

Approximate Bayesian Computation with proper scoring rules

Erlis Ruli, Nicola Sartori and Laura Ventura

Abstract We illustrate a novel approach for developing robust posterior distributions using Approximate Bayesian Computation (ABC) methods with proper scoring rules. This is formally motivated by the use of unbiased estimating functions as automatic informative summary statistics in ABC. Examples with the Tsallis score are illustrated, and comparisons with robust M -estimating functions are considered. **Abstract** *In questo lavoro si propone un approccio per ottenere distribuzioni a posteriori robuste attraverso i metodi ABC con le scoring rules proprie. L'approccio è formalmente motivato dall'uso di funzioni di stima non distorte come statistiche nelle procedure ABC. Il metodo proposto è illustrato attraverso due esempi con la funzione score di Tsallis e, ove possibile, si considerano anche confronti con le equazioni di stima di tipo M .*

Key words: Likelihood-free inference, M -estimators, robustness, summary statistic, unbiased estimating function.

1 Introduction

The summary of the data offered by the likelihood function is the basis of Bayesian methods. However, Bayesian likelihood-based inference cannot be performed when the likelihood is analytically or computationally intractable. Moreover, the stability of Bayesian likelihood-based procedures generally requires strict adherence to the model assumptions, and it is well-known that they may behave quite poorly under slight violations of model assumptions (see e.g. [5]).

When the likelihood function is computationally intractable, ABC methods, also known as likelihood-free techniques, allow us to make inference about the parameters of the posterior distribution (see e.g. [7]). ABC methods rely on a function which measures the distance between some empirical summary statistics and their simulation based counterparts. The most crucial point of ABC algorithms is the

Dept. of Statistics, University of Padova, Italy, e-mail: ventura@stat.unipd.it

selection of these summary statistics, particularly regarding their sufficiency properties. In this respect, in [8] it is shown that composite likelihood score functions can be fruitfully used as automatic informative summary statistics in ABC in order to obtain accurate approximations to the posterior distribution. Composite likelihoods are suitable surrogates of the full likelihood when the latter is intractable, and their use has been widely advocated in different complex applications ([10]). The use of composite likelihood score functions as summary statistics is formally motivated by the use of the score function of the full likelihood, and it may be extended to general unbiased estimating functions.

The aim of this paper is to discuss ABC with proper scoring rules (see e.g. [4], and references therein). Full likelihood and composite likelihood estimation are special cases of the general estimation technique based on proper scoring rules. Such rules supply unbiased estimating equations for any statistical model, which can be chosen to increase robustness or for ease of computation. The appeal of scoring rule estimation lies in the potential adaptation of the scoring rule to the problem at hand, and it forms a special case of M -estimation (see e.g. [6]). This contribution focuses on proper scoring rules with robustness properties in order to obtain robust posterior distributions, and the comparison with classical robust M -procedures is also considered.

In Section 2 some background on proper scoring rules and ABC with composite score functions is given. The proposed ABC algorithm, based on proper scoring rules, is presented in Section 3, together with two illustrative numerical examples. Concluding remarks are given in Section 4.

2 Statistical methods

2.1 Proper scoring rules

Let X be a random variable, taking values in a sample space $\mathcal{X}$. A *scoring rule* (see e.g. [3]) is a loss function $S(x, Q)$ measuring the quality of a quoted probability distribution Q for X , in the light of the realized outcome x of X . It is *proper* if, for any distribution P for X , the expected score $S(P, Q) := E_{X \sim P} S(X, Q)$ is minimized by quoting $Q = P$. There is a wide variety of proper scoring rules ([4]). A prominent example is the *log score* $S(x, Q) = -\log q(x)$, with $q(\cdot)$ the density (or the probability mass function) of X , which is simply the negative log likelihood. Another useful example is the *Tsallis score* ([9]), given by

$$S(x, Q) = (\gamma - 1) \int q(y)^\gamma d\mu(y) - \gamma q(x)^{\gamma-1}, \quad (1)$$

with $\gamma > 1$, also called the *density power score* in [1]. Setting $\gamma = 2$ in (1) yields the *quadratic* or *Brier score* ([2]). Other examples of some special proper scoring rules are given in [4].

In estimation problems, proper scoring rules provide attractive loss and utility functions that can be tailored to a scientific problem. Suppose that we wish to fit a parametric statistical model $F_\theta = F(x; \theta)$ based on the sample $(x_1, \dots, x_n)$. To estimate θ , we might consider the goodness-of-fit by the total empirical score $S_n(\theta) = \sum_{i=1}^n S(x_i, F_\theta)$. Then, asymptotic arguments indicate that $\hat{\theta}_S = \arg \max_{\theta} S_n(\theta) \rightarrow \theta_0$ as $n \rightarrow \infty$, where θ_0 is the true parameter value. This suggests a general approach to estimation: choose a proper scoring rule that is tailored to the problem at hand and use $\hat{\theta}_S$ as the *optimal score estimator* based on the scoring rule. Maximum likelihood estimation, as well as composite likelihood estimation, are special cases of optimal score estimation when $S_n(\theta)$ is the corresponding log likelihood, and optimal score estimation forms a special case of M -estimation ([6]). Indeed, consider

$$s(x, \theta) = \frac{\partial S(x, F_\theta)}{\partial \theta}. \quad (2)$$

Then, θ can be estimated by $\hat{\theta}_S$, the root of the estimating equation $\sum_{i=1}^n s(x_i, \theta) = 0$. It can be shown that, for any differentiable scoring rule and any smooth statistical model, $E_\theta(s(X, \theta)) = 0$, i.e. $s(x, \theta)$ is an unbiased estimating function ([4]). In particular, it will typically deliver a consistent and asymptotically normal estimator for θ , with mean θ and variance $V(\theta) = K(\theta)^{-1}J(\theta)K(\theta)^{-1}$, with $J(\theta) = E_\theta(s(X, \theta)s(X, \theta)^T)$ and $K(\theta) = E_\theta(\partial s(X, \theta)/\partial \theta^T)$. The form of $V(\theta)$ is due to the failure of the information identity since, in general $K(\theta) \neq J(\theta)$.

From the general theory of M -estimators, the influence function of the estimator $\hat{\theta}_S$ is given by

$$\text{IF}_S(x; \theta) = K(\theta)^{-1}s(x, \theta),$$

and it measures the effect on the estimator of an infinitesimal contamination at the point x , standardized by the mass of the contamination. The estimator $\hat{\theta}_S$ is B-robust if and only if $s(x, \theta)$ is bounded in x . Robustness properties of the Tsallis score are discussed in [1].

2.2 ABC with unbiased estimating equations

Let $\pi(\theta)$ be a prior distribution for θ , $L(\theta)$ the likelihood function based on data $x = (x_1, \dots, x_n)$ and $\pi(\theta|x) \propto \pi(\theta)L(\theta)$ the posterior distribution. Suppose that $L(\theta)$ is unavailable, but that it is possible to simulate from the full model F_θ . The accept-reject ABC algorithm works as follows:

1. generate θ^* from the prior $\pi(\theta)$
2. generate x^* from F_{θ^*}
3. if $\rho(t(x), t(x^*)) \leq \varepsilon$ accept θ^* , otherwise reject

where $t(\cdot)$ is a suitable summary statistic, $\rho(\cdot)$ is a given distance, and ε is a fixed tolerance level. One of the crucial points of the ABC algorithm is the choice of $t(\cdot)$. In this respect, in [8] it is suggested to consider the composite score function,

evaluated at the maximum composite likelihood estimate, as the summary statistic for the ABC procedure. This is formally motivated by the use of the score function of the full likelihood, and it may be extended to general unbiased estimating functions.

3 ABC with proper scoring rules

Starting from a proper scoring rule $S_n(\theta)$ and extending results in [8], we propose $\sum_{i=1}^n s(x_i, \theta)$ as a summary statistic in ABC. In terms of the accept-reject ABC algorithm in Section 2.2, we propose to replace step 3. with

$$\rho \left(\sum_{i=1}^n s(x_i, \hat{\theta}_S), \sum_{i=1}^n s(x_i^*, \hat{\theta}_S) \right) \leq \varepsilon .$$

The evaluation of $\sum s(x_i, \theta)$ in $\hat{\theta}_S$ is computationally convenient since $\sum s(x_i, \hat{\theta}_S) = 0$ and we only need to evaluate $\sum s(x_i^*, \hat{\theta}_S)$.

If $s(x, \theta)$ is bounded in x , i.e. if the estimator $\hat{\theta}_S$ is B-robust, then the approximation to the posterior obtained with this algorithm is resistant with respect to slight violations of the model assumptions.

Example 1: equi-correlated normal model Let F_θ be a q -variate normal distribution $N_q(\mu, \Sigma)$, with μ vector of means and Σ covariance matrix, with the main diagonal equal to σ^2 and the off-diagonal entries equal to $\rho\sigma^2$.

Let us focus on the correlation coefficient ρ and let us assume $\mu = 0$ and $\sigma^2 = 1$. As an illustration we consider a sample of size $n = 50$ from F_θ , with $\theta = \rho = 0.5$ and $q = 20$. With an uniform prior for ρ , the full posterior $\pi(\rho|x)$ based on the full likelihood is compared with the ABC posterior based on the Tsallis score, for $\gamma = 1.1, 1.25, 1.5, 1.95$, also with contaminated data. The contamination is obtained by adding the vector $c\mathbf{1}_n$, to one of the q dimensions of the observed data, where $\mathbf{1}_n$ is the $(n \times 1)$ unit vector and c is a positive constant (fixed to 0, 7.5, 15, 22.5 and 30).

The comparisons are shown in Figure 1. On the left side, we notice that, when there is no contamination, e.g. $c = 0$, and γ increases, the ABC posterior based on Tsallis' score flattens out, and is slightly shifted with respect to the full posterior. On the right side, instead, we notice that contaminating the dataset by c leaves the shape of the Tsallis score essentially unchanged. On the other hand, the full score tends to shift to the right, as the value of c increases. Hence, while both scores are not B-robust (increasing the range of ρ reveals an S-like shaped function for Tsallis' score), the Tsallis score is nevertheless more resistant to outlying observations.

Example 2: Normal model Let F_θ be the normal distribution $N(\mu, \sigma^2)$, with mean μ and variance σ^2 . Let $\theta = (\mu, \sigma)$. As an example let us consider a random sample of size $n = 50$ from $N(0, 1)$, and let us assume that μ is a priori $N(0, 9)$ and that σ is uniformly distributed in $(0, 13)$ and independent from μ .

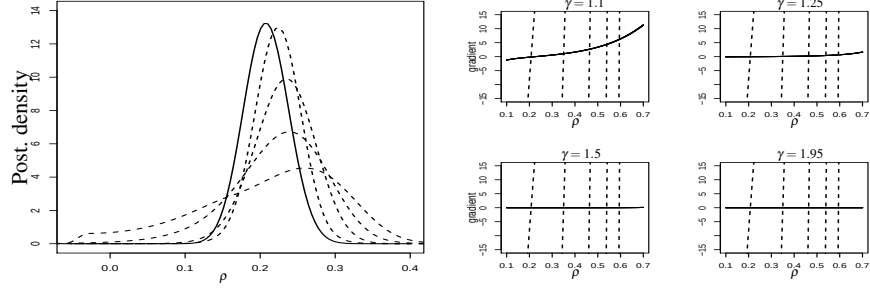


Fig. 1 Equi-correlated normal model. Left: ABC posteriors (dashed) based on the Tsallis score with several values of γ (increasing values of γ correspond to flatter densities) compared with the full posterior (solid). Right: full (dashed lines) and Tsallis (solid) scores for different γ and c values (for each quadrant, the dashed lines corresponds, from left to right, to c equal to 0, 7.5, 15, 22.5 and 30).

Figure 2 illustrates the posterior based on the full likelihood, the ABC posterior based on the Huber's location and scale estimating equations and the ABC posterior based on the Tsallis scoring rule with $\gamma = 1.25$. We notice that the marginal poste-

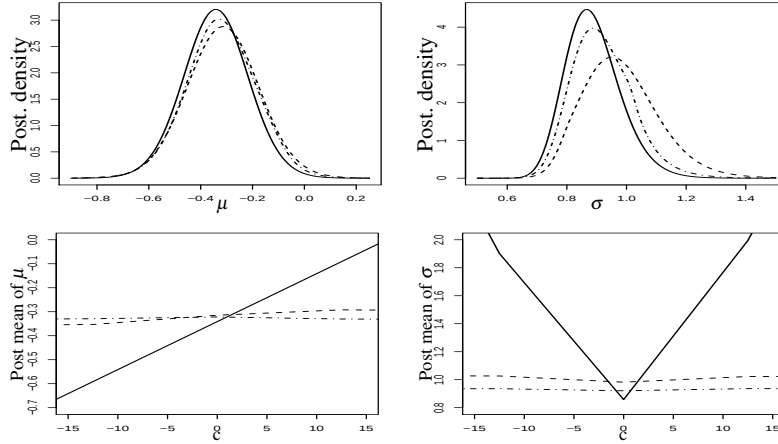


Fig. 2 $N(\mu, \sigma)$ model. Top: marginal ABC posteriors for μ (left) and σ (right) with Tsallis' score (dot-dashed) and ABC posterior with Huber's score (dashed) compared with the full posterior (solid). Bottom: sensitivity analyses of posterior means under contaminated data.

riors of σ are quite different and this is reasonable as the three posteriors are based on different functions of the observed data x . Indeed the ABC posteriors are based on suitable functions of the data which are guaranteed to deliver robust parameter estimates, whereas the full posterior is based on the sufficient summary statistic.

We consider again datasets with several contaminations $c = (\pm 50, \pm 37.5, \pm 25.0, \pm 12.5)$ and for each of the contaminated datasets we compute the expectations of the full posterior and the two ABC posteriors based on Huber's and Tsallis'

score, respectively. From the sensitivity analyses of the posterior means plotted on the bottom part of Figure 2, we notice that, while the full posterior mean is quite sensitive to outlying observations, the means of the two ABC posteriors are robust.

4 Final remarks

The Bayesian literature on robust posteriors, with respect to data, is not well developed. A typical approach to robustness is by means of robust estimating functions which, under suitable regularity conditions, can deliver robust point estimates. However, the use of these robust unbiased estimating functions in Bayesian inference is generally possible only through the use of quasi- and the empirical likelihoods in the Bayes formula (see e.g. [5]). However, quasi-likelihood is available only for scalar parameters of interest while empirical likelihood can be unstable for small sample sizes.

This article contributes to Bayesian robustness theory by proposing an approximation method that is based on unbiased estimating function. The method delivers a posterior distribution which is guaranteed to be robust, when the underlying scoring rule or estimating function is robust.

Acknowledgements This work was supported by a grant from the University of Padua (Progetti di Ricerca di Ateneo 2011) and by the Cariparo Foundation Excellence-grant 2011/2012.

References

1. Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C.: Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85**, 549–559 (1998)
2. Brier, G.W.: Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3 (1950)
3. Dawid, A.P.: Probability forecasting. In: *Encycl. Stat. Sci.* (ed. Kotz, Johnson and Read), 210–218 (1986)
4. Dawid, P., Musio, M.: Theory and applications of proper scoring rules. *Metron*, to appear (2014)
5. Greco, L., Racugno, W., Ventura L.: Robust likelihood functions in Bayesian inference. *J. Stat. Plann. Inf.* **138**, 1258–1270 (2008)
6. Huber, P.J., Ronchetti, E.M.: *Robust Statistics*. Wiley (2009)
7. Marin, J.M., Pudlo, P., Robert, C.P., Ryder, R.J.: Approximate Bayesian computational methods. *Stat. Comput.* **22**, 1167–1180 (2012)
8. Ruli, E., Sartori, N., Ventura, L.: Approximate Bayesian Computation with composite score functions. *arXiv:1311.7286* (2014)
9. Tsallis, C.: Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **52**, 479–487 (1988)
10. Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. *Stat. Sin.* **21**, 5–42 (2011)