on λ given ψ, and, as already stressed, it is not possible to use improper priors in the ABC approach. A solution may be using a prior which mimics the reference prior: we have taken $\theta_1, \theta_2 \overset{\text{iid}}{\sim} \text{Ga}(0.1, 0.1)$. Notice that, in the
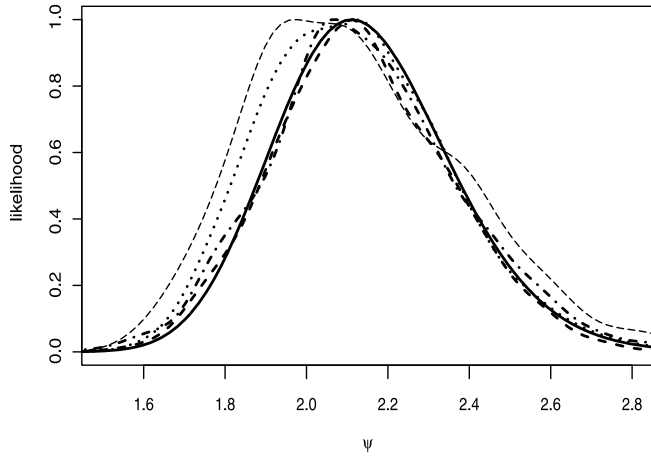
Figure 1. The integrated likelihood of ψ (solid line) and its approximations for different tolerance levels: $\varepsilon = 0.001$ (dashed), $\varepsilon = 0.01$ (dotted), $\varepsilon = 0.1$ (dotdashed) and $\varepsilon = 0.5$ (longdashed).

economy of the method, only the prior on λ, not on ψ is important. The ABC algorithm has been implemented to obtain approximations for the posterior distributions of θ_1 and θ_2. The distance ρ has been taken as the Euclidean distance, different tolerance levels have been compared — $\varepsilon = (0.001, 0.01, 0.1, 0.5)$ — and the sample means of the two samples have been taken as summary (sufficient) statistics. Samples of 1,000 simulations have been obtained to approximate the posterior distributions. The approximation to the posterior distribution of ψ is then simply obtained as the ratio between the accepted values for θ_1 and θ_2 via ABC. Given a sample from the prior distribution of ψ, the approximation of its integrated likelihood is obtained through the ratio between the kernel density estimates of both the prior and the posterior distribution.

Figure 1 shows the approximations with different choices of the tolerance level: the approximations are close together and they are all close to the integrated likelihood; the choice for the tolerance level does not seem to have a strong influence; it is mostly a matter of computational power: the acceptance rate is generally very low (often under 1%), nevertheless it grows with the tolerance level. As the threshold goes to zero, the approximation is closer to the integrated likelihood, although the computational time increases.

Simulations have been repeated for different scenarios, by changing the sample size and the number of simulations, however the results do not seem to change in a significant way. In particular, as expected, the algorithm does not depend on the (induced) prior on ψ.

Example 4.2. [Neyman and Scott's class of problems]. This is a famous class of problems, where the number of parameters increases with the sample size (Neyman and Scott [20]; Lancaster [17]). Here we consider a specific example, already

discussed in Davison [11] and Liseo [18], namely matched pairs of Bernoulli observations: every subject is assigned to treatment or control group and the randomization occurs separately within each pair, i.e. each data point in one data set is related to one and only one data point in the other data set. Let Y_{ij}'s be Bernoulli random variables, where $i = 1, \ldots, k$ represents the stratum and $j = 0, 1$ indicates the observation within the pair. The probability of success p_{ij} follows a logit model:

$$(10) \qquad \text{logit } p_{ij} = \lambda_i + \psi_j$$

For identifiability reasons, ψ_0 is set equal to 0, while $\psi_1 = \psi$ is considered constant across the k strata; ψ is the parameter of interest. To formalize the problem, assume (R_{i0}, R_{i1}) are k independent matched pairs such that, for each i:

$$(11) \qquad R_{i0} \sim \text{Be}\left(\frac{e^{\lambda_i}}{1 + e^{\lambda_i}}\right), \quad R_{i1} \sim \text{Be}\left(\frac{e^{\lambda_i + \psi}}{1 + e^{\lambda_i + \psi}}\right).$$

The complete likelihood for $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$ and ψ is

$$(12) \qquad L(\psi, \boldsymbol{\lambda}) = \frac{e^{\sum_{i=1}^{k} \lambda_i S_i + \psi T}}{\prod_{i=1}^{k}(1 + e^{\lambda_i})(1 + e^{\lambda_i + \psi})}$$

where $S_i = R_{i0} + R_{i1}$ for $i = 1, \ldots, k$ and $T = \sum_{i=1}^{k} R_{i1}$ is the number of successes among the cases. It is easy to show that the conditional maximum likelihood estimate of λ_i is infinite when $S_i = 0$ or $S_i = 2$. The classical solution to this problem is to eliminate the pairs where $S_i = 0$ or $S_i = 2$ from the analysis. Nevertheless this is certainly a loss of information, because the fact that a pair gives the same result under both treatments may suggest a "not-so-big" difference between groups.

It is easy to show that the conditional maximum likelihood estimator is $[\hat{\lambda}_{i,\psi} \mid (S_i = 1)] = -\psi/2$; also, let b be the number of pairs with $S_i = 1$. The profile likelihood of ψ is

$$(13) \qquad \tilde{L}(\psi \mid S_i = 1) = \frac{e^{\psi T}}{(1 + e^{\frac{\psi}{2}})^{2b}}$$

This likelihood function is not useful, since the maximum likelihood estimate for ψ is inconsistent (see Davison [11], Example 12.13): as b increases, $\hat{\psi} \to 2\psi$. The modified version of the profile likelihood, proposed by Barndorff-Nielsen [4] uses a multiplying factor:

$$(14) \qquad M(\psi) = \left| J_{\lambda\lambda}\left(\psi, \hat{\lambda}_\psi\right) \right|^{-\frac{1}{2}} \left| \frac{\partial \hat{\lambda}}{\partial \hat{\lambda}_\psi} \right| = \frac{e^{\frac{b\psi}{4}}}{(1 + e^{\frac{\psi}{2}})^b}$$

where $J_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)$ is the lower right corner of the observed Fisher information matrix.

The conditional distribution of T given $S_1 = S_2 = \cdots = S_b = 1$ is Binomial and depends on ψ only. That is $T \mid [S_1 =$

$S_2 = \cdots = S_b = 1, \psi] \sim \text{Bin}(b, \frac{e^\psi}{1+e^\psi})$; we can use it to get a conditional likelihood function:

$$(15) \qquad L_C(\psi) \propto \binom{b}{T} \frac{e^{\psi T}}{(1 + e^\psi)^b}$$

which leads to a consistent maximum conditional likelihood estimator.

A Bayesian approach has the advantage that it does not need to discard the pairs with $S_i = 0$ or 2. The likelihood contribution for the i-th pair is simply

$$(16) \qquad L(\psi, \lambda_i) = \frac{e^{\lambda_i S_i + \psi R_{i1}}}{(1 + e^{\lambda_i})(1 + e^{\lambda_i + \psi})}.$$

With a change of parametrization $\omega_i = e^{\lambda_i}/(1 + e^{\lambda_i})$ and using a (proper) Jeffreys' prior for $\omega_i|\psi$ (namely a $Beta(\frac{1}{2}, \frac{1}{2})$), the integrated likelihood is

$$(17) \qquad L_i(\psi) = e^{\psi R_{i1}} \int_0^1 \frac{\omega_i^{S_i - \frac{1}{2}}(1 - \omega_i)^{\frac{3}{2} - S_i}}{1 - \omega_i(1 - e^\psi)} d\omega_i$$

where the integral is one of the possible representation of the Hypergeometric or Gauss series, as shown in Abramowitz and Stegun [1] (formula 15.3.1, pag. 558). Therefore, the integrated likelihood is proportional to

$$(18) \qquad L_i(\psi) \propto {}_2F_1\left(1, S_i + \frac{1}{2}, 3, 1 - e^\psi\right) e^{\psi R_{i1}}$$

Define $L_{jl}(\psi)$ as the integrated likelihood function associated with the i-th pair for which $(R_{i0}, R_{i1}) = (j, l)$ and n_{jl} the number of pairs for which $(R_{i0}, R_{i1}) = (j, l)$, then the integrated likelihood function for ψ is

$$(19) \qquad L_{int}(\psi) \propto \prod_{j,l=0,1} L_{jl}(\psi)^{n_{jl}}.$$

It is worthwhile to notice that this likelihood is not, in some sense, comparable with profile and conditional likelihoods, because it also considers the pairs discarded by non-Bayesian methods.

The ABC approach has been used with simulated data, with a sample size n equal to 30. Simulations were performed by setting $\psi = 1$, a value which is quite frequent in applications, when similar treatments are compared. The values of $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ have been generated by setting $\xi_i = \lambda_i/(1 + \lambda_i)$ and drawing the ξ's from a $U(0, 1)$ distribution. Again, we have used the Euclidean distance between summary statistics and different tolerance levels $\varepsilon = (0.001, 0.01, 0.1, 0.5)$. The summary statistics are the sample means for $\mathbf{R}_0$ and $\mathbf{R}_1$. We have also assumed a normal prior for ψ with zero mean and standard deviation equal to 10. The proposed values for λ_i's have been generated from a $Beta(\frac{1}{2}, \frac{1}{2})$ distribution for the above defined transformations ξ_i's.
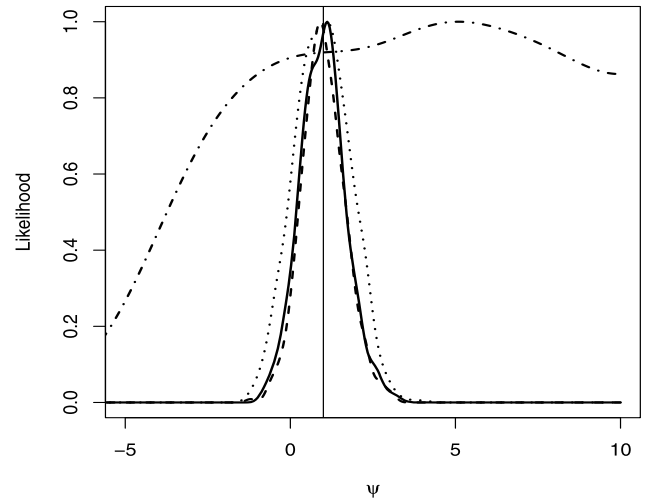

Figure 2. ABC approximations of the integrated likelihood for ψ with different tolerance levels: $\varepsilon = 0.001$ (solid line), $\varepsilon = 0.01$ (dashed), $\varepsilon = 0.1$ (dotted), $\varepsilon = 0.5$ (dotdash).

With a sample from the posterior distribution of ψ for each tolerance level and a sample from its prior distribution, we have obtained an approximation of the likelihood of ψ via density kernel estimation. The results are shown in Figure 2: the approximations are quite good for tolerance levels below 0.1; on the other hand, when the threshold grows to 0.5 the approximate likelihood function is very flat and multi-modal, i.e. too many proposed values, even very different from the true value of ψ, are misleadingly accepted; for example, a value of ψ around 43 has been accepted in one of our simulations.

Once again, a fair comparison between Bayesian and non-Bayesian approaches is not strictly possible, nevertheless the various proposals are shown in Figure 3: all the proposed solutions are concentrated relatively close to the true value, although the profile likelihood seems to be biased towards large values of ψ: this behaviour is also present, although to a minor extent in the modified profile and the integrated likelihood solutions. The ABC approximation closely mimics the integrated likelihood, obtained via a saddle-point approximation of the Hypergeometric series Butler and Wood [10].

Similar conclusions are valid for different choices of the prior distribution, different sample sizes, and different numbers of simulations. Just like in Example 4.1, the acceptance rates are typically very low (always under 1% for tolerance levels under 0.01 and about 5% for a tolerance level of 0.1). Acceptance rates dramatically increase to about 60%, for $\varepsilon = 0.5$; however in these cases, approximations get much worse.

Example 4.3 [Likelihood function for the quantiles of a g-and-k distribution]. Quantile distributions are, in general,
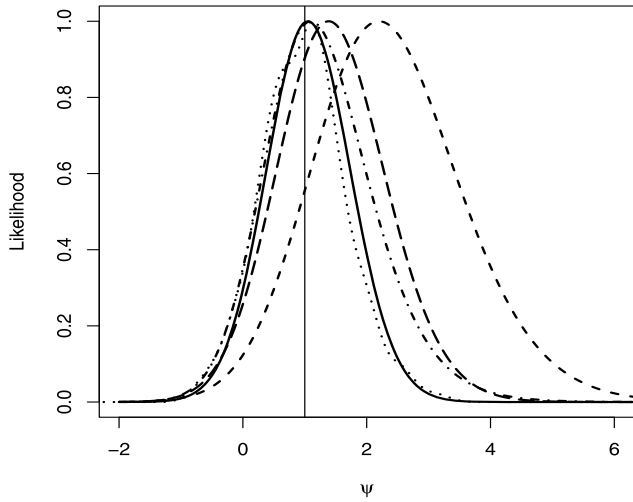
Figure 3. Likelihood functions for ψ based on different solutions ($n = 30$): the profile likelihood (dashed line), the conditional likelihood (dotdash line), the modified profile likelihood (with the correction of Barndorff-Nielsen, longdashed line), the integrated likelihood (solid line, drawn by using the Laplace approximation of ${}_2F_1$ by Butler and Wood [10]) and the ABC approximation ($n_{sim} = 1,000$, prior $\psi \sim N(1, 10)$ and tolerance level $\varepsilon = 0.001$, dotted line).

Figure 4. Empirical cumulative distribution function (above) and histogram (below) of the simulated data from a g-and-k distribution (with $A = 3$, $B = 1$, $g = 2$, $k = 0.5$).

defined by the inverse of their cumulative distribution function. They are characterized by a great flexibility of shapes obtained by varying parameters values. They may easily model kurtotic or skewed data with the great advantage that they typically have a small number of parameters, unlike mixture models which are usually adopted to describe this kind of data. An advantage of quantile distributions is that it is extremely easy to simulate from them by means of a simple inversion. However, there is no free lunches, and the above advantages are paid with the fact that their probability density functions (and therefore, the implied likelihood functions) are often not available in a closed form expression.

One of the most interesting examples of quantile class of distributions is the so-called g-and-k distribution, described in Haynes et al. [13], whose quantile function Q is given by

(20)

$$Q(u; A, B, g, k) = A + B\left[1 + c\frac{1 - \exp\{-g\,z(u)\}}{1 + \exp\{-g\,z(u)\}}\right] \cdot \left\{1 + z(u)^2\right\}^k z(u)$$

where $z(u)$ is the u-th quantile of the standard normal distribution; parameters A, B, g and k represent location, scale, skewness and kurtosis respectively; c is an additional parameter which measures the overall asymmetry and it is generally fixed at 0.8, following Rayner and MacGillivray [22]. The class of Normal distributions is a proper subset
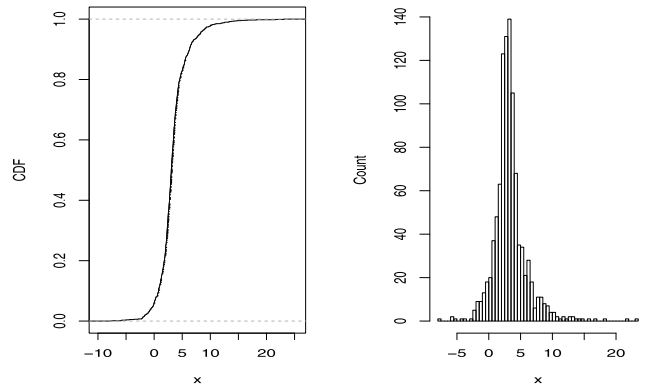
of this class; it is obtained by setting $g = k = 0$. Suppose we are interested in one or more quantiles using this model. There are no easy solution (S) to the problem of constructing a partial likelihood for these quantiles. The fact that the likelihood function is not available makes any classical approach practically impossible to implement. Rayner and MacGillivray [23] propose a numerical maximum likelihood approach; however they also explain that very large sample sizes are necessary to obtain reliable estimates of the parameters. On the other hand, even though the quantile distributions have no explicit likelihood, simulation from these models is easy, and an approximate Bayesian computation approach, also for producing an integrated likelihood of the parameters of interest, seems reasonable.

For this specific problem, two types of ABC algorithms have been compared: the former is the usual ABC algorithm based on simulations from the prior distributions (with 10^3 iterations); the latter is an ABC-MCMC algorithm (10^6 iterations, with a burn-in of 10^5 simulations). Two versions of ABC-MCMC have been used, the former described in Marin et al. [19] (see Algorithm 2) and the latter described in Allingham et al. [2] (see Algorithm 3). The main difference between these two versions of ABC-MCMC algorithm is that, in the first case, there is no rejection step; at each iteration a value is accepted (either the new proposed value or the value accepted in the previous iteration); in the second case, instead, it is possible to discard the current value and to propose a new one, so the chain always "moves".

Data have been simulated from a g-and-k distribution with parameters $A = 3$, $B = 1$, $g = 2$ and $k = 0.5$. As previously said, c is considered known and set equal to 0.8. The sample size, has been set equal to $n = 1,000$. The empirical cumulative distribution function and the histogram of the simulated data are shown in Figure 4.

The transition kernel of the ABC-MCMC algorithm needs to be chosen having in mind two conflicting objectives: on one hand, full exploration of the parameter space,

Algorithm 2 Likelihood-free MCMC sampling	**Algorithm 3** Likelihood-free MCMC sampling
Initialization	Initialization

Algorithm 2 Likelihood-free MCMC sampling

Initialization

A1) Generate θ' from the prior distribution $\pi(\cdot)$

A2) Generate a data set $\mathbf{z}' \sim f(\mathbf{z} \mid \theta')$, where f is the model of the data

A3) If $\rho\{\eta(\mathbf{y}), \eta(\mathbf{z}')\} \leq \varepsilon$, set $(\theta^{(0)}, \mathbf{z}^{(0)}) = (\theta', \mathbf{z}')$, otherwise return to A1)

MCMC-step

for $t = 1, \ldots, T$ {

1) Generate θ^{prop} from the Markov kernel $q(\cdot \mid \theta^{(t-1)})$

2) Generate $\mathbf{z}'$ from the model $f(\cdot \mid \theta^{prop})$

3) Calculate $h(\theta^{(t-1)}, \theta^{prop}) = \min(1, \frac{\pi(\theta^{prop}) \, q(\theta^{(t-1)}|\theta^{prop})}{\pi(\theta^{(t-1)}) \, q(\theta^{prop}|\theta^{(t-1)})})$

4) **if** $\rho\{\eta(\mathbf{y}), \eta(\mathbf{z}')\} \leq \varepsilon$, set $(\theta^{(t)}, \mathbf{z}^{(t)}) = (\theta^{prop}, \mathbf{z}')$ with probability h,

else $(\theta^{(t)}, \mathbf{z}^{(t)}) = (\theta^{(t-1)}, \mathbf{z}^{(t-1)})$

}

Algorithm 3 Likelihood-free MCMC sampling

Initialization

A1) Generate θ' from the prior distribution $\pi(\cdot)$

A2) Generate a data set $\mathbf{z}' \sim f(\mathbf{z} \mid \theta')$, where f is the model of the data

A3) If $\rho\{\eta(\mathbf{y}), \eta(\mathbf{z}')\} \leq \varepsilon$, set $(\theta^{(0)}, \mathbf{z}^{(0)}) = (\theta', \mathbf{z}')$, otherwise return to A1)

MCMC-step

for $t = 1, \ldots, T$ {

1) Generate θ^{prop} from the Markov kernel $q(\cdot \mid \theta^{(t-1)})$

2) Generate $\mathbf{z}'$ from the model $f(\cdot \mid \theta^{prop})$

3) Calculate $h(\theta^{(t-1)}, \theta^{prop}) = \min(1, \frac{\pi(\theta^{prop}) \, q(\theta^{(t-1)}|\theta^{prop})}{\pi(\theta^{(t-1)}) \, q(\theta^{prop}|\theta^{(t-1)})})$

4) **if** $\rho\{\eta(\mathbf{y}), \eta(\mathbf{z}')\} \leq \varepsilon$, set $(\theta^{(t)}, \mathbf{z}^{(t)}) = (\theta^{prop}, \mathbf{z}')$ with probability h,

else return to 1)

}

and, on the other hand, a reasonably high acceptance rate, which increases for proposals mostly concentrated where the posterior mass is present. As described in Allingham et al. [2] uniform priors with bounds (0, 10) have been chosen for each parameter and a random walk-normal kernel with variance 0.1 has been used together with a large number of iterations (10^6) so that the parameter space is likely to be fully investigated. The vector of summary statistics consists of the sample mean, the standard deviation, and the sample skewness and kurtosis indexes. The Euclidean distance has been used to compare summary statistics.

The tolerance level ε has been chosen in a recursive way: first, a very large value has been selected, and a histogram of all the distances has been drawn. A reasonable value has been taken from the 5% left tail of this histogram. Then, the chosen threshold has been compared with smaller values. In particular, a threshold equal to 3 corresponds to 3.9% left tail. This has been compared with tolerance levels equal to 2 and 0.5.

The analysis of the approximate posterior distributions shows that three out of four parameters (A, B and k) are well identified, while the posterior distribution of g is rather flat. In general, as the tolerance level decreases, results improve and posterior distributions tend to be more concentrated. Nevertheless, even using the lowest tolerance level the posterior distribution of g does not seem to concentrate around any value. This suggests that the algorithm needs an even smaller value of the threshold. A simulation with tolerance level equal to 0.25 has been then performed using Algorithm 2: the approximation of the posterior distribution of g is still not centered around its true value, even if there is a mode around it; nevertheless the problem with this so low tolerance level and this type of algorithm is that the acceptance rate of new proposed values is very low and the chain does not move too much. This tolerance level is also so low to make the application of the other algorithms prohibitive in terms of computational time.
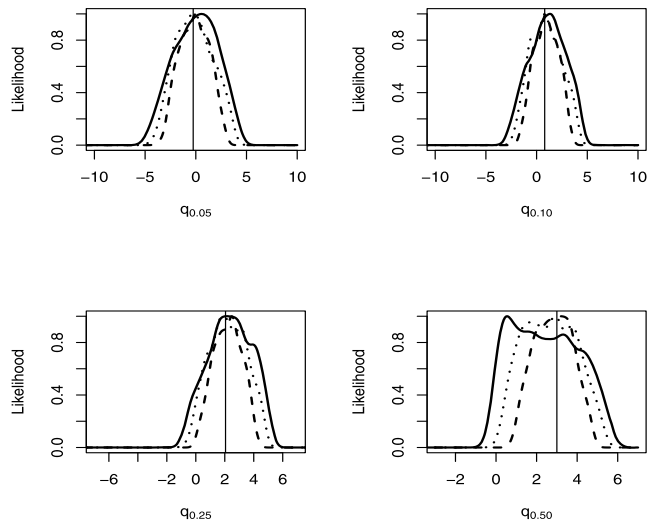

Figure 5. Likelihood approximations of the quantiles of a g-and-k distribution parameters for simulated data, obtained with an ABC algorithm which simulates proposal values from the prior distributions ($\mathrm{U}(0, 10)$ *for each parameter): tolerance levels equal to* 3 *(solid line),* 2 *(dashed line) and* 0.5 *(dotted line).*

Our main goal of the analysis was to find an approximation of the integrated likelihood function for a given quantile: in particular, we have considered the percentiles of order 0.05, 0.10, 0.25 and 0.50. Notice that, in the g-and-k distribution model, the median is always equal to A.

The results are shown in Figure 5, 6 and 7. The performance is in general very good: the approximations are always concentrated around the true values.

The ABC algorithm with simulations from the prior distribution has some apparent problems of multi-modality, which are however absent using Algorithm 2. However, in this case, the obtained approximations are not very smooth,
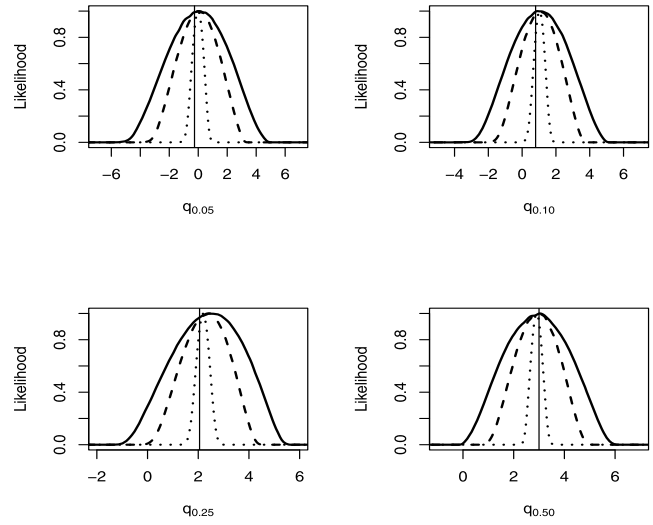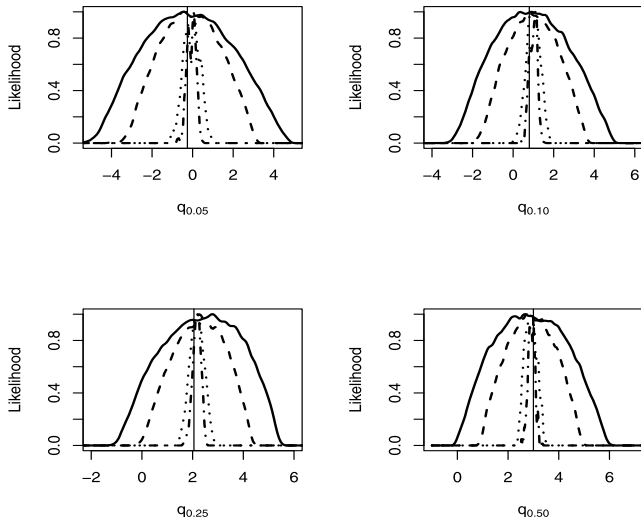
Figure 6. Likelihood approximations of the quantiles of a g-and-k distribution parameters for simulated data, obtained with an ABC-MCMC Algorithm 2 with Gaussian kernel: tolerance levels equal to 3 (solid line), 2 (dashed line) and 0.5 (dotted line) and 0.25 (dotdashed line).

Figure 7. Likelihood approximations of the quantiles of a g-and-k distribution parameters for simulated data, obtained with an ABC-MCMC Algorithm 3 with Gaussian kernel: tolerance levels equal to 3 (solid line), 2 (dashed line) and 0.5 (dotted line).

and they show more irregularities as the tolerance level decreases: as we have already remarked, a too low threshold leads to very low acceptance rates and this means that the chains do not move too much.

In this example, Algorithm 3 has the best overall performance: the approximations are smooth and all concentrated around the true quantile values. As the tolerance level decreases, the likelihood approximations are more concentrated; obviously the computational time gets larger.

The acceptance rates of these algorithms are in general very low:

- the basic ABC algorithm has an acceptance rate of 0.138% when the threshold is equal to 3, and it goes down to 0.041% and 0.007% with tolerance levels of 2 and 0.5 respectively;
- the ABC-MCMC Algorithm 2 needs, respectively, 187, 1487, about $500K$ and more than 3 millions of simulations for the initialization step for the different tolerance levels 3, 2, 0.5 and 0.25. The acceptance rates of the proposed values is also very low: 18.41%, 9.90%, 0.47% and 0.046% respectively; it is clear that the acceptance rates relative to the smaller thresholds cannot lead to smooth approximations;
- the ABC-MCMC Algorithm 3 needs 1104, 4383 and about $400K$ simulations for the initialization step for tolerance levels 3, 2 and 0.5 respectively; in this case every accepted value is a "new" value, and this solves the problems in Algorithm 2.

In conclusion, ABC-MCMC seems to perform better, although the versions we have implemented present some cons:

the algorithm in Marin et al. [19] is faster but it must be calibrated in terms of the tolerance level, which has to be low in order to achieve good approximations, and the MCMC acceptance rate, which has to be sufficiently high in order to allow the chains to move.

Example 4.4 [Semiparametric regression]. Consider the following model

$$(21) \qquad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \gamma(\mathbf{z}) + \varepsilon,$$

where $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ is a vector of n real-valued variables, and $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$, $\gamma(\mathbf{z}) = (\gamma(z_1), \gamma(z_2), \ldots, \gamma(z_n))'$ and $\mathbf{z} = (z_1, z_2, \ldots, z_n)'$ are observed constants respectively taking values in $\mathbb{R}^p$ and $\mathcal{Z}$, ε is the usual random component that we assume having multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Omega}_\phi$ which depends on some parameters ϕ, $\boldsymbol{\beta}$ is a vector of unknown parameters taking values in $\mathbb{R}^p$ and $\gamma : \mathcal{Z} \to \mathbb{R}$ is an unknown function.

If the analysis is focused on $\boldsymbol{\beta}$ or $\boldsymbol{\Omega}_\phi$, γ may be considered a nuisance parameter and a method to remove it from the analysis is needed. In particular, if a weight function for γ based on a zero-mean Gaussian stochastic process with covariance function $K_\lambda(\cdot, \cdot)$ with parameter λ is used, the vector $(\gamma(z_1), \ldots, \gamma(z_n))$ has a multivariate Normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_\lambda$ and the integrated likelihood function of $\boldsymbol{\beta}$ is

$$(22)$$

$$|\boldsymbol{\Omega}_\phi + \boldsymbol{\Sigma}_\lambda|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)'(\boldsymbol{\Omega}_\phi + \boldsymbol{\Sigma}_\lambda)^{-1}(\mathbf{Y} - \mathbf{X}\beta) \right\}$$

where $\boldsymbol{\Sigma_\lambda}$ is the $n \times n$ matrix with $K_\lambda(z_i, z_j)$ in the (i, j) element. This form may be obtained because of the assumption on the Normal distribution of the errors and the use of a Gaussian process weight function for γ; more general cases are not so straightforward to handle outside the normal set-up.

In He and Severini [14] the Authors show that, for a given choice of $K_\lambda(\cdot, \cdot)$, when the dispersion parameter, say $\boldsymbol{\eta} = (\phi, \lambda)$ is known, $\boldsymbol{\beta}$ can be estimated by the generalized least-squares estimator: $\hat{\boldsymbol{\beta}} = \mathbf{X}^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}$ where $V = \boldsymbol{\Omega}_\phi + \boldsymbol{\Sigma}_\lambda$; if the dispersion parameter is unknown, β can be estimated as a function of an estimator of $\boldsymbol{\eta}$, $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\eta}})$.

The method has been used with data from a survey of the fauna on the sea bed lying between the Queensland coast and the Great Barrier Reef; the response variable analysed is a score, on a log weight scale, which combines information across the captured species; this score value is considered dependent on the latitude $\mathbf{x}$ in a linear way and on the longitude $\mathbf{z}$ in an unknown way; see Bowman and Azzalini [9] for more details. The model is

$$(23) \qquad Y_j = \beta_0 + x_j\beta_1 + \gamma(z_j) + \varepsilon_j, \quad j = 1, \ldots, n$$

where $\varepsilon_1, \ldots, \varepsilon_j$ are independent normal errors with mean 0 and constant variance σ_ε^2. Using the integrated likelihood approach, a Gaussian covariance function

$$(24) \qquad K(z, \tilde{z}) = \tau^2 \exp\left(-\frac{1}{2}\frac{|z - \tilde{z}|^2}{\alpha}\right)$$

and a restricted maximum likelihood estimate (REML, Harville [12]) for the nuisance parameters, the estimates of β_1 is 1.020, with a standard error of 0.356 (see He and Severini [14]).

We have used our ABC approximation in order to find an integrated likelihood for $\boldsymbol{\beta}$. It is then necessary to define proper prior distributions for all the parameters of the model, i.e. $\boldsymbol{\beta}$, σ_ε^2 and the parameter of the covariance function of the Gaussian process, α and τ^2.

For $\boldsymbol{\beta}$ a g-prior has been chosen such that $\boldsymbol{\beta} \sim N_2(\mathbf{0}, g\sigma_\varepsilon^2(\mathbf{X^TX})^{-1})$, where $g \sim U(0, 2n)$ and $\sigma_\varepsilon^2 \sim IG(a, b)$ with a, and b suitably small (as an approximation of the Jeffreys' prior). A Gaussian process with squared exponential covariance function has been used as prior process for the function $\gamma(\cdot)$. The hyper-parameters of the Gaussian process have the following prior distributions: $\tau^2 \sim IG(a, b)$, with $a = b = 0.01$ and $\alpha \sim IG(2, \nu)$ with $\nu = \rho_0/(-2\log(0.05))$ and $\rho_0 = \max_{i,j=1\ldots n} |z_i - z_j|$; see Schmidt and Gelfand [24] and Banerjee et al. [3] for more details.

The choice of the summary statistics is not straightforward, because it is necessary to find statistics that take into account both the parametric and the nonparametric parts of the model, nevertheless sufficiency is not guaranteed. A function of $\mathbf{z}$ has been considered and the maximum likelihood estimates of the coefficients of the new model have
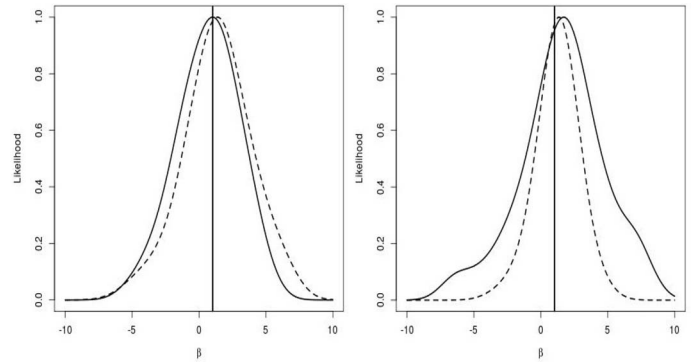

Figure 8. ABC approximation of the integrated likelihood of β_1 in the semiparametric model. The approximations are obtained by using a Uniform prior (left) and an approximation of the Jeffreys' prior (right) for σ_ε^2. Two different choices of summary statistics are compared: the maximum likelihood estimates of the model, with linear (solid lines) and quadratic (dashed lines) approximations of $\gamma(\cdot)$.

been used as summary statistics. In particular, two choices of function has been considered: $g(z_j) = z_j$ and $h(z_j) = z_j^2$ for $j = 1, \ldots, n$. An analysis of the maximum likelihood estimates has shown that the estimate of the constant β_0 is particularly unstable, therefore only the estimates for the predictor variables' coefficients contribute to the approximation as summary statistics.

In the MCMC step, normal transitional kernels have been used for all the parameters of the model, centered at the values accepted on the previous step and with small variance.

The results are shown in Figure 8: the ABC approximation with 10^6 simulations are concentrated around the estimates obtained by maximizing the integrated likelihood of the model. In this case, the ABC approach may be seen as a way to properly account for the uncertainty on the nuisance parameters that is not considered when REML estimates are used. Figure 8 compares different choices of summary statistics and prior distributions for the variance σ_ε^2: on the left a $U(0, 10)$ is used and on the right a proper approximation of the Jeffreys' prior is used ($Ga(a, b)$ with a, b small). All the approximations are smooth and concentrated around the maximum likelihood estimate. Moreover, Figure 8 shows that using the summary statistics based on a quadratic approximation of $\gamma(\cdot)$ leads to better results, because they are all smooth. On the other hand the approximations obtained by considering a linear model with respect to $\mathbf{z}$ present slight multimodality problems.

The number of simulations for the initialization step depends on the choice of the tolerance level: the approximation of the likelihood of β_1 by using a Uniform prior for σ_ε^2 needs 368, 2,053 and 10,945 simulations to accept the first value for tolerance levels of 1, 0.5 and 0.25 respectively; the approximation with Gamma prior with small parameters for σ_ε^2 needs 40, 34 and 81 simulations to accept the first value.

These results refer to the summary statistics obtained with the quadratic approximation of $\gamma(\cdot)$, the other choice of summary statistics considered has shown similar values.

The acceptance rates of ABC-MCMC algorithm are in general low, in particular with the lowest tolerance levels; they are around 25% for the highest thresholds considered.

5. DISCUSSION

We have explored the use of ABC methodology, relatively new computational tools for Bayesian inference in complex models, in a rather classical inferential problem, namely the elimination of nuisance parameters. We stress the fact that there are many situations where it is practically impossible to obtain a likelihood function for the parameter of interest in a closed form: in those cases the proposed method can be a competitive alternative to numerical methods.

As a technical aside one should note that, in many situations the prior $\pi(\boldsymbol{\psi})$ might be available in a closed form, so the kernel approximation of the prior is not necessary, and the accuracy of our method is even better. However, we have preferred to present the method in its generality.

Another issue related to this last point is the approximation of the marginal posterior density of some components of the parameter. Also in this case, the problem is made simpler by the fact that no approximation is needed for the prior and standard asymptotic arguments for kernel estimators hold for the approximation obtained from the posterior sample. The main drawback of the present approach is that it requires the use of proper prior densities. This can be a problem, especially when the nuisance parameter is high-dimensional and the elicitation process would be difficult. A practical solution in these case is to adopt proper priors which approximate the appropriate improper noninformative prior for that model.

ACKNOWLEDGEMENTS

REFERENCES

[1] ABRAMOWITZ, M. and STEGUN, I. A., *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. U.S. Government Printing Office, Washington, D.C., 1964. MR0167642

[2] ALLINGHAM, D., KING, R. A. R., and MENGERSEN, K. L., Bayesian estimation of quantile distributions. *Statistics and Computing*, 19(2):189–201, 2009. MR2486231

[3] BANERJEE, S., GELFAND, A. E., and CARLIN, B. P., *Hierarchical Modelling and Analysis for Spatial Data*. Chapman and Hall CRC, 2004.

[4] BARNDORFF-NIELSEN, O., On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70(2):343–365, 1983. MR0712023

[5] BASU, D., On the elimination of nuisance parameters. *Journal of the American Statistical Association*, 72(358):355–366, 1977. MR0451477

[6] BEAUMONT, M., Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41:379–406, 2010.

[7] BERGER, J. O., LISEO, B., and WOLPERT, R. L., Integrating likelihood methods for eliminating nuisance parameters. *Statistical Science*, 14(1):1–28, 1999. MR1702200

[8] BLUM, M. G. B., Approximate Bayesian Computation: A nonparametric perspective. *Journal of the American Statistical Association*, 105(491):1178–1187, 2010. MR2752613

[9] BOWMAN, A. W. and AZZALINI, A., *Applied Smoothing Techniques for Data Analysis*. Oxford Univ. Press., 1997.

[10] BUTLER, R. W. and WOOD, A. T. A., Laplace approximations for hypergeometric functions with matrix argument. *The Annals of Statistics*, 30(4):1155–1177, 2002. MR1926172

[11] DAVISON, A. C., *Statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2003. MR1998913

[12] HARVILLE, D. A., Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977. MR0451550

[13] HAYNES, M. A., MACGILLIVRAY, H. L., and MENGERSEN, K. L., Robustness of ranking and selection rules using generalised g-and-k distributions. *Journal of Statistical Planning and Inference*, 65(1):45–66, 1997. MR1619666

[14] HE, H. and SEVERINI, T. A., Integrated likelihood estimation in semiparametric regression models. *Technical report, Northwestern University Department of Statistics*, 2013.

[15] KALBFLEISCH, J. D. and SPROTT, D. A., Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society*, **B**, 32(2):175–208, 1970. MR0270474

[16] KENDALL, M., STUART, A., and ORD, J. K., *Kendall's advanced theory of statistics. Vol. 1.* The Clarendon Press Oxford University Press, New York, fifth edition, 1987. ISBN 0-19-520561-8. Distribution theory. MR0902361

[17] LANCASTER, T., The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2):391–413, 2000. ISSN 0304-4076. MR1752336

[18] LISEO, B., The elimination of nuisance parameters. In *Bayesian thinking: modeling and computation*, volume 25 of *Handbook of Statistics*, pages 193–219. Elsevier, North-Holland, Amsterdam, 2005. MR2490527

[19] MARIN, J. M., PUDLO, P., ROBERT, C. P., and RYDER, R. J., Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012. MR2992292

[20] NEYMAN, J. and SCOTT, E. L., Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32, 1948. MR0025113

[21] PACE, L. and SALVAN, A., *Principles of statistical inference. From a neo-Fisherian perspective.* World Scientific Publishing Co. Inc., River Edge, NJ, 1997. MR1476674

[22] RAYNER, G. D. and MACGILLIVRAY, H. L., Weighted quantile-based estimation for a class of transformation distributions. *Computational Statistics and Data Analysis*, 39(4):401–433, 2002. MR1910020

[23] RAYNER, G. D. and MACGILLIVRAY, H. L., Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing*, 12(1):57–75, 2002. MR1877580

[24] SCHMIDT, A. M. and GELFAND, A., A Bayesian Coregionalization Model for Multivariate Pollutant Data. *Journal of Geophysics Research*, 108(D24), 2003.

[25] SEVERINI, T. A., Integrated likelihood functions for non-Bayesian inference. *Biometrika*, 94(3):529–542, 2007. ISSN 0006-3444. MR2410006

[26] SEVERINI, T. A., Likelihood ratio statistics based on an integrated likelihood. *Biometrika*, 97(2):481–496, 2010. ISSN 0006-3444. MR2650752

[27] SEVERINI, T. A., Frequency properties of inferences based on an integrated likelihood function. *Statistica Sinica*, 21(1):433–447, 2011. ISSN 1017-0405. MR2796870

[28] SEVERINI, T. A., *Likelihood methods in statistics*. Oxford University Press, Oxford, 2000. MR1854870

[29] SISSON, S. A. and FAN, Y., Likelihood-free MCMC. In *Handbook of Markov chain Monte Carlo*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 313–335. CRC Press, 2011. MR2858454

Clara Grazian
Dipartimento di Scienze Statistiche
Sapienza Università di Roma
Piazzale Aldo Moro 5
00185, Roma
Italy
E-mail address: clara.grazian@uniroma1.it

Université Paris-Dauphine
Paris
France

CREST
Paris
France

Brunero Liseo
MEMOTEF
Sapienza Università di Roma
Via del Castro Laurenziano 9
00161, Roma
Italy
E-mail address: brunero.liseo@uniroma1.it